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

Siyao Peng **Yang Janet Liu** **Amir Zeldes**
Department of Linguistics, Georgetown University
`{sp1184, yl879, amir.zeldes}@georgetown.edu`

Abstract

A lack of large-scale human-annotated data has hampered the hierarchical discourse parsing of Chinese. In this paper, we present GCDT, the largest hierarchical discourse treebank for Mandarin Chinese in the framework of Rhetorical Structure Theory (RST). GCDT covers over 60K tokens across five genres of freely available text, using the same relation inventory as contemporary RST treebanks for English. We also report on this dataset's parsing experiments, including state-of-the-art (SOTA) scores for Chinese RST parsing and RST parsing on the English GUM dataset, using cross-lingual training in Chinese and English with multilingual embeddings.

1 Introduction

Hierarchical discourse parsing has shown its importance in document-level natural language understanding (NLU) tasks, such as text summarization (Yoshida et al., 2014; Goyal and Eisenstein, 2016; Xu et al., 2020; Xiao et al., 2020; Huang and Kurohashi, 2021) and sentiment analysis (Bhatia et al., 2015; Markle-Hus et al., 2017; Kraus and Feuerriegel, 2019; Huber and Carenini, 2020). Among discourse frameworks, Rhetorical Structure Theory (RST, Mann and Thompson 1988) is a document-level discourse analysis formalism that assumes a single-rooted, labeled constituent tree for each document. Unlike the Penn Discourse Treebank (PDTB, Miltsakaki et al. 2004), which primarily focuses on local discourse relations and for which more data exists in Chinese, RST builds a document tree using nested relations within a sentence, across sentences, and across paragraphs. RST is thus particularly significant at the macro-level, which is more challenging for understanding discourse organization than at the micro-level (Jia et al., 2018; Hou et al., 2020; Zhang et al., 2020).

Despite the complexity of RST and the human labor required, many new datasets have come out

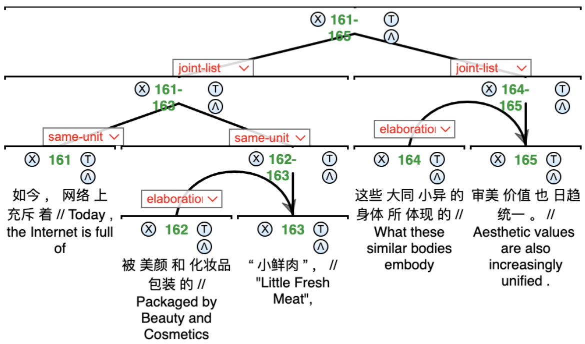

Figure 1: A RST subtree with two relative clauses annotated as *elaboration-attribute* and *same-unit* in `GCDT_academic_dingzhen` with automatic *zh* → *en* translations appended after the source Chinese texts.

in the past decades (Zeldes et al., 2019, 2021), including English (Carlson et al., 2001; Zeldes, 2017), Basque (Iruskieta et al., 2013), Bangla (Das and Stede, 2018), Brazilian Portuguese (Cardoso et al., 2011), Dutch (Redeker et al., 2012), German (Stede and Neumann, 2014), Persian (Shahmohammadi et al., 2021), Russian (Toldova et al., 2017), Spanish (da Cunha et al., 2011), and the Spanish-Chinese parallel corpus (Cao et al., 2018).

However, a substantial gap remains in the availability of document-level hierarchical discourse datasets for non-European languages, particularly Chinese, of sufficient magnitude for training contemporary neural parsers. Aside from the small parallel Spanish-Chinese dataset by Cao et al. (2018, see below) with only 400+ discourse relation instances, there are no available Chinese treebanks in the RST framework. Thus, neither monolingual nor multilingual RST constituent parsers are trained in Chinese and cannot benefit downstream tasks.

