任务导向对话的数据和平台建设

朱祺

清华大学计算机系

Introduction

- Task-oriented dialogue system

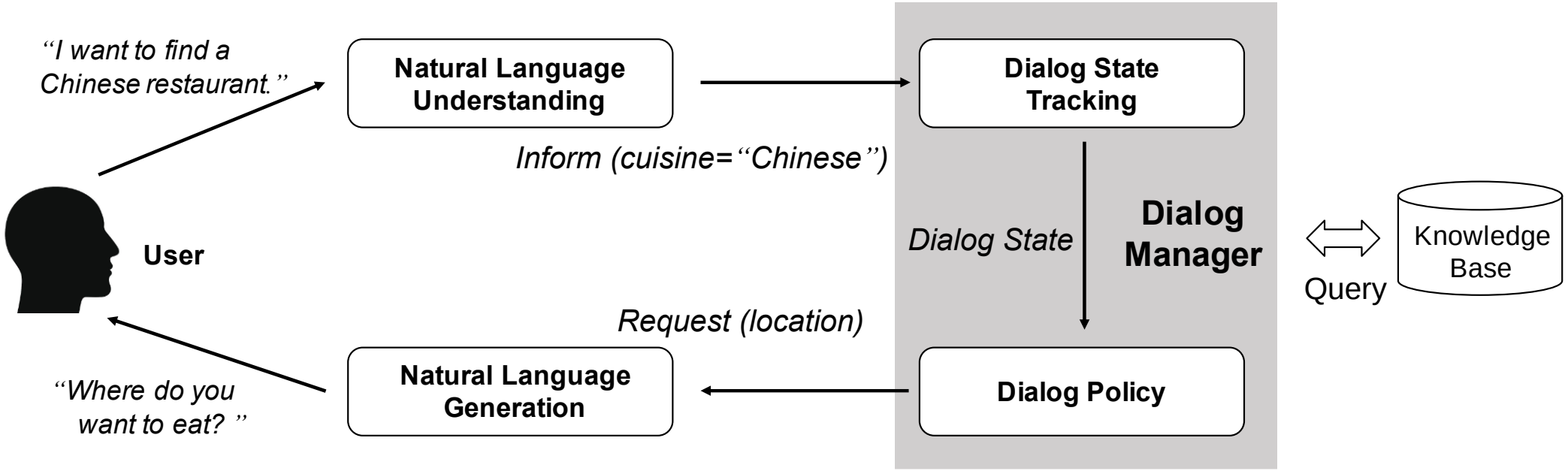


 Figure from "Recent Advances and Challenges in Task-oriented Dialog System" (Zheng et al., 2020)

任务导向对话的数据和平台建设

- Dataset and Toolkit advance the research.
  - CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset
    - First large-scale **Chinese** multi-domain task-oriented dataset.
    - Rich annotations supports a wide range of tasks.
    - Challenging inter-domain dependency.
  - ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems
    - Inherits ConvLab's framework but integrates SOTA models and more datasets.
    - Provides an analysis tool and an interactive tool to assist researchers in diagnosing dialogue systems.
    - End-to-end benchmark on MultiWOZ.

Resources

- CrossWOZ
  - Paper: https://arxiv.org/abs/2002.11893 (TACL 2020)
  - Code and data: https://github.com/thu-coai/CrossWOZ
  - Will be used in DSTC-9 (Track: Multi-domain Task-oriented Dialog Challenge II) for cross-lingual DST task.
- ConvLab-2
  - Paper: https://arxiv.org/abs/2002.04793 (ACL 2020 demo track)
  - Code: https://github.com/thu-coai/ConvLab-2
  - Tutorial on colab: link
  - Demo video of interactive tool: link
  - Will be used in DSTC-9 (Track: Multi-domain Task-oriented Dialog Challenge II) as the development platform.

CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, Minlie Huang

TACL 2020

Outline

- **Introduction**
- Data Collection
- Corpus Statistics
- Benchmark
- Conclusion

Introduction

- Background: multi-domain task-oriented dialogue.
  - System assists user to accomplish a task.
  - Involving multiple domains.

CrossWOZ
usr: 你好，可以帮我推荐一个评分是4.5分以上的景点吗? Hello, could you recommend an attraction with a rating of 4.5 or higher? sys: 天安门城楼，簋街小吃和北京欢乐谷都是很不错的地方呢。 Tiananmen, Gui Street, and Beijing Happy Valley are very nice places. usr: 我喜欢北京欢乐谷，你知道这个景点周边的酒店都是什么吗? I like Beijing Happy Valley. What hotels are around this attraction? sys: 那可多了，有A酒店, B酒店, C酒店。 There are many, such as hotel A, hotel B, and hotel C. usr: 太好了，我正打算在**景点附近**找个酒店住宿呢，知道哪家评分是4分以上，提供叫醒服务的不? Great! I am planning to find a hotel to stay **near the attraction**. Which one has a rating of 4 or higher and offers wake-up call service?

Introduction

- Motivation
  - Advance multi-domain task-oriented dialogue modeling.
  - Alleviate the shortage of Chinese task-oriented datasets.
- CrossWOZ:
  - **Large-scale**: 6k sessions, 102k utterances, 5 domains.
  - **Chinese**: first large-scale Chinese task-oriented dataset.
  - **Cross-domain**: dependency between domains is challenging.
  - **Rich annotation**: dialog act, system state, and user state.

Introduction

- Data statistics
  - Large scale Chinese human-to-human dialogue.
  - Involving more domains and slots in dialogue.

