Dialogue Learning with Human Teaching and Feedback in End-To-End Trainable Task-Oriented Dialogue Systems

Bing Liu[1], Gokhan Tür[2], Dilek Hakkani-Tür[2], Pararth Shah[2], Larry Heck[2]
[1]Carnegie Mellon University, [2]Google Research
liubing@cmu.edu, {gokhant,dilekh,pararth,larryheck}@google.com

Abstract

- This work focuses on **interactive** learning of task-oriented dialogue systems.
- Learning dialogue policy online from scratch with reinforcement learning (RL) requires a large number of interactive learning sessions with users.
- People thus often pre-train the dialogue agent using dialogue corpora before doing online interactive learning.
- Model with such pre-training may suffer from the mismatch of dialogue state distribution between offline supervised training and online interactive learning:
  - Agent's response at each turn has a direct influence on the distribution of dialogue state during user interaction
  - A small mistake from the agent may lead to compounding errors in dialogue due to this covariate shift
- We propose a hybrid imitation and RL method with human teaching and feedback in addressing this challenge.
- The proposed neural dialogue model can be optimized end-to-end for natural language understanding, dialogue state tracking, and dialogue policy learning.

Model Architecture

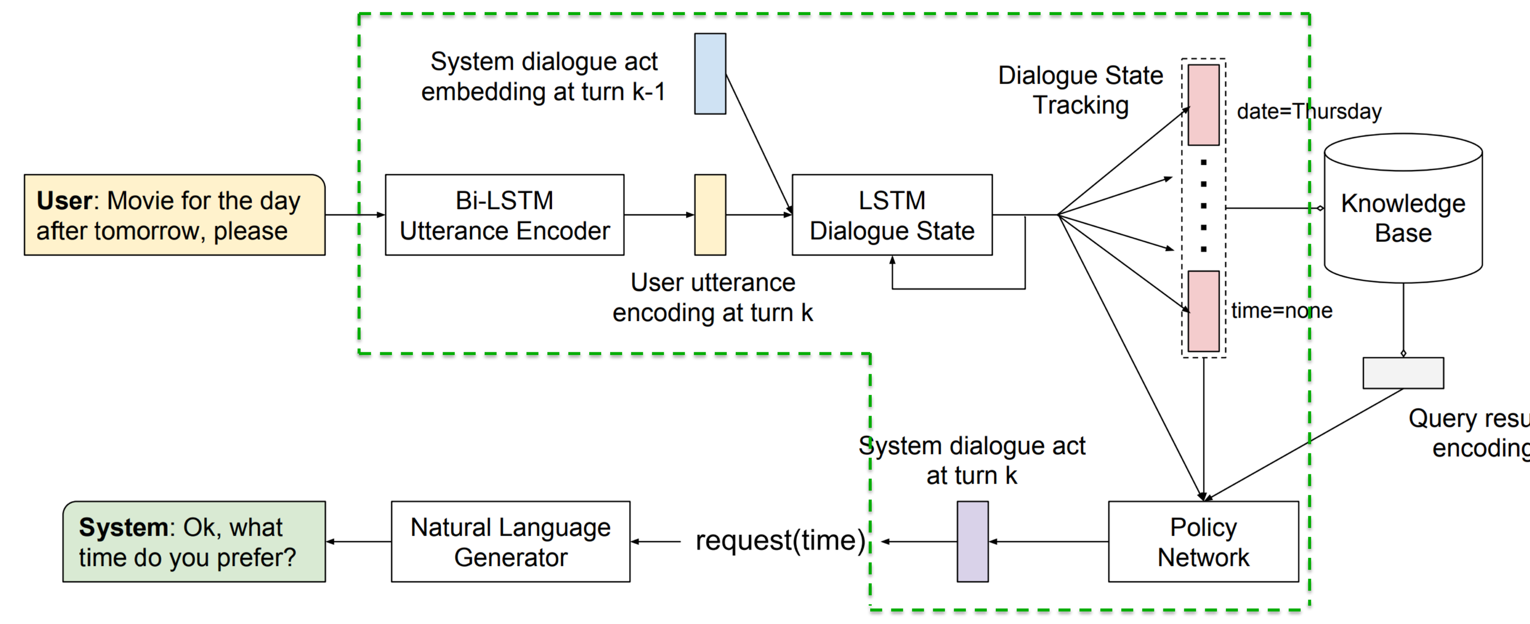


Figure 1. Proposed end-to-end task-oriented dialogue model architecture.

