On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems

Pei-Hao (Eddy) Su Milica Gašić Nikola Mrkšić Lina Rojas-Barahona
Stefan Ultes David Vandyke Tsung-Hsien Wen Steve Young

Dialogue Systems Group

UNIVERSITY OF CAMBRIDGE

Task-oriented Dialogue System

UNIVERSITY OF CAMBRIDGE

With paid users

Task:
- Find a ~~restaurant~~, ~~**Chinese**~~, ~~**cheap**~~, ~~**west**~~
- Ask **phone**, ~~**address**~~

Hi, How may I help you?

I want a cheap Chinese Restaurant.

Where in the city would you like?

Somewhere in the west, please.

Yim Wah is a nice Chinese place.

Great, can you give me the address?

It is at 2-4 Lensfild Road.

Ok, thank you, bye!

Thanks, goodbye.

Success evaluation

By task: Fail
(no **phone**)

By user: Success
(get all he asked)

Goal

UNIVERSITY OF CAMBRIDGE

Defining suitable learning objective (**reward**)
to train a dialogue system on-line from **real users**

Statistical Spoken Dialogue System

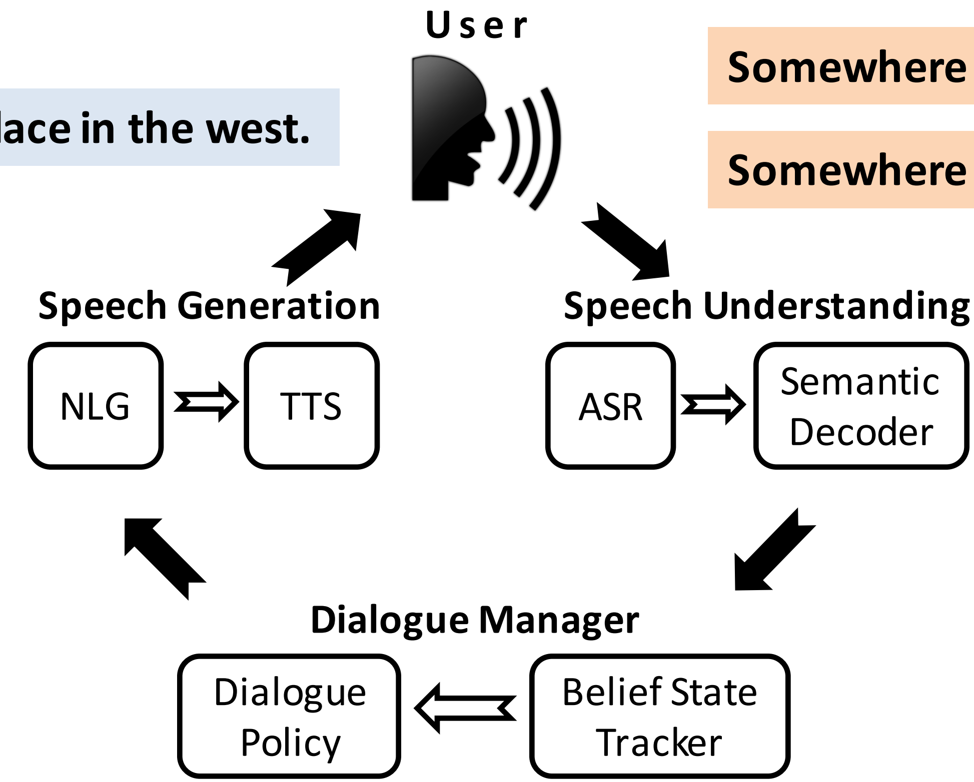
UNIVERSITY OF CAMBRIDGE

Yim Wah is a nice place in the west.

Somewhere in the west, please.

Somewhere in the wet, please.

