End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning

Authors: Jason D. Williams and Geoffrey Zweig
Speaker: Hamidreza Shahidi

Outline

- Introduction
- Model description
- Optimizing with supervised learning
- Optimizing with reinforcement learning
- Conclusion

Task-oriented dialogue systems

A dialog system for:

- Initiating phone calls to a contact in an address book

How can I help you?
Call Jason
Which type of phone: mobile or work?
Oh, actually call Mike on his office phone
Calling Michael Seltzer, work.
`PlaceCall`

- Ordering a taxi
- Reserving a table at a restaurant

Task-oriented dialogue systems

A dialog system for:

- Initiating phone calls to a contact in an address book

How can I help you?
Call Jason
Which type of phone: mobile or work?
Oh, actually call Mike on his office phone
Calling Michael Seltzer, work.
`PlaceCall`

- Ordering a taxi
- Reserving a table at a restaurant

Reinforcement learning Setting

State = (user’s goal, dialogue history)

Actions =
- Text actions $\longrightarrow$ *“Do you want to call <name>?”*
- API calls $\longrightarrow$ *PlacePhoneCall(<name>)*

Reward = 1 for successfully completing the task, and 0 otherwise

Reinforcement learning Setting

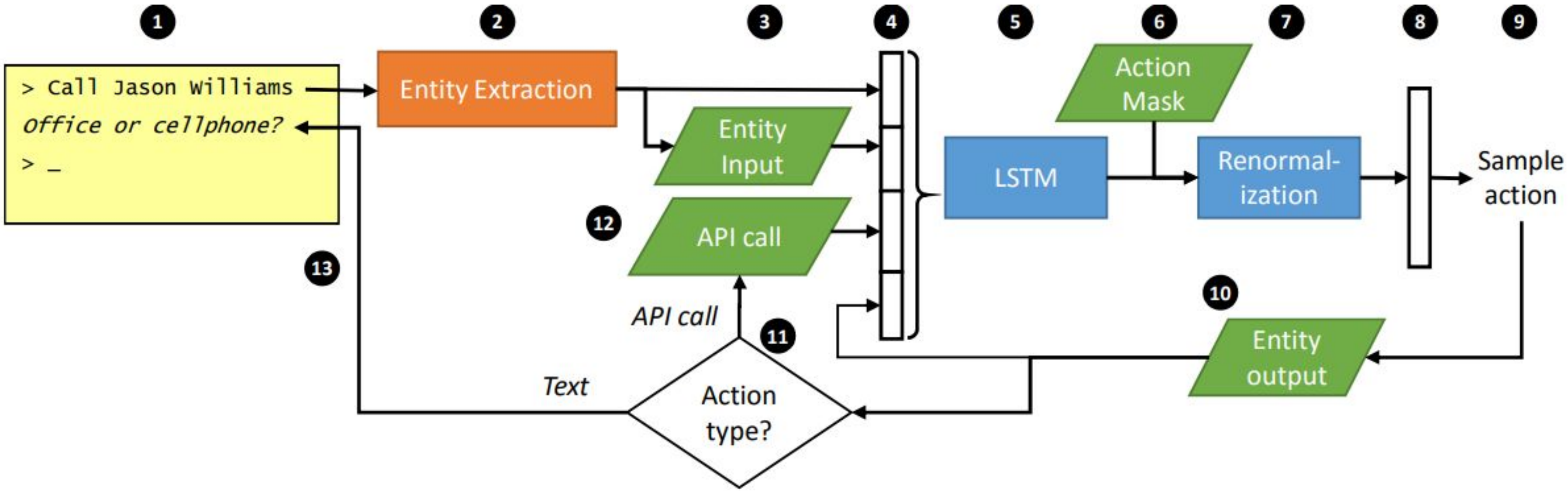
State = (user's goal, dialogue history)

Actions =
- Text actions ⟶ *"Do you want to call <name>?"*
- API calls ⟶ *PlacePhoneCall(<name>)*

