Empathetic Response Generation via Regularized Q-Learning

Jen-Tzung Chien and Yi-Chien Wu
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

***Abstract*—This study presents a new offline reinforcement learning to improve the diversity of dialogue generation while the issue of limited training data is tackled. In particular, a regularized Q-function is introduced to generate the empathetic response where the distribution shift between the learned policy and the behavior policy is penalized and the resulting out-of-distribution response is mitigated. Importantly, an emotion aware reward function is designed to guide the agent to sufficiently learn various dialogue traits and emotions for empathetic response generation. In addition, a soft sampling method is implemented to draw the action to ensure the agent to reach those candidate responses which are marginal but appropriate. The flexibility of response generation is accordingly enhanced through the reward function by considering the informative dialogue features. The experiments on the emotional conversation generation show the merit of the proposed method.**

I. Introduction

Traditional response generation using rule-based or knowledge-based methods generally guarantee the high-quality responses but usually result in some errors when the user inputs are not matched with the pre-defined templates. With the rapid growth of deep learning from massive social media data, the conversational AI has been successfully developed to implement for deployment of various dialogue applications. Large language models (LLMs) based on the generative pre-trained transformer (GPT) have been popularly developed to generate flexible and human-like expressions with diverse responses. However, purely considering the prediction of word sequence based on probability distribution using GPT may not really meet the human expectations. It is crucial to adjust the existing foundation model to build a large-scaled pre-trained backbone to meet a specific desired expectation. For example, how to attract users to keep interacting with the dialogue system is a critical demand whereas the empathy is seen as an important consideration as it significantly affects the user satisfaction.

To enhance the diversity of natural language generation in an empathetic conversation, the generation of sentences is treated as a sequential decision-making process where reinforcement learning (RL) can provide an useful solution. The advantage of utilizing RL methods to build a dialogue agent rather than using supervised learning methods is due to the fact that RL methods more likely integrate the specific rewards which can be designed by customers to meet different tasks and thus better achieve true goals of the agents. Besides, by considering RL schemes, it is possible to incorporate long-term influences and dynamics from a generated response into an ongoing conversation. This paper presents an offline RL method that utilizes an emotion aware reward function to guide the agent to learn the desired features for response generation in a conversation. Particularly, the regularized Q-learning is developed to prevent the language model based on GPT2 from producing the out-of-distribution responses that may harm the overall conversation performance. The goal of this study is to enhance the empathy in dialogue generation and increase the user desire to interact with the system. Some analytical experiments on natural language generation are reported to illustrate the usefulness of this method.

II. Dialogue and Behavior Modeling

A. Learning for dialogue generation

Reinforcement learning (RL) is a powerful technique to simulate the action transitions [1] which are implemented by designing or crafting the reward function where the interactive dialogue system considering the whole trajectory for selecting dialogue acts or generating conversational responses [2] can be flexibly constructed [3], [4]. In general, dialogue generation can be seen as a sequential decision-making process where the agent takes an action a_t based on the policy π_θ, transits to a new state s_{t+1}, and then receives a feedback r_t in a form of reward which is maximized along a trajectory to improve its response performance. Hence, the learning objective is to maximize the expected cumulative future reward

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\gamma^t r_t\right] \tag{1}$$

where θ is the policy parameter and $\gamma \in (0, 1]$ is the discount factor. The goal of the whole procedure is to learn how to continuously choose an action that maximize $J(\theta)$.

Integrating RL techniques in a dialogue system involves learning the dialogues which are obtained incrementally through the tricks of self-play or human-in-the-loop. Typically, such a learning procedure is costly and time-consuming, unlike the supervised learning which trains a dialogue model from a data collection. The existing RL methods depend heavily on interacting with a learned model of a human [3], [5] to generate low-cost experience, but this may result in unintelligible language to generate incorrect response. There are twofold solutions to handle this issue. One is to impose the strong prior to ensure the generated response which is closely similar to that contained in the dataset [3], [6], and the other is to utilize the dialogue management or the template-based method to directly reuse the samples in the dataset [5].

