On-line Dialogue Policy Learning with Companion Teaching

EACL 2017 @ VALENCIA, SPAIN

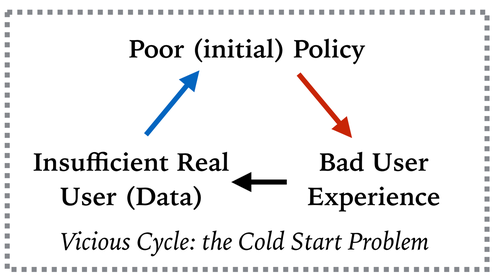
Lu Chen, **Runzhe Yang**, Cheng Chang, Zihao Ye, Xiang Zhou and Kai Yu

SpeechLab, Shanghai Jiao Tong University

SJTU SPEECH LAB
上海交通大學智能語音實驗室

How to Build Evolvable Conversational Agent in Real World Scenarios?

- The **off-line trained policy** is not guaranteed to work well in real world scenarios.
- The **on-line dialogue policy learning** is essential to making conversational agent evolvable.
- However, simply deploying the existing framework of dialogue system CANNOT live up to our expectations, because the **Cold Start Problem** has not been well addressed in the old frameworks.
- The cold start problem can be illustrate as following vicious cycle.
- **In this work, we try to propose a practical framework to address the cold start problem.**



Vicious Cycle: the Cold Start Problem

Possible Solutions to break the vicious cycle

- Inefficient Learning Process (Solvable) ✓
- Unsafe Policy Behavior (Solvable) ✓
- Individual Rationality (Unsolvable) ✗

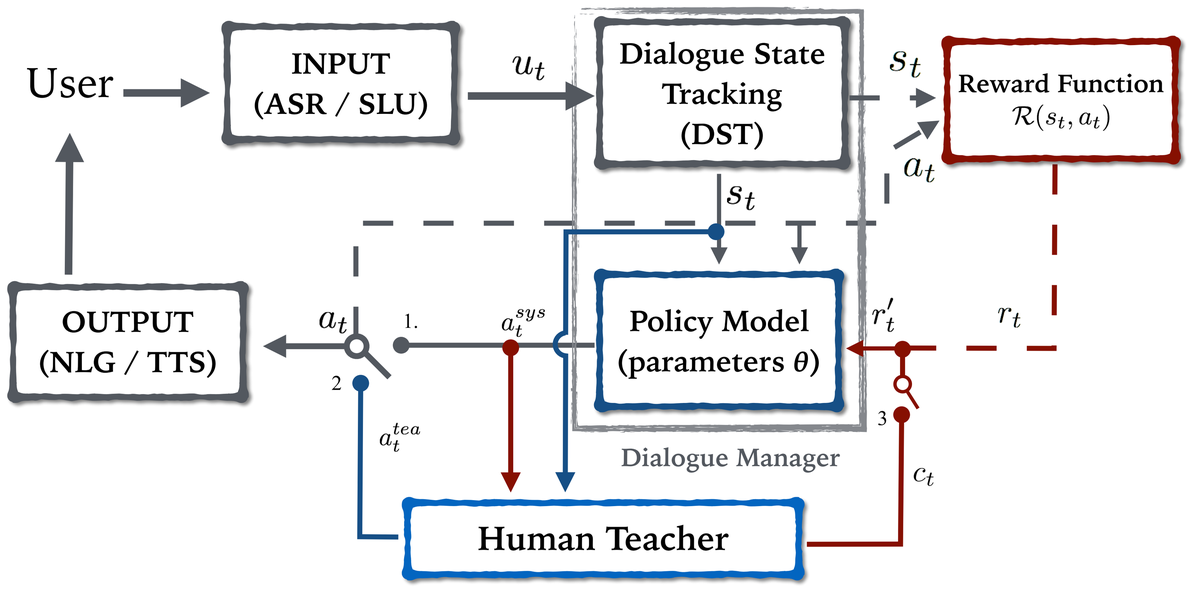


Figure 1: Companion Teaching Framework for On-line Policy Learning

1. The **ASR/SLU** module receives an acoustic input signal from the human user.
2. The **Dialogue State Tracker** keeps the dialogue state up-to-date in the form of dialogue act.
3. The **Human Teacher** then determines **whether to teach the policy model or not:**
   - if yes, then the teacher chooses a **Teaching Strategy** to guide the learning of the policy model.
4. Once the **Policy Model** gets a training signal, it can update the policy parameters using **Reinforcement Learning**.
5. The **NLG/TTS** module sends back the response to the human user.

Therefore, an ideal on-line policy learning framework should be measured using following two criteria:

Efficiency reflects how long it takes for the on-line policy learning algorithm to reach a satisfactory performance level.

