iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 Issue: V Month of publication: May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5288

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Ontology based Chatbot for General Conversation using Deep Natural Language Processing

Dr. Ganesh D[1], Mr. Praveen Kumar Shah[2]
[1]Asst. Prof, School of CS & IT, JAIN Deemed-to-be- University, Bangalore
[2]Student, Dept. of MSc IT Jain Deemed-to-be- University, Bangalore

***Abstract:** Chatbots are replacing a number of the roles that were traditionally performed by human workers, like online customer service agents and educators. From the initial stage of rule-based chatbots to the time of rapid development in AI, the performance of chatbots keeps improving.*

The aim of this research is to develop a chatbot for general conversation using Cornell movie corpus dataset, a dataset of more than 600 movies containing thousands of conversations between lots of characters. Moreover, the model can be used to train different datasets to create chatbots in any domain such as chatbots for movie buffs, weather forecasting, taking online appointment with doctor as and more. It deals with building of a super powerful chatbot but by implementing a state of the art and Deep Natural Language processing model. The seq2seq model will be implemented with one of the best API to build deep learning applications or artificial intelligence, which will be tensor flow and generate a chatbot for general conversation like a friend.

Keywords: Big data, Machine Learning, Deep Learning, RNN

I. INTRODUCTION

A chatbot or conversational agent (CA) may be a package, which will interact or "chat" with a personality's user in linguistic communication like English (1). The primary chatbot ELIZA was designed by Joseph Weizenbaum in 1966 (2). It was originally created to simulate a psychotherapist (3). After ELIZA, there are different chatbots created, such as A.L.I.C.E. and Mitsuku. Most of those chatbots are developed using AI terminology (AIML), which may be an artificial language that permits chatbots to acknowledge patterns within the input sentences. It will respond with sentences from a template ranging from 1991, many chatbots are developed for the Loebner Prize competition, which is that the oldest Turing Test contest to seek out the chatbot considered by the judges to be the foremost human-like. Mitsuku and Rose are the two-chatbot winners in recent years. a good range of literature review was done on previous work associated with chatbot. Chatbots have gotten popular and that they are now creeping into our smartphones. People spend a lot of time on the apps installed in the smartphones every day. According to a recent report by Flurry Analytics (4), the overall app usage in 2016 grew by 11 percent compared to 2015, and therefore the time that users spent in apps grew by 69 percent. The time spent in messaging and social apps increased by 394 percent in 2016 compared to 2015. In China, senior users of mobile phones might not know the way to use every apps; however, most of them use the chatting app, like WeChat, frequently (5). The purpose of this research is to develop an ontology- based chatbot for general conversation using Cornell movie corpus dataset, a dataset of more than 600 movies containing thousands of conversations between lot of characters. Moreover, the model can be used to train different datasets to create chatbots in any domain such as chatbots for movie buffs, weather forecasting, taking online appointment with doctor as and more. It deals with building of a super powerful chatbot but by implementing a state of the art and Deep Natural Language processing model.

It will be the seq2seq, which will be implemented with one of the best API to build deep learning applications, or artificial intelligence, which will be tensor flow and generate a chatbot for general conversation like a friend. The plan of attack involves the common five parts for generating applications of artificial intelligence. The first part involves installing anaconda and getting the dataset, anaconda is the Ide that is used to build the chatbot and to be more precise spyder inside anaconda because it is like a studio with great tools that is useful while implementing the chatbot. Then get the dataset that is Cornell movie corpus dataset a dataset of more than 600 movies containing thousands of conversations between lots of characters. The chatbot will be trained on this dataset because the main objective is to build chatbot that can have general conversation with humans like a friend to give movies. However, the model that will be developed will be used to train different datasets for some other purposes. For example, we will be able to train the same chatbot on a more specific dataset to build chatbot as calendar assistant or a navigation assistant. Moreover, chatbot is trained to talk about everyday conversations and that is why movies are perfect because in movies you have many random conversations, general conversations between friends. The second part involves with data preprocessing, which is inevitable whenever we build an AI or whenever you build a machinery model, you have to make the dataset compatible with the model you are going to build. Since, we are building a neural network based model and therefore the data will have to have a special format especially for the inputs. Besides, you have to clean the text because the less you clean it and simplify it the more difficult it will be for a model to train itself to talk like a human. The third part deals with building the Seq2Seq model which is a state of the art deep NLP model which deals with building a brain composed of encoder and decoder and then you have to assemble all of them to build a final brain which is not trained in this part. The fourth part deals with training this seq2seq model, training the brain made in part three. Therefore, we will train it, set up the functions, get the optimizer and then apply some to update the weight of the neurons of the brain so that it improves its ability to talk with us. Finally, the fifth part of this implementation; we will test the chatbot, which deals with testing the seq2seq model. We will be designing some kind of code which when executed have an interface so we can ask some questions and then the chatbot will answer and then we will test the chatbot by observing its answers and see how its capable of conserving with us. Even though they will not understand how to type text messages on the mobile phones, they will use voice messages and pictures to precise their ideas.

Recent chatbots are embedded in chatting apps or web content, which enable tasks to be accomplished through conversations, on one mobile device rather than having an oversized computer and other peripheral devices, like mouse and keyboard. These are among the important reasons why chatbots became popular. Research is conducted on chatbots using the quantitative bibliometric analysis to assist researchers to spot research gaps for the longer-term research agenda in chatbots. The results of the analysis found a possible research opportunity in chatbots because of the emergence of the deep learning technology. This new technology may change the direction of future research in chatbots (6).

