SLT 2018 Special Session – Microsoft Dialogue Challenge: Building E2E Task-Completion Dialogue Systems.

Agenda

1:00 - 1:10PM	Opening, Jianfeng Gao (MSR)
1:10 – 1:40PM	“Past, Present, and Future of Conversational AI”, Gokhan Tur (Uber)
1:40 – 2:10PM	“Towards Building More Intelligent Conversational System: Semantics, Consistency & Interactiveness”, Minlie Huang (Tsinghua)
2:10 – 2:40PM	“Towards Open-Domain Conversational AI”, Vivian Chen (NTU)
2:40 – 3:00PM	Break
3:00 – 3:20PM	“MS dialogue challenge: result and outlook”, Sungjin Lee (MSR)
3:20 – 3:35PM	“Universe Model: A Human-like User Simulator Based on Dialogue Context”, Sihong Liu (Beijing University of Posts and Telecommunications)
3:35 – 3:50PM	“Double dueling Agent for Dialogue Policy Learning”, Yu-An Wang (NTU)
3:50 – 4:30PM	Panel discussion - Alex Acero (Apple), Vivian Chen (NTU), Minlie Huang (Tsinghua), Sungjin Lee (MSR), Spyros Matsoukas (Amazon), Gokhan Tur (Uber)

Topics we’d like to discuss…

- E2E system evaluation: simulated users, paid users and unpaid users.

- A unified modeling framework for dialogues: rule-based, SL, RL; fully data-driven vs. hybrid.

- Scalable training for task-oriented dialogues

- Dealing with heterogeneous data: chitchat, goal-oriented, non-conversational data

- Incorporating EQ (or empathy) into dialogue: recognize user emotion, generate empathetic responses

- Towards human-level intelligence: understand humans and their surrounding physical world

- Towards an ethical AI bot.

- …

Gao, Galley, Li. 2018. Neural Approaches to Conversational AI. arxiv.org/pdf/1809.08267.pdf

