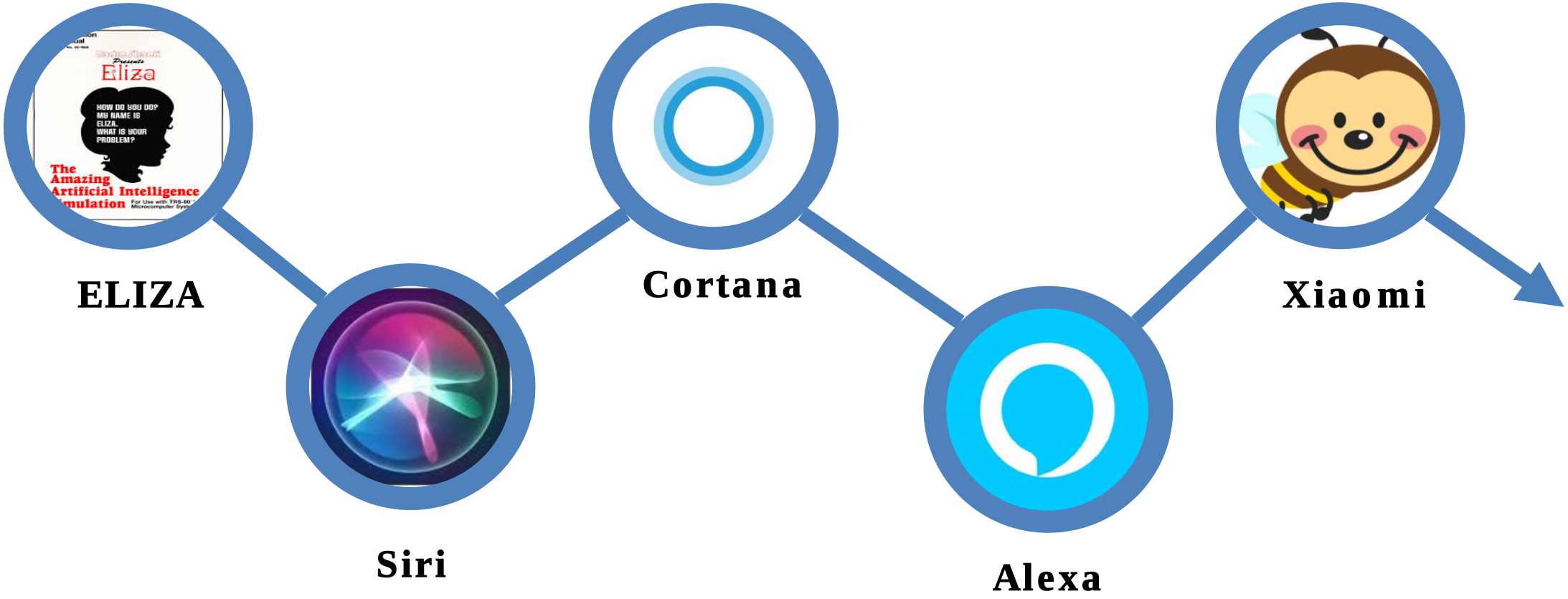
任务导向型对话系统评测

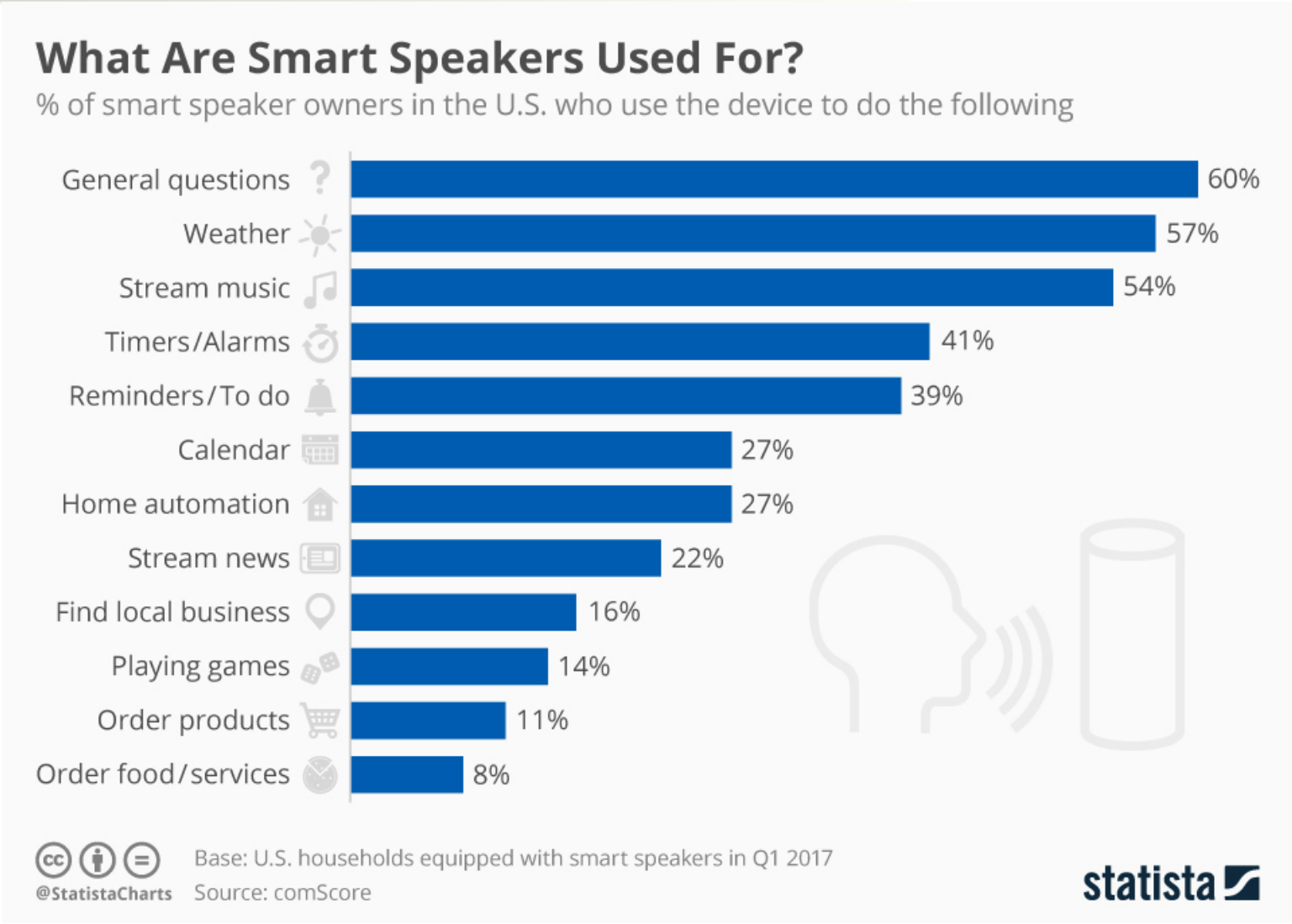
高信龙一 (Ryuichi Takanobu)