Many investors understood the potential of AI and they have made significant investments to harness the technology. Investors include Li Ka-Shing, the urban center billionaire. He invested in several startups that specialize in AI (7); Jack Ma, the founding father of Alibaba. He invested in an Israeli startup using AI to evolve ecommerce search technologies (8); Dr. Kai-Fu Lee, a famous tech investor. He invested in several investments on AI startups that target the event of AI (9).Study shows two main goals were achieved from the automation process. One was the pliability to come back up with different versions of the chatbot in numerous languages, bringing chatbot technology to languages with few if any NLP resources: the corpus-based learning techniques transferred straightforwardly to develop chatbots for Afrikaans and Qur'anic Arabic. The second achievement was the pliability to be told a extremely sizable amount of categories within a quick time, saving effort and errors in doing such work manually which generated over 1,000,000 AIML categories or conversation-rules from the BNC corpus, 20 times the scale of existing AIML rule-sets, and possibly the most important AI Knowledge-Base ever (10). Corpora are widely employed by linguists to develop and refine "language models", descriptions of lexis, grammar, dialogue, etc. Language models also can be automatically extracted or machine-learnt from corpora, to drive language analysis systems; as an example, machine learning of Part-of-Speech taggers from PoS-tagged corpora, machine learning to automatically cluster words in a very corpus into grammatical classes (11).

II. HISTORY OF CHATBOTS

The idea of chatbot systems originated within the Massachusetts Institute of Technology (12), where Weizenbaum implemented the ELIZA chatbot to emulate a psychotherapist. The concept was simple and supported keyword matching. The input is inspected for the presence of a keyword. If such a word is found, the sentence is mapped in step with a rule related to the keyword; if not, a connected free remark, or under certain conditions an earlier transformation, is retrieved. For instance, if the input includes the keyword "mother", ELIZA can respond, "Tell me more about your family". This rule is inspired by mother and family are central to psychological problems so a therapist could encourage a patient to open up about their family but the ELIZA program doesn't really 'understand' this psychological strategy, it merely matches the keyword and regurgitates a regular response. The following major program was PARRY. In contrast to ELIZA, rather than simulating a psychotherapist, PARRY modelled a paranoid patient during an interview together with his therapist (13). Both ELIZA and PARRY use certain tricks to be ready to successfully perform in conversations. ELIZA directs the conversation far from herself by asking questions. ELIZA uses parts of the user's input within the output questions and seems to be following the conversations. Additionally to those techniques, PARRY has little stories to tell and have an inclination to insert these within the conversation. Moreover, A.L.I.C.E. is that the Artificial Linguistic Internet Computer Entity, first implemented by Wallace in 1995. ALICE knowledge about English conversation patterns is stored in AIML files. Artificial Intelligence Mark-up Language could also be a derivative of Extensible Mark-up Language (XML), developed by Wallace and so the Alicebot free software community during 1995–2000 to enable people to input dialogue pattern knowledge into chatbots supported the ALICE open-source software technology (14).

III. ARTIFICIAL NEURAL NETWORKS

A wide range of research goes on in Artificial Neural Networks (ANNs) to resolve a spread of problems in pattern recognition, prediction, optimization, associative memory, and control. Customary methodologies are proposed for taking care of those issues. Although effective applications will be found in certain well-constrained situations, none is adaptable enough to perform well outside its space. ANNs give energizing options, and diverse applications could profit by utilizing them (15). Human beings are endowed characteristics not present in modern parallel computers which incorporates massive parallelism distributed representation and computation, intelligence, generalization ability, additivity, inherent contextual scientific discipline, fault tolerance and low energy consumption. (16). However, humans can easily be sure of complex perceptual issues (like perceiving a person during a group from a minor have a look at his face) at such a rapid and degree on overshadow the world's quickest PC. For what reason is there such a momentous contrast in their exhibition? Inspired by biological neural networks, ANNs are massively parallel computing systems consisting of an especially huge number of straightforward processors with numerous interconnections. ANN models endeavor to utilize some "hierarchical" standards accepted to be utilized within the human. Figure 1 below shows A neuron (or nerve cell) could be a unique organic cell that forms data (see Figure 3.1). It is made out of a cell body, or soma, and two varieties of out-reaching at tree-like branches: the axon and the dendrites. The cell body features a nucleus that contains data about genetic or hereditary characteristics, a plasma that holds the atomic gear for creating material required by the neuron. A neuron gets signals (driving forces) from different neurons through its dendrites (receivers) and transmits sign created by its cell body along the axon (transmitter), which within the long-term branches into strands and substrands. At the terminals of those strands are the neural connections. A neurotransmitter may be a basic structure and useful unit between two neurons (an axon strand of 1 neuron and a dendrite of another), When the impulse receives at the neural connection's terminal, certain chemical substances called neurotransmitters are discharged. The neurotransmitters diffuse across the synaptic gap, to reinforce or inhibit, looking on the kind of the synapse, the receptor neuron's own tendency to emit electrical impulses that sheds light to understand the functionality of the algorithms. The synapse's effectiveness are often adjusted by the signals passing through it so the synapses can learn from the activities during which they participate. This dependence on history acts as a memory, which is possibly chargeable for human memory.

