An End-to-End Trainable Neural Network Model with Belief Tracking for Task-Oriented Dialog

Bing Liu, Ian Lane
Carnegie Mellon University
liubing@cmu.edu, lane@cmu.edu

Outline

- **Background & Motivation**
- Proposed Method
- Experiments & Results
- Conclusions

Task-Oriented Spoken Dialogue Systems

- Unlike open domain chit-chat type of dialogs, task-oriented dialog systems enable user to perform everyday tasks in specific domains.
- Tracking dialogue state overall multiple turns, asking questions to clarify a user request, grounding with information from external resources, etc.
- Current systems are highly handcrafted, using pipeline approach with connected modules for SLU, DST, Policy, and NLG.

Limitations of Current Approach

- **Lack of flexibility**
  - Module input depends on preceding module outputs
- **Credit assignment problem**
  - Error propagates from upstream to downstream system modules
- **Misaligned optimization targets**
  - Each system modules (SLU, DST, Policy) has its own optimization target, which may not directly aligned with final system optimization criteria (e.g. task success rate, etc.)

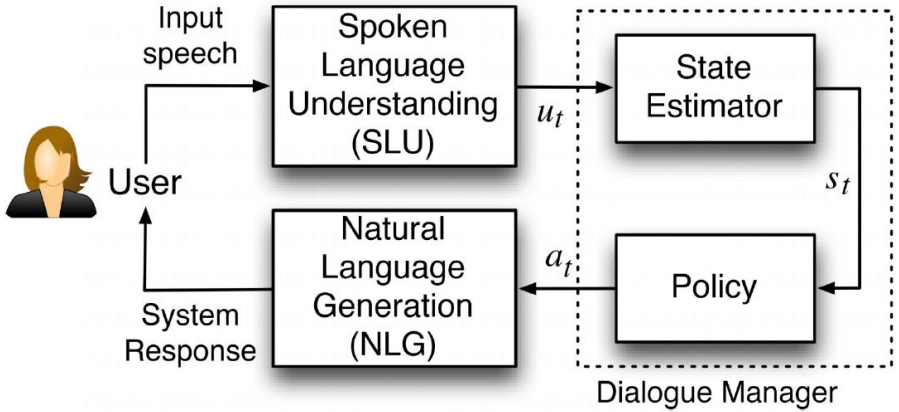


Figure from "*Li Deng at AI Frontiers:Three Generations of Spoken Dialogue Systems*"

Motivation

Can we design an end-to-end trainable system with an unified model for dialogue state tracking, knowledge base (KB) operation, and response generation?

