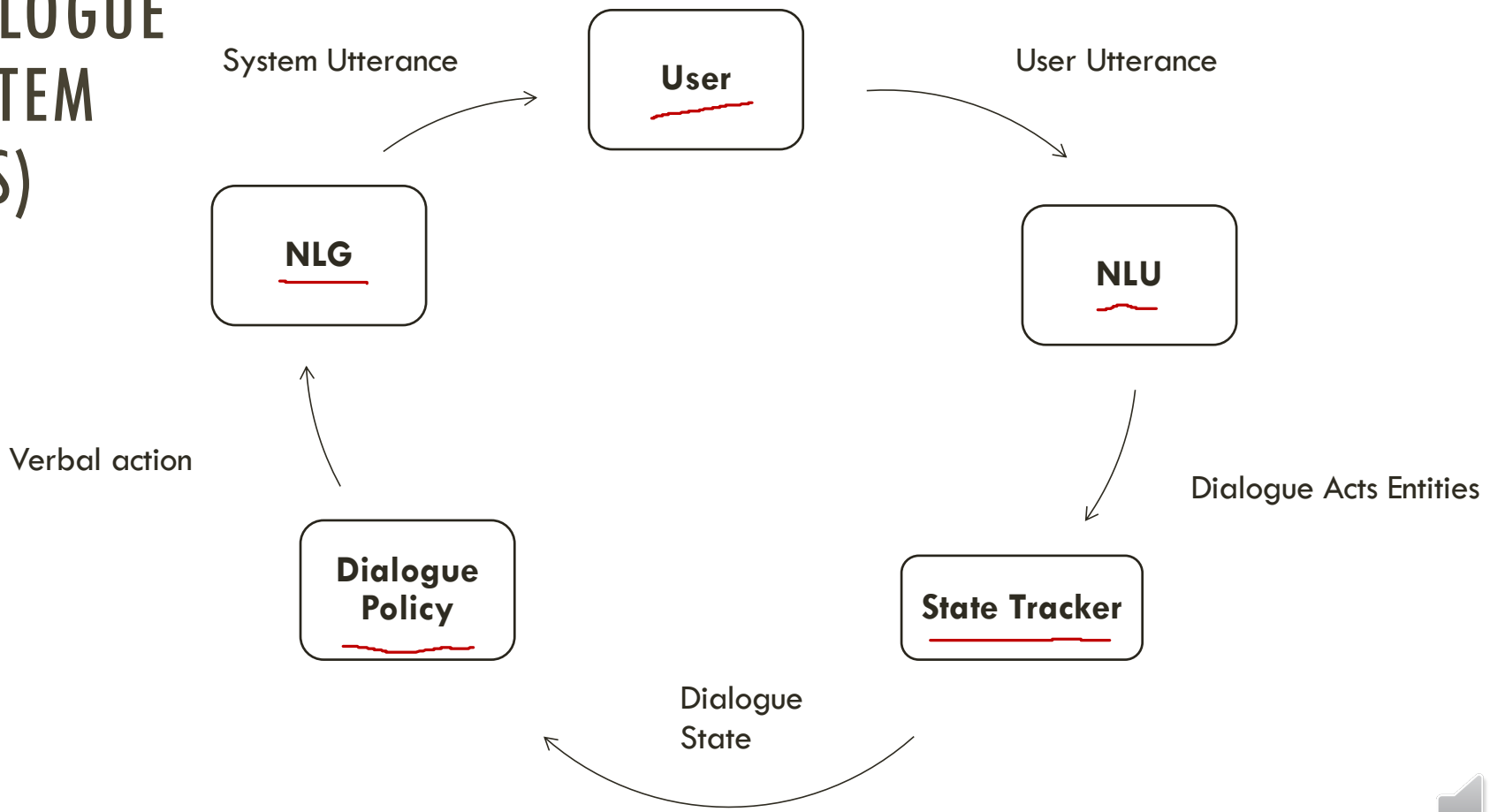
REINFORCEMENT LEARNING FOR TASK-ORIENTED DIALOGUE SYSTEMS

"Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning"

Tiancheng Zhao and Maxine Eskenazi
Carnegie Mellon University

Presented by: Hytham Farah
CS885 – University of Waterloo – July 12

SPOKEN DIALOGUE SYSTEM (SDS)



