International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

A ChatBot using Seq2Seq and Bag of Words Model

Dr. Devraj Verma[1], Rakshith Yadhav [2], Sanjana M[3], Varsha T P[4], Vineeth[5]

[1] *Assistant Professor, Dept. of Computer Science and Engineering, Jain University, Kanakapura, Karnataka, India*
[2]*Dept. of Computer Science and Engineering, Jain University, Kanakapura, Karnataka, India*
[3]*Dept. of Computer Science and Engineering, Jain University, Kanakapura, Karnataka, India*
[4]*Dept. of Computer Science and Engineering, Jain University, Kanakapura, Karnataka, India*

Abstract - *Chatbots are useful in automating redundant tasks in an business organization .Chatbot is a project to further learn the ideas and implementation of deep natural language processing also known as DNLP. This chatbot uses the most advanced DNLP model known as sequence2sequence model which are specific types Recurrent neural networks models which produces great results from the chatbot once the chatbot goes through proper training*

1. INTRODUCTION

This project involves in building a ChatBot using two models, from scratch and training it with proper datasets and understanding how recurrent neural networks work. First model is Sequence2Sequence consists of three parts, first part is Data preprocessing which means to clean the data properly so that it could be fed into the network, second part is to build the Sequence2Sequence model which consists of two steps .First step is to develop the encoder of the LSTM's which takes in the input and next is to develop the decoder for the LSTM's which produces us the desired output. The third step is to train the seq2seq model with the dataset and some hyper-parameters like sequence length ,batch size, epochs, Rnn size, learning rate for the model, keep probability, learning rate decay and so on.

And the second model is bag-of-words (BoW), is a way of extracting characteristics/features from text so that it can be used for modeling, for example machine learning algorithms .Bag of Words is basically the representation of words in the document to vector format which includes three basic steps, first a vocabulary of known words and second number of times the known words occurred. And this basically has three simple steps first collect data, second design the vocabulary and third creating the document vector.

All the above mentioned features of the model can be tweaked to get a much more efficient Chatbot it is in the hands of developer to tweak these features to enhance the quality of the chat bot. As mentioned the Chatbot uses RNN which are somewhat different from artificial neural network (ANN's)

Since Recurrent neural networks (RNN's) have the ability to remember the immediate output of their previous layer of neurons which would be very much helpful for our Chatbot to create a kind of context before outputting the result.

1.1 Problem Definition

A business without current trends or technologies used would be really difficult to reach the customers, if required. And also for each and every queries are handled by human that is all the services of the business is taken care by the employees which is time consuming and it would be costly. Due to the above issues most of the businesses are trying to inculcate the current technologies , chatterbot is one among the technologies which is used to provide basic services to the customers(online), which also improves the customer experience and enhances the efficiency of the business and business development.

And a chatbot is a software that simulates human conversation through text or text to speech. Chatbot, is an Artificial Intelligence (AI) feature that can be embedded and used through any major messaging applications. There are many models using which a chatbot can be built, So the problem is to develop a Chatbot using Bag Of Words(BoW) and Sequence to Sequence model to obtain good results And also to have a in depth understanding of Recurrent neural networks , LSTM's and Bag Of Words and Sequence to Sequence model works . For the Chatbot to attain good results it needs to be trained with good CPU for a long time which could be a major challenge.

And once the two models are built successfully, they are compared on which among those two is more efficient and how much human intervention is required.

2. EXISTING SYSTEMS

2.1 If-else:

Very inefficient and redundancy is more. Scope is determined by the developer.

2.2 Audio frequency component analysis:

Sound waves and there frequency are compared to previously recorded frequencies.

2.3 Bag of words model

Bag of words is very popular NLP model. A very brief explanation is given below. Imagine there is a bag containing words now to relate with an example look at the table all the words with a positive result is given 1 and negative result

are given 0. These words are used as a reference to determine future word occurrence designated contacts too. The location is sent via SMS if the GPRS is not functioning.

2.4 Sequence to Sequence Model:

Sequence to Sequence Models are a specific types of recurrent neural networks which allow the Chatbot to retain some type of context so it can have better results from the Chatbot. Its two layers to its architecture one Encoder and one Decoder.

