Composite task completion bot with Hierarchical Reinforcement Learning

Jianfeng Gao

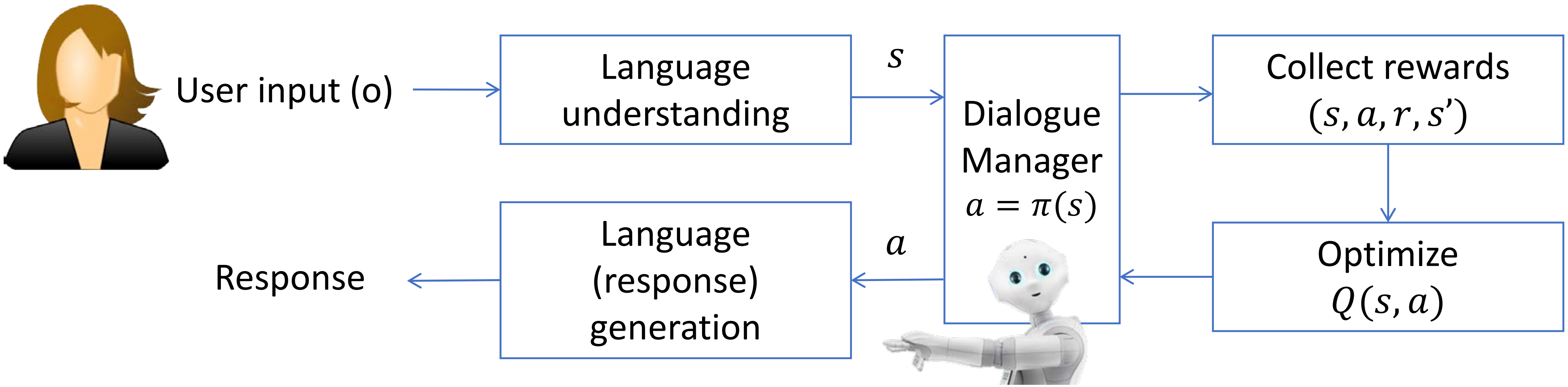
Joint work with **Baolin Peng, Xiujun Li, Lihong Li,**