清华大学 计算机系

Conversational AI



Conversational AI



Outline

- 任务导向型对话系统背景
- 对话系统最新进展
- 对话系统评测
- 实验分析与讨论

Conversational System

- Dialog 1
  - M: Hello, welcome to the Cambridge dialog system. What kind of food would you like?
  - H: Yeah I want to find a restaurant that serves European food.
  - M: Hotel du Vin and Bistro is a nice place. It serves European food.
  - H: Uh what is the address and phone number?
  - M: Their address is XXX and their phone is YYY.
  - H: Thank you, goodbye.

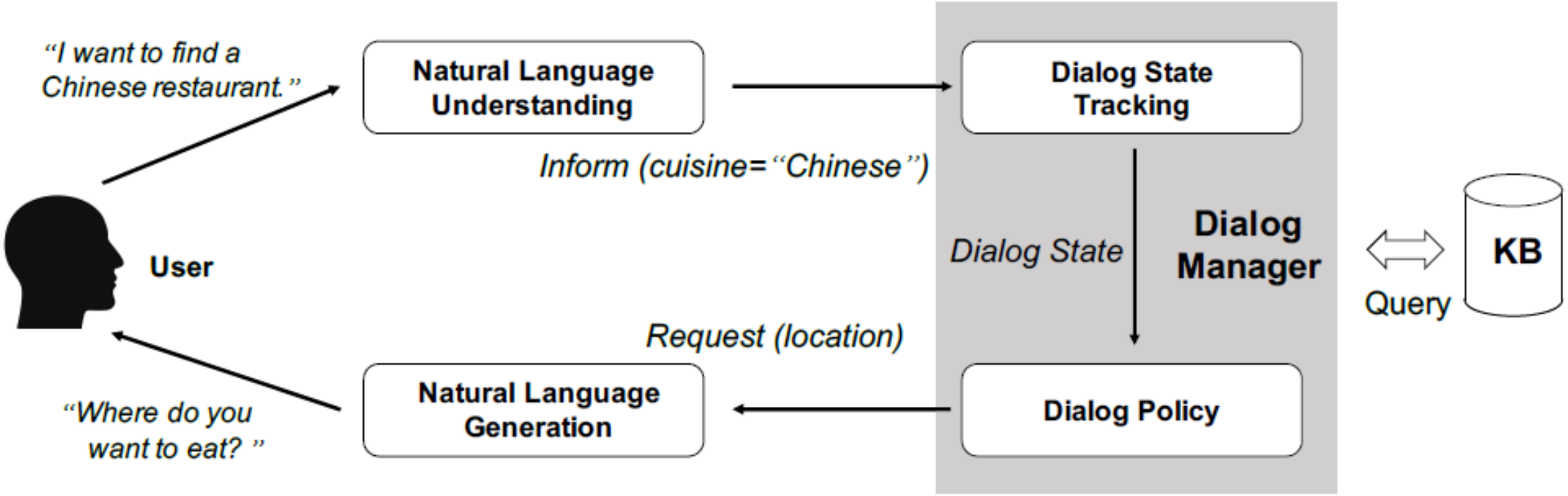
Conversational System

- Dialog 2
  - A: How old are you?
  - B: I’m 16. Why are you asking?
  - A: I thought you were 12.
  - B: What made you think so?
  - A: I don’t know what you are talking about
  - B: You don’t know what you are saying.

Conversational System

- Task-Oriented Dialog System
  - Goal-oriented
  - Require precise understanding, it's hard to collect data
  - Modular, highly hand-crafted, restricted ability, but meaningful and useful
- Chat-Based Conversational Agent
  - Chit-chat (no goal)
  - Large amounts of data (but probably not helpful so much)
  - End-to-end, highly data-driven, but meaningless/inappropriate and unreliable

Task-Oriented Dialog



The common pipeline architecture of a task-oriented dialog system

Zheng Z, et al. 2020. Recent advances and challenges in task-oriented dialog systems.

Component-wise Evaluation

- NLU
  - identify user intents and extract associated information (e.g. slots and values) from users' raw utterances
  - Slot F1, Intent F1
- DST
  - encode the extracted information as a compact set of dialog state
  - informable slots with user constraints, and requestable slots
  - Slot acc., Joint goal acc.

