香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Transfer Reinforcement Learning for Task-oriented Dialogue Systems

PhD Thesis Defense

Kaixiang Mo

Supervisor: Professor Qiang Yang

Committee: Prof. Pascale Fung (ECE), Prof. Mei Ling Meng (Sys Engg & Engg Mgmt, CUHK),

Prof. Lei Chen, Prof. Xiaojuan M., Chairman: Prof. Fugee Tsung.

Department of Computer Science and Engineering, 23rd August, 2018

Outline

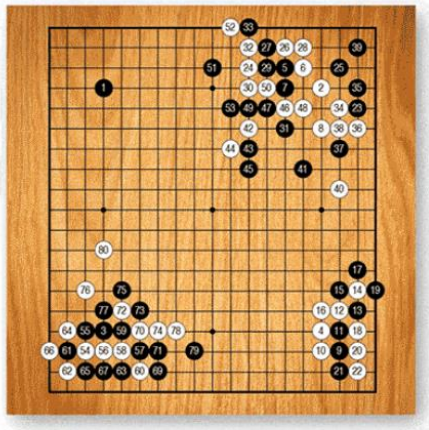
- Background
- Transfer Reinforcement Learning Problem
- Summarization of Related Work
- Transfer Reinforcement Learning through Personalized Q-function
- Transfer Reinforcement Learning through Personal Word Gating
- Transfer Reinforcement Learning through Speech-Act and Slot alignment
- The Unified Model
- Conclusion

Outline

- Background
- Transfer Reinforcement Learning Problem
- Summarization of Related Work
- Transfer Reinforcement Learning through Personalized Q-function
- Transfer Reinforcement Learning through Personal Word Gating
- Transfer Reinforcement Learning through Speech-Act and Slot alignment
- The Unified Model
- Conclusion

Background

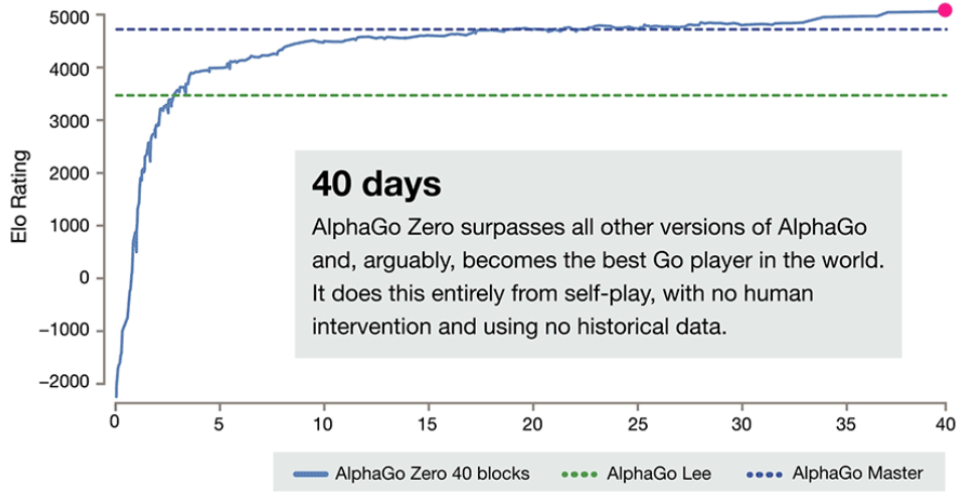
- Reinforcement Learning
  - Alpha Go Zeros can beat the most skillful human player in the world
  - Require a predefined state space
  - Require a large number of self-play training data



Captured Stones

70 hours

AlphaGo Zero plays at super-human level. The game is disciplined and involves multiple challenges across the board.



40 days

AlphaGo Zero surpasses all other versions of AlphaGo and, arguably, becomes the best Go player in the world. It does this entirely from self-play, with no human intervention and using no historical data.

Background

- Transfer Learning
  - Transfer knowledge to target domain with limited labelled data
  - Can transfer from source domain with different distributions
  - Do not consider states

Notation

Notation	Description	Notation	Description
Data			
$\mathcal{D}^s = \{\{X_i^s, Y_i^s, r_i^s\}_{i=1}\}$	Source Dialogues	$\mathcal{D}^t = \{\{X_i^t, Y_i^t, r_i^t\}_{i=1}\}$	Target Dialogues
X^s	Source User Utterance	X^t	Target User Utterance
Y^s	Source Agent Reply	Y^t	Target Agent Reply
$H_i^s = \{\{X_j^s, Y_j^s\}_{j=1}^{i-1}, X_i^s\}$	Dialogue Context/State for the i-th turn in Source	$H_i^t = \{\{X_j^t, Y_j^t\}_{j=1}^{i-1}, X_i^t\}$	Dialogue Context/State for the i-th turn in Target
A^s	Source Speech-act	A^t	Target Speech-act
S^s	Source Slot Set	S^t	Target Slot Set
V^s	Source Slot Value Set	V^t	Target Slot Value Set
$\pi^s : H_i^s \mapsto Y_i^s$	Source Dialogue Policy	$\pi^t : H_i^t \mapsto Y_i^t$	Target Dialogue Policy
Distribution			
$p(V^s)$	Probability distribution of Source Slot Value Set	$p(V^t)$	Probability distribution of Target Slot Value Set

Task-oriented Dialogue System

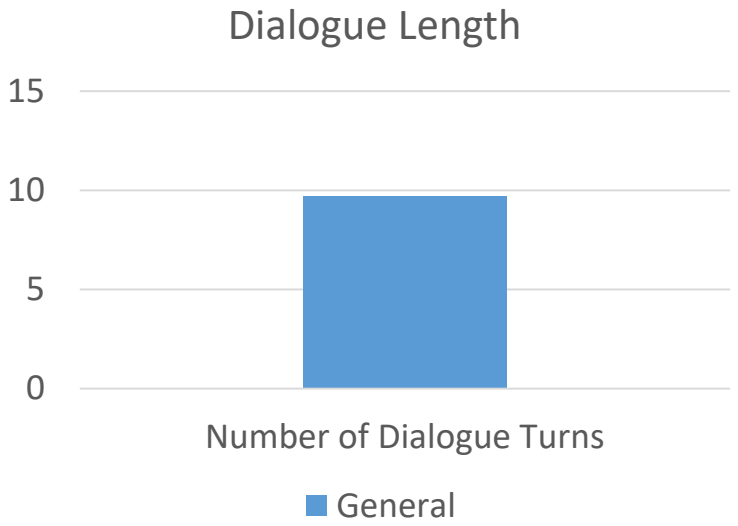
- Dialogue systems have 2 major categories
  - Open-domain dialogue systems are used for daily chatting, and no feedback is required.
  - Task-oriented dialogue systems help users to finish certain tasks via dialogues, **user feedback** is required.

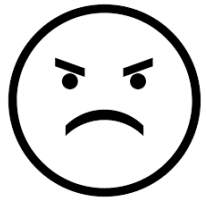
	Open-domain Dialogue
X_1	nothin much, and how's the book?!
Y_1	its good but i'm only like halfway through cuz i don't feel like reading. i'm so bored ...

	Task-oriented Dialogue
X_1	Can I have coffee please?
Y_1	What coffee would you like?
X_2	I would like a cup of Mocha.
Y_2	Hot Mocha or Iced Mocha?
X_3	Iced Mocha, with extra cream.
Y_3	What cup size do you want?
X_4	Tall.
Y_4	What is your address?
X_5	No.1199 Mingsheng Road.
Y_5	Please pay.
Y_6	Payment Completed.
R	<Task-Completion-Feedback>

Task-oriented Dialogue System: Challenges

- Cannot adapt to each user's preference.
- Dialogues are lengthy.
- Manual dialogue states design / require a lot of data from target user.



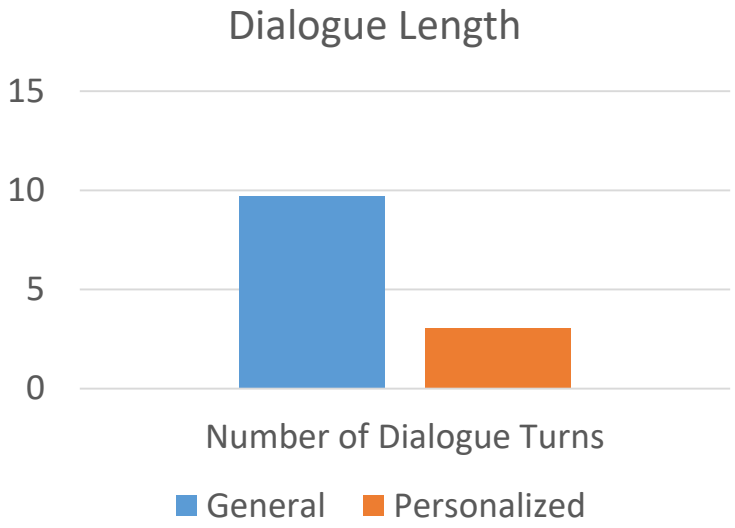


Customer in a hurry:
Don't you remember me?
Do I have to tell you every time?

	Coffee Shopping Dialogue (X=customer, Y=system)
X_1	Can I have coffee please?
Y_1	What coffee would you like?
X_2	I would like a cup of Mocha.
Y_2	Hot Mocha or Iced Mocha?
X_3	Iced Mocha.
Y_3	What cup size do you want?
X_4	Tall.
Y_4	What is your address?
X_5	No.1199 Mingsheng Road.
Y_5	Please pay.
Y_6	Payment Completed.
R	<Task-Completion-Feedback>

Transfer Reinforcement Learning (TRL) for Personalized Task-oriented Dialogue System

- Adapt to each user's preference.
- Reduce dialogue length.
- Learn dialogue states with source data and a few target data



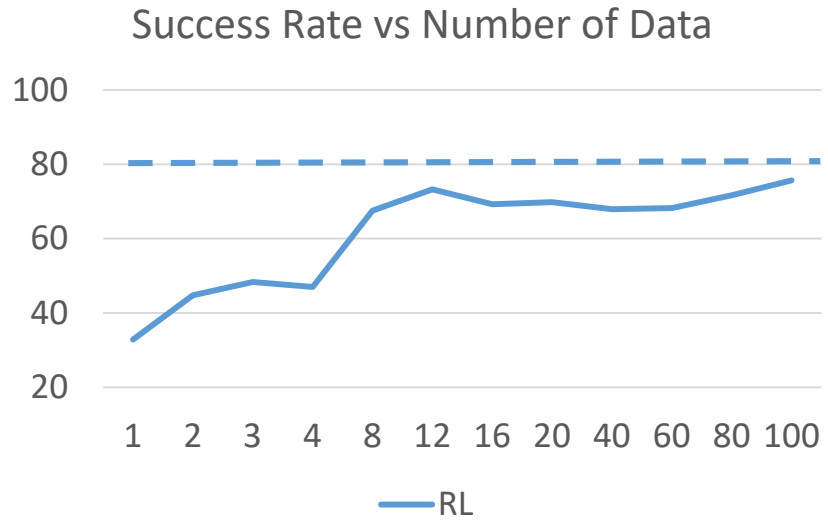
Customer in a hurry:
You can always remember!

	Personalized Coffee Shopping Dialogue (X=customer, Y=system)
X_1	Can I have coffee please?
Y_1	Same as before? Tall Iced Mocha and deliver to No.1199 Mingsheng Road?
X_2	Yes, as always.
Y_2	Please pay.
X_3	Payment completed.
R	<Task-Completion-Feedback>

Personalized dialogue is much shorter!

Task-oriented Dialogue System: Challenges

- Requires a lot of training data and outside feedback in a new domain
- Previous policy cannot be reused

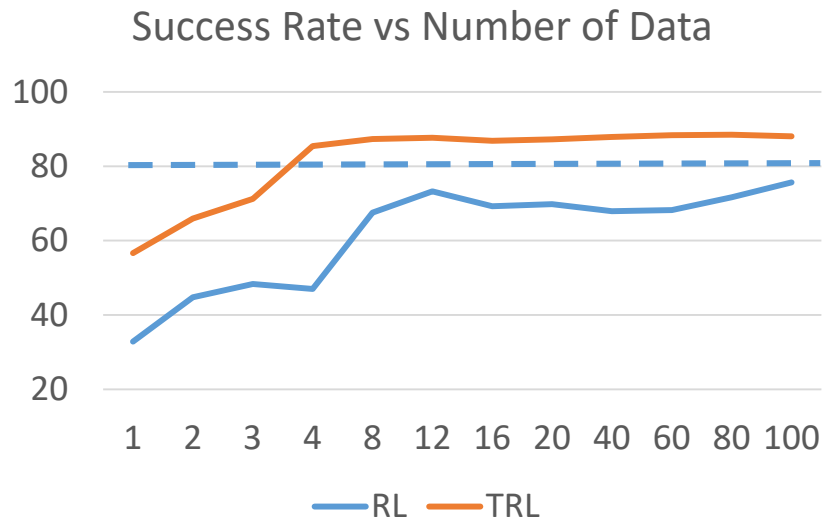


	Coffee Shopping Dialogue (X=customer, Y=system)
X_1	Can I have coffee please?
Y_1	What coffee would you like?
X_2	I would like a cup of Mocha.
Y_2	Hot Mocha or Iced Mocha?
X_3	Iced Mocha.
Y_3	<?>

Candidate Reply Set
Y_{c1}: * What cup size do you want?
Y_{c2}: What is your address?
Y_{c3}: Please pay.

Transfer Reinforcement Learning (TRL) for Cross-domain Dialogue Policy Transfer

- Less training data / Better performance
- Leverage previous policy



	Coffee Shopping Dialogue (X=customer, Y=system)
X_1	Can I have coffee please?
Y_1	What coffee would you like?
X_2	I would like a cup of Mocha.
Y_2	Hot Mocha or Iced Mocha?
X_3	Iced Mocha.

State: [Coffee, Temp]

Transfer Dialogue Policy from Source

Source: Restaurant
State: [food]
Action: request(area)

Action: request(size)

Y_3	**What cup size do you want?**

Target Q-function: 4.09

Outline

- Background
- Transfer Reinforcement Learning Problem
- Summarization of Related Work
- Transfer Reinforcement Learning through Personalized Q-function
- Transfer Reinforcement Learning through Personal Word Gating
- Transfer Reinforcement Learning through Speech-Act and Slot alignment
- The Unified Model
- Conclusion

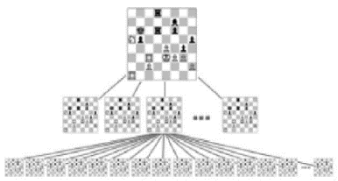
Transfer Reinforcement Learning Problem

- Reinforcement Learning
  - training and testing data has the same state space, action space, state transition function and reward function.
- Transfer Supervised Learning
  - training and testing data have different features/labels, but do not consider states.
- Transfer Reinforcement Learning
  - training and testing data can have different state space, action space, state transition function or reward function.

	No States	**Have States**
No Transfer	Supervised Learning	Reinforcement Learning
Transfer	Transfer Supervised Learning	Transfer Reinforcement Learning

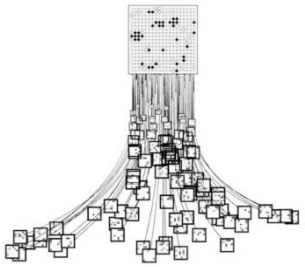
Transfer Reinforcement Learning **Examples**

- Policy Transfer from Chess to Go
  - Different **states**, E.g. different board.
  - Different **actions**, E.g. different moves.
  - Different **transitions**, E.g. pieces can move / stones in Go cannot move.
  - Different **reward** function, E.g. kill the rival King / occupy more space.



Which piece to move and where to move?

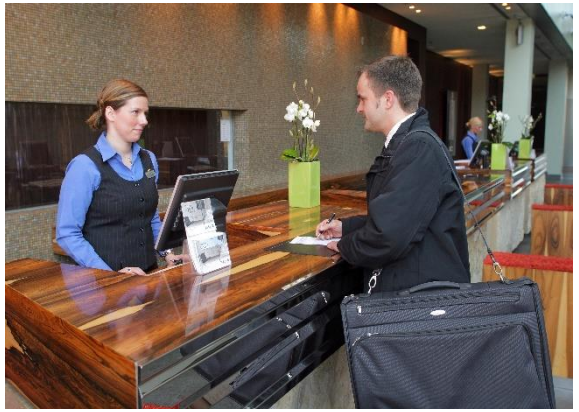
Transfer Learning



Where to put the next stone?

Transfer Reinforcement Learning Problem in Task-oriented dialogue system

- How to transfer dialogue policy across users, domains and systems?



Much data

I want to book a room. : User
Agent: When are you going to check in?
Tomorrow. : User
Agent: When are you leaving?
The day after tomorrow.: User

Transfer Learning

Little data

Can I have a cup of coffee? : User
Agent: What coffee would you like?
Latte Please.: User
Agent: Please pay.

Task-oriented Dialogue System: Basic Concepts

- Speech-Act A: User Intention (system-wise)
  - e.g. “request” in system A might be “ask” in system B
- Slots S: Information (domain-wise)
- Slot-value V: Choice (user-wise).