- **Utterance Encoding**
  - Bidirectional LSTM utterance reader: $U_k = [\overrightarrow{h_{T_k}^{U_k}}, \overleftarrow{h_1^{U_k}}]$
- **Dialogue State Tracking**
  - Encode a continuous form of dialogue state in a dialogue-level LSTM
  - Produce a probability distribution over values for each tracked goal slot

$$s_k = \text{LSTM}(s_{k-1}, [U_k, A_{k-1}]) \qquad P(l_k^m \mid \mathbf{U}_{\leq k}, \mathbf{A}_{<k}) = \text{SlotDist}_m(s_k)$$

- **KB Operation**
  - Generate KB query by filling estimated goal values to query templates
- **Dialogue Policy**
  - Emit a system action based on the dialogue state and KB query results.

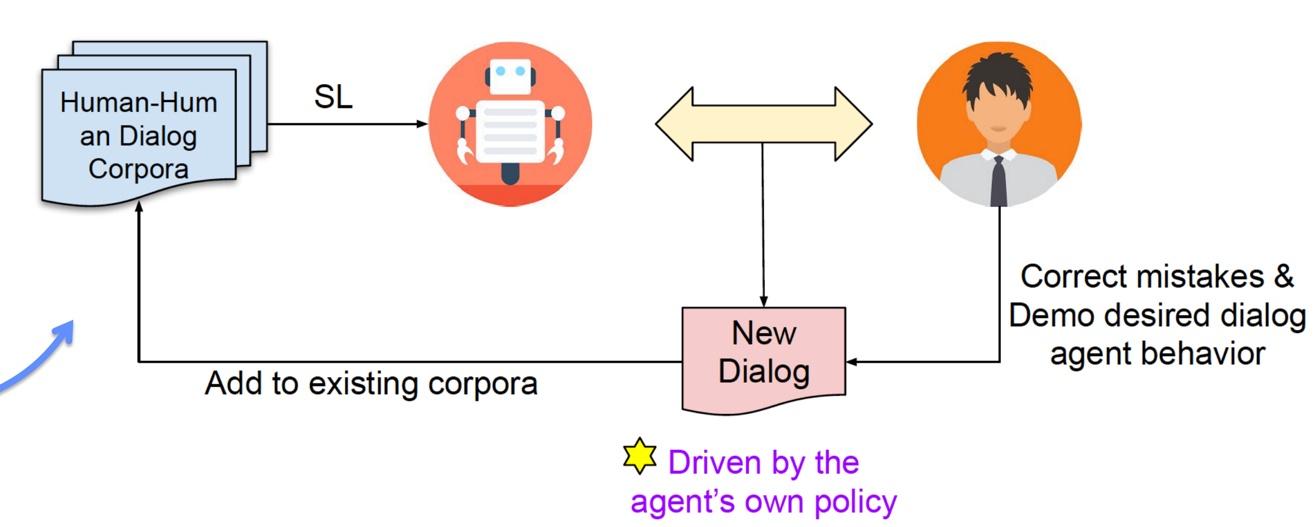
$$P(a_k \mid U_{\leq k}, A_{<k}, E_{\leq k}) = \text{PolicyNet}(s_k, v_k, E_k)$$

Model Training

- **Supervised Pre-training**
  - Train model end-to-end on dialogue samples D with MLE and obtain an initial policy $\pi_\theta(a|s)$

