Guided *K*-best Selection for Semantic Parsing Annotation

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

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

Semantic Parsing

User utterance

When's the lecture scheduled for in May?

Canonical utterance

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

SCFG

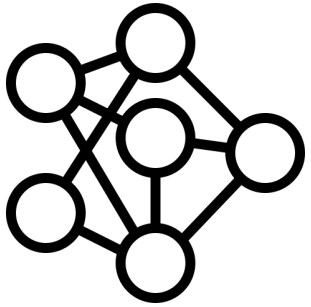
Meaning representation

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

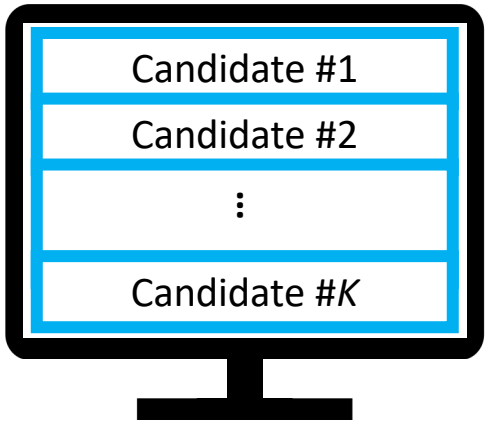
Semantic Parsing Annotation via *K*-best

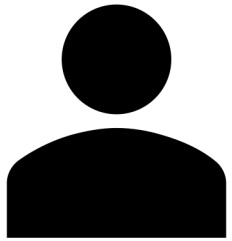
User utterance

When's the lecture scheduled for in May?



Low-Res (1k instances) Semantic Parsing Model





Canonical utterance

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

SCFG

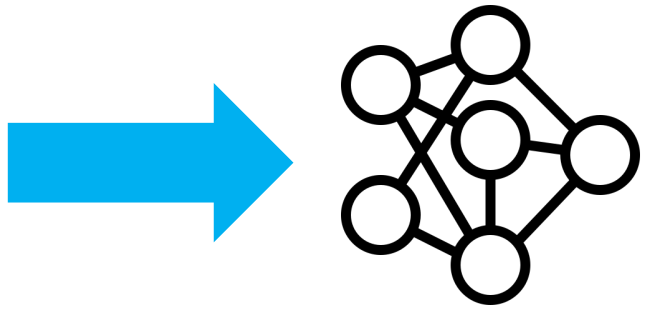
Meaning representation

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

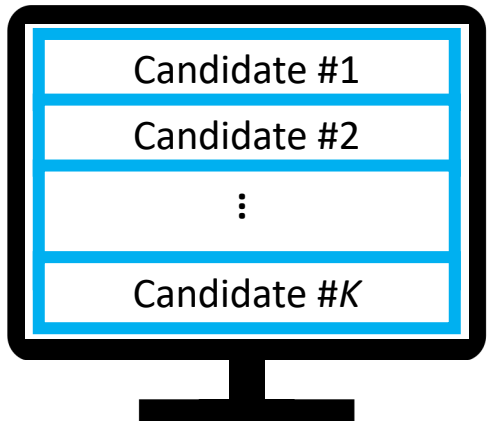
Semantic Parsing Annotation via Guided *K*-best

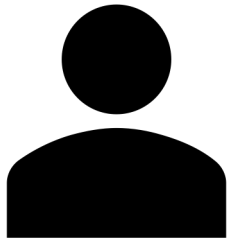
User utterance

When's the lecture scheduled for in May?