Type	Single-domain goal					Multi-domain goal		
Dataset	DSTC2	WOZ 2.0	Frames	KVRET	M2M	MultiWOZ	Schema	**CrossWOZ**
Language	EN	EN	EN	EN	EN	EN	EN	**CN**
Speakers	H2M	H2H	H2H	H2H	M2M	H2H	M2M	**H2H**
# Domains	1	1	1	3	2	7	16	**5**
# Dialogues	1,612	600	1,369	2,425	1,500	8,438	16,142	**5,012**
# Turns	23,354	4,472	19,986	12,732	14,796	115,424	329,964	**84,692**
Avg. domains	1	1	1	1	1	1.80	1.84	**3.24**
Avg. turns	14.5	7.5	14.6	5.3	9.9	13.7	20.4	**16.9**
# Slots	8	4	61	13	14	25	214	**72**
# Values	212	99	3,871	1363	138	4,510	14,139	**7,871**

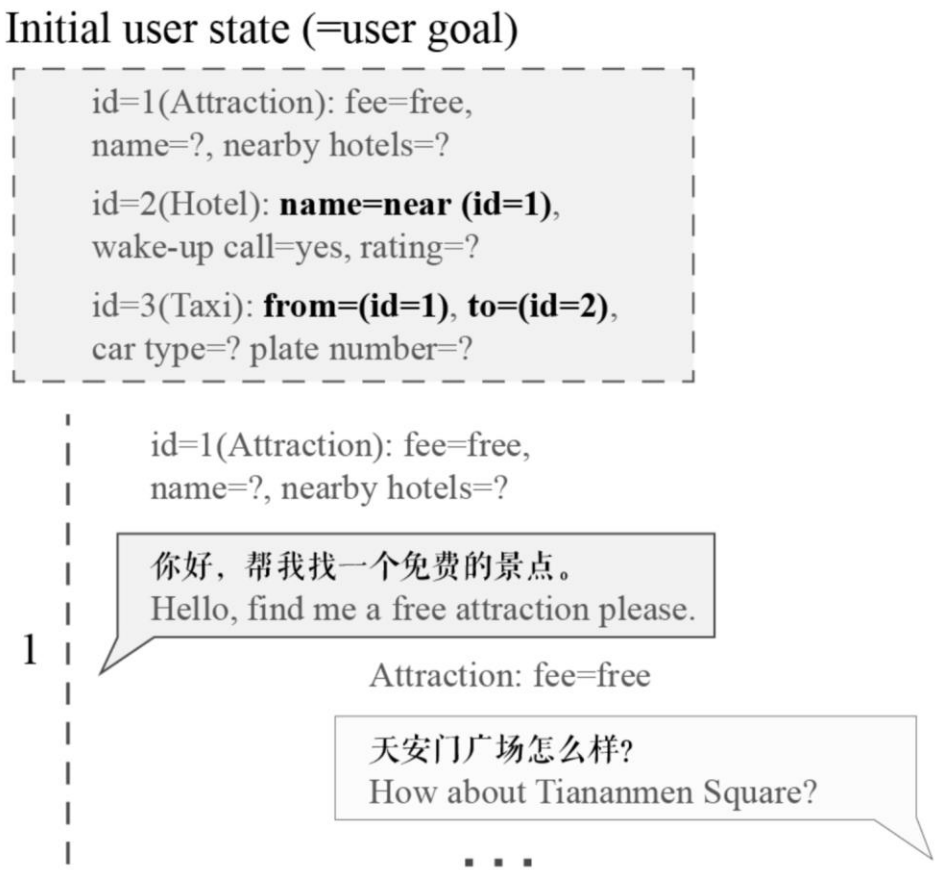
Introduction

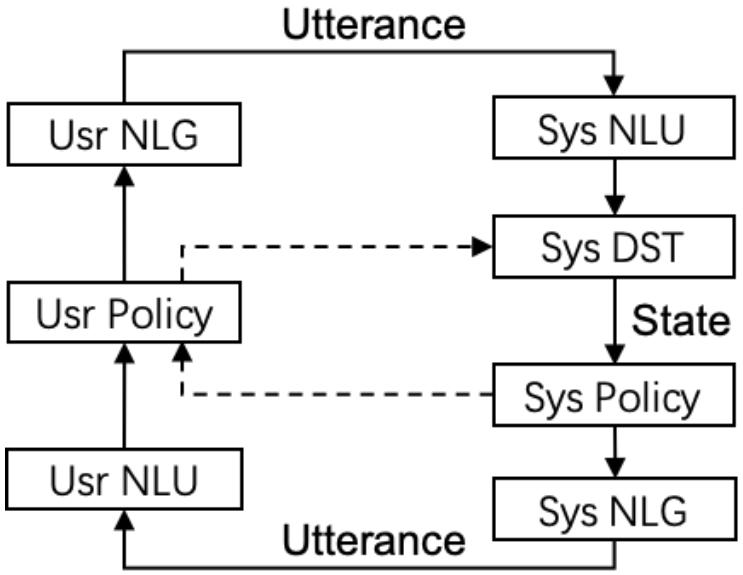
- Cross-domain dependency
  - Cross-domain constraints are **pre-specified** in MultiWOZ and Schema, while determined **dynamically** in CrossWOZ.

MultiWOZ	CrossWOZ
usr: I'm looking for a college type attraction. ... usr: I would like to visit in town centre please. ... usr: Can you find an Indian restaurant for me that is **also in the town centre**?	usr: 你好，可以帮我推荐一个评分是4.5分以上的景点吗? Hello, could you recommend an attraction with a rating of 4.5 or higher? sys: 天安门城楼，簋街小吃和北京欢乐谷都是很不错的地方呢。 Tiananmen, Gui Street, and Beijing Happy Valley are very nice places. usr: 我喜欢北京欢乐谷，你知道这个景点周边的酒店都是什么吗? I like Beijing Happy Valley. What hotels are around this attraction? sys: 那可多了，有A酒店, B酒店, C酒店。 There are many, such as hotel A, hotel B, and hotel C. usr: 太好了，我正打算在**景点附近**找个酒店住宿呢，知道哪家评分是4分以上，提供叫醒服务的不? Great! I am planning to find a hotel to stay **near the attraction**. Which one has a rating of 4 or higher and offers wake-up call service?
Schema	
usr: I want a hotel in San Diego and I want to check out on Thursday next week. ... usr: I need a one way flight to **go there**.	

Introduction

- Rich annotation supports a variety of tasks





Outline

- Introduction
- **Data Collection**
- Corpus Statistics
- Benchmark
- Conclusion

Data Collection

- Database Construction
- Goal Generation
- Dialogue Collection
- Dialogue Annotation

Data Collection

- 1 Database Construction
  - Travel information: Hotel, Attraction, and Restaurant. (HAR domains)
  - Metro: derive from the travel information.
  - Taxi: no database.

Domain	Attract.	Rest.	Hotel
# Entities	465	951	1133
# Slots	9	10	8+37*
Avg. nearby attract.	4.7	3.3	0.8
Avg. nearby rest.	6.7	4.1	2.0
Avg. nearby hotels	2.1	2.4	-

*: hotel services

Data Collection

- 2 Goal Generation
  - Sample domains & slots.
  - Build cross-domain constraints.
  - Sample values from database as normal constraints.
  - Blank values are requirements.
  - Generate an equivalent natural language description.

