



JoTR: A Joint Transformer and Reinforcement Learning Framework for Dialogue Policy Learning

Wai-Chung Kwan^{1*}, Huimin Wang^{2*}, Hongru Wang¹, Zezhong Wang¹
Bin Liang¹, Xian Wu², Yefeng Zheng², Kam-Fai Wong¹



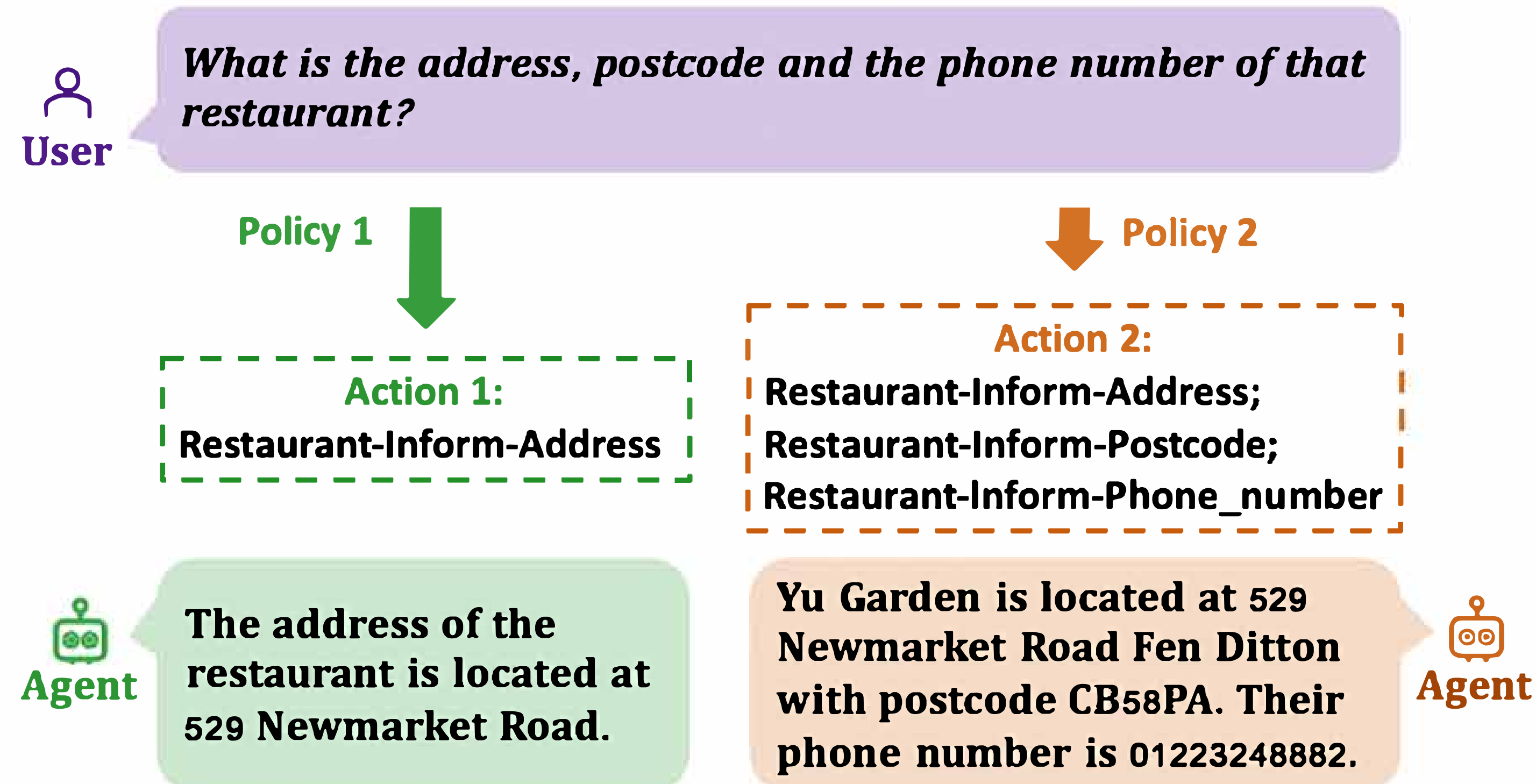
腾讯天衍实验室
TENCENT JARVIS

¹The Chinese University of Hong Kong ²Jarvis Lab, Tencent

Introduction

Dialog policy learning (DPL) plays a crucial role in pipeline task-oriented dialog systems by determining the next abstracted system action.