Low-Res (1k instances) Semantic Parsing Model



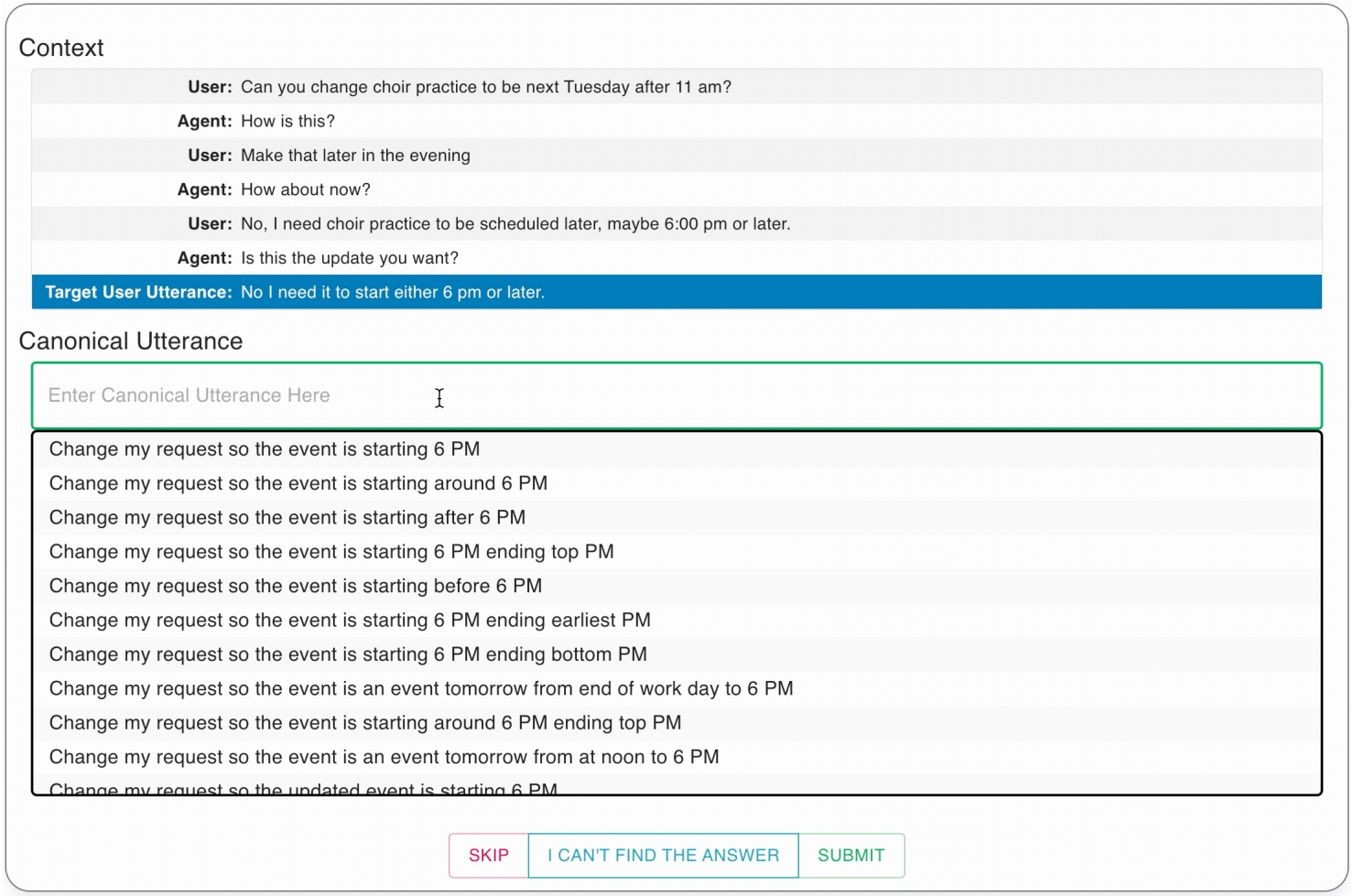


Canonical utterance

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

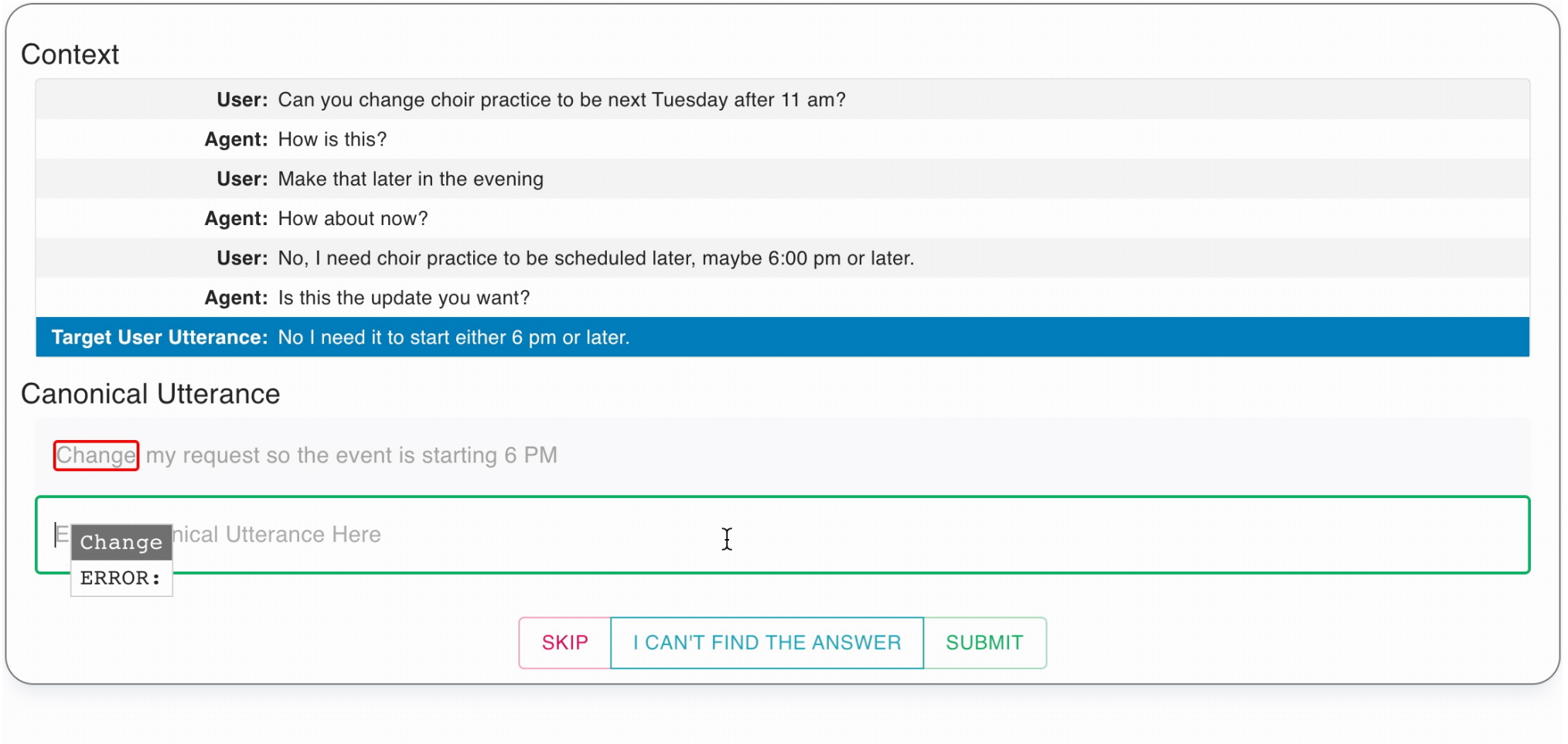
When *K* gets large, can we **guide** annotators to select the correct parse both **fast** and **accurately**?

