Why Can't Discourse Parsing Generalize? A Thorough Investigation of the Impact of Data Diversity

Yang Janet Liu and Amir Zeldes
Department of Linguistics, Georgetown University
{yl879, amir.zeldes}@georgetown.edu

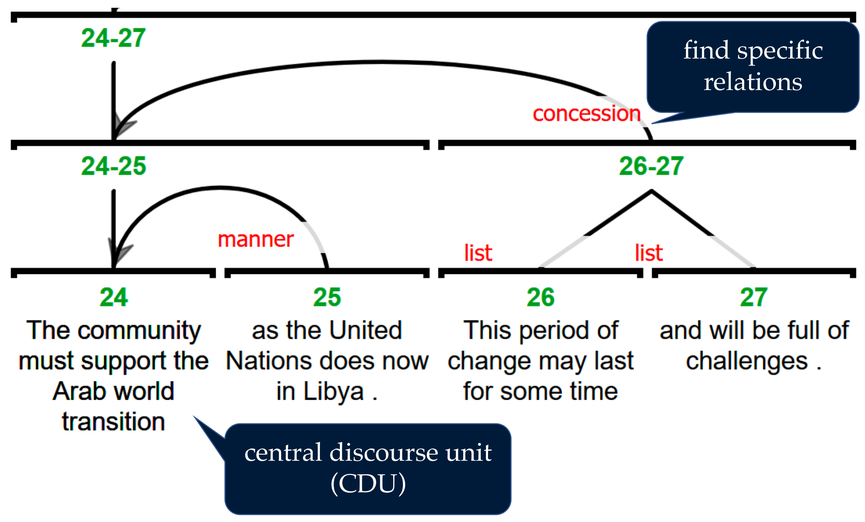
GEORGETOWN UNIVERSITY

SCAN ME

EACL 2023 @ Dubrovnik, Croatia

RST Parsing & Generalizability

- Discourse parsing is the task of identifying and classifying the coherence relations that hold between different parts of a text.
- Rhetorical Structure Theory (RST, Mann and Thompson 1989) is a functional theory of text organization that constructs hierarchical structures in text, which have numerous applications.
- This study is the first to fully evaluate cross-genre RST parsing generalizability on complete trees in datasets with the same EDU segmentation.
- Overall, we find that **diverse training data leads to better generalization on unseen genres regardless of model architecture**. RST parsing work should devote more attention to **multi-genre corpora as benchmarks**.



English RST Corpora

RST Discourse Treebank (RST-DT, Carlson et al. 2003)

the standard English RST benchmark, with data from the 1989 Wall Street Journal (WSJ) section of the Penn Treebank (PTB, Marcus et al. 1993)

GUM (Zeldes, 2017)

- a multi-genre corpus covering 12 written and spoken genres
- continuously growing, with new data added in each version
- for this paper: GUM v8

Evaluation Metrics

Original Parseval eval scores on binary trees with gold EDU segmentations, following the recommendation of Morey et al (2017), on the following levels:

- **S**pan: whether subtrees span the right EDUs
- **N**uclearity: whether edges point the right way
- **R**elation: whether labels are correct

Experiments & Results & Findings

Exp1: Cross-Corpus Generalization (RST-DT & GUM)

hypothesis: since GUM contains many genres, models trained on it will degrade less when testing on RST-DT than in the opposite scenario

parser 1: Guz and Carenini (2020, BOTTOM-UP) ✅
parser 2: Liu et al. (2021, TOP-DOWN)

setup: train the parsers on the TRAIN partition of each dataset and report scores on the TEST set

train	test	S	N	R
RST-DT	RST-DT	76.5	65.9	54.8
	GUM	65.3 (-11.2)	49.5 (-16.4)	–
	GUM *news*	71.0 (-5.5)	57.5 (-8.4)	–
GUM	GUM	69.9	57.0	48.5
	RST-DT	72.7 (+2.8)	57.4 (+0.4)	–
	GUM *news*	71.6	58.5	49.5

Table 3: Cross-Corpus Results (5 run average) of the BOTTOM-UP Parser from Guz and Carenini (2020).

train	test	S	N	R
RST-DT	RST-DT	76.5	65.2	54.2
	GUM	66.2 (-10.3)	50.8 (-14.4)	–
	GUM *news*	67.9 (-8.6)	55.8 (-9.4)	–
GUM	GUM	68.6	54.9	46.1
	RST-DT	71.1 (+2.5)	55.9 (+1.0)	–
	GUM *news*	73.4	63.3	57.2

Table 4: Cross-Corpus Results (5 run average) of the TOP-DOWN Parser from Liu et al. (2021).