⋮



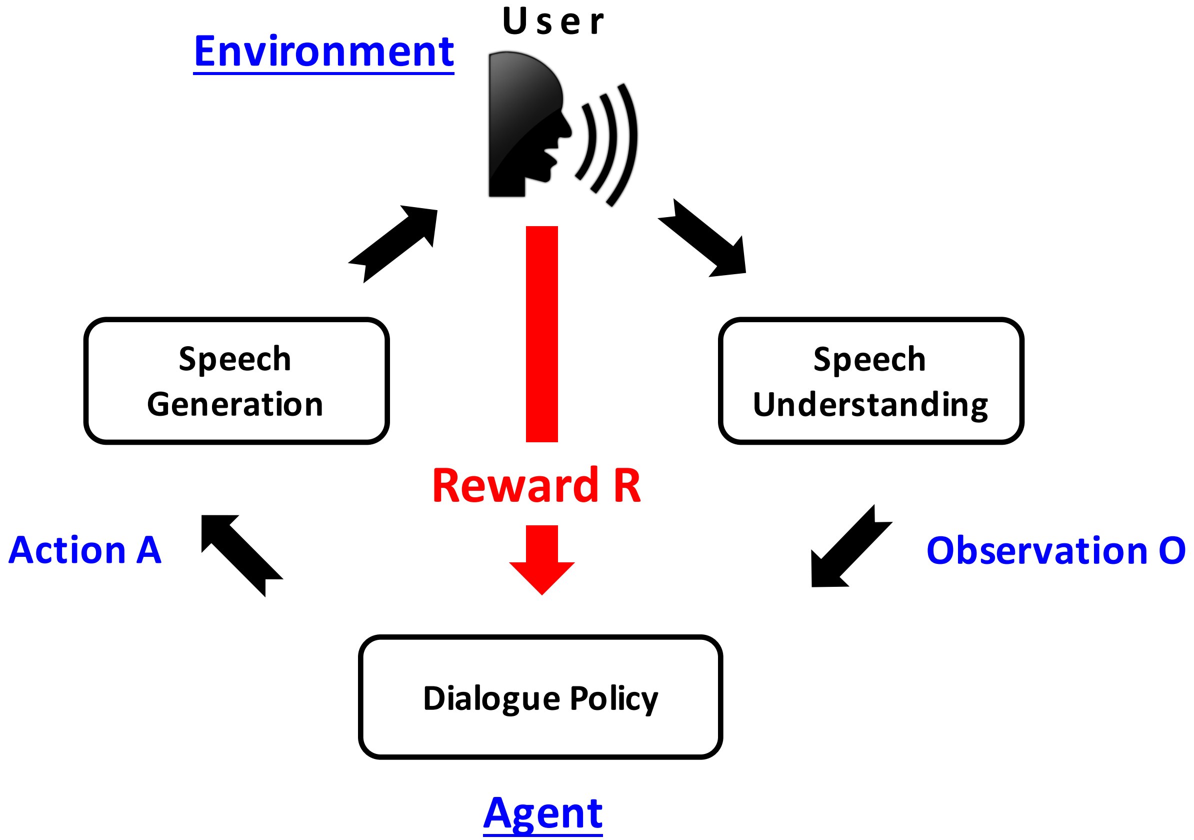
Inform(name=Yim Wah, area=west)

Area

East	West	...	None
0.01	0.94	0	0.05

Dialogue Manager in RL framework

UNIVERSITY OF CAMBRIDGE



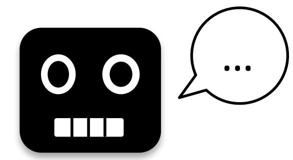
Correct rewards are a crucial factor in dialogue policy training

The Reinforcement Signal in SDS

UNIVERSITY OF CAMBRIDGE

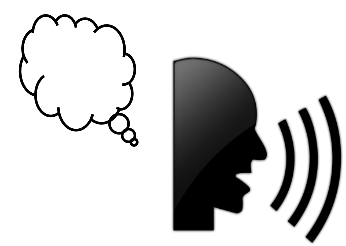
Typical Reward Function:

- per turn penalty -1
- Large reward at completion if successful



➢ Typically requires prior knowledge of the task

- ✔ Simulated user
- ✔✘ Paid users (Amazon Mechanical Turk)
- ✘ Real users



The Reinforcement Signal in SDS

UNIVERSITY OF CAMBRIDGE

How to learn policy from real users?