Annotation Interface: Scroll

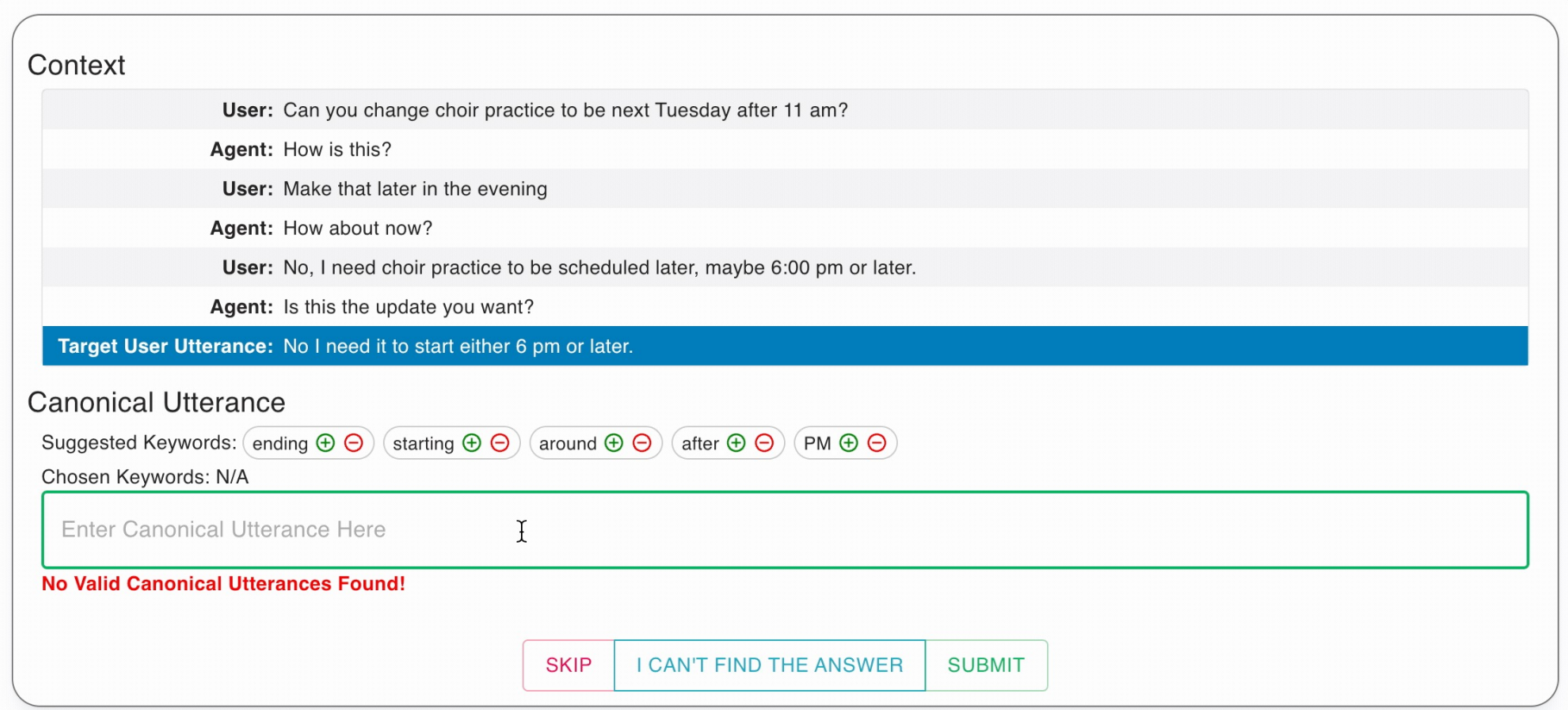


Example taken from SMCalFlow (Semantic Machines et al., 2020) dev set.

