Dialogue Discourse Dependency Parsing from Pre-Trained and Fine-Tuned Language Models

Anonymous ACL submission

Abstract

Discourse parsing suffers from data sparsity, especially for dialogues. As a result, we explore approaches to build naked discourse structures for dialogues, based on attention matrices from Pre-trained Language Models (PLMs). We investigate multiple auxiliary tasks for fine-tuning and show that the dialogue-tailored Sentence Ordering (SO) task performs best. For the crucial step of selecting the best attention head in PLMs, we propose unsupervised and semi-supervised methods. On the Strategic Conversation (STAC) corpus, we reach F_1 scores of 57.2 for the unsupervised and 59.3 for the semi-supervised methods - SOTA for both settings. Restricting our evaluation to projective trees, scores improve to 63.3 and 68.1, respectively.

1 Introduction

Dialogues correspond to an exchange between two or more people. As such, they are generally opposed to monologues, typically authored by a single person. Dialogues can generally take place in person (e.g. meetings, chit-chats), via calls (e.g. customer or medical services), or through text, such as in online forums or direct messages. Recently, the rise of reliable transcription methods and a spike in online communication led to an astonishing explosion of dialogue data. As a result, the need for automatic systems to process dialogues has increased dramatically. For example, summarization of meetings or exchanges with customer service agents could be used to enhance collaborations or analyze customers issues (Li et al., 2019; Feng et al., 2021); machine reading comprehension in the form of question-answering could improve dialogue agents' performance and help knowledge graph construction (He et al., 2021; Li et al., 2021).

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.

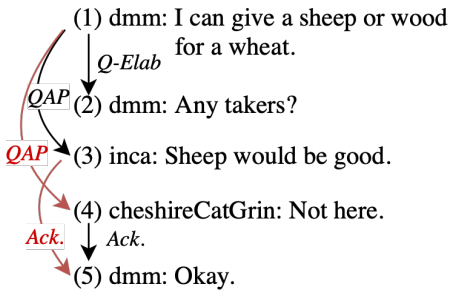

Figure 1: Excerpt of dependency structures in file *s2-leagueM-game4*, STAC. Red links are non-projective.

Several discourse frameworks have been proposed, underlying a variety of annotation projects. For dialogues, data has been primarily annotated within the Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003). Discourse structures are thereby represented as dependency graphs with arcs linking spans of text and additional, semantico-pragmatic relations (e.g. *Acknowledgment* (Ack), *Contrast* or *Question-Answer Pair* (QAP)). Figure 1 shows an example from the STAC corpus (Asher et al., 2016). In this work, we focus on naked structures in SDRT. We are aware that relations are important for downstream tasks. Nevertheless, deriving the structure is the first crucial and valuable step.

Data sparsity has always been an issue for discourse parsing both in monologues and dialogues: the largest and most commonly used corpus annotated under the Rhetorical Structure Theory RST-DT (Carlson et al., 2001) contains $21,789$ discourse units, against $10,678$ for STAC. Restricted to domain and size, the performance of supervised discourse parsers is still low, especially for dialogues, with at best 73.8% F_1 for the naked structure on STAC (Wang et al., 2021). Several transfer learning approaches have thus been proposed, mainly focused on monologues. Previous work

demonstrated that discourse information can be extracted from related tasks, like sentiment analysis (Huber and Carenini, 2020) and summarization (Xiao et al., 2021), or from pre-trained and fine-tuned language models (Huber and Carenini, 2022). While a heuristic-based weakly supervised approach has been recently applied to dialogues (Badene et al., 2019b), we are the first to propose semi- and unsupervised strategies, which can effectively uncover discourse information captured in large pre-trained language models (PLMs).

We take inspiration from previous work (Koto et al., 2021; Pandia et al., 2021; Huber and Carenini, 2022) showing that document-level discourse information can be captured in PLMs like BERT (Devlin et al., 2019) and BART (Lewis et al., 2020), and can be further enhanced by related fine-tuning tasks. We find, however, that the fine-tuning tasks proposed in previous work are not performing well, since they are not designed for dialogues. Dialogues are generally less structured, interspersed with more informal linguistic usage (Sacks et al., 1978), and have structural particularities (Asher et al., 2016). Thus, we propose a new task specially tailored to dialogues: Sentence Ordering (SO), by extending the original proposal by Barzilay and Lapata (2008) with several novel shuffling strategies, enhancing the pair-wise, inter-speech block, and inter-speaker discourse information in PLMs.

A key issue in using PLMs to extract document-level discourse information is how to choose the best attention head. We are the first to tackle this issue in dialogues by proposing both an unsupervised and a semi-supervised approach. The former is based on a novel "Dependency Attention Support" (DAS) metric. This metric calculates the degree of support for the dependency trees generated by each head. We select high-DAS head(s). On the other hand, the semi-supervised approach picks the heads with the best performance on a small annotated validation dataset.

Experimental results on the STAC dataset reveal that our unsupervised and semi-supervised methods outperform a strong baseline LAST (F_1 56.8%, Sec. 4.2), delivering substantial gains on the complete STAC dataset (F_1 59.3%, Sec. 5.2) and show further improvements on the tree-structured subset (F_1 68.1%, Sec. 6.3).

To summarize, our contributions in this work are: (1) Detecting the presence of dialogue discourse information stored in PLMs and fine-tuned models with our newly proposed sentence ordering task; (2) Unsupervised and semi-supervised methods for discourse parsing based on fine-tuned PLMs; (3) An experimental comparison with the strong LAST baseline and other approaches, followed by a detailed quantitative and qualitative analysis of the extracted structures.

2 Related Work

Discourse structures for complete documents have been mainly annotated within the Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003) or the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), with the latter leading to the largest corpora and many discourse parsers for monologues, while SDRT is the main theory for dialogue corpora, i.e. STAC (Asher et al., 2016) and Molweni (Li et al., 2020). In SDRT, discourse structures are dependency graphs with possibly non-projective links (see Figure 1) compared to constituent tree structures in RST. Early approaches to discourse parsing on STAC used varied decoding strategies, such as Maximum Spanning Tree algorithm (Muller et al., 2012; Li et al., 2014; Afantenos et al., 2012) or Integer Linear Programming (Perret et al., 2016). Shi and Huang (2019) first proposed a neural architecture based on hierarchical Gated Recurrent Unit (GRU) and reported 73.2% F_1 on STAC for naked structures. Recently, Wang et al. (2021) adopted Graph Neural Networks (GNNs) and reported marginal improvements (73.8% F_1).

Data sparsity being the issue, a new trend towards semi- and unsupervised discourse parsing has emerged, almost exclusively for monologues. Huber and Carenini (2019, 2020) leveraged sentiment information and showed promising results in cross-domain setting with the silver-standard labeled corpus. Xiao et al. (2021) extracted discourse trees from neural summarizers and confirmed the existence of discourse information in self-attention matrices. Another line of work proposed to enlarge training data with a combination of several parsing models (Jiang et al., 2016; Kobayashi et al., 2021; Nishida and Matsumoto, 2022). As for dialogues, transfer learning approaches are rare. Badene et al. (2019a,b) investigated a weak supervision paradigm where expert-composed heuristics, combined to a generative model, are applied to unseen data. Their method, however, requires domain-dependent annotation and a relatively large validation set for rule verification. Another study

by Liu and Chen (2021) focused on cross-domain transfer using STAC (chats in a game) and Molweni (chats in Ubuntu forum). They applied simple adaptation strategies (mainly lexical information) on a SOTA discourse parser and show improvement compared to bare transfer (train on Molweni and test on STAC F_1 increase from 42.5% to 50.5%). Yet, their model failed to surpass simple baselines.Very recently, Nishida and Matsumoto (2022) investigated bootstrapping methods to adapt BERT-based parsers to out-of-domain data with some success. In comparison to all this previous work, to the best of our knowledge, we are the first to propose a fully unsupervised method and its extension to a semi-supervised setting.