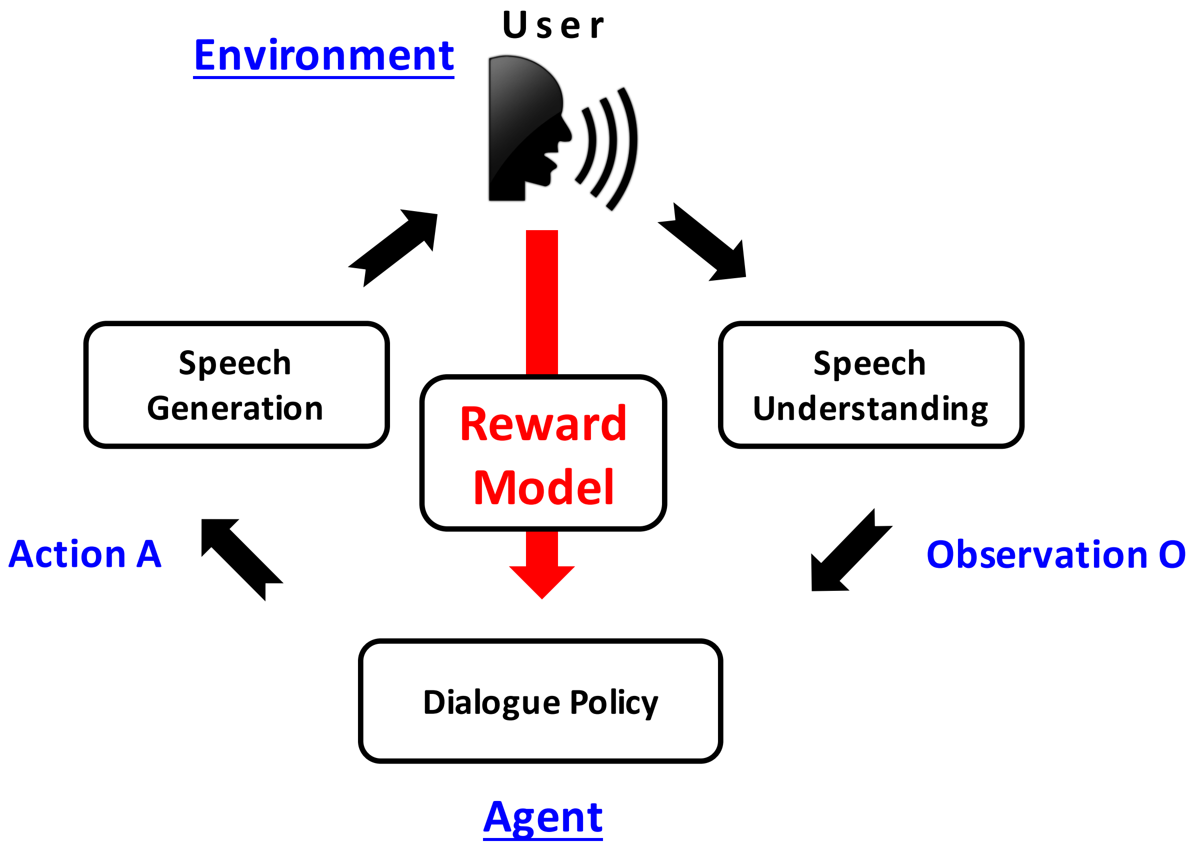
- Infer success (reward) directly from dialogues
  - Train a reward estimator from data

    (Su & Vandyke et. al, Interspeech 2015)

- ~~User rating~~ Robust reward model on user rating
  - Noisy → Gaussian Process with uncertainty
  - Difficult/Costly to obtain → Active Learning