In this paper, we present the Georgetown Chinese Discourse Treebank (GCDT) corpus,[1] a new,

[1] The source texts, annotations, and guidelines are open-source (CC-BY) and available at `https://github.com/logan-siyao-peng/GCDT`. The corpus is also searchable in the ANNIS interface (Krause and Zeldes, 2014) at `https://gucorpling.org/annis/#_c=R0NEVA==`.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 382–391
November 20–23, 2022. ©2022 Association for Computational Linguistics

freely available, multi-genre RST corpus of 50 medium to long documents for Mandarin Chinese, as the sample subtree shown in Figure 1. The corpus covers over 60K tokens and 9K Elementary Discourse Units (EDUs). In addition to presenting the SOTA parsing results in monolingual settings for the dataset, we jointly train a model with both English and Chinese datasets, testing finetuning and automatic translation-based approaches to improve performance in both Chinese and English as the target language. Experimenting with different monolingual and multilingual embeddings, we find that joint training and translation improve performance on the smaller Chinese and larger English datasets. However, finetuning only helps with the smaller Chinese data. Finally, we show that monolingual RoBERTa embeddings outperform multilingual embeddings in applicable settings. Still, the best overall performance is achieved using Chinese and English data in a multilingual training regime.

2 Previous Work

RST Datasets in English and Chinese The English RST-DT corpus (Carlson et al., 2001) is the primary benchmark in the RST framework. The large corpus (205K tokens) includes only news articles from the Penn Treebank (Marcus et al., 1993). Another English RST corpus is GUM (Zeldes, 2017), a multi-genre corpus growing in size yearly and currently (V8.0.0) contains 180K tokens from 12 written or spoken genres. GUM is thus slightly smaller in the token count but has a larger number of discourse relation instances due to a shorter average unit length in tokens. Moreover, the dynamic aspect of GUM makes it different to set up benchmark scores compared to other RST corpora. To our knowledge, this paper publishes the first set of RST parsing performances on GUM V8.0.0.

The Spanish-Chinese parallel corpus (Cao et al., 2018) is a small Chinese RST corpus (15K tokens) constructed for translation studies. To support this goal, its EDUs are adjusted to align between Spanish and Chinese rather than staying faithful to the syntax of the individual languages. Its relation inventory is also distinct from inventories used for English corpora, as are the segmentation criteria used in the corpus, limiting its compatibility with other datasets. Another older Chinese RST corpus was reported in Yue (2008) with 97 news commentaries annotated. However, to our knowledge, the dataset is no longer accessible or used in RST parsing or other tasks (Cao, 2018).

Other Hierarchical Chinese Discourse Datasets There are a few other hierarchical discourse corpora in Mandarin Chinese, but none of them annotate single-rooted RST trees for longer documents. The CDT-CDTB corpus (Li et al., 2014b) uses connectives to build up discourse trees only within paragraphs, for 500 news documents from the Chinese Penn Treebank (Xue et al., 2005). Not only are many of the connectives ambiguous in Chinese (Li et al., 2014a; Lu et al., 2018), discourse trees in CDT-CDTB are also small (only 4.5 EDUs/tree). This dataset, therefore, differs substantially from the expected structure of an RST treebank, in which EDUs are expected to be all clauses in the text with functionally motivated relation labels, such as *cause* or *background*.

MCDTB (Jiang et al., 2018) further utilizes a set of discourse relations to connect between paragraphs within 720 documents. The design choice to use specific inter-paragraph-only annotations creates an interesting distinction between micro-level versus macro-level relations (Sporleder and Lascarides, 2004; Wang et al., 2017), but also deviates from RST's fundamental idea of constructing a single tree for an entire document, in which the same inventory of labels is used for all nodes.

Moving beyond constituent-based discourse trees, Cheng and Li (2019) annotated 108 scientific abstracts in their Sci-CDTB corpus using Discourse Dependency Structure (DDS; Hirao et al., 2013; Morey et al., 2018). Cheng et al. (2021) further converted other Chinese discourse corpora into the DDS representation. Even though DDS simplifies parsing and is more similar to other linguistic annotation schemes, such as Universal Dependencies (Nivre et al., 2016) for syntax, the dependency-style discourse annotation loses significant information on the ordering or scope of satellite attachments. For example, whether a unit with *cause* and *attribution* satellites means that both the cause and the result are attributed to someone, as in Appendix C, or that something caused an attributed statement. In other words, when multiple discourse units modify the same nucleus, the relative importance of the satellites and their scopes are ignored.

