Faculty of Computer Science and Information Technology

Improving Attentive Sequence-to-Sequence Generative-Based Chatbot Model Using Deep Neural Network Approach

Wan Solehah Binti Wan Ahmad

Master of Science
2022

Improving Attentive Sequence-to-Sequence Generative-Based Chatbot Model Using Deep Neural Network Approach

Wan Solehah Binti Wan Ahmad

A thesis submitted

In fulfillment of the requirements for the degree of Master of Science

(Computer Science)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2022

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

……

Signature

Name: Wan Solehah Binti Wan Ahmad

Matric No.: 19020167

Faculty of Computer Science and Information Technology

Universiti Malaysia Sarawak

Date : 3/8/2022

ACKNOWLEDGEMENT

First and foremost, all praises and greatest thanks to Allah s.w.t for giving me the strength and determination to complete my study while I am going through a difficult phase during this research journey. I also would like to express my gratitude to my supervisor, Associate Professor Ts. Dr. Mohamad Nazim Bin Jambli for giving me all the guidance, advices, and support in my journey to complete this research. The same goes to all people who have indirectly contributed in the way of discussion and knowledge exchanges. Furthermore, I would like to thank Universiti Malaysia Sarawak (UNIMAS) for their financial support under Zamalah Graduate Scholarship.

Last but not least, I would like to dedicate this special thanks to my parents and my family, who have become my backbone, giving me utmost moral support when things get tough and really hard for me to continue this study. I'm grateful to them, especially to my sisters, because of their never-ending support and encouragement, I had a speedy recovery and was finally able to write up my thesis to completion.

ABSTRACT

Deep Neural Network (DNN) is a combination method between two different subfields of Machine Learning application, including the Artificial Neural Network (ANN) and Deep Learning (DL). An example of the DNN model is the Attentive Sequence-to-Sequence (Seq2Seq) model that was first created to tackle a problem setting in language processing. One of the applications is the chatbot model that works explicitly to accurately respond to users' inquiries. Through the years, a chatbot application has seen some improvement, from generating hard-generic responses to more flexible response. The adoption of DNN method into chatbot application produces a new generation chatbot that called as Generative-Based Chatbot. However, it is difficult to create and train a Generative-Based chatbot model that can maintain relevancy of dialogue generation in a long conversation. Hence, this research's objective aimed to propose an optimization strategy based on Structural Modification and Optimizing Training Network for improving the lacking of accuracy of response in the chatbot application, to propose the algorithm enhancement to improve the current attention mechanism in the Attentive Sequence-to-Sequence model and the network's training optimization of its inability to memorize the dialogue history, and lastly, to evaluate the accuracy of response of the proposed solution through data training on loss function and real data testing. The structural modification that is based on a slight modification in Additive Attention mechanism. The method is by adding a scaling factor for the dimension of the decoder hidden state. The other one is the network training's environment optimization that is done through hyperparameter optimization by selecting and fine-tuning high impact parameters which include Optimizer, Learning Rate and Dropout to reduce error rate (loss function). The strategies applied showed that the final accuracy obtained through the training after implementing a modification in the algorithm is at 81% accuracy rate compared to the

basic model that recorded its final accuracy at 79% accuracy rate. Meanwhile, after modification and optimization, the model's performance recorded its final value of accuracy and loss rate at 87% and 0.51, respectively. The result indicates the performance of the optimized model outperforms the baseline model.

Keywords: DNN, Generative-Based Chatbot, Attentive Seq2Seq model, structural modification, hyperparameter optimization

Penambahbaikan Model Attentive Seq2Seq bagi Generative-based Chatbot Menggunakan Pendekatan Deep Neural Network

