

Guided K-best Selection for Semantic Parsing Annotation

Anton Belyy^{*1}, Chieh-Yang Huang^{*2}, Jacob Andreas³, Emmanouil Antonios Platanios³, Sam Thomson³, Richard Shin³, Subhro Roy³, Aleksandr Nisnevich³, Charles Chen³, Benjamin Van Durme³

¹Johns Hopkins University, ²Pennsylvania State University, ³Microsoft Semantic Machines ¹abel@jhu.edu, ²chiehyang@psu.edu, ³sminfo@microsoft.com

* Equal Contribution. Work performed during an internship at Microsoft Semantic Machines.

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN



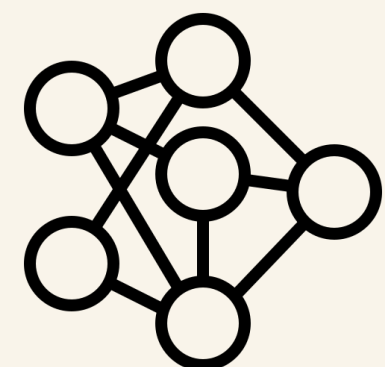
Scan for the
Demo Video

Introduction

- Collecting data for **Conversational Semantic Parsing (CSP)** is time-consuming and demanding.
- K-best selection approach to help?
 - Generate a set of candidates.
 - Ask annotators to traverse the set and select the correct parse.
- How to improve the annotation **speed** and the **accuracy**?
→ **Guided K-best selection.**

Natural utterance

When's the lecture scheduled for in May?



Model trained on 1k Dialogues:
Accuracy@1 = 0.63

Low-Resource Semantic Parsing Model

Guided K-best Selection

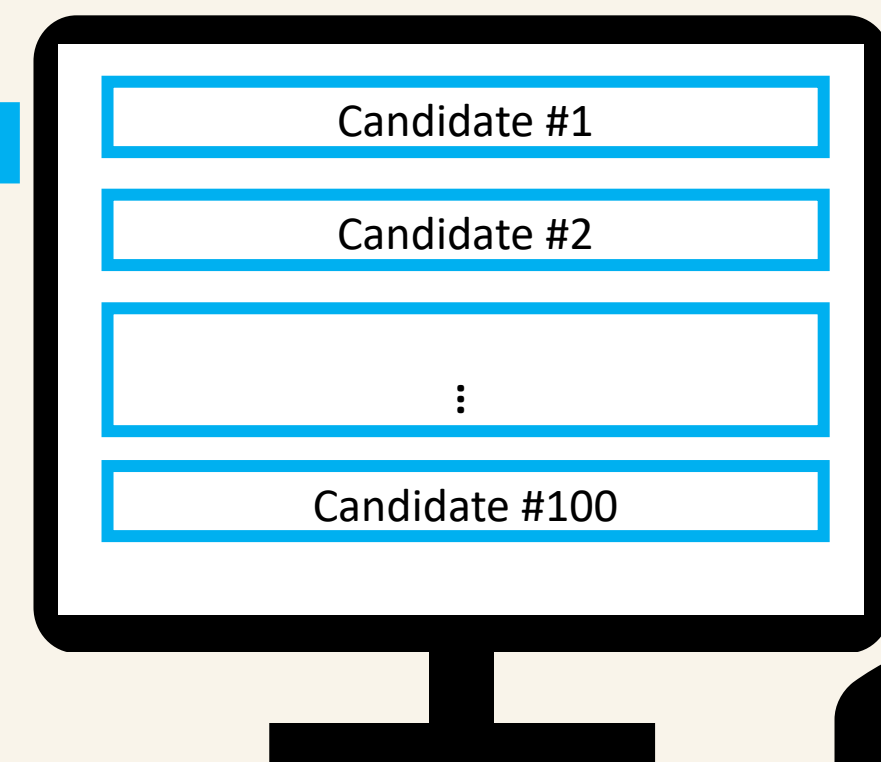
Canonical utterance

start time of find event called something like "lecture" during May

Interchangeable using an SCFG

```
(Yield :output (:start (singleton  
:results (FindEventWrapperWithDefaults  
:constraint (EventDuringRange :event  
(Constraint[Event] :subject (?~  
#(String "lecture")))) :range  
(FullMonthofMonth :month #("MAY"))))))
```

Meaning representation



Model trained on 1k Dialogues + K-Best
UI: Accuracy@1 = 0.74

Guided K-best Selection Interfaces

A Context

User: Can you change choir practice to be next Tuesday after 11 am?
Agent: How is this?
User: Make that later in the evening.
Agent: How about now?
User: No, I need choir practice to be scheduled later, maybe 6:00 pm or later.
Agent: Is this the update you want?
Target User Utterance: No I need it to start either 6 pm or later.