Safety* reflects whether the initial policy can satisfy the quality-of-service requirement in real-world scenarios during on-line policy learning period.

* Most previous studies of on-line policy learning have been focused on the ***efficiency* issue**, such as
- Gaussian process reinforcement learning (GPRL) (Gasic et al., 2010),
- Deep reinforcement learning (DRL) (Fatemi et al., 2016; Williams and Zweig, 2016; Su et al., 2016), etc.

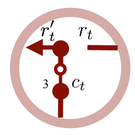
* However, ***safety* is a prerequisite for the efficiency to be achieved**.
- **Reason**: an unsafe on-line learned policy can consequently fail to attract sufficient real users to continuously improve the policy, no matter how efficient the algorithm is.
- **Urgency**: on the ***safety* issue** which little work has been done.

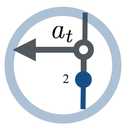
Our Solution: Human-in-the-loop

In this work, we propose a **human-machine hybrid RL framework**, ***Companion Teaching***, which includes a human teacher in the on-line dialogue policy training loop. The involved human teacher accompanies the machine and provides **immediate hands-on guidance at turn level** during on-line policy learning period. This will lead to a ***safer*** policy learning process since the learning is done before any possible dialogue failure at the end.

Teaching Strategies

Teaching via Critic Advice (CA) corresponds to the **right switch (position 3)** in Figure 1. The key idea is to give the policy model an ***extra immediate reward signal*** from teacher, which differentiates between good actions and bad actions.





Teaching via Example Action (EA) corresponds to the **left switch (position 2)**. The human teacher ***directly gives an example action*** at a particular state. The system can learn from teacher's action by considering the action as its own exploration action.

Teaching via Example Action with Predicted Critique (EAPC) take advantages of both EA and CA. The human teacher ***gives an example action*** and meanwhile, an ***extra reward*** will be given to the policy model as well. And this extra reward signal lasts even in teacher's absence. To form this extra reward, the example actions with corresponding states will be collected to train a ***weak action prediction model***.

Experiments & Results

- *Dataset:* Dialogue State Tracking Challenge 2 (DSTC2) dataset
- *DST*: a Rule-based Tracker (Sun et al., 2014)
- ***Policy Model***: a Deep Q-Network (DQN) (Mnih et al., 2015)
  - Two hidden layers to map a belief state s_t to the values of the possible actions a_t at that state, $Q(s_t, a_t; \theta)$.
  - a target network with weight vector θ^- is used.
- ***Reward Design:*** consisting of three parts
  - Length penalty: *-1* at each turn;
  - Success bonus: *+30* at the end of the session;
  - Extra reward: $1 \le c_t \le 20$.
- *User Simulator: an agenda-based user simulator (Schatzmann et al., 2007)*
- ***Simulated Teacher:*** a **well-trained policy model** with success rate 0.78 in our experiment.
- ***Teaching Budget:*** 1500 turns

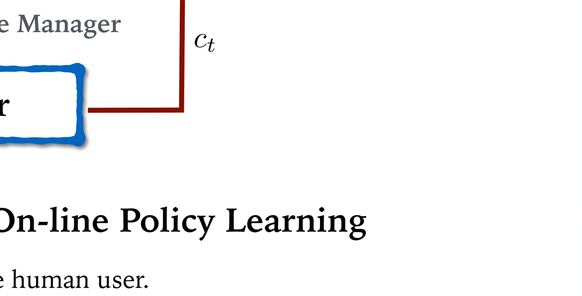


Figure 2: The training curves of moving average success rate.

Figure 3: The testing curves of moving average success rate.

Evaluating Safety:
The moving success rate-#dialogue curve in training (Figure 2), in which the real performance experienced by users when training our system on-line with different companion teaching strategies is reflected.

Evaluating efficiency:
How fast our system can learn from user interaction and human teaching. It can be evaluated by the number of dialogues required to achieve a reasonable performance in the testing curve (Figure 3).

Conclusion In this paper, we propose a novel framework, **Companion Teaching**, to include a human teacher in the dialogue policy training loop to make the learning process ***safe*** and ***efficiency***. Three teaching ways are realized and compared: **critic advice (CA)** where the teacher gives a reward, **example action (EA)** where the teacher gives an action, and a **combination of both (EAPC)**. The experiments shows that EAPC teaching strategy with a small number of teaching can achieve the requirement for on-line dialogue policy learning.

Acknowledgement This work was supported by the Shanghai Sailing Program No. 16YF1405300, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the China NSFC projects (No. 61573241 and No. 61603252) and the Interdisciplinary Program (14JCZ03) of Shanghai Jiao Tong University in China.