B. Learning with behavior regularization

Offline RL is useful in a scenario where the frequent interactions with an environment are not allowed, such as the scenarios in healthcare tasks and self-driving cars. However, the training dataset $\mathcal{D}$ for offline RL follows the visitation distribution of the behavior policy π_β. This circumstance causes the demand to occasionally evaluate the state-action pairs that may not be well covered from the training data when computing the loss. In addition, the estimated Q-function tends to overestimate those out-of-distribution (OOD) behaviors [7]. The shift in visitation distribution leads to a significant extrapolation error which will be propagated and accumulated across the decision-making chain process, and seriously harm the model performance for dialogue behavior. The solutions to address the distribution shift in offline RL mainly fall into two categories including policy regularization [7], [8], [9] and critic penalty [10], [11], [12].

1) Policy regularization: Policy regularization aims to guide the agent to learn an optimal policy π under specific constraints, such as minimizing the distance of the selected actions to those in the training dataset and keeping a state-action visitation similar to that appearing in the dataset so as to alleviate the extrapolation error in offline RL. One classic method, called the batch-constrained deep Q-learning (BCQ) [7], employed the conditional variational auto-encoder (VAE) $G_\omega(\cdot)$ to generate the plausible candidate actions with perturbations ξ_ϕ to increase the diversity of generation. The Q-function $Q(\cdot)$ measures the scores of n candidate actions a_i sampled from the generative model $\{a_i \sim G_\omega(s)\}_{i=1}^n$ and the one with the highest score is selected as the final output. This scheme effectively regularizes the learned policy $\pi_\theta(s)$ given by state s via

$$\pi_\theta(s) = \underset{a_i+\xi_\phi(s,a_i,\Phi)}{\operatorname{argmax}} Q\left(s, a_i + \xi_\phi\left(s, a_i, \Phi\right)\right) \tag{2}$$

to be as close as the behavioral policy π_β. For the application in natural language generation, the pre-trained language models such as those GPT variants [13], [14], [15] can provide a language prior $\mu(\cdot|s)$ that prevents RL model π from selecting inappropriate words. This language prior is useful in offline RL with limited data.

2) Critic penalty: Considering the conservative Q-learning (CQL) [10], the critic penalty is merged in the Q-function where the conservative Q-value is estimated to act as the lower bound of Q-value by adding the regularization term. The Q-network as a critic is updated to calculate the Q-value according to the minimax objective

$$\begin{aligned}\min_Q \max_\pi \ & \alpha \cdot \left(\mathbb{E}_{s\sim\mathcal{D},a\sim\pi}[Q(s,a)] - \mathbb{E}_{s\sim\mathcal{D},a\sim\widehat{\pi}_\beta}[Q(s,a)]\right) \\ & + \frac{1}{2}\mathbb{E}_{s,a,s'\sim\mathcal{D}}\left[\left(Q(s,a) - \widehat{\mathcal{B}}^{\pi_k}\widehat{Q}^k(s,a)\right)^2\right] + \mathcal{R}(\mu)\end{aligned} \tag{3}$$

where α is the learning rate, $\widehat{\pi}_\beta(a|s)$ is the empirical behavior policy, $\widehat{\mathcal{B}}^{\pi_k}$ is the empirical Bellman operator, and $\widehat{Q}^k$ is the updated Q-value in iteration k. In Eq. (3), the second term is seen as the temporal difference error and the others are treated as the regularization terms. The term $\mathbb{E}_{s\sim\mathcal{D},a\sim\pi(a|s)}[Q(s,a)]$ is used as a penalty to minimize the Q-value for the action a chosen by the currently learned policy π at the state s contained in the dataset, and the term $-\mathbb{E}_{s\sim\mathcal{D},a\sim\widehat{\pi}_\beta(a|s)}[Q(s,a)]$ is utilized to encourage the Q-network to provide relatively high scores for those actions a that exist in the dataset with the state s. Such a regularization, known as the CQL regularizer, is added during the training process, which is helpful to reduce the chance of obtaining the overestimated Q-value. As for the final term $\mathcal{R}(\mu)$, it can be chosen as a particular regularizer according to the need of a task. For instance, the Kullback-Leiblier (KL) divergence is viewed as a popular choice of measuring the closeness between the current distribution $\mu = \pi$ and its prior distribution ρ, namely

$$\mathcal{R}(\mu) = -D_{\mathrm{KL}}(\mu\|\rho). \tag{4}$$

This regularization term is minimized to find a dialogue policy π which is as close as ρ.