Annotation Interface: Autocomplete

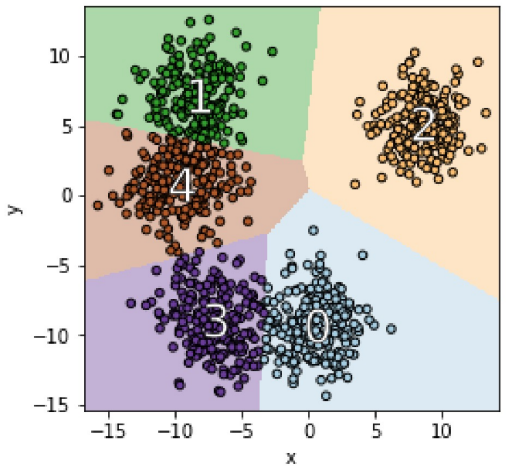


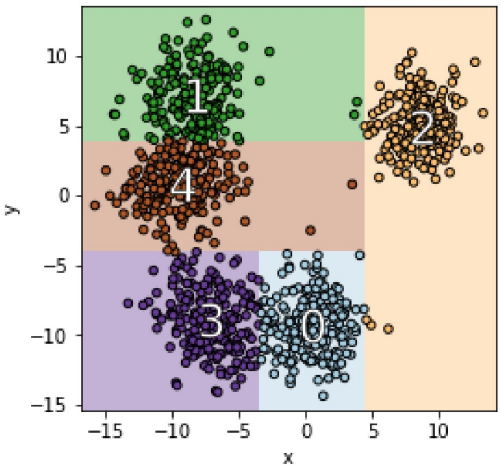
Annotation Interface: Search

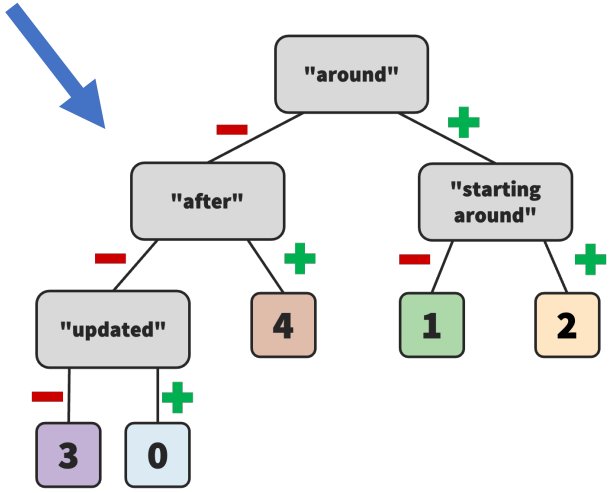


Guiding Annotators via Keyword Suggestion

- Inspired by clustering-based methods for diverse paraphrasing (Hu et al, 2019)
- Instead of picking a cluster, users choose to include (+) or exclude (-) **keywords**
- We use *explainable k-means* (Dasgupta et al, 2020) to distill *k* clusters into ***k' < k* keywords,** which are generated from the intermediate nodes of the explanation tree







Hu et al. Large-scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering. CoNLL 2019
Dasgupta et al. Explainable k-Means and k-Medians Clustering. ICML 2020