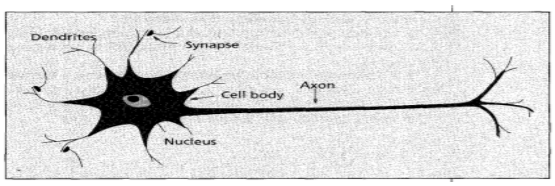


Fig 3.1 a sketch of biological neuron (16)

The cerebral cortex in people around two to three millimeters thick with a surface zone of about 2,200 cm^2, about double the territory of a regular PC console. The cerebral cortex contains about 10"neurons, which is approximately the amount of stars within the Milky Way Galaxy." Neurons are massively associated, substantially more amazing and thick than phone systems. Every neuron is related to 1000 to l0000 different neurons. Altogether, the human cerebrum contains roughly 10^{14} to $l0^{15}$ interconnections. Neurons impart through an exceptionally short train of beats, commonly milliseconds in duration. The message is regulated on the beat-transmission frequency. This recurrence can change from a pair to some hundred hertz, which could be a million times slower than the quickest exchanging pace in electronic circuits. In any case, complex perceptual choices as an example face acknowledgment that are commonly made by humans within a pair of hundred milliseconds. These choices are made by a network of neurons whose operational speed is simply a pair of .milliseconds (16).

IV. RECURRENT NEURAL NETWORKS (RNNs)

The recurrent neural network-based language model gives further speculation: instead of considering only some going before words, neurons with contribution from recurrent connections are assumed to represent immediate memory. The model takes in itself from the data the way to speak to memory. While shallow feed forward neural networks (those with just one hidden layer), can just bunch comparative words, repetitive neural system (which may be considered as a profound design can perform clustering of comparable histories that enables as an example efficient representation of patterns with variable length (17). There has been plenty of effort within the field of statistical language modeling. Among models of linguistic communication, neural network based models perceived to outperform most of the competition, and were showing steady improvements in state of the art speech recognition systems (18).

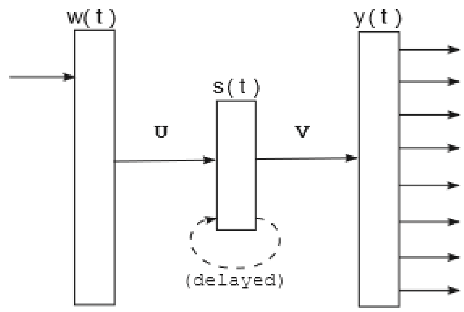


Figure 4.1 Simple recurrent neural network (18)

Among many following modifications of the first model, the recurrent neural network-based language model provides further generalization: rather than considering just several preceding words, neurons with input from recurrent connections are assumed to represent remembering. The model learns itself from the information the way to represent memory. While shallow feed forward neural networks (those with only one hidden layer) can only cluster similar words, recurrent neural network, which may be considered as a deep architecture, can perform clustering of similar histories (19). This permits for example efficient representation of patterns with variable length. During this work, it highlights the importance of the Backpropagation through time algorithm for learning appropriate remembering. Then, the way to further improve the first RNN LM by decreasing its computational complexity is discussed. In the end, we briefly discuss possibilities of reducing the scale of the resulting model. The recurrent neural network described in is additionally called Elman network (20). Its architecture is shown in figure 4.1. The vector x (t) is formed by concatenating the vector w (t) that represents the current word while using one of N coding (thus its size is equal to the size of the vocabulary) and vector s (t − 1) that represents output values in the hidden layer from the previous time step. The network is trained by using the standard backpropagation and contains input, hidden and output layers. Values in these layers are computed follows:

$$x(t) - [w(t)^T s(t-1)^T]^T \ldots\ldots\ldots\ldots (1)$$

$$s_j(t) - f\left(\sum_i x_i(t) u_{ji}\right) \ldots\ldots\ldots\ldots (2)$$

$$y_k(t) = g\left(\sum_j s_j(t) v_{kj}\right) \ldots\ldots\ldots\ldots (3)$$

V. Seq2Seq MODELLING

The ability to grasp one's environment, essential for intelligence, is not static. The order within which events occur are often even more important than the events themselves, and an intelligent system, whether or not it's a frog, a robot, or a human, must be able to detect this ordering and to breed this ordering on some cue. Yet many attempts to model neural networks, like associative memory (21) and therefore the Boltzmann machine (22), dealt only with static equilibrium instead of with ordering of patterns. Generally, a temporal sequence S is defined as: p1 - p2_... - pm. Each p, (i = 1.., m) is called a component of S (sometimes we call it a spatial pattern, or just a symbol). The length of a sequence is the number of components in the sequence. In general, a sequence may include repetitions of the same subsequence in different contexts. For example, S: C-A-B-D-A-B-€ contains repetitions of subsequence A-B, and such a subsequence is called a recurring subsequence. The correct successor will be determined only by knowing symbols before the present one. We talk to the prior subsequence required to breed the present symbol p, in Sas the context of p, and therefore the length of this prior subsequence because the degree of p. The symbol D in SI, as an example, incorporates a degree of three. The degree of a sequence is defined because the maximum degree of its components. A 1-degree sequence is termed a straightforward sequence, and otherwise a sequence may be a complex sequence. If a recurring subsequence of S contains in itself another recurring subsequence, e.g. A-B-A in A-B-A-C-A-B-A-D, S is called a high-order complex sequence, otherwise a first-order complex sequence. Neural networks to store and recognize a temporal sequence of input stimuli are previously studied. Grossberg (23) demonstrated one neural network called the outstar avalanche which will be wont to generate temporal patterns. The outstar avalanche consists of n sequential outstars. Any outstar 3n, can store a spatial pattern and be activated by a sign within the vertex v, these vertices are connected as v1 -+ v2 + . * . -+ v, and a sign from v, arrives with some delay at v,+,. Soan initial signal at v1 can produce sequentially the spatial patterns stored in 3n1, 3n2, . * , 3n, respectively. It supported the anatomy of the dentate gyrus region of the mammalian hippocampus, Stanley and Kilmer (24) designed a network called the wave model, which may learn sequences of inputs separated by certain time intervals and reproduce these sequences when cued by their initial subsequences. Recently, employing a bidirectional associative memory built from two fields of fully connected neurons, Kosko (25) showed that by feeding the spatial pattern output from one field back to the opposite field, the network could generate a sequence of patterns over time that alternates between the two fields.