III. Empathetic Response Generation

This study presents a new approach to empathetic response generation [16] where the Chinese sentiment analysis (cnsenti) library in [17], [18] was adopted to perform the sentiment analysis and the emotion detection from the user's text input and directly concatenate the detected emotion class to the input. The cnsenti library was also used to label the emotions for the replies in the training corpus to address the lack of high-quality corpus with Chinese training data for various emotion classes. This corpus was used in the experiments on RL dialogues by fine-tuning LLM based on GPT-2.

A. Emotion aware reward

A challenging issue for a successful dialogue system is to learn how to draw people to continue interacting with the system. This study takes into account the following six empathetic features $\{r_i\}_{i=1}^6$ [3], [6] and builds an emotion aware reward to learn an agent to generate the empathetic responses with diversity and safety.

- *length of the agent response:* r_1 is to encourage the agent to generate longer responses while avoiding very short responses that end the conversation easily.
- *consistency of user sentiment with agent emotion:* r_2 is to encourage the agent to provide the empathetic replies by considering the detected emotion of the user input and incorporating it into the generated response.
- *asking questions:* r_3 is to encourage asking the follow-up questions related to the inputs to show that the agents have actively listened to the user's concerns and are engaged in the conversation.
- *context consistency:* r_4 is to enhance the coherence of the generated reply and avoid the situation where the generated reply receives a high return but its content is ungrammatical or semantically incoherent by calculating

$$r_4 = \frac{1}{N_a}\log p_{\mathrm{s2s}}^{\mathrm{fw}}(a|s) + \frac{1}{N_s}\log p_{\mathrm{s2s}}^{\mathrm{bw}}(s|a) \tag{5}$$

where $p^{\text{fw}}_{\text{s2s}}(a|s)$ and $p^{\text{bw}}_{\text{s2s}}(s|a)$ denote the forward (fw) and backward (bw) probabilities for generating a response a given user input s and producing the previous sentence s given a response a, and N_a and N_s denote the lengths of a and s in a sequence-to-sequence (s2s) generation, respectively. Bidirectional context coherence is measured.

– *detection of offensive words:* r_5 is to measure the penalty when the generated response contains some inappropriate words which are predefined in a given list and received by a negative reward.

– *suppression of non-sense reply production:* r_6 is to measure how the sentences generated by the agent have the sensitivity or potential to continue the conversation by computing the negative log likelihood of a non-sense sentence s via

$$r_6 = -\frac{1}{N_{\mathbb{S}}} \sum_{s \in \mathbb{S}} \frac{1}{N_s} \log p_{\text{s2s}}(s|a). \tag{6}$$

In Eq. (6), $N_{\mathbb{S}}$ is the cardinality of a set of generated sentences $\mathbb{S}$, and N_s is the number of tokens in a non-sense response s.

Finally, the emotion aware reward function r_t of a response a at time t is a weighted sum of the rewards over empathetic features $r = \sum_{i=1}^{6} \lambda_i r_i$ by using the weights $\{\lambda_i\}_{i=1}^{6}$.

B. Regularized Q-function

This study presents the behavior regularization in dialogue generation by controlling the extrapolation errors due to the distribution shift in offline RL. A regularized Q-function is implemented to constrain the dialogue agent from choosing OOD actions. At the same time, the confident actions are up-weighted to allow the agent to focus on the successful emotion aware samples.

1) Extrapolation error control: In the implementation, the pre-trained language model is fine-tuned by using a specific dataset to obtain the probability $p(a|s)$ of generating the response a given a dialogue input s, which acts as a prior probability to avoid selecting those OOD actions [7]. The KL divergence between current policy q and prior policy p is added as the regularization term in the objective for Q-learning. The learning objective is formulated as

$$\mathbb{E}_{q(\tau)}[r(\tau)] - D_{\text{KL}}[q(\tau) \| p(\tau)] \tag{7}$$

where τ is a trajectory of dialogue actions, and the distributions of learned policy and prior policy are calculated by $q(\tau) = \prod_{t=1}^{T} \pi_\theta(a_t|s_t)$ and $p(\tau) = \prod_{t=1}^{T} p(a_t|s_t)$, respectively, and

$$D_{\text{KL}}[q\|p] = \sum_{\tau} q(\tau)(\log q(\tau) - \log p(\tau)). \tag{8}$$

