Transformer Decoder Based Reinforcement Learning Approach for Conversational Response Generation

Farshid Faal, Jia Yuan Yu, Ketra Schmitt
Concordia Institute for Information System Engineering
Concordia University
Montreal, Canada
f_faal@encs.concordia.ca, jiayuan.yu@concordia.ca, ketra.schmitt@concordia.ca

***Abstract*—Developing a machine that can hold an engaging conversation with a human is one of the main challenges in designing a dialogue system in the field of natural language processing. Responses generated by neural conversational models with log-likelihood training methods tend to lack informativeness and diversity. We address the limitation of log-likelihood training in dialogue generation models, and we present the Reinforce Transformer decoder model, our new approach for training the Transformer decoder based conversational model, which incorporates proximal policy optimization techniques from reinforcement learning with the Transformer decoder architecture. We specifically examine the use of our proposed model for multi-turn dialogue response generation in a real word human to a human dataset. To verify the effectiveness of our proposed framework, we evaluate our model on the Reddit dialogues data, which is a real word human to a human dataset. Experiments show that our proposed response generating model in a dialogue achieves significant improvement over recurrent sequence-to-sequence models and also the state of the art Transformer based dialogue generation models based on diversity and relevance evaluation metrics.**

***Index Terms*—Reinforcement Learning, Transformers , Deep Neural Network, Open-Domain Dialogue Generation Systems, Proximal Policy Optimization**

I. Introduction

Developing an intelligent dialogue system, a system with the ability to understand natural language (hold the conversation with humans) has been one of the longest-running goals in the field of Natural Language Processing (NLP) and Artificial Intelligence (AI) in some sense dating back to the ELIZA project at MIT in the 1960s [1], [2]. By advancement in machine learning techniques and large amounts of conversational data becomes available for training, the field of natural language understanding and natural language generation got the attention of lots of researchers in academics and different industries.

Dialogue systems are categorized into two main types of systems based on their functionality: task-oriented [3]–[5] and open-domain [6]–[8]. The goal in task-oriented dialogue systems is to assist the user in performing a specific task (such as booking a restaurant). Open-domain dialogue systems are not limited to a specific task or specific domain, and the dialogue performs in the open domain. The goal of open-domain dialogue systems is to generate a meaningful response given a dialogue context [8], [9].

Among different approaches for developing open-domain dialogue systems, the Sequence-to-Sequence (Seq2Seq) framework with Recurrent Neural Network (RNN), and Maximum Likelihood Estimation (MLE) objectives deliver promising results in designing the dialogue systems [10]–[12]. Despite the success of these methods in modeling the dialogue systems, still, some limitations are indicated in previous works. The first limitation that points in several studies is that the Seq2Seq-RNN models fail to capture long-term temporal dependencies across conversation turn. The gradient vanishing problem limits the ability of Seq2Seq-RNN models to capture long-term temporal dependencies across conversation turns. The second limitation is exposure bias in these models. The most popular method used method to train the standard Seq2Seq-RNN models is the teacher-forcing algorithm [13]. During training in this algorithm, the decoder uses two inputs to generate the next word, the previous output state from RNN and ground-truth word. However, at the test time, the decoder only uses its own generated word at a previous time step to predict a new word since the ground-truth data is not available anymore. This discrepancy is referred to as exposure bias and limits the informativeness of the generated responses since the decoding error compounds rapidly during inference [13]. The third limitation observed in these models is a training objective for these models. Most of the existing dialogue models learn the conditional distribution of the response given the context from the MLE objective [14], [15]. The human dialogue data is usually redundant, and training a Seq2Seq-RNN model on these datasets with MLE objective, provide a simple mapping between the context and response, which yields generic and dull responses.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

All these works are categorized as a supervised learning approach. One of the limitations of this approach is lack of relatedness between training data and online scenarios. This limitation makes it difficult to optimize the dialogue systems toward its goals, generating diverse and informative responses and reducing blandness. Furthermore, in supervised methods, the objective is to optimize for an immediate reward rather than a long-term reward, which makes the dialogue system having a bland response and fails to promote long-term engagement with the user.