Reward = 1 for successfully completing the task, and 0 otherwise

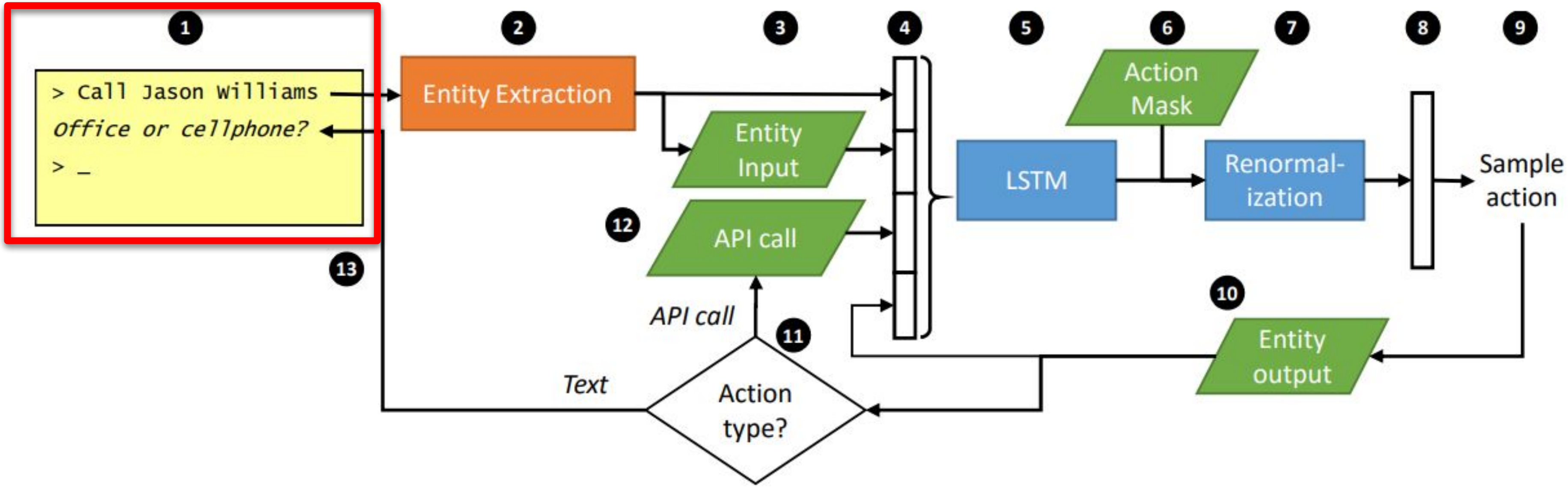
Model description

Model



1
2
3
4
5
6
7
8
9
> Call Jason Williams
Office or cellphone?
> _
Entity Extraction
Entity Input
API call
12
13
Action Mask
LSTM
Renormal-ization
Sample action
10
Entity output
API call
11
Text
Action type?

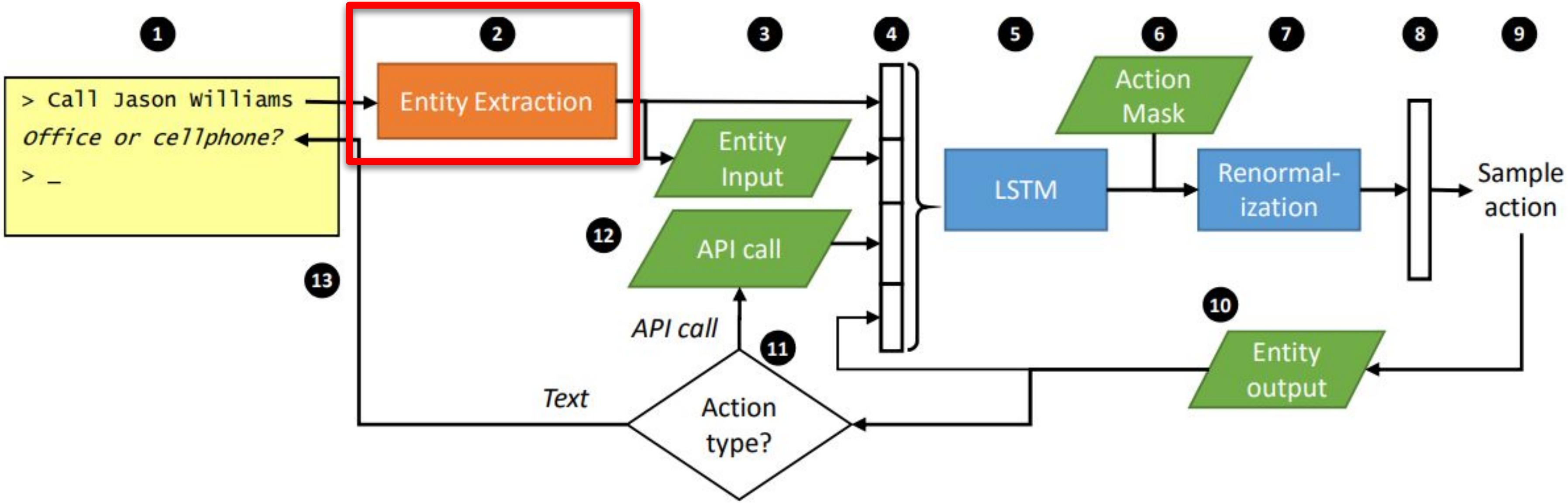
User Input



1
2
3
4
5
6
7
8
9
> Call Jason Williams
Office or cellphone?
> _
Entity Extraction
Entity Input
Action Mask
LSTM
Renormal-ization
Sample action
12
API call
13
10
API call
11
Entity output
Text
Action type?

