Reward Estimation for Dialogue Policy Optimisation

Pei-Hao (Eddy) Su

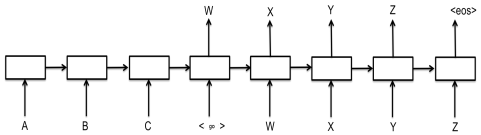
DeepHack.Turing , 25 July 2017

UNIVERSITY OF CAMBRIDGE

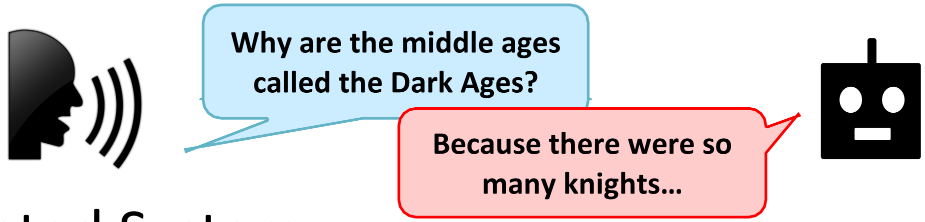
Dialogue Systems

UNIVERSITY OF CAMBRIDGE

- Chat-based Agents
  - Hope to talk about everything (open domain)
  - No specific goal, focus on conversation flow



Variants of Seq2Seq model:
[Vinyals and Le 2015]
[Serban et al 2016]
[Al-Rfou et al. 2016]
[Li et al. 2016]



- Task-oriented System
  - Achieve a certain task (closed domain)
  - Combination of rules and statistical components
  - Ground language using a knowledge base (ontology)
    - Pipeline dialogue systems [Henderson et al. 2005 , Williams and Young 2007]
    - End-to-End dialogue systems [Antoine et al. 2017, Wen et al. 2017]

Task-oriented Dialogue System

UNIVERSITY OF CAMBRIDGE

With paid subjects

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

Not Practical

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

Objective: Fail
(no **phone**)

Subjective: Success
(get all he asked)

Ambiguity

Goal

UNIVERSITY OF CAMBRIDGE

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

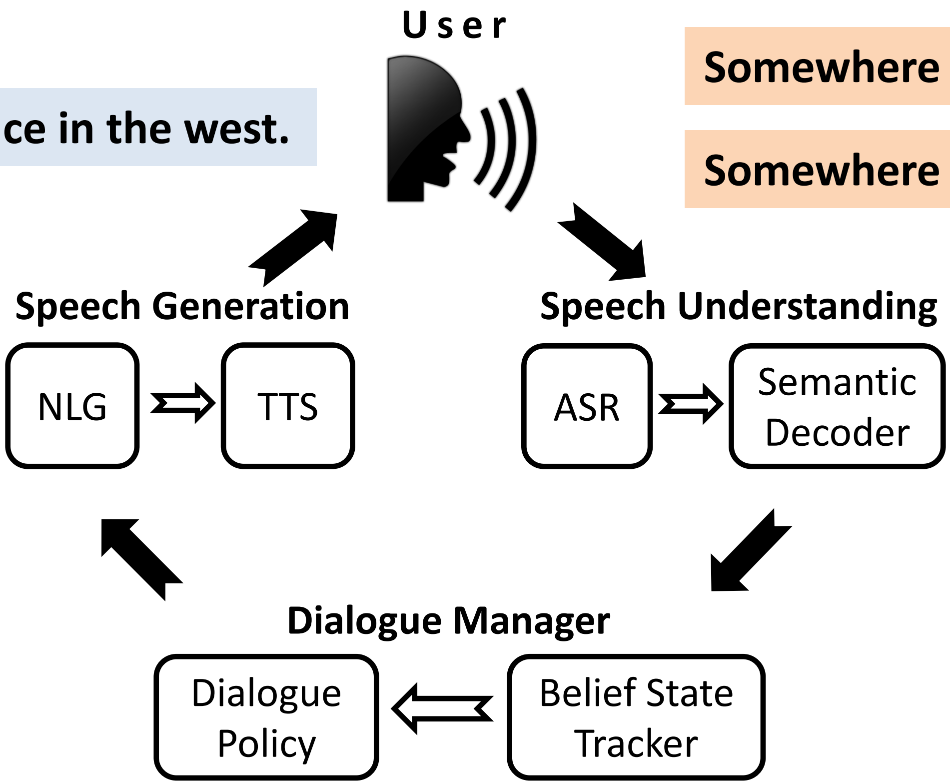
- Tasks
  - Evaluate the dialogue (reward modelling)
  - Deal with unreliable user rating
  - Learn a dialogue policy
- Models
  - Recurrent neural networks, Gaussian processes
- Methods
  - Reinforcement learning, On-line learning, Active learning

Outline

Pipeline 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.

⋮

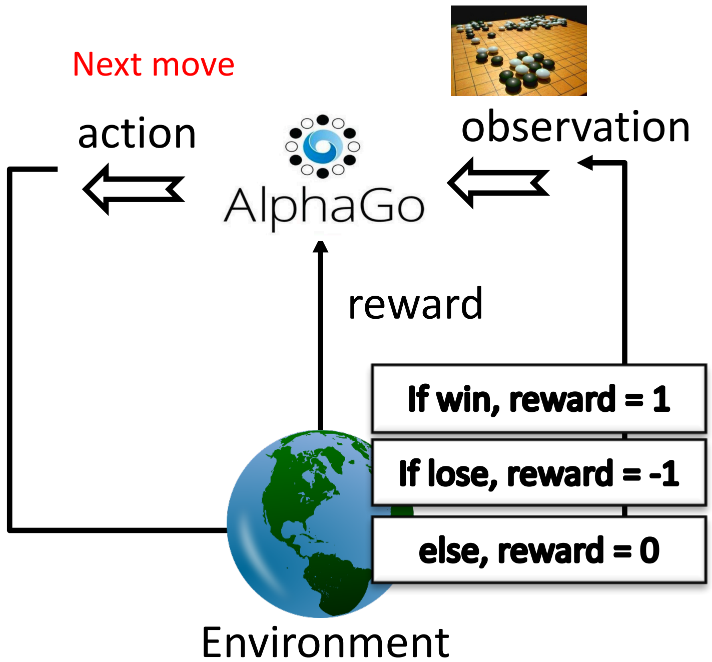
Inform(name=Yim Wah, area=west)