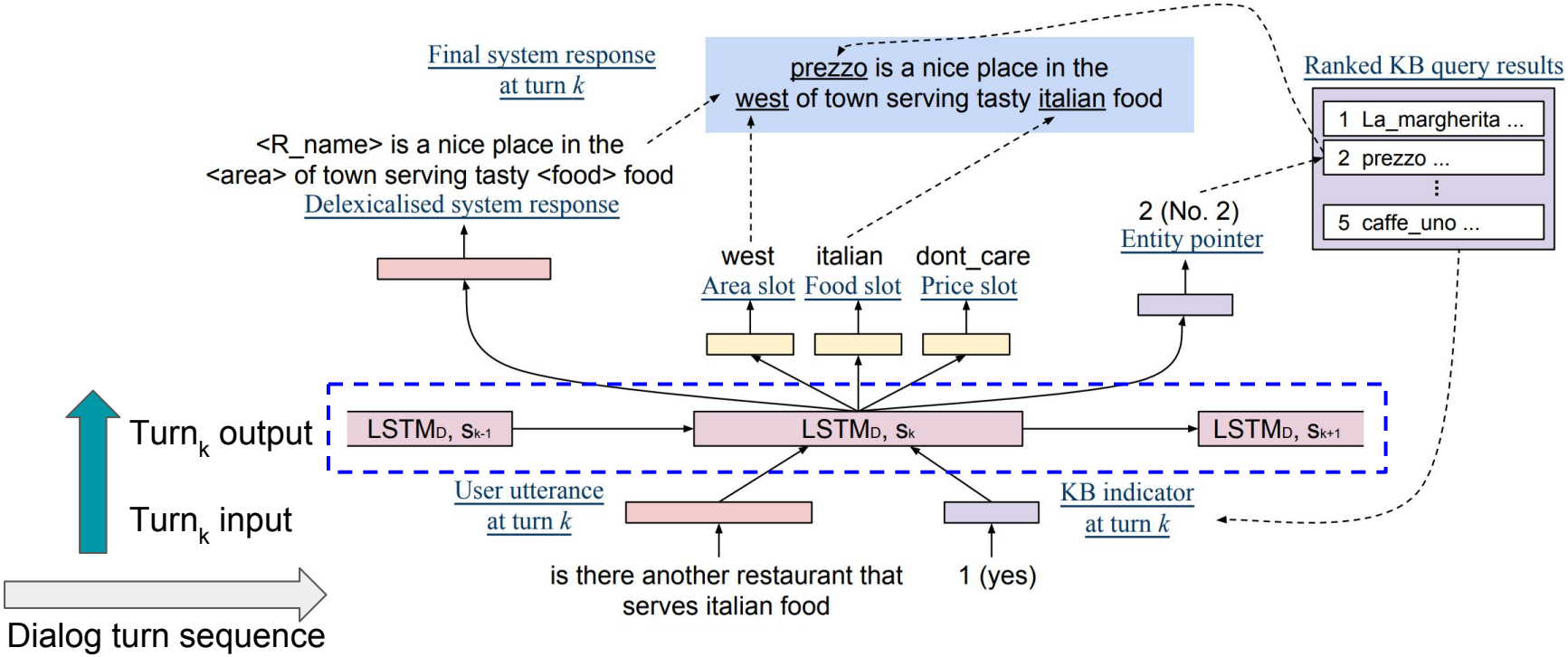
- Fully explore knowledge that can be shared among different components
- Optimization can be made towards the final objective in an end-to-end fashion

Outline

- Background & Motivation
- **Proposed Method**
- Experiments & Results
- Conclusions