Multilingual RST Parsers RST parsing is a task that merges a sequence of gold or predicted EDUs and forms a labeled tree structure for the entire document. Since RST datasets share the same unlabeled constituent tree structure, based on the principle that more prominent units should serve as nuclei to less prominent satellite units, multilingual joint training has achieved SOTA results in multilingual RST parsing in several languages. Translating EDUs across languages (Cheng and Li, 2019; Liu et al., 2020, 2021) and mapping word embeddings into the same space (Braud et al., 2017; Iruskieta and Braud, 2019; Liu et al., 2020, 2021) are two common approaches to encoding EDUs across languages in joint training. Among this line of work, Liu et al. (2021) presented a SOTA multilingual RST parser with a pointer-network decoder for top-down depth-first span splitting. The model uses the multilingual *xlm-roberta-base* (Conneau et al., 2020) and trains jointly with six languages: English, Portuguese, Spanish, German, Dutch, and Basque. The current work uses the parser from Liu et al. (2021) for training between the Chinese GCDT corpus and the English GUM corpus.

3 GCDT: Georgetown Chinese Discourse Treebank

GCDT is an open-source multi-genre RST dataset in Mandarin Chinese. Following the design of GUM (Zeldes, 2017), GCDT contains 50 documents, 10 from each of 5 genres which also appear in GUM: academic articles, biographies (*bio*), interview conversations, news, and how-to guides (*whow*), as shown in Table 1. Unlike existing Chinese discourse corpora, GCDT focuses on building larger discourse trees for medium-to-long documents. We select documents with an average of 1K+ tokens to provide more training data for learning higher-level discourse structures.

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.

EDU Segmentation Elementary Discourse Unit (EDU) segmentation is fundamental to RST. We deviate from previous corpora that predominately use potentially ambiguous punctuation (Li et al., 2014a) to segment EDUs, regardless of the surrounding structures. Instead, our Chinese EDU segmentation mirrors the syntactic criteria established in the English RST-DT and GUM corpora (Carlson and Marcu, 2001; Carlson et al., 2001; Zeldes, 2017), largely equating EDUs with the propositional structure of clauses. We use the Penn Chinese Treebank (Xue et al., 2005) as our syntactic guidelines. We first manually tokenize according to Xia (2000b) and conduct EDU segmentation based on parts-of-speech defined in Xia (2000a).

Most notably, we segment relative clauses in GCDT, following the practice in English and other corpora (Carlson et al., 2001; Zeldes, 2017; Das and Stede, 2018; Cardoso et al., 2011; Redeker et al., 2012; Toldova et al., 2017). Chinese relative clauses present a unique feature in the existing RST treebanks. To our knowledge, GCDT is the first RST corpus in any language in which prenominal relative clauses are annotated for discourse relations. Cross-referencing Dryer (2013a,b) with languages of existing RST corpora suggests that only Basque also exhibits the Relative-Noun order found in Chinese. Yet, relative clauses are not segmented in the Basque RST dataset (Iruskieta et al., 2015). Moreover, since relative clauses intervene between Verb-Object in Chinese, the pseudo-relation *same-unit* is used to express discontinuous EDUs, as shown in Figure 1. Segmenting and annotating discourse relations for relative clauses is one of the reasons that GCDT has relatively short EDUs, on average 6.5 tokens/EDU.

Relation Annotation GCDT builds up constituent discourse trees based on gold EDUs using rstWeb (Zeldes, 2016). We use the enhanced two-level relation labels from GUM V8.0.0 with 15 coarse and 32 fine-grained relations (see Appendix A for relation distributions in GCDT and GUM).

Data Split We provide an 8-1-1 `train-dev-test` split per genre to facilitate future RST parsing experiments. Both human inter-annotator agreements and parsing results are assessed on the five `test` documents, with one from each of the five genres.

