Policy optimisation of POMDP-based dialogue managers without state-space compression

Milica Gašić, Matthew Henderson, Blaise Thomson, Pirros Tsiakoulis and Steve Young
Cambridge University Engineering Department

1 INTRODUCTION

POMDP-based approach to dialogue management

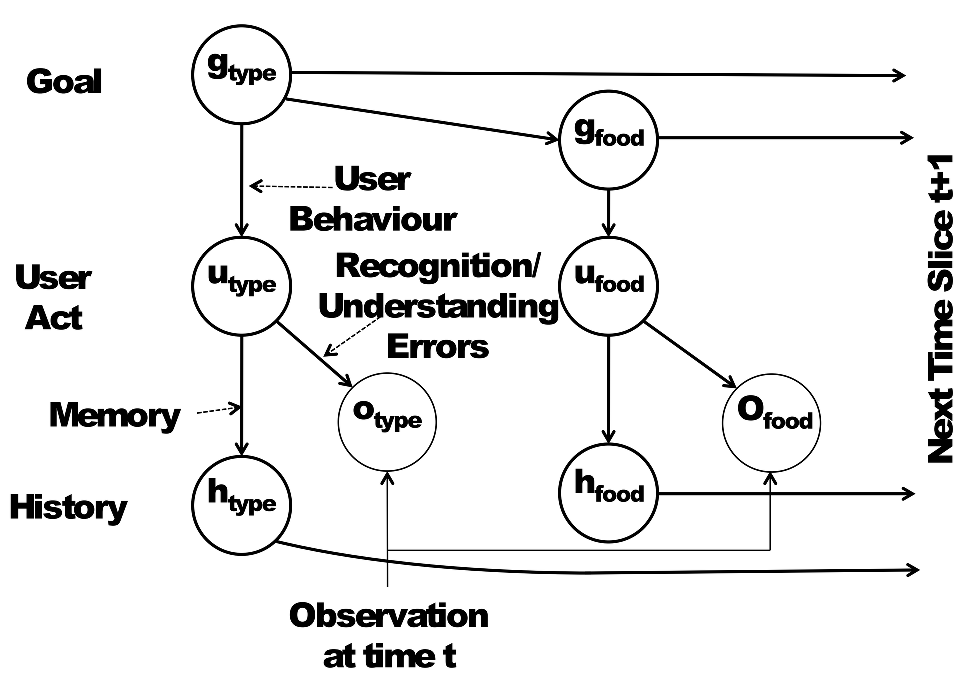
- Maintains a distribution over possible dialogue states – *the belief state*
- Enables robustness to speech recognition errors
- Policy optimisation requires:
  - Designer's effort to specify a state-space compression
  - A large number of dialogues

Gaussian process-based optimisation

- Enables efficient learning by exploiting correlations in nearby parts of the space

2 BAYESIAN UPDATE OF DIALOGUE STATE DIALOGUE MANAGER

The dialogue state is factored into conditionally independent elements.



The belief state $\mathbf{b} \in \mathcal{B}$ consists of marginal probabilities of the goal nodes and the history nodes.

3 POLICY OPTIMISATION

Policy optimisation is performed via reinforcement learning from interaction with a simulated user.

Natural actor-critic algorithm

- Features of the belief space $\mathcal{B}$ important for learning are extracted into summary space $\mathcal{C}$
- Summary space $\mathcal{C}$ and summary action space $\mathcal{A}$ are mapped into feature space $\mathcal{F}$
- Policy is parameterised as a linear function of features from $\mathcal{F}$ passed through a soft-max nonlinearity

$$\pi(a|\mathbf{c},\theta) = \frac{e^{\theta f_a(\mathbf{c})}}{\sum_a e^{\theta f_a(\mathbf{c})}}$$

- Gradient methods are used for policy optimisation

Gaussian process Sarsa algorithm

- The Q-function is the expected cumulative reward

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

- Optimising the Q-function Q^{π} is equivalent to optimising the policy π
- The Q-function is modelled as a Gaussian process

$$Q^{\pi}(\mathbf{b},a) \sim \mathcal{GP}(0, k((\mathbf{b},a),(\mathbf{b},a))),$$

where $k((\mathbf{b},a),(\mathbf{b}',a')) = k_{\mathcal{B}}(\mathbf{b},\mathbf{b}')k_{\mathcal{A}}(a,a')$

