Using Planning to Improve Semantic Parsing of Instructional Texts

Vanya Cohen and **Ray Mooney**

TEXAS
The University of Texas at Austin

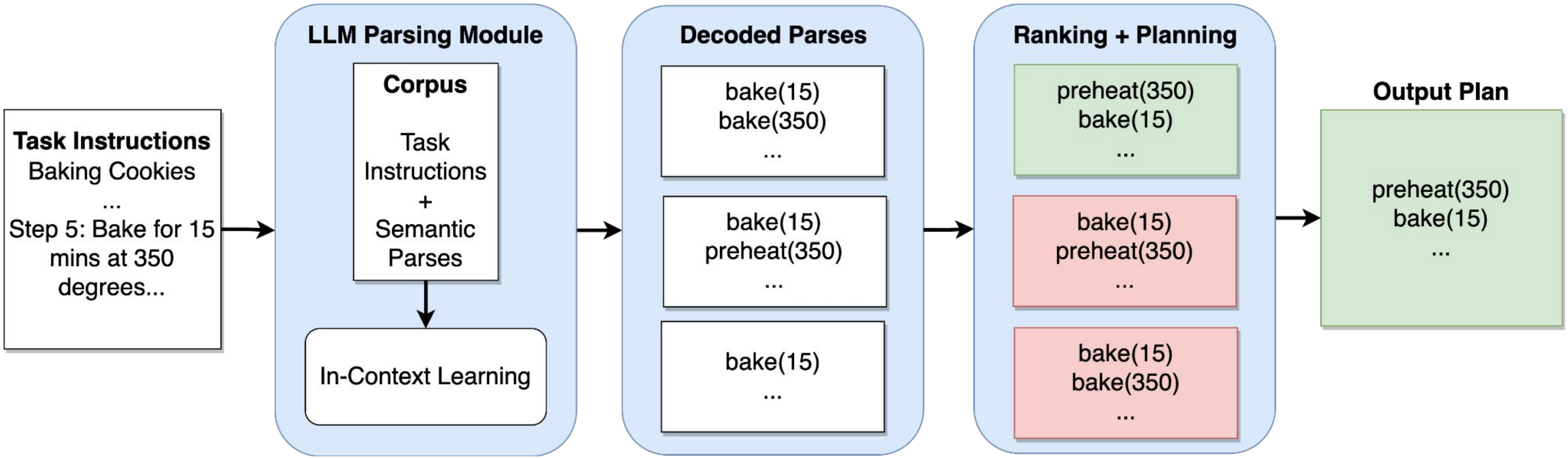
Introduction

- Few-shot semantic parsing of long-form instructional texts poses unique challenges
  - Long context dependencies and ambiguous, domain-specific language
- Utilize ***planning domain information*** to improve quality of generated ***semantic parses (plans)***
  - Add structured reasoning to LLM-based semantic parsing
- Planning Augmented Semantic Parsing
  - Symbolic-planning-based decoder
  - Ranks and corrects candidate parses
  - Combines strength of LLMs and classical AI planning

Background

- LLM for few-shot semantic parsing (Shin and Van Durme 2021)
    - **Davinci Codex**
- Datasets: cooking recipes with plans
    - Describe the steps needed to make the recipe
    - **Bolllini et al. 2013** and **Tasse and Smith 2008**
- Recipe parses are composed of actions
    - STRIPs-like operators with preconditions and postconditions
    - Actions are only ***executable*** if their ***preconditions*** are satisfied
- Want to output semantic parses which are executable, while capturing recipe semantics

Method


Task Instructions
Baking Cookies
...
Step 5: Bake for 15 mins at 350 degrees...
LLM Parsing Module
Corpus
Task Instructions + Semantic Parses
In-Context Learning
Decoded Parses
bake(15)
bake(350)
...
bake(15)
preheat(350)
...
bake(15)
...
Ranking + Planning
preheat(350)
bake(15)
...
bake(15)
preheat(350)
...
bake(15)
bake(350)
...
Output Plan
preheat(350)
bake(15)
...

Method

- Few-shot semantic parsing with in-context example selection
  - Cosine similarity with a paragraph embedding model
- Rank candidate parses (ten candidates)
  - Minimize syntax errors ***(SE)***, precondition errors ***(PE)***
  - Minimize the number of steps that need to be added to make the parsed plan executable ***(AS)***
  - Maximize the probability of all the plan steps (***ln*** $\boldsymbol{P_t}$)
- Output the highest scoring plan with added steps to make it executable (if possible)

$$score = \ln(1.0 - \frac{SE + PE + AS}{N} + \epsilon) + \frac{1}{T}\sum_{t=1}^{T} \ln P_t$$

Method

- Few-shot semantic parsing with in-context example selection
    - Cosine similarity with a paragraph embedding model
- Rank candidate parses (ten candidates)
    - Minimize syntax errors ***(SE)***, precondition errors ***(PE)***
    - Minimize the number of steps that need to be added to make the parsed plan executable ***(AS)***
    - Maximize the probability of all the plan steps (***ln*** P_t)
- Output the highest scoring plan with added steps to make it executable (if possible)

$$score = \boxed{\ln(1.0 - \frac{SE + PE + AS}{N} + \epsilon)} + \boxed{\frac{1}{T}\sum_{t=1}^{T} \ln P_t}$$