- I want a cup of latte.
- Inform(coffee=Latte)

Speech-Act	Slot	Slot-value
A	S	V
Intention	Information	Choice

Domain		
Speech-act A	Request Inform Bye Confirm	Ask Tell End MakeSure
Slots S	Coffee-type Temperature Size Address	Room-Type Area HasParking Stars PriceRange
Slot-value V	Latte Mocha Macchiatto	KingSize Single Room Double Bed

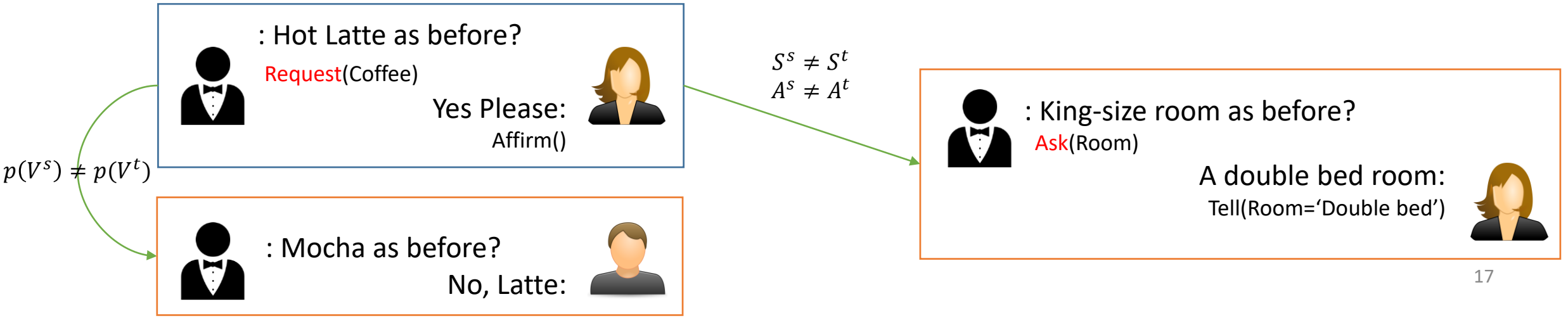
Challenges of TRL in Task-oriented dialogues

- Different person have different preferences
- Common knowledges mixed with personal information
- Different domains have different slots
- Different systems have different speech-acts

$$p(V^s) \neq p(V^t)$$

$$S^s \neq S^t$$

$$A^s \neq A^t$$



Types of TRL in Task-oriented dialogues

- Separating common dialogue states and personal preferences

$$Q^{\pi_u} = Q_g + Q_p$$

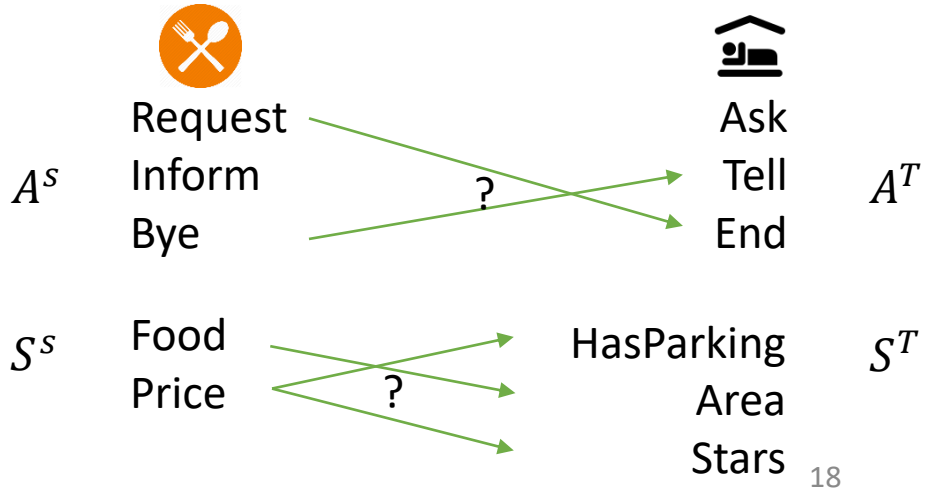
Common Personal

- Transfer end-to-end dialogue policy at finer-granularity

Still in the same old way, hot latte?

Common Personal

- Learning to map dialogue policy across domains with different slots and different speech-acts.



Unified Policy Transfer Framework

- The target domain Q-function can be represented by a function of state H^t and action Y^t as
- $Q(H^t, Y^t) =$

$$(1 - O^p(H^t))Q_g(f(H^t), f(Y^t)) + O^p(H^t)Q_p(H^t, Y^t)$$

Source Domain Q-function	Cross Domain Mapping function	Personal Gate	Personal Q-function

Note that mode fine-tuning is one of our special cases.

State1=[Food=?, Area=?, Price=?] → State1=[Type=?, Stars=?, Area=?, Park=?]

State2=[Food=Chinese, Area=? Price=?] → State2=[Type=Hotel, Stars=3, Area=?, Park=?]

Action=Request(Food) → Action=Request(Stars)

Dialogue Policy Transfer Examples

- Transfer Personalized Dialogue Policy across different users

X_1	Can I have coffee please?
Y_1	Same as before? **Medium Hot Latte** and deliver to **No.101 Shandong Road**?
X_2	Yes, as always. <Success-Feedback>

State=[X_1]
Action=Y_1 → State=[X_1, Y_1, X_2]
Action=Y_2 <Success>

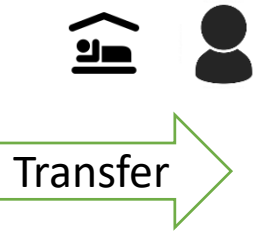
X_1	Can I have coffee please?
Y_1	Same as before? **Tall Iced Mocha** and deliver to **No.1199 Mingsheng Road**?

State=[X_1]
Action=Y_1

- Transfer across domains

X_1	I am looking for a restaurant.
Y_1	What food would you like?

State=[Food=?, Area=?, Price=?]
Action=Request(Food)



X_1	I want to find a hotel.
Y_1	How many stars should the hotel have?

State=[Type=?, Stars=?, Area=?, Park=?]
Action=Request(Stars)

Outline

- Background
- Transfer Reinforcement Learning Problem
- Summarization of Related Work
- Transfer Reinforcement Learning through Personalized Q-function
- Transfer Reinforcement Learning through Personal Word Gating
- Transfer Reinforcement Learning through Speech-Act and Slot alignment
- The Unified Model
- Conclusion

Summarization of Related Works

- Our works

	Same Domain $\mathbf{S^s = S^t}$		**Cross Domain $\mathbf{S^s \neq S^t}$**	
General $\mathbf{p(V^s) = p(V^t)}$ $\mathbf{A^s = A^t}$	None-transfer methods. (Williams 2008a,b; Young et al. 2010; Lefèvre et al. 2009; Li et al. 2009; Gašic and Young 2014; Gašic et al. 2013a; Daubigney et al. 2012b),		Same Speech-Act $\mathbf{A^s = A^t}$	Gaussian Process Transfer (Gašic et al. 2014; Gašic et al. 2013b; Gašic et al. 2015a) Bayesian Committee Machine Transfer (Gašic et al. 2015b)
Personal $\mathbf{p(V^s) \neq p(V^t)}$ $\mathbf{A^s = A^t}$	Fine-tuning	Linear Model Transfer (Genevay and Laroche 2016). Gaussian Process Transfer (Casanueva et al. 2015)	Different Speech-Act $\mathbf{A^s \neq A^t}$	Simultaneous Policy Transfer across Slots and Speech-acts Novelty: Learn cross domain slot mapping and speech-act mapping without external database or human knowledge.
	Personal Q-fun	(Mo et al. 2016) Novelty: Model Q-function with a general and a personal part.		
	Personal Word Gate	Novelty: First Phrase-Level Transfer Method.		

Mo, Kaixiang, et al. "Personalizing a Dialogue System with Transfer Reinforcement Learning." *AAAI-2018*.

Related Works: Datasets and Evaluation Metrics

- Datasets
  - User Simulators
  - Real-User records
- Evaluation Metrics
  - Live Evaluation
    - Averaged Reward
    - Success Rate
    - Number of Dialogue Turns
  - Static Evaluation
    - AUC
    - BLEU (for end-to-end dialogue policy, for each dialogue turn)

Dataset	Task	Evaluation Metrics
TownInfo (Thomson and Young 2010)	Policy	Reward, Success Rate, Number of Dialogue Turns
Cambridge Restaurant (Gašic and Young 2014)	Policy, NLG	Reward, Success Rate, Number of Dialogue Turns, BLEU
SF Restaurant (Gašic et al. 2015b)	Policy, NLG	Reward, Success Rate, Number of Dialogue Turns, BLEU
SF Hotel (Gašic et al. 2015b)	Policy, NLG	Reward, Success Rate, Number of Dialogue Turns, BLEU
L11 (Gašic et al. 2015b)	Policy, NLG	Reward, Success Rate, Number of Dialogue Turns, BLEU
Babi Task (Bordes et al. 2016)	NLG	BLEU

Datasets used for dialogue policy learning are listed,
and the full dataset table for all dialogue tasks is in the appendix.

Outline

- Background
- Transfer Reinforcement Learning Problem
- Summarization of Related Work
- Transfer Reinforcement Learning through Personalized Q-function
- Transfer Reinforcement Learning through Personal Word Gating
- Transfer Reinforcement Learning through Speech-Act and Slot alignment
- The Unified Model
- Conclusion

Personalized Q-function Problem Example

- How to transfer personalized policy across users with different preferences?

Source Domain  

	Alice's 21st Coffee Shopping Dialogue
X_1	Can I have coffee please?
Y_1	What coffee would you like?
X_2	I would like a cup of **Latte**.
Y_2	**Hot Latte** deliver to **No.101 Shandong** Road?
X_3	Yes, exactly!

Alice's Dialogues 

Bob's Dialogues 

...

Transfer

Target Domain  

	John's 3rd Coffee Shopping Dialogue
X_1	Can I have coffee please?
Y_1	What coffee would you like?
X_2	I would like a cup of **Mocha**.
Y_2	

Candidate Reply Set

Y_{c1}: **Cold Mocha** deliver to **No.1199 Mingsheng** Road?

Y_{c2}: What is your address?

Y_{c3}: Hot Mocha or Iced Mocha?

Learning Common Dialogue States and Personalized Q-function

- How to transfer personalized policy across users with different preferences?
- Input:
  - Abundant historical dialog sessions $\{\{X_i^{u_s}, Y_i^{u_s}\}\}$ from the source users u_s, where X_i^u is the user utterance of user u and Y_i^u is corresponding the system reply in the i-th turn.
  - A few historical dialog sessions $\{\{X_i^{u_t}, Y_i^{u_t}\}\}$ of a target user u_t.
- Output:
  - Dialogue Policy π^{u_t} for the target user that guide the target user to finish a task as soon as possible?

Many source user dialogues with different preference.
[Latte, Macchiato, …]

Transfer

Target users with limited data
[Mocha]

Learning Common Dialogue States and Personalized Q-function

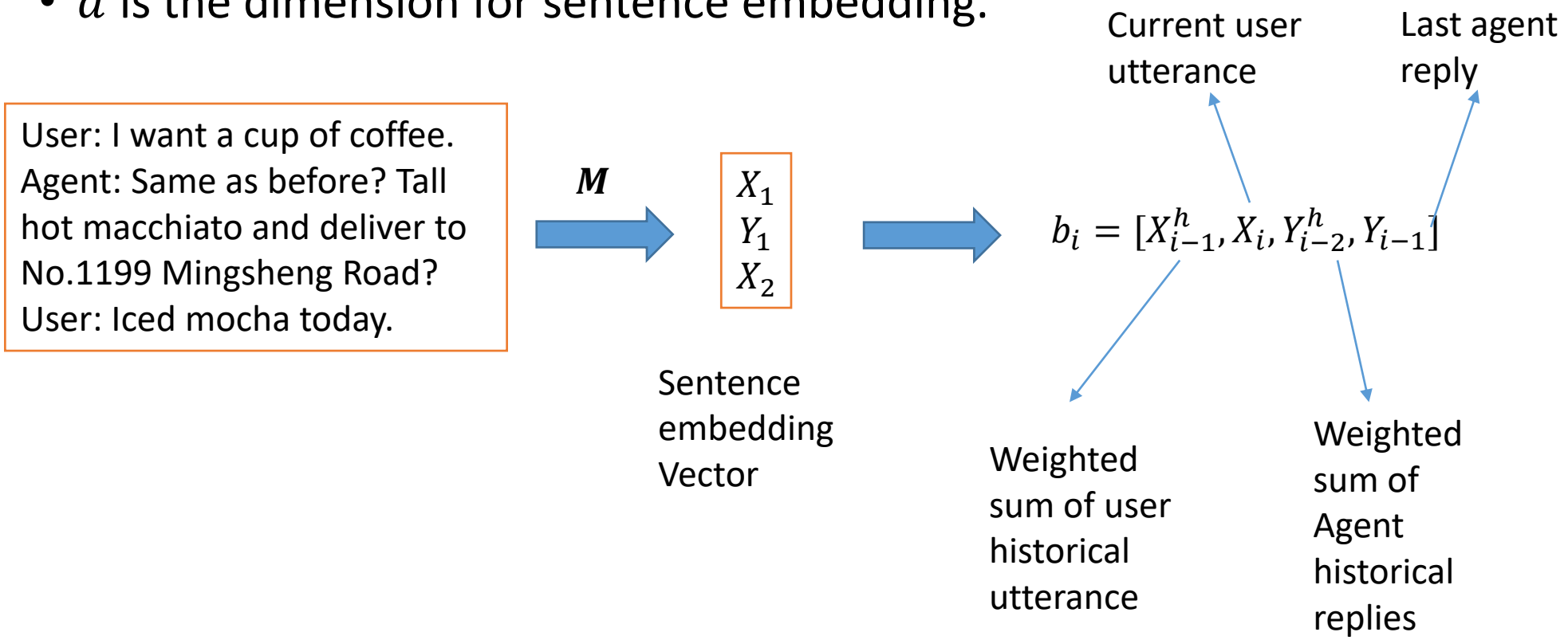
- The common dialogue states are learned in source domain
  - Belief state vector $b_i = f(H_i; \boldsymbol{M})$, where dialogue history $H_i = \{\{X_j, Y_j\}_{j=1}^{i-1}, X_i\}$
- The personalized Q-function
  - $\mathrm{Q}^{\pi_u}(H_i, Y_i|\Theta) = Q_g(H_i, Y_i|\Theta^g) + Q_p\left(H_i, Y_i\middle|\Theta_u^p\right)$

    General part Personal part

  - General part $Q_g(H_i, Y_i|\Theta^g)$ for the future general reward.
  - Personal part $Q_p\left(H_i, Y_i\middle|\Theta_u^p\right)$ for the future personal reward.
  - Only sum of general and personal reward can be observed during training.
- The dialogue states are learn jointly with the Q-function.

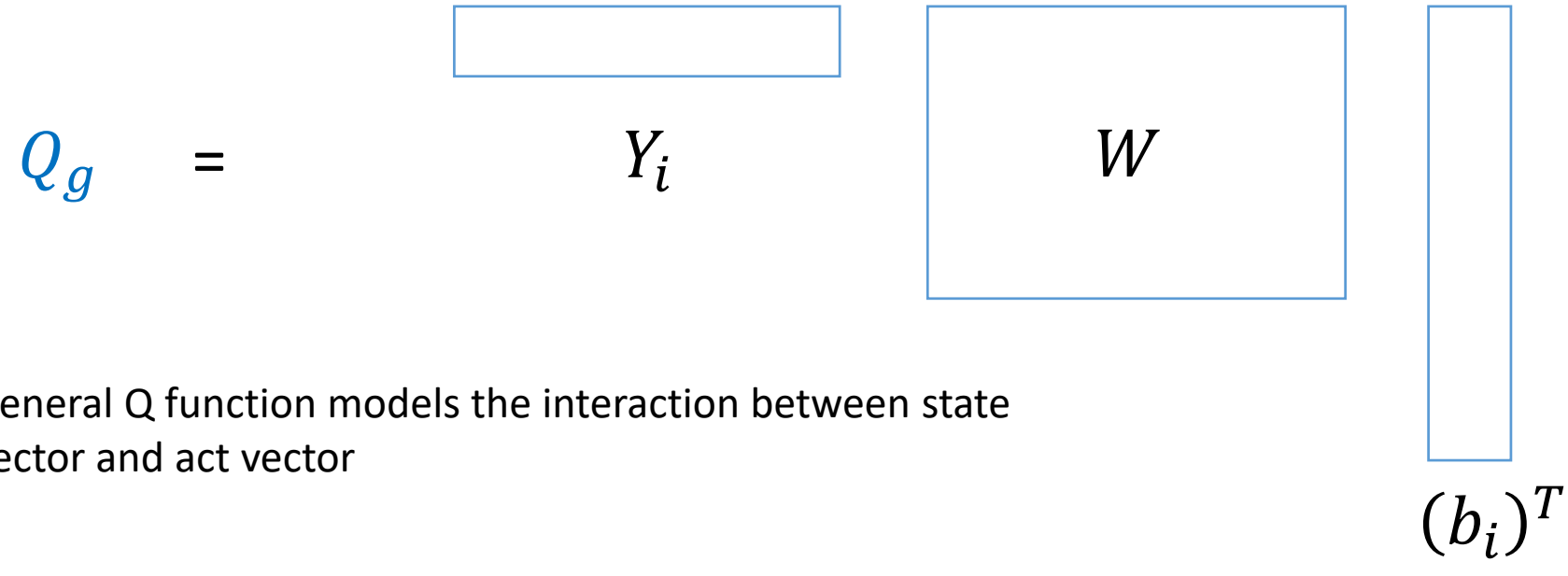
Learning Common Dialogue States and Personalized Q-function

- Common Belief State model Belief state vector $b_i = f(\mathcal{H}_i; \boldsymbol{M})$
  - $M \in \mathbb{R}^{v \times d}$ is the state projection matrix
  - v is the vocabulary size
  - d is the dimension for sentence embedding.



