

# Reduced Search Spaces for Efficient Reinforcement Learning of Dialogue Strategies

Heriberto Cuayahuitl<sup>1</sup>, Steve Renals<sup>1</sup>, Oliver Lemon<sup>2</sup> and Hiroshi Shimodaira<sup>1</sup>  
 CSTR<sup>1</sup>, HCRC<sup>2</sup>, School of Informatics, University of Edinburgh

{h.cuayahuitl,s.renals,olemon,h.shimodaira}@ed.ac.uk

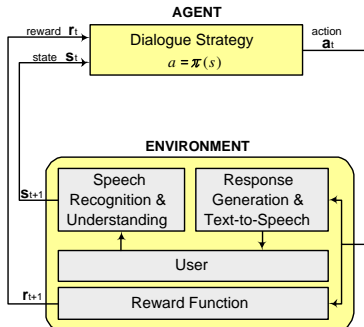


Introduction

## Contribution:

A method to generate reduced search spaces for learning dialogue strategies using reinforcement learning.

## Reinforcement Learning for Spoken Dialogue Systems



**Agent**  
Takes actions for every situation in the conversation by following a dialogue strategy.

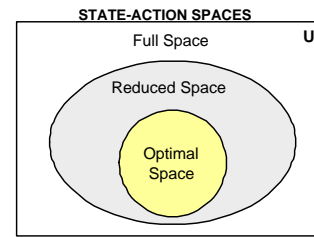
**Environment**  
Usually described as a Markov Decision Process with states  $S$ , actions  $A$ , transition function  $T$  and reward function  $R$ .

The task of a dialogue strategy  $\pi$  is to choose actions in order to maximize the total reward received in the conversation.

## Learning Dialogue Strategies

**Problem:** Expensive computational cost due to the fact that search spaces grow exponentially! (assuming no constraints)

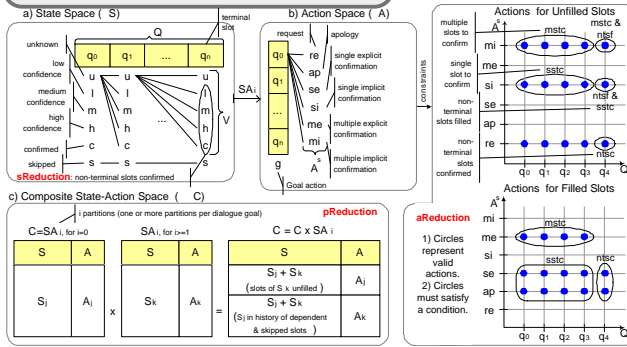
**Example:** A dialogue system with 7 slots ( $Q$ ), 5 state variables ( $V$ ), and 6 single actions ( $A^*$ ) has  $|S \times A| \approx 3.3$  million state-actions, with  $|S| = |V|^{|Q|}$  states and  $|A| = |Q| * |A^*|$  actions per state.



**Idea**  
To avoid unnecessary learning by using prior knowledge that reduces the search space to only valid state-actions.

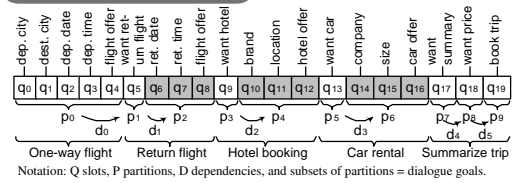
Search Space Reduction

## Proposed Method: sapReduction



- sReduction:** Avoid invalid states (slot-state variable combinations).
- aReduction:** Avoid invalid actions (slot-single action combinations).
- pReduction:** Merge multiple search spaces using dependent partitions.

