Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1019–1024
Marseille, 11–16 May 2020

Chinese Discourse Parsing: Model and Evaluation

Chuan-An Lin[1], Shyh-Shiun Hung[1], Hen-Hsen Huang[2], Hsin-Hsi Chen[1]

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[2]Department of Computer Science, National Chengchi University, Taipei, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
calin@nlg.csie.ntu.edu.tw, shhung@nlg.csie.ntu.edu.tw, hhhuang@nccu.edu.tw, hhchen@ntu.edu.tw

Abstract

Chinese discourse parsing, which aims to identify the hierarchical relationships of Chinese elementary discourse units, has not yet a consistent evaluation metric. Although Parseval is commonly used, variations of evaluation differ from three aspects: micro vs. macro F1 scores, binary vs. multiway ground truth, and left-heavy vs. right-heavy binarization. In this paper, we first propose a neural network model that unifies a pre-trained transformer and CKY-like algorithm, and then compare it with the previous models with different evaluation scenarios. The experimental results show that our model outperforms the previous systems. We conclude that (1) the pre-trained context embedding provides effective solutions to deal with implicit semantics in Chinese texts, and (2) using multiway ground truth is helpful since different binarization approaches lead to significant differences in performance.

Keywords: Discourse parsing, CKY, Shift-reduce, PARSEVAL

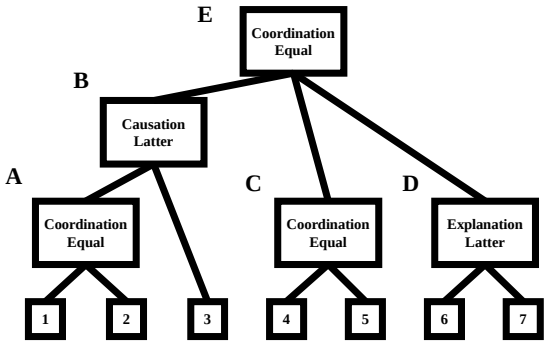
1. Introduction

As mentioned by Rhetorical Structure Theory (Mann and Thompson, 1988), a discourse is composed of elementary discourse units (EDUs), which can be formed into a hierarchical structure by relating each other with discourse relations. This kind of discourse parsing tree provides a deep understanding of an article and benefits downstream NLP tasks.

The majority of Chinese discourse relation is "implicit", lacking obvious lexical cues to discriminate the relation type (Li et al., 2014b). It makes the task of Chinese discourse parsing more challenging as the parser has to catch the implicit meaning from the text instead of relying on lexical information.

According to the Connective-Driven Dependency Tree (CDT) scheme (Li et al., 2014b), there are four subtasks in Chinese discourse parsing, including EDU segmentation, tree structure construction, relation sense labeling, and relation center labeling. Figure 1 illustrates an example of Chinese discourse parsing tree. Note that a coordination relation may have more than two arguments while the other kinds of relations are always binary. Details of Chinese discourse parsing can be found in (Li et al., 2014b). The CDT annotates the hierarchical discourse structure of a given article, which is different from the PDTB-style scheme proposed by (Zhou and Xue, 2012).

Although the annotation of the Chinese Discourse TreeBank (CDTB) is well-defined, the evaluation is divergent. Generally, PARSEVAL (Black et al., 1991) is used to evaluate the quality of a predicted parsing tree. For a golden standard discourse tree, we have a set of non-leaf nodes $\boldsymbol{N} = \{n_1, n_2, ..., n_k\}$. We also have a set of text spans $\boldsymbol{T} = \{t_1, t_2, ..., t_k\}$ where n_i dominates t_i for all i (e.g., the node B in Figure 1 dominates the text spanning from EDU 1 to EDU 3, so if we use n_j to represent node B, then t_j should represent the text span from EDU 1 to 3). Similarly, for a predicted discourse parsing tree, we have non-leaf nodes $\boldsymbol{N'} = \{n_1^{'}, n_2^{'}, ..., n_h^{'}\}$ and text spans