As pre-trained language models (PLMs) such as BERT (Devlin et al., 2019), BART (Lewis et al., 2020) or GPT-2 (Radford et al., 2019) are becoming dominant in the field, *BERTology* research has gained much attention as an attempt to understand what kind of information these models capture. Probing tasks, for instance, can provide fine-grained analysis, but most of them only focus on sentence-level syntactic tasks (Jawahar et al., 2019; Hewitt and Manning, 2019; Kim et al., 2019; Jiang et al., 2020). As for discourse, by applying probing tasks, Zhu et al. (2020) and Koto et al. (2021) showed that BERT and BART encoder networks capture more discourse information than other models, like GPT-2. Very recently, Huber and Carenini (2022) introduced a novel way to encode long documents and explored the effect of different fine-tuning tasks on PLMs, confirming that pre-trained and fine-tuned PLMs both can capture discourse information. Inspired by all these studies on monologues, we investigate how latent information in PLMs can be leveraged for dialogue discourse parsing here.

3 Method: from Attention to Discourse

3.1 Problem Formulation and Simplifications

Given a dialogue with n *Elementary Discourse Units* (EDUs), which are the minimal spans of text (mostly clauses, at most a sentence) to be linked by discourse relations: $D = \{e_1, e_2, e_3, ..., e_n\}$, the goal is to extract a Directed Acyclic Graph (DAG) connecting the n EDUs that best represents its SDRT discourse structure from attention matrices in PLMs[1] (see Figure 2 for an overview of the process). In our proposal, we make a few simplifications, partially adopted from previous work. We do not deal with SDRT *Complex Discourse Units* (CDUs) following Muller et al. (2012); Afantenos et al. (2015) and do not tackle relation type assignment. Furthermore, similar to Shi and Huang (2019), our solution can only generate discourse trees. Extending our algorithm to non-projective trees ($\approx 6\%$ of edges are non-projectives in tree-like examples) and graphs ($\approx 5\%$ of nodes with multiple incoming arcs) are left as future work.

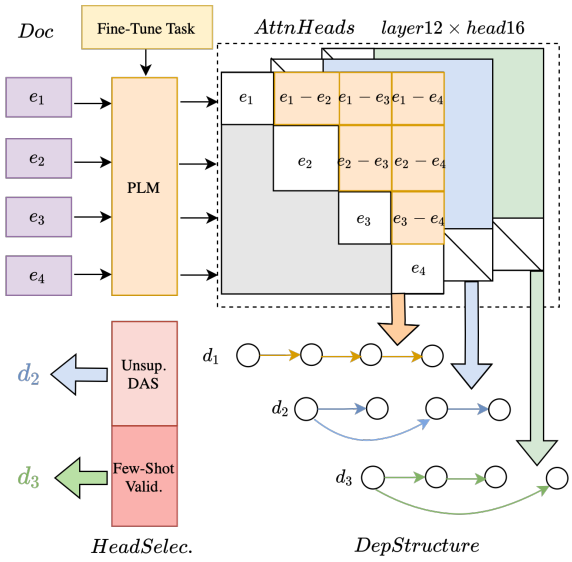

Figure 2: Pipeline for discourse structure extraction.

3.2 Which kinds of PLMs to use?

We explore both vanilla and fine-tuned PLMs, as they were both shown to contain discourse information for monologues (Huber and Carenini, 2022).

Pre-Trained Models: We select BART (Lewis et al., 2020), not only because its encoder has been shown to effectively capture discourse information, but also because it dominated other alternatives in preliminary experiments, including DialoGPT (Zhang et al., 2020) and DialogLM (Zhong et al., 2022), LMs pre-trained with conversational data[2].

Fine-Tuning Tasks: We fine-tune BART on three discourse-related tasks:

Summarization: we use BART fine-tuned on the popular CNN-DailyMail (CNN-DM) news corpus (Nallapati et al., 2016), as well as on the SAMSum dialogue corpus (Gliwa et al., 2019).

Question-Answering: we use BART fined-tuned on the latest version of the Stanford Question

[1] For more details on extracting discourse information from attention mechanisms see Liu and Lapata (2018).

[2] See Appendix E for additional results with further PLMs.

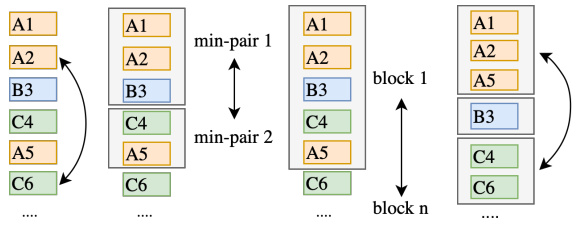

Figure 3: Shuffling strategies (left to right: partial, minimal-pair, block, speaker-turn) on a sequence of utterances 1 to 6, with A, B, C as the speakers.

Answering Dataset (SQuAD 2.0) (Rajpurkar et al., 2018).

Sentence Ordering: we fine-tune BART on the Sentence Ordering task, reordering a set of shuffled sentences to their original order. Considering the complexity of dialogues, we additionally define several shuffling strategies so that the learning is more gradual and effective. Specifically, as shown in Figure 3, we explore: (a) *partial-shuf*: randomly picking 3 utterances (2 for short dialogues with less than 4 utterances) in a dialogue and shuffling them while keeping the context unchanged. (b) *minimal-pair-shuf*: shuffling minimal pairs, comprising of a pair of speech turns from 2 different speakers with at least 2 utterances. A speech turn represents the beginning of a new speaker. (c) *block-shuf*: shuffling a block containing multiple speech turns. We divide one dialogue into $[2, 5]$ blocks based on the number of utterances[3] and shuffle between blocks. (d) *speaker-turn-shuf*: grouping all speech productions of one speaker together. The sorting task consists of ordering speech turns from different speakers' production. We evenly combine all permutations mentioned above to create our **mixed-shuf** data set and conduct the SO task as the third auxiliary task to fine-tune BART.

Choice of Attention Matrix: The BART model contains three kinds of attention matrices: encoder, decoder and cross attention. We use the encoder attention in this work, since it has been shown to capture most discourse information (Koto et al., 2021) and outperformed the other alternatives in preliminary experiments on a validation set.

[3]Block size is designed to be as twice or 3 times bigger than "min-pair", we thus set criteria aiming to have ≈ 6 EDUs per block: $|utt.| < 12 : b = 2$, $|utt.| \in [12, 22] : b = 3$, $|utt.| \in [22, 33] : b = 4$, $|utt.| \geq 33 : n = 5$.

3.3 How to derive trees from attention heads?

Given an attention matrix $A^t \in \mathbb{R}^{k \times k}$ where k is the number of tokens in the input dialogue, we derive the matrix $A^{edu} \in \mathbb{R}^{n \times n}$, with n the number of EDUs, by computing $A^{edu}(i, j)$ as the average of the submatrix of A^t corresponding to all the tokens of EDUs e_i and e_j, respectively. As a result, A^{edu} captures how much EDU e_i depends on EDU e_j and can be used to generate a tree connecting all EDUs by maximizing their dependency strength. Concretely, we find a Maximum Spanning Tree in the fully-connected dependency graph A^{edu} using the Eisner algorithm (Eisner, 1996). Conveniently, since an utterance cannot be anaphorically and rhetorically dependent on following utterances in a dialogue, as they are previously unknown (Afantenos et al., 2012), we can further simplify the inference by applying the following hard constraint to remove all backward links from the attention matrix A^{edu}: $a_{ij} = 0$, if $i > j$.

3.4 How to find the best heads?

Xiao et al. (2021) and Huber and Carenini (2022) showed that discourse information is not evenly distributed between heads and layers. However, they do not provide a strategy to select the head(s) containing most discourse information. Here, we propose two effective selection methods: fully unsupervised or semi-supervised.