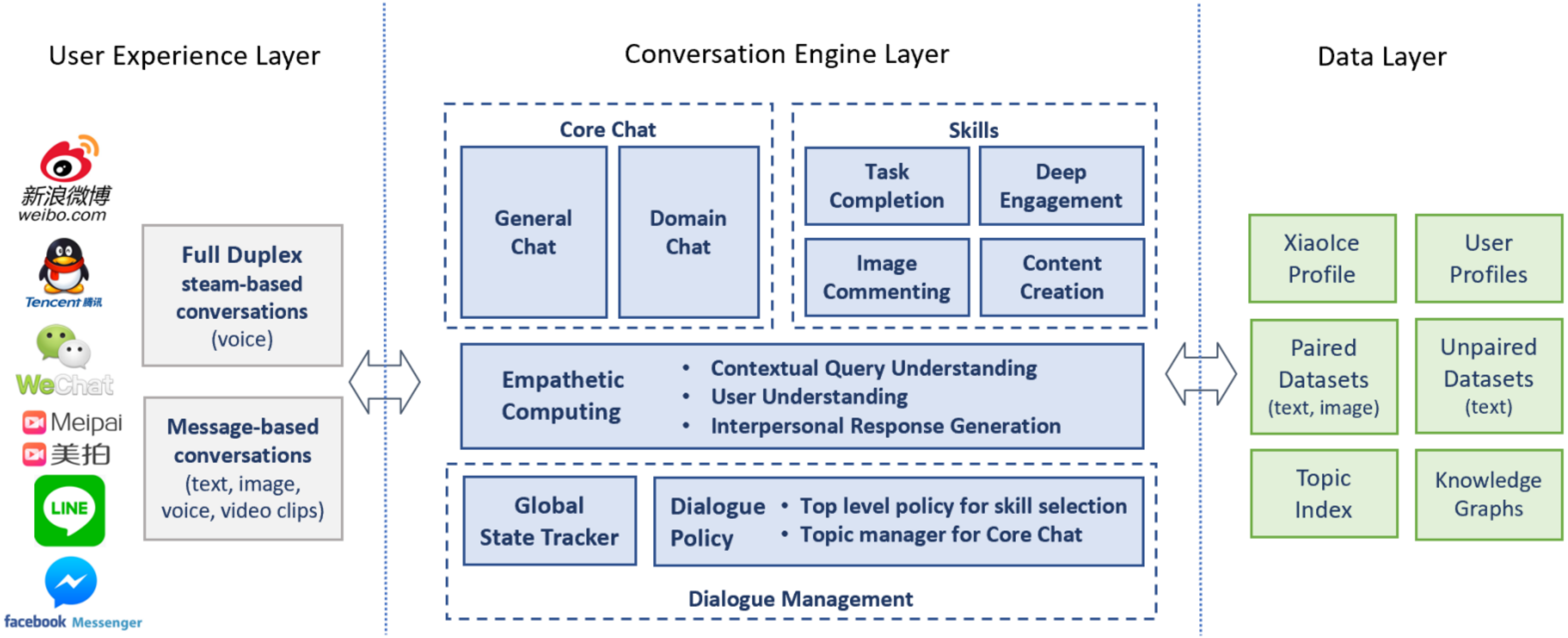


Figure 6.7: XiaoIce system architecture. Figure credit: Zhou et al. (2018)

Gao, Galley, Li. 2018. Neural Approaches to Conversational AI. arxiv.org/pdf/1809.08267.pdf

Microsoft Research

清華大學 Tsinghua University

MS Dialogue Challenge

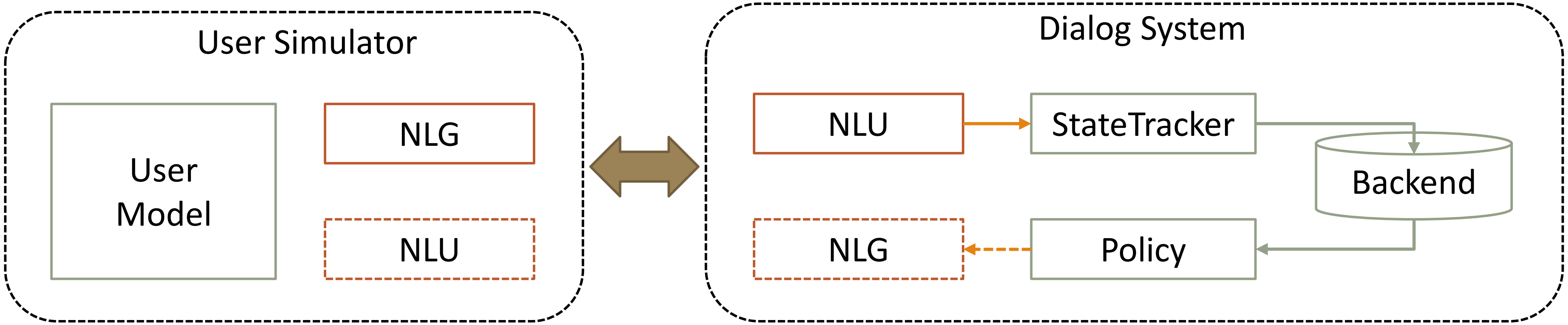
SUNGJIN LEE, JIANFENG GAO, XIUJUN LI, JJ LIU, SARAH PANDA

MS Dialogue Challenge

Challenges help the dialogue research community evaluate on common testbeds and advance the technology together.

