A Course Project Report

On

Deep Reinforcement Learning-Based Chatbot

BY

Abhinav Verma 2020A7PS1093H

Aman Raj Singh 2019B1A31483H

Shantanu Kumar 2019B3A70375H

Vavilala Hrushikesh Reddy 2020A7PS0030H

Sathvik Bhaskarpandit 2019A7TS1200H

Under the supervision of

Dr. Paresh Saxena

for

Reinforcement learning

CS - F317

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

HYDERABAD CAMPUS

(April 2023)

ACKNOWLEDGMENTS

I would like to use this opportunity to express my sincere gratitude to Dr. Paresh Saxena BITS Pilani for his support, invaluable guidance and illuminating views on issues throughout the course of the project.

I would also like to thank Sir for his support, comments and constructive suggestions during the project work.

Finally, I am sincerely grateful to my prestigious institute BITS Pilani, for giving me this opportunity to gain knowledge, and apply my skills through this project.

ABSTRACT

In recent years, chatbots have become increasingly popular due to their ability to provide human-like interactions and automate customer service processes. However, developing a chatbot that can converse naturally and intelligently with humans is a challenging task. Reinforcement learning, a subfield of machine learning, provides a promising approach to address this challenge. In this method, a chatbot is trained through trial and error by interacting with humans and receiving rewards for good conversational responses. This abstract provides an overview of the process of building a chatbot using reinforcement learning. It covers the basics of reinforcement learning, the steps involved in building a chatbot, and some of the challenges faced in the process. The abstract also discusses the benefits of using reinforcement learning, including the ability to learn from experience and adapt to changing environments. Overall, building a chatbot using reinforcement learning has the potential to create more natural and engaging conversational experiences for users.

CONTENTS

Introduction

"Deep Reinforcement Learning for Dialogue Generation" is a research paper authored by Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. The paper explores the use of deep reinforcement learning for generating dialogues between a user and a machine. The authors argue that while dialogue systems have been around for decades, they have not yet achieved human-level performance in terms of generating natural and coherent conversations. To address this, the paper proposes a new approach that leverages deep reinforcement learning techniques to improve dialogue generation. The authors provide a detailed description of their approach and evaluate its effectiveness through a series of experiments. The results show that their approach outperforms existing state-of-the-art methods and represents a significant step forward in the field of dialogue generation. Overall, this paper is a valuable contribution to the ongoing effort to develop more advanced and effective dialogue systems.

Related Work

Here are some related works to the paper "Deep Reinforcement Learning for Dialogue Generation":

1. "A Neural Conversational Model" by Oriol Vinyals and Quoc Le. This paper proposes a sequence-to-sequence model with attention for generating dialogue responses. It utilizes a hierarchical encoder-decoder architecture that can learn to generate diverse and coherent responses.

2. "Seq2Seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models" by Jiacheng Xu et al. This paper introduces a visual debugging tool for sequence-to-sequence models, which can help researchers and practitioners better understand the inner workings of such models.

3. "Neural Responding Machine for Short-Text Conversation" by Lifeng Shang et al. This paper proposes a neural responding machine that uses a sequence-to-sequence model with an attention mechanism to generate short-text responses. It also utilizes reinforcement learning to improve response quality and diversity.

4. "Towards Persona-Based Empathetic Conversational Models" by Emily Dinan et al. This paper proposes a persona-based conversational model that uses a memory network to store information about the user's persona and context. It also uses a reinforcement learning approach to encourage the model to be more empathetic in its responses.

5. "Hierarchical Reinforcement Learning for Open-Domain Dialog" by Xiujun Li et al. This paper proposes a hierarchical reinforcement learning approach for open-domain dialog, which learns to decompose a dialog into a sequence of subtasks and optimize each subtask separately. It also utilizes a knowledge graph to provide additional information to the model.

Deep Reinforcement Learning-Based Chatbot(Original Work)

Research Paper - Deep Reinforcement Learning for Dialogue Generation by Jiwei Li1, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao and Dan Jurafsky

Problem definition:

In recent years, the use of chat technology, including dialogue systems and chatbots, has become increasingly prevalent in human-computer interaction. Chatbots are computer programs designed to simulate conversation with human users, often in natural language. They have been widely adopted in various industries, including customer service, healthcare, and education, as they offer a cost-effective and efficient way to provide personalized assistance and support.However, the current state-of-the-art neural models for dialogue generation tend to be shortsighted, as they predict utterances one at a time but ignore their influence on future predictions. This limitation can lead to incoherent or irrelevant responses that can frustrate users and decrease the effectiveness of chatbots in their respective applications.To address this issue, the proposed solution is to implement a deep reinforcement learning inspired chatbot that can model future reward while generating texts. Reinforcement learning is a type of machine learning that involves training an agent to make decisions by maximizing a reward signal. In the context of chatbots, the reward signal can be defined as the quality of the dialogue, including factors such as coherence, relevance, and naturalness. By modeling future reward, the chatbot can consider the potential outcomes of its dialogue and generate responses that are more coherent and contextually appropriate.

The implementation of a more intelligent and coherent chatbot has potential applications in various industries, such as customer service, healthcare, and education. For example, in customer service, an intelligent chatbot can provide personalized assistance and support to customers, leading to increased customer satisfaction and loyalty. In healthcare, chatbots can assist patients in scheduling appointments, providing medical advice, and monitoring their health, leading to improved healthcare outcomes. In education, chatbots can provide personalized assistance and support to students, leading to improved learning outcomes.Overall, the proposed solution has the potential to improve the performance of chatbots and enhance the user experience in human-computer interaction, leading to more effective and efficient applications in various industries.

Main findings

The paper proposes a novel approach to training dialogue agents using deep reinforcement learning, which allows the agent to consider the long-term consequences of its actions. The authors evaluated their model on a large dataset of human-human dialogues and found that it performs better than existing state-of-the-art methods in generating natural and engaging conversations. They also demonstrated that their model is effective in interactive settings, where it can generate coherent and engaging dialogues with human users. Overall, the paper contributes to the ongoing effort to develop more advanced and effective dialogue systems and highlights the potential of deep reinforcement learning for improving dialogue generation.

Methodology

The research paper "Deep Reinforcement Learning for Dialogue Generation" proposes a novel approach for generating natural and engaging dialogues using deep reinforcement learning. The authors first introduce the problem of dialogue generation and discuss the limitations of existing methods, such as rule-based systems and template-based approaches. They argue that these methods often produce stilted and repetitive responses and fail to capture the nuances of natural language.

To address these challenges, the authors propose a new approach that combines a neural network model with reinforcement learning. The neural network model used in this approach is an encoder-decoder architecture with an attention (Bahdanau et al. 2015) mechanism, which is capable of generating responses that are both contextually relevant and semantically coherent. The encoder takes the input dialogue context and converts it into a fixed-length representation, while the decoder generates the response based on the encoded representation and an attention mechanism that focuses on relevant parts of the context.

The authors use reinforcement learning to optimize the quality of the generated responses. The reinforcement learning algorithm used in this approach is based on the idea of maximizing a reward signal, which evaluates the quality of the generated dialogue responses. The authors use a specific type of reward function, which is based on the similarity between the generated response and a human-generated response. The goal of the reinforcement learning algorithm is to learn a policy that maximizes the expected reward, which is achieved by iteratively generating dialogue responses and adjusting the parameters of the model.

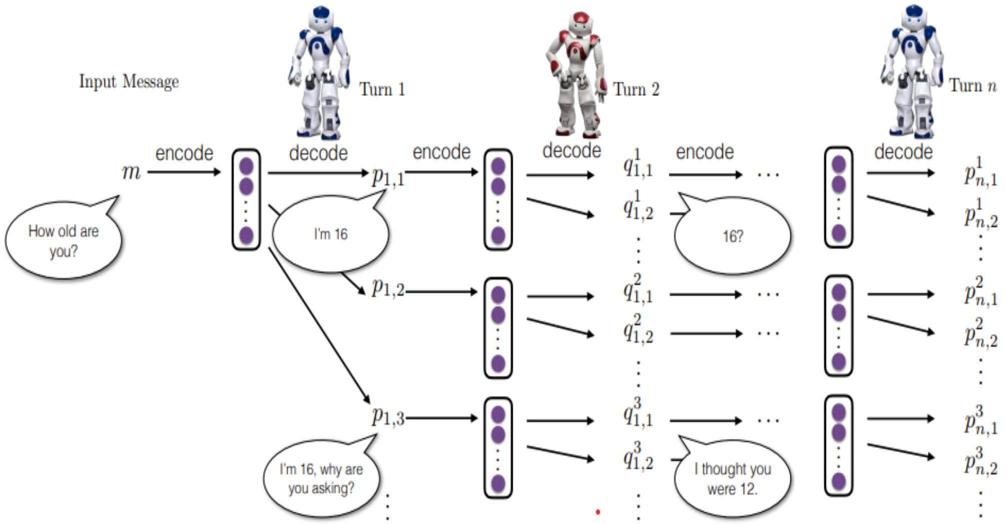


Figure 1: Dialogue simulation between the two agents.

To evaluate the performance of their proposed approach, the authors conduct experiments on several dialogue generation tasks, including chit-chat and task-oriented dialogues. The results show that the proposed approach outperforms existing methods in terms of the quality and diversity of the generated responses. For example, the model is able to generate responses that are more informative and engaging, and less repetitive and generic, compared to other methods.

Overall, the methodology of this research paper involves using a combination of a neural network model with an attention mechanism and a reinforcement learning algorithm to generate natural and engaging dialogues. The approach is evaluated on several dialogue generation tasks, and the results demonstrate its effectiveness in generating high-quality responses that capture the nuances of natural language.

Reinforcement Learning aspect of the original paper:

The learning system simulates dialogues between two agents using policy gradient methods to reward sequences that exhibit three valuable conversational properties: informativeness, coherence, and ease of response. We use p for sentences produced by the first agent and q for those produced by the second.

The agents engage in mutual conversation as they circulate.
A dialogue is represented by a series of alternating sentences produced by the two agents: p1, q1, p2, q2,..., pi, qi. Encoder-decoder neural network paradigm for language.
Using policy search, the network parameters are optimized to maximize the expected future reward. Policy gradient methods are more suitable than Q-learning for our scenario because we can initialize the encoder-decoder RNN with MLE parameters that already generate plausible responses, before altering the objective and tuning towards a policy that maximizes long-term reward.
Q-learning, on the other hand, explicitly estimates the future expected reward of each action, which may differ by orders of magnitude from the MLE objective, thereby deeming MLE parameters inappropriate for initialization.

Action: An action generates a dialogue utterance.
The action space is infinite because sequences of arbitrary length can be generated.

State: The previous two dialogue turns [pi, qi] denote a state. The dialogue history is converted to a vector representation in an LSTM encoder model by feeding in the concatenation of pi and qi.

Policy
The parameters of a policy take the form of an LSTM encoder-decoder (i.e., pRL(pi+1|pi, qi)). Note that instead of a deterministic policy which would produce a

discontinuous objective that is challenging to optimize with gradient-based methods, a stochastic representation of the policy is utilized.

Reward

Reward r represents the reward for each action. In this subsection, we describe how these parameters can be approximated in computable reward functions: Ease of Answering, Information Flow and Semantic Coherence.

Ease of answering reduces the likelihood of generating dull responses by penalizing utterances in S (the list of dull responses). It is measured as negative log likelihood of generating a response from the set S for a dialog. The reward function can be defined as follows:

$$r_1 = -\frac{1}{N_{\mathbb{S}}} \sum_{s \in \mathbb{S}} \frac{1}{N_s} \log p_{\text{seq2seq}}(s|a)$$

$p_{seq2seq}$ represents the likelihood output of the seq2seq model. The reward function

Information Flow makes sure that the conversation keeps moving forward by penalizing any semantic resemblance between successive turns from an agent. The reward function can be defined as follows:

$$r_2 = -\log \cos(h_{p_i}, h_{p_{i+1}}) = -\log \cos \frac{h_{p_i} \cdot h_{p_{i+1}}}{\|h_{p_i}\| \|h_{p_{i+1}}\|}$$

h_{pi} and h_{pi+1} denote encoder representations for two consecutive turns p_i and p_{i+1}

Semantic Coherence penalizes utterances in situations where they are rewarded without being grammatical or coherent. The reward function can be defined as follows:

$$r_3 = \frac{1}{N_a} \log p_{\text{seq2seq}}(a|q_i, p_i) + \frac{1}{N_{q_i}} \log p_{\text{seq2seq}}^{\text{backward}}(q_i|a)$$

To ensure that the responses are coherent and grammatically correct, the mutual information between the given input and the action a are used. The model is also reverse trained on the probability of the input given the current response.

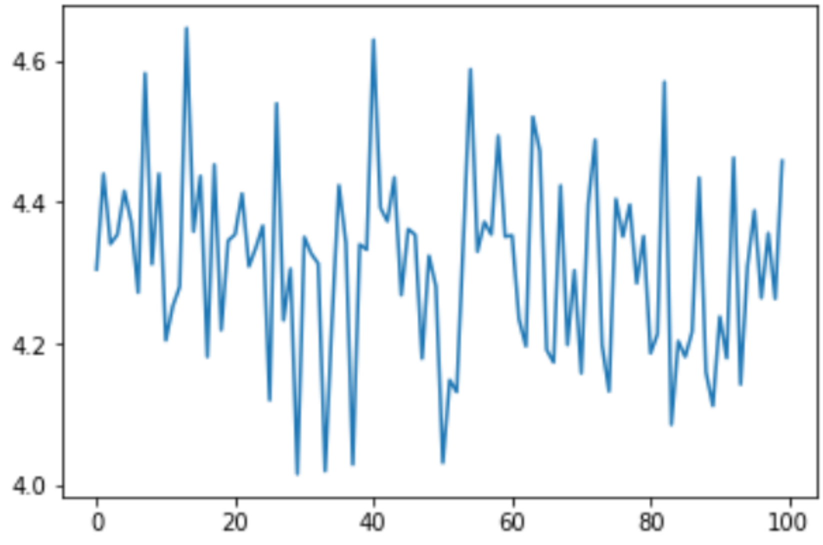
The final reward function for action a is a weighted sum of the above three rewards given by:

$$r(a, [p_i, q_i]) = \lambda_1 r_1 + \lambda_2 r_2 + \lambda_3 r_3$$

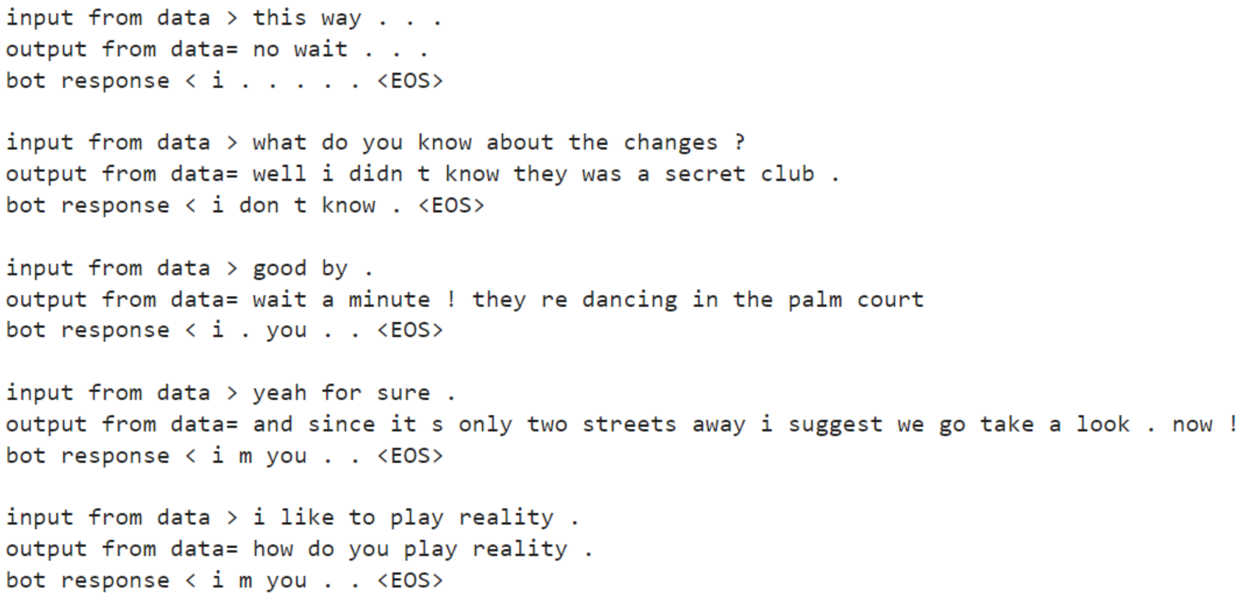
Results from Original paper:

- The original paper used a Seq2Seq model having a unidirectional RNN with attention mechanism
- We trained the same model on Cornell movie dialogue corpus which contains 220,579 conversational exchanges between 10,292 pairs of movie characters.
- Our decision to use this dataset stems from its size and diversity, with a wide range of language formality, time periods, sentiment, and more. We aim to leverage this diversity to improve the robustness of our model, allowing it to handle various forms of inputs and queries effectively.
- During a single training iteration, we applied a couple of techniques to expedite convergence. Firstly, we utilized teacher forcing to act as training wheels for the decoder, leading to more efficient training. However, during inference, this technique may cause instability in the model as the decoder may not have enough opportunities to generate its own output sequences. Secondly, we implemented gradient clipping, which is a commonly used technique to counteract the issue of the "exploding gradient." This approach involves setting a maximum threshold value for gradients, preventing them from growing uncontrollably and leading to overflow (NaN) or overshooting steep cliffs in the cost function.

- The model was trained for 10,000 iterations. The figure below shows average loss over the training.



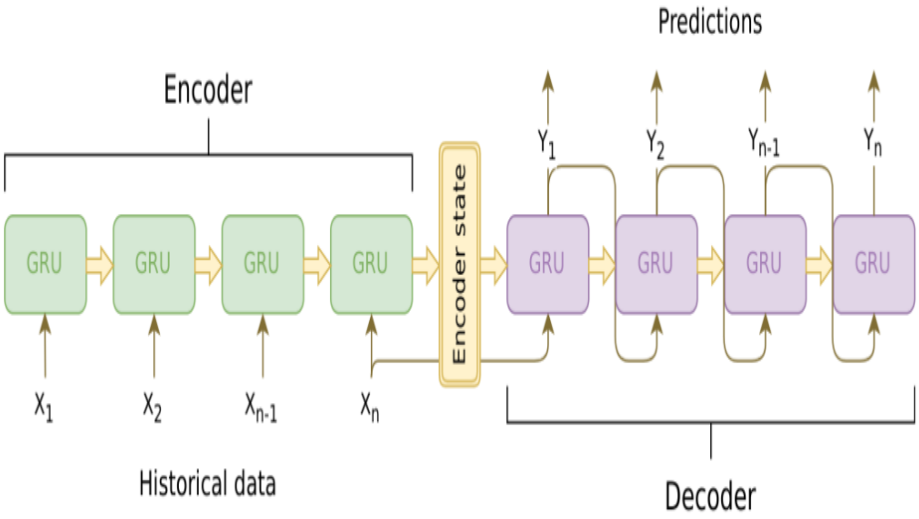
Example conversation from chatbot:



```
input from data > this way . . .
output from data= no wait . . .
bot response < i . . . . . <EOS>

input from data > what do you know about the changes ?
output from data= well i didn t know they was a secret club .
bot response < i don t know . <EOS>

input from data > good by .
output from data= wait a minute ! they re dancing in the palm court
bot response < i . you . . <EOS>

input from data > yeah for sure .
output from data= and since it s only two streets away i suggest we go take a look . now !
bot response < i m you . . <EOS>

input from data > i like to play reality .
output from data= how do you play reality .
bot response < i m you . . <EOS>
```

Inference: Responses generated were very vague and irrelevant, One reason could be due to a small dataset which only contains movie script of a particular type.
In order to improve the responses, we incorporated a few changes in the model.

PROPOSED IMPROVEMENTS

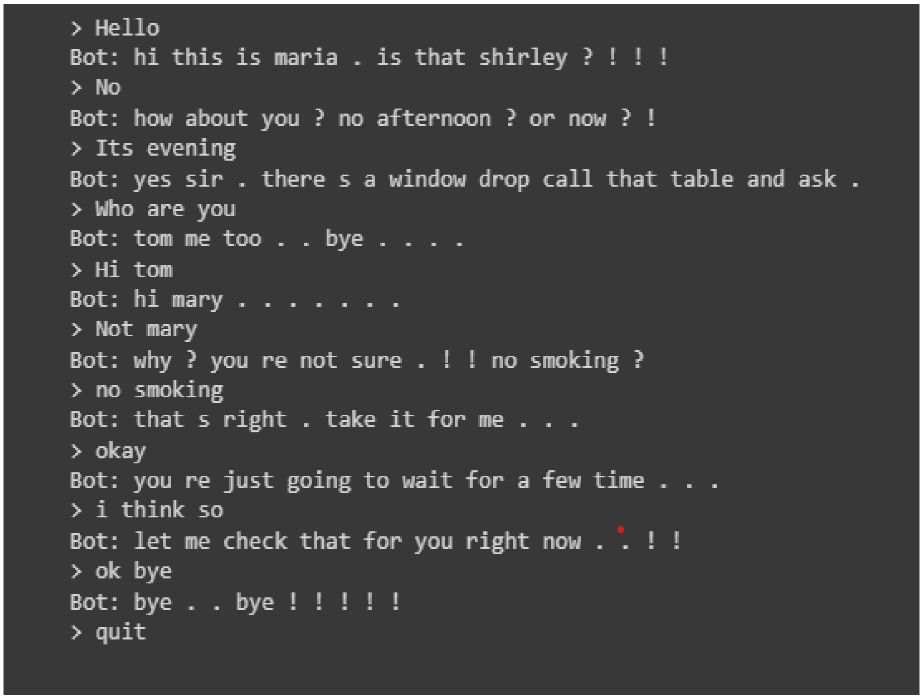
- Using global attention along with Bidirectional RNNs.
  - We use a seq2seq model to accept a variable-length input sequence and return a variable-length output sequence using a fixed-sized model. The diagram below depicts the fundamental structure of our model.
  - 

The encoder RNN iterates over the input sentence token by token (e.g., word), outputting a "output" vector and a "hidden state" vector at each time step. Cho et al. describe the use of a bidirectional multilayered Gated Recurrent Unit. The primary reason for this is that it allows us to encode both past and future contexts.

The method used by the decoder RNN to generate the response sentence is token-by-token, utilizing both the encoder's context vectors and internal hidden states to produce the next word in the sequence. However, relying solely on the context vector to encode the input sequence's complete meaning can lead to information loss, particularly when handling long input sequences. To address this issue, the Bahdanau attention mechanism was introduced. However, our model employs the global attention mechanism introduced by Luong et al. This method differs from Bahdanau's local attention in that it considers all of the encoder's hidden states instead of just the one from the current time step. Additionally, global attention only uses the decoder's hidden state from the current time step to calculate attention weights, whereas Bahdanau's attention requires knowledge of the decoder's state from the previous time step.

- The improved model was trained on 30k iterations(without RL) and 10k iterations(with RL). The results obtained from the model with RL gave much better results even after being trained for less epochs

Without RL

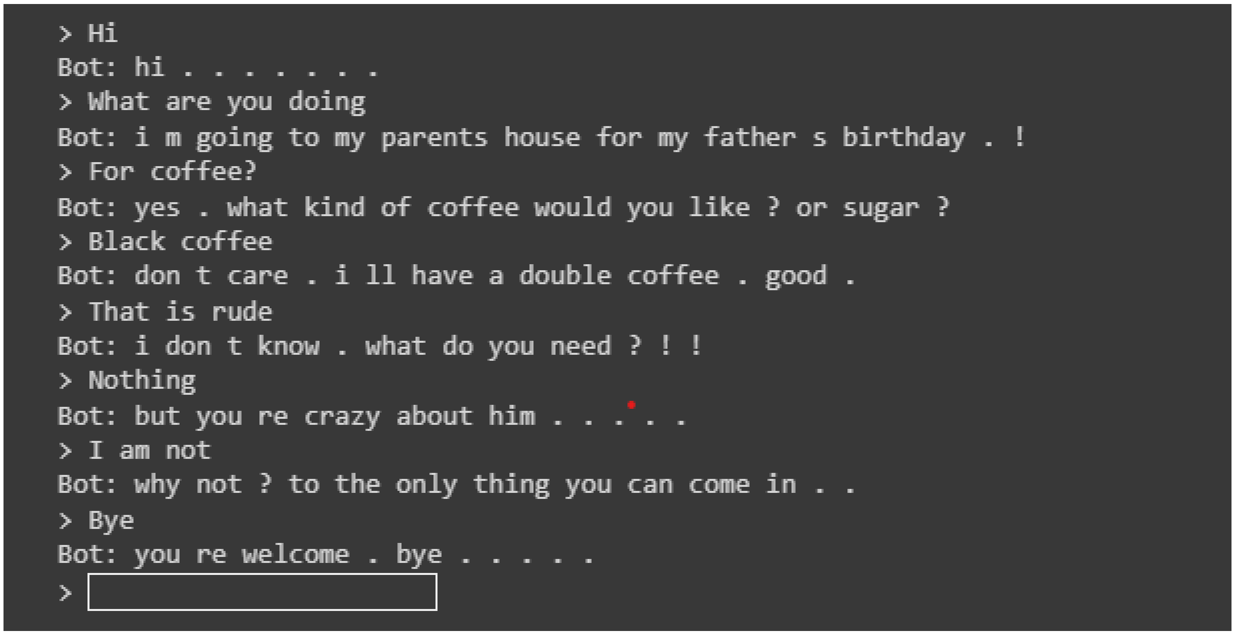


```
> Hello
Bot: hi this is maria . is that shirley ? ! ! !
> No
Bot: how about you ? no afternoon ? or now ? !
> Its evening
Bot: yes sir . there s a window drop call that table and ask .
> Who are you
Bot: tom me too . . bye . . . .
> Hi tom
Bot: hi mary . . . . . . .
> Not mary
Bot: why ? you re not sure . ! ! no smoking ?
> no smoking
Bot: that s right . take it for me . . .
> okay
Bot: you re just going to wait for a few time . . .
> i think so
Bot: let me check that for you right now . . ! !
> ok bye
Bot: bye . . bye ! ! ! ! !
> quit
```

We found that the results had improved as compared to the original settings(even with RL). The data was more meaningful and made more sense.

With RL

- We found the results had improved even more compared to without RL and Basis settings.
- 10 episodes could be generated in one iteration
- If the response is less than a particular length/repetitive, we break the iteration in that case.
- Final reward = (0.25*ease of answering + 0.25*Information flow + 0.50*Semantic coherence),



```
> Hi
Bot: hi . . . . . . .
> What are you doing
Bot: i m going to my parents house for my father s birthday . !
> For coffee?
Bot: yes . what kind of coffee would you like ? or sugar ?
> Black coffee
Bot: don t care . i ll have a double coffee . good .
> That is rude
Bot: i don t know . what do you need ? ! !
> Nothing
Bot: but you re crazy about him . . . . .
> I am not
Bot: why not ? to the only thing you can come in . .
> Bye
Bot: you re welcome . bye . . . . .
>
```

Conclusion

Our proposed approach involves utilizing a reinforcement learning framework to generate neural responses in dialogues by having two agents simulate a conversation. This approach combines the benefits of both neural SEQ2SEQ systems and reinforcement learning for dialogues. We extend on the work done in the paper by introducing a bi-directional RNN using global attention instead of local attention. We are able to generate better responses which are more semantically coherent and are not dull.

References

Deep Reinforcement Learning for Dialogue Generation
Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, Dan Jurafsky

Neural Machine Translation by Jointly Learning to Align and Translate
Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio

Effective Approaches to Attention-based Neural Machine Translation
Minh-Thang Luong, Hieu Pham, Christopher D. Manning

Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs
Cristian Danescu-Niculescu-Mizil, Lillian Lee