



Joint System and User Modelling

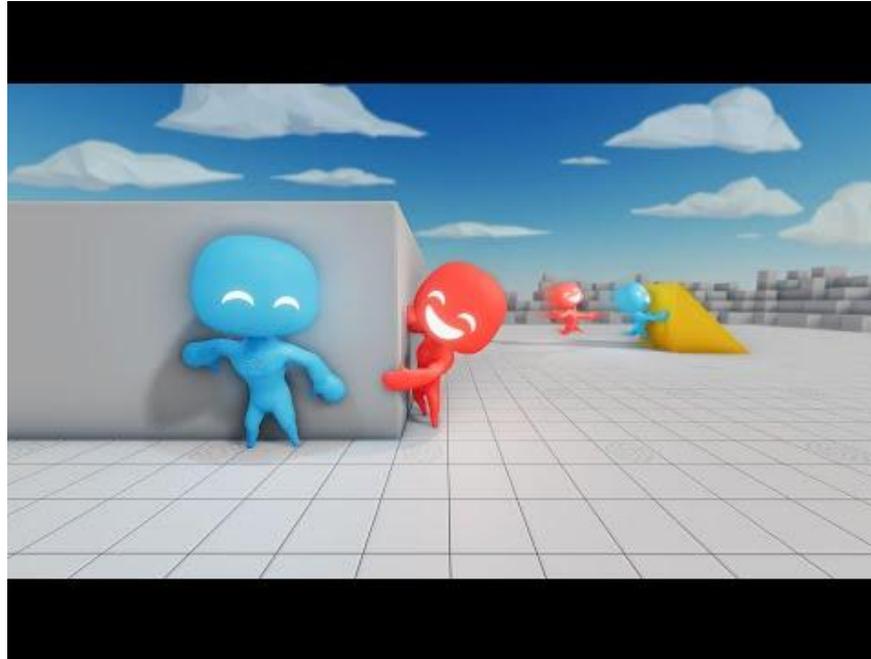
Hsien-chin Lin and Christian Geishauser
Dialogue Systems and Machine Learning Group

18.06.2021

- Part I: Multi-agent reinforcement learning for dialogue
 - What is multi-agent RL?
 - How has it been applied to dialogue?
 - Emergent communication
- Part II: Data augmentation and evaluation by machine-to-machine conversation
 - How to collect more dialogue data?
 - Evaluate dialogue systems by machine-to-machine interaction



Part I: Multi-agent reinforcement learning



Source: YouTube, Multi-Agent Hide and Seek

Why train a user agent with RL?

- Building a rule-based user simulator is labour-intensive and hard to maintain
- supervised trained user agents only know how to act in situations seen in the data

Why train a user agent with RL?

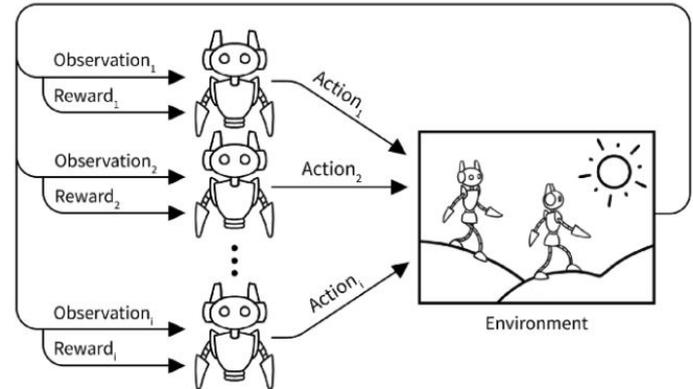
- Building a rule-based user simulator is labour-intensive and hard to maintain
- supervised trained user agents only know how to act in situations seen in the data
- Advantages of using RL for improving user agent:
 - can explore situations that haven't been observed in the data and learn how to act there
 - can optimise discrete metrics, for instance:
 - were all constraints and requests communicated?
 - did it inform the constraints before requesting information?
 - utilize RL algorithms
 - optimise for diversity to obtain different user agents, optimise for maximum entropy for diverse behaviour,...