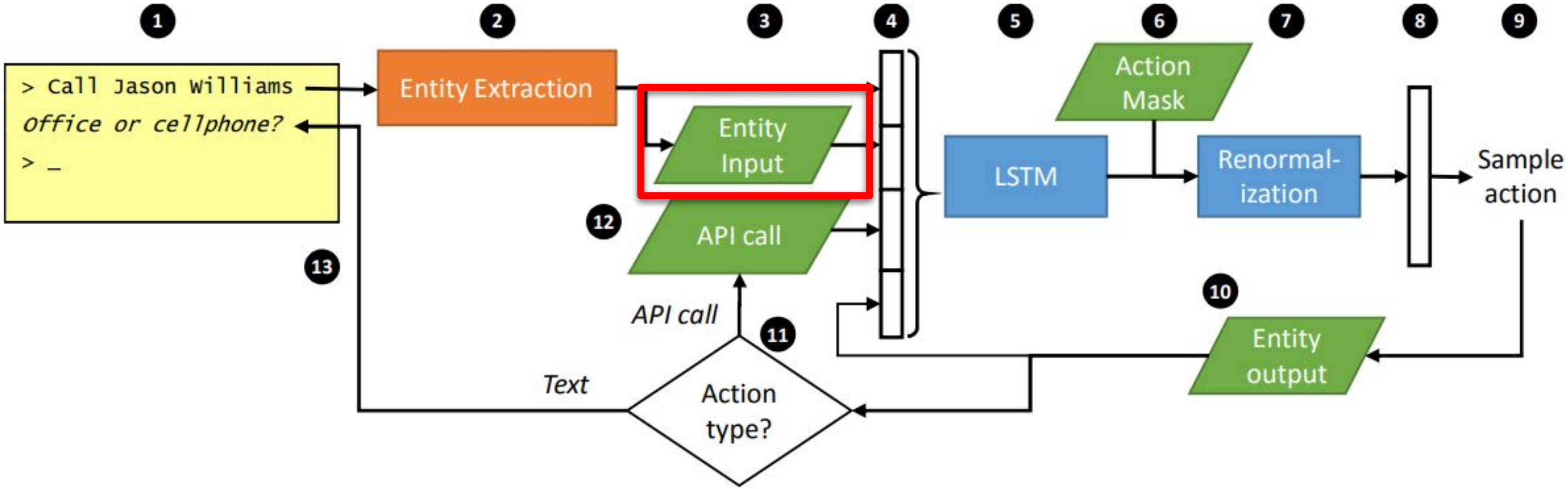
Entity Extraction

For example: identifying “Jason Williams” as a <name> entity

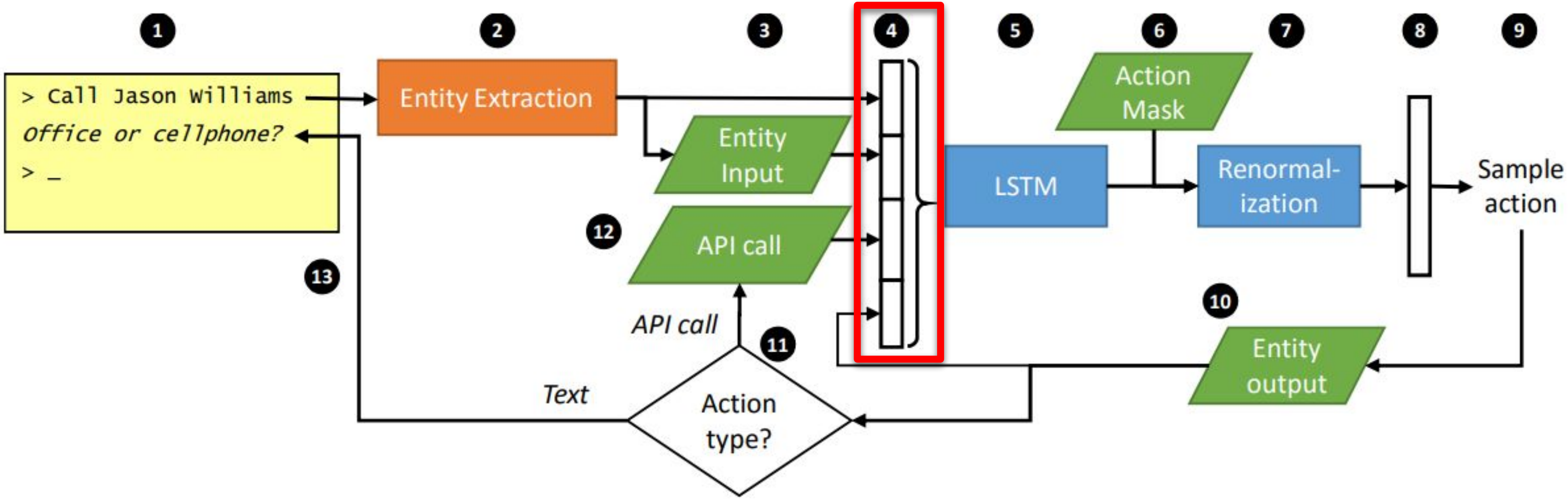


Entity Input

For example: Maps from the text “Jason Williams” to a specific row in a database

