Robust Unstructured Knowledge Access In Conversational Dialogue With ASR Errors

Yik-Cheung Tam, Jiacheng Xu, Seeger Zou, Zecheng Wang, Tinglong Liao, Shuhan Yuan

{yt2267, jx1038, jz3714, zw2374, tl2564, sy2448}@nyu.edu

New York University Shanghai

ICASSP 2022 Poster ID: SPE-24.4

NYU SHANGHAI 上海 纽约大学

Motivation

- In spoken dialogue system, user may ask a question:

  “ok do they deliver food?”

  - Knowledge access is required to search for answer in knowledge base
- Challenge: Query contains automatic speech recognition (ASR) errors
  - “ok do they delver food”
- ASR errors can deteriorate accuracy of language understanding

$\Rightarrow$Require a robust knowledge access towards ASR errors

Tasks

Part of the DSTC10 challenge

Kim et. al. 2022. DSTC10 Track Proposal: Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations. In DSTC10 Workshop @ AAAI

Knowledge Turn Detection (KTD)

Knowledge selection (KS) with knowledge access

Dialogue		Ground-truth Annotations	
Speaker	Utterance	Sub-track #1	Sub-track #2
User	hi ummm i'm looking for a place at uhhh to stay at fisherman's wharf at a hotel in a moderate price range	hotel-area: fisherman's wharf hotel-pricerange: moderate	
Agent	sure let me see what can i find for you ok unfortunately we're not showing any result is there any specification that i can change to possibly find you something		
User	is there wholesale in the expensive pressure engine student	hotel-area: fisherman's wharf hotel-pricerange: expensive	
Agent	sure let me see ok so there is one called the suites at fisherman's wharf is that something that would be interesting to you		
User	can you tell me how much parking coast		Q: What is the cost of parking at the site? A: Parking costs $25 per day at this property.
Agent	sure let me see okay this hotel charges twenty five dollars per day		
User	okay how about can i book a room for thursday for seven nights	hotel-area: fisherman's wharf hotel-pricerange: expensive hotel-book_day: thursday hotel-book_stay: seven	
Agent	ok all right yeah so we do have thursday available so i'll book that for seven nights for you is there anything else you need from that		
User	awesome i'm can i also request a place to dine in the same area and the moderate price rain and that sort of sandwiches	restaurant-area: fisherman's wharf restaurant-pricerange: moderate restaurant-food: sandwiches	
Agent	yeah let me check that for you ok so we do have two options one is called boudin bakery and cafe and the other one is pier market seafood restaurant		
User	in can I go with the 1st auction and can you check whether they accept apple pay		Q: Can I use Apple Pay for my purchases at Boudin Bakery & Cafe? A: Apple pay is not accepted at Boudin Bakery & Cafe.
Agent	yeah let me see ok so it's showing that they don't accept apple pay		
User	okay well can i book a table for three on thursday at one pm	restaurant-area: fisherman's wharf restaurant-pricerange: moderate restaurant-food: sandwiches restaruant-name: Boudin Bakery&Cafe restaurant-book_people: three restaurant-book_day: thursday restaurant-book_time: 13:00	
Agent	ok let me yeah so they do have it for one o'clock i'll book that for three people for you		

Figure 2: An example conversation with the ground-truth labels for both sub-tracks.

Challenge and proposal

- Training dialogues are in written style without ASR errors
- Dev set only has 263 dialogues containing ASR errors (not enough data)
- Knowledge selection by brute force can be time consuming: P(y=1|context, k_j) for all possible knowledge k_j in knowledge base

Proposal:

- Generate more noisy training data from clean data via ASR error simulator
- Pre-select promising clusters via knowledge cluster classification
- Re-ranking knowledge titles in promising clusters

ASR error simulator

- Obtain word confusion pairs from ASR confusion:
  - “okay do they (delver|deliver|delo|over|lover|del|dolo)?”
- Perform letter-letter alignment among word candidates
  - d e l * v e r
  - d e l i v e r
- Estimate P(t | s, i) for substitution, deletion, insertion errors
- Randomly replace/insert letters in a clean word to obtain a noisy word:

Table 1. Samples of ASR error simulation.

clean word	noisy versions
’alcohols’	’alchols’, ’alcohos’, ’alcohls’, ’alchos’
’alimentum’	’alimentm’, ’alimetum’, ’alimenum’, ’alietum’
’wifi’	’wifr’, ’wivi’, ’wrfi’, ’wifie’