Inter-Annotator Agreement (IAA) We evaluate agreement on the five `test` documents to obtain human ceiling scores for parser performance. One Chinese native-speaker linguist annotated the entire corpus, and another read the guidelines and conducted independent EDU segmentation. We measured segmentation agreement, adjudicated segmentation between the two annotators, and then separately annotated relation trees on gold EDUs to measure relation agreement. We also release the double annotations in GCDT for future experiments on annotation disagreements. On segmentation, we obtained a token-wise agreement of 97.4% and Cohen's κ =0.89. The agreements on micro-averaged original Parseval F1 of Span, Nuclearity, and Relation are 84.27, 66.15, and 57.77 respectively. The IAA of GCDT is similar to that of the English RST-DT benchmark – 78.7, 66.8, and 57.1 – when evaluated using the original Parseval (Morey et al., 2017). The results show that the GCDT annotation agreement is highly satisfactory even though the documents are much longer and exhibit more genre diversity than RST-DT.

4 Experiments

We present benchmark results on GCDT using the SOTA multilingual parser, DMRST (Liu et al., 2021). Results are shown in two experimental settings: ***monolingual*** training using only one dataset (either Chinese GCDT or English GUM V8.0.0) and ***multilingual*** training using data from both corpora (GCDT+GUM). Besides directly combining corpora from the two languages, we also experiment with finetuning and automatic EDU-wise translation. We use the same set of hyperparameters as reported in Liu et al. (2021). Similarly, we also report monolingual and multilingual parsing performance on GUM V8.0.0.

Datasets Cross-genre adaptability remains a bottleneck in RST parsing (Nishida and Matsumoto, 2022; Atwell et al., 2021). To isolate cross-lingual versus cross-genre influences, we conduct monolingual and multilingual experiments using the following data compositions: 1) **GCDT**: 50 Chinese documents from 5 genres; 2) **GUM-12**: 193 English documents from 12 genres; 3) **GUM-5**: 99 GUM documents from the same 5 genres in GCDT.

Language Models We test different monolingual and multilingual BERT and RoBERTa embeddings (see Appendix B for details).

Metrics We use the 15 coarse relation classes shared between GCDT and GUM and follow the recommendation of Morey et al. (2017) to use the original Parseval micro-averaged F1 for Span, Nuclearity (Nuc), and Relation (Rel).

Multilingual Training Setups In addition to training with combinations of the original GCDT and GUM datasets using multilingual embeddings (see Appendix D for specific data partitions used in the GCDT+GUM-combined experiments), we also experiment with two techniques to improve performances on both target datasets. Specifically, to improve on Chinese GCDT:

1) **Finetuning**: we first train models with both English and Chinese data and then continue training only on the training partition of the target dataset (i.e., GCDT).

2) **Automatic EDU-wise Translation**: we use GoogleTranslator[2] to automatically translate EDUs from the other dataset to the target language (i.e., EDU-wise en→zh translations of GUM) and train on the original GCDT and translated GUM data. The advantage of the translation approach is that we can replace the multilingual embeddings with higher-performing monolingual embeddings.

5 Results

We present monolingual and multilingual results on GCDT and GUM in Tables 2 and 3, as well as genre-wise performance on GCDT in Table 4.

Monolingual Parsing Similar to previous observations (Staliūnaitė and Iacobacci, 2020; Naseer et al., 2021; Tarunesh et al., 2021), Table 2 shows that RoBERTa outperforms BERT in both languages. Monolingual RoBERTa embeddings achieve the best performance when training with monolingual data, e.g., *hfl/chinese-roberta-wwm-ext* obtained 51.76 on the relation level on GCDT.

Multilingual Parsing Our multilingual parsing experiments include joint training, finetuning, and automatic EDU-wise translation. Based on the monolingual results, we use the best-performing multilingual embedding *xlm-roberta-base* (Conneau et al., 2020) with the GCDT+GUM-combined multilingual data. Different aspects of the multilingual parsing results are shown in Table 3.

