ISSN 2347 - 3983
Volume 8. No. 5, May 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter35852020.pdf
https://doi.org/10.30534/ijeter/2020/35852020

Chatterbot implementation using Transfer Learning and LSTM Encoder-Decoder Architecture

Kolla Bhanu Prakash[1], YVR Nagapawan [2], N Lakshmi Kalyani [3], V Pradeep Kumar [4]
[1]Assoc.Prof.- Department of CSE, Koneru Lakshmaiah Education Foundation , India, drkbp1981@gmail.com
[2]Assi.Prof.- Department of CSE, Anurag Engineering College, India, ynpawan@gmail.com
[3] Assi.Prof.- Department of CSE,VNR Vignana Jyothi Institute of Engineering & Technology, India, neerukondait@gmail.com
[4] Assi.Prof.- Department of CSE, B. V. Raju Institute of Technology, India, pradeepvadla@gmail.com

ABSTRACT
The goal of this project is to develop a chatbot using deep learning models. Chatterbot is an existing research area whose main goal is to appear as human as possible and most of the current models(which use RNN and related sequential learning models) are unable to achieve this task to relate over long dependencies. Adding onto that , NLP tasks require a lot of data which can be hard to collect for smaller projects/tasks which inspired to try out sequence to sequence learning model using LSTM. For that we have used a movie dialog corpus of 220,579 conversation exchanges among which about 50,000 conversational exchanges are only used as a training corpus to our model since training on more conversation exchanges requires high computation power than we have.

Key words : Deep learning, Chatterbot, RNN, NLP, LSTM.

1. INTRODUCTION

A chatbot is a software program that uses text or text-to-speech to perform an online chat conversation instead of direct communication with a live human user. Built to convincingly mimic how a human being will haves conversation with partner, chatbot systems usually testing and tuning in continuous way, and many in development remain unsuccessful effectively pass after conversing the standard Turing test . In 1994, Michael Mauldin (creator of the first Verbot) originally coined the word "ChatterBot" to describe such conversational programs.

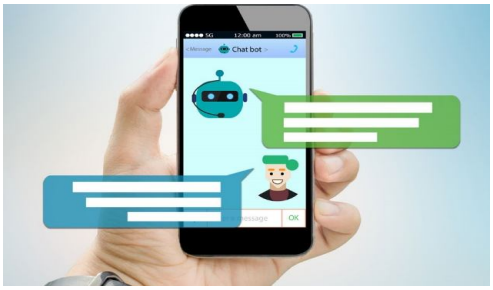

Figure 1: Chatting Interface (Chatbot Interacts)

The Figure 1 shows the chatting interface for the user and chatbot interaction and the Figure 2 shows platforms/apps that uses chatbot. Natural language processing is one important area in AI research. AI fields that are weak usually implement advanced applications or programming languages designed specific to the limited role they require. For example, markup language used by A.L.I.C.E. (AIML) that is unique to its use as a chatting agent and has been used by Alicebots. Nevertheless, A.L.I.C.E. is still is focused on pattern matching techniques and no practical experience, as back in 1966 ELIZA used the same methodology [12].

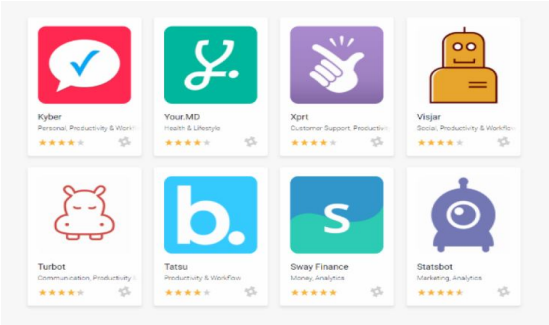

Figure 2: Platforms/Apps that uses chatbot

Integrating NLP in the chatbots means more human presence. If a chatbot is created and deployed , this is for general usage, so you also see people asking questions about it. It seems very much in line with human nature that users would try to fool and throw off the chatbot. You may attempt to solve that by adding default responses, but that tends to fall short very much as it is almost difficult to anticipate which questions will be answered and how they will be raised as shown in Figure 3. Latest research focused mainly on unsupervised and semi-supervised algorithms in computing.

