HEALTHCARE QUESTION AND ANSWER SYSTEM BASED ON SEQUENCE TO SEQUENCE MODEL

EI ZIN PHYO

M.C.Sc. **DECEMBER 2022**

HEALTHCARE QUESTION AND ANSWER SYSTEM BASED ON SEQUENCE TO SEQUENCE MODEL

BY

EI ZIN PHYO

B.C.Sc.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Computer Science

(M.C.Sc.)

UNIVERSITY OF COMPUTER STUDIES, YANGON

DECEMBER 2022

ACKNOWLWDGEMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis. To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to express my deepest gratitude and my thanks to **Dr. Mie Mie Khin**, Rector, University of Computer Studies, Yangon, for her kind permission to submit this thesis.

I would like to express my appreciation to **Dr. Si Si Mar Win**, Professor, Course Coordinators (M.C.Sc.(Thesis)/ M.I.Sc.(Thesis)), Faculty of Computer Science of the University of Computer Studies, Yangon, for their superior suggestion, administrative supports and encouragement during my academic study.

My thanks and regards go to my supervisor, **Dr.Tin Zar Thaw**, Professor, Faculty of Information Science, University of Computer Studies, Yangon, for her support, guidance, supervision, patience and encouragement during the period of study towards completion of this thesis.

I also wish to express my deepest gratitude to, **Daw Aye Aye Khine**, Associate Professor, Department of English, University of Computer Studies, Yangon, for her editing this thesis from the language point of view.

Moreover, I would like to extend my thanks to all my teachers who taught me throughout the master's degree course and my friends for their cooperation.

Last but not least, I especially thank my parents, all of my colleagues, and friends for their encouragement and help during my thesis.

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

---------- ----------

Date Ei Zin Phyo

ABSTRACT

The healthcare sector is one of the most important domains that impacts the entire global population and is closely linked to the development of any country. There are millions and billion pieces of healthcare information available, but making the right information accessible when needed is very important. The advent of Question Answering System has been applied as a promising solution and an efficient approach for retrieving significant healthcare information easily and time saving. The deep learning algorithms are used to train the data and bring output within a specific range by using statistical analysis. Recurrent Neural Network (RNN) based Sequence-to-sequence (Seq2Seq) model is one of the most commonly researched model to implement artificial intelligence question answering system. This system has been implemented using neural network where it use bidirectional RNN as encoder and Luong Attention RNN as decoder. The system presented a healthcare question-answering system to provide healthcare information based on three attention sequence to sequence models: general, dot and concat with three sub datasets: NCI, NIHSeniorHealth and NHLBI in MedQuAD.

Keywords: Healthcare dataset, recurrent neural network, sequence to sequence model, bidirectional RNN, Luong Attention RNN

CONTENTS

LIST OF FIGURES

FIGURE **PAGE**

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

The nation of economy and people of lives both rely heavily on medical services. Despite the fact that there are more healthcare facilities than ever before, access issues and congestion in the patient flow remain serious concerns. Particularly in providing healthcare services for rural and distant areas, is important because the healthcare situation and focusing on this realistic issue in developing countries. Health is something not to control. Consequently, people have not visited the hospital because of a minor issue. Therefore, question and answer system is advancing rapidly around the world. Conversational user interface is a software that runs simplify and structurally.

With the help of Question and Answer System, patients can connect with one another quickly. And when appropriately applied, they can assist healthcare professionals in exceeding patient expectations and enhancing patient outcomes. However, AI solutions can lack the human touch, which is crucial for providing high-quality healthcare. Due to the shift away from in-person encounters to digital settings, digital health platforms, particularly those focusing on mental health issues, have experienced significant growth in recent years. Platforms that allow users to ask and receive answers about their mental health are becoming more and more common online alternatives for those who cannot receive therapy from a human physician. These platforms' effectiveness has been repeatedly questioned by critics, due to the lack of face to face connection and empathy between patient and clinician. Due to technologies like AI, ML, and NLP, it is said that question and answer system has reached a level where they can gauge human sentiments. The uniqueness in every individual's behavior can confuse question and answer system. Some people could enjoy an informal discourse, while others would prefer a formal one.

The use of question and answer system has spread from consumer customer service to matters of life and death. The question and answer system has entered the healthcare industry and can help solve many of its problems. So, healthcare facilities in general are a crucial resource for poor nations, but they are frequently challenging to set up due to a lack of awareness and infrastructure development. Many internet users rely on it to find

answers to their questions about healthcare. The user can more easily obtain medical advice and become familiar with the various disorders and available diagnoses. They have implemented a number of strategies to improve communication a question and answer system for disease prediction. This study proposes the disease prediction question and answer system using the concepts of NLP and deep learning algorithms. The prediction is carried out using the sequence to sequence algorithm.

1.1 Overview of the System

The future of doctor-patient communication, healthcare planning, and management will be dominated by question and answer system. An artificial tool called a question and answer system is intended to mimic an intelligent discussion with human users. These question and answer system offer users a practical way to conduct information searches and are capable of handling straightforward questions with ease. These self-service tools are frequently a more intimate means of dealing with healthcare services than using an external call center or visiting a website. In order to help patients and prevent problems from occurring during regular business hours, such as spending a lengthy time on hold or making appointments that do not fit into their busy schedules, question and answer system are used. Hospitals will benefit from this healthcare system's 24/7 online health care support, providing detailed and broad information. Patients are helped to direct what they want by a series of inquiries.

1.2 Objectives of the Thesis

The following factors are described as the objective of the thesis.

- To provide users with correct information in medical domain
- To save time and cost of users and to know more healthcare information
- To apply recurrent neural networks to the healthcare question answering system
- To analyze the evaluation results of three attention models based on sequence to sequence models

1.3 Related Work

The paper, some references are used from previous proposes papers. Rojas, I., Joya, G., Catala, A. (eds) proposed an attention-based approach to short text classification for Twitter mining for public health monitoring [13]. The proposed

system was to automatically filter Tweets which are relevant to the syndrome of asthma/difficulty breathing based on bi-directional Recurrent Neural Network architecture with an attention layer (termed ABRNN). The system further distinguished between two variants of the ABRNN based on the Long Short Term Memory and Gated Recurrent Unit architectures respectively, termed the ABLSTM and ABGRU and hybrid. That found that the ABLSTM outperforms the other models with accuracy and an *F*1-score.

In the paper [9], the lexicon-based technique and the machine learning-based approach are thoroughly compared. Comparing the lexicon-based technique to the machine learning approach, the former was simpler to implement and comprehend. When compared to the Lexicon technique, the accuracy rate offered by the machine learning-based strategy was significantly greater, showing excellent performance. The development of chatbots used the Python programming language. The system concluded that when it comes to implementing machine learning techniques in chatbots, they are more effective than lexicon-based approaches. Himanshu Gadge, Vaibhav Tode, Sudarshan Madane, Prateek Kachare and Anuradha Deokar [7] had presented an Artificially Intelligent Chat-bot using applications of Deep Learning to fight COVID-19 including various viral diseases faced by human being in day Today life. Authors had covered some solutions to the user's query which will be beneficial for proper understanding of the patients by using neural network. Vishwanath Karad [4] discussed the chat bot application helped the student to know about the admission process of the college from anywhere with internet connection and receive fast replies. The system was proposed by utilizing the Python library's chatterbot algorithm, which makes it simple to create automated responses to user input.

The suggested solution consisted of an internet application that responds to questions presented by college administrators. The proposed system used pattern-matching, natural language processing and data mining. Jiao Liu, Yanling Li and Min Lin discussed intent detection is a crucial task that falls under spoken language interpretation, which is a crucial component of the human-machine interaction system [11]. The performance of semantic slot filling was closely related to the accuracy of intent recognition, and it was useful for the conversation system's subsequent research. The rich semantic information in user language cannot be understood by the conventional machine learning method due to the difficulty of purpose detection in

human-machine dialogue systems. In order to advance the study of multi-intent detection methods based on deep neural networks, the proposed system primarily analyzed, compared, and summarized the deep learning methods used in the research of intent detection in recent years. It also considered how to apply deep learning model to multi-intent detection task.

In the paper, Jurgita Kapociute-Dzikiene [8] discussed accurate generative chatbots were usually trained on large datasets of question and answer pairs. Despite such datasets not existing for some languages, it did not reduce the need for companies to have chatbot technology in their websites. However, companies usually owned small domain-specific datasets about their products, services, or used technologies. That found effective solutions to create generative seq2seq-based chatbots from very small data. Encoder–decoder LSTM-based approaches was suitable for English language than other languages for a morphologically complex language. The RNN was the top choice for sequential modeling jobs since it can handle sequences of any length. The RNN, like the LSTM, has demonstrated greater performance while maintaining contextual data. As a result, it was discovered that the LSTM was substantially more accurate in capturing spatial and temporal connections [16]. The conventional feed-forward network, in contrast, was unable to handle long-term dependencies and fails to preserve contextual information. Due to the high dimensionality of sequential data (like text), it might present a significant issue. BiLSTM, which eliminated several long and short dependencies, was the solution for such spatial data. For text classification, many researchers had combined CNN and RNN models.

1.4 Organization of the Thesis

The thesis is organized in five chapters. They are as follows: Chapter 1 describes introduction of the system, overview of system, objectives of the thesis, related works and thesis organization. Chapter 2 discusses the theoretical background. Chapter 3 presents the overview of the system architecture, dataset collection, training the models, testing the models and evaluation metric. Chapter 4 expresses the design and implementation of the proposed system and show the experimental result. Finally, Chapter 5 presents the conclusions, limitation and further extension.

CHAPTER 2
BACKGROUND THEORY

A complex interconnection of relatively simple processing nodes arranged in successive layers, including an input layer (where data is ingested), an output layer (where predictions are generated), and hidden layers, gave rise to deep learning models (those in between the input and output layers). An artificial intelligence technique called a neural network instructs computers to analyze data in a manner modeled after the human brain. Deep learning is a sort of machine learning that employs interconnected neurons or nodes in a layered framework to mimic the human brain. It develops an adaptive system that computers utilize to continuously learn from their errors and improve. Therefore, artificial neural networks aim to resolve complex issues, such as summarizing documents or recognizing faces, with greater accuracy.

2.1 Introduction of Deep Learning

A subset of machine learning is deep learning. Deep learning systems can perform better with access to more data, which is the machine equivalent of more experience, in contrast to typical machine learning algorithms, many of which have a finite ability to learn regardless of the amount of data they obtain [1]. Machines can be trained to perform specific activities such as driving a car, spotting weeds in a field of crops, diagnosing diseases, checking machinery for flaws, and other jobs once they have acquired sufficient experience through deep learning. Deep learning networks discover intricate patterns in the data they analyze to learn new things. By building computational models with numerous processing layers, the networks can establish different levels of abstraction to represent the data.

For many tasks, such as computer vision, speech recognition (also known as natural language processing), machine translation, and robotics, deep learning systems outperform typical machine learning systems by a large margin. This does not mean that creating regular machine learning systems is easier than creating deep learning systems. Despite the fact that feature recognition in deep learning is autonomous, thousands of hyperparameters (knobs) must be adjusted for a model to be effective.

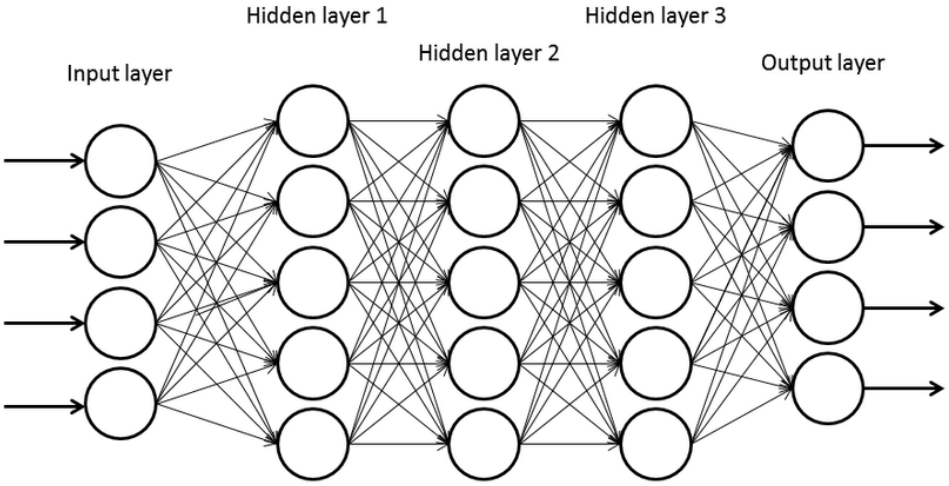

Figure 2.1 Deep Learning Neural Network

2.2 Neural Network Overview

Artificial intelligence deep learning is represented by neural networks. Traditional machine learning methods cannot handle some application cases because they are either complex or too broad. Neural networks, as they are generally called, step in and close the gap in these situations. The biological neurons in the human body that activate under specific conditions and cause a related action to be taken by the body in response are the model for artificial neural networks. Artificial neural networks are made up of several interconnected layers of artificial neurons that are activated by activation functions that enable ON/OFF switching [7]. Similar to conventional machine algorithms, neural networks learn specific values during the training stage.

