Discourse Structure Extraction from Pre-Trained and Fine-Tuned Language Models in Dialogues

Findings of EACL 2023

Chuyuan Li, Patrick Huber, Wen Xiao, Maxime Amblard, Chloé Braud, Giuseppe Carenini

Dialogues

Context & Motivation

- Explosion of dialogue data
    - Form: In person, calls, texts (online forums)
    - Objective: chit-chats, task-specific (e.g.: restaurant reservation)

- Simple surface-level features not sufficient (Qin et al., 2017)
  → Need semantic & pragmatic relations, for instance **discourse analysis**

Fig: Dialog forms, from Internet

Dialogues

Context & Motivation

- Explosion of dialogue data
  - Form: In person, calls, texts (online forums)
  - Objective: chit-chats, task-specific (e.g.: restaurant reservation)

- Simple surface-level features not sufficient (Qin et al., 2017)
  → Need semantic & pragmatic relations, for instance **discourse analysis**

- Issue: data sparsity
  - RST-DT (Wall Street Journal): 21.8k discourse units
  - STAC (The Settlers of Catan board game, Asher et al., 2016): ~10k discourse units

Fig: Dialog forms, from Internet

Discourse Structure in Dialogues

Segmented Discourse Representation Theory

- SDRT Framework ([Asher et al., 2003](#))
  - Presented as **graph**, with <u>nodes</u> represent discourse units (DU) and <u>edges</u> rhetorical relations

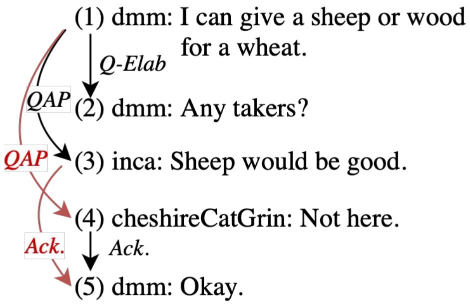


Fig: Excerpt s2-leagueM-game4, STAC.

Dialogue Specificities

- Generally <u>less structured</u>, informal linguistic usage ([Sacks et al., 1978](#))
- Structural <u>particularities</u>, e.g., *lozenge*-shape

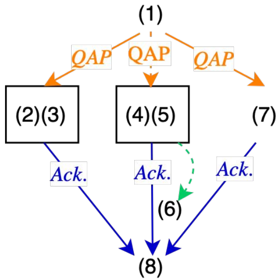


Fig: Lozenge-shaped discourse structure, STAC.

Discourse Structure in PLMs

EMPIRICAL INSPIRATION

- BERTology Research
    - Discourse probing/structure extraction tasks in Pre-Trained Language Models (PLMs):
      Koto et al., 2021, Pandia et al.. 2021, Huber&Carenini 2022

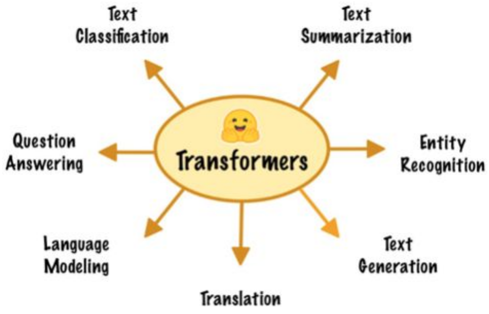


Fig: Top: illustration of depdency structure in SDRT; Bottom: Transformer-based model and tasks

Discourse Structure in PLMs

EMPIRICAL INSPIRATION

- BERTology Research
  - Discourse probing/structure extraction tasks in Pre-Trained Language Models (PLMs): Koto et al., 2021, Pandia et al.. 2021, Huber&Carenini 2022
- Structure extraction from attention matrices: Liu&Lapata2018

⇒ Our Task: extract discourse structure in dialogues from PLMs

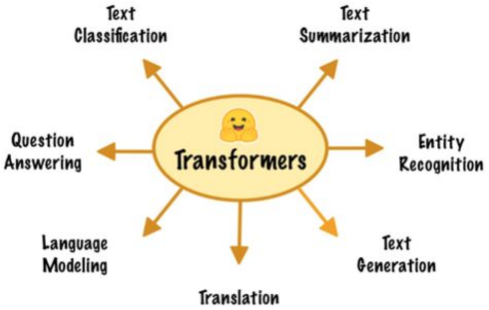
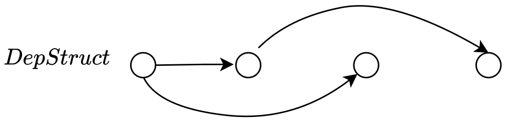


Fig: Top: illustration of depdency structure in SDRT; Bottom: Transformer-based model and tasks