Id	Domain	Slot	Value
1	Attraction	fee	*free*
1	Attraction	name	_____
1	Attraction	nearby hotels	_____
2	Hotel	name	**near (id=1)**
2	Hotel	wake-up call	*yes*
2	Hotel	rating	_____
3	Taxi	from	**(id=1)**
3	Taxi	to	**(id=2)**
3	Taxi	car type	_____
3	Taxi	plate number	_____

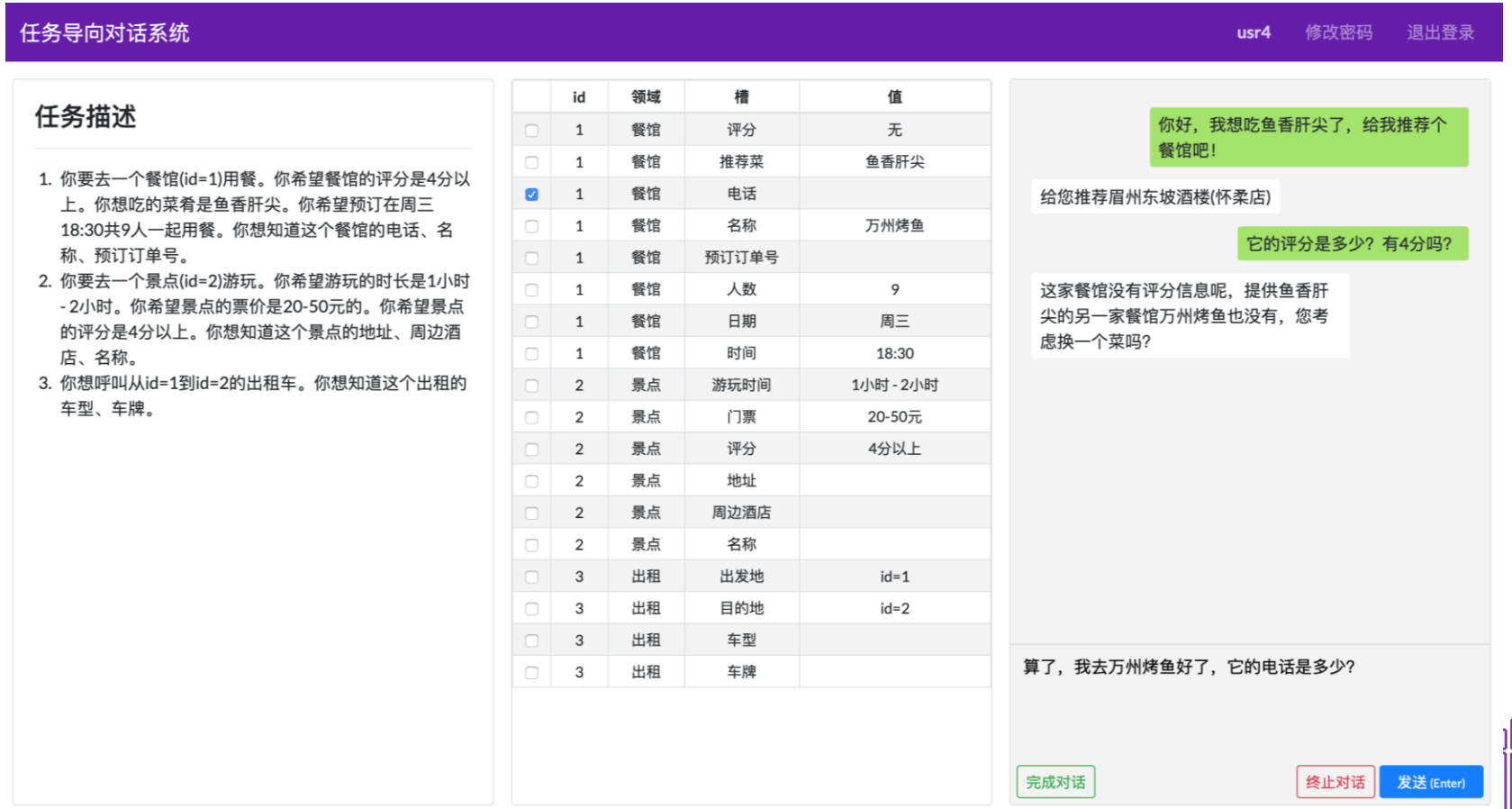
Table 4: A user goal example (translated into English). Slots with bold/italic/blank value are cross-domain informable slots, common informable slots, and requestable slots. In this example, the user wants to find an attraction and one of its nearby hotels, then book a taxi to commute between these two places.

Data Collection

- 3 Dialogue Collection
  - Paired workers converse synchronously online.
  - User side:
    - Update user state according to the system response.
    - Select some semantic tuples in the user state, which indicates the dialogue acts.
    - Compose the utterance according to the selected semantic tuples.
    - When “NoOffer”, users are encouraged to relax the constraints manually.