In a nutshell, each neuron receives a multiplied version of inputs and random weights, which are then coupled with a static bias value (specific to each neuron layer). This information is then delivered to an appropriate activation function, which determines the final value to be given out of the neuron. Depending on the type of input values, a variety of activation functions are possible. The loss function (input vs. output) is calculated after the last neural net layer generates the output, and backpropagation is used to change the weights to minimize the loss. The overall procedure is centered on determining the best weight values. For better comprehension, kindly refer to the following. Although there are several varieties of neural networks, they generally fall

into three categories: Feedforward neural networks, Convolutional neural networks and Recurrent neural networks.

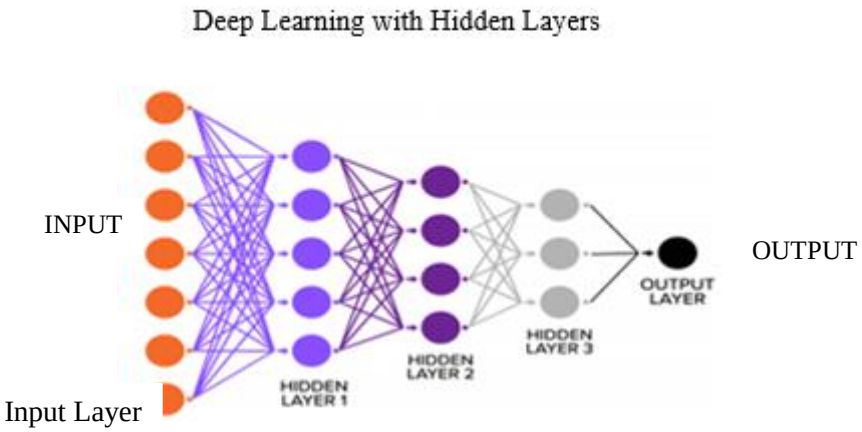

Figure 2.2 Neural Network

2.2.1 Feedforward Neural Network

A feed forward neural network is a type of artificial network in which there is no cycle in the connections between the nodes. Artificial neural networks called feed-forward networks do not have looping nodes. Due to the fact that all input is simply sent forward, this kind of neural network is also referred to as a multi-layer neural network. A recurrent neural network, in which particular paths are cycled, is the reverse of a feed forward neural network. Because input is only processed in one direction, the feed forward model is the simplest type of neural network [18]. The data may go via a number of hidden nodes, but it always proceeds forward and never backward.

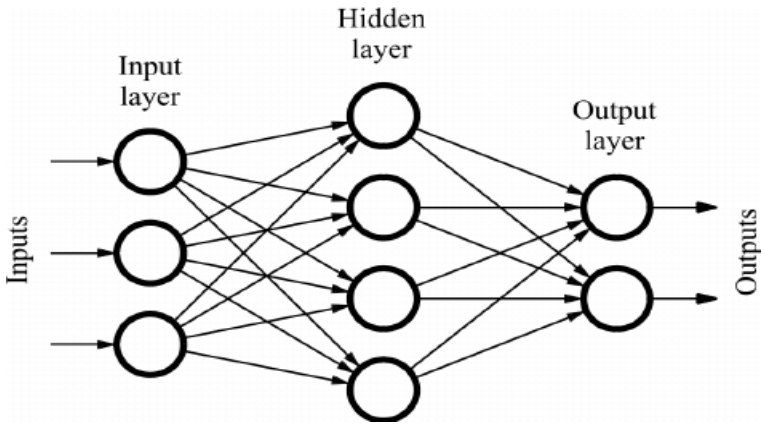

Figure 2.3 Feed Forward Neural Network

The single-layer perceptron is frequently used as a feed-forward neural network model for classification. Single-layer perceptrons can also incorporate machine learning. Neural networks can compare their outputs with the intended values by training them to modify their weights depending on a characteristic known as the delta rule.Gradient decline is the product of training and learning. Multi-layered perceptrons also adjust their weights. However, this method is called back-propagation. If this is the case, the hidden layers of the network will be modified in accordance with the output values that the final layer generates.

2.2.2 Convolutional Neural Network

Convolutional neural network (CNN), a class of artificial neural networks that has become dominant in various computer vision tasks, is attracting interest across a variety of domains. CNN is designed to automatically and adaptively learn spatial hierarchies of features through backpropagation by using multiple building blocks, such as convolution layers, pooling layers, and fully connected layers.

The convolutional layer is a crucial component of a convolutional neural network. A convolutional layer can be thought of as a collection of convolutional kernels, or tiny square templates, that move across an image and search for patterns. When that area of the image complies with the pattern of the kernel, the kernel provides a large positive number; otherwise, it returns zero or a lesser value. To save on processing costs, pooling layers makes the convolutional feature map smaller [17]. Global patterns for cars are produced by fully connected layers, which learn global patterns based on the high-level characteristics produced by the convolutional and pooling layers. The last layer activates the model after the fully connected layer has processed the input data, and the neural network then makes predictions.

Convolutional neural network is composed of multiple building blocks, such as convolution layers, pooling layers, and fully connected layers, and is designed to automatically and adaptively learn spatial hierarchies of features through a backpropagation algorithm.

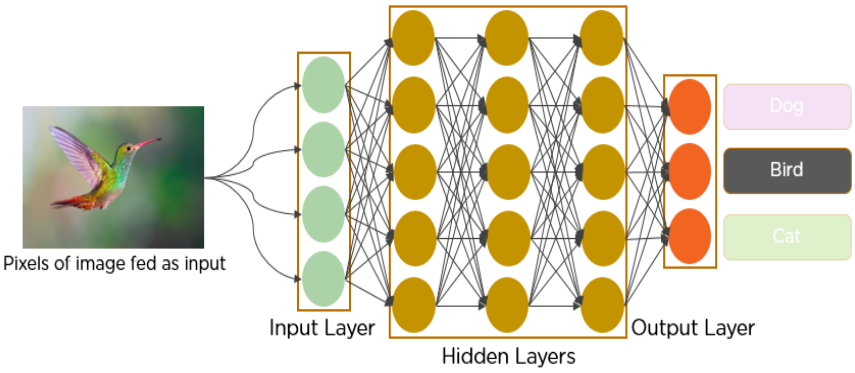

Figure 2.4 Convolutional neural networks

2.2.3 Recurrent Neural Network

Recurrent neural networks, or RNNs, are a particular kind of neural network designed for data sequences. Due to its internal memory, it is the first algorithm to recall its input, making it ideal for machine learning issues involving sequential data. It is one of the algorithms that helped deep learning accomplish some incredible successes over the past several years. In this article, the author will examine over the fundamental ideas behind recurrent neural networks, as well as the main problems they face and how to fix them.

Because traditional neural networks contain discrete input and output layers, they are ineffective for handling sequential data. To preserve the outcomes of earlier outputs in internal memory, a new neural network called the Recurrent Neural Network was created. The network then uses these results as inputs. Thus, it can be applied to tasks like time series prediction, speech and voice recognition, natural language processing, and pattern recognition. The outputs of a layer in a loop are stored in memory using hidden layers in RNN. The following are the four most popular varieties of recurrent neural networks:

1. One to One RNN

It is employed to address general machine learning issues with a single input and output. It is also known as Vanilla Neural Network. It is used to solve regular machine learning problems.

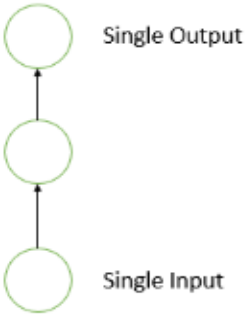

Figure 2.5 One to One RNN

2. One to Many RNN

A one-to-one recurrent neural network has one input and many outputs. This is demonstrated in the diagram below.

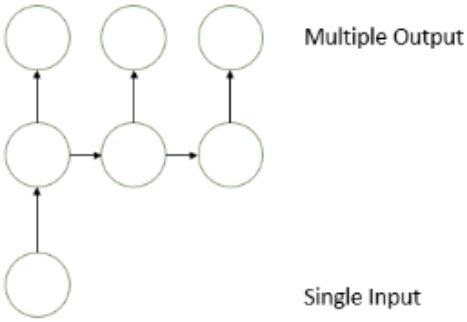

Figure 2.6 One to Many RNN

3. Many to One RNN

A typical example is the many-to-one RNN architecture (Tx>1, Ty=1) used in sentiment analysis models. As the name implies, this type of model is utilized when more than one input is needed to produce a single output.

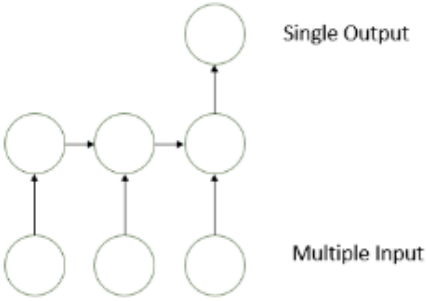

Figure 2.7 Many to One RNN

Consider the Twitter sentiment analysis model as an illustration. A text input (words as several inputs) in that model provides its fixed sentiment (single output). Another illustration would be a system for assigning movie ratings that uses review texts as input to provide a number between 1 and 5 to a film.

4. Many to Many RNN

The most prevalent use of this type of RNN architecture is in machine translation. Many-to-Many architecture can also be represented in models where the input and output layers are of different sizes. For instance, the three wonderful English words "I Love You" are translated into only two in Spanish, "te amo." Because a non-equal Many-to-Many RNN architecture is at work in the background, machine translation models are therefore capable of returning words that are either more or less than the input string.

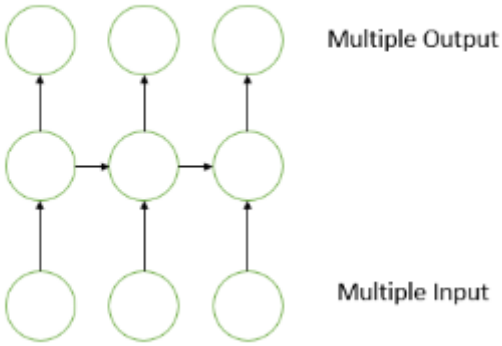

Figure 2. 8 Many to Many RNN

An artificial neural network that employs sequential data or time series data is known as a recurrent neural network (RNN). These deep learning techniques are frequently applied to ordinal or temporal issues, including speech recognition, image captioning, language translation, and nlp (natural language processing).

- Recurrent Neural Networks are used in several domains.
- In Natural Language Processing (NLP), they've been used to generate handwritten text, perform machine translation and speech recognition. But their applications are not restricted to processing language.
- Generally, the advantage of recurrent networks is that they share weights for each position of the input vector.

- Also, a model can process sequences with different lengths by sharing the weights.
- Another advantage is that it reduces the number of parameters (weights) that the network needs to learn.
- The basic principle in recurrent networks is that the input vector and some information from the previous step (generally a vector) are used to calculate the output and information passed to the next step.
- In general, the formulas used to calculate the output values in each step are called units (blocks).

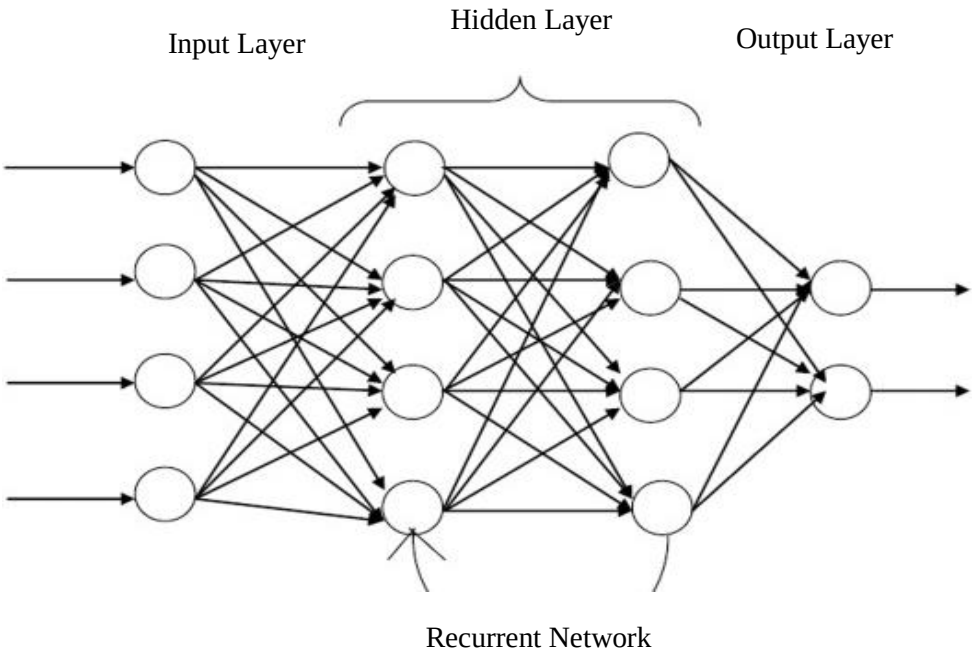

Figure 2.9 Recurrent Neural Network

Types of Recurrent Neural network(RNN):

1. **Bidirectional recurrent neural networks**

Another variety of RNNs known as bidirectional recurrent neural networks (BRNNs) simultaneously learn the forward and backward directions of input flow. Compared to ordinary RNNs [6], BRNNs can only learn information in one way, this is different. Bidirectional information flow refers to the simultaneous learning of both directions. A typical artificial neural network uses backward projections to assess the past and forward projections to forecast the future. However, they are not combined as in a BRNN. Bidirectional RNNs, or BRNNs, are used to permit straight (past) and

backward traversal of input (future). A BRNN is made up of two RNNs, one of which starts at the beginning of the data sequence and moves forward, and the other of which starts at the end and moves backward.