Discourse Structure as DAG in Dialogues

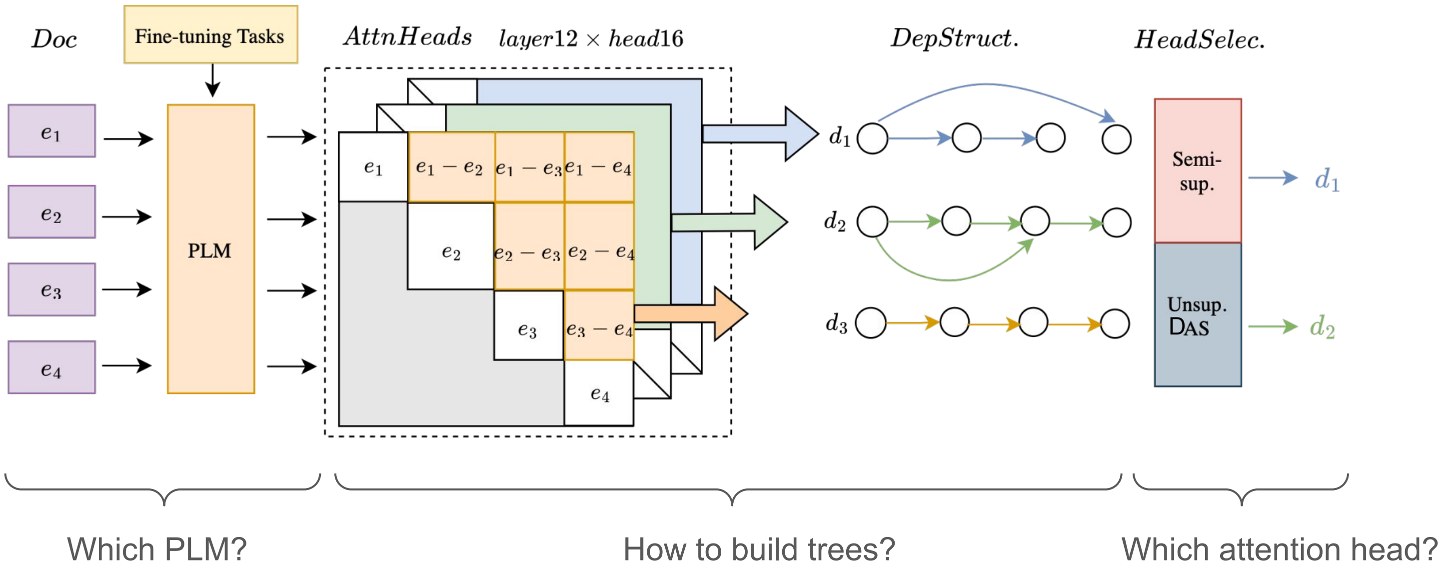
Task Formulation

- Dialogue with *n* elementary discourse units (EDUs) *D={e1, e2, …, en}*
- Extract a Directed Acyclic Graph (DAG) connecting the *n* EDUs that best represent SDRT structure

- Simplifications
  - Complex discourse units (CDUs) → EDUs
  - DAG → Dependency Trees, as in Muller2012, Li2014, Afantenos2012, Shi2019, Wang2021 (note that Perret2016 predict DAGs)

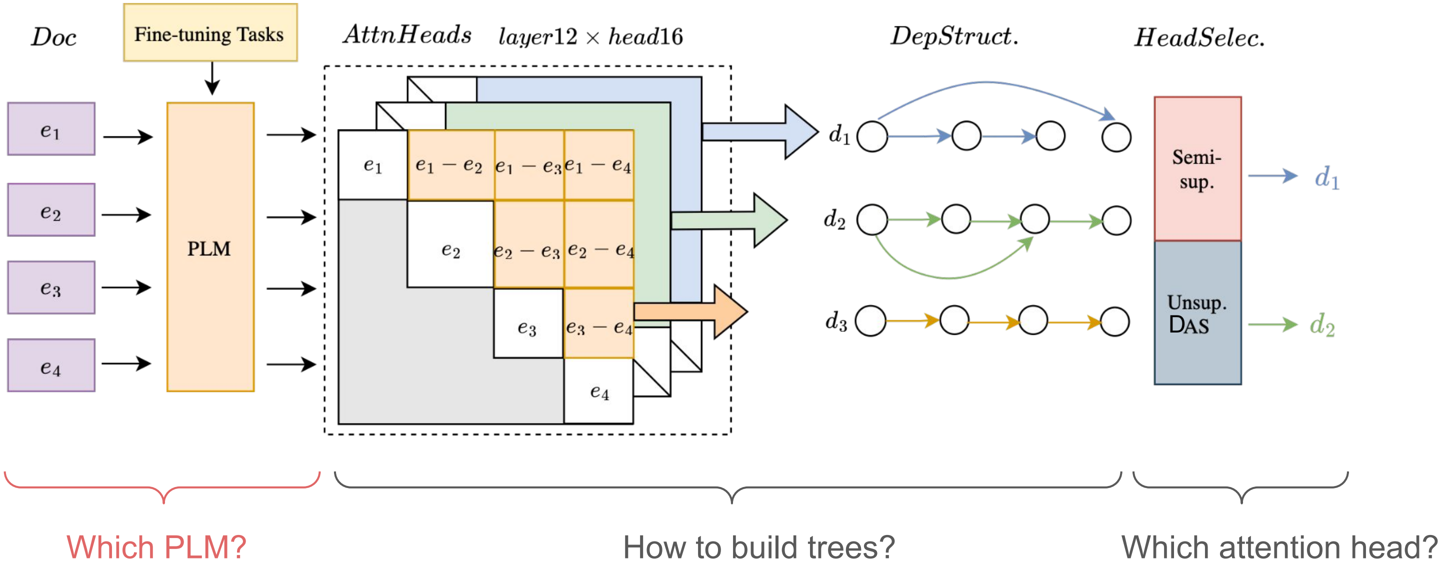
Discourse Structure in Dialogues from PLMs

PIPELINE



Discourse Structure in Dialogues from PLMs

PIPELINE



Discourse Structure in Dialogues from PLMs

METHODS (1) – WHICH KINDS OF PLMS TO USE?

- Pre-Trained Models
  - BART (Lewis et al., 2019): encoder-decoder
  - Others: DialoGPT (Zhang et al., 2020), DialogLM (Zhong et al., 2022)

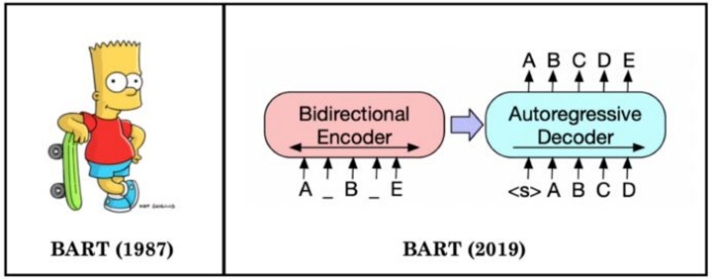


Fig: BART from The Simpsons; BART model. Source.

Discourse Structure in Dialogues from PLMs

METHODS (1) – WHICH KINDS OF PLMS TO USE?