3.4.1 Unsupervised Best Head(s) Selection

Dependency Attention Support Measure (DAS): Loosely inspired by the confidence measure in Nishida and Matsumoto (2022), where the authors define the confidence of a teacher model based on predictive probabilities of the decisions made, we propose a DAS metric measuring the degree of support for the maximum spanning (dependency) tree (MST) from the attention matrix. Formally, given an attention matrix A^g (i.e., A^{edu} for the dialogue g) with n EDUs, the MST T^g is built by selecting $n - 1$ attention links l_{ij} from A^g based on the tree generation algorithm. Please note that DAS can be easily adapted for a general graph by removing the restriction to $n - 1$ arcs. DAS measures the strength of all those connections by computing the average score of all the selected links:

$$DAS(T^g) = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} Sel(A^g, i, j) \quad (1)$$

with $Sel(A^g, i, j) = A^g_{ij}$, if $l_{ij} \in T^g$, 0 otherwise.

Selection Strategy: With DAS, we can now compute the degree of support from each attention head h on each single example g for the generated tree $DAS(T_h^g)$. We therefore propose two strategies to select attention heads based on the DAS measure, leveraging either global or local support. The **global** support strategy selects the head with highest averaged DAS score over all the data examples:

$$H_{global} = \arg\max_h \sum_{g=1}^{M} DAS(T_h^g) \quad (2)$$

where M is the number of examples. In this way, we select the head that has a generally good performance on the target dataset. The second strategy is more adaptive to each document, by only focusing on the **local** support. It does not select one specific head for the whole dataset, but instead selects the head/tree with the highest support for each single example g, i.e.,

$$H_{local}^g = \arg\max_h DAS(T_h^g) \quad (3)$$

3.4.2 Semi-Supervised Best Head(s) Selection

We also propose best heads selection using a few annotated examples. In conformity with real-world situations where labeled data is scarce, we sample three small subsets with $\{10, 30, 50\}$ data points (i.e., dialogues) from the validation set. We examine every attention matrix individually, resulting in 12 layers $\times$ 16 heads candidate matrices for each dialogue. Then, the head with the highest micro-F_1 score on the validation set is selected to derive trees in the test set. We also consider layer-wise aggregation, with details in Appendix A.

4 Experimental Setup

4.1 Datasets

We use the multi-party dialogue STAC corpus (Asher et al., 2016), annotated following the SDRT framework, to evaluate our approach on the discourse dependency structure prediction task. Including 300 strategic conversations of players trading goods during the board game *The Settlers of Catan*, this corpus contains some high-frequency game-related words such as *sheep, clay* and *wood*.

To evaluate a variety of fine-tuned PLMs (see sec 3.2), we use publicly available HuggingFace models for the summarization and question-answering tasks. For the newly proposed sentence ordering (SO) task, we train the BART model on two dialogue datasets: (1) the STAC corpus itself (raw text) (2) DailyDialog (Li et al., 2017), covering various topics for English learners (10 categories), from ordinary life to finance. We select this corpus due to its large size, diversity of topics and high quality. We summarize the key dataset statistics for STAC and DailyDialog in Table 1. STAC has a separation of 82%, 9%, 9% for train, validation, and test sets resp.; DailyDialog 85%, 8%, 8%. We purposely exclude the Molweni corpus (Li et al., 2020) in this work, due to major quality issues found in preliminary dataset exploration, with details in Appendix B.

Dataset	#Doc	#Utt/doc	#Tok/doc	#Spk/doc	Domain
DailyDialog	13,118	13	119	2	Daily
STAC	1,161	11	50	3	Game

Table 1: Key statistics of datasets. Utt = sentences in DD or EDUs in STAC; Tok = tokens; Spk = speakers.

4.2 Baselines and Supervised Dialogue Discourse Parsers

We compare against the simple yet strong unsupervised LAST baseline (Schegloff, 2007), attaching every EDU to the previous one. Furthermore, to assess the gap between our approach and supervised dialogue discourse parsers, we compare with the Deep Sequential model by Shi and Huang (2019) and the Structure Self-Aware (SSA) model by Wang et al. (2021).

4.3 Evaluation Metrics

We report the micro-F_1 for discourse parsing and the Unlabeled Attachment Score (UAS) for the generated naked dependency structures.

4.4 Implementation Details

We base our work on the transformer HuggingFace library (Wolf et al., 2020) (see Appendix F) and follow the *text-to-marker* framework proposed in Chowdhury et al. (2021) for the SO fine-tuning procedure. We use the original separation of train, validation, and test sets; set the learning rate to $5e-6$; use a batch size of 2 for DailyDialog and 4 for STAC, and train for 7 epochs. All other hyper-parameters are set following Chowdhury et al. (2021). We do not do any hyper-parameter tuning. We omit 5 documents in DailyDialog during training since the documents lengths exceed the token limit. We replace speaker names with markers (e.g. Sam $\rightarrow$ "spk1"), following the preprocessing pipeline for dialogue utterances in PLMs.

5 Results

5.1 Results with Unsupervised Head Selection

Results using our novel unsupervised DAS method on STAC are shown in Table 2 for both the global (H_g) and local (H_l) head selection strategies. These are compared to: (1) the unsupervised LAST baseline (at the top), which only predicts local attachments between adjacent `EDUs`. LAST is considered a strong baseline in discourse parsing (Muller et al., 2012), but has the obvious disadvantage of completely missing long-distance dependencies which may be critical in downstream tasks. (2) The supervised Deep Sequential parser by Shi and Huang (2019) and Structure Self-Aware model by Wang et al. (2021) (center of the table), trained on STAC, reaching resp. 71.4%[4] and 73.8% in F_1.

In the last sub-table we show unsupervised scores from pre-trained and fine-tuned LMs on three auxiliary tasks: summarization, question-answering and sentence ordering (SO) with the mixed shuffling strategy. We present the global head (H_g) and local heads (H_l) performances selected by the DAS score (see section 3.4.1). The best possible scores using an oracle head selector (H_{ora}) are presented for reference.

Comparing the values in the bottom sub-table, we find that the pre-trained BART model underperforms LAST, with global head and local heads achieving similar performance. Noticeably, models fine-tuned on the summarization task ("+CNN", "+SAMSum") and question-answering ("+SQuAD2") only add marginal improvements compared to BART. In the last two lines of the subtable, we explore our novel sentence ordering fine-tuned BART models. We find that the BART+SO approach surpasses LAST when using local heads. As commonly the case, the intra-domain training performs best, which is further strengthened in this case due to the special vocabulary in STAC. Importantly, our PLM-based unsupervised parser can capture some long-distance dependencies compared to LAST (Section 6.2). Additional analysis regarding the chosen heads is in Section 6.1.

5.2 Results with Semi-Sup. Head Selection

While the unsupervised strategy only delivered minimal improvements over the strong LAST baseline, Table 3 shows that if a few annotated examples are provided, it is possible to achieve substantial gains. In particular, we report results on the vanilla BART model, as well as BART model fine-tuned on DailyDialog ("+SO-DD") and STAC itself ("+SO-STAC"). We execute 10 runs for each semi-supervised setting ($[10, 30, 50]$) and report average scores and the standard deviation.

[4] We re-train the model, scores are slightly different due to different train-test splits, as in Wang et al. (2021).

Model			
Unsupervised Baseline			
LAST			56.8
Supervised Models			
Deep-Sequential (2019)			71.4
SSA-GNN (2021)			73.8
Unsupervised PLMs	H_g	H_l	H_{ora}
BART	56.6	56.4	57.6
+ CNN	56.8	56.7	57.1
+ SAMSum	56.7	56.6	57.6
+ SQuAd2	55.9	56.4	57.7
+ SO-DD	56.8	57.1	58.2
+ SO-STAC	56.7	**57.2**	59.5

Table 2: Micro-F_1 on STAC for supervised SOTA models and PLMs. H_g: global best head. H_l: local best heads. H_{ora}: oracle head. Best (non-oracle) score in the 3^{rd} block in bold.