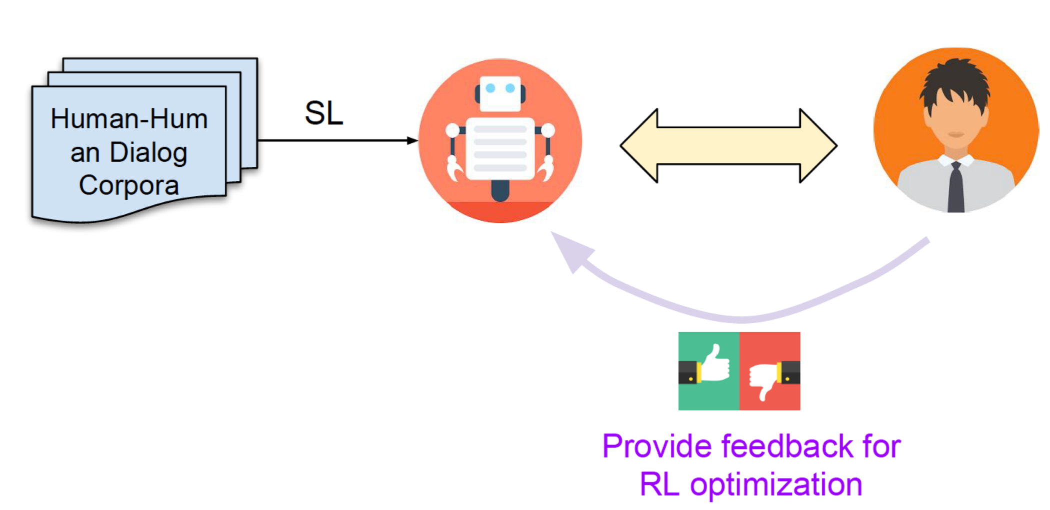
$$\min_\theta \sum_{k=1}^{K} -\Big[\sum_{m=1}^{M} \lambda_{l^m} \log P(l_k^{m*} | \mathbf{U}_{\leq k}, \mathbf{A}_{<k}, \mathbf{E}_{<k}; \theta) + \lambda_a \log P(a_k^* | \mathbf{U}_{\leq k}, \mathbf{A}_{<k}, \mathbf{E}_{\leq k}; \theta)\Big]$$

- **Imitation Learning with Human Teaching**
  1) Run the current policy $\pi_\theta(a|s)$ with user to collect new dialogue samples D_π
  2) Ask user to correct the agent's mistakes in user goal estimation for each dialogue turn in D_π
  3) Add the corrected dialogue samples to the existing corpora: $D \longleftarrow D \cup D_\pi$
  4) Train model end-to-end on D with MLE and obtain an updated policy $\pi_\theta(a|s)$; back to 1)



- **Reinforcement Learning with Human Feedback**
  - i. Run the current policy $\pi_\theta(a|s)$ with user for a new dialogue and collect user's feedback
  - ii. Train model end-to-end with REINFORCE and obtain an updated policy; back to i)

$$\nabla_\theta J_k(\theta) = \nabla_\theta \mathbb{E}_\theta [R_k] = \mathbb{E}_{\theta_a} [\nabla_\theta \log \pi_\theta(a_k|s_k) R_k]$$



Experiments & Results

- **Dataset**:
  - DSTC2 in restaurant search domain
  - In-house collected corpus in movie booking domain
- **Offline Evaluation** (on dialogue state tracking)

Table 1: Belief tracking results on DSTC2 corpus (with ASR hypothesis as input)

Model	Area	Food	Price	Joint
RNN [24]	92	86	86	69
NBT [6]	90	84	94	72
Our end-to-end model	90	84	92	72

Table 2: Belief tracking results on movie booking dataset

Model	Num_ticket	Movie	Theater	Date	Time	Joint
Our end-to-end model	98.22	91.86	97.33	99.31	97.71	84.57

- **Interactive Evaluation**

Figure 2: Interactive learning curves on *task success rate*.

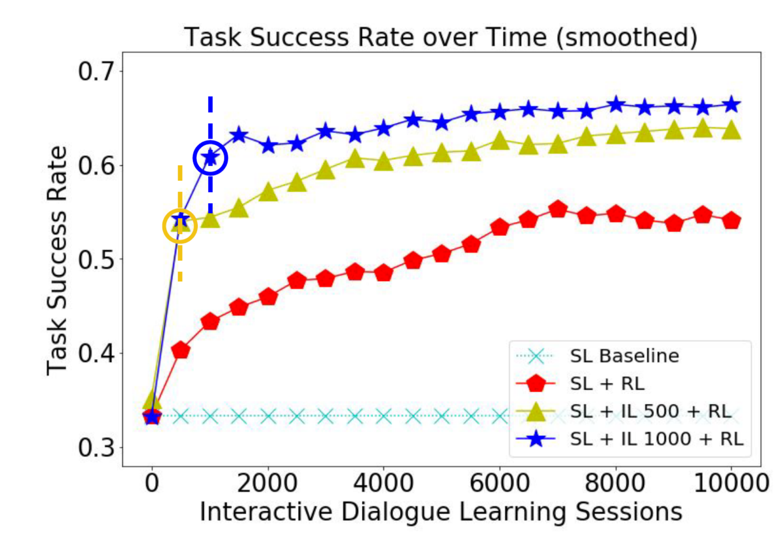


Figure 3: Interactive learning curves on *average dialogue turn size*.