3. METHODOLOGY

To accomplish the research objectives, a systematic process is followed. The research process begins with the identification of the research of research topic where studies was carried out to obtain enough information in the topic. A literature review was carried out to study how artificial intelligence is to create technology that allows computers and machines to function in an intelligent manner. The literature review further looks into the role of a chatbot, which simulates a real interaction with users via a chat interface. And chatbots is being widely absorbed into businesses these days due to its multiple benefits in achieving seamless customer interaction and also the availability of chatbots keep the customer connected whenever is required.

After completing literature review, a survey using close and open-ended questionnaires was carried out to identify whether the chatbots is actually really helpful in businesses. After gathering all the information how chatbots are worked and really useful, this research identifies the importance of building a chatbot by also learning the technology involved in it. Finally, based on overall findings, a chatbot was successfully built using using seq2seq model and Bag of Words model. Then we could able to compare which among them is better and efficient

4. SYSTEM ARCHITECTURE

4.1. Seq2Seq Model:

The Chatbot using Seq2Seq Model which uses long short term memory also known as LSTM's blocks (a type of RNN architecture). As the name suggests, seq2seq takes input as sequence of words (sentence or sentences) and generates an output sequence of words. It does so by use of the recurrent neural network (RNN). Although the vanilla version of RNN is rarely used, its more advanced version i.e. LSTM or GRU are used. This is because RNN suffers from the problem of vanishing gradient. LSTM is used in the version proposed by Google. It develops the context of the word by taking 2 inputs at each point of time. One from the user and other from its previous output, hence the name recurrent (output goes as input). And there are different types of architectures in RNN.

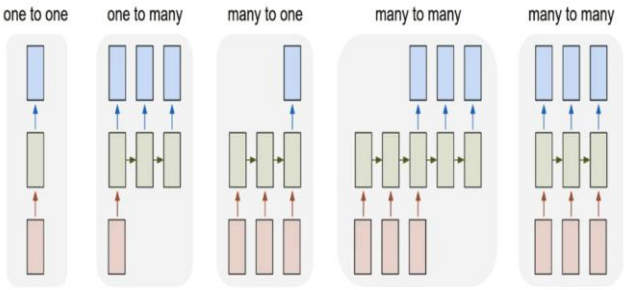

Fig -1: Flowchart of Seq2Seq Architecture

- One to One
- One to Many
- Many to Many

Since the chatbot using seq2seq model requires a variable size input and variable size output, we chose many-to-many RNN architecture. In this model, first a vector of words must be created using numbers and the numbers are associated with the position of the words. The vector will be started with a constant known as SOS (Strat of String) and EOS (End of String). Once the vectors are created, feed the vectors into the RNN which produces the result.

4.2. Bag of Words Model

BoW is basically the way of extracting the characteristics/features from text so that it can be used in building models for example machine learning algorithm. The approach is very easy and flexible. BoW is a representation of words which includes two main things, document of known words and frequency of known words. The BoW can be simple or complex, the complexity comes in how we handle both creating the vocabulary of known words and occurrence of known words. Which has three important steps first collecting the data, second- designing the vocabulary and the last step is creating the document vector.

Step 1: Collecting the data:

It can be any data given by the user, we can treat each and every line as different document.

Step 2: Designing the vocabulary:

Here we can collect the list of all the unique words ignoring case sensitive and punctuation and put that into the model vocabulary.

Step 3: Creating the document vector

Here we check the frequency of words in each document and the main goal of this step is to turn each word in the document into the vector form so that it can be easily used as an input to the other machine learning algorithms. The easiest way of scoring is to mark the presence of words as a Boolean value, 0 for absent and 1for present and then convert it into a binary vector.

5. RESULTS AND DISCUSSIONS

In this chapter we would be discussing the results and outcomes generated by the chatbot for both seq2seq model and also bag of words model and analyze them and also discuss on ways to improve the model for better results.

5.1 Seq2Seq Result

The Chatbot primary model was Seq2Seq and it was trained with appropriate dataset of movie corpus for 2 whole weeks since the dataset was quite big it took a considerable amount of time to show satisfy able results. The current model was planned to run for at least 100 epochs but since that was to CPU intensive it was bought down to around 30 epochs for better CPU utilization and also good results. Finally in the time span of around 2 weeks the model was trained for a total of around 40 epochs.

Once properly trained we run the chatbot in the Anaconda command prompt by using this command **python main.py --test interactive** which then run the main.py file in the interactive mode so we can chat with the chatbot. These are the results which we got from running the tests.