VI. LITERATURE SURVEY

Nguyen et al. (26) presented a paper titled “A Neural Chatbot with Personality”. They experiment building open-domain response generator with personality and identity. They built chatbots that imitate characters in popular TV shows: Barney from How I Met Your Mother, Sheldon from the massive Bang Theory, Michael from The Office, and Joey from Friends. Lin et al. (27) presented a paper titled “A Web-based Platform for Collection of Human-Chatbot Interactions”. They describe a classy platform for evaluating and annotating human-chatbot interactions, its main features and goals, similarly because the longer term plans we have got for it. During this paper, they have introduced WebChat, up to the foremost effective of our knowledge, the first crowd-source initiative to assemble and annotate human-chatbot interactions. Caspi et al. (28) presented a paper titled “A step Towards Sequence- to-Sequence Alignment”. This paper presents an approach for establishing correspondences in time and in space between two different video sequences of the identical dynamic scene, recorded by stationary uncalibrated video cameras. Jain et al. (16) presented an editorial titled “Artificial Neural Networks: A Tutorial “. This text is for those readers with little or no knowledge of ANNs to assist them understand the opposite articles during this issue of Computer. They discuss the motivations behind the event of A”s, describe the essential biological neuron and therefore the artificial computational model, outline net- work architectures and learning processes, and present a number of the foremost commonly used ANN models. Io et al. (6) presented a paper “Chatbots and Conversational Agents: A Bibliometric Analysis”. The contribution of this research is to assist researchers to spot research gaps for the longer-term research agenda in chatbots. The results of the analysis found a possible research opportunity in chatbots because of the emergence of the deep learning technology. This new technology may change the direction of future research in chatbots. Several recommendations for future research are provided supported the results obtained from their analysis. Wang et al. (29) presented a paper “Complex Temporal Sequence Learning supported Short Term Memory”. They design neural networks to be told, recognize, and reproduce complex temporal sequence, with remembering (STM) modeled by units comprising recurrent excitatory connections between two neurons (a dual neuron model). Shawat et al. (30) presented a paper “Using corpora in machine-learning chatbot systems”. This paper presents a program to find out from spoken transcripts of the Dialogue Diversity Corpus of English, the Minnesota French Corpus, the Corpus of Spoken Afrikaans, the Qur’an Arabic-English parallel corpus, and the British National Corpus of English; Mikolov et al. (31) presented a paper “Extensions of Recurrent Neural Network Language Model”. They need shown approaches that result in quite 15 times speedup for both training and testing phases. Next, they need shown importance of employing a back-propagation through time algorithm. An empirical comparison with feed forward networks is additionally provided. In the end, they need to discuss possibilities the way to reduce the number of parameters within the model. The resulting RNN model can thus be smaller, faster both during training and testing, and more accurate than the essential one. The comparison to plain feed forward neural network based language models, furthermore as comparison to BP trained RNN models show clearly the potential of the presented model. Furthermore, they need shown the way to obtain significantly better accuracy of RNN models by combining them linearly. Shum et al. (32) presented a paper “From Eliza to XiaoIce: challenges and opportunities with social chatbots”. This study deals with Conversational systems that have come a protracted way since their inception within the 1960s. After decades of research and development, we have seen progress from Eliza and Parry within the 1960s and 1970s, to task-completion systems as within the Defense Advanced Research Projects Agency (DARPA) communicator program within the 2000s, to intelligent personal assistants like Siri, within the 2010s, to today’s social chatbots like XiaoIce. Elibol et al. (33) presented a paper “Learning Language Models of Movie Characters”. They investigated learning language models of individual movie characters. They train a recurrent neural net-based model on an out sized dataset of movie scripts with no character specificity to find out a general dialogue model first. Then, they need to transfer the parameters from this pre-trained model to initialize another model and learn a personality specific model from one show. Chen et al. (34) presented a paper “Mitigating the impact of Speech Recognition Errors on Chatbot using Sequence-to-Sequence Model”. They applied sequence-to-sequence model to mitigate the impact of speech recognition errors on open domain end-to-end dialog generation. Moreover, they cast the task as a site adaptation problem where ASR transcriptions and original texts are in two different domains. During this paper, their proposed model includes two individual encoders for every domain data and make their hidden states kind of like make sure the decoder predict the identical dialog text. The strategy demonstrates that the sequence-to-sequence model can learn the ASR transcriptions and original text pair having the identical meaning and eliminate the speech recognition errors. Experimental results on Cornell movie dialog dataset demonstrate that the domain adaption system help the spoken dialog system generate more similar responses with the initial text answers. Zhou et al. (35) presented a paper "Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network". They investigate matching a response with its multi-turn context using dependency information based entirely on attention. The recently proposed Transformer in machine translation inspires their solution and that they extended the eye mechanism in two ways. Zhou et al. (36) presented a paper "Predicting effects of non-coding variants with deep learning–based sequence model". To predict the non-coding variant effects *de novo* from sequence, they developed a deep learning–based algorithmic framework, DeepSEA that directly learns a regulatory sequence code from large-scale chromatin-profiling data, enabling prediction of chromatin effects of sequence alterations with single-nucleotide sensitivity

VII.PROBLEM STATEMENT

The previous methods accustomed build chatbot using sequence-to-sequence modelling or another machine learning algorithms generated chatbots for specific domains, and totally on small dataset that were defined manually during training. On the opposite hand, this proposed work can generate chatbot for general conversations for huge amount of dataset. Moreover, the proposed model can generate chatbots for other domains like calendar assistant or a navigation assistant etcetera.