1. 仅去年中国银行就累计向外商投资企业提供了六百九十多亿元的人民币贷款， Last year alone, the Bank of China provided more than 6.9 billion yuan in RMB loans to foreign-invested enterprises.
2. 另外还向外商投资企业发放外汇现汇贷款四十多亿美元， In addition, it also issued foreign exchange loans of more than US$4 billion to foreign-invested enterprises.
3. 这些贷款重点支持基础原材料、化工、机械等行业。 These loans focus on supporting basic raw materials, chemicals, machinery and other industries.
4. 据统计，到去年底，中国银行向外商投资企业累计发放的外汇现汇贷款和人民币贷款达到二百一十亿美元和二千五百九十三亿元，According to statistics, by the end of last year, the Bank of China's foreign exchange loans and RMB loans reached RMB 21 billion and RMB 259.3 billion.
5. 人民币贷款余额已近四百五十四亿元。The balance of RMB loans has reached nearly 4,540 million yuan.
6. 目前，约有十五万家外商投资企业在中国银行开立帐户， At present, about 150,000 foreign-invested companies open accounts with Bank of China.
7. 其中二万多家获得中国银行的贷款支持。More than 20,000 of them have received loans from the Bank of China.

Figure 1: An example of Chinese discourse parsing tree from the Chinese Discourse Treebank (Li et al., 2014b)

$\boldsymbol{T}^{'} = \{t_1^{'}, t_2^{'}, ..., t_h^{'}\}$. Assuming that we do not consider the relation label of each node, the recall, precision and F1 score can thus be calculated following the PARSEVAL criteria.

$$Recall = \frac{|\{\hat{t} \mid \hat{t} \in T, \hat{t} \in T^{'}\}|}{|T'|} \quad (1)$$

$$Precision = \frac{|\{\hat{t} \mid \hat{t} \in T, \hat{t} \in T^{'}\}|}{|T|} \quad (2)$$

$$F1 = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision} \quad (3)$$

Based on this metric, there are the following evaluation scenarios:

Micro vs. macro F1 scores: a micro F1 is computed by taking each node across all test examples into account at once, while a macro F1 is simply the averaged F1 of all predicted parsing trees.

Binary vs. multiway ground truth: Since both transition-based and chart-based models often construct binary discourse parsing trees, we have to consider whether to use the original multiway golden standard tree or a binarized version for comparison. While only Coordination relation may have more than two arguments with equal weight, multi-nucleus relations may exist, e.g., about 9% in the experimental corpus of this paper. The choice of ground truth type may thus significantly affect the fairness of evaluation.

right-heavy vs. left-heavy binarization: We need to choose the way of binarization even early in the preprocessing stage due to the same limitation for the models to learn and predict binary structures. When evaluating, we either binarize the multiway golden standard tree for comparison or try to reverse the predicted tree to a multiway tree. Therefore, the choice of binarization affects not only evaluation but also model training. Figure 2 illustrates the two ways of binarization. In the left-heavy version, the children of a multiway node $\boldsymbol{M}$ are re-merged from left to right. The left two children are merged to form a new binary node recursively until $\boldsymbol{M}$ also becomes binary. Right-heavy binarization is the reversed version of the left-heavy one where the children are re-merged from right to left.

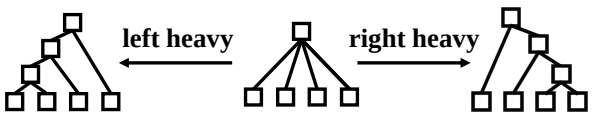


Figure 2: left-heavy and right-heavy binarization.

In this work, we develop a neural network model with pre-trained context embedding to learn implicit semantics in Chinese discourse. Further, we directly compare our model to prior works with divergent evaluation scenarios. Our experiments lead to two main contributions:

1. Our model successfully utilizes the pretrained transformer to reach state-of-the-art performance.
2. We give suggestions for future researches based on our analysis of different scenarios.

