

# GDTB: Genre Diverse Data for English Shallow Discourse Parsing across Modalities, Text Types, and Domains





Yang Janet Liu<sup>2,3,†\*</sup>, Tatsuya Aoyama<sup>1\*</sup>, Wesley Scivetti<sup>1\*</sup>, Yilun Zhu<sup>1\*</sup>, Shabnam Behzad<sup>1</sup>, Lauren Elizabeth Levine<sup>1</sup>, Jessica Lin<sup>1</sup>, Devika Tiwari<sup>1</sup>, Amir Zeldes<sup>1</sup>

<sup>2</sup>MaiNLP, Center for Information and Language Processing, LMU Munich, Germany

<sup>3</sup>Munich Center for Machine Learning (MCML)



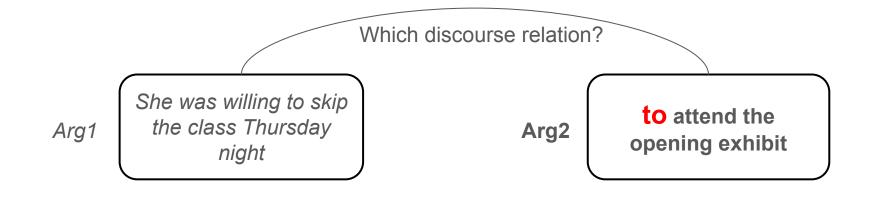
# Introduction

• Understanding discourse relations is essential for exploring the structure of natural languages, for model pre-training and advancing NLP tasks, including discourse parsing and other applications

<sup>1</sup>Corpling Lab, Georgetown University

- Discourse Relation Frameworks: **PDTB** (Prasad et al. 2014), **RST** (Mann and Thompson 1988), SDRT (Asher and Lascarides 2003) etc.
- Existing datasets for training shallow discourse parsing systems, like PDTB-3, lack diversity of domains (limited to newswire)
- We present a new high-quality, PDTB-style benchmark GDTB based on the GUM corpus with 16 diverse genres, valuable resource for out-of-domain PDTB-style shallow discourse parsing

The Shallow Discourse Parsing Task



## **►** GDTB: The Benchmark

	GDTB	PDTB v3
Tokens	228,399	1,156,308
Docs	235	2,161
Genres	16	1
AltLex	224	1,498
AltLexC	13	140
EntRel	553	5,538
Explicit	7,202	24,238
Hypophora	465	146
Implicit	4,503	21,781
Norel	662	287
All	13,622	53,628

Table 1: Relation Type Counts: GDTB vs. PDTB v3.

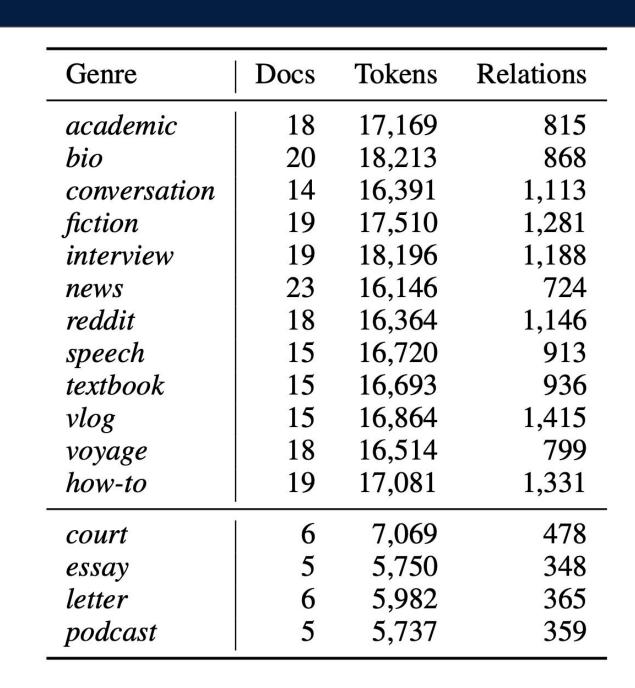


Table 2: Genre Breakdown for GDTB. The bottom four 'growing' genres are still being collected for GUM and counts represent sizes as of GUM v10.

# **►** Construction of GDTB

Steps
Sense Mapping
Explicit Module
Implicit Module
AltLex Module
AltLexC Module
Hypophora Module
EntRel Module
Argument Span Module

Relation Scores (exact label and span match)				
type	P	R	<b>F1</b>	
altLex	0.9500	0.7600	0.8444	
altLexC	1.0000	1.0000	1.0000	
<b>EntRel</b>	0.7593	0.8913	0.8200	
<b>Explicit</b>	0.9812	0.9874	0.9843	
Hypophora	0.8750	0.8537	0.8642	
<b>Implicit</b>	0.8784	0.8205	0.8485	
NoRel	0.7887	0.9180	0.8485	
micro-avg.	0.9277	0.9161	0.9218	

micro-avg.	0.9277	0.9161	0.9218	
Span Scores (incl. relation type but not sense)				
altLex	0.9500	0.7600	0.8444	
altLexC	1.0000	1.0000	1.0000	
<b>EntRel</b>	0.7778	0.9130	0.8400	
<b>Explicit</b>	0.9935	1.0000	0.9967	
Hypophora	0.8750	0.8537	0.8642	
<b>Implicit</b>	0.9824	0.9176	0.9489	
NoRel	0.7887	0.9180	0.8485	
micro-avg.	0.9678	0.9554	0.9616	
	<u> </u>			

Table 3: Test Set Accuracy (manual correction).

