A Semi-supervised Dialogue Discourse Parsing Pipeline

Chuyuan Li, Maxime Amblard, Chloé Braud

A Semi-supervised Dialogue Discourse Parsing Pipeline

Chuyuan Li[1] Maxime Amblard[1] Chloé Braud[2]
(1) LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
lisa27chuyuan@gmail.com, maxime.amblard@loria.fr
(2) IRIT, Université de Toulouse, CNRS, ANITI, Toulouse, France
chloe.braud@irit.fr

RÉSUMÉ

Analyse semi-supervisée du discours dans les dialogues

L'analyse du discours joue un rôle crucial dans le Traitement Automatique des Langues (TAL) et a démontré son utilité dans diverses applications telles que le résumé et les systèmes de questions-réponses. Dans cet article, nous abordons ce problème difficile en raison de la rareté des données annotées : l'analyse du discours dans les dialogues. Notre approche de l'analyse du discours comporte deux étapes : tout d'abord, nous prédisons la structure du discours, puis nous identifions les relations au sein de la structure. En utilisant seulement 50 exemples comme données d'entraînement, nos méthodes obtiennent des résultats compétitifs par rapport à l'état de l'art supervisé dans le même domaine et de bien meilleures performances inter-domaines, avec également une meilleure stabilité.

ABSTRACT

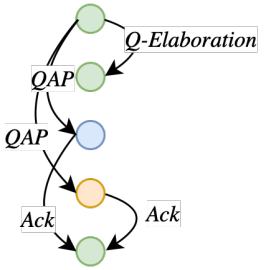
Discourse analysis plays a crucial role in Natural Language Processing (NLP) and has demonstrated its usefulness in various downstream applications like summarization and question answering. In this work, we study discourse in dialogues : an under-explored setting due to significant data scarcity challenge. We conduct discourse parsing within a pipeline : first, we predict the discourse structure, and then we identify the relations within the structure. Using only 50 examples as gold training data, our methods achieve competitive results compared to supervised state-of-the-art in-domain and much stronger performance cross-domain, with also better stability.

MOTS-CLÉS : Analyse du discours, apprentissage automatique, dialogue.

KEYWORDS: Discourse analysis, machine learning, dialogue.

1 Introduction

Discourse analysis aims to uncover the inherent structure of documents and has been shown useful for many applications, from sentiment analysis or fake news detection (Bhatia *et al.*, 2015; Karimi & Tang, 2019), to summarization or machine translation (Chen & Yang,



e_1 dmm: I can give a sheep or wood for a wheat.

e_2 dmm: Any takers?

e_3 inca: Sheep would be good.

e_4 cheshireCatGrin: Not here.

e_5 dmm: Okay.

FIGURE 1 – An SDRT graph structure (left) of a dialogue (right) in STAC corpus (*s2-leagueM-game4*). e are EDUs. QAP : *question answer pair* ; Ack : *acknowledgment* ; Q-Elaboration : *question elaboration*. Graphic extracted from Li (2023).

2021; Chen *et al.*, 2020). In recent years, the availability of accurate transcription methods and the increase in online communication have led to a tremendous rise in dialogue data, necessitating the development of automatic analysis systems. However, simple surface-level features are oftentimes not sufficient to extract valuable information from conversations (Qin *et al.*, 2017). Rather, we need to understand the semantic and pragmatic relationships organizing the dialogue, for example through the use of discourse information.

Discourse parsing task consists of retrieving a structure from documents, where spans of text are linked by semantic-pragmatic relations (such as *Explanation, Acknowledgment, Contrast...*). It is a hard task, with low performance especially for multi-party dialogues involving intricate relations between speakers. Hence, on the English chat corpus STAC (Asher *et al.*, 2016) (board game) – annotated under the Segmented Discourse Representation framework (Asher & Lascarides, 2003) with graph structures and 16 relations –, State-Of-The-Art (SOTA) supervised parser Structured-Joint (Chi & Rudnicky, 2022) reports 59.6 at best on the full structure, with a drop of about 20 points for cross-domain when testing on the Molweni corpus (Li *et al.*, 2020) (Ubuntu forum). A main challenge in discourse parsing is data scarcity, along with limitations of supervised approaches in cross-domain scenarios (Liu & Chen, 2021) or incomplete parsing that overlooks the important relation information (Badene *et al.*, 2019; Huber & Carenini, 2022; Li *et al.*, 2023). In this work, we propose the first semi-supervised full discourse parsing pipeline that sequentially conducts parsing tasks. We show that with minimal supervision, our pipeline can achieve comparable results to supervised models both in in-domain and cross-domain scenarios.

