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

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona,
Stefan Ultes, David Vandyke, Tsung-Hsien Wen and Steve Young
ACL 2016

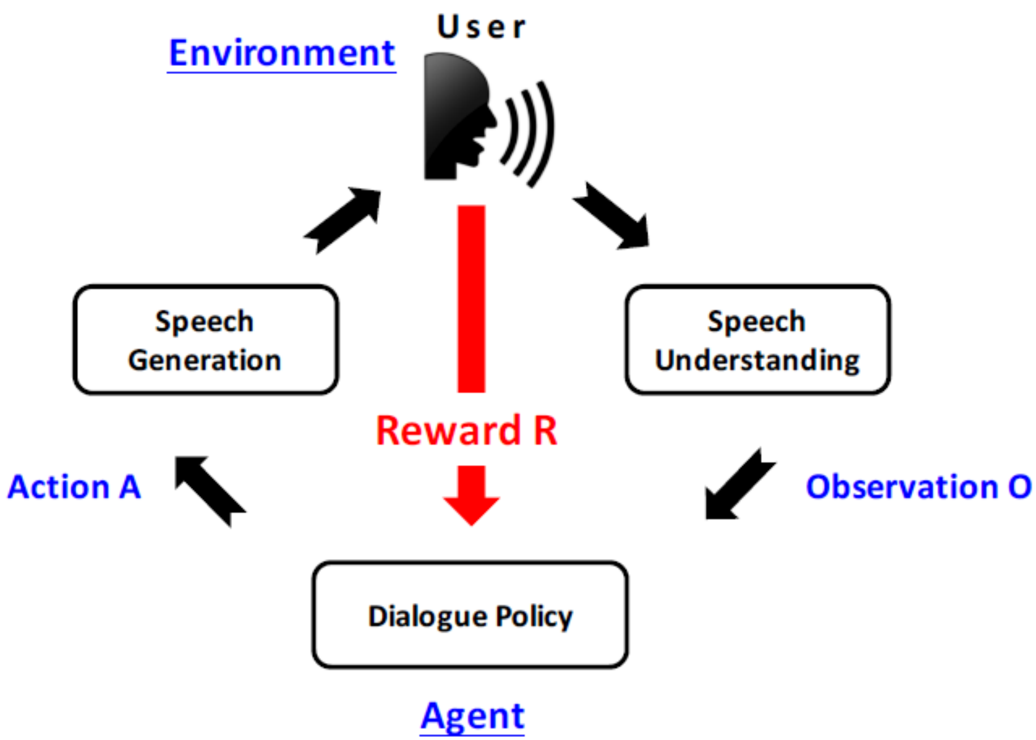
Juho Kim

Goal

- Design a suitable learning objective (**reward**) to train an RL-based dialogue system online from real users