Asli Celikyilmaz, Sungjin Lee, Kam-Fai Wong

Microsoft AI & Research

Dec. 8, 2017, Long Beach

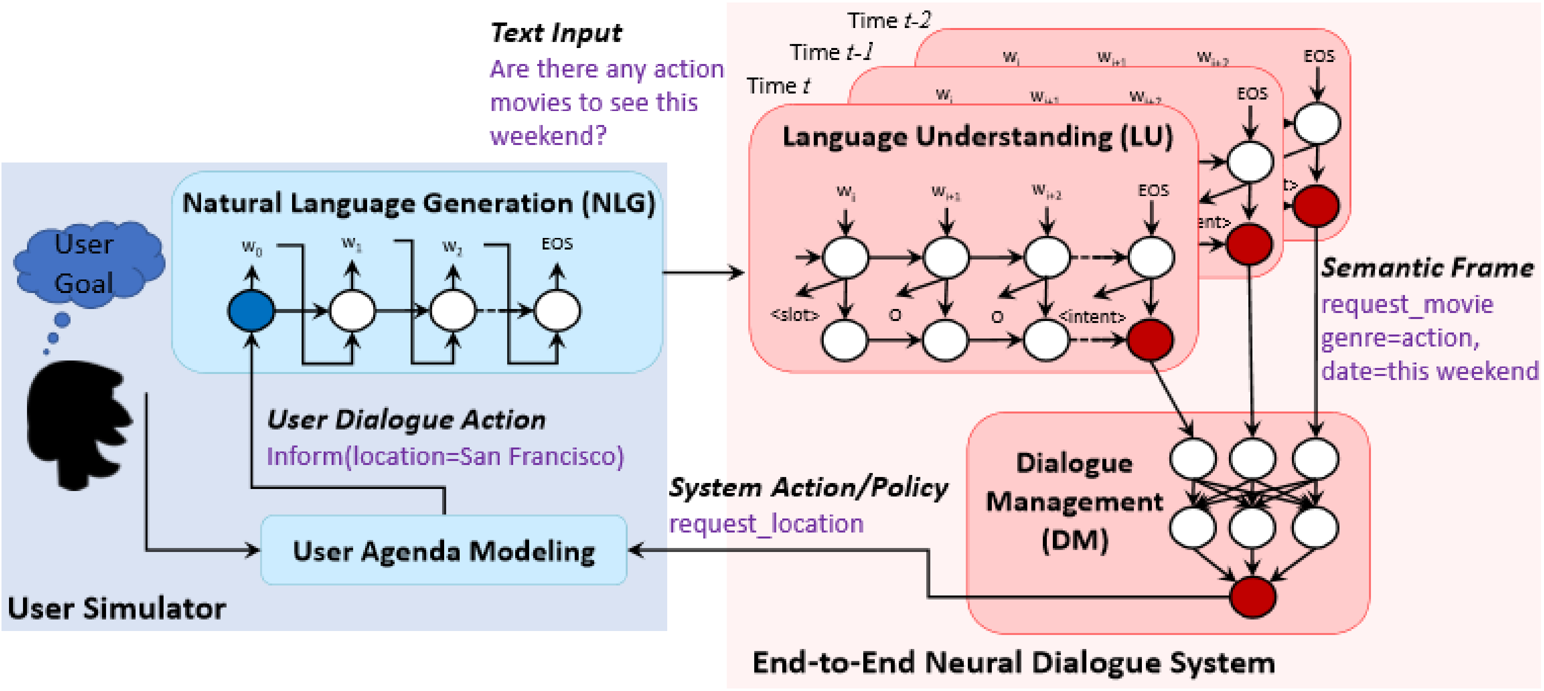
Conversation as Reinforcement Learning



Types of Bots	State	Action	Reward
Task Completion Bots (DSTC7 proposal) (e.g., Movies, Restaurants, Travel)	User input + Context	Dialog act + slot_value	Task success rate # of turns
Info Bots (e.g., Q&A bot over KB, Web)	Question + Context	Clarification questions, Answers	Relevance of answer # of turns
Social Bot (DSTC7 proposal) (e.g., XiaoIce, Zo)	Conversation history	Response	Engagement(?)

A user simulator for RL and evaluation

- DSTC7 proposal: E2E Task-Completion Dialogue Track

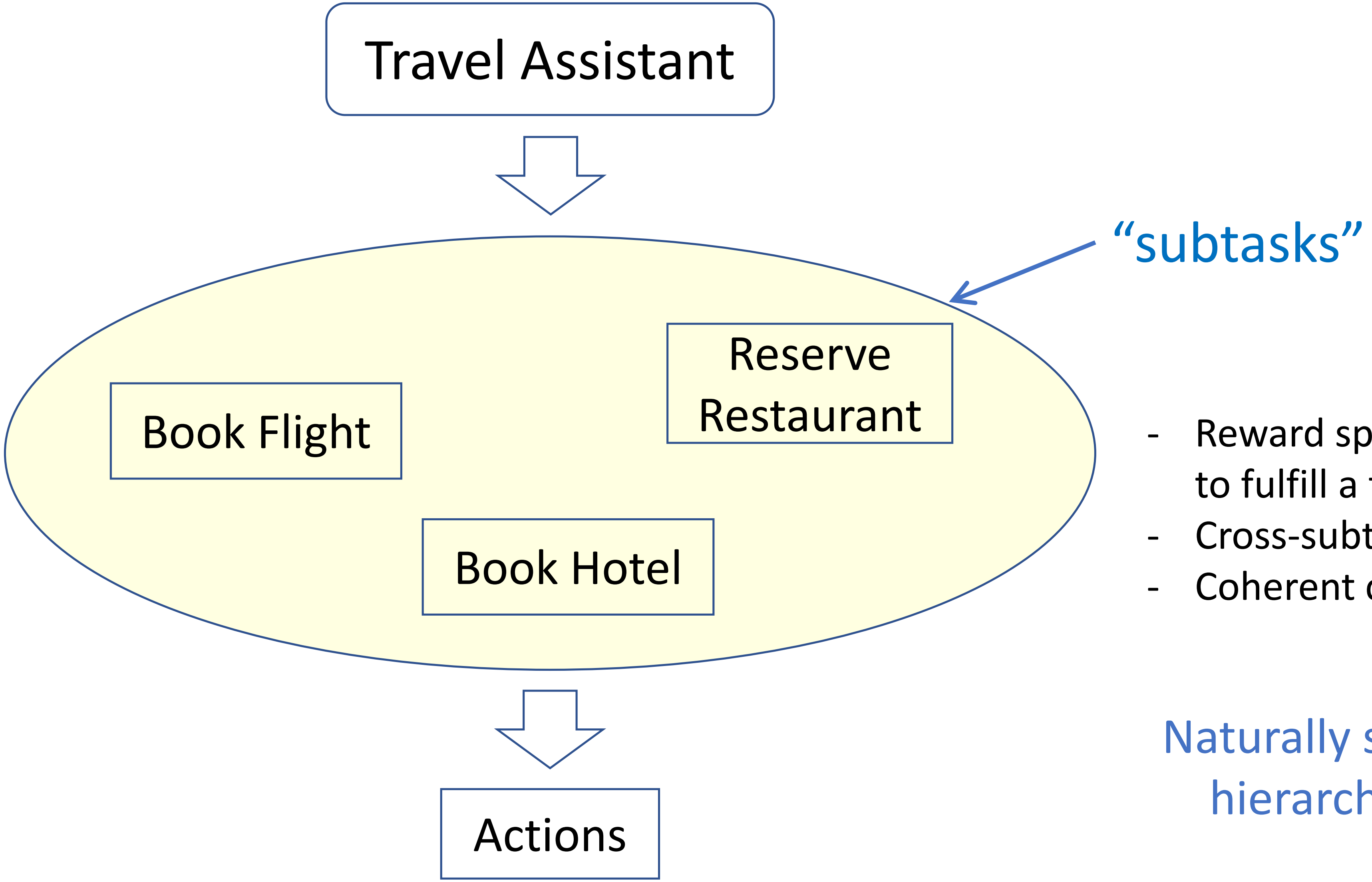


- Robustness: automatic action selection based on uncertainty by RL
- Flexibility: allow user-initiated behaviors
- Reproducibility: a R&D setting that allows consistent comparisons of competing methods