Train on → Test with ↓	BART F_1	+ SO-DD F_1	+ SO-STAC F_1
LAST BSL	56.8	56.8	56.8
Gold H	57.6	58.2	59.5
Unsup H_g	56.6	56.8	56.7
Unsup H_l	56.4	57.1	57.2
Semi-sup 10	$57.0_{0.012}$	$57.2_{0.012}$	$57.1_{0.026}$
Semi-sup 30	$57.3_{0.005}$	$57.3_{0.013}$	$59.2_{0.009}$
Semi-sup 50	$\mathbf{57.4_{0.004}}$	$\mathbf{57.7_{0.005}}$	$\mathbf{59.3_{0.007}}$

Table 3: STAC micro-F_1 scores from BART and fine-tuned models with unsupervised and semi-supervised approaches. Subscription is standard deviation.

With oracle attention heads (Gold H in the table), all three models achieve superior performance compared to LAST. Furthermore, using a small scale validation set (50 examples) to select the best attention head remarkably improves the F_1 score from 56.8% (LAST) to 59.3% (+SO-STAC).

F_1 improvements across increasingly large validation-set sizes are consistent, accompanied by smaller standard deviations, as would be expected. The semi-supervised results are very encouraging: with 30 annotated examples, we already reach a

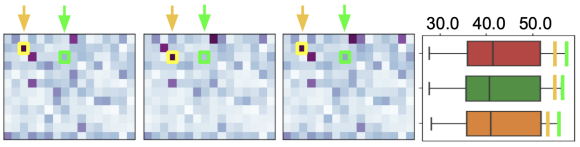

Figure 4: Heatmaps: DAS score matrices (layers: top to bottom=12 to 1, heads: left to right=1 to 16) for BART, BART+SO-DD, BART+SO-STAC. Darker purple=higher DAS score.
Boxplot: Head-aggregated UAS scores for model BART (orange), BART+SO-DD (green) and BART+SO-STAC (red). Light green=head with highest UAS. Yellow=head with highest DAS score.

performance close to the oracle result, and with more examples we can further reduce the gap.

6 Analysis

6.1 Effectiveness of DAS

We now take a closer look at the performance degradation of our unsupervised approach based on DAS in comparison to the upper-bound defined by the performance of the oracle-picked head. To this end, Figure 4 shows the DAS score matrices (left) for three models with the oracle heads and DAS selected heads highlighted in green and yellow, respectively. It becomes clear that the oracle heads do not align with the DAS selected heads. Making a comparison between models, we find that discourse information is consistently located in deeper layers, with the oracle heads (light green) consistently situated in the same head for all three models. However, while not aligning with the oracle, the top performing DAS heads (in yellow) are among the top 10% best heads in all three models, as shown in the box-plot on the right. Hence, we confirm that the DAS method is a reasonable approximation to find discourse intense self-attention heads among the 12×16 attention matrices.

6.2 Document and Arc Lengths

The inherent drawback of the simple, yet effective LAST baseline is its inability to predict indirect arcs. To test if our approach can reasonably predict distant arcs of different length in the dependency trees, we analyze our results in regards to the arc lengths. Additionally, since longer documents tend to contain more distant arcs, we also examine the performance across different document lengths compared to LAST.

Arc Distance: To examine the discourse parsing performance for data sub-sets with specific arc lengths, we present the UAS score plotted against different arc lengths on the left side in Figure 5. Our analysis thereby shows that direct arcs achieve high UAS score ($> 80\%$), independent of the model used. We further observe that the performance drops considerably for arcs of distance two and onwards, with almost all models failing to predict arcs longer than 6. BART+SO-STAC model correctly captures an arc of distance 13. Please note that the presence for long-distance arcs (≥ 6) is limited, accounting for less than 5% of all arcs.

We further analyze the precision and recall scores when separating dependency links into *direct* (adjacent forward arcs) and *indirect* (all other non-adjacent arcs), following Xiao et al. (2021). For direct arcs, all models perform reasonably good. The precision is higher ($\approx +6\%$) and recall is lower than the baseline (100%), indicating that our models predict less direct arcs but more precisely. For indirect arcs, the best model is BART+SO-STAC (20% recall, 44% prec.), closely followed by original BART model (details in Appendix C.1).

Document Length: Longer documents tend to be more difficult to process because of the growing number of possible discourse parse trees. Hence, we analyze the UAS performance of documents in regards to their length, here defined as the number of `EDUs`. Results are presented on the right side in Figure 5, comparing the UAS scores for the three selected models and LAST for different document lengths. We split the document length range into 5 even buckets between the shortest (2 `EDUs`) and longest (37 `EDUs`) document, resulting in 60, 25, 16, 4 and 4 examples per bucket.

For documents with less than 23 `EDUs`, all fine-tuned models perform better than LAST, with BART fine-tuned on STAC reaching the best result. For documents between 23 and 30 `EDUs`, the

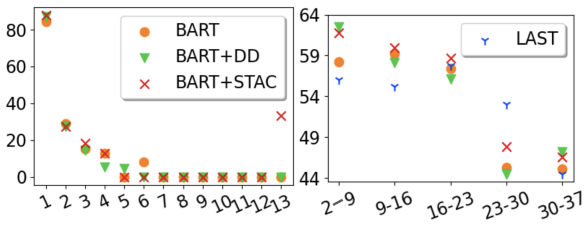

Figure 5: Left: UAS and arcs' distance. x axis: arc distance. Right: averaged UAS for different length of document. x axis: #`EDUs` in a document. y axis: UAS.

	#Doc	#EDUs		#Arcs	
		Single-in	Multi-in	Proj.	N-proj.
(1) Non-Tree	48	706	79	575	170
(2) Tree	61	444	0	348	35
- **Proj. tree**	**48**	314	0	266	0

Table 4: STAC test set ground-truth tree and non-tree statistics. "Single-in" and "multi-in" means EDU with single or multiple incoming arcs.

PLMs under-perform the LAST baseline, likely over-predicting distant arcs, while ground-truth distant arcs only start to appear more frequently in longer documents, with 30 or more EDUs. As a result, we see that longer documents ($\geq$ 23) are indeed more difficult to predict than short documents, with even the performance of our best model (BART+STAC) strongly decreasing.

6.3 Projective Trees Examination

Given the fact that our method only extracts projective tree structures, we now conduct an additional analysis, exclusively examining the subset of STAC containing projective trees, on which our method could in theory achieve perfect accuracy.

Table 4 gives key statistics for this subset ("proj. tree"). For the 48 extracted tree examples, the document length decreases from an average of 11 to 7 EDUs, however, still contains $\approx$ 40% indirect arcs, keeping the parsing difficulty comparable. Discourse parsing results are presented in Table 5. As shown, all three unsupervised models outperform LAST. The best model is still BART fine-tuned on STAC, followed by the inter-domain fine-tuned +SO-DD and BART models. Using the semi-supervised approach, we see further improvement with the F_1 score reaching 68% (+6% than LAST). Degradation for direct and indirect edges' precision and recall scores see Appendix C.2.

Following Ferracane et al. (2019), we analyze key properties of the 48 gold trees compared to our extracted structures using the semi-supervised method. To test the stability of the derived trees, we use three different seeds to generate the shuffled datasets to fine-tune BART. Table 6 presents the averaged scores and the standard deviation of the trees. In essence, while the extracted trees are generally "thinner" and "taller" than gold trees and contain slightly less branches, they are well aligned with gold discourse structures and don't contain "vacuous" trees, where all nodes are linked to one of the first two EDUs.

Train on $\rightarrow$ Test with $\downarrow$	BART F_1	+ SO-DD F_1	+ SO-STAC F_1
LAST BSL	62.0	62.0	62.0
Gold H	64.8	67.4	68.6
Unsup H_g	62.5	62.5	62.1
Unsup H_l	62.1	62.9	63.3
Semi-sup 10	$54.6_{0.058}$	$59.2_{0.047}$	$61.6_{0.056}$
Semi-sup 30	$60.3_{0.047}$	$60.3_{0.044}$	$65.6_{0.043}$
Semi-sup 50	$\mathbf{64.8_{0.000}}$	$\mathbf{66.3_{0.023}}$	$\mathbf{68.1_{0.014}}$

Table 5: Micro-F_1 scores on STAC projective tree subset with BART and SO fine-tuned BART models.