In this paper, we propose the Reinforced Transformer decoder (R-TD) model, the combination of the Transformer decoder architecture and Proximal Policy Optimization (PPO) method from Reinforcement Learning (RL) algorithm for multiturn dialogue modeling that addresses the limitations in modeling the dialogue systems. The R-TD is formulated as an autoregressive language model and uses multi-layer Transformer decoder as a model architecture. Recent advances in large-scale Transformer based architectures [16]–[18] have achieved great empirical success in different natural language understanding tasks such as question answering, named entity recognition, sentence classification, and sentence similarity. One of the key point in success of these models is their ability to capture long-term temporal dependencies in the input context. This ability also makes them a great candidate to model multi-turn dialogue systems. Transformer architecture in the R-TD model allows us to capture long-term temporal dependencies in a context of dialogue data better than RNN based models; however, the original transformer models trained based on MLE objective and still sufferers from some of its limitations like short answer generation. To alleviate this limitation, we incorporate reinforcement learning training for transformers to yield longer and more informative answers. In order to stabilize the training of the Transformer decoder in our proposed model, we incorporate PPO techniques that constrain the policy to control it and make the policy to be stable. The results show that sentences generated by our proposed R-TD model are diverse and contain information specific to the source prompt. The effectiveness of our approach is validated empirically on the Reddit social media dataset.

II. Related Works

The main idea behind the earliest conversation models is inspired by statistical and neural machine translation [19]–[21]. One of the first attempt in casting the conversation models as a machine translation problem was [6], which applied a phrase-based translation method to extracted dialogues from Twitter dataset [7]. The representation of data in these works is in the form of (query, response). This representation creates a significant limitation for generating contextually appropriate responses. Also, the dialogue generated with these approaches is usually short and not informative.

To tackle the above limitations, the RNN based approaches for answer generation proposed in [22], [23] that generate longer answers in dialogue systems. Long-Short-Term-Memory (LSTM) [24] and Gated Recurrent Unit (GRU) [25], are two most favourite extensions of RNN that are used in modelling the dialogue systems [26]. The LSTM/GRU models have been shown effective in encoding the textual data; however, they have the limitation for dealing with long context (usually more than 500 words [27]). To address this limitation and exploit the longer-term context, hierarchical models proposed in [9], [14], [28]. Among these methods, one of the popular models is the Hierarchical Recurrent Encoder-Decoder (HRED), which was proposed by [28]. In the HRED model, a two-level hierarchy that combines two RNNs are used, one for a word level and one for the dialogue utterance level. This architecture helps to reduce the vanishing gradient problem, a problem that limits RNN's ability to model very long word sequences. Despite the success of LSTM/GRU models in language generation tasks, their encoding of the entire source sequence into a fixed-size vector brings some limitations, especially when dealing with long source sequences. Attention-based models [28], [29] is another approach that proposed to reduce this limitation. The attention algorithm allows the model to condition on just parts of input context that are relevant to predict the next word. Attention models and variants have contributed to significant progress in the state-of-the-art in machine translation. In a dialogue system, also attention models use to avoid word repetitions in generated responses [30].

A transformer-based architecture like Open-AI GPT and GPT-2 [17], [18], which uses a multi-layer self-attentive mechanism to allow fully-connected cross-attention to the full context in a computationally efficient manner, seems like a natural choice for exploring a more general solution. Transformer models, for example, allow long-term dependency information to be better be preserved across time thereby improving content consistency [18]. They also have higher model capacity due to their deep structure and are more effective in leveraging large-scale datasets than RNN-based approaches [29].

To address the limitations in supervised approaches, some researchers have investigated reinforcement learning for dialogue systems [8], [31], [32]. The RL approach was investigated in both goal-oriented and open-domain dialogue systems recently. In an open-domain dialogue system, user goal is not explicitly defined, hence defining appropriate metrics for evaluating success such as reward functions is the main challenge in such dialogue systems. Li et al. [8] have made the first attempt to use RL in open-domain dialogue systems. In this approach, the dialogue system trained with data generated by conversing two computer agents in a simulator. The author used a combination of three reward functions, to alleviate the problems of the supervised Seq2Seq model.

III. Methodology

A. Model Architecture

We can represent the dialogue system as an alternating sequence between user and machine. The dialogue starts with a query from a user and the machine responses to that query, and this conversation continues until the "end of the dialogue."