ABSTRAK

Rangkaian Neural Dalaman (DNN) adalah kaedah gabungan antara dua sub-bidang aplikasi Machine Learning yang berbeza, termasuk Artificial Neural Network (ANN), dan Deep Learning (DL). Contohnya, model DNN adalah model Attentive Sequence-to-Sequence (Seq2Seq) yang pertama kali dibuat untuk mengatasi masalah dalam aplikasi pemprosesan bahasa. Salah satu aplikasi tersebut adalah model chatbot yang secara khusus berfungsi untuk melaksanakan tugas memberikan respon terhadap pertanyaan dari pengguna dengan tepat. Selama bertahun-tahun, aplikasi chatbot telah melihat beberapa peningkatan dari segi kelancaran memberi respon daripada menghasilkan respon yang statik kepada respon yang lebih fleksibel. Penerapan kaedah DNN ke dalam aplikasi chatbot menghasilkan chatbot generasi baru yang disebut sebagai Generative-based Chatbot. Walaubagaimanapun, sukar untuk mencipta dan melatih model chatbot Generative-based yang boleh mengekalkan respon yang relevan dalam perbualan yang panjang. Oleh itu, objektif penyelidikan ini bertujuan untuk mencadangkan strategi pengoptimuman berdasarkan Pengubahsuaian Struktur model dan Pengoptimuman Rangkaian Latihan untuk menambah baik kekurangan ketepatan respon aplikasi chatbot, untuk mencadangkan penambahbaikan algoritma model Attentive Sequence-to-Sequence dan latihan rangkaian pengoptimuman ke atas ketidakupayaannya untuk menghafal sejarah dialog, serta akhir sekali, untuk menilai ketepatan respon melalui latihan data yang diukur dengan Loss Function dan ujian data sebenar. Pengubahsuaian struktur adalah berdasarkan pada sedikit pengubahsuaian dalam mekanisme Additive Attention dengan menambahkan faktor penskalaan untuk dimensi jaringan tersembunyi di dalam dekoder, dan pengoptimuman

lingkungan latihan jaringan yang dilakukan melalui pengoptimuman hiperparameter dengan memilih dan menyesuaikan parameter berimpak tinggi merangkumi Pengoptimum, Kadar Pembelajaran dan Penurunan untuk tujuan mengurangkan kadar ralat (fungsi pengurangan kadar ralat). Strategi yang diterapkan menunjukkan ketepatan akhir yang diperoleh setelah melaksanakan pengubahsuaian dalam algoritma pada kadar ketepatan 81% berbanding model asas yang mencatat ketepatan terakhirnya pada kadar ketepatan 79%. Sementara itu, prestasi model setelah pengubahsuaian dan pengoptimuman mencatatkan nilai akhir ketepatan dan kadar kerugian masing-masing pada 87% dan 0.51. Hasilnya menunjukkan prestasi model yang dioptimumkan mengatasi model yang sedia ada.

Kata kunci: *DNN, Generative-based Chatbot, model Attentive Seq2Seq, modifikasi struktur, pengoptimuman hiperparameter*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AdaGrad	Adaptive Gradient Algorithm
Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
ANN	Artificial Neural Network
DL	Deep Learning
DNN	Deep Neural Network
EA	Enhanced Attention
FP	False Positive
RNN	Recurrent Neural Network
LR	Learning Rate
NAG	Natural Answer Generation
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
NMT	Neural Machine Translation
RMSProp	Root Mean Square Propagation
Seq2Seq	Sequence-to-Sequence
TP	True Positive

CHAPTER 1

INTRODUCTION

1.1 Research Background

In recent years, especially in the pandemic era, the trend of chatbot application usage has seen incredible growth. According to a study in 2016 by Aspect Software Research, 44% of consumers said they would prefer to interact with a chatbot over a human customer service representative, (Sweezy, 2018), and the numbers are expected to lead its way in coming years.

While it may be a scary coincidence, the chatbot's topic trends increase in these two years due to the Covid-19 pandemic. As if, it becomes essential in the industry as the pressure in maintaining the services while requiring minimal involvement of live interaction between humans is increased significantly in customer services. For example, when the government of Malaysia announced its first movement control order in March 2020, all flights were banned from flying except the essential one. The aviation industry is struggling to handle hectic changes in flight schedules when mass inquiries and complaints come from upset customers.