Previous challenges were largely focused on particular components, trained and tested on static datasets

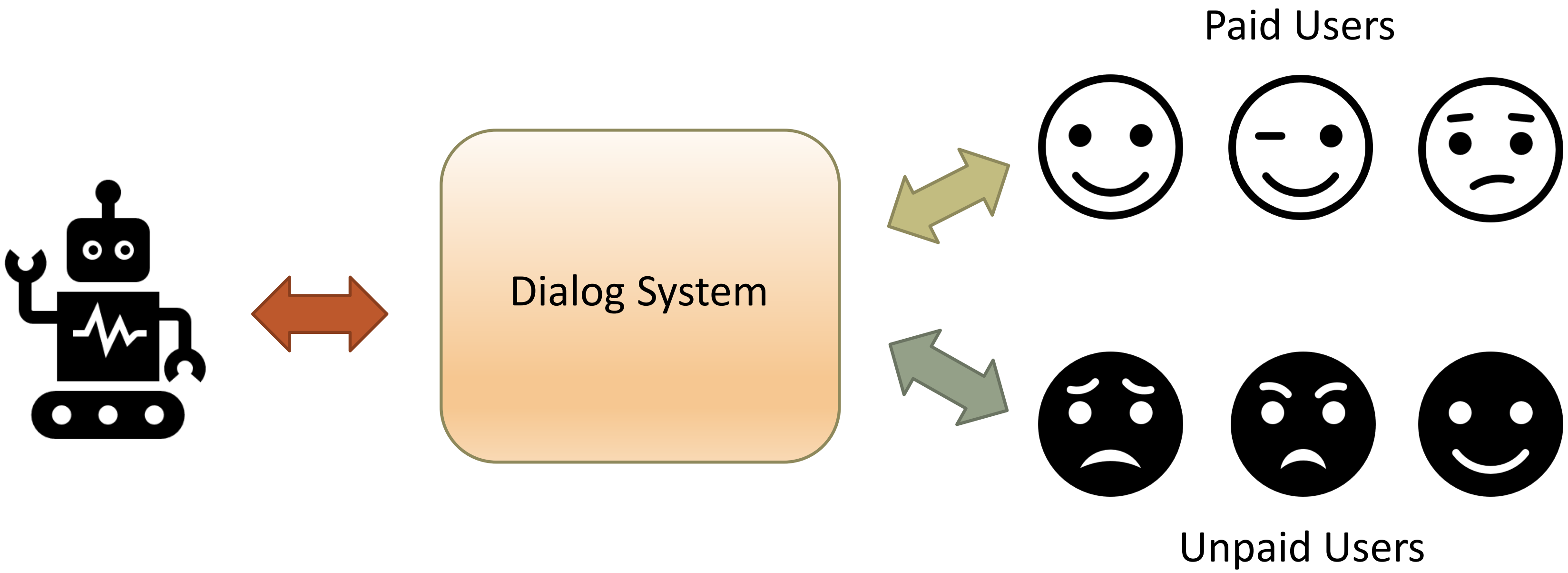
MS Dialogue Challenge is unique in aiming for **end-to-end** system evaluation.



Real User Evaluation

Real users are vastly different from simulated users.

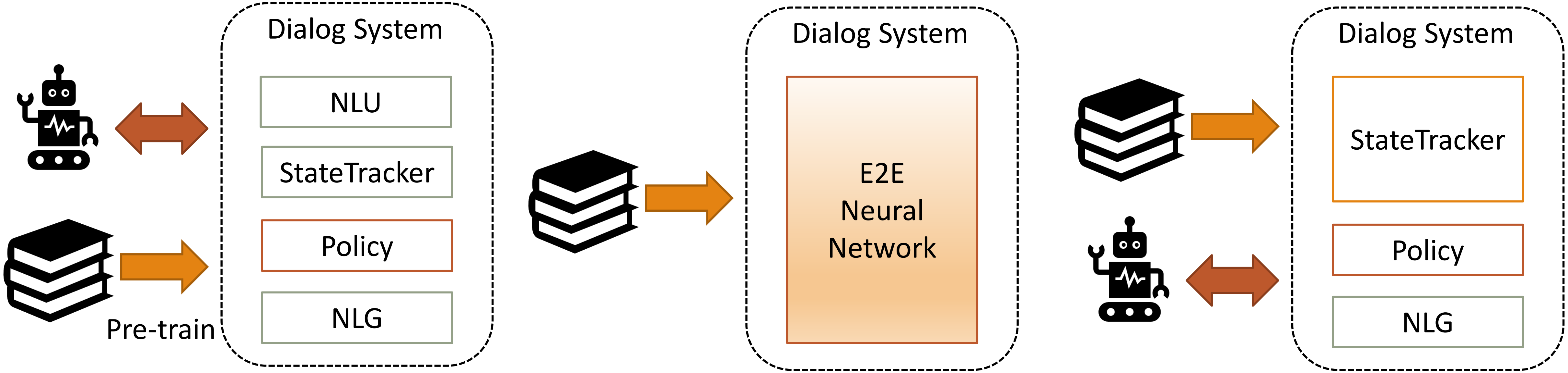
Unpaid users are again significantly different from paid users *(from The Spoken Dialog Challenge)*.