Multi-agent Reinforcement Learning

- -> Train a user agent with RL
- Actual goal is to obtain a good dialogue system policy
- -> Train both user agent and system agent using RL
- We have two RL agents that need to be trained -> multi-agent RL setting

- Single agent RL
 - single agent that interacts with the environment
 - optimal strategy: maximise expected return for a given environment
- Multi agent RL:
 - multiple agents that interact with the environment
 - optimal strategy: varies depending on the behaviour of other agents that might change over time
 - -> environment becomes non-stationary for each individual agent
 - constantly need to adapt to environment changes

- Stochastic game (Markov game)
 - N : number of agents
 - S : state space
 - $A = A_1 \times \dots \times A_N$ joint action space
 - R : reward function, emitting reward for every agent i
 - P : transition probability distribution
 - $O = O_1 \times \dots \times O_N$ joint observation space
 - γ : discount factor



- Centralised learning
 - if all agents observe the full state, we can model cooperative multi-agent system as single meta-agent

- Centralised learning
 - if all agents observe the full state, we can model cooperative multi-agent system as single meta-agent
 - action space grows exponentially with number of agents
 - not applicable when each agent receives different observations
 - Multi-action policies can be considered as these (one agent per action, domain, ...)

- Centralised learning
 - if all agents observe the full state, we can model cooperative multi-agent system as single meta-agent
 - action space grows exponentially with number of agents
 - not applicable when each agent receives different observations
 - Multi-action policies can be considered as these (one agent per action, domain, ...)
- Independent learners
 - each agent independently learns its own policy, treating other agents as part of the environment

- Centralised learning
 - if all agents observe the full state, we can model cooperative multi-agent system as single meta-agent
 - action space grows exponentially with number of agents
 - not applicable when each agent receives different observations
 - Multi-action policies can be considered as these (one agent per action, domain, ...)
- Independent learners
 - each agent independently learns its own policy, treating other agents as part of the environment
 - Example: Independent Q-learning (IQL)
 - Does not suffer from exponential growth of action space
 - each agent only needs its local observation
 - Problem: non-stationarity of the environment (experience replay methods not straightforward)

- Centralized training with decentralized execution (CTDE)
 - assume you can observe the full state during training, but only local observations during testing
 - Critic is only necessary during training

- Centralized training with decentralized execution (CTDE)
 - assume you can observe the full state during training, but only local observations during testing
 - Critic is only necessary during training
 - -> learn centralised critic operating on the full space
 - -> learn actors independently on their respective smaller space (important in dialogue)

- Centralized training with decentralized execution (CTDE)
 - assume you can observe the full state during training, but only local observations during testing
 - Critic is only necessary during training
 - -> learn centralised critic operating on the full space
 - -> learn actors independently on their respective smaller space (important in dialogue)
 - helps to resolve the non-stationarity of the critic

$$g = \nabla_{\theta} \log \pi(u|\tau_t)(r + \gamma V(s_{t+1}) - V(s_t))$$

local observation  global observation 

- Dialogue system and user simulator are both agents that can be optimised using RL
 - common goal: successfully complete the conversation, i.e. fulfilling the user goal
 - user must provide information about the problem to be solved
 - system must solve the problem using goal information that was provided by the user

- Dialogue system and user simulator are both agents that can be optimised using RL
 - common goal: successfully complete the conversation, i.e. fulfilling the user goal
 - user must provide information about the problem to be solved
 - system must solve the problem using goal information that was provided by the user

- Games like Hide-and-Seek:
 - agents can come up with their own behaviour to solve the game
 - in dialogue: agents should not come up with an artificial language to complete the conversation

- Liu, Lane, 2017 (IterDPL)
 - train agents iteratively to deal with non-stationarity
- Papangelis et. al., 2019
 - Model the dialogue as stochastic game and optimise via WoLF-PHC
 - slow learning rate when winning, high learning rate when losing
 - train agents concurrently
- Both show improved performance over supervised trained agents
 - using DSTC2 data for supervised training