2. Related work

Several benchmark datasets have been used to develop discourse parsers for English and Chinese respectively. For English, the most commonly used one is the Rhetorical Structure Theory Discourse Treebank (RST-DT) (Carlson et al., 2001). RST-DT follows the Rhetorical Structure Theory (RST) and is annotated from 385 Wall Street Journal articles. For Chinese, the Chinese discourse treebank (CDTB) (Li et al., 2014b) is a hierarchically annotated corpus. We will use this corpus to conduct our experiments. CDTB follows the CDT scheme, where 500 Xinhua newswire documents selected from Chinese Treebank are annotated.

Although many works have been done on RST-DT (Yu et al., 2018) (Heilman and Sagae, 2015) (Ji and Eisenstein, 2014) (Li et al., 2016) (Li et al., 2014a) (Joty et al., 2013), related researches focusing on Chinese are relatively fewer. Sun and Kong (2018) propose a transition based neural network model to construct a discourse parsing tree based on the given EDUs and their POS features. For the end-to-end system development, Kong and Zhou (2017) build a pipeline framework, with each stage utilizing sparse features. They use a greedy bottom-up approach to construct a parsing tree. Lin et al. (2018) propose a unified framework based on recursive neural network to jointly parse the EDUs and the discourse structure with a probabilistic CKY-like algorithm. All the three proposed models construct binary parsing trees and thus require either a de-binarization step or binarizing ground truth for comparison.

Morey et al. (2017) note that there is a discrepancy in evaluation among different works on RST-DT even though these works are also based on PARSEVAL. They thus reproduce these methods to make direct comparisons. For CDTB, there are two main branches of evaluation scenarios. Kong and Zhou (2017) and Lin et al. (2018) adopt micro F1 score, multiway gold parsing tree, and left-heavy binarization. In contrast, Sun and Kong (2018) use macro F1, binary gold tree, and right-heavy binarization.

Recently, Devlin et al. (2018) introduce a neural language representation model called Bidirectional Encoder Representations from Transformers (BERT). It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context. BERT has been proved to perform prominently well on many NLP tasks such as language understanding, question answering, and commonsense inference (Devlin et al., 2018). The pre-trained model is also suitable for tasks that have to understand the in-depth meaning of language but with training data of small scale like RST-DT or CDTB.

3. Model Description

This work is improved from our previous work (Lin et al., 2018), and the resulting framework is shown in Figure 3. We modify the original RvNN-based CKY-like construction process to be a new CKY phase (the cycle in the middle of Figure 3). Our core motivation is to utilize the pretrained neural transformers in the discourse tree construction process while keeping the new model still compatible with the original training procedure. Besides, we test different versions of binarization and de-binarization for comparisons.

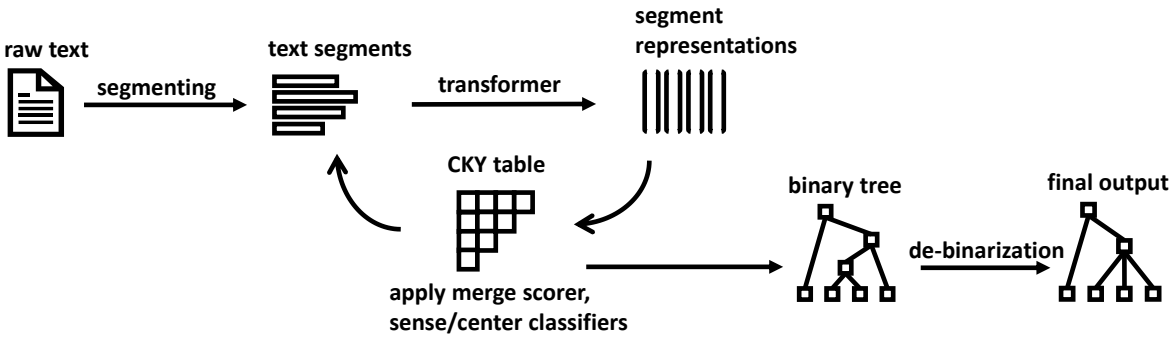


Figure 3: The framework of our model.

We discuss these two parts of our model in the following subsections.