Learning Resources

MS Dialogue Challenge provides different types of learning resources to allow for training various approaches.

State-of-the-art components are available.



First Challenge Domains

Domains: Movie, Restaurant, Taxi

Human-human dialogs using Amazon Mechanical Turk.

Fully annotated with intents and slot-values

Task	#Ints	#Slots	#Dials	#Utters	#Turns
Movie	11	29	2890	21656	7.5
Rest.	11	30	4103	29720	7.3
Taxi	11	19	3094	23312	7.6

Downloadable at https://github.com/xiul-msr/e2e_dialog_challenge/tree/master/data

user	Please find me a Japanese **restaurant** that is open until at least midnight tonight in downtown seattle.
agent	I'm sorry - this is a movie booking service, I am not able to search for restaurants.
user	Oh, show me theaters in southeast portland oregon that are showing star wars the force awakens tonight
agent	You could still catch Star Wars tonight at Regal Lloyd Center, Century Eastport 16, Or Regal Movies on TV Stadium 16. Would you like to book tickets at any of those?
user	**That's odd. Someone another person got crossed with me.** But sure. Is there a very late showing I could still catch at any of those?
agent	How many tickets would you like?
user	**Well, I think it's a little late for tonight.** How about 4 tickets, to Regal Lloyd, anytime after 7pm tomorrow night.
agent	Great, I was able to purchase 4 tickets to Star Wars for 3/8 at 9:50 pm at REGAL LLOYD CENTER 10 & IMAX MOVIE TIMES + TICKETS 1510 N.E. Multnomah St., Portland, OR 97232.
user	Perfect. Thank you so much.

Evaluation

Simulation Evaluation

- Metrics: **Success Rate**, Average Turns, Average Reward
- Reward:
  - Success: $\underline{2 \times max\_turn - turn\_len}$
  - Failure: $\underline{-max\_turn - turn\_len}$
- Average over 5 Runs
- Each run simulated 2000 dialogues

Real User Evaluation

- Metrics: **Success Rate**, Average Turns, Average Reward, and User Rating (1-5).
- Hired human judges to evaluate the agents
- Overall, 2648 conversations are collected
- On average, 295 dialogues per human judge are collected.

First Challenge Result

Movie Domain Entry	Automatic Success Rate	Human Success Rate	Human Rating
NTU-Double-Q	41.8%	**31.1%**	**2.65**
DQN	**44.1%**	30.8%	2.62
NTU-HDQN	33.3%	27.3%	2.49*
BUPT-Transfer-DDQ	11.5%	9.66%	2.24*
Rule	6.13%	6.42%	1.78*

*Statistically significant with p<0.05

4 reinforcement learning-based agents and 1 rule-based agent.

Overall, RL-based agents outperform rule-based agent.

The first and second systems in automatic evaluation switch positions in human evaluation.

More details available from participants’ presentation.