Area

East	West	...	None
0.01	0.94	0	0.05

Reinforcement Learning 101

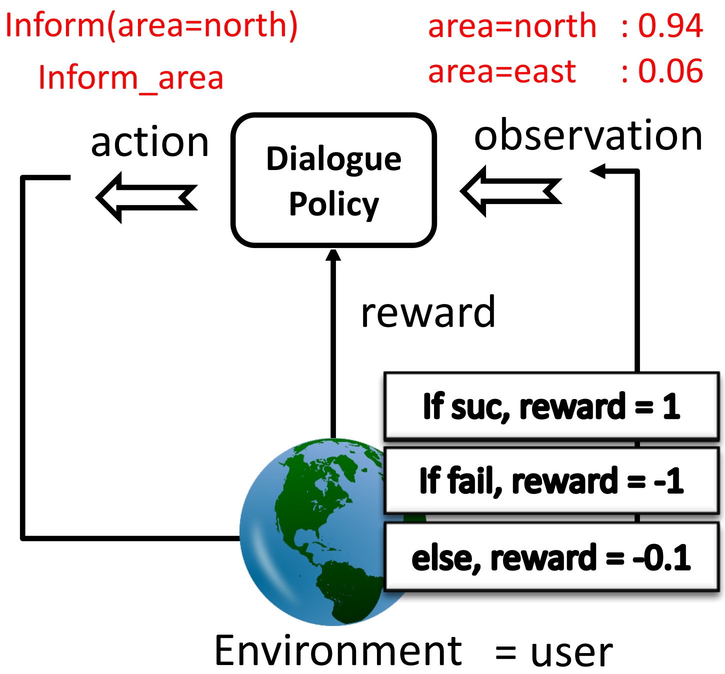
UNIVERSITY OF CAMBRIDGE



Agent learns to take actions
to maximise total reward

It beat GO champions in 2016 and 2017

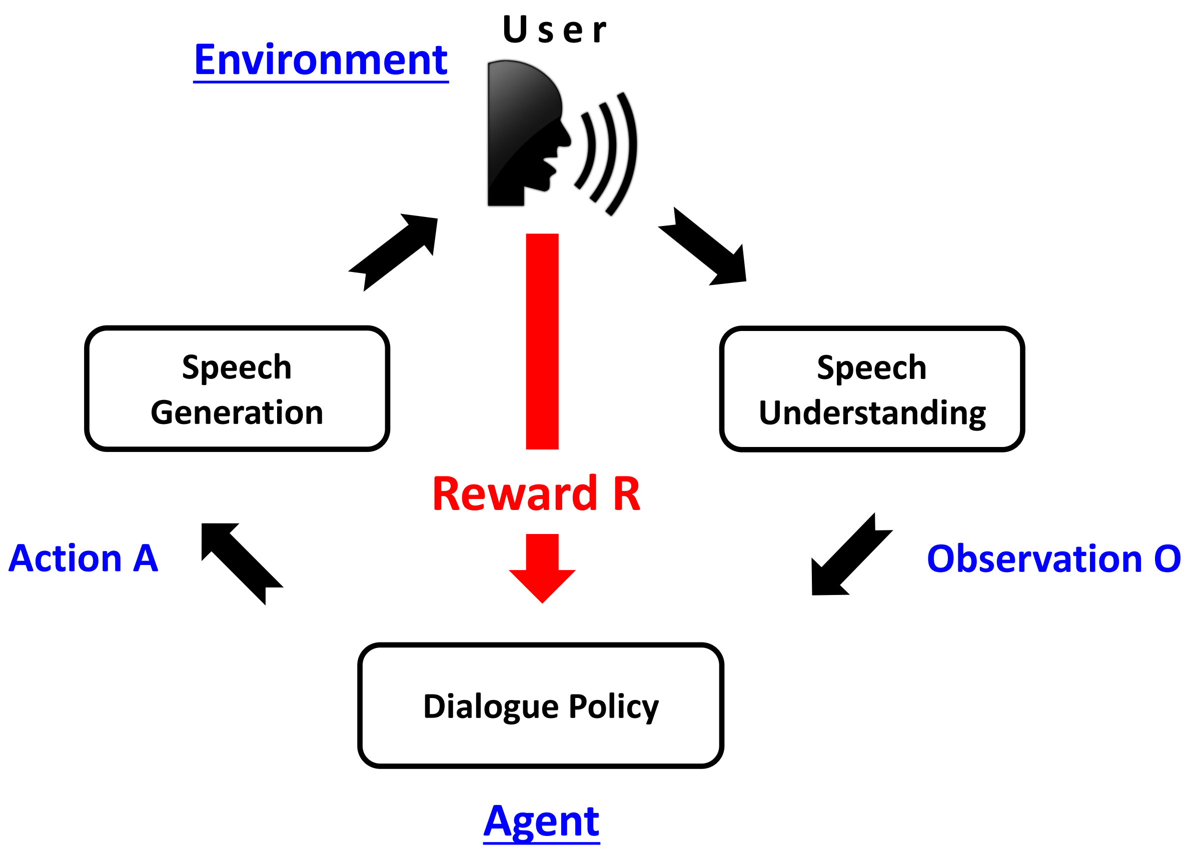
Reinforcement Learning 101



Agent learns to take actions
to maximise total reward

Dialogue Manager in RL framework

UNIVERSITY OF CAMBRIDGE



Correct rewards are a crucial factor in dialogue policy training

Reward for RL $\cong$ Evaluation for SDS

UNIVERSITY OF CAMBRIDGE

- Dialogue is a special RL task:
  - Human involves in **interaction** and **rating (evaluation)** of a dialogue
  - **Human**-in-the-loop framework: human is troublesome but useful

- Rating: correctness, appropriateness, and adequacy

- Expert rating	high quality, high cost
- User rating	unreliable quality, medium cost
- Objective rating	Check desired aspects, low cost

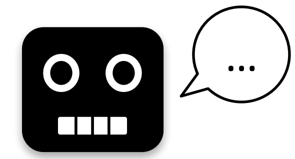
The Reinforcement Signal in SDS

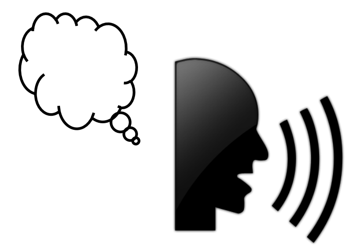
UNIVERSITY OF CAMBRIDGE

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



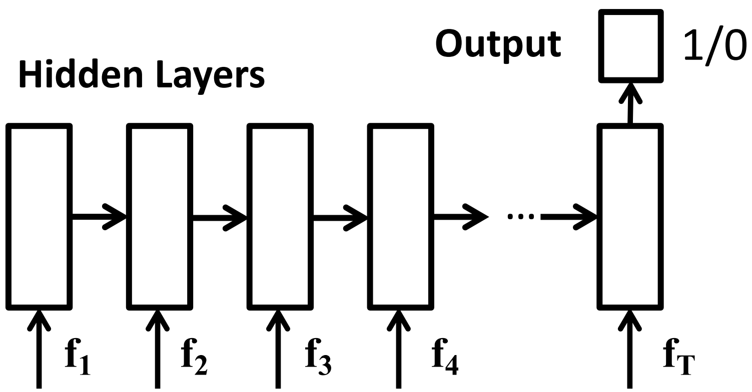


The Reinforcement Signal in SDS

UNIVERSITY OF CAMBRIDGE

How to learn policy from real users?

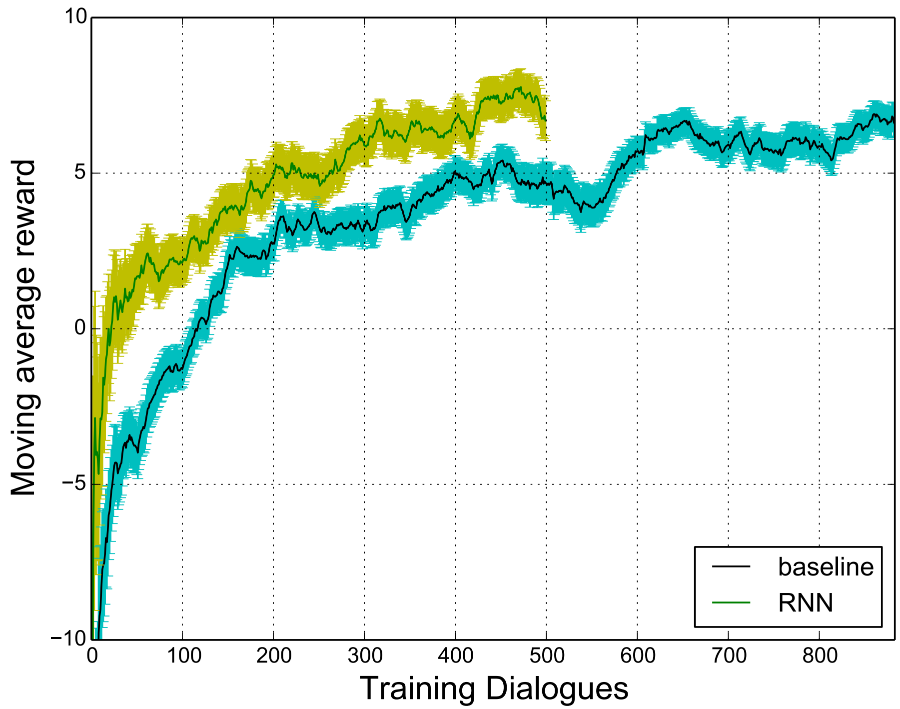
- Infer success (reward) directly from dialogues
  - Train a reward estimator from data (Su et al. 2015)



The Reinforcement Signal in SDS

UNIVERSITY OF CAMBRIDGE

RNN Reward Estimator for Policv Learning



Objective-Baseline

- Needs task info.
- Learns only from *Obj=Subj* dialogue (500 out of ~900)

RNN-system

- No task/user feedback
- Learns from every dialogue (all 500)

RNN-system learnt policy more practically and efficiently than *Objective*-baseline

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 et al. 2015)

- User rating
  - Noisy
  - Difficult/Costly to obtain

- Robust user rating model (Su et al. 2016)
  - ~~Noisy~~ → Gaussian Process with uncertainty
  - ~~Difficult/Costly~~ → Active Learning

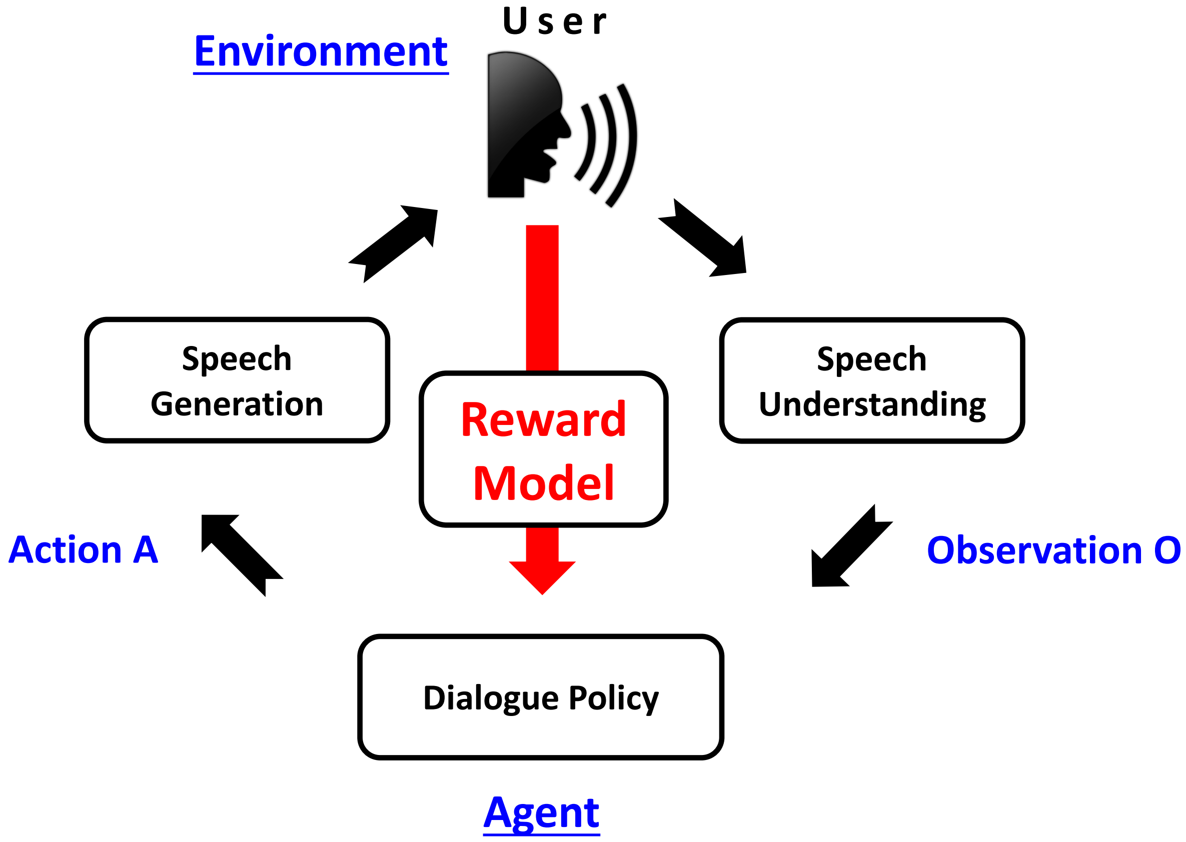
Outline

UNIVERSITY OF CAMBRIDGE

1. Motivation – Learning from human users
2. Proposed Framework
3. Experiment
4. Conclusion

System Framework

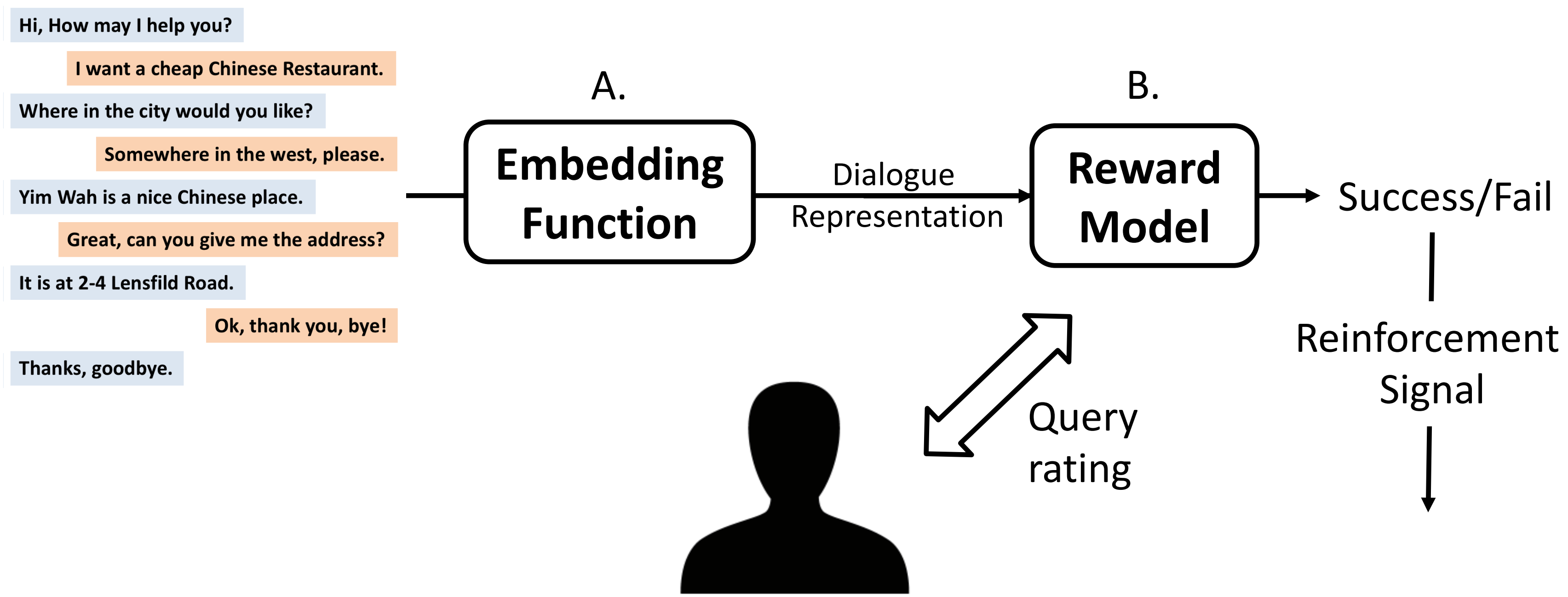
UNIVERSITY OF CAMBRIDGE



System Framework

UNIVERSITY OF CAMBRIDGE

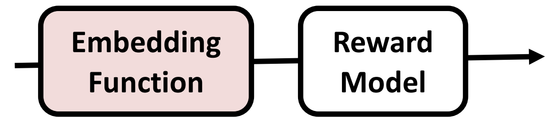
Reward modelling on user binary success rating

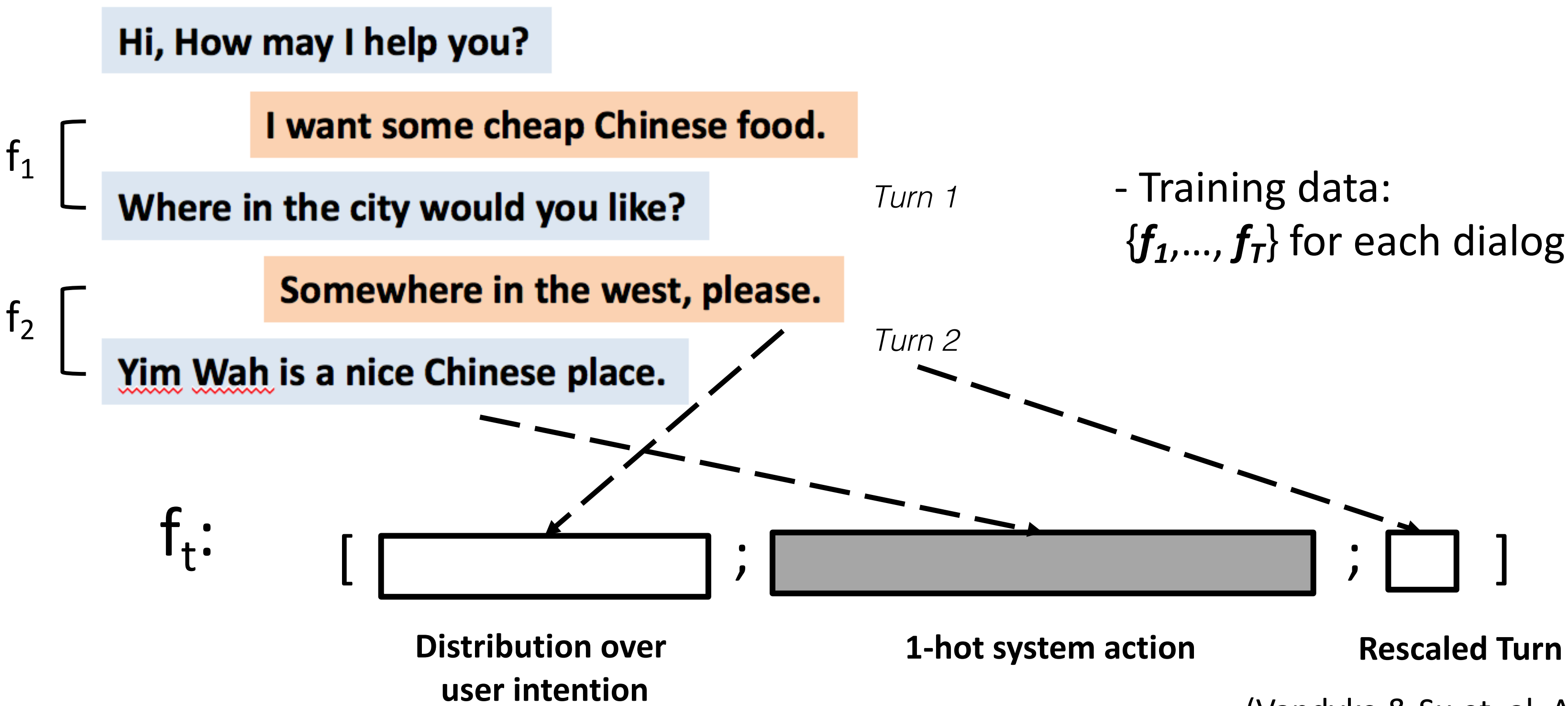


A. Dialogue Embedding

UNIVERSITY OF CAMBRIDGE

Maps a dialogue seq to a fixed-length vector



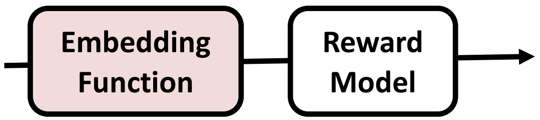


- Training data: $\{f_1, \ldots, f_T\}$ for each dialogue

(Vandyke & Su et. al, ASRU 2015)

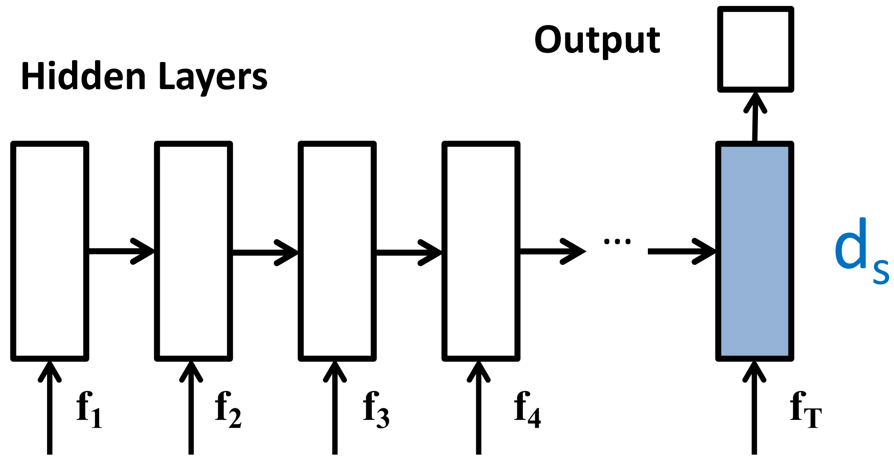
A. Dialogue Embedding - Supervised

UNIVERSITY OF CAMBRIDGE



Re-use the supervised RNN

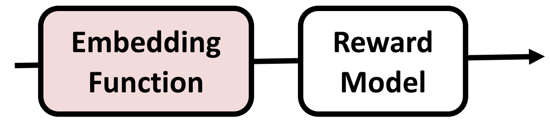
- Last hidden layer as dialogue representation



A. Dialogue Embedding - Unsupervised

UNIVERSITY OF CAMBRIDGE

Bi-LSTM Encoder-Decoder (Seq2Seq)



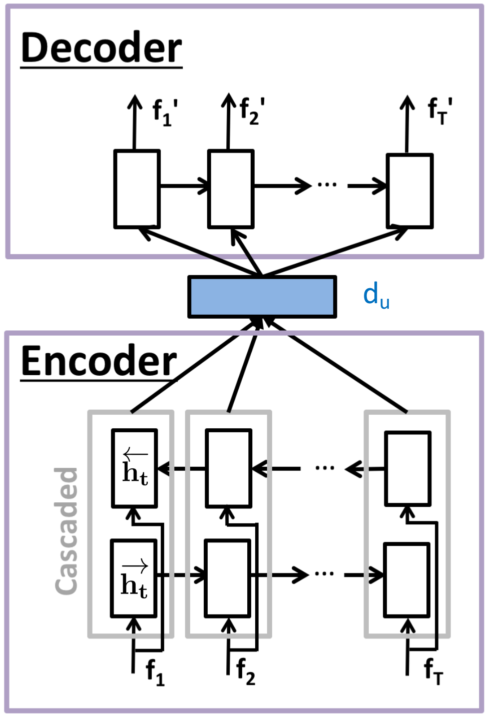
- Reconstruct inputs with variable-lengths
- $\mathbf{h_t}=[\overrightarrow{\mathbf{h_t}};\overleftarrow{\mathbf{h_t}}]$ captures forward-backward info
- Bottleneck d_u is the dialogue representation

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

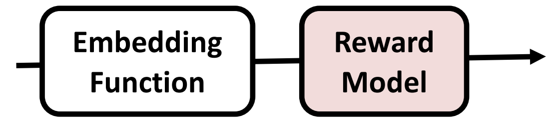
- MSE training criterion:

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

  - $\boldsymbol{f_t}$: input/target, $\boldsymbol{f'_t}$: prediction



B. Active Reward Learning Model

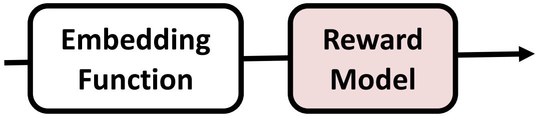


- Determine class probability: $p(y|\boldsymbol{d}, D)$, given $D = \{(\boldsymbol{d}_i, y_i)\}_{i=1}^{n}$
  - where $y = \{+1, -1\}$
- Handle the issue of noisy and costly user rating
- **Gaussian process** (GP) with active learning

B. Active Reward Learning Model

UNIVERSITY OF CAMBRIDGE

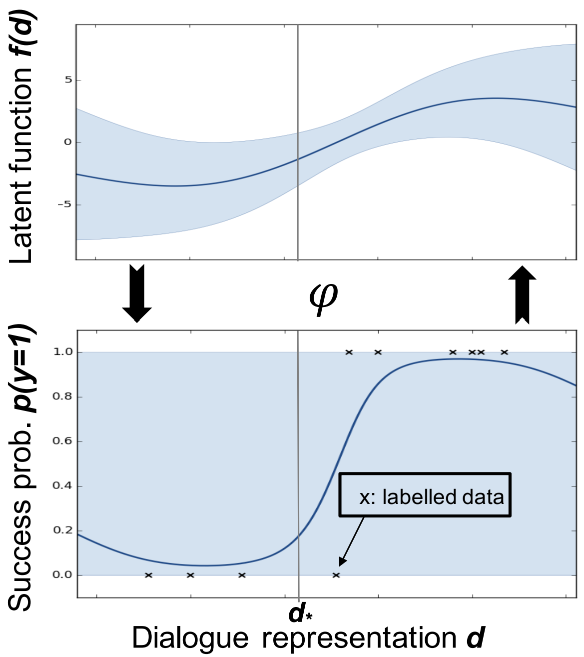
Gaussian process classifier for success rating



- GP is shown useful in policy learning (Gasic ’14, Casanueva ’15)
  - Learn from few observations
  - Provides a measure of uncertainty

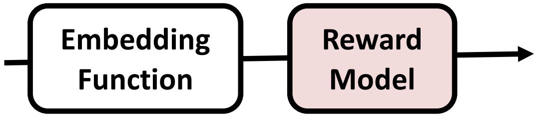
- $p(y=1|\boldsymbol{d},D)=\varphi(f(\boldsymbol{d}|D))$
  - f : latent function: $R^{\dim(\boldsymbol{d})}\rightarrow R$
  - φ: probit function: $R\rightarrow[0,1]$

- $f(\boldsymbol{d})\sim GP(m(\boldsymbol{d}),k(\boldsymbol{d},\boldsymbol{d}'))$
  - $k(\mathbf{d},\mathbf{d}')=p^2\exp(-\frac{||\mathbf{d}-\mathbf{d}'||^2}{2l^2})$



B. Active Reward Learning Model

Gaussian process classifier for success rating



- Prior:

  $f(\boldsymbol{d}) \sim GP(m(\boldsymbol{d}), k(\boldsymbol{d}, \boldsymbol{d}'))$

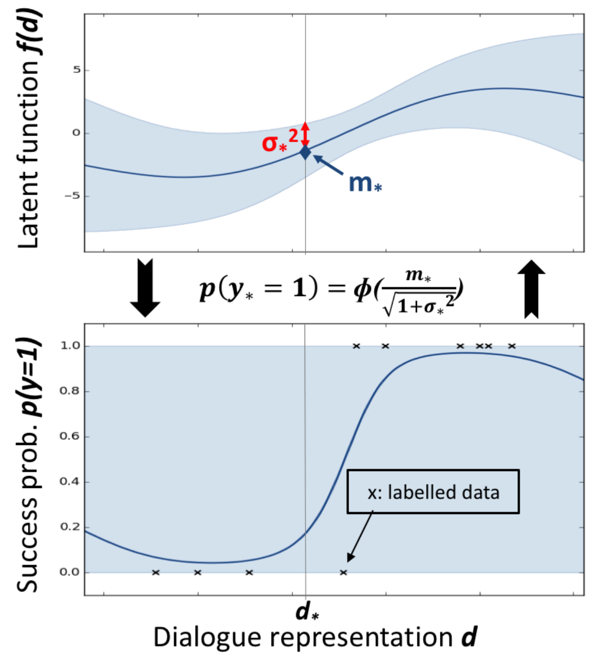
- Predictive distribution:

  $p(y=1 \,|\, \boldsymbol{d}, D) = \varphi(f(\boldsymbol{d} \,|\, D))$

- Prediction on $\boldsymbol{d}_*$:

  $p(y_* = 1|\boldsymbol{d}_*, D) = \varphi(m_*/\sqrt{1+\sigma_*{}^2})$

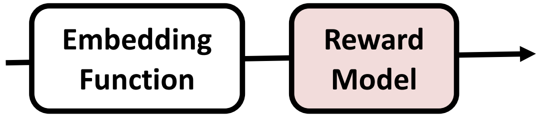
  $(\frac{m_*}{\sqrt{1+\sigma_*{}^2}} \rightarrow 0 \Rightarrow \varphi(\cdot) \rightarrow 0.5)$



B. Active Reward Learning Model

UNIVERSITY OF CAMBRIDGE

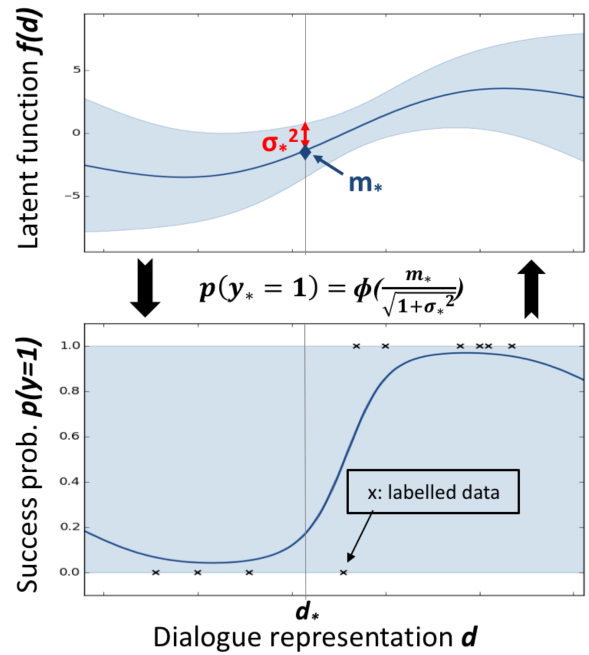
Gaussian process classifier for success rating



Handle the issue of noisy and costly user rating

- Add **Noise term** in the RBF kernel
  - More noise -> less certain
- **Active learning**: threshold on prob.
  - λ: when to query user rating

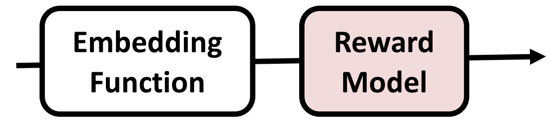
$$k(\mathbf{d},\mathbf{d}') = \underbrace{p^2 \exp(-\frac{||\mathbf{d}-\mathbf{d}'||^2}{2l^2})}_{\text{Input correlation}} + \underbrace{\sigma_n^2}_{\text{User rating noise}}$$

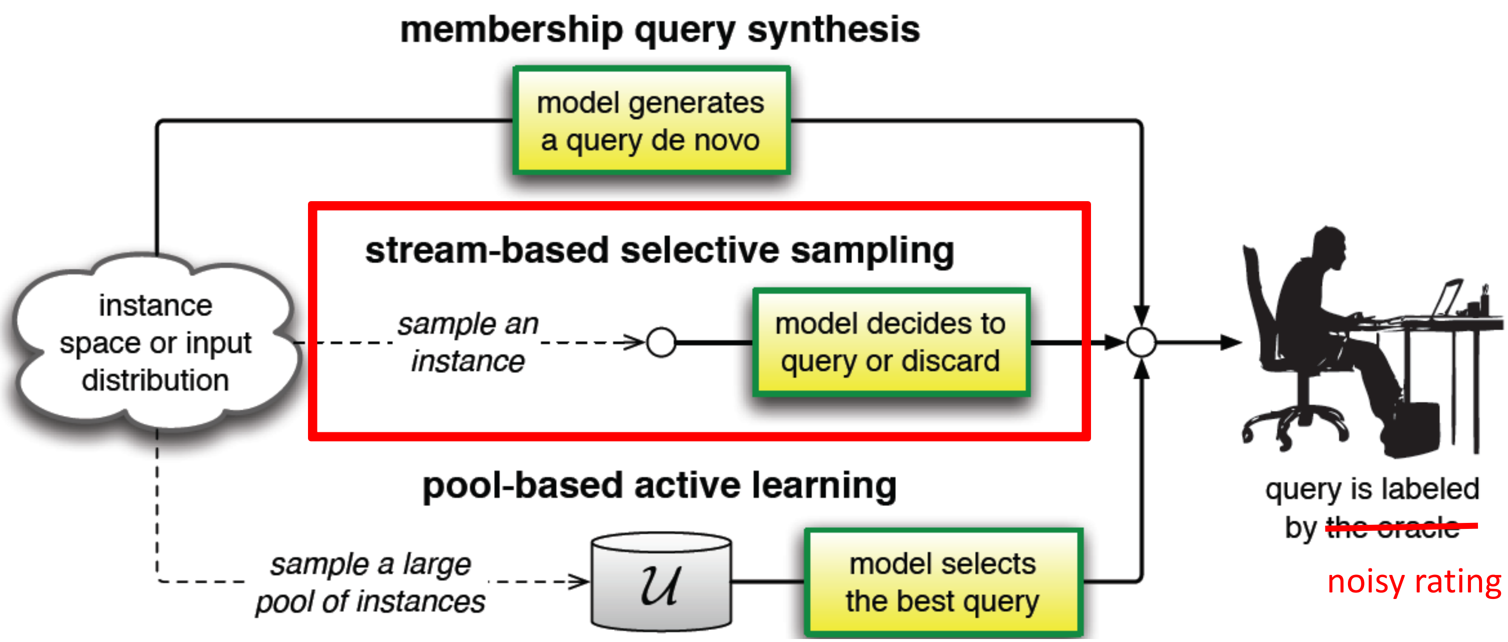


B. Active Reward Learning Model

UNIVERSITY OF CAMBRIDGE

Categories of Active Learning





Settles. Active Learning Literature Survey. 2009

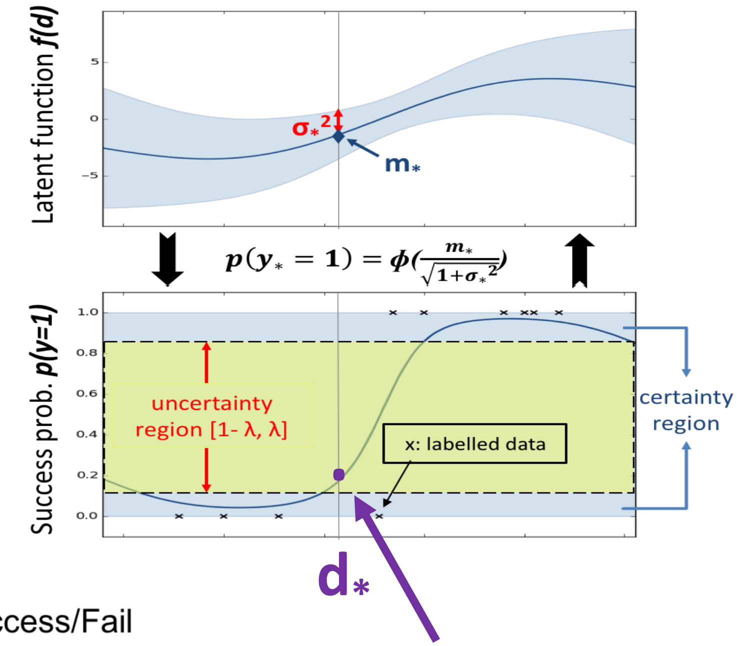
System Framework

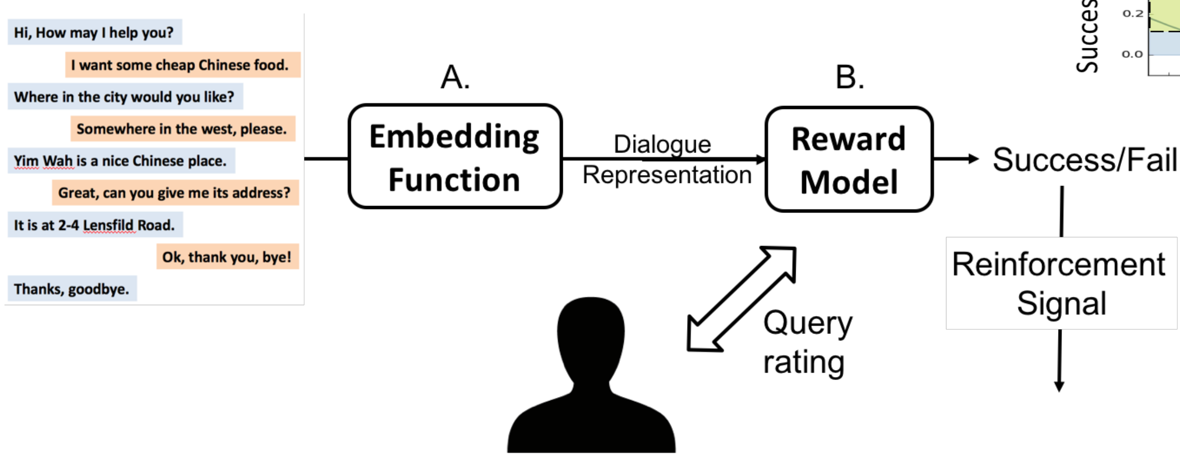
UNIVERSITY OF CAMBRIDGE

Active Reward Model in the loop

$$D = \{(\mathbf{d}, y)\}$$

$$\{f_1, \ldots, f_T\} \;—\; \sigma(f_{1:T}) \longrightarrow \mathbf{d}_*$$





In green area, query!
-> User rates: Failed
-> Reward: -1*scalar

Outline

❶ Motivation – Learning from human users

❷ Proposed Framework

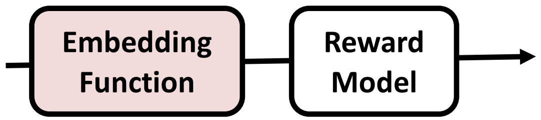
❸ Experiment

❹ Conclusion

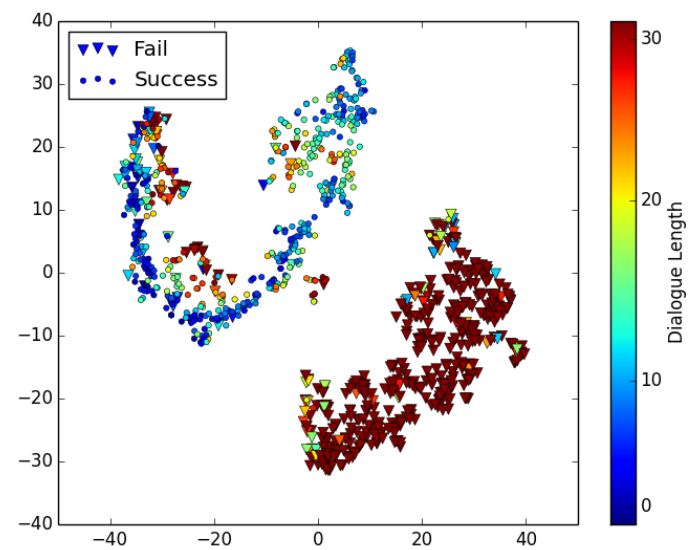
Dialogue Representation - Supervised

UNIVERSITY OF CAMBRIDGE

Visualising dialogue distribution



- Labelled restaurant dialogue data
  - train:valid:test = 1000:1000:1000
  - dim(d_s) = 32
- Analysis using t-SNE on d_s
  - Two clusters: Successful v.s. Failed
  - Successful: short, Failed: time-out
  - **Highly affected by training labels**

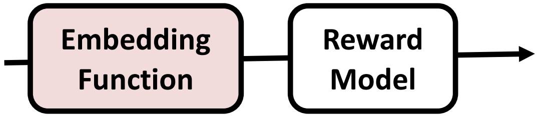


t-SNE plot

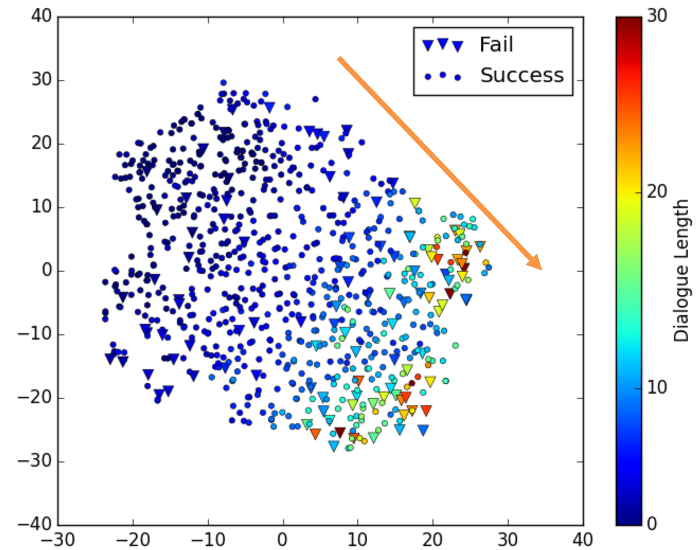
Dialogue Representation - Unsupervised

UNIVERSITY OF CAMBRIDGE

Visualising dialogue distribution



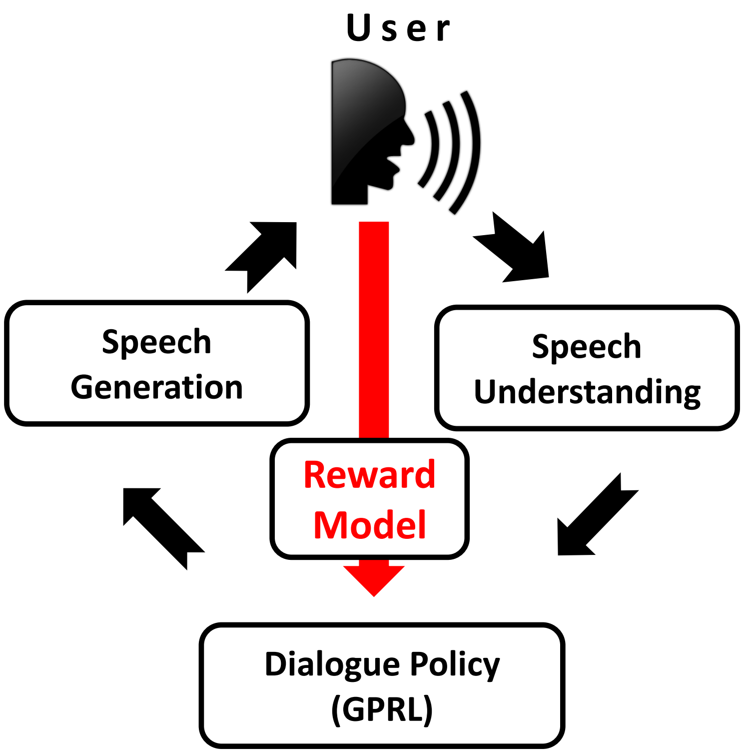
- Un-labelled restaurant dialogue data
  - train:valid:test = 8565:1199:650
  - $\dim(d_u) = 64$
- Analysis using t-SNE on d_u
  - Colour gradient: short → long length
  - Successful dialogues < 10 turns
    - Users don't engage in longer dialogues
  - *length* correlates highly to *success*

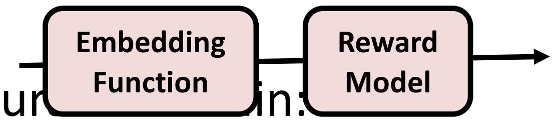


t-SNE plot

System Setup

Embed the reward model in SDS





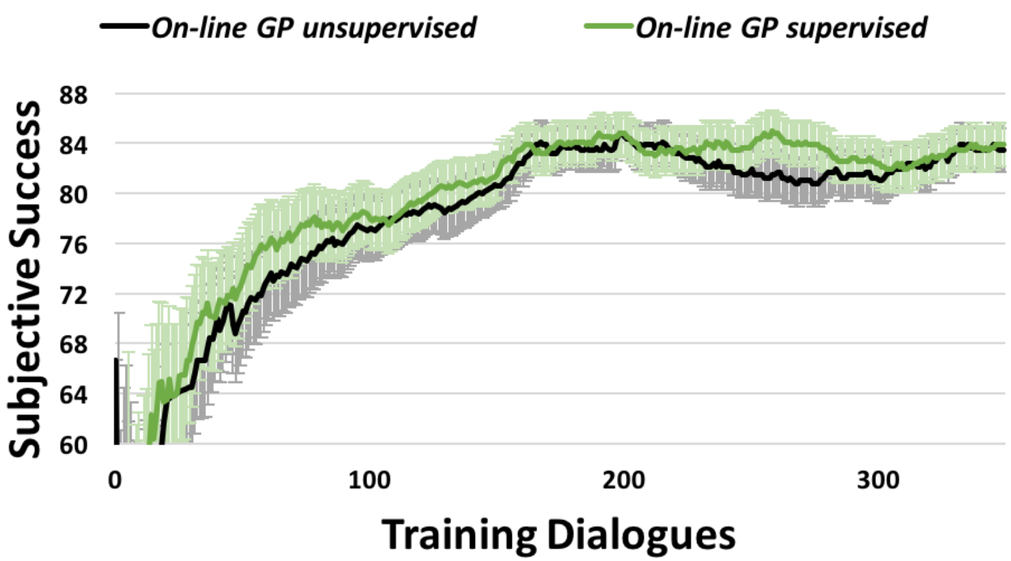
- Cambridge restaur… in:
  - ~100 venues
  - 3 informable slots: area, price range, food
  - 3 requestable slots: addr, phone, postcode
- Reward:
  - per turn -1,
  - When dialogue ends, binary (0/1) * 20:

- 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

Dialogue policy learning with real users

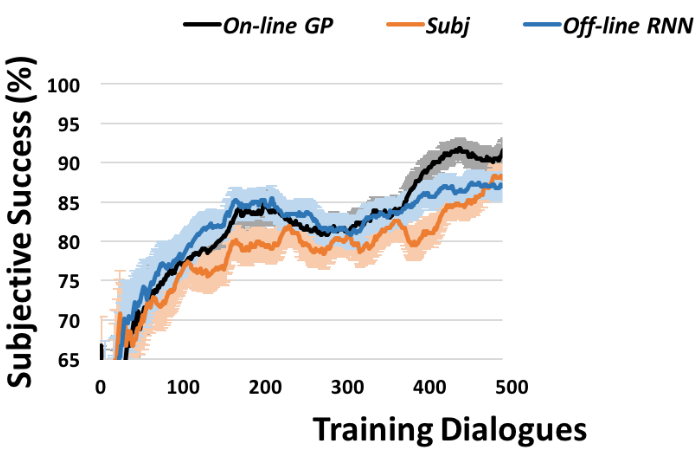


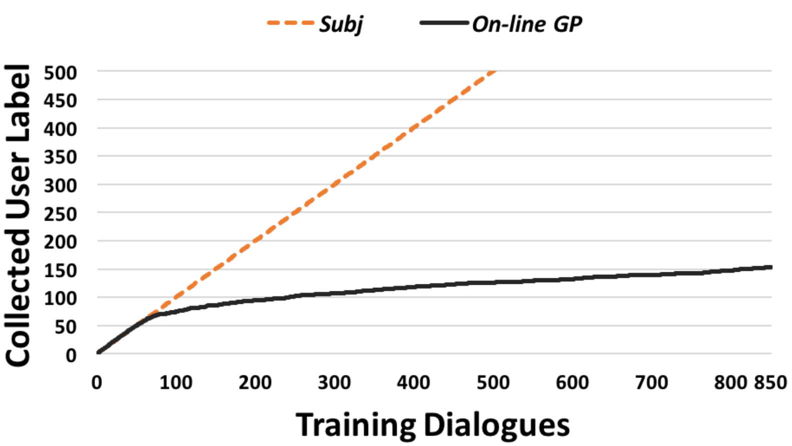
- Similar performance
- However, Supervised embedding requires additional labels
- Unsupervised method is thus more desirable

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

Outline

1. Motivation – Learning from human users
2. Proposed Framework
3. Experiment
4. Conclusion

Conclusion

UNIVERSITY OF CAMBRIDGE

- Proposal: an **on-line active reward learning** framework
  - Unsupervised Dialogue Embedding: Bi-LSTM Encoder-Decoder
  - On-line 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

Discussion

UNIVERSITY OF CAMBRIDGE

- Extend the reward model to (ordinal) regression/multi-class task
  - Currently handles only binary classification
- Methods for evaluating the dialogue embedding
  - Mostly measured by downstream tasks

Discussion

UNIVERSITY OF CAMBRIDGE

- Transfer knowledge across domains [1]
- Handle ambiguous meaning of languages [2]
- Learn to reply in richer context [3]
- Get high-quality data [4]

[1] Gašić et. al, Policy Committee for adaptation in multi-domain spoken dialogue systems, *ASRU 2015*
[2] Mrkšić, et. al, Counter-fitting Word Vectors to Linguistic Constraints. NAACL 2016
[3] Su et. al, Sample-efficient Actor-Critic Reinforcement Learning with Supervised Data for Dialogue Management, *SIGDIAL 2017*
[4] Wen et. al, A Network-based End-to-End Trainable Task-oriented Dialogue System, *EACL 2017*

Acknowledgement

- Past & Present Group members:
  - Steve Young (Supervisor)
  - Milica Gasic (Advisor)
  - Dongho Kim
  - Pirros Tsiakoulis
  - Matt Henderson
  - David Vandyke
  - Nikola Mrksic
  - Shawn Wen
  - Lina Rojas Barahona
  - Stefan Ultes
  - Pawel Budzianowski
  - Inigo Casanueva

- Financial supports:
  - Taiwan Cambridge PhD Scholarship
  - Funding from Engineering Department


תודה
Dankie
Gracias
Спасибо
شكراً
Merci
Takk
Köszönjük
Terima kasih
Grazie
Dziękujemy
Dėkojame
Ďakujeme
Vielen Dank
Paldies
Kiitos
Täname teid
谢谢
Thank You
Tak
感謝您
Obrigado
Teşekkür Ederiz
감사합니다
Σας Ευχαριστούμ
ขอบคุณ
Bedankt
Děkujeme vám
ありがとうございます
Tack

References

UNIVERSITY OF CAMBRIDGE

1. Pei-Hao Su, 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. Pei-Hao Su, 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, Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Tsung-Hsien Wen and Steve Young, **"Multi-Domain Dialogue Success Classifiers for Policy Training"**. *In Proceeding of ASRU 2015*

References

UNIVERSITY OF CAMBRIDGE

Chat-based Systems

- Oriol Vinyals, Quoc Le, **"A Neural Conversational Model"**. *In arXiv 1506.05869*
- Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, Joelle Pineau, **"Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models"**. *In AAAI 2016*
- Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, Dan Jurafsky, "**Deep Reinforcement Learning for Dialogue Generation**". *In EMNLP 2016*
- Al-Rfou et al., "**Conversational Contextual Cues: The Case of Personalization and History for Response Ranking**". *In arXiv 2016*

Task-oriented Dialogue Systems

- James Henderson, Oliver Lemon, Kallirroi Georgila, **"Hybrid Reinforcement/Supervised Learning for Dialogue Policies from COMMUNICATOR data"**. *In* IJCAI Workshop 2005
- Jason Williams and Steve Young, **"Partially observable Markov decision processes for spoken dialog systems"**. *In CSL 2007*
- Antoine Bordes, Y-Lan Boureau, Jason Weston, "**Learning End-to-End Goal-Oriented Dialog**". *In ICLR 2017*
- Wen et. al, "**A Network-based End-to-End Trainable Task-oriented Dialogue System**", in EACL 2017

Questions?

-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.

phs26@cam.ac.uk
http://mi.eng.cam.ac.uk/~phs26/

UNIVERSITY OF CAMBRIDGE

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.