Simple RNNs, GRUs, or LSTMs are all acceptable types of network blocks for BRNNs. BRNN is useful for the following applications: Handwriting Recognition, Speech Recognition, Dependency Parsing, Natural Language Processing. A bidirectional RNN is created by combining two RNNs that train the network in opposing directions-one from the beginning to the end of a sequence and the other from the end to the beginning. By allowing the model to learn about more than just the past and present, it aids in the analysis of future events.

2. Long short-term memory units

The vanishing gradient problem, which occurs when a neural network cannot be adequately trained, is a disadvantage of typical RNNs. Deeply layered neural networks, which are employed to handle complex data, cause this. The performance of conventional gradient-based learning RNNs decreases with increasing size and complexity [6]. Effective parameter tuning at the first layers becomes computationally and time intensive. Long short-term memory (LSTM) networks are one solution to the issue. Data is divided into short-term and long-term memory cells by RNNs constructed with LSTM units.

3. Gated recurrent units

A type of recurrent neural network unit that can be used to simulate sequential data is called a gated recurrent unit (GRU). Sequential data can also be modeled using LSTM networks, however they are less effective than conventional feed-forward networks. Networks can benefit from the strengths of both units the LSTM's capacity to learn from long-term connections and the GRU's capacity to learn from short-term patterns by combining an LSTM and a GRU.

Additionally, GRUs handle the issue with vanishing gradients that affects regular recurrent neural networks (values used to update network weights). Grading may become too little to have an impact on learning if it shrinks over time as it back propagates, rendering the neural network untrainable. RNNs can basically "forget" lengthier sequences if a layer in a neural net is unable to learn. The update gate and reset gate are two gates that GRUs utilize to address this issue. These gates can be

trained to retain information from further back and determine what information is allowed through to the output. As a result, it can transfer important information along an event chain to improve its forecasts.

2.3 Sequence to Sequence Model

Sequence to Sequence model is a machine translation and language processing technique based on encoder-decoders that maps an input sequence to an output sequence with a tag and attention value. The goal is to attempt to forecast the following state sequence using two RNNs that will cooperate using a unique token. Grouping to arrangement models depend on RNN design and comprises of two RNNs: an encoder and a decoder. The encoder's errand is to process the info, and the decoder to process the yield [14]. Grouping to arrangement models can be thought of as one decoder hub creating yield relating to one encoder hub. This model has direct application in machine interpretation as a comparing word for the yield language can be produced by decoder effectively by taking a gander at single word of information language at once.

Sequence to sequence creates an output series of words from an input sequence of words (sentence or sentences). Utilizing the recurrent neural network, it accomplishes this (RNN). LSTM or GRU, the more sophisticated variant of RNN, are utilized more frequently than the more basic version, which is rarely used. This is due to the vanishing gradient issue that RNN has. The Google-recommended version makes use of LSTM. By requiring two inputs at each instant, it creates the word's context. Recurrent refers to two outputs, one from the user and the other from past output (output goes as input).

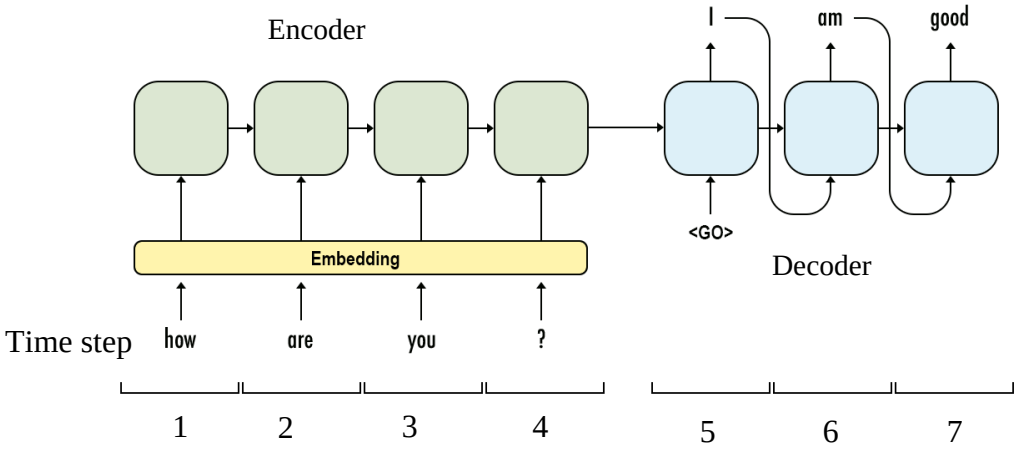

Figure 2.10 Sequence to Sequence Model

2.4 Encoder-Decoder Model

Recurrent neural networks are used in the encoder-decoder model to solve sequence-to-sequence prediction issues. Although it was first created for machine translation issues, it has also demonstrated success in other related sequence-to-sequence prediction issues like text summarization and query resolution. The method uses two recurrent neural networks: the encoder, which encodes the input sequence; and the decoder, which decodes the encoded input sequence into the target sequence. To be in the proper format, data must be encoded. In the example of Pictionary, the research is turned a word (text) into a drawing (image). In the context of machine learning, transforms a list of Spanish words into a two-dimensional vector, sometimes referred to as the hidden state. Recurrent neural networks are stacked to create the encoder (RNN)[5]. The author employs this kind of layer since the model can comprehend the context and temporal relationships of the sequences thanks to its structure. The previous RNN time step's state is the concealed state, the encoder's output.

A two-dimensional vector representing the entire meaning of the input sequence is the encoder's output. The number of cells in the RNN determines the length of the vector. A message that has been encoded must be decoded before it can be understood. The Pictionary team's second member will transcribe the image into a word. The decoder will transform the two-dimensional vector into the output sequence, which is the English phrase, in the machine learning model. In order to predict the English term, it is also constructed with RNN layers and a thick layer.

The possibility of different input and output sequence lengths is one of this model's key advantages. This makes way for some really intriguing uses, such question-and-answer sessions or video captioning. This straightforward encoder-decoder approach has a significant limitation in that all data must be condensed into a one-dimensional vector, which can be very challenging for lengthy input sequences. Having said that, as encoder decoder models provide the foundation for attention models and transformers, understanding them is essential for the most recent developments in NLP. We shall follow the construction of a translation model with an encoder decoder structure in the subsequent article. The attention mechanism will then be discussed in order to increase accuracy.

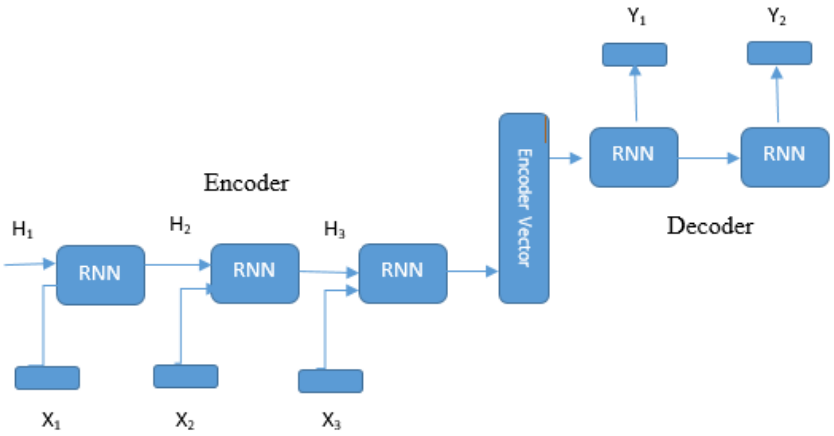

Figure 2.11 Encoder-Decoder Model

2.5 Encoder with Bidirectional Recurrent Neural Networks

Recurrent Neural Networks, or RNNs, are a specialized class of neural networks used to process sequential data. Sequential data can be considered a series of data points. A specialized kind of neural networks called recurrent neural networks, or RNNs, are used to analyze sequential input. A set of data points can be thought of as sequential data. Video is sequential because it is made up of a series of video frames, music is sequential because it is made up of a series of sound elements, and text is sequential because it is made up of a series of letters. It is necessary to preserve the data learned from earlier instances while modeling sequential data. The prior argument presented by the participants must be taken into account in a discussion, for instance, if you are to correctly predict the following argument during that debate. The argument is crafted so that it follows the direction of the discussion. Similarly, an RNN learns and retains the information in order to create a judgment, and this depends on the prior learning. BRNNs encoder algorithm is the following [19]:

Inputs:

- input_seq: batch of input sentences; shape=(max_length, batch_size)
- input_lengths: list of sentence lengths corresponding to each sentence in the batch;
- hidden: hidden state; shape=(n_layers x num_directions, batch_size, hidden_size)

Computation Graph:

- Convert word indexes to embedding.

- Pack padded batch of sequences for RNN module.
- Forward pass through GRU.
- Unpack padding.
- Sum bidirectional GRU outputs.
- Return output and final hidden state

Outputs:

- outputs: output features from the last hidden layer of the GRU (sum of bidirectional outputs); shape=(max_length, batch_size, hidden_size)
- hidden: updated hidden state from GRU; shape=(n_layers x num_directions, batch_size, hidden_size)

2.6 Attention

The attention mechanism was introduced to improve the performance of the encoder-decoder model for machine translation. The Attention Mechanism has proved itself to be one necessary component of RNN to deal with tasks like Neural Machine Translation, Question Answering and Natural Language Inference. If the context vector is the only vector exchanged between the encoder and decoder, the entire sentence must be encoded in that one vector. For each step of the decoder's own outputs, attention enables the decoder network to "focus" on a different portion of the encoder's outputs. It is suggested that attention can be used to align and translate a lengthy sequence of information that is not need to be a fixed length. While translation is the act of employing the pertinent information to choose the suitable output, alignment is the machine translation problem that determines which elements of the input sequence are relevant to each word in the output [2]. These issues can be easily resolved with the use of attention models, which also offer flexibility to translate long sequences of information. The specific mechanics behind Attention, it must be noted that there are 2 different major types of Attention:

- Bahdanau Attention
- Luong Attention

2.6.1 Bahdanau Attention

The Bahdanau attention was proposed to address the performance bottleneck of conventional encoder-decoder architectures, achieving significant improvements over

the conventional approach. Bahdanau attention takes concatenation of forward and backward source hidden state (Top Hidden Layer). But in the Bahdanau at time t we consider about t-1 hidden state of the decoder. Then the author calculate alignment, context vectors as above. But then concatenate this context with hidden state of the decoder at t-1. So before the softmax this concatenated vector goes inside a GRU. Bahdanau has only concat score alignment model. Bahdanau recommend uni-directional encoder and bi-directional decoder [13]. The entire step-by-step process of applying Attention in Bahdanau attention is as follows:

1. Generating the Hidden States of the Encoder-The encoder generates the hidden states of each element in the input sequence.
2. Calculating Alignment Scores between each hidden state of the encoder and the previous hidden state of the decoder (Note: The last encoder hidden state can be used as the first hidden state in the decoder).
3. Softmaxing the Alignment Scores: Each encoder's hidden alignment scores are pooled, represented by a single vector, and then softmaxed.
4. Context vector calculation: the encoder hidden states are multiplied by their corresponding alignment scores to create the context vector.
5. Decoding the Output: To produce a new output, the context vector is passed into the decoder RNN for that time step together with the preceding decoder's hidden state.
6. The procedure (steps2-5) is repeated for each decoder time step until a token is generated or the output exceeds the set maximum length.

$$\text{Score}(h_t, h_{\bar{s}}) for\ concat = h_a^T tanh(W_a[h_t; h_s^T]) \quad \textbf{Equation 2.1}$$

2.6.2 Luong Attention

Deep Learning's attention mechanism gives particular aspects of the data it processes more attention. By introducing two new classes of attentional mechanisms a global approach that pays attention to all source words and a local approach that only pays attention to a particular subset of words in predicting the target sentence, the Luong attention sought to make a number of improvements over the Bahdanau model

for neural machine translation [2]. The second type of attention is called Luong attention, sometimes known as multiplicative attention, and it was developed on top of the Bahdanau Attention Mechanism. In both the encoder and the decoder, Luong attention exploited top hidden layer states. There are three different kinds of alignments in Luong. Both are uni-directional in Luong's view. Additionally, Luong advises only using the outputs from the top layer because their model is generally simpler. Applying Attention in Luong attention involves the following detailed process:

1. Generating the Hidden States of the Encoder-The encoder generates the hidden states of each element in the input sequence.
2. Decoder RNN: To create a new hidden state for that time step, the previous decoder hidden state and decoder output are transmitted via the Decoder RNN.
3. Calculating Alignment Scores-Alignment scores are computed using the new decoder hidden state and the encoder hidden state.
4. Softmaxing the Alignment Scores: Each encoder's hidden alignment scores are pooled, represented by a single vector, and then softmaxed.

