Advanced Cutting Edge Research Seminar

Dialogue Management using Reinforcement Learning

Assistant Professor
Koichiro Yoshino
Course works

1. Basis of spoken dialogue systems
   – Type and modules of spoken dialogue systems

2. Deep learning for spoken dialogue systems
   – Basis of deep learning (deep neural networks)
   – Recent approaches of deep learning for spoken dialogue systems

3. Dialogue management using reinforcement learning
   – Basis of reinforcement learning
   – Statistical dialogue management using intention dependency graph

4. Dialogue management using deep reinforcement learning
   – Implementation of deep Q-network in dialogue management
Problem of Q-learning

Initialize every pair of $Q(s, a)$
set $\varepsilon (0<\varepsilon<1)$
while update $< \text{threshold}$

observe $s^t$
if rand() $< \varepsilon$
the system takes the action $a^t$ according to $\max_{a^t} Q(s^t, a^t)$
else
randomly select $a^t$

decide $s^{t+1}$
receive reward $R(s^t, a^t, s^{t+1})$
update $Q(s^t, a^t)$

$(1 - \alpha)Q(s^t, a^t) + \alpha \left( R(s^t, a^t, s^{t+1}) + \gamma \max_{a^{t+1}} Q(s^{t+1}, a^{t+1}) \right)$
end

How do we decide $s$?
How to define rewards?
How do we apply for $b$?
User simulator

- **User simulator decides** $s$ **at each turn**
  - Very simple model works on $P(s^{t+1}|s^t, a^t)$
    - But we can calculate actual $Q^*(s,a)$ without Q-learning, if we know $P(s^{t+1}|s^t, a^t)$

- **There are several approaches as other modules**
  - **Heuristics**
    - Agenda-based: The simulator assumes goal and agenda
    - IDG system: The simulator assumes goal and transitions to goals
  - **Data driven approaches:**
    - N-gram based utterance generator,
    - neural network based
Seq2seq model for user simulation

- Train the action sequence of the user given a sequence of context
  - It just simulate actions (intentions) of the user
  - Simple encoder-decoder model is used

Learning convergence

- **Q-learning requires many trials to converge**
  - Even if only train on small number of states and actions
  - For 7 states and 3 actions, it requires 100,000 steps
- **Good user simulator is necessary**
  - Required number of data for simulator is smaller than Q-learning
  - If we know the behavior of the user, it is possible to build the model with less training data
Reward definition

• Task completion
  – This is the most simple and easy to understand
  – E.g. system receives +10 rewards in task successes, -10 penalties in task fails and -1 penalty on each turn

• User satisfaction
  – Regression result of the user satisfaction
  – It works comparably to the task completion

• Inverse reinforcement learning
  – Calculate the reward from the expert data (dialogue data of wizard of Oz; dialogue system acted by human)


Reward definition and policy

- Relation between rewards and trained policy
  - Higher penalty increase the number of “confirmation” (conservative)
  - Q-learning trained more progressive policy if the penalty is small

-10 for mistake • -50 for mistake
  - 1 do
  - 2 do
  - 3 goal
  - 4 do
  - 5 goal
  - 6 do
  - 7 goal

Relation between rewards and trained policy
Higher penalty increase the number of “confirmation” (conservative)
Q-learning trained more progressive policy if the penalty is small

-10 for mistake • -50 for mistake
  - 1 do
  - 2 do
  - 3 goal
  - 4 do
  - 5 goal
  - 6 do
  - 7 goal
Other remaining problems

• How do we decide belief point to be sampled as $b$?
  – It is hard to Update any $Q(b, a)$, because $b$ is not a point as $s$
    • It will be hyper plain
• Grid-based value iteration
  – Decide belief points with grid
• Point-based value iteration
  – Decide belief points from sampling of data
• Regression (Q-network)
  – If we can develop a regression to calculate $Q(b, a)$, it can calculate $Q(b', a)$ (if the model is successfully trained)
States are defined from dialogue frames, and simple actions are set up
  - Of course, you can develop much more complicated intention graph

User simulator can be developed under very simple assumption
  - Page 28

Belief update can be modeled with prior that can be acquired from the domain knowledge
  - Supervised learning requires large scale training data
    - $h^t = \tanh(W_{Xh}X^t + W_{hh}h^{t-1} + c_h)$
    - $b^t = \text{softmax}(W_{hb}h^t + c_b)$

Reward functions are task completion

Problem of belief point sampling will be solved by Q-network
If we use simple assumption in IDG...

- The user firstly decide their goal $g$
  - The goal is 3 in the example

- The user says the first utterance as any node between the goal node and the root node
  - It will be 1, 2 or 3 in the example
  - $P(s'|g,a)$: user simulator

- If the system confirms with the user, the user repeats the previous utterance (node) again
  - Required if the system assumes belief (=confidence of each state)
If we use simple assumption in IDG...

• If we know the current state (node), we also can estimate possible goals, which are children of the current state
  – \( P(g|s) \): goal model

• State transition can be approximated with the goal model and the user simulator
  – \( P(s|s, a) = \sum_g P(s|g, s, a)P(g|s) \)
  – \( \approx \sum_g P(s|g, a)P(g|s) \)
  – The benefit of this model is that we start from the zero-resource
Let’s see the source code...

https://github.com/ahclab/Q-learning-DM
Future directions of spoken dialogue systems

- Controllable dialogue systems
  - Especially for neural conversation models

Future directions of spoken dialogue systems

- **Multi-modal, affective computing**
  - Considering non-verbal user states such as emotion will improve the experience of conversation for the user
  - Multi-modal information is important to observe such states
  - Systems are also required to use their own emotional, friendly, kind expressions to user

Eliciting Positive Emotion through Affect-Sensitive Dialogue Response Generation: A Neural Network Approach, Lubis et al., In Proc AAAI2018
Future directions of spoken dialogue systems

- **Interaction with real world**
  - Grounding
  - Relations between real objects and concepts
  - Knowledge acquisition from conversation
  
  - How to learn from the conversation?

- **Connections to IoT**
  - Smart speaker
  - But we need to take care about malicious users: Microsoft Tay

Lexical Acquisition through Implicit Confirmations over Multiple Dialogues, Ono et al., In Proc SIGDIAL2017
Future directions of spoken dialogue systems

• Many new tasks will appear
  – Conventional task-oriented tasks → more complicated multi domain task that requires knowledges of several tasks and domains
  – Chat-oriented system → chatting system that can keep what they talk with the user by dynamically changing the topic or behaviors to keep the user attention

• New learning theory
  – Deep reinforcement learning: actor-critic
  – Bayesian deep reinforcement learning: gives some priors to reduce the number of learning data
Choose one from following works:

1. Read one original paper that is introduced in 4 classes and submit A4 x 2 pages summarization.
2. Try the source code on Github, work on your original domain and submit the source code and summary.

Deadline: Feb 13 23:59JST

- Email: koichiro@is.naist.jp
- Subject: ACE-report-[your student number]-[your name]
- Current Q-learning code may contain some bugs, I’ll try to fix that by the end of this Friday...