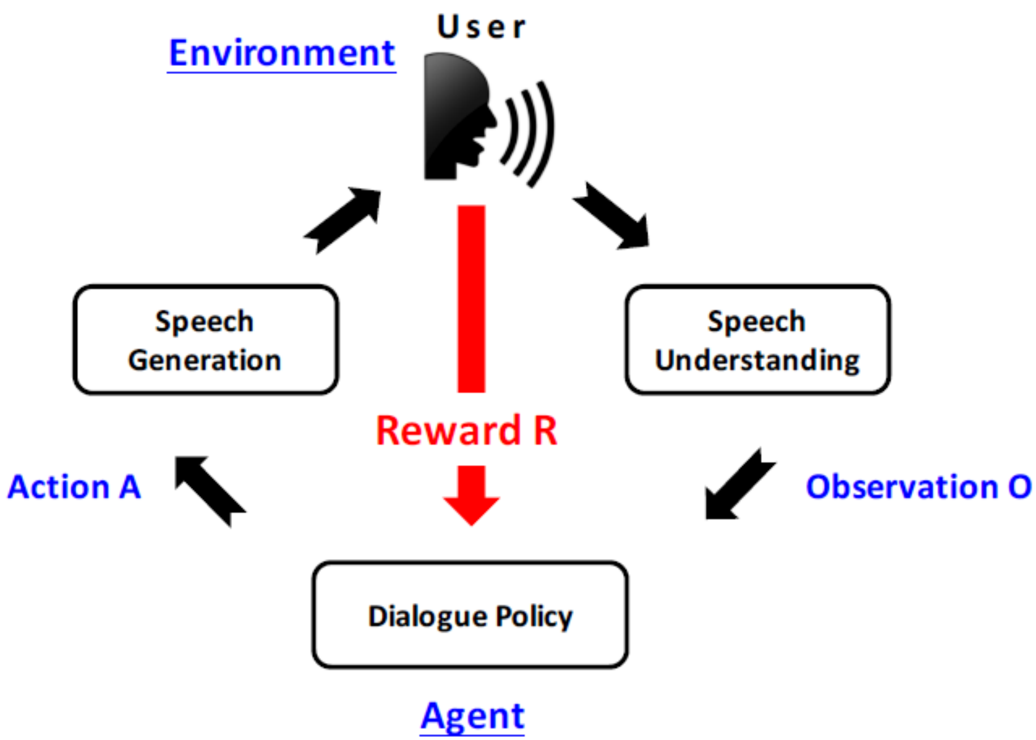
Goal

- Design a suitable learning objective (**reward**) to train an RL-based dialogue system online from real users



Goal

- Design a suitable learning objective (**reward**) to train an RL-based dialogue system online from real users



Correct rewards are critical in dialogue policy training

Reinforcement signals in dialogue systems

How to learn policy from real users?

Reinforcement signals in dialogue systems

How to learn policy from real users?

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

Reinforcement signals in dialogue systems

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

Reinforcement signals in dialogue systems

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
  - Difficult/costly to obtain
  - Noisy

Reinforcement signals in dialogue systems

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~~ Reward modeling on user rating
  - Difficult/costly to obtain → Active learning
  - Noisy → Gaussian process with uncertainty

Proposed method

Reward modeling on user binary success rating

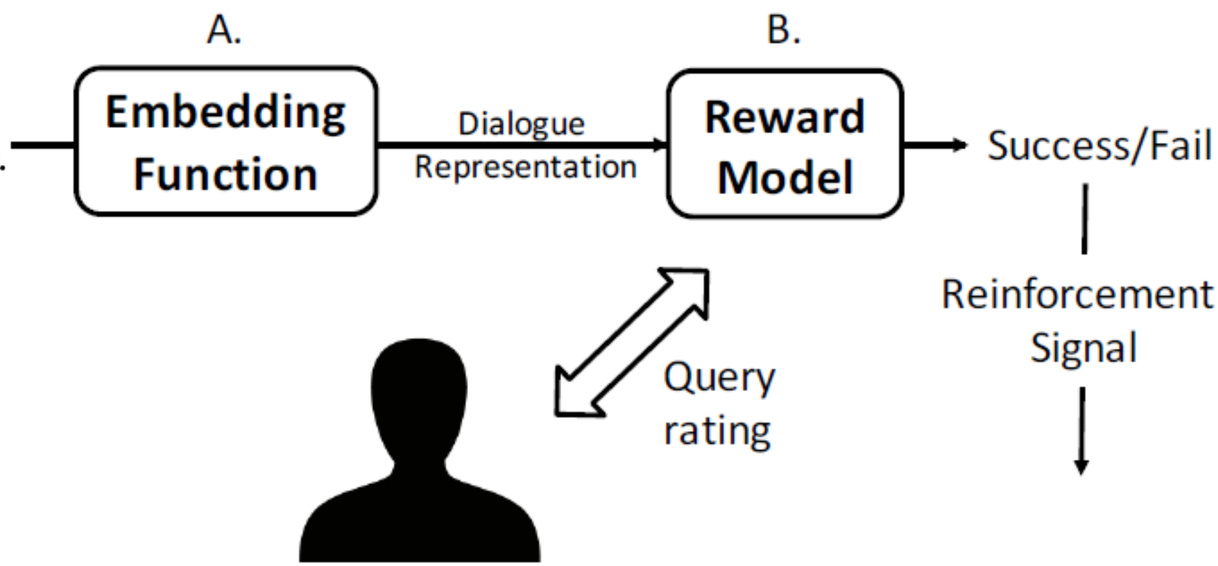
Proposed method

Reward modeling on user binary success rating

Dialogue (input)

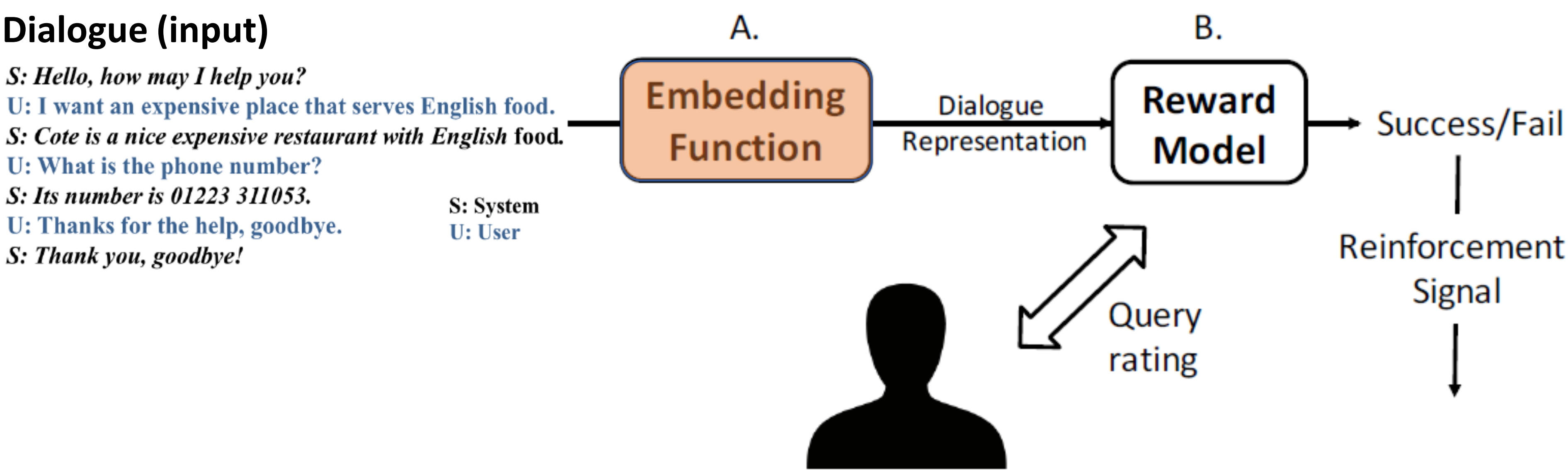
S: Hello, how may I help you?
U: I want an expensive place that serves English food.
S: Cote is a nice expensive restaurant with English food.
U: What is the phone number?
S: Its number is 01223 311053.
U: Thanks for the help, goodbye.
S: Thank you, goodbye!

S: System
U: User



Proposed method

Reward modeling on user binary success rating



A. Dialogue embedding

Mapping a dialogue sequence to a fixed-length vector

S: Hello, how may I help you?

Turn 1 f_1 U: I want an expensive place that serves English food.

S: Cote is a nice expensive restaurant with English food.

Turn 2 f_2 U: What is the phone number?

S: Its number is 01223 311053.

S: System
U: User

f_t : concatenated vector of

- user intention determined the semantic decoder
- distribution over each concept defined in the ontology
- one-hot encoding of the system's reply action
- turn number

(Vandyke et al., ASRU 2015)