The situation worsens when the workloads become increased as the workers' capacity at one time is reduced. This example only highlights the situation in the aviation industry, but in reality, every industry suffers from this problem in terms of customer services or management. Every industry needs to develop a solution to automate almost everything by using a third-party app that can be an agent between the parties that gives the services and receives the services so that 'a new norm practices' is convenient for everyone.

Evidently, a chatbot application becomes very much aligned with the solution needed from the industry. But building an intelligent chatbot that can perform efficiently according to user's demands, however, is quite challenging as it requires an algorithm that works effectively in understanding the context of the inputs from the users, text entailment, and language understanding (Wu et al., 2018). So, to achieve a satisfactory functional chatbot, one must understand well the architecture of the chatbot to find the loophole that exists in the current chatbots.

Before going in deeper, it is important to comprehend the definition of chatbot application and how it evolves from traditional chatter bot to deep neural network chatbot. First of all, a chatbot can be classified as an application that builds under the machine learning field, which can be said as the new interest of study in these recent years. It is considered as a Human-Computer Interaction (HCI) model and an artificial intelligence programme (Bansal & Khan, 2018) that simulates a conversation with human users.

The application involves users' intent, the input-output processing, and the response generation method that retrieves the information from the knowledge base. Figure 1.1 shows the most basic architecture of the chatbot model.

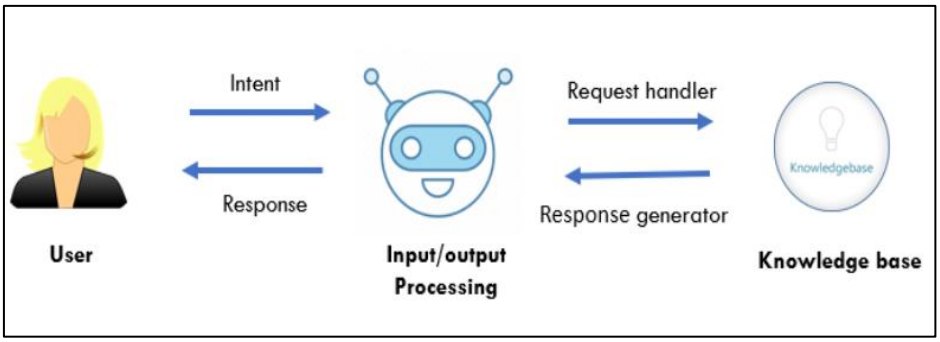

Figure 1.1: Basic architecture of chatbot model

In particular, most of the chatbot development emphasizes the part of the response generation method, and it has two types of methods that indirectly classify the types of chatbot. The first one is categorized as rule-based, and the other one is artificial intelligent-based. Figure 1.2 presents the response generation method's types.

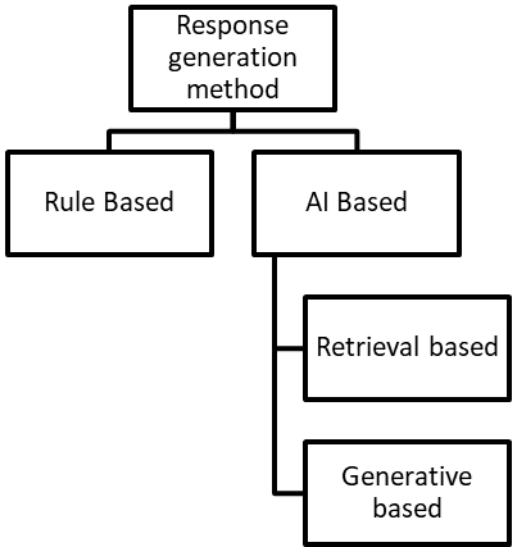

Figure 1.2: Type of response generation model

In the earlier years, a chatbot does not have the capability to generate a human-like response as it lacks language processing knowledge. During those years, the response generation method from the chatbot called a finite state or rule-based, generated hard-generic responses. The method uses a pattern matching technique that is widely derived from Artificial Intelligence Markup Language (AIML) (Trivedi et al., 2019). AIML is considered a knowledge category that consists of input and template that contains answers of chatbot (Ranoliya et al., 2017). Figure 1.3 below shows the example of the AIML tag.

```
<category>
  <pattern> HELLO
  </pattern>
  <template>
    Hi, user!!!!
  </template>
</category>
```