#### **Dataset Conversion**

- Use GUM v10 and its multi-layer annotations (gold syntax, coreference, eRST)
- Process
  - Explicit, implicit, AltLex, AltLexC on the right
  - Hypophora Module: generated from each RST
     TOPIC-QUESTION relation
  - EntRel Module: If no relation specified for two adjacent sentences, ENTREL for coreference in elaborative relations, otherwise NOREL
  - Argument Span Module: align target and source
     EDU spans to PDTB-style

### **Quality Evaluation**

- System outputs vs. manually corrected test set (1531 rels)
- Two scenarios: exact match & span-only match
- micro-F1 score of 92 for overall quality, above human agreement scores in previous research
- Argument spans are relatively reliable compared to sense prediction, especially for implicit cases

#### Sense Mapping **Explicit Module** She was willing to skip PDTB v3 Relations contingency.cause.resul the class Thursday eRST Relations contingency.cause.reaso Contingency. JOINT-LIST contingency.condition purpose CAUSAL-CAUSE SAME-UNIT TOPIC-QUESTION to attend the She was willing to in order to opening exhibit skip the class attend the opening (Rank using DiscoDisco, Gessler Thursday night exhibit et al. 2020) **GDTB** eRST **Implicit Module** RESTATEMENT-REPETITION She couldn't care about tiny details Connective Prediction (flan-t5 large FT) Her mind was on Expansion. She couldn't care Equivalence the broader about tiny details in other words pictures — Her mind was on the in other words broader pictures — Connective Prediction model eRST **GDTB** Sense Mapping **AltLex Module** PDTB v3 Relations the number of transactions will be eRST Relations <u>contingency.cause.result</u> CAUSAL-RESULT \_\_\_\_ contingency.cause.reaso CAUSAL-RESULT Contingency. JOINT-LIST contingency.purpose cause.result PURPOSE-GOAL SAME-UNIT which will result in TOPIC-QUESTION which will result in unfavorable economic the number of performance. transactions will be unfavorable economic performance. eRST **AltLexC Module** Rule-based Mapping Had it happened five hours earlier or four hours earlier, eRST Relations CONTINGENCY-CONDITION Contingency. condition PDTB v3 Relations I think the death toll Auxiliary Inversion Had it happened I think the death toll contingency.condition would have been

# Experiments & Results

- 3 setups: within-corpus, cross-corpus, joint-training
- Results
  - o cross-corpus degradation observed, especially for **implicit** relations
  - Best-performing genres for each model are news and academic the worst-performing genre is court

	Test Set		
Training	GDTB	PDTB v3	
within-corpus	0.6447	0.7572	
cross-corpus	0.5660	0.4457	
joint-training	0.6440	0.7390	

Table 4: Overall Accuracy Scores (within-corpus=train set is from the corpus of the test set; cross-corpus=train set from opposite corpus; joint=train on both).

Train	Test	Explicit	Implicit	altLex	altLexC	Hypophora
GDTB	GDTB	0.7645	0.4579	0.4400	1	0.8780
	PDTB v3	0.6114	0.2842	0.3333	0.5000	0.7500
PDTB v3	GDTB	0.6794	0.4048	0.3600	1	0.5854
IDID VS	PDTB v3	0.8817	0.6020	0.8986	0.9167	0.8750
GDTB &	GDTB	0.7374	0.4908	0.4400	1	0.9512
PDTB v3	PDTB v3	0.8679	0.5683	0.8261	0.8333	0.8750

Table 5: Accuracy by Relation Types.

	<b>GDTB-trained</b>	PDTB-trained	joint-training
TED-MDB (English)	0.5214	0.5556	0.5641

Table 6: Accuracy Scores of TED-MDB (English).

overall explicit implicit

1.00

0.75

0.50

0.25

0.00

academic discontinuous contraction could be seen existent existent exist get get exist get exist get exist get exist get exist get exist get get exist get exist get exist get exist get exist get exist get get exist get exist get exist get exist get exist get exist get

0.75
0.50
0.25
0.00

Saddenic bo out esset kitcherien ette kens patret kett kens patret kent kens patret kens patret kens patret kent kens patret kens patr

Figure 2: GDTB Scores by Genres and Relation Types.

(b) cross-corpus model.

# **►** Conclusion

eRST

would have been

more than a thousand

- Introducing GDTB, a valuable, high-quality PDTB-style dataset covering 16 English spoken and written genres for open-domain shallow discourse parsing
- Demonstrate reliable conversion from RST relations to PDTB-style annotations
- Cross-corpus experiments reveal PDTB's current inadequacy for relation classification in open domain settings



five hours earlier or

four hours earlier.

- Extend to the RST-PDTB conversion for other resources
- Contribute to theoretical studies of
  - discourse relation variation across genres
- the comparison of alignments between PDTB & RST/eRST



Paper

more than a thousand

**GDTB** 