Proposed knowledge cluster classifier

A double-headed model $L(\Theta) = L_{cls}(W_{bert}, W_{cls}) + \lambda \cdot L_{lm}(W_{bert}, W_{lm})$

Huggingface model name:
wilsontam/bert-base-uncased-dstc10-knowledge-cluster-classifier

Source code:
https://github.com/yctam/dstc10_track2_task2

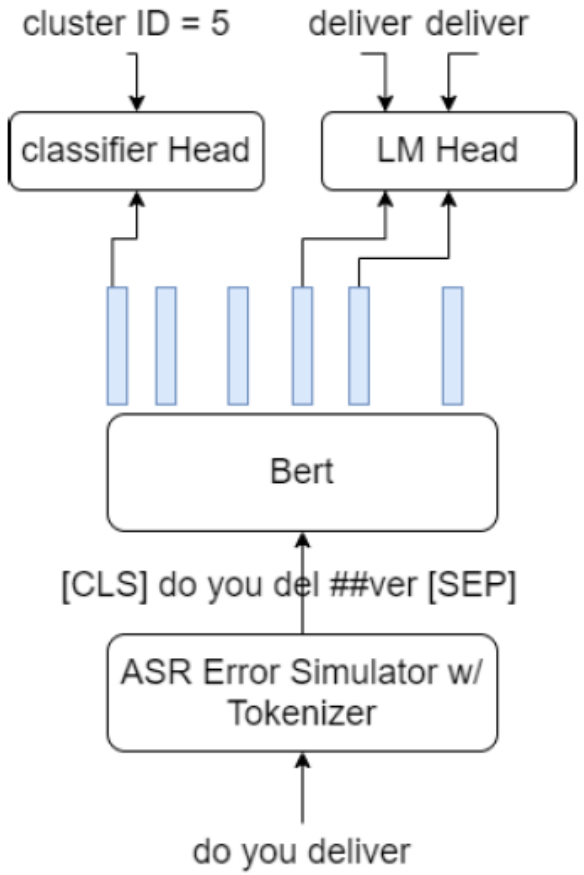


Fig. 1. Proposed system architecture with an ASR error simulator.

Knowledge clustering

- Similar titles appear in knowledge base
- Goal: Group similar knowledge titles
  - provide automatic label for knowledge cluster classification

Proposal:

- Train a Q-A Bert cross-encoder to validate if a title-body pair is valid
- Iterative merge clusters via title-body validation