	Avg.branch	Avg.height	%leaf	Norm. arc
GT	1.67	3.96	0.46	0.43
BART	1.20	5.31	0.31	0.34
+SO-DD	$1.32_{0.014}$	$5.31_{0.146}$	$0.32_{0.019}$	$0.37_{0.003}$
+SO-STAC	$1.27_{0.076}$	$5.28_{0.052}$	$0.32_{0.011}$	$0.35_{0.015}$

Table 6: Statistics for ground truth projective trees and extracted trees from oracle attention heads in BART and fine-tuned BART models.

Further, qualitative analysis of inferred structures is presented in Appendix D. Tellingly, on two STAC examples our model succeeds in predicting $>$ 82% of projective arcs, some of which span across 4 EDUs. This is encouraging, providing anecdotal evidence that our method is suitable to extract reasonable discourse structures.

7 Conclusion

Since dialogue discourse parsing suffers from extreme data sparsity, we explore approaches to build naked discourse structures from PLMs attention matrices. We show sentence ordering to be the best fine-tuning task and our unsupervised and semi-supervised methods for selecting the best attention head outperform a strong baseline, delivering substantial gains especially on tree structures. Interestingly, discourse is consistently captured in deeper PLMs layers, and more accurate for shorter links.

In the near future, we intend to explore graph-like structures from attention matrices, for instance, by extending treelike structures with additional arcs of high DAS score and applying linguistically motivated constraints, as in Perret et al. (2016). We would also like to expand shuffling strategies for SO and to explore other auxiliary tasks. We plan to infer full discourse structures by adding the prediction of rhetorical relation types in the long term.

Limitations

Similarly to previous work, we have focused on generating only projective tree structures. This not only covers the large majority of the links ($\approx 94\%$), but it can also provide the backbone for accurately inferring the remaining non-projective links in future work. We focus on the naked structure, as it is a significant first step and a requirement to further predict relations for discourse parsing.

We decided to run all our experiments on the only existing high quality corpus, i.e., STAC. In essence, we traded-off generalizability for soundness of the results. A second corpus we considered, Molweni, had to be excluded due to serious quality issues.

Lastly, since we work with large language models and investigate every single attention head, computational efficiency is a concern. We used a 4-core GPU machine with the highest VRAM at 11MiB. The calculation for one discourse tree on one head was approximately 0.75 seconds (in STAC the averaged dialogue length is 11 `EDUs`), which quickly summed up to 4.5 hours with only 100 data points for 192 candidate trees in one LM. When dealing with much longer documents, for example AMI and conversational section in GUM (in average > 200 utterances/dialogue), our estimation shows that one dialogue takes up to ≈ 2 minutes, which means 6.5 hours for 192 candidate trees. Even though we use parallel computation, the exhaustive "head" computation results in a tremendous increase in time and running storage. One possibility is to investigate only those "discourse-rich" heads, mainly in the deeper layers, for future work.

Ethical Considerations

We carefully select the dialogue corpora used in this paper to control for potential biases, hatespeech and inappropriate language by using human annotated corpora and professionally curated resources. Further, we consider the privacy of dialogue partners in the selected datasets by replacing names with generic user tokens.

Since we are investigating the nature of the discourse structures captured in large PLMs, our work can be seen as making these models more transparent. This will hopefully contribute to avoid unintended negative effects, when the growing number of NLP applications relying on PLMs are deployed in practical settings.

In terms of environmental cost, the experiments described in the paper make use of RTX 2080 Ti GPUs for tree extraction and A100 GPUs for BART fine-tuning. We used up to 4 GPUs for the parallel computation. The experiments on corpus STAC took up to 1.2 hours for one language model, and we tested a dozen models. We note that while our work is based on exhaustive research on all the attention heads in PLMs to obtain valuable insights, future work will able to focus more on discourse-rich heads, which can help to avoid the quadratic growth of computation time for longer documents.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaıs Cadilhac, Cedric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, et al. 2012. Modelling strategic conversation: model, annotation design and corpus. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (Seinedial), Paris*.

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 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)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019a. Data programming for learning discourse structure. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 640–645, Florence, Italy. Association for Computational Linguistics.

Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019b. Weak supervision for learning discourse structure. In *EMNLP*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. 2021. Is everything in order? a simple way to order sentences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10769–10779.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2019. Evaluating discourse in structured text representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 646–653, Florence, Italy. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Yuchen He, Zhuosheng Zhang, and Hai Zhao. 2021. Multi-tasking dialogue comprehension with discourse parsing. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 551–561, Shanghai, China. Association for Computational Lingustics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Patrick Huber and Giuseppe Carenini. 2019. Predicting discourse structure using distant supervision from sentiment. 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)*, pages 2306–2316, Hong Kong, China. Association for Computational Linguistics.

Patrick Huber and Giuseppe Carenini. 2020. MEGA RST discourse treebanks with structure and nuclearity from scalable distant sentiment supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7442–7457, Online. Association for Computational Linguistics.

Patrick Huber and Giuseppe Carenini. 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*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Kailang Jiang, Giuseppe Carenini, and Raymond Ng. 2016. Training data enrichment for infrequent discourse relations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2603–2614, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sang-goo Lee. 2019. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. In *International Conference on Learning Representations*.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2021. Improving neural rst parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1600–1612.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiaqi Li, Ming Liu, Zihao Zheng, Heng Zhang, Bing Qin, Min-Yen Kan, and Ting Liu. 2021. Dadgraph: A discourse-aware dialogue graph neural network for multiparty dialogue machine reading comprehension. *arXiv preprint arXiv:2104.12377*.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu and Nancy Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 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*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India. The COLING 2012 Organizing Committee.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379.

Jérémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 99–109, San Diego, California. Association for Computational Linguistics.

Kechen Qin, Lu Wang, and Joseph Kim. 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)*, pages 974–984, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Emanuel A Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge university press.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2021. Predicting discourse trees from transformer-based neural summarizers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4139–4152, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

Zining Zhu, Chuer Pan, Mohamed Abdalla, and Frank Rudzicz. 2020. Examining the rhetorical capacities of neural language models. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 16–32, Online. Association for Computational Linguistics.

A Semi-sup. Layer-Wise Results

We consider both **layer-wise** attention matrices - averaging 16 attention heads for every layer which gives 12 candidate layers -, and **head-wise** attention matrices - taking each attention matrix individually which results in 192 candidate matrices. Here we show results completed with layer-wise matrices for the whole test set and treelike examples in Table 7 and Table 8.

B Molweni Corpus Quality Investigation

Molweni (Li et al., 2020) is a corpus derived from Ubuntu Chat Corpus (Lowe et al., 2015). It contains 10,000 short dialogues between 8 to 15 utterances, annotated in SDRT framework.

Considering the complexity of Ubuntu chat logs (multiple speakers, entangled discussion with various topics), we first conduct an examination of the corpus. Disappointingly, we found heavy repetition within sequential documents and inconsistency in discourse annotation among the same utterances. We thus decide not to include it in this work.

Train on → Test with ↓	BART F_1	+ SO-DD F_1	+ SO-STAC F_1
Gold H	57.6	58.2	59.5
Semi-sup-10 1L	$55.8_{0.008}$	$55.7_{0.010}$	$55.6_{0.009}$
Semi-sup-30 1L	$55.8_{0.006}$	$56.5_{0.004}$	$56.3_{0.004}$
Semi-sup-50 1L	$56.2_{0.002}$	$56.4_{0.007}$	$56.4_{0.001}$
Semi-sup-10 1H	$57.0_{0.012}$	$57.2_{0.012}$	$57.1_{0.026}$
Semi-sup-30 1H	$57.3_{0.005}$	$57.3_{0.013}$	$59.2_{0.009}$
Semi-sup-50 1H	$\mathbf{57.4_{0.004}}$	$\mathbf{57.7_{0.005}}$	$\mathbf{59.3_{0.007}}$

Table 7: Micro-F_1 scores on STAC test set with BART and fine-tuned models. H = "head", L = "layer". Best semi-supervised score is in bold. Subscription is std. deviation.