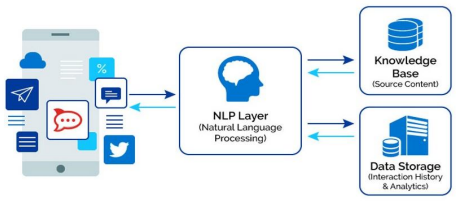

Figure 3: NLP for Chatbots

Today's AI systems are capable of interacting with consumers, assessing their desires, identifying their preferences and recommending an acceptable course of action with minimal or no intervention from person. Many conversational agents accessible now, such as Apple's Siri, Microsoft's Cortana, Amazon's Search Assistant and Alexa. The basic principle of chatbots is to have the best answer to every question it receives. [5]. The Figure 4 and 5 below shows Chatbot's conceptual map using Deep learning.

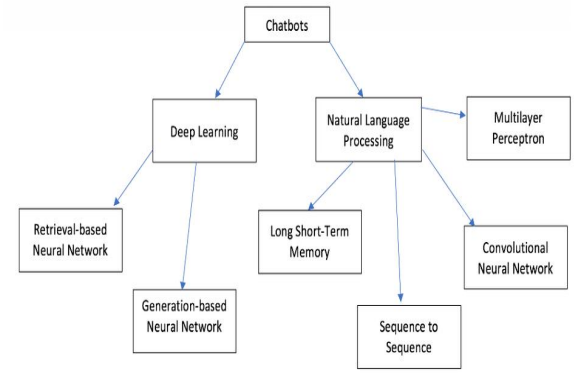

Figure 4: Chatbot conceptual map

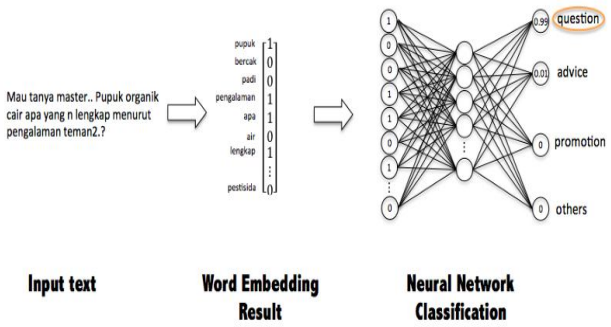

Figure 5: Classification using Neural Network

Nearly all chatbots use deep learning, some version of a Sequence to Sequence (Seq2Seq) pattern. In 2014, with a paper named "Sequence to Sequence Learning with Neural Networks," Ilya Sutskever, Oriol Vinyals and Quoc Le published the foundational work in this area.

RNNs are a neural networks that can manage sequential data, such as videos (frame series), and more commonly, sequences of texts or virtually collection of symbols. The benefit of this is that the network doesn't need to ask what the icons say. Think of the condition as network memory as in this memory unit (internal state), which is constantly modified during each process, RNN devours a sequence (sentence), word by word, important information about the sentence. The main distinction between a normal RNN network and LSTM / GRU networks is the architecture having the memory unit Figure 6. An LSTM cell has multiple gates to keep track of useful information, to forget unnecessary information and to carefully expose information at every move.

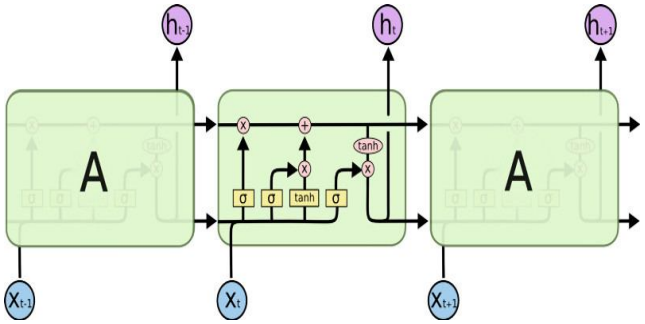

Figure 6: LSTM Architecture

The Figure 7 Sequence to Sequence model(seq2seq) is made up of Two RNN encoders, and one decoder. The encoder reads the list of inputs word for word and transmits a description, ideally catching the input sequence's context. The decoder generates the output sequence. The goal is to jointly optimize the log probability of the output series according to the input series [6].