Learning Common Dialogue States and Personalized Q-function

- $\mathrm{Q}^{\pi_u}(H_i, Y_i|\Theta) = Q_g(H_i, Y_i|\Theta^g) + Q_p\left(H_i, Y_i\middle|\Theta_u^p\right)$
- $Q_g(H_i, Y_i|\Theta^g) = Y_i W (b_i)^T$, where $W \in \mathbb{R}^{d*4d}$, b_i is the current belief state defined in the previous slide



General Q function models the interaction between state vector and act vector

Learning Common Dialogue States and Personalized Q-function

- Personal part of the Q-function

Categorical distribution with $p_u =$
$\{s_0: \{Mocha: 0.5, latte: 0.5\},\ s_1: \{hot: 0.1, cold: 0.9\}, \ldots\}$

s_{ij}^u is the slot value of slot j in Y_i^u
s_{i0}^u: Mocha
s_{i1}^u: Cold

- $Q_p\left(H_i^u, Y_i^u \middle| p_u, w_p\right) = w_p \sum_{j=1} f\left(s_{ij}^u; p_u\right) \delta(S_j, H_i^u)$ =1.5*(0.5*1.0+0.9*1.0)=2.1

User: I want a cup of coffee.	Q_p
Reply Candidates	**Personal Q-fun**
In the same old ways? **Cold Mocha**? (2 suitable values detected for the target user)	+2.1

Personal weight learned in target domain training data.
Larger value means the user has stable preference.

S_j is the j-th slot
S_0: Coffee
S_1: Temperature

Is this info type S_j first mentioned in this system reply?

$\delta(S_0, H_i^u) = 1.0, \delta(S_1, H_i^u) = 1.0$ since it is the first recommendation of coffee type and temperature.

Learning Common Dialogue States and Personalized Q-function

- Q-learning
  - $L(\Theta) = E[\left(r_i + \lambda max_{\mathcal{A}_{i+1}} Q(H_{i+1}, Y_{i+1}|\Theta) - Q(H_i, Y_i|\Theta)\right)^2]$
- Algorithm
  - Step1: Learning common dialogue states and Q-function from source domain
    - Learn the $Q_g(H_i, Y_i|\Theta^g)$ + $Q_p\left(H_i, Y_i\middle|\Theta_u^p\right)$ in **source** domain, where Q_g is retained and Q_p is discarded.
  - Step2: Learn personalized Q-function for target user
    - $Q_g(\mathcal{H}_i, \mathcal{A}_i|\Theta^g)$ is transferred from **source** domain
    - The personal Q-function part $Q_p\left(\mathcal{H}_i, \mathcal{A}_i\middle|\Theta_u^p\right)$ is learned from **target** domain

Baselines

	Source	User Similarity	Way to transfer	Has Personal Q-function
NoTL	-	-	-	No
Sim (Casanueva et al. 2015)	The most similar user	Predefined	Pre-train on source	No
Bandit (Genevay and Laroche 2016)	The most similar user	Bandit	Pre-train on source	No
PriorSim (Gašic et al. 2013)	The most similar user	Predefined	Use source as Prior	No
PriorAll (Gašic et al. 2013)	All source users	-	Use source as Prior	No
All	All source users	-	Pre-train on source	No
PETAL (Proposed)	All source users	-	Pre-train on source	Yes

Our work:

Real-world Experiment

- Evaluation: AUC (Williams and Zweig 2016)
  - For each dialogue turn
    - The ground truth human reply, label=1.
    - Randomly sample 10 replies, label=0.
    - The dialogue policy will output a score for each candidate.
    - The AUC for this turn is calculated.
  - The averaged AUC is calculated for all dialogue turns, with 5 random seeds.

Dialogue history H_2

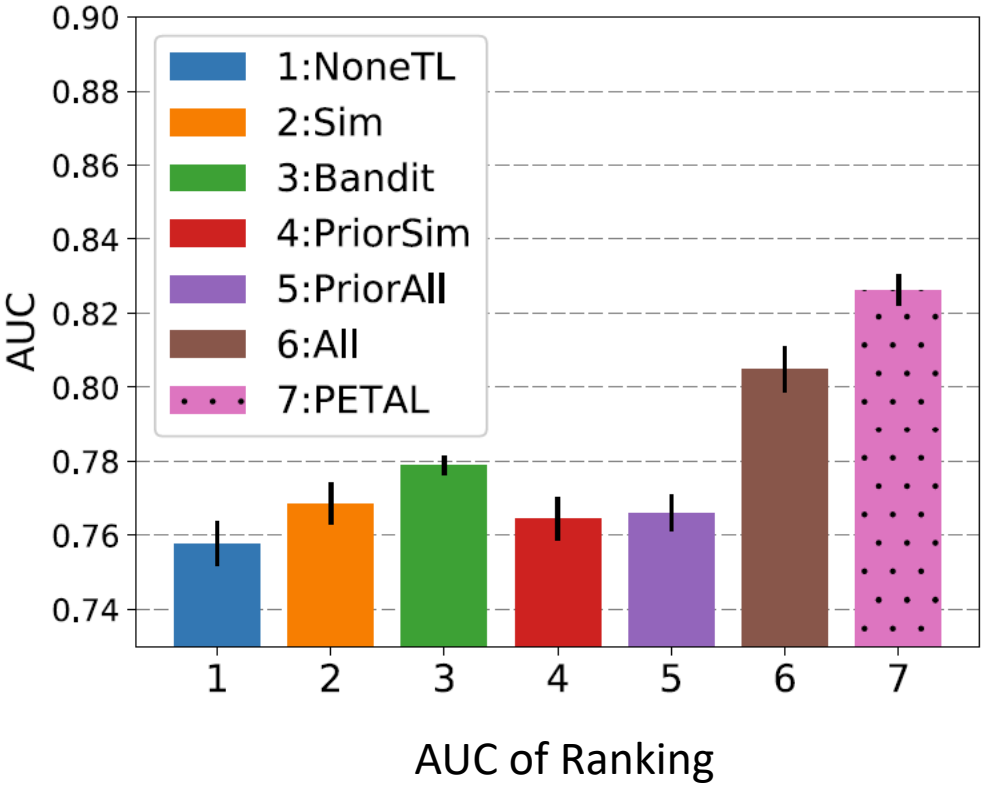
X_1:需要一杯星巴克中杯拿铁
Y_1:**拿铁(中杯)**；请问冰的还是热的？
送到 **海淀大街34号海置创投大厦5层**
对吧，配送费每单5元起
X_2 :现在有配送费啦？$[r_1]$

Label	Candidate Reply Set
1	Y_{c1}: * 一直都有呀 (Ground Truth Human Reply1)
0	Y_{c2}: 请问想喝什么？ (Randomly sampled Reply2)
0	Y_{c3}: 拿铁没有冷的。 (Randomly sampled Reply3)

[Williams and Zweig 2016] Williams, J. D., and Zweig, G.2016. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. arXiv preprint arXiv:1606.01269

Real-world Experiment

- Setting: Coffee ordering
  - Collected in O2O company, between real customers and human personal assistants.
  - 52 source users and 20 target users, 2000+ multi-turn dialogues.
  - Evaluation: AUC



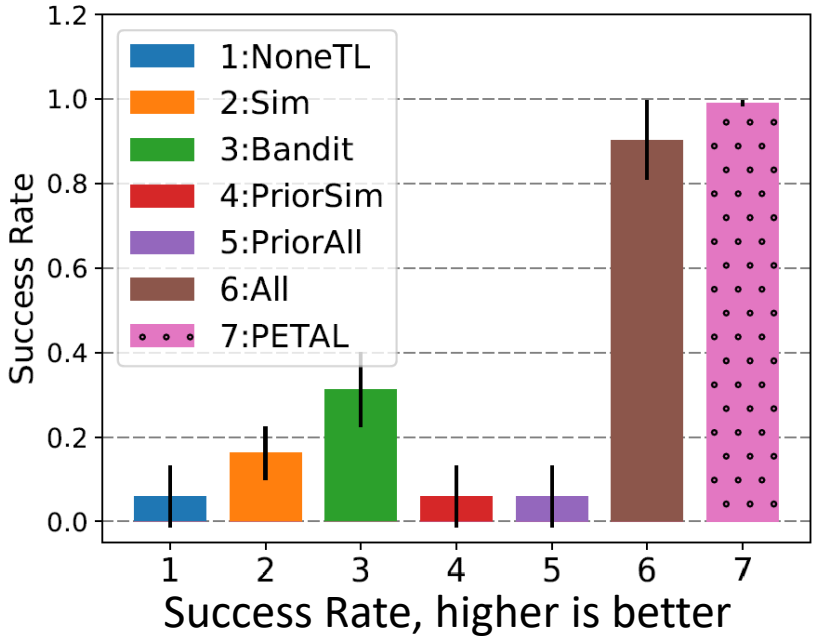
AUC of Ranking

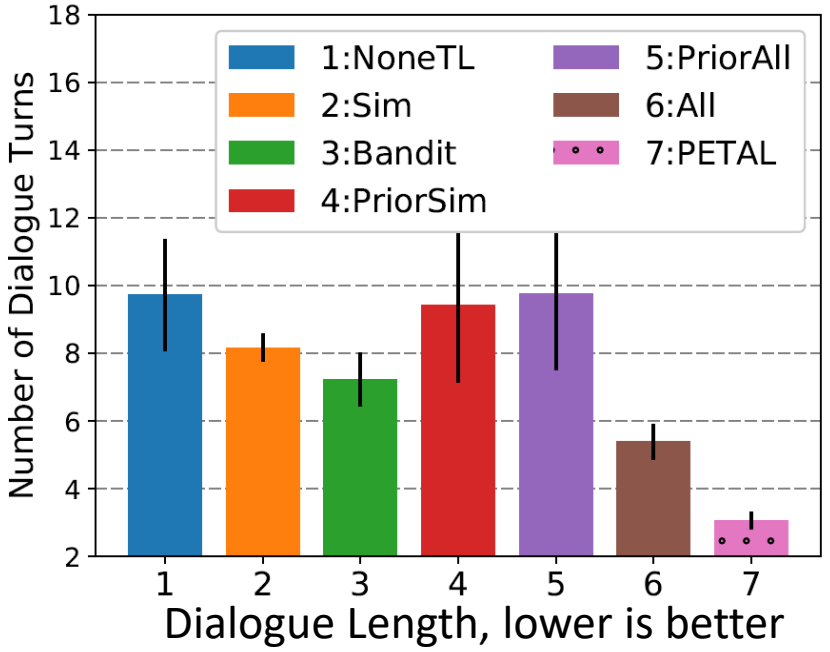
	Source Domain		Target Domain	
	Users	Dialogues	Users	Dialogues
Real Data	52	1859	20	329
Simulation	11	176000	5	100

User utterance :	I want a cup of coffee.
All	PETAL	Response Candidates
0.86	1.36	* Same as before? Tall hot americano and deliver to Central Conservatory of Music?
0.99	0.92	All right, deliver to No.1199 Beiyuan Road, Chaoyang District, Beijing?
0.72	0.69	What's your address?

Simulation Experiment

- User Simulator
  - Order coffee according to its preference
- Evaluation: Averaged Reward, Success Rate and Dialogue Length
  - Source: Model is trained with 176000 dialogues from 11 source user simulators with/without preference
  - Adapt: adapt on total 100 dialogues from 5 target test user.
  - Test: 5 target users with fixed preference will order coffee for 1000 times.
  - Success rate is the probability of the agent finishing the dialogue within 20 dialogue turns.





Simulation Case Study

Baseline

User:	I want a cup of coffee.
Agent:	Hot coffee?
User:	Yes.
Agent:	What would you like to drink?
User:	Hot macchiato, please.
Agent:	What's your address?
User:	No.1199 Minsheng Road, Pudong District, Shanghai.
Agent:	Alright. Tall, Grande, or Venti?
User:	Tall.
Agent:	Please pay.
User:	Payment completed.

PETAL: Personalized POMDP

User:	I want a cup of coffee.
Agent:	Same as before? Tall hot macchiato and deliver to No.1199 Minsheng Road, Pudong District, Shanghai?
User:	I want iced mocha today.
Agent:	Sure, please pay.
User:	Payment completed.

Reduced Length!

Outline

- Background
- Transfer Reinforcement Learning Problem
- Summarization of Related Work
- Transfer Reinforcement Learning through Personalized Q-function
- Transfer Reinforcement Learning through Personal Word Gating
- Transfer Reinforcement Learning through Speech-Act and Slot alignment
- The Unified Model
- Conclusion

Personal Word Gating Problem Example

- How to transfer only common dialogue knowledges?

Source Domain

	Alice's 21st Coffee Shopping Dialogue
X_1	Can I have coffee please?
Y_1	What coffee would you like?
X_2	I would like a cup of **Latte**.
Y_2	**Hot Latte** deliver to **No.101 Shandong** Road?
X_3	Yes, exactly!

Alice's Dialogues

Bob's Dialogues

...

Transfer

Target Domain

	John's 3rd Coffee Shopping Dialogue
X_1	Can I have coffee please?
Y_1	What coffee would you like?
X_2	I would like a cup of **Mocha**.
Y_2	?

Candidate Reply Generation

Y_2: **Cold Mocha** deliver to <word5> <word6> ...

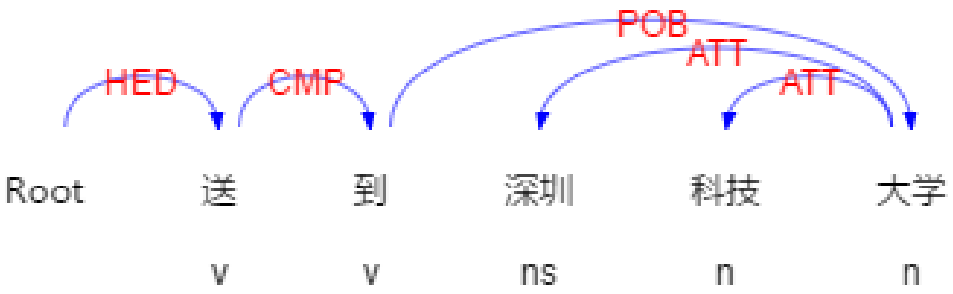
? ?

Fine-grain Transfer Learning with Personal Word Gating

- How to transfer an end-to-end personalized dialogue policy and avoid negative transfer due to personal information leakage?
- Input:
  - Historical dialog sessions $\{\{X_n^u, Y_n^u, r_n^u\}\}$ of each user.
  - Personal word labels O_n^u for each Y_n^u
    - Still deliver to **Kingdee Building**, right? O_n^u=[0 0 0 1 1 0]
- Output:
  - Personalized End-to-end dialogue policy $Y_n^u = f(X_n^u, \{X_i^u, Y_i^u\}_{i=0}^{n-1})$ for each user u.