[Li+ 17] https://github.com/MiuLab/TC-Bot

Composite task completion bot with Hierarchical RL

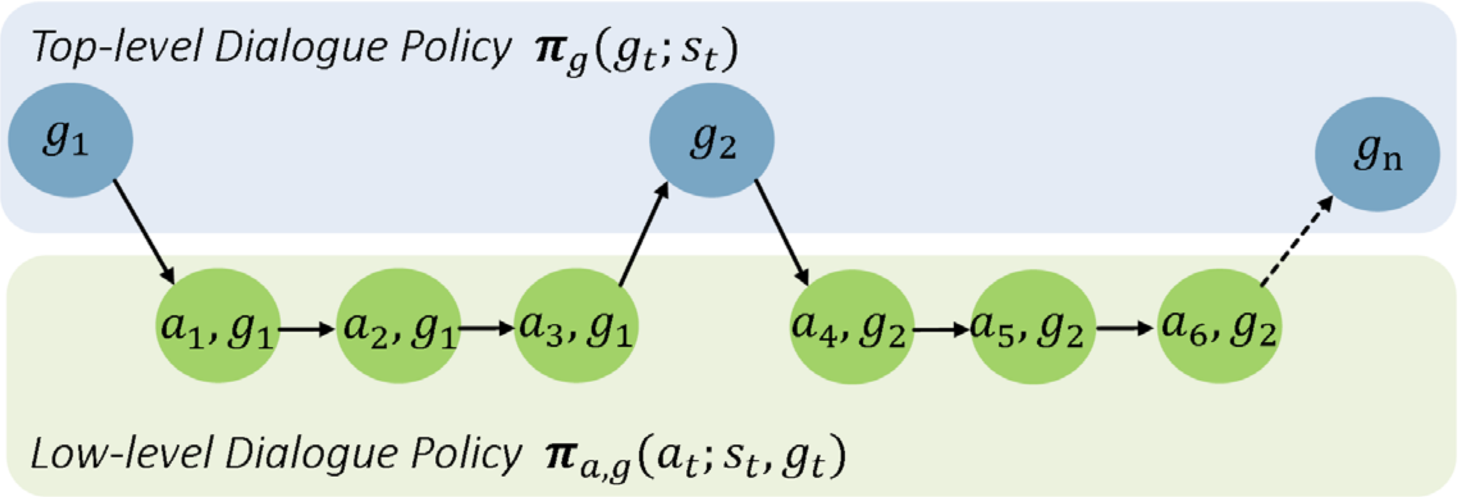
[Peng+ 17]

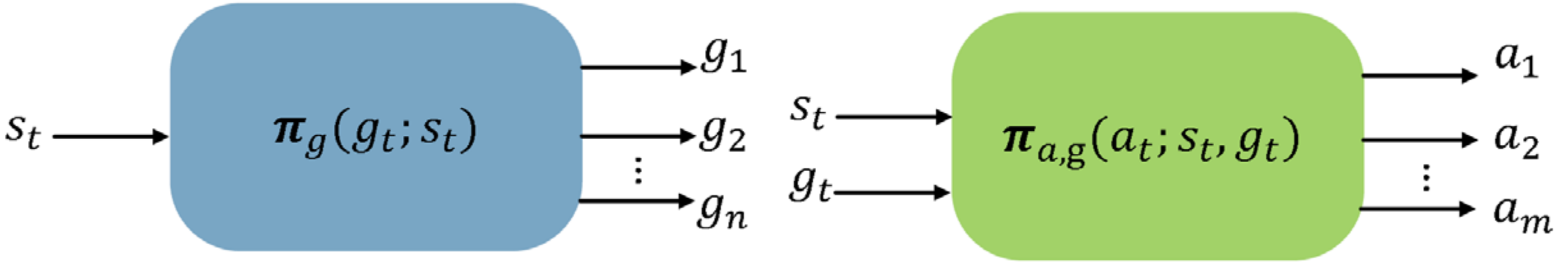


- Reward sparsity: takes many turns to fulfill a task.
- Cross-subtask slot constraints
- Coherent conversation

