Discourse Relation Prediction and Discourse Parsing in Dialogues with Minimal Supervision

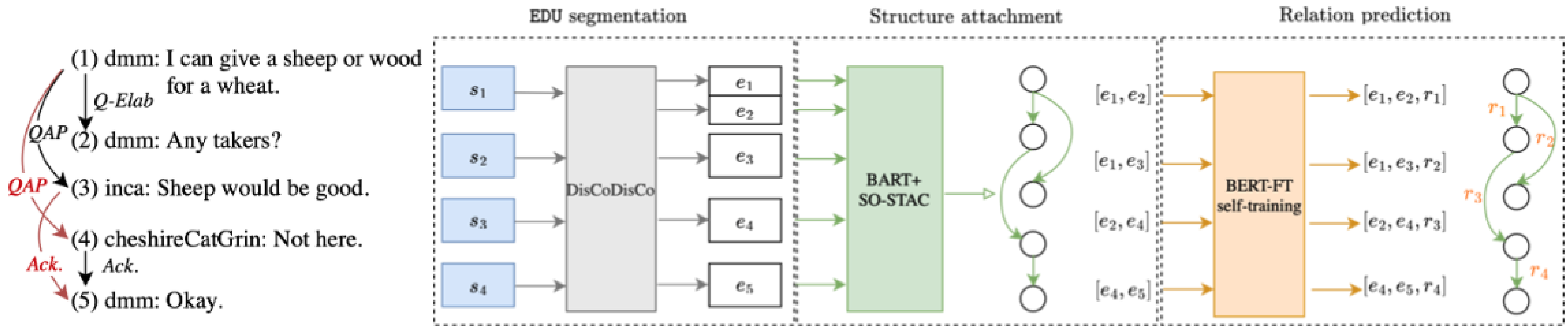
Chuyuan Li[1], Chloé Braud[2], Maxime Amblard[3], Giuseppe Carenini[1]

[1]{cli2711, carenini}@cs.ubc.ca, [2]chloe.braud@irit.fr, [3]maxime.amblard@loria.fr

Context & Motivation

- Explosion of online **dialogue data** brings increasing need for automatic dialogue analysis systems
- Simple surface-level features are not sufficient, we need semantic & pragmatic information, for instance **discourse analysis**
- However, discourse analysis faces **data scarcity**, e.g., SDRT-framework [2] annotated STAC corpus [1] $\approx 10k$ elementary discourse units (EDUs)
- **Semi-supervised** approaches: leverage information from PLMs for structure extraction [5]; self-training techniques on monologues [6]
- Focus: discourse relation prediction in dialogues + integration into a full pipeline: EDU segmentation → structure attachment → relation prediction

Pipeline Design



Left: a dialogue example. Nodes are EDUs; edges are relations. Right: *structure-then-relation* pipeline. s are speech turns; e are EDUs; r are relations.

1. EDU segmentation: Off-the-shelf segmenter DisCoDisCo [4], achieves an F_1 score of 94.8%.

2. Structure attachment with fine-tuned BART [5]

- Sentence-Ordering (SO) pre-training task to enhance pair-wise, inter-speech turns, and inter-speaker discourse information in BART
- Attention matrices are regarded as fully connected graphs, Maximum Spanning Tree algorithm is used to extract dependency structures
- Examine each attention matrix individually in BART encoder, use a small set of annotated dialogues to locate the best attention matrices

3. Relation prediction with BERT and self-training

- Classifier $\mathcal{M}$: fine-tuned BERT, input follows Next Sentence Prediction pattern: [CLS] EDU_1 [SEP] EDU_2
- Sample selection criteria: (a) top-k: top k pseudo-labeled data (b) top-class-k: most confident pseudo-labeled data in each class and together results in k examples so that the label ratio is maintained; $k \in [800, 1800, ...7800]$
- Iterative training: $\mathcal{M}$ is trained iteratively with the combination of 700 pairs of gold annotated data and k augmented pseudo-labeled data

Relation Prediction Results (left) & Full Parsing in-domain and cross-domain Results (right)