3.1. CKY Phase

After segmenting by punctuation from a given paragraph, we have a series of text segments $\boldsymbol{S} = \{s_1, s_2, ..., s_n\}$. Let $sp_{i,j}$ denote the text span ranging from s_i to s_j. In each iteration of the CKY phase, given (i, j, k), $1 \leq i \leq j < k \leq n$, we apply the transformer to calculate the representation of text span $sp_{i,j}$ and $sp_{j+1,k}$:

$$\begin{aligned} r_{i,j} &= \boldsymbol{Transformer}(sp_{i,j}) \\ r_{j+1,k} &= \boldsymbol{Transformer}(sp_{j+1,k}) \end{aligned} \tag{4}$$

Let T_L denote the candidate tree derived from the text span $sp_{i,j}$, T_R denote the candidate tree derived from $sp_{j+1,k}$, and T_M denote the tree derived from merging T_L and T_R. In other words, we relate the root nodes of T_L and T_R with a discourse relation to form a new tree T_M. T_M is thus a candidate tree of $sp_{i,k}$, and T_L and T_R are the left child and the right child of the root of T_M, respectively. We can calculate the probability of this candidate tree based on conditional probability:

$$\begin{aligned} P(T_M) &= P(T_M|T_L, T_R) \cdot P(T_L, T_R) \\ &= P(T_M|T_L, T_R) \cdot P(T_L) \cdot P(T_R) \end{aligned} \tag{5}$$

where we assume T_L and T_R are independent since they correspond to disjoint text spans. $T_L, T_R, P(T_L)$, and $P(T_R)$ are stored in the CKY table for dynamic programming, so we just need to calculate the probability of merging T_L and T_R with our merge scorer:

$$P(T_M|T_L, T_R) = \boldsymbol{MergeScorer}(r_{i,j}, r_{j+1,k}) \tag{6}$$

In this way, our model can perform CKY-like dynamic programming on (i, j, k) to find the candidate tree with the highest probability. The sense classifier and the center classifier are used to label discourse relations as well as EDU as in (Lin et al., 2018). We use BERT (Devlin et al., 2018) as the transformer, which is suitable for fine-tuning with rather uncomplicated neural networks. Note that the CKY phase is designed to simplify the original RvNN framework that calculates discourse representations recursively according to the tree structure. Therefore, BERT, along with its pre-trained linguistic knowledge, can learn the underlying discourse structure itself with raw text segments as inputs. We apply the implementation of (Wolf et al., 2019), which contains a Chinese version of pre-trained BERT.

3.2. Binarization and De-binarization

We adopt different binarization and de-binarization approaches for preprocessing before training and evaluation. We first need to perform binarization before generating training instances. After the CKY phase, we need either to de-binarize the output binary parsing tree or to binarize the multiway golden standard tree for comparison.

Similar to binarization, de-binarization can be left-heavy or right-heavy. For left-heavy de-binarization, we traverse the binary tree from the root node. If we find a node N labeled with discourse relation of coordination and with its arguments equally weighted, we check its left child L to see whether the discourse relation of L is labeled in the same way. If so, we merge N and L to be a new multiway node. We do it recursively until we cannot find such L. The right-heavy approach is the reverse version of the left-heavy one. The choice of left-heavy/right-heavy should be consistent between training and evaluation.

4. Experiments

4.1. Settings

We conduct our experiments in CDTB and split the data into a training set and a test set following the policy of (Lin et al., 2018). We use cross-entropy loss function with 0.01 weighted L2 regularization for training. The learning rate is set to be 10^{-6}, and the batch size is set to be 10.

4.2. Results and Discussions

The performances of our model compared to the previous works on two main evaluation scenarios are shown in Table 1 and Table 2, respectively. While Table 1 stands for the evaluation scenario of micro F1 score, left-heavy binarization, and multiway ground truth, Table 2 corresponds to macro F1 score, right-heavy binarization, and binary ground truth. We denote the two variations of our models as Ours-L and Ours-R to represent the choice of left-heavy or right-heavy binarization while training.

From both results, we can see that our model outperforms the previous works significantly. For comparing further the

scores with gold EDUs or under an end-to-end parsing scenario, we can know that the improvements mainly come from better structure prediction of discourse parsing trees. These results show the effectiveness of the pretrained context and its ability to learn underlying discourse structures without explicit cues.