Train on → Test with ↓	BART F_1	+ SO-DD F_1	+ SO-STAC F_1
Gold H	64.8	67.4	68.6
Semi-sup-10 1L	$59.4_{0.028}$	$60.6_{0.029}$	$58.3_{0.018}$
Semi-sup-30 1L	$62.1_{0.002}$	$61.8_{0.012}$	$59.8_{0.009}$
Semi-sup-50 1L	$62.1_{0.000}$	$62.3_{0.003}$	$59.9_{0.006}$
Semi-sup-10 1H	$54.6_{0.058}$	$59.2_{0.047}$	$61.6_{0.056}$
Semi-sup-30 1H	$60.3_{0.047}$	$60.3_{0.044}$	$65.6_{0.043}$
Semi-sup-50 1H	$\mathbf{64.8_{0.000}}$	$\mathbf{66.3_{0.023}}$	$\mathbf{68.1_{0.014}}$

Table 8: Micro-F_1 scores on STAC projective tree subset with BART and SO fine-tuned BART models.

Clus ID	Doc ID	#Theor =arc	#Err arc	#Theor =rel	#Err rel
1	{1, 2, 3}	18	2	16	2
2	{7, 8, 9}	18	0	18	7
3	{10, 11, 12, 13, 14}	80	4	76	25
...					
105	500	4787	284	4503	606
-	-	100%	5.9%	100%	13.5%

Table 9: Quantitative resume of link and relation inconsistency in Molweni test set. "Theor =arc": number of arcs between the same utterances, *a priori* should be linked in the same way; "Theor =rel": number of relations between the linked utterances.

Clusters: Among 500 dialogues in discourse augmented test set, we found 105 "clusters". One cluster groups all the documents with only one or two different utterances. For instance, document id 10 and 11 are in the same cluster since only the second utterance is different (Figure 10). A similar situation is attested in the documents {1, 2, 3}, {7, 8, 9}, {19, 20, 21}, to name a few.

Annotation Inconsistency: A closer examination of the annotation in similar examples reveals inconsistency for both discourse links and rhetorical relations. Precisely, we investigate every *document pair* (two documents in the same cluster) in all 105 clusters in the test set. A visualization of inconsistency for documents 10 and 11 is shown in Figure 10: apart from $\texttt{EDU}_2$, we expect the same links and relations among other `EDUs`. However, we observe one link inconsistency (in red) and two relation inconsistencies (in blue). In total, we find 6% of link errors (#Err arc) within the same `EDUs` and 14% of relation errors (#Err rel) in the test set[5]. The scores are shown in Table 9.

The Ubuntu Chat Corpus contains long dialogues with entangled discussion. A pre-processing had been made to generate shorter dialogues. While these slightly different short dialogues could be interesting for other dialogue studies in the field. Our focus on the discourse structure request more various data points and most importantly, the coherent discourse annotation.

C Precision and Recall Scores for Direct and Indirect Arcs in STAC

C.1 STAC Test Set

We show the precision and recall of direct and indirect arcs for the test set in Figure 6. Each color

[5]For validation and train sets we find similar error rates.

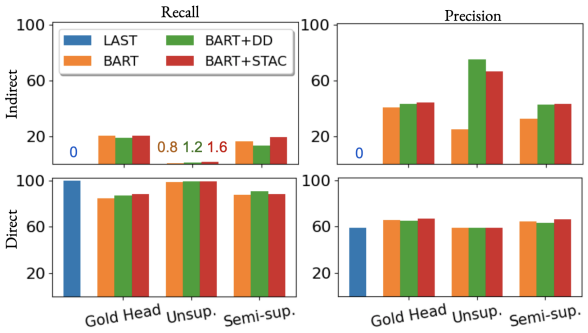

Figure 6: Comparison of recall (left) and precision (right) of indirect (top) and direct (bottom) links in LAST baseline and SO fine-tuned models on STAC.

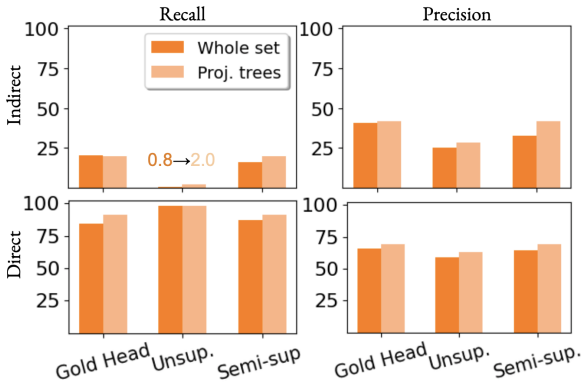

Figure 7: Recall and precision metrics in whole test set (darker color) *vs.* projective tree subset (brighter color), with BART model.

represents one model, with blue represents LAST.

C.2 STAC Projective Tree Set

To compare the performance of the whole test set and tree-structured subset, we now present the recall and precision scores of BART (Fig. 7), BART+SO-DD (Fig. 8), and BART+SO-STAC (Fig. 9) separately.

D Qualitative Analysis in STAC

We show a few concrete tree examples: 3 well predicted (Figure 11, 12, 13), 3 badly predicted (Figure 14, 15, 16), and 2 random examples (Figure 17, 18). Some patterns observed from badly predicted structures: (1) chain-style prediction: as shown in Figure 15 and 18 where only adjacent `EDUs` are linked together; (2) inaccurate indirect arc prediction: especially for long documents such as the one in Figure 16.

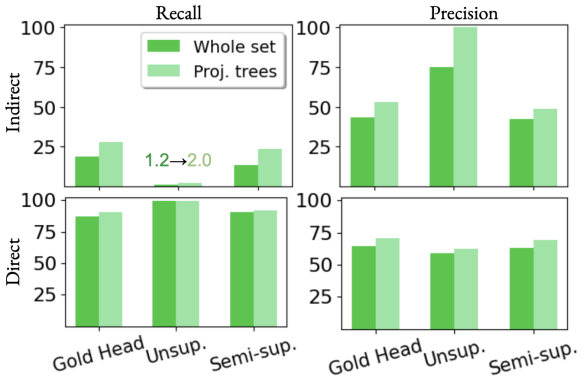

Figure 8: Recall and precision metrics in whole test set (darker color) *vs.* projective tree subset (brighter color), with BART+SO-DD model.

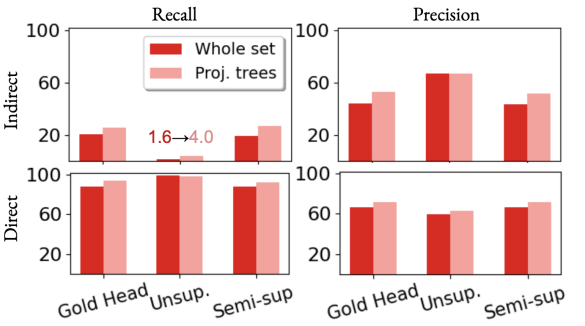

Figure 9: Recall and precision metrics in whole test set (darker color) *vs.* projective tree subset (brighter color), with model BART+SO-STAC.