Luong Attention's Alignment functions:

$$\text{Score}\,(h_t, \bar{h_s})\text{for dot} = h_s^T * h_s^T \qquad \textbf{Equation 2. 2}$$

$$\text{Score}\,(h_t, \bar{h_s})\text{for general} = h_s^T * W_a * h_s^T \qquad \textbf{Equation 2. 3}$$

$$\text{Score}\,(h_t, \bar{h_s})\text{for concat} = h_a^T \tanh(W_a[h_t + h_s^T]) \qquad \textbf{Equation 2. 4}$$

- Dot is the simplest of the functions; all that is required to generate the alignment score is to multiply the secret states of the encoder and the hidden states of the decoder.
- The general function is comparable to the dot function, with the addition of a weight matrix to the equation.
- Concat is the process in which the hidden states of the encoder and decoder are combined before being sent through a linear layer. This implies that, unlike

in Bahdanau Attention, the decoder hidden state and encoder hidden state will not have separate weight matrices but rather a common one. The output of the Linear layer will first undergo a tanh activation function before being multiplied by a weight matrix to create the alignment score.

2.7 Luong Attention with Decoder

Multiplicative attention is another name for Luong's attention. By performing basic matrix multiplications, it converts encoder states and decoder states into attention scores. It is quicker and uses less space when a simple matrix multiplication is used. Luong attention mechanism uses the current decoder's hidden state to compute the alignment vector [3].

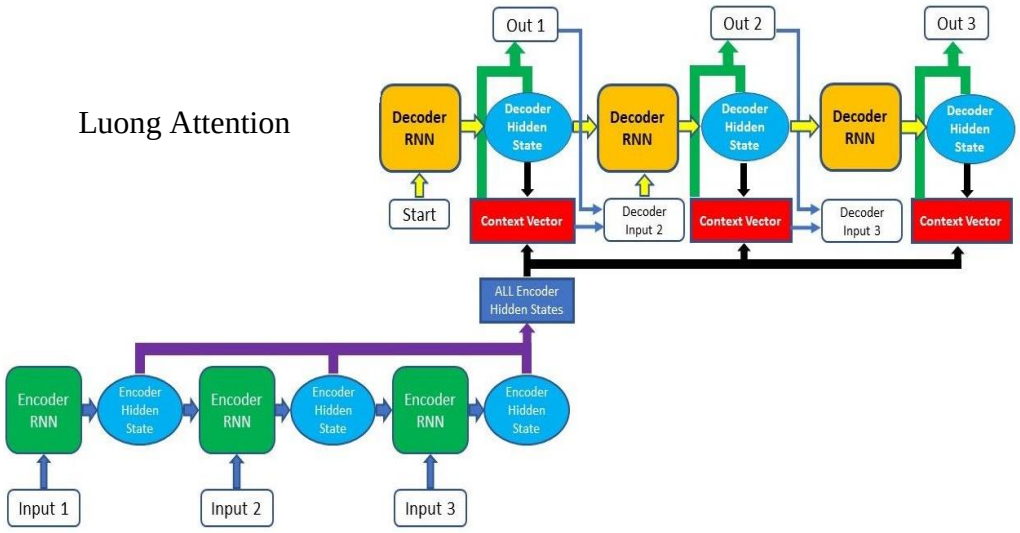

Figure 2.12 Luong Attention Overview

Inputs:

- input_step: one time step (one word) of input sequence batch; shape= (1, batch_size)
- last_hidden: final hidden layer of GRU; shape=(n_layers x num_directions, batch_size, hidden_size)
- encoder_outputs: encoder model's output; shape=(max_length, batch_size, hidden_size)

Computation Graph:

- Get installing of current info word.
- Forward through unidirectional GRU.

- Calculate consideration loads from the current GRU yield from (2).
- Multiply consideration loads to encoder yields to get new "weighted entirety" setting vector.
- Concatenate weighted setting vector and GRU yield.
- Predict next word (without softmax).
- Return output and last shrouded state.

Outputs:

- output: softmax normalized tensor giving probabilities of each word being the correct next word in the decoded sequence; shape=(batch_size, voc.num_words)
- hidden: final hidden state of GRU; shape= (n_layers x num_directions, batch_size, hidden_size) [3]

2.8 Long-Short Term Memory

Recurrent neural networks of the Long Short Term Memory (LSTM) variety are capable of learning order dependency. The current RNN step uses the output from the previous step as its input. It addressed the problem of RNN long-term reliance, in which the RNN can predict words based on recent data but cannot predict words kept in long-term memory. As the gap length increases, RNN's performance is inefficient. It is employed in the processing, prediction [16], and classification of time-series data. Four neural networks and a large number of memory cells, which are arranged in a chain pattern, make up the LSTM. A cell, an input gate, an output gate, and a forget gate make up a typical LSTM unit. Three gates regulate the information flow into and out of the cell, and the cell retains values for arbitrary time periods. Time series with indeterminate duration can be categorized, examined, and predicted with the LSTM algorithm. Three gates regulate the information flow into and out of the cell, and the cell retains values for arbitrary time periods. Time series with indeterminate duration can be categorized, examined, and predicted with the LSTM algorithm.

Forget Gate

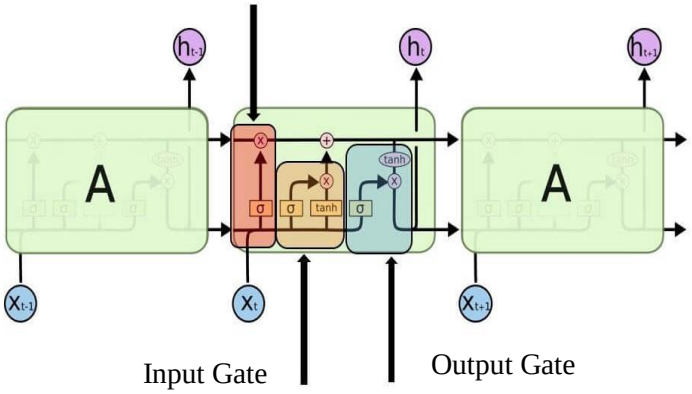

Figure 2.13 Gate of Long Short Term Memory

- Input Gate: It chooses which input values should be applied to the memory change. It is decided whether to pass through 0 or 1 data using the sigmoid function. The tanh function also adds weight to the supplied data, ranking its significance on a scale from -1 to 1.
- Forget Gate: It identifies the information that needs to be taken out of the block. A sigmoid function makes the decision. It examines the prior state (ht-1), the content input (Xt), and each number in the cell state Ct-1 to produce a number between 0 (omit this), and 1. (keep this).
- Output Gate: The output is determined by the input and memory of the block. It is decided whether to pass through 0 or 1 data using the sigmoid function. And which numbers can pass through 0 and 1 is determined by the tanh function. Additionally, the tanh function gives the input values weight by multiplying them by the sigmoid output to determine their importance on a scale from -1 to 1.

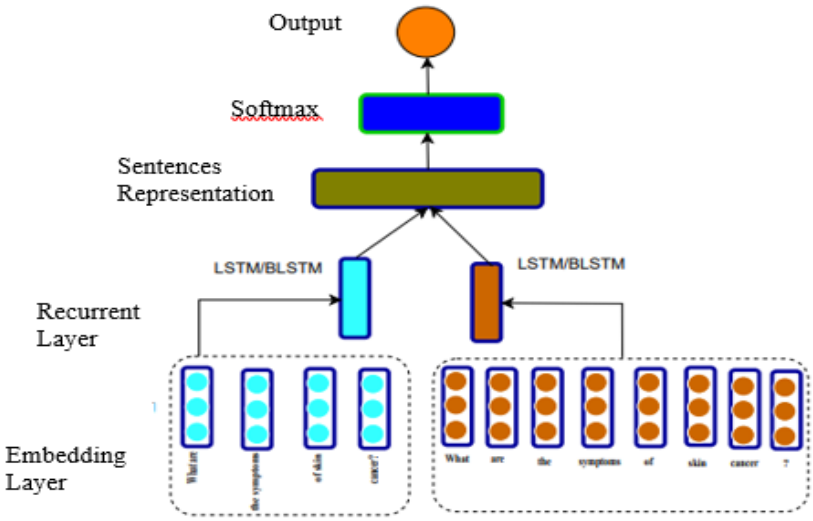

Figure 2.14 Architecture of Long Short Time Memory

2.9 Chapter Summary

The technological theories required for this thesis are discussed in this chapter. Understanding how an artificial neural network (ANN) functions generally includes a detailed explanation of its process flow. Additionally, the training algorithms and learning processes are discussed. The several recurrent neural network layers and contemporary RNN architectures and Attention mechanism are covered in the next chapter. And, the topic of biological neurons and artificial neurons was discussed in this chapter. Additionally, explained are the activation functions of luong attention mechanism and the artificial neural network's work flow. Additionally, a deep artificial neural network and its use in applications for classification and clustering are demonstrated. The sorts of learning techniques used to train these networks were then covered in this chapter. Recurrent neural networks' various training techniques and operational flow are described. Last but not least, the architectures of contemporary RNN layers are also shown, and each graphic makes obvious the structures.

CHAPTER 3

THE PROPOSED SYSTEM ARCHITECTURE

A question and answer system is a computer program that can communicate in natural language with users. Because of the large amount of information on the internet, question and answer system can deliver precise and effective information based on the user's needs. Question and answer system is utilized for casual communication as well as in fields like customer support, virtual assistance, online trainings, and online reservations [7]. The suggested healthcare question and answer system can engage with users and simulate a conversation with a medical professional. There are a few healthcare question and answer system that already exist, but they do not give users medication for any illnesses; instead, they connect them to a medical QA forum and display questions that are similar to their symptoms and may have already been addressed by professionals. The aim is to demonstrate that the proposed medical question and answer system could be a better alternative to many already existing question and answer system in the domain of medicine.

3.1 Dataset Collection

Query answering models are helpful for finding answers in documents because they can extract the answer to a question from a text. Some question-answering models are capable of producing replies devoid of context. So, the dataset is very important to build such as a system. Deep learning can be applied to any data type. The data types and the data gathered will depend on the problem which is trying to solve.

MedQuAD [3] includes 47,457 medical question-answer pairs created from 12 NIH websites (e.g. cancer.gov, niddk.nih.gov, GARD, MedlinePlus Health Topics). The collection covers 37 question types (e.g. Treatment, Diagnosis, Side Effects) associated with diseases, drugs and other medical entities such as tests. Additional annotations any included in the XML files.

There used 12 trusted websites to construct a collection of question-answer pairs. They are provided below:

1.National Cancer Institute (NCI): It contains free text from 116 articles on various cancer types and 729 question and answer pairs.

2.Genetic and Rare Diseases Information Center (GARD): This website contains all disease question/answer pairs from 4278 topics and 5394 question and answer pairs.

3.Genetics Home Reference (GHR): It contains consumer-oriented information about the effects of genetic variation on human health ,1099 articles about and 5430 question and answer pairs.

4.MedlinePlus Health Topics: It contains information on symptoms, causes, treatment and prevention for diseases, health conditions, and wellness issues of 981 articles and 981 question and answer pairs.

5.National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK): It contains text from 174 health information pages on diseases studied by this institute of 1192 question and answer pairs.

6.National Institute of Neurological Disorders and Stroke (NINDS): It contains free text from 277 information pages on neurological, stroke-related diseases from this resource and 1104 question and answer pairs.

7.NIHSeniorHealth: This website contains health and wellness information for older adults and 71 articles and 769 question and answer pairs.