Data Collection

- 3 Dialogue Collection
  - User side

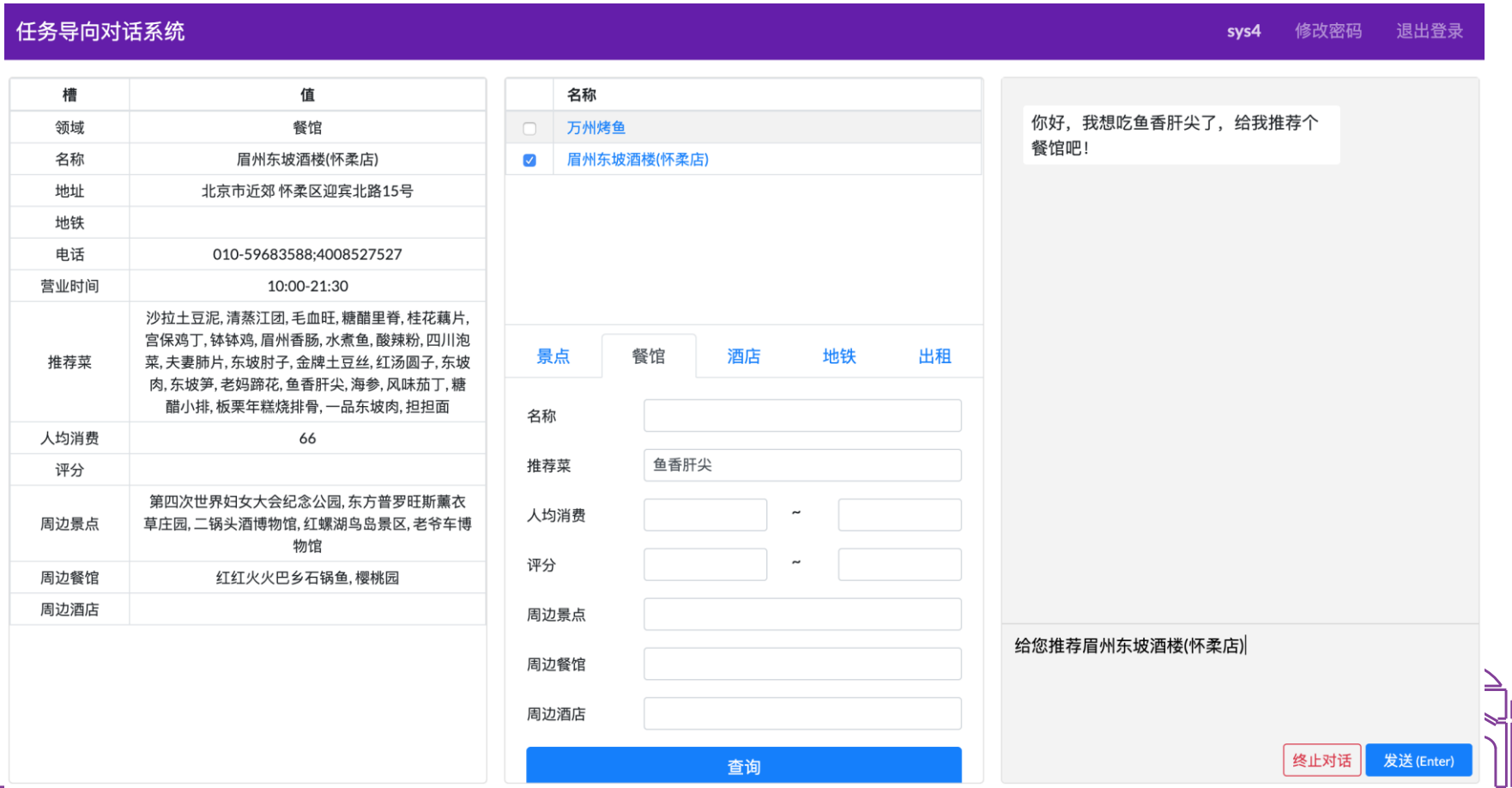


Data Collection

- 3 Dialogue Collection
  - System side:
    - Fill the query according to the previous user response.
    - Search the database and select the retrieved entities.
    - Respond in natural language based on the information of the selected entities.
    - When “NoOffer”, systems will try to recommend an alternative.

Data Collection

- 3 Dialogue Collection
  - System side



Data Collection

- 4 Dialogue Annotation
  - User & system state already available.
  - Derive dialogue act from state update and dialogue history.
  - 3 experts annotate 50 dialogues for verification.

Outline

- Introduction
- Data Collection
- **Corpus Statistics**
- Benchmark
- Conclusion

Corpus Statistics

- Overall statistics
  - The average number of sub-goals is 3.24, which is much larger than that in MultiWOZ (1.80) and Schema (1.84).

	Train	Valid	Test
# Dialogues	5,012	500	500
# Turns	84,692	8,458	8,476
# Tokens	1,376,033	137,736	137,427
Vocab	12,502	5,202	5,143
Avg. sub-goals	3.24	3.26	3.26
Avg. STs	14.8	14.9	15.0
Avg. turns	16.9	16.9	17.0
Avg. tokens	16.3	16.3	16.2

Corpus Statistics

- 5 goal types:
  - **S**: only one sub-goal in HAR domains.
  - **M**: multiple sub-goals in HAR domains.
  - **M+T**: multiple sub-goals in HAR domains and at least one sub-goal in the metro or taxi domain.
  - **CM**: multiple sub-goals in HAR domains with cross-domain constraints.
  - **CM+T**: multiple sub-goals in HAR domains with cross-domain constraints and at least one sub-goal in the metro or taxi domain.

HAR: Hotel, Attraction, and Restaurant

Corpus Statistics

- Data statistics
  - CM and CM+T are more challenging because additional cross-domain constraints.

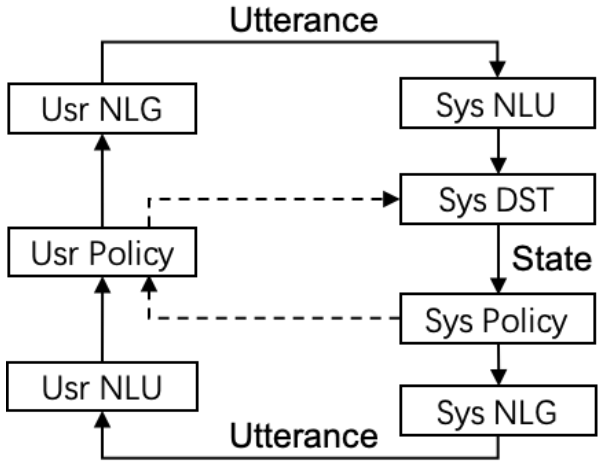
Goal type	S	M	M+T	CM	CM+T
# Dialogues	417	1573	691	1759	572
NoOffer rate	0.10	0.22	0.22	0.61	0.55
Multi-query rate	0.06	0.07	0.07	0.14	0.12
Goal change rate	0.10	0.28	0.31	0.69	0.63
Avg. dialogue acts	1.85	1.90	2.09	2.06	2.11
Avg. sub-goals	1.00	2.49	3.62	3.87	4.57
Avg. STs	4.5	11.3	15.8	18.2	20.7
Avg. turns	6.8	13.7	16.0	21.0	21.6
Avg. tokens	13.2	15.2	16.3	16.9	17.0

Outline

- Introduction
- Data Collection
- Corpus Statistics
- **Benchmark**
- Conclusion

Benchmark

- Cross-domain constraints are challenging for all these tasks.