Fine-grain Transfer Learning with Personal Word Gating

- Personal Word Label
  - Well defined as words related to all possible choices in the domain slots.
  - Choices related to coffee-type, temperature, cup size, address
- Annotation
  - Keyword matching
    - For enumerable slot-value
  - Semantic tree parser
    - For slot-value that has patterns
  - Human checking



Fine-grain Transfer Learning with Personal Word Gating

- The dialogue policy generates word Y_t for user u:
- $Q(H_t, Y_t) =$

$$(1 - O_t^u)Q_g(H_t, Y_t|\Theta^g) + O_t^u Q_p(H_t, Y_t|\Theta_u^p)$$

Shared RNN

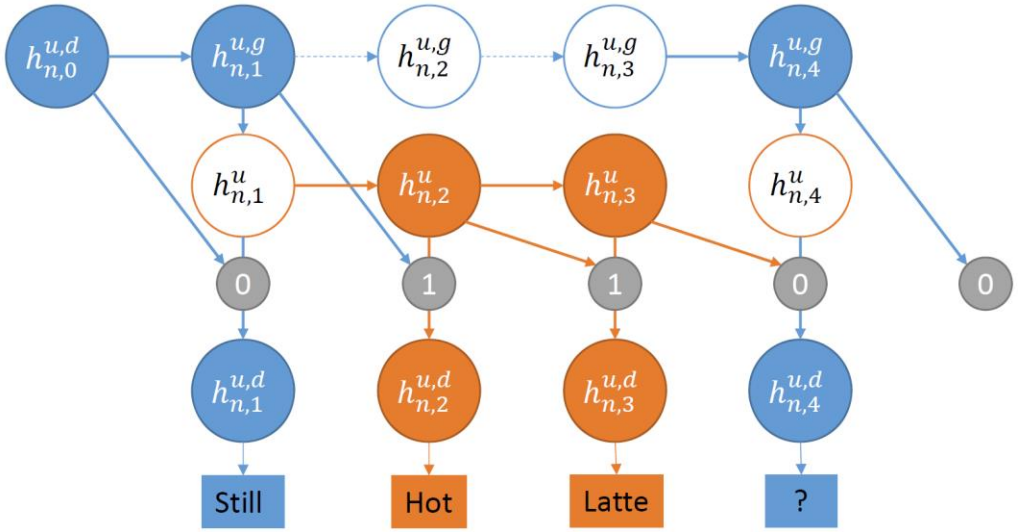
Personal Gate

Personal RNN

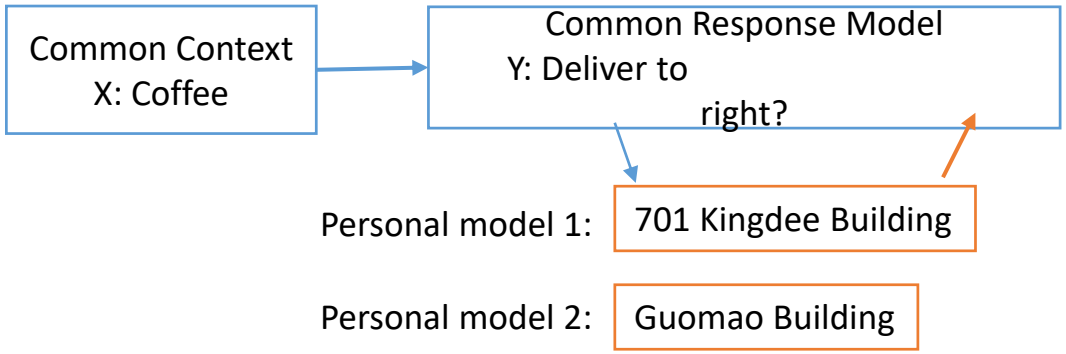
- $O_t^u = g(H_{t-1}, Y_{t-1}, O_{t-1}^u)$ is the **binary** personal word gate.
- $H_t = \{h^g, h^u\}$ is the hidden states of the RNN models.

Fine-grain Transfer Learning with Personal Word Gating

- Transfer fine-granularity **phrase-level** knowledge.
- The personal word gate selects different model for different words.
- Easily combined with various dialogue models, e.g. S2S, HRED.



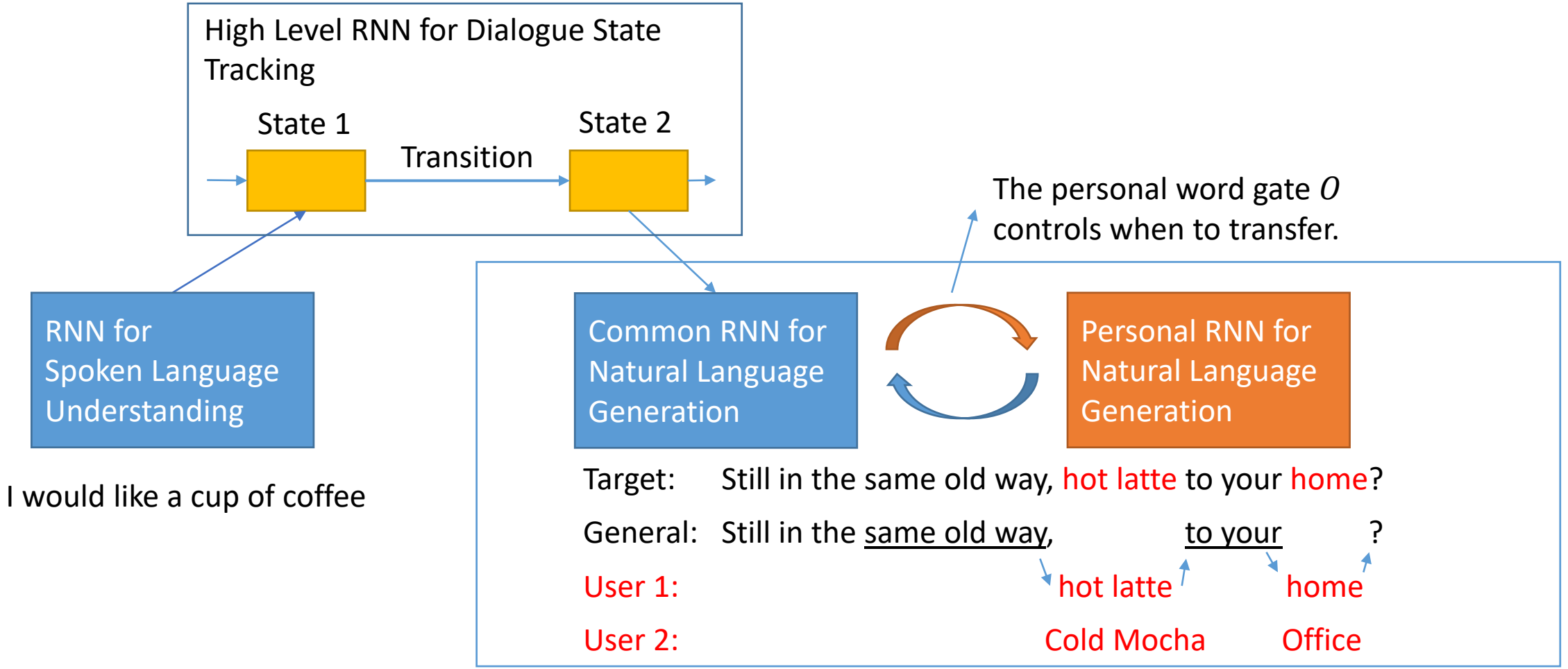
Personalized Decoder



Example of different reply for different person.

Fine-grain Transfer Learning with Personal Word Gating

Model: Personalized HRED



Fine-grain Transfer Learning with Personal Word Gating

- Reinforce Algorithm
  - $L(\Theta) = -\int_{D^u} p(D^u|\Theta) J^{D^u} dD^u$, where $D^u = \{X_n^u, O_n^u, Y_n^u, r_n^u\}$ is a training dialogue record, J^{D^u} is the total reward for D^u.
- Algorithm
  - For each dialogue $D^u = \{X_n^u, O_n^u, Y_n^u, r_n^u\}$ of each user:
    - Simultaneous update
      - the control gate function g
      - the shared RNN Θ^g
      - the Personal RNN Θ_u^p for user u.

Baselines

	Base Model	Multi-turn	Multi-task	Transfer Granularity	Shared Component	Has Personal Q-function
S2S (Sutskever, Vinyals, and Le 2014)	S2S	No	-	-	-	No
ST-S2S (Luong et al. 2015)	S2S	No	Yes	Sentence	Encoder, Decoder	No
ST-E-S2S (Luong et al. 2015)	S2S	No	Yes	Sentence	Encoder	No
PT-S2S (Proposed)	S2S	No	Yes	Phrase	Encoder, Common Decoder	Yes
HRED (Serban et al. 2015)	HRED	Yes	-	-	-	No
ST-HRED	HRED	Yes	Yes	Sentence	Encoder, Decoder	No
PT-HRED (Proposed)	HRED	Yes	Yes	Phrase	Encoder, Common Decoder	Yes

Our works:

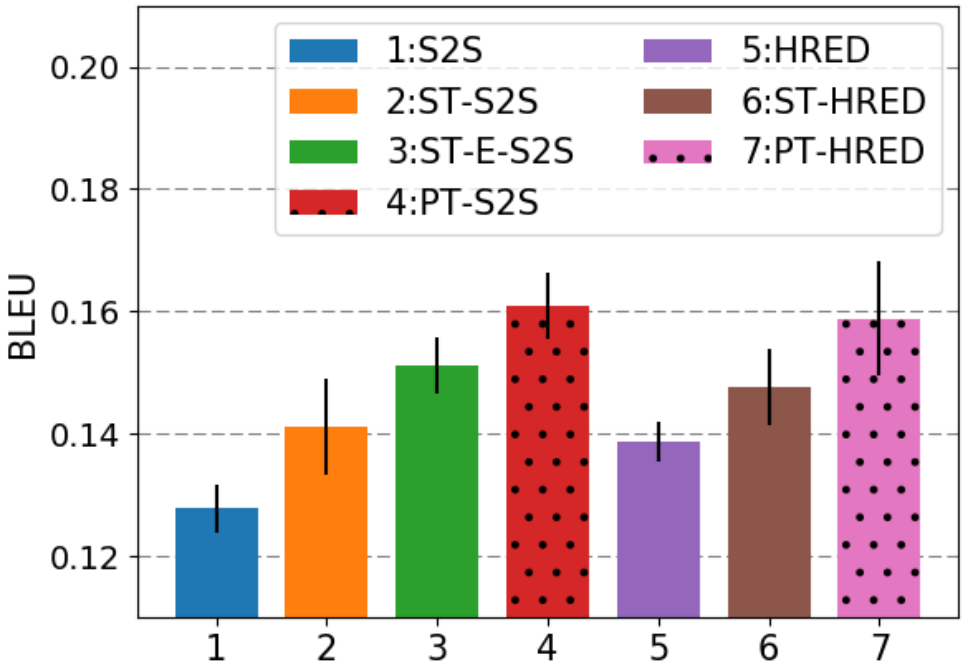
Experiments: Real-data

- Real-Dataset
  - O2O personalized coffee shopping dialogue
  - 52 users
  - 464 Training Dialogue and 831 Testing Dialogue
- Evaluation:
  - BLEU:
    - for each turn, 5 randomly generated replies will be compared with the ground truth reply.

Dialogue history H_2

X_1:需要一杯星巴克中杯拿铁
Y_1 :**拿铁(中杯)**；请问冰的还是热的？送到 **海淀大街34号海置创投大厦5层** 对吧，配送费每单5元起
X_2 :现在有配送费啦？[r_1]

Y_2: 一直都有呀
$\bar{Y}_2$: <Predicted Sentence>

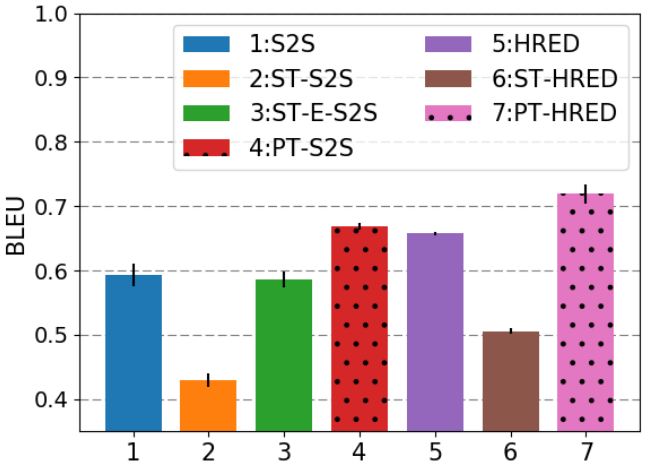


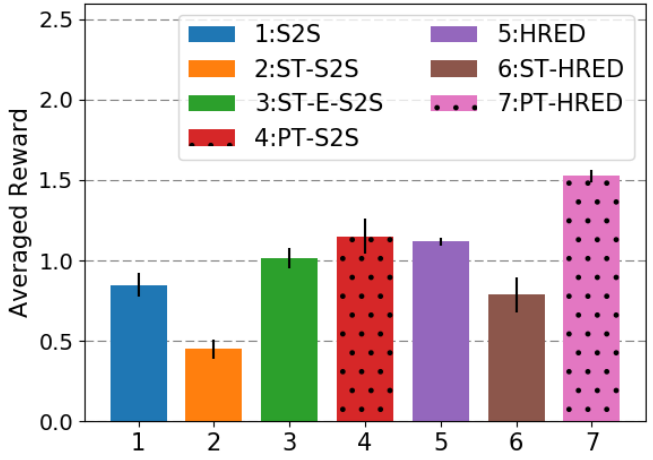
Experiment: Simulation

- Simulation
  - Coffee ordering with 4 slots
  - 10 simulated users
  - 50 training dialogues
  - 2000 testing dialogues
  - Online Testing for 1000 orders
- Evaluation Metrics:
  - BLEU
  - Online Test Reward
    - Candidate replies are ranked by generation probability and the top reply is used.

Dialogue history H_2

X_1: 想喝杯咖啡了 $[r_1]$
Y_1: 还是老样子 **中杯 热 新鲜调制咖啡**, 送到**北京市海淀区丹棱街1号互联网金融中心23层**吗？
X_2: 来一杯英式红茶拿铁送到深圳市南山区科技园飞亚达大厦10楼 $[r_1 = -0.4]$
Y_2: 麻烦您支付一下
X_3:已经付款了 $[r_2 = 1]$





Case study

- Sentence-level transfer HRED transferred the wrong address from the source users
- Phrase-level transfer HRED transferred the answer pattern with the correct address.

Sentence-level transfer HRED (ST-HRED)
X: 麻烦 太妃 榛果 拿铁 热 venti 送到 公司 #
Y: 请问 冰 的 还是 热 的 ？ 送到 **海淀区 清上园 UNK - UNK - UNK** 对 吧

Transfer wrong address.

Phrase-level transfer HRED (PT-HRED)
X: UNK 大杯 焦糖玛奇朵 到 公司 #
Y: 请问 冰 的 还是 热 的 ？ 送到 **朝阳区 三元 桥 国际 港 A座 UNK** 对 吧

Correct personalized address.

Outline

- Background
- Transfer Reinforcement Learning Problem
- Summarization of Related Work
- Transfer Reinforcement Learning through Personalized Q-function
- Transfer Reinforcement Learning through Personal Word Gating
- Transfer Reinforcement Learning through Speech-Act and Slot alignment
- The Unified Model
- Conclusion

Cross-domain Dialogue Policy Transfer Problem Example

- How to transfer dialogue policy across different Speech-Acts and domains?

Source Domain 

	Restaurant Booking Dialogue
X_1	I want to find a restaurant.
Y_1	What food would you like?
X_2	I would like Chinese food.
Y_2	Which area?
X_3	In central.

Transfer

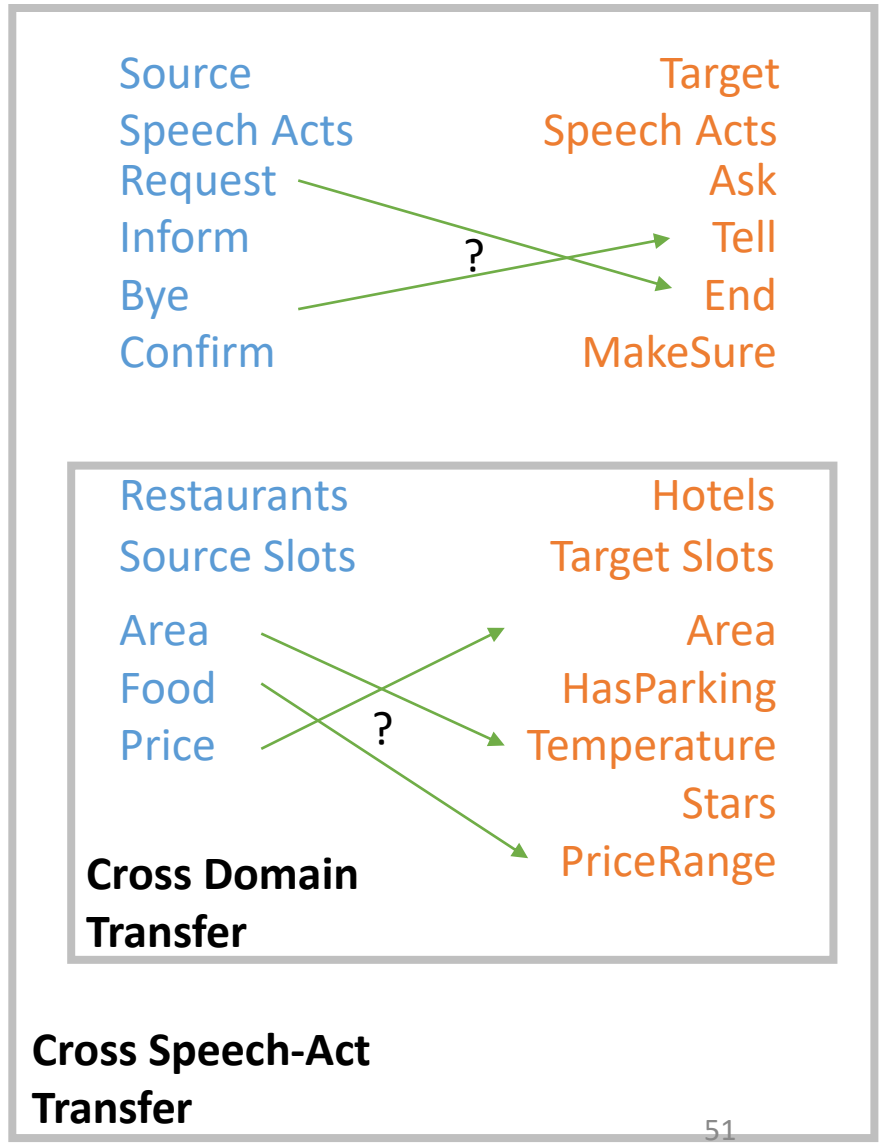
Target Domain 

	Hotel Booking Dialogue
X_1	I am looking for a place to stay.
Y_1	A hotel or a guest house?
X_2	I would like a hotel.
Y_2	

Abstract Reply Set
Y_{c1} Request(Area): Which area?
Y_{c2} Request(Price): What price range?

