A Complete Shift-Reduce Chinese Discourse Parser with Robust Dynamic Oracle

Shyh-Shiun Hung,[1] Hen-Hsen Huang,[2,3] and Hsin-Hsi Chen[1,3]
[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[2] Department of Computer Science, National Chengchi University, Taiwan
[3] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
shhung@nlg.csie.ntu.edu.tw, hhhuang@nccu.edu.tw,
hhchen@ntu.edu.tw

Abstract

This work proposes a standalone, complete Chinese discourse parser for practical applications. We approach Chinese discourse parsing from a variety of aspects and improve the shift-reduce parser not only by integrating the pre-trained text encoder, but also by employing novel training strategies. We revise the dynamic-oracle procedure for training the shift-reduce parser, and apply unsupervised data augmentation to enhance rhetorical relation recognition. Experimental results show that our Chinese discourse parser achieves the state-of-the-art performance.

1 Introduction

Discourse parsing is one of the fundamental tasks in natural language processing (NLP). Typical types of discourse parsing include hierarchical discourse parsing and shallow discourse parsing. The former is aimed at finding the relationships among a series of neighboring elementary discourse units (EDUs) and further building up a hierarchical tree structure (Mann and Thompson, 1988). Instead of establishing a tree structure, the latter finds the across-paragraph relations between all text units in a paragraph or a document. Based on Rhetorical Structure Theory Discourse Treebank (RST-DT) (Carlson et al., 2001a), hierarchical discourse parsing in English has been well-studied.

This paper focuses on hierarchical discourse parsing in Chinese. Previous work on hierarchical Chinese discourse parsing is mostly based on the RST-style Chinese Discourse Treebank (Li et al., 2014). To distinguish from the other Chinese Discourse Treebank (Zhou and Xue, 2012), which is annotated with the PDTB-style for shallow discourse parsing, we use the term CDTB-14 to refer to the RST-style one and the term CDTB-12 to refer to the PDTB-style one. Kong and Zhou (2017) propose a pipeline framework and generate the discourse parsing tree in a bottom-up way. Lin et al. (2018) propose an end-to-end system based on a recursive neural network (RvNN) to construct the parsing tree with a CKY-like algorithm. Sun and Kong (2018) use transition-based method with the stack augmented parser-interpreter neural network (SPINN) (Bowman et al., 2016) as the backbone model, helping their model make a better prediction with the previous information.

In this work, we attempt to construct a complete Chinese discourse parser, which supports all the four sub-tasks in hierarchical discourse parsing, including EDU segmentation, tree structure construction, nuclearity labeling, and rhetorical relation recognition. Given a paragraph, our parser extracts all EDUs, builds the tree structure, identifies the nucleuses, and recognizes the rhetorical relations of all internal nodes. We propose a revised dynamic-oracle procedure (Yu et al., 2018) for training the shift-reduce parser. Because of the limited training instances in CDTB-14, we also address the data sparsity issue by introducing unsupervised data augmentation (Xie et al., 2019). Experimental results show that our methodology is effective, and our model outperforms all the previous models. The contributions of this work are three-fold shown as follows.

1. We explore the task of Chinese discourse parsing with a variety of strategies, and our parser achieves the state-of-the-art performance. Our robust dynamic-oracle procedure can be applied to other shift-reduce parsers.

2. Our complete Chinese discourse parser handles a raw paragraph/document directly and performs all the subtasks in hierarchical discourse parsing. No pre-processing procedures such as Chinese word segmentation, POS-tagging, and syntactic parsing are required.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 133–138
July 5 - 10, 2020.

3. We release the pre-trained, standalone, ready-to-use parser as a resource for the research community.[1]

2 Methodology

Figure 1 gives an overview of our parser. Five stages are performed to transform a raw document into a parse tree: EDU segmentation, tree structure construction, rhetorical relation and nuclearity classification, binary tree conversion, and beam search.