B Top-5 Candidates

☐ Change my request so the event is starting 6 PM
☐ Change my request so the event is starting around 6 PM
☐ Change my request so the event is starting after 6 PM
☐ Change my request so the event is starting 6 PM ending top PM
☐ Change my request so the event is starting before 6 PM

C Canonical Utterance

Suggested Keywords: around after updated starting around starting after
Chosen Keywords: starting ending

6 pm

D This input is NOT valid.

E SKIP I CAN'T FIND THE ANSWER SUBMIT

C-1 no-kbest

Change my request so the updated event is starting

C-2 scroll

Change my request so the event is starting at noon

Change my request so the event is starting 6 PM
Change my request so the event is starting around 6 PM
Change my request so the event is starting after 6 PM
Change my request so the event is starting 6 PM ending top PM
Change my request so the updated event is starting 6 PM

C-3 autocomplete

Change my request so the event is starting 6 PM

Change my request so the event is starting is starting
is an
is ending

C-4 search

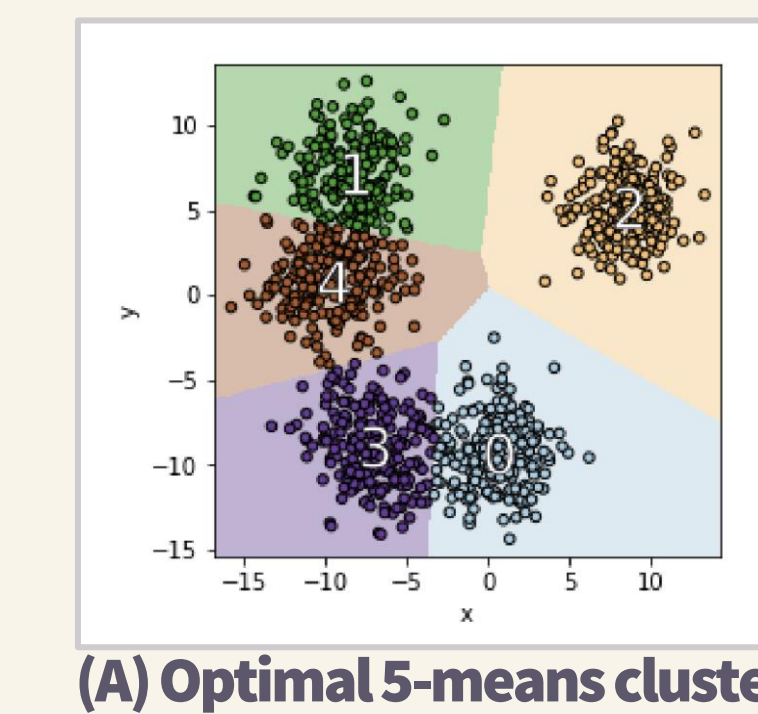
event after

Change my request so the event is starting after 6 PM
Change my request so the event is starting after 6 PM ending top PM
Change my request so the event is starting after 6 PM ending top PM
Change my request so the event is starting after 6 PM ending top PM
Change my request so the event is starting after 6 PM ending top PM