Cross domain & system policy transfer with Speech-Act and Slot alignment

- How to transfer dialogue policy across different Speech-Acts and domains?
- Input:
  - Plenty of data $\{\{X^s, Y^s, r^s\}\}$ in source domain, where X^s, Y^s are composed of source speech-act A^s and source slots S^s.
  - Little data $\{\{X^t, Y^t, r^t\}\}$ in target domain, where X^t, Y^t are composed of target speech-act A^t and target slots S^t. Note that source and target speech-acts and slots might be totally disjoint.
- Output:
  - Policy π^t in target domain



Cross domain & system policy transfer with Speech-Act and Slot alignment

- The target domain Q-function is
- $Q(H^t, Y^t) =$

$$Q_g(f(H^t), f(Y^t))$$

Source Domain Q-function

Cross Domain Mapping function

H^t and Y^t are state and action in the target domain, while $f(H^t)$ and $f(Y^t)$ are corresponding state and action in the source domain.

Cross domain & system policy transfer with Speech-Act and Slot alignment

- Dialogue Action Vector $\bar{X} = [\bar{a}, \bar{s}], \bar{Y} = [\bar{a}, \bar{s}]$
  - X: I want to find a Chinese restaurant.
  - $\bar{X} = \mathrm{SLU}(X)$
  - $\bar{X} = [\{\mathrm{inform}: 1, \mathrm{req}: 0, \mathrm{reqalt}: 0\}, \{\mathrm{type}: 1, \mathrm{food}: 1, \mathrm{price}: 0, \mathrm{area}: 0\}]$

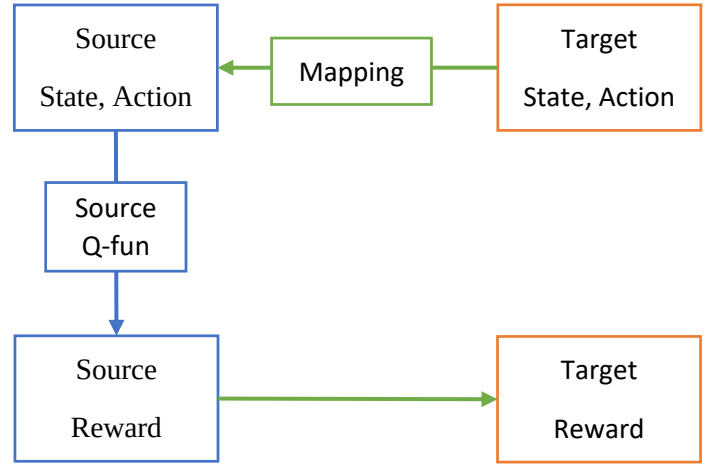
    $\bar{a}$: User Speech-Act; $\bar{s}_I$: Slots

- Dialogue State vector $\bar{H} = [\bar{v}_{db}, \bar{a}, \bar{s}_I, \bar{s}_R]$
  - $\bar{H}_i = \mathrm{DST}(H_i)$, where dialogue history $H_i = \{\{X_j, Y_j\}_{j=1}^{i-1}, X_i\}$
  - $\bar{H}_i = [\{\#\mathrm{ent}: 3, \mathrm{more}_{\mathrm{cons}}: 1\}, \{\mathrm{inform}: 1, \mathrm{request}: 0\}, \mathrm{I}\{\mathrm{type}: 1, \mathrm{food}: 1\}, \mathrm{R}\{\mathrm{Phone}: 0\}]$

    $\bar{v}_{db}$: Database Vector; $\bar{a}$: User Speech-Act; $\bar{s}_I$: Informed Slots; $\bar{s}_R$: Requested Slots
  - Examples: What is the **phone** of this **Chinese** restaurant?

Cross domain & system policy transfer with Speech-Act and Slot alignment

- state/action are translated with a slot similarity matrix M_S and an Speech-Act similarity matrix M_A.
  - $\bar{H}^s = f(\bar{H}^t; M_s, M_A)$
  - $\bar{Y}^s = f(\bar{Y}^t; M_s, M_A)$



- Target Q-function $Q(\bar{H}^t, \bar{Y}^t)$ can be expressed with source Q-function $Q(\bar{H}^s, \bar{Y}^s)$, M_S and M_A.
  - $Q(\bar{H}^t, \bar{Y}^t) = Q_g(f(\bar{H}^t; M_s, M_A), f(\bar{Y}^t; M_s, M_A))$
  - $\pi^t(\bar{H}^t) = \mathrm{argmax}_{a^t} Q(\bar{H}^t, \bar{Y}^t)$

State:[hasParking, Price]
Action:[request hasParking]

State:[food]
Action:[request food]
Source Q-function: 4.09

Target Q-function: 4.09

Cross domain & system policy transfer with Speech-Act and Slot alignment

- Q-learning
  - $L(\Theta) = E\left[\left(r_i + \gamma\, max_{\bar{Y}_{i+1}^t} Q\left(\bar{H}_{i+1}^t, \bar{Y}_{i+1}^t | \Theta\right) - Q\left(\bar{H}_i^t, \bar{Y}_i^t \middle| \Theta\right)\right)^2\right] + R(\Theta)$
- Regularization $R(\Theta)$ for the Speech-Act mapping matrix, for improved stability.
  - Similar Speech-Acts have similar occurrence probability.
  - Similar Speech-Acts occurred with similar slots. Hello()≠Request(**Phone**)
- Algorithm
  - Step1: Build Source policy with Q-learning
  - Step2: Build Speech-Act and Slot alignment
    - For each $\{\bar{H}_i^t, \bar{Y}_i^t, r_i\}$ in the target domain
      - $\Theta = \Theta + \alpha * \Delta L(\Theta)$

Baselines

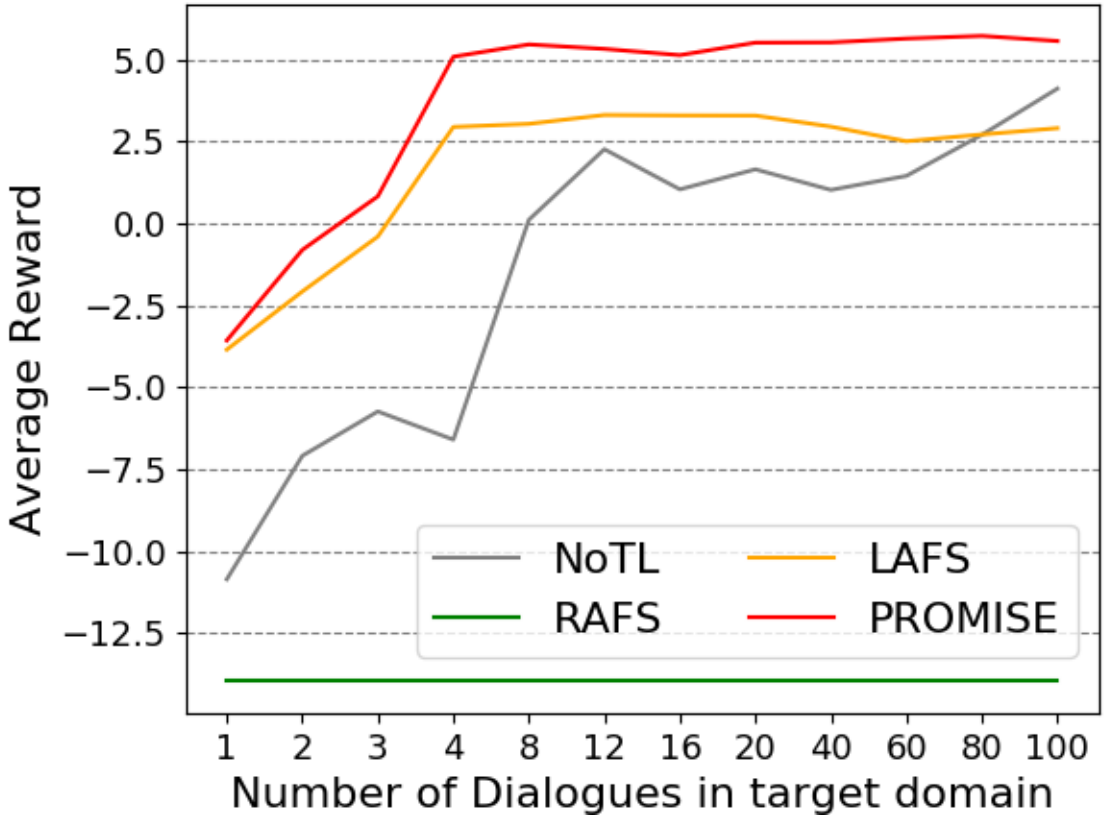
	Base Policy	Source	Speech-act mapping	Slot mapping	Auxiliary Database
NoTL	GP	-	-	-	-
RAFS	GP	Restaurant	Random	Heuristic (Gašic et al., 2015a)	Yes
LAFS	GP	Restaurant	Learned	Heuristic (Gašic et al., 2015a)	Yes
PROMISE (Proposed)	GP	Restaurant	Learned	Learned	No
FAFS (**Upper Bound**)	GP	Restaurant	Ground-Truth	Heuristic (Gašic et al., 2015a)	Yes
FALS (**Upper Bound**)	GP	Restaurant	Ground-Truth	Learned	No

Our work:

Cross domain & system policy transfer with Speech-Act and Slot alignment

- Simulation with Pydial
  - Restaurant Booking -> Hotel Booking.
  - Train/evaluate with simulators.
- Reward
  - -1 for each turn
  - +20 for success
  - **Final** Reward Only
- Dataset

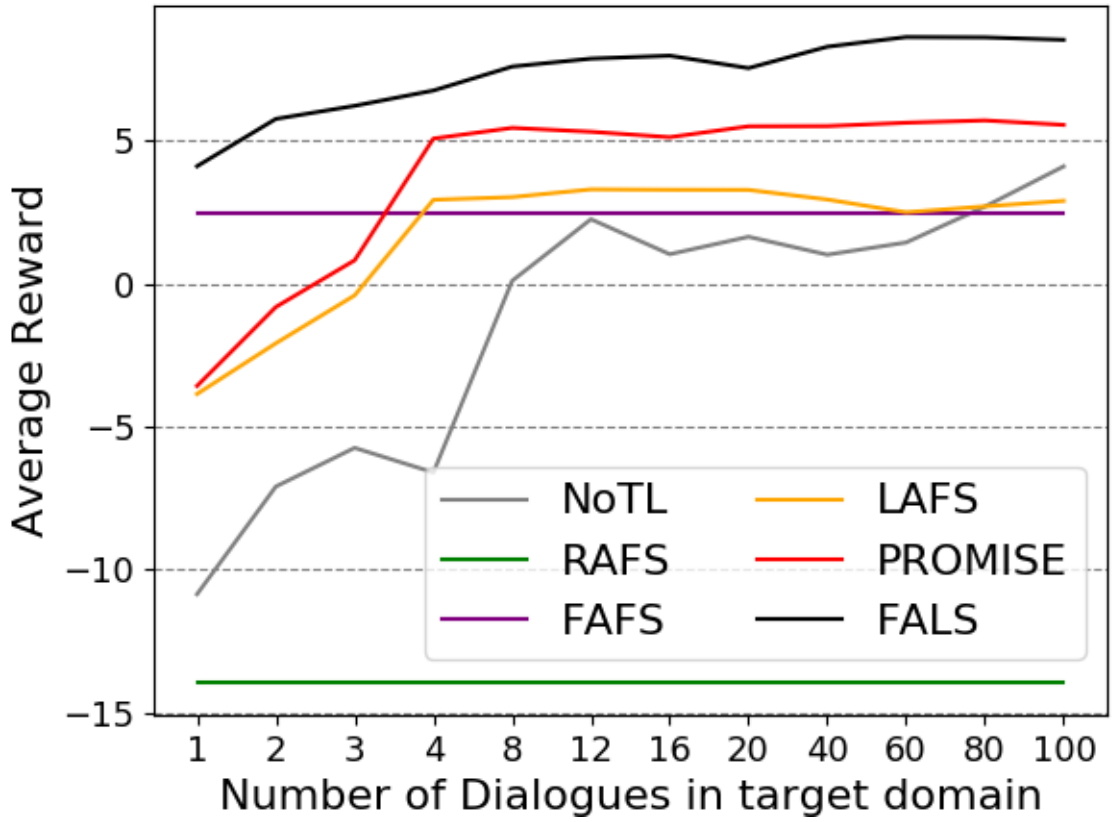
	Acts	Slots	Dialogues
Source	22	9	1000
Target	22	11	1-100



Speech-Act Learning is necessary for dialogue transfer.

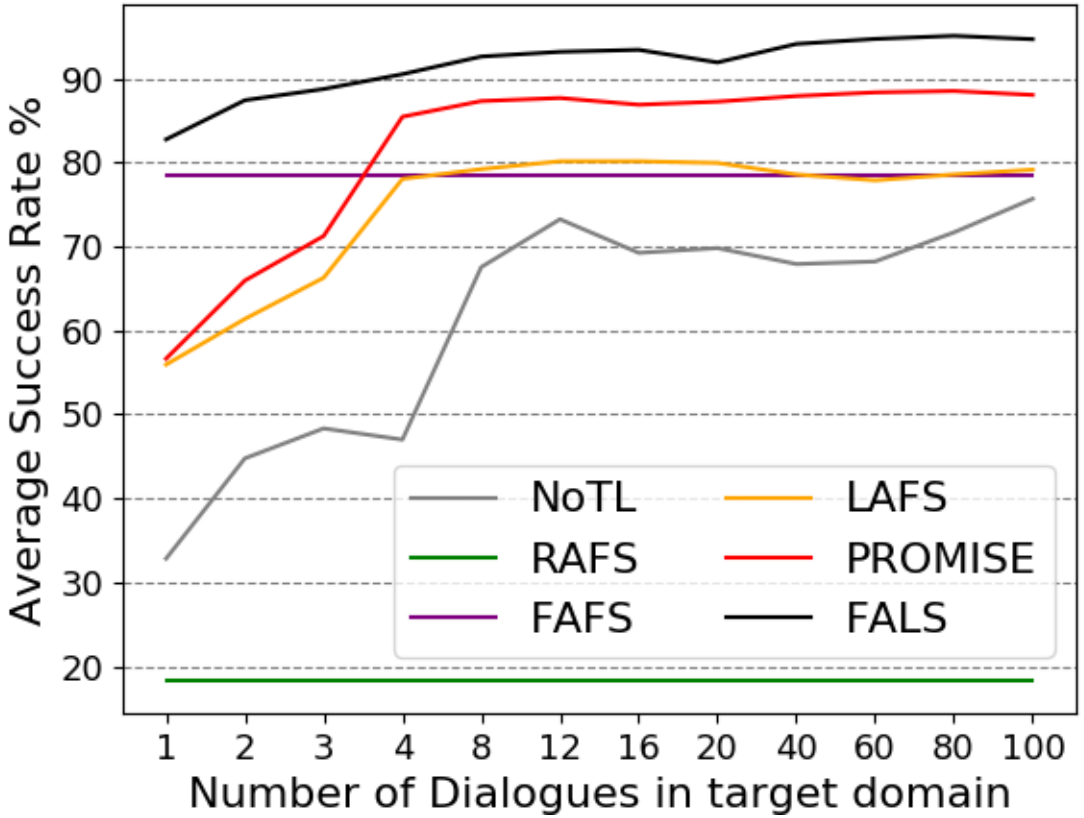
Cross domain & system policy transfer with Speech-Act and Slot alignment

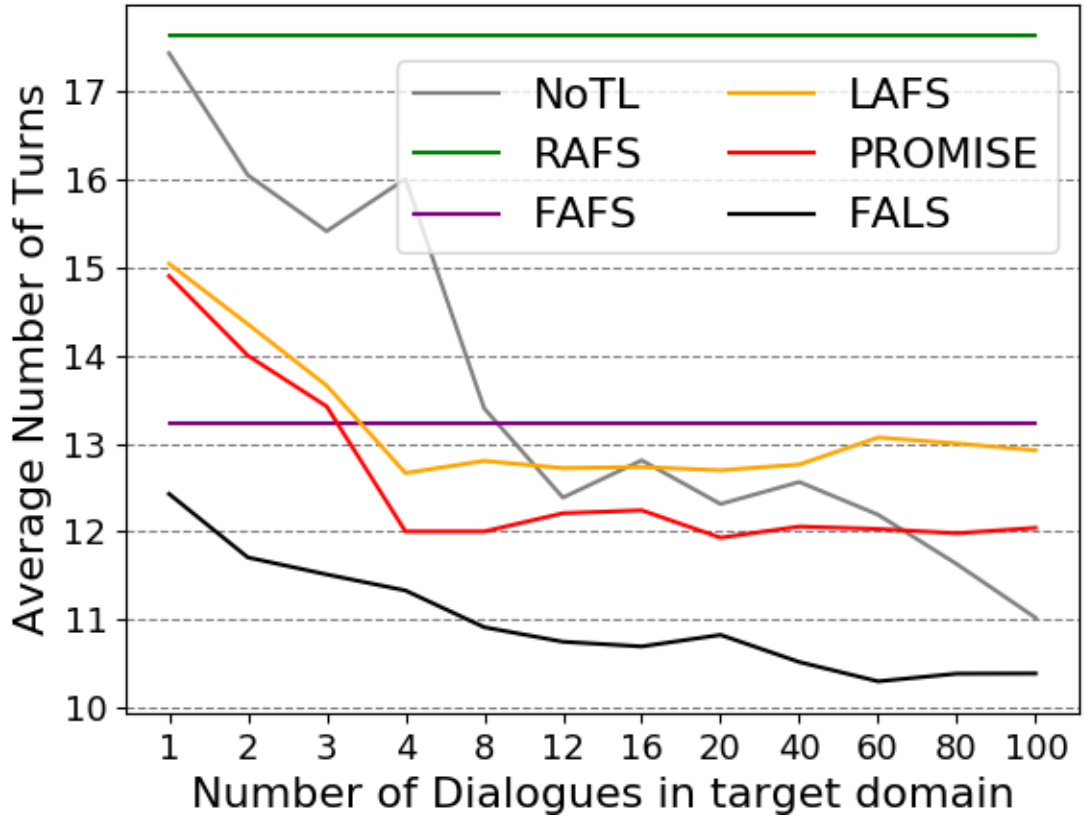
- Performance upper bound: “FAFS” and “FALS”
  - Ground truth speech-act mapping
- Result: PROMISE > LAFS, FALS > FAFS
  - Both Speech-Act Learning and Slot learning are necessary for dialogue policy transfer.