Such a regularized Q-function guides the agent to prefer choosing the actions in the dataset and prevent selecting the OOD actions. To fulfill Eq. (7), the regularized Q-function (denoted as the RQ score) is yielded by

$$\begin{aligned} Q\left(s_t, a_t\right) = \mathbb{E}_\pi \Bigg[& \sum_{t'=t}^{T} r\left(s_{t'}, a_{t'}\right) + \log p\left(a_{t'}|s_{t'}\right) \\ & - \log \pi_\theta\left(a_{t'}|s_{t'}\right)\Bigg] \end{aligned} \tag{9}$$

which is maximized in RL to select the generated responses. In addition to reward maximization, the term $\mathbb{E}_\pi[\log p\left(a_{t'}|s_{t'}\right)]$ is maximized to encourage the agent to select the actions under the prior and prevent the unrealistic OOD actions via KL regularization, and the term $\mathbb{E}_\pi[-\pi_\theta\left(a_{t'}|s_{t'}\right)]$ is seen as an entropy function which is maximized to maintain the diversity of dialogue agent.

2) Importance weight approximation: Importance weight approximation is employed to up-weight the actions with high confidence which help the agent focus on successful dialogue trajectories. In the offline RL setting, we could not directly sample the trajectories from the learned policy π_θ to calculate the total expected reward. Accordingly, the behavior policy π_β is employed to sample the trajectories $\tau = \{s_t, a_t\} \sim \pi_\beta$ for learning from the fixed dataset in an offline manner. Importance sampling is used to address the difference between the sample policy π_β and the updated policy π_θ. The unbiased estimation of policy gradient given by the importance weight

$$w_t = \prod_{t'=1}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\pi_\beta(a_{t'}|s_{t'})} \tag{10}$$

can be obtained by

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\beta}\left[\sum_{t=1}^{T} w_t \nabla_\theta \log \pi_\theta(a_t|s_t) Q(s_t, a_t)\right] \tag{11}$$

which represents the gradient with the total expected reward regularized from the state s_t.

However, this process may be sensitive during optimization. The convergence time will be increased when the importance weight is multiplied over each action for numerous time steps. To tackle this issue, the importance weight is approximated by $w_t \approx \frac{\pi_\theta(a_t|s_t)}{\pi_\beta(a_t|s_t)}$. This is equivalent to maximize the total expected reward under the on-policy action distribution of π_θ and the off-policy state distribution produced by π_β. Despite of being biased, this estimator has been demonstrated empirically to reduce the variance and perform well when π_β and π_θ are close. As for how to obtain the value of π_β, an intuitive way is to estimate π_β directly from the dataset. Here, a simple approximation $\pi_\beta(\tau) \approx 1/N$ is made when τ is included in the dataset, otherwise $\pi_\beta(\tau) = 0$. With this approximation, the denominator, i.e. π_β, can be neglected in the optimization since it is a constant. Then, the approximation to policy gradient $\nabla_\theta J(\theta)$ in Eq. (11) is further derived by

$$\begin{aligned} & \mathbb{E}_{\tau \sim \pi_\beta}\left[\sum_{t} \prod_{t'=1}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\pi_\beta(a_{t'}|s_{t'})} \nabla_\theta \log \pi_\theta(a_t|s_t) Q(s_t, a_t)\right] \\ & \approx \mathbb{E}_{\tau \sim \pi_\beta}\left[\sum_{t} \frac{\pi_\theta(a_t|s_t)}{\pi_\beta(a_t|s_t)} \nabla_\theta \log \pi_\theta(a_t|s_t) Q(s_t, a_t)\right] \\ & \approx \sum_{i=1}^{N} \sum_{t=1}^{T} \pi_\theta(a_t^i|s_t^i) \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) Q(s_t^i, a_t^i) \end{aligned} \tag{12}$$

where i denotes the trajectory index, and $\pi_\theta(a_t^i|s_t^i)$ is the model confidence to up-weight the actions preferred by current policy.