Majority class				27.1
BERT (base 700)				$40.1_{0.8}$
BERT-ft (base 700)				$56.6_{1.0}$
Self-training	Top-k	Top-class-k		
#Pair	loop1	loop1	loop2	loop3
+ 800	$54.1_{3.0}$	$57.7_{1.1}$	$55.9_{1.1}$	$\underline{\mathbf{58.1_{1.2}}}$
+ 1800	$53.6_{3.6}$	$57.3_{1.6}$	$\underline{\mathbf{58.4_{1.0}}}$	$57.4_{2.1}$
+ 2800	$55.7_{1.9}$	$57.6_{0.3}$	$57.5_{1.5}$	$\underline{58.1_{2.2}}$
+ 3800	$56.6_{2.1}$	$\underline{57.6_{1.6}}$	-	-
+ 4800	$56.8_{0.5}$	$\underline{57.8_{1.2}}$	-	-
+ 5800	$\underline{\mathbf{58.1_{0.8}}}$	$\mathbf{58.0_{0.7}}$	-	-
+ 6800	$57.8_{1.0}$	$57.9_{0.9}$	-	-
+ 7800	$\underline{57.8_{0.7}}$	$57.0_{2.3}$	-	-

- top-class-k (*vs.* top-k) selection consistently brings improvement; iterative training helps especially for *infrequent* relation types

Train / Test	Train	STAC/STAC				STAC/Molweni-clean		
	#Doc	EDU	Link	Rel	Full	Link	Rel	Full
Structured joint [3]	947	-	$70.7_{0.5}$	$77.3_{1.2}$	$54.6_{0.7}$	$61.5_{3.4}$	$59.5_{4.3}$	$36.6_{3.8}$
Structured joint	50	-	$55.1_{3.5}$	$61.1_{2.1}$	$33.6_{2.2}$	$51.1_{6.4}$	$33.6_{9.5}$	$17.2_{5.3}$
Arc-factored	50	-	$42.7_{2.8}$	$56.4_{2.5}$	$24.0_{1.0}$	$53.7_{2.1}$	$38.8_{2.9}$	$20.9_{1.1}$
$\text{GPT3.5}_{\text{few shot}}$	3	-	20.7	24.1	7.3	-	-	-
$\text{GPT3.5}_{\text{zero shot}}$	-	-	20.0	22.8	4.4	-	-	-
Ours (gold EDU)	50	-	$\mathbf{59.3_{0.7}}$	$\mathbf{62.0_{1.1}}$	$\mathbf{38.6_{0.7}}$	$\mathbf{75.6_{0.7}}$	$\mathbf{41.3_{3.8}}$	$\mathbf{31.2_{2.9}}$
Ours (pred EDU)	50	94.8	$52.2_{0.4}$	$61.2_{1.6}$	$32.8_{0.9}$	∼	∼	∼

- In-domain (board game): our pipeline largely outperforms SOTA supervised parser SJ in link attachment (+4%), relation prediction (+1%), and full parsing (+5%)
- Cross-domain (board game → Ubuntu chat): superior performance compared to SJ: link (+24%), relation (+8%), and full parsing (+14%)

Conclusion

- Versatile pipeline for sequentially addressing all tasks in discourse parsing
- Strong performance in both in-domain and cross-domain settings
- **Future work**: improve relation prediction by using more out-of-domain raw data in self-training; evaluate the pipeline on spoken dialogue data and test on other discourse frameworks

References

[1] N. Asher and et al. the STAC corpus. In *LREC*, 2016.

[2] N. Asher and A. Lascarides. *Logics of conversation.* Cambridge University Press, 2003.

[3] T. Chi and A. Rudnicky. Structured dialogue discourse parsing. In *SIGDial*, 2022.

[4] L. Gessler and et al. Discodisco at the disrpt2021 shared task. In *DISRPT*, 2021.

[5] C. Li and et al. Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In *EACL Findings*, 2023.

[6] N. Nishida and Y. Matsumoto. Out-of-domain discourse dependency parsing via bootstrapping. *TACL*, 2022.