Title-body validation

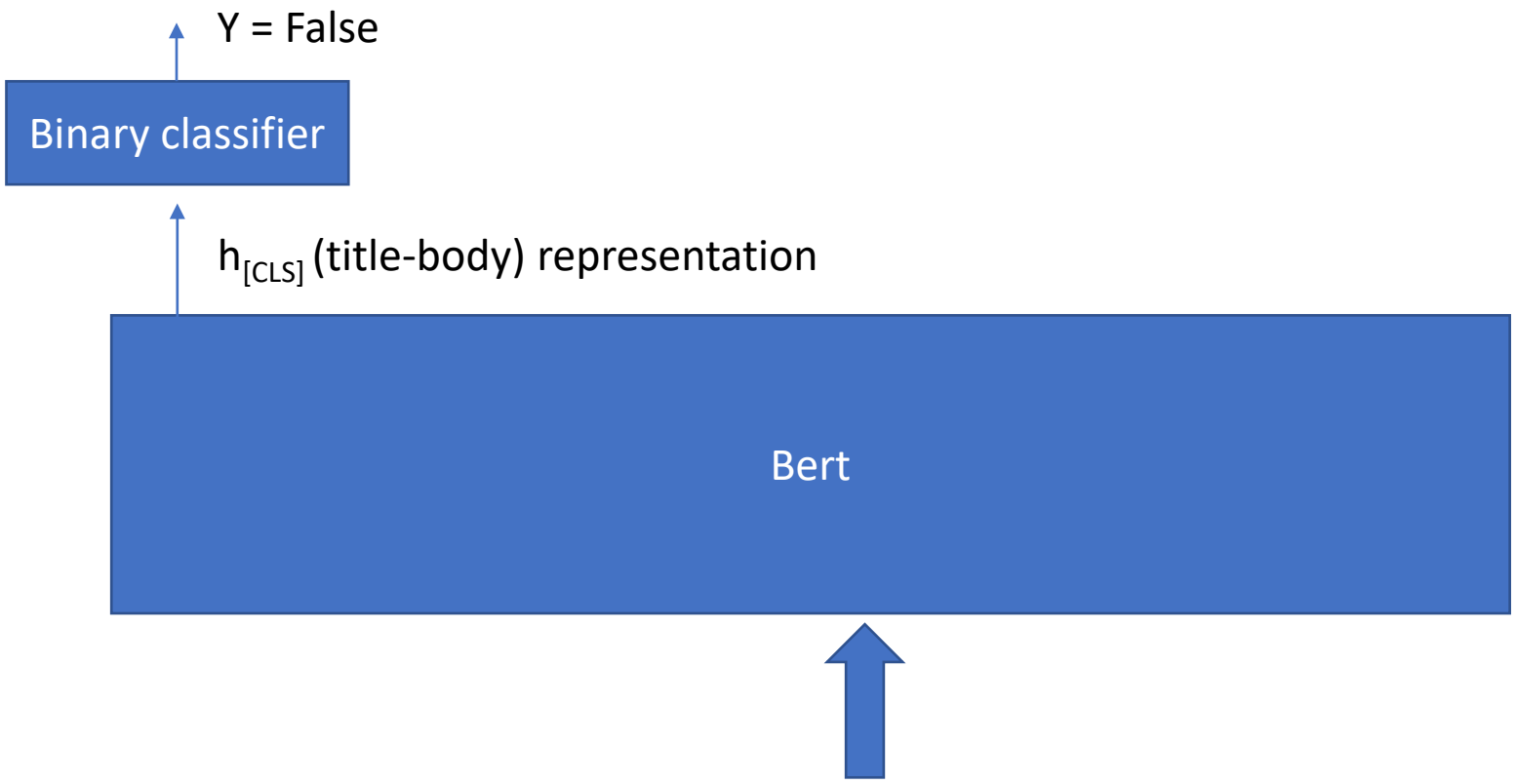
- 2-class classification
- Positive example from knowledge base:
  - Entity: comfort inn by the bay hotel san francisco
  - Title: are there any wifi at the site?
  - Body: free wifi is available at the ~~[comfort inn by the bay hotel san Francisco]~~ hotel