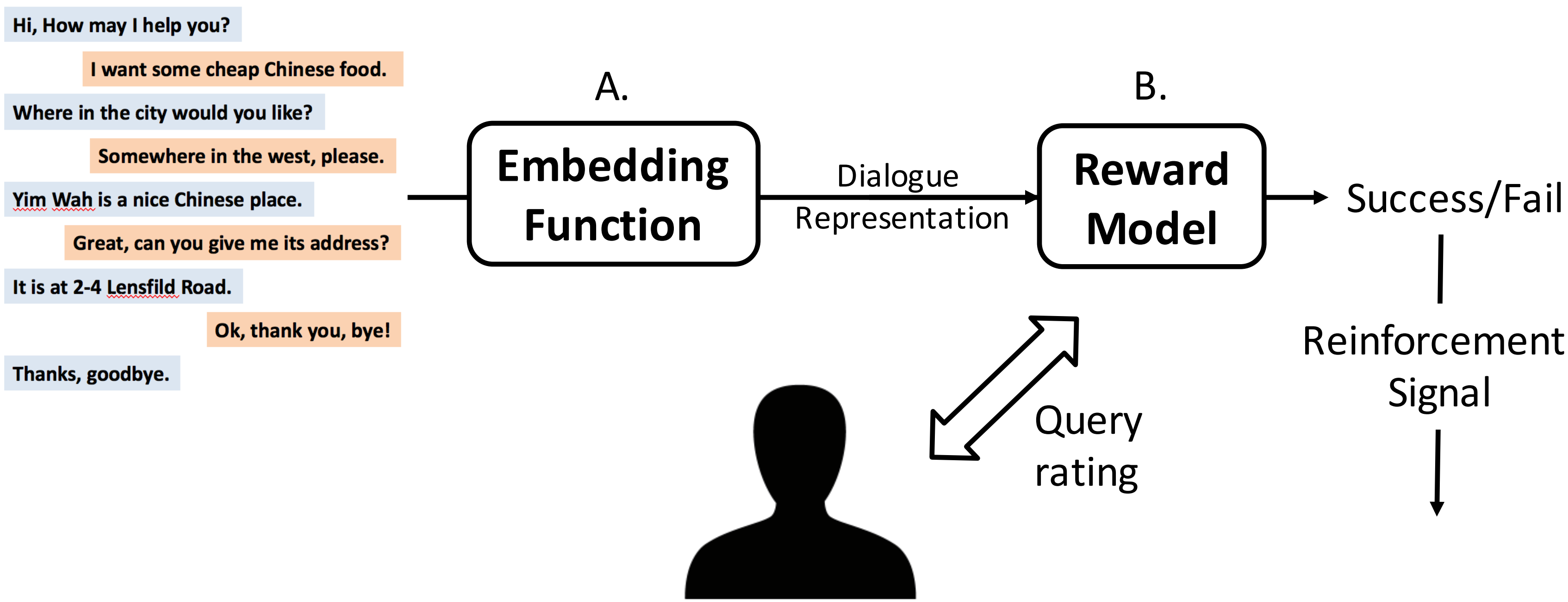
System Framework

UNIVERSITY OF CAMBRIDGE



System Framework

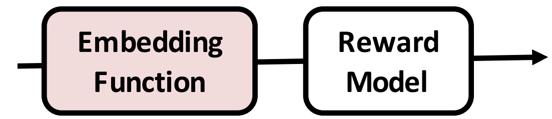
Reward modelling on user binary success rating