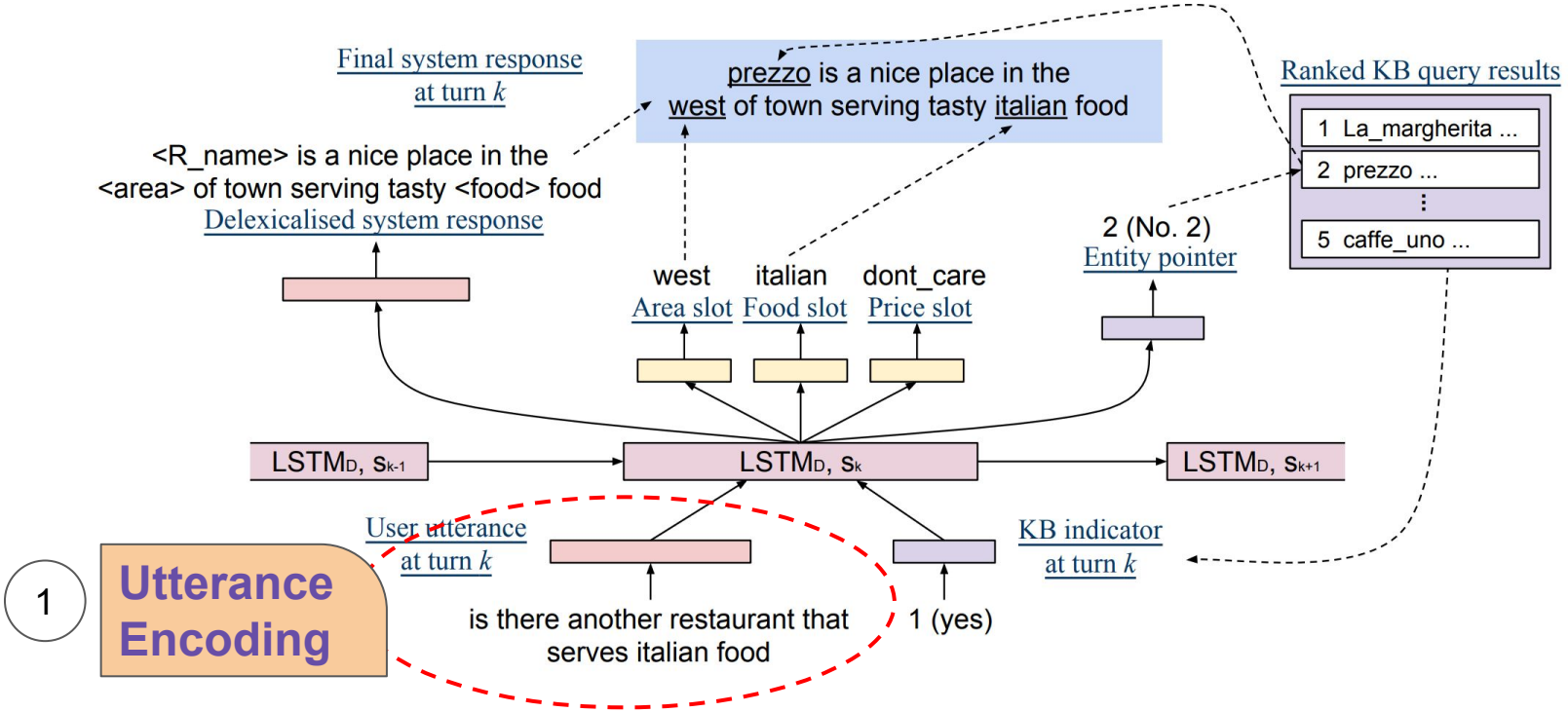
End-to-End Neural Dialog Model

Hierarchical LSTM with multi-task learning



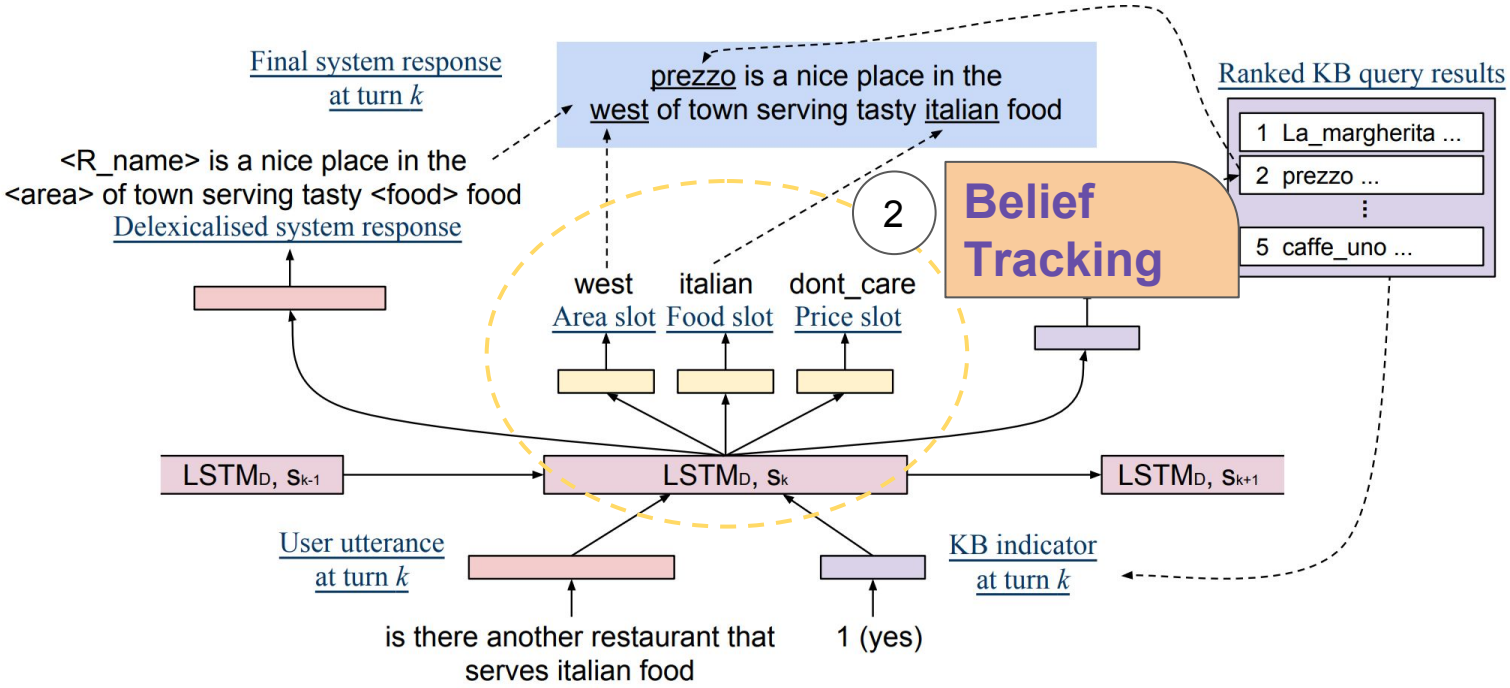
End-to-End Neural Dialog Model

Hierarchical LSTM with multi-task learning