- CTDE for dialogue
 - user simulator has own observation such as the goal
 - dialogue system has own observation such as database results
 - -> no meta-agent that operates on the shared space
- Critic that has information from user and system can better estimate task success

- CTDE for dialogue
 - user simulator has own observation such as the goal
 - dialogue system has own observation such as database results
 - -> no meta-agent that operates on the shared space
- Critic that has information from user and system can better estimate task success
- Takanobu et. al., 2020 (MADPL)
 - use centralised critic to estimate the shared goal of task success
 - use dedicated critics for the private goals of user and dialogue system
 - penalty when requested info is not provided immediately; task success rate based on user agent description
 - penalty when requesting too early; reward when all constraints and requests are communicated
 - work in MultiWOZ setting

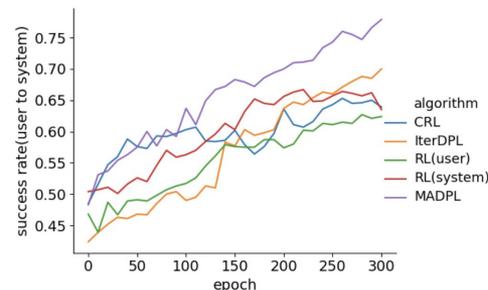
- Takanobu et. al., 2020 (MADPL)

$$\nabla_{\phi} \log \pi_{\phi}(a^S | s^S) [A^S(s^S) + A^G(s)]$$

$$\nabla_{\omega} \log \mu_{\omega}(a^U | s^U) [A^U(s^U) + A^G(s)]$$

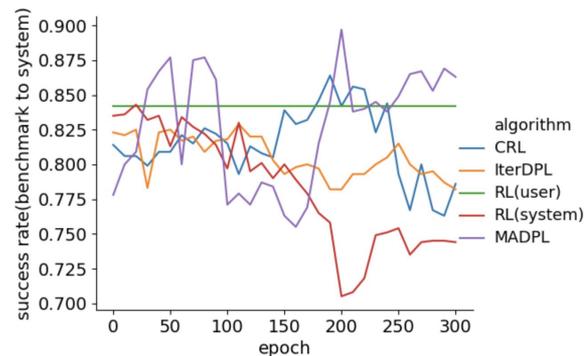
- CRL uses centralised critic
- IterDPL trains the two agents iteratively instead of concurrently
- All algorithms are able to improve success

System	User	Turns	Inform	Match	Success
SL	SL	6.34	73.08	82.58	49.7
SL	RL	8.75	76.86	76.28	60.2
RL	SL	6.20	72.84	79.15	51.1
RL	RL	7.92	75.96	70.37	58.7
CRL		8.13	68.29	89.71	66.6
IterDPL		8.79	74.01	81.04	64.6
MADPL		8.96	76.26	90.98	70.1



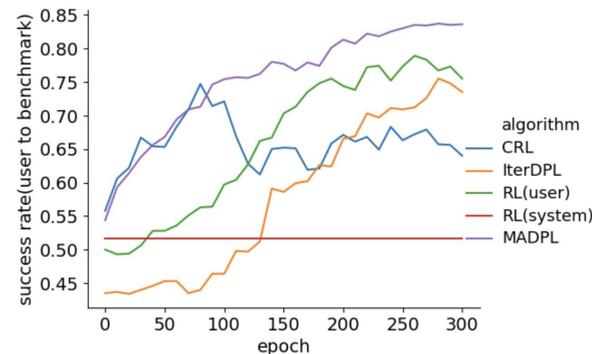
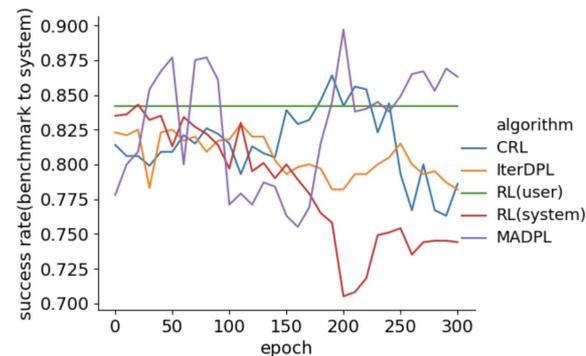
Multi-agent RL for Dialogue

- Evaluation with **rule-based user simulator** during multi-agent training
 - erratic behaviour
 - large action space of system
 - more possible optimal solutions?



