On-line Dialogue Policy Learning with Companion Teaching

Lu Chen, Runzhe Yang, Cheng Chang, Zihao Ye, Xiang Zhou and Kai Yu
Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Eng.
SpeechLab, Department of Computer Science and Engineering
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, P. R. China
{chenlusz, yang_runzhe, kai.yu}@sjtu.edu.cn

Abstract

On-line dialogue policy learning is the key for building evolvable conversational agent in real world scenarios. Poor initial policy can easily lead to bad user experience and consequently fail to attract sufficient real users for policy training. We propose a novel framework, *companion teaching*, to include a human teacher in the on-line dialogue policy training loop to address the cold start problem. Here, dialogue policy is trained using not only user's reward but also teacher's example action as well as estimated immediate reward at turn level. Simulation experiments showed that, with a small number of human teaching dialogues, the proposed approach can effectively improve user experience at the beginning and smoothly lead to good performance with more user interaction data.

1 Introduction

Statistical dialogue management has attracted great interest in both academia and industry due to its promise of data-driven interaction policy learning. Since policy learning is a sequential decision problem, *reinforcement learning* (RL) has been widely used for policy training. *Partially observable Markov decision process* (POMDP) (Kaelbling et al., 1998), as the mainstream approach, has been reported to achieve impressive performance gain compared to rule-based DM (Williams and Young, 2007; Young et al., 2010). However, it is still rarely used in real world scenarios. This is largely because most POMDP based policy learning research is usually carried out using either a user simulator or unreal users (such as lab users). The off-line trained policy is not guaranteed to work well in real world scenarios. Therefore, on-line policy learning has been of great interest. We believe that an ideal on-line policy learning framework should be measured using 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. On the other side, *safety* is a pre-requisite for the efficiency to be achieved. This is because, no matter how efficient the algorithm is, an unsafe on-line learned policy can lead to bad user experience at the beginning of learning period and consequently fail to attract sufficient real users to continuously improve the policy. Therefore, it is important to address the safety issue, on which little work has been done.

In this paper, a novel safe on-line policy learning framework is proposed, referred to as *companion teaching*. This is a human-machine hybrid RL framework. Different from the whole dialogue based human demonstration approach (Chinaei and Chaib-draa, 2012), here a 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.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 198–204,
Valencia, Spain, April 3-7, 2017.

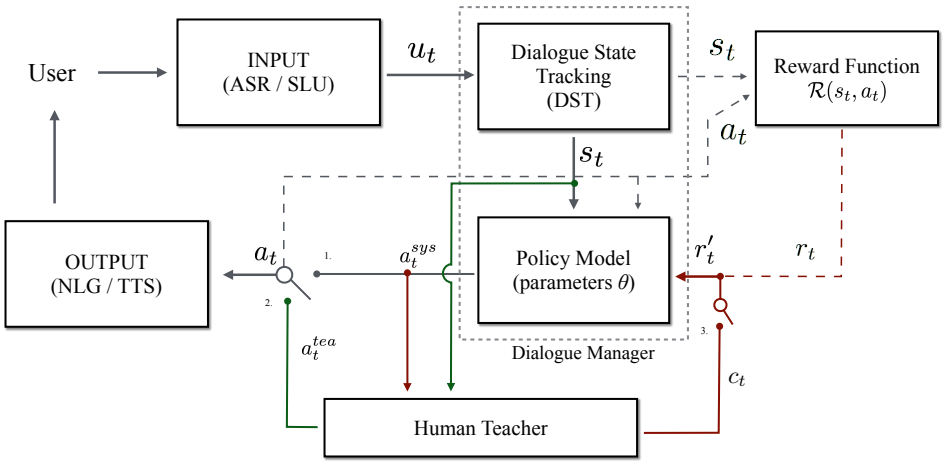

Figure 1: Companion Teaching Framework for On-line Policy Learning

A major contribution of the paper is to introduce example actions of the human teacher to guide on-line policy learning of the agent. Furthermore, we combine example action based guidance with an additional action prediction model to continuously give extra supervision reward signal in teacher's absence. Simulated experiments using deep Q-learning show that the combined teaching strategy significantly improves both *safety* and *efficiency* within a fixed time budget of the human teacher.