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

utterance appears from user. In our proposed model, the multi-turn dialogue history considered as a long text and the sequence generating task considered as language modeling. Let's consider an unsupervised corpus consist of L tokens as $\mathcal{W} = \{w_1, \cdots, w_L\}$. The standard language modeling task objective on corpus $\mathcal{W}$ is defined as maximizing the following likelihood:

$$L(\mathcal{W}) = \sum_{i=1}^{L} \log P_\theta \left(w_i | w_{1 \cdots i-1}\right) \tag{1}$$

Where the prefix $w_{1 \cdots i-1} := w_1, \cdots, w_{i-1}$ is for convenience. where k is the size of the context window, and the conditional probability P is a generative model with parameters θ. We adopt the input representation of the unsupervised model to switch it to the supervised conversational dataset that we have for training our model. In a single turn conversation, if we define the first utterance $\mathbf{x}^i = \{x_1^i, \cdots, x_M^i\}$ as a source sequence (input sequence), with M number of tokens, and the second utterance $\mathbf{y}^i = \{y_1^i, \cdots, y_N^i\}$ as a target sequence (or a ground-truth), with N number of tokens, then the dataset $\mathcal{W}^c$ consists of $(\mathbf{x}^i, \mathbf{y}^i)$ pairs, $\mathcal{W}^c = \{(\mathbf{x}^1, \mathbf{y}^1), \cdots, (\mathbf{x}^D, \mathbf{y}^D)\}$, where for each source utterance $\mathbf{x}^i$, there is a ground-truth $\mathbf{y}^i$ and D is the number of pairs in a dataset. The conditional probability of generating target sequence given source sequence in a single turn conversation can be written as the product of a series of conditional probabilities:

$$P_\theta(\mathbf{y}^i | \mathbf{x}^i) = \prod_{j=2}^{n} P_\theta \left(y_j^i | y_{1 \cdots j-1}^i, x_{1 \cdots m}^i\right) \tag{2}$$

For multi-turn conversation, after generating the first response $\mathbf{y}_1$, it will concatenate with the source sequences to create a dialogue history $\mathbf{x} = \{\mathbf{x}^1, \mathbf{y}^1\}$. In next turn, to generate the response $\mathbf{y}^2$ associated for input utterance $\mathbf{x}^2$, the dialogue history $\mathbf{x} = \{\mathbf{x}^1, \mathbf{y}^1, \mathbf{x}^2\}$ is considered as a source utterance. The conversation continues until the "end of dialogue" utterance appears from the user. In a multi-turn dialogue generation task, given the dialogue history $\mathbf{x}$, the dialogue response generation task can be defined as generating a response $\hat{y}$ with G number of tokens $\hat{\mathbf{y}} = \{\hat{y}_1, \cdots, \hat{y}_G\}$ where the distribution of the generated tokens is defined as follows:

$$P_\theta \left(\hat{\mathbf{y}} | \mathbf{x}\right) = \prod_{i=2}^{g} P_\theta \left(\hat{y}_i | \hat{y}_{1 \cdots i-1}, \mathbf{x}\right) \tag{3}$$

The generative model that is used in our proposed R-TD model as a policy network $\pi_\theta = P_\theta$, is a multilayer Transformer decoder based on the GPT-2 architecture where θ in the parameters of GPT-2 architecture. The GPT-2 model applies a multi-headed self-attention operation over the input tokens followed by position-wise feedforward layers to produce an output distribution over target tokens [17], [18]. We used a 12-layer GPT-2 model with masked self-attention heads (12 attention heads) and hidden state size of 768 dimensional states with maximum sequence length of 1024 tokens. The training objective for generating sequences in GPT-2 model is defined as maximizing the following likelihood:

$$L(\mathcal{W}^c) = \sum_{j=2}^{n} \log \pi_\theta \left(y_j^i | y_{j-1}^i, \mathbf{x}\right) \tag{4}$$

B. Sequence Generation as an RL Problem

In our proposed model, as discussed previously, the Transformer decoder is viewed as an "agent" that interacts with an external "environment". The parameters of the model, θ, define the policy π_θ, that predicts the next word as an action at each time step. Let us define $\mathcal{S}$ as a possible infinite set of states the environment can be in, $\mathcal{A}$ is a possibly set of actions $\hat{y}_t \in \mathcal{A}$ the agent can take in a state, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is a transition probability of the environment reaches to state s_{t+1} after taking action in state s and R is a reward that agent received at the end of trajectory. The interaction between agent and environment is modeled as a discrete-time Markov decision process (MDP) [33] that is described by a tuple $M = \langle \mathcal{S}, \mathcal{A}, p, R, \gamma \rangle$. The agent observes the environment's current state $s_t \in \mathcal{S}$ and takes an action $\hat{y}_t$ according to a policy $\pi_\theta \left(\hat{y}_t | s_t\right) : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, then the environment transitions to a next state s_{t+1} according to transition probabilities $p(s_{t+1} | s, \hat{y}_t)$. Upon generating the last token (end of sequence token), the agent receives the reward based on reward function definition. In our work we consider the discount factor $\gamma = 1$.