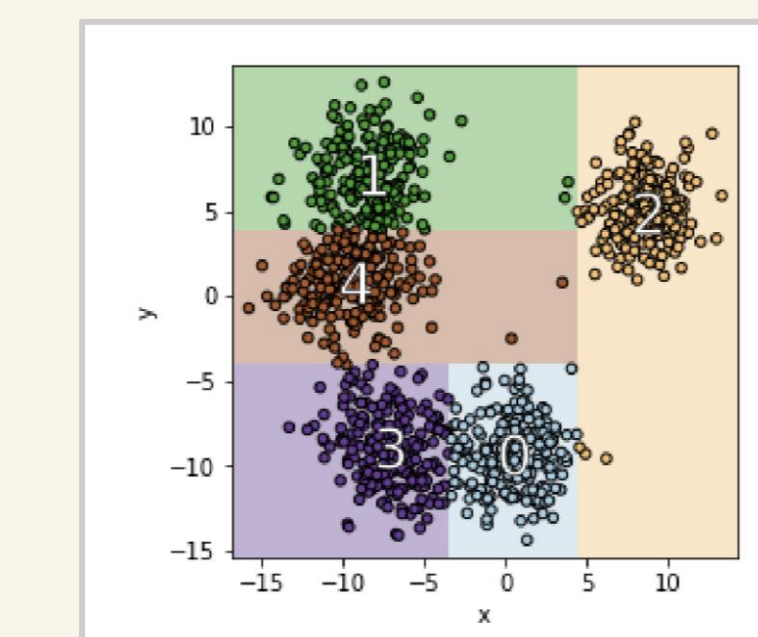
- 5 variants** of guided K-best selection interface.
- no-kbest** **C-1** and **scroll** **C-2** are baselines.
- autocomplete** **C-3** allows users to get suggestion autoregressively.
- search** **C-4** allows users to query candidates in arbitrary order.
- search-keywords** **C** extends **search** by showing 5 discriminative keywords. Users can choose to include (+) or exclude (-) the keywords.
- D** indicates whether the current input is grammatical or not.

Keyword Suggestion

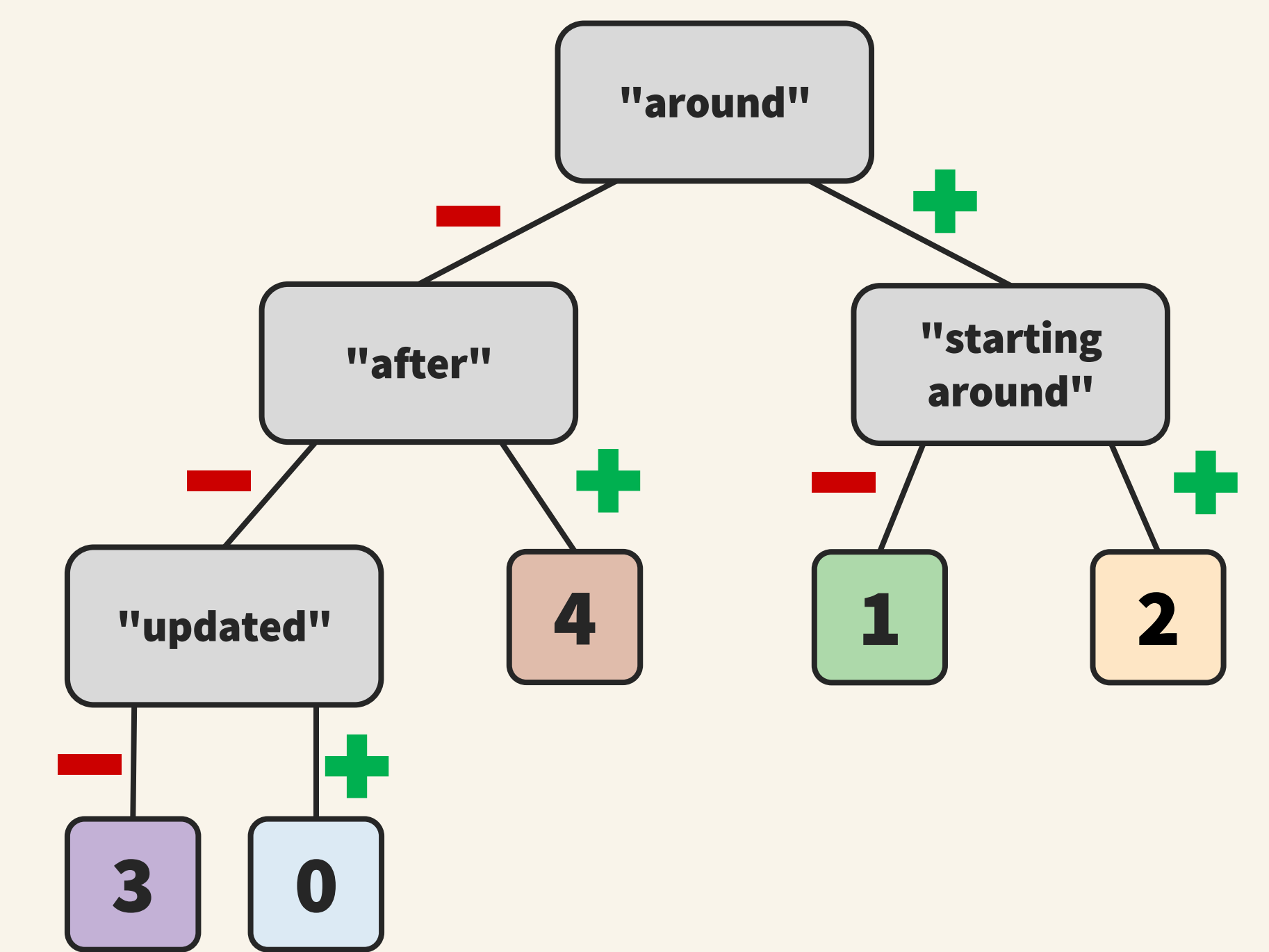
- Given a **K-best list**, we extract keywords by:
 - Apply k-means to obtain **k clusters**.
 - We choose one candidate to represent each cluster, resulting **k diverse candidates**.
 - Using n-gram (n=1, 2, 3) as features, we employ a cluster explanation technique [1] to distill the k diverse candidates into **k' keywords**. This can be shown as a binary tree (see the example below).