Naturally solved by hierarchical RL

A hierarchical policy learner



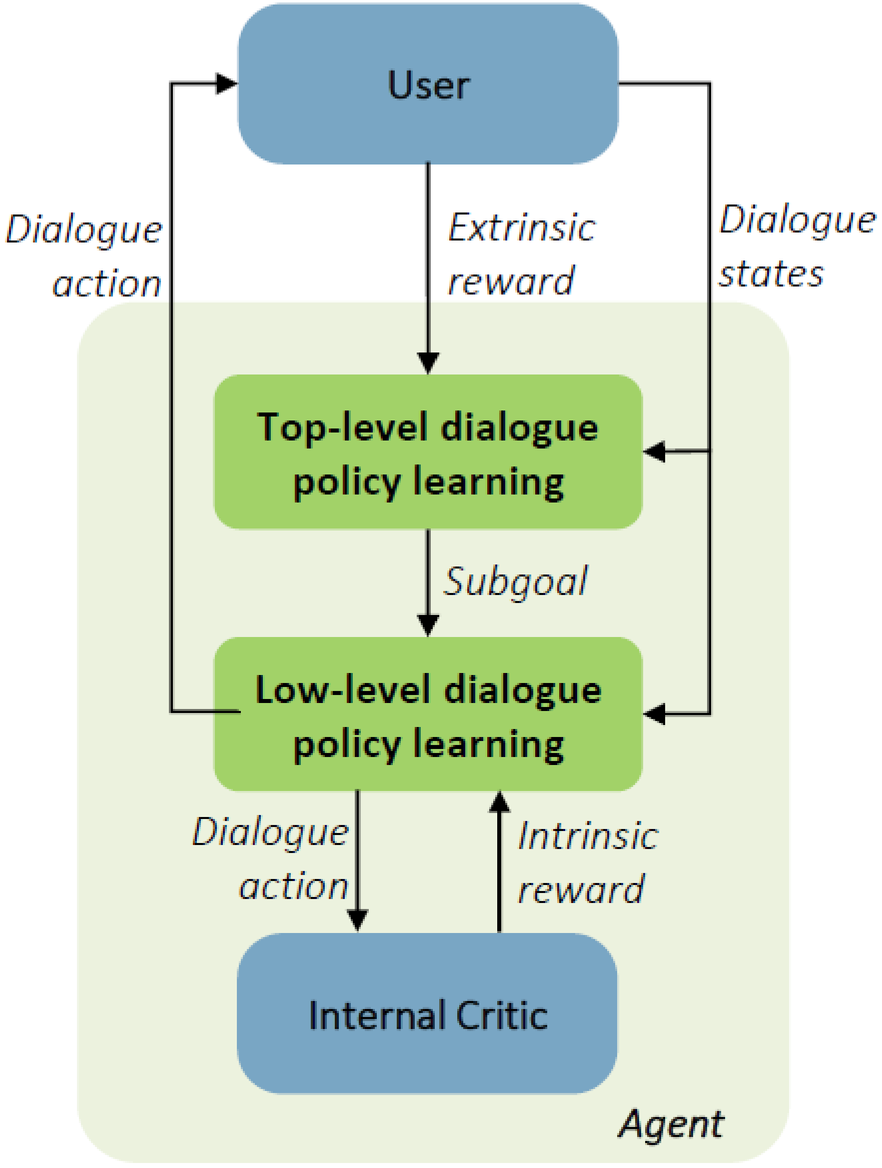


Global state tracker: ensure all cross-subtask constraints be satisfied

Similar to HAM [Parr & Russell 98] and hierarchical DQN [Kulkarni+ 16]

Exploiting structural Info for efficient exploration

- Internal Critic gives intrinsic reward
  - How likely a subtask is completed
- Encourage agent to complete a subtask before moving to another subtask
- Better user experience: less freq switch btw subtasks



User simulators

- Type A:
  - Have hard constraints for the *flight* and *hotel* (single value for all informed slots)
  - No preference on which subtask to accomplish first
- Type B:
  - At least one of informed slots in *book-flight-ticket* have multiple values
  - Prefer to start with *book-flight-ticket*
- Type C:
  - At least one of informed slots in *reserve-hotel* have multiple values
  - Prefer to start with *reserve-hotel*

Results on simulated users

	Type A			Type B			Type C		
Agent	Succ.	Turn	Reward	Succ.	Turn	Reward	Succ.	Turn	Reward
Rule	.322	46.2	-24.0	.240	54.2	-42.9	.205	54.3	-49.3
Rule+	*.535*	*82.0*	*-3.7*	*.385*	*110.5*	*-44.95*	*.340*	*108.1*	*-51.85*
RL	.437	45.6	-3.3	.340	52.2	-23.8	.348	49.5	-21.1
HRL	**.632**	**43.0**	**33.2**	**.600**	**44.5**	**26.7**	**.622**	**42.7**	**31.7**

(a) Success Rate of User Type A (b) Success Rate of User Type B (c) Success Rate of User Type C

Figure 4: Learning curves of dialogue policies for different User Types under simulation