C. Implementation procedure

The overall workflow of response generation and candidate selection is addressed. The whole process is divided into three stages. In the first stage, the cnsenti library [17] was employed as an emotion detector where the emotion class of an user input was estimated and concatenated with the input as an augmented user input to feed in language model based on GPT-2 [14]. This GPT-2 language model was fine-tuned by using the emotional conversation generation (ECG) [19] dataset which contained 1M samples of user inputs and their responses. The adapted GPT-2 was accordingly used to generate several candidate responses as the possible empathetic responses for subsequent selection. In the second stage, Q-function acted as a critic to provide the evaluation of Rq scores over individual candidate responses based on the emotion aware reward function. This implementation considered the double Q-network [20] where one Q-network was responsible for estimating the Q-value of each candidate response while the other Q-network was used as a target network to calculate the target Q-value. The updating process for double Q-network was based on the minimization of mean square error loss between the estimated Q-value and target Q-value. In particular, each Q-network was implemented by using BERT encoder [21]. In the third stage, the best candidate was selected according to the RQ scores. Importantly, the diversity of the selected response was assured since those suboptimal responses also had a chance to be selected during the RL procedure. Here, the selection was based on the distribution over Q-value

$$\pi_\theta(a|s) \propto \exp Q(s,a). \tag{13}$$

This implementation was carried out for evaluation the proposed method in the following experiments.

IV. Experiments

This study conducted the evaluation on empathetic response generation based on the GPT-2 (Chinese) which was fine-tuned by using ECG dataset in the first stage and then fine-tuned by using the proposed RL fine-tuning in the second stage.

A. Experimental settings

In the implementation, the dataset ECG [19] was used for GPT-2 fine-tuning in the first stage while the other dataset called the large-scale cleaned Chinese conversation (LCCC) [23] was used for RL-based fine-tuning where LCCC was rigorously filtered to eliminate various kinds of noises. There are 6.8M single-turn and multi-turn conversations in LCCC-base [23] dataset where their emotions were obtained through cnsenti [17]. The emotion classes include 喜歡 (like), 悲傷 (sad), 噁心 (disgusted), 憤怒 (angry), 開心 (happy), and 其它 (other). The simplified Chinese was converted to the traditional Chinese for all text data in the experiments. In the learning procedure, AdamW optimizer was used. The initial learning rate was 1.5e-4 and 5e-5, the batch size was 4 and 8, the number of parameters was 102M and 311M, and the training time was 4 and 60 hours in first and second stages, respectively, where GPU with GeForce RTX 4090 24GB was used and 90K text corpus was collected in RL training. For comparative study, the perplexity (PPL) was used to measure the performance for generation and prediction in the evaluation. BLEU-2 and BLEU-4 [24] were evaluated for Chinese response generation where the complete meanings of Chinese vocabulary with two and four words were evaluated, respectively. Inter-distinct unigram (Dist-1) and bigram (Dist-2) [25] were used to measure the diversity of the generated sentences. These metrics were used to conduct the automatic evaluation for diversity and emotion factor in the generated

TABLE I
RESPONSE GENERATIONS IN THE FIRST CASE STUDY BY USING DIFFERENT METHODS.

Type	Context	RQ Score
Input	明天要出去玩，興奮地睡不著覺 (Go out to play tomorrow, and too excited to sleep.)	
Pre-trained	哈哈，我也覺得興奮 (Ha! Ha! I also feel excited.)	–
CDial-GPT	哈哈，我也是 (Ha! Ha! Me too.)	–
RQL Cand 1	我也是興奮的睡不著 (I am also too excited to sleep.)	2.51
Cand 2	明天要是沒有下雨我也興奮的睡不著，明天要出去玩，晚安 (If there is no rain tomorrow, I am excited and sleepless. Go out to play tomorrow. Good night.)	1.44
Cand 3	你明天不是要上班麼 (You aren't going to work tomorrow, aren't you?)	1.58
Cand 4	興奮地睡不著，因為明天又是週末了 (So excited and could not sleep because weekend is coming again tomorrow.)	1.76
Cand 5	是啊，我今天也要去玩了，興奮得不行 (Yes! I also go playing today. I am super excited.)	0.54
Final reply	興奮地睡不著，因為明天又是週末了 (So excited and could not sleep because weekend is coming again tomorrow.)	1.76

TABLE II
RESPONSE GENERATIONS IN THE SECOND CASE STUDY BY USING DIFFERENT METHODS.