- both parsers show a very significant degradation when training on RST-DT to parse OOD data from GUM.
- by contrast, the GUM-trained model actually scores better on RST-DT than on GUM.

Exp2: Joint Training (RST-DT)

1) Simple Concatenation (**CONCAT**)

2) Model Stacking
   - **FLAIR-LABEL**: train an LSTM using FLAIR (Akbik et al., 2019) to predict EDU dependency labels
   - **SR-LABEL**: train a full shift-reduce parser on GUM, generate predictions for RST-DT in the GUM scheme, and collapse such labels into dependencies
   - **SR-GRAPH**: featurize each EDU's predicted dependency attachment direction and EDU distance to the parent EDU

3) PLM Fine-tuning (**SR-FT**): fine-tune SpanBERT on full parsing of GUM

Findings

1. All scenarios except for SR-FT are virtually equivalent to training on RST-DT alone, suggesting that added features are more distracting than helpful.
2. Complex global structure and pragmatic inferences still cause errors not prevented by more genres with different vocabulary

	S	N	R	*architecture*
Zhang et al. (2021)*	76.3	65.5	**55.6**	TOP-DOWN
Liu et al. (2021)♢	**76.5**	65.2	54.2	TOP-DOWN
Guz and Carenini (2020)♢	**76.5**	**65.9**	54.8	BOTTOM-UP
this paper (CONCAT)♠	75.9	64.8	54.1	
this paper (FLAIR-LABEL)♠	75.8	65.6	55.3	
this paper (SR-LABEL)♠	76.2	66.0	55.3	BOTTOM-UP
this paper (SR-GRAPH)♠	75.8	65.5	54.7	
this paper (SR-FT)♢	**76.3**	**66.2**	**55.5**	
Human (Morey et al., 2017)	78.7	66.8	57.1	–

Table 5: Joint Training Performance on RST-DT. * = original paper score. ♢ = 5 run avg.; ♠ = 3 run avg.

Exp3: OOD Multi-Genre Degradation (GUM)

RQ: how badly a multi-genre trained model will degrade on unseen genres, when the annotation scheme remains identical?

1. to explore OOD degradation, we conducted 10 experiments, comparing the normal genre-balanced scenario (GUM-test) with testing on each genre when it is not in "train" (one-vs-all, OVA)
2. since data for the smaller 4 growing genres may be less reliable and non-comparable, we separately report scores for training on all 8 large genres (ALL-LARGE), tested on each of the four growing genres: *conversation, speech, textbook, vlog*

The *degradation* column shows that the parser suffers when a genre is removed from training across the board, except for *news* and the Span level of *reddit*, suggesting that collecting more news data may not be a priority.

(See section 3.3 for more discussion. We also conducted a thorough error analysis on the worst performing genre, how-to guides, and categorized errors in section 4.)

	GUM test			ova			*degradation*		
non-growing	S	N	R	S	N	R	S	N	R
academic	77.0	68.5	59.8	75.2	66.2	55.7	1.7	2.3	4.1
bio	70.4	58.2	51.2	68.8	53.9	43.2	1.6	4.3	**8.0**
fiction	66.3	53.1	43.7	64.5	50.1	42.1	1.8	3.0	1.7
interview	73.3	59.0	50.9	73.0	56.7	49.7	0.3	2.2	1.2
news	71.7	58.4	49.1	72.2	59.2	51.3	-0.5	-0.8	-2.2
reddit	66.0	52.3	44.2	66.6	51.9	43.3	0.6	0.4	0.8
voyage	78.3	62.1	51.8	77.4	59.7	49.3	0.9	2.4	2.4
how-to	76.5	63.6	54.6	67.1	54.3	44.8	**9.3**	**9.3**	**9.9**
	GUM test			ALL-LARGE			*degradation*		
growing	S	N	R	S	N	R	S	N	R
conversation	45.4	34.5	26.7	42.7	31.4	21.8	2.7	3.1	**4.9**
speech	76.0	64.4	55.2	76.4	62.9	54.8	-0.4	1.5	0.4
textbook	77.4	66.8	57.3	76.2	64.3	54.5	1.2	2.6	2.9
vlog	64.8	49.0	42.8	63.3	49.0	40.4	1.5	0.0	2.5

Table 6: Per Genre Scores for GUM test vs. the OVA or ALL-LARGE Experiments (3 run average).