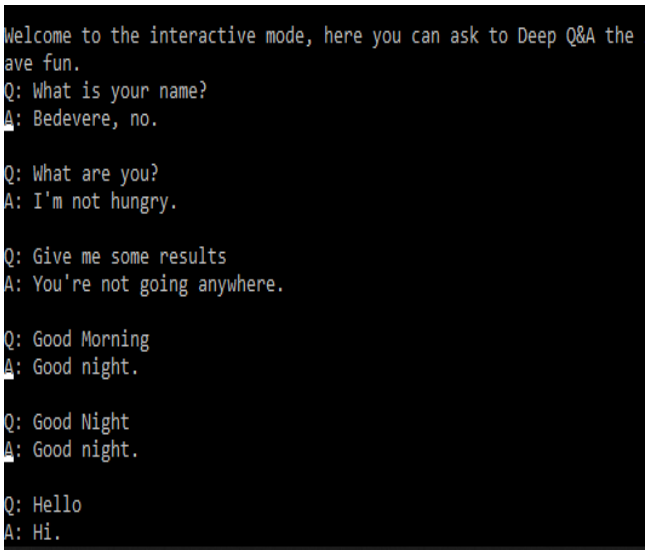

Fig -2: Result of Seq2Seq architecture model

5.2 Result of Bag of Words Model

Since Bag of words model is not a typical neural network based model it doesn't require weeks of training to get the result but an efficient probability machine learning algorithm would suffice. To run this model we need to open Anaconda command prompt and move to it code directory and in the command prompt we need to type python bow.py. On running the command we could interact with the chatbot in command line.

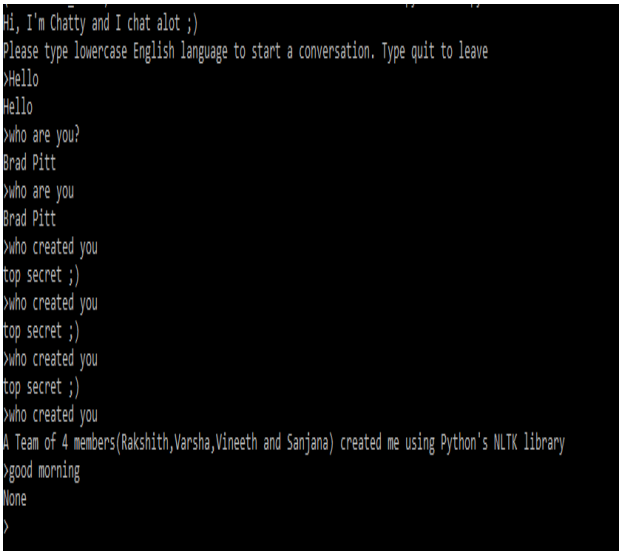

Fig -3: Result of BOW Model

6. CONCLUSIONS

Based on the research objective that had been stated in the earlier information the following are the achievements

The first objective has been successfully completed, which is first to investigate about all the topics related to seq2seq model and BoW model and analyses how it works and its complexity in learning and implementing. The problem that is the chatbot using both the models has been successfully completed

The next objective is to be able to have a deep knowledge about how RNN works which was eventually gotten by implementing the models and during the training of the model we got to know that, the more we train the model the better will it work which is basically educating the bot

The final objective, to be able to build the chatbot using both the models which is successfully done. We could build successfully only because all the relevant information and features were gathered

The last objective was also able to conduct after building both the models successfully

And hereby after building both the models we conclude unlike seq2seq model BoW model doesn't learn with how many ever samples but always chooses the result from specific set of words hence “bag of words”.

REFERENCES

[1] Alex Sherstinsky,"Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network”, Submitted on 9 Aug 2018.

[2] Zachary C. Lipton, University of California, San Diego,"A Critical Review of Recurrent Neural Networks for Sequence Learning", June 5th, 2015.

[3] Iiya Sutskever, Oriol Vinyals, Quoc V Le ,"Sequence to Sequence Learning with Neural Networks", Submitted on 10 Sep 2014, last revised 14 Dec 2014, Cornell University

[4] Sepp Hochreiter Fakultat fur Informatik Technische Universitat Munchen 80290 Munchen, Germany,Jurgen Schmidhuber IDSIA Corso Elvezia 366900 Lugano, Switzerland "LONG SHORT-TERM MEMORY",1997

.

.