Figure 1.3: Example of AIML tag

As the technology evolves with the introduction of Machine Learning (ML) algorithms, the chatbot application has seen an improvement by manipulating language processing. The development of the Natural Language Processing (NLP) method used in chatbots in later years has resulted in a better set of algorithms that successfully changes from a finite state machine response to a retrieval-based response generation method. The illustration of how retrieval-based chatbot works from the application of NLP techniques are shown in Table 1.1 below.

Table 1.1: Application of NLP technique

No	Question	Answer	Entity
1.	What is your name?	My name is Jason.	name
2.	What do you want to eat?	I want to eat fried rice.	eat

The NLP works by matching the predefined entity in the input from users with the predefined answers in the database. A special classification of ‘entity’ makes the chatbot responses more flexible than the finite-state method.

In the recent development of chatbots, most of the current chatbots use Artificial Neural Network (ANN) method combined with a Deep Learning environment as ANN has been proved can generate an action that can imitate a human being. In other words, ANN becomes a ‘brain’ for the chatbot application to learn how to generate a proper response. An example of a popular model that adopts the ANN and Deep Learning method is the Sequence-to-Sequence model. It is widely used in machine translation and chatbot application that involves a language processing.

The existence of ANN and Deep Learning fields then combined with Natural Language Processing (NLP) technique produces a new generation chatbot called a Generative-Based Chatbot. An illustration of the generative based type is present below in Figure 1.4.

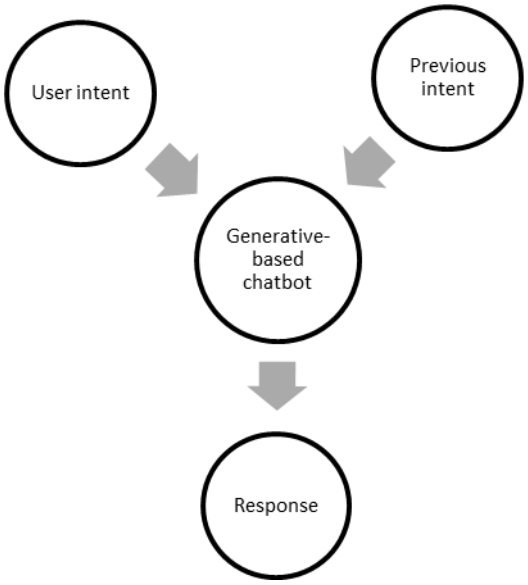

Figure 1.4: Generative-Based Chatbot

A Generative-Based Chatbot is considered as an ultimate high-functional chatbot that can generate a dynamic response without bounding with a hard borderline of domain knowledge by leveraging the Natural Language Generation (NLG) technique (Shang et al., 2015). However, the researchers admit that creating and training such a model is difficult because it requires a large amount of training data to create a successful conversation (Hien et al., 2018). Often in a long conversation, the chatbot start to loss the relevancy in dialogue generation (Sojasingarayar, 2020).

1.2 Problem Statement

As mentioned in the introduction, it is difficult to create and train a Generative-Based chatbot model. Most of the current chatbots applied the existing technology that used different areas and models to produce an almost-like generative chatbot. For example, the hybrid methods that evaluate the retrieved to generated response and select the most effective one (Song et al., 2018). By any means, the current chatbot still has the limitations and can be described as below:

i. Most existing chatbot models have a bad tendency to generate a generic or inconsistent answer. This is because the current chatbot still heavily rely on the rules that are predefined in the algorithm technique (Nuruzzaman et al., 2018). In general, a chatbot has their own domain knowledge that usually function as a database that store the predefined of question-answer pair. But, even when a huge database is used, there is usually a restriction for the chatbot to recognize the entity needed in their domain knowledge.

ii. Chatbot model do not perform well on the longer sentences, as the information got lost in the process, resulting in low response accuracy