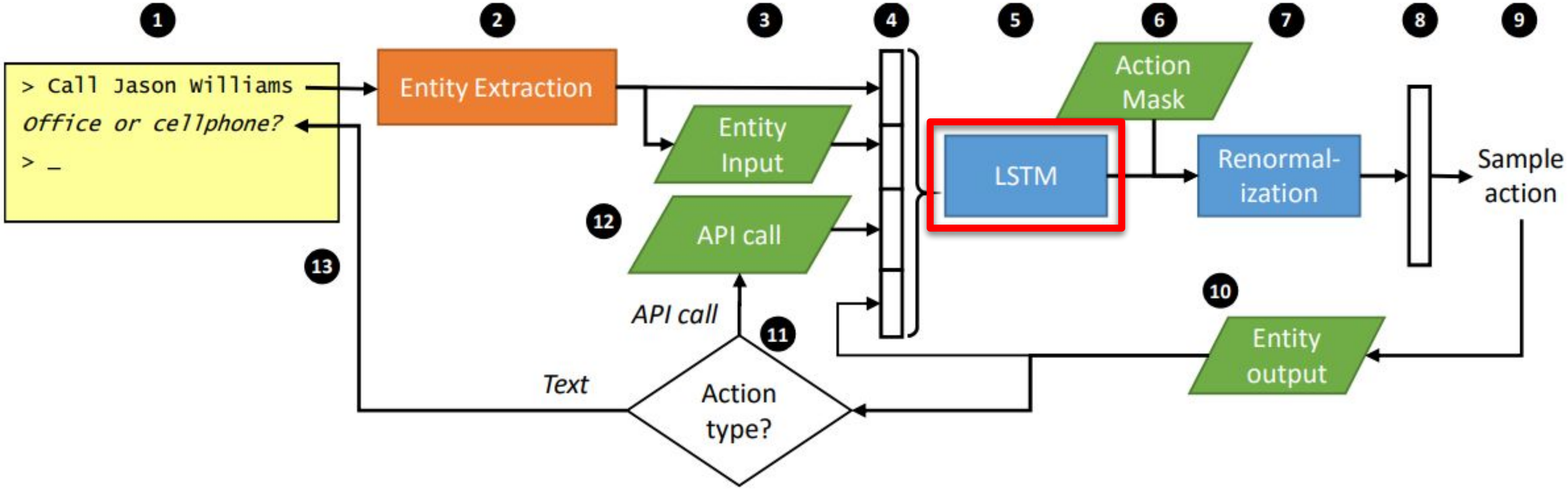


Feature Vector



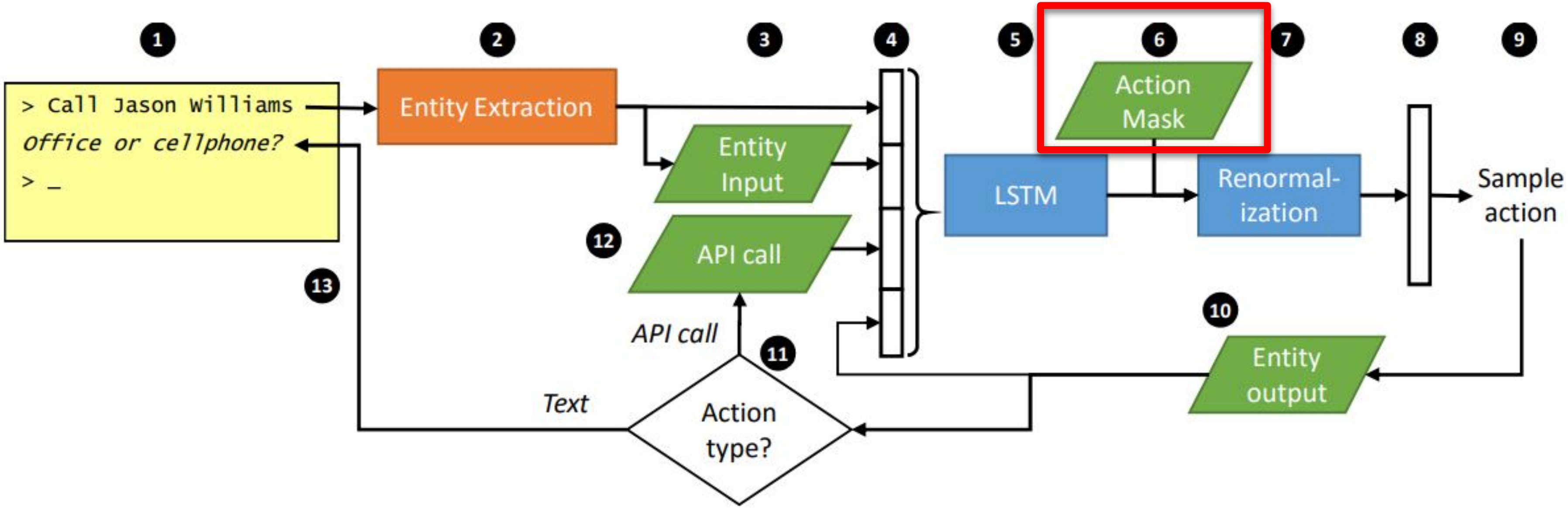
Recurrent Neural Network

LSTM neural network is used because it has the ability to remember past observations arbitrarily long.



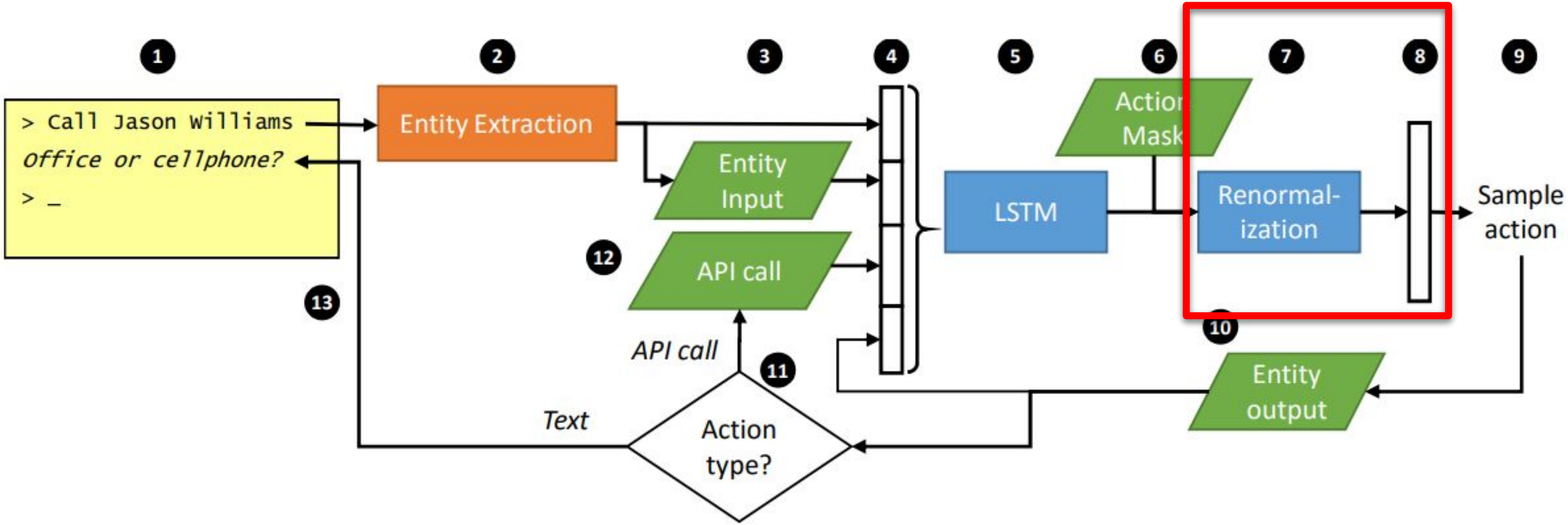
Action Mask

If a target phone number has not yet been identified, the API action to place a phone call may be masked.



Re-normalization

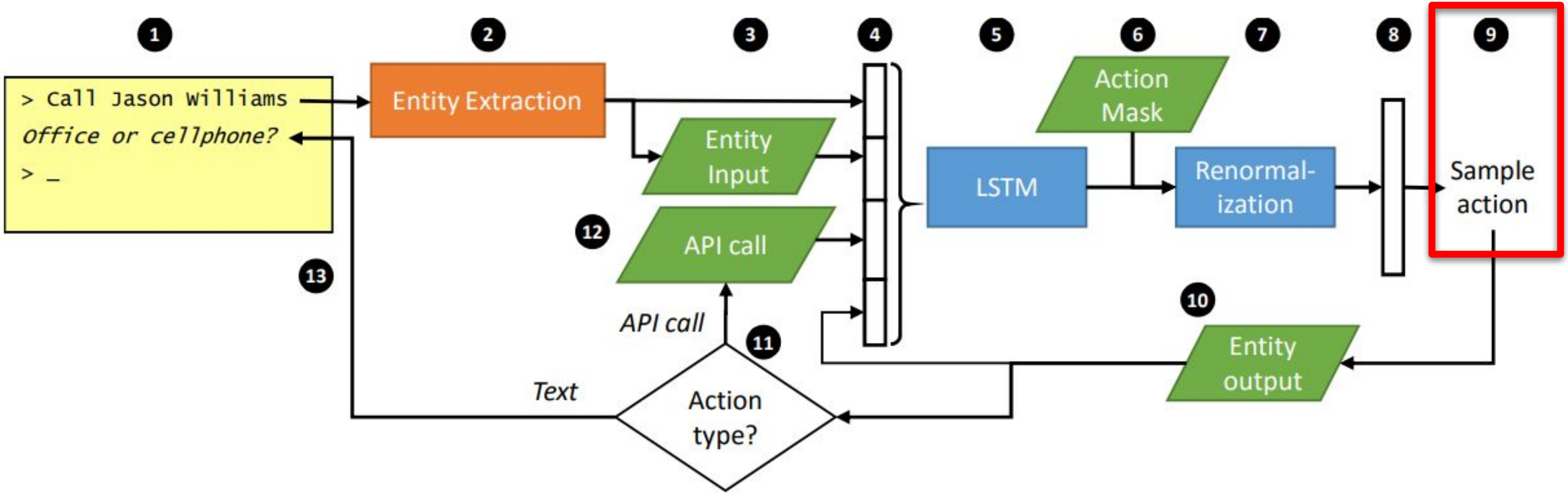
Pr{masked actions} = 0 ⟶ Re-normalize into a probability distribution



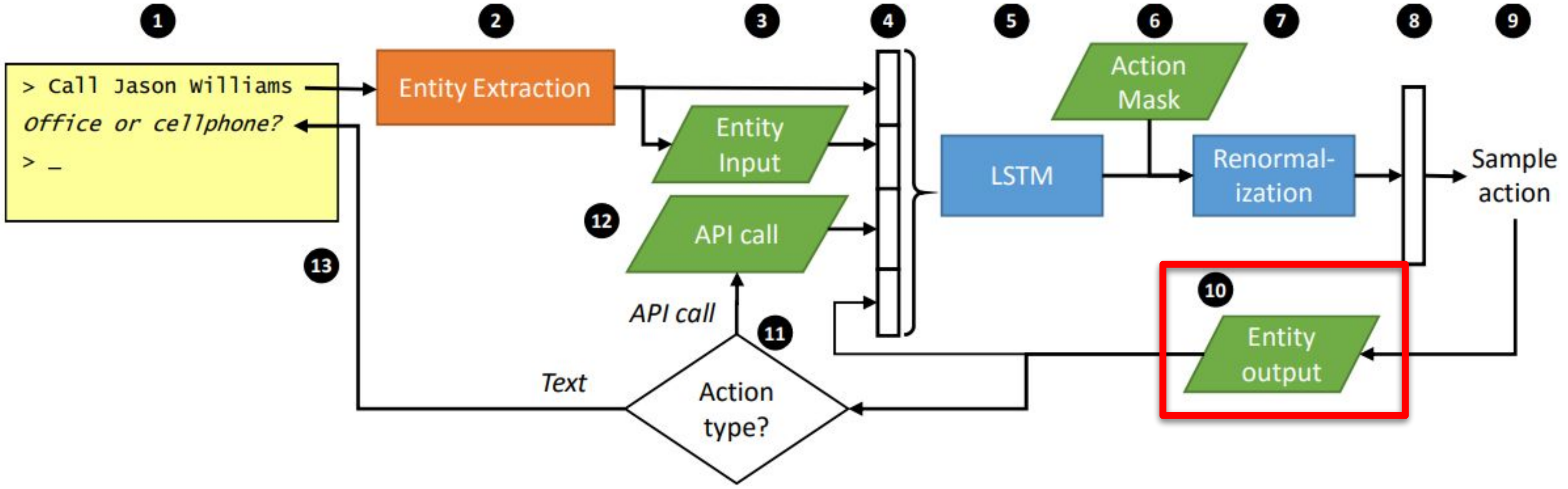
Sample Action

RL: sample from the distribution

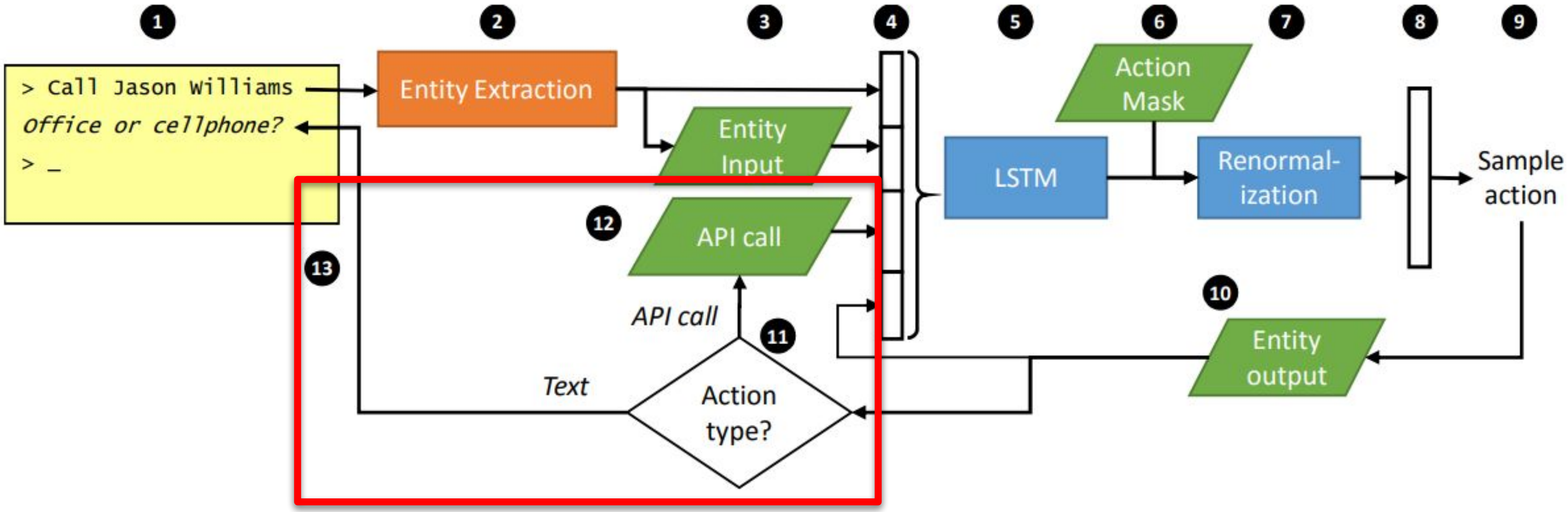
SL: select action with highest probability



