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

Motivation

Proposal

Experiments

Conclusion

With paid users

Hi, How may I help you?

I want a cheap Chinese Restaurant.

Where in the city would you like?

Yim Wah is a nice Chinese place.

Great, can you give me the address?

It is at 2-4 Lensfield Road.

Ok, thank you, bye!

Thanks, goodbye.

Success evaluation

By task: Fail (no phone)

By user: Success (get all he asked)

Task:
- Find a restaurant, Chinese, cheap, west
- Ask phone, address
Defining suitable learning objective (*reward*) to train a dialogue system on-line from *real users*
Statistical Spoken Dialogue System

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)

<table>
<thead>
<tr>
<th>Area</th>
<th>East</th>
<th>West</th>
<th>...</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01</td>
<td>0.94</td>
<td>0</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Correct rewards are a crucial factor in dialogue policy training.
The Reinforcement Signal in SDS

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
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

- **Environment**
- **Speech Generation**
- **Speech Understanding**
- **Action A**
- **Observation O**
- **Reward Model**
- **Dialogue Policy**
- **Agent**
System Framework

Reward modelling on user binary success rating

Hi, How may I help you?
I want some cheap Chinese food.
Where in the city would you like?
Somewhere in the west, please.
Yim Wah is a nice Chinese place.
Great, can you give me its address?
It is at 2-4 Lensfield Road.
Ok, thank you, bye!
Thanks, goodbye.

A. Embedding Function

B. Reward Model

Dialogue Representation

Success/Fail

Reinforcement Signal

Query rating

On-line Dialogue Reward and Policy Learning
A. Unsupervised Dialogue Embedding

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

Bi-directional LSTM Encoder-Decoder

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

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

- MSE training criterion:

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

- $f_t$: input/target, $f'_t$: prediction
B. Active Reward Learning Model

Gaussian process classifier for success rating

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

\[
k(d, d') = p^2 \exp\left(-\frac{\|d - d'\|^2}{2l^2}\right) + \sigma_n^2
\]

- Input correlation
- User rating noise

Graph showing success probability \( p(y=1) \) as a function of dialogue representation \( d \) with labelled data at \( d^* \).
B. Active Reward Learning Model

**Gaussian process** classifier for success rating

- **Noise term** in the RBF kernel affects uncertainty
  - More noise -> less certain
- **Active learning**: uncertainty + threshold
  - \( \lambda \): when to actively query user rating

\[
k(d, d') = \sigma_n^2 \exp(-\frac{||d - d'||^2}{2l^2}) + \sigma_n^2
\]

- Input correlation
- User rating noise

![Diagram of Gaussian process classifier](image)

- Success prob. \( p(y=1) \)
- Uncertainty region \([1-\lambda, \lambda] \)
- Certainty region
- \( x \): labelled data
- Dialogue representation \( d \)
System Setup

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:

<table>
<thead>
<tr>
<th>On-line GP</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj</td>
<td>User rating only</td>
</tr>
<tr>
<td>Off-line RNN (Su. et al. 2015)</td>
<td>RNN with 1K simulated data</td>
</tr>
</tbody>
</table>

- Crowd-sourced users from Amazon Mechanical Turk
On-line Dialogue Reward & Policy Learning

Dialogue policy learning with real users

<table>
<thead>
<tr>
<th>Dialogues</th>
<th>Reward Model</th>
<th>Subjective (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400 - 500</td>
<td>Off-line RNN</td>
<td>89.0 +/- 1.8</td>
</tr>
<tr>
<td></td>
<td>Subj</td>
<td>90.7 +/- 1.7</td>
</tr>
<tr>
<td></td>
<td>On-line GP</td>
<td>91.7 +/- 1.6</td>
</tr>
<tr>
<td>500 - 850</td>
<td>Subj</td>
<td>87.1 +/- 1.0</td>
</tr>
<tr>
<td></td>
<td>On-line GP</td>
<td>90.9 +/- 0.9</td>
</tr>
</tbody>
</table>

- All reached > 85% after 500 dialogues
- On-line GP is more robust than Subj in longer run
On-line Dialogue Reward & Policy Learning

Dialogue policy learning with real users

- All reached > 85% 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

- **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?


phs26@cam.ac.uk


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
-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

**-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] results()

**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.