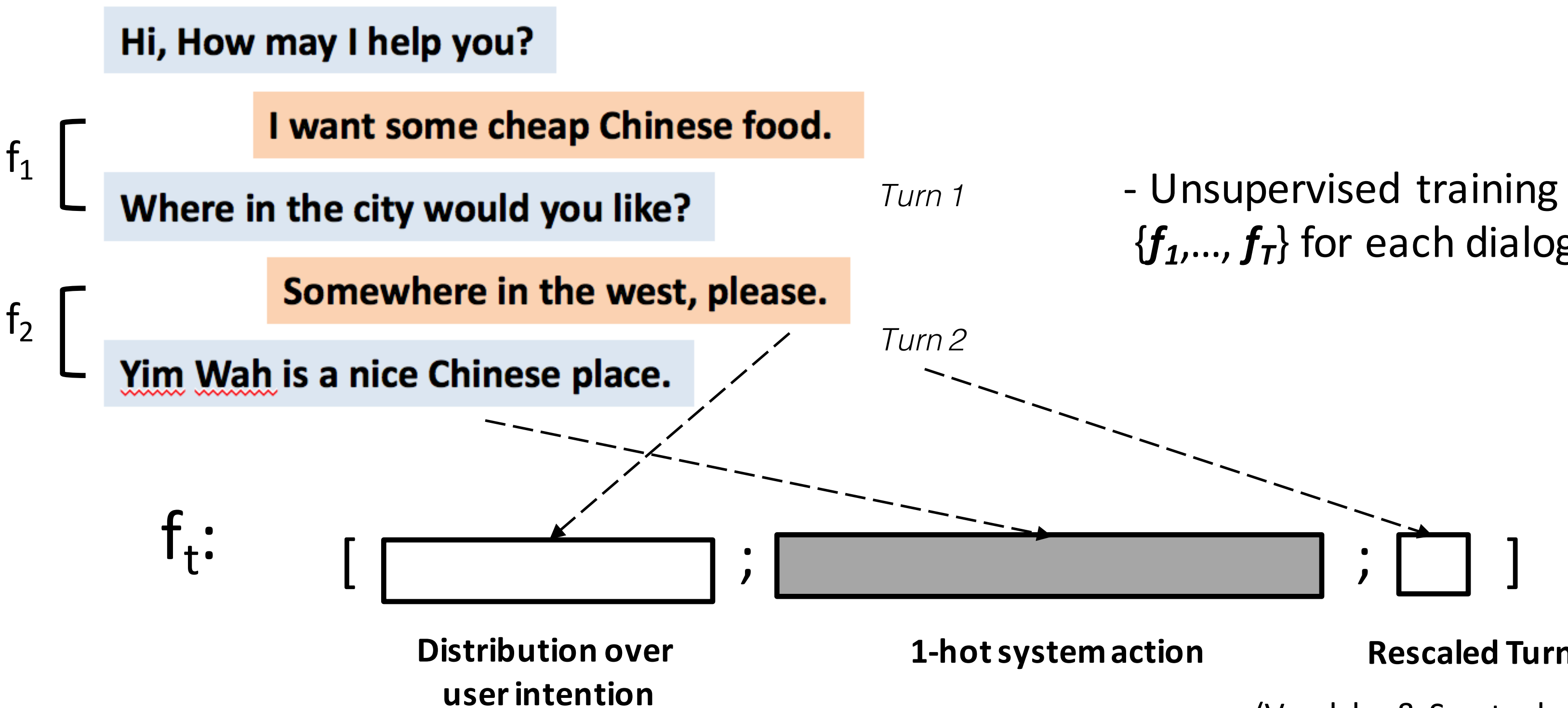


A. Unsupervised Dialogue Embedding

UNIVERSITY OF CAMBRIDGE

Mapping a dialogue sequence to a fixed-length vector





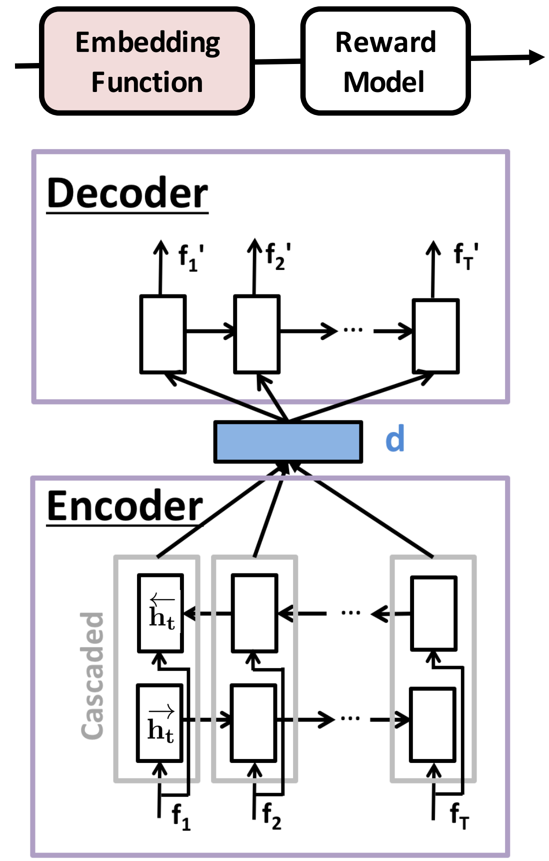
- Unsupervised training data: $\{f_1, \ldots, f_T\}$ for each dialogue

(Vandyke & Su et. al, ASRU 2015)

A. Unsupervised Dialogue Embedding

UNIVERSITY OF CAMBRIDGE

Bi-directional LSTM Encoder-Decoder



- Reconstruct inputs with variable-lengths
- $\mathbf{h_t}=[\overrightarrow{\mathbf{h_t}};\overleftarrow{\mathbf{h_t}}]$ captures forward-backward info
- Bottleneck d is the dialogue representation

$$\mathbf{d} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h_t}$$

- MSE training criterion:

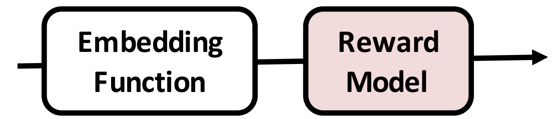
$$MSE = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}||\mathbf{f}_t - \mathbf{f}'_t||^2$$