The goal of training is to maximize the expected reward $J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[R\left(\tau\right)\right]$ where τ is the trajectory (generated words in a sequence). Note that, since in the task of dialogue generation, the trajectory is finite (the length of generated sentences are finite) then we described our policy gradient in the form of bounded-length trajectory case with the length of H as described in $\tau = (s_1, \hat{y}_1, s_2, \hat{y}_2, \ldots, s_H, \hat{y}_H)$.

The policy gradient (PG) [34] algorithms, are family of the algorithms that tries to optimize the policy directly. The gradient $\nabla_\theta J(\theta)$ is computed as follows:

$$\nabla_\theta J\left(\pi_\theta\right) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{H} R(\tau) \nabla_\theta \log \pi_\theta \left(\hat{y}_t | s_t\right)\right] \tag{5}$$

The vanilla policy gradient update described in (5) has no bias but high variance. In order to reduce the expose high variance, we add a baseline function $b(s_t)$ to (5) as follows:

$$\nabla_\theta J\left(\pi_\theta\right) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{H} \nabla_\theta \log \pi_\theta \left(\hat{y}_t | s_t\right) \left(R(\tau) - b(s_t)\right)\right] \tag{6}$$

The baseline can be an arbitrary function, as long as it does not depend on the "action", hence the baseline does not change the expected gradient, but importantly, it can reduce the variance of the gradient estimate.

C. Choice of Baseline in Policy Gradient

The general form of policy gradient can be defined as follows:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{H} \nabla_\theta \log \pi_\theta \left(\hat{y}_t | s_t\right) A^{\pi_\theta}\right] \tag{7}$$

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

where A^{π_θ} is advantage function and could be defined as following form:

$$A^{\pi_\theta} = R(\tau) - b\left(s_t\right) \tag{8}$$

Considering the value function $V^{\pi_\theta}\left(s_t\right)$ and Q-function $Q^{\pi_\theta}\left(s_t, \hat{y}_t\right)$, the other valid choice for advantage function can be defined as:

$$A^{\pi_\theta} \;=\; Q^{\pi_\theta}\left(s_t, \hat{y}_t\right) - V^{\pi_\theta}\left(s_t\right) \tag{9}$$

The choice of advantage function in (9) has the lowest possible variance; however, the advantage function in practice is not known and must be estimated [35]. For advantage function approximation, usually, the neural network is used as a function approximator. The first challenge for using a neural network as a function approximator is that it required a large number of samples, and also it is difficult to obtain stable and steady improvement in training the neural network despite the non-stationary of the incoming data in dialogue systems. To reduce the bias in (6) and not to deal with training the second approximator that causes insatiability, we baseline the REINFORCE algorithm with the reward obtained by the current model under the inference algorithm used at test time. In this method, the baseline obtained by performing a greedy search over model output probability distribution at each time step. Let's define the greedy output selection as $(\hat{y}_1^g, \cdots, \hat{y}_H^g)$. Hence, the advantage in (9) defined as:

$$A^{\pi_\theta} = R(\hat{y}_1, \cdots, \hat{y}_H) - R(\hat{y}_1^g, \cdots, \hat{y}_H^g) \tag{10}$$

This approach avoids all the inherent training difficulties associated with actor-critic methods, where a second critic network must be trained to estimate value functions, and the actor must be trained on estimated value functions rather than actual rewards. A similar approach was used in the context of obtaining baseline with the reward obtained by the current model under the inference algorithm used at test time for image captioning [36], and to our knowledge, this is the first time that this approach incorporated for optimizing the Transformer decoder policy network for dialogue generation task.

D. Proximal Policy Optimization

We apply proximal policy optimization [37] to ensure to take the biggest possible improvement step on a policy without causing the instability in performance. PPO is modified from Trust Region Policy Optimization (TRPO) [38] by using a clipped surrogate objective while retaining similar performance. In TRPO, the policy updates by taking the largest step possible to improve the performance, while satisfying the KL-Divergence constraint that specified how close the new and old policies allowed to be. Since a single bad step can unstable the policy and collapse the policy performance, avoiding this kind of collapse is help to improve the process of training. The PPO only relies on clipping in the objective function to heuristically constrain the KL-divergence and limit the improvement of the new policy not to get far from the old policy. let's define the probability ratio between old and new policies as follows:

$$r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}$$

The objective function of PPO is defined as follows:

$$\theta_{new} = \arg\max_\theta \mathbb{E}_{s,a\sim\pi_{\theta_{old}}}\left[L\left(s, a, \theta_{old}, \theta\right)\right] \tag{11}$$

