Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 12, No. 3, 2024, 349-358.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Transformer-based Generative Chatbot Using Reinforcement Learning

Nura Esfandiari, Kourosh Kiani *, and Razieh Rastgoo

Electrical and Computer Engineering Department, Semnan University, Semnan, Iran.

Article Info

Article History:
Received 07 May 2024
Revised 11 August 2024
Accepted 19 September 2024

DOI:10.22044/jadm.2024.14466.2549

Keywords:
Chatbot, Generative Chatbot, Transformer Model, Reinforcement Learning, Dialogue-based System, Conversation System.

**Corresponding author: Kourosh.kiani@semnan.ac.ir (K. Kiani).*

Abstract

A chatbot is a computer program system designed to simulate human-like conversations and interact with users. It is a form of conversational agent that utilizes Natural Language Processing (NLP) and sequential models to understand user input, interpret their intent, and generate appropriate answer. This approach aims to generate word sequences in the form of coherent phrases. A notable challenge associated with previous models lies in their sequential training process, which can result in less accurate outcomes. To address this limitation, a novel generative chatbot is proposed, integrating the power of Reinforcement Learning (RL) and transformer models. The proposed chatbot aims to overcome the challenges associated with sequential training by combining these two approaches. The proposed approach employs a Double Deep Q-Network (DDQN) architecture with utilizing a transformer model as the agent. This agent takes the human question as an input state and generates the bot answer as an action. To the best of our knowledge, this is the first time that a generative chatbot is proposed using a DDQN architecture with the embedded transformer as an agent. Results on two public datasets, Daily Dialog and Chit-Chat, validate the superiority of the proposed approach over state-of-the-art models employing various evaluation metrics.

1. Introduction

A chatbot, in simple terms, is a virtual conversation partner powered by Artificial Intelligence (AI) and NLP[1]. Powered by artificial intelligence (AI), chatbots are rapidly becoming ubiquitous across numerous industries. They act as virtual agents, transforming how we interact with various sectors such as customer service and marketing [2, 3], education [4], Deaf people [4-6], banking and insurance [5-7], data collection and management [8], Unified Modeling Language (UML) modeling [9], reading comprehension [10], and health [10, 11]. The key to designing an effective chatbot is generating bot answers that are semantically relevant to the human question [12]. In recent years, with the development of Large Language Models (LLMs), chatbots have made significant advancements in their sophistication. There are several well-known chatbots based on LLMs such as: ChatGPT, Gemini, etc [13]. Each of them has its own strengths and areas of expertise. Alongside their numerous advantages, these types of chatbots have various challenges such as: computational cost especially for real-time chatbot applications and lack of explainability and transparency [14]. In recent years, various AI-based models have been developed that can be divided into three main categories: Retrieval-based, Generative-based, and Hybrid approach.

The retrieval-based method entails selecting the most similar answer from a dataset of predefined bot phrases, using functional scoring metrics [15].

The key to retrieval-based chatbots is question-answer matching. To compute the similarity between the human question and bot answer, some approaches are presented. Lowe et al. [16] proposed a TF-IDF vectors of both the human and bot phrases that are computed by concatenating the TF-IDF scores. The candidate bot answer with the highest cosine similarity to the human question vector is selected as the final answer from the bot. While TF-IDF is simple to use and doesn't require training on a dataset, it lacks accuracy. Recent advancements in computer vision using deep neural networks (DNNs) have inspired researchers to explore similar approaches for NLP tasks like question-answer matching. Models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRUs) have been applied to this problem. Among these, LSTM models have demonstrated superior performance compared to both RNNs and TF-IDF in answer selection [17]. Additionally, Shu et al. [18] proposed a retrieval-based model that combines keyword extraction modules and a two-stage transformer to search for bot answers related to the human question. Retrieval-based chatbots, though widely studied, have significant drawbacks. Unable to create original answers, they are limited to reusing phrases from their stored database. This constraint restricts the chatbot's ability to provide diverse answers and necessitates a substantial database for effective operation.

Generative-based approach learns patterns from training data using deep learning models to generate relevant answer. This approach typically is built on Sequence to Sequence (Seq2Seq) learning models that can lead to increasing the generalization and performance compared to retrieval-based models [19-22]. Lin et al.[23] presented an end-to-end empathetic chatbot called CAiRE. This model utilizes a large-scale pre-trained language model as its foundation, which is then fine-tuned using a combination of objectives. However, the challenge of dealing with the sequential nature of models still persists, which can potentially result in less accurate outcomes. To address this challenge, recent advancements in transformer models [24-28] can be utilized to take advantage of parallel computing and the self-attention mechanism in these models [29]. Zhang et al. [30] introduced powerful model for generating natural language answers in dialogues called Dialogue Generative Pre-trained Transformer (DIALOGPT). Zhou, H., et al. [31] leverage emotion states to train their model, allowing it to capture and express a range of emotions in its responses. The pre-trained model PLATO [32] analyzes conversations using underlying representations to capture their essence. It then utilizes attention mechanisms to combine this extracted information with the inherent qualities of the dialogue. Yanjie Gou et al. [33] proposed the context-aware memory enhanced transformer framework (COMET) as an approach that incorporates knowledge bases into end-to-end task-oriented dialogue systems. While generative-based models can generate new answers, they often struggle with generalization. When faced with inputs or contexts significantly different from their training data, these models may generate inaccurate or irrelevant answers.

To overcome these challenges, some articles explore hybrid approaches. The approach presented in [34] integrates bidirectional recurrent neural networks (BiRNNs) with attention mechanisms and deep reinforcement learning techniques, including deep Q-networks (DQN) and quantile regression deep Q-networks (QR-DQN), to minimize the generation of irrelevant answers. Tran and Le [35] combined BERT2BERT model and RL algorithm to improve chatbot answer generation by exploring bi-directional context. Zhu et al. [36] leveraged the N-best retrieved bot phrases as a basis for computing the reward signal for their generator model. SeqGAN [37] and StepGAN [38] are two notable approaches that combine the Generative Adversarial Network (GAN) framework with RL techniques for sequence generation tasks. In SeqGAN, the RL reward signal, determined by the discriminator based on the complete sequence, is fed back to intermediate state-action steps using Monte Carlo search. However, the Monte Carlo Tree Search (MCTS) employed in SeqGAN suffers from high computational costs. StepGAN approaches evaluate the GAN sequentially. The discriminator is trained to assess and score individual sub-sequences generated by the generator. This allows for automatic evaluation of the suitability and quality of these sequences. In general, these approaches suffer from slow convergence due to high variance and low processing speed. To overcome this limitation, Esfandiari et al. [39] proposed an approach based on Conditional Wasserstein Generative Adversarial Networks (CWGAN) using transformer model which processes parallelly during the training phase. In [40] a new embedding vector is proposed. This paper focused on computationally efficient word embedding method.

While these previous researches have obtained the promising results, the challenges regarding the sequential nature as well as the semantic coherence of conversations still remained that require further investigation.

To enhance the accuracy in sequential models and improve the quality of the generated phrases in human-bot conversations, this paper proposes a novel hybrid approach for chatbots by integrating RL and transformer models. The transformer model has an agent role that we aim to learn it by RL to generate bot answer that is semantically relevant to human question.

The main contributions of this paper can be summarized as follows:

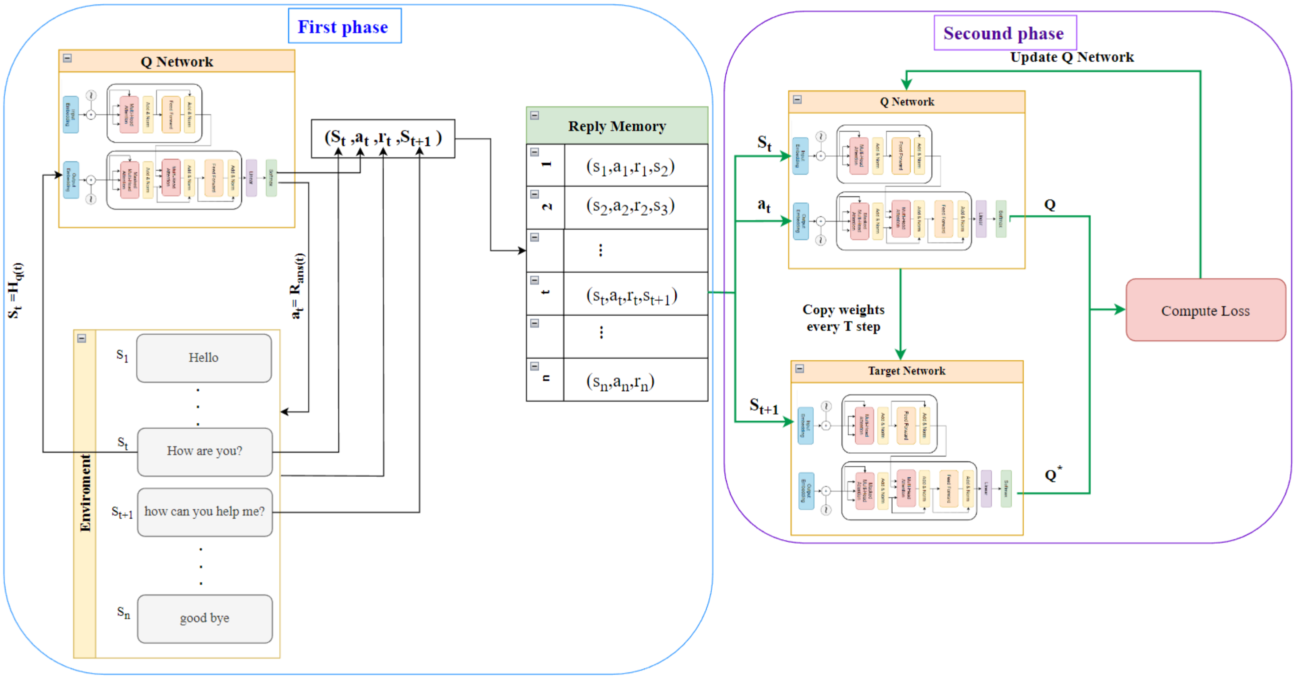

Figure 1. General architecture of proposed approach.

1) Model: A generative chatbot using the embedded transformer model as an agent in RL framework is proposed. DDQN is used as RL method. By leveraging the parallel computing capabilities of the transformer model and its ability to handle long-length dependencies between input sequences, the proposed model successfully improves the quality of the generated phrases in human-bot conversations. To the best of our knowledge, this is the first time that such a model has been proposed in the context of chatbot models.
2) Learning methodology: The proposed model is learned using the DDQN methodology that can assist the model to provide a more semantic bot answer related to human question in conversation.
3) Performance: We conducted experiments on two challenging datasets to evaluate the performance of our approach. The results clearly demonstrate that our model surpasses state-of-the-art methods in terms of performance.

The structure of this paper is as follows: Section 2 presents a detailed description of our proposed approach. In Section 3, we present the results obtained from conducting experiments on two challenging datasets. The findings are further discussed in Section 4. Finally, Section 5 concludes the paper by summarizing the key contributions of our work and highlighting potential avenues for future research in this field.

2. Proposed Approach

In this section, we will provide a comprehensive description of the proposed model, highlighting its key components and their interactions. We propose a generative chatbot using the embedded transformer model as an agent in RL network. DDQN is used as RL method. General architecture of proposed approach is illustrated in Figure 1. The DDQN architecture has three main components: reply memory, Q network, and target network that is trained over multiple time steps across numerous episodes. The Q network is the agent that is trained to produce the optimal state-action value. The target network was used to estimate the Q-values for actions to prevent overestimation bias. The proposed model consists of two phases: preparing reply memory and train agent.

In the first phase, the reply memory interacts with environment to gather data that is used to train the Q network. Replay memory is a technique used to store and randomly sample past experiences, providing diverse training data to improve the Q-network's learning. No training happens during this phase. The environment is a collection of conversations, each containing pairs of questions and answers. In this paper, we treat questions as states and answers as actions. For each observation, the replay memory selects an action from the current state generated by the agent. This action is then returned to the environment, which provides the corresponding reward and next state. The replay memory stores this observation as a sample (current state [S_t], action [a_t], reward[r_t], next state [S_{t+1}]) in the training dataset.

In the proposed approach, the agent is a pre-trained full transformer model that is trained on two separate datasets.

As shown in Figure 2, for each observation, pretrained model processes the question as the environment's state and generates the answer as an action to be stored in the reply memory. For transformer input, the real questions are preprocessed, tokenized, indexed, and embedded using the BERT model. To accelerate training, a linear layer was employed to reduce the feature dimensionality to 64, as shown in Equation (1):

$$y = f\left(\sum\nolimits_{i=0}^{n} w_i x_i + \theta\right) \qquad (1)$$

Once the replay memory is filled with training data, the second phase involves training the Q-network as an agent to estimate the optimal state-action value. Firstly, a random batch of samples from the reply memory is taken, then, it is inputted to both networks. The Q network takes the current state and the action from each data sample, predicting the Q-value for that particular action. Q-value is the maximum output value of the decoder. Also, the target network takes the next state from each data sample and predicts the Q^*-value. Q network and target network are the same and both are transformer model. The target-value is Q^* plus the reward from the sample as shown in Equation (2):

$$\text{target} - \text{value} = r + (\gamma * Q^*) \qquad (2)$$

Where $\gamma \in [0,1]$ is the discount factor and it accounts for the uncertainty in future rewards. γ is usually above 0.9 and smaller than 1. The predicted Q-value and target-value are used to compute the loss to train the Q network in each time-step. For this goal, the Mean Squared Error (MSE) loss using the difference between the target-value and the predicted Q-value is employed as shown in Equation (3):

$$\text{Loss} = MSE(\text{predicted Qvalue, target value}) \qquad (3)$$

Using the back-propagating algorithm, the weights of the Q Network are updated. However, the target network is not trained, leading to not computing the loss. So, the back-propagation is not done.

This completes the processing for this time-step. The processing repeats for the next time-step.

After T time-steps, the Q network weights are copied to the target network. This lets the target network get the improved weights so that it can also predict more accurate Q values.

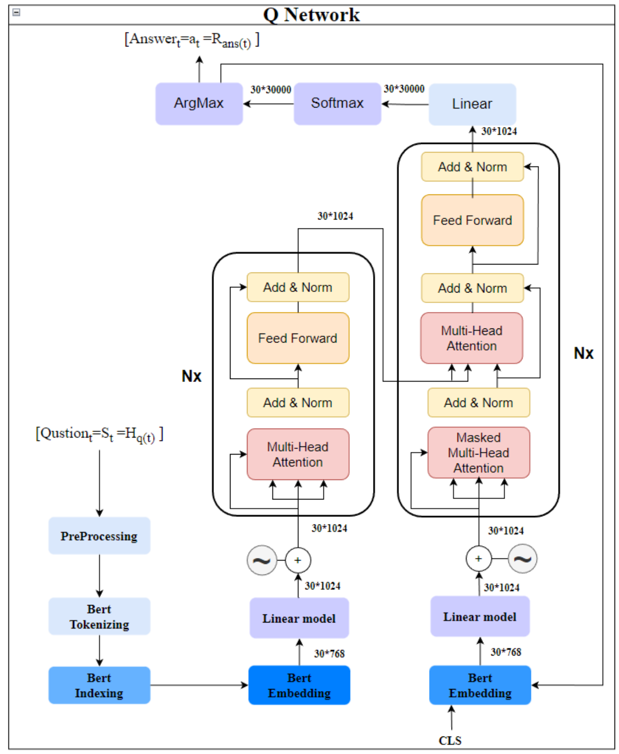

Figure 2. General architecture of pretrained model.

Generally, in RL, it is best to allow some randomness in the action selection at the beginning of the training. This randomness is determined by the epsilon parameter. The epsilon variable usually starts somewhere close to 1, slowly decaying to somewhere around 0 during training. To do this process in the proposed model, random answer was selected in random model. The real answer for the next state is used as the random action. Since the data is shuffled beforehand, this inherently ensures the selected response will be random. Following the training phase, only the agent component is deployed as a chatbot generator. This component receives user questions and generates corresponding bot answers.

We consider the following conditions for RL learning:

Policy π_θ: The policy π_θ is a mapping function $\pi: S \rightarrow A$ where $\pi_\theta(s_t)$ determines the action a_t that can be performed by an agent in state s_t. In this paper, a pretrained transformer model is considered as policy π_θ.

State s_t: Each sample of question in conversation is defined as a state. The agent receives the question $H_q(t)$ as the state s_t from environment.

Action a_t: The action is a generated answer by agent ($R_{ans}(t)$) that is the argmax output of the transformer decoder.

Reward r_t: To calculate the reward, firstly a cross-entropy function is used to measure the distance between the answer generated by the agent (p) and the real answer in the environment (q) as shown in Equation (4):

$$CE(p,q) = -\sum_{t \in T} p(t) \log q(t) \qquad (4)$$

where t is the index vector of tokens in the answer phrase. The output of the cross entropy is normalized in the interval of 0 to 1 by SoftMax function. This function is a mathematical function used to convert a vector of numbers into a probability distribution. This means that the output values will all be between 0 and 1, and their sum will always be 1.

In cross entropy, the output close to 1 indicates more distance and less similarity and vice versa. Whereas, for the reward function, the output closer to 1 indicates less distance and more similarity. This means that cross entropy and reward behave oppositely to each other. To equate these two criteria by considering two thresholds of 0.75 and 0.45, the normalized output of cross entropy is assigned to three reward groups 1, 0 and -1 as follows:

$$\begin{cases} \text{N\_CE}>=0.75 & r=-1 \\ 0.45<\text{N\_CE}<0.75 & r=0 \\ \text{N\_CE}<=0.45 & r=1 \end{cases} \qquad (5)$$

N_CE is a normalized output of cross entropy between p and q.

3. Results

This section aims to present details on the model evaluation. To this end, after explaining the implementation details and working environment, a brief introduction is presented on two datasets used in the model evaluation. After that, the assessment metrics are explained. Finally, the results of the comparative methods are presented.

3.1. Implementation details

The assessments are performed on a Microsoft Windows 11 OS, utilizing python software on hardware featuring an NVIDIA GeForce RTX 3090 and a Core (TM) i5-12600K with 128GB of RAM. The model was implemented using the PyTorch library. Table 1 lists the implementation parameters used in the implementation.

Table 1. Particularity of the parameters employed in the proposed architecture.

Parameters	value	Parameters	value
Learning rate	0.00005	Transformer layers	8
Batch size	32	Dataset split ration for test data	20%
Epoch numbers	100	Sentence Max Length	30
Processing way	GPU	Gamma	0.99
Dropout	0.5	Reply memory size	100000

3.2. Datasets

To evaluate the proposed model, two datasets are utilized: The Daily Dialog and Chit-Chat. The Daily Dialog dataset encompasses a diverse collection of daily conversations, categorized into three primary domains: work (14.49%), ordinary life (28.26%), and relationships (33.33%). Comprising 13,118 multi-turn dialogues, it aims to capture the richness and variety of topics encountered in daily interactions. The second dataset from the Chit-Chat challenge of the BYU perception, control, and cognition laboratory consists of 7,168 conversations involving 1,315 unique participants. These interactions yielded 258,145 utterances, providing valuable insight into everyday conversation.

3.3. Assessment metrics

To assess the efficiency of the proposed model, two evaluation metrics are employed: BiLingual Evaluation Understudy (BLEU) [40] and Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L)[41]. BLEU, originally designed for machine translation evaluation, quantifies the level of agreement between the generated text and reference translations based on n-gram overlap [40]. In contrast, ROUGE-L focuses on recall, measuring the similarity between the generated text and reference summaries by considering the longest common subsequences [41].

3.4. Experimental results

The model's performance is evaluated on The Daily Dialog and Chit-Chat datasets using BLEU and ROUGE-L metrics (Table 2). As these tables show, the proposed model performs better on Chit-Chat dataset. This comes from this point that the Chit-Chat dataset is based on chat and conversation while The Daily Dialog dataset relies on the multi-turn dialogues.

Table 2. Performance of proposed architecture on two.

Dataset	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE
Daily Dialog	0.524	0.478	0.459	0.446	0.562
Chit-Chat	0.976	0.971	0.967	0.965	0.976

Several state-of-the-art models, including Seq2Seq [18], BERT2BERT [42], COHA [31], PLATO [32] and BERT2BERT+RL [35] are used to compare the proposed model on the Daily Dialog dataset (Table 3) and MLE [38], SeqGAN [37], StepGAN [38], and CaiRE [23] on the Chit-Chat dataset (Table 4) with average BLEU and ROUGE metrics.

We selected these models to comparison for the following reasons:

1. **Common Data and Benchmarks:** All models were evaluated using the same dataset and metrics, allowing for a fair comparison.
2. **Published Results:** The selected models have published results that can be cited and referenced.
3. **Similar Methodologies:** The basic approaches of these models are comparable to ours. For instance, both BERT2BERT + RL and our approach utilize a combination of transformer models and RL. However, our approach differs in how the transformer is used within the DDQN network and aims to specifically improve the accuracy of generated responses in sequence models.

Table 3. The comparison results of the proposed model with the different methods on the Daily Dialog dataset.

Approach	Average BLEU	ROUGE
Seq2Seq	0.075	-
BERT2BERT	0.314	0.302
COHA	0.178	-
PLATO	0.438	-
BERT2BERT +RL	0.39	0.375
Proposed	**0.476**	**0.552**

Table 4. The comparison results of the proposed model with the different methods on the Chit-Chat dataset.

Approach	Average BLEU	ROUGE
Seq2Seq	0.258	-
SeqGAN	0.261	-
StepGAN	0.239	-
CAiRE	0.703	-
CWGAN	0.961	0.965
Proposed	**0.969**	**0.976**

For investigating the generated answers of the proposed chatbot, various questions from different domains were asked from the chatbot. Table 5 and Table 6 show the results of the chatbot trained using the Daily Dialog and Chit-Chat dataset, respectively.

Table 5. Some examples of the generated answers using the proposed chatbot on the Daily Dialog dataset.

Human Question: **i will go to bed early tonight**
Real Bot Answer: ***Do you have an airsickness*** **Generated Bot Answer:** ***I you have to hour***
Human Question: **of course, please wait a moment i shall go and get it**
Real Bot Answer: **thank you** **Generated Bot Answer:** **yes you**
Human Question: **let me check yes, we have**
Real Bot Answer: **may i try it on** **Generated Bot Answer:** **i i help it on**
Human Question: **you cannot come any earlier than that**
Real Bot Answer: **i am sorry that has the only available time** **Generated Bot Answer:** **i am sorry i has right same time**

Table 6. Some examples of the generated answers using the proposed chatbot on the Chit-Chat dataset.

Human Question: ***what has your opinion of Hal***
Real Bot Answer: ***i bet we had got along great*** **Generated Bot Answer:** ***we bet we had got along great***
Human Question: ***is there anyone who tells you what to do***
Real Bot Answer: ***i do not report to anyone*** **Generated Bot Answer:** ***i do not report to anyone***
Human Question: ***do you think i look bad today***
Real Bot Answer: ***i do not have a way to know how you look but i really enjoy talking with you*** **Generated Bot Answer:** ***honestly cannot not have a way to know how you look but i really enjoy talking with you***
Human Question: ***are you going to have lunch***
Real Bot Answer: ***i do not need to eat*** **Generated Bot Answer:** ***i do not need to eat***

4. Discussion

One prevalent challenge in contemporary chatbots lies in their sequential processing, which can sometimes lead to suboptimal accuracy.

Our proposed model addresses this limitation by leveraging a hybrid architecture that combines the strengths of RL and transformer models. Based on experimental results, our model is capable of generating bot answers that are more accurate and semantically relevant when compared to human question. In this section, the proposed approach will be discussed from three different perspectives:

- **Training and Stability:** Generally, the evaluation metric in RL is defined as the total reward that the agent periodically collects in an episode during the training phase. While Figures 3 and 4 illustrate the typical behavior of the average total reward metric during training on Daily Dialog and Chit-Chat, it is important to note that this metric can be quite noisy. This noise arises because small adjustments to a policy's weights can significantly alter the distribution of states the policy encounters during training.

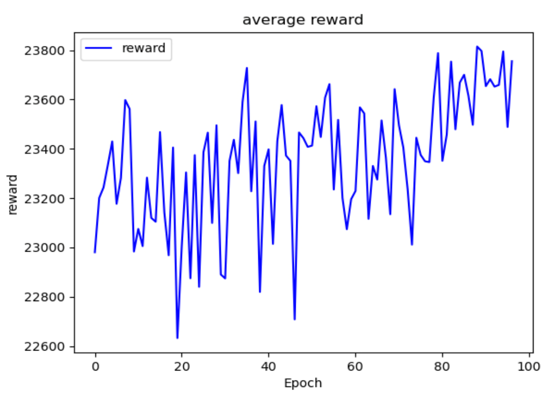

Figure 3. Average total reward on the Daily Dialog dataset.

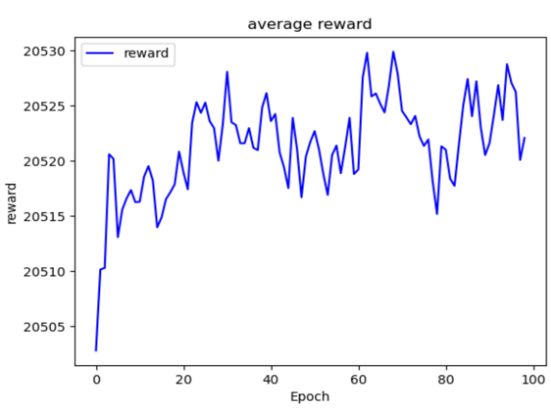

Figure 4. Average total reward on the Chit-Chat dataset.

A more reliable performance measure is the agent's estimated action-value function, denoted as Q. This function approximates the long-term reward an agent can expect by following its chosen policy from any given state.

The average predicted Q-value, as depicted in the Figures 5 and 6, exhibits a much smoother trajectory compared to the average total reward obtained by the agent. Furthermore, during training, we observed a consistently smooth increase in the predicted Q-value without encountering any divergence issues in any of our experiments.

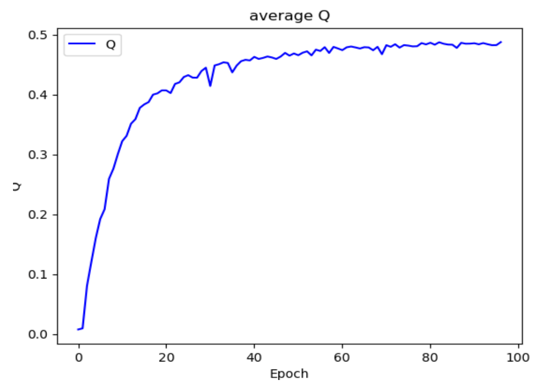

Figure 5. Average Q value on the Daily Dialog dataset.

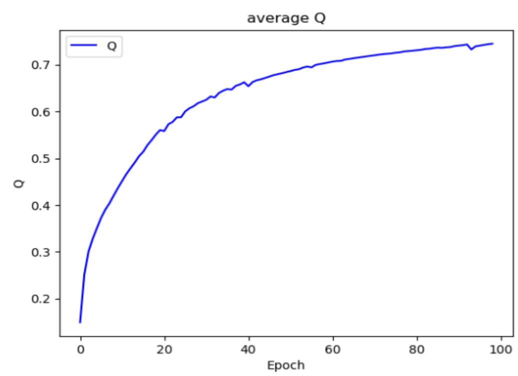

Figure 6. Average Q value on the Chit-Chat dataset.

- **Ablation analysis:** In this section, we delve into the examination of the impact of different parameters in the transformer model in Table 7 (the Daily Dialog dataset) and Table 8 (the Chit-Chat dataset). For this purpose, three distinct categories have been identified. Category A investigates the head size relative to the model, category B examines the number of layers, and category C assesses the dropout rate. According to the base attention paper [29], the head size relative to the model should be set to 64. So, this ratio has been considered.
- **Compared methods analysis:** According to experimental results, the proposed model can generate more accurate and semantically relevant answers for the chatbot dialogue due to the extracting richer features by transformer as well as learning by RL in both datasets.

To demonstrate the superiority of our proposed method over traditional sequence-based models, we compared its performance to Seq2Seq models on both datasets. The baseline Seq2Seq models utilized LSTM units in both the encoder and decoder, along with an attention mechanism. Furthermore, on the Daily Dialog dataset, we compared our approach to the BERT2BERT method. BERT2BERT is a transformer-based model that uses a transformer architecture for both the encoder and decoder. By comparing our method to BERT2BERT, we aimed to demonstrate the advantages of combining transformer models with RL over using transformer models alone.

Table 7: Investigating the impact of transformer parameters for the Daily Dialog dataset.

Model	Type	Head	Layer	Dropout	d_{model}	BLEU4
Transformer (encoder+ decoder)	A	16	8	0.5	1024	**44.6**
		4	8	0.5	512	42.2
		32	8	0.5	2048	43.3
	B	16	4	0.5	1024	42.1
		16	16	0.5	1024	42.9
	C	16	8	0.1	1024	43.9
		16	8	0.7	1024	44.1

Table 8: Investigating the impact of transformer parameters for the Chit-Chat dataset.

Model	Type	Head	Layer	Dropout	d_{model}	BLEU4
Transformer (encoder+ decoder)	A	16	8	0.5	1024	**0.965**
		4	8	0.5	512	0.953
		32	8	0.5	2048	0.961
	B	16	4	0.5	1024	0.951
		16	16	0.5	1024	0.959
	C	16	8	0.1	1024	0.960
		16	8	0.7	1024	0.958

5. Conclusion and Future work

In this paper, a new chatbot model which employs the transformer model as an agent of RL approach has been proposed. The agent in this model is the implemented using a full transformer model. The main objective of our approach was to generate bot answer and improve the semantic efficiency of conversations by learning a mapping between the human question and bot answer using RL. Moreover, this method has enhanced the naturalness of the generated answer and led to more engaging interactions by the transformer. proposed model was evaluated on two datasets using various evaluation metrics. The results confirmed the proposed model's superiority over state-of-the-art approaches based on BLEU and ROUGE-L metrics. Leveraging the capabilities of RL and the transformer model, the proposed model generated accurate, semantically relevant, and human-like answers.

As a future work, large language models can be used to enhance the model in various domains and increase the diversity of the generated answers.

References

[1] G. Caldarini, S. Jaf, and K. McGarry, “A literature survey of recent advances in chatbots,” *Information*, vol. 13, pp. 41, 2022.

[2] A. D. Tran, J. I. Pallant, and L. W. Johnson, “Exploring the impact of chatbots on consumer sentiment and expectations in retail,” *Journal of Retailing and Consumer Services*, vol. 63, pp. 102718, 2021.

[3] C. W. Okonkwo and A. Ade-Ibijola, “Chatbots applications in education: A systematic review” *Computers and Education: Artificial Intelligence*, vol. 2, pp. 100033, 2021.

[4] R. Rastgoo, K. Kiani, and S. Escalera, “Word separation in continuous sign language using isolated signs and post-processing,” *Expert Systems with Applications*, vol. 249, pp. 123695, 2024.

[5] R. Rastgoo, K. Kiani, and S. Escalera, “Sign language recognition: A deep survey,” *Expert Systems with Applications*, vol. 164, pp. 113794, 2021.

[6] R. Rastgoo, K. Kiani, S. Escalera, V. Athitsos, and M. Sabokrou, “A survey on recent advances in sign language production,” *Expert Systems with Applications*, vol. 243, pp. 122846, 2024.

[7] D. Mangla, R. Aggarwal, and M. Maurya, “Measuring perception towards AI-based chatbots in the insurance sector,” in *2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, 2023.

[8] M.-H. Tsai, C.-H. Yang, J.-Y. Chen, S.-C. Kang, “Four-stage framework for implementing a chatbot system in disaster emergency operation data management: A flood disaster management case study,” *KSCE Journal of Civil Engineering*, vol. 25, 2020.

[9] R. Ren, J. W. Castro, A. Santos, O. Dieste and S. T. Acuña, "Using the SOCIO Chatbot for UML Modelling: A Family of Experiments," in IEEE Transactions on Software Engineering, vol. 49, no. 1, pp. 364-383, 1 Jan. 2023,
Oscar; Silvia T.

[10] Sh. Foolad, K. Kiani, and R. Rastgoo, “Recent advances in multi-choice machine reading comprehension: A survey on methods and datasets,” *arXiv:2408.02114*, 2024.

[11] P. I. Prayitno, R. P. Pujo Leksono, F. Chai, R. Aldy and W. Budiharto, "Health Chatbot Using Natural Language Processing for Disease Prediction and Treatment," 2021 1st International Conference on

Computer Science and Artificial Intelligence (ICCSAI), Jakarta, Indonesia, 2021, pp. 62-67.

[12] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, pp. 100006, 2020.

[13] H. Naveed, A.U Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[14] O. Caelen, M.-A Blete, "Developing apps with GPT-4 and ChatGPT," *O'Reilly Media, Inc.*, Aug. 2023.

[15] Y. Zhu, J.-Y Nie, K. Zhou, P. Du, H. Jiang, Z. Dou, "Proactive retrieval-based chatbots based on relevant knowledge and goals," in *SIGIR '21*, pp. 2000–2004, 2021

[16] R. Lowe, N. Pow, I. Serban, J. Pineau, "The Ubuntu Dialogue Corpus: A large dataset for research in unstructured multi-turn dialogue systems," in *SIGDIAL Conference*, pp. 285–294, 2015.

[17] Z. Peng and X. Ma, "A survey on construction and enhancement methods in service chatbots design," *CCF Transactions on Pervasive Computing and Interaction*, vol. 1, no. 3, pp. 204–223, 2019.

[18] C. Shu, Z. Zhang, Y. Chen, J. Xiao, J.H. Lau, Q. Zhang, Z. Lu, "Open Domain Response Generation Guided by Retrieved Conversations," in IEEE Access, vol. 11, pp. 99365-99375, 2023

[19] M. Dhyani and R. Kumar, "An intelligent chatbot using deep learning with bidirectional RNN and attention model," *Materials Today: Proceedings*, vol. 34, pp. 817–824, 2021.

[20] Y. Wang, W. Rong, Y. Ouyang and Z. Xiong, "Augmenting Dialogue Response Generation with Unstructured Textual Knowledge," in IEEE Access, vol. 7, pp. 34954-34963, 2019

[21] Y. Peng, Y. Fang, Z. Xie, G. Zhou, "Topic-enhanced emotional conversation generation with attention mechanism," *Knowledge-Based Systems*, vol. 163, pp. 429–437, 2019.

[22] M. Yang, W. Tu, Q. Qu, Z. Zhao, X. Chen, J. Zhu, "Personalized response generation by dual-learning based domain adaptation," *Neural Networks*, vol. 103, pp. 72–82, 2018.

[23] Z. Lin, P. Xu, G.I. Winata, F.B. Siddique, Z. Liu, J. Shin, P. Fung, "Caire: An end-to-end empathetic chatbot," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09), pp.13622-13623, 2020.

[24] T.-H. Lin, Y.-H. Huang, and A. Putranto, "Intelligent question and answer system for building information modeling and artificial intelligence of things based on the bidirectional encoder representations from transformers model," *Automation in Construction*, vol. 142, pp. 104483, 2022.

[25] A. K. M. Masum, S. Abujar, S. Akter, N. J. Ria and S. A. Hossain, "Transformer Based Bengali Chatbot Using General Knowledge Dataset," 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, 2021, pp. 1235-1238.