2 Preliminaries

2.1 Segmented Discourse Representation Theory

The Segmented Discourse Representation Theory (SDRT) (Asher, 1993; Lascarides & Asher, 1993; Asher & Lascarides, 2003) is a dynamic representation theory of discourse. In SDRT,

	Split			#DU/doc		#Tok/sent		#Tok/doc		#Spk/doc		Rel
Dataset	train	dev	test	max	avg	max	avg	max	avg	max	avg	type
STAC	947	105	109	105	13.0	13	4.4	607	50	6	3.0	16
Molweni	9,000	500	500	14	8.8	17	11.9	208	105	9	3.5	16

TABLE 1 – Statistics in STAC and Molweni corpora. Numbers of discourse units per document (DU/doc), tokens per sentence (tok/sent), tokens per document (tok/doc), speakers per document (spk/doc) are given. Both corpora have the same relation types.

the basic elements of analysis are clause-like text spans, known as Discourse Units (DUs). The smallest units of DUs are Elementary Discourse Units (short in EDUs). The coherence of a document is obtained via a structure – oftentimes tree-like or graph-like – of rhetorically connected discourse units. Figure 1 gives an SDRT-annotated example where nodes and edges represent EDUs and relations, respectively.

2.2 Discourse Datasets

We utilize two English dialogue corpora in this study, both annotated under the SDRT framework. Some key statistics are shown in Table 1.

The Strategic Conversations corpus (STAC) (Asher *et al.*, 2016) is currently the most commonly used dialogue corpus to train SDRT-style parsers. It contains 45 online multi-party strategic chat logs during the board game *The Settlers of Catan*, where players discuss and exchange resources to build roads and cities. The vocabulary in this corpus is thus special, with a high frequency of words such as *sheep, clay, wood*. The corpus is manually annotated and divided into 1161 sub-documents. We follow the split in Shi & Huang (2019) : 947 for training, 105 for validation, and 109 for testing.

The Molweni Corpus (Li *et al.*, 2020) is derived from the Ubuntu Chat Corpus (Lowe *et al.*, 2015), where 10,000 short multi-party technical chat logs are annotated for discourse analysis and machine reading comprehension. Despite its large size, a large portion of the documents are highly repetitive. The original annotation suffers from quality issues such as inconsistency, making the results less reliable. Therefore, we revised the annotation of a small subset (50 documents) to ensure a more robust evaluation (test only).

3 Proposition and Experiments

3.1 A Pipeline Design

A standard discourse parsing involves three tasks : EDU segmentation, link attachment, and relation prediction. Most previous work in dialogue discourse parsing starts with gold-

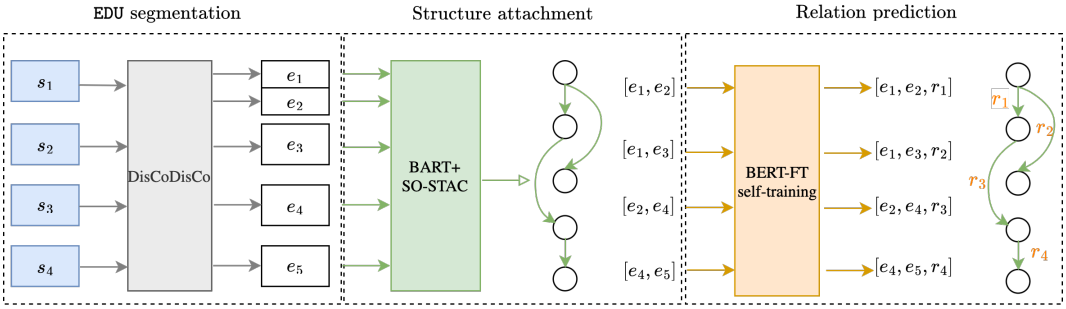


FIGURE 2 – Semi-supervised discourse parsing pipeline proposition. s are utterances ; e are EDUs ; r are rhetorical relations. DisCoDisCo model is proposed in Gessler *et al.* (2021). BART+SO-STAC is BART model fine-tuned on Sentence Ordering task (Li *et al.*, 2023). BERT-FT is BERT model fine-tuned with self-training for relation prediction.

standard EDU annotations and applies a *structure-then-relation* approach (Afantenos *et al.*, 2015; Shi & Huang, 2019; Liu & Chen, 2021; Wang *et al.*, 2021). We follow this pipeline by first predicting naked structures and then providing relations, as depicted in Figure 2. Remarkably, we train the system using only 50 documents in STAC, with an average of 13 EDUs per document, making it the first semi-supervised discourse parsing system for dialogues.