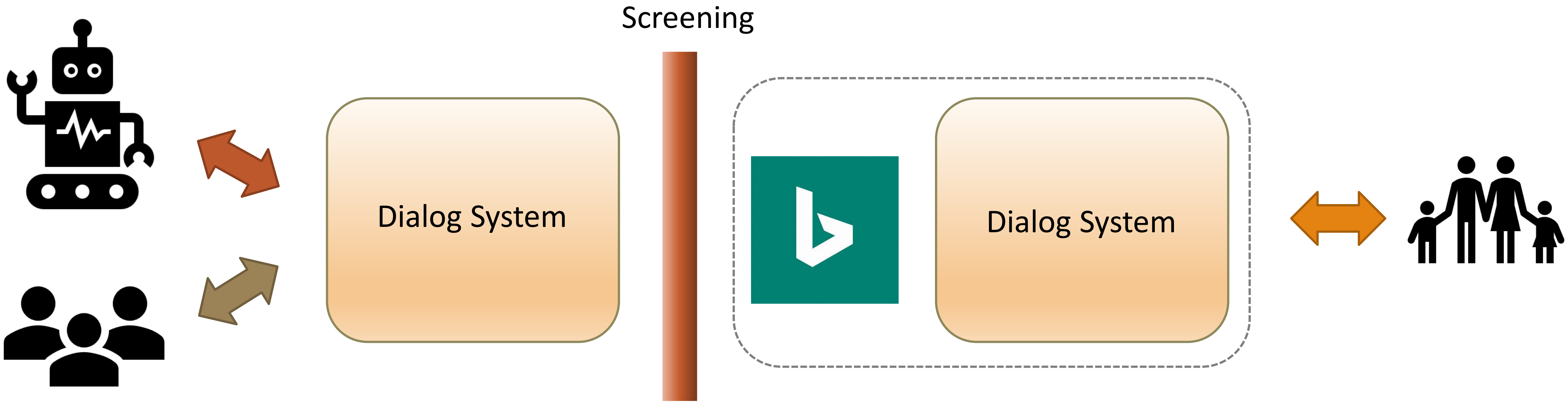
No entries for Restaurant and Taxi domain.

Next Challenge with Tsinghua University

Movie Bot Interacting with **Real** Users

Unpaid users are again significantly different from paid users.

Evaluate top-performing movie bots with unpaid real users

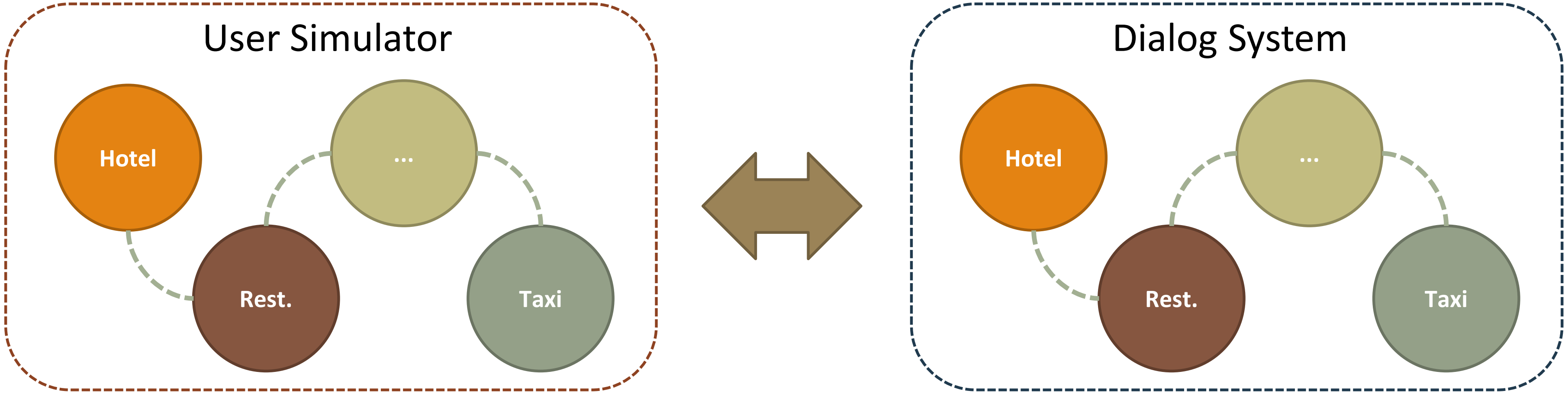


Multi-domain E2E Dialogue System

Multi-domain user simulator based on MultiWOZ extended with uncooperative behavior.

