Guided K-best Selection for Semantic Parsing Annotation

Anton Belyy*[1], Chieh-Yang Huang*[2], Jacob Andreas[3], Emmanouil Antonios Platanios[3], Sam Thomson[3], Richard Shin[3], Subhro Roy[3], Aleksandr Nisnevich[3], Charles Chen[3], Benjamin Van Durme[3]

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

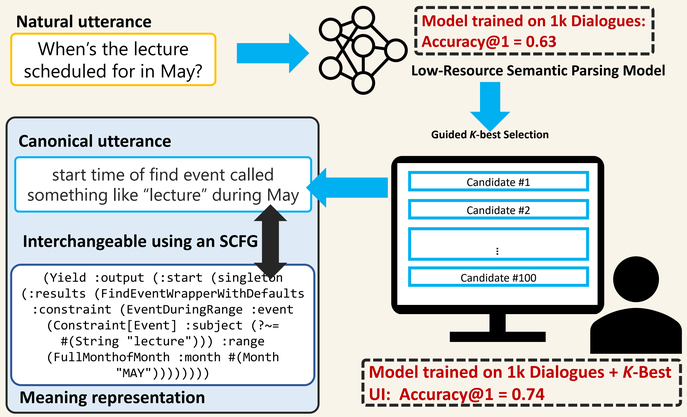
* Equal Contribution. Work performed during an internship at Microsoft Semantic Machines.

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

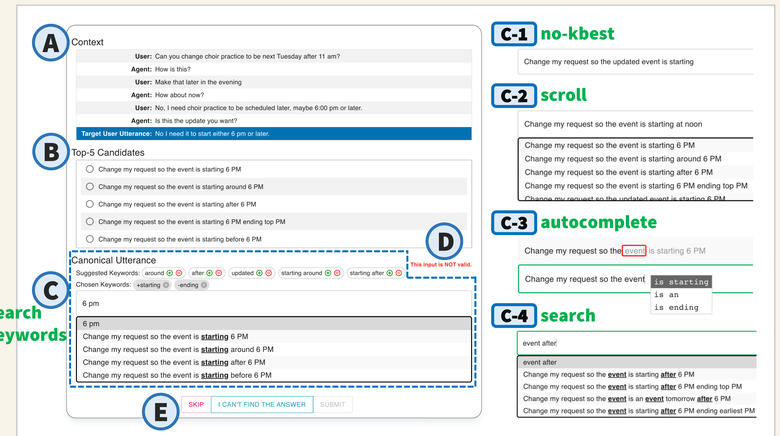
Scan for the Demo Video

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.**



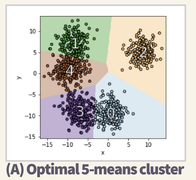
Guided K-best Selection Interfaces



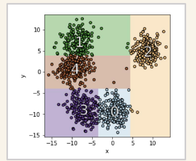
- **5 variants** of guided K-best selection interface.
  - **no-kbest** C-1 and **scroll** C-2 are baselines.
  - **autocomplete** C-3 allows users to get suggestion autoregressively.
  - **search** C-4 allows users to query candidates in arbitrary order.
  - **search-keywords** (C) extends **search** 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

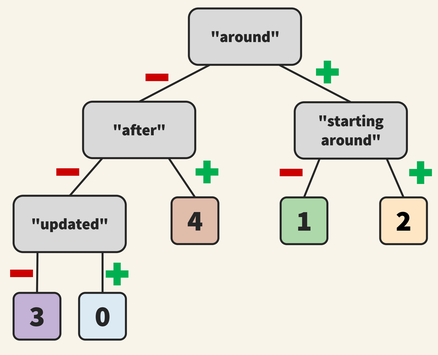
- Given a **K-best list**, we extract keywords by:
  1. Apply k-means to obtain **k clusters**.
  2. We choose one candidate to represent each cluster, resulting **k diverse candidates**.
  3. Using n-gram (n=1, 2, 3) as features, we employ a cluster explanation technique [1] to distill the k diverse 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 SMCalFlow [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%

[Participants]

- A total of 5 participants joined the experiment.
- Participants where randomly assigned to a particular interface and data split.

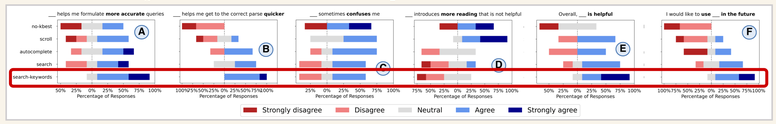
[Result]

- **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.



[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.

Experiment – Guidance Comparison

[Experiment]

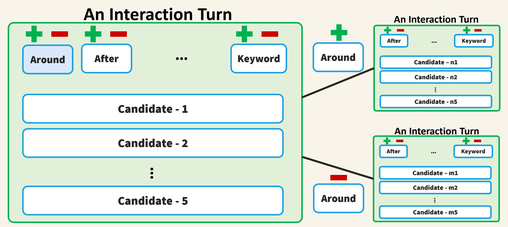
- Compared Keyword Suggestion with PDC algorithm [4].

[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	3.10	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

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