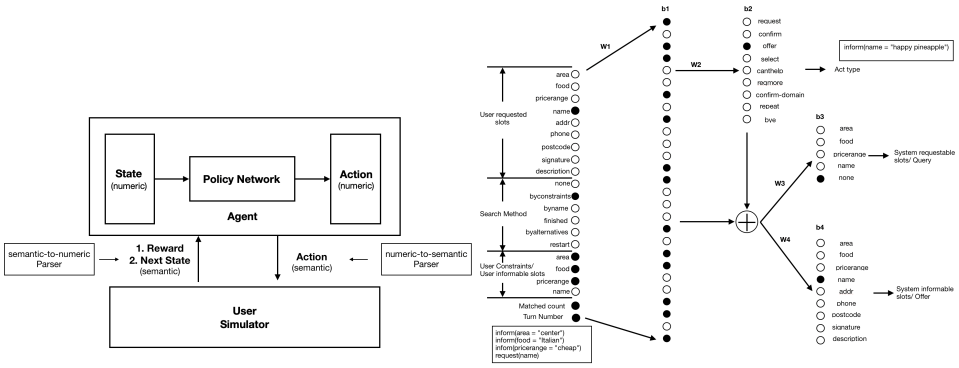
Learning What and How to Talk with Reinforcement Learning

A spoken dialogue system (SDS) allows users to talk directly to the system and, in turn, receive an acoustic response. The study of the SDS can bring practical impact, as it can assist users to fulfill tasks in a hands-off and eyes-free manner, for example an in-car speech assistant which can help the driver to find a nearest parking lot. The agent's strategy in conversing with a human user is modelled with the dialogue policy (DP) component in the dialogue management (DM) module.

The scope of this study is to investigate the DM module in a task-oriented SDS with a reinforcement learning (RL) algorithm, see Figure 1(a). More specifically, the policy gradient method is adopted in modelling and optimizing dialogue policy in a simulated environment. Typically, the term dialogue act is used to represent utterances. A dialogue act contains an intent (e.g., *request, inform, confirm ...*) and one or more slot-value pairs (e.g., *food=Italian*, *pricerange=cheap*)

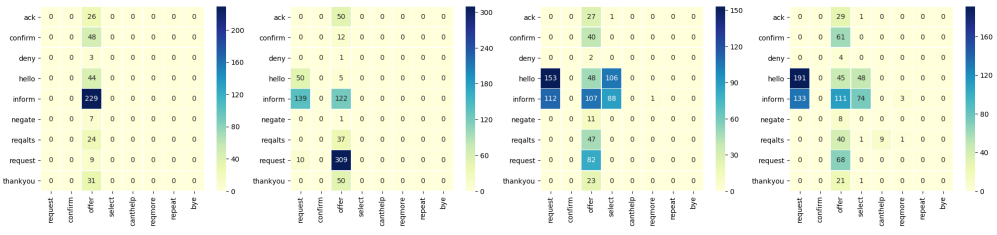


(a) Reinforcement Learning for Dialogue Management

(b) Policy Network used in this study

Figure 1: Pipeline

Most RL studies on dialogue policy modelling required a pre-defined state and action space , which limits what the agent can experience and what actions the agent can take. However, this way of learning is not compatible with human interaction. Humans can experience different states and take any action. In fact, what such RL agents learn is a simple mapping from a pre-defined set of states and a pre-defined set of action, rather than a real strategy. A novel feature representation is proposed in this study, see Figure 1(b). The agent can effectively learn a domain-independent DP and the learning process can be visualized in a direct and intuitive way (Figure 2). Along the course of training, the agent learns to select different types of intent through continuous exploration.



(a) 1^{st} (b) 40^{th} (c) 117^{th} (d) 229^{th}

Figure 2: Learning process at the 1, 40, 117 and 229 validation epoch), with rows indicate user intent and columns indicate agent intent as a response