Model		Unsup		Semi-sup		
	H_{ora}	H_g	H_l	Semi10	Semi30	Semi50
BART	57.6	56.6	56.4	$57.0_{0.012}$	$57.3_{0.005}$	$57.4_{0.004}$
+ SO-DD	58.2	56.8	57.1	$57.2_{0.012}$	$57.3_{0.013}$	$57.7_{0.005}$
+ SO-STAC	59.5	56.7	57.2	$57.1_{0.026}$	$59.2_{0.009}$	**$\underline{59.3}_{0.007}$**
RoBERTa	57.4	56.8	56.8	$55.6_{0.013}$	$56.8_{0.002}$	$\underline{56.9}_{0.003}$
DialoGPT	56.2	42.7	36.2	$52.9_{0.043}$	$55.1_{0.017}$	$\underline{56.2}_{0.000}$
DialogLED	57.2	56.8	56.7	$54.6_{0.026}$	$54.7_{0.061}$	$\underline{56.6}_{0.019}$
+ SO-DD	57.7	56.4	56.6	$55.0_{0.028}$	$56.1_{0.024}$	$\underline{57.3}_{0.009}$
+ SO-STAC	58.4	56.8	57.1	$57.7_{0.001}$	$\underline{58.2}_{0.005}$	$57.7_{0.001}$

Table 10: Micro-F_1 on STAC with other PLMs. Best score (except H_{ora}) in each row is underlined.

E Results with other PLMs

We test with RoBERTa (Liu et al., 2019), DialoGPT (Zhang et al., 2020), and DialogLED (DialogLM with Longformer) (Zhong et al., 2022) to see how different language models encode discourse information. As shown in Table 10, the most discourse-rich head in RoBERTa slightly underperform BART (-0.2%), so does the DialogLED (-0.4%) and DialoGPT (-1.4%). Sentence ordering fine-tuned DialogLED model outperforms the original one, proving that our proposed SO task can help encoding the discourse information.

F Huggingface Models

Table 11 shows the models and the sources we obtained from Huggingface library (Wolf et al., 2020).

Model
BART-large https://huggingface.co/facebook/bart-large
BART-large-cnn https://huggingface.co/facebook/bart-large-cnn
BART-large-samsum https://huggingface.co/linydub/bart-large-samsum
BART-large-finetuned-squad2 https://huggingface.co/phiyodr/bart-large-finetuned-squad2
RoBERTa-large https://huggingface.co/roberta-large
DialoGPT-small https://huggingface.co/microsoft/DialoGPT-small
DialogLED-large-5120 https://huggingface.co/MingZhong/DialogLED-large-5120

Table 11: Huggingface models and URLs.

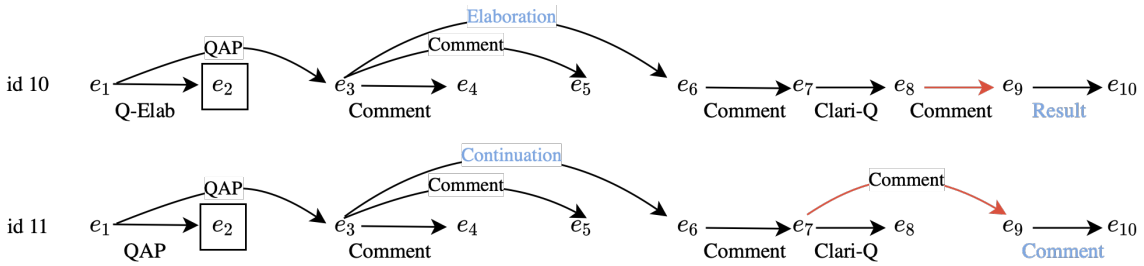

Figure 10: Similar documents in the same cluster. Circled EDUs are different. In red: inconsistent discourse arcs; in blue: inconsistent rhetorical relation.

test id 10:

[e_1] matthew99857: so do i need additional hardware to fix it ?
[e_2] vocx: ca n't you disable the raid from the bios ? check your motherboard manual .
[e_3] ikonia: just use the disk as an individual disk
[e_4] sugi: vocxi : oh i am sorry . i misunderstood you . thank i will try it now
[e_5] vocx: you need to word better your answers , seems like nobody in getting you today .
[e_6] sugi: vocx : iso 9660 cd-rom filesystem data udf filesystem data (unknown version , id 'nsr01 ')
[e_7] ikonia: looks like that should work as a loop back file system
[e_8] sugi: -mount -o loop but instead of .iso .mdf ? or the .mds file ?
[e_9] ikonia: try it , linux see 's it as a " image " so it may work
[e_{10}] sugi: vocx : wow it worked , i feel retard for nto

test id: 11

[e_1] matthew99857: so do i need additional hardware to fix it ?
[e_2] ikonia: no you need to stop using raid
[e_3] ikonia: just use the disk as an individual disk
[e_4] sugi: vocxi : oh i am sorry . i misunderstood you . thank i will try it now
[e_5] vocx: you need to word better your answers , seems like nobody in getting you today .
[e_6] sugi: vocx : iso 9660 cd-rom filesystem data udf filesystem data (unknown version , id 'nsr01 ')
[e_7] ikonia: looks like that should work as a loop back file system
[e_8] sugi: -mount -o loop but instead of .iso .mdf ? or the .mds file ?
[e_9] ikonia: try it , linux see 's it as a " image " so it may work
[e_{10}] sugi: vocx : wow it worked , i feel retard for nto

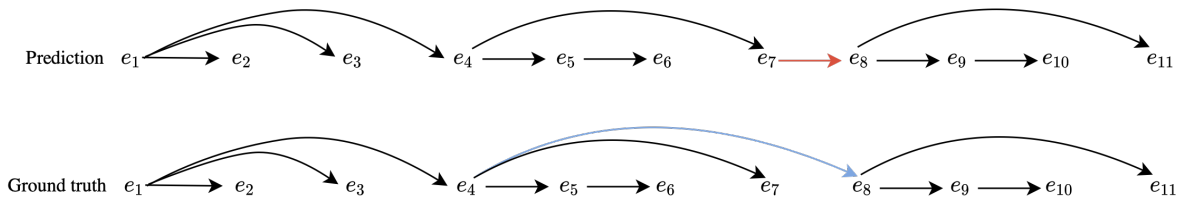

Figure 11: Well predicted example: *pilot02-4*. #EDUs: 11. UAS: 90%. In red: FP arcs; in blue: FN arcs.
[e_1] Cat: anyone would give me clay? [e_2] Thomas: none here [e_3] william: no [e_4] Cat: I have one wood to exchange [e_5] Cat: any takers? [e_6] william: no [e_7] Cat: for sheep, wheat or clary [e_8] Thomas: can I buy a sheep for two ore? [e_9] william: have none [e_{10}] Thomas: kk [e_{11}] Cat: no sheep

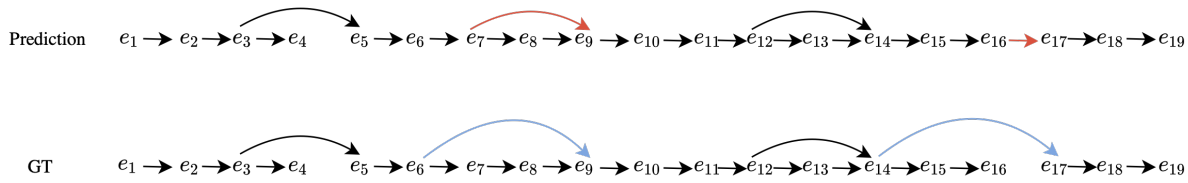

Figure 12: Well predicted example: *pilot02-18*. #EDUs: 19. UAS: 88.9%. In red: FP arcs; in blue: FN arcs. [e_1] william: hi markus. [e_2] william: how many people are we waiting for? [e_3] Thomas: think it's 1 more [e_4] william: ok [e_5] Markus: yes, one more [e_6] Markus: seems there's a hickup logging into the game ... [e_7] Thomas: that's ok, I not on a schedule [e_8] Thomas: *I'm [e_9] Markus: I guess you two had no problems joining the game? [e_{10}] william: nope [e_{11}] Markus: Ah great! [e_{12}] Markus: So, one of you can now start the game. [e_{13}] Markus: Have fun! [e_{14}] william: the arrow is pointing at me [e_{15}] william: but i cant press roll [e_{16}] william: oh sorry [e_{17}] Thomas: u can place a settlement [e_{18}] Thomas: first [e_{19}] Thomas: u roll later

