Autonomous Sub-domain Modeling for Dialogue Policy with Hierarchical Deep Reinforcement Learning

Giovanni Yoko Kristianto[1], Huiwen Zhang[23], Bin Tong[1], Makoto Iwayama[1], Yoshiyuki Kobayashi[1]

[1]Hitachi Central Research Laboratory, Tokyo, Japan

[2]Shenyang Institute of Automation, Shenyang, China

[3]University of Chinese Academy of Sciences, Beijing, China

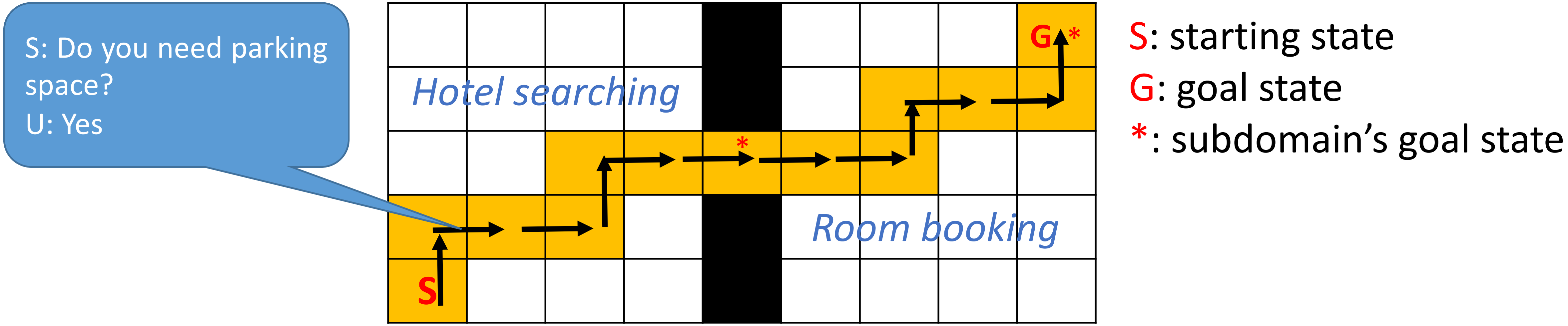
Introduction: Motivation

- Dialogue system for composite task-oriented domains
  - Hotel reservation: hotel searching + room booking
  - Travel planning: flight reservation + hotel reservation + car renting

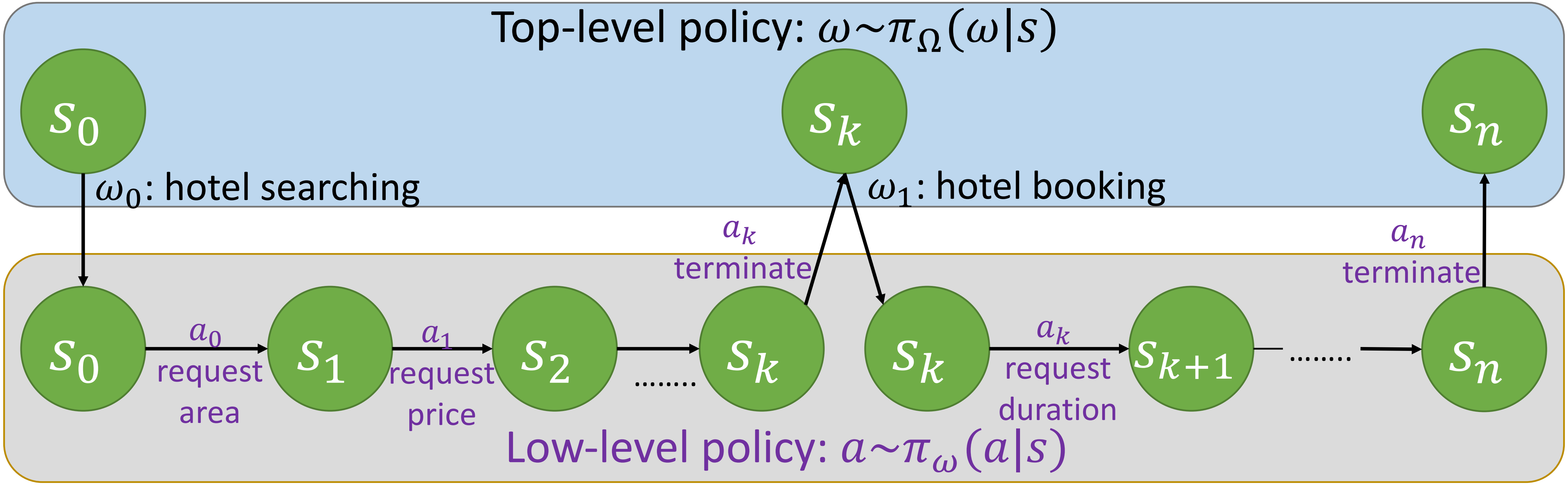
- Challenges in reinforcement learning for composite domain
  - Composite dialogues have larger state-actions space
  - The state trajectory is longer (i.e. need more turns)
  - Sparser rewards