Human Evaluations

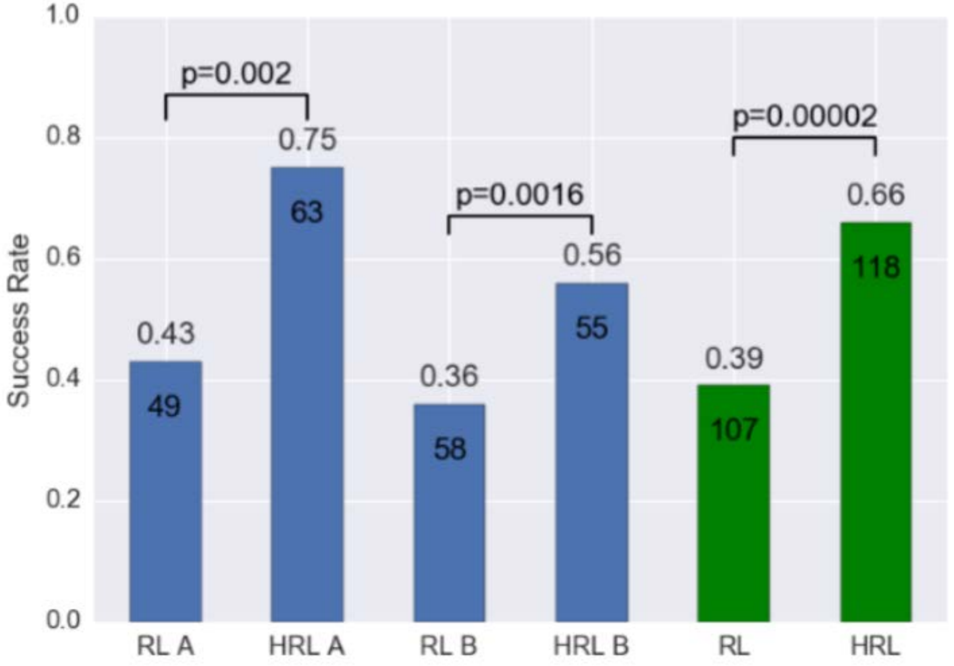


Figure 5: Performance of HRL agent versus RL agent tested with real users: success rate, number of tested dialogues and p-values are indicated on each bar; the rightmost green ones are for total (difference in mean is significant with $p < 0.01$).

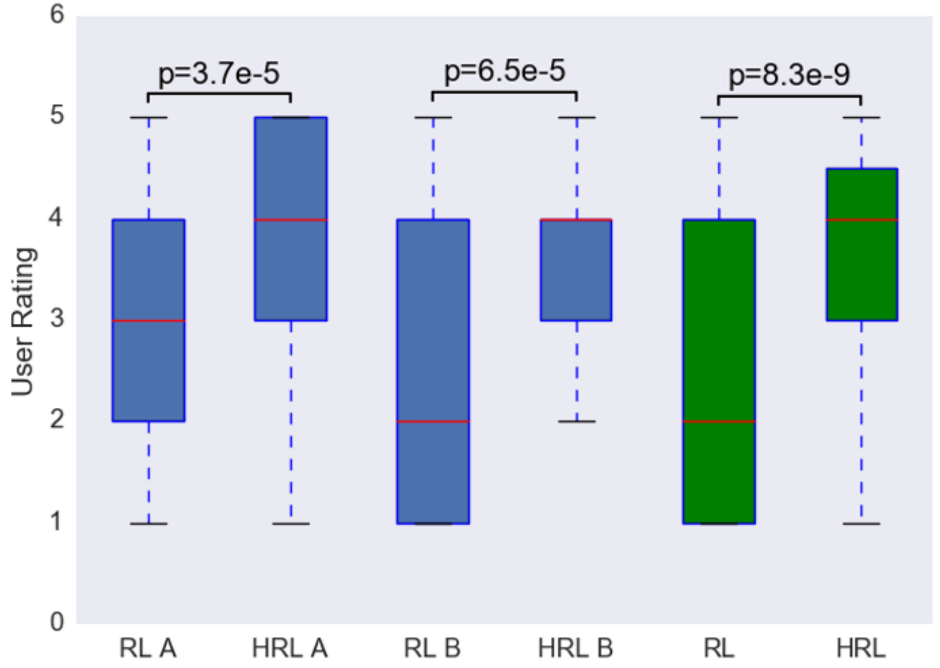


Figure 6: Distribution of user ratings for HRL agent versus RL agent, and total.

User goal – ***Two tickets for "Deadpool" tomorrow 9PM at AMC Pacific Place 11 theater, Seattle.***

RULE BASED SYSTEM

User: Do you have any ticket for movie 'Deadpool' at 9:00 pm?
Agent: Which theater do you want?
User: AMC pacific place 11 theater, please!
Agent: How many tickets do you want?
User: Two!
Agent: What date?
User: Tomorrow!
Agent: Which city do you want?
User: Seattle!
Agent: Do you have any specification for zip code?
User: No, I do not care.
Agent: Do you have any distance constraint?
User: No, I do not care.
Agent: How about video format? 2d? Imax 3d?
User: Any kind is okay.
Agent: Do you have any constraint to price?
User: No!
Agent: Okay, I will issue 2 tickets for you, tomorrow 9:00 pm at AMC pacific place 11 theater, Seattle,
movie 'Deadpool'.
User: Okay, thanks.
Agent: Thanks!

Wasteful!
(Very likely the user does not care.)

REINFORCEMENT LEARNING SYSTEM

User: I want two tickets for movie 'Deadpool'.
Agent: What is the starttime you prefer?
User: 9:00 PM
Agent: Which day?
User: Tomorrow
Agent: Which city?
User: Seattle
Agent: Which theater do you prefer?
User: AMC Pacific Place 11 theater.
Agent: Okay, I will issue 2 tickets for you, tomorrow 9:00 pm at AMC pacific place 11 theater, Seattle,
movie 'Deadpool'.
User: Okay, thanks.
Agent: Thanks!