- Fine-Tuning Tasks & Corpora
    - Summarization: CNN-Dailymail, SAMSum
    - Question-Answering: SQuAD2
    - **Sentence Ordering (SO)**: STAC, DailyDialog

Discourse Structure in Dialogues from PLMs

Methods (1) – Which Kinds Of PLMs To Use?

- Fine-Tuning Tasks & Corpora
  - Summarization: CNN-Dailymail, SAMSum
  - Question-Answering: SQuAD2
  - **Sentence Ordering (SO)**: STAC, DailyDialog
    - Barzilay&Lapata 2008, Chowdhury et al., 2021
    - Mixed shuffling strategies: pair-wise, inter-block, inter-speaker shuffling

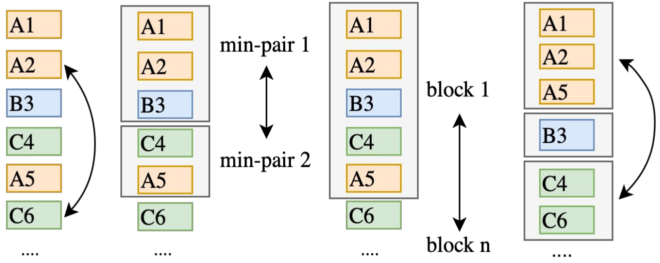
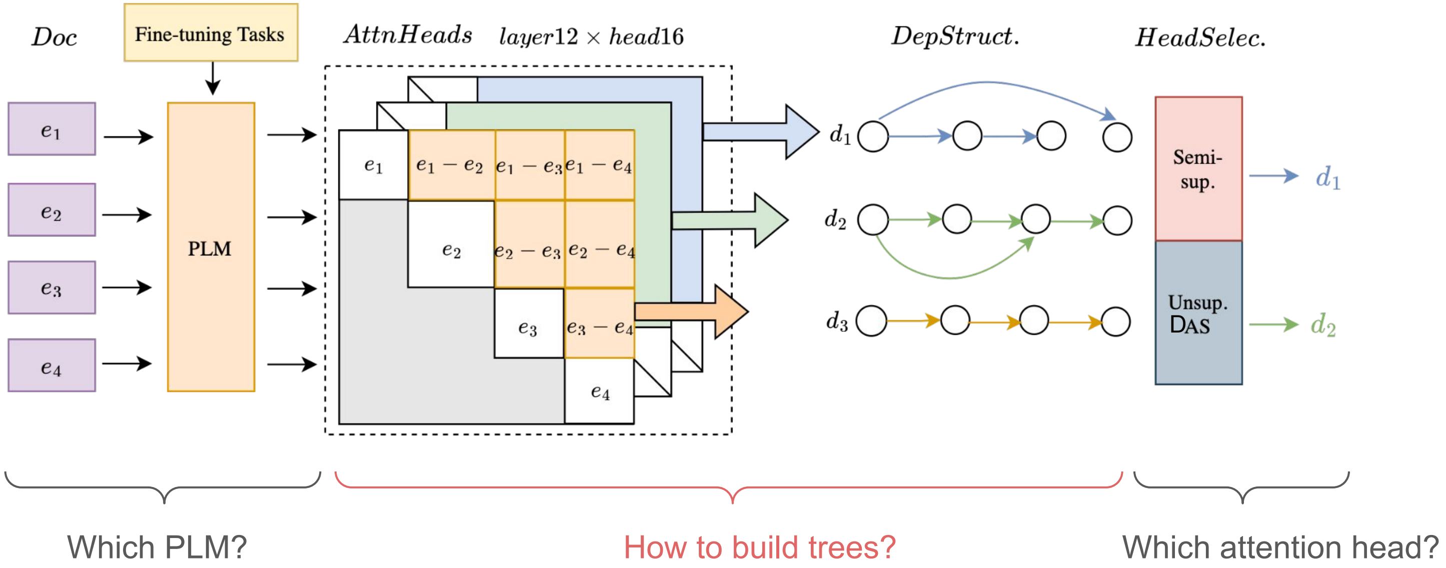


Fig: partial, minimal-pair, block, speaker-turn shuffling strategies.

Discourse Structure in Dialogues from PLMs

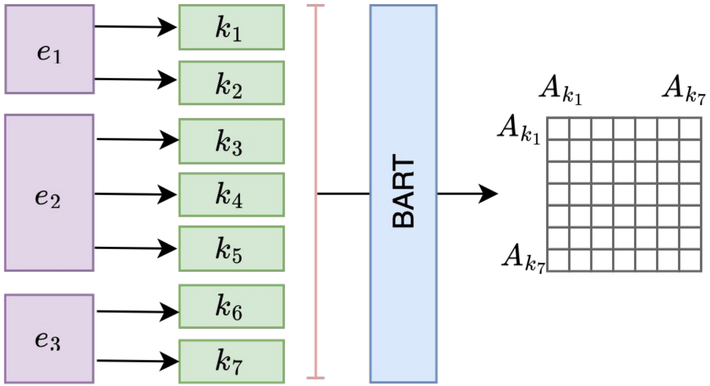
PIPELINE



Discourse Structure in Dialogues from PLMs

METHODS (2) – HOW TO DERIVE TREES FROM ATTENTION HEADS?

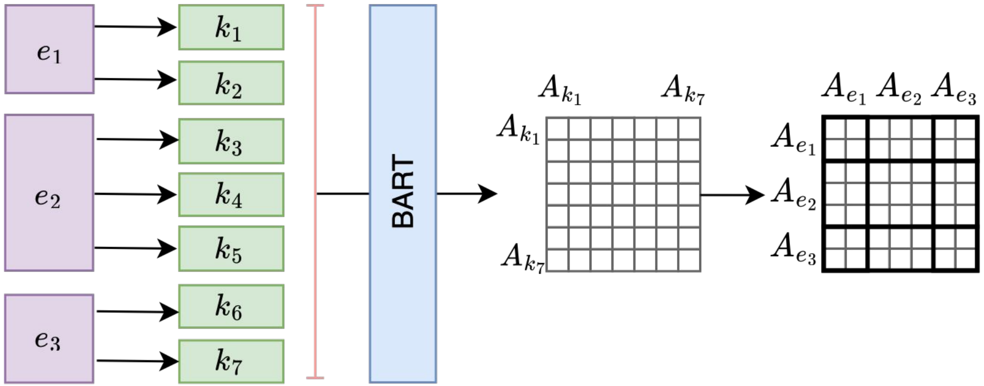
- From each attention matrix



Discourse Structure in Dialogues from PLMs

METHODS (2) – HOW TO DERIVE TREES FROM ATTENTION HEADS?