Type	Context	RQ Score
Input	我想念以前住在鄉下的日子，雖然不方便但至少一家人都在一起 (I miss the time when living in the suburb. Although it was inconvenient, at least all family members could live together.)	
Pre-trained	唉，這日子沒法過了！ (Oh! It was hard to live!)	-
CDial-GPT	想念我的床想念我的床 (Miss my bed, Miss my bed.)	-
RQL Cand 1	我也是，想念以前的日子 (Me too. Miss the old time.)	2.46
Cand 2	對啊，我們都要好 (Yes! We were all fine.)	1.14
Cand 3	不過我覺得我們這裡也是這樣，一個人在外面一定要小心啊，我現在一個人在外地也很害怕 (I feel that we here are the same. Be careful when living outside. I am now frightened when living alone and outside.)	1.98
Cand 4	是啊，我也想念以前一起玩的時光啊 (Yes! I also miss the old time when playing together.)	0.99
Cand 5	這是我第一次離開家，真是太想念了 (This is the first time that I left from home. Really miss a lot!)	1.08
Final reply	是啊，我也想念以前一起玩的時光啊 (Yes! I also miss the old time when playing together!)	0.99

TABLE III
EVALUATION OF DIFFERENT METHODS UNDER VARIOUS METRICS.

Model	Size	PPL (↓)	BLEU-2 (↑)	BLEU-4 (↑)	Dist-1 (↑)	Dist-2 (↑)
Transformer [22]	113M	22.10	6.72	3.14	0.0088	0.1397
GPT2-chitchat [23]	88M	–	2.28	0.54	0.0103	0.1625
Pre-trained GPT2 [23]	88M	21.27	5.96	2.71	0.0080	0.1172
CDial-GPT [23]	104M	22.76	5.69	2.50	0.0077	0.1087
RQL w/o IWA	108M	20.08	6.98	3.21	0.0100	0.1623
RQL w/o EEC	108M	19.31	7.15	**3.29**	0.0098	0.1585
RQL w IWA & EEC	108M	**18.01**	**7.20**	3.20	**0.0106**	**0.1695**

responses. The size of parameters was given. An empirical selection of λ_i was done. The multinomial sampling was used to select the final reply.

B. Experimental results

First of all, Table I shows the first case study on the comparison of evaluating the response generations by using the pre-trained model via GPT-2, the baseline model via CDial-GPT [23] and the proposed model via regularized Q-learning (hereafter is denoted by RQL). CDial-GPT is a Chinese GPT-2 which was purely trained on the LCCC dataset while the pre-trained model was fine-tuned by using the ECG dataset. In general, the pre-trained and baseline models generate the appropriate responses. But, both of them relatively lack richness and precision in the response generation when compared with that generated by the proposed RQL. Five candidate responses (ranked by GPT-2 in the first stage) and their RQ scores (calculated by the regularized Q-function in Eq. (9) in the second stage) are shown. The proposed RQL not only pays attention to the emotional word 興奮 (excited) in the input sentence but also considers the other information mentioned in the response such as 睡不著 (could not sleep) and 出去玩 (go out to play), making the response more diverse and informative. Additionally, the proposed RQL extends the information which is not mentioned in the user input like 上班 (go to work) and 週末 (weekend), making the response even more diverse. Notably, the final output is selected by the multinomial sampling which can increase the diversity of the generated response. It is noted that, next to the first candidate (Cand 1), the fourth candidate (Cand 4) receives higher regularized Q-value than those in the second, third and fifth candidates since the fourth response is more empathetic and accurate than the others. This fourth candidate captures the correct tokens 週末 (weekend) which are reached in the other candidates. It is noticed that some emotions are more explicit in guiding the model's response while the others are more implicit and related to the specific emotions and dialogue in the training data. In the second case study as shown in Table II, the related work CDial-GPT model focuses on the single term 想念 (miss), but generates less coherent response, while the proposed model provides more informative candidate replies. However, some responses may not be so appropriate, as seen in the third candidate. Here, the use of multinomial sampling in selecting the final output helps the model choose a response that may not have the top RQ score but is more suitable for the situation.

Table III shows the results of automatic evaluation by using the proposed RQL and the related methods which include the transformer decoder [22], the GPT2-chitchat [23], the pre-trained language model (Chinese variant of GPT-2) [23], and the CDial-GPT [23]. The best result in the comparison is shown by bold. Relative to the other methods, the proposed method conducts a two-stage implementation where the supervised fine-tuning and the regularized Q-learning are performed for domain adaptation and empathetic generation where the emotional domain and the diverse reply are concerned, respectively. There is no RL learning in previous works. For ablation study, the proposed RQL is also implemented by ignoring the modifications based on importance weight approximation (RQL w/o IWA) and extrapolation error control (RQL w/o EEC). There are several findings in RQL. First, the increase of model size is moderate. Notably, BLEU measures the matching between the generated result and the reference answer, and may not reflect the capability of generating diverse responses. Although the BLEU values are generally low for various methods, the proposed RQL is still better than the other methods. In addition, RQL with IWA & EEC obtains the highest Dist-1 and Dist-2 in the comparison. It is obvious that RQL with the schemes of importance weight approximation and extrapolation error control achieves the best results among different methods in most of evaluation metrics.