VIII. EXISTING SYSTEM

Some previous work's sort of a step Towards Sequence- to- Sequence Alignment dole out by Irani, Yaron Caspi and Michal an approach for establishing correspondences in time and in space between two different video sequences of the identical dynamic scene, recorded by stationary uncalibrated video cameras. The strategy simultaneously estimates both spatial alignment in addition as temporal synchronization (temporal alignment) between the two sequences, using all available spatio-temporal information. However, they used sequence to sequence modelling for alignment without deep learning. Other previous work includes Mitigating the impact of Speech Recognition Errors on Chatbot using Sequence-to-Sequence Model where sequence-to-sequence modelling was wont to mitigate the impact of speech recognition errors on open domain end-to-end dialog generation. Moreover, they cast the task as a site adaptation problem where ASR transcriptions and original texts are in two different domains. During this paper, their proposed model includes two individual encoders for every domain data and make their hidden states like make sure the decoder predict the identical dialog text. The strategy demonstrates that the sequence-to-sequence model can learn the ASR transcriptions and original text pair having the identical meaning and eliminate the speech recognition errors. Experimental results on Cornell movie dialog dataset demonstrate that the domain adaption system help the spoken dialog system generate more similar responses with the initial text answers.

IX. PROPOSED SYSTEM

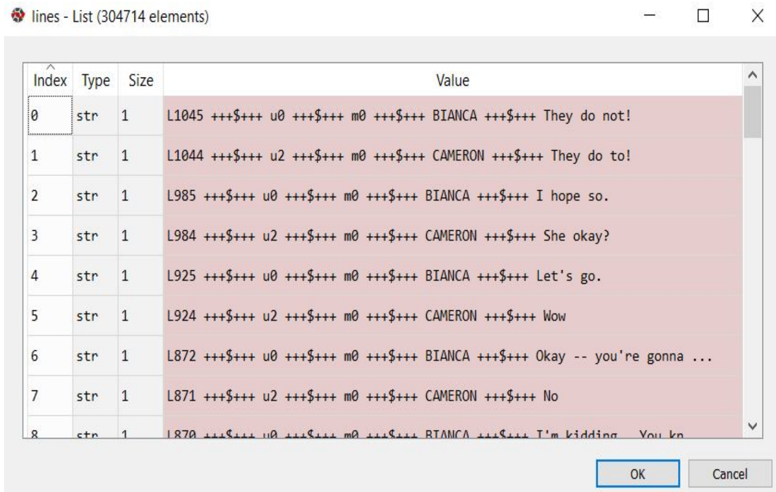
The proposed model deals with building of a super powerful chatbot but by implementing a state of the art and Deep Natural Language processing model that will be the seq2seq and it will be implemented with one of the best API to build deep learning applications or artificial intelligence, which will be tensor flow.

X. RESULTS AND DISCUSSION

This section presents significant results produced through data preprocessing, training and testing the model. Final results was produced to generate a chatbot in a IPythonConsole using Cornell movie corpus dataset, a dataset of more than 600 movies containing thousands of conversations between lot of characters and chatbot will be trained on this dataset because the main objective was to build chatbot that can have general conversation with humans like a friend. Figure 10.1 & 10.2 shows importing the datasets into spyder IDE. Word2Count Dictionary is generated that maps each word to its number of occurrences by removing not so much frequent words of our corpus. It will optimize the training we need essential word of the vocabulary then translation of all the questions and the answers into integers and replacing all the words that were filtered out by token OUT. Sorting of clean questions and answers which will speed up the training and reduce the amount of padding. The length of questions sorted the questions and answers. Figure 10.3 & 10.4 shows splitting the questions and answers into training sets and figure 10.5 & 10.6 shows splitting the questions and answers into validation sets. Eighty-five percent data were used for training and 15 percent for validation by shuffling the data sets. However, due to some inconsistencies in results, the data sets were not divided into sixty by forty ratio. Below results shown in figure 10.7 shows our final chatbot, which is obtained by tuning the model many numbers of times. Since large amount of datasets were trained for three days using GPU instead of CPU. It is able to provide results but not all answers are grammatically correct. Moreover, it is quite a time-consuming process.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

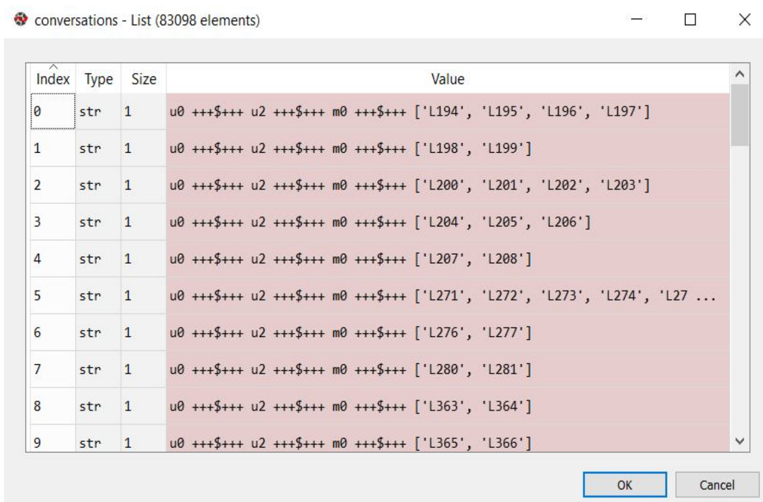
A. Datasets


lines - List (304714 elements)

