R13/Ishank Juneja/16D070012

March 14, 2020

Spoken dialogue management (DM) systems consist of a combination of an automatic speech recognition (ASR) component and a sequential decision making module responsible for picking the most appropriate response for every query. Thus given an ASR system, the DM problem can be considered to be a sequential decision making task with the aim being to comply with the user's requests in as few dialogue exchanges as possible.

Even before the present paper by Roy and others, Markov Decision Processes (MDPs) had been effectively employed as the model to learn a good action selection policy through planning. Under the MDP model, the set of states represent the preceding dialogue as a whole, while the actions correspond to responses produced by the robotic system. However, under this model the correct way of encoding dialogue into state is ambiguous. Further, the reliability of such MDP based systems is poor in the face of noisy environments. These shortcomings of MDP based models are exacerbated in situations where speech signals reaching the agents microphone are not clear. The authors point out that the problem of low quality speech is quite prevalent when users interact with anthropomorphised mobile robots. To overcome these shortcomings, the authors propose moving to the partially observable MDP (POMDP) model. Under the POMDP model described in the paper, the underlying MDP state (which is not directly observable) is the user's *intention* for the dialogue task. The DM agent's goal under the POMDP model is to find an optimal dialogue strategy given certain observations and dialogue history.

The underlying MDP for the POMDP is given by $(\mathcal{S}, \mathcal{A}, T, R)$. Where, the set of states $\mathcal{S}$ is the state of the user i.e. the user's intent, the set of actions $\mathcal{A}$ is the set of prepared responses the system can deliver, the transition probabilities T connect the underlying states in a desired directed graph and the reward function R is hand crafted to encourage desirable behaviour. To this description, the POMDP model adds the following -

- Set of possible observations $\mathcal{O}$, and a distribution over them - $O(o, s, a) = P(o|s, a)$. The set $\mathcal{O}$, consists of keywords extracted from user utterances.
- An initial belief state, $P(s : s \in \mathcal{S})$
- A modified reward function conditioned on observations, i.e R becomes the map $R : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \mathcal{R}$

First the POMDP model created undergoes a planning phase to determine an optimal or near-optimal policy. Most often, exact planning is computationally intractable under POMDP models and the only recourse is to look for approximately optimal policies. To do this, the authors create an augmented MDP that represents the POMDP well. This simplification is achieved under the assumption that uncertainty in state (i.e. user intent) is *domain specific and localised*. Under these somewhat stringent assumptions, the state of the system is completely captured by the tuple $(\arg\max_s p(s), H(p(s)))$. Where $H(p(s))$ is the entropy of the belief state distribution $p(s)$.

Experiments and Results

The authors conduct experiments with the *Flo* (Short for Florence Nightingale) robot with the aim of performing tasks typical in an assisted living environment. Tasks that users can inquire about include their medication schedule and TV program schedules for a few stations. The model that allows *flo* to solve its tasks consists of 13 states and 20 actions. Among the 20 programmed actions, 10 actions are for the primary abilities of *flo* and the remaining 10 are clarification questions. The set of observations corresponds to 16 possible key words and a dummy observation for unintelligible utterances. The reward scheme for *flo* is such that there is a large +ve reward (+100) for providing the correct response, a small -ve reward (-1) for asking a clarifying question and a large -ve reward (-100) for asking the user to repeat themself.
On comparing the reward accumulated by the planning methods of - (1) exact POMDP solution, (2) approximate/augmented POMDP solution and the (3) MDP based solution, the authors found that both POMDP models provide much better returns than the MDP based model. The difference between the exact POMDP solution and the approximate POMDP solution was quite small with the exact solution performing slightly better. However, the approximate solution more than compensated for its demerits through a planning time nearly three orders of magnitude lower than that of the exact solution.

The authors attribute the success of the POMDP methods to the fact that it can compensate for the variable reliability of the ASR stage. This becomes possible under the POMDP model since as recognition degrades, the model can actively gather information from the user.

A question I have about the paper is -

- It is not immediately obvious to me how does the entropy $H(p(s))$ becomes a sufficient statistic for an entire belief state.