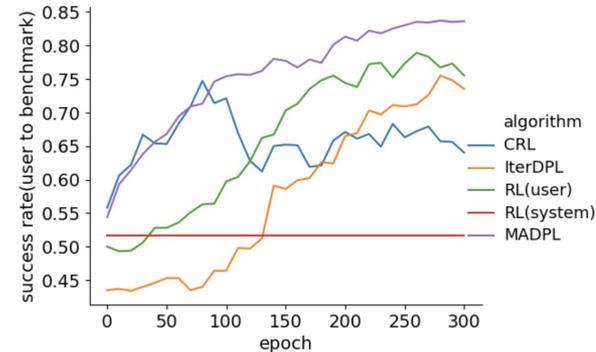
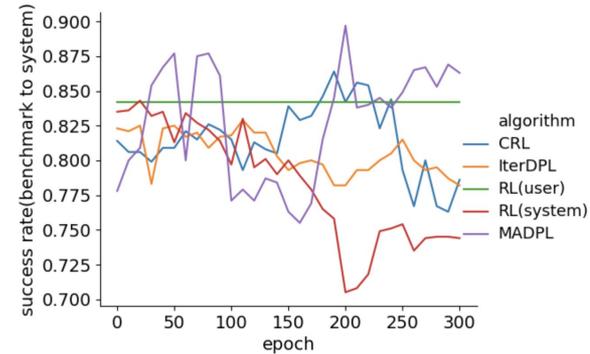
Multi-agent RL for Dialogue

- Evaluation with **rule-based user simulator** during multi-agent training
 - erratic behaviour
 - large action space of system
 - more possible optimal solutions?
- Evaluation with **rule-based system** during multi-agent training
 - rule-based system is able to interact with user agent
 - easier to learn a good user agent



Multi-agent RL for Dialogue

- Evaluation with **rule-based user simulator** during multi-agent training
 - erratic behaviour
 - large action space of system
 - more possible optimal solutions?
- Evaluation with **rule-based system** during multi-agent training
 - rule-based system is able to interact with user agent
 - easier to learn a good user agent
- Need rule-based methods for this evaluation
 - exactly what should be omitted



- Human evaluation
 - humans compare dialogues generated by different algorithms and give preference

- Human evaluation
 - humans compare dialogues generated by different algorithms and give preference
- Role of supervised pre-training
 - training only a system agent from scratch with a rule-based simulator is already challenging
 - training both concurrently makes it much more difficult
 - supervised pre-training makes RL training possible in the first place
 - supervised pre-training gives already a good bias towards the solution one wants to have
 - can we squeeze more out of the data? (Imitation learning, Inverse RL, ...)

- Human evaluation
 - humans compare dialogues generated by different algorithms and give preference
- Role of supervised pre-training
 - training only a system agent from scratch with a rule-based simulator is already challenging
 - training both concurrently makes it much more difficult
 - supervised pre-training makes RL training possible in the first place
 - supervised pre-training gives already a good bias towards the solution one wants to have
 - can we squeeze more out of the data? (Imitation learning, Inverse RL, ...)
- Can we train entirely without data?
 - and use that bootstrapped policy to interact/learn with real humans

- Emergent communication: Language arises because agents must communicate in order to solve a task
 - language that is learned has solely a functional purpose

- Emergent communication: Language arises because agents must communicate in order to solve a task
 - language that is learned has solely a functional purpose
- Common task: Speaker and listener communicate
 - speaker needs to navigate listener to a certain spot in the map
 - speaker needs to describe a target image that the listener has to select
 - user agent and system communicate goal and information?

