GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing

Siyao Peng, Yang Janet Liu & Amir Zeldes

Department of Linguistics @ Georgetown University
Washington D.C., USA

AACL 2022 (Online) @ October 31, 2022

Table of Contents

Rhetorical Structure Theory

- RST (Mann and Thompson, 1988) creates a discourse tree over Elementary Discourse Units (EDUs) within a document;
- Significant at the document level and important for text summarization (Xu et al., 2020; Xiao et al., 2020; Huang and Kurohashi, 2021) and sentiment analysis (Kraus and Feuerriegel, 2019; Huber and Carenini, 2020).

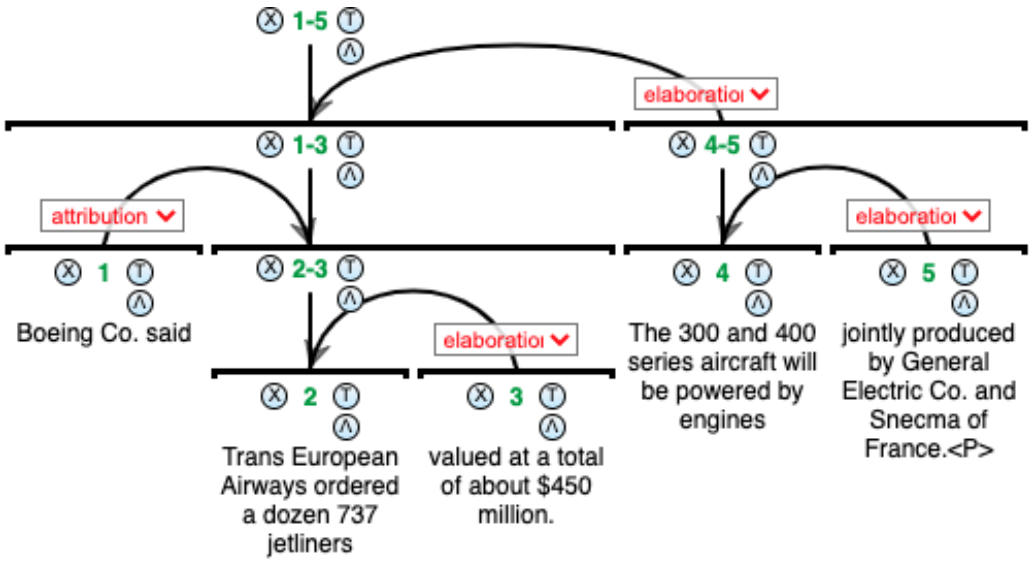

Figure 1: An example from *wsj_1153* in RST-DT.

Lack of Chinese RST Data

RST Datasets

Many datasets came out in the past two decades:

- English: RST-DT (Carlson et al., 2001) and GUM (Zeldes, 2017b);
- German (Stede and Neumann, 2014), Dutch (Redeker et al., 2012), Spanish (da Cunha et al., 2011), etc.

However, there is a gap for Chinese.

Chinese Discourse Datasets

- CDT-CDTB (Li et al., 2014): small discourse trees within paragraphs;
- MCDTB (Jiang et al., 2018): discourse trees between paragraphs;
- Sci-CDTB (Cheng and Li, 2019): Discourse Dependency for abstracts;
- Spanish-Chinese corpus (Cao et al., 2018): translation-aligned EDUs.

However, none annotates a single-rooted tree for longer documents.

Table of Contents

Georgetown Chinese Discourse Treebank (GCDT)

- Largest RST dataset in Mandarin Chinese;
- 50 documents evenly across five genres: academic articles, biographies, Wikipedia interviews, Wikinews, and how-to guides;
- Open-source and extensive guidelines.

Genre	#Docs	#Toks	# EDUs	Source
academic	10	14,168	2,033	hanspub.org/
bio	10	13,485	2,018	zh.wikipedia.org/
interview	10	11,464	1,810	zh.wikinews.org/
news	10	11,249	1,652	zh.wikinews.org/
whow	10	12,539	2,197	zh.wikihow.com/
Total	50	62,905	9,710	

Table 1: GCDT corpus statistics.