Introduction: Solution

- To decompose the composite domain into multiple sub-domains
  e.g. hotel reservation -> hotel searching + room booking
- To provide <u>internal reward function</u> for each sub-domain, so that the rewards are less sparse



Dialogue Policy Execution using Sub-domains



Previous Works

- Peng et al. 2017. *Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep RL*. In EMNLP ’17
- Budzianowski et al. 2017. *Sub-domain Modelling for Dialogue Management with HRL*. In SIGDIAL ’17

Summary:

- Manually defining composing sub-domains (and the internal reward functions)

Our Approach

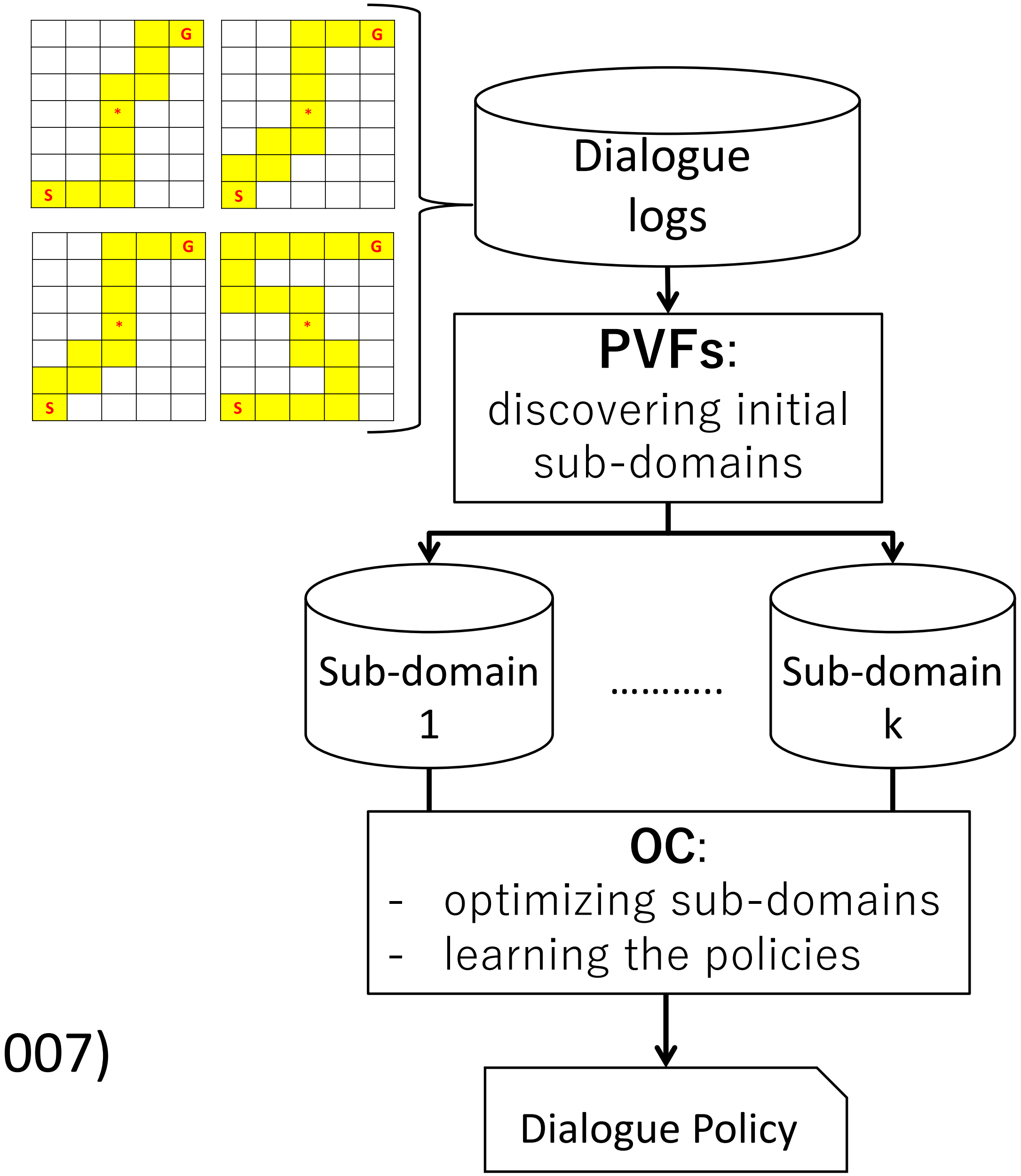
- To autonomously discover sub-domains (and their internal reward functions), and
- To utilize them for training dialogue policy