A. Dialogue embedding

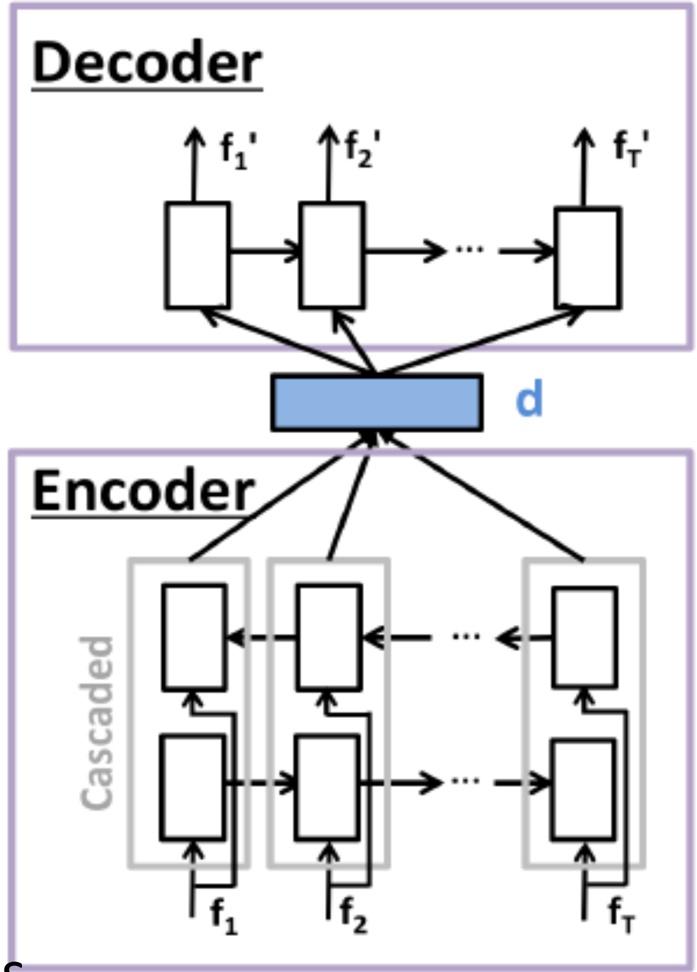
Bi-directional LSTM encoder-decoder

- Inputs are turn-level features
- $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$ captures forward and backward information
- Dialogue representation

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

- Mean squared error training:

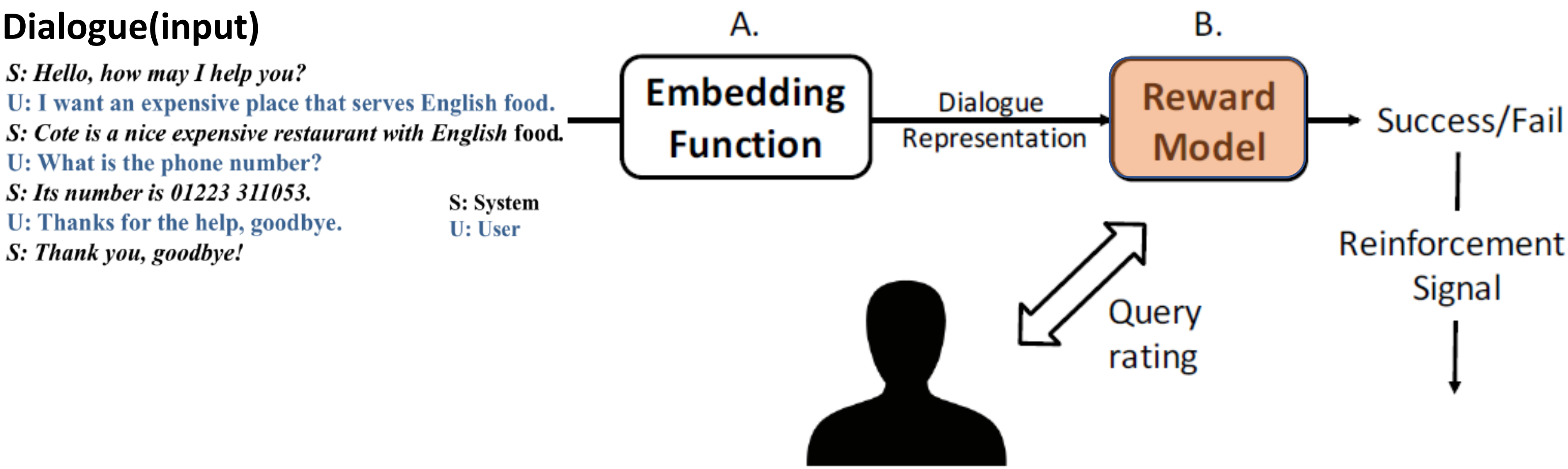
$$MSE = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}||\mathbf{f}_t - \mathbf{f}'_t||^2$$



T: number of turns, N: number of all dialogues

Proposed method

Reward modeling on user binary success rating



B. Active reward learning model

Model dialogue success using Gaussian process regression

$$p(y = 1|\mathbf{d}, \mathcal{D}) = \phi(f(\mathbf{d}|\mathcal{D}))$$

B. Active reward learning model

Model dialogue success using Gaussian process regression

$$p(y = 1|\mathbf{d}, \mathcal{D}) = \underbrace{\phi}_{\text{cumulative Gaussian}}(\underbrace{f(\mathbf{d}|\mathcal{D})}_{\mathrm{GP}(m(\mathrm{d}),\, k(\mathrm{d},\mathrm{d}'))})$$