Exp4: Genre Variety in a Fixed-Size Sample (GUM)

hypothesis: ideally, we want to compare scores on a fixed OOD test set for equal-sized training corpora, divided into fewer or more genres

- If there are not enough recurring examples of infrequent phenomena, because data is so diverse, learning might fail due to sparseness
  - If having too many small genres is harmful, we expect cohort 3 (C3) to perform worst;
  - By contrast, if diversity is helpful, C3 should perform best.

ID	genres	docs	EDUs	ID	genres	docs	EDUs
C1	*academic*	18	1,970	C3	*academic*	9	1,004
	bio	19	1,981		*bio*	9	930
	news	23	1,760		*news*	10	635
	total	60	5,711				
C2	*fiction*	15	1,941		*fiction*	8	1,027
	interview	15	1,931		*interview*	8	1,199
	how-to	15	1,840		*how-to*	8	917
	total	45	5,712		**total**	52	5,712

Table 7: Composition of 3 Fixed-Size Training Cohorts with Different Genre Contents.

	C1			C2			C3			C3−C1			C3−C2			mean_C3_gain		
test	S	N	R	S	N	R	S	N	R	S	N	R	S	N	R	S	N	R
conversation	34.8	23.4	13.9	40.3	27.9	18.0	37.9	26.4	18.0	3.0	3.0	4.1	-2.5	-1.5	0.0	0.3	0.7	2.0
reddit	60.3	45.3	36.0	63.5	46.9	37.6	61.8	47.6	37.3	1.5	2.3	1.4	-1.7	0.7	-0.3	-0.1	1.5	0.6
speech	72.5	58.2	46.9	72.6	59.3	47.7	71.6	57.1	48.0	-0.9	-1.1	1.1	-1.0	-2.1	0.3	-0.9	-1.6	0.7
textbook	73.6	59.0	48.9	70.9	55.0	45.6	74.0	60.5	51.4	0.5	1.5	2.5	3.1	5.5	5.9	1.8	3.5	4.2
vlog	57.8	41.3	35.0	58.8	44.5	35.3	57.7	43.4	34.8	-0.1	2.1	-0.2	-1.1	-1.1	-0.5	-0.6	0.5	-0.3
voyage	76.6	58.1	47.5	76.5	57.4	46.4	78.0	59.1	50.2	1.5	1.0	2.7	1.6	1.7	3.8	1.5	1.4	3.3
macro_avg	62.6	47.6	38.0	**63.8**	48.5	38.4	63.5	**49.0**	**40.0**	0.9	1.5	1.9	-0.3	0.5	1.5	0.3	1.0	1.7
micro_avg	58.7	44.2	34.8	**60.5**	45.7	35.7	59.8	**45.9**	**36.9**	1.1	1.7	2.1	-0.6	0.2	1.2	0.2	1.0	1.6

Table 8: Performance of 3 Fixed-Size Train Cohorts with Different Genre Contents (5 run average).

Analysis & Discussion & Takeaways

1. Training on multiple genres, each with comparatively fewer documents, can lead to good performance with only minor degradation on the very narrow WSJ domain from RST-DT.
2. Adding a second dataset for joint training creates **a "break-even" effect**: the benefit of more data helps about as much as the disparate domains harm within-corpus performance.
3. Errors are skewed by genre: 1) **Evaluation** is problematic in *fiction* and *interview*, 2) **Explanation** and **Organization** are surprisingly hard to predict in 3 genres each.
4. For CDU detection, which can benefit summarization or long-form QA systems, in the cross-corpus setting, an RST-DT trained model captures only a single GUM CDU correctly (acc=0.042 vs. 0.375 for a GUM-trained model); scores on RST-DT are much higher: acc=0.842 for SR-FT trained on RST-DT vs. 0.553 for a GUM-trained model.
5. More training genres with smaller portions each promotes OOD generalization, and development of more diverse multi-genre data should take priority over building up material in existing genres to promote generalizable parsing.

References

- Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.
- Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167, Online. Association for 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.
- Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3): 581–612.