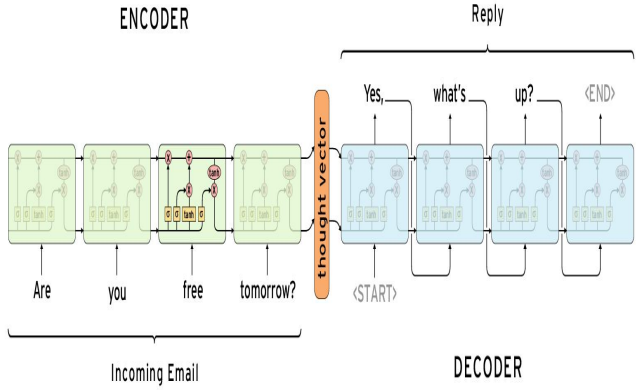

Figure 7: Sequence to Sequence Model

Cho and. Al. suggested a novel form of the neural network named RNN Encoder-Decoder consisting of two recurrent neural networks (RNN). One RNN codes a symbol sequence into a representation of a fixed-length matrix, and the other decodes the representation into another symbol sequence.

1.1 Transfer Learning in Chatbot

In training deep neural networks, AI engineers have been increasingly excellent at correctly mapping from inputs to outputs involving broad data sets whether they are words, images, mark predictions etc. However, the ability to take a broad perspective of situations that aren't similar to those they experience during training is often missing in these models. Real world simulations can be chaotic, and engineers are forced to make incorrect assumptions if they apply these models to data sets that have not been observed. The Figure 8 Transfer learning should solve the problem.

Transfer learning is basically a model's capacity to transform information into new environments. This work implements a modern Transfer-Transfo method, for conversational generative data-driven systems such as chatbots. The algorithm incorporates programming based on transfer.

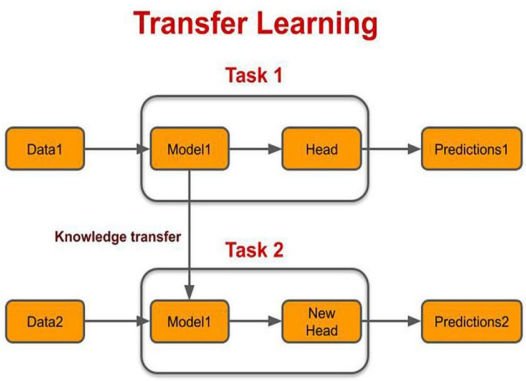

Figure 8: Transfer Learning Model

learning as well as transformer models Like Figure 9. It performs fine-tuning for the algorithm by using a multi-task objective which incorporates a variety of unsupervised prediction procedures. Supervised learning, powered by deep learning, is largely responsible for the wave of AI applications that we have seen in the last few years [21]. Transfer learning will make these developments significantly greater.

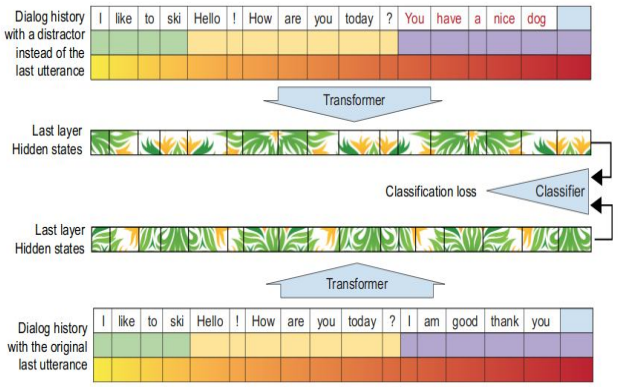

Figure 9: Transfer-Transfor input representation

Pretrained language models were used to produce state-of-the-art outcomes on a wide variety of NLP tasks (e.g., sequence marking and grouping of sentences). Any of the latest works using pre-trained language models include the transformer ULMFit, ELMo, GLoMo and OpenAI. These language modeling systems perform extensive pre-training of hierarchical representations or graph-based representations of whole structures. This idea represents a change from the usage of unary word embedding, which has been used extensively over the past few years to solve many NLP functions, to the use of more complex, abstract representations [9].

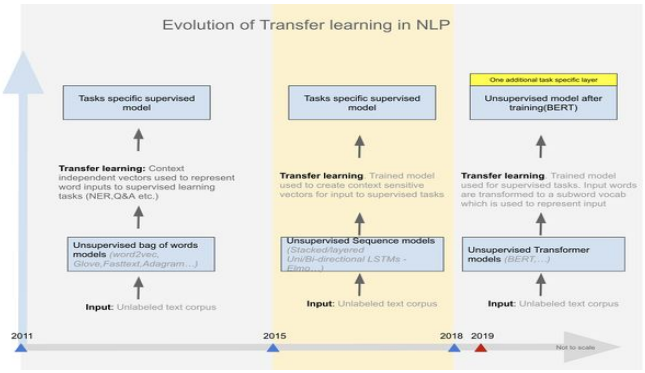

Figure 10: Evolution of Transfer Learning in NLP

Pre-training allows a model to catch and absorb from a large-scale corpus a range of linguistic anomalies, such as long-term dependency and negation [22]. This information is then used (transferred) to initialize and then train another model for performing well on a similar NLP task, such as classifying sentiments as shown in Figure10.