  - f_t: input/target, f'_t: prediction

B. Active Reward Learning Model

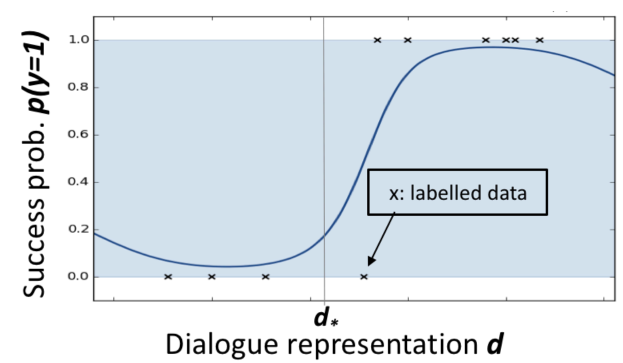
UNIVERSITY OF CAMBRIDGE

Gaussian process classifier for success rating



- **<u>Noise term</u>** in the RBF kernel affects uncertainty
  - More noise -> less certain

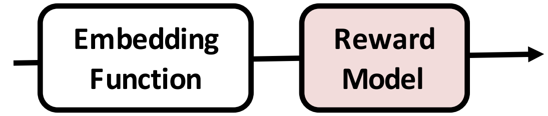
$$k(\mathbf{d}, \mathbf{d}') = \underbrace{p^2 \exp(-\frac{||\mathbf{d} - \mathbf{d}'||^2}{2l^2})}_{\text{Input correlation}} + \underbrace{\sigma_n^2}_{\text{User rating noise}}$$



B. Active Reward Learning Model

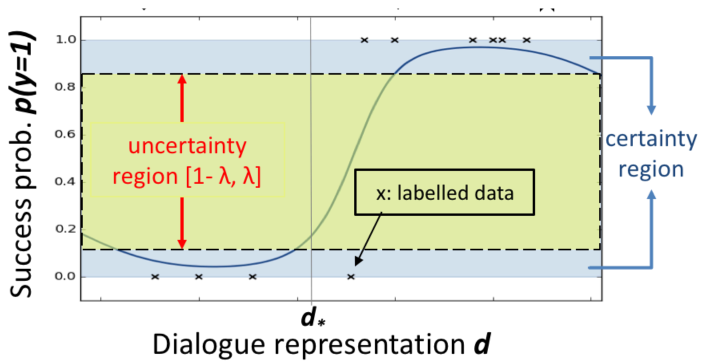
UNIVERSITY OF CAMBRIDGE

Gaussian process classifier for success rating



- **<u>Noise term</u>** in the RBF kernel affects uncertainty
  - More noise -> less certain
- **<u>Active learning</u>**: uncertainty + threshold
  - λ: when to actively query user rating

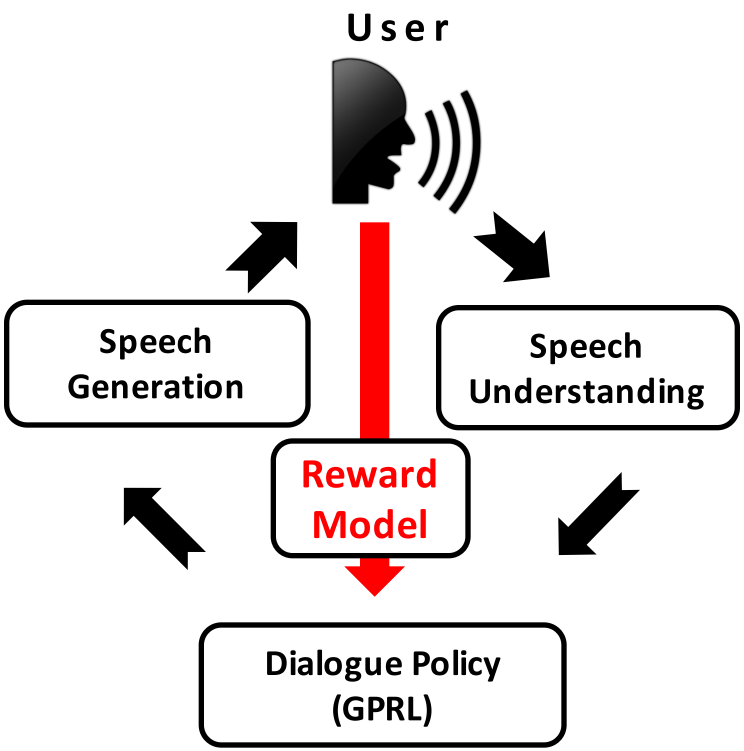
$$k(\mathbf{d}, \mathbf{d}') = \underbrace{p^2 \exp\left(-\frac{||\mathbf{d} - \mathbf{d}'||^2}{2l^2}\right)}_{\text{Input correlation}} + \underbrace{\sigma_n^2}_{\text{User rating noise}}$$