Firstly, joint training outperformed monolingual results in all three test scenarios: GCDT, GUM-5, and GUM-12. For example, training on GCDT+GUM-12 using XLM RoBERTa achieved an F_Rel of 52.61 on GCDT, higher than the 50.45 trained with only GCDT, and the same embedding.

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

Thirdly, pretraining on the GCDT+GUM-combined training sets and training on the training set of the target corpus improves performance on Chinese GCDT but deteriorates on the English GUM. We hypothesize that with more English training data available, there is less headroom for improvement. In contrast, finetuning for the smaller

[2]https://github.com/nidhaloff/deep-translator

corpus	monolingual embedding	Span	Nuc	Rel	multilingual embedding	Span	Nuc	Rel
GCDT	*bert-base-chinese*	73.15±0.53	55.71±0.66	50.81±0.65	*bert-base-multilingual-cased*	67.34±1.32	47.66±0.73	43.97±0.93
	hfl/chinese-roberta-wwm-ext	**75.51±0.68**	**57.08±0.81**	**51.76±0.97**	***xlm-roberta-base***	**74.35±0.54**	**54.17±1.20**	**50.45±1.09**
GUM-5	*bert-base-cased*	64.61±1.42	49.58±1.51	40.43±1.56	*bert-base-multilingual-cased*	64.52±2.68	51.63±2.07	44.96±1.46
	roberta-base	**73.85±0.70**	**58.95±0.79**	**50.35±1.18**	***xlm-roberta-base***	**72.45±0.97**	**56.78±0.80**	**47.69±0.88**
GUM-12	*bert-base-cased*	60.93±0.63	47.92±0.62	40.20±0.40	*bert-base-multilingual-cased*	64.47±0.50	50.69±0.32	43.25±0.35
	roberta-base	**68.59±0.58**	**55.32±0.27**	**46.29±0.46**	***xlm-roberta-base***	**66.12±0.59**	**52.58±0.52**	**45.06±0.45**

Table 2: Monolingual parsing results on the test sets of GCDT, GUM-5, and GUM-12 with Chinese, English, and multilingual BERT and RoBERTa embeddings (mean±std over five runs).

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**			
joint training w/ XLM RoBERTa	74.24±0.48	56.68±0.86	52.21±0.83	joint training w/ XLM RoBERTa	72.56±0.71	60.63±0.43	52.57±0.77
+finetuning w/ XLM RoBERTa	76.97±0.32	57.94±0.82	53.38±0.51	+finetuning w/ XLM RoBERTa	73.44±0.36	59.40±0.56	50.57±0.97
+en→zh trans. w/ XLM RoBERTa	74.80±0.78	56.58±0.98	51.18±1.15	+zh→en trans. w/ XLM RoBERTa	72.21±1.11	60.07±1.25	52.32±1.05
+en→zh trans. w/ ZH RoBERTa	**77.66±0.42**	**59.29±0.59**	**54.66±0.76**	**+zh→en trans. w/ EN RoBERTa**	**74.73±0.40**	**62.65±0.72**	**54.32±0.82**
Train on GCDT+GUM-12 and Dev/Test on GCDT				**Train on GUM-12+GCDT and Dev/Test on GUM-12**			
joint training w/ XLM RoBERTa	74.33±0.49	57.24±0.99	52.61±1.13	joint training w/ XLM RoBERTa	70.32±0.37	57.49±0.73	49.14±0.34
+finetuning w/ XLM RoBERTa	**76.95±0.65**	**59.40±0.64**	**55.28±0.23**	+finetuning w/ XLM RoBERTa	66.00±0.24	53.13±0.22	45.47±0.42
+en→zh trans. w/ XLM RoBERTa	73.99±0.79	56.31±1.43	51.51±1.34	+zh→en trans. w/ XLM RoBERTa	70.28±0.55	57.63±0.55	49.26±0.39
+en→zh trans. w/ ZH RoBERTa	78.11±0.39	59.42±0.90	54.41±1.23	**+zh→en trans. w/ EN RoBERTa**	**71.41±0.47**	**59.17±0.35**	**50.63±0.48**

Table 3: Multilingual parsing results with finetuning and automatic translation on the test sets of GCDT+GUM combinations with highest-performing Chinese (ZH), English (EN), and multilingual (XLM) RoBERTa embeddings.