DIALOGUE ACT EXAMPLE

Tag	Description	Agreement
H	**Hint:** The tutor gives advice to help the student proceed with the task	.50
DIR	**Directive:** The tutor explicitly tells the student the next step to take	.63
ACK	**Acknowledgement:** Either the tutor or the student acknowledges previous utterance; conversational grounding	.73
RC	**Request for Confirmation:** Either the tutor or the student requests confirmation from the other participant (e.g., *"Make sense?"*)	Insufficient data
RF	**Request for Feedback:** The student requests an assessment of his performance or his work from the tutor	1.0
PF	**Positive Feedback:** The tutor provides a positive assessment of the student's performance	.90
LF	**Lukewarm Feedback:** The tutor provides an assessment that has both positive and negative elements	.80
NF	**Negative Feedback:** The tutor provides a negative assessment of the student's performance	.40
Q	**Question:** A question which does not fit into any of the above categories	.95
A	**Answer:** An answer to an utterance marked Q	.94
C	**Correction:** Correction of a typo in a previous utterance	.54
S	**Statement:** A statement which does not fit into any of the above categories	.71
O	**Other:** Other utterances, usually containing only affective content	.69

Table taken from Young Ha et al. 2020

CHALLENGES OF SDS

1 – Credit assignment Problem

- Hard to locate source of error

2 - Process Independence

- Optimizing one module may make the other modules sub-optimal

SOLUTION: COMBINE THE MODULES

Model the SDS as an RL agent with of an underlying POMDP

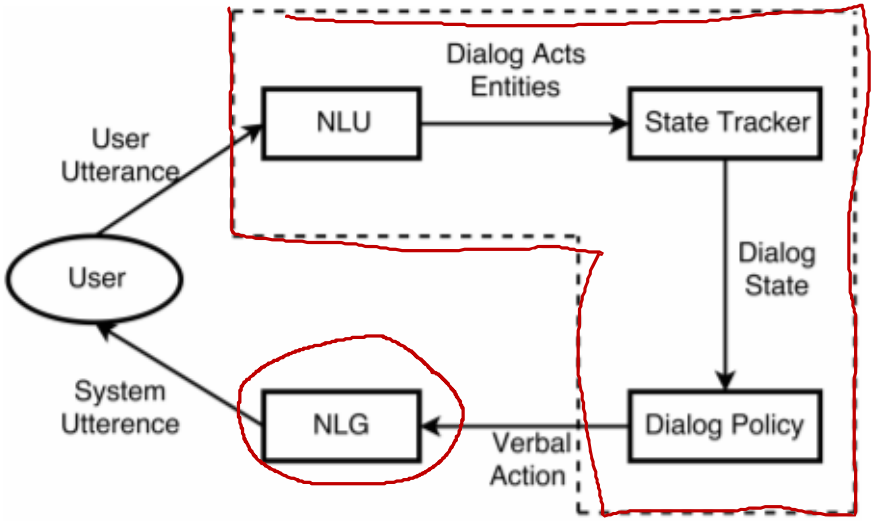
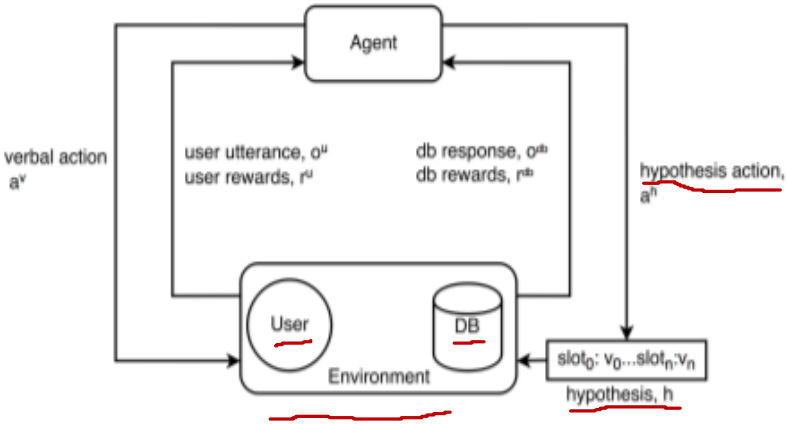


Figure 1: Conventional pipeline of an SDS. The proposed model replaces the modules in the dotted-line box with one end-to-end model.