System Setup

UNIVERSITY OF CAMBRIDGE

Embed the reward model in SDS



- Cambridge restaurant domain:
  - 150 venues
  - 3 informable slots: area, price range, food
  - 3 requestable slots: addr, phone, postcode
- Reward:
  - per turn -1,
  - When dialogue ends, binary (0/1) * 20:

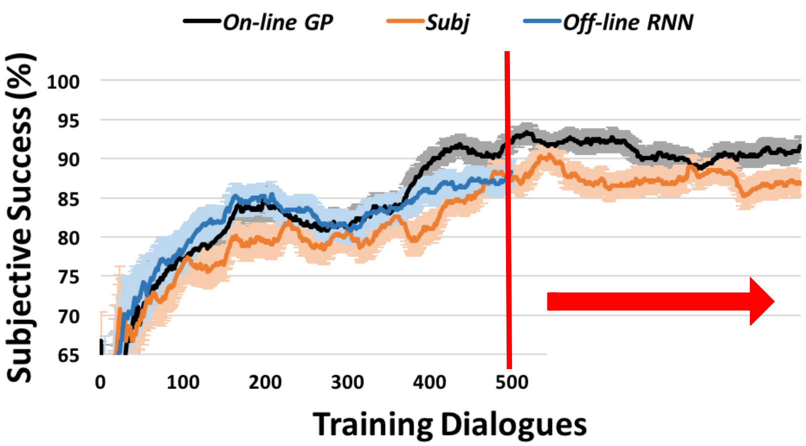
- On-line GP	Proposed method
- Subj	User rating only
- Off-line RNN (Su. et al. 2015)	RNN with 1K simulated data

- Crowd-sourced users from Amazon Mechanical Turk

On-line Dialogue Reward & Policy Learning

UNIVERSITY OF CAMBRIDGE

Dialogue policy learning with real users



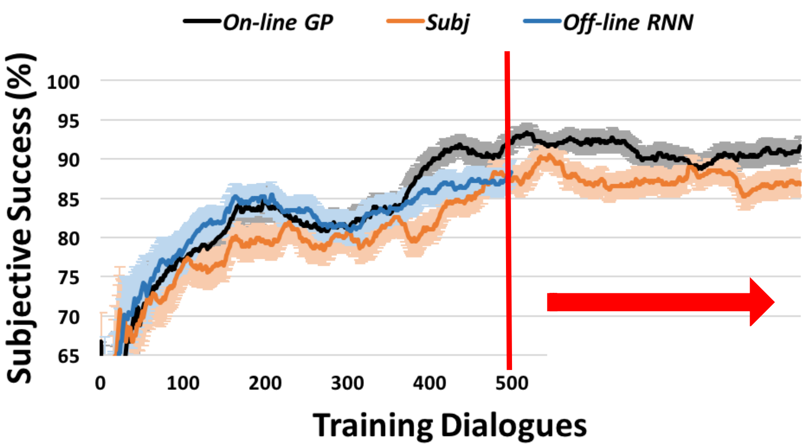
Dialogues	Reward Model	Subjective (%)
400 - 500	*Off-line RNN*	89.0 +- 1.8
	Subj	90.7 +- 1.7
	On-line GP	91.7 +- 1.6
500 - 850	*Subj*	87.1 +- 1.0 *
	On-line GP	**90.9 +- 0.9** *

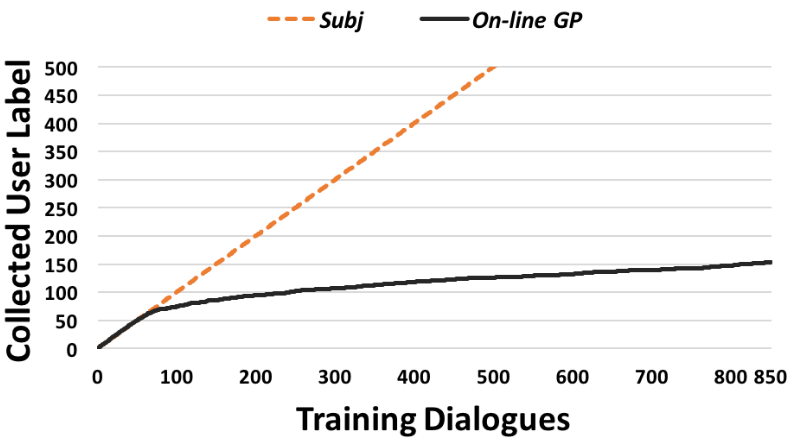
- All reached **> 85 %** after 500 dialogues
- *On-line GP* is more robust than *Subj* in longer run

