R13/Sanit Gupta/160100010

August 6, 2020

In this paper, the authors develop dialogue systems. They point out that a big flaw of the existing systems is that they don't take into account that the state representation is not completely reliable. This is because the environment can be noisy and it might not even be possible to infer the dialogue state perfectly. Also, the system developed should work independent of who the speaker is. Therefore, they believe that modeling the problem as a POMDP is a good idea because considering it to be an MDP would lead the agent to have unwarranted confidence in its decisions.

As this is a POMDP, a belief state i.e. a probability distribution over potential true states is maintained. Although finding the optimal policy for a POMDP is intractable, a near optimal policy can be computed much more easily.

They make the assumption that the uncertainty is localised and hence can be represented simply by the most likely true state and the entropy of the belief state which is approximately a sufficient statistic for the entire belief state. Therefore, instead of learning a policy which maps each belief states to action, they learn one which maps this approximate representation to actions.

The example they consider is that of a mobile robot meant to serve as a nursing home assistant. They handcraft a model of the environment with 13 states and 20 actions (10 for different abilities, and 10 others to clarify). Reward is given when a user request is fulfilled and most other actions get penalized (the clarification actions aren't penalized much).

The Incremental Improvement algorithm was used to generate the POMDP policy (exact). It was unsuccessful in computing one for the full problem, so it was used for a smaller dialogue model. For this smaller problem, the exact POMDP solution was unsurprisingly the best one although it did take very long to compute. For the full problem, the approximate POMDP solution beat the MDP solution by a significant margin. Also, although the agents were not deployed for enough users to say anything with certainty, the preliminary results did agree with what was expected: the POMDP approach did better for the person for who the word accuracy was lower and both approaches performed similarly well for a user with high recognition rate.

A few questions I had:

- Why are the rewards conditioned on the observations?

- If extensive experiments with humans were carried out later, what were the results? Else, why were they not performed?

.