COMPARING THE TWO PIPELINES

New SDS



Conventional SDS

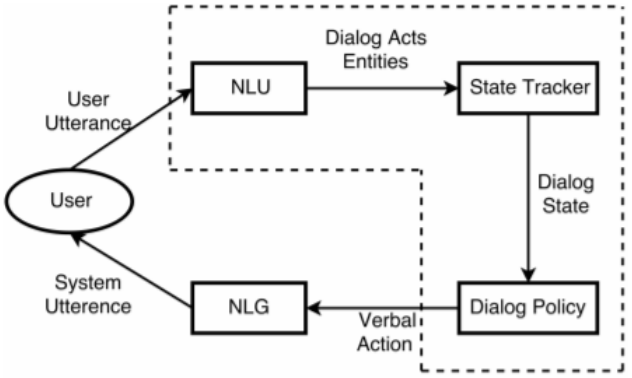
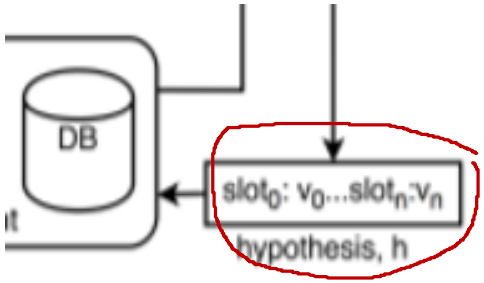


Figure 1: Conventional pipeline of an SDS. The proposed model replaces the modules in the dotted-line box with one end-to-end model.

2 ENVIRONMENTS

Environment	Action	Observation	State	Reward
User (E^u)	Dialogue Action	User Utterance	User Intention	User Rewards (eg. user satisfaction, completion of intended task)
Database (E^{db})	Hypothesis Action	All objects in the database which satisfy the hypothesis	The object that needs to be retrieved from the database	Database Rewards

THE HYPOTHESIS…



Determines what information is retrieved from the database

Tracks the state of the dialogue

THE HYPOTHESIS ACTION…

Modifies the value for one of the slots

Movie Recommender Example: change the value of the **genre** slot from **unknown** to **horror.**

DATABASE REWARD

Add a pseudo reward function:

$$\bar{R}(s, a, s') = R(s, a, s') + F(s, a, s')$$
$$F(s, a, s') = \gamma\phi(s') - \phi(s)$$

Where:

- D is the total number of elements in the database
- d_t is the number of elements being considered at state s_t
- P_{max} is a constant.

$$\phi(s_t) = P_{max}(1 - \frac{d_t}{D}) \quad \text{if } d_t > 0$$
$$\phi(s_t) = 0 \quad \text{if } d_t = 0$$

Intuition: Reward the agent for narrowing down the database as quickly as possible.

BELIEF STATE WITH LSTM

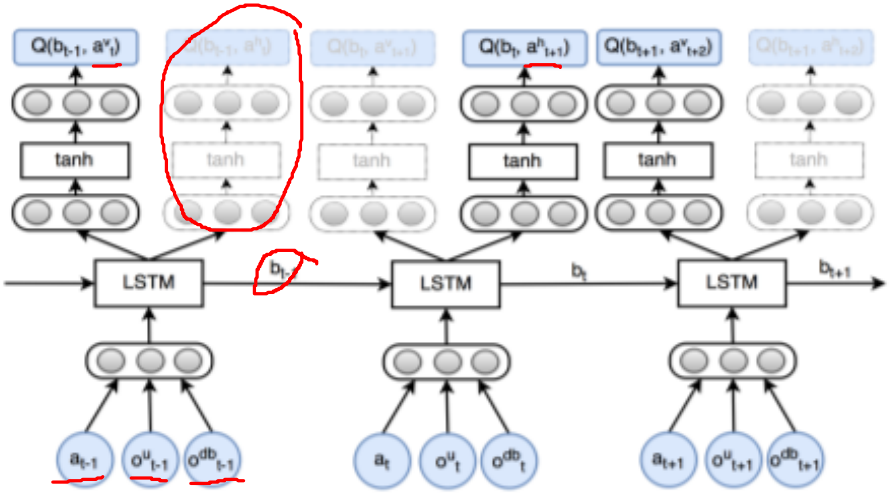


Figure 3: The network takes the observation o_t at turn t. The recurrent unit updates its hidden state based on both the history and the current turn embedding. Then the model outputs the Q-values for all actions. The policy network in grey is masked by the action mask

WORKED EXAMPLE: 20 QUESTIONS GAME

The user thinks of a famous person

The agent may ask the user 20 yes/no questions

User answers honestly with any natural language utterance:

- Yes ("I think so", "Yep", "He does", "Certainly" etc…)
- No ("He is not", "Nope", "Certainly not", etc…)
- I don't know ("Not sure", "Maybe", "No clue", etc…)

The game terminates when:

- The agent guesses the correct answer (+30)
- Max game length (100 steps, $\gamma = .99$) is reached OR no people in the database match the current hypothesis (-30)

Wrong guesses (max 10) will result in a penalty (-5)

IMPLEMENTATION DETAILS

100 famous people:

6 Attributes / Slots:

- Birthplace, Birthday, Degree, Gender, Profession, Nationality.

Agent has 31 Questions to select

Simulator:

- Samples person uniformly at random.
- Has a 5% chance that it will consider an attribute as unkown.
- Natural language response generated from SWDA corpus. Sampled using the same distribution.

Attribute	Q_a	Example Question
Birthday	3	Was he/she born before 1950?
Birthplace	9	Was he/she born in USA?
Degree	4	Does he/she have a PhD?
Gender	2	Is this person male?
Profession	8	Is he/she an artist?
Nationality	5	Is he/she a citizen of an Asian country?

Table 1: Summary of the available questions. Q_a is the number of questions for attribute a.

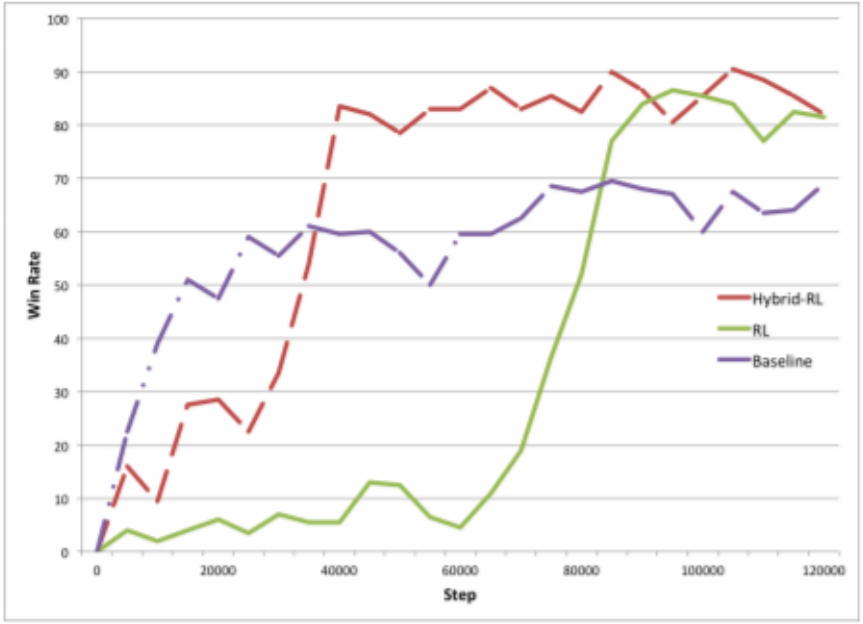
Intent	SWDA tags
Yes	Agree, Yes answers, Affirmative non-yes answers
No	No answers, Reject, Negative non-no answers
Unknown	Maybe, Other Answer

Table 2: Dialog act mapping from SWDA to *yes/no/unknown*

RESULTS: WIN RATE

	Win Rate (%)	Avg Turn
Baseline	68.5	12.2
RL	85.6	21.6
Hybrid-RL	90.5	19.22

Table 3: Performance of the three systems



RESULTS: STATE TRACKING

	Unknown	Yes	No
Baseline	0.99/0.60	0.96/0.97	0.94/0.95
RL	0.21/0.77	1.00/0.93	0.95/0.51
Hybrid-RL	0.54/0.60	0.98/0.92	0.94/0.93

Precision / Recall

Precision = Correct Guesses / Total Guesses
i.e. The higher the less false positives

Recall =
Correct Guesses / All possible correct Gueses.
i.e. The higher the more likely it is to find correct labels.

CONCLUSION

Related Work:

- Dialogue State Tracking
- End-to-End SDS

Future Challenges:

- Improving sample efficiency
- More thorough empirical evaluation
- Techniques that allow integration of domain knowledge