8.National Heart, Lung, and Blood Institute (NHLBI): It contains text from 135 articles on diseases, tests, procedures, and other relevant topics on disorders of heart, lung, blood, and sleep. It contains 559 question and answer pairs.

9.Centers for Disease Control and Prevention (CDC): It contains 152 articles on diseases and conditions with 270 question and answer pairs.

10.MedlinePlus A.D.A.M. Medical Encyclopedia: It contains 4366 articles about conditions, tests, and procedures. 17,348 question and answer pairs were extracted from this resource.

11.MedlinePlus Drugs: We extracted free text from 1316 articles about Drugs and generated 12,889 question and answer pairs.

12.MedlinePlus Herbs and Supplements: We extracted free text from 99 articles and generated 792 question and answer pairs.

All of the above of websites are dataset in MedQuAD [19] but the author used three websites to construct a collection of question-answer pairs. They are provided below: National Cancer Institute (NCI), NIHSeniorHealth and National Heart, Lung, and Blood Institute (NHLBI).

The National Cancer Institute (NCI) oversees a broad range of research, training, and information dissemination activities that reach across the entire country, meeting the needs of all demographics rich and poor, urban and rural, and all

racial/ethnic populations. The NCI is the leader of the cancer research enterprise, collectively known as the National Cancer Program, and the largest funder of cancer research in the world. NCI concentrates on two major responsibilities in particular**:** Cancer Research and Training and Support for Cancer Researchers. The cause, diagnosis, prevention, treatment, and rehabilitation from cancer, as well as the ongoing care of cancer patients and their families, are all areas in which the National Cancer Institute conducts and supports research. It also provides training, health information distribution, and other programs in these areas. The Nation's medical and behavioral research is overseen by NIHSeniorHealth. Its goal is to discover fundamental truths about the nature and function of living systems and to use those truths to improve health, extend life, and lessen disease and impairment.

The National Heart, Lung, and Blood Institute (NHLBI) offers leadership for a research, education, and training program that promotes the prevention and treatment of heart, lung, and blood illnesses as well as improving overall health so that people can live longer and more satisfying lives. The NHLBI promotes the training and mentorship of young scientists and physicians as well as the dissemination of research advancements to the general public. It also permits the conversion of basic discoveries into clinical practice. In collaboration with both public and private entities, including as academic institutions, business, and other governmental agencies, it builds and sustains a strong, cooperative research infrastructure. In order to encourage the application of research findings and make the most of available resources to address public health needs, the Institute works in conjunction with patients, families, healthcare professionals, scientists, professional societies, patient advocacy groups, community organizations, and the media. The NHLBI also works with foreign groups to lessen the impact of blood, lung, and heart illnesses all across the world.

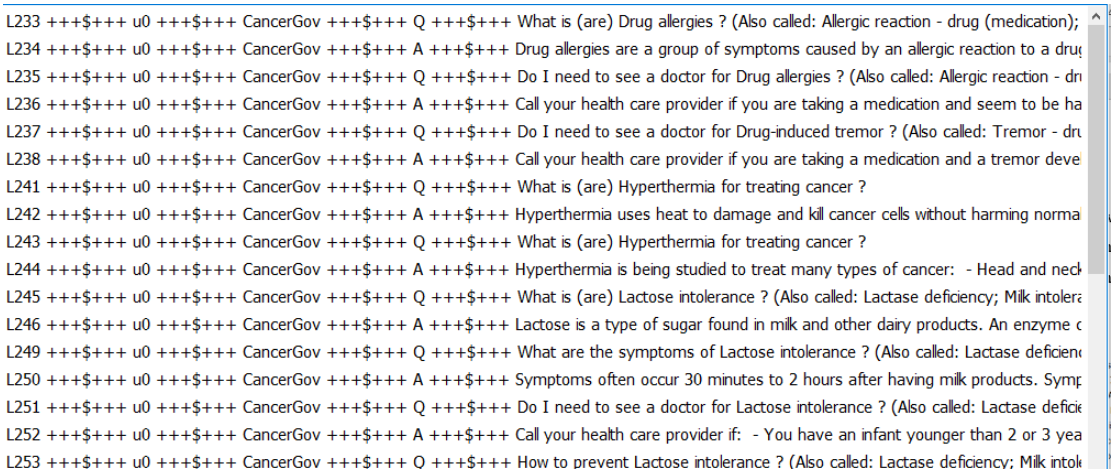

Figure 3.1 Sample Dataset Form

3.2 Training the Models

In the data science development lifecycle, the model training phase is where practitioners attempt to match the ideal weights and bias to a machine learning algorithm in order to minimize a loss function over the prediction range. Building the best mathematical representation of the relationship between data features and a target label (in supervised learning) or between the features themselves is the goal of model training (unsupervised learning). Since they specify how to optimize the deep learning algorithms, loss functions are a crucial component of model training. Different types of loss functions are used by data scientists depending on the goal, type of data, and technique.

The essential phase in deep learning that produces a model ready for validation, testing, and deployment is model training. The quality of the apps created using the model is determined by its performance. During the model training phase, the caliber of the training data and the training algorithm are both crucial components. Training, validation, and testing data are typically separated. The end use case influences the selection of the training algorithm. The ideal algorithm must balance a variety of factors, including model complexity, interpretability, performance, computation requirements, etc. Model training is a complex and crucial step in the entire deep learning growth cycle as a result of all these factors. In system, the work flow of the three models (General, Dot, Concat), three model can be trained with increasing no. of hidden layers to make it more accurate and no. of iterations in model training. The eighty percent of training and twenty percent of testing is used in these models.

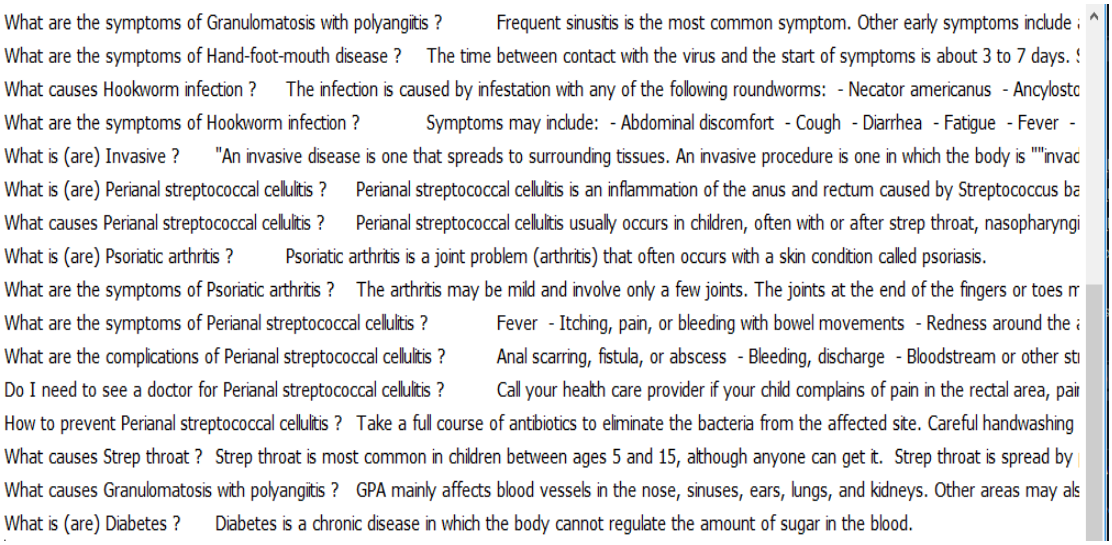

Figure 3.2 Training Dataset Sample Form

- Start preparing training data ...
- Reading lines...
- Read sentence pairs
- Trimmed to sentence pairs
- Counting words...
- Counted words:
- Iteration: 1; Percent complete: 10.0%; Average loss: 7.3758
- Iteration: 2; Percent complete: 20.0%; Average loss: 7.3393
- Iteration: 3; Percent complete: 30.0%; Average loss: 6.0982
- Iteration: 4; Percent complete: 40.0%; Average loss: 5.9207
- Iteration: 5; Percent complete: 50.0%; Average loss: 5.6064

Parameter Setting for RNN:

This Table 3.1 shows the parameter setting of the system. It shows the step by step process to create the models. This system includes three models(general, dot, concat) and which are create 2 encoder and 2 decoder, hidden size is 500 and batchsize is 64 , set the iteration is 1200 to create a good model, sentence maximum length is 200 and minimun length is 3.

Table 3.1 Parameter Setting

Encoder_layer	2
Decoder_layer	2
Hidden size	500
Batch size	64
Sentence_max_length	200
Sentence_min_length	3
dropout	0.1
checkpoint_iteration	1200

3.3 Testing the Models

Once more, a portion of the data from the original data source makes up the test data. It simply consists of data points that were not used in the creation or validation of the model. Only after comparing a model to test data and statistical analysis demonstrates a reasonable match do we deem the model ready for usage. In model, change over all letters to lowercase and trim all non-letter characters with the exception of essential accentuation. Trimmed to all sentence pairs of given dataset and counting Unique Words. To accomodate sentences of various sizes in a similar clump, the author will make the bunched information tensor of shape (max_length, batch_size), where sentences shorter than the max_length are zero cushioned after an EOS_token.

Question > do i need to see a doctor for hypersensitivity vasculitis ? also called allergic vasculitis leukocytoclastic vasculitis cutaneous vasculitis

Real Answer > call your provider if you have symptoms of hypersensitivity vasculitis .

['call', 'your', 'provider', 'if', 'you', 'have', 'symptoms', 'of', 'hypersensitivity', 'vasculitis', '.', 'EOS',

'EOS', 'EOS']

Bot: call your provider if you have symptoms of hypersensitivity vasculitis .

Blue Score: 1.0

Question > how to prevent lactose intolerance ? also called lactase deficiency milk intolerance disaccharidase deficiency dairy product intolerance

Real Answer > there is no known way to prevent lactose intolerance . you can prevent symptoms by avoiding foods with lactose .

['there', 'is', 'no', 'known', 'way', 'to', 'prevent', 'prevent', 'prevent', 'prevent', 'prevent', 'prevent', 'prevent', 'symptoms', 'symptoms', 'by', 'avoiding', 'foods', 'lactose', 'lactose', 'EOS', …., 'EOS', 'EOS', 'EOS', 'EOS', 'EOS', 'EOS', 'EOS', 'EOS']

Bot: there is no known way to prevent prevent prevent prevent prevent prevent prevent symptoms symptoms by avoiding foods lactose lactose

Blue Score: 0.75

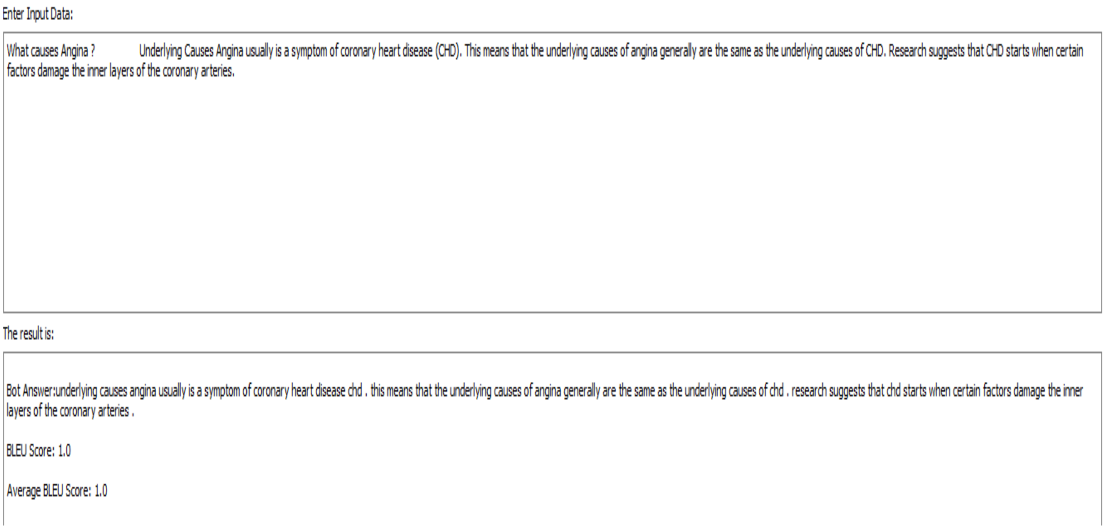

Figure 3.3 Testing Dataset Sample Form

3.4 Evaluation Metric

The system uses data from the test dataset to assess the new model's quality and accuracy once it has been trained. The BLEU (Bilingual Evaluation Understudy) score, which measures how similar the candidate text is to the reference texts, is used in this approach to quantify the model quality. Scores closer to one imply more similar texts. By exporting the test dataset with the model predictions, the BLEU score may examine the model output for particular data items in addition to providing an overall evaluation of model quality [18]. Both the reference text (from the original dataset) and the candidate text for the model are included in the exported data.