Although most previous work begins with gold EDUs, we consider it crucial to evaluate in a deployed scenario where the parser performs segmentation first. We thus integrate an off-the-shelf segmenter DisCoDisCo (Gessler *et al.*, 2021) – a straightforward sequence tagging model – to perform EDU segmentation. DisCoDisCo achieves an F_1 score of 94.8%.

Next, the predicted EDUs are put into a fine-tuned BART model (Lewis *et al.*, 2020) for **Structure Attachment**. BART is fine-tuned with dialogue-tailored Sentence Ordering task (Barzilay & Lapata, 2008) to enhance the pair-wise, inter-speech block, and inter-speaker discourse information. We hypothesize that the location of discourse information in the network may vary, possibly influenced by the length and complexity of the dialogues. Therefore, we investigate each attention head individually. In one attention head, the attention values among EDUs can be seen as edge weights (Liu & Lapata, 2018). Thus, by using Maximum Spanning Tree algorithms such as Eisner (Eisner, 1996), we obtain a discourse tree structure. For the key issue on choosing the best attention head, we use a small validation set of {10, 30, 50} annotated documents. We find that with just 50 examples, the optimal attention head can be consistently located.

The last step is **Relation Prediction** : with predicted EDU pairs, the goal is to assign a rhetorical relation among 16 candidates. Here we choose BERT model (Devlin *et al.*, 2019) as backbone as it has shown superior performance in discourse-related classification tasks (Chen *et al.*, 2019; Atwell *et al.*, 2021). We prepare the input relation pairs by following the Next Sentence Prediction pattern in BERT, inspired by Shi & Demberg (2019) : a [CLS]

Model	#Train	Segment	Link	Relation	Full
SJ (Chi & Rudnicky, 2022)	1000	-	$70.7_{0.5}$	$77.3_{1.2}$	$54.6_{0.7}$
SJ (Chi & Rudnicky, 2022)	50	-	$55.1_{3.5}$	$61.1_{2.1}$	$33.6_{2.2}$
Ours w gold EDU & link	50	-	-	$58.4_{1.3}$	-
Ours w gold EDU	50	-	**$59.3_{0.7}$**	**$62.0_{1.1}$**	**$38.6_{0.7}$**
Ours w pred EDU	50	94.8	$52.2_{0.4}$	$61.2_{1.6}$	$32.8_{0.9}$

TABLE 2 – Semi-supervised parsing results with the reproduction of SOTA supervised parser Structured Joint (SJ) and our semi-supervised pipeline. Scores are average micro-F_1 over 10 runs. In 50 train setup, best scores are in bold. - not applicable.

token begins the sequence, followed by the first EDU, [SEP], and the second EDU. We loosely translate the output probabilities in BERT model as its predictive confidence, enabling sorting predicted pairs. We select the top k pairs of most confident pseudo-labeled data in each relation type, in which way we maintain the label ratios. This is a simple yet effective sample selection criterion. Through iterative self-training, our classifier is enhanced with the combination of gold and pseudo-labeled data.

3.2 Full Parsing Results

The full parsing results on STAC test set are displayed in Table 2. For comparison, we replicate the SOTA supervised model Structured Joint (SJ) (Chi & Rudnicky, 2022) which uses RoBERTa-base model (Liu *et al.*, 2019) as backbone and employs 3-dimension attention to encode links and relations jointly. In the upper part of the Table, we show SJ performance with 1000 and 50 training data. In the lower part, we detail relation prediction results (gold EDU & link), parsing with gold segmentation (gold EDU), and parsing with predicted segments (pred EDU). Anecdotally, the extracted structures on STAC corpus are found to be similar to the gold SDRT-graphs, achieving an F_1 score of 59.3 and outperforming a strong baseline by 3 points (Li *et al.*, 2023). For relation prediction, our self-trained BERT classifier achieves an accuracy of 58.4% at best. When applying the deployed pipeline, we obtain 32.8% micro-F_1, as displayed in the last line of Table 2. Under the same training size, our pipeline exhibits much better performance compared to SJ model in both link attachment (59.3% vs. 55.1%) and relation prediction (62.0% vs. 61.1%) tasks, bringing a noteworthy improvement of 5 points in full parsing.

In order to test the generalizability of our proposal, we apply SJ model and our pipeline in a cross-domain setup : training on 50 documents from STAC and evaluate on 50 re-annotated dialogues in Molweni test set (Molweni-clean). Preliminary results on Molweni-clean show that our pipeline achieves superior performance on all tasks, surpassing SJ model on link (+24%), relation (+8%), and full parsing (+14%). On relation prediction, SJ considers the

tree structure and relation jointly, while our approach focuses on individual relation pairs. As documents across different genres exhibit diverse structures, our method, despite being more localized, is better suited for general applicability. Moreover, our model exhibits greater stability, whereas the SJ model is heavily biased towards one domain.