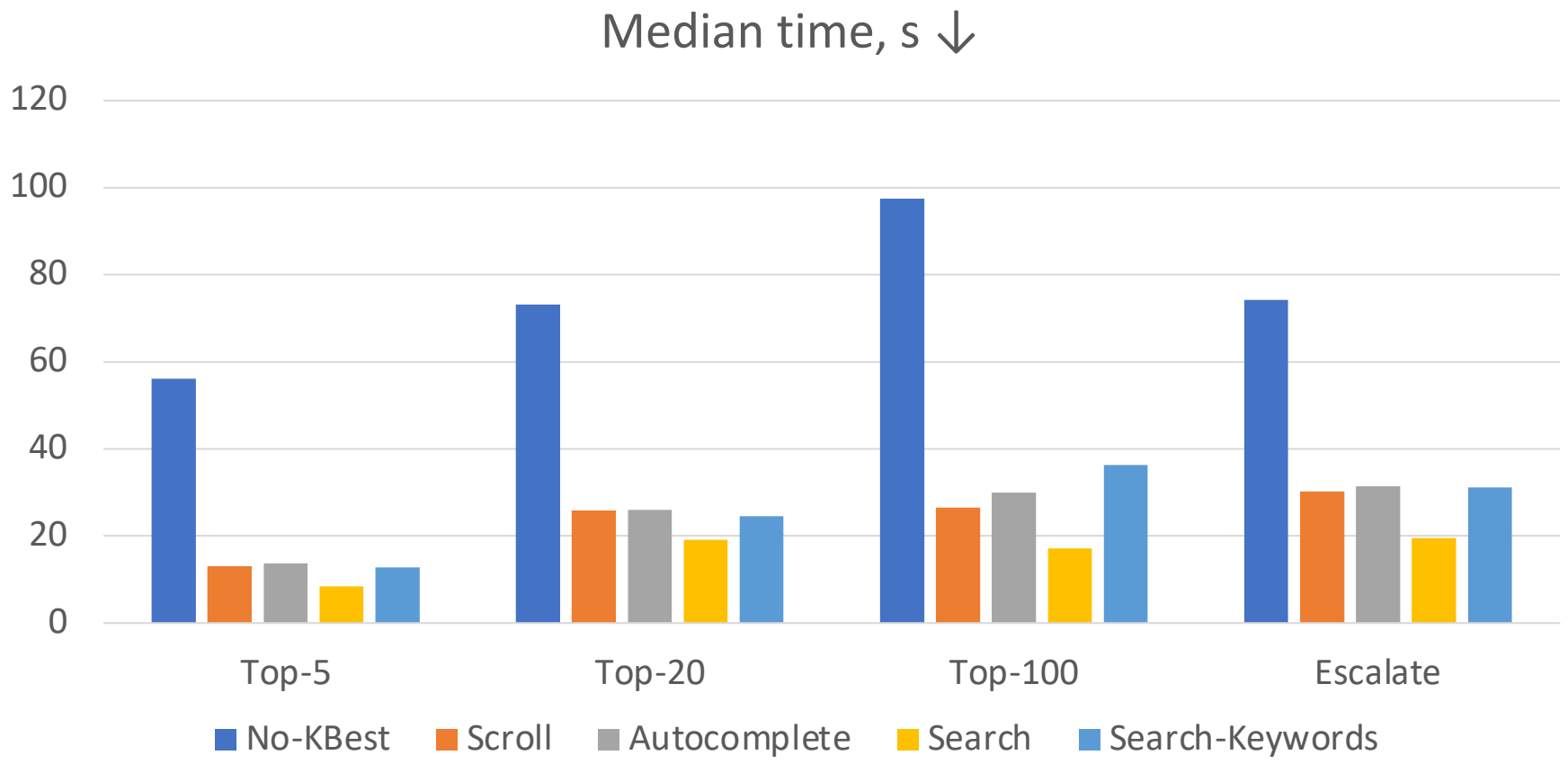
Experimental Settings

- **Interface comparison**
  - **Model**: VACSP (Platanios et al, 2021), generates *K*=100 best parses
    - **Train set**: SMCalFlow (Semantic Machines et al, 2020), 1000 utterances (sampled uniformly at random)
    - **Eval set**: SMCalFlow (Semantic Machines et al, 2020), 300 utterances (Top-5, Top-20, Top-100, Escalate)
  - **Participants**: 5 annotators, randomly assigned to the specific (UI, data split)

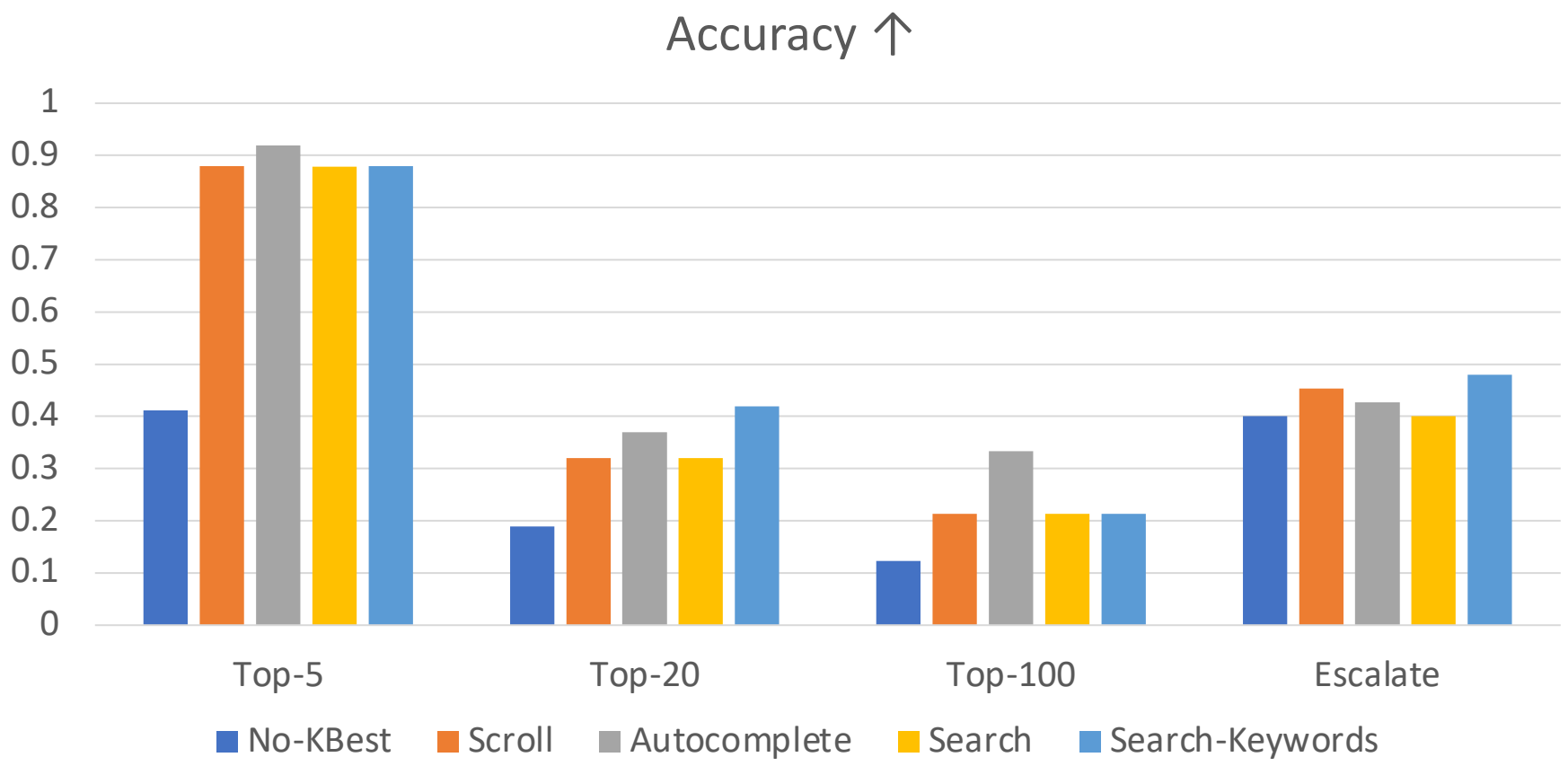
- **Guidance comparison**
  - **Model**: VACSP (Platanios et al, 2021), generates *K*=100 best parses
    - **Train set**: SMCalFlow (Semantic Machines et al, 2020), 1000 utterances (sampled uniformly at random)
    - **Eval set**: SMCalFlow (Semantic Machines et al, 2020), 12000 utterances