		S	M	M+T	CM	CM+T	Overall
BERTNLU	Dialogue act F1	96.69	96.01	96.15	94.99	95.38	95.53
– context		94.55	93.05	93.70	90.66	90.82	91.85
RuleDST	Joint state accuracy (single turn)	84.17	78.17	81.93	63.38	67.86	71.33
TRADE	Joint state accuracy	71.67	45.29	37.98	30.77	25.65	36.08
SL policy	Dialogue act F1	50.28	44.97	54.01	41.65	44.02	44.92
	Dialogue act F1 (delex)	67.96	67.35	73.94	62.27	66.29	66.02
Simulator	Joint state accuracy (single turn)	63.53	48.79	50.26	40.66	41.76	45.00
	Dialogue act F1 (single turn)	85.99	81.39	80.82	75.27	77.23	78.39
DA Sim	Task finish rate	76.5	49.4	33.7	17.2	15.7	34.6
NL Sim (Template)		67.4	33.3	29.1	10.0	10.0	23.6
NL Sim (SC-LSTM)		60.6	27.1	23.1	8.8	9.0	19.7

Benchmark

- The transition between related domains is especially challenging to model.
  - NLU: dialogue act F1.

	General	Inform	Request	Recom	NoOffer	Select
BERTNLU	99.45	94.67	96.57	98.41	93.87	82.25
– context	99.69	90.80	91.98	96.92	93.05	68.40

  - DST:
    - TRADE joint acc.: 36%->12%.
  - Policy:
    - SL policy dialogue act F1 (delex): 66%->54%.

Outline

- Introduction
- Data Collection
- Corpus Statistics
- Benchmark
- **Conclusion**

Conclusion

- The first large-scale Chinese Cross-Domain task-oriented dataset.
- The rich annotation supports a wide range of tasks.
- The dependency between domains is more challenging and requires more accurate context understanding.

ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems

Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu,

Jinchao Li*, Baolin Peng*, Jianfeng Gao*, Xiaoyan Zhu, Minlie Huang

ACL 2020 demo track

*: Collaboration with Microsoft Research

Outline

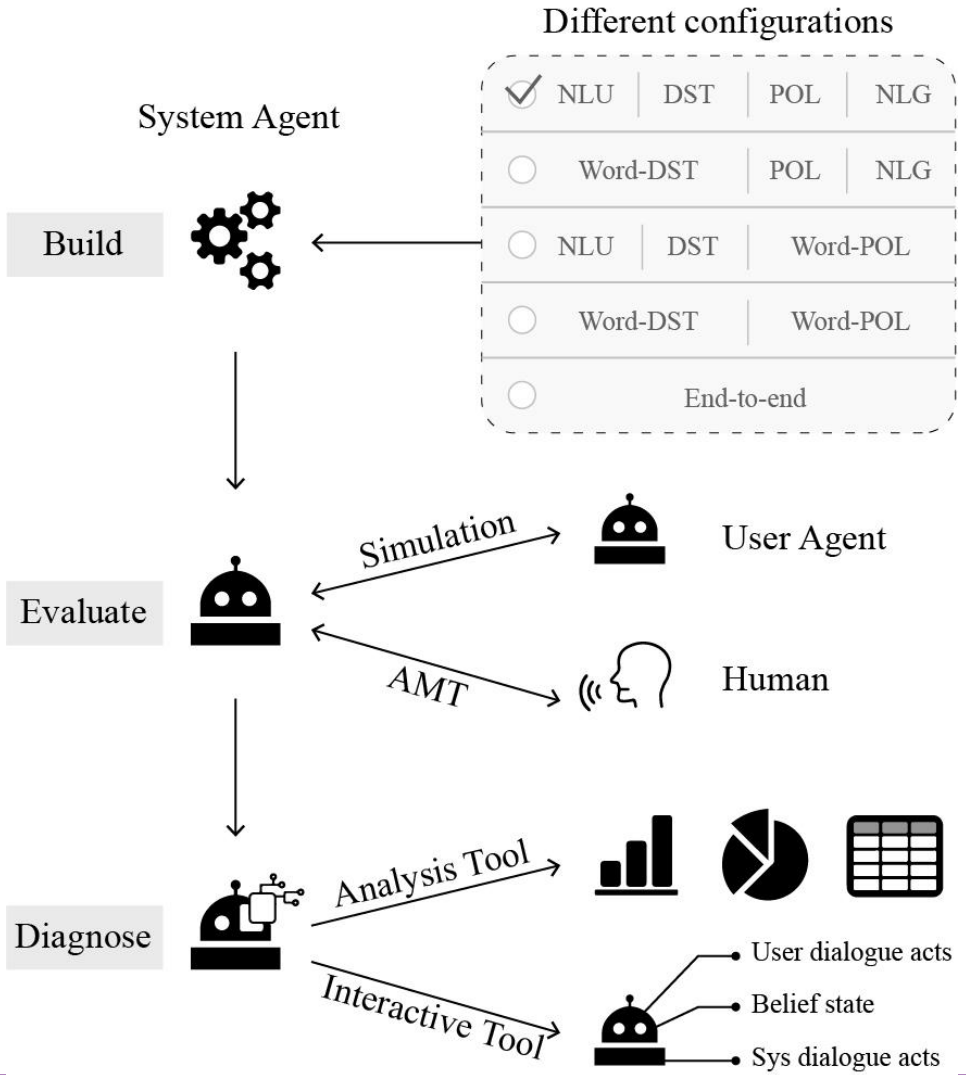
- **Introduction**
- Framework
- Supported Models & Datasets
- Analysis Tool
- Interactive Tool
- Demo

Introduction

- Existing Platforms
  - PyDial (Research):
    - Focuses on reinforcement learning dialogue policy.
  - ParlAI (Research):
    - Supports a variety of tasks and thus need to customize for modular dialogue system.
  - **ConvLab** (Research):
    - Multi-domain dialogue platform which supports end-to-end evaluation.
  - Rasa (Production):
    - Designed for non-specialist developer.
  - Plato (Production):
    - Modular and flexible framework. Supports multi-agent interaction.

Introduction

- ConvLab-2
  - Inherits ConvLab's framework but integrates SOTA dialogue models and supports more datasets.
  - Provides an analysis tool and an interactive tool to assist researchers in diagnosing dialogue systems.
  - End-to-end benchmark on MultiWOZ.