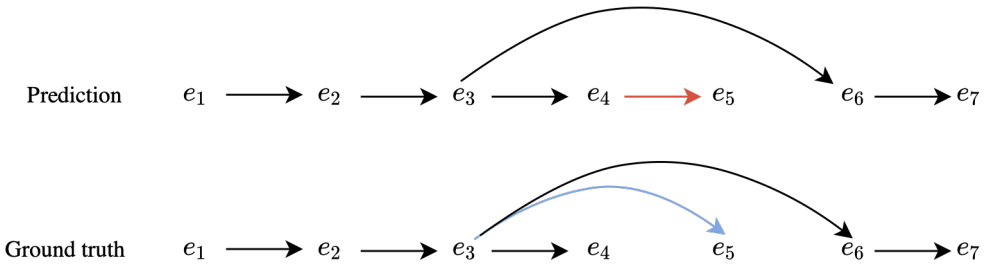

Figure 13: Well predicted example: *s1-league3-game3*. #EDUs: 7. UAS: 83.3%. In red: FP arcs; in blue: FN arcs. [e_1] Gaeilgeoir: ? [e_2] yiin: build road [e_3] inca: think we're meant to negotiate trades in the chat before offering [e_4] yiin: oop [e_5] yiin: ok then [e_6] inca: part of the guys' experiment [e_7] yiin: oh i see

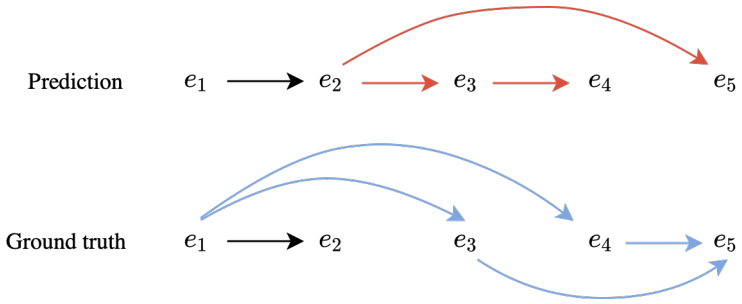

Figure 14: Badly predicted example: *s2-leagueM-game4*. #EDUs: 5. UAS: 20%. In red: FP arcs; in blue: FN arcs. [e_1] dmm: i can give a sheep or wood for a wheat. [e_2] dmm: any takers? [e_3] inca: sheep would be good. [e_4] CheshireCatGrin: Not here. [e_5] dmm: okay.

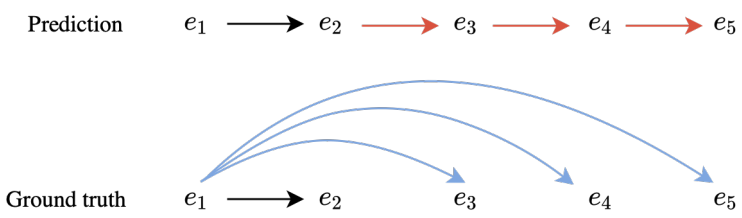

Figure 15: Badly predicted example: *s1-league3-game3*. #EDUs: 5. UAS: 25%. In red: FP arcs; in blue: FN arcs. [e_1] nareik15: anyone have ore. [e_2] nareik15: I have some wood to trade. [e_3] yiin: no sorry. [e_4] inca: nope, sorry. [e_5] Gaeilgeoir: no, sorry.

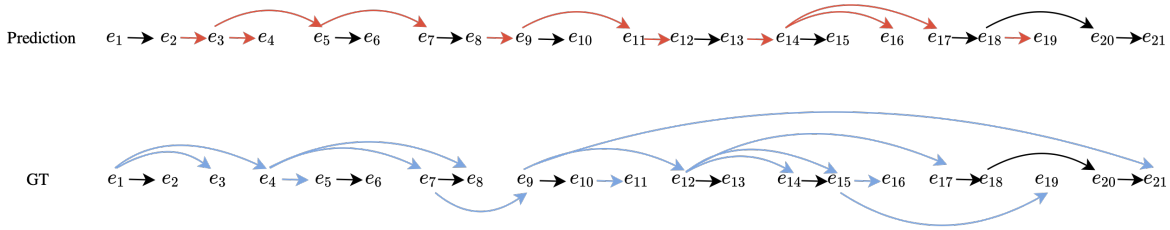

Figure 16: Badly predicted example: *s1-league4-game2*. #EDUs: 21. UAS: 30%. In red: FP arcs; in blue: FN arcs. [e_1] Shawnus: need wheat [e_2] Shawnus: want..clay? [e_3] ztime: you odo? [e_4] ztime: yer.. [e_5] ztime: I need clay.. [e_6] ztime: can give wheat [e_7] Shawnus: k [e_8] Shawnus: this might be where i lose my road card a? [e_9] ztime: er.. [e_{10}] ztime: I think the trade is wrong? [e_{11}] ztime: did you want wheat? [e_{12}] Shawnus: yes [e_{13}] Shawnus: for clay [e_{14}] ztime: it said you wanted clay... [e_{15}] somdechn: We all want wheat man [e_{16}] somdechn: and clay... [e_{17}] ztime: ok [e_{18}] ztime: thanks.. [e_{19}] Shawnus: haha [e_{20}] Shawnus: thanks [e_{21}] somdechn: That happens in the real game as well.

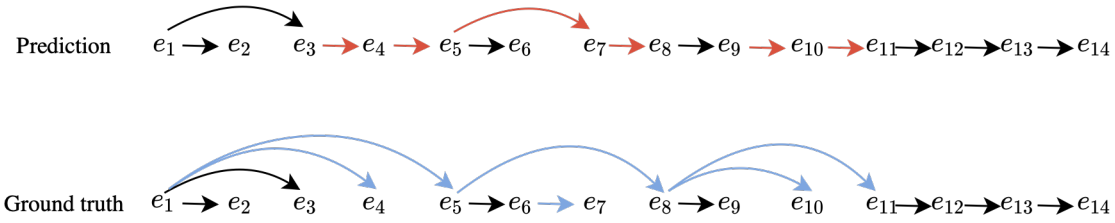

Figure 17: Random example: *s2-league4-game2*. #EDUs: 14. UAS: 53.9%. In red: FP arcs; in blue: FN arcs. [e_1] ztime: 7!!!! [e_2] somdechn: Yeah right... [e_3] ztime: what... is this a fix? [e_4] Shawnus: hahaha [e_5] ztime: ok anyone want wheat? [e_6] Shawnus: nope [e_7] Shawnus: just someone to roll 9's.. [e_8] somdechn: Yes [e_9] somdechn: I can give you wood. [e_{10}] ztime: was that yes to a trade somdech? [e_{11}] ztime: OK.. cool.. for 1 wheat? [e_{12}] somdechn: and an ore.. :) [e_{13}] ztime: err.. don't have ore.. [e_{14}] ztime: thanks..

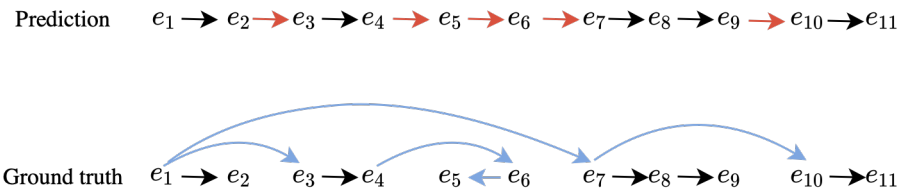

Figure 18: Random example: *s1-league3-game3*. #EDUs: 11. UAS: 50%. In red: FP arcs; in blue: FN arcs. [e_1] nareik15: anyone have wood to trade. I have sheep [e_1] yiin: no [e_1] Gaeilgeoir: Sorry, [e_1] Gaeilgeoir: I need wood too [e_1] Gaeilgeoir: I have wheat [e_1] Gaeilgeoir: if you want [e_1] inca: do you have wheat kieran? [e_1] inca: if so [e_1] inca: i can trade wood [e_1] nareik15: sorry, [e_1] nareik15: plenty of sheep though :)