A Semi-supervised Dialogue Discourse Parsing Pipeline
Chuyuan Li, Maxime Amblard, Chloé Braud

To cite this version:
Chuyuan Li, Maxime Amblard, Chloé Braud. A Semi-supervised Dialogue Discourse Parsing Pipeline. Journées Scientifiques du GDR Lift - LIFT 2023, Nov 2023, Nancy, France. hal-04356416

HAL Id: hal-04356416
https://inria.hal.science/hal-04356416
Submitted on 20 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
A Semi-supervised Dialogue Discourse Parsing Pipeline

Chuyuan Li¹  Maxime Amblard¹  Chloé Braud²
(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,
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,
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-
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).
<table>
<thead>
<tr>
<th>Model</th>
<th>#Train</th>
<th>Segment</th>
<th>Link</th>
<th>Relation</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>SJ (Chi &amp; Rudnicky, 2022)</td>
<td>1000</td>
<td>-</td>
<td>70.7_{0.5}</td>
<td>77.3_{1.2}</td>
<td>54.6_{0.7}</td>
</tr>
<tr>
<td>SJ (Chi &amp; Rudnicky, 2022)</td>
<td>50</td>
<td>-</td>
<td>55.1_{3.5}</td>
<td>61.1_{2.1}</td>
<td>33.6_{2.2}</td>
</tr>
<tr>
<td>Ours w gold EDU &amp; link</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>58.4_{1.3}</td>
<td>-</td>
</tr>
<tr>
<td>Ours w gold EDU</td>
<td>50</td>
<td>-</td>
<td>59.3_{0.7}</td>
<td>62.0_{1.1}</td>
<td>38.6_{0.7}</td>
</tr>
<tr>
<td>Ours w pred EDU</td>
<td>50</td>
<td>94.8</td>
<td>52.2_{0.4}</td>
<td>61.2_{1.6}</td>
<td>32.8_{0.9}</td>
</tr>
</tbody>
</table>

**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


DOI : 10.18653/v1/P19-1061.


DOI : 10.18653/v1/D15-1263.


DOI : 10.18653/v1/N19-1423.


