Online Semantic Parsing for Latency Reduction in Task-Oriented Dialogue

Jiawei (Joe) Zhou, Jason Eisner, Michael Newman, Emmanouil Antonios Platanios, Sam Thomson

jzhou02@g.harvard.edu,
{jason.eisner,mike.newman,anthony.platanios,samuel.thomson}@microsoft.com
Harvard University, Microsoft Semantic Machines
Task-Oriented Dialogue

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Sure. Is this what you are looking for?
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Sure. Is this what you are looking for?
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Sure. Is this what you are looking for?

Can we start generating the program and executing it before the user finishes speaking?

Sure. Is this what you are looking for?
Online Prediction/Decision Problems

E.g.:

• Simultaneous translation
• Text Auto-completion
• Uber pool
• Etc.
Online Prediction/Decision Problems

E.g.:

- Simultaneous translation
- Text Auto-completion
- Uber pool
- Etc.

Beneficial to start making decisions before seeing all the input!
Online Prediction/Decision Problems

E.g.:
- Simultaneous translation
- Text Auto-completion
- Uber pool
- Etc.

Beneficial to start making decisions before seeing all the input!

Ours:
- Online Semantic Parsing
  - Learn the anticipation?
  - How to formally evaluate?
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Online Semantic Parsing

Assumptions:

- Execution time dominates ⇒ predict early
- Consistent parsing history unnecessary (unlike simultaneous MT) ⇒ reparse from scratch after each token (like re-translation: Arivazhagan et al., 2020)
Online Semantic Parsing

Assumptions:

- Execution time dominates ⇒ predict early
- Consistent parsing history unnecessary (unlike simultaneous MT) ⇒ reparse from scratch after each token (like *re-translation*: Arivazhagan et al., 2020)

We propose a two-step approach

- **Propose**: predict a complete graph from the current utterance prefix
- **Select**: select the graph nodes (function invocations) that are worth executing at this time
Propose a Program/Graph

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Approach (a)

LMComplete + FullToGraph

utterance prefix

⇒

full utterance

⇒

full program
Propose a Program/Graph

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Add a pool party with Barack Obama <MASK>

⇓ (fine-tuned BART)

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

⇓ (full parser)

Approach (a)

LMComplete + FullToGraph
Propose a Program/Graph

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Approach (b)

PREFIX_TO_GRAPH

utterance prefix
down
full program
Propose a Program/Graph

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Approach (b)

PrefixToGraph

Add a pool party with Barack Obama <MASK>  
⇓ (specialized parser)
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Graph-based Semantic Parser

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Graph-based Semantic Parser

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Yield Create Event
0 1
Graph-based Semantic Parser

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Yield CreatEvent -RA-(0,:arg0) subject -RA-(1,:arg0)
<str> pool party </str> -RA-(3,:arg0) start
attendees contains FindPerson “Barack Obama”
attendees contains FindPerson “Joe”
Graph-based Semantic Parser

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Yield CreateEvent -RA- (0,:arg0) subject -RA- (1,:arg0)

<str> pool party </str> -RA- (3,:arg0) start ...

FindPerson -RA- (22,:arg1) <str> Joe </str> -RA- (31,:arg0)
Graph-based Semantic Parser

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM

Model: Transformer with self-pointing mechanism, similar to Zhou et al. (2021)
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Subgraph Selection

Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Add a pool party with Barack Obama and Joe for tomorrow at 9:00 AM.
Data and Base Models

| Dataset                        | SMCalFlow | TreeDST |
|-------------------------------|-----------|---------|
| # utterances in training      | 121,024   | 121,652 |
| # utterances in validation    | 13,496    | 22,910  |
| Best reported accuracy†       | 80.4      | 88.3    |
| FULLToGRAPH accuracy          | 80.7      | 90.8    |
| Prefix BLEU (no completion)   | 38.04     | 37.54   |
| LMCOMplete BLEU               | 53.51     | 55.93   |

† both from Platanios et al. (2021)
PrefixToGraph performance on SMCalFlow validation data of varying prefix lengths
Final Latency Reduction vs. Cost

Timing measured by the number of source tokens

[SMCalFlow] Execution Time: 1.0
Best Possible Reduction: 1.85

[TreeDST] Execution Time: 1.0
Best Possible Reduction: 2.21
Final Latency Reduction vs. Cost

Faster Execution

[SMCalFlow] Execution Time: 0.2
Best Possible Reduction: 0.37

- offline full-to-graph
- prefix with full-to-graph
- LM-completion + full-to-graph
- prefix-to-graph

[TreeDST] Execution Time: 0.2
Best Possible Reduction: 0.45

- offline full-to-graph
- prefix with full-to-graph
- LM-completion + full-to-graph
- prefix-to-graph
Final Latency Reduction vs. Cost

Slower Execution

[SMCalFlow] Execution Time: 3.0
Best Possible Reduction: 4.91

- offline full-to-graph
- prefix with full-to-graph
- LM-completion + full-to-graph
- prefix-to-graph

[TreeDST] Execution Time: 3.0
Best Possible Reduction: 5.61

- offline full-to-graph
- prefix with full-to-graph
- LM-completion + full-to-graph
- prefix-to-graph
Average Latency Reduction per Function

| Function Call                      | offline latency | latency reduction |
|------------------------------------|-----------------|-------------------|
| FindEventWrapperWithDefaults       |                 |                   |
| Yield                              |                 |                   |
| DeletePreflightEventWrapper        |                 |                   |
| DeleteCommitEventWrapper           |                 |                   |
| CreatePreflightEventWrapper        |                 |                   |
| CreateCommitEventWrapper           |                 |                   |
| RecipientWithNameLike              |                 |                   |
| UpdatePreflightEventWrapper        |                 |                   |
| UpdateCommitEventWrapper           |                 |                   |
| FindManager                        |                 |                   |
| EventAttendance                    |                 |                   |
| RecipientAvailability              |                 |                   |
| FindReports                        |                 |                   |
Conclusion

• We propose a new task: Online Semantic Parsing, with a rigorous latency reduction evaluation metric

• We show it is possible to reduce latency by 30% – 63% using a strong graph-based semantic parser, either
  • trained to parse the prefix directly, or
  • combined with a language model for utterance completion

• Similar approaches could be applied to other executable semantic representations.

Thanks
Arivazhagan, N., Cherry, C., Te, I., Macherey, W., Baljekar, P., and Foster, G. (2020). Re-translation strategies for long form, simultaneous, spoken language translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7919–7923. IEEE.

Platanios, E. A., Pauls, A., Roy, S., Zhang, Y., Kyte, A., Guo, A., Thomson, S., Krishnamurthy, J., Wolfe, J., Andreas, J., and Klein, D. (2021). Value-agnostic conversational semantic parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3666–3681, Online. Association for Computational Linguistics.
Zhou, J., Naseem, T., Fernandez Astudillo, R., Lee, Y.-S., Florian, R., and Roukos, S. (2021). Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6279–6290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.