Genre	Trained on GCDT			Trained on GCDT+GUM-5 w/ zh→en trans.			Trained on GCDT+GUM-12 w/ zh→en trans.			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 4: GCDT genre-wise performances on sample models trained on GCDT, as well as translation-augmented GCDT+GUM-5 and GCDT+GUM-12 combinations using *hfl/chinese-roberta-wwm-ext*.

Chinese dataset added to the comparatively little information available to the parser.

Lastly, results show that augmenting with automatic translation and using monolingual embeddings achieved the best performance on three of the four test scenarios, while the best result on GCDT was achieved by training together with GUM-12 and finetuning on GCDT.

Genre-wise Analysis We further select three models trained in the monolingual GCDT and translation-augmented scenarios, GCDT+GUM-5 and GCDT+GUM-12, using the Chinese RoBERTa embedding (Cui et al., 2021). Table 4 provides per-genre parsing results of the models on the five `test` genres. On the one hand, the average performance on how-to guides (*whow*) is much higher than *academic* articles for both models and humans. This demonstrates a good human-model alignment regarding which genre is the hardest or easiest (cf. Zeldes and Simonson 2016). On the other, model results are the farthest from the human ceiling scores on the highest performing *whow* genre. We hypothesize that characteristics of genres triggered the different performances. Future multi-genre experiments could be conducted across datasets to study out-of-domain effects in multilingual RST parsing scenarios.

6 Conclusion

This paper presents GCDT, the largest RST dataset for Mandarin Chinese, which closely follows established RST guidelines and is highly comparable to existing English RST corpora. Besides evaluating annotation quality and establishing SOTA results on this dataset in monolingual settings, we also jointly train on GCDT and a similar English corpus—GUM—and demonstrate that multilingual training and automatic EDU translation boost parser performance. However, finetuning is only helpful when targeting the smaller Chinese dataset. We further conduct per-genre analyses and show that parsing performance varies widely between some genres but less between others. We hope that this dataset can alleviate the lack of training resources for hierarchical discourse parsing in Chinese and facilitate multilingual and multi-genre RST parsing, as well as other downstream NLP tasks.

Acknowledgements

We thank Nianwen Xue for providing insights on Chinese syntax, which helped refine the EDU segmentation guidelines. We also thank the anonymous reviewers and Nathan Schneider for their insightful comments.

References

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and Online. Association for Computational Linguistics.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Shuyuan Cao. 2018. *Using annotated discourse information of a RST Spanish-Chinese treebank for translation and language learning tasks*. Ph.D. thesis, Universitat Pompeu Fabra.

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)*, pages 156–166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Paula CF Cardoso, Erick G Maziero, Mara Luca Castro Jorge, Eloize MR Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago AS Pardo. 2011. CSTNews - A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54(2001):56.

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*.

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*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.

Yi Cheng, Sujian Li, and Yueyuan Li. 2021. Unifying discourse resources with dependency framework. In *Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings*, pages 257–267, Berlin, Heidelberg. Springer-Verlag.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

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*, pages 1–10, Portland, Oregon, USA. Association for Computational Linguistics.

Debopam Das and Manfred Stede. 2018. Developing the Bangla RST Discourse Treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer. 2013a. Order of relative clause and noun. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013b. Relationship between the order of object and verb and the order of relative clause and noun. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Naman Goyal and Jacob Eisenstein. 2016. A Joint Model of Rhetorical Discourse Structure and Summarization. In *Proceedings of the Workshop on Structured Prediction for NLP*, pages 25–34, Austin, TX. Association for Computational Linguistics.

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*, pages 1515–1520, Seattle, Washington, USA. Association for Computational Linguistics.

Shengluan Hou, Shuhan Zhang, and Chaoqun Fei. 2020. Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications. *Expert Systems with Applications*, 157:113421.