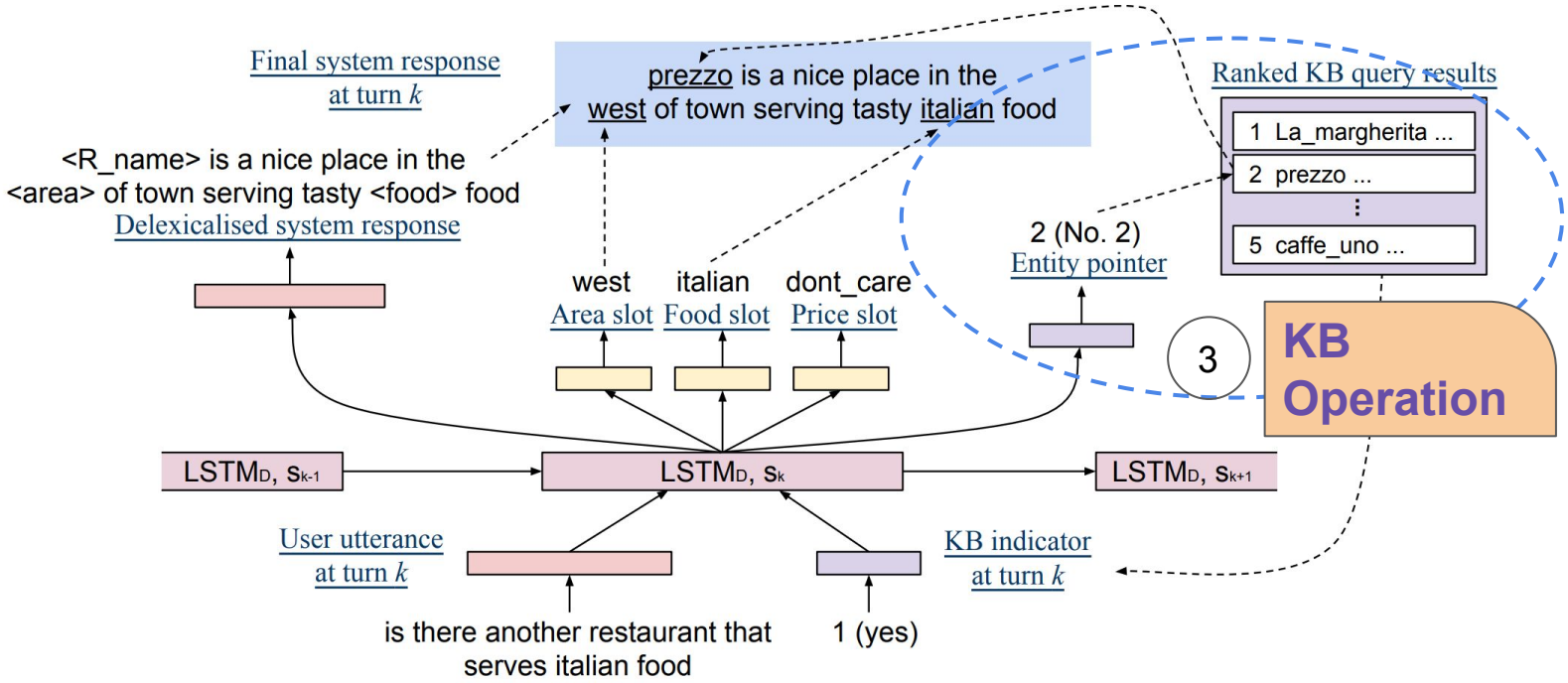
End-to-End Neural Dialog Model

Hierarchical LSTM with multi-task learning



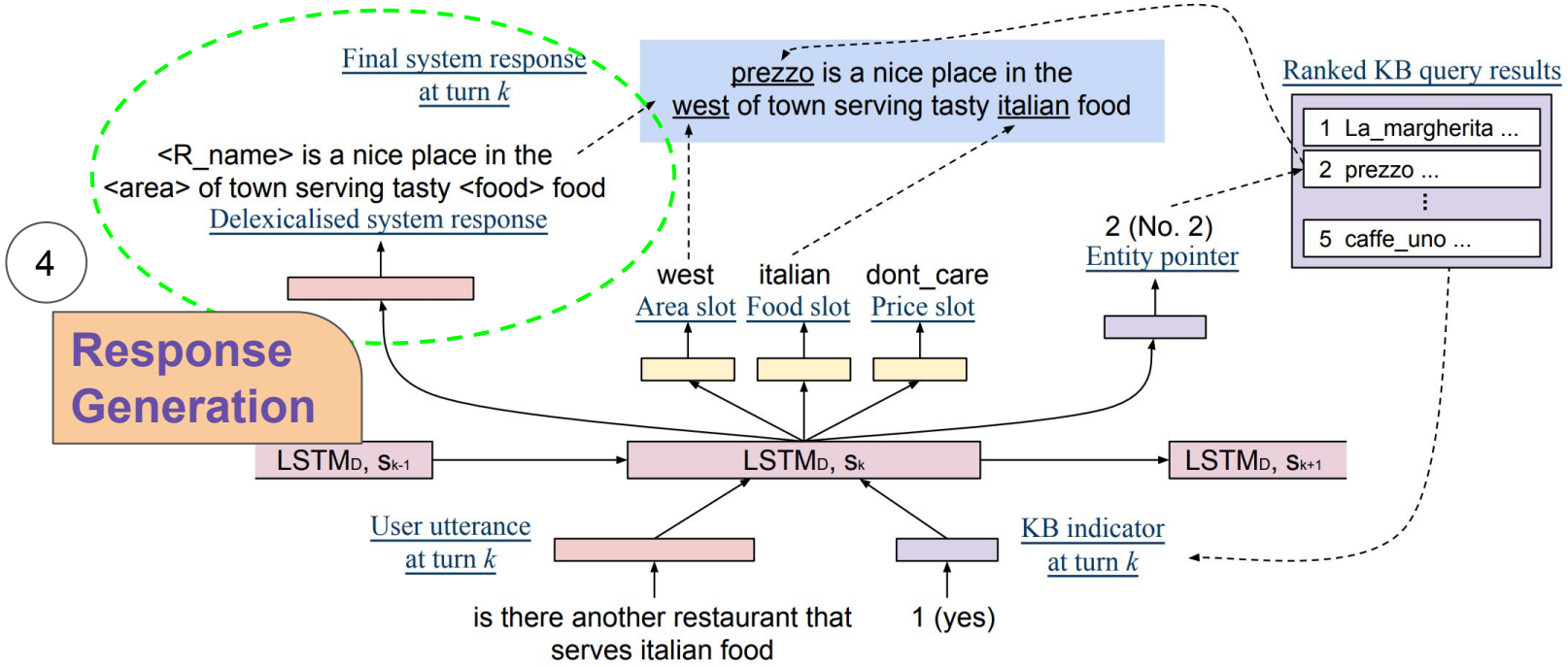
End-to-End Neural Dialog Model

Hierarchical LSTM with multi-task learning



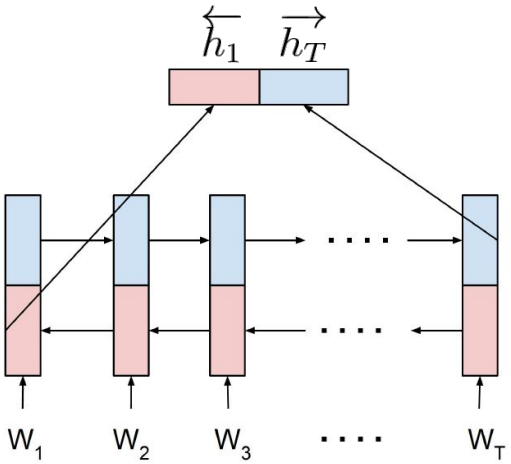
