Curiosity-driven Reinforcement Learning for Dialogue Management

Paula Wesselmann, Yen-Chen Wu, Milica Gašić

Introduction

- Dialogue Manager (DM) is the brain of the dialogue system
- DM tracks believes and determines behaviour of system
- DM uses Reinforcement Learning (RL) to learn a policy
- RL is learning from feedback/ rewards

Motivation

- Hard to obtain user feedback/ external reward
- Explore more efficiently
- Improve policy learning

Policy learning $\pi(b) : \mathbb{B} \rightarrow \mathbb{A}$

- Choose policy that maximises total Reward
- = Policy with optimal Q function

$$Q : \mathbb{B} \times \mathbb{A} \rightarrow \mathbb{R}$$

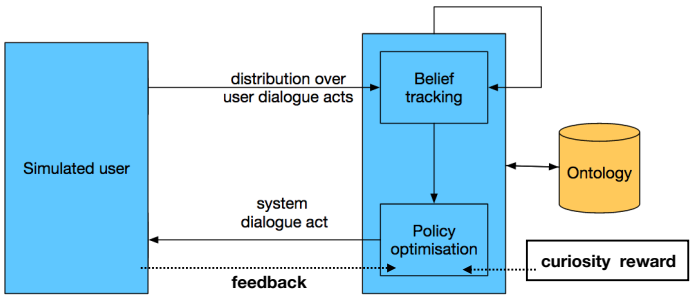
$$Q^{\pi}(b,a) = E_{\pi}\{\sum_{k=0}^{T-t} \gamma^k r_{t+k} | b_t = b, a_t = a\}$$

$$\pi^*(b) = \arg\max_a Q^*(b,a)$$

- Q-function represented by deep Neural Network
- Policy optimisation with DQN
- For DQN policy is used eps-greedily to determine action

PyDial

- CUED Python Statistical Dialogue System



The Dialogue Manager:
1. Updates believe state of system b_t
2. Selects action a_t

Intrinsic Reward Signal

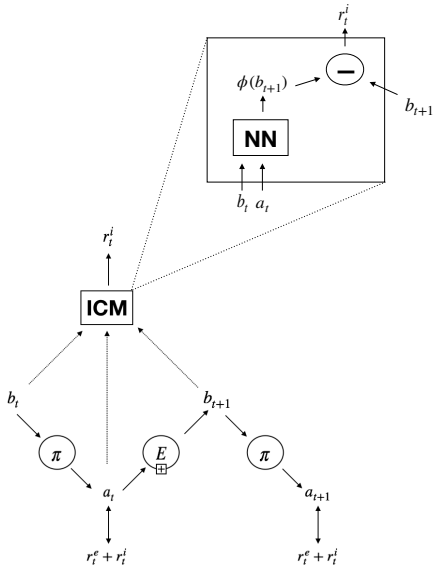
- RL relies on reward signals (usually external feedback)
- For Dialogue systems those reward signals are often hard to obtain, not accurate or even absent
- Intrinsic reward systems such as curiosity, can replace external feedback or be used in addition to external rewards
- Explore more efficiently by actively seeking new knowledge, no random exploring

Intrinsic Curiosity Module (ICM)

- State prediction error as curiosity reward (Pathak et al. 2017)
- No random exploration needed anymore i.e. no eps-greedy

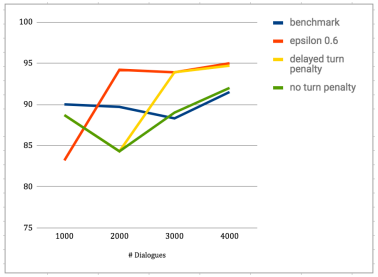
$$r_t^i = \frac{\eta}{2} || \phi(b_{t+1}) - b_{t+1} ||_2^2 = \eta L_F$$

$$\min_{\theta_P, \theta_F} \left[-\lambda \mathbb{E}_{\pi(s_t;\theta_P)}[\sum_t r_t] + \beta L_F \right]$$

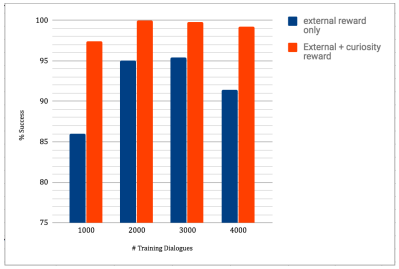


Handcoded Curiosity Experiments

- Increased initial exploration (random)
- Vary the use of turn penalty as reward signal



Preliminary Results



Most simple environment, only one seed;

Next Steps

- Tuning the reward signal and other parameters
- Intrinsic reward signals only
- Predicting using larger (more specific) action space
- Alternative curiosity rewards to state prediction error
- Implement in hierarchical framework

Reference:
Pathak, D. and Agrawal, P. and Efros, A. A. and Darrell, T.
Curiosity-driven Exploration by Self-supervised Prediction (2017)