Cross domain & system policy transfer with Speech-Act and Slot alignment

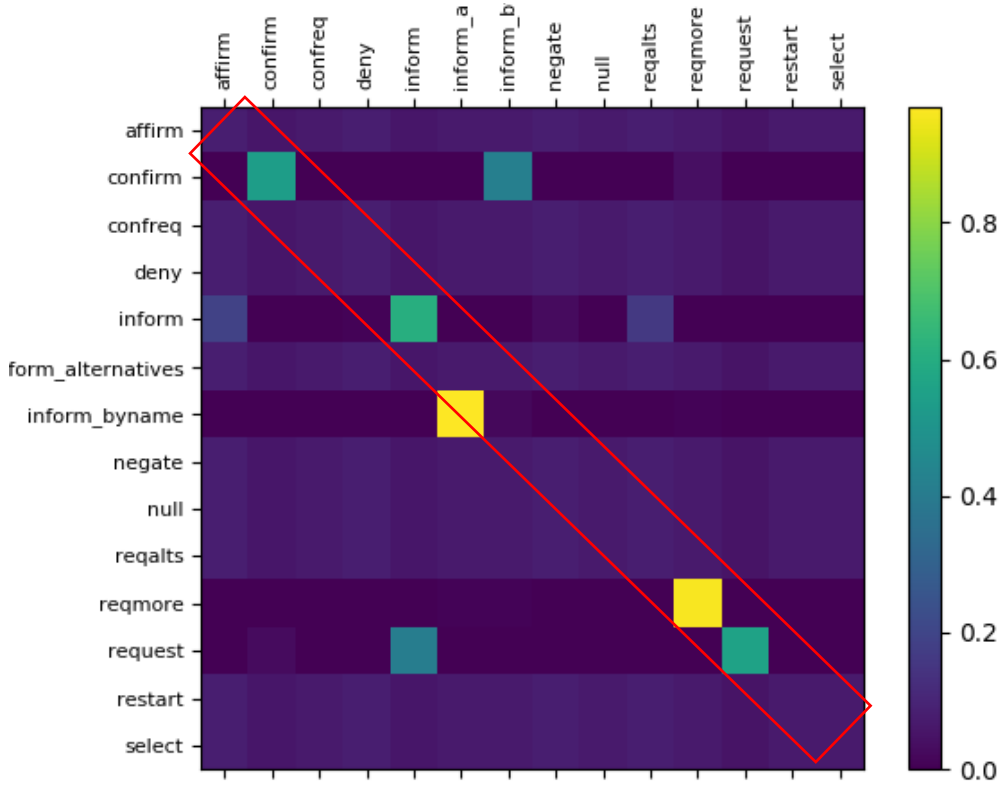
- Averaged Success Rate and Averaged Dialogue Length



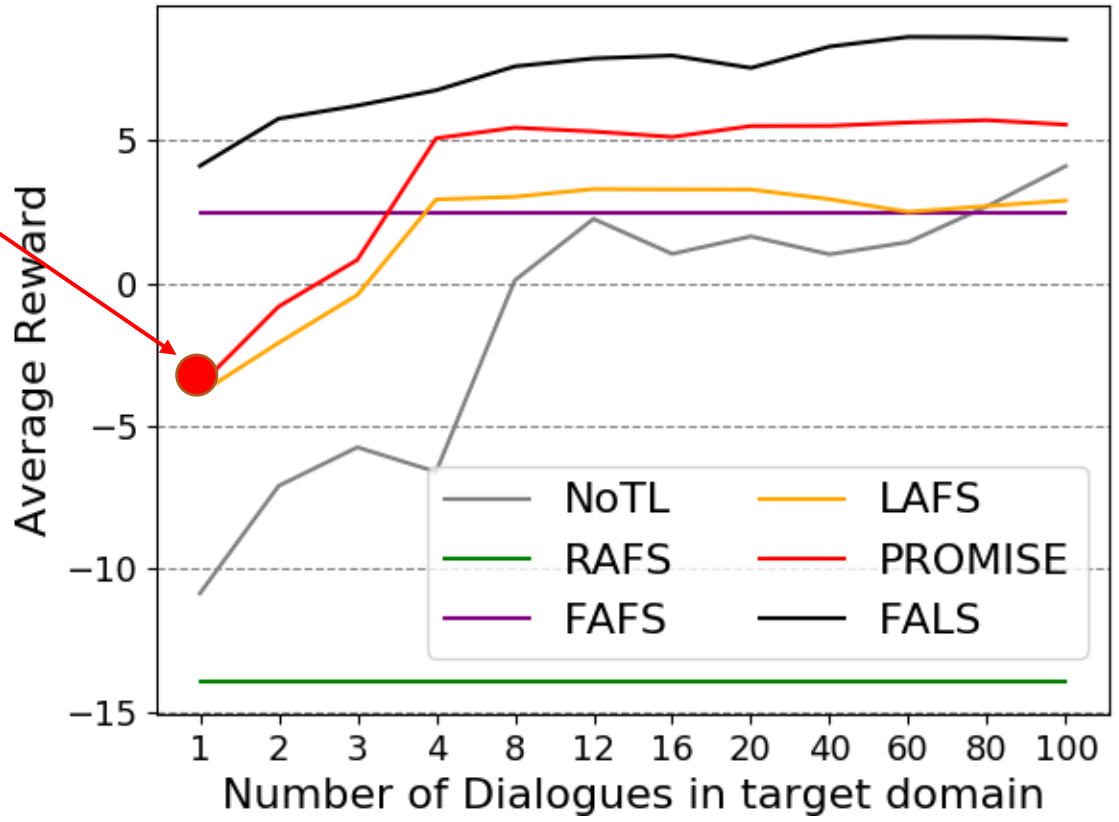


Cross domain & system policy transfer with Speech-Act and Slot alignment

- Speech-Act Mapping when target has 1 dialogue, shown in

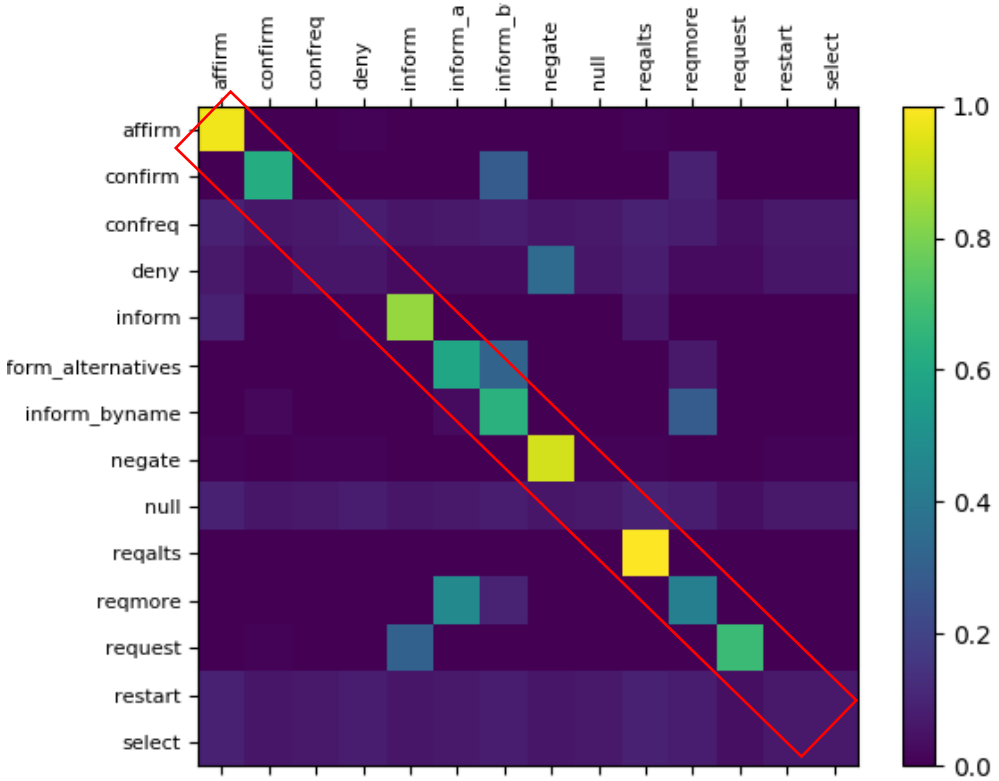


Ground Truth Mapping is diagonal matrix

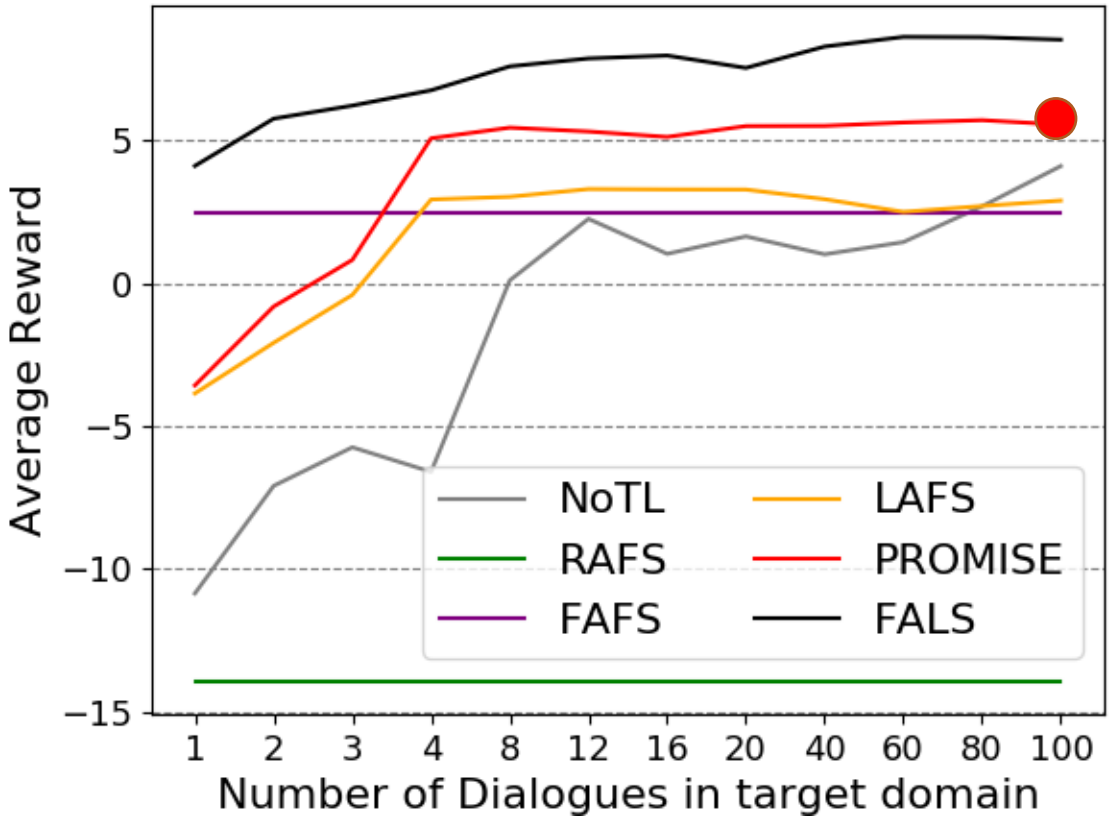


Cross domain & system policy transfer with Speech-Act and Slot alignment

- Speech-Act Mapping when target has 100 dialogues



Ground Truth Mapping is diagonal matrix



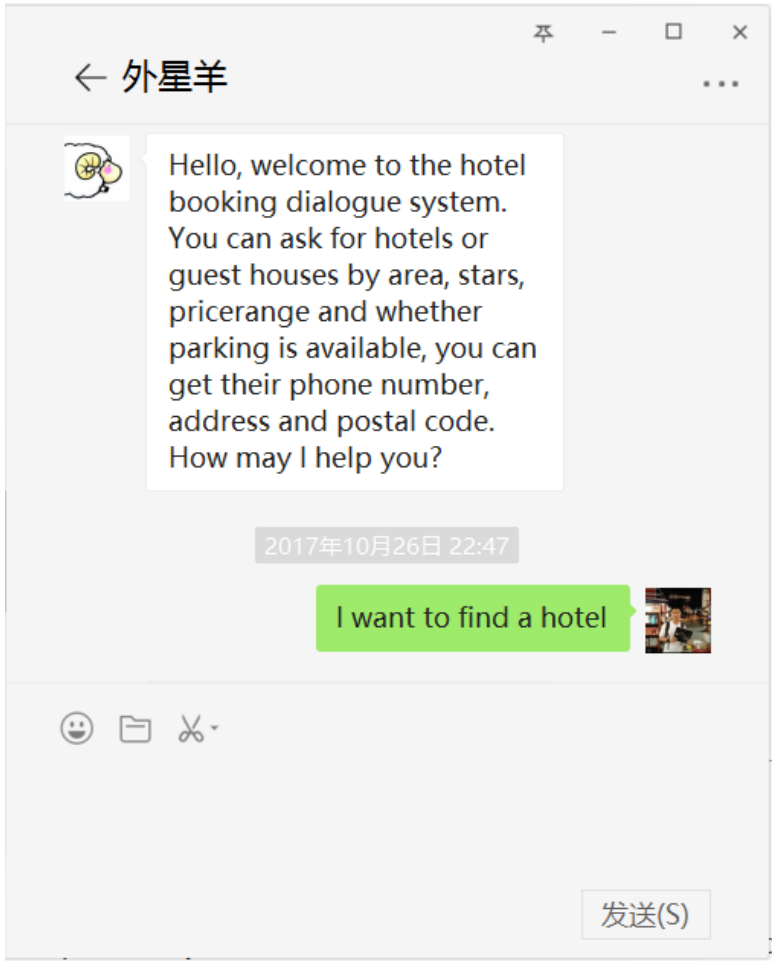
Ground truth Speech-Act mapping can actually be learned.

Cross domain & system policy transfer with Speech-Act and Slot alignment

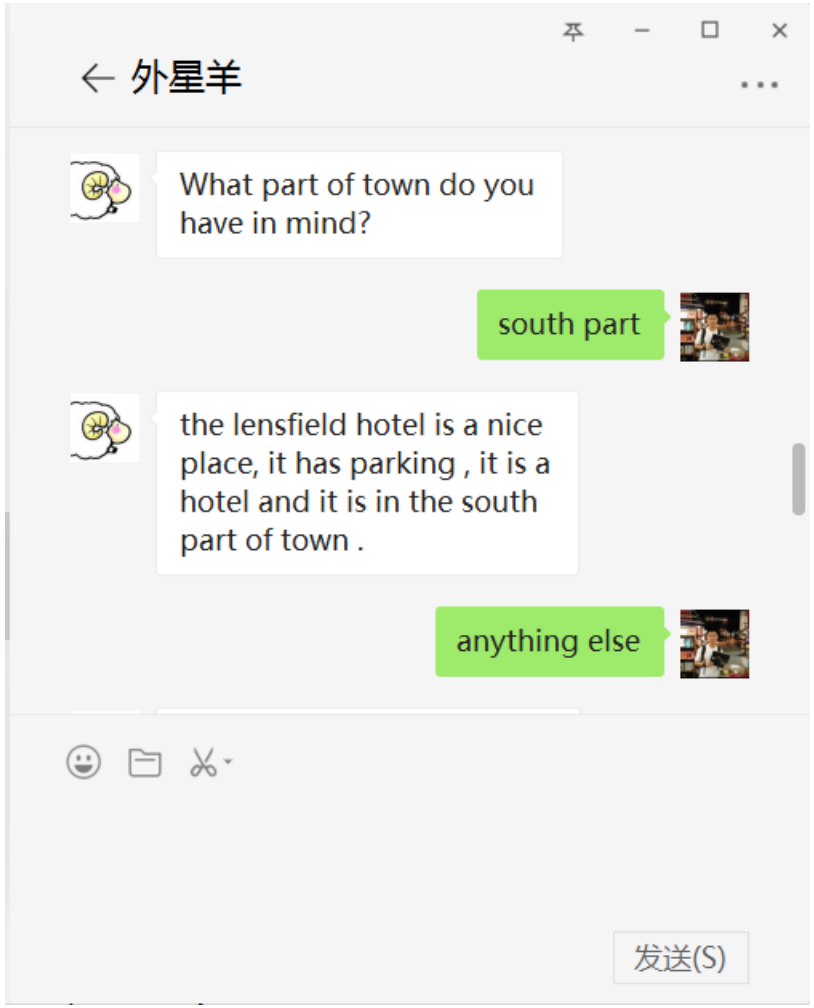
- Real-world experiment
  - Restaurant Booking -> Hotel Booking.
- Dataset
  - Source: Real dialogues on Restaurants booking domain.
  - Target: Real dialogues on Hotel booking domain.
- Evaluation:
  - Static
    - AUC of reply candidate ranking.
  - Live
    - Evaluated with user simulators.
    - Averaged Reward, Success Rate and Averaged Dialogue Length.