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*, pages 3046–3052, Online. Association for Computational Linguistics.

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*, pages 185–197, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mikel Iruskieta, Marıa J Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th workshop RST and discourse studies*, pages 40–49.

Mikel Iruskieta and Chloé Braud. 2019. EusDisParser: improving an under-resourced discourse parser with cross-lingual data. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 62–71, Minneapolis, MN. Association for Computational Linguistics.

Mikel Iruskieta, Iria da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*, 49(2):263–309.

Yanyan Jia, Yuan Ye, Yansong Feng, Yuxuan Lai, Rui Yan, and Dongyan Zhao. 2018. Modeling discourse cohesion for discourse parsing via memory network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 438–443, Melbourne, Australia. Association for Computational Linguistics.

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*, pages 3493–3504, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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):65–79.

Thomas Krause and Amir Zeldes. 2014. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014a. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classification system. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 577–587, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. Multilingual neural RST discourse parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Yao-jie Lu, Mu Xu, Chang-xing Wu, De-yi Xiong, Hong-ji Wang, and Jin-song Su. 2018. Cross-lingual implicit discourse relation recognition with co-training. *Frontiers of Information Technology & Electronic Engineering*, 19(5):651–661.

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.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Joscha Markle-Hus, Stefan Feuerriegel, and Helmut Prendinger. 2017. Improving Sentiment Analysis with Document-Level Semantic Relationships from Rhetoric Discourse Structures. Accepted: 2016-12-29T00:30:14Z.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

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*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

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

Muchammad Naseer, Muhamad Asvial, and Riri Fitri Sari. 2021. An Empirical Comparison of BERT, RoBERTa, and Electra for Fact Verification. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 241–246.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-Domain Discourse Dependency Parsing via Bootstrapping: An Empirical Analysis on Its Effectiveness and Limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

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)*, pages 2820–2825, Istanbul, Turkey. European Language Resources Association (ELRA).

Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. Persian rhetorical structure theory. *arXiv preprint arXiv:2106.13833*.

Caroline Sporleder and Alex Lascarides. 2004. Combining hierarchical clustering and machine learning to predict high-level discourse structure. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 43–49, Geneva, Switzerland. COLING.

Ieva Staliūnaitė and Ignacio Iacobacci. 2020. Compositional and lexical semantics in RoBERTa, BERT and DistilBERT: A case study on CoQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7046–7056, Online. Association for Computational Linguistics.

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)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. 2021. Trusting RoBERTa over BERT: Insights from Checklisting the Natural Language Inference Task. *arXiv preprint arXiv:2107.07229*.

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Rhetorical relations markers in Russian RST treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33, Santiago de Compostela, Spain. Association for Computational Linguistics.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

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

Fei Xia. 2000b. The segmentation guidelines for the Penn Chinese Treebank (3.0).

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*, pages 124–134, Online. Association for Computational Linguistics.

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*, pages 5021–5031, Online. Association for Computational Linguistics.

Nianwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Nat. Lang. Eng.*, 11(2):207–238.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1839, Doha, Qatar. Association for Computational Linguistics.

Ming Yue. 2008. Rhetorical structure annotation of Chinese news commentaries. *Journal of Chinese Information Processing*, 22(4):19–23.

Amir Zeldes. 2016. rstWeb - a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, San Diego, California. Association for Computational Linguistics.

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

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN. Association for Computational Linguistics.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amir Zeldes and Dan Simonson. 2016. Different flavors of GUM: Evaluating genre and sentence type effects on multilayer corpus annotation quality. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW X)*, pages 68–78, Berlin.

Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou. 2020. A top-down neural architecture towards text-level parsing of discourse rhetorical structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6386–6395, Online. Association for Computational Linguistics.

A Label Distributions

Table 5 gives descriptive statistics of the distribution of relations in GCDT, as well as numbers for comparison from the GUM corpus, which uses the same inventory of relations and covers all and more genres in the GCDT dataset.

