0.1 Deep Reinforcement Learning with Hierarchical Recurrent Encoder-Decoder

In this research, we propose a dialogue generation model using deep reinforcement learning (DRL) and HRED. Li et al. succefully integrated the LSTM sequence-to-sequence model (Sutskever et al.) and policy gradient opitmization methods for dialogue generation (DRL-SEQ2SEQ). However, one of major limitations of this model is that a policy learned by RL choosing an action consider only the previous two utterances rather than the entire past conversation. Therefore, it may be difficult to continue the conversation while maintaining its relevant topics through multiple utterances. Our model is mainly different from the DRL-SEQ2SEQ model in three aspects. First, we pre-train a model using HRED p_{HRED} as mentioned in the previous section and use this for initilizing the RL policy. Second, instead of using only two previous dialogue utterances to define a state, we define a state at time t as a hidden state of the high-level context RNN at time t, $s_t = c_t = g(c_{t-1}, u_t)$. We use the same definition for actions - generating a dialogue utterance. Thus, the RL policy, π, is defined as $p_{RL}(u_{t+1}|c_t)$. Finally, our reward functions are similar to ones in the DRL-SEQ2SEQ model, but we replace the SEQ2SEQ model with the pre-trained HRED model for the probability of generating a response given a state information.