End-to-End Neural Dialog Model

Hierarchical LSTM with multi-task learning



1. Utterance Encoding

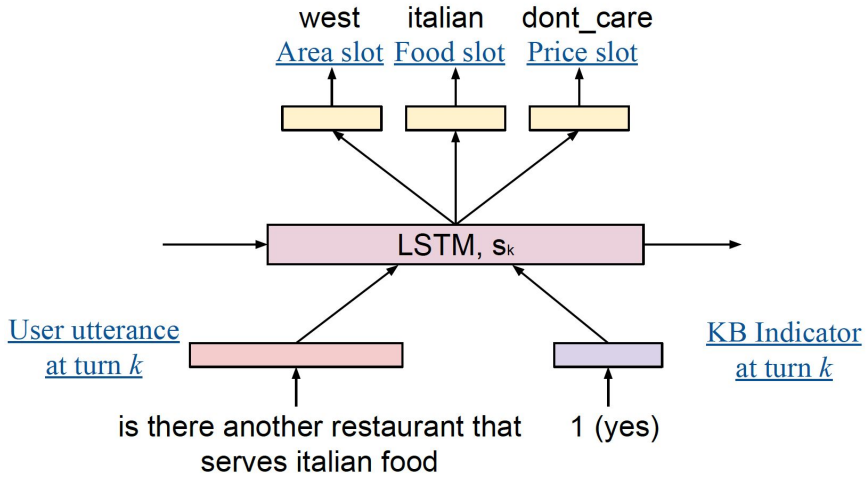
- Utterance encoding with bidirectional LSTM



$$U_k = [\overrightarrow{h_{T_k}^{U_k}}, \overleftarrow{h_1^{U_k}}]$$