- Emergent communication: Language arises because agents must communicate in order to solve a task
 - language that is learned has solely a functional purpose
- Common task: Speaker and listener communicate
 - speaker needs to navigate listener to a certain spot in the map
 - speaker needs to describe a target image that the listener has to select
 - user agent and system communicate goal and information?
- Typical questions
 - Is the emerged language interpretable?
 - is it of compositional nature?

- Kottur et. al. 2017
 - Natural language does not emerge “naturally” in multi-agent dialog
 - large enough vocabulary results in symbols mapped to instances
 - small vocabulary still non-compositional and hard to interpret

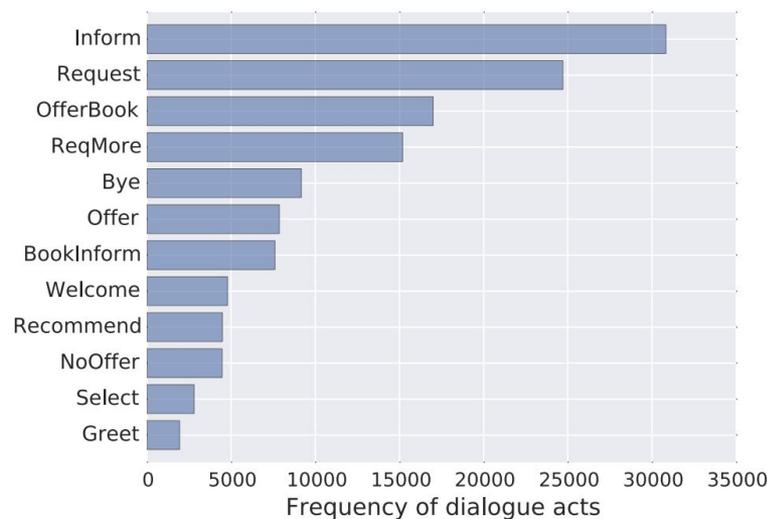
- Kottur et. al., 2017
 - Natural language does not emerge “naturally” in multi-agent dialog
 - large enough vocabulary results in symbols mapped to instances
 - small vocabulary still non-compositional and hard to interpret
- Lazaridou et. al., 2020
 - language models infer structural properties of language from text corpora
 - ignore the functional aspect of communication
 - multi-agent communication focuses only on using language as a utility
 - no “natural” language
 - functional learning: “what to say”
 - structural learning: “how to say”

Part II: Data augmentation and evaluation

- Learning from human
 - Time consuming and costly
 - Cold start

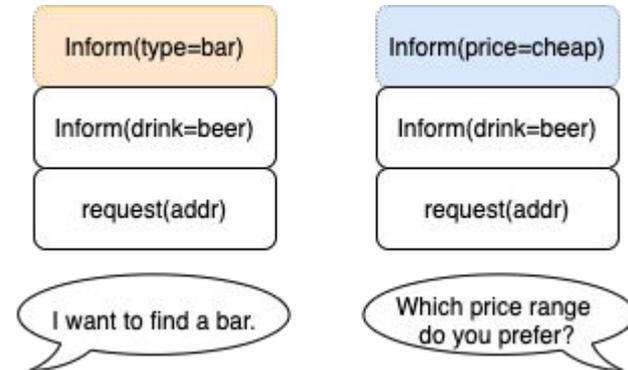
- Learning from human
 - Time consuming and costly
 - Cold start
- Data collection
 - Annotation (MultiWOZ 2.0, 2.1, 2.2, 2.3, 2.4)

- Learning from human
 - Time consuming and costly
 - Cold start
- Data collection
 - Annotation
- Coverage and bias



MultiWoz 2.0 (Budzianowski et al., 2018)

- Rule-based user simulator
- Pros
 - data free
 - interpretable
- Cons
 - Require expert knowledge
 - Domain dependent
 - Not human like
 - Hard to build for complicated domains



Agenda-based user simulator
(Schatzmann et al. 2007)