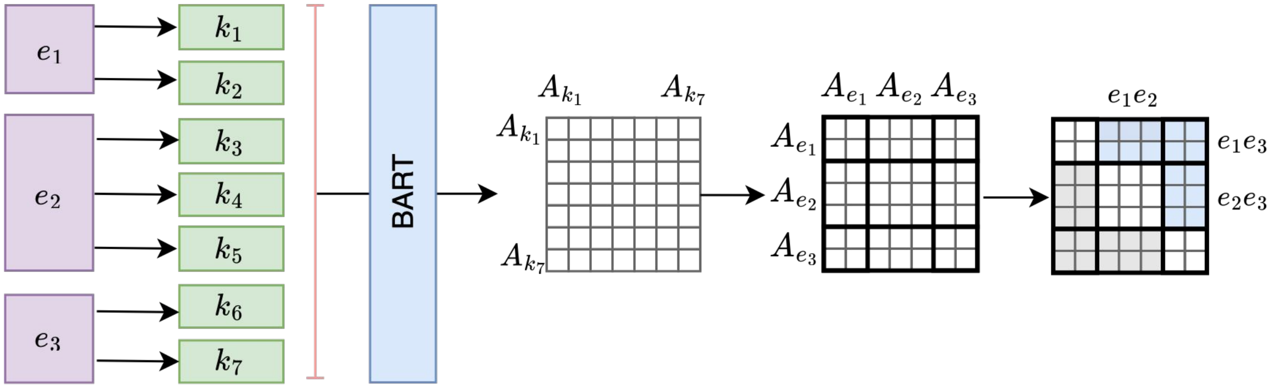
- From each attention matrix



Discourse Structure in Dialogues from PLMs

METHODS (2) – HOW TO DERIVE TREES FROM ATTENTION HEADS?

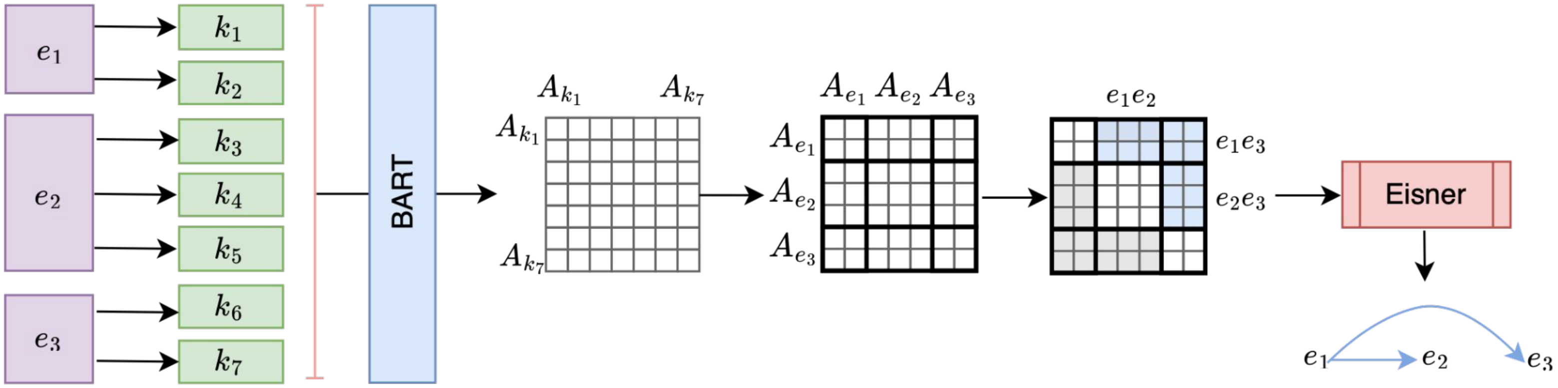
- From each attention matrix



Discourse Structure in Dialogues from PLMs

METHODS (2) – HOW TO DERIVE TREES FROM ATTENTION HEADS?

- From each attention matrix

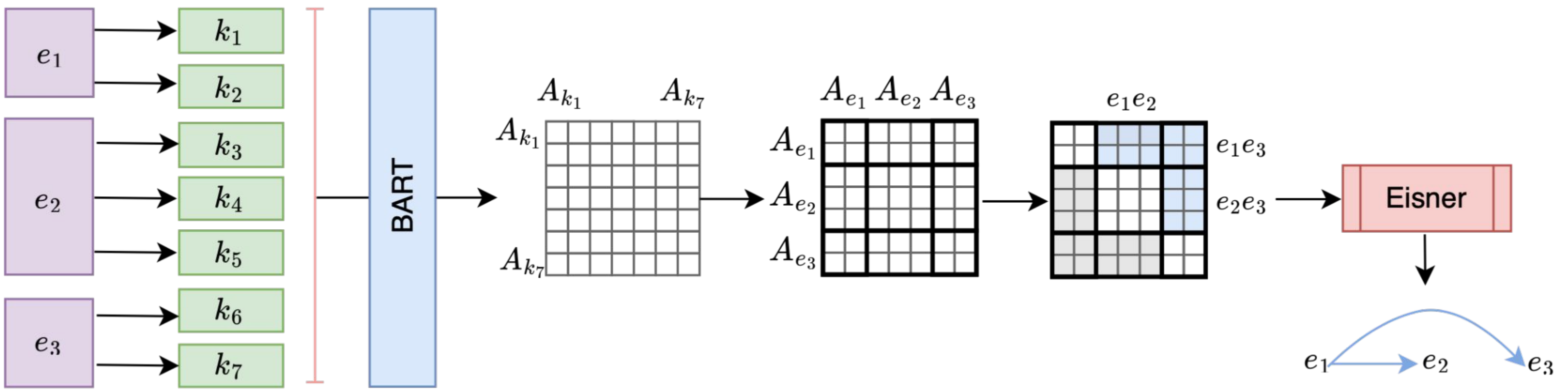


Discourse Structure in Dialogues from PLMs

METHODS (2) – HOW TO DERIVE TREES FROM ATTENTION HEADS?