Where L is defined as follows:

$$L\left(s, a, \theta_{old}, \theta\right) = \min\left(r(\theta)A^{\pi_{\theta_{old}}}(s, a), clip\left(\epsilon, A^{\pi_{\theta_{old}}}(s, a)\right)\right)$$

The parameter ϵ is a hyperparameter and the function clip$(\epsilon, A^{\pi_\theta})$ is defined as follows:

$$\text{clip}\left(\epsilon, A^{\pi_\theta}\right) = \begin{cases} (1+\epsilon)A^{\pi_\theta} & A^{\pi_\theta} \geq 0 \\ (1-\epsilon)A^{\pi_\theta} & A^{\pi_\theta} < 0 \end{cases}$$

The hyperparameter ϵ determines how far away the new policy can improve from old policy while still profiting the objective. Our implementation of PPO for training the policy is based on [39]. We consider 1M episodes with four PPO epochs per batch and one minibatch each, we select $\epsilon = 0.2$ and default value for other parameters according to [39].

E. Reward Function for RL Training

One of the challenges in learning dialogue models with RL is how to define an effective reward function for training the agent in an environment. Defining the proper reward function is an expert domain challenge; i.e., it is required to know the problem definition accurately and have a vast knowledge of distinguishing different actions by an agent. One of the main limitation of the open-domain dialogue system is generating bland and uninformative responses. To address this problem, we implement a mutual information scoring function [26] as a reward function for our RL training approach. Intuitively, maximizing the mutual information, help the model to avoid assigning a high reward to sequences that are ungrammatical or not coherent and allows the model to generate responses that are more specific to the source, while generic responses are largely down-weighted. We can define the mutual information reward function between two consecutive utterances X_i and X_{i+1} as follows:

$$R = (1-\lambda)\log\pi_\theta\left(\mathbf{x}_i|\mathbf{x}_{i-1}\right) + \lambda\log\pi_\theta^{bw}\left(\mathbf{x}_{i-1}|\mathbf{x}_i\right) \tag{12}$$

Where π_θ^{bw} is a backward probability of generating the previous utterance $\mathbf{x}_{i-1}$ given an utterance $\mathbf{x}_i$ and λ is a hyperparameter. In our work we select $\lambda = 0.5$. In (12), the reward function R, employs a pretrained backward model to predict source sentences from given responses. To compute the reward, 8 hypotheses are generated for input source sentence by the policy π_θ by using the top-K sampling method [40] (we set k=10), and then according to (12), the reward associated with each sample is calculated. Maximizing backward model likelihood penalizes the bland hypotheses, as frequent and repetitive hypotheses, can be associated with many possible queries, thus yielding a lower probability for any specific query. The backward pretrained model π_θ^{bw} is trained using the

same π_θ by just interchanging the source and target responses in a training dataset and conditioning the π_θ to generate the source sequence $\mathbf{x}_{i-1}$, given the target sequence $\mathbf{x}_i$.

F. Applying RL training

The algorithms 1 describes our proposed framework to train a Transformer decoder with reinforcement learning algorithm in details.

Algorithm 1: PPO policy optimization

Result: Optimized policy with updated parameter θ^*
Initialized the policy π_θ with parameter θ and clipping threshold $\epsilon = 0.2$;
foreach *epoch* **do**
  foreach *batch* **do**
    Sample the policy to generate set of sequences;
    Calculate the reward R_t^k ;
    Obtain the baseline b_t by greedy-sampling the policy;
    Compute the advantage $A^{\pi_\theta} = R(\tau) - b(s_t)$;
    Assignee the current policy to the old policy : $\theta_{old} \leftarrow \theta$
    foreach *PPO iteration (4 iteration)* **do**
      Compute policy update:
      $\theta^* = \arg\max_\theta \mathbb{E}_{s,a\sim\pi_{\theta_{old}}} [L(s, a, \theta_{old}, \theta)]$
    end
  end
end

In dialogue generation task, since the size of action space is equal to the size of vocabulary in a dataset; hence we deal with a large action space problem. To deal with a randomly initialized poor policy that let to slow convergence or even instability in training, we first train the generative model with cross-entropy loss using the ground truth sequences and then, we handle the transition between cross-entropy loss to RL loss. At the beginning of the training, the model completely relies on the cross-entropy loss. We pretrain the policy until the score on the development set stops improving and then, the training completely relies on RL loss. In next step after pretraining, the policy sampled to generate set of sequences. The advantage associated to these sequences is then calculated using and this advantage considered to compute the policy update with PPO algorithm. In our work, we consider 4 iteration in PPO algorithm for updating the policy at each batch.