V. CONCLUSIONS

We have presented a regularized Q-learning approach with the custom-designed emotion aware reward function to train a dialogue agent to generate empathetic responses with specific emotions. The Kullback-Leibler control term, included in the Q-function, helped to penalize the agent for generating the OOD responses, reducing the extrapolation errors caused by the distribution shifts. Using the designed reward functions facilitated the agent to learn the desired features and improved the quality of dialogue responses. This study employed the soft sampling to enhance the diversity of the final output reply. The experiments on Chinese empathetic generation showed the merit of the proposed method. However, this approach may not be enough to handle multi-turn dialogue or integrate the previous conversation history in a long dialogue. Therefore, it will be a future direction to introduce the additional memory to handle the pieces of information that were mentioned in the previous turns. To enhance the empathy in the generated responses, we will also need more precise and appropriate reward functions to guide the empathetic agent to learn better.

REFERENCES

[1] Dhawal Gupta, Yinlam Chow, Azamat Tulepbergenov, Mohammad Ghavamzadeh, and Craig Boutilier, "Offline reinforcement learning for mixture-of-expert dialogue management," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[2] Tien-Ching Luo and Jen-Tzung Chien, "Variational dialogue generation with normalizing flows," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7778–7782.

[3] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao, "Deep reinforcement learning for dialogue generation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1192–1202.

[4] Verena Rieser and Oliver Lemon, *Reinforcement Learning for Adaptive Dialogue Systems: A Data-Driven Methodology for Dialogue Management and Natural Language Generation*, Springer Science & Business Media, 2011.

[5] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang, "Decoupling strategy and generation in negotiation dialogues," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2333–2343.

[6] Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard, "Human-centric dialog training via offline reinforcement learning," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 3985–4003.

[7] Scott Fujimoto, David Meger, and Doina Precup, "Off-policy deep reinforcement learning without exploration," in *Proc. of International Conference on Machine Learning*, 2019, pp. 2052–2062.

[8] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine, "Stabilizing off-policy Q-learning via bootstrapping error reduction," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] Scott Fujimoto and Shixiang Shane Gu, "A minimalist approach to offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[10] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine, "Conservative Q-learning for offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1179–1191, 2020.

[11] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma, "MOPO: Model-based offline policy optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14129–14142, 2020.

[12] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn, "COMBO: Conservative offline model-based policy optimization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28954–28967, 2021.

[13] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving language understanding by generative pre-training," *OpenAI blog*, 2018.

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[16] Ge Li, Mingyao Wu, Chensheng Wang, and Zhuo Liu, "DQ-HGAN: A heterogeneous graph attention network based deep Q-learning for emotional support conversation generation," *Knowledge-Based Systems*, vol. 283, pp. 111201, 2024.

[17] Linhong Xu, Hongfei Lin, Yu Pan, Hui Ren, and Jianmei Chen, "Constructing the affective lexicon ontology," *Journal of China Society for Scientific and Technical Information*, vol. 27, pp. 180–185, 2008.

[18] Jen-Tzung Chien and Chih-Jung Tsai, "Amortized mixture prior for variational sequence generation," in *Proc. of International Joint Conference on Neural Networks*, 2020, pp. 1–6.

[19] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc of AAAI Conference on Artificial Intelligence*, 2018, vol. 32, pp. 730–738.

[20] Hado Van Hasselt, Arthur Guez, and David Silver, "Deep reinforcement learning with double Q-learning," in *Proc. of AAAI conference on Artificial Intelligence*, 2016, vol. 30, pp. 2094–2100.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. of International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[23] Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang, "A large-scale chinese short-text conversation dataset," in *Proc. of CCF International Conference on Natural Language Processing and Chinese Computing*, 2020, pp. 91–103.

[24] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 654–664.

[25] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan, "A diversity-promoting objective function for neural conversation models," in *Proc. of Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 110–119.