- Performance of the models are evaluated with BLEU score method.
- Bilingual Evaluation Understudy (BLEU) is a metric that is used to compute the quality of system generated answers.
- The output of BLEU score range is always between 0 and 1, value close to 1 show that the system generated answer is more analogous to the expert generated answer and 0 is no match at all.

$$\text{BLEU} = \min\left(1, \frac{\text{hypothesis length}}{\text{reference length}}\right) \frac{Max\ number\ of\ words\ occurs\ in\ reference}{Total\ no\ of\ words\ in\ \text{hypothesis}}$$

Equation 3. 1

Where hypothesis means dataset's expert answer and reference is QA system's answer. Then, the system shows the comparative results of the models.

Table 3.2 Training and Testing of the Dataset

Dataset Name	QA Pairs	Training Pairs (80%)	Testing Pairs (20%)	Unique Words	Dot Accuracy	General Accuracy	Concat Accuracy
NIH Senior Health	769	616	153	6249	0.42	0.48	0.31
NCI	729	584	145	6272	0.48	0.41	0.39
NHLBI	559	448	111	5901	0.52	0.48	0.42

Table 3.2 shows the result of blue scores of nine different Sequence to Sequence models for three datasets. According to the experimental results, the QA model on dot function is the best than other models. The QA model on concat produces less blue score than other models. So the healthcare QA system should use Dot and general functions to get more accurate answers. The three models performed best on three datasets from the National Heart, Lung, and Blood Institute (NHLBI), the National Cancer Institute (NCI), and NIHSeniorHealth were trained and tested on the various corpora. On the NHLBI dataset, the dot model provided the highest accuracy. The General model trained on NIHSeniorHealth performed the best when tested on our test set (153 medical NIHSeniorHealth pairs). When evaluated on all datasets, the concatenation model did not perform well. The varied forms of input could cause incorrect internal conceptualizations of medical phrases and inquiries in the deep brain layers, which would account for this.

The test consumer health questions' complexity, which often consists of multiple subquestions, contextual information, and the potential for misspellings and grammatical errors, could possibly be to blame for the decline in performance. Due to the stringent semantics, the open-domain concept of textual entailment may not entirely apply to question entailment. While the definitions of the three datasets for question-answering depend on the relationship between the sets of responses to the compared questions, the general textual entailment definitions merely refer to the premise and hypothesis.

3.5 Chapter Summary

In this chapter, a thorough explanation of the healthcare question and answer system's step-by-step execution is provided. Each step in the system flow diagram and its description have been provided. This chapter covered a thorough discussion of the dataset, RNN architectures, the training of these models, loss functions, and system optimization strategies.

CHAPTER 4
IMPLEMENTATION AND EXPERMENTAL RESULTS OF THE SYSTEM

4.1 Implementation of the System

Modern consumers are looking for novel ways to obtain information quickly and easily. Therefore, the majority of healthcare providers have begun utilizing question and answer system virtual healthcare assistants to improve patient engagement and streamline interoperability. The main objective of this system is to make health information easily accessible to many people. In this system, the dataset is taken from 3 websites and this dataset used is divided into train and test. Data is obtained after training and testing and this is used in Dot, General, Concat models, the final result is output.

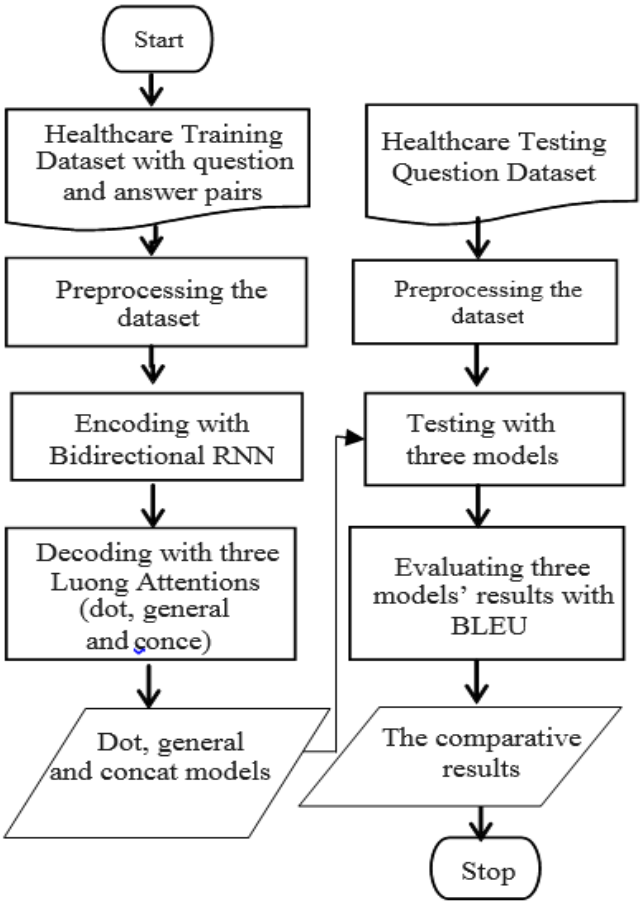

Figure 4.1 Overview of the System

The Figure 4.1 shows the system flow of the proposed system. The eighty percent of training and twenty percent of testing is used in the presented system. The system takes healthcare training dataset with question and answer pairs as input. The input pairs are preprocessed to change over all letters to lowercase and trim all non-letter characters with the exception of essential accentuation. After that all sentence pairs of given dataset are trimmed to count Unique Words. To accommodate sentences of various sizes in a similar clump, sentences are being padding with zero cushion to lounge information tensor of shape (max_length, batch_size). TensorFlow [8], an open-source machine learning library, was used to implement the models. The proposed methodology was evaluated on three healthcare datasets, which contains healthcare question and answer pairs. Most of the answer sentence length are long. Although longer input sequence length increases the accuracy of the model in the training process, it increase execution time of the training process. In this system, the sentence length of the pairs is set with 200. The default hyper parameters used to train nine models based on three datasets and three loung attention alignment functions are as follows.

After preprocessing step, the system encode the dataset with Bidirectional RNN and decode lounge attention decoder according the three lounge attention alignment: dot, general and concat. The system trains these three sequence to sequence models on each three dataset: NCI, NIHSeniorHealth and National Hearth. For three dataset, the system has produced nine models. After creating nine models, the system is asked healthcare information and answers the appropriate answer. To evaluate three models based on three datasets, Bilingual Evaluation Understudy (BLEU) metric is used to compute the quality of system generated answers. The output of BLEU score range is always between 0 and 1, value close to 1 show that the system generated answer is more analogous to the expert generated answer and 0 is no match at all.

4.2 Experimental Result

This is the main form of the proposed system. Firstly, the user enters the system, the home page will look like the picture shown below. The system presented a healthcare question-answering system to provide healthcare information based on three attention based sequence to sequence models: general, dot and concat with three sub datasets: NCI, NIHSeniorHealth and NHLBI in MedQuAD. NCI dataset is a dataset that describes content about cancer, NIHSeniorHealth includes the dataset of health instructions relevant to the older and NHLBI dataset includes health questions and

answers related to heart, blood, lung and sleep issues. In this form, three tabs, radio buttons and buttons are contributed.

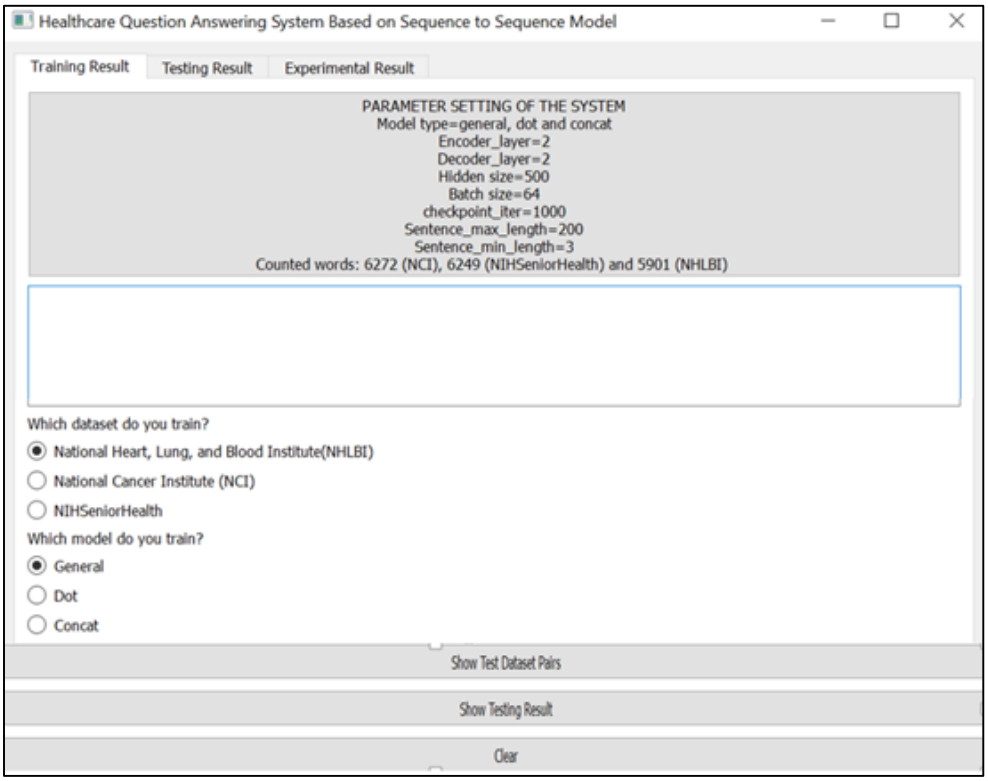

Figure 4.2 Home page of System

NCI dataset is a dataset that describes content about cancer, NIHSeniorHealth includes the dataset of health instructions relevant to the older and NHLBI dataset includes health questions and answers related to heart, blood, lung and sleep issues. This form contains three tabs: training result tab is used to display three testing datasets by clicking the show test dataset pairs and particular dataset name button that shown in figure 4.4 and to display testing result of desire model and dataset by clicking the show testing result button and by clicking desire dataset and model buttons that shown in figure 4.5.