2. Belief Tracking

- Belief tracking with hierarchical LSTM



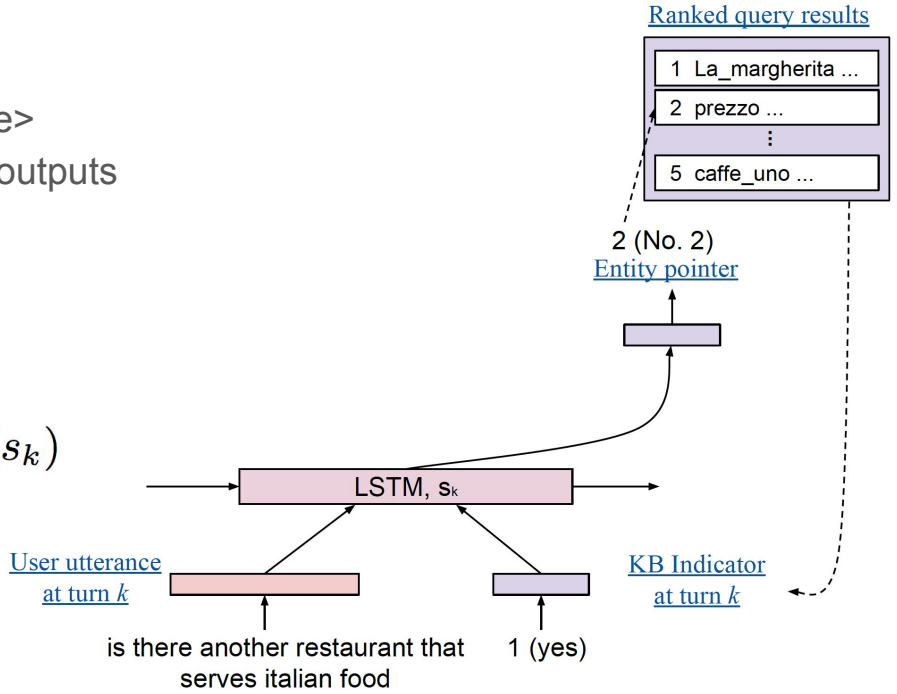
$$s_k = \mathrm{LSTM}_{\mathrm{D}}(s_{k-1},\ [U_k,\ I_k])$$

$$P(S_k^m \mid \mathbf{U}_{\leq k},\ \mathbf{I}_{\leq k}) = \mathrm{SlotDist}_m(s_k)$$