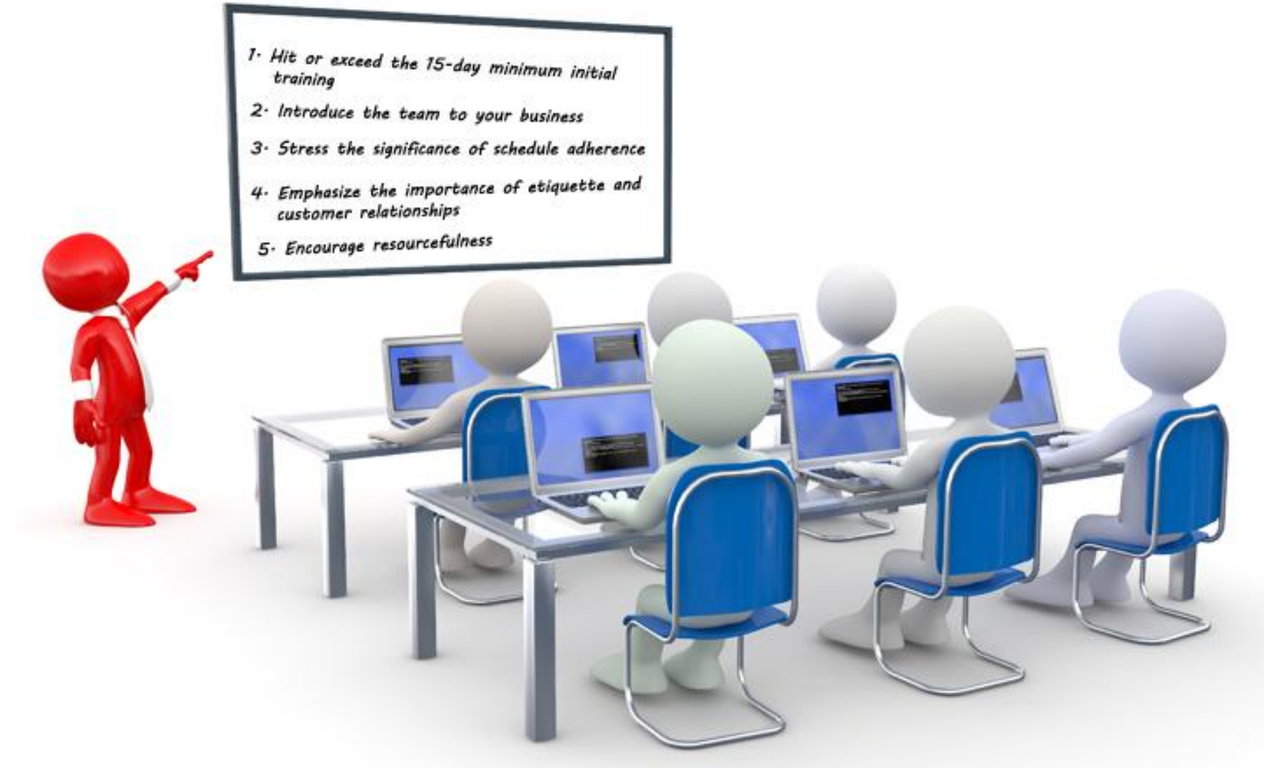
Entity Output



Taking Action



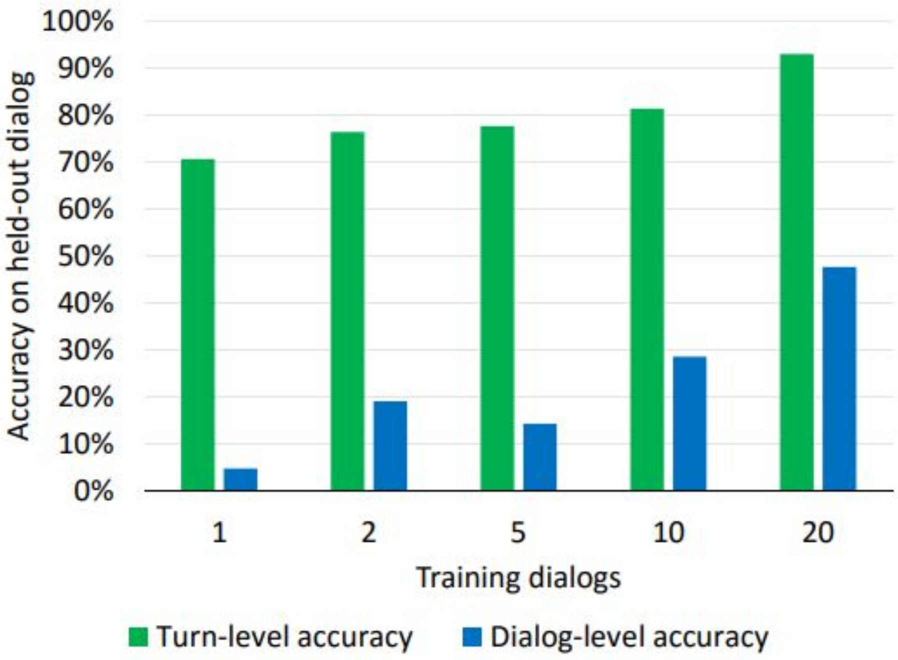
Training the Model

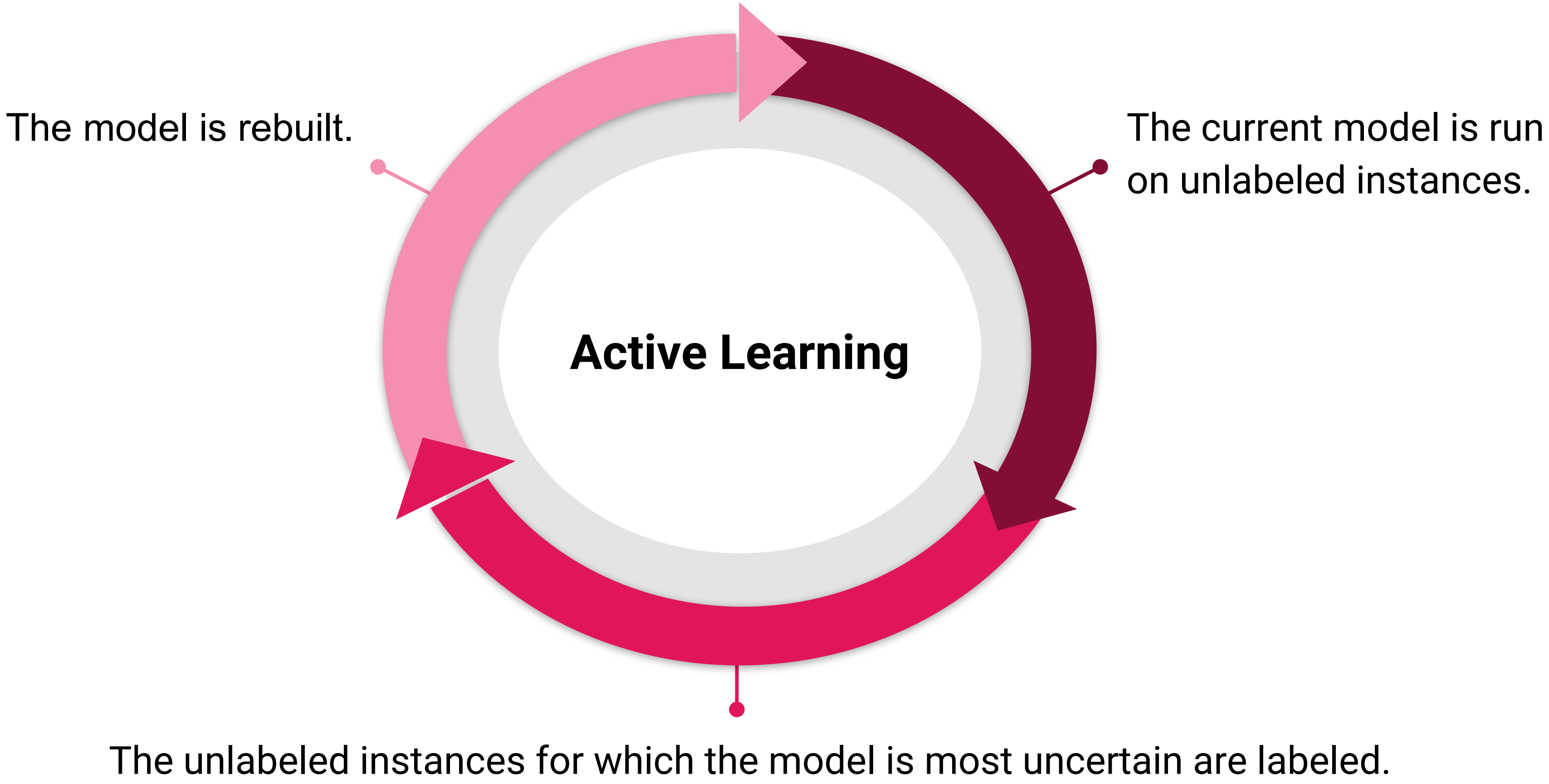


Optimizing with supervised learning

Prediction accuracy

- Loss = categorical cross entropy
- Training sets = 1, 2, 5, 10, and 20 dialogues
- Test set = one held out dialogue




The model is rebuilt.
The current model is run on unlabeled instances.
Active Learning
The unlabeled instances for which the model is most uncertain are labeled.

Active learning

- For active learning to be effective, the scores output by the model must be a good indicator of correctness.
- 80% of the actions with the lowest scores are incorrect.
- Re-training the LSTM is fast

→ **Labeling low scoring actions will rapidly improve the performance.**