Index	Type	Size	Value
0	str	1	L1045 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ They do not!
1	str	1	L1044 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ They do to!
2	str	1	L985 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ I hope so.
3	str	1	L984 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ She okay?
4	str	1	L925 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Let's go.
5	str	1	L924 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ Wow
6	str	1	L872 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Okay -- you're gonna ...
7	str	1	L871 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ No

OK Cancel

Figure 10.1 Lines dataset

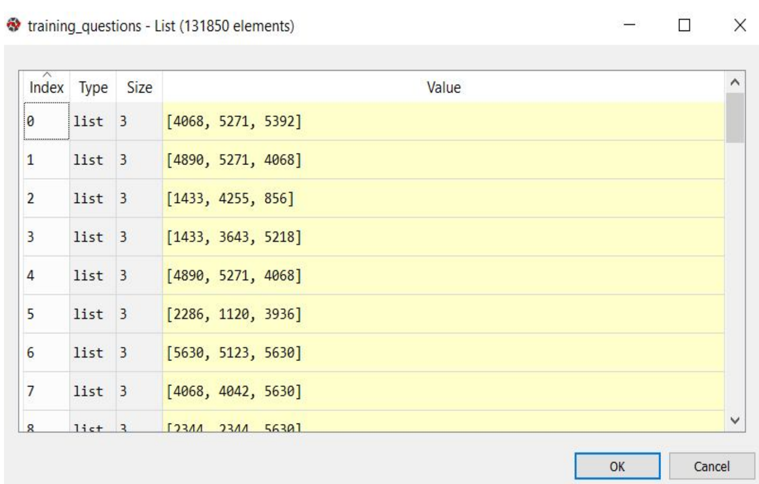

conversations - List (83098 elements)

Index	Type	Size	Value
0	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L194', 'L195', 'L196', 'L197']
1	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L198', 'L199']
2	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L200', 'L201', 'L202', 'L203']
3	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L204', 'L205', 'L206']
4	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L207', 'L208']
5	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L271', 'L272', 'L273', 'L274', 'L27 ...
6	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L276', 'L277']
7	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L280', 'L281']
8	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L363', 'L364']
9	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L365', 'L366']

OK Cancel

Figure 10.2: Conversations dataset

B. Training Questions


training_questions - List (131850 elements)

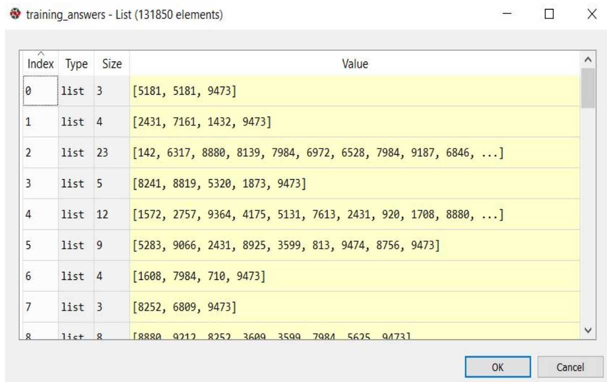
Index	Type	Size	Value
0	list	3	[4068, 5271, 5392]
1	list	3	[4890, 5271, 4068]
2	list	3	[1433, 4255, 856]
3	list	3	[1433, 3643, 5218]
4	list	3	[4890, 5271, 4068]
5	list	3	[2286, 1120, 3936]
6	list	3	[5630, 5123, 5630]
7	list	3	[4068, 4042, 5630]

OK Cancel

Figure 10.3: Training questions

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

C. Training Answers

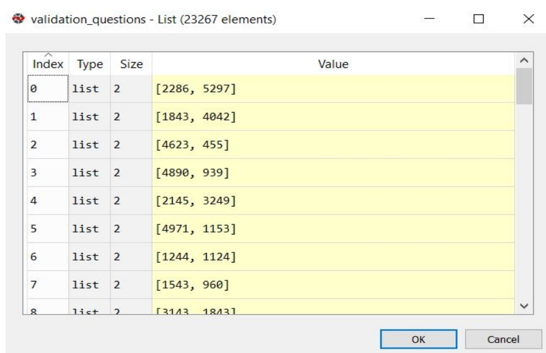

training_answers - List (131850 elements)

Index	Type	Size	Value
0	list	3	[5181, 5181, 9473]
1	list	4	[2431, 7161, 1432, 9473]
2	list	23	[142, 6317, 8880, 8139, 7984, 6972, 6528, 7984, 9187, 6846, ...]
3	list	5	[8241, 8819, 5320, 1873, 9473]
4	list	12	[1572, 2757, 9364, 4175, 5131, 7613, 2431, 920, 1708, 8880, ...]
5	list	9	[5283, 9066, 2431, 8925, 3599, 813, 9474, 8756, 9473]
6	list	4	[1608, 7984, 710, 9473]
7	list	3	[8252, 6809, 9473]
8	list	8	[8880, 9212, 8252, 3609, 3599, 7984, 5625, 9473]