Input:

- Dialogue logs
- k: number of sub-domains

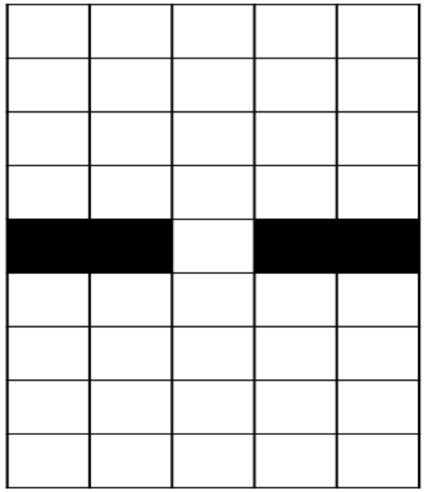
Methods:

- Proto-value functions (PVFs; Mahadevan, 2007)
- Option Critic (OC; Bacon, 2017)

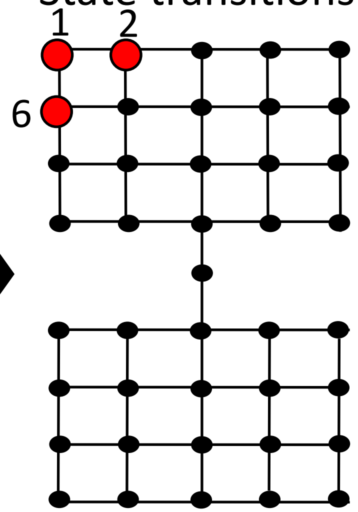


Proto-value function (PVFs) (Mahadevan, 07; Machado, 17)