Replace entity as hotel

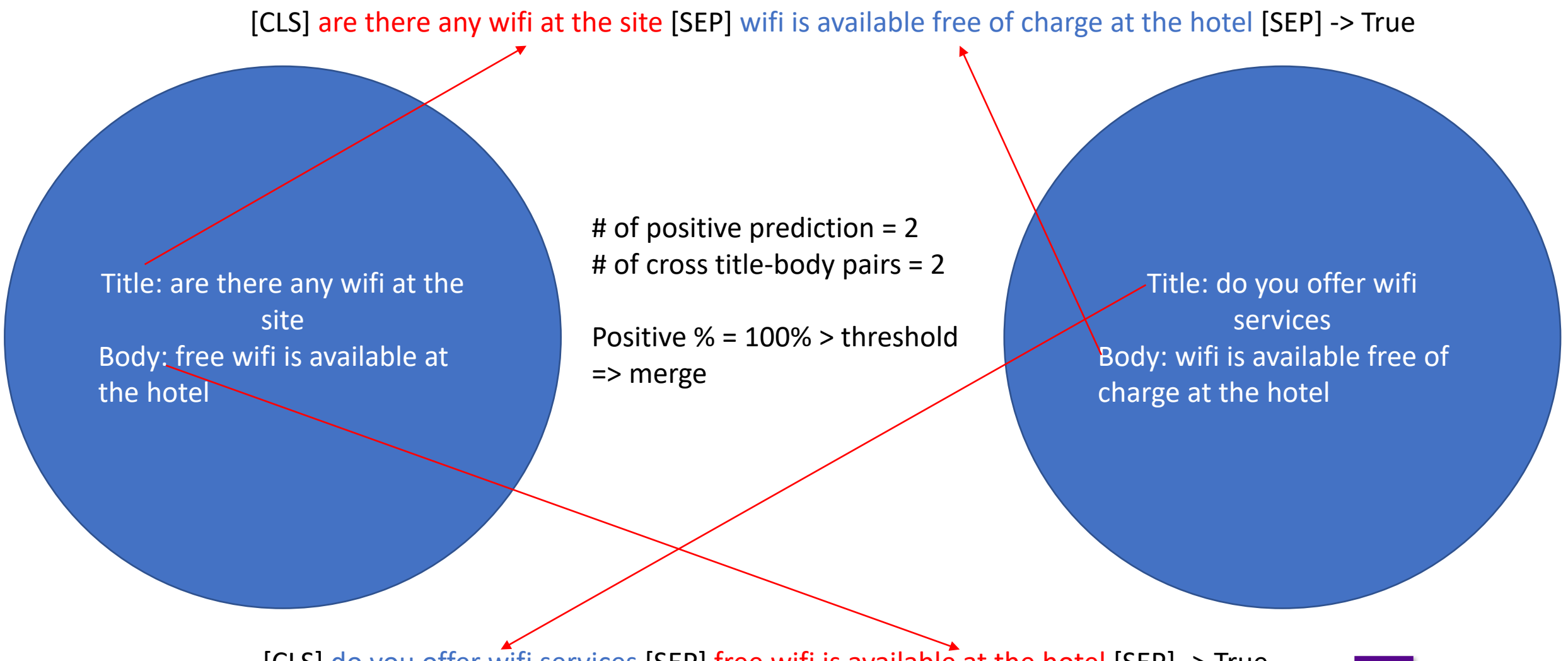
- Negative sample via random sampling of knowledge titles from the same entity (assuming titles are mostly unique within an entity)
  - Entity: comfort inn by the bay hotel san francisco
  - Title: are there any wifi at the site?
  - Body: no, non-smoking rooms