Existing methods employ classification approaches that rely on a predefined action list, preventing flexible action generation.



Method

We propose JoTR, a transformer-based reinforcement learning framework that learns a token-grained generative policy.

Training

We pre-train the model and then fine-tune it using PPO through interactions with a user simulator.

We incorporate **reward shaping**, which provides additional rewards when the system informs or requests desired slots.

We evaluate our approach on two popular task-oriented dialogue datasets: MultiWOZ and SGD.

Main Results



| Model | MultiWOZ | | | | SGD | | | |
|----------------------------|-------------|-------------|--------------|------------|-------------|--------------|--------------|------------|
| | Succ.↑ | Turn↓ | Rew.↑ | #Acts↑ | Succ.↑ | Turn↓ | Rew.↑ | #Acts↑ |
| JOIE | 0.91 | 18.90 | 40.82 | 147 | 0.51 | 11.10 | 15.32 | 210 |
| MLP _{ppo} | 0.56 | 30.72 | -26.76 | 162 | 0.54 | 23.43 | 16.50 | 233 |
| SimpleTOD‡ | 0.62 | - | - | 186 | 0.50 | - | - | 361 |
| DASP‡ | 0.85 | - | - | - | 0.70 | - | - | - |
| ChatGPT | 0.73 | 13.10 | 41.05 | 165 | 0.50 | 11.04 | 15.48 | 242 |
| JoTR | 0.93 | 9.94 | 68.46 | 249 | 0.79 | 15.23 | 49.25 | 494 |
| JoTR _{w/o rs} | 0.89 | 9.95 | 66.42 | 207 | 0.72 | 16.53 | 38.84 | 429 |
| JoTR _{w/o ppo} | 0.67 | 18.44 | 32.18 | 189 | 0.55 | 24.76 | 14.62 | 357 |
| JoTR _{pretrained} | 0.76 | 14.19 | 44.87 | 195 | 0.64 | 19.25 | 28.18 | 372 |

It uses **fewer turns** (9.94 vs 18.9) and generates **more diverse actions** (249 vs 147) to fulfill the user's goal.

Reward shaping improves success rate from 0.89 to 0.93 in MultiWOZ and from 0.72 to 0.79 in SGD.

Without PPO (JoTR w/o ppo), the success rate drops 28% and the number of unique actions decreases 27%.

Case Study

I would like to find a park in the east of town .

Yes certainly. We have one. Cherry hinton water play is a good choice.

Sounds great. Can you please provide me with the address, postcode and phone number?

Their phone number is 012200, they are located at Cherry Hinton Hall, Cherry Hinton Road, and their postcode is CB18DW.

I am also looking for a guesthouse with free parking, preferably in the center.

Alexander Bed and Breakfast is a good place. You can contact them at 012725.

Can you please make a booking for 8 people and 5 nights starting from Saturday

Booking was successful. The reference number is 00000002.

