On-line Dialogue Policy Learning with Companion Teaching
Lu Chen, Runzhe Yang, et al. Advised by Prof. Kai Yu

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 agents evolve.
- 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 illustrated as following vicious cycle.
  - In this work, we try to propose a practical framework to address the cold start problem.

Possible Solutions to break the vicious cycle
- Insufficient Learning Process (Solvable)
- Unlucky 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. Then 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.

Experiments & Results
- Dataset: Dialogue State Tracking Challenge 2 (DIST2) 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, a) = \theta.
  - A target network with weight vector $\theta$ 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 if $o \in o^{*}$
  - User Simulator: an agenda-based user simulator (Schuster et al., 2007).
  - Trained Teacher: a well-trained policy model with success rate 0.78 in our experiment.
  - Teaching Budget: 1500 turns.

Evaluating Safety: The moving success rate dialogues curve is 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 efficient. These 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 (EACP). The experiments show that EACP teaching strategy with a small number of teaching can achieve the requirements for on-line dialogue policy learning.

Acknowledgement
This work was supported by the Shanghai Sailing Program No. 18YF1405300, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the China NSFC projects (No. 61773241, and No. 61632052) and the Interdisciplinary Program(14JCZ03) of Shanghai Jiao Tong University in China.

个人信息：计算机科学 杨闰哲
邮箱：yang_runzhe@sjtu.edu.cn