2.1 Elementary Discourse Unit Segmentation

Typically, EDU segmentation is a sequence labeling task (Wang et al., 2018; Peters et al., 2018). We propose a model for labeling each Chinese character in a raw document. The Begin-Inside scheme is employed that the word beginning with a new EDU will be labeled as *B*, and the rest of the words will be labeled as *I*. Our model is based on the pre-trained text encoder BERT (Devlin et al., 2018). More specifically, we adopt the version *BERT-base, Chinese* since this is the only pre-trained BERT dedicated to Chinese so far. As the BERT for Chinese is character-based, we feed each Chinese character into a BERT layer to obtain its contextual embedding. Then, we fine tune the representation with an additional dense layer and measure the probability of each label of each character with a softmax layer. The model is further trained as conditional random fields (CRFs) (Lafferty et al., 2001) for finding the global optimal label sequence.

2.2 Tree Construction

We propose a shift-reduce parser for building the structure of the discourse parse tree. A shift-reduce parser maintains a stack and a queue for representing a state during parsing, and an action classifier is trained to predict the action (i.e., shift or reduce) for making a transition from the given state to the next state. In the initial state, the stack is empty, and the queue contains all the EDUs in a raw document. In the final state, the queue is empty, and the stack contains only one element, i.e., the discourse parse tree of the whole paragraph.

To decide whether to shift or to reduce, we propose an action classifier by considering the information of the top two elements of the stack s_1 and s_2 (i.e., the two most recent discourse units) and the first element of the queue q (i.e., the next

[1] https://github.com/jeffrey9977/Chinese-Discourse-Parser-ACL2020

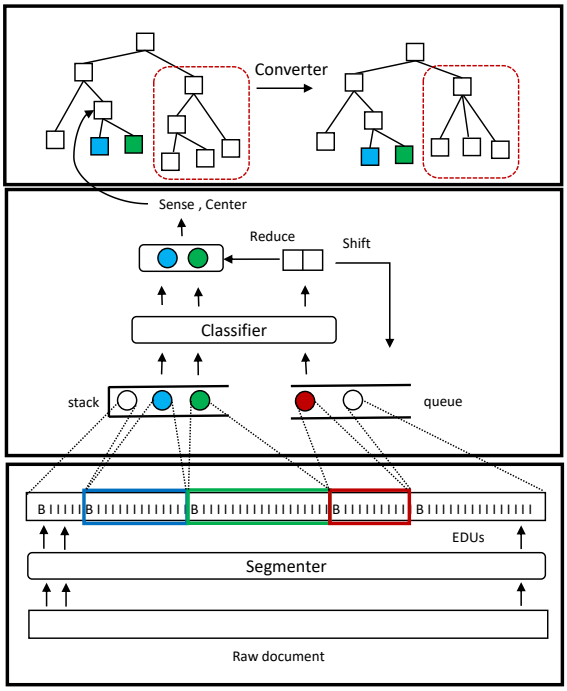

Figure 1: Overview of our Chinese discourse parser.

EDU). The textual form of each of these three discourse units will be fed into the BERT encoder for representing as $Enc(s_1)$, $Enc(s_2)$, and $Enc(q)$. Next, we concatenate the max pooling of $Enc(s_1)$, $Enc(s_2)$, and $Enc(q)$ and feed the resulting vector into a dense layer to predict the next action.

Since shift-reduce is a greedy algorithm, it can hardly recover from an error state. The shift-reduce parser is typically trained with the teacher mode, where only correct states are given, and the resulting parser may perform poor when it reaches unfamiliar states. For this reason, we propose a revised dynamic-oracle procedure (Yu et al., 2018) for training our discourse parser. One drawback of the original dynamic oracle is that some golden training examples may be neglected. Because CDTB-14 has relatively few action steps to build a tree, the probability of falling into a wrong state is much small compared to that of RST-DT. In our revision, we want to guarantee all correct states have been trained. As shown in Algorithm 1, the document will be gone through twice when training a document example. We first follow the golden actions, and choose action predicted by the model with a probability α at the second time. We refer to them as teacher mode and student mode, respectively. Note that we follow the suggestion of Yu et al. (2018) to set α to 0.7.

Algorithm 1 Training Procedure for Our Shift-Reduce Discourse Parser.