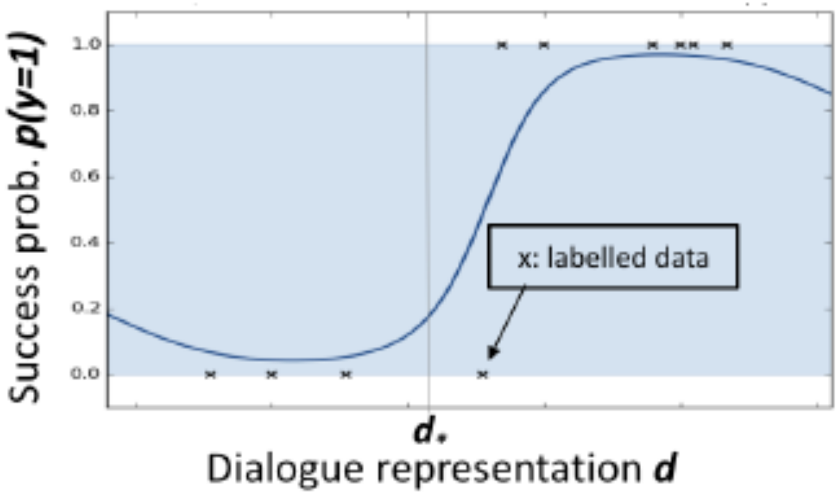
B. Active reward learning model

Model dialogue success using Gaussian process regression

$$p(y=1|\mathbf{d},\mathcal{D}) = \phi(f(\mathbf{d}|\mathcal{D}))$$

- **Noise term** in the RBF kernel affects uncertainty

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



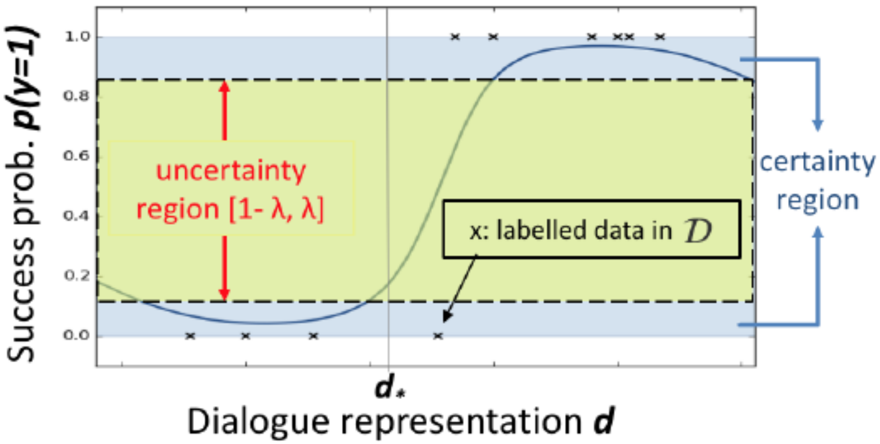
B. Active reward learning model

Model dialogue success using Gaussian process regression

$$p(y=1|\mathbf{d},\mathcal{D}) = \phi(f(\mathbf{d}|\mathcal{D}))$$

- **Noise term** in the RBF kernel affects uncertainty
- **Active learning:** uncertainty + threshold
  - Model is uncertain → query user rating actively

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



Experiments

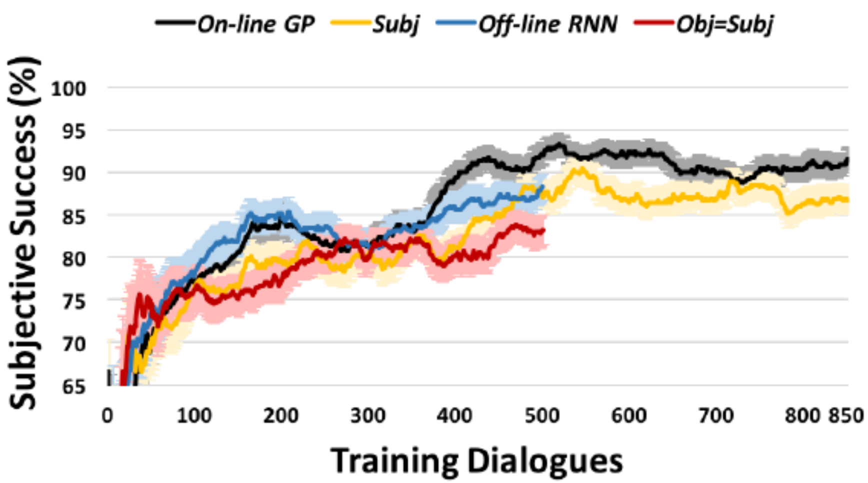
- Dataset: Cambridge restaurant domain

  - 150 venues
  - 3 information slots: area, price range, food
  - 3 request slots: address, phone, postcode

- Reward for success/failure

  - Per turn: -1
  - When dialogue ends, binary(0/1) * 20

Experiments



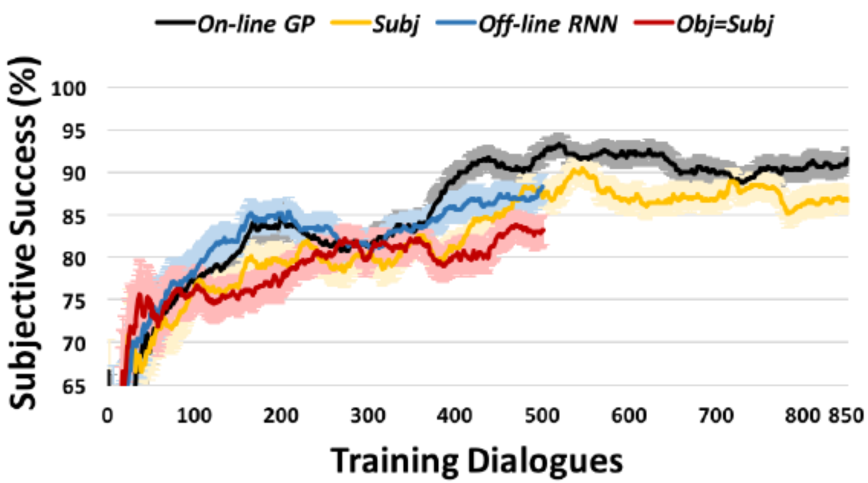
Dialogues	**Reward Model**	***Subjective*** (%)
400-500	*Obj=Subj*	85.0 ± 2.1
	off-line RNN	89.0 ± 1.8
	Subj	90.7 ± 1.7
	on-line GP	91.7 ± 1.6
500-850	*Subj*	87.1 ± 1.0
	on-line GP	$\mathbf{90.9 \pm 0.9}$*

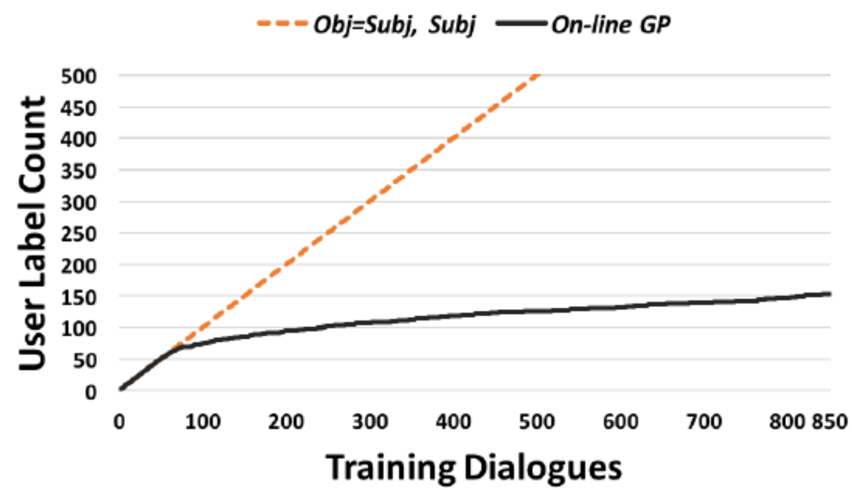
$* \, p < 0.05$

- All reached > 85% after 500 dialogues
- Proposed method is better than others in the longer run

On-line GP: proposed method
Subj: method that optimizes the policy using only user assessment
Off-line RNN: RNN with 1K simulated data
Obj=Subj: method using the dialogues that user's subjective assessment is consistent to the objective one

Experiments





- All reached > 85% after 500 dialogues
- Proposed method is better than others in the longer run
- Proposed method needs smaller queries from user rating

On-line GP: proposed method
Subj: method that optimizes the policy using only user assessment
Off-line RNN: RNN with 1K simulated data
Obj=Subj: method using the dialogues that user's subjective assessment is consistent to the objective one

Conclusion

- Propose method: on-line active reward learning
  - Dialogue embedding: Bi-LSTM Encoder and Decoder
  - Active reward model: GP regression with uncertainty threshold
  - Reduce data annotation costs and model noisy user rating

- Achieve online policy learning from real users w/o task information