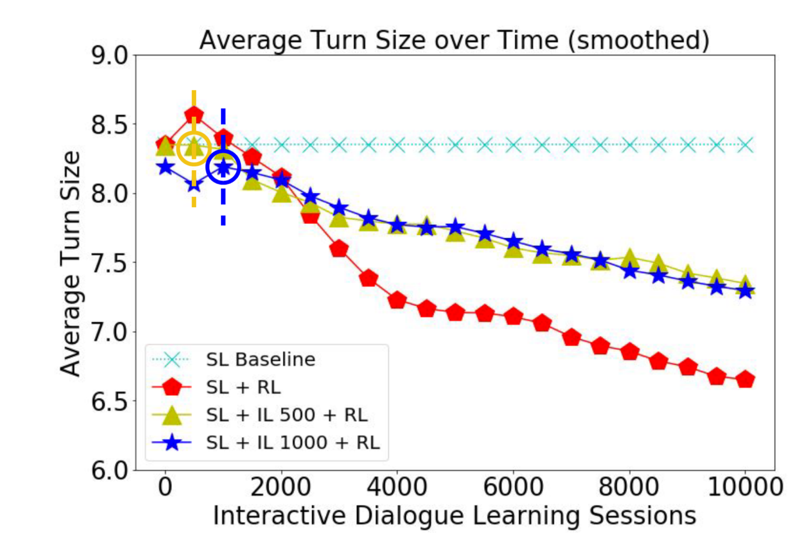
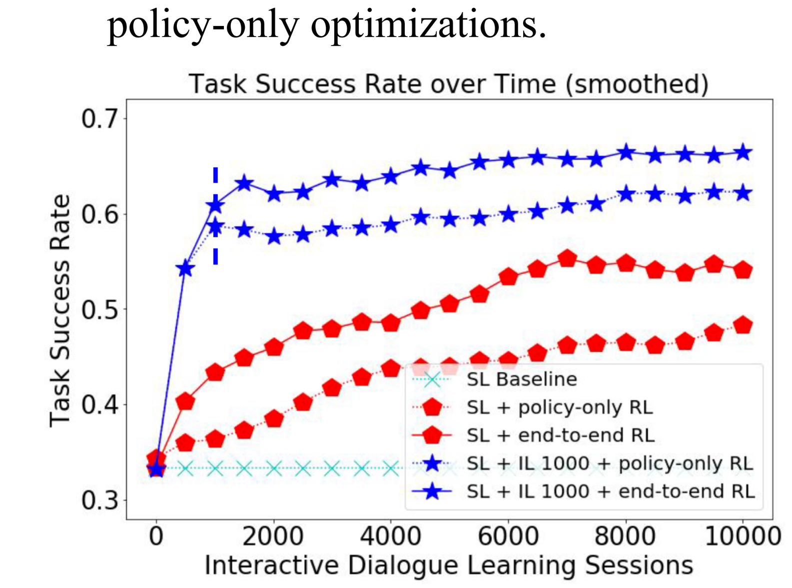


Figure 4: Interactive learning curves on *task success rate* with end-to-end vs. policy-only optimizations.



SL: Supervised Learning
IL: Imitation Learning with human teaching
RL: Reinforcement Learning with human feedback

Figure 4: **Human evaluation** results. Mean and standard deviation of satisfaction scores from crowd workers (between 1 to 5).

Model	Score
SL	3.987 ± 0.086
SL + IL 1000	4.378 ± 0.082
SL + IL 1000 + RL	4.603 ± 0.067

Conclusions

- In this work, we focus on training task-oriented dialogue systems through user interactions, where a dialogue agent improves through communicating with users and learning from the mistake it makes.
- We show that our neural dialogue agent can effectively learn from user teaching with the proposed imitation learning method. Learning with RL on user feedback after IL improves the model performance further.