$S, Q \leftarrow$ empty stack, elementary discourse units
while Q is not empty $\vee$ S has more than 1 unit **do** ▷ Teacher mode
  $predicted, golden \leftarrow$ ACTIONCLASSIFIER($S.top_1(), S.top_2(), Q.front()$) , GOLDENACTION
  COMPUTELOSSANDUPDATE($predicted, golden$)
  PERFORMACTION($golden$)
$S, Q \leftarrow$ empty stack, elementary discourse units
while Q is not empty $\vee$ S has more than 1 unit **do** ▷ Student mode
  $predicted, golden \leftarrow$ ACTIONCLASSIFIER($S.top_1(), S.top_2(), Q.front()$) , GOLDENACTION
  COMPUTELOSSANDUPDATE($predicted, golden$)
  if $rand() > \alpha$ **then** PERFORMACTION($golden$) **else** PERFORMACTION($predicted$)

2.3 Rhetorical Relation Recognition

If two discourse units are decided to be merged during the tree construction stage, a new internal node will be generated and the relationship of the two discourse units will be determined. Predicting the relation between two textual arguments is a typical classification task in NLP. We propose a BERT-based classifier, which predicts the relation of two arguments separated by the symbol `[SEP]`, with additional dense layers as the output.

In CDTB-14, the "coordination" relation accounts for 59.6% of the training data, while minor relations suffer from data sparseness. To address this issue, we introduce unsupervised data augmentation (UDA) (Xie et al., 2019) to enhance the performance. We adopt the discourse pairs in CDTB-12 as the material for UDA. Note that other unlabeled text pairs can also be used for UDA. We chose those from CDTB-12 simply because the format is convenient for us to use.

The original loss is shown as Eq. 1. Given a span of text x, our main model $P(\cdot)$ predicts the rhetorical relation y_c. Eq. 2 shows the additional consistency loss to enforce the smoothness of our main model, and $\hat{x}$ stands for the augmented unlabeled sentence pair. L and U stand for labeled data and unlabeled data, respectively. As shown in Eq. 3, we train both objectives at the same time with a weight λ to adjust the effect of UDA.

$$H = -\frac{1}{N}\sum_{x\in L}^{N}\sum_{c=1}^{M} y_c \log\left(P(y_c|x)\right) \quad (1)$$

$$D_{KL} = -\frac{1}{N}\sum_{x\in U}^{N} P\left(y|x\right)\log\left(\frac{P(y|x)}{P(y|\hat{x})}\right) \quad (2)$$

$$\mathcal{L}\left(\theta\right) = H + \lambda\, D_{KL} \quad (3)$$

The UDA procedure first generates the augmented unlabeled sentence pairs. Various approaches to paraphrasing can be employed. In this work, we utilize the back-translation strategy (Sennrich et al., 2016), where we translate the Chinese sentence pair to English and then translate back to Chinese. This is equivalent to add noises to the original inputs. As the original and the back-translated sentence pairs express the same meaning, our model is expected to predict the same label for both pairs. By minimizing the consistency loss, our model can behave consistently no matter whether an original instance or its paraphrases are given. In this way, the model can be more generalized and robust. Besides, when our model is able to predict the same label for both sentence pairs, it means that our model has also learned their label.

2.4 Nuclearity Labeling

Nuclearity labeling is aimed at determining the nucleus from a sentence pair. The nuclearity of two sentences has a correlation with their relationship, thus we jointly train the rhetorical relation and the nuclearity classifiers, where the loss for back-propagation is the sum of the losses of both classifiers. Similar to the imbalance issue of rhetorical relation recognition, the 'Equal' class accounts for 51% of training data. We also employ UDA for performance enhancement.

2.5 Binary Tree Conversion

For simplicity, our shift-reduce parser constructs a binary tree. However, the parse trees annotated in CDTB-14 are not always binary. In the training and the test sets, 8.9% and 10% of the internal nodes have more than two children, respectively. Most of the previous works do not handle the binary tree conversion, and some of the work further convert the golden trees into binary trees to calculate their scores, resulting in less accurate evaluation. In the

training stage, we convert the multiway trees to their corresponding left-heavy binary trees (Morey et al., 2018). In the testing stage, we convert the binary tree constructed by our parser to the corresponding multiway tree. For example, a three-way node, $A \rightarrow XYZ$, will be converted to $A \rightarrow A'Z$ and $A' \rightarrow XY$. The conversion is deterministic and bidirectional, so it is free from ambiguity.