To fairly compare the binarization policy, we evaluate the performances of Our-L and Our-R under gold multiway ground truth. We can see from Table 3 that Ours-R outperforms Ours-L on almost all parts with micro or macro F1 scores. The gaps in macro scores are especially more significant than those on micro scores.

Model	EDU	Span	+S	+C	Join
Zhou	gold	52.3	33.8	23.9	23.2
Lin		64.6	42.7	38.5	35.0
Ours-L		**76.5**	**50.8**	**48.5**	**43.1**
Zhou	**93.8**	46.4	28.8	23.1	22.0
Lin	87.2	49.5	32.6	28.8	26.8
Ours-L	92.4	**68.9**	**43.3**	**42.0**	**37.0**

Table 1: Performance with micro F1 score, left-heavy binarization, and multiway ground truth. The upper rows show the results where gold EDUs are given. **Span** is the F1 score of structure prediction. **+S** is the F1 score of both structure and relation senses are predicted correctly. **+C** measures the F1 score that both structure and relation centers are predicted correctly. **Join** corresponds to predictions that are correct for both structure, senses, and centers.

Model	EDU	Span	+S	+C	Join
Sun	gold	84.0		53.9	
Ours-R		**87.2**	**61.4**	**60.1**	**55.0**
Sun	93.0	78.2		53.2	
Ours-R	**93.5**	**81.3**	**56.9**	**54.6**	**50.0**

Table 2: Performance with macro F1 score, right-heavy binarization, and binarized ground truth.

Model	mi/ma	EDU	Span	+S	+C	Join
Ours-L	micro	gold	**76.5**	50.8	48.5	43.1
	macro		**83.9**	54.8	52.4	45.5
Ours-R	micro	gold	75.1	**51.5**	**49.5**	**44.8**
	macro		82.8	**57.3**	**56.0**	**50.7**
Ours-L	micro	92.4	68.9	43.3	42.0	37.0
	macro		76.6	48.3	46.4	40.6
Ours-R	micro	**93.5**	**69.3**	**45.9**	**43.1**	**38.6**
	macro		**77.8**	**53.3**	**50.9**	**46.1**

Table 3: Comparison between models trained with left-heavy and right-heavy binarization policies.

It is known that under micro F1 evaluation, different paragraphs are weighted in proportion to the number of their nodes in the discourse trees, while each paragraph is equally weighted under macro evaluation. Therefore, we can infer that Our-R takes advantage of predicting local structures. This strength leads to higher performances in some paragraphs with small discourse trees. This explanation is supported by a better EDU score 93.5 of Ours-R compared to the 92.4 of Ours-L since EDUs, which are constructed from merging proper segments in the CKY-like process, can also be seen as local structures. Both Ours-L and Ours-R grasp the general structure of discourses, so the performance gap under macro evaluation is rather smaller.

We further analyze the distribution of relation senses predicted by both models in Figure 4 and Figure 5. We find that Ours-R is less biases to the majority of relation sense, which is Coordination. This tendency occurs even before the de-binarization process. We can see from Figure 6 that Ours-R's judgments lead to entirely higher performances on all relation senses.

Overall, our experiments show that the right-heavy binarization policy makes the model learn to parse more effectively. Since different binarization choices fundamentally affect how models learn the knowledge about discourse structures, we suggest that multiway ground truth is more suitable for evaluation in order to allow future researches to explore different polices of binarization as well as de-binarization.

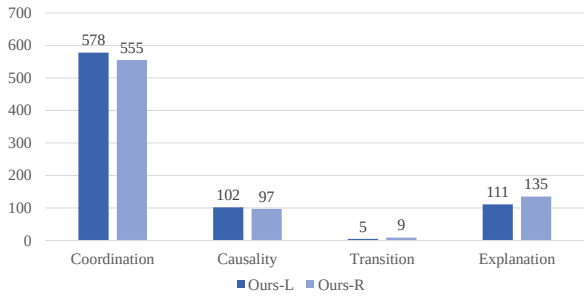


Figure 4: Distribution of predicted relations before de-binarization.