(A) Optimal 5-means cluster



(B) Tree based 5-means cluster



(C) Threshold tree

[1] Sanjoy Dasgupta, et al. Explainable k-means and k-medians clustering. ICML 2020.

Experiment - Interface Comparison

[Data]

- We sampled **300 utterances** from SMCaFlow [2].
- VACSP-1k [3] is used to generate **K=100 candidate parses**.
- Stratified sampling is used to control the distribution of gold answers.

	Top-5	Top-20	Top-100	Escalate
Stratified	25%	25%	25%	25%
True	76%	6%	4%	14%

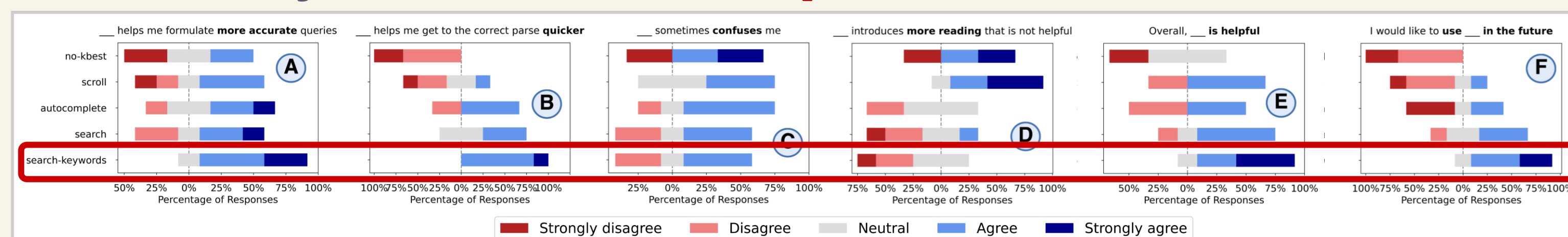
[Result]

- Autocomplete** achieves the **highest accuracy**.
- Search** helps **reduce time usage** up to 35% compared to **Scroll**.
- Search-keywords** strikes the **balance** between accuracy and time usage.

	Exact Match Accuracy ↑							Median Time (sec) ↓				
	Top-5	Top-20	Top-100	Escalate	Escalate _m	All	True	Top-5	Top-20	Top-100	Escalate	All
No-KBest	.411	.189	.123	.400	.067	.197	.339	56.13	73.17	97.48	74.29	69.43
Scroll	.880	.320	.213	.453	.067	.370	.706	13.00	25.84	26.47	30.23	24.73
Autocomplete	.919	.370	.333	.427	.067	.422	.743	13.71	26.01	30.02	31.47	25.53
Search	.878	.320	.213	.400	.080	.373	.707	8.48	19.09	17.16	19.55	16.02
Search-Keywords	.880	.419	.213	.480	.093	.401	.716	12.78	24.51	36.26	31.15	23.91

[User Feedback]

- Search-keywords** is rated as the **top choice** across all criteria.