2.6 Beam Search

To decode a transition sequence during the testing stage, the standard method is to choose the action that has the maximum probability of the current time step as the input for the next time step. However, this greedy approach might fail to find the sequence that has the maximum overall probability only because one of the action probability is small in that sequence. Beam search (Wiseman and Rush, 2016) is a heuristic search algorithm that explores a graph by maintaining the top k results at every time step. This approach helps keep a number of potential candidates from discarding. Note that the greedy approach is equivalent to beam search with a beam width $k = 1$.

When performing the shift-reduce parsing, two kinds of states have only one action to choose: (1) less than two elements in the stack, and (2) no element in the queue. Under the above two conditions, the probability of the selected action will be 1, making our model to be overly biased on those sequences having many non-optional stages. For this reason, we apply an alternative way to compute the sequence probability during beam search. Our modified beam search is still fulfilled by maintaining the top k sequences, but the score of a sequence is calculated by the average probabilities of the selected actions that have more than one choice.

3 Experiments

3.1 Experimental Settings

Following the setting of Kong and Zhou (2017), we divide CDTB-14 into the training set, including 450 articles (2,125 paragraphs), and test set, including 50 articles (217 paragraphs). We keep 10% of the training data for validation. PARSEVAL (Carlson et al., 2001b) is used for evaluation.

3.2 Experimental Results

Table 1 shows the performances of our parser in micro-averaged F-score, compared with previous work **Zhou** (Kong and Zhou, 2017) and **Lin** (Lin et al., 2018). We also implement **BERT-CKY**, a CKY parser by using BERT for representation, as an additional baseline model. The evaluation is based on multiway trees.

Model	EDU	+T	+R	+N	All
Zhou	Given	52.3	33.8	23.9	23.2
Lin		64.6	42.7	38.5	35.0
BERT-CKY		76.5	50.8	48.5	43.1
Ours		**82.8**	**57.6**	**56.0**	**50.5**
Zhou	93.8	46.4	28.8	23.1	22.0
Lin	87.2	49.5	32.6	28.8	26.8
BERT-CKY	92.4	68.9	43.3	42.0	37.0
Normal	**97.4**	78.8	54.6	52.0	47.1
Dynamic	**97.4**	78.9	54.5	51.8	47.1
Ours	**97.4**	**80.0**	**55.9**	**53.6**	**48.9**

Table 1: Performances of EDU segmentation (EDU), tree construction (T), rhetorical relation recognition (R), nuclearity labeling (N), and all subtasks, reported in Micro-averaged F-score.

Both the performances with and without golden EDUs are measured. The results show that BERT is highly competitive and has the ability to catch the potential relations between discourse units since **Lin** and **BERT-CKY** basically use the same approach while the latter model uses BERT as the text encoder. Our parser outperforms all the baseline models and achieves a significant improvement without the golden EDUs given. Note that **BERT-CKY** is based on Lin et al. (2018), which has its own EDU segmentation module different from ours, hence the EDU score is different.

We examine the performance of three different training techniques for shift-reduce parsing. As mentioned in Section 2.2, **Normal** stands for action classifier trained with gold standard actions, **Dynamic** stands for Dynamic Oracle introduced by Yu et al. (2018), and **Ours** stands for our revised dynamic-oracle procedure where the model is trained with both gold standard actions and dynamic oracle actions.

Compared to **Normal**, experimental results show no improvement made by the original dynamic oracle, while our revised dynamic oracle outperforms the other two strategies. Our strategy does not ignore the golden action in every correct state and also has the chance to explore error states.

In order to compare with **SUN** (Sun and Kong, 2018), we convert the golden standard trees into binary trees and measure the performances on bi-

Model	EDU	+T	+R	+N	All
Sun	93.0	78.2		53.2	
Ours	**97.4**	**83.3**	**58.1**	**55.7**	**52.0**

Table 2: Performances measured on binary trees, reported in macro-averaged F-score.

nary trees in macro-averaged F-score. The results are shown in Table 2. Sun and Kong (2018) do not address all subtasks in Chinese discourse parsing, and our model outperforms **SUN** in every subtask.