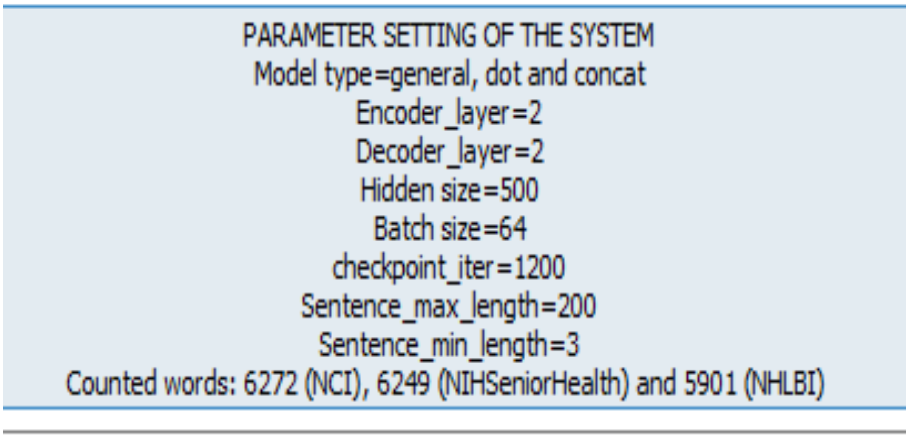

Figure 4.3 Parameter setting of the System

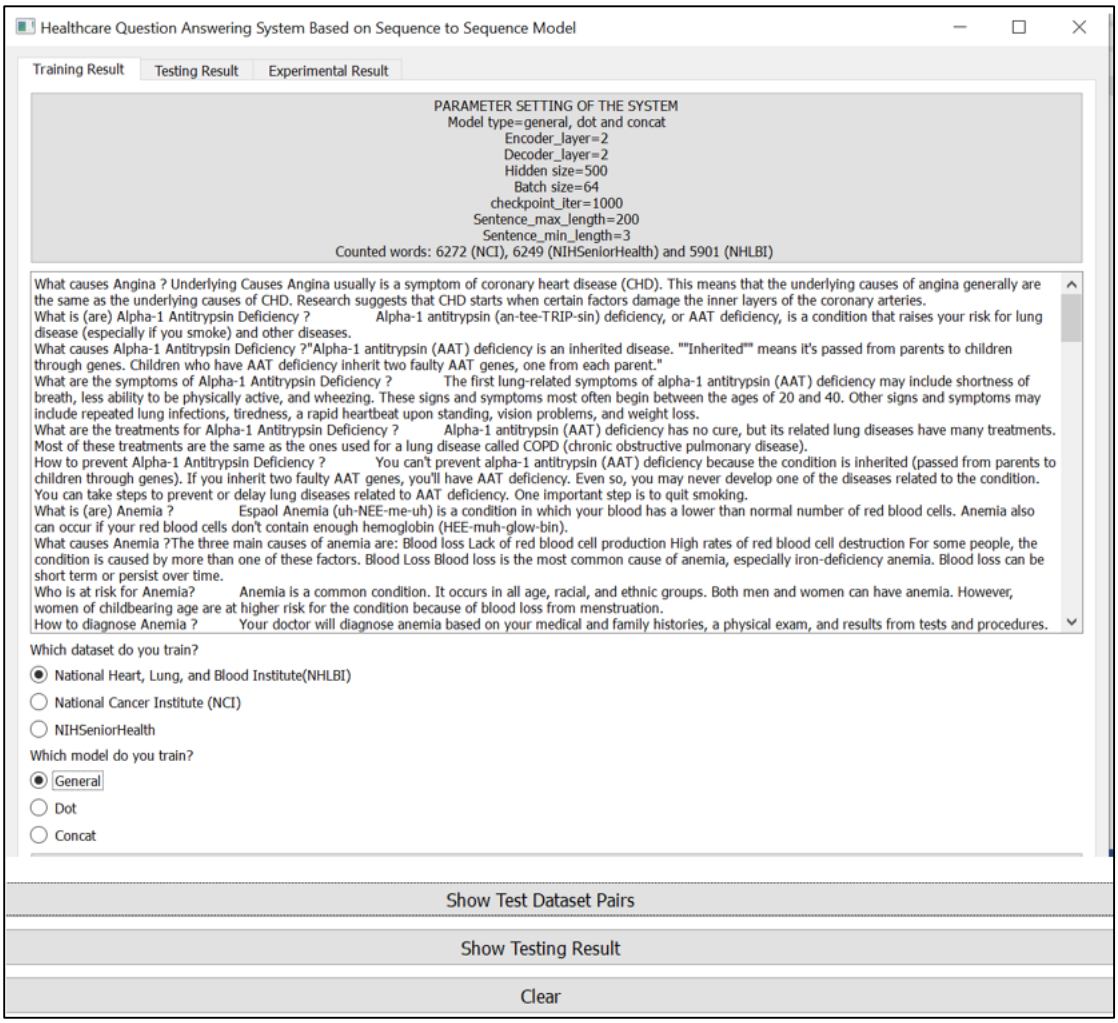

Figure 4.4 Test Dataset Pairs

The Figure 4.3 shown in parameter setting of the system. It demonstrates the method used to build the models step by step. The proposed system is compared with the nine sequence to sequence models based on three datasets. Most of the answer sentence length are long. This system has three models (generic, dot, and concat), which construct two encoders and two decoders, a hidden size of 500, a batch size of 64, an iteration setting of 1200 to create a good model, and sentence of maximum lengths has 200 and minimum lengths has 3 respectively. Although longer input sequence length increases the accuracy of the model in the training process, it increases execution time of the training process. In this system, the sentence length of the pairs is set with 200. Figure 4.4 shows NHLBI test dataset pairs and NCI and NIHSeniorHealth datasets can also be shown by clicking their dataset names.

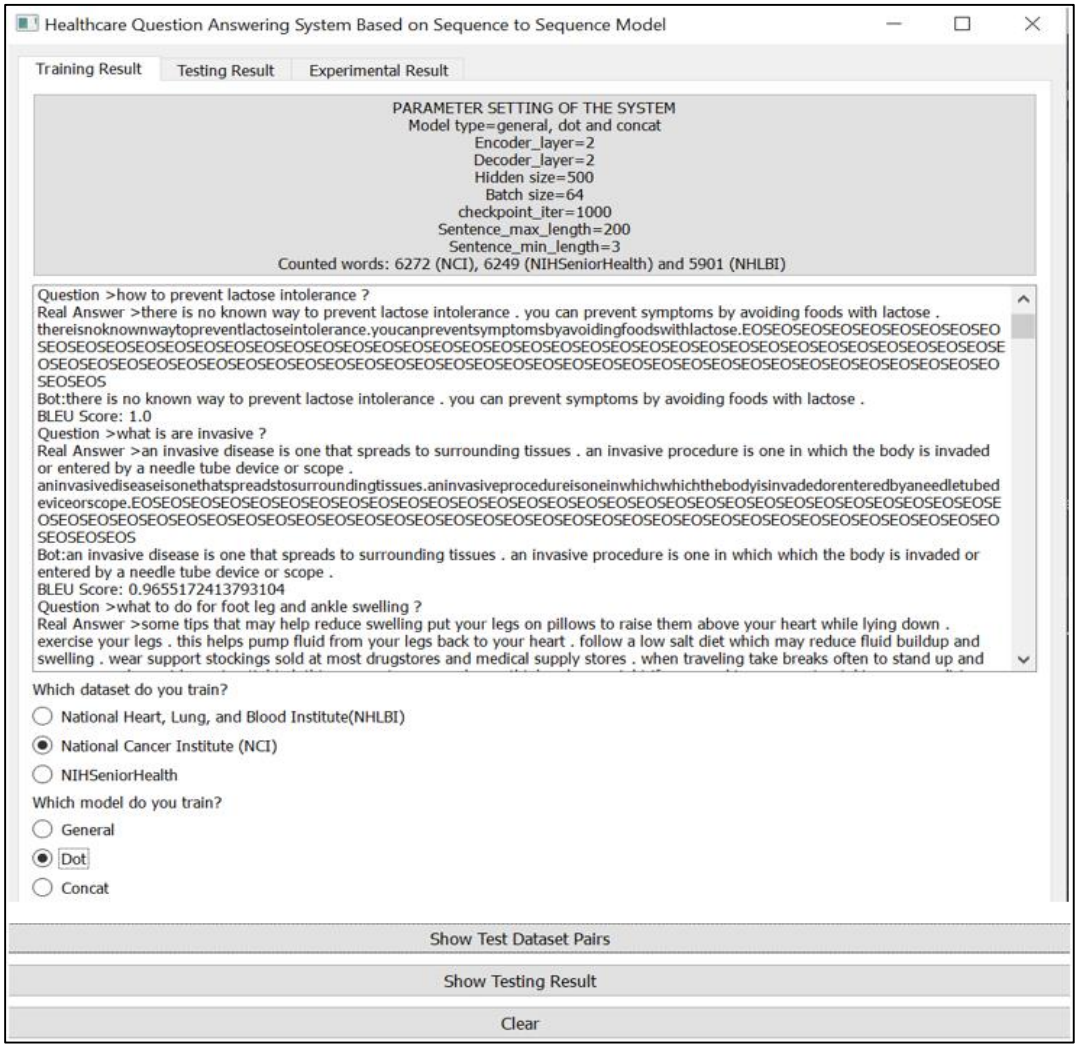

Figure 4.5 Testing Result of Training Model

Figure 4.5 displays testing result of NCI dataset training model according to BLEU score value. Other eight model results can be displayed by clicking particulate model and dataset radio buttons. The average BLEU score of NCI with 145 question and answer pairs can be shown. The average BLEU score of NHLBI with 111 question and answer pairs can be displayed and the average BLEU score of NIHSeniorHealth with 153 question and answer pairs can be described.

In the testing tab, the system test user's input test sentences of three datasets with nine models: NHLBI general sequence to sequence model, NHLBI dot sequence to sequence model, NHLBI concat sequence to sequence model, NCI general sequence to sequence model, NCI dot sequence to sequence model, NCI concat sequence to sequence model, NIHSeniorHealth general sequence to sequence model, NIHSeniorHealth dot sequence to sequence model and NIHSeniorHealth concat sequence to sequence model.

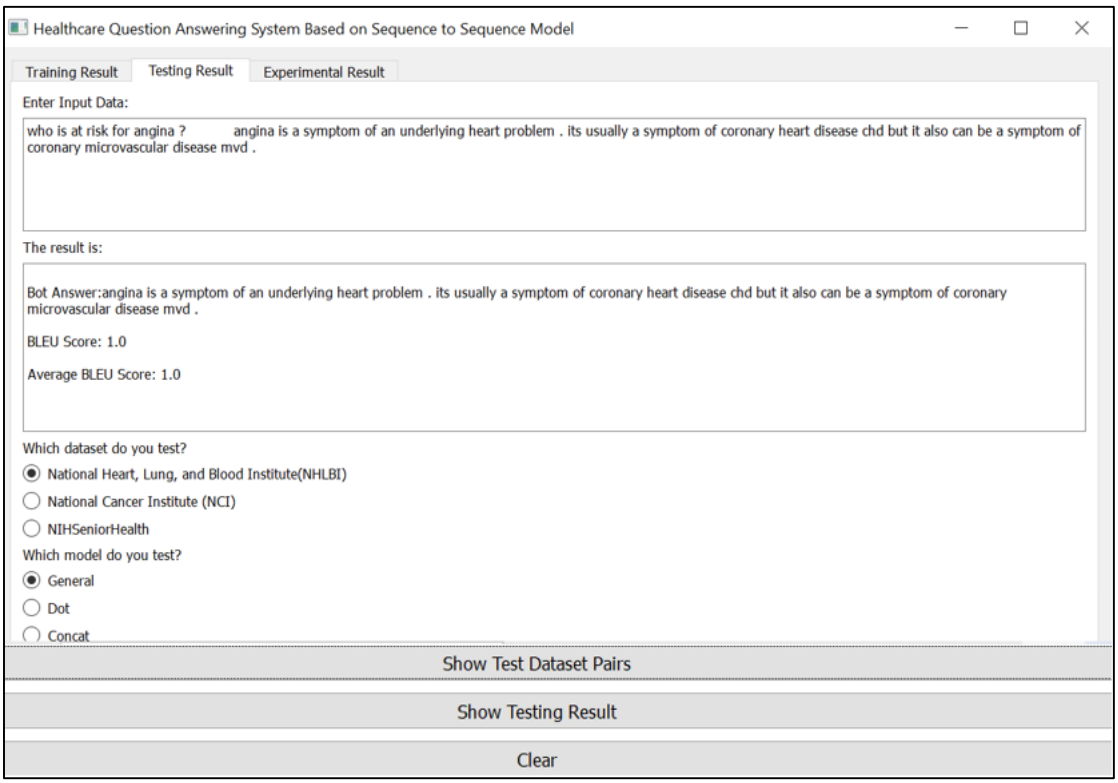

Figure 4.6 Testing Result

In Figure 4.6, NHLBI general sequence to sequence model is used to answer the question: "Who is at risk for angina?". The system can answer this question accurately and BLEU sore of the result is one. Other eight model are used to test healthcare questions about three datasets

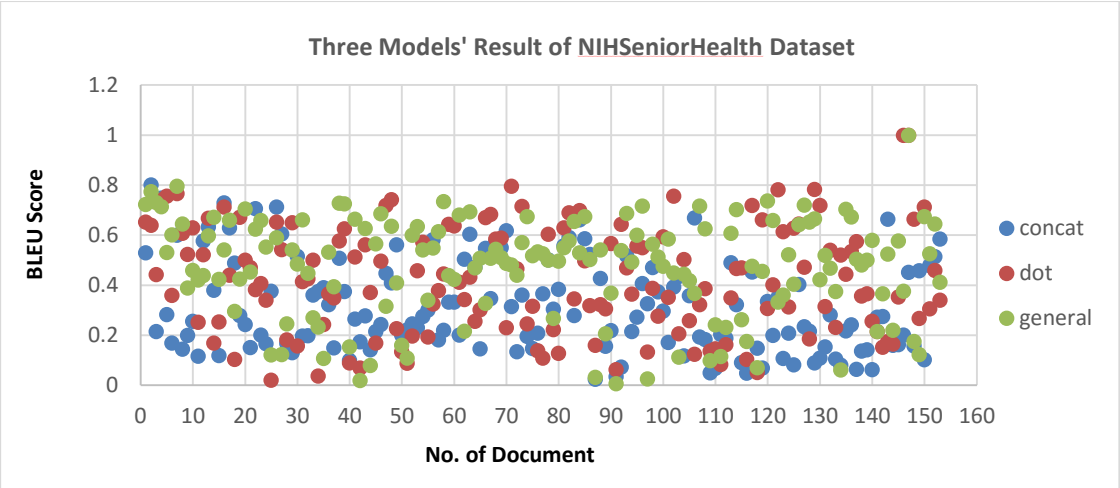

Figure 4.7 Experimental Result of NIHSeniorHealth Dataset

In Figure 4.8, the bot answers 0, 2 and 1 sentences based on concat, dot and general models respectively with 100%.The bot answers 2, 10 and 3 sentences based on concat, dot and general models respectively with greater than or equal to 75%.The bot answers 33, 61 and 84 based on concat, dot and general models respectivelys with greater than or equal to 50%.