Platanios et al. Value-Agnostic Conversational Semantic Parsing. ACL 2021
Semantic Machines et al. Task-Oriented Dialogue as Dataflow Synthesis. TACL 2020

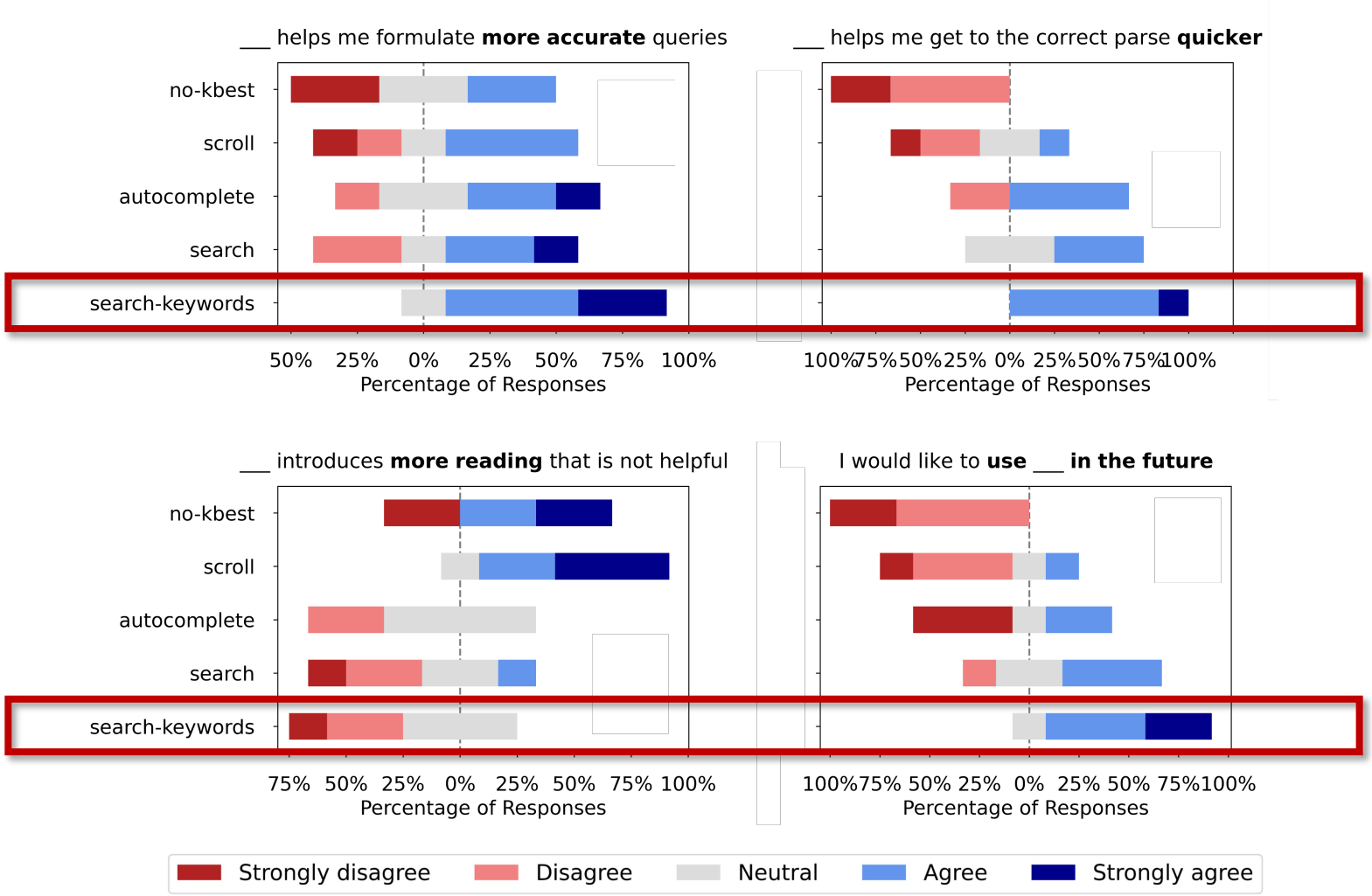
Results: Annotation Speed



Results: Annotation Accuracy



Results: Annotators' Feedback



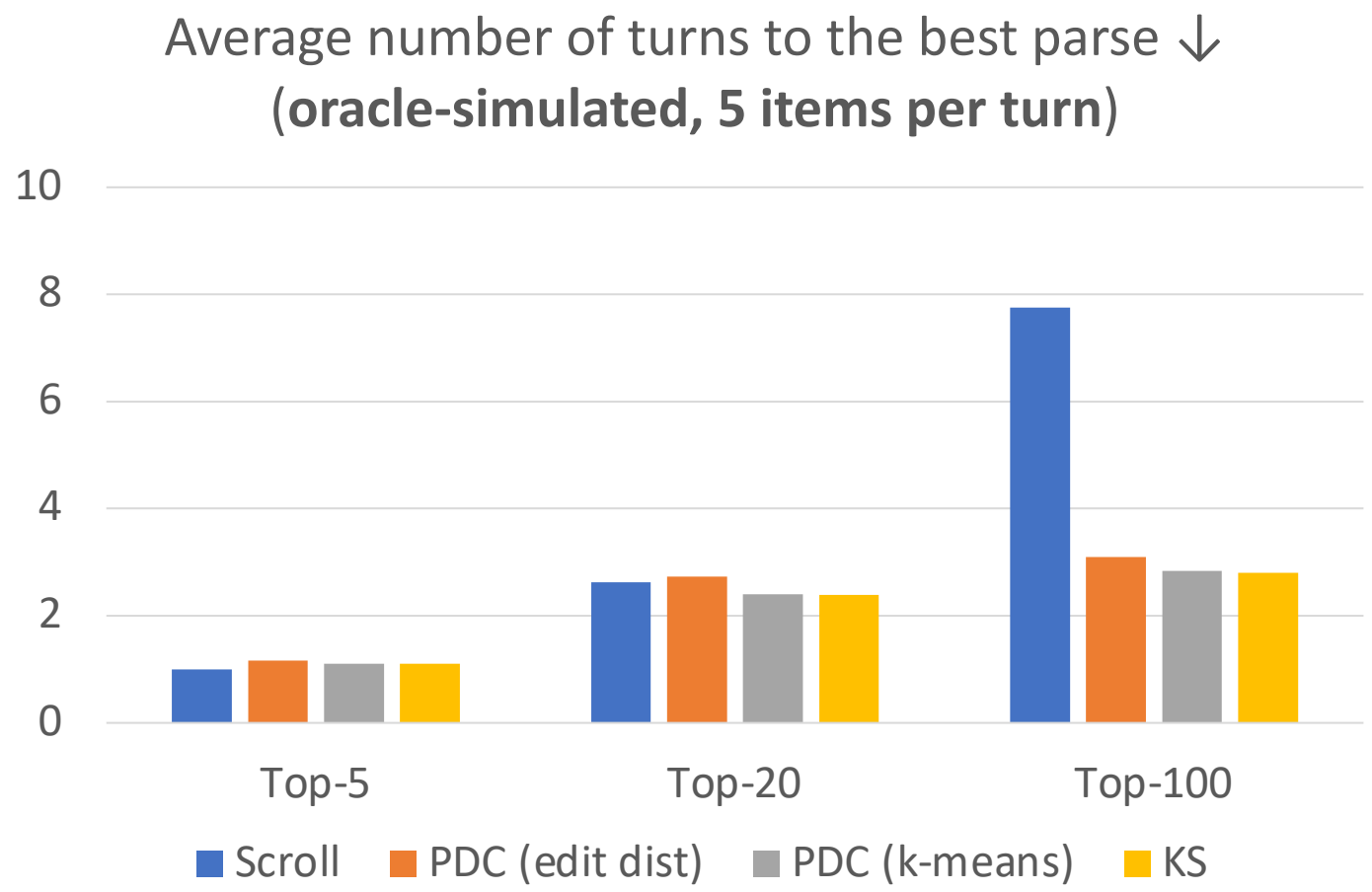
Experimental Settings

- **Interface comparison**
  - **Model**: VACSP (Platanios et al, 2021), generates *K*=100 best parses
    - **Train set**: SMCalFlow (Semantic Machines et al, 2020), 1000 utterances (sampled uniformly at random)
    - **Eval set**: SMCalFlow (Semantic Machines et al, 2020), 300 utterances (Top-5, Top-20, Top-100, Escalate)
  - **Participants**: 5 annotators, randomly assigned to the specific (UI, data split)

- **Guidance comparison**
  - **Model**: VACSP (Platanios et al, 2021), generates *K*=100 best parses
    - **Train set**: SMCalFlow (Semantic Machines et al, 2020), 1000 utterances (sampled uniformly at random)
    - **Eval set**: SMCalFlow (Semantic Machines et al, 2020), 12000 utterances