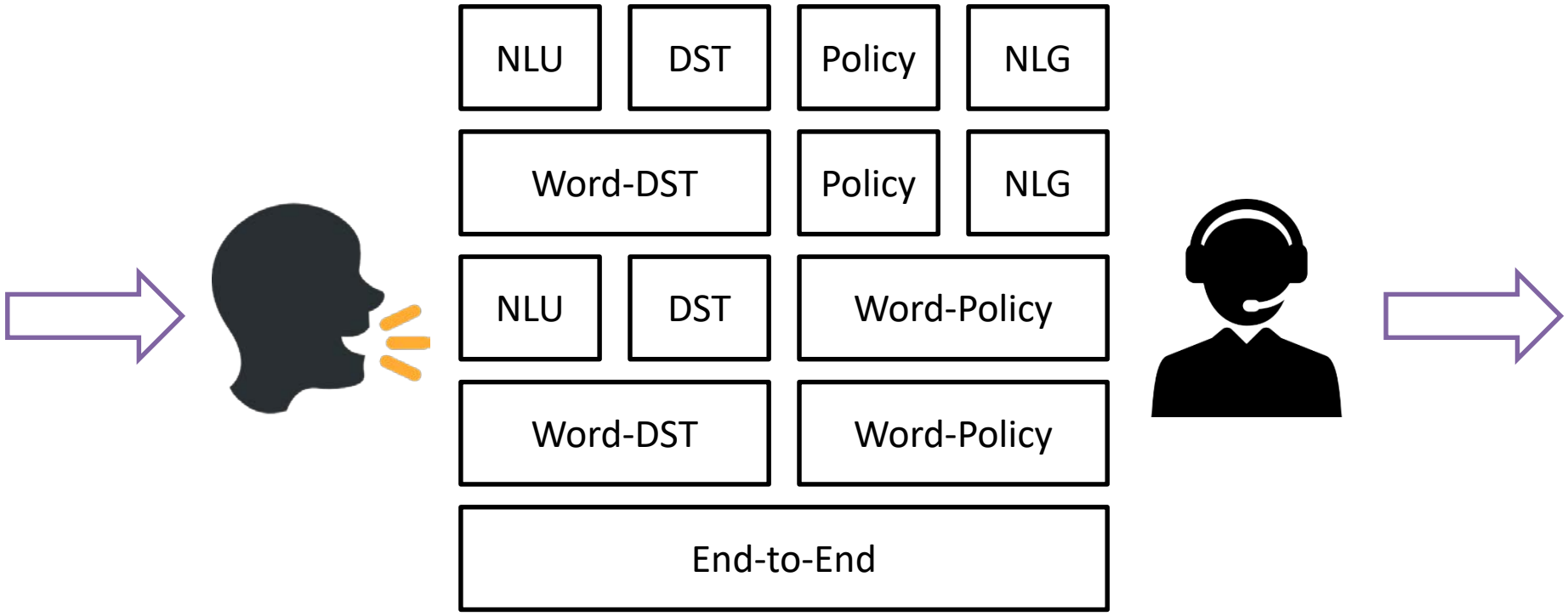
Component-wise Evaluation

- Policy
  - rely on the dialog state to select a system action
  - Inform, Match, Success
- NLG
  - generate a natural language response from a structured representation
  - BLEU, Perplexity, Slot error rate

Each module is evaluated separately
Almost all metrics are in the single-turn setting

Architecture

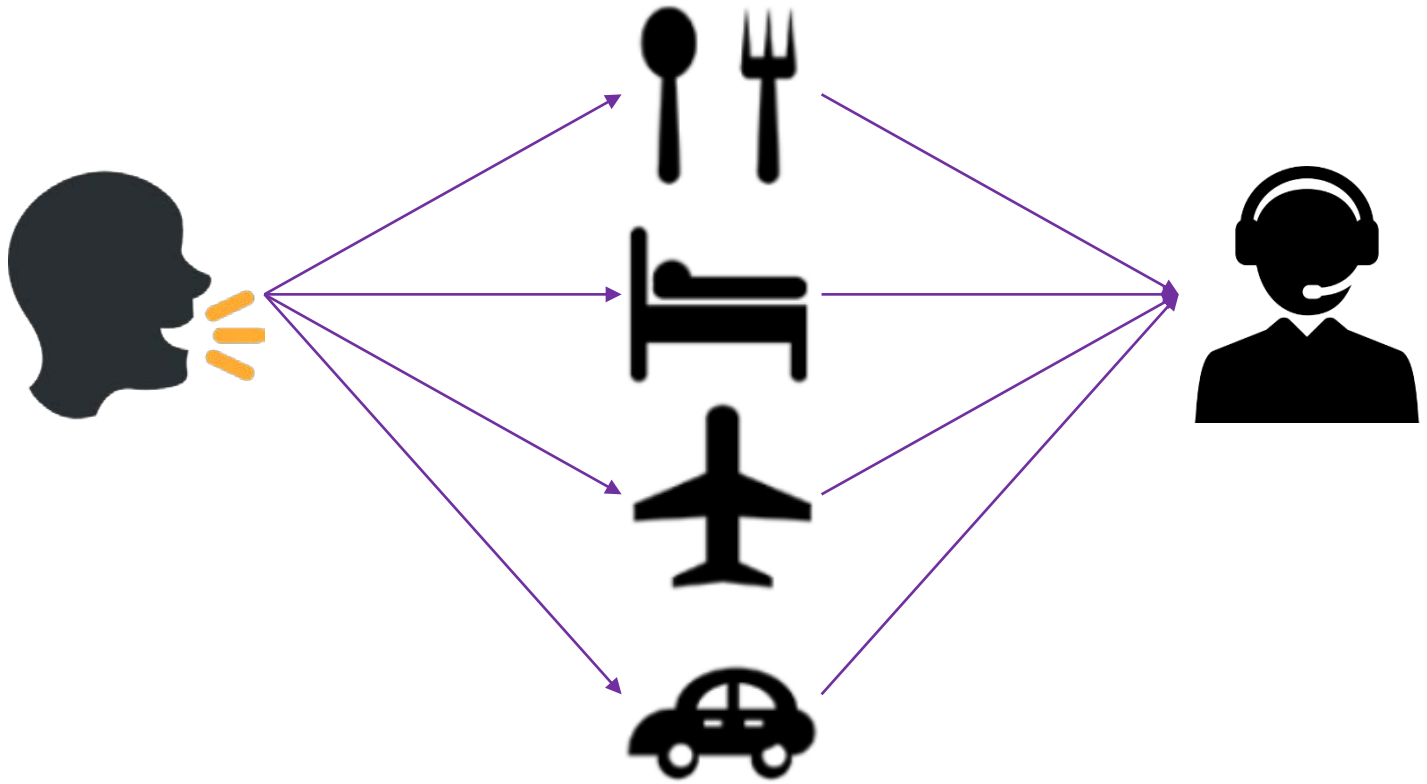
- Pipeline or end-to-end?



A wide variety of system configurations settings

Task Complexity

- Scalability and robustness



Multi-Turn Evaluation

- Both A and B have 80% accuracy, which one is better?

Turn	A	B
1	✘	✔
2	✔	✔
3	✔	✔
4	✔	✔
5	✔	✘

Contribution

- Multi-turn, system-wise simulated & human evaluation over a wide variety of configurations & settings to investigate overall performance
  - Which configurations lead to better goal-oriented dialog systems?
  - Whether the component-wise, single-turn metrics are consistent with system-wise, multi-turn metrics for evaluation?
  - How does the performance vary when a system is evaluated using tasks of different complexities?
  - Does simulated evaluation correlate well with human evaluation?

Ryuichi T, et al. 2020. Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation. SIGDIAL 2020 **Best Paper**

Data

- MultiWOZ
  - Dialog state
  - System dialog acts
  - User goal
  - Lots of baselines for different sub-tasks
- Augment corpus with NLU annotation
  - User dialog acts

Metric	DSTC2	SFX	WOZ2.0	FRAMES	KVRET	M2M	MultiWOZ
# Dialogues	1,612	1,006	600	1,369	2,425	1,500	**8,438**
Total # turns	23,354	12,396	4,472	19,986	12,732	14,796	**115,424**
Total # tokens	199,431	108,975	50,264	251,867	102,077	121,977	**1,520,970**
Avg. turns per dialogue	14.49	12.32	7.45	**14.60**	5.25	9.86	13.68
Avg. tokens per turn	8.54	8.79	11.24	12.60	8.02	8.24	**13.18**
Total unique tokens	986	1,473	2,142	12,043	2,842	1,008	**24,071**
# Slots	8	14	4	**61**	13	14	25
# Values	212	1847	99	3871	1363	138	**4510**

Budzianowski P, et al. 2018. MultiWOZ: A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling
EMNLP 2018 **Best Resource Paper**

User Goal

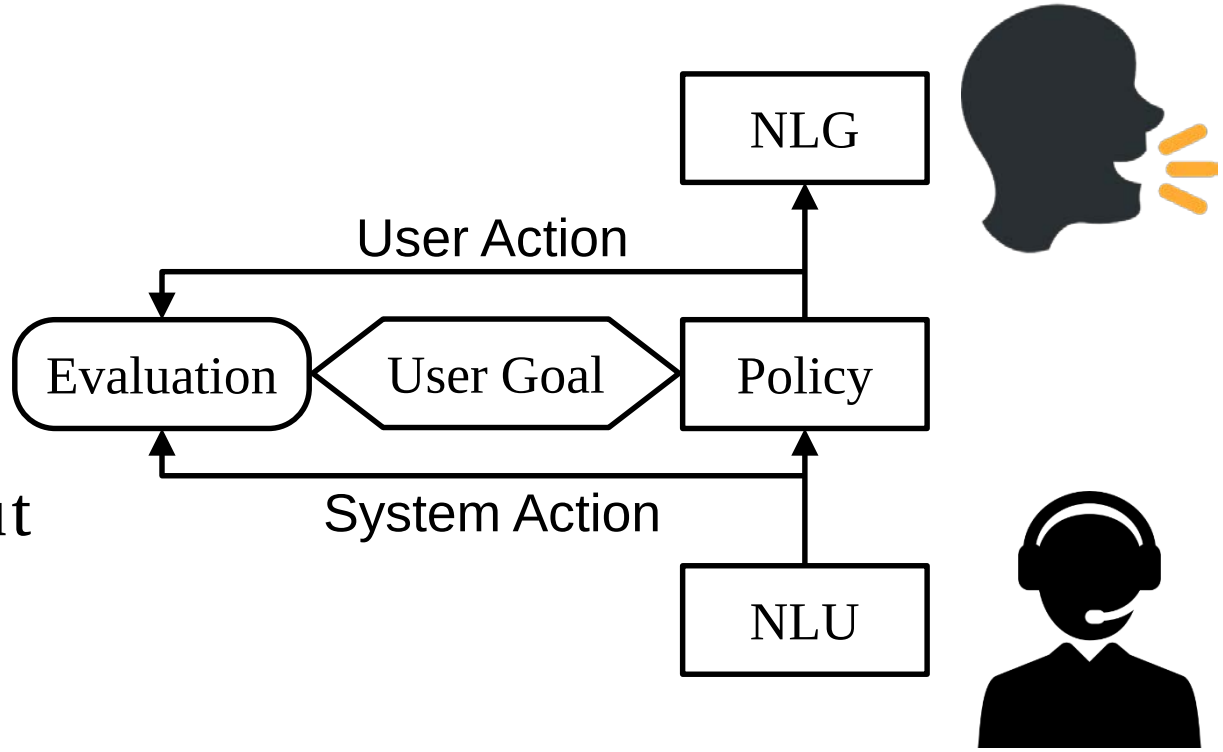
- Description of the state that a user wants to reach in a conversation

Attraction		Hotel			Taxi	
Info	Reqt	Info	Reqt	Book	Info	Reqt
type=museum area=centre	entrance fee address postcode	price range=expensive stars=4 type=hotel parking=yes	address area	people=8 day=saturday stay=5	arrive by=19:45	car type phone

- A fixed set of 1,000 user goals for evaluation

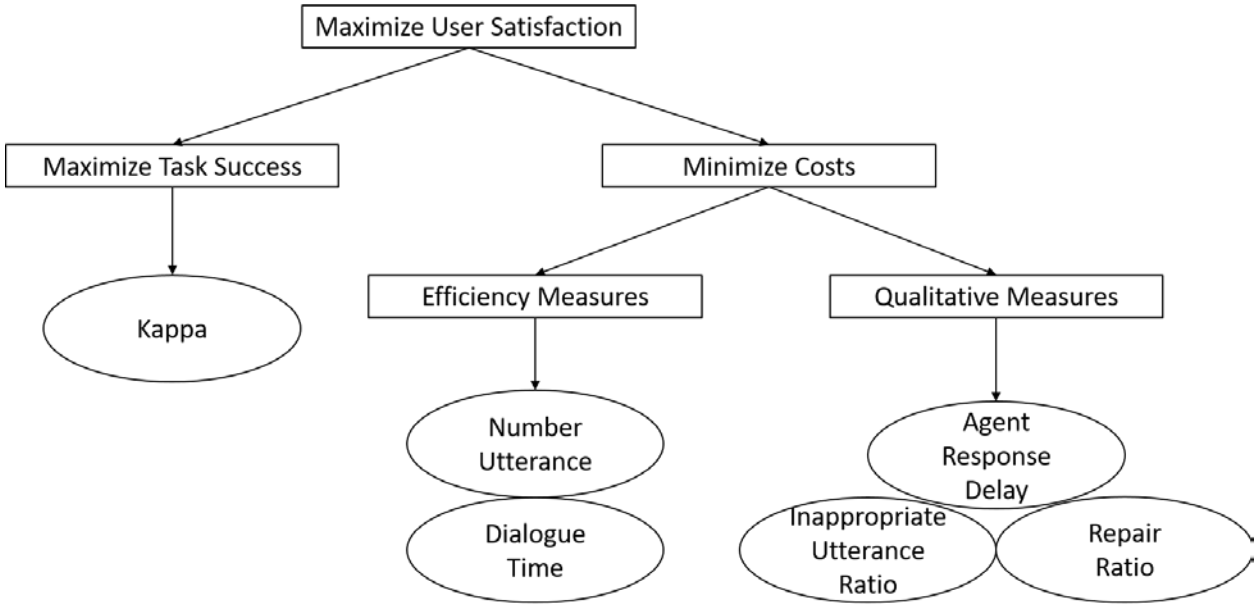
Platform & Simulator

- ConvLab with benchmarks
- Agenda-based user simulator
  - NLU: MILU (RNN-based)
  - Policy: Agenda
  - NLG: Retrieval
- Dialog acts from the input and output of user policy are used for evaluation



System-wise Evaluation

- Multi-turn evaluation
- Dialog cost
  - Dialog turn (user + system)
- Task success
  - Inform F1: requests
  - Match rate: constraints



Walker M A, et al. 1997. PARADISE: a framework for evaluating spoken dialogue agents.

System Configurations & Models

- 4 pipeline, 10 joint, 2 end-to-end systems
- NLU
  - joint tagging scheme: combine with domain annotation
  - BERT, MILU
- DST
  - DA-level: Rule
  - Word-level: MDBT, TRADE, SUMBT, COMER

System Configurations & Models

- Policy
  - DA-level: Rule, GDPL
  - Word-level: MDRG, HDSA, LaRL
- NLG
  - Retrieval, SCLSTM
- E2E
  - TSCP, DAMD

Perf. under Different Settings

- SYSTEM 1/2 achieves high overall performance
- Pipeline > Joint/End-to-end

ID	Configuration				Turn	Inform			Match	Succ.
	NLU	DST	Policy	NLG		Prec.	Rec.	F1		
1	BERT	rule	rule	retrieval	6.79	0.79	0.91	0.83	90.54	80.9
2	MILU	rule	rule	retrieval	7.24	0.76	0.88	0.80	87.93	77.6
3	BERT	rule	GDPL	retrieval	10.86	0.72	0.69	0.69	68.34	54.1
4	BERT	rule	rule	SCLSTM	13.38	0.64	0.58	0.58	51.41	43.0
5	MDBT		rule	retrieval	16.55	0.47	0.35	0.37	39.76	18.8
6	SUMBT		rule	retrieval	13.71	0.51	0.44	0.44	46.44	27.8
7	TRADE		rule	retrieval	9.56	0.39	0.41	0.37	38.37	22.4
8	COMER		rule	retrieval	16.79	0.30	0.28	0.28	29.06	17.3
9	BERT	rule	MDRG		17.90	0.35	0.34	0.32	29.07	19.2
10	BERT	rule	HDSA		15.91	0.47	0.62	0.50	39.21	34.3
11	BERT	rule	LaRL		13.08	0.40	0.68	0.48	68.95	47.7
12	SUMBT		HDSA		18.67	0.27	0.32	0.26	14.78	13.7
13	SUMBT		LaRL		13.92	0.36	0.64	0.44	57.63	40.4
14	TRADE		LaRL		14.44	0.35	0.57	0.40	36.07	30.8
15	TSCP				18.20	0.37	0.32	0.31	13.68	11.8
16	DAMD				11.27	0.64	0.69	0.64	59.67	48.5

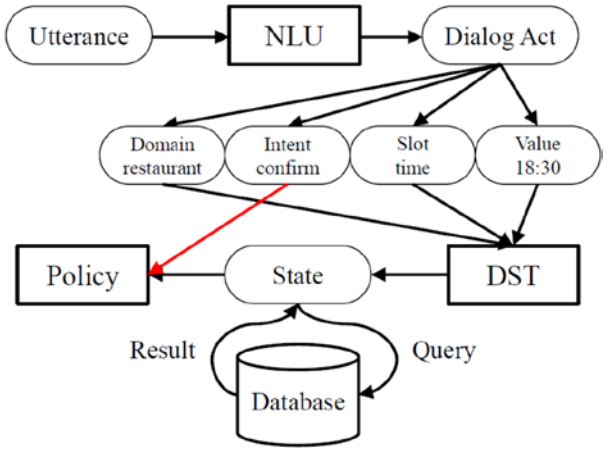
Perf. under Different Settings

- Database query
  - A large-scale external database
- Some non-pipeline systems perform relatively well (11, 13, 16)

ID	Configuration				Turn	Inform			Match	Succ.
	NLU	DST	Policy	NLG		Prec.	Rec.	F1		
1	BERT	rule	rule	retrieval	6.79	0.79	0.91	0.83	90.54	80.9
2	MILU	rule	rule	retrieval	7.24	0.76	0.88	0.80	87.93	77.6
3	BERT	rule	GDPL	retrieval	10.86	0.72	0.69	0.69	68.34	54.1
4	BERT	rule	rule	SCLSTM	13.38	0.64	0.58	0.58	51.41	43.0
5	MDBT		rule	retrieval	16.55	0.47	0.35	0.37	39.76	18.8
6	SUMBT		rule	retrieval	13.71	0.51	0.44	0.44	46.44	27.8
7	TRADE		rule	retrieval	9.56	0.39	0.41	0.37	38.37	22.4
8	COMER		rule	retrieval	16.79	0.30	0.28	0.28	29.06	17.3
9	BERT	rule	MDRG		17.90	0.35	0.34	0.32	29.07	19.2
10	BERT	rule	HDSA		15.91	0.47	0.62	0.50	39.21	34.3
11	BERT	rule	LaRL		13.08	0.40	0.68	0.48	68.95	47.7
12	SUMBT		HDSA		18.67	0.27	0.32	0.26	14.78	13.7
13	SUMBT		LaRL		13.92	0.36	0.64	0.44	57.63	40.4
14	TRADE		LaRL		14.44	0.35	0.57	0.40	36.07	30.8
15	TSCP				18.20	0.37	0.32	0.31	13.68	11.8
16	DAMD				11.27	0.64	0.69	0.64	59.67	48.5

Perf. under Different Settings

- NLU + DST >> Word-DST

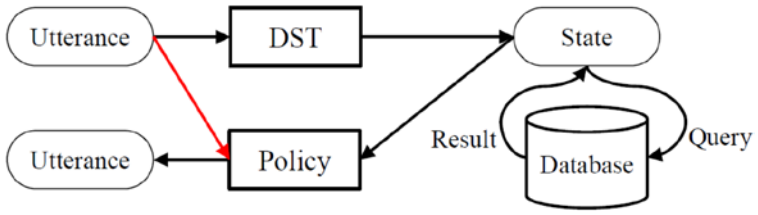


Missing intent info. in word-DST

ID	Configuration				Turn	Inform			Match	Succ.
	NLU	DST	Policy	NLG		Prec.	Rec.	F1		
1	BERT	rule	rule	retrieval	6.79	0.79	0.91	0.83	90.54	80.9
2	MILU	rule	rule	retrieval	7.24	0.76	0.88	0.80	87.93	77.6
3	BERT	rule	GDPL	retrieval	10.86	0.72	0.69	0.69	68.34	54.1
4	BERT	rule	rule	SCLSTM	13.38	0.64	0.58	0.58	51.41	43.0
5	MDBT		rule	retrieval	16.55	0.47	0.35	0.37	39.76	18.8
6	SUMBT		rule	retrieval	13.71	0.51	0.44	0.44	46.44	27.8
7	TRADE		rule	retrieval	9.56	0.39	0.41	0.37	38.37	22.4
8	COMER		rule	retrieval	16.79	0.30	0.28	0.28	29.06	17.3
9	BERT	rule	MDRG		17.90	0.35	0.34	0.32	29.07	19.2
10	BERT	rule	HDSA		15.91	0.47	0.62	0.50	39.21	34.3
11	BERT	rule	LaRL		13.08	0.40	0.68	0.48	68.95	47.7
12	SUMBT		HDSA		18.67	0.27	0.32	0.26	14.78	13.7
13	SUMBT		LaRL		13.92	0.36	0.64	0.44	57.63	40.4
14	TRADE		LaRL		14.44	0.35	0.57	0.40	36.07	30.8
15	TSCP				18.20	0.37	0.32	0.31	13.68	11.8
16	DAMD				11.27	0.64	0.69	0.64	59.67	48.5

Perf. under Different Settings

- Policy + NLG > Word-Policy
  - Word-DST + Word-Policy performs relatively well



Utterances are encoded again in word-policy

ID	Configuration				Turn	Inform			Match	Succ.
	NLU	DST	Policy	NLG		Prec.	Rec.	F1		
1	BERT	rule	rule	retrieval	6.79	0.79	0.91	0.83	90.54	80.9
2	MILU	rule	rule	retrieval	7.24	0.76	0.88	0.80	87.93	77.6
3	BERT	rule	GDPL	retrieval	10.86	0.72	0.69	0.69	68.34	54.1
4	BERT	rule	rule	SCLSTM	13.38	0.64	0.58	0.58	51.41	43.0
5	MDBT		rule	retrieval	16.55	0.47	0.35	0.37	39.76	18.8
6	SUMBT		rule	retrieval	13.71	0.51	0.44	0.44	46.44	27.8
7	TRADE		rule	retrieval	9.56	0.39	0.41	0.37	38.37	22.4
8	COMER		rule	retrieval	16.79	0.30	0.28	0.28	29.06	17.3
9	BERT	rule	MDRG		17.90	0.35	0.34	0.32	29.07	19.2
10	BERT	rule	HDSA		15.91	0.47	0.62	0.50	39.21	34.3
11	BERT	rule	LaRL		13.08	0.40	0.68	0.48	68.95	47.7
12	SUMBT		HDSA		18.67	0.27	0.32	0.26	14.78	13.7
13	SUMBT		LaRL		13.92	0.36	0.64	0.44	57.63	40.4
14	TRADE		LaRL		14.44	0.35	0.57	0.40	36.07	30.8
15	TSCP				18.20	0.37	0.32	0.31	13.68	11.8
16	DAMD				11.27	0.64	0.69	0.64	59.67	48.5

Component-wise vs. System-wise Evaluation

- NLU

Model	Slot	Intent	Overall
MILU	81.90	85.82	83.27
BERT	84.25	89.84	86.21

(a) NLU

- DST
  - Not consistent

Model	Slot Acc.	Joint Acc.
MDBT†	89.53	15.57
SUMBT†	96.44	46.65
TRADE†	96.92	48.62
COMER	95.52	48.79

(b) Word-level DST

ID	Configuration				Turn	Inform			Match	Succ.
	NLU	DST	Policy	NLG		Prec.	Rec.	F1		
1	BERT	rule	rule	retrieval	6.79	0.79	0.91	0.83	90.54	80.9
2	MILU	rule	rule	retrieval	7.24	0.76	0.88	0.80	87.93	77.6
3	BERT	rule	GDPL	retrieval	10.86	0.72	0.69	0.69	68.34	54.1
4	BERT	rule	rule	SCLSTM	13.38	0.64	0.58	0.58	51.41	43.0
5	MDBT		rule	retrieval	16.55	0.47	0.35	0.37	39.76	18.8
6	SUMBT		rule	retrieval	13.71	0.51	0.44	0.44	46.44	27.8
7	TRADE		rule	retrieval	9.56	0.39	0.41	0.37	38.37	22.4
8	COMER		rule	retrieval	16.79	0.30	0.28	0.28	29.06	17.3
9	BERT	rule	MDRG		17.90	0.35	0.34	0.32	29.07	19.2
10	BERT	rule	HDSA		15.91	0.47	0.62	0.50	39.21	34.3
11	BERT	rule	LaRL		13.08	0.40	0.68	0.48	68.95	47.7
12	SUMBT		HDSA		18.67	0.27	0.32	0.26	14.78	13.7
13	SUMBT		LaRL		13.92	0.36	0.64	0.44	57.63	40.4
14	TRADE		LaRL		14.44	0.35	0.57	0.40	36.07	30.8
15	TSCP				18.20	0.37	0.32	0.31	13.68	11.8
16	DAMD				11.27	0.64	0.69	0.64	59.67	48.5

Component-wise vs. System-wise Evaluation

- Policy

Model	BLEU	Inform	Succ.
MDRG†	18.8	71.3	61.0
HDSA†	23.6	82.9	68.9
LaRL†	12.8	82.8	79.2

(c) Word-level Policy

- NLG
  - Not consistent

Model	BLEU	SER
Retrieval	33.1	–
SCLSTM	51.6	3.10

(d) NLG

ID	Configuration				Turn	Inform			Match	Succ.
	NLU	DST	Policy	NLG		Prec.	Rec.	F1		
1	BERT	rule	rule	retrieval	6.79	0.79	0.91	0.83	90.54	80.9
2	MILU	rule	rule	retrieval	7.24	0.76	0.88	0.80	87.93	77.6
3	BERT	rule	GDPL	retrieval	10.86	0.72	0.69	0.69	68.34	54.1
4	BERT	rule	rule	SCLSTM	13.38	0.64	0.58	0.58	51.41	43.0
5	MDBT		rule	retrieval	16.55	0.47	0.35	0.37	39.76	18.8
6	SUMBT		rule	retrieval	13.71	0.51	0.44	0.44	46.44	27.8
7	TRADE		rule	retrieval	9.56	0.39	0.41	0.37	38.37	22.4
8	COMER		rule	retrieval	16.79	0.30	0.28	0.28	29.06	17.3
9	BERT	rule	MDRG		17.90	0.35	0.34	0.32	29.07	19.2
10	BERT	rule	HDSA		15.91	0.47	0.62	0.50	39.21	34.3
11	BERT	rule	LaRL		13.08	0.40	0.68	0.48	68.95	47.7
12	SUMBT		HDSA		18.67	0.27	0.32	0.26	14.78	13.7
13	SUMBT		LaRL		13.92	0.36	0.64	0.44	57.63	40.4
14	TRADE		LaRL		14.44	0.35	0.57	0.40	36.07	30.8
15	TSCP				18.20	0.37	0.32	0.31	13.68	11.8
16	DAMD				11.27	0.64	0.69	0.64	59.67	48.5

Single-turn vs. Multi-turn Evaluation

- E2E

Model	BLEU	Inform	Succ.
TSCP	15.5	66.4	45.3
DAMD	16.6	76.3	60.4

(e) E2E

- Error propagation to downstream modules and effect on following turns

ID	Configuration				Turn	Inform			Match	Succ.
	NLU	DST	Policy	NLG		Prec.	Rec.	F1		
1	BERT	rule	rule	retrieval	6.79	0.79	0.91	0.83	90.54	80.9
2	MILU	rule	rule	retrieval	7.24	0.76	0.88	0.80	87.93	77.6
3	BERT	rule	GDPL	retrieval	10.86	0.72	0.69	0.69	68.34	54.1
4	BERT	rule	rule	SCLSTM	13.38	0.64	0.58	0.58	51.41	43.0
5	MDBT		rule	retrieval	16.55	0.47	0.35	0.37	39.76	18.8
6	SUMBT		rule	retrieval	13.71	0.51	0.44	0.44	46.44	27.8
7	TRADE		rule	retrieval	9.56	0.39	0.41	0.37	38.37	22.4
8	COMER		rule	retrieval	16.79	0.30	0.28	0.28	29.06	17.3
9	BERT	rule	MDRG		17.90	0.35	0.34	0.32	29.07	19.2
10	BERT	rule	HDSA		15.91	0.47	0.62	0.50	39.21	34.3
11	BERT	rule	LaRL		13.08	0.40	0.68	0.48	68.95	47.7
12	SUMBT		HDSA		18.67	0.27	0.32	0.26	14.78	13.7
13	SUMBT		LaRL		13.92	0.36	0.64	0.44	57.63	40.4
14	TRADE		LaRL		14.44	0.35	0.57	0.40	36.07	30.8
15	TSCP				18.20	0.37	0.32	0.31	13.68	11.8
16	DAMD				11.27	0.64	0.69	0.64	59.67	48.5

Perf. of Task with Different Complexities

- Different single domain

ID	Restaurant				Train				Attraction		
	Turn	Info.	Match	Succ.	Turn	Info.	Match	Succ.	Turn	Info.	Succ.
1	2.82	0.94	96.9	98	3.06	1.0	100	100	3.12	0.69	63
2	2.84	0.92	100	98	2.99	1.0	94.2	97	3.70	0.73	65
3	8.68	0.70	69.4	70	6.07	0.80	67.3	75	5.61	0.67	62
4	6.00	0.77	68.8	78	11.53	0.71	67.3	55	12.57	0.57	46
6	9.41	0.64	72.7	60	5.13	0.97	90.4	93	14.79	0.23	9
11	9.91	0.39	66.7	61	4.02	0.86	88.5	97	4.73	0.68	80
13	8.35	0.40	65.6	60	4.19	0.85	94.2	96	6.06	0.60	73
15	14.72	0.37	11.5	27	16.02	0.46	11.5	25	16.12	0.51	24
16	6.36	0.80	92.2	90	10.21	0.61	55.8	58	8.32	0.69	67

Most systems achieve better performance in *Restaurant* and *Train* than *Attraction*

Perf. of Task with Different Complexities

- Different number of domains

ID	Single				Two				Three			
	Turn	Info.	Match	Succ.	Turn	Info.	Match	Succ.	Turn	Info.	Match	Succ.
1	3.22	0.84	84.7	87	6.96	0.81	94.9	78	8.15	0.82	88.4	69
2	3.90	0.78	79.7	82	6.74	0.76	95.3	72	10.54	0.79	85.0	66
3	9.18	0.67	66.7	60	12.38	0.60	42.9	42	13.55	0.50	44.6	21
4	8.65	0.66	58.3	62	17.24	0.38	28.0	14	18.03	0.46	24.4	13
6	10.35	0.44	60.4	41	14.74	0.44	50.9	17	15.97	0.25	20.9	0
11	8.79	0.45	72.2	55	13.37	0.52	74.0	59	19.30	0.39	50.4	0
13	8.48	0.45	62.5	61	14.08	0.45	61.0	47	18.95	0.36	40.7	0
15	15.09	0.33	10.0	26	19.10	0.25	17.8	8	20.00	0.19	0.0	1
16	8.89	0.66	68.1	65	13.48	0.52	57.1	34	18.59	0.58	45.5	12

All systems have performance drop as the number of domains increases

Simulated vs. Human Evaluation

- A subset of 100 goals for human evaluation
- 5 systems across different architectures
- Other metrics
  - Language understanding
  - Response appropriateness