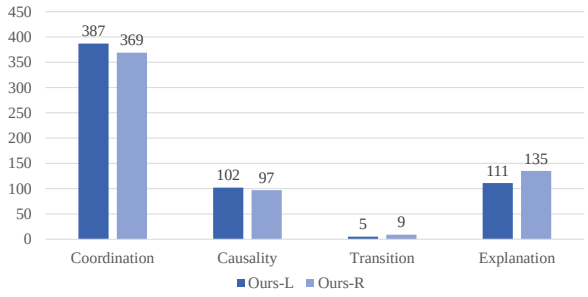


Figure 5: Distribution of predicted relations after de-binarization.

5. Conclusion

In this work, we unify a pretrained neural context embedding and CKY-like construction algorithm to reach state-of-the-art performance on Chinese discourse parsing. Further, we point out that current evaluation scenarios on CDTB are

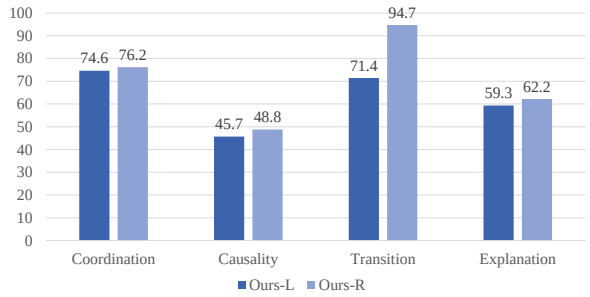


Figure 6: Micro F1 scores of predicted relations after debinarization.

still divergent. By experimenting across different scenarios, we find that the choice of binarization is critical to the learning process. We thus suggest that multiway ground truth is more suitable for evaluation.

For future work, we will continue exploring how underlying mechanisms of Chinese discourse structure interact with different parsing policies, from left-heavy/right-heavy binarization choice to more fundamental transition-based/chart-based parsing approaches. We aim to conduct more extensive experiments to gain more insights into Chinese discourse parsing.

6. Acknowledgements

This research was partially supported by the Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-108-2634-F-002-008-, MOST-108-2218-E-009-051-, and MOST-109-2634-F-002-034 and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

7. Bibliographical References

Black, E., Abney, S., Flickenger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., and Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Heilman, M. and Sagae, K. (2015). Fast rhetorical structure theory discourse parsing. *CoRR*, abs/1505.02425.

Ji, Y. and Eisenstein, J. (2014). Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland, June. Association for Computational Linguistics.

Joty, S., Carenini, G., Ng, R., and Mehdad, Y. (2013). Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria, August. Association for Computational Linguistics.

Kong, F. and Zhou, G. (2017). A cdt-styled end-to-end chinese discourse parser. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 16(4):26:1–26:17, July.

Li, J., Li, R., and Hovy, E. (2014a). Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar, October. Association for Computational Linguistics.

Li, Y., Feng, W., Sun, J., Kong, F., and Zhou, G. (2014b). Building chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 2105–2114, Doha, Qatar, October. Association for Computational Linguistics.

Li, Q., Li, T., and Chang, B. (2016). Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371, Austin, Texas, November. Association for Computational Linguistics.

Lin, C.-A., Huang, H.-H., Chen, Z.-Y., and Chen, H.-H. (2018). A unified RvNN framework for end-to-end Chinese discourse parsing. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 73–77, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Morey, M., Muller, P., and Asher, N. (2017). How much progress have we made on rst discourse parsing? a replication study of recent results on the rst-dt. In *EMNLP*.

Sun, C. and Kong, F. (2018). A transition-based framework for chinese discourse structure parsing. *Journal of Chinese Imformation Processing*, 32(12):48.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yu, N., Zhang, M., and Fu, G. (2018). Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 69–77, Jeju Island, Korea, July. Association for Computational Linguistics.

8. Language Resource References

Carlson, L., Marcu, D., and Okurovsky, M. E. (2001). Building a discourse-tagged corpus in the framework of

rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Li, Y., Feng, W., Sun, J., Kong, F., and Zhou, G. (2014). Building chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 2105–2114, Doha, Qatar, October. Association for Computational Linguistics.