Experiment – Guidance Comparison

[Experiment]

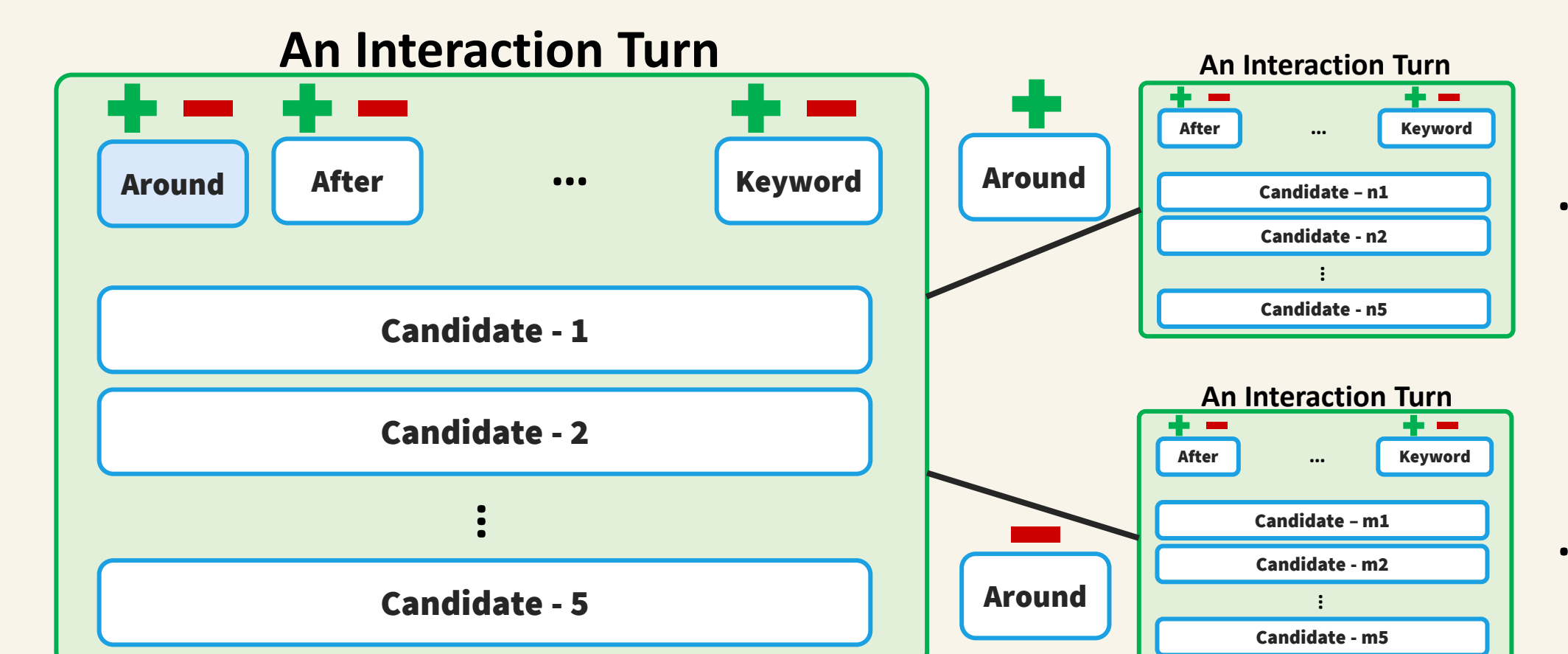
- Compared Keyword Suggestion with PDC algorithm [4].

[Oracle Simulation]

- Optimal Solution.
- KS and PDC significantly **reduce the number of turns** in Top-100.
- Adding explanations (KS) **doesn't hurt the performance**.

[Human Annotation]

- Adding explanations (KS) can further **reduce the annotation time** compared to PDC.



	Oracle simulation results (k = 5) Average number of turns ↓				Human annotation results (k = 5) Median time (sec) ↓			
	Top-5	Top-20	Top-100	All	Top-5	Top-20	Top-100	All
KS (ours)	1.10	2.39	2.80	1.24	15.30	46.99	48.20	36.71
PDC (k-means, canonical)	1.11	2.40	2.84	1.24	25.09	73.23	55.42	52.94
PDC (agglomerative, canonical)	1.16	2.73	1.31	1.31	—	—	—	—
PDC (agglomerative, meaning)	1.15	2.68	2.91	1.29	—	—	—	—
Scroll	1.00	2.63	7.75	1.33	24.18	42.30	56.37	37.21

[2] Semantic Machines, et al. "Task-oriented dialogue as dataflow synthesis." TACL 2020.

[3] Platanios, Emmanouil Antonios, et al. "Value-Agnostic Conversational Semantic Parsing." ACL 2021.

[4] Ippolito, Daphne, et al. "Comparison of diverse decoding methods from conditional language models." ACL (2019)