MDPs



State transitions

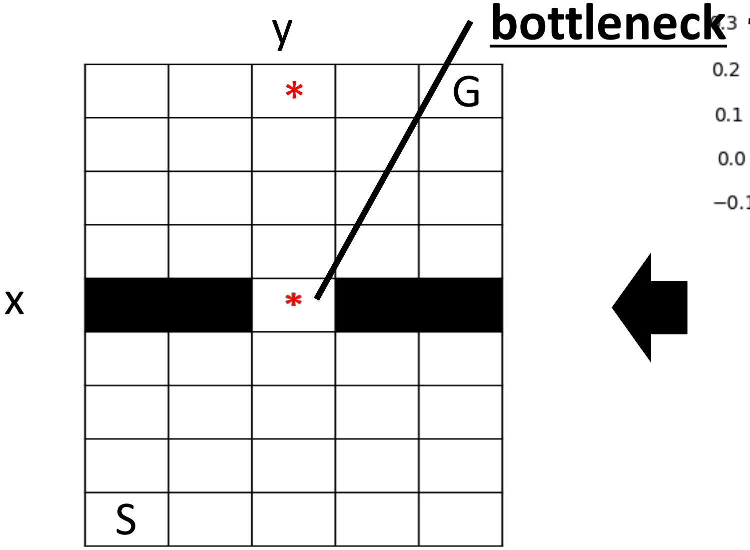


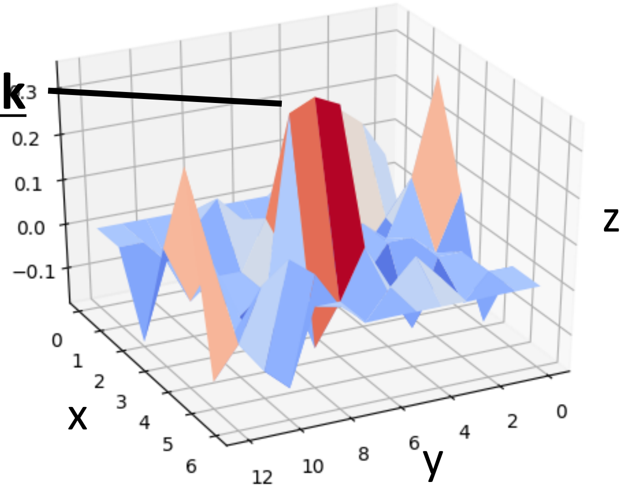
# of states =41

Adjacency matrix (A)

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \\ 39 \\ 40 \\ 41 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & ..\,41 \end{array} \\ \begin{bmatrix} 0 & \color{red}{1} & 0 & 0 & 0 & \color{red}{1} & \\ \color{red}{1} & 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 1 & 0 & 0 & \\ & \vdots & & & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \end{bmatrix} \end{array}$$

1. Get Laplacian Matrix
   $L = D - A$
2. Obtain basis functions from L by eigen-decomposition





a basis function (internal critic)

- PVFs use no external reward
- Discovered sub-goals are task-independent
  - *But, we need to maximize external rewards specific to our task!*

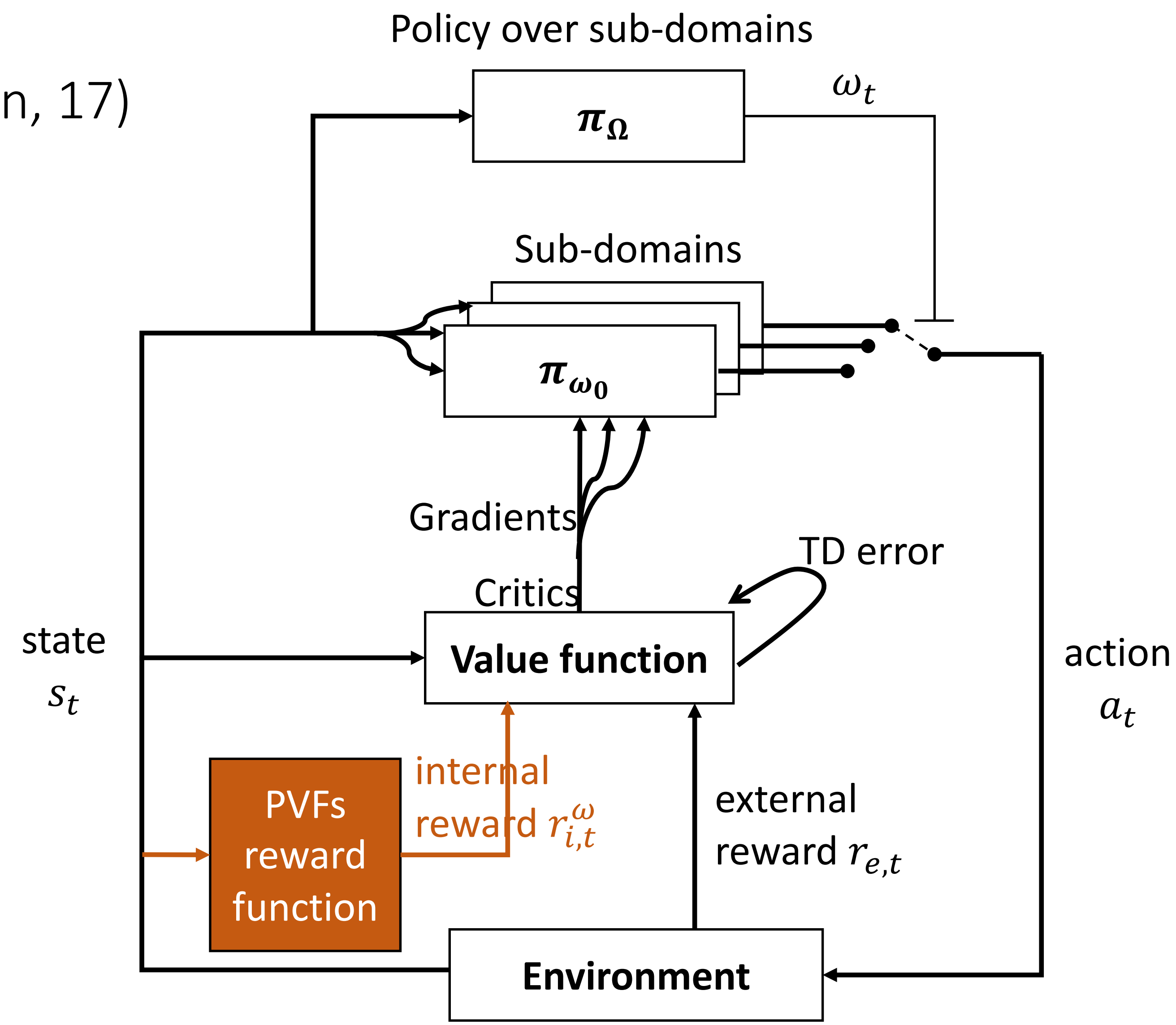
Our Approach: Option-Critic (OC) (Bacon, 17)

