Jenish Maharjan **Week 11**

Paper: *A Deep Reinforcement Learning Chatbot*

Summary:

Serban et al describe their implementation of a deep reinforcement learning chatbot named MILABOT in this paper. MILABOT was developed for the Amazon Alexa Prize competition and has the capability of making conversation with humans via speech and text. They describe their model as a large-scale ensemble system that uses deep learning and reinforcement learning. The system is trained to select a response from the models in its ensemble using crowdsourced data and real-world user interactions. The proposed model makes use of a off-policy model-based learning procedure showing considerable improvements in A/B testing experiments with real-world users.

In contrast to the rule-based systems, the MILABOT system uses a statistical machine learning approach which allows for very less assumptions to be made about the process of understanding and generating natural language. The system only uses a very number of hand-crafted states and rules. Inspired by the success of ensemble-based machine learning systems that consist of independent sub-models combined intelligently together, the system consists of an ensemble of response models - total of 22 models are used in the system - each of which takes an input dialogue history and outputs a response in natural language text. Each model is designed to generate responses on a diverse set of topics using a variety of strategies. The responses from all the models are then combined by the dialogue manager along with the corresponding confidence values of automatic speech recognition system (ASR confidence). The dialogue manager generates a set of candidate responses. If a priority response is available in the set, the response is returned. In any other case, then the response is selected by the model selection policy.

Selection of the response by the dialogue manager is a sequential decision-making problem which has to make a trade-off between immediate and long-term user satisfaction. In the paper, they describe five approaches they investigated to learn the model selection policy, each of which was evaluated with real-world users. The agent in this scenario is the dialogue manager which takes actions of selecting the responses in order to maximize rewards. The model is parameterized in two ways – action-value parameterization and stochastic policy parameterization. The first approach is the Supervised Learning with crowdsourced labels where the action-value function is estimated using supervised learning on crowdsourced labels. This approach also works as the initialization for all other approaches. Human evaluators are shown a dialogue along with 4 candidate responses and they score each candidate response on a 1-5 Likert-type scale. The model parameters are optimized with respect to log-likelihood using mini-batch SGD to predict the 4^{th} layer which represents the labels. The second approach titled Off-policy REINFORCE used examples of dialogues recorded between the system and the real-world users to directly learn a stochastic policy. This method is analogous to learning from trial and error.

Another approach, Learned Reward Function, trains a linear regression model to predict the user score from a given dialogue so that with a given dialogue history and a candidate response, the model predicts the corresponding user score. The final approach, Q-learning with the Abstract Discourse Markov Decision Process, uses a learning through a simplified Markov decision process, called the Abstract Discourse MDP which is similar to training with a user simulator. The score predicted by the Supervised model is used as the per time-step reward function. Given the MDP, Q-learning with experience replay is used to learn the policy.

A/B testing experiments were carried out to evaluate the dialogue manager policies for selecting the response model. The A/B testing was carried out in 3 different periods. In the first period, all the approaches were evaluated over a heuristic baseline policy of Evibot and Alicebot. In the second period, Off-policy REINFORCE and Q-learning AMT were evaluated after making minor system improvements with respect to Initiatorbot and filtering profanities. In the third period, the experiments were continued during with about three hundred user ratings were collected after discarding returning users. From these experiments, it was observed that Q-learning AMT performed best among all polcies with respect to Alexa user scores in the first and third experiment periods. Both Q-learning AMT and Off-policy REINFORCE demonstrated substantial improvements over all other policies.

Strengths:

- Using 22 models for generating responses is almost an exhaustive approach of making use of and benefiting from the various strategies that have been introduced in Natural Language generation.
- The A/B testing experiments performed in a span of 3 testing periods adds to the confidence on each of the policies that they tried out.
- The description of each policy is well summarized.

Weaknesses:

- The structure and organization of the content could have been improved.
- A flow chart or a diagram showing the full pipeline would have been really helpful.

Confusions:

- What is the purpose of parameterization?
- How were the models eliminated after the first A/B testing experiment period?

Discussion Questions:

- What kind of policies govern collecting data from users like the data collected from Alexa users?
- Wouldn’t priority responses hinder in a more natural conversation pattern?
- What is Likert-type scale?