OK Cancel

Figure 10.4: Training answers

D. Validation Questions

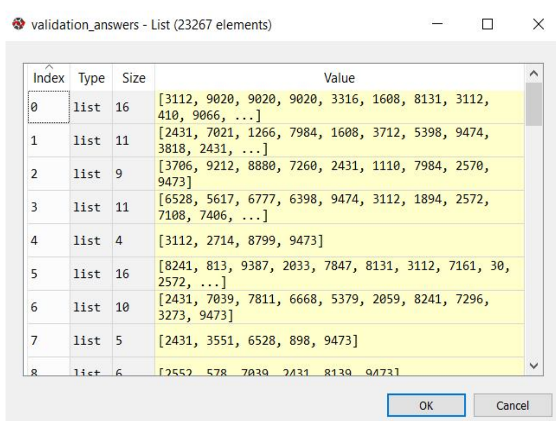

validation_questions - List (23267 elements)

Index	Type	Size	Value
0	list	2	[2286, 5297]
1	list	2	[1843, 4042]
2	list	2	[4623, 455]
3	list	2	[4890, 939]
4	list	2	[2145, 3249]
5	list	2	[4971, 1153]
6	list	2	[1244, 1124]
7	list	2	[1543, 960]
8	list	2	[3143, 1843]

OK Cancel

Figure 10.5: Validation questions

E. Validation Answers


validation_answers - List (23267 elements)

Index	Type	Size	Value
0	list	16	[3112, 9020, 9020, 9020, 3316, 1608, 8131, 3112, 410, 9066, ...]
1	list	11	[2431, 7021, 1266, 7984, 1608, 3712, 5398, 9474, 3818, 2431, ...]
2	list	9	[3706, 9212, 8880, 7260, 2431, 1110, 7984, 2570, 9473]
3	list	11	[6528, 5617, 6777, 6398, 9474, 3112, 1894, 2572, 7108, 7406, ...]
4	list	4	[3112, 2714, 8799, 9473]
5	list	16	[8241, 813, 9387, 2033, 7847, 8131, 3112, 7161, 30, 2572, ...]
6	list	10	[2431, 7039, 7811, 6668, 5379, 2059, 8241, 7296, 3273, 9473]
7	list	5	[2431, 3551, 6528, 898, 9473]
8	list	6	[2552, 578, 7039, 2431, 8139, 9473]

OK Cancel

Figure 10.6: Validation answers

The training loss error and validation loss error were in close proximity with each other that resulted in perfect fitting because both values end up to be roughly the same. To reduce computation, small amounts of datasets were trained for lesser amount of iterations, however, could not lead to significant results. The training loss error is less than validation loss error, the training data but not at all the validation data, in other words it is not generalizing correctly to unseen data. Setting and final tuning of Hyperparameters for training and testing the data sets was quite a time consuming process and high amount of research was done to find the right Hyperparameters and high amount of iteration is done to get feasible results and reduce the training and testing time. The final size of Hyperparameters is clearly represented in Table 1 below:

Hyperparameters	Size
epocs	50
batch size	32
RNN size	1024
encoding embedding size	1024
decoding embedding size	1024
learning rate	0.001
learning rate decay	0.9
Keep_probability	0.5

Table 1: Setting the Hyperparameters

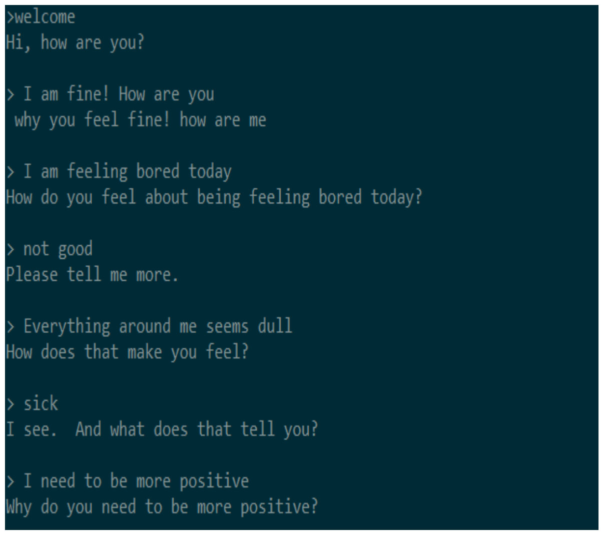


Figure 10.7 Chatbot

XI. CONCLUSION

This project deals with building a chatbot but by implementing a state of the art and Deep Natural Language processing model that will be the seq2seq. It will be implemented with one of the best API to build deep learning applications or artificial intelligence, which will be tensor flow. It involves installing anaconda and getting the dataset, which is Cornell movie corpus dataset a dataset of more than 600 movies containing thousands of conversations between lot of characters and chatbot will be trained on this dataset because the main objective is to build chatbot that can have general conversation with humans like a friend to give movies related information. However, the model developed will be used to train different datasets for some other purposes. For example, we will be able to train the same chatbot on a more specific dataset like a calendar assistant or a navigation assistant. The chatbot developed showed perfect example of perfect fitting because the training loss error and validation loss error were in close proximity with each other that resulted in perfect fitting because both values end up to be roughly the same. However, yet not the best chatbot as we can notice grammatical errors. However, computation was not an easy task as it took three days to train the model and small data sets were not able to produce significant results. Other datasets should be analyzed such as twitter datasets, calendar, navigation and medical datasets and results should be observed whether the model produces feasible results in other domains. Moreover, the model should be tuned again to reduce the training and testing time using cloud computing. Windows azure that provides access to tools such as R, Python, Hadoop frameworks.