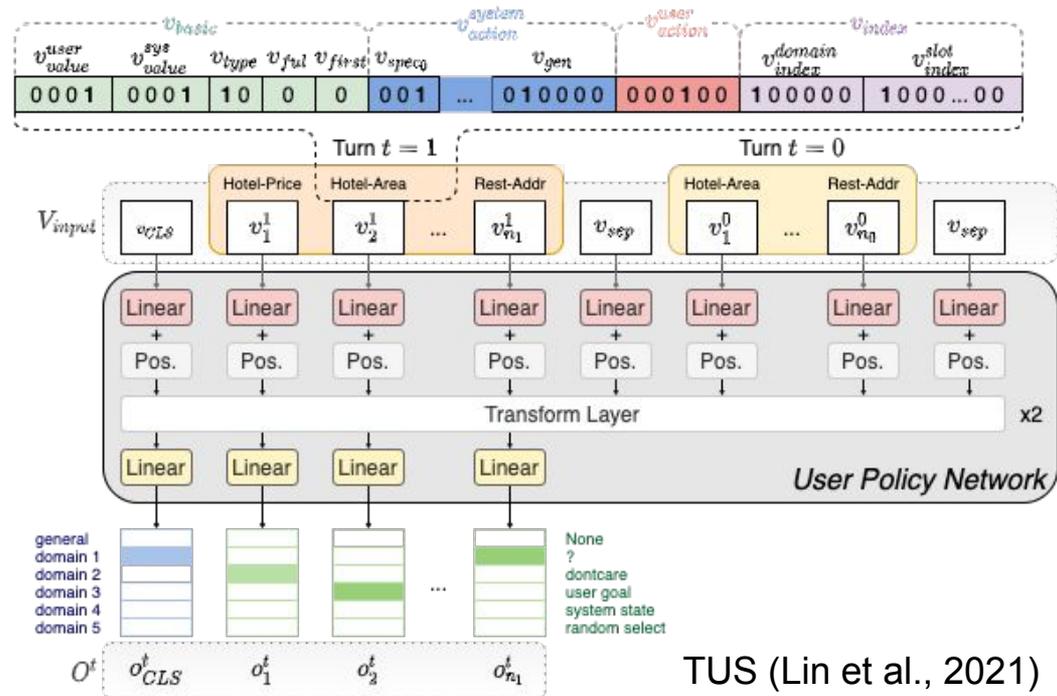
- Data-driven user simulator

- Pros

- Learn from data
- Domain independent (TUS)

- Cons

- Bias on the corpus
- Exploration limited by the corpus
- Zero-shot transfer is still challenging



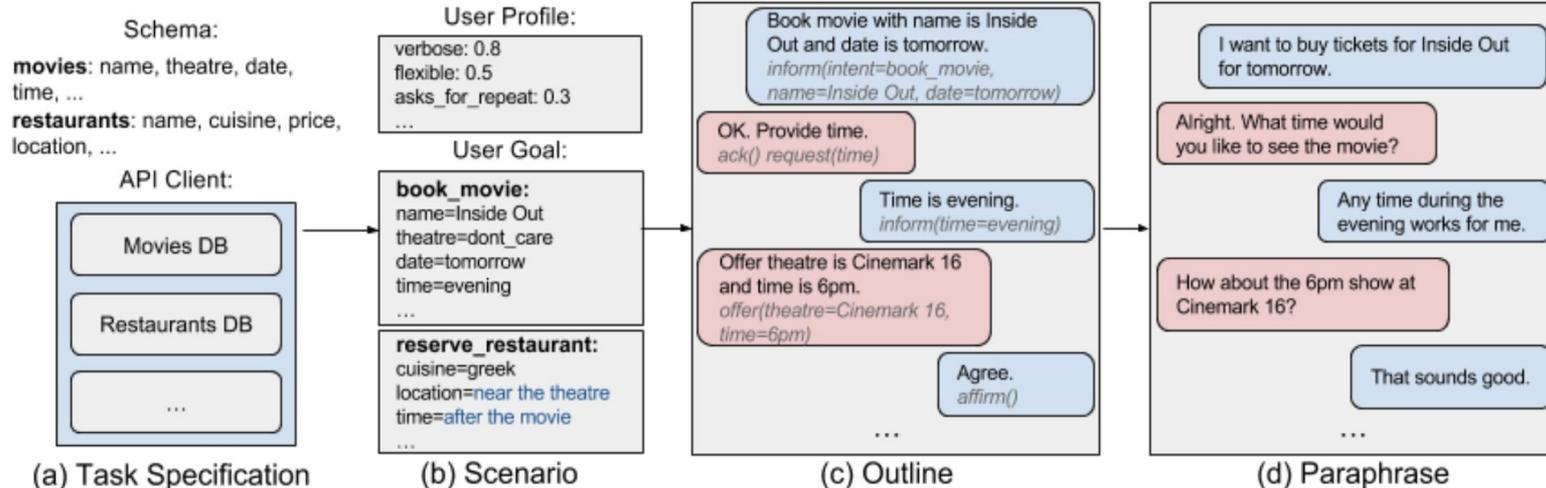
TUS (Lin et al., 2021)