- OC learns sub-domains using external reward only

<u>Combining PVFs and OC:</u>

- To update the critics by

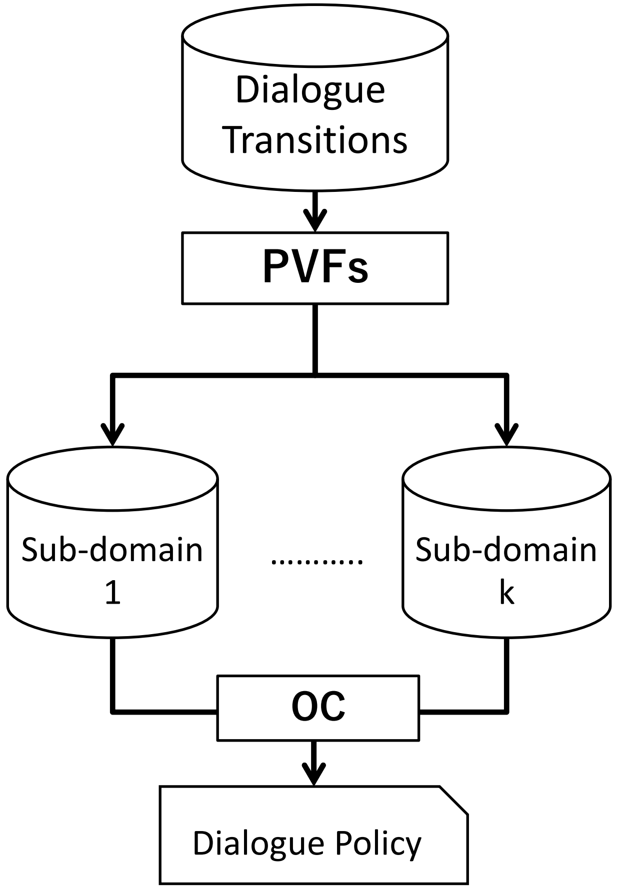
$$\alpha r_{i,t}^{\omega} + (1-\alpha) r_{e,t}$$



Experiment

- Impact on success rate
  - Policy without sub-domain (Reinforce)
  - Policy with manual sub-domain (Gaussian Process RL)
  - Policy with autonomous sub-domains (proposed approach)