REFERENCES

[1] Different measurements metrics to evaluate a chatbot system. B. A. Shawar, and E. Atwell. 2007, Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies.
[2] A Computer Program For the Study of Natural Language Communication Between Man And Machine. J.Weizenbaum. 1966, Commun. ACM, Vol. 9, pp. 36-45.
[3] A comparison between ALICE and Elizabeth chatbot systems. Atwell, B. A. Shawar and E. 2002, Raport instytutowy, University of Leeds.
[4] On Their Tenth Anniversary, Mobile Apps Start Eating Their Own. Khalaf, S. 2016, Flurry Analytics.
[5] Seniors prove WeChat is not just for young. Jinran, C. Yin and Z. 2017, CHINA CHANNEL.
[6] Chatbots and Conversational Agents: A Bibliometric Analysis . Lee, H. N. Io and C. B. 2017, Department of Accounting and Information Management, University of Macau, China .

[7] Li Ka-Shing Buys Canada's Reliance Home for $2.1 Billion. Deveau, S. 2017, Bloomberg.
[8] Alibaba Invests in AI Startup. Galvin, P. 2016, Tech Exec.
[9] 'Sinovation Ventures' Dr. Kai-Fu Lee is betting big on artificial intelligence. Dillet, R. 2016, TechCrunch.
[10] Using corpora inmachine-learning chatbotsystems. Atwell, Bayan Abu Shawar and Eric. University of Leeds.
[11] Usingcorporaforlanguageresearch:Studies in the honour of Geoffrey Leech . Atwell, E. 1996, Harlow: Longman, pp. 151-166.
[12] Weizenbaum, J. ELIZA-A computer program for the study of natural language communication between man and machine. Communications of the ACM. 1966, pp. 36-45.
[13] Colby, K. Computer models of thought and language . Simulation of belief systems. San Francisco : Freeman, 1973, pp. 251-286.
[14] ALICE. .A.L.I.C.EAIFoundationwebsite. 2002.
[15] Introduction to the Theory of Neural Computation. J. Hertz, A. Krogh, and R.G. Palmer. 1991, Addison-Wesley, Reading, Mass.
[16] Artificial Neural Networks: ATutorial. Anil K.Jain, Jianchang Mao and Mohiuddin K. M. IBM Almaden Research Center.
[17] Scalinglearning algorithms towards AI. InLarge-ScaleKernelMachines. Y. Bengio, Y. LeCun. s.l. : MITPress, 2007, InLarge-ScaleKernelMachines.
[18] Training Neural Network Language Models On Very Large Corpora. Jean-LucGauvain, HolgerSchwenk and. 2005, Joint Conference HLT/EMNLP.
[19] Scaling learning algorithms towards AI. Bengio, Y. LeCun and Y. 2007, In Large-Scale Kernel Machines, MIT Press.
[20] Finding Structure in Time. Elman, Jeffrey L. 1990, Cognitive Science.
[21] Neural networks and physical systems with emergent collective computational abilities. Hopfield, J. j. 1982, Proc. Natl.Acad. Sci, Vol. 79, pp. 2554-2558.
[22] A learning algorithm for Boltzmann machines. D. H. Ackley, G. E. Hinton and T. J. Sejnowski. 1985, Cognitive Science.
[23] Some networks that can learn, remember, and reproduce any number of complicated space patterns. Grossberg, S. 1969, Math. and Mechan, pp. 53-91.
[24] A wave model of temporal sequence learning. Kilmer, C. Stanley and W. L. 1975, lnt. J. Man-MachineStud, pp. 397-412.
[25] Bidirectional associative memory. Kosko, B. 1988, IEEE Trans.Sys.Man.Cybern, Vol. 18, pp. 49-69.
[26] A Neural Chatbot with Personality. Huyen Nguyen, David Morales and Tessera Chin. 2014, Computer Science Department, Stanford University.
[27] A Web-based Platform for Collection of Human-Chatbot Interactions . Lue Lin, Luis Fernando D'Haro and Rafael E. Banchs. 2017, Biopolis, Singapore.
[28] A Step Towards Sequence-to-Sequence Alignment. Irani, Yaron Caspi and Michal. 2000, Proceedings IEEE Conference on Computer Vision and Pattern Recognition.
[29] Complex Temporal Sequence Learning Based on Short-term Memory . ARBIB, Deliang Wang and Micheal A. 1990, Proceedings of the IEEE.
[30] Using corpora inmachine-learning chatbotsystems. Atwell, Bayan Abu Shawar and Eric. 2005, International Journal of Corpus Linguistics.
[31] Extensions of Recurrent Neural Network Language Model. Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Honza and SanjeevKhudanpur. 2011, Precedings of IEEE.
[32] From Eliza to XiaoIce: challenges and opportunities with social chatbots . Heung-yeung SHUM, Xiao-dong HE and Di LI. 2018, Frontiers of Information Technology and Electronic Engineering.
[33] Learning Language Models of Movie Characters. Gholami, Oguz H. Elibol and Milad. 2017, Preceedings of IEEE.
[34] Mitigating the impact of Speech Recognition Errors on Chatbot using Sequence-to-Sequence Model. Pin-Jung Chen, I-Hung Hsu, Yi-Yao Huang and Hung-Yi Lee. 2017, IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).
[35] Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, Hua Wu. 2018, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 1118-1127.
[36] Predicting effects of noncoding variants with deep learning–based sequence model. Troyanskaya, Jian Zhou & Olga G.