4 Conclusion

In this work, we propose a versatile pipeline for sequentially addressing all tasks in discourse parsing. In conformity with real-world situations with limited labeled data, we leverage information from Pre-trained Language Models such as BART and BERT and utilize semi-supervised techniques. Our method shows strong performance in both in-domain and cross-domain settings.

For future work, we intend to improve the derived discourse structures and explore the use of more, possibly out-of-domain raw data, and investigate other bootstrapping approaches for relation prediction. We would also like to evaluate our pipeline on different kinds of data such as transcribed spoken dialogues and another discourse-annotated framework such as the Rhetorical Structure Theory.

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the PIA project "Lorraine Université d'Excellence", ANR-15-IDEX-04-LUE, as well as the CPER LCHN (Contrat de Plan État- Région - Langues, Connaissances et Humanités Numériques). It was partially supported by the ANR (ANR-19-PI3A-0004) through the AI Inter-disciplinary Institute, ANITI, as a part of France's "Investing for the Future — PIA3" program, and through the project AnDiAMO (ANR-21-CE23- 0020). We would like to thank the Grid'5000 community (https ://www.grid5000.fr/).

Références

AFANTENOS S., KOW E., ASHER N. & PERRET J. (2015). Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 928–937, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1109.

ASHER N. (1993). *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.

ASHER N., HUNTER J., MOREY M., FARAH B. & AFANTENOS S. (2016). Discourse structure and dialogue acts in multiparty dialogue : the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2721–2727, Portorož, Slovenia : European Language Resources Association (ELRA).

ASHER N. & LASCARIDES A. (2003). *Logics of conversation*. Cambridge University Press.

ATWELL K., LI J. J. & ALIKHANI M. (2021). Where are we in discourse relation recognition ? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 314–325.

BADENE S., THOMPSON K., LORRÉ J.-P. & ASHER N. (2019). Data programming for learning discourse structure. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 640–645, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1061.

BARZILAY R. & LAPATA M. (2008). Modeling local coherence : An entity-based approach. *Computational Linguistics*, **34**(1), 1–34.

BHATIA P., JI Y. & EISENSTEIN J. (2015). Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 2212–2218, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1263.

CHEN J., LI X., ZHANG J., ZHOU C., CUI J., WANG B. & SU J. (2020). Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, p. 30–36.

CHEN J. & YANG D. (2021). Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1380–1391, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.109.

CHEN M., CHU Z. & GIMPEL K. (2019). Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 649–662.

CHI T.-C. & RUDNICKY A. (2022). Structured dialogue discourse parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 325–335.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EISNER J. (1996). Three new probabilistic models for dependency parsing : An exploration. In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*.

GESSLER L., BEHZAD S., LIU Y. J., PENG S., ZHU Y. & ZELDES A. (2021). Discodisco at the disrpt2021 shared task : A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, p. 51–62.

HUBER P. & CARENINI G. (2022). Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*.

KARIMI H. & TANG J. (2019). Learning hierarchical discourse-level structure for fake news detection. 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)*, p. 3432–3442.

LASCARIDES A. & ASHER N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and philosophy*, **16**(5), 437–493.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

LI C. (2023). *Facing Data Scarcity in Dialogues for Discourse Structure Discovery and Prediction*. Thèse de doctorat, Université de Lorraine.

LI C., HUBER P., XIAO W., AMBLARD M., BRAUD C. & CARENINI G. (2023). Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In *Findings of the Association for Computational Linguistics : EACL 2023*, p. 2517–2534.

LI J., LIU M., KAN M.-Y., ZHENG Z., WANG Z., LEI W., LIU T. & QIN B. (2020). Molweni : A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 2642–2652, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.238.

LIU Y. & LAPATA M. (2018). Learning structured text representations. *Transactions of the Association for Computational Linguistics*, **6**, 63–75. DOI : 10.1162/tacl_a_00005.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

LIU Z. & CHEN N. (2021). Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, p. 122–127, Punta Cana, Dominican Republic and Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.codi-main.11.

LOWE R., POW N., SERBAN I. & PINEAU J. (2015). The Ubuntu dialogue corpus : A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 285–294, Prague, Czech Republic : Association for Computational Linguistics. DOI : 10.18653/v1/W15-4640.

QIN K., WANG L. & KIM J. (2017). Joint modeling of content and discourse relations in dialogues. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 974–984, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1090.

SHI W. & DEMBERG V. (2019). Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, p. 5790–5796.

SHI Z. & HUANG M. (2019). A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, p. 7007–7014.

WANG A., SONG L., JIANG H., LAI S., YAO J., ZHANG M. & SU J. (2021). A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirtieth International Conference on International Joint Conferences on Artificial Intelligence*.