User goal – ***Two tickets for "Hail Caesar" tomorrow 7:15PM at AMC Lowes Oak tree 6, Seattle.***

RULE BASED SYSTEM

User: Do you any ticket for movie 'Hail Caesar' at theater AMC Lowes Oak Tree 6?
Agent: What start time do you prefer?
User: **7:15 pm** Tomorrow
Agent: Which city you are looking for?
User: Seattle
Agent: How many tickets do you want?
User: Two
Agent: Okay, I have issued two tickets for movie 'Hail Caeser' tomorrow **7:50 pm** at theater AMC Lowes Oak Tree 6, Seattle.
User: Oh, no.
Agent: Thanks!

Task failed!
(Cannot correct the LU mistake.)

REINFORCEMENT LEARNING SYSTEM

Agent: What start time do you prefer?
User: **7:15 pm** Tomorrow
Agent: Which city you are looking for?
User: Seattle
Agent: Okay, I will issue you tickets for movie 'Hail Caeser' tomorrow **7:50 pm** at theater AMC Lowes Oak Tree 6, Seattle.
User: Oh, no.
Agent: **It is 7:50 pm, right?**
User: Two tickets for 7:15 pm, please.
Agent: Okay, I will issue you two tickets for movie 'Hail Caeser' tomorrow **7:15 pm** at theater AMC Lowes Oak Tree 6, Seattle.
User: Thanks!
Agent: Thanks!

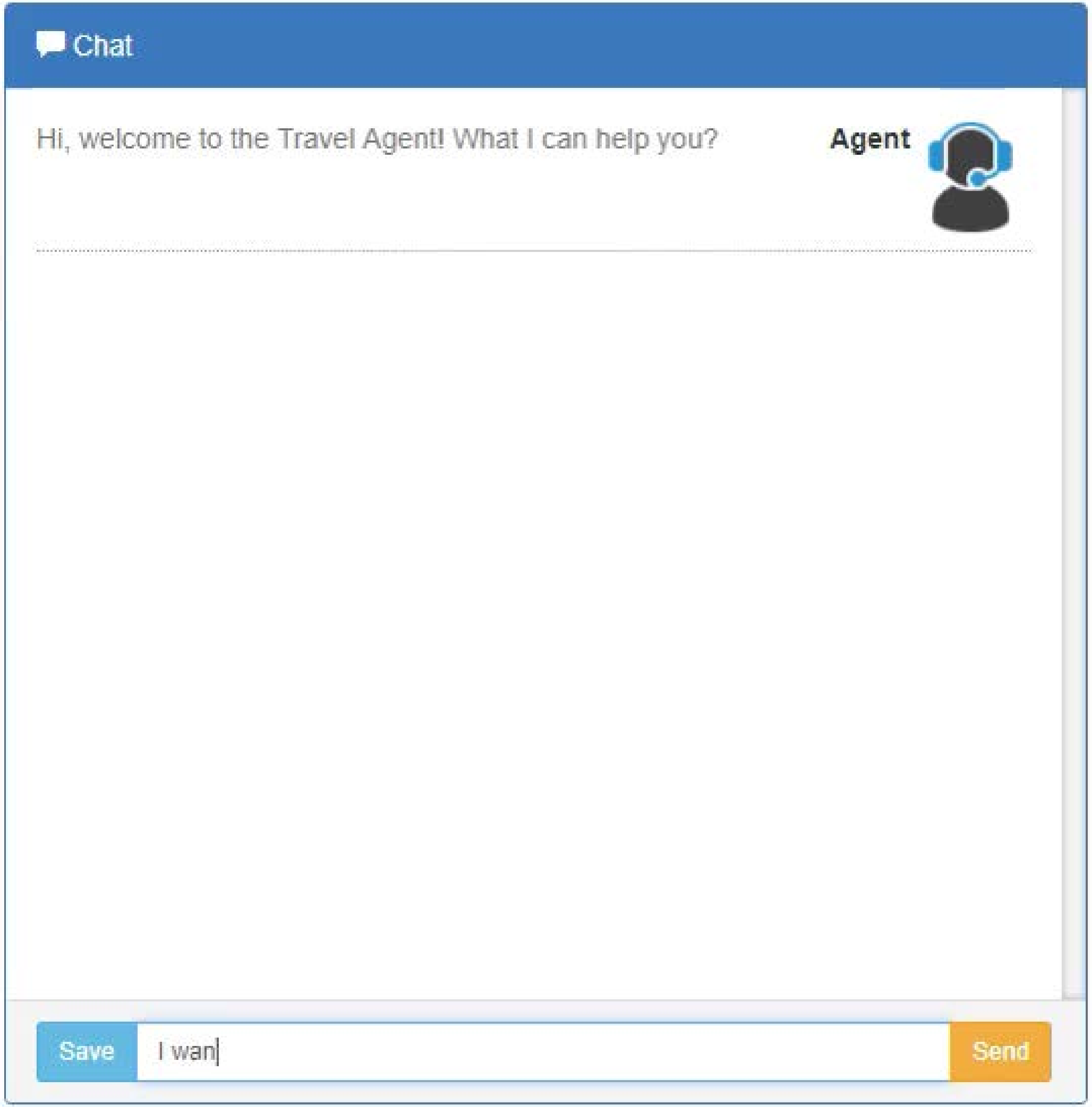
User Goal

reserve-hotel subtask:	*book-flight-ticket* subtask:

{
"request_slots": {
"hotel_price": "?"
"hotel_date_checkout": "?"
"hotel_name": "?"
},
"inform_slots": {
"hotel_city": "Cancun",
"hotel_numberofpeople": "3",
"hotel_date_checkin": "09/20"
}
}
	{
	"request_slots": {
	"price": "?"
	"return_time_dep": "?"
	"return_date_dep": "?"
	"depart_time_dep": "?"
	"seat": "?"
	}
	"inform_slots": {
	"or_city": "Campinas",
	"dst_city": "Cancun",
	"numberofpeople": "3",
	"depart_date: "9/20"
	}
	}

Bolded slots are the joint constraints between two subtasks

RL agent	**HRL agent**
Agent: Welcome to our travel agent. May I know your original city, destination city and departure date? *User*: I depart from Campinas to Cancun and prefer to leave on 09-20. *Agent*: There is a return flight on 9/26 and hotel checkout date is on 9/26 too. *User*: My wife and son will travel with me, three people in total. *Agent*: There is a hotel called Hotel Tropic available for you. *User*: What depart time is available? *Agent*: Departure time is at 10:00AM. *User*: What is the departure time of return flight? *Agent*: It is 4:00PM. *User*: Which seat is available? *Agent*: Only Business class available. *User*: What's the price for the flight? *Agent*: Total is 1399 dollars for flight. *User*: What's the price for the hotel? *Agent*: It is 1091 dollars for hotel. *User*: Could you please help me book the tickets? *Agent*: Departure flight on 9/20 and hotel checkin date on 9/20 are available. *User*: Could you please help me book the tickets? *Agent*: Okay. Tickets are issued. *User*: Thanks! *Agent*: Thanks!	*Agent*: Welcome to our travel agent. May I know your original city, destination city and departure date? *User*: I depart from Campinas to Cancun and prefer to leave on 09-20. *Agent*: Only Business class seat is available. *User*: What is the depart time? *Agent*: Departure time is at 10:00AM. *User*: My wife and son will travel with me, three people in total. *Agent*: There is one return flight departed at 4:00PM. *User*: What is the date of the return flight? *Agent*: It is on 9/26 and hotel checkout date is also 9/26. *User*: What's the price for the flight? *Agent*: Total is 1399 dollars for flight. *User*: Which hotel is available? *Agent*: A hotel called Hotel Tropic is available for you. *User*: What's the price for the hotel? *Agent*: It is 1091 dollars for hotel. *User*: Could you please help me book the tickets? *Agent*: Okay. Tickets are issued. *User*: Thanks! *Agent*: Thanks!
Success: but the agent does not behave naturally where is an inter-subtask transition from *book-flight-ticket* to *reserve-hotel*.	**Success**


Chat
Hi, welcome to the Travel Agent! What I can help you?
Agent
Save
I wan
Send

Subgoal discovery for HRL:

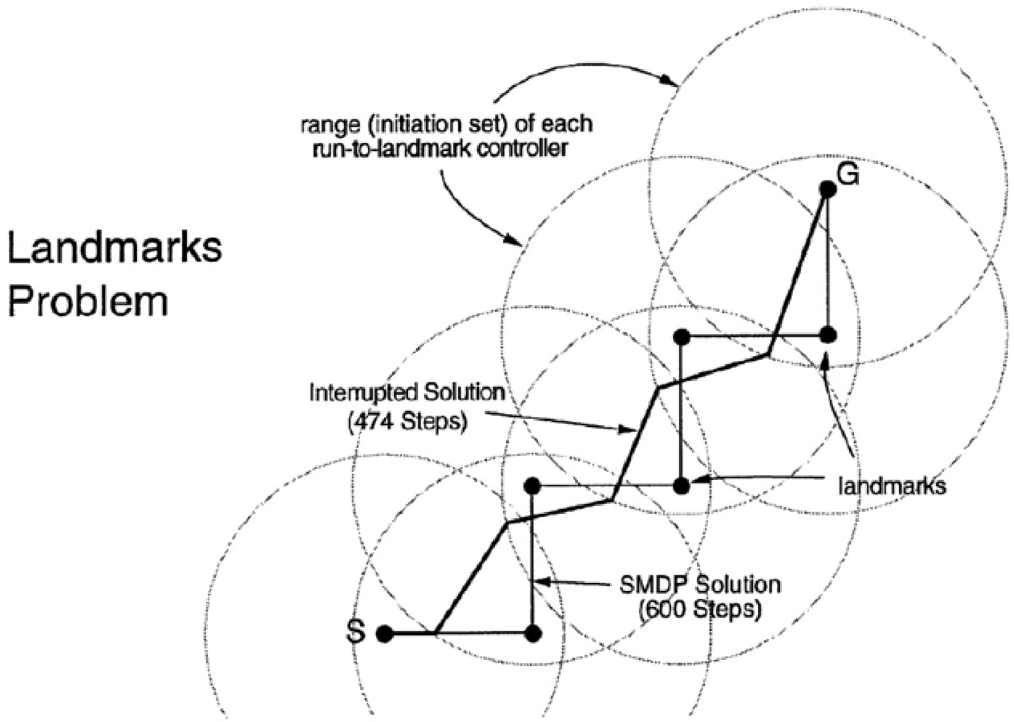


Figure 3: Subgoals for the landmarks problem (Sutton et al., 1999). Though the solution with subgoals may not be optimal, having the subgoals could usually reduce the search space, and potentially accelerate the learning efficiency.