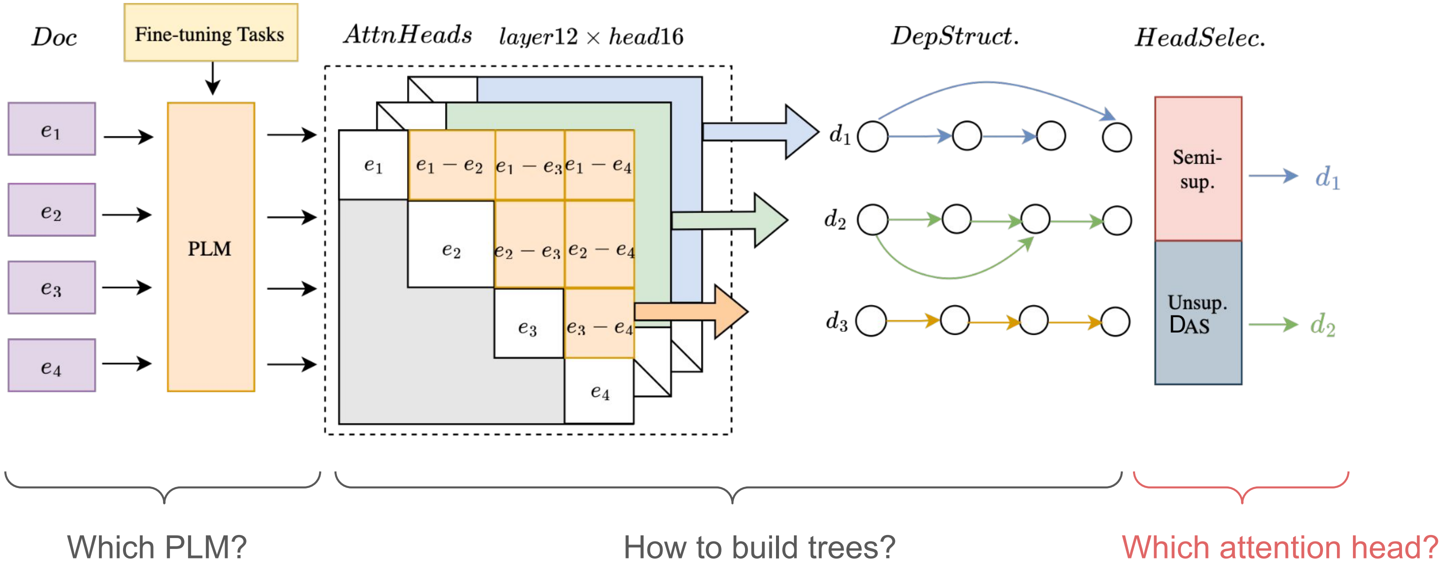
- From each attention matrix

→ *Heads x Layers* candidates



Discourse Structure in Dialogues from PLMs

PIPELINE



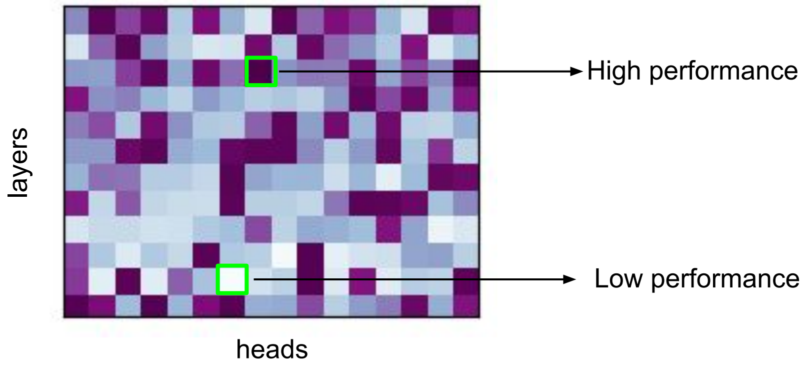
Discourse Structure in Dialogues from PLMs

METHODS (3) – HOW TO FIND THE BEST HEADS?

- Discourse extraction method operates on single self-attention matrices

  → BART:192 candidate matrices (16 heads x 12 layers)

- <u>Question</u>: which heads / layers contain most discourse information?



Discourse Structure in Dialogues from PLMs

Methods (3) – How To Find The Best Heads?

- Unsupervised Selection
  - *Dependency Attention Support* (**DAS**) score

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

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

Where Tg is Eisner extracted Tree for dialog g.

Discourse Structure in Dialogues from PLMs

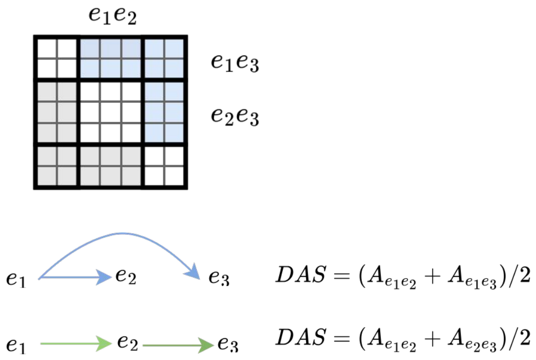
METHODS (3) – HOW TO FIND THE BEST HEADS?

- Unsupervised Selection
  - *Dependency Attention Support* (**DAS**) score

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

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

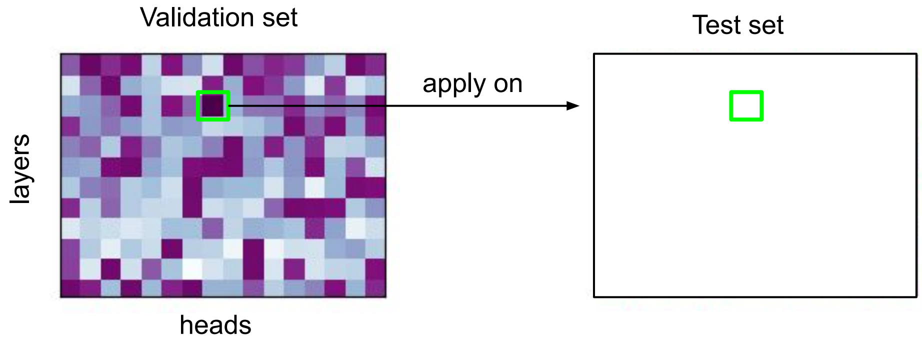
Where Tg is Eisner extracted Tree for dialog g.



Discourse Structure in Dialogues from PLMs

METHODS (3) – HOW TO FIND THE BEST HEADS?

- Semi-supervised Selection
  - Use annotated subset of {10, 30, 50} examples in validation set
  - Obtain best performing head, apply on test set
  - Execute 10 runs for each subset



Discourse Structure in Dialogues from PLMs

EXPERIMENTAL SETTINGS

- Datasets: STAC (Settlers of Catan board game)
- PLM: BART
- Baselines & Supervised Discourse Parsers
  - *LAST* – unsupervised baseline
  - Deep Sequential (Shi2019), Graph Neural Network (Wang2021) – gap with supervised parsers
- Evaluation Metrics
  - Micro-F1
  - Unlabeled attachment score (UAS)

Discourse Structure in Dialogues from PLMs

RESULTS (1) – UNSUPERVISED DAS

- LAST: unsupervised baseline
- H_g: global head
- H_l: local head
- H_ora: oracle head

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

- BART underperform LAST
- FT on summarization (+CNN, +SAMSum) and QA (+SQuAd2): marginal improvements
- FT on SO (+SO-DD, +SO-SATC) surpass LAST, but less than oracle head

Discourse Structure in Dialogues from PLMs

RESULTS (2) – SEMI-SUPERVISED METHOD

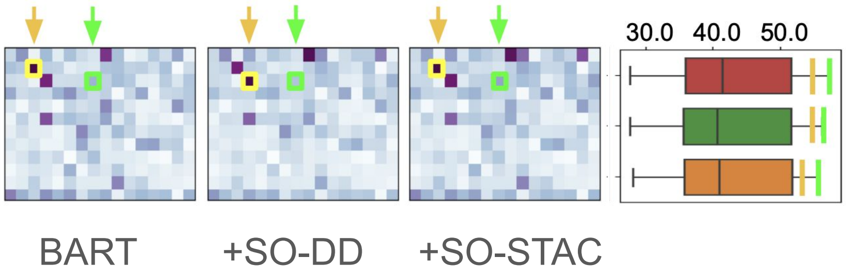
- Use a few (10/30/50) annotated examples in validation set to help find the best attention head
  - All 3 models > LAST
  - With 50 examples, F1 improve from 56.8 → 59.3, achieve almost oracle performance (59.5)
  - Improvement is consistent acros different models and validation sizes, with smaller std-dev.

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	<u>56.6</u>	56.8	56.7
Unsup H_l	56.4	<u>57.1</u>	<u>57.2</u>
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	$\mathbf{57.4_{0.004}}$	$\mathbf{57.7_{0.005}}$	$\boxed{\mathbf{59.3_{0.007}}}$

Discourse Structure in Dialogues from PLMs

Analysis (1) – Effectiveness Of Das

- DAS score matrices
  - Yellow: DAS selected heads
  - Green: Oracle heads

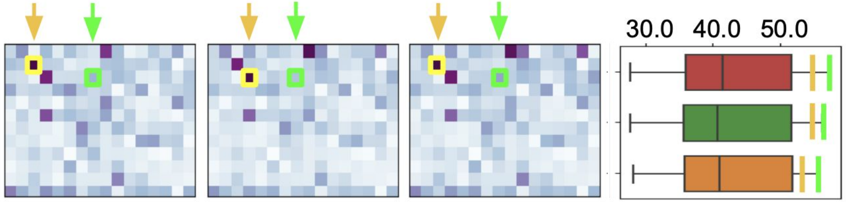


Heatmap: top to bottom: layer 12 to 1, left to right: head 1 to 16.
Boxplot: head-aggregated UAS scores. Red: BART model; green: BART+SO-DD; orange: BART+SO-STAC.

Discourse Structure in Dialogues from PLMs

ANALYSIS (1) – EFFECTIVENESS OF DAS

- DAS score matrices
  - Yellow ☐: DAS selected heads
  - Green ☐: Oracle heads



BART +SO-DD +SO-STAC

Heatmap: top to bottom: layer 12 to 1, left to right: head 1 to 16.
Boxplot: head-aggregated UAS scores. Red: BART model; green: BART+SO-DD; orange: BART+SO-STAC.