- Data-driven user simulator
- Pros
 - Domain independent (TUS)
 - Learn from data
- Cons
 - Bias on the corpus
 - Exploration limited by the corpus
 - Zero-shot transfer is still challenging

US for training	removed data(%)	ABUS						TUS						mean
		Attr.	Hotel	Rest.	Taxi	Train	all	Attr.	Hotel	Rest.	Taxi	Train	all	
TUS-noAttr	32.20	0.69	0.64	0.81	0.65	0.75	0.77	0.71	0.58	0.66	0.61	0.69	0.69	0.73
TUS-noTaxi	19.60	0.63	0.61	0.81	0.61	0.70	0.74	0.69	0.60	0.69	0.64	0.68	0.69	0.72
TUS-noRest	45.21	0.62	0.66	0.80	0.56	0.75	0.76	0.71	0.60	0.64	0.65	0.64	0.68	0.72
TUS-noTrain	36.95	0.64	0.65	0.78	0.67	0.62	0.73	0.67	0.54	0.63	0.64	0.58	0.64	0.68
TUS-noHotel	40.15	0.59	0.59	0.76	0.61	0.54	0.69	0.64	0.52	0.61	0.61	0.55	0.62	0.66
TUS	0	0.69	0.68	0.81	0.66	0.77	0.79	0.73	0.59	0.66	0.68	0.64	0.69	0.74

- To collect data for a new domain...
- Problems of Wizard-of-Oz setup
 - might not cover all the expected interactions
 - might contain dialogues unfit for use (too simple or too complex)
 - annotation errors
- Including automation and crowdsourcing to collect datasets

Generating datasets through self-play



machine-to-machine (Shah et al. 2018)

- Crowd workers are asked to rewrite the machine-generated conversations
- They are encouraged to use linguistic phenomena like coreference (“Reserve that restaurant”) and lexical entrainment
- Second round of crowdsourcing for validation the annotation and utterances

- Crowd workers are asked to rewrite the machine-generated conversations
- They are encouraged to use linguistic phenomena like coreference (“Reserve that restaurant”) and lexical entrainment
- Second round of crowdsourcing for validation the annotation and utterances

	Machine-to-Machine	Wizard-of-Oz
conversation policy	based on models	real users
annotation	easy	expensive
coverage	controllable	interesting

MultiWOZ 2.0

u: I am looking to book a train that is leaving from Cambridge to Bishops Stortford on Friday.

s: There are a number of trains leaving throughout the day. What time would you like to travel?

u: I want to get there by 19:45 at the latest.

s: Okay! The latest train you can take leaves at 17:29, and arrives by 18:07. Would you like for me to book that for you?

u: Yes please. I also need the travel time, departure time, and price.

...

Human trial

u: book a train

s: Where did you want to depart from?

u: from cambridge

s: There are 1414. The address is Parkside, Cambridge.

u: to birmingham new street

s: Can you tell me what day you would like to travel, please? When would you like to travel? I have 133 trains that meet your criteria. The address is Parkside, Cambridge .

u: please book for friday

...