divided and conquer

The 4-room game

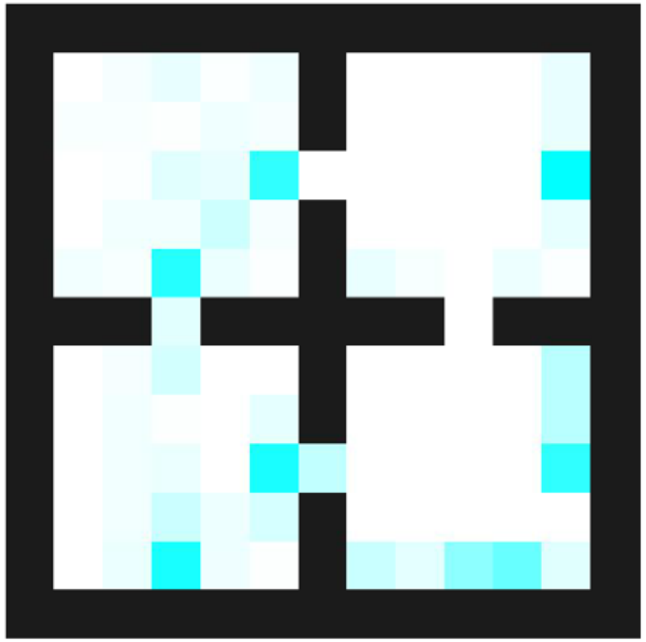


Figure 7: Termination probability visualization for the 4-room experiment. Each time the agent travels from the upper-left corner cell to the lower-right corner cell. The visualization shows the termination probabilities of the RNN generative models in the HRL training after the sequence segmentation process. Darker colors mean higher probabilities.

Summary

- An intelligent, human-like, open-domain conversational system
- Dialogue as RL
- A case study: Composite task completion bot with Hierarchical RL
- Ongoing research: subgoal discovery for hierarchical RL

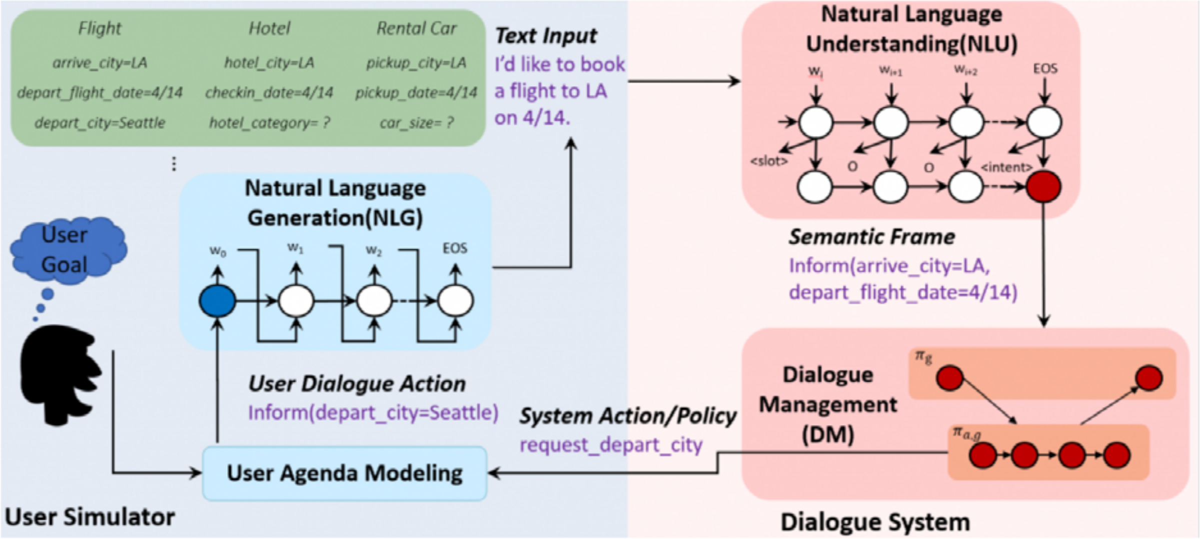
Contact Information:

- www.microsoft.com/en-us/research/people/jfgao/
- Learn more at deep RL for goal-oriented dialogues

Deep Reinforcement Learning for Goal-Oriented Dialogues

Established: April 18, 2016

This project aims to develop intelligent dialogue agents to help users effectively accomplish tasks via natural language conversation. A typical goal-oriented dialogue system contains three major components: natural language understanding (NLU), natural language generation (NLG), and dialogue management (DM) that consists of state tracking and policy learning. Our research focus is on deep reinforcement learning approaches for dialogue management in goal-oriented dialogue settings, including movie ticket booking, trip planning, sales assistant etc.



Composite Task Completion Dialogue System

People

Research Team

Asli Celikyilmaz
Researcher

Jianfeng Gao
Partner Research Manager

Sungjin Lee
Senior Researcher

Xiujun Li

Jason Williams
Principal Researcher

Past Interns & Visitors