→ Disocurse information consistently located in deeper layers

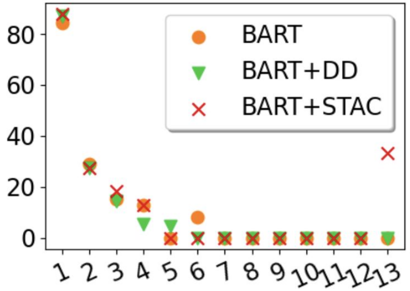
→ Oracle heads situated in the same attention matrices for 3 models

→ DAS != Oracle, but among top 10% best heads, reasonable approximation

Discourse Structure in Dialogues from PLMs

ANALYSIS (2) – DOCUMENT & ARC LENGTHS

- Test if our approach can predict distant edges (compared to LAST with 0 disant edge)



←: UAS and arcs' distance,
x-axis distance, y-axis: UAS

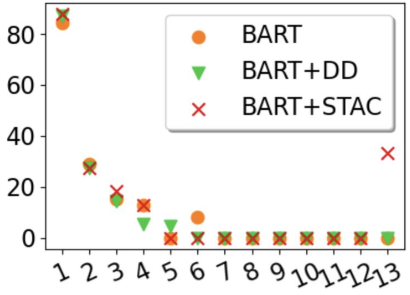
Arc Distance

- Direct arcs: high UAS score (>80%)
- Dist >=2, performance drops
- Dist > 6, almost all fail

Discourse Structure in Dialogues from PLMs

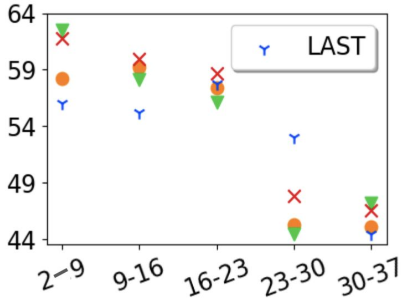
Analysis (2) – Document & Arc Lengths

- Test if our approach can predict distant edges (compared to LAST with 0 disant edge)



←: UAS and arcs' distance,
x-axis distance, y-axis: UAS

→: averaged UAS for different length of document,
x-axis: document length, y-axis: UAS.



Arc Distance

- Direct arcs: high UAS score (>80%)
- Dist >=2, performance drops
- Dist > 6, almost all fail

Document Length

- 5 even intervals [2, 37]
- |doc| < 23 EDUs, all models better than LAST
- [23, 30] worse than bsl, over-predict distant arcs

Discourse Structure in Dialogues from PLMs

ANALYSIS (3) – EXAMINATION ON PROJECTIVE TREES

- Proportion of trees *vs.* graphs in STAC
    - Simplified assumptions
    - Direct and fair comparison

		#EDUs		#Arcs	
	#Doc	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: Trees and non-tree statistics in STAC.

Discourse Structure in Dialogues from PLMs

Analysis (3) – Examination On Projective Trees

- Proportion of trees *vs.* graphs in STAC
  - Simplified assumptions
  - Direct and fair comparison

		#EDUs		#Arcs	
	#Doc	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: Trees and non-tree statistics in STAC.

- Unsupervised and Semi-supervised Experiments
  - Results are improved: F1 from 59% → **68%**
  - Tree Properties (Ferracane et al., 2019)
    - Avg. branch, height, % of leaf, normalized arc, "vacuous" trees (details in appendix)
    - → Well aligned with gold trees
    - → "Thinner" and "taller"

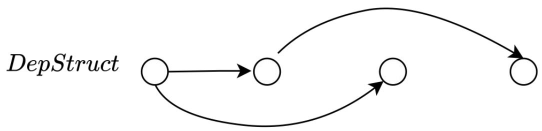
Train on → Test with ↓	BART F_1	+ SO-DD F_1	+ SO-STAC F_1
LAST BSL	62.0	62.0	62.0
Gold H	64.8	67.4	68.6
Unsup H_g	<u>62.5</u>	62.5	62.1
Unsup H_l	62.1	<u>62.9</u>	<u>63.3</u>
Semi-sup 10	$54.6_{0.058}$	$59.2_{0.047}$	$61.6_{0.056}$
Semi-sup 30	$60.3_{0.047}$	$60.3_{0.044}$	$65.6_{0.043}$
Semi-sup 50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

Table: Micro-F1 on STAC projective tree subset.

Discourse Structure in Dialogues from PLMs

CONCLUSION & FUTURE WORK

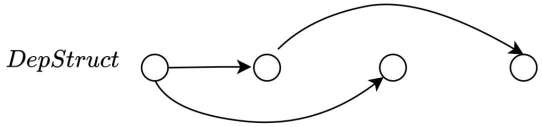
- Detection the presence of discourse information in PLMs
- Design of sentence-ordering fine-tuned task tailored for dialogue structures
- Extraction of naked discourse structure with unsupervised and semi-supervised strategies



Discourse Structure in Dialogues from PLMs

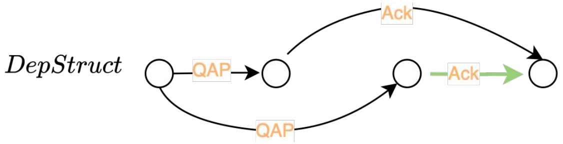
Conclusion & Future Work

- Detection the presence of discourse information in PLMs
- Design of sentence-ordering fine-tuned task tailored for dialogue structures
- Extraction of naked discourse structure with unsupervised and semi-supervised strategies



Future work

- Explore **graph-like** structures by extending treelike structures
- Perform full discourse parsing by adding **relation prediction**



Discourse Structure Extraction from Pre-Trained and Fine-Tuned Language Models in Dialogues

Thank you!

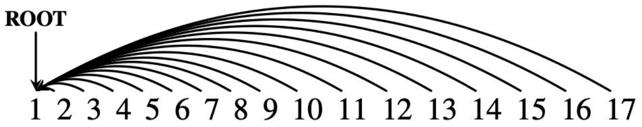
Appendices

Properties of 48 projective dependency trees GT vs. extracted trees from PLMs

	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.

Illustration of “vacuous” trees (Ferracane 2018)



Qualitative investigation of well predicted example

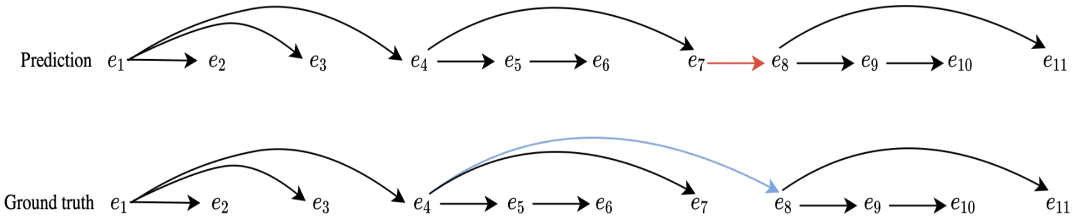


Fig: Well predicted: pilot02-4, STAC. UAS: 90%. In red: false positive; in blue: false negative.