- Qualitative analysis of the discovered sub-domains



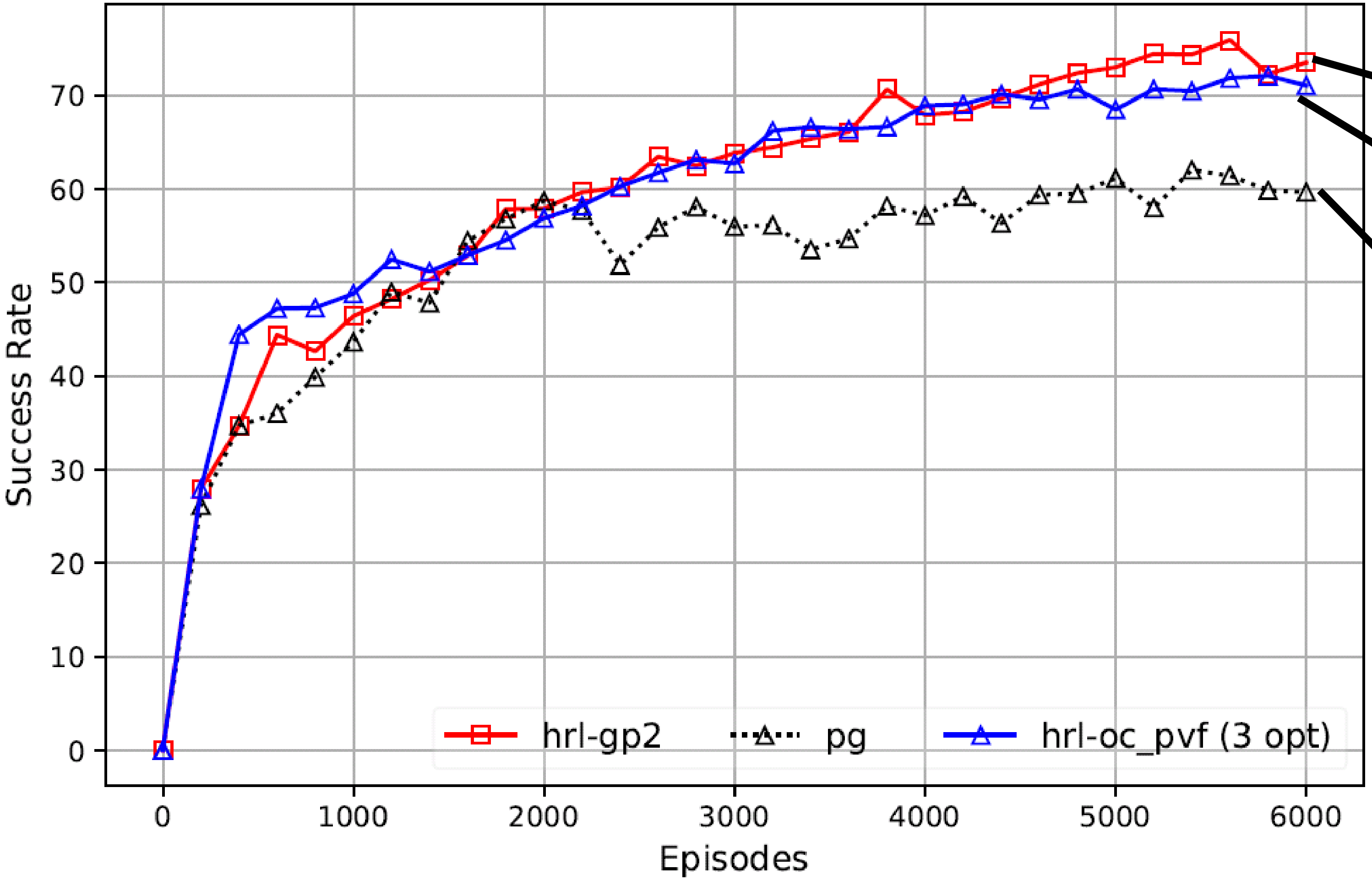
Experiment: Setup

- Dialogue domain: Hotel Reservation
  - # of constraint slots: 13
  - # of system actions: 44

- Training and testing utilized simulator
  - Behavior: searching for a hotel, followed by booking and/or payment

- Implementation
  - Sample transitions used by PVFs: 1,000 (sampled from 100 dialogue samples)

Results: Success Rate



Manual sub-domains **75.9%**

Auto sub-domains (k=3) **72.1%**

No sub-domain **62.0%**

Results: goals of discovered sub-domains are intuitive

…. U: request(address) S: inform(address=xxx) U: hello(day=…, duration=…, name=…, task=booking) …. user ends 'searching', then enters 'booking'	…. U: inform(address=…, cardNo=…, name=…, task=payment) S: inform(paid=true) U: bye() user ends 'payment'
…. U: hello(day=…, duration=…, name=…, task=booking) S: inform() U: request_alternative() …. user ask alternative information	…. U: inform(peopleNo=…, task=booking) S: inform(booked=true) U: bye() user ends 'booking'

Conclusion

- We proposed a dialogue policy learning framework consisting of proto-value functions and option-critic
  - It performs comparable with using manually modeled sub-domains
  - The discovered sub-domains are intuitive

- Future works
  - Human evaluation
  - Transfer learning by exploiting sub-domains