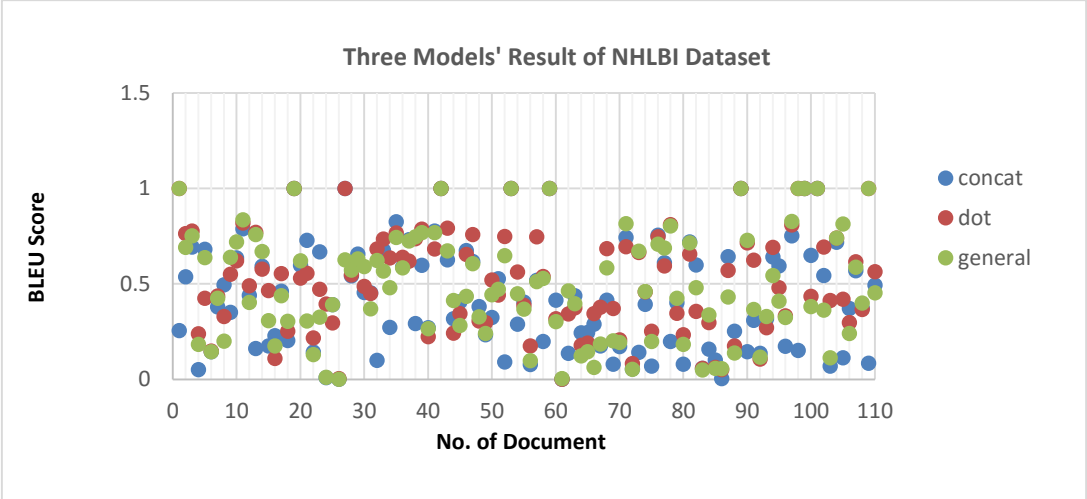

Figure 4.8 Experimental Result of NLHLBI Dataset

In Figure 4.9, the bot answers 8, 11 and 10 sentences based on concat, dot and general models respectively with 100%. The bot answers 14, 25 and 21 sentences based on concat, dot and general models respectively with greater than or equal to 75%.The

bot answers 46, 57 and 49 based on concat, dot and general models respectively with greater than or equal to 50%.

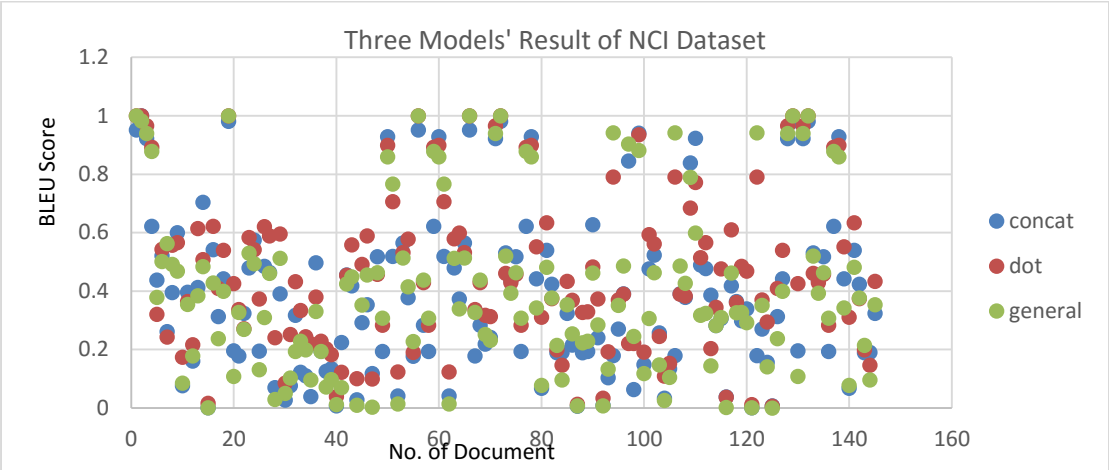

Figure 4.9 Experimental Result of NCI Dataset

In Figure 4.10, the bot answers 1, 8 and 7sentences based on concat, dot and general models respectively with 100%.The bot answers 20, 25 and 28 sentences based on concat, dot and general models respectively with greater than or equal to 75%.The bot answers 43, 58 and 40 based on concat, dot and general models respectively with greater than or equal to 50%.

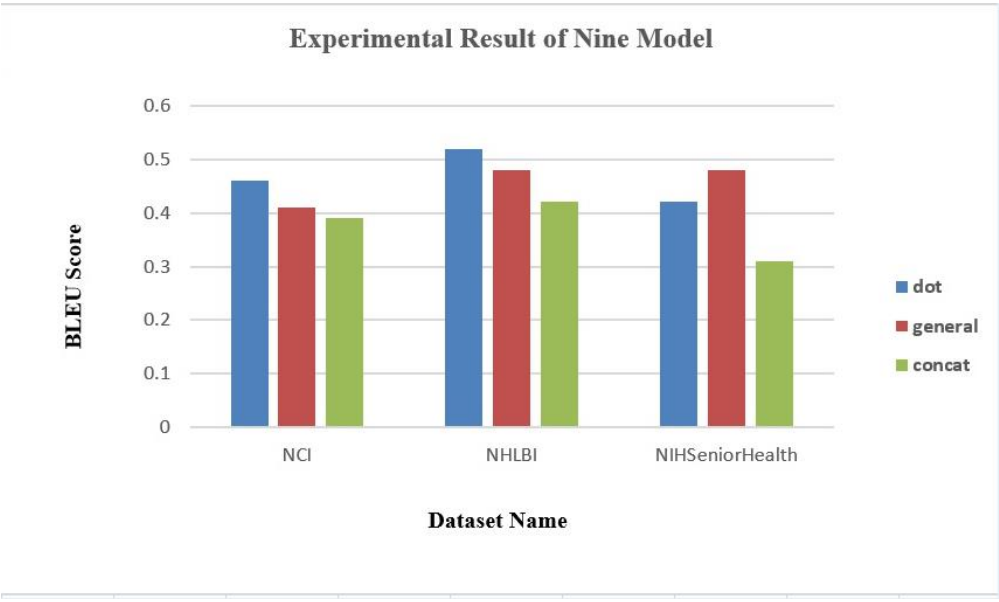

Figure 4.10 Comparing Three Model Result

When you have at least one category or discrete variable, use bar charts to compare model accuracy on different datasets. Longer bars denote greater values, and

each bar represents a summary value for one discrete level. Counts, sums, averages, and standard deviations are a few examples of summary values. Bar graphs are another name for bar charts. Figure 4.11 shown when trained and tested on the same corpus, the three model gave the best results on three datasets (National Cancer Institute (NCI), NIHSeniorHealth and National Heart, Lung, and Blood Institute (NHLBI). Figure 4.11 is shown comparing three model result on the different three datasets. On the NCI datasets, the Dot model is the best performance. On NHLBI datasets, the Dot model is the best performance and the General model is the best performance on NIHSeniorHealth datasets.

4.3 Language

Python was selected as the primary programming language to create the question and answer system after taking the project's environment into consideration. There are a few libraries that have been utilized to put the question-and-answer method into action [10]. The first is the torch library, which main objective is to offer a standard and user-friendly interface for various deep learning frameworks like PyQt5 [15], etc. On top of the PyQt5 library, torch is used in the question and response system. It is helpful for handling a lot of data. Some utilities, including dividing a dataset into train validation and test sets, are employed in this context.

4.4 Chapter Summary

The system's performance changes over time and across various dataset partitions. This chapter displays the results of experiments using the models. Additionally, this chapter displays the accuracy, and weighted average. Also shown the confusion matrices for the multi-class categorization. In summary, these analytical findings have been discussed and the comparison between three models in various splitting data over several training epochs can be shown.

CHAPTER 5
CONCLUSION

The healthcare question answering will be implemented seq2seq model of deep learning to answer right information and adapt self-learning. The system will learn using neural networks where bidirectional RNN one is used as encoder and Luong Attention RNN is used as decoder. According to three Loung Attention Alignment functions: dot, general and concat, this system created nine healthcare question answering models based on three sub datasets: NIHSeniorHealth, NCI and NHLBI in the MedQuAD dataset. This system evaluates three models' performance using BLEU score. According to the experimental result, the accuracy of concat models are lower than the other accuracy of other models. The accuracy of Dot models are higher than the other models. This system uses the same no. of iterations but the models need to be changed no. of iterations according to the size of datasets to make it more accurate. Generally, the model can be trained with increasing no. of hidden layers to make it more accurate and no. of iterations in model training.

5.1 Advantages

A question and answer system is beneficial for patients as they regularly send text messages or emails regarding operation dates, schedules, and appointments. Question and answer system never leaves the patients unattended. They win patients' trust by providing an efficient and prompt response. Question and answer system has good efficiency to transform the healthcare industry. They can considerably boost proficiency besides enhancing the accuracy of detecting symptoms, post-recovery care, preventive care, and feedback procedures.

5.2 Limitation and Further Extension

Although MedQuAD dataset contain 12 sub datasets, the proposed system uses three sub datasets: NCI, NHBLI and NIHSeniorHealth. Another nine sub datasets will be train and test by the proposed system in the future. Future plans include using the same model with the healthcare dataset to develop the other system modules, such as feature extraction, text classification, online healthcare services and response generation. In order to make the question and answer system work in real time, the system will also be tested to identify any areas that need to be addressed. Future work will construct the frontend utilizing the API flask library of python. This model can be

used in future work to accomplish the self-learning part of the fully functional question and answer.

REFERENCES

[1] Chung, K. and R. C. Park , "Chatbot-based heathcare service with a knowledge based on Deep Learning", 22(1), 1925–1937, (2019).

[2] Dhang Luong Hieu Pham Christopher "Effective Approaches to Attention-based Neural Machine Translation", Minh- D.Manning Computer Science Department, Stanford University, Stanford, CA 94305, 2015.

[3] Dataset Link: MedQuaD Collection of 47k QA pairs (Ben Abacha and Demner-Fushman, 2019): https://paperswithcode.com/dataset/medquad.

[4] Dr. Vishwanath Karad , "Research Paper on Chatbot Development for Educational Institute", 8 Jun 2021.

[5] E. Mutabazi, J. Ni, G. Tang, and W Cao, "A review on medical textual question answering systems based on deep learning Approaches," Applied Sciences, vol. 11, no. 12, p. 5456, 2021.

[6] Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J., "LSTM: A search space odyssey". *IEEE Trans. Neural Netw. Learn. Syst.* 2016, *28*, 2222–2232.

[7] Himanshu Gadge, Vaibhav Tode, Sudarshan Madane, Prateek Kachare, Prof. Anuradha Deokar, "A Chatbot for Medical Purpose using Deep Learning(Neural Network)" /e International Journal on Artificial Intelligence Tools, vol. 29, no. 06, Article ID 2050019, 2020.

[8] Jurgita Kapočiūtė-Dzikienė, "A Domain-Specific Generative Chatbot Trained fromLittle Data", March 2020, online Applied Sciences.

[9] Karthik Konar, "A Comparative Study on Chatbot Based on Machine Learning and Lexicon Based Technique" International Journal of May 5,2020.

[10] "Keras, The Python Deep Learning library". Available online, https://keras.io.

[11] Liu, J., Li Y., Lin M., "Review of Intent Detection Methods in the Human-Machine Dialogue System", Conf. Ser. 2019, 1267, 012059."

[12] Oh, K.-J., D. Lee, B. Ko, and H.-J. Choi," A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation". In 2017 18th IEEE International Conference on Mobile Data Management (MDM). IEEE, 2017.

[13] Rojas, I., Joya, G., Catala, A. (eds), “Attention-Based Recurrent Neural Networks (RNNs) for Short Text Classification: An Application in Public Health Monitoring”, Advances in Computational Intelligence. IWANN 2019.

[14] S. Ilya, V. Oriol, and V. L. Quoc, “Sequence to sequence learning with neural networks,” 2014, https://arxiv.org/abs/ 1409.3215.

[15] “TensorFlow,” Available at https://www.tensorflow.org/.

[16] Yu Wang, “A new concept using LSTM Neural Networks for dynamic system identification”, IEEE, Department of Electrical Engineering, Yale University, New Haven.

[17] Zhang, Y., Li, D., Wang, Y., Fang, Y.,Xiao, W.,” Abstract Text Summarization with a Convolutional Seq2seq Model”. Appl. Sci. 2019, 9, 1665.

[18] Zhou, Bolei, et al., "Learning deep features for scene recognition using places database", Advances in neural information processing systems. 2014.

[19] Zsma Ben Abacha, Dina Demner-Fushman, “A Question-Entailment Approach to Question Answering”, 23 Jan 2019.

PUBLICATION

[1] Ei Zin Phyo, Tin Zar Thaw, “Healthcare Question and Answer System Based On Sequence to Sequence Model”, University of Computer Studies, Yangon, Myanmar, 2022.