References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz: A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *2007 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–152.

A User Goal

In the task-oriented dialog setting, the entire conversation is around a user goal $G = (C, R)$ implicitly, where C denotes the constraint and R is the requests (Schatzmann et al., 2007). The user goals appeared in the original corpus are all extracted into one database. Note that each user goal in MultiWOZ (Budzianowski et al., 2018) may consist of a "real" user goal that describes what the user wants in the end, along with a different "failed" user goal that indicates what the user wants at first. Every time a dialog is launched, the user goal is initialized by the user simulator at the beginning of a dialog session, by randomly sampling the *constraint slots* and *requests slots* from the user goal database. Each slot is sampled according to its frequency in the dataset, and a "failed" user goal would be also initialized with a certain probability to simulate the changes in a user goal.

B Case Study

Table 1 shows two dialog sessions with the same user goal in PPO and GDPL, respectively. The user goal consists of *attraction*, *hotel* and *taxi* domains. The system is required to answer all the information in *Reqt* based on the constraints in *Info* and book a right entity that meets the needs in *Book*. When the user asks for some information about the museum, GDPL manages to provide all the requested information, while PPO only answers part of the information, which causes the user to inquire about entrance fee again (red line). Additionally, GDPL answers the area of the hotel (blue line) and the dialog proceeds successfully, however, PPO fails.

C Reward Visualization

Fig. 1 shows the reward visualization of the session discussed in the *reward evaluation* subsection with its original dialog acts.

Attraction		Hotel			Taxi	
Info	Reqt	Info	Reqt	Book	Info	Reqt
type=museum area=centre	entrance fee address postcode	price range=expensive stars=4 type=hotel parking=yes	address area	people=8 day=saturday stay=5	arrive by=19:45	car type phone

PPO	GDPL
U: Hi, I would like to visit a museum in the centre area. S: I found 1 attraction matching your requirements. U: May I have its address, postcode, and entrance fee? S: The address of the *regency gallery* is *the grafton centre, east road*, and postcode is *cb11ps*. U: What about the entrance fee? S: Admission is free. U: I also need to book an expensive 4-star hotel that has parking. S: There are 2 hotels matching your requirements. Its name is *huntingdon marriott hotel*. U: Could you tell me the address and area of one? S: The address of *huntingdon marriott hotel* is *kingfisher way, hinchinbrook business park, huntingdon*. U: What is the area of the hotel? S: The name is *huntingdon marriott hotel*. U: What is the area of the hotel? S: The name is *huntingdon marriott hotel*. U: What is the area of the hotel? S: The name is *huntingdon marriott hotel*. ⋮ U: What is the area of the hotel? S: The name is *huntingdon marriott hotel*.	U: Hi, I would like to visit a museum in the centre area. S: The *regency gallery* matches your requirements. U: May I have its address, postcode, and entrance fee? S: Its address is *the grafton centre, east road*, postcode is *cb11ps*, and admission is free. U: I also need to book an expensive 4-star hotel that has parking. S: There are 2 hotels matching your requirements. U: Could you tell me the address and area of one? S: The address of *huntingdon marriott hotel* is *kingfisher way, hinchinbrook business park, huntingdon*. U: What is the area of the hotel? S: It's located on the centre of city. U: Please book it for 8 people and 5 nights on saturday. S: Your reference number is 5NAWGJDC. U: Could you also book me a taxi that arrives by 19:45? S: Booking is successful. It will be a yellow honda and the phone number is 3736171789. U: Thank you. That will be all. S: You are welcome. Goodbye.
Failure: Agent does not tell the area of the hotel properly.	**Success**

Table 1: A sample dialog session comparison between PPO and GDPL.

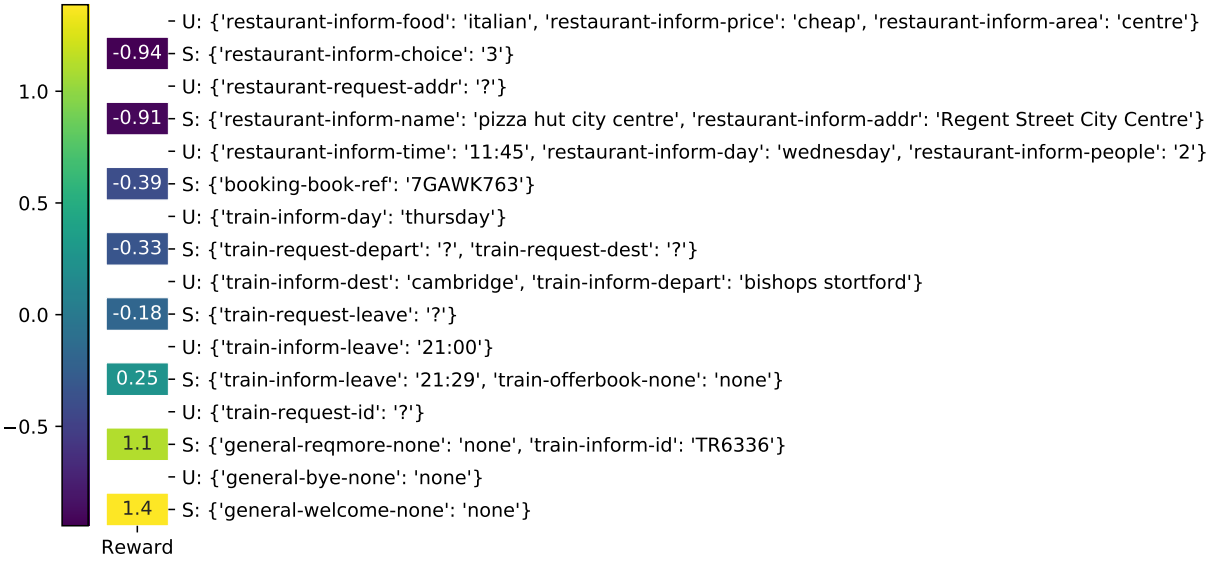

Figure 1: Reward visualization with dialog acts.