IV. Experiments

A. Dataset

We evaluate our proposed model on the dialogue corpus that is extracted from Reddit conversations. Reddit is a massive collection of forums where people can post social news, discuss different topics, share their ideas, and comment on other people's posts. The contents in Reddit organize their subjects into subreddits, which cover a wide range of topics, including sports, news, politics, movies, science, and social media. These subreddits are monitored by moderators and filled with quality content, and the grammatical quality of the sentences extracted from Reddit is very high. Including a wide range of topics with grammatical quality sentences make Reddit well suited for grounded open-domain conversational modeling. The extracted dataset from Reddit dialogues contains about 3M dialogues that are randomly sampled 50K dialogues as a development data and 50K dialogues as test data.

B. Setup

The policy network in our model inherits from OpenAI GPT-2, 12 layers Transformer decoder with masked self-attention heads (12 attention heads) and hidden state size of 768 dimensional states. The model uses learned positional embeddings with sequence length of 1024 tokens. We model a multiturn dialogue session as a long text and frame the generation task as language modeling. We used a bytepair encoding (BPE) [41] vocabulary with 50257 merges and for regularization we used residual, embedding, and attention dropouts with a rate of 0.1. We used the Adam optimization with a max learning rate of 1e-5 and the learning rate was increased linearly from zero over the first 2000 updates and annealed to 0 using a cosine schedule. We first train the Transformer decoder with MLE objective until its score on the development dataset stops improving and then continue training with PPO objective. We evaluate our proposed model compare with 3 baseline generative methods, which we describe below:

- **Seq2Seq**
  The first generative model that we consider as a baseline is a work presented in [15] that is a 4-layers LSTM encoder-decoder with 1000 cells at each layer and 1000 dimensional word embeddings.
- **RL-Seq2Seq**
  This model is proposed by [8]. In this scheme the policy gradient method for optimizing the Seq2Seq policy is implemented and manually tailored reward function considered.
- **Transformer decoder with MLE**
  We also consider Transformer decoder as a generative model with the MLE objective without training with the PPO objective. For this model, we trained the Transformer decoder with MLE objective, and the training will be continued until there is no improvement observed in the validation dataset.

C. Automatic Evaluation

To evaluate a quality of generated responses in our proposed dialogue generation model, we performed automatic evaluation based on relevance and diversity metrics. To evaluate diversity, we use distinct unigrams (Distinct-1)/bigrams (Distinct-2) [8] and Entropy (Ent-n) [42] metrics in our work. Models with higher a number of distinct n-grams and Entropy tend to produce more diverse responses.

For relevance evaluation, we adopt contextualized embedding metrics. Contextualized representations from a transformer-based model like BERT [16] are recently shown

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

to be beneficial in many NLP tasks. In our work, we consider three embedding-based metrics and we use BERT to have contextualized representations for each word. The three embedding metrics that we consider in our work are Average metric [43], Greedy metric [44] and Extreme metric [45]. In the Average metric, two separated vector achieves by taking the mean over word embeddings in model generated response and ground-truth response, and then the cosine similarity between these two vectors computes. In the greedy metric, the responses embeds by taking the maximum cosine similarity over embeddings of two utterances. The Extreme metric obtains sentence representation by taking the largest extreme values among the embedding vectors of all the words it contains, then calculates the cosine similarity of the sentence representations.

The last evaluation metric we consider is Normalized Average Length (NAL) metric. The NAL metric measures the average number of words in model-generated responses normalized by the average number of words in the ground truth. To compute the NAL score, we consider the length of ground-truth and generated responses and compute the ratio between these two sequences.

The results obtained using diversity evaluation metrics are summarized in Table I.

TABLE I
DIVERSITY METRICS FOR RESPONSE GENERATION IN REDDIT DIALOGUES DATASET

Model	Dist-1	Dist-2	Ent-4
Seq2Seq	0.761%	1.912%	6.832
RL-Seq2Seq	1.820%	4.233%	8.112
MLE-TD	6.561%	25.426%	9.117
R-TD	11.173%	47.348%	10.876

The results in Table I demonstrate that the R-TD model achieves the highest diversity scores among other models. comparing the R-TD with MLE-TD and RL-Seq2Seq, we observe substantial improvements on diversity due to use of mutual information reward function and incorporating the PPO update from reinforcement learning algorithm in training the Transformer decoder.