[26] B. Peng, M. Galley, P. He, C. Brockett, L. Liden, E. Nouri, Z. Yu, B. Dolan, J. Gao, "GODEL: Large-scale pre-training for goal-directed dialog", *arXiv:2206.11309*, 2022.

[27] T. Shao, Y. Guo, H. Chen and Z. Hao, "Transformer-Based Neural Network for Answer Selection in Question Answering," in IEEE Access, vol. 7, pp. 26146-26156, 2019.

[28] S. Shang, J. Liu and Y. Yang, "Multi-Layer Transformer Aggregation Encoder for Answer Generation," in IEEE Access, vol. 8, pp. 90410-90419, 2020.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need," in *NIPS '17*, pp. 6000–6010, 2017

[30] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, B. Dolan, "DIALOGPT: Large-scale generative pre-training for conversational response generation," in *Annual Meeting of the Association for Computational Linguistics*, pp. 270–278, 2019.

[31] H. Zhou et al H. Zhou, M. Huang, T. Zhang, X. Zhu, B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 730–738, 2018.

[32] S. H. Bao, H. Wang, F. Wu, and H. Wang, "PLATO: Pre-trained dialogue generation model with discrete latent variable," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 85–96, 2022.

[33] Y. Gou, Y. Lei, L.o Liu, Y. Dai, C. Shen, "Contextualize knowledge bases with transformer for end-to-end task-oriented dialogue systems," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[34] R. Keerthana, G. Fathima, and L. Florence, "Evaluating the performance of various deep reinforcement learning algorithms for a conversational chatbot," in *2nd International Conference for Emerging Technology*, pp. 1–8, 2021.

[35] Q.-D. L. Tran and A.-C. Le, "Exploring bi-directional context for improved chatbot response generation using

deep reinforcement learning," *Applied Sciences*, vol. 13, no. 8, pp. 5041, 2023.

[36] Q. Zhu, L. Cui, W.-N. Zhang, F. Wei, T. Liu, "Retrieval-enhanced adversarial training for neural response generation," in *Annual Meeting of the Association for Computational Linguistics*, pp. 3763–3773, 2018.

[37] L. Yu, W. Zhang, J. Wang, Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2852 – 2858, 2017.

[38] Y.-L. Tuan and H.-Y. Lee, "Improving conditional sequence generative adversarial networks by stepwise evaluation," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1346–1355, 2018.

[39] N. Esfandiari, K. Kiani, and R. Rastgoo, "A conditional generative chatbot using transformer model," *arXiv:0234.02074*, 2023.

[40] F. Jafarinejad. "Benefiting from Structured Resources to Present a Computationally Efficient Word Embedding Method", Journal of AI and Data Mining, 10, 4, 2022, 505-514.

[41] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[42] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain, 2004.

[43] C. Chen, "BERT2BERT: Towards reusable pretrained language models," in *Association for Computational Linguistics*, Dublin, Ireland, 2022.

چت‌بات مولد مبتنی بر ترنسفورمر با استفاده از یادگیری تقویتی

نورا اسفندیاری، کوروش کیانی* و راضیه راستگو

دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.

ارسال ۲۰۲۴/۰۵/۰۷؛ بازنگری ۲۰۲۴/۰۸/۱۱؛ پذیرش ۲۰۲۴/۰۹/۱۹

چکیده:

چت‌بات یک سیستم برنامه‌نویسی کامپیوتری است که جهت شبیه سازی مکالمات شبیه انسان و تعامل با کاربران طراحی شده است. چت‌بات، نوعی عامل گفتگویی است که از پردازش زبان طبیعی و مدل‌های متوالی برای درک ورودی کاربر، تفسیر منظور آنها و تولید پاسخ مناسب استفاده می‌کند. هدف این رویکرد تولید کلمات متوالی به شکل عبارات منسجم است. یک چالش قابل توجه مرتبط با مدل‌های قبلی در فرآیند آموزش متوالی آنها نهفته است که می‌تواند منجر به کاهش دقت نتایج شود. برای رفع این محدودیت، با ادغام قدرت دو رویکرد یادگیری تقویتی و مدل های ترانسفورمر، یک چت‌بات جدید مولد پیشنهاد می شود. هدف چت‌بات پیشنهادی، برطرف کردن چالش‌های مرتبط با آموزش متوالی با ترکیب این دو رویکرد می باشد. رویکرد پیشنهادی از یک معماری شبکه Q دوگانه عمیق (DDQN) با استفاده از یک مدل ترنسفورمر به عنوان عامل استفاده می‌کند. این عامل پرسش انسان را به عنوان حالت ورودی (state) می‌گیرد و پاسخ ربات را به عنوان عمل (action) تولید می‌کند. تا جایی که ما می‌دانیم، این اولین باری است که یک چت‌بات مولد با استفاده از یک معماری DDQN با ترنسفورمر جاسازی شده به عنوان عامل پیشنهاد می شود. نتایج بر روی دو مجموعه داده عمومی، Daily Dialog و Chit-Chat، برتری رویکرد پیشنهادی نسبت به مدل‌های موجود با استفاده از معیارهای ارزیابی مختلف را تأیید می‌کنند.

کلمات کلیدی: چت بات، چت بات مولد، مدل ترنسفورمر، یادگیری تقویتی، سیستم مبتنی بر دیالوگ، سیستم مکالمه.