On-line policy optimisation of spoken dialogue systems via live interaction with human subjects

Milica Gašić, Filip Jurčiček, Blaise Thomson, Kai Yu and Steve Young

Cambridge University Engineering Department

{*mg436,fj228,brmt2,ky219,sjy*}*@eng.cam.ac.uk*

1 INTRODUCTION

- Hidden Information State system
  - POMDP-based dialogue manager that maintains a distribution over possible states at every dialogue turn
  - Optimises the policy in a smaller summary space
  - Requires a large number of dialogues to train a policy
  - Relies on the use of a user simulator
- Gaussian process reinforcement learning
  - POMDP dialogue policy π maps (summary) states $\mathbf{b}$ into actions a so that the total reward is maximal:

$$Q(\mathbf{b}, a) = \max_\pi E_\pi \left(\sum_{\tau=t+1}^{T} \gamma^{\tau-t-1} r_\tau | \mathbf{b}_t = \mathbf{b}, a_t = a \right)$$

  - Q-function can be modelled as a Gaussian process (GP), which for every summary state-action pair $(\mathbf{b}, a)$ gives a Gaussian distribution $\mathcal{N}\left(\overline{Q}(\mathbf{b}, a), cov((\mathbf{b}, a), (\mathbf{b}, a))\right)$
  - This enables faster policy optimisation.
- We investigate policy optimisation directly from human interaction
  - Using a low risk learning technique based on GPs,
  - Via Amazon Mechanical Turk service,
  - To replace the need for a user simulator.

2 LOW-RISK POLICY MODEL

On-line learning requires manual balancing of *exploitation* of current estimate of the Q-function and *exploration* of unexplored actions.

- We propose a stochastic policy model which
  - Automatically balances exploration/exploitation via sampling from Gaussian distributions for $Q(\mathbf{b}, a)$ for every action a and taking the action which has the highest sampled Q-value:

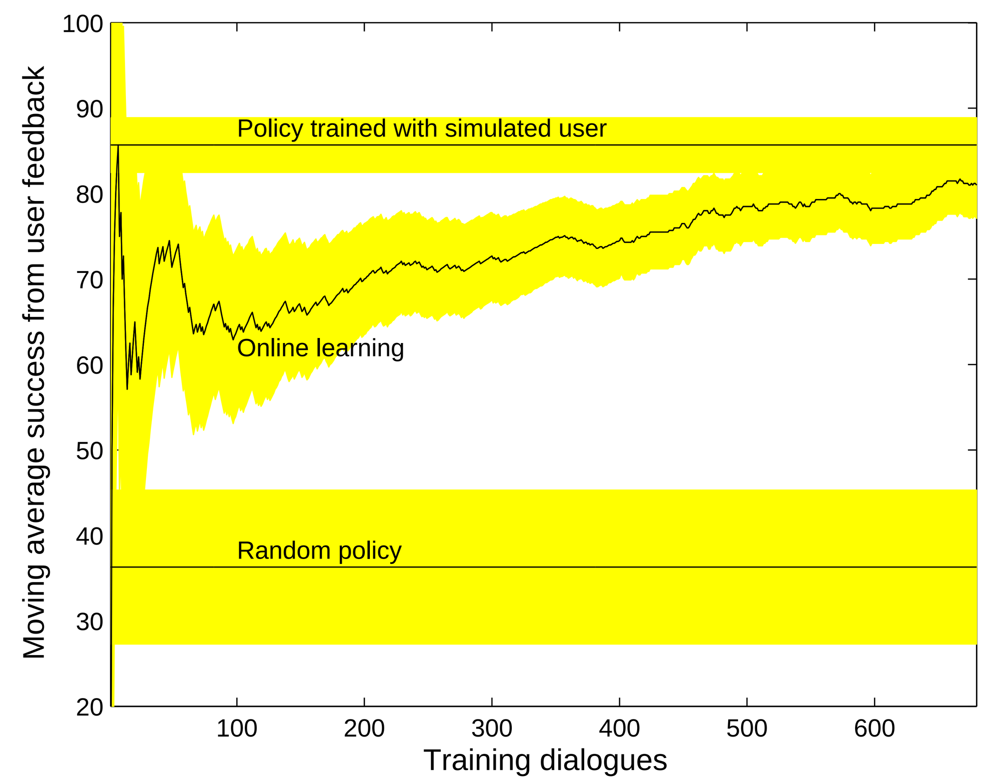
$$Q^i(\mathbf{b}, a_i) \sim \mathcal{N}(\overline{Q}(\mathbf{b}, a_i), cov((\mathbf{b}, a_i), (\mathbf{b}, a_i))))$$