Relation		P	R	F
Coordination	-UDA	84.3	77.8	80.9
	+UDA	90.7	76.9	83.2
Causality	-UDA	38.7	43.2	40.8
	+UDA	38.7	55.4	45.6
Transition	-UDA	80.0	80.0	80.0
	+UDA	80.0	88.9	84.2
Explanation	-UDA	46.0	57.6	51.1
	+UDA	45.2	70.9	55.2

Table 3: Performances of the four rhetorical relations (%) with and without UDA. Occurrences of these relations are 59.6%, 17.1%, 1.6%, and 21.7%, respectively.

3.3 Discussions

To examine the effectiveness of UDA, Table 3 shows the performances of rhetorical relation recognition with and without UDA. Experimental results show that application of UDA successfully enhances the recall scores of the three minor classes with a little trade-off in the recall score of the dominant class, Coordination. In addition, the F-scores of all the four relations are increased. In other words, applying UDA deals with the data imbalance issue and improves the overall performance. Applying UDA to nuclearity classification also has a similar improvement as Table 3.

Theoretically, beam search with a larger beam width helps find a better solution. As shown in Table 4, however, our parser is worse when a larger beam width is used, which means the sequence having higher overall score does not ensure the better decoding result. Our experiment only shows the beam widths up to five because the scores of worse sequences are already higher than that of the correct sequence in some cases. That is, the larger beam widths seem to be unnecessary.

Beam Size	EDU	+T	+R	+N	All
$k = 1$	Given	**82.8**	**57.6**	**56.0**	**50.5**
$k = 2$		81.8	56.8	55.1	49.7
$k = 5$		81.7	56.7	54.9	49.6

Table 4: Performances of beam search with different beam widths.

The reason may be that beam search is not really suitable for the shift-reduce paradigm. For example, a sequence might fall into a seriously bad stage but the rest of actions can be easily determined so that the sequence will get a high overall probability. This assumption also implies that unlike beam search applied on sequence to sequence model, we cannot judge a transition sequence is good or bad by solely considering its overall score. In addition, for longer textual units such as paragraph, human readers and writers may not follow the assumption of overall optimization. Instead, human beings read and write sequentially, similar to the greedy nature.

We also evaluate our approach in English discourse parsing. The famous dataset, RST-DT, is used. Our model achieves F-scores of 85.0%, 58.8%, 69.9%, and 56.7% in tree construction, rhetorical relation recognition, nuclearity labeling, and all subtasks, respectively. The overall performance is similar to that of the state-of-the-art model (Yu et al., 2018).

4 Conclusion

This work proposes a standalone, complete Chinese discourse parser. We integrate BERT, UDA, and a revised training procedure for constructing a robust shift-reduce parser. Our model is compared with a number of previous models, and experimental results show that our model achieves the state-of-the-art performance and is highly competitive with different setups. We will explore cross-lingual transfer learning for supporting more languages.

Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-109-2634-F-002-040-, MOST-109-2634-F-002-034-, MOST-108-2218-E-009-051-, and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

References

Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1477, Berlin, Germany. Association for Computational Linguistics.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001a. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001b. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue (SIGDIAL'01)*, pages 1–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yancui Li, Wenhe Feng, Jing Sun, Fang Kong, and Guodong Zhou. 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)*, pages 2105–2114, Doha, Qatar. Association for Computational Linguistics.

Chuan-An Lin, Hen-Hsen Huang, Zi-Yuan Chen, and Hsin-Hsi Chen. 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. Association for Computational Linguistics.

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

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on rst discourse parsing and evaluation. *Comput. Linguist.*, 44(2):197–235.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. *CoRR*, abs/1606.02960.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Nan Yu, Meishan Zhang, and Guohong Fu. 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. Association for Computational Linguistics.

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)*, pages 69–77, Jeju Island, Korea. Association for Computational Linguistics.