Annotation Procedures

XML and Metadata Annotation

- Document metadata and gold sentence, paragraph, & section breaks.

Tokenization

- Fundamental to EDU boundaries;
- Following CTB's (Xue, 2005) guidelines (Xia, 2000a).

EDU Segmentation

- Essential to RST annotation, and we equate EDUs with the propositional structure of clauses using POS guidelines (Xia, 2000b);
- First RST corpus that segments prenominal relative clauses as EDUs.

Relation Annotation

- Two-level relation labels from GUM 8.0.0 (Zeldes, 2017a) with 15 coarse and 32 fine-grained relations.

GCDT example annotations

This example presents two prenominal relative clauses as *elaboration-attribute* in a RST tree.

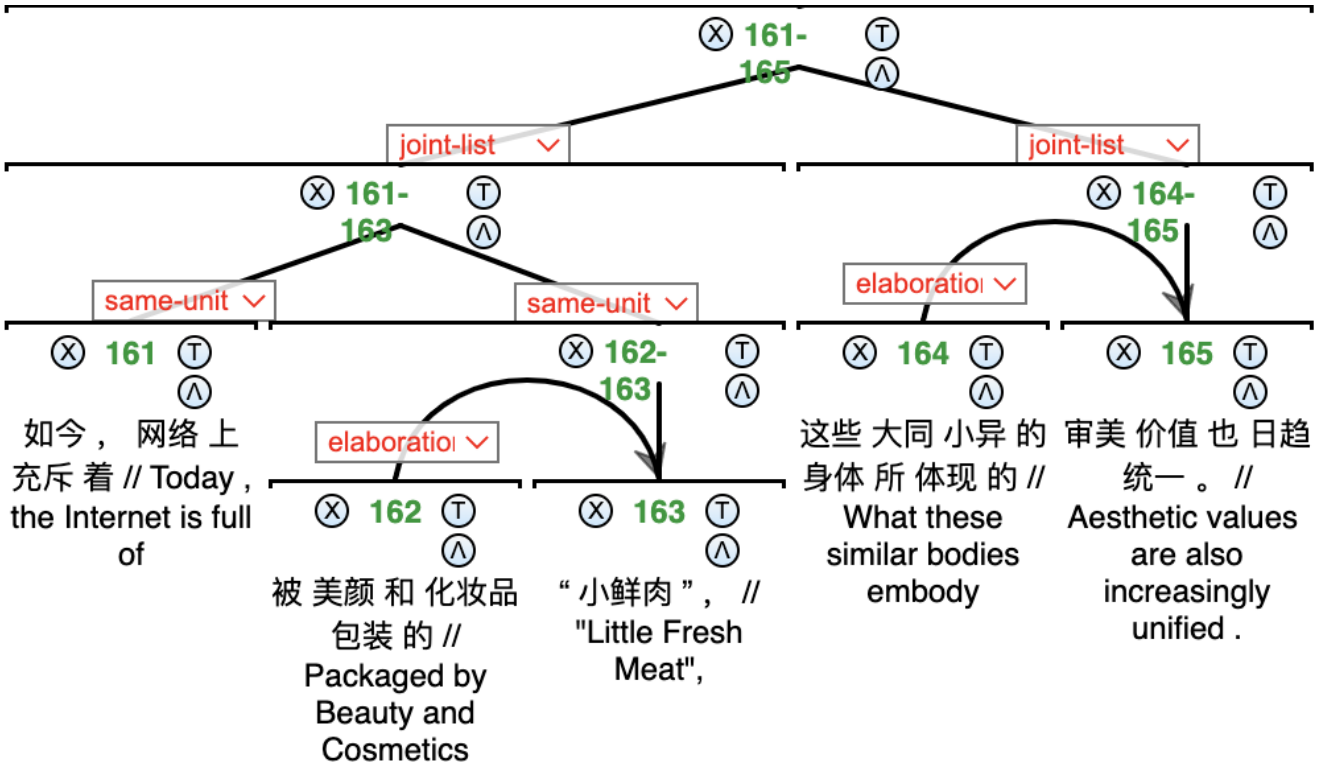

Figure 2: A RST subtree with two relative clauses annotated as *elaboration-attribute* with EDU-wise zh→en automatic translations.

Data Management & Agreement

Data Split

- 8-1-1 train-dev-test split per genre;
- Both human inter-annotator agreements and parsing results are assessed on the five test documents, one from each genre.

Inter-Annotator Agreement

Token-wise Segmentation		Original Parseval		
Accuracy	Cohen's κ	Span	Nuclearity	Relation
97.4%	0.89	84.3	66.2	57.8

Table 2: GCDT's Inter-Annotator Agreement (IAA).

Table of Contents

Experiments

Highlights

- Benchmark parsing results on the new GCDT and the GUM V8.0.0;
- Multilingual training on GCDT and GUM with fine-tuning and automatic translation;
- Isolating cross-lingual versus cross-genre influences in GUM.

Setups

- **Parser**: SOTA multilingual parser (DMRST, Liu et al. 2021);
- **Datasets:**
  - **GCDT**: 50 Chinese documents from 5 genres;
  - **GUM-12**: 193 English documents from 12 genres;
  - **GUM-5**: 99 GUM documents from the same five genres in GCDT;
- **Metrics:** Original Parseval for Span, Nuclearity, and Relation on binarized trees (Morey et al., 2017);
- **Language Models:** Chinese, English, and multi-lingual models.

Monolingual Results

corpus	monolingual embedding	Span	Nuc	Rel
GCDT	*bert-base-chinese*	73.15	55.71	50.81
	bert-base-multilingual-cased	67.34	47.66	43.97
	hfl/chinese-roberta-wwm-ext	**75.51**	**57.08**	**51.76**
	xlm-roberta-base	74.35	54.17	50.45
GUM-12	*bert-base-cased*	60.93	47.92	40.20
	bert-base-multilingual-cased	64.47	50.69	43.25
	roberta-base	**68.59**	**55.32**	**46.29**
	xlm-roberta-base	66.12	52.58	45.06

Table 3: Monolingual results on GCDT and GUM-12 with Chinese, English, and multilingual BERT and RoBERTa embeddings (averaged over five runs).

- RoBERTa outperforms BERT in both English and Chinese;
- Monolingual RoBERTa embeddings achieve SOTA.

Multilingual Experiment Setups

Language Models

- Using best-performing language embeddings from monolingual experiments;
  - Chinese (ZH): *hfl/chinese-roberta-wwm-ext*;
  - English (EN): *roberta-base*;
  - Multilingual (XLM): *xlm-roberta-base*;

Finetuning

- Firstly pre-train on GUM+GCDT and then continue to train only on the training set of the target corpus;

Automatic EDU-wise Translation

- Conducting EDU-wise automatic translations using Google Translate and merge the translated RST trees with the target dataset.

Experiments: Multilingual Results 1

Experiment	Span	Nuc	Rel
Test on GCDT			
Train on GCDT	74.35	54.17	50.45
Train on GCDT+GUM-12	**74.33**	**57.24**	**52.61**
Human Agreement	84.27	66.15	57.77
Test on GUM-5			
Train on GUM-5	72.45	56.78	47.69
Train on GUM-5+GCDT	**72.56**	**60.63**	**52.57**
Test on GUM-12			
Train on GUM-12	66.12	52.58	45.06
Train on GUM-12+GCDT	**70.32**	**57.49**	**49.14**

Table 4: Multilingual parsing results with GCDT+GUM using *xlm-roberta-base* (averaged over five runs).

- Firstly, joint training outperformed monolingual results on GCDT, GUM-5, and GUM-12.

Experiments: Multilingual Results 2

Experiment	Span	Nuc	Rel
Train on GCDT+GUM-5	74.24	56.68	52.21
Train on GCDT+GUM-12	**74.33**	**57.24**	**52.61**
Human Agreement	84.27	66.15	57.77

Table 5: Multilingual parsing results with GCDT+GUM-5 versus GCDT+GUM-12 when tested on GCDT using *xlm-roberta-base* (averaged over five runs).

- Secondly, more genres from GUM (GCDT+GUM-12) achieved slightly better performance than training only using the same genres (GCDT+GUM-5) when tested on GCDT.

Experiments: Multilingual Results 3

Experiment	Span	Nuc	Rel
Test on GCDT			
Train on GCDT+GUM-5	74.24	56.68	52.21
Train on GCDT+GUM-12	74.33	57.24	52.61
Train on GCDT+GUM-5 and finetune on GCDT	76.97	57.94	53.38
Train on GCDT+GUM-12 and fine-tune on GCDT	**77.25**	**59.43**	**55.41**
Human Agreement	84.27	66.15	57.77
Test on GUM-5			
Train on GUM-5+GCDT	**72.56**	**60.63**	**52.57**
Train on GUM-5+GCDT and finetune on GUM-5	73.44	59.40	50.57
Test on GUM-12			
Train on GUM-12+GCDT	**70.32**	**57.49**	**49.14**
Train on GUM-12+GCDT and finetune on GUM-12	66.00	53.13	45.47

Table 6: Multilingual parsing results with finetuning on GCDT+GUM combinations using *xlm-roberta-base* (averaged over five runs).

- Thirdly, pretraining on the GCDT+GUM-combined training sets and finetuning on the training set of the target corpus improves performance on Chinese GCDT but deteriorates on the English GUM.

Experiments: Multilingual Results 4

Experiment	Span	Nuc	Rel	Experiment	Span	Nuc	Rel
Train on GCDT+GUM-5 and Dev/Test on GCDT				**Train on GUM-5+GCDT and Dev/Test on GUM-5**			
multitrain w/ XLM RoBERTa	74.24	56.68	52.21	multitrain w/ XLM RoBERTa	72.56	60.63	52.57
+finetuning w/ XLM RoBERTa	76.97	57.94	53.38	+finetuning w/ XLM RoBERTa	73.44	59.40	50.57
+en→zh trans. w/ XLM RoBERTa	74.80	56.58	51.18	+zh→en trans. w/ XLM RoBERTa	72.21	60.07	52.32
+en→zh trans. w/ ZH RoBERTa	**77.66**	**59.29**	**54.66**	**+zh→en trans. w/ EN RoBERTa**	**74.73**	**62.65**	**54.32**
Train on GCDT+GUM-12 and Dev/Test on GCDT				**Train on GUM-12+GCDT and Dev/Test on GUM-12**			
multitrain w/ XLM RoBERTa	74.33	57.24	52.61	multitrain w/ XLM RoBERTa	70.32	57.49	49.14
+finetuning w/ XLM RoBERTa	**77.25**	**59.43**	**55.41**	+finetuning w/ XLM RoBERTa	66.00	53.13	45.47
+en→zh trans. w/ XLM RoBERTa	73.99	56.31	51.51	+zh→en trans. w/ XLM RoBERTa	70.28	57.63	49.26
+en→zh trans. w/ ZH RoBERTa	78.11	59.42	54.41	**+zh→en trans. w/ EN RoBERTa**	**71.41**	**59.17**	**50.63**
Human Agreement	**84.27**	**66.15**	**57.77**				

Table 7: Multilingual parsing results with finetuning and automatic translation on GCDT+GUM combinations (averaged over five runs).

- Lastly, augmenting with automatic translation and using monolingual embeddings achieved the best performance in most scenarios.

Genre-wise Results

Genre	Trained on GCDT			Trained w/ trans on GCDT+GUM-5			Trained w/ trans on GCDT+GUM-12			Human Agreement		
	Span	Nuc	Rel	Span	Nuc	Rel	Span	Nuc	Rel	Span	Nuc	Rel
academic	74.64	54.07	48.33	72.25	47.37	43.54	75.12	51.20	44.98	80.38	59.33	49.76
bio	72.87	54.26	52.71	74.81	57.75	53.49	77.52	59.69	55.43	81.57	63.92	55.69
interview	74.68	56.33	52.53	80.38	61.39	55.70	77.85	56.96	48.73	83.55	62.50	54.61
news	76.63	56.52	50.54	83.15	64.13	57.07	78.80	60.33	54.35	80.98	61.96	54.35
whow	77.89	57.76	54.79	80.20	66.34	62.71	80.20	65.68	61.06	91.99	77.70	69.34
Overall	75.45	55.85	52.07	77.97	59.71	55.04	78.06	59.44	53.87	84.27	66.15	57.77

Table 8: Genre-wise performances of sample models trained on GCDT, as well as translation-augmented GCDT+GUM-5 and GCDT+GUM-12 combinations using *hfl/chinese-roberta-wwm-ext*.