$$a = \arg\max_{a_i} Q^i(\mathbf{b}, a_i)$$

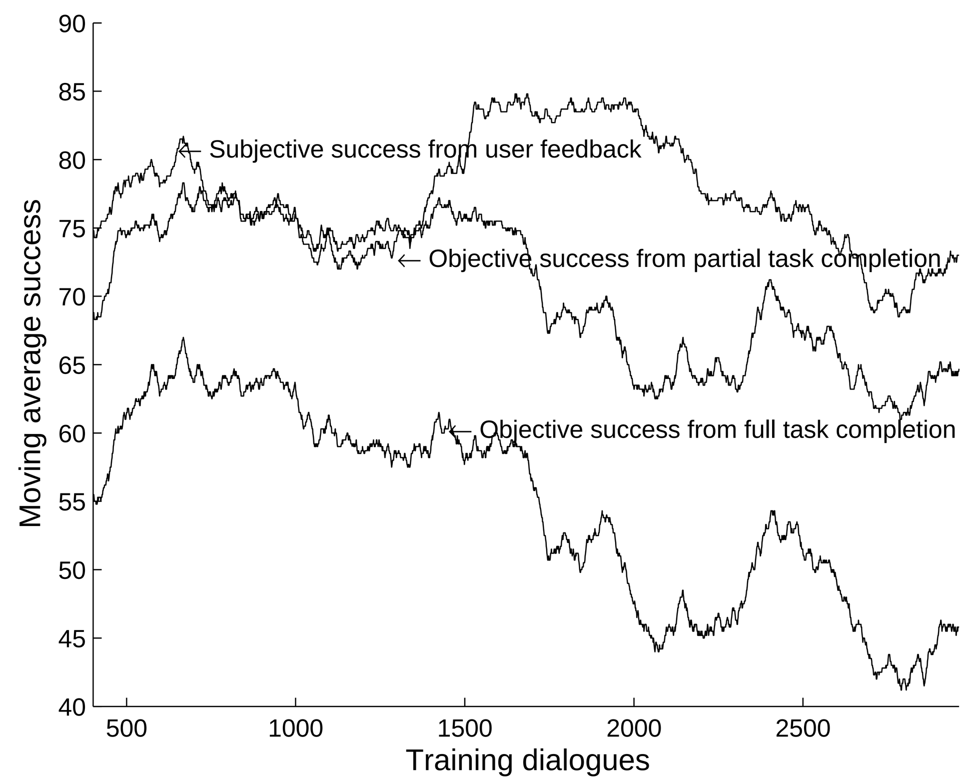
  - This reduces the risk of taking bad actions during learning.

3 ON-LINE LEARNING

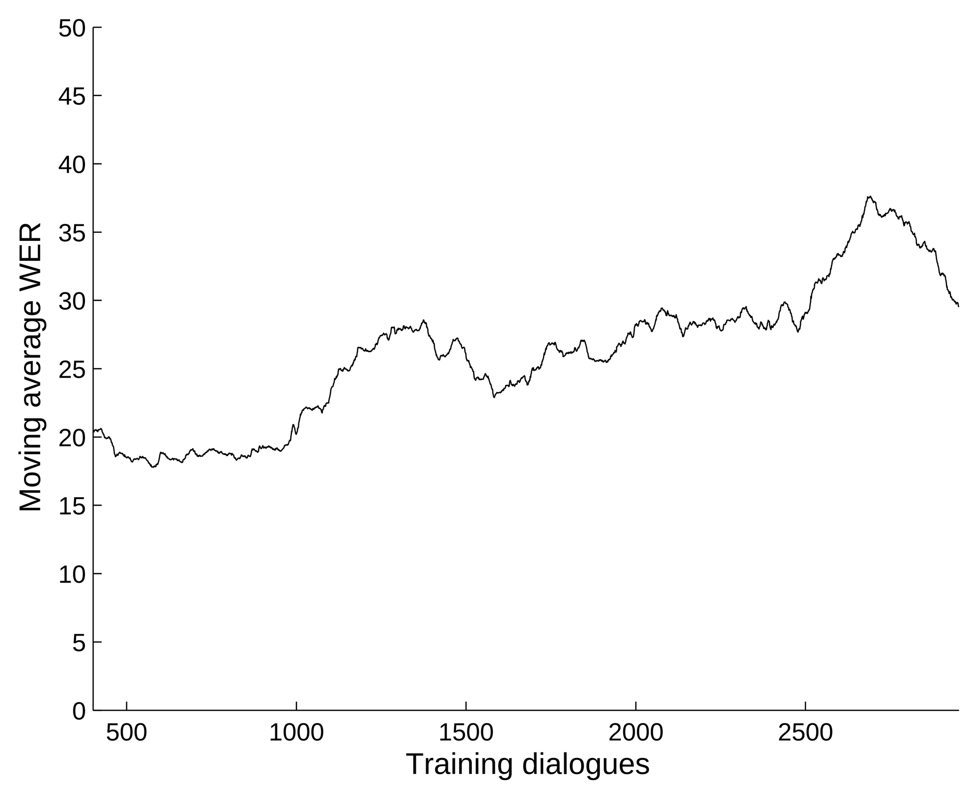
- Experimental Set-up
  - 252 users were recruited via Amazon Mechanical Turk and provided with a dialogue task.
  - They called a telephone-based dialogue system for Cambridge restaurant domain.
  - 2960 dialogues were collected.
  - Users gave a binary feedback at the end of every dialogue for on-line learning reward.
- Initial training
  - Performance of the policy that is learning on-line during initial 680 dialogues:



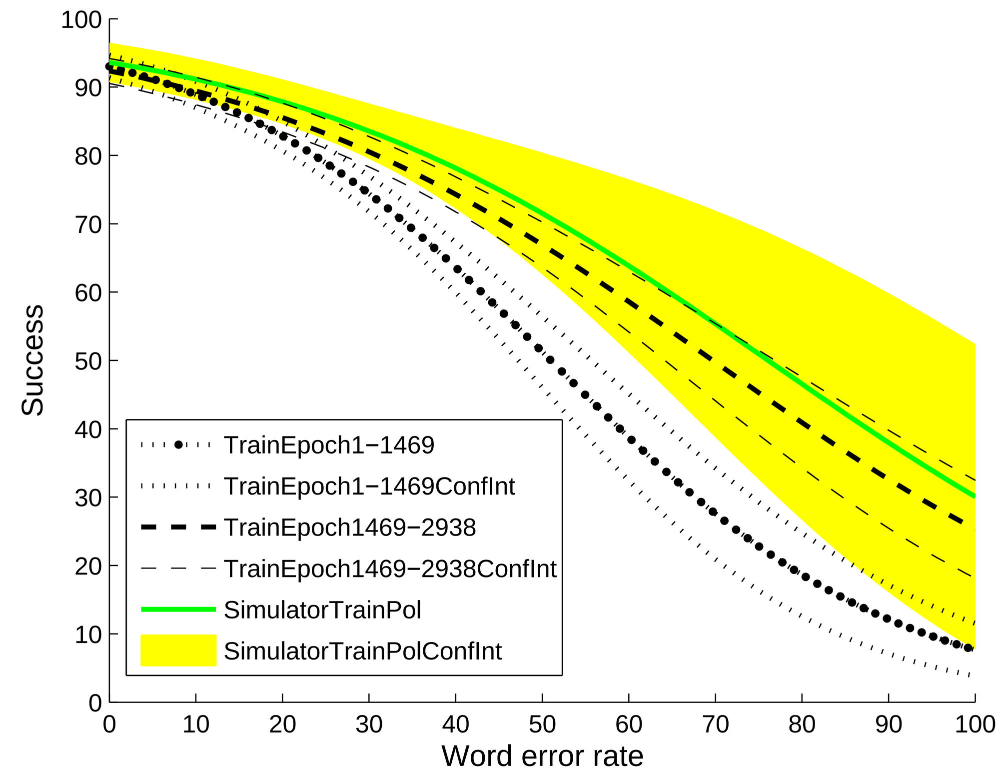
- Long-term adaptation
  - Performance based on user feedback is compared to performance based on objective measures
    - full task completion – the system fully completed the dialogue task
    - partial task completion – the system partially completed the dialogue task
  - Exhibits
    - Variability in performance
    - Divergence between objective and subjective measures



- Influence of word error rate
  - The word error rate was not stable over the course of learning.



  - Regressing performance against different error rates shows that the policy produced after the second stage of learning is
    - more robust
    - statistically indistinguishable from the policy trained on the simulated user



- Human perception of dialogue success

	Random policy	Online learning	Simulator trained
Positive user feedback	36.3	76.9	85.7
Fully completed task	17.7	53.8	63.7
p(feedbck= 1|complt= 1)	0.80	0.94	0.94
p(feedbck= 1|complt= 0)	0.26	0.57	0.68
Total dialogues	114	2960	466

  - Retraining the policy offline on dialogues where subjective and objective score are the same gave better performance on the simulated user.

4 CONCLUSION

- Policy trained on-line reached the performance of a policy trained on a simulated user.
- Once the policy reached reasonable behaviour it is difficult for the users to estimate the reward accurately.
- While the framework deals well with noisy inputs from the recogniser, inaccuracy of the assigned reward can lead to variability in performance.