Please also get me a taxi at the hotel arriving by 18:15

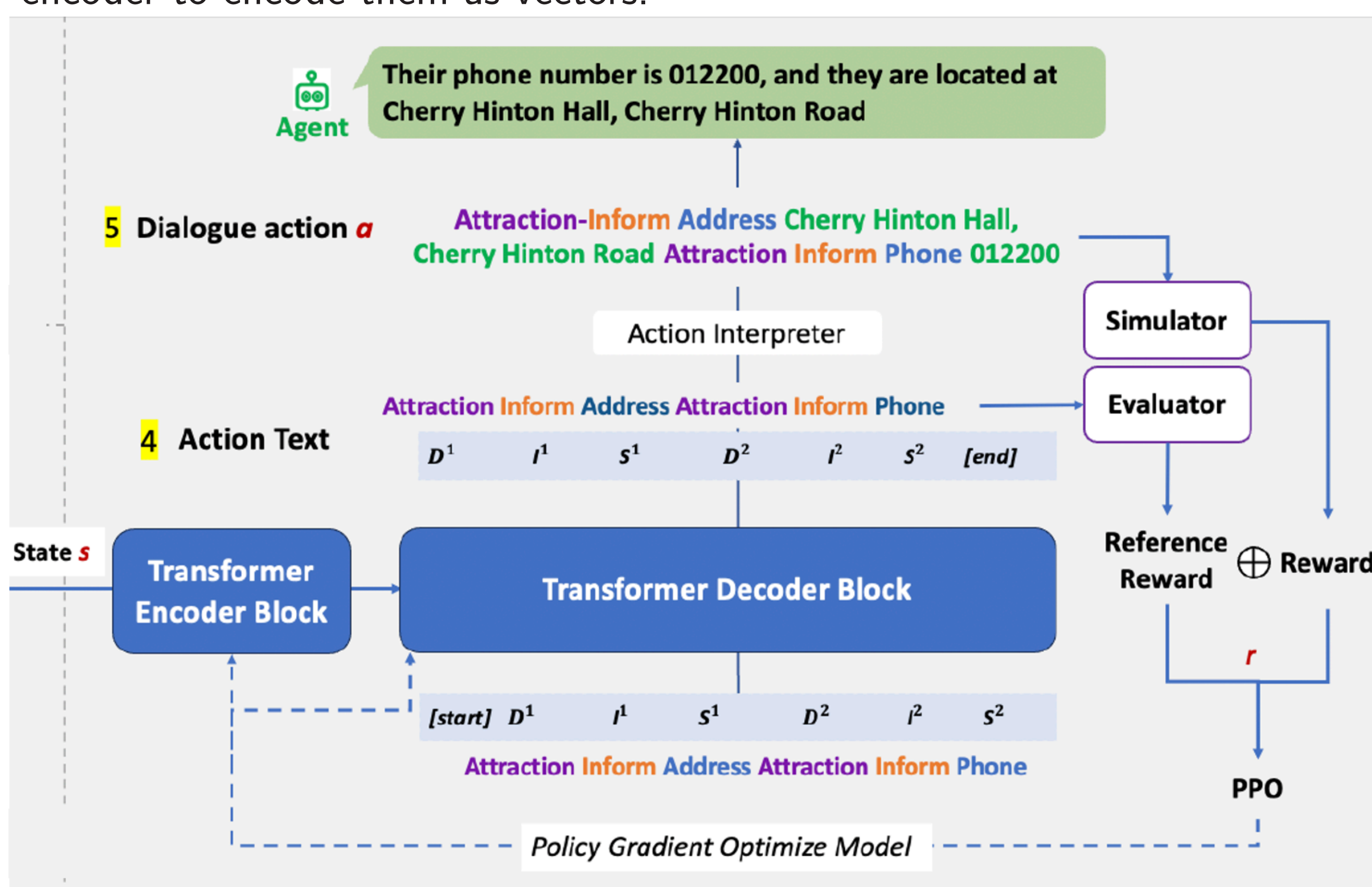
I have successfully booked you a skoda. The contact number is 87593830634.

Finished Successfully

JoTR can generate appropriate and efficient actions, as highlighted in yellow.

It provides useful additional information without explicitly being asked, as shown in pink.

We first flatten the dialogue state into text and then use a transformer encoder to encode them as vectors.



We use a transformer decoder to iteratively generate the domain, intent, and slot values, one word at a time, until it reaches a completion token based on the encoded state.