Paper: *A Deep Reinforcement Learning Chatbot*

Nirajan Koirala

CSC 9010_001

Summary

Authors present a deep reinforcement learning chat-bot which consists of an ensemble of natural language generation and retrieval modals, including neural network and template-based models. The data they use is crowd-sourced and the system is trained to select an appropriate response from the models in its ensemble. Evaluation is done using A/B testing and it is able to perform significantly better than other systems. In section 2, overview of their system is provided. It consists of an ensemble of response model and each model takes input a dialogue history and outputs a response in natural language text. For the generation of a response, first all the models generate a set of candidate responses and if there exists a priority response in the set of candidate responses, this response is returned by the system. If no priority response is found then the response is selected by the model selection policy. They use 22 response models in the system and provide the model selection policy details in section 4. In order to make the trade-off between immediate and long-term user satisfaction they consider selecting the appropriate response as a sequential decision making problem. The reinforcement learning frame work of Sutton and Barto is used where the main task of the agent is maximizing the discounted sum of rewards.

For the parametrization of agent's policy they use two different approaches, action-value parametrization and stochastic policy parametrization which parameterizes a distribution over actions. For learning the parameters, they use 5 different machine learning approaches. The first one they use is called Supervised Learning AMT and it estimates the action-value function using supervised learning on crowsourced labels. The next approach learns a stochastic policy directly from examples of dialogues recorded between the system and real world users. Later two approaches trains a linear regression model to predict the user score from a given dialogue. The last approach is based on learning a policy through a simplified Markov decision process, called the Abstract Discourse MDP.

Section 5 discusses the experiments carried out using A/B testing to evaluate the dialogue manager policies for selecting the response model. Three different experiments are carried out using the different methods. For the first experiment they test various dialogue manager policies using greedy variants for the off-policy reinforce policies. A heuristic baseline policy EvilBot and Alicebot are also used based on their availability. In the second and third A/B testing experiment, they test two policies, Off-policy reinforce and Q-learning AMT. It is observed that Q-learning AMT is able to perform best among all policies w.r.t Alexa user scores in the first and third experiments. After computing several linguistic statistics for the policies in the experiments, the two policies Q-learning AMT and Off-policy reinforce demonstrate substantial improvements over all other policies. The experiments show that ensemble approaches work better than others and it works using different models which output natural language responses and the system policy selects one response among them.

Strengths

- The ensemble method used for natural language generation is performing better than systems using any other techniques.
- Many different approaches are used for model selection and carrying out experiments.

Weaknesses

- Organization of the paper could have been improved.

Points of Confusion

- What does risk tolerance means in context of Q-learning AMT?
- Authors suggest that number of turns per conversation is a good criteria to evaluate the models. I am not sure how would they evaluate the quality of those conversations?

Discussion Questions

- Would BLEU scores work as a evaluation method on the output of these models?
- Would using RNNs in the models help make the conversation between the system and user more context based and natural?
- A discussion on section 5.2 would he helpful.