Data Collection Wechat Interface

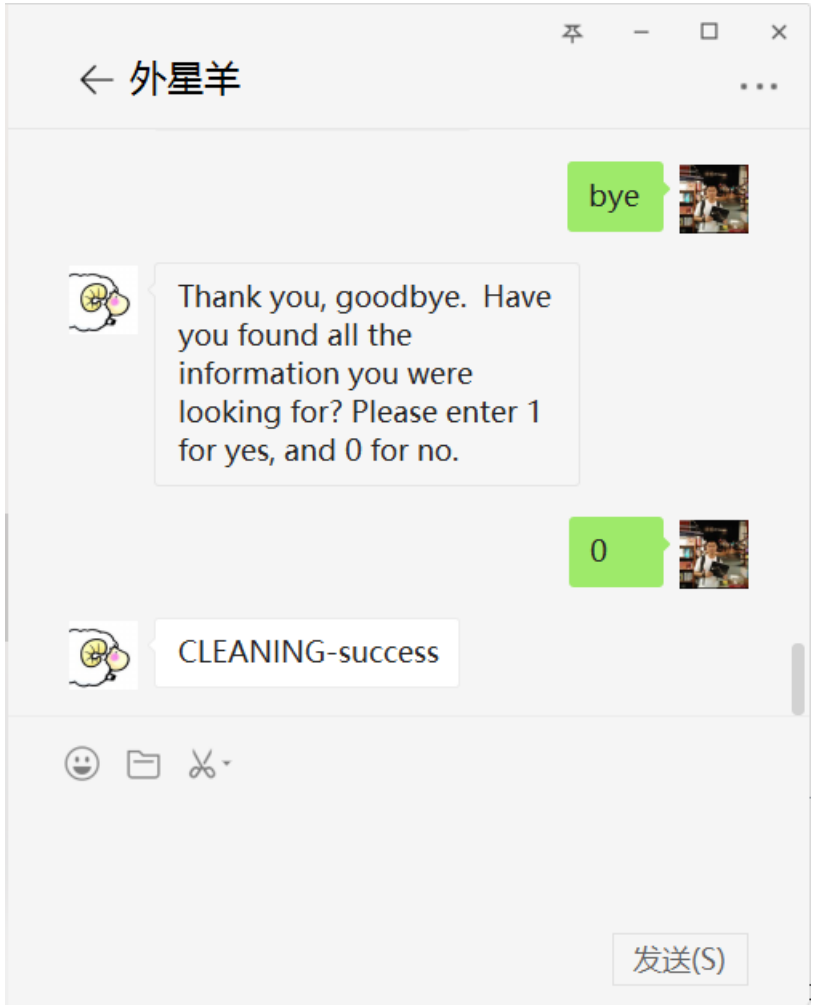
Dialogue begin



Dialogue in progress



User subjective feedback

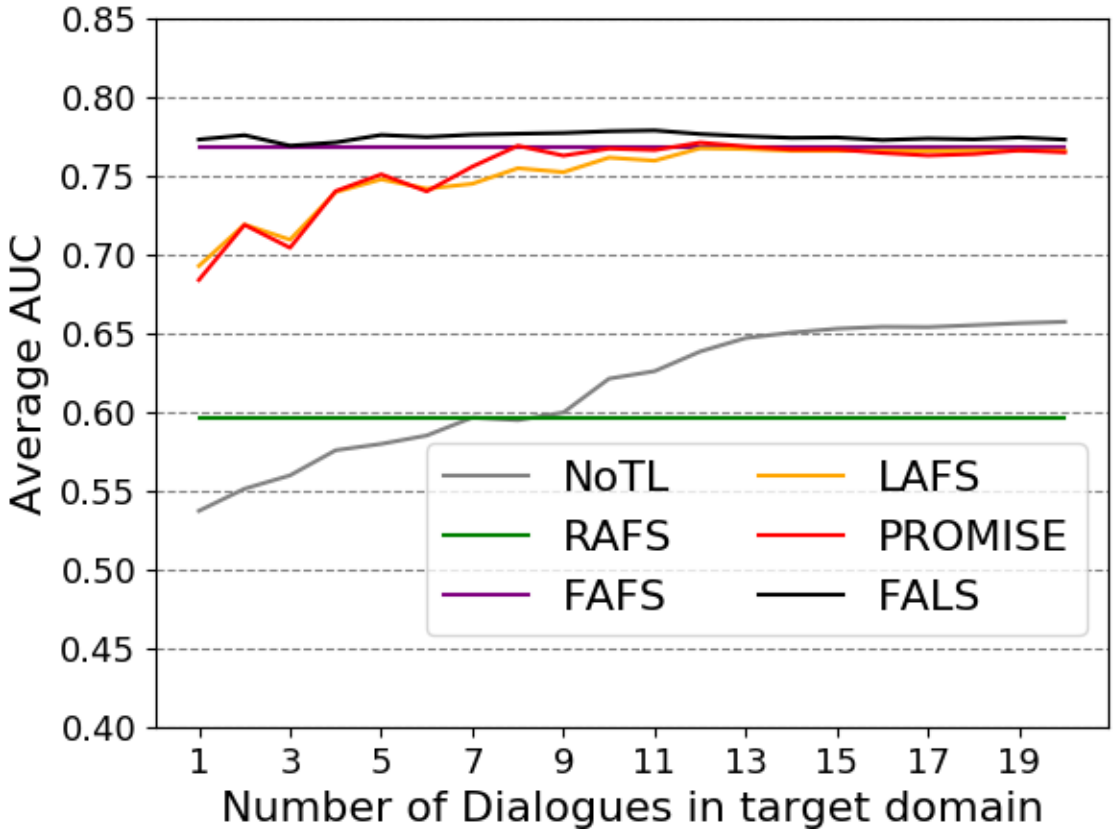


Cross domain & system policy transfer with Speech-Act and Slot alignment

- Real-word Dataset
  - Train: Real Customer VS Robot Agent
  - Test: Real Customer VS Real Human Servant
  - **Immediate** Reward

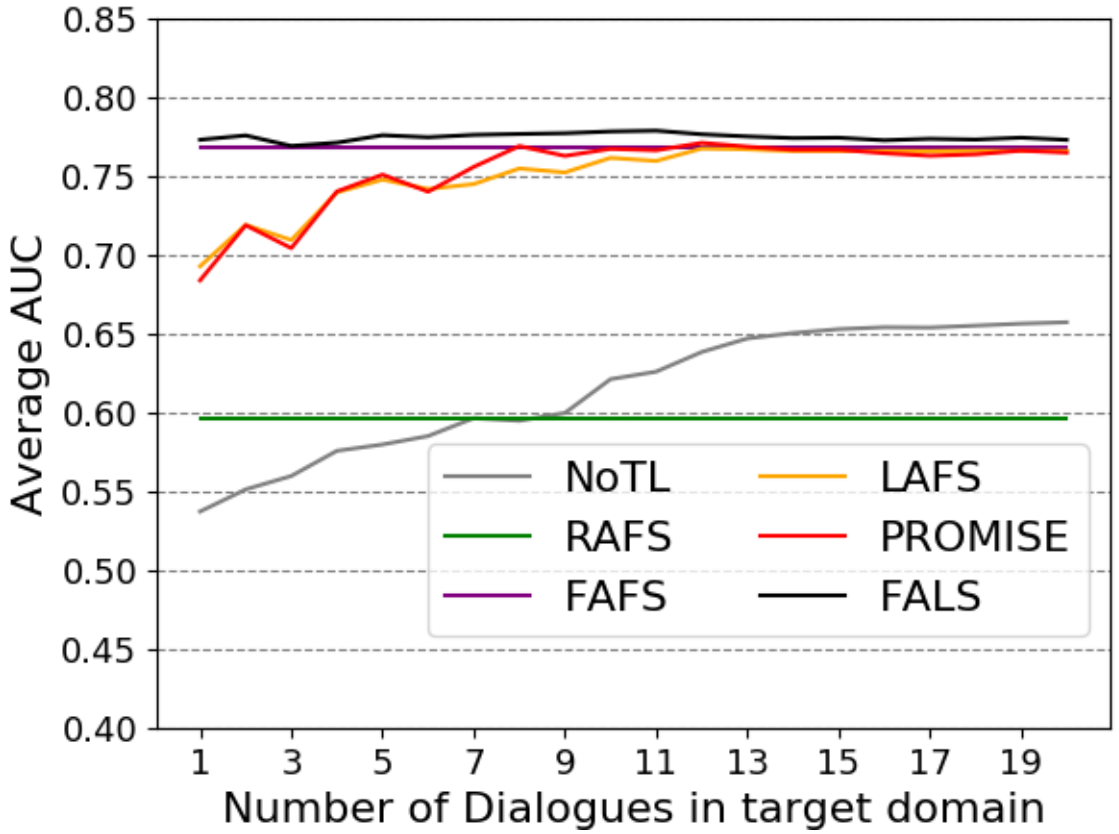
	Train Src	Train Tgt	Test Tgt
Real-data	40	1-20	60

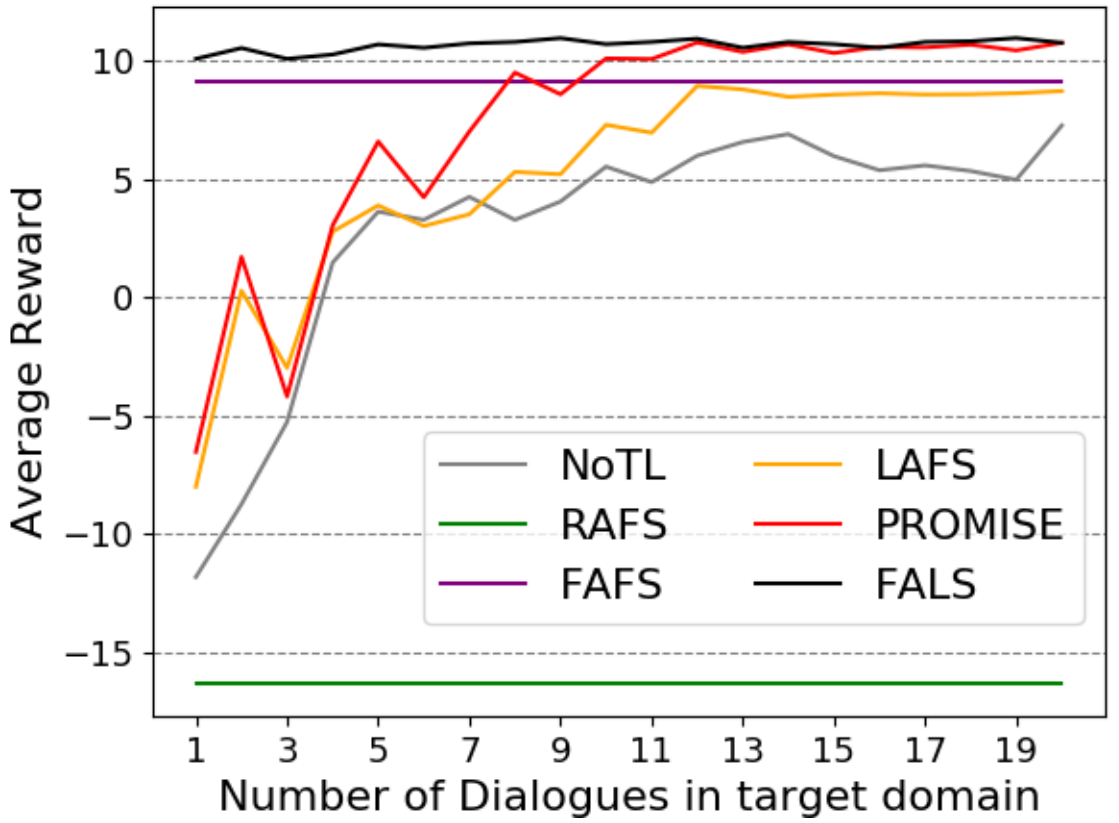
- Evaluation: AUC
  - In each turn, all possible replies are ranked according to the sampled Q-score.
- Transfer learning helps a lot for task-oriented dialogue systems.
  - NoTL has large variance.



Cross domain & system policy transfer with Speech-Act and Slot alignment

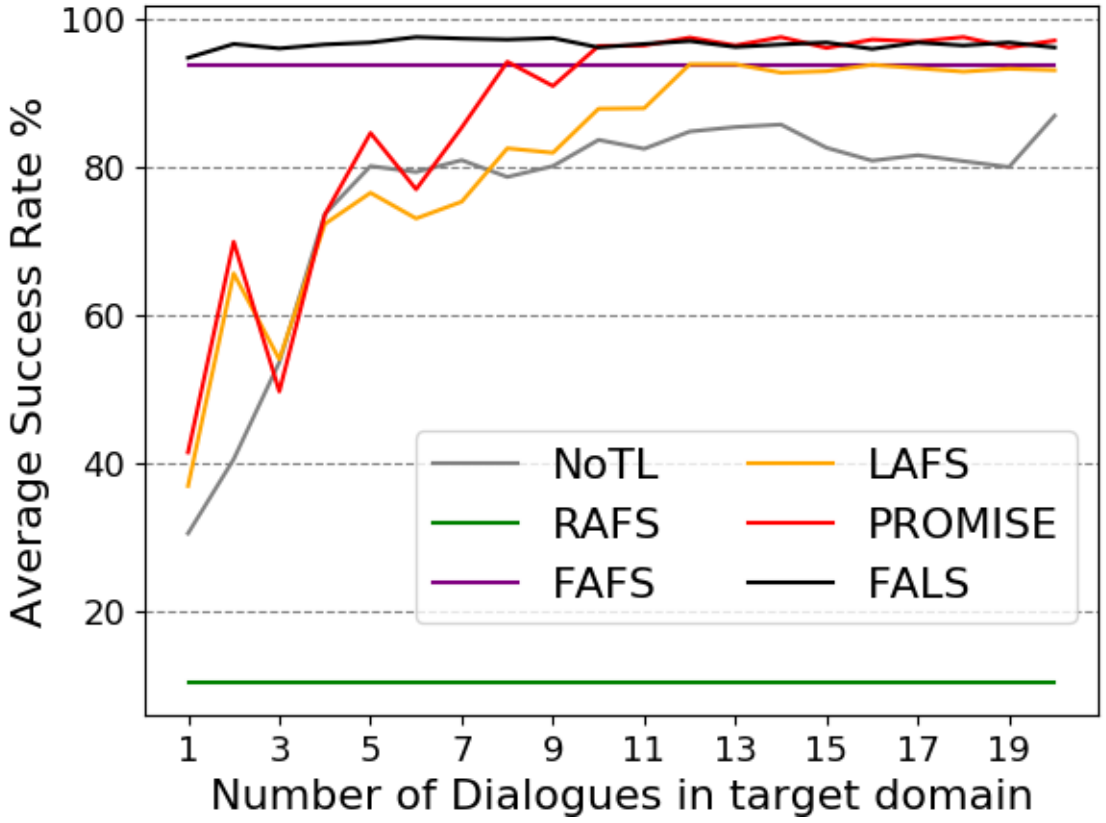
- Averaged AUC in static evaluation, and Averaged Reward in live evaluation

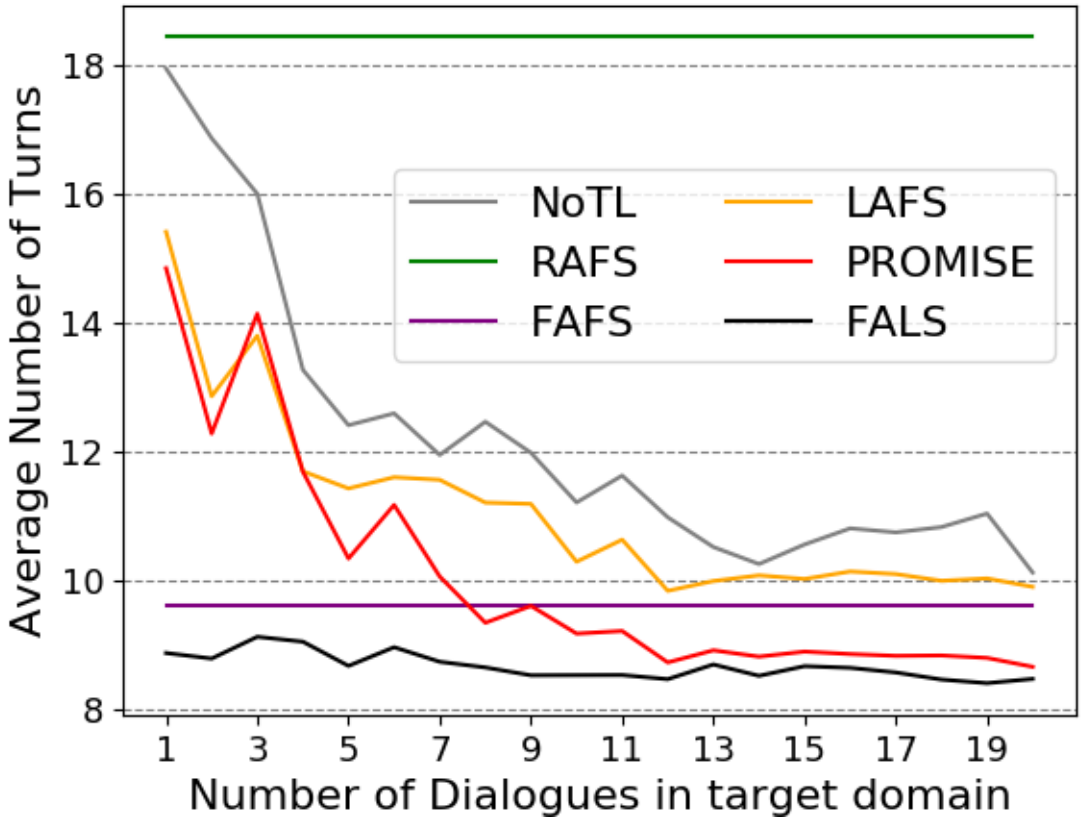




Cross domain & system policy transfer with Speech-Act and Slot alignment

- Averaged Success Rate and Averaged Dialogue Length in Live Evaluation





Outline

- Background
- Transfer Reinforcement Learning Problem
- Summarization of Related Work
- Transfer Reinforcement Learning through Personalized Q-function
- Transfer Reinforcement Learning through Personal Word Gating
- Transfer Reinforcement Learning through Speech-Act and Slot alignment
- The Unified Model
- Conclusion

Unified Policy Transfer Framework

- The target domain Q-function can be represented by a function of state H^t and action Y^t as
- $Q(H^t, Y^t) =$

$$(1 - O^p(H^t))Q_g(f(H^t), f(Y^t)) + O^p(H^t)Q_p(H^t, Y^t)$$

Source Domain Q-function	Cross Domain Mapping function	Personal Gate	Personal Q-function

Note that mode fine-tuning is one of our special cases.

