Reward Estimation for Dialogue Policy Optimisation

Pei-Hao (Eddy) Su

DeepHack.Turing, 25 July 2017
Dialogue Systems

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

- 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

**Motivation**

**Proposal**

**Experiments**

**Conclusion**

---

With paid subjects

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

**Not Practical**

Success evaluation

*Objective*: Fail (no *phone*)

*Subjective*: Success (get all he asked)

---

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 Lensfield Road**.

Ok, thank you, bye!

Thanks, goodbye.
Goal

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

1. Motivation – Learning from real users
2. Proposed Framework
3. Experiment
4. Conclusion
Pipeline 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>
Reinforcement Learning 101

Agent learns to take actions to maximise total reward

AlphaGo

It beat GO champions in 2016 and 2017
Agent learns to take actions to maximise total reward
Motivation

Dialogue Manager in RL framework

Environment

Speech Generation

Speech Understanding

Reward R

Action A

Observation O

Dialogue Policy

Correct rewards are a crucial factor in dialogue policy training
**Motivation**

Reward for RL $\equiv$ Evaluation for SDS

- 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

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 et al. 2015)
The Reinforcement Signal in SDS

RNN Reward Estimator for Policy Learning

**Objective-Baseline**
- Needs task info.
- Learns only from \( \text{Obj}=\text{Subj} \) 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
**Motivation**

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 $\rightarrow$ Gaussian Process with uncertainty
  - Difficult/Costly $\rightarrow$ Active Learning
# Outline

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

Environment

Speech Generation

Reward Model

Speech Understanding

Action A

Observation O

Dialogue Policy

Agent
System Framework

Reward modelling on user binary success rating

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 Lensfield Road.
Ok, thank you, bye!
Thanks, goodbye.

A. Embedding Function

B. Reward Model

Dialogue Representation

Success/Fail

Reinforcement Signal

Query rating

Motivation

Conclusion

Proposal

Experiments
A. Dialogue Embedding

Maps a dialogue seq to a fixed-length vector

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

(Vandyke & Su et al, ASRU 2015)
A. Dialogue Embedding - Supervised

Re-use the supervised RNN

- Last hidden layer as dialogue representation
A. Dialogue Embedding - Unsupervised

**Bi-LSTM Encoder-Decoder (Seq2Seq)**

- Reconstruct inputs with variable-lengths
- $h_t = [\vec{h}_t ; \hat{h}_t]$ captures forward-backward info
- Bottleneck $d_u$ 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

- Determine class probability: $p(y|d, D)$, given $D = \{(d_i, y_i)\}_{i=1}^n$
  - where $y = \{+1, -1\}$

- Handle the issue of noisy and costly user rating

- **Gaussian process** (GP) with active learning
B. Active Reward Learning Model

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

- \( p(y = 1|d, D) = \phi(f(d|D)) \)
  - \( f \): latent function: \( R^{\text{dim}(d)} \to R \)
  - \( \phi \): probit function: \( R \to [0,1] \)

- \( f(d) \sim GP(m(d), k(d, d')) \)
  - \( k(d, d') = p^2 \exp(-\frac{||d - d'||^2}{2l^2}) \)
Gaussian process classifier for success rating

- Prior:
  \[ f(d) \sim GP(m(d), k(d, d')) \]

- Predictive distribution:
  \[ p(y=1 \mid d, D) = \phi(f(d \mid D)) \]

- Prediction on \( d_* \):
  \[ p(y_* = 1 \mid d_*, D) = \phi(m_*/\sqrt{1 + \sigma_*^2}) \]
  \[ \left( \frac{m_*}{\sqrt{1+\sigma_*^2}} \rightarrow 0 \Rightarrow \phi(\cdot) \rightarrow 0.5 \right) \]
B. Active Reward Learning Model

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.
  - $\lambda$: when to query user rating

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

Input correlation  \hspace{2cm} User rating noise

**Motivation**

**Proposal**

**Experiments**

**Conclusion**
B. Active Reward Learning Model

Categories of Active Learning

- **Membership Query Synthesis**
  - Model generates a query de novo

- **Stream-based Selective Sampling**
  - Sample an instance
  - Model decides to query or discard

- **Pool-based Active Learning**
  - Sample a large pool of instances
  - Model selects the best query

Settles. Active Learning Literature Survey. 2009
Active Reward Model in the loop

\[ D = \{(d, y)\} \]

\[ \{f_1, \ldots, f_T\} \rightarrow \sigma(f_{1:T}) \rightarrow d^* \]

**Motivation**

**Proposal**

**Experiments**

**Conclusion**

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

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

- Labelled restaurant dialogue data
  - train:valid:test = 1000:1000:1000
  - $\text{dim}(d_s) = 32$

- Analysis using t-SNE on $d_s$
  - Two clusters: Successful v.s. Failed
  - Successful: short, Failed: time-out
  - Highly affected by training labels
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 $\rightarrow$ long length
  - Successful dialogues < 10 turns
    - Users don’t engage in longer dialogues
  - length correlates highly to success
System Setup

Embed the reward model in SDS

- Cambridge restaurant domain:
  - ~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
    - Subj
    - Off-line RNN (Su. et al. 2015)

- Crowd-sourced users from Amazon Mechanical Turk

<table>
<thead>
<tr>
<th>Method</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-line GP</td>
<td>Premium method ..........................</td>
</tr>
<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>
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

Dialogues Reward Model

Subjective (%)

| 400-500 Off-line RNN | Subj | 89.0 + 1.8 | 90.7 + 1.7 | 91.7 + 1.6 |
| 500-850 Subj | On-line GP | 87.1 + 1.0 | 90.9 + 0.9 |

- 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
# Outline

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

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

- 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

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

• **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

---

**Acknowledgement**
References


Chat-based Systems
- Oriol Vinyals, Quoc Le, “A Neural Conversational Model”. In arXiv 1506.05869
- AI-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)

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

phs26@cam.ac.uk
http://mi.eng.cam.ac.uk/~phs26/
<table>
<thead>
<tr>
<th>Motivation</th>
<th>Proposal</th>
<th>Experiments</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

Motivation

Proposal

Experiments

Conclusion
Example Dialogues – Low Noise

-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] 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.