2 Companion Teaching for On-line Dialogue Policy Learning

Including human in the loop has been recognized as an effective way to accelerate on-line policy learning (Thomaz and Breazeal, 2006; Khan et al., 2011; Cakmak and Lopes, 2012; Loftin et al., 2016). Most previous approaches employ teaching signals at the end of dialogues, either the whole human-to-human dialogue history or a single reward to evaluate the human-machine dialogue performance (Su et al., 2016; Ferreira and Lefèvre, 2015). Here, we propose a new three-party turn-level human-machine hybrid learning framework to address both the safety and the efficiency issues of on-line policy learning.

2.1 Companion Teaching Framework

In the *companion teaching* framework, there are three intelligent participants: machine dialogue manager (agent), human user and human teacher. Dialogue manager consists of dialogue state tracker and policy model. The goal of on-line policy learning is to learn policy from data via interaction with human users in real scenarios. Here, *human teacher* is the extra party compared with the classic statistical dialogue manager architecture (Young et al., 2013). The human teacher, as a companion of the agent, guides policy learning at each turn, hence, referred to as *companion teaching*. The framework is depicted in figure 1:

At each turn, the ASR/SLU module receives an acoustic input signal from the human user and the dialogue state tracker keeps the dialogue state up-to-date in the form of dialogue act. In this paper, it is assumed that the dialogue states from the tracker are transparent to both policy model and human teacher. The human teacher then determines whether to teach the policy model or not and chooses an appropriate way to guide the learning of the policy model. Once the policy model gets a training signal, either from the teacher or from the user, it can update the policy parameters using reinforcement learning. Since the "teaching" is carried out at turn level with immediate effect, it is likely that bad choices resulting from the poor or unstable policy can be effectively reduced.

Note that the assumption of dialogue state sharing between policy model and the human teacher is consistent with realism for two reasons. First, under the real work model of customer service, call-center people needs to refer to database query results given by the system, which must contain the information of dialogue states inferred by the system. Second, when support staffs reply to clients, they often choose replies among several recommended candidates rather than type answers. This fact implies human can observe system's dialogue act and even reply in this format.

2.2 Teaching Strategy

As indicated in figure 1, there are two switches representing two strategies of teaching.

Teaching via Critic Advice (CA) corresponds to the right switch in figure 1. The key idea is for the human teacher to give the policy model an extra immediate reward signal which differentiates between good actions and bad actions. CA is also referred to as turn-level *reward shaping*, which has been investigated in various applications (Wiewiora et al., 2003; Thomaz and Breazeal, 2008; Judah et al., 2010). Previous works show that teaching agent via additional turn-level critic advice can make agent significantly outperform those under pure RL. A major problem of Critic Advice based teaching is that the critique signal can only be given after a hazardous action is taken by the system. It may not be able to dramatically improve system policy immediately. Hence, it is hard to avoid unsafe situations while system is trying to do exploration, especially, at the beginning of learning.

To address the shortcoming of CA, we propose **Teaching via Example Action (EA)**. It corresponds to the left switch in figure 1. Here, 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 within the RL framework. Note that this strategy is distinctly different from imitation learning in (Abbeel and Ng, 2004). The goal of imitation learning is to figure out the teacher's reward function rather than updating the system's policy parameters. In contrast, in the companion teaching framework, the role of human teacher's example action is more like a guidance to agent exploration and agent will still get a corresponding reward from the environment. This training method is pragmatic since it prevents unsafe situations during starting period by guiding agent's exploration. However, this EA approach requires more time cost of the human teacher than the CA approach.