The results obtained using relevance evaluation metrics are summarized in Table II. The results in Table II demonstrate

TABLE II
RELEVANCE METRICS FOR RESPONSE GENERATION IN REDDIT DIALOGUES DATASET

Model	Average	Greedy	Extreme	Avg Length
Seq2Seq	0.529	0.393	0.366	8.31
RL-Seq2Seq	0.683	0.431	0.401	11.16
MLE-TD	0.793	0.544	0.502	12.68
R-TD	0.823	0.661	0.644	16.72

that the R-TD model achieves the highest scores based on relevance metrics; which means that it capable of capturing the topic of the target response than other models. In multi-turn dialogue generation system, if the generated response deviates too much from the target topic, the response content will not be relevant to the dialogue context and it deserves a lower relevancy score; and Similarly, if the generated response is highly related to the target topic, the diversity score becomes lower. This trade-off between relevance and diversity must be considered during the evaluation process. The results in Table 1 and Table II show that the R-TD model is capable of generating diverse response with high relevance score to the topic, compare with all three dialogue systems. If we compare the R-TD model with MLE-TD in Table II, we observe that the average length of the generated response by R-TD is higher than MLE-TD. This indicates the MLE-TD suffer from the impact of being trained with only the maximum likelihood objective.

D. Qualitative Evaluation

Table III and Table IV demonstrate samples of the model generated dialogues. The dialogue presented in Table III is an open topic dialogue between user and bot. The questions ask from the bot are not topic-specific, and they are general questions. The questions asked from a bot that presented in Table IV are related to a specific topic (social media in this case), and the bot answer the user specific questions about social media. The results in Table III and Table IV demonstrate that our model is capable of handling different conversations with a high level of coherence in the generated responses. The responses generated by our model are related to the questions with a high level of diversity in responses. We observe that the dull responses like "I don't know" and "I am not sure" are not generated by the model, and instead, the system generates an alternative, reasonable answer.

V. CONCLUSION

In this paper, we propose a novel Transformer-based conversational model, Reinforce Transformer decoder (R-TD), an autoregressive Transformer decoder model that trained with Proximal Policy Optimization from reinforcement learning algorithm for training response generation to promote informative and diverse conversations between human and dialogue agent. We evaluated our model on the Reddit dataset. The results demonstrate that the R-TD model improves the proportion of high-quality responses without losing the ability to generate fine quality replies in comparison to existing methods. The results reveal that the combination of Transformer architecture with reinforcement learning training algorithm is responsible for the performance improvement over simple Transformer decoder architecture that trained based on maximum likelihood estimation objective.

REFERENCES

[1] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, p. 36–45, Jan. 1966.

[2] H. Shum, X. He, and D. Li, "From eliza to xiaoice: challenges and opportunities with social chatbots," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 10–26, 2018.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE III
AN INTERACTIVE EXAMPLE OF GENERAL TOPICS CONVERSATION

Role	Response
User	hi how are you?
Bot	I'm good! Just had the worst hangover I've ever had. What about you?
User	I am good, thanks for asking, why you had hangover?
Bot	I went to a friends house and was eating too many things so I ate too much food, so I fell sick.
User	I hope you will be better soon
Bot	thank you, me too. I am a very tired person, so I am also going to stay sick.
User	do you go to the gym regularly?
Bot	yes I try to go once a week. I go every 4 days.
User	I think once a week is not enough, you must go to the gym 3 times a week!
Bot	well that was a bit harsh. maybe 2 3 times
User	Do you prefer to go to the gym in the morning or afternoon?
Bot	I usually go to the gym after 8 or 9 pm. sometimes I go to my own gym after 5.
User	I like to go to the gym at noon
Bot	that's the only time I go to gym. it's only an hour or so away.
User	what is your job?
Bot	I work as a chef
User	cool! so you know about food!
Bot	yeah, I know about stuff like sushi
User	how long do you work as a chef?
Bot	5 years I've been a chef since high school. I'm a student now.
User	so you work and study at the same time!
Bot	I work and study at the same time. That makes me hard
User	good, I wish success in your career as a chef
Bot	And I wish you success in your career as well!
User	Thanks, have a good day!
Bot	Np, see you later!