Bert-based knowledge validation

- Huggingface model name: wilsontam/bert-base-uncased-dstc10-kb-title-body-validate

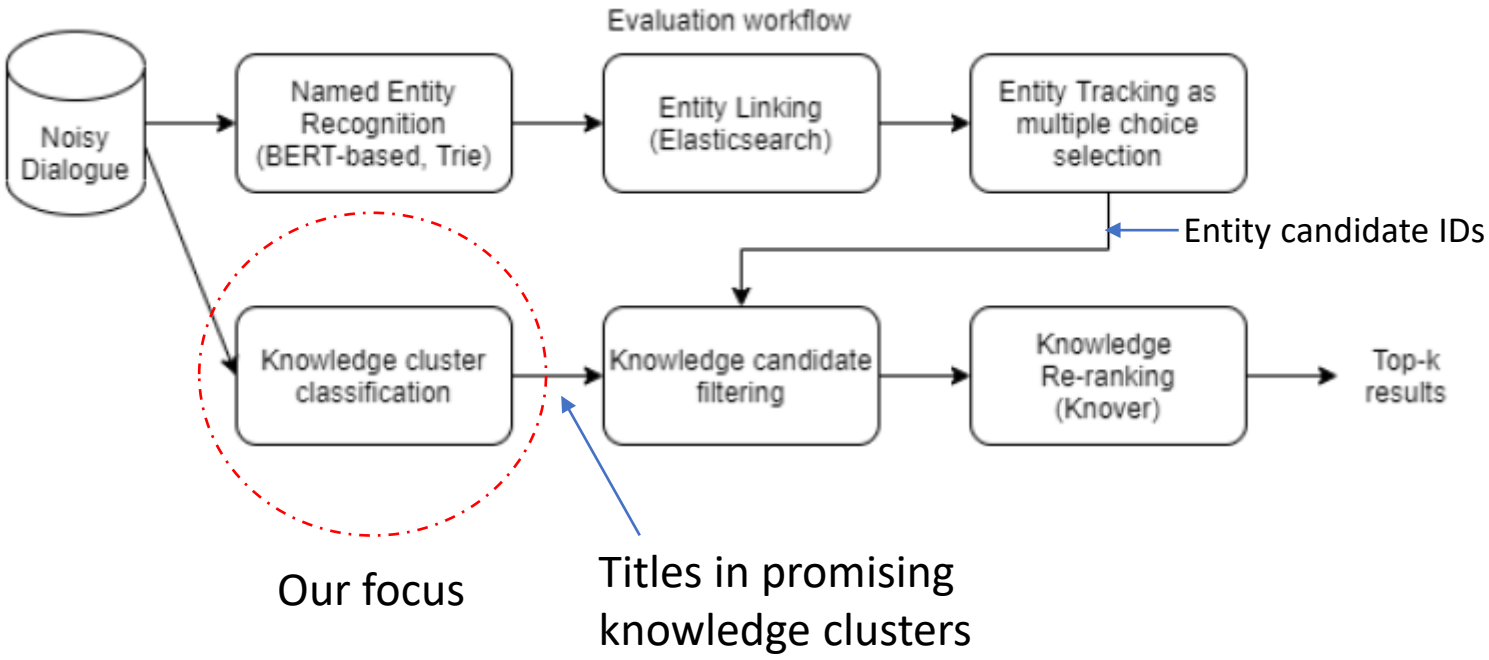


Validate & merge two clusters



NYU SHANGHAI 上海 纽约大学

Proposed system workflow for evaluation



Experimental setup

- Official DSTC10 dev set for dev & devtest (1:1 split) as noisy data (263 dialogues)
- DSTC9 evaluation data excluding DSTC10 portion as clean data (3867 dialogues)
- Knowledge titles as the last turn for training (12k title-body pairs)
- Apply ASR error simulation on clean dialogues & knowledge titles to generate noisy training set for knowledge cluster classifier training
- Official DSTC10 test set for test evaluation (ASR word error rate is about 24%, 1988 dialogues)
- Metrics: F1, Recall@{1,5} Mean reciprocal rank (MRR@5)

Dev & Devtest results

- Observation: Large number of training steps are required for better performance (LM weight > 0)
- On-the-fly ASR error simulation provide more variety for robust training

Table 4: Knowledge-turn detection and knowledge cluster classification results on the development set using the proposed approach with various LM weights.

LM weight	F1 (KTD)	R@1	R@5	MRR@5	Steps
Baseline	0.9306	0.8118	0.8712	0.8333	14925
0.0	**0.96**	0.90	**0.94**	0.915	13930
0.1	**0.96**	**0.94**	**0.94**	**0.94**	84575
0.5	**0.96**	**0.94**	**0.94**	**0.94**	59700
1.0	0.94	0.92	0.92	0.92	32835

Table 6: Knowledge-turn detection and knowledge selection results on the test set using the proposed approach with various LM weights after knowledge re-ranking using Knover (He et al. 2021)

LM weight	F1 (KTD)	R@1	R@5	MRR@5
Baseline	0.9433	0.7358	0.8301	0.7798
0.0	0.9714	0.7238	0.8571	0.7873
0.1	0.9714	0.7619	0.8952	0.8253
0.5	**0.9904**	**0.7809**	**0.9333**	**0.8460**
1.0	0.9411	0.7647	0.9019	0.8300
0.5 (w/o entity tracker)	**0.9904**	0.6666	0.8380	0.7333
0.5 (w/o Knover)	**0.9904**	0.7619	0.8952	0.8101

Official DSTC10 evaluation results

Table 7: DSTC10 evaluation results on knowledge cluster classification using the proposed approach.

System	R@1	R@5	MRR@5
Knover baseline	0.6923	0.7374	0.7091
Submitted system	**0.7901**	**0.8263**	**0.8045**
Post-eval (last-turn)	0.7838	0.8256	0.7998
Post-eval (all-turns)	0.6973	0.7496	0.7170

Table 8: DSTC10 evaluation results on knowledge-turn detection using the proposed approach.

System	P	R	F1
Knover baseline	0.8967	0.6735	0.7692
Submitted system	0.8852	0.8697	0.8774
Post-eval (last-turn)	**0.8878**	0.8696	**0.8786**
Post-eval (all-turns)	0.7636	**0.9033**	0.8276
Unsup knowledge clusters	0.8494	0.9004	0.8742

Table 9: DSTC10 evaluation results on knowledge selection using the proposed approach.

System	R@1	R@5	MRR@5
Knover baseline	0.495	0.6472	0.5574
Submitted system	0.7105	0.7976	0.7493
Post-eval (last-turn)	**0.7144**	**0.8032**	**0.7541**
Post-eval (all-turns)	0.6586	0.7391	0.6950
Unsup knowledge clusters	0.6979	0.7846	0.7369

Conclusion & future work

- ASR error simulation is effective
- Knowledge clustering enables fast knowledge retrieval

Future work:

- Compare with different ASR error simulators
- Apply ASR error simulator on dialogue state tracking

Thank you!

Poster ID: SPE-24.4

SPE-24: Dialog System II: General Topics

Mon, 9 May, 21:00 - 21:45 Singapore Time (UTC +8)
Mon, 9 May, 15:00 - 15:45 France Time (UTC +2)
Mon, 9 May, 13:00 - 13:45 UTC
Mon, 9 May, 09:00 - 09:45 New York Time (UTC -4)

Location: Gather Area E
Track: Speech and Language Processing

NYU SHANGHAI 上海 纽约大学