Optimizing with reinforcement learning

Policy gradient

Dialog history at time t

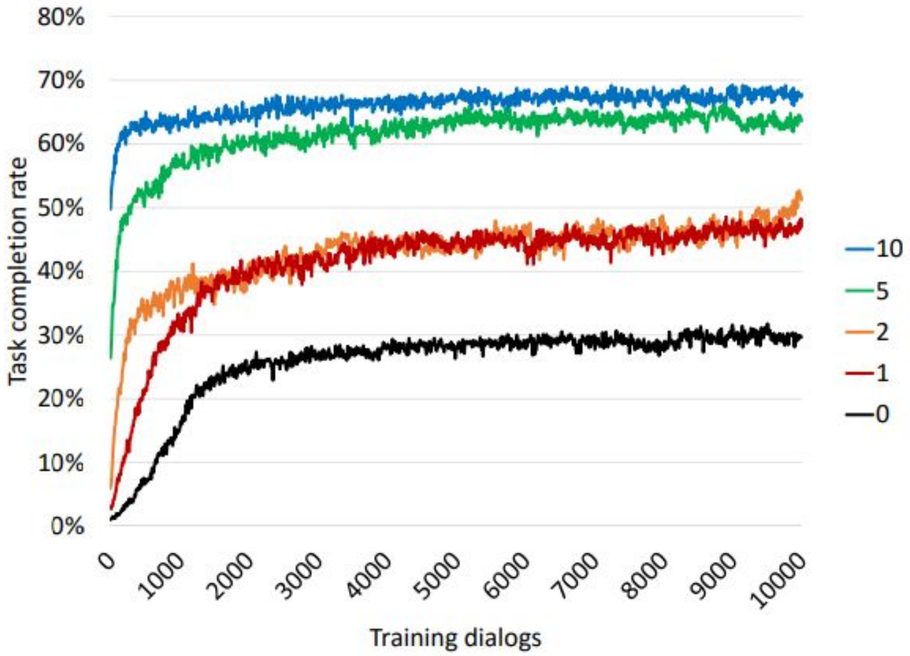
Return of the dialogue

$$\mathbf{w} \leftarrow \boxed{\mathbf{w}} + \alpha\Big(\sum_t \nabla_w \log \boxed{\pi(a_t|\boxed{h_t}; w)}\Big)(\boxed{R} - b)$$

Weights of the LSTM

The LSTM which outputs a distribution over actions

RL Evaluation



(a) TCR mean.

(b) TCR standard deviation.

Conclusion

1. This paper has taken a first step toward an end-to-end learning for task-oriented dialog systems.
2. The LSTM automatically extracts a representation of the dialogue state (no hand-crafting).
3. Code provided by the developer can enforce business rules on the policy.
4. The model is trained using both SL & RL.

Thank you