The critic advice method can make the learning more effective and the example action method can make the learning process safer. In order to take advantages of both EA and CA, we further propose to combine the two, i.e. **Teaching via Example Action with Predicted Critique (EAPC)**. Here, the human teacher gives an example action and meanwhile, an extra reward c_t 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 dialogue states will be collected to train a weak action prediction model. The input of this model is the dialogue state, and the output is the probabilities for each action. When the human

Algorithm 1 EAPC Algorithm

Require:
Observe N_o steps teaching before training the action prediction model $\mathcal{P}$. the interval N_i of updating $\mathcal{P}$, the maximal extra reward $\delta > 0$.

Initialize policy model π and action prediction model $\mathcal{P}$
Initialize replay memory $\mathcal{D} = \{\}$ and teacher experience $\mathcal{E} = \{\}$
for *episode* $= 1, N$ **do**
  Update the dialogue state s_0
  for $t = 0, T$ **do**
    Set extra reward $c_t \leftarrow 0$
    Get system action $a_t^{sys} \sim \pi(\cdot|s_t)$
    $a_t \leftarrow a_t^{sys}$
    if $human\ teaching$ is $true$ **then**
      Teacher gives the action a_t^{tea}
      $a_t \leftarrow a_t^{tea}$
      Set extra reward $c_t \leftarrow \delta$
      Store the pairs (s_t, a_t^{sys}) in $\mathcal{E}$
      if $|\mathcal{E}| > N_o$ and $N_i \% |\mathcal{E}| = 0$ **then**
        Supervised training $\mathcal{P}$ on dataset $\mathcal{E}$
      end if
    else
      if $|\mathcal{E}| > N_o$ **then**
        $\mathcal{P}(s_t)$ predicts a a_t^{pred} and tells the estimated probability p
        if $a_t^{sys} = a_t^{pred}$ **then**
          $c_t \leftarrow \delta p$
        else
          $c_t \leftarrow -\delta p$
        end if
      end if
    end if
    Give the action a_t to the environment, observe the reward r_t and update the dialogue state s_{t+1}
    $r'_t = r_t + c_t$
    Store $\{s_t, a_t, r'_t, s_{t+1}\}$ in $\mathcal{D}$
    Update the policy model π by RL
  end for
end for
return policy π

teacher is not involved in, the supervised model will predict the most probable teacher action under the current dialogue state. If the predicted action is same as the action given by the policy model, the extra reward δ discounted by the probability of the predicted action will be given to the policy model. Otherwise, the extra reward $-\delta$ discounted by the probability of the predicted action will be given to the policy model. This method is shown as algorithm 1.

2.3 Reinforcement Learning Algorithm

The *companion teaching* framework does not depend on a specific reinforcement learning algorithm, hence is compatible with all existing algorithms. In this paper, we implement a Deep Q-Network (DQN) (Mnih et al., 2015) with two hidden layers to map a belief state s_t to the values of the possible actions a_t at that state, $Q(s_t, a_t; \theta)$, where θ is the weight vector of the neural network.

In DQN, two techniques were proposed to overcome the instability of neural network training, namely experience replay and the use of a target network (Mnih et al., 2015). At every turn, the transition including the previous state s_t, previous action a_t, corresponding reward r'_t and current state s_{t+1} is put in a finite pool $\mathcal{D}$. When the teaching method EA is used in the t-th turn, $a_t = a_t^{tea}$, otherwise $a_t = a_t^{sys}$. When CA is used, $r'_t = r_t + c_t$, otherwise $r'_t = r_t$. Once the pool has reached its maximum size, the oldest transition will be deleted. During training, a mini-batch of transitions is uniformly sampled from the pool, i.e. $(s_t, a_t, r'_t, s_{t+1}) \sim U(D)$. This method removes the instability arising from strong correlation between the subsequent transitions of a dialogue. Additionally, a target network with weight vector θ^- is used. This target network is similar to the Q-network except that its weights are only copied every K steps from the Q-network, and remain fixed during all the other steps. The loss function for the Q-network at each iteration takes the following form:

$$L(\theta) = \mathrm{E}_{(s_t, a_t, r'_t, s_{t+1}) \sim U(\mathcal{D})} \left[\left(r'_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta^-) - Q(s_t, a_t; \theta) \right)^2 \right]$$

where $\gamma \in [0, 1]$ is the discount factor.

