Optimising spoken dialogue systems using Gaussian process reinforcement learning for a large action set

Thomas F W Nicholson (tfwn2), Milica Gašić (mg436)
Cambridge University Engineering Department

**Dialogue System Policy Optimisation**

View as MDP:

- **Actions**: dialogue actions
- **State**: user intent belief state
- **Reward**: successful completion

Learn a policy that takes action that maximises long term reward

**GPSARSA**

Expected long-term reward as a function of belief and action:

$$Q^\pi(s, A) = E(L_t | s_t = s, A_t = A)$$

Model using a Gaussian Process with product of kernels:

- **Belief state**: squared exponential
- **Action**: kroenecker delta

Optimise Q function using online policy updates. Action is then chosen by policy:

$$\pi(A) = \operatorname{argmax}_A Q(s, A)$$

**Action modeling problems**

1) Learning takes place in summary action space

- **Summary action**: inform_name
- **Master action**: inform(name="Curry House")

2) All or nothing distance between actions

**Solution**

View actions as trees:

- inform(type="Chinese")
- inform(type="Chinese", name="Wanli")

Use tree kernel between actions:

- Count common sub/subset trees between them:
  - type: Chinese
  - name: Wanli

  $$k(s, s') = 3$$

**Intricacies**

- Large number of actions means large Gram matrix
  - Gaussian Process over each dialogue intent
  - Invert each one separately

Finding highest scoring action is expensive

- Don't want to evaluate all actions (~70,000)
- Can onstruct tree per-layer to maximise Q-value

**Extensions**

1) Per-layer weighting
   - Some layers in tree more important

2) Distributed representation for action values
   - Bangladeshi similar to Indian but not to Italian