Results

- How-to guides (*whow*) performs much better than *academic* for both models and humans – a good human-model alignment;
- Model results on *whow* are the farthest from the human ceiling.

Summary

- The largest Rhetorical Structure Theory (RST) dataset for Mandarin Chinese – GCDT;
- SOTA parsing results on GCDT and the similar English GUM corpus;
- Multilingual training, fine-tuning, and automatic EDU translation;
- Genre diversity and genre-wise results;
- Data, guideline and code available at `https://github.com/logan-siyao-peng/GCDT`.


Questions?
Comments?

References I

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The RST Spanish-Chinese Treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 156–166. https://aclanthology.org/W18-4917

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*. https://aclanthology.org/W01-1605

Yi Cheng and Sujian Li. 2019. Zero-shot Chinese Discourse Dependency Parsing via Cross-lingual Mapping. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*. Association for Computational Linguistics, Tokyo, Japan, 24–29. https://doi.org/10.18653/v1/W19-8104

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the Development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*. Association for Computational Linguistics, Portland, Oregon, USA, 1–10. https://aclanthology.org/W11-0401

Yin Jou Huang and Sadao Kurohashi. 2021. Extractive Summarization Considering Discourse and Coreference Relations based on Heterogeneous Graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 3046–3052. https://doi.org/10.18653/v1/2021.eacl-main.265

References II

Patrick Huber and Giuseppe Carenini. 2020. From Sentiment Annotations to Sentiment Prediction through Discourse Augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 185–197. https://doi.org/10.18653/v1/2020.coling-main.16

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. MCDTB: A Macro-level Chinese Discourse TreeBank. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3493–3504. https://aclanthology.org/C18-1296

Mathias Kraus and Stefan Feuerriegel. 2019. Sentiment Analysis Based on Rhetorical Structure Theory: Learning Deep Neural Networks from Discourse Trees. *Expert Syst. Appl.* 118, C (mar 2019), 6579. https://doi.org/10.1016/j.eswa.2018.10.002

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)*. Association for Computational Linguistics, Doha, Qatar, 2105–2114. https://doi.org/10.3115/v1/D14-1224

References III

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*. Association for Computational Linguistics, Punta Cana, Dominican Republic and Online, 154–164. https://doi.org/10.18653/v1/2021.codi-main.15

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

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1319–1324. https://doi.org/10.18653/v1/D17-1136

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multi-Layer Discourse Annotation of a Dutch Text Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 2820–2825. http://www.lrec-conf.org/proceedings/lrec2012/pdf/887_Paper.pdf

References IV

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 925–929. http://www.lrec-conf.org/proceedings/lrec2014/pdf/579_Paper.pdf

Fei Xia. 2000a. The Segmentation Guidelines for the Penn Chinese Treebank (3.0). (2000), 33.

Fei Xia. 2000b. The Part-of-Speech Guidelines for the Penn Chinese Treebank (3.0). (2000).

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2020. Do We Really Need That Many Parameters In Transformer For Extractive Summarization? Discourse Can Help !. In *Proceedings of the First Workshop on Computational Approaches to Discourse*. Association for Computational Linguistics, Online, 124–134. https://doi.org/10.18653/v1/2020.codi-1.13

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-Aware Neural Extractive Text Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5021–5031. https://doi.org/10.18653/v1/2020.acl-main.451

Nianwen Xue. 2005. Annotating Discourse Connectives in the Chinese Treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*. Association for Computational Linguistics, Ann Arbor, Michigan, 84–91. https://www.aclweb.org/anthology/W05-0312

References V

Amir Zeldes. 2017a. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation* 51, 3 (Sept. 2017), 581–612. https://doi.org/10.1007/s10579-016-9343-x

Amir Zeldes. 2017b. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation* 51, 3 (2017), 581–612.