Maximize the plan probability

Minimize the fraction of plan steps with errors, and added steps

Experiments

- Evaluation Metrics
  - Longest Common Subsequence (LCS)
  - Plan Steps F1: harmonic mean of precision and recall of generated steps
  - Precondition Errors (PE) and Syntax Errors (SE)

- Experimental Settings
  - Rank (PPL): selects the plan with the lowest perplexity
  - Rank: ranks the plans by the scoring function without correcting precondition errors
  - Rank + Plan: our full ranking method with planning to correct errors

Results (Bollini et al. 2013)

Models	(Bollini et al., 2013)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.897 ± 0.008	0.962 ± 0.685	0.042 ± 0.008	0.784 ± 0.004
Davinci Codex, E=5	0.949 ± 0.005	0.198 ± 0.009	0.002 ± 0.004	0.863 ± 0.003
Re-Rank				
Davinci Codex, E=1	0.901 ± 0.008	0.382 ± 0.037	0.025 ± 0.008	0.798 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	0.120 ± 0.015	0.002 ± 0.004	$\mathbf{0.868 \pm 0.002}$
Re-Rank + Plan				
Davinci Codex, E=1	0.903 ± 0.008	0.143 ± 0.033	0.025 ± 0.008	0.807 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	$\mathbf{0.033 \pm 0.000}$	0.002 ± 0.004	$\mathbf{0.870 \pm 0.002}$

Results (Bollini et al. 2013)

Models	(Bollini et al., 2013)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.897 ± 0.008	0.962 ± 0.685	0.042 ± 0.008	0.784 ± 0.004
Davinci Codex, E=5	0.949 ± 0.005	0.198 ± 0.009	0.002 ± 0.004	0.863 ± 0.003
Re-Rank				
Davinci Codex, E=1	0.901 ± 0.008	0.382 ± 0.037	0.025 ± 0.008	0.798 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	0.120 ± 0.015	0.002 ± 0.004	$\mathbf{0.868 \pm 0.002}$
Re-Rank + Plan				
Davinci Codex, E=1	0.903 ± 0.008	0.143 ± 0.033	0.025 ± 0.008	0.807 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	$\mathbf{0.033 \pm 0.000}$	0.002 ± 0.004	$\mathbf{0.870 \pm 0.002}$

Results (Bollini et al. 2013)

Models	(Bollini et al., 2013)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.897 ± 0.008	0.962 ± 0.685	0.042 ± 0.008	0.784 ± 0.004
Davinci Codex, E=5	0.949 ± 0.005	0.198 ± 0.009	0.002 ± 0.004	0.863 ± 0.003
Re-Rank				
Davinci Codex, E=1	0.901 ± 0.008	0.382 ± 0.037	0.025 ± 0.008	0.798 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	0.120 ± 0.015	0.002 ± 0.004	$\mathbf{0.868 \pm 0.002}$
Re-Rank + Plan				
Davinci Codex, E=1	0.903 ± 0.008	0.143 ± 0.033	0.025 ± 0.008	0.807 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	$\mathbf{0.033 \pm 0.000}$	0.002 ± 0.004	$\mathbf{0.870 \pm 0.002}$

Results (Tasse and Smith 2008)

Models	(Tasse and Smith, 2008)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.692 ± 0.003	0.827 ± 0.086	0.875 ± 0.199	0.443 ± 0.002
Re-Rank				
Davinci Codex, E=1	0.695 ± 0.004	0.293 ± 0.016	$\mathbf{0.226 \pm 0.024}$	0.446 ± 0.001
Re-Rank + Plan				
Davinci Codex, E=1	0.695 ± 0.003	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.237 \pm 0.018}$	0.446 ± 0.001

Results (Tasse and Smith 2008)

Models	(Tasse and Smith, 2008)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.692 ± 0.003	0.827 ± 0.086	0.875 ± 0.199	0.443 ± 0.002
Re-Rank				
Davinci Codex, E=1	0.695 ± 0.004	0.293 ± 0.016	$\mathbf{0.226 \pm 0.024}$	0.446 ± 0.001
Re-Rank + Plan				
Davinci Codex, E=1	0.695 ± 0.003	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.237 \pm 0.018}$	0.446 ± 0.001

Results (Tasse and Smith 2008)

Models	(Tasse and Smith, 2008)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.692 ± 0.003	0.827 ± 0.086	0.875 ± 0.199	0.443 ± 0.002
Re-Rank				
Davinci Codex, E=1	0.695 ± 0.004	0.293 ± 0.016	$\mathbf{0.226 \pm 0.024}$	0.446 ± 0.001
Re-Rank + Plan				
Davinci Codex, E=1	0.695 ± 0.003	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.237 \pm 0.018}$	0.446 ± 0.001

Conclusion and Future Work

- Our neuro-symbolic approach generates semantic parses with more valid plans
- Reduces precondition errors while maintaining content similarity to ground-truth plans
- Future work:
    - Automatically generating planning domain definitions
    - Testing in other planning domains (e.g. ALFRED)
    - Use more capable LLMs (GPT-4)