- Human evaluation
 - Time-consuming and costly (500 dialogue/day)
 - coverage

- Human evaluation
 - Time-consuming and costly
 - coverage
- Self-play evaluation
 - The user model is far from the real users

- Human evaluation
 - Time-consuming and costly
 - coverage
- Self-play evaluation
 - The user model is far from the real users
- How to speed up the evaluation?

Self-play for chit chat bots

- The chatbot can talk to itself
- Automatic matrix (Ghandeharioun et. al. 2019)
 - perplexity
 - embedding matrix
 - KL-Divergence between the posterior and the prior distribution.
- The dialogue can also be rated by crowd workers
 - engagingness
 - interestingness
 - knowledge
 - humanness

Hi how are you today

Hi, I'm pretty good! Just listening to some aerosmith, they're my fave :) whatre you up to?

I am listening to some italian music

Italian music, nice! What do you do for work?

I'm a professional wrestler

Oh wow! I'm a construction worker, I've been doing it since I was 16.

What do you do for a living

I'm a construction worker haha, just told you! Whats your favorite type of food?

My favorite is garlic bread

Yum! I love mexican food, so good :) do you have any hobbies?

I like to workout and go to the gym

We're a bit different- I love watching nascar and ufc. They're so fun!

Hello there, how are you?

I am doing great. How are you?

I am great, I did something crazy for me and colored my hair blue!

I have a daughter and a son who also love blue colored balls. You should meet them

Well that neat, I got a new car my mother gave so maybe I could see them!

It is a beautiful city. And, I try to be... Just cannot afford a bigger house atm.

I am sorry to hear that, I feel bad going out of town for spring break now.

Ok. I going to school in the spring for casino manager

Well I turn 29 next week, I wonder if that is a good age to apply as one.

My grandmother just died from lung cancer, sucks

Who would you prefer to talk to for a long conversation?

I would prefer to talk to **Speaker 1** I would prefer to talk to **Speaker 2**

Please provide a brief justification for your choice (a few words or a sentence)

Please enter here...

ACUTE-EVAL (Li et al. 2019)

- With complete control of synthetic data generated by dialogue self-play, we can generate unseen patterns in the test set (Majumdar et al. 2019)

- With complete control of synthetic data generated by dialogue self-play, we can generate unseen patterns in the test set (Majumdar et al. 2019)
- Out of template

- With complete control of synthetic data generated by dialogue self-play, we can generate unseen patterns in the test set (Majumdar et al. 2019)
- Out of template
- Out of Pattern
 - Turn compression
 - New api
 - Reordering the slot filling
 - Another Slot (irrelevant slots)
 - Audit more: When the system requests a new slot, the user provides another slot-value pair in addition

■ Experiment

- Single End-to-End Memory Network (SMN) (Bordes et al. (2017))
- Multiple End-to-End Memory Network (MMN)

Test Case	IT		OOT	
	<i>SMN</i>	<i>MMN</i>	<i>SMN</i>	<i>MMN</i>
Non OOP	88.62	90.17	87.39	88.27
Turn Comp.	27.80	55.00	27.90	54.70
New API	7.42	7.83	8.17	6.67
Reordering	54.50	45.50	54.00	41.50
Another Slot	38.00	25.00	41.50	27.50
Audit More	15.50	34.00	16.00	35.00
OOP Avg.	28.64	33.47	29.51	33.07

- Liu, Lane, 2017: Iterative Policy Learning in End-to-End trainable task-oriented neural dialog models
- Papangelis et. al., 2019: Collaborative Multi-Agent Dialogue Model Training via Reinforcement Learning
- Takanobu et. al., 2020: Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition
- Kottur et. al., 2017: Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog
- Lazaridou et. al., 2020: Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning
- Baker et. al., 2020: Emergent Tool Use from Multi-Agent Autocurricula
- Shah et. al. 2018: Building a conversational agent overnight with dialogue self-play
- Ghandeharioun et al. 2019: Approximating interactive human evaluation with self-play for open-domain dialog systems
- Li et al. 2019: Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons
- Majumdar et al. 2019: Generating Challenge Datasets for Task-Oriented Conversational Agents through Self-Play