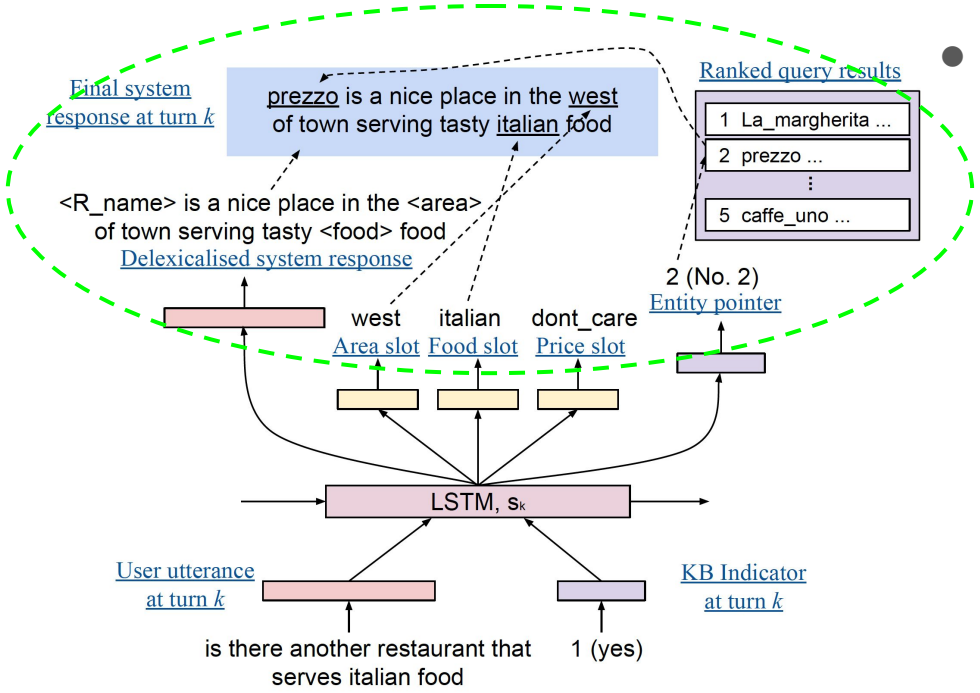
3. KB Operation

- Issuing API Calls
  - E.g. api_call <area> <food> <pricerange>
  - Fill slots with values from belief tracker outputs
- KB Results Processing
  - KB results as a list of structured entities
  - Emits an entity pointer

$$P(E_k \mid \mathbf{U}_{\leq k},\ \mathbf{I}_{\leq k}) = \mathrm{EntityPointerDist}(s_k)$$



4. Response Generation

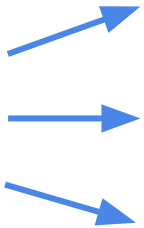


- Response generation
  - Select a delexicalised response template
  - Final system response generated by replacing slot tokens with values from belief tracker outputs and KB query results

$$P(R_k \mid \mathbf{U}_{\leq k},\ \mathbf{I}_{\leq k}) = \mathrm{ResponseDist}(s_k)$$

Alternative Model Designs

Hierarchical LSTM
+ feed de-lex res (1)
+ feed goal slots (2)
+ feed both (3)



(1) Feed *previous emitted system response template* back to dialog-level LSTM state

(2) Feed *previous emitted slot values* back to dialog-level LSTM state

(3) Feed *both* back to dialog-level LSTM state

Model Training

- Linear interpolation of multi-task cross-entropy losses

$$\min_{\theta} \sum_{k=1}^{K} -\Big[\sum_{m=1}^{M} \lambda_{S^m} \log P(S_k^{m*} \mid \mathbf{U}_{\leq k},\ \mathbf{I}_{\leq k}; \theta) + \lambda_E \log P(E_k^* \mid \mathbf{U}_{\leq k},\ \mathbf{I}_{\leq k}; \theta) + \lambda_R \log P(R_k^* \mid \mathbf{U}_{\leq k},\ \mathbf{I}_{\leq k}; \theta) \Big]$$