On-line Dialogue Reward & Policy Learning

UNIVERSITY OF CAMBRIDGE

Dialogue policy learning with real users





- All reached **> <u>85 %</u>** after 500 dialogues
- *On-line GP* is more robust than *Subj* in longer run
- *On-line GP* needs only 150 queries from user rating

Conclusion

UNIVERSITY OF CAMBRIDGE

- Proposal: an **on-line active reward learning** framework
  - Unsupervised Dialogue Embedding: Bi-LSTM Encoder Decoder
  - Active Reward Model: GP Classifier with uncertainty threshold
  - Reduce **data annotation** and mitigate **noisy user rating**
  - No need of **labelled data** and **user simulator**

- Achieve **truly on-line policy learning** from real users w/o task info

Thank You!
Questions?

Data available at *http://goo.gl/EdM99V*

phs26@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

References

UNIVERSITY OF CAMBRIDGE

1. <u>Pei-Hao Su</u>, Milica Gašić, Nikola Mrkšić, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen and Steve Young, **"On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems"**. *In Proceeding of ACL 2016*

2. <u>Pei-Hao Su</u>, David Vandyke, Milica Gašić, Dongho Kim, Nikola Mrkšić, Tsung-Hsien Wen and Steve Young, **"Learning from Real Users: Rating Dialogue Success with Neural Networks for Reinforcement Learning in Spoken Dialogue Systems"**. *In Proceeding of Interspeech 2015*
3. David Vandyke, <u>Pei-Hao Su</u>, Milica Gašić, Nikola Mrkšić, Tsung-Hsien Wen and Steve Young, **"Multi-Domain Dialogue Success Classifiers for Policy Training"**. *In Proceeding of ASRU 2015*

Example Dialogues – Low Noise

UNIVERSITY OF CAMBRIDGE

-Low noise condition- (User Subjective Evaluation: *Success*, **on-line GP Evaluation:** *Success*)

Machine : Hello, welcome to the Cambridge dialogue system, What kind of food would you like?
Human : [Top ASR] Yeah I want to find a restaurant that serves European food.
[Top SEM] inform(food=european,type=restaurant)
Machine : Hotel du Vin and Bistro is a nice place. It serves European food.
Human : [Top ASR] Uh what is the address and phone number?
[Top SEM] request(phone,addr)
Machine : Hotel du Vin and Bistro is a nice place.
Their address is 15-19 Trumpington Street and their phone is 01223 227330.
Human : [Top ASR] Thank you goodbye.
[Top SEM] bye()
Machine : Thank you for using this system.

Example Dialogues – High Noise

UNIVERSITY OF CAMBRIDGE

-Higher noise condition- (User Subjective Evaluation: *Success*, **on-line GP Evaluation:** *Success***)**

Machine : Hello, welcome to the Cambridge dialogue system, What kind of food would you like?
Human : [Top ASR] And I want to find a expensive restaurant it should be in the south part of town.
[Top SEM] inform(area=south,pricerange=expensive,type=restaurant)
Machine : Gourmet burger kitchen is a nice place it is in the expensive price range.
Human : [Top ASR] And do you have something else?
[Top SEM] null()
[2nd SEM] reqalts()
Machine : Did you say you are looking for a venue in the area of south ?
Human : [Top ASR] Yes.
[Top SEM] affirm()
Machine : Frankie and Benny's is a nice place, it is in the south part of town and it is in the expensive price range.
Human : [Top ASR] Okay can I have the uh address phone number called for any.
[Top SEM] request(addr)
[2nd SEM] request(phone,addr)
Machine : Frankie and Benny's is a nice place.
Their address is Clifton Way and their phone number is 01223 412430.
Human : [Top ASR] What is the pub called for free you any?
...
[5th ASR] What is the post code for free you any?
[Top SEM] request(postcode)
Machine : Frankie and Benny's is a nice place. Their postcode is C.B 1, 7 D.Y.
Human : [Top ASR] Thank you goodbye.
[Top SEM] bye()
Machine : Thank you for using this system.