Baseline system based on hierarchical reinforcement learning.

Human evaluation with crowdsourced users.

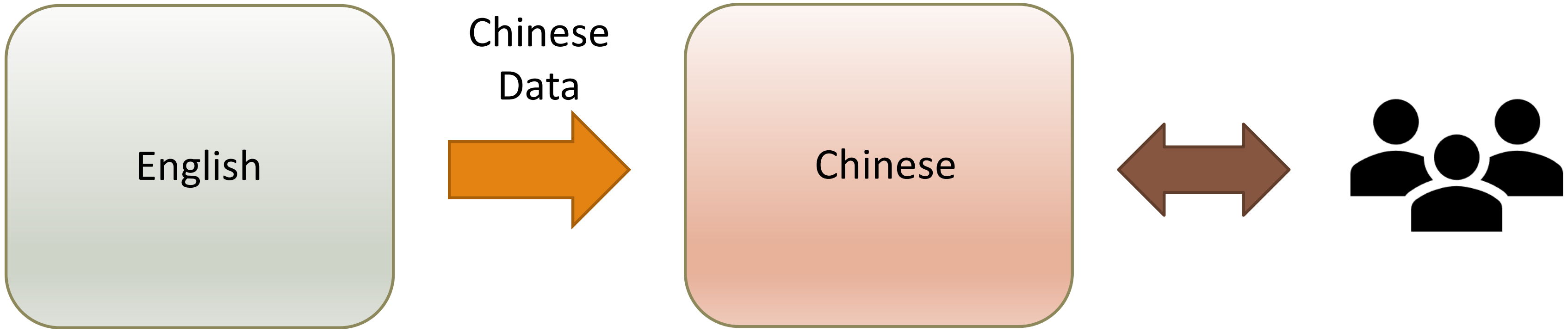


Cross-lingual Settings

Chinese as low resource language in the Movie domain.

Build a Chinese bot with a small amount of Chinese dialog data by leveraging all the resources available for English bot.

Human evaluation with crowdsourced users.



Post-Challenge

Data will remain publicly available.

Keep codebase updated with state-of-the-art models

Maintain a leaderboard and continuously evaluate top systems with real users.

Logistics

✓	Jul. 2019	Challenge overview and resource release (ACL/Sigdial)
✓	Nov. 2019	Evaluation, paper submission
✓	Dec 2019	Workshop (ASRU/NeurIPS)

Special thanks to…

SLT for hosting this special session

Our advisory board

Challenge participants

Evaluation

Domain	Agent	Simulation Evaluation			Human Evaluation			
		Success	Reward	Turn	Success	Reward	Turn	Rating
Movie	Double Q	41.8%	-2.73	19.59	**31.1%**	-18.15	31.31	**2.65**
	DQN	**44.1%**	0.93	18.85	**30.8%**	-18.08	26.09	**2.62**
	HDQN	33.3%	-15.60	32.39	27.3%	-23.47	32.39	2.49
	Transfer-DDQ	11.5%	-35.75	16.18	9.66%	-36.51	16.13	2.24
	Rule	6.13%	-42.38	20.0	6.42%	-41.29	18.0	1.78
Restaurant	DQN	**30.18%**	0.70	22.32	**22.9%**	-22.53	26.67	**2.35**
	Rule	7.22%	-24.50	18.00	6.85%	-32.02	16.02	1.94
Taxi	DQN	**43.5%**	0.26	22.60	**25.2%**	-19.88	25.09	**2.38**
	Rule	12.2%	-31.00	22.00	8.70%	-32.21	19.98	1.71