→ Belief Tracking Losses

→ KB Operation Loss

→ Response Prediction Loss

Outline

- Background & Motivation
- Proposed Method
- **Experiments & Results**
- Conclusions

Dataset

- Converted from DSTC2 corpus [1]
- Added (a) API calls to KB, & (b) KB query results, as in [2]
- Evaluation is on accuracy for belief tracking, entity pointer prediction, response matching (delexicalised & final)

Num of train & dev / test dialogs	2118 / 1117
Num of turns per dialog in average (including API call commands)	7.9
Num of area / food / pricerange options	5 / 91/ 3
Num of delexicalised response candidates	78

[1] M. Henderson, B. Thomson, and J. Williams, “The second dialog state tracking challenge,” in SIGDIAL, 2014.
[2] A. Bordes and J. Weston, “Learning End-to-End Goal-Oriented Dialog,” in ICLR, 2017.

Model Configuration and Training

- Word embedding size = 300
- Utterance-level LSTM state size = 150
- Dialog-level LSTM state size = 200

- Mini-batch training: batch size = 32
- Optimizer: Adam, initial learning rate = 1e-3
- Dropout: on non-recurrent connections, keep prob = 0.5
- Gradient clipping: 5

Results and Analysis

On different utterance encoding methods

- Bi-LSTM > BoW Emb
- Semantic similarities of words captured in the pre-trained word vectors are helpful in generating a better representation of user input

Model	Entity Pointer	Joint Goal	De-lex Res	Final Res
BoW Emb Encoder	93.5	72.6	55.4	51.2
+ word2vec	93.6	74.3	55.9	51.5
Bi-LSTM Encoder	93.8	**77.2**	55.8	52.6
+ word2vec	**94.4**	76.6	**56.6**	**52.8**

Results and Analysis

On different model designs

- Contrary to our intuition, feeding previous system output did not help.
- Reason: data sparsity issue of the dataset → model overfits training set

Model	Entity Pointer	Joint Goal	De-lex Res	Final Res
Hierarchical LSTM	**94.4**	**76.6**	**56.6**	**52.8**
+ feed de-lex res (1)	93.6	74.8	55.4	51.8
+ feed goal slots (2)	94.1	75.3	55.3	51.8
+ feed both (3)	93.7	72.7	55.3	51.6

Results and Analysis

Comparing to other models on **belief tracking**

- Used live ASR hypothesis as model input.

Model	Area Goal	Food Goal	Price Goal	Joint Goal
RNN	92	86	86	69
RNN + sem. dict	91	86	93	73
NBT-DNN (Mrkšić et al., 2016)	90	84	94	72
NBT-CNN (Mrkšić et al., 2016)	90	83	93	72
Hierarchical LSTM	90	84	93	73

Results and Analysis

Comparing to other models on **generating system response**

- Follow the per-response accuracy metric used in prior work

Model	Per-res Accuracy
Memory Networks (Bordes and Weston, 2016)	41.1
Gated Memory Networks (Perez and Liu, 2016)	48.7
Sequence-to-Sequence (Eric and Manning, 2017)	48.0
Query-Reduction Networks (Seo et al., 2017)	51.1
Hierarchical LSTM	**52.8**

User Study

- Small scale user study with 10 users
- Feedback on appropriateness of the system responses
  - 52.8% → 73.6%
- Per-response accuracy metric may not well correlate with human judgments
- Better dialogs evaluation measurements should to be further explored

Outline

- Background & Motivation
- Proposed Method
- Experiments & Results
- **Conclusions**

Conclusions

- A novel end-to-end trainable neural network model for task-oriented dialog
- System is capable of tracking dialog state, interfacing with KB by issuing API calls, and incorporating query results into system responses to successfully complete task-oriented dialogs
- Robust performance in belief tracking and system response prediction.
- Next step: end-to-end reinforcement learning for task-oriented dialogs (in submission)

Thanks & Questions