3 Experiments

Simulation experiments were performed to assess the proposed companion teaching framework and three different teaching strategies.

We implement an agenda-based user simulator (Schatzmann et al., 2007) to emulate the behavior of the human user, and use a well-trained policy model with success rate 0.78 serving as the human teacher in our experiment. As for data set, we use the Dialogue State Tracking Challenge 2 (DSTC2) dataset (Henderson et al., 2014), which is in a restaurant information domain. This domain has 7 slots of which 4 can be used by the system to constrain the database search. The summary action space consists of 16 summary actions. We use a rule-based tracker (Sun et al., 2014) for dialogue state tracking.

As the reward, at each turn, a reward of -1 was given to the policy model, and at the end of the dialogue a reward of +30 was given if the dialogue finishes successfully. The maximal extra reward δ is 1, and the maximum of turns is 20.

During training, the teacher has a fixed time budget of 1500 turns to perform teaching at the beginning. Intermediate policies were recorded at every 500 dialogues. Each policy was then evaluated using 1000 dialogues when testing.

3.1 Evaluation Metrics

We mainly care about *safety* and *efficiency* in the comparison of different teaching strategies of companion teaching for dialogue policy learning.

The degree of *safety* can be assessed by investigating the moving success rate-#dialogue curve in training, which reflects the real performance experienced by users when training our system on-line with different teaching strategies. If the success ratio keeps high in the curve, we think it is safe.

The *efficiency* should be evaluated by the learning speed: 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.

3.2 Experiment Results

We compared the moving average success rate [1] for three different teaching strategies and the results are given in Figure 2. We can figure out that

[1] For each point on the curve, the success rate is the average of previous 1000 dialogues when training.

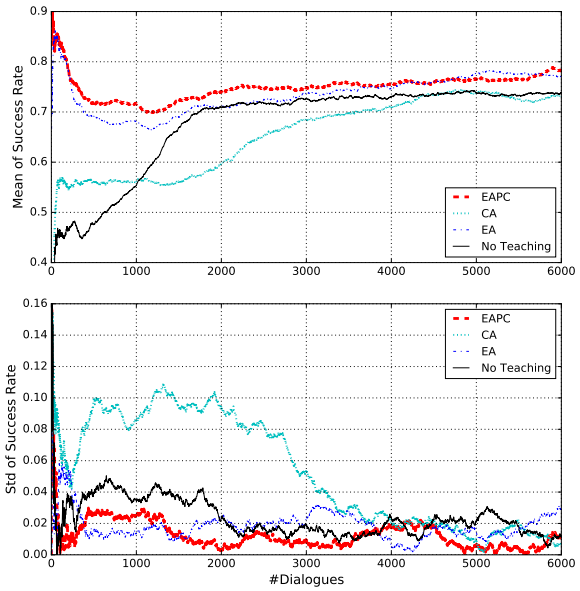

Figure 2: The training curves of moving average success rate. The top and the bottom are the means and standard deviations of success rate respectively for 3 trials.

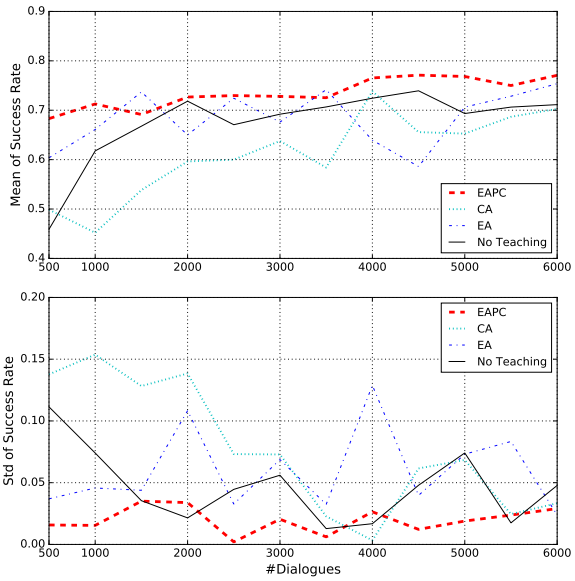

Figure 3: The testing curves of success rate. The top and the bottom are the means and standard deviations of success rate respectively for 3 trials.

the policy with EAPC teaching strategy performs best when training, with always more than 70% average success rate, which means that the learning with EAPC is safer. Better still, the standard deviation is also the smallest, which indicates a stable learning process. Besides, EA has similar performance with EAPC, both of them can achieve the requirement of safety when training.