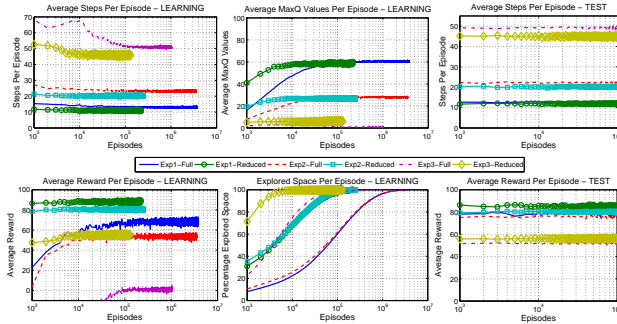
## Experimental Setup



**Experiments:** Exp1 ( $p_0$ ), Exp2 ( $p_0 - p_2$ ), Exp3 ( $p_0 - p_9$ ).  
**Search Spaces:** Reduced used *sapReduction*, full used step 3.  
**Transition Function:** Deterministic, given by the simulated user responses and speech recognition confidence levels.  
**Reward Function:** +100 if all slots were confirmed or skipped, -20 if nothing to confirm/apologize, and -1 otherwise.  
**Learning Setup:** Q-Learning,  $\epsilon$ -greedy (20% exploration), initial Q-values = 0, and convergence = explored space  $\geq 99.9\%$  & avg. MaxQ values of last against previous  $10^4$  episodes  $\leq 0$ .

## Results

Average results of reduced versus full spaces: 94% space reduction, 93% faster convergence, 8.4% less time-steps, and 7% higher reward.



Performance results (X-axis in log scale) in the learning and test phases, data points are averages of groups of 10000 episodes.

## Sizes of Search Spaces

Experiment	Full (S)	Reduced (S)	Full (SA)	Reduced (SA)	% of SA Reduction
Exp1	3126	630	93722	5636	93.98
Exp2	3255	663	95979	5760	93.99
Exp3	4125	957	110097	6672	93.93

## Sample Dialogue

TURN	STATE	ACTIONS* (reduced set)	UTTERANCE
SYS	q0.uq1.uq2.uq3.uq4.u	q0.re.q1.re.q2.re.q3.re	How may i help you?
USR			I want to fly from Edinburgh to London on October third in the morning.
SYS	q0.hq1.hq2.hq3.hq4.u	q0.se.q2.me.q1.se.q2.se.q1.me.q1.mq3.m.e.q3.se.q0.ap.q2.ap.q3.ap.q1.me.q1.ap	I'm sorry, I didn't catch the date. Could you repeat it please?
USR			October third.
SYS	q0.hq1.hq2.hq3.hq4.u	q0.se.q2.me.q1.se.q2.se.q1.me.q1.mq3.m.e.q3.se.q0.ap.q2.ap.q3.ap.q1.me.q1.ap	Okay, flights from Edinburgh to London departing between 5:00 a.m. and noon on Wednesday October third. One moment please. I have flights on British Airways, KLM, and Easy Jet. The earliest British Airways flight leaves at 6:20 a.m.
USR			No i want to fly out later than seven a.m.
SYS	q0.cq1.cq2.cq3.cq4.u	q3.se.q4.se.q3.ap	Okay, flights between 7:00 a.m. and noon. One moment please. What about a KLM flight leaving at 7:10 a.m. I also have a British Airways flight leaving at 7:40 a.m. Would that work?
USR			I will take the British Airways leaving at seven forty.
SYS	q0.cq1.cq2.cq3.cq4.u	q4.ap.q4.se	Okay, British Airways flight 1631 departing Edinburgh at 7:40 a.m. and arriving in London at 9:10 a.m. Do you want to add this to your itinerary?
USR			Yes.
SYS	q0.cq1.cq2.cq3.cq4.c	g	...

\*The learnt dialogue strategy takes the actions in bold font.

## Conclusions and Future Work

- The proposed method is generic to optimize confirmation, can be extended, and does not require significant dev. effort.
- Benefits of reduced spaces: 1) less computer memory, 2) faster convergence, and 3) potential better performance.
- Future work: More complex and larger systems using hierarchical learning and dialogue simulators learnt from data.

Results and Conclusions