Relation Name	GCDT%	GUM%
Nucleus-Satellite Relations		
elaboration-attribute	7.71%	4.60%
attribution-positive	4.37%	3.08%
elaboration-additional	4.25%	4.81%
explanation-evidence	4.15%	2.08%
context-background	2.89%	2.66%
context-circumstance	2.69%	2.40%
organization-preparation	2.36%	1.83%
causal-cause	1.98%	1.63%
organization-heading	1.78%	1.49%
contingency-condition	1.77%	1.67%
adversative-concession	1.68%	2.04%
purpose-goal	1.54%	1.63%
restatement-partial	1.32%	1.13%
evaluation-comment	1.15%	2.29%
mode-means	1.09%	0.55%
explanation-justify	0.87%	1.60%
causal-result	0.87%	1.54%
adversative-antithesis	0.58%	1.47%
mode-manner	0.52%	0.89%
topic-question	0.44%	1.10%
organization-phatic	0.27%	1.37%
attribution-negative	0.23%	0.30%
purpose-attribute	0.21%	0.87%
explanation-motivation	0.2%	0.71%
topic-solutionhood	0.01%	0.20%
Multi-Nucleus Relations		
joint-list	22.28%	12.90%
same-unit	18.69%	4.71%
joint-sequence	4.99%	4.41%
joint-other	4.83%	4.48%
adversative-contrast	3.32%	2.40%
joint-disjunction	0.64%	1.13%
restatement-repetition	0.32%	1.82%

Table 5: Distribution of 32 relations (15 classes, including *same-unit*) in GCDT and GUM V8.0.0.

B Specific PLMs Used in the Experiments

Table 6 shows the Chinese, English, and multilingual BERT and RoBERTa pretrained language models used in the experiments described in §4.

Type	Details
BERT	Chinese: *bert-base-chinese* (Devlin et al., 2019)
	English: *bert-base-cased* (Devlin et al., 2019)
	Multilingual: *bert-base-multilingual-cased* (Devlin et al., 2019)
RoBERTa	Chinese: *hfl/chinese-roberta-wwm-ext* (Cui et al., 2021)
	English: *roberta-base* (Liu et al., 2019)
	Multilingual: *xlm-roberta-base* (Conneau et al., 2020)

Table 6: An overview of pretrained BERT and RoBERTa language models used in the experiments.

C A Fragment of RST Annotation in GCDT

Figure 2 presents a relation hierarchy of *attribution-positive* scoping over *causal-cause*.

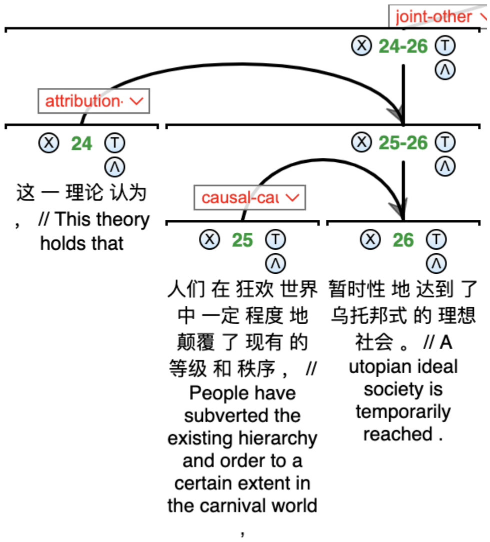

Figure 2: A RST subtree with *attribution-positive* scoping over *causal-cause* from `GCDT_academic_dingzhen` with automatic *zh* → *en* translation.

D Data Splits for Multilingual Experiments

Table 7 presents the `train/dev/test` splits when jointly training with GCDT and GUM in multilingual experiments.

	`train`: GCDT+GUM `dev/test`: GUM	`train`: GCDT+GUM `dev/test`: GCDT
`train`	GUM-train + GCDT-train + GCDT-dev	GCDT-train + GUM-train + GUM-dev
`dev`	GUM-dev	GCDT-dev
`test`	GUM-test	GCDT-test

Table 7: An overview of the `train/dev/test` splits of GCDT and GUM used for training in the multilingual experiments.