ID	Turn	LU	RA	Succ.	Corr.
1	18.58	3.62	3.69	62	0.57
6	20.63	2.85	2.91	27	0.72
11	19.98	2.36	2.41	23	0.53
13	19.26	2.17	2.49	14	0.46
16	16.33	2.61	2.65	23	0.55

- Simulated evaluation correlates moderately well with human evaluation
- Task success rate of most systems decreases significantly

Case Study: Simulated Evaluation

Hotel			Restaurant	
Info	Reqt	Book	Info	Book
parking=yes pricerange=moderate	postcode address	stay=2 day=sunday	food=british pricerange=moderate area=centre	day=sunday time=16:30 people=8

- SYSTEM-1: Success
- SYSTEM-6: Failure
  - fail to track the value of price range in the dialog state
  - do not answer the postcode

SYSTEM-1	SYSTEM-6
U: I would prefer it in the centre area . Actually I need a moderately priced restaurant . Are there any fitting that description ? This restaurant should serve british food too .	U: I would prefer it in the centre area . Actually I need a moderately priced restaurant . Are there any fitting that description ? This restaurant should serve british food too .
D: *restaurant-food:british, restaurant-pricerange:moderate, restaurant-area:centre*	D: *restaurant-food:british, restaurant-pricerange:moderate, restaurant-area:centre*
S: I have 4 options for you ! How about the copper kettle ? It is located in the centre .	S: I have 4 options for you ! It is in the centre area . That is a british restaurant . the copper kettle is just your thing . It is in the centre area .
U: I need a table for 8 people . Can you try to book it at 16:30 . We will be there sunday night .	U: It will be for 8 people . Can you try for 16:30 ? I am also looking for a restaurant to book for sunday.
D: *restaurant-people:8, restaurant-day:sunday, restaurant-time:16:30*	D: *restaurant-people:8, restaurant-day:sunday, restaurant-time:16:30*
S: Your reference number is 00000013 .	S: All set . Your reference number is 00000013 .
U: I would like to keep it in the moderate range , please . Oh , i need parking.	U: Hello , I would like to find a hotel that includes free parking . I also need a moderate place to stay .
D: *hotel-parking:yes, hotel-pricerange:moderate*	D: *hotel-parking:yes*
S: I have 14 different options for you ! How about hamilton lodge ? Fits your request perfectly . It has free wifi .	S: Reference number is : 00000013 .
U: That sounds fine , I just need the postcode though .	U: What is the postcode ?
D:	D:
S: The postal code for that hotel is cb41da .	S: Thank you for using our services .
U: I need a room starting on sunday . There is going to be 8 of us. How about only 2 nights .	U: That sounds fine , I just need the postcode though .
D: *hotel-day:sunday, hotel-people:8, hotel-stay:2*	D:
S: Here is the booking information : Booking was successful . Reference number is : 00000001.	S: Glad to have been of help . Thank you for using the service . Enjoy the rest of your day !
U: Thank you very much , goodbye .	U: I just need the postcode .
D:	D:
S: You are very welcome. Goodbye.	S: Ok , thank you . Have a good day .

Case Study: Human Evaluation

Train			Attraction	
Info	Reqt	Book	Info	Reqt
destination=peterborough day=saturday departure=cambridge leaveAt=13:00	duration trainID	people=7	area=centre	phone postcode

- SYSTEM-1: Success
- SYSTEM-6: Failure
  - do not answer the phone number of the requested tourist attraction

SYSTEM-1	SYSTEM-6
U: I am looking for a train that leaves at 13:00	U: Hi! I'm looking for a train departing from Cambridge after 13:00
S: What day would you like ? Where would you like your train to take you ? Where did you want to depart from ?	S: What day did you have in mind ? Where are you heading to ?
U: I would like to leave on Saturday and depart from Cambridge	U: Saturday and heading to peterborough
S: What is your destination ?	S: There is a train arriving at 13:00 would you like me to book tickets for that one ? I can book you on saturday. Would you like me to book a train to peterborough for you ? Would you like me to book a train from cambridge for you ?
U: I need to go to Peterborough	U: Yes, please book for 7 people
S: We can book you for the train leaving at 13:00 . I can book you on Saturday. Would you like me to book a train to peterborough for you ? Would you like me to book a train from cambridge for you ?	S: All set . Your reference number is 00000066 .
U: Yes. I need the train ID and booking for 7 people	U: Thanks! I'm looking forward to seeing local tourist attractions
S: Reference number is : 00000054 .	S: Thank you for using our services
U: I want to see the tourist attractions so it should be near the Centre	U: Do you know of any attractions in the centre?
S: There are 44 . Would you like me to recommend one for you ? Would you like vue cinema ?	S: We have 44 options , can i recommend for you ? I would suggest sheep's green and lammas land park fen causeway .
U: Yes. I also want the phone number and postcode for the cinema	U: Can I get the phone number for Sheep's Green?
S: The phone number is 08712240240 and postcode is cb11ps . They are located within the centre .	S: What time would you like to leave ?
U: Do you have me booked for a party of 7 people?	U: Around noon
S: Here is the booking information : Booking was successful . Reference number is : 00000054.	S: Have a good day .
U: Are there 7 seats booked?	U: Can you give me the phone number of the tourist attraction?
S: Reference number is : 00000054 .	S: Thank you for contacting the help desk . Have a great day .
U: Thank You for your help and Goodbye!	U: Can you give me a list of attractions?
S: You are welcome . Have a good day !	S: Have a good day .

Conclusion

- Pipeline systems often achieve better overall performance than joint models and end-to-end systems
- Results of component-wise, single-turn evaluation are not always consistent with that of system-wise, multi-turn evaluation
- Performance of dialog systems drops significantly with the increase of task complexity, while pipeline systems are relatively robust
- Simulated evaluation correlates moderately well with human evaluation

Discussion

- Good semantic parsing (e.g. dialog acts) is essential to build a dialog system
- Pipeline systems perform better, but require fine-grained annotations
- Proposed models should be assembled into a complete system for meaningful evaluation and fair comparison via a standardized platform
- Scalability and robustness are important and should be improved
- Simulated evaluation is still a valid alternative despite the discrepancy

Thanks for your attention

Paper: https://arxiv.org/abs/2005.07362

Homepage: https://truthless11.github.io

Contact: gxly19@mails.tsinghua.edu.cn