Specification of the Unified Model

- $Q(H^t, Y^t) = (1 - O^p(H^t))Q_g(f(H^t), f(Y^t)) + O^p(H^t)Q_p(H^t, Y^t)$

Specification

- Section 4: $\mathrm{Q}^{\pi_u}(H_i, Y_i|\Theta) = Q_g(H_i, Y_i|\Theta^g) + Q_p\left(H_i, Y_i\middle|\Theta_u^p\right)$
  - where $f(H^t) = H^t, f(Y^t) = Y^t, O^p(H^t) = O^p$

- Section 5: $Q(H_t, Y_t) = (1 - O_t^u)Q_g(H_t, Y_t|\Theta^g) + O_t^u Q_p(H_t, Y_t|\Theta_u^p)$
  - where $f(H^t) = H^t, f(Y^t) = Y^t, O_t^u = g(H_{t-1}, Y_{t-1}, O_{t-1}^u)$

Still Cold Mocha? → Still Hot Latte?

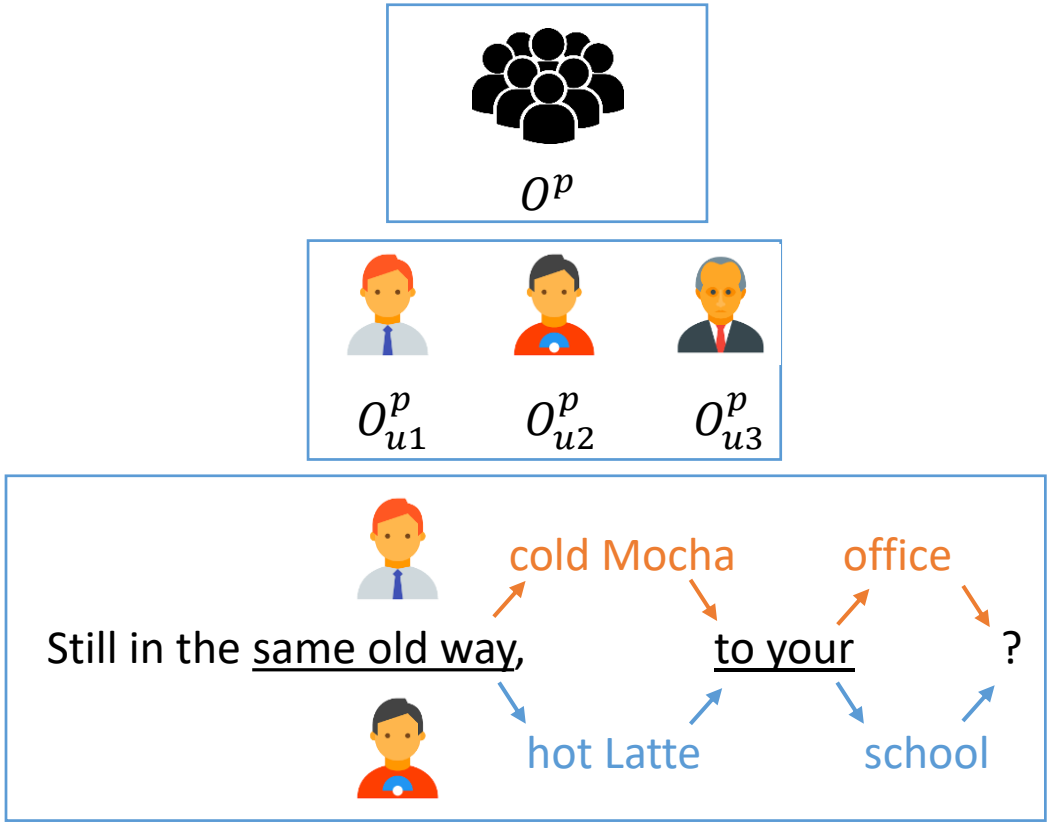
- Section 6: $Q(H^t, Y^t) = Q_g(f(H^t), f(Y^t))$
  - where $O^p(H^t) = 0$

Personal Specification at different granularities

- $Q(H^t, Y^t) = (1 - O^p(H^t))Q_g(f(H^t), f(Y^t)) + O^p(H^t)Q_p(H^t, Y^t)$

Specification

- Global Personal Weight: same O^p for all user
- User-level Personal Weight: $O^p(u)$ varies for each user
- Word-level Personal Weight function: $O^p(H^t)$ varies for each word and for each user



Outline

- Background
- Transfer Reinforcement Learning Problem
- Summarization of Related Work
- Transfer Reinforcement Learning through Personalized Q-function
- Transfer Reinforcement Learning through Personal Word Gating
- Transfer Reinforcement Learning through Speech-Act and Slot alignment
- The Unified Model
- Conclusion

Conclusion

- We have systematically studied dialogue policy transfer problems in
  - transfer across users with different preference;
  - transfer common fine-granularity knowledges;
  - transfer across domains with different speech-acts.
- We have proposed models and algorithms for the about problems, and we conclude with a unified transfer reinforcement learning framework.
- We have applied the above models on different applications and verify the effectiveness of the proposed models.

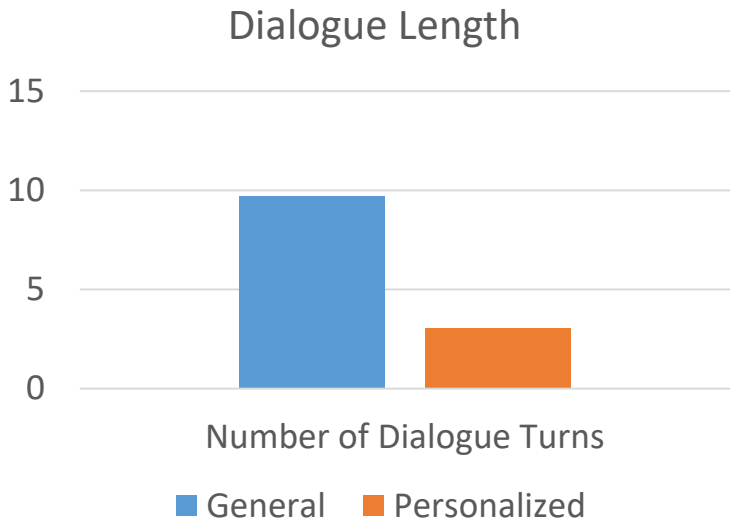
Comparison of Transfer Reinforcement Learning (TRL) works for Task-oriented Dialogue Systems

	Transfer Granularity	Personalized Q-function	End-to-End	Cross Domain	Different Speech-act	Auxiliary Database
Fine-tuning (Genevay and Laroche 2016; Casanueva et al. 2015)	Sentence	No	No	No		
TRL with Personal Q-fun	Sentence	Yes	No	No		
S2S with shared Encoder (Luong et al. 2015)	Sentence	No	Yes	No		
TRL with Personal Word Gate	Phrase	Yes	Yes	No		
Gaussian Process Transfer (Gašic et al. 2014; Gašic et al. 2013b; Gašic et al. 2015a)	Speech-act / Slot	No	No	Yes	No	Yes
TRL with Speech-Act and Slot alignment	Speech-act / Slot	No	No	Yes	Yes	No

Our works:

TRL for Personalized Task-oriented Dialogue System

- Adapt to each user's preference.
- Reduce dialogue length.
- Learn dialogue states with source data and a few target data



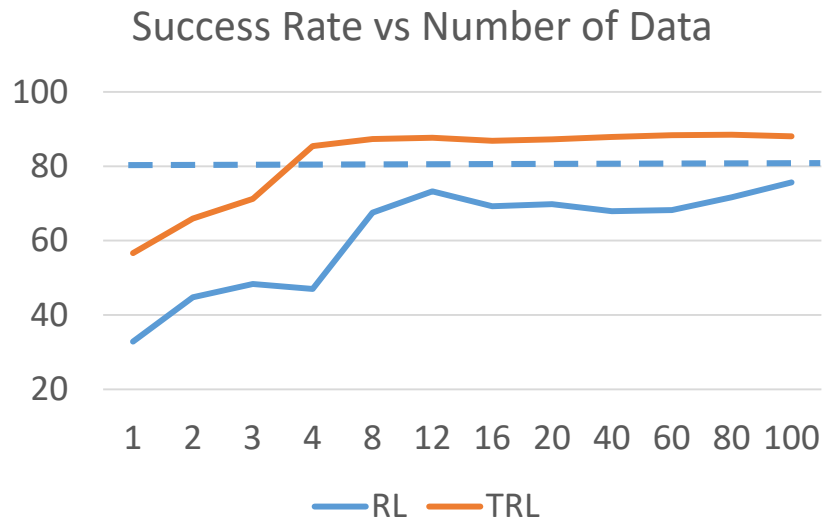
Customer in a hurry:
You can always remember!

	Personalized Coffee Shopping Dialogue (X=customer, Y=system)
X_1	Can I have coffee please?
Y_1	Same as before? Tall Iced Mocha and deliver to No.1199 Mingsheng Road?
X_2	Yes, as always.
Y_2	Please pay.
X_3	Payment completed.
R	<Task-Completion-Feedback>

Personalized dialogue is much shorter!

TRL for Cross-domain Dialogue Policy Transfer

- Less training data / Better performance
- Leverage previous policy



	Coffee Shopping Dialogue (X=customer, Y=system)
X_1	Can I have coffee please?
Y_1	What coffee would you like?
X_2	I would like a cup of Mocha.
Y_2	Hot Mocha or Iced Mocha?
X_3	Iced Mocha.

State: [Coffee, Temp]

Transfer Dialogue Policy from Source

Source: Restaurant
State: [food]
Action: request(area)

Action: request(size)

Y_3	**What cup size do you want?**

Target Q-function: 4.09

Thesis Summary

Task		Progress
Personalized Coffee Shopping Dialogue System		Industrial Demo
A survey of Task-oriented Dialogue Systems		Ph.D Qualification Exam
Personal Dialogue System	Personal Q-function	Done [AAAI 2018]
	Personal Word Gate	Done [IJCAI 2018 submitted]
Simultaneous Dialogue Policy Transfer across Domains and Speech-Acts		Done [IJCAI 2018 submitted]

Thank you

Appendix: All Datasets and Evaluation Metrics

- Spoken Language Understanding (SLU)
- Dialogue State Tracking (DST)
- Dialogue Policy (Policy)
- Natural Language Generation (NLG)

Dataset	Task	Evaluation Metrics
ATIS [9]	SLU	Accuracy, F1
TouristInfo [5]	SLU, NLG	Accuracy, F1, BLEU
DARPA Communicator [62]	SLU	Accuracy, F1
DSTC 1 [68]	DST	Accuracy, L2
DSTC 2 [23]	DST	Accuracy, L2
DSTC 3 [24]	DST	Accuracy, L2
TownInfo [59]	DST, Policy	Accuracy, L2, Reward, Success Rate, Number of Dialog Turns
Cambridge Restaurant [18]	Policy, NLG	Reward, Success Rate, Number of Dialog Turns, BLEU
SF Restaurant [14]	Policy, NLG	Reward, Success Rate, Number of Dialog Turns, BLEU
SF Hotel [14]	Policy, NLG	Reward, Success Rate, Number of Dialog Turns, BLEU
L11 [14]	Policy, NLG	Reward, Success Rate, Number of Dialog Turns, BLEU
Babi Task [6]	Policy, NLG	Accuracy, BLEU

Reference

- [Williams and Zweig 2016] Williams, J. D., and Zweig, G.2016. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. arXiv preprint arXiv:1606.01269
- [Williams 2008a] Williams, Jason D. "The best of both worlds: unifying conventional dialog systems and POMDPs." INTERSPEECH. 2008.
- [Williams 2008b] Williams, Jason D. "Integrating expert knowledge into POMDP optimization for spoken dialog systems." Proceedings of the AAAI-08 Workshop on Advancements in POMDP Solvers. Vol. 2. No. 3. 2008.
- [Young et al. 2010] Young, Steve, et al. "The hidden information state model: A practical framework for POMDP-based spoken dialogue management." Computer Speech & Language 24.2 (2010): 150-174.
- [Lefèvre et al. 2009] Lefèvre, Fabrice, et al. "k-Nearest neighbor Monte-Carlo control algorithm for POMDP-based dialogue systems." Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics, 2009.
- [Li et al. 2009] Li, Lihong, Jason D. Williams, and Suhrid Balakrishnan. "Reinforcement learning for dialog management using least-squares Policy iteration and fast feature selection." INTERSPEECH. 2009.
- [Gaši´c and Young 2014] Gasic, Milica, and Steve Young. "Gaussian processes for pomdp-based dialogue manager optimization." IEEE/ACM Transactions on Audio, Speech, and Language Processing 22.1 (2014): 28-40.
- [Gaši´c et al. 2013a] Gašić, Milica, et al. "On-line policy optimisation of bayesian spoken dialogue systems via human interaction." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.
- [Daubigney et al. 2012b] Daubigney, Lucie, Matthieu Geist, and Olivier Pietquin. "Off-policy learning in large-scale POMDP-based dialogue systems." Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.

Reference

- [Gasic et al. 2014] Gašic, Milica, et al. "Incremental on-line adaptation of pomdp-based dialogue managers to extended domains." In Proceedings on InterSpeech (2014).
- [Gašic et al. 2013b] Gasic, Milica, et al. "POMDP-based dialogue manager adaptation to extended domains." Proceedings of the SIGDIAL 2013 Conference. 2013.
- [Gaši´c et al. 2015a] Gašić, Milica, et al. "Distributed dialogue policies for multi-domain statistical dialogue management." Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015.
- [Gaši´c et al. 2015b] Gašić, M., et al. "Policy committee for adaptation in multi-domain spoken dialogue systems." Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE, 2015.
- [Genevay and Laroche 2016] Genevay, Aude, and Romain Laroche. "Transfer learning for user adaptation in spoken dialogue systems." Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2016.
- [Casanueva et al. 2015] Casanueva, Inigo, et al. "Knowledge transfer between speakers for personalised dialogue management." Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2015.
- [Mo et al. 2016] Mo, Kaixiang, et al. "Personalizing a Dialogue System with Transfer Learning." arXiv preprint arXiv:1610.02891 (2016).

Reference

- [Bordes et al. 2016] Bordes, Antoine, and Jason Weston. "Learning end-to-end goal-oriented dialog." *arXiv preprint arXiv:1605.07683* (2016).
- [Thomson and Young 2010] Blaise Thomson and Steve Young. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588, 2010.
- [Gašic and Young 2014] Milica Gaši´c and Steve Young. Gaussian processes for pomdp-based dialoguemanager optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):28–40, 2014.
- [Gaši´c *et al.* 2015b] M Gaši´c, N Mrkši´c, PH Su, David Vandyke, Tsung-HsienWen, and S Young. Policy committee for adaptation in multi-domain spoken dialogue systems. 2015.