(Sojasingarayar et al., 2020). In the deep learning field, there is a sub-area that specifically help the application to perceive and adjust the environment it needs to find the focal point accordingly. The sub-area is called an attention mechanism. Attention mechanism becomes a vital tool in chatbot application, as the nature of sentences can have many different meanings according to their focal point. As the current chatbots still rely on predefined rules to identify the entity in the input and the database, the attention mechanism will take part in the process by identifying the context of the sentence in finding the underlying meaning based on the sequence of characters and words in the input sentence. But, there is still a limitation in current model that can only focus on the short distance of word's arrangement in a sentence due to an inability to memorize the dialogue history (Caldarini et al., 2022).

iii. The capacity of the model's network in training algorithm is not optimum for chatbot training as it still cannot maximize the performance of the model (Elgeldawi et al., 2021). In building a neural network chatbot, a deep learning environment is needed to provide a relationship between the neural response and stimulus that exist in the neural networks. The environment is controlled by hyperparameters in which have the functions to determine the networks' structure and how the networks are trained (Elgeldawi et al., 2021). Every hyperparameter needs to be set before training, and it is important to set the variables with appropriate values. As the environment can vary, how the hyperparameter reacts with the environment will also set a different optimum value. The current chatbot application mostly fails to recognize the optimum capacity of the network environment due to its complex nature.

1.3 Motivation

Several studies have been conducted studying the performance of the response generation method and natural answer generation (NAG) in chatbots. In a recent review done on a topic of Enhancements to the Sequence-to-Sequence-Based Natural Answer Generation Models by Palasundram et al. (2020), it highlights a lot of topic area that is still needed to be improved on various factors. The factors such as insufficient information in questions addressed in another works by Li et al. (2016), Zhou et al. (2018), and Ashgar et al. (2018). The exposure bias's factor addressed by Mou et al. (2016) and Bahdanau et al. (2017). The cross-entropy loss function's factor addressed by Jiang et al. (2019) and Wei & Zhang (2019). Lastly the long sentences' factor emphasized by Shang et al. (2015). The proposed enhancement from various studies that include implementing attention mechanism, Tao et al. (2018), additional encoders, Xing et al. (2017), and beam search, Shao et al. (2017) also stated there. However, despite all of these studies, there is still room for improvement from the current models that are proposed by the researchers. Motivated by the cited works, this study aims to unlock the full potential of chatbot applications by doing further research and advancement in a certain area of neural network chatbot.

Throughout the research, a lot of existing works linked to the DNN chatbot model are reviewed that lead to the conclusion that the DNN chatbot model still can be improved. Hence, the very first idea of this research is to adopt an improvement strategy to the existing model, as in any way, it can open up to a lot of potential of another enhancement method. Inspired from many reviewed works, it is decided that the enhancement in this research can be categorized into two categories, including the structural modification and the optimization of the network training's environment, as both can contribute a significant improvement to the model.

1.4 Research Objectives

The main objective of this research is to improve the chatbot functional capacity by optimizing the capability of the chatbot model in capturing the focus of user intent in the query sentence and optimizing the model network training's capacity. To achieve this, there are specifics objective need to be done as defined below:

i. To propose an optimization strategy based on Structural Modification and Optimizing Training Network for improving the lacking of accuracy of response in the chatbot application.

ii. To propose the algorithm enhancement to improve the current attention mechanism in the Attentive Sequence-to-Sequence model and the network's training optimization because of its inability to memorize the dialogue history.

iii. To evaluate the accuracy of response of the proposed solution through data training on loss function and real data testing.

1.5 Research Scope

This study focused on a specific neural network chatbot model, which is a Sequence-to-Sequence (Seq2Seq) which was originally proposed for natural machine translation (NMT), (Elbayad, 2020). Chatbot application, also known as a conversational agent, operates via language processing. Thus, the Seq2Seq model would be the best benchmarking model in evaluating future chatbot enhancement. In addition, the scope of enhancement focused on attention mechanisms specifically involving Additive Attention and Scaled-dot product Attention. The attention mechanism is considered as one of the important