In figure 3, we compared the testing curves and investigated the learning efficiency of different strategies. The results show that the learning with EAPC is more efficient and maintains the lowest derivation during learning. After 500 dialogues interaction, it can obtain nearly 70% success rate, 22.4% higher compared with the one without teaching. And it is even about 10% higher than that of only using EA method.

Taken together, the teaching strategy EAPC can achieve the requirement *safety* and *efficiency* of on-line dialogue policy learning.

4 Conclusion and Future Work

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*. 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 demonstrated that our proposed EAPC teaching strategy with a small number of teaching can achieve the requirement of both *safety* and *efficiency* for on-line dialogue policy learning.

Currently, the evaluation of our proposed framework was only done in simulation experiments. We expect to deploy our proposed framework with real human teachers in real-world scenarios to verify the effectiveness of companion teaching. Furthermore, in this paper, the teaching were all done at the beginning of on-line training. This may be too simplistic and uneconomic in real world applications. Further work will be needed to answer the question of when for the human to teach.

Acknowledgments

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.

References

Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pages 1–8.

Maya Cakmak and Manuel Lopes. 2012. Algorithmic and human teaching of sequential decision tasks. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial IntelligenceI*, pages 1536–1542.

Hamid R. Chinaei and Brahim Chaib-draa. 2012. An inverse reinforcement learning algorithm for partially observable domains with application on healthcare dialogue management. In *2012 Eleventh International Conference on Machine Learning and Applications (ICMLA)*, pages 144–149. IEEE.

Mehdi Fatemi, Layla El Asri, Hannes Schulz, Jing He, and Kaheer Suleman. 2016. Policy networks with two-stage training for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 101–110, Los Angeles, September. Association for Computational Linguistics.

Emmanuel Ferreira and Fabrice Lefèvre. 2015. Reinforcement-learning based dialogue system for human–robot interactions with socially-inspired rewards. *Computer Speech and Language*, 34(1):256–274, November.

Milica Gasic, Filip Jurcicek, Simon Keizer, Francois Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2010. Gaussian processes for fast policy optimisation of pomdp-based dialogue managers. In *Proceedings of the SIGDIAL 2010 Conference*, pages 201–204, Tokyo, Japan, September. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Kshitij Judah, Saikat Roy, Alan Fern, and Thomas Dietterich. 2010. Reinforcement learning via practice and critique advice. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 481–486.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, May.

Faisal Khan, Bilge Mutlu, and Xiaojin Zhu. 2011. How do humans teach: On curriculum learning and teaching dimension. In *Proceedings of the Advances in Neural Information Processing Systems 24*, pages 1449–1457.

Robert Loftin, Bei Peng, James MacGlashan, Michael L. Littman, Matthew E. Taylor, Jeff Huang, and David L. Roberts. 2016. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous Agents and Multi-Agent Systems*, 30(1):30–59.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152, Rochester, New York, April. Association for Computational Linguistics.

Pei-Hao Su, Milica Gašić, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*.

Kai Sun, Lu Chen, Su Zhu, and Kai Yu. 2014. The sjtu system for dialog state tracking challenge 2. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 318–326, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Andrea L. Thomaz and Cynthia Breazeal. 2006. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, volume 6, pages 1000–1005.

Andrea Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737, April.

Eric Wiewiora, Garrison Cottrell, and Charles Elkan. 2003. Principled methods for advising reinforcement learning agents. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 792–799.

Jason D. Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422, April.

Jason D. Williams and Geoffrey Zweig. 2016. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269*.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174, April.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.