Qualitative investigation of badly predicted examples

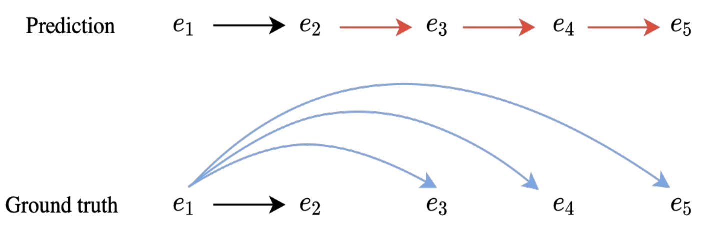


Fig: Badly predicted: s1-league3-game3, STAC. UAS: 25%. ***Failed in predicting distant edges.***

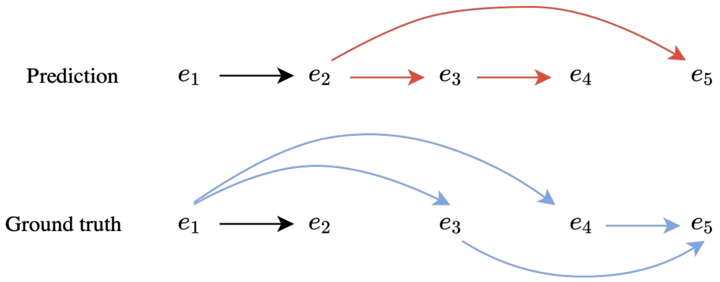


Fig: Badly predicted: s2-leagueM-game4, STAC. UAS: 20%. ***Failed in predicting "lozenge" shape.***

Results with other PLMs

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

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

Recall and Precision of indirect and direct edges in LAST and FT models

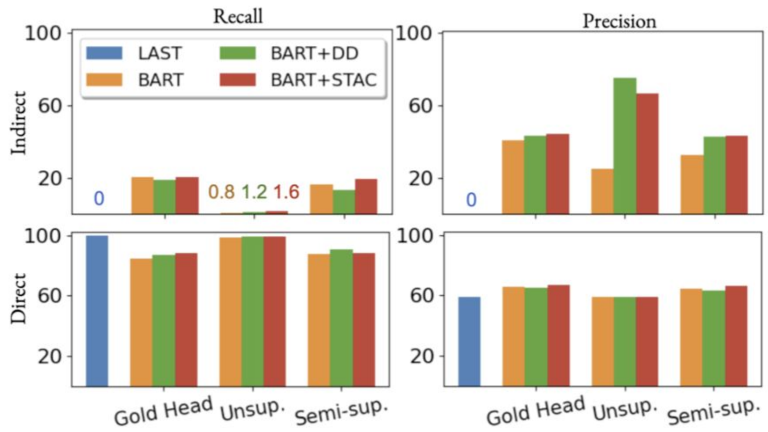


Figure 6: Comparison of recall (left) and precision (right) of indirect (top) and direct (bottom) links in LAST baseline and SO fine-tuned models on STAC.

Recall and Precision of indirect and direct edges in LAST and FT models, whole test *vs.* trees

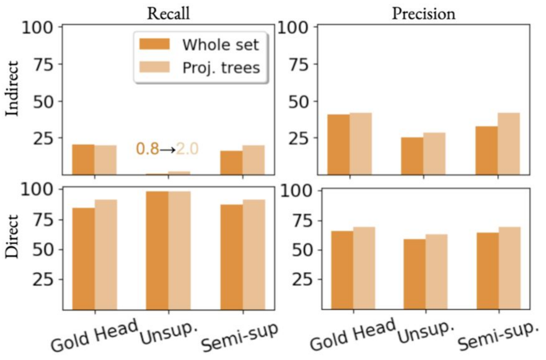


Figure 7: Recall and precision metrics in whole test set (darker color) *vs.* projective tree subset (brighter color), with BART model.

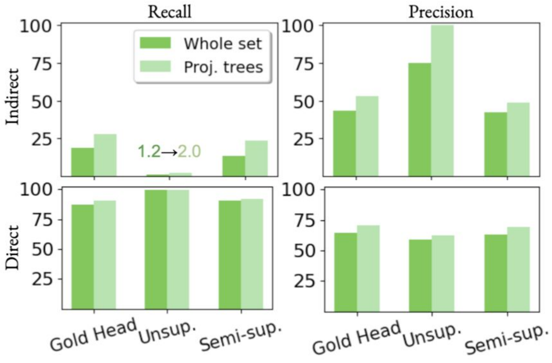


Figure 8: Recall and precision metrics in whole test set (darker color) *vs.* projective tree subset (brighter color), with BART+SO-DD model.

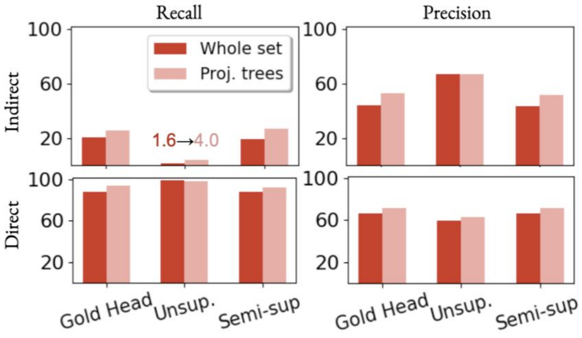


Figure 9: Recall and precision metrics in whole test set (darker color) *vs.* projective tree subset (brighter color), with model BART+SO-STAC.