TABLE IV
AN INTERACTIVE EXAMPLE OF SPECIFIC TOPIC CONVERSATION

Role	Response
User	Hi how are you?
Bot	I'm fine, how are you?
User	I am good, thanks for asking!
Bot	No problem. I am glad you are doing ok!
User	I want to talk about social media with you
Bot	Yeah, I wanna hear from you about social media. Maybe we could make it a bit like this, social media is like an internet thing but for gaming.
Uses	do you have any accounts in social media?
Bot	I used to have a twitter account, but I don't anymore.
User	why?
Bot	Because I was too lazy to change my old password
User	what do you think about the role of social media in our society?
Bot	Not much. I was very active on it but not to the level of most things. I think I would be in my prime as a journalist though
User	What type of content we can share on social media?
Bot	We're all social media, it just takes time. You can't do everything in one post like on twitter, but you have to start somewhere and be consistent
User	do you think social media will be dangerous in future?
Bot	Not a lot but it will be good, social media is pretty much the new way to get information
User	what is the risk of social media for kids?
Bot	They're the same risk as they are for adults.
User	could you name some risks of social media for kids?
Bot	Social media for kids doesn't exist.
User	How much time do you spend replying and reacting to all your Social mentions?
Bot	I usually only respond to one of my social mentions a week, and respond to others when I'm feeling down and or depressed.
User	Does your mood affect your behaviour on Social Media?
Bot	Nope, just personal. Social Media is a social game and I'm just a social player, I really enjoy my time with people.
User	Has Social media made your life better?
Bot	yeah, I think so, I've gotten over so much stress and the only problem I ever had to deal with is the one time I broke up with someone.
User	nice to chat with you! take care!
Bot	Thanks, you too, man!

[3] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu *et al.*, "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 530–539, 2014.

[4] L. Deng, G. Tur, X. He, and D. Hakkani-Tur, "Use of kernel deep convex networks and end-to-end learning for spoken language understanding," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 210–215.

[5] A. Deoras and R. Sarikaya, "Deep belief network based semantic taggers for spoken language understanding." in *Interspeech*, 2013, pp. 2713–2717.

[6] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 583–593.

[7] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems: The journal version," *Dialogue & Discourse*, vol. 9, no. 1, pp. 1–49, 2018.

[8] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1192–1202.

[9] C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou, "Hierarchical recurrent attention network for response generation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[11] M.-T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 11–19.

[12] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[13] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. [Online]. Available: http://arxiv.org/abs/1511.06732

[14] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-*

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

assets/researchcovers/languageunsupervised/language understanding paper. pdf, 2018.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[19] M. Simard, N. Ueffing, P. Isabelle, and R. Kuhn, "Rule-based translation with statistical phrase-based post-editing," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2007, pp. 203–206.

[20] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational linguistics*, vol. 30, no. 4, pp. 417–449, 2004.

[21] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1700–1709.

[22] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 196–205.

[23] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1577–1586.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[26] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 110–119.

[27] U. Khandelwal, H. He, P. Qi, and D. Jurafsky, "Sharp nearby, fuzzy far away: How neural language models use context," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 284–294.

[28] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie, "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 553–562.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[30] H. Mei, M. Bansal, and M. R. Walter, "Coherent dialogue with attention-based language models," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[31] J. Li, A. H. Miller, S. Chopra, M. Ranzato, and J. Weston, "Learning through dialogue interactions by asking questions," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=rkE8pVcle

[32] J. Li, A. H. Miller, S. Chopra, and M. Ranzato, "Dialogue learning with human-in-the-loop," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=HJgXCV9xx

[33] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[34] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[35] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *CoRR*, vol. abs/1506.02438, 2015.

[36] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.

[37] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[38] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, 2015, pp. 1889–1897.

[39] P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, Y. Wu, and P. Zhokhov, "Openai baselines," https://github.com/openai/baselines, 2017.

[40] Q. Huang, Z. Gan, A. Celikyilmaz, D. Wu, J. Wang, and X. He, "Hierarchically structured reinforcement learning for topically coherent visual story generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8465–8472.

[41] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1715–1725.

[42] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan, "Generating informative and diverse conversational responses via adversarial information maximization," in *Advances in Neural Information Processing Systems*, 2018, pp. 1810–1820.

[43] J. Mitchell and M. Lapata, "Vector-based models of semantic composition," in *proceedings of ACL-08: HLT*, 2008, pp. 236–244.

[44] V. Rus and M. Lintean, "An optimal assessment of natural language student input using word-to-word similarity metrics," in *International Conference on Intelligent Tutoring Systems*. Springer, 2012, pp. 675–676.

[45] G. Forgues, J. Pineau, J.-M. Larchevêque, and R. Tremblay, "Bootstrapping dialog systems with word embeddings," in *Nips, modern machine learning and natural language processing workshop*, vol. 2, 2014.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE