Learning Reasoning Paths over Semantic Graphs for Video-grounded Dialogues

Hung Le, Nancy F. Chen, Steven C.H. Hoi

Presented at 9th International Conference on Learning Representations (ICLR 2021)

We proposed a novel framework of *Reasoning Paths in Dialogue Context* (**PDC**) to discover information flows among dialogue turns and predict reasoning paths to generate dialogue responses.

1 **Q**: is it just one person in the video ? **A**: There is one visible person , yes .

2 **Q**: what is he carrying in his hand ? **A**: **he** is looking down at his cellphone and laughing while **walking** forward in **a living room** .

3 **Q**: Is there any noise in the video ? **A**: No there is no noise in the video .

4 **Q**: can you tell if he's watching a video on his phone ? **A**: I can't tell what he's watching . **he walks** into **a table** from not paying attention .

5 Q: does **he** just **walk** back and forth in the video?

A: **he** walks towards the back of **the living room** , and walks right into **the table** .

(1) Sequential propagation

Sequential reasoning approaches fail to detect long-distance dependencies (e.g. between the current turn and the 2nd turn)

1 **Q**: is it just one person in the video ? **A**: There is one visible person , yes .

2 **Q**: what is he carrying in his hand ? **A**: **he** is looking down at his cellphone and laughing while **walking** forward in **a living room** .

3 **Q**: Is there any noise in the video ? **A**: No there is no noise in the video .

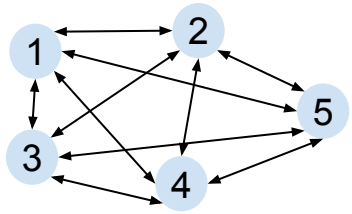
4 **Q**: can you tell if he's watching a video on his phone ? **A**: I can't tell what he's watching . **he walks** into **a table** from not paying attention .

5 Q: does **he** just **walk** back and forth in the video?

A: **he** walks towards the back of **the living room** , and walks right into **the table** .

(1) Sequential propagation

(2) Graph-based propagation



In graph-based reasoning approaches process, many irrelevant signals (e.g. from 1st and 3rd turn) are directly forwarded to the current turn.

1 **Q**: is it just one person in the video ? **A**: There is one visible person , yes .

2 **Q**: what is he carrying in his hand ? **A**: he is looking down at his cellphone and laughing while **walking** forward in **a living room** .

3 **Q**: Is there any noise in the video ? **A**: No there is no noise in the video .

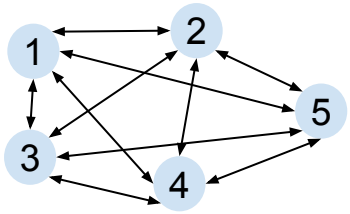
4 **Q**: can you tell if he's watching a video on his phone ? **A**: I can't tell what he's watching . he **walks** into **a table** from not paying attention .

5 Q: does he just **walk** back and forth in the video?

A: he walks towards the back of **the living room** , and walks right into **the table** .

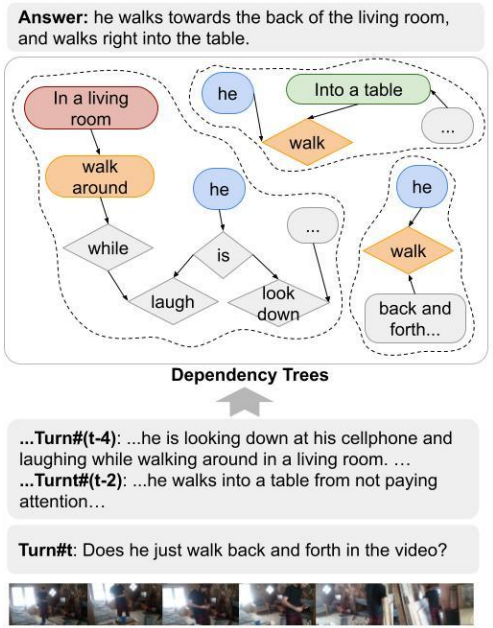
(1) Sequential propagation

(2) Graph-based propagation



(3) Path-based propagation

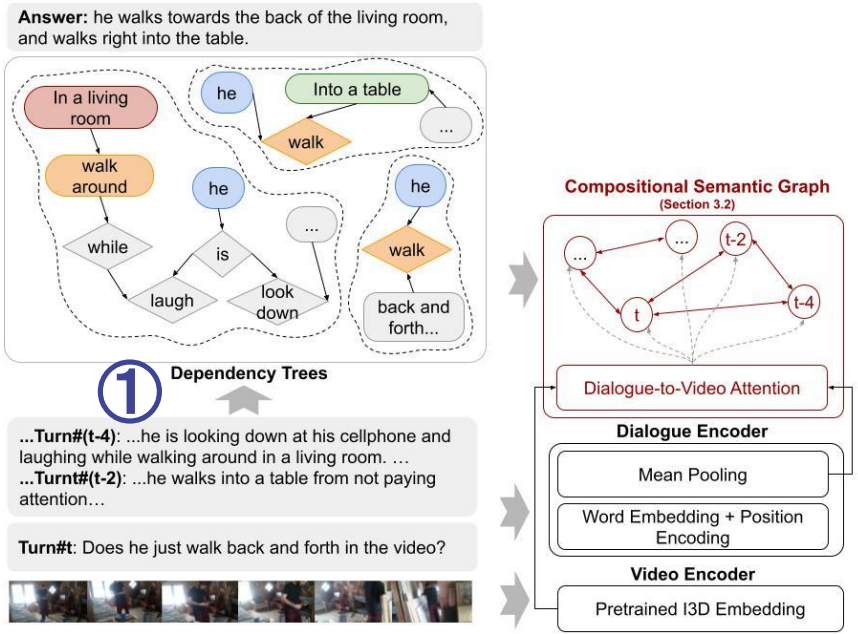
PDC learns to construct reasoning paths from past turns to current turn



①

First, each dialogue turn (question+answer) is decomposed by syntactic dependency parser

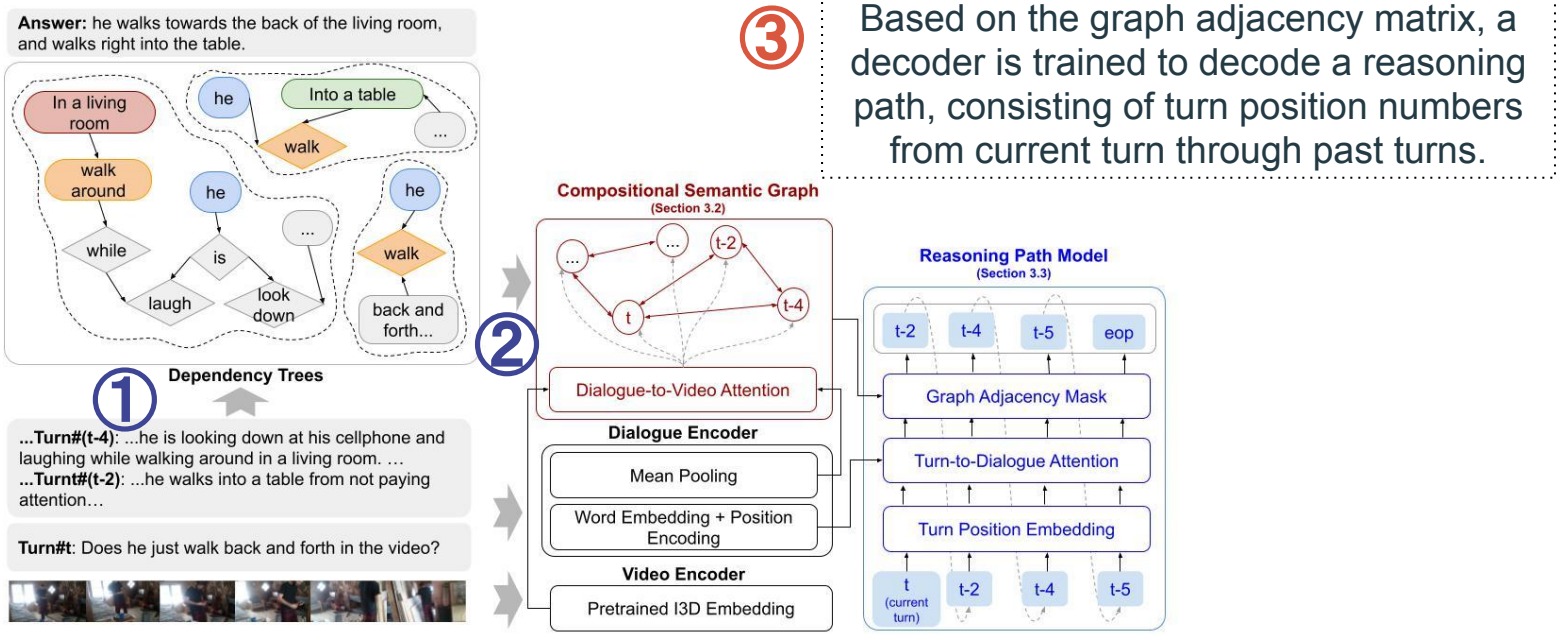
PDC learns to construct reasoning paths from past turns to current turn



②

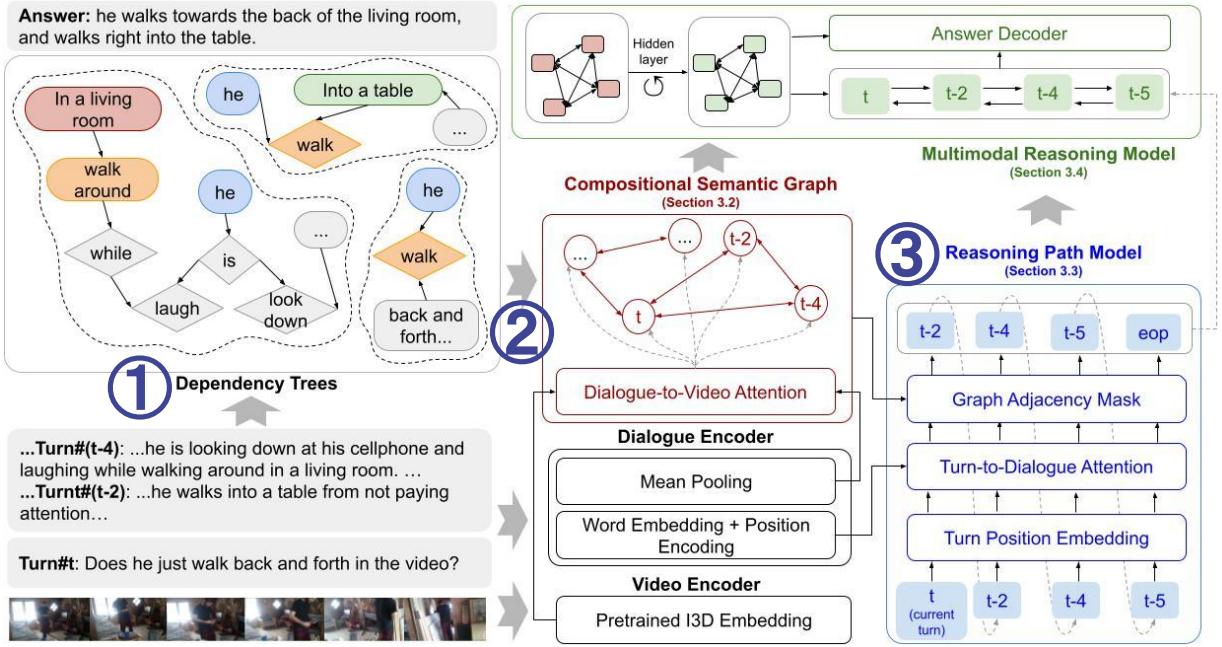
A turn-based semantic graph is built in which turns are nodes and edges connects turns that contain semantically similar subnodes

PDC learns to construct reasoning paths from past turns to current turn



Based on the graph adjacency matrix, a decoder is trained to decode a reasoning path, consisting of turn position numbers from current turn through past turns.

PDC learns to construct reasoning paths from past turns to current turn

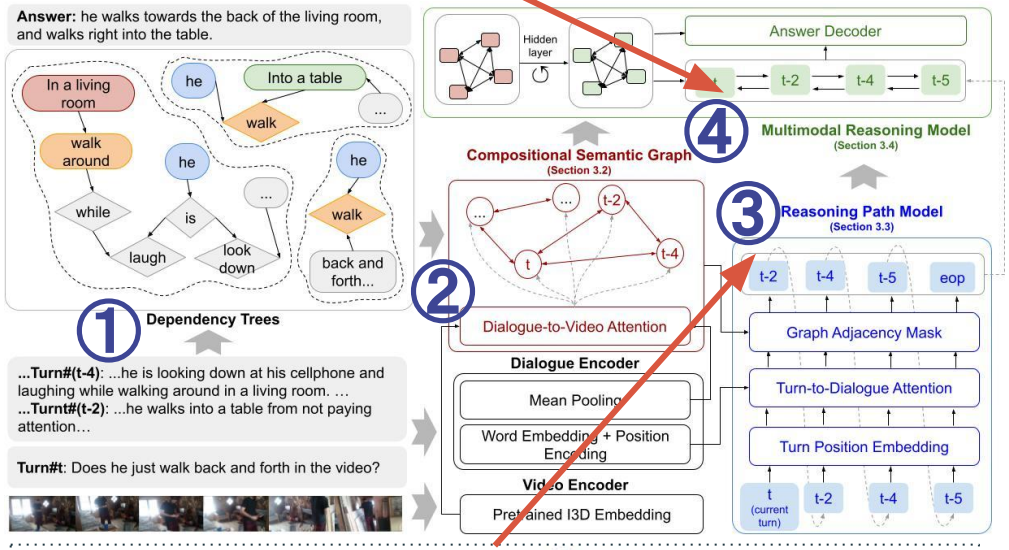


④

A recurrent network or transformer network is used to traverse dialogue turns based on the decoded reasoning path.

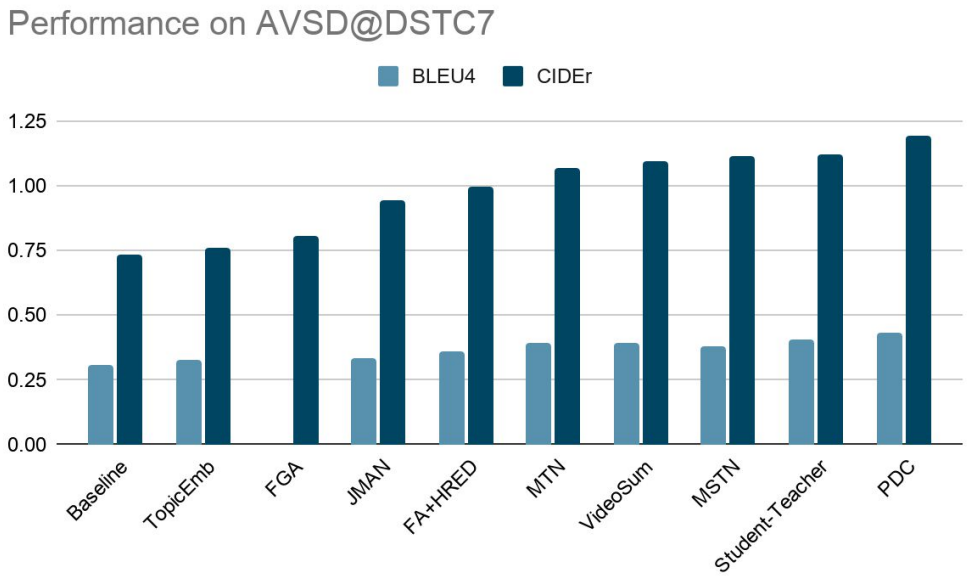
PDC learns to construct reasoning paths from past turns to current turn

$$\hat{\mathcal{A}}_t = \arg\max_{\mathcal{A}_t} P(\mathcal{A}_t|\mathcal{I}, \mathcal{C}_t, \mathcal{Q}_t; \boldsymbol{\theta}) = \arg\max_{\mathcal{A}_t} \prod_{m=1}^{L_{\mathcal{A}}} P_m(w_m|\mathcal{A}_{t,1:m-1}, \mathcal{I}, \mathcal{C}_t, \mathcal{Q}_t; \boldsymbol{\theta})$$

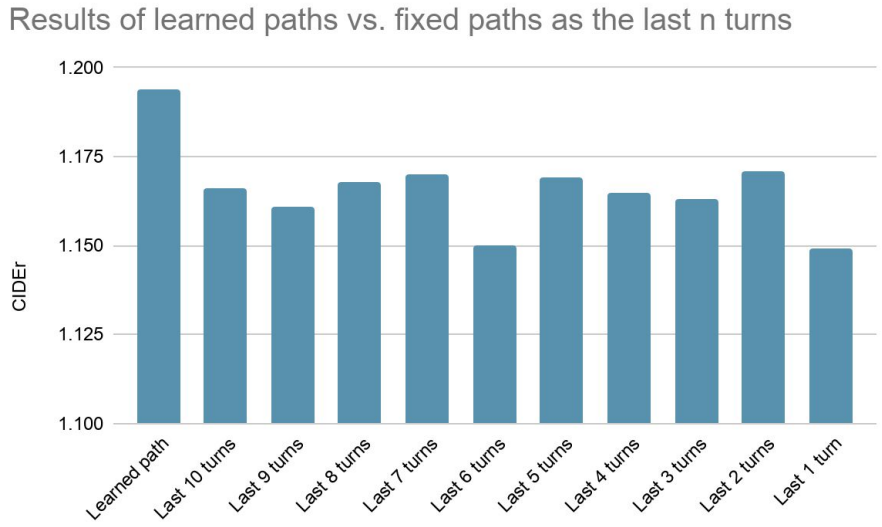


$$\hat{\mathcal{R}}_t = \arg\max_{\mathcal{R}_t} P(\mathcal{R}_t|\mathcal{C}_t, \mathcal{Q}_t; \phi) = \arg\max_{\mathcal{R}_t} \prod_{m=1}^{L_{\text{path}}} P_m(r_m|\mathcal{R}_{t,1:m-1}, \mathcal{C}_t, \mathcal{Q}_t; \phi)$$

PDC outperforms prior approaches on the AVSD benchmark



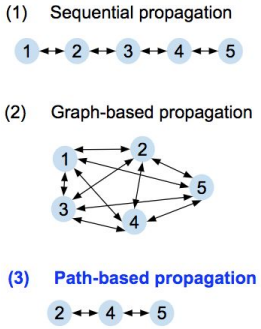
PDC can learn dynamic reasoning paths rather than using a fixed temporal-ordered path



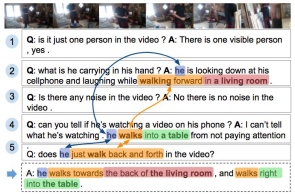
Not all information in the dialogue history is relevant.

PDC improves model transparency and is less dependent on the distribution of dialogue context size (~ 5 turns in AVSD).

Summary



PDC can learn reasoning paths to forward the most relevant contextual signals from past turns to the current turn.



PDC improves model transparency and is more dynamic to the dialogue context distribution.

Learning Reasoning Paths over Semantic Graphs for Video-grounded Dialogues

Hung Le, Nancy F. Chen, Steven C.H. Hoi

Thank you for your attention and interest in this paper!