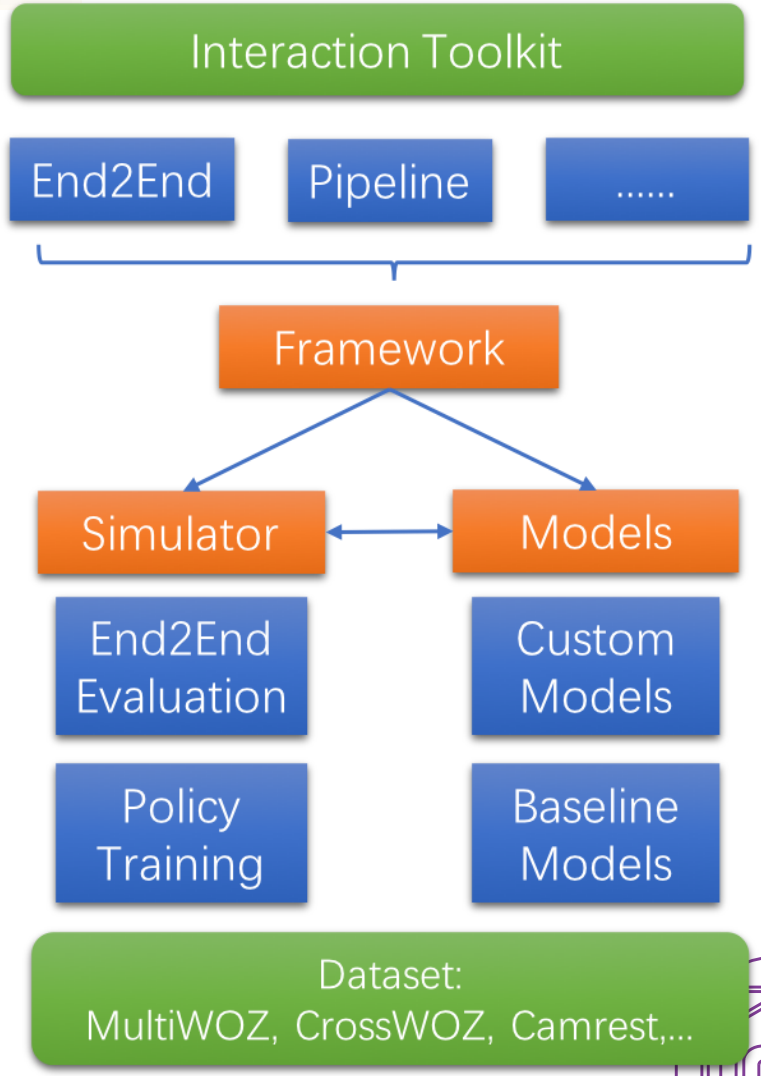


Outline

- Introduction
- **Framework**
- Supported Models & Datasets
- Analysis Tool
- Interactive Tool
- Demo

Framework

- Dialogue Agent
  - Pipeline/end-to-end/custom
  - Takes utterance as input and replies.
- Module
  - NLU/DST/POL/NLG/custom
  - Used to build dialogue agent.
- Evaluator/Analysis Tool
  - Evaluate the dialogue between two agents.
- Interactive Tool
  - Deploys the agent to a server and provides a graphical interface for interaction.



Outline

- Introduction
- Framework
- **Supported Models & Datasets**
- Analysis Tool
- Interactive Tool
- Demo

Supported Models & Datasets

- Models
  - NLU: SVM, MILU, **BERTNLU**
  - DST: rule, MDBT, **TRADE, SUMBT**
  - Policy: rule, Imitation, REINFORCE, PPO, **GDPL**, MDRG, **HDSA, LaRL**
  - Simulator policy: Agenda, VHUS
  - NLG: Template, SCLSTM
  - End2End: Sequicity, **DAMD**, **ROLL-OUTS RL**
- Datasets
  - CamRest676, MultiWOZ 2.1, CrossWOZ, DealOrNoDeal

Compared to ConvLab, newly integrated models in ConvLab-2 are marked in bold.

Supported Models & Datasets

- End-to-end Performance on MultiWOZ (partial results)

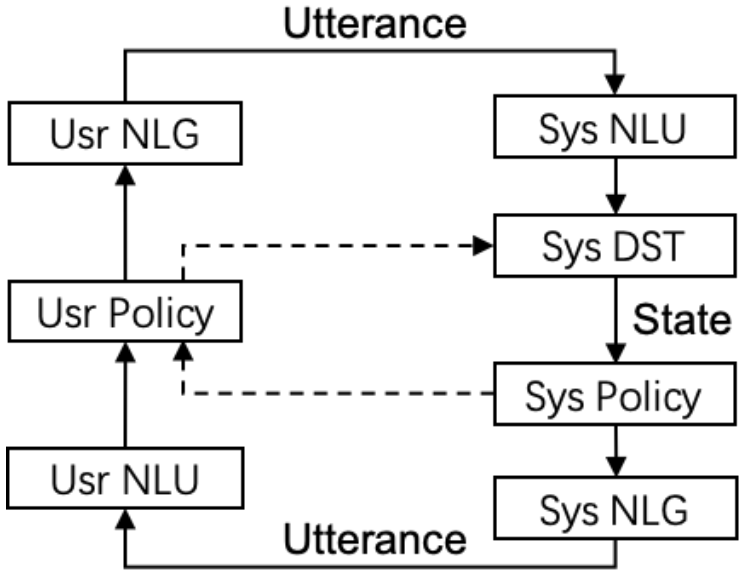
NLU	DST	Policy	NLG	Complete rate	Success rate	Book rate	Inform P/R/F1
BERTNLU	RuleDST	RulePolicy	TemplateNLG	92.1	85.5	91.5	79.8/92.8/83.8
BERTNLU	RuleDST	RulePolicy	**SCLSTM**	40.1	41.0	51.5	68.5/56.5/59.1
BERTNLU	RuleDST	**PPOPolicy**	TemplateNLG	69.7	56.6	56.6	64.8/79.0/68.1
None	**SUMBT**	RulePolicy	TemplateNLG	34.7	33.8	57.8	52.3/50.6/47.3
BERTNLU	RuleDST	**LaRL**	None	40.6	34.0	45.6	47.8/54.1/47.6
None	**SUMBT**	**LaRL**	None	39.4	33.1	39.5	48.5/56.0/48.8
None	None	**DAMD***	None	38.5	33.6	50.9	62.1/60.7/57.4

Outline

- Introduction
- Framework
- Supported Models & Datasets
- **Analysis Tool**
- Interactive Tool
- Demo

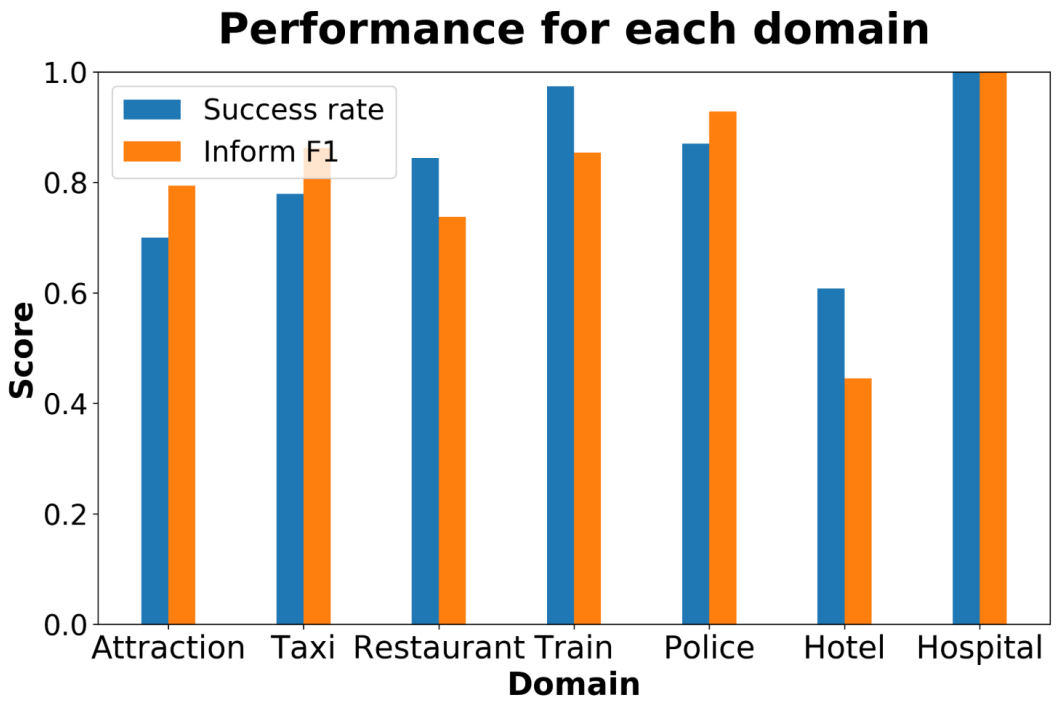
Analysis Tool

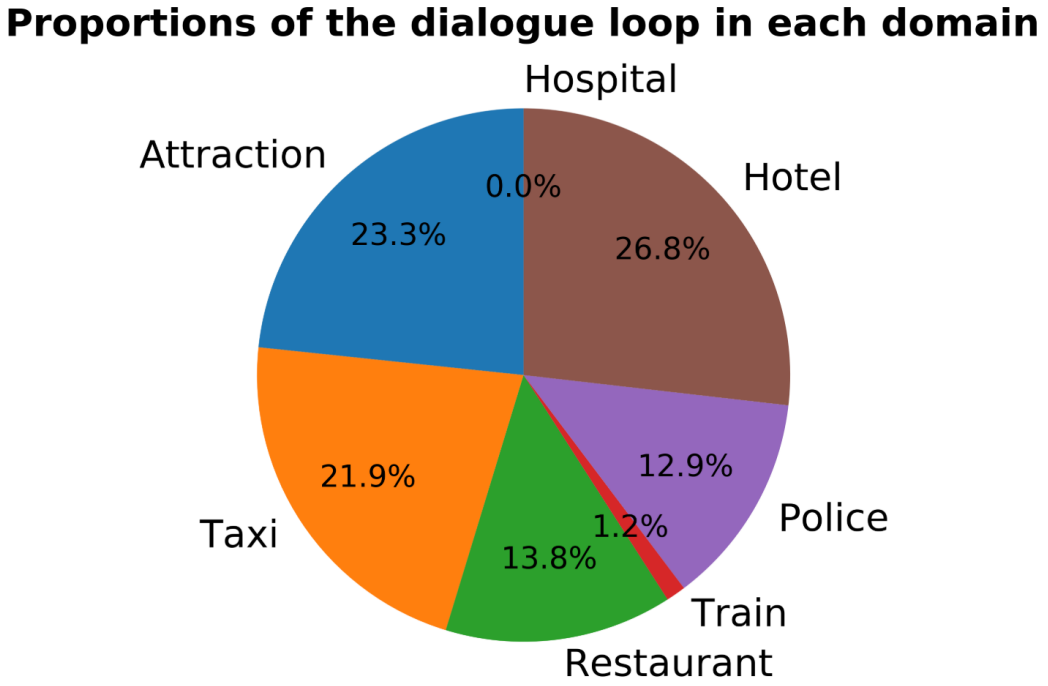
- Need more information to evaluate a dialogue system comprehensively
  - Metrics for overall performance
  - Common errors of the NLU component
  - Frequent invalid, redundant, and missing system dialogue acts predicted by the dialogue policy.
  - The system dialogue acts from which the NLG component generates responses that confuse the user simulator.
  - The causes of dialogue loops. Dialogue loop is the situation that the user keeps repeating the same request but gets no answer.
- Generate an HTML report with charts and tables.



Analysis Tool

- Partial results





Analysis Tool

- Partial results

Overall results:
Success Rate: 60.8%; inform F1: 44.5%

Most confusing user dialogue acts:
Request-Hotel-Post-?
- 34%: Request-Hospital-Post-?
- 32%: Request-Attraction-Post-?

Request-Hotel-Addr-?
- 29%: Request-Attraction-Addr-?
- 28%: Request-Restaurant-Addr-?

Request-Hotel-Phone-?
- 26%: Request-Restaurant-Phone-?
- 26%: Request-Attraction-Phone-?

Invalid system dialogue acts:
- 31%: Inform-Hotel-Parking
- 28%: Inform-Hotel-Internet

Redundant system dialogue acts:
- 34%: Inform-Hotel-Stars

Missing system dialogue acts:
- 25%: Inform-Hotel-Phone

Most confusing system dialogue acts:
Recommend-Hotel-Parking-yes
- 21%: Recommend-Hotel-Parking-none
- 18%: Inform-Hotel-Parking-none

Inform-Hotel-Parking-yes
- 17%: Inform-Hotel-Parking-none

Inform-Hotel-Stars-4
- 16%: Inform-Hotel-Internet-none

User dialogue acts that cause loop:
- 53% Request-Hotel-Phone-?
- 21% Request-Hotel-Post-?
- 14% Request-Hotel-Addr-?