4 KERNEL CHOICE

The optimisation can take place either in $\mathcal{C}$ or $\mathcal{B}$

- Kernel on the summary space $\mathcal{C}$

$$k_{\mathcal{C}}(\mathbf{c},\mathbf{c}') = \langle\mathbf{c},\mathbf{c}'\rangle + 1 \quad (1)$$

- Kernel on the belief space $\mathcal{B}$ is the sum of individual kernels on each of the l hidden nodes

$$k_{\mathcal{B}}(\mathbf{b},\mathbf{b}') = \sum_l \langle\mathbf{b}_l,\mathbf{b}_l'\rangle, \quad (2)$$

- Kernel on the summary action space $\mathcal{A}$

$$k_{\mathcal{A}}(a,a') = 1 - \delta_a(a') \quad (3)$$

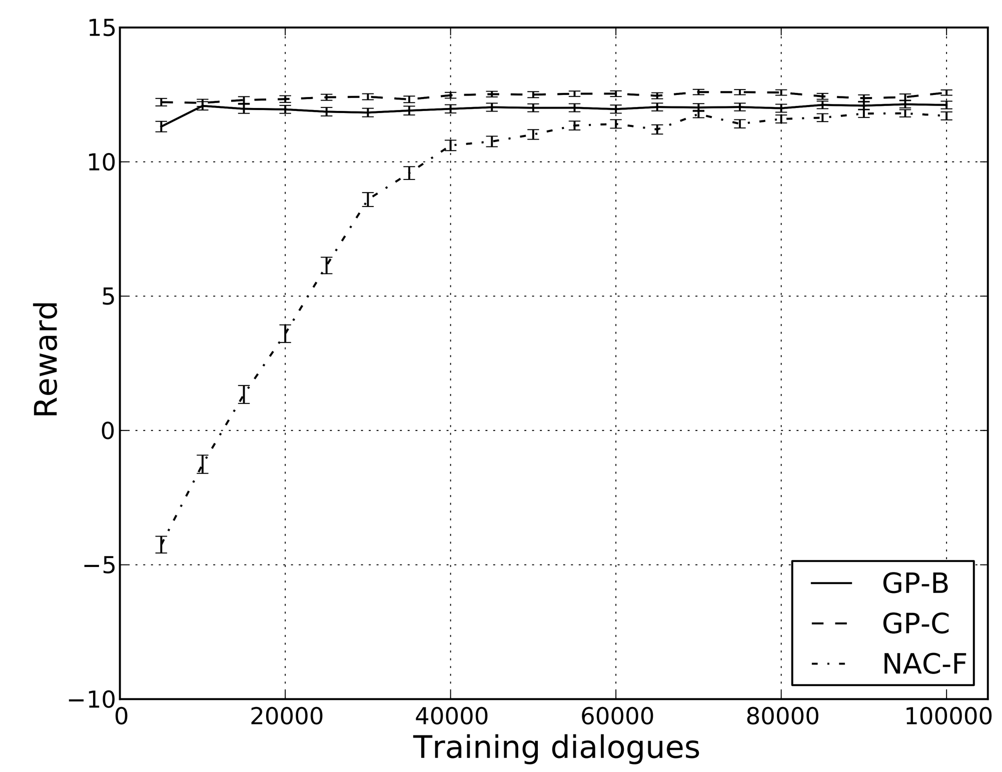
5 EXPERIMENTS

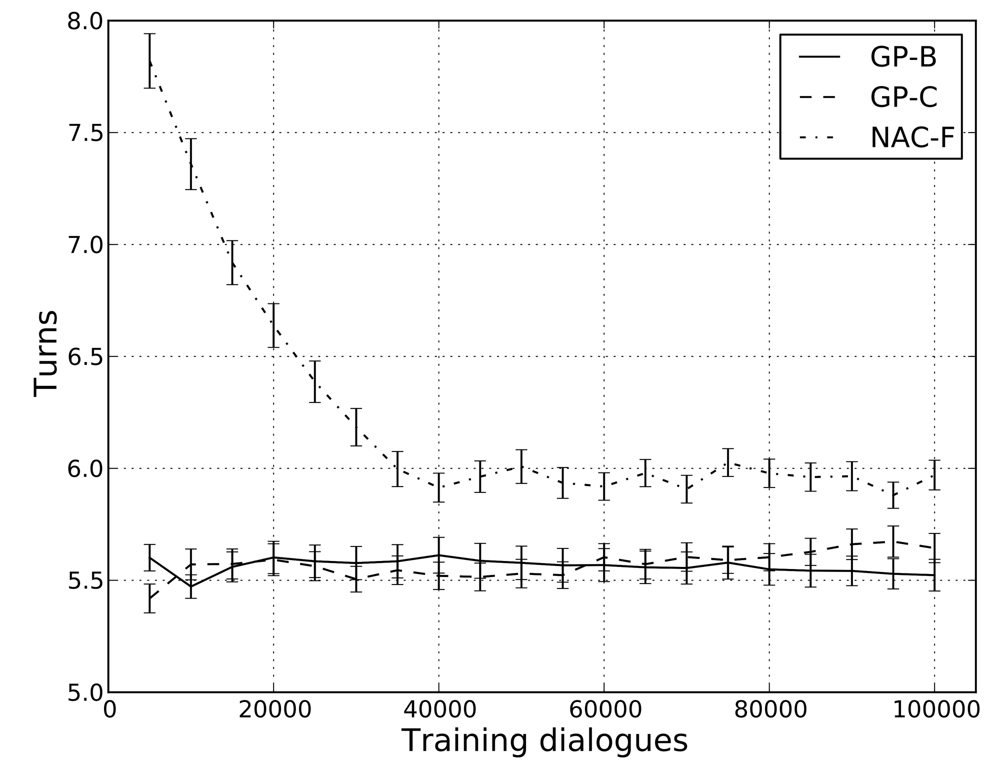
Set-up

- TopTable restaurant information domain for Cambridge
- Contains 150 venues and each has 8 attributes
- Summary space $\mathcal{C}$: 200 binary features
- Belief space $\mathcal{B}$: 25 hidden nodes and each represents a distribution over 3 to 150 values
- Summary action space $\mathcal{A}$: 16 summary actions

Training procedure

- Learning from interaction with the simulated user
- Learning curves:





- Qualitative policy examination:

	GP-B	GP-B-PART	GP-C	NAC-F
Inform	71.1%	75.2%	58.7%	59.1%
Select	2.9%	1.6%	4.0%	0.9%
Confirm	0.3%	0.0%	0.2%	2.3%
Request	25.7%	23.2%	37.1%	37.7%

User testing

- Subjects recruited via Amazon M-Turk
- They talked to the system and provided feedback
  - Policy performance:

	#N	Success	Ave User Turns
NAC-F	252	94.4 [90.9, 96.9]	7.4 (0.2)
GP-C	249	**95.2 [91.7, 97.5]**	**6.7 (0.2)**
GP-B-PART	249	91.6 [87.4, 94.7]	7.0 (0.3)
GP-B	265	93.6 [89.9, 96.2]	7.0 (0.2)

  - Qualitative policy examination:

	GP-B	GP-B-PART	GP-C	NAC-F
Inform	78.7%	69.1%	72.1%	69.6%
Select	2.6%	1.8%	8.3%	4.0%
Confirm	1.3%	4.6%	1.2%	2.0%
Request	17.4%	24.5%	18.4%	24.4%

6 CONCLUSION

- Possible to train policy directly on the full belief state
- Discrepancies between real users and simulated users
- Future work includes
  - Improved kernel function design
  - Removing the need for the summary action space

7 ACKNOWLEDGEMENTS

This work was supported by PARLANCE (www.parlance-project.eu), an EU Seventh Framework Programme project (grant number 287615).