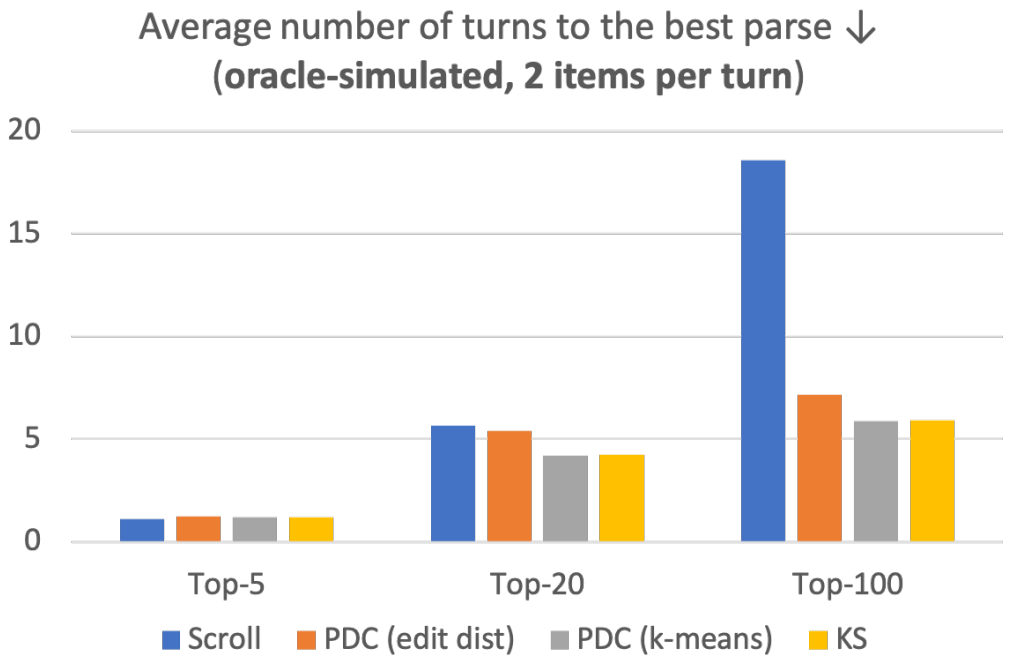
Platanios et al. Value-Agnostic Conversational Semantic Parsing. ACL 2021
Semantic Machines et al. Task-Oriented Dialogue as Dataflow Synthesis. TACL 2020

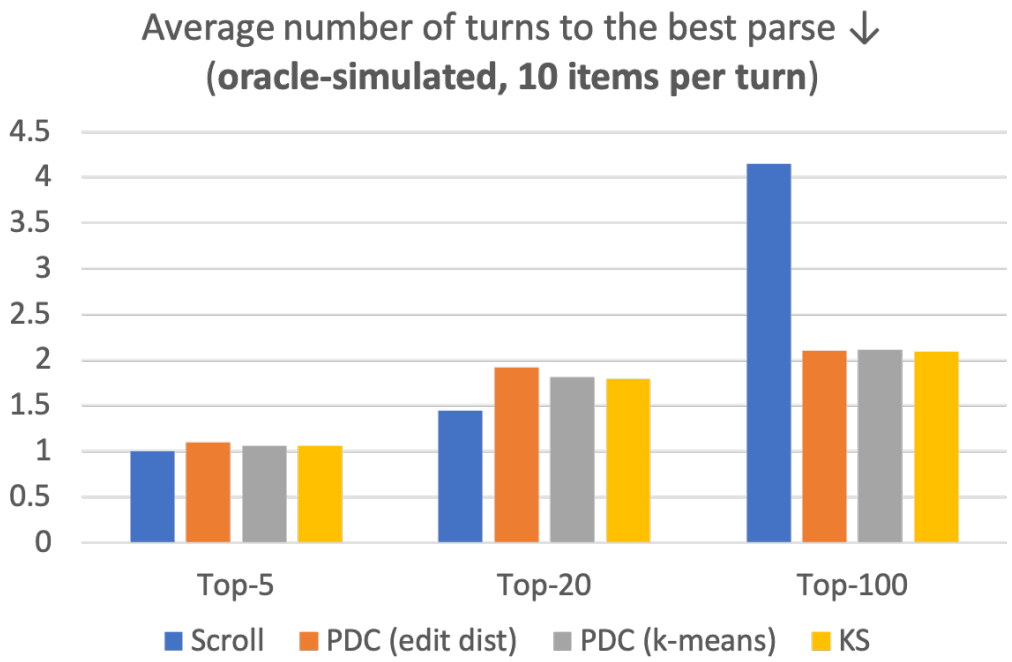
Results: Guidance Comparison



Ippolito et al. Comparison of Diverse Decoding Methods from Conditional Language Models. ACL 2019

Results: Guidance Comparison (Cont'd)





Takeaways

- Proposed human-in-the-loop annotation UIs that leverage **guidance mechanisms** over ***K*-best lists** generated by a low-resource model
- Guidance mechanisms are based on **autocompletion** (via prefix match) and search **keyword suggestion** (via explainable *k*-means clustering)
- Guidance-based UIs achieve **higher annotation speed** and **accuracy** than non-guided baseline interfaces, i.e. `No-KBest` and `Scroll`
- Among guidance-based UIs, `Autocomplete` achieves the **highest accuracy**, `Search` achieves the **highest speed**, and `Search-Keywords` receives the **most positive feedback** while balancing accuracy and speed