Outline

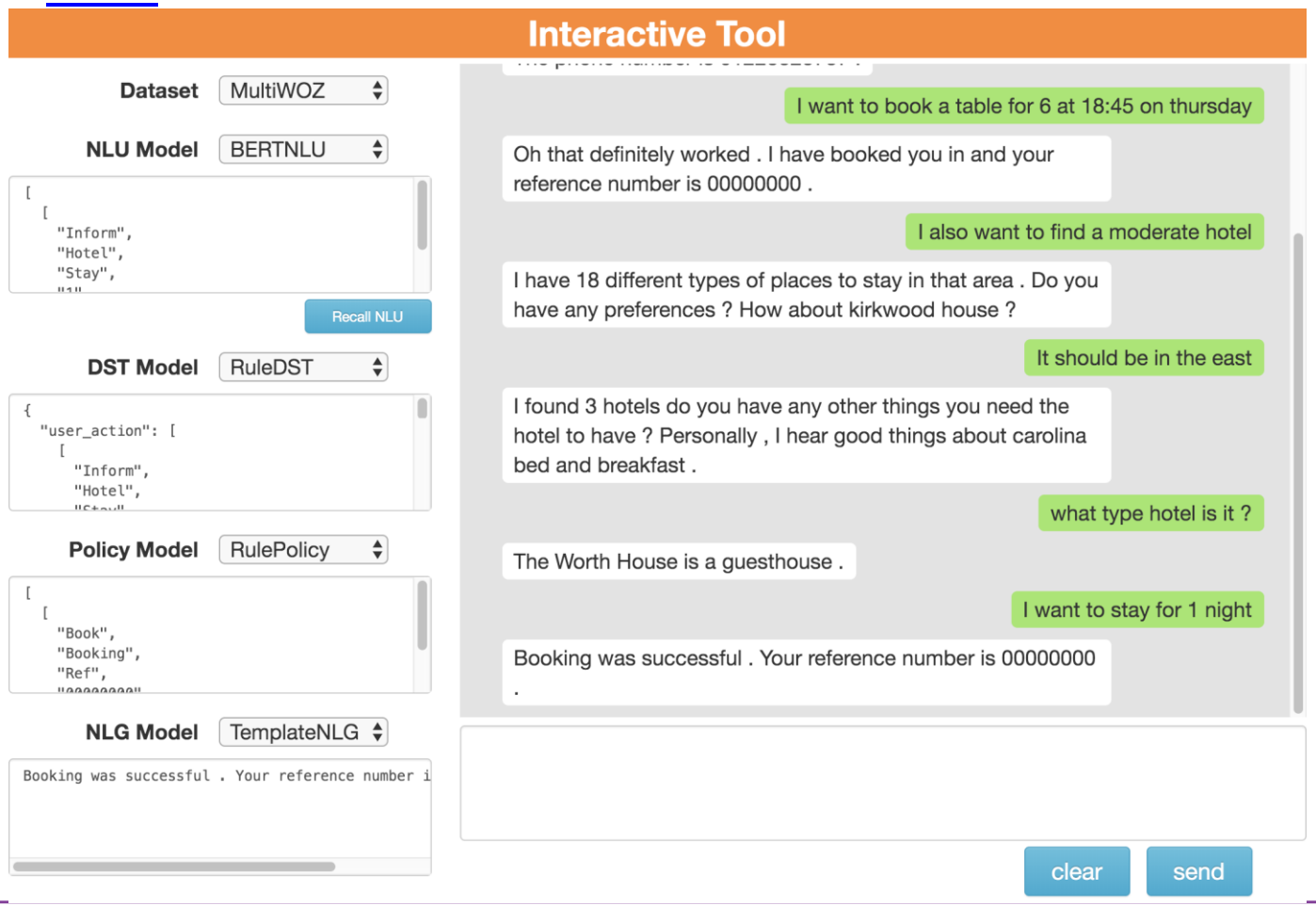
- Introduction
- Framework
- Supported Models & Datasets
- Analysis Tool
- **Interactive Tool**
- Demo

Interactive Tool

- Converse with a dialogue system through a graphical user interface.
  - Deploy the customized dialogue system to a server.
  - Accessed the dialogue system via a web browser.
  - The intermediate output of each module is displayed and can be modified manually.
- Can be used for:
  - Debugging each module of a dialogue system.
  - Collecting human-machine dialogue.
  - Iteratively train a model with user feedback.

Interactive Tool

- Demo video: link



Outline

- Introduction
- Framework
- Supported Models & Datasets
- Analysis Tool
- Interactive Tool
- **Demo**

Demo

- Use ConvLab-2 to build, evaluate, and diagnose a traditional pipeline dialogue system developed on the MultiWOZ dataset.
- More detailed tutorial on [colab]

```python
import ... # import necessary modules
# Create models for each component
# Parameters are omitted for simplicity
sys_nlu = BERTNLU(...)
sys_dst = RuleDST(...)
sys_policy = RulePolicy(...)
sys_nlg = TemplateNLG(...)
# Assemble a pipeline system named "sys"
sys_agent = PipelineAgent(sys_nlu, sys_dst,
    sys_policy, sys_nlg, name="sys")
# Build a user simulator similarly but without DST
user_nlu = BERTNLU(...)
user_policy = RulePolicy(...)
user_nlg = TemplateNLG(...)
user_agent = PipelineAgent(user_nlu, None,
    user_policy, user_nlg, name="user")
# Create an evaluator and a conversation environment
evaluator = MultiWozEvaluator()
sess = BiSession(sys_agent, user_agent, evaluator)
# Start simulation
sess.init_session()
sys_utt = ""
while True:
    sys_utt, user_utt, sess_over, reward = sess.
    next_turn(sys_utt)
    if sess_over:
        break
print(sess.evaluator.task_success())
print(sess.evaluator.inform_F1())
# Use the analysis tool to generate a test report
analyzer = Analyzer(user_agent, dataset="MultiWOZ")
analyzer.comprehensive_analyze(sys_agent,
    total_dialog=1000)
# Compare multiple systems
sys_agent2 = PipelineAgent(MILU(...), sys_dst,
    sys_policy, sys_nlg, name="sys")
analyzer.compare_models(agent_list=[sys_agent,
    sys_agent2], model_name=["bertnlu", "milu"],
    total_dialog=1000)
```

Build system agent

Build user agent

One simulation

Use analysis tool

Tsinghua University

Thanks!

Any question?