2. Related Work

2.1. A Chatbot Using LSTM-based Multilayer Embedding for Elderly Care

Because of population changes, programs targeted to elderly people are now more important than ever, and increasingly available. The details from the MHMC's daily encounters with older people have been gathered in this report. Because individuals may say the method, the sentences that are recorded in the preprocessing phase to cover the variation of the conversational sentences are translated into patterns. A multi-layer embedding technique based on LSTM is then used to extract contextual details between terms and phrases using several sentences in a single turn when communicating with the elderly. Ultimately, the Euclidean distance is used to choose an appropriate type of question, which is then used to choose the correct response for the elderly [8].

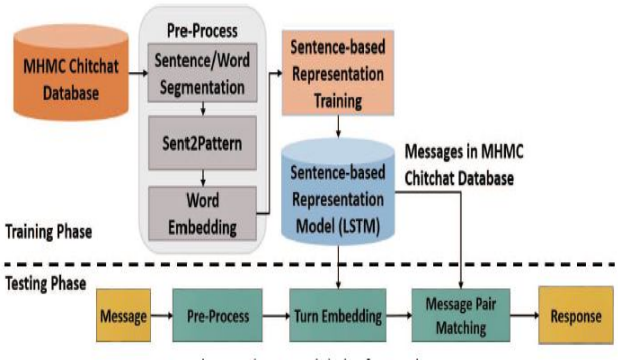

Figure 11: Chatbot Framework

This research launched a chatbot for elderly treatment, which can be used to assist Or talk with seniors to relieve their loneliness. They compiled the collection of chitchat from 40

topics in order to perform a daily discussion as training material. All the messages were classified and patterned in the MHMC chitchat database As shown in Figure 11. Then they used a two-layer LSTM model to address the interaction between words and phrases in the semantics. Finally they choose the corresponding response by contrasting pairs of message through Euclidean distance. The outcome indicates that a stronger answer can be generated by the suggested approach, achieving 79.96 per cent matching the top 1 message pair [10].

2.2. A Chatbot by Combining Finite State Machine, Information Retrieval, and Bot-Initiative Strategy

They developed a conversational framework in this work by integrating a finite state process, a model of retrieval and a technique of dialog between computer and initiative. Unless the user's utterance is beyond the range of modelled dialogues, the input is interpreted by the retrieval algorithm. This model of retrieval has demonstrated greater consumer satisfaction in comparative experiments than the one centred on the neural-network. Yi et. Al used data from different sources to create a conversational model. Firstly, they used Twitter scrapped Evi and dialog details to handle everyday conversations.[2]. They identified 31 transitions and made 12 states using model of ASK purpose to model dialogues [2]. It enters 'Start' state when their program is initialized. A solution is selected between the candidate sentences depending on the transition triggered by a user request, and the particular state that stores the background knowledge. There are two types of dialogs that are handled using finite state method: multi-round and one-round dialogue [15].

2.3 Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

In this model they proposed a new neural network architecture named RNN Encoder– Decoder composed of two recurrent neural networks (RNN). One RNN encodes a series of symbols into a representation of a matrix of defined length, and the other decodes the representation into another symbol set. The performance of a predictive machine translation system is empirically shown to be increasing utilizing the conditional probabilities of phrase pairs computed by the RNN Encoder – Decoder as additional component of the existing log-linear model [3].

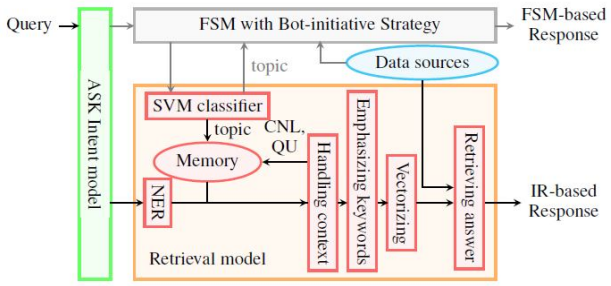

Figure 12: FSM Chatbot Architecture

Cho et Al taught the model to realize the probability of translating an English word into a matching French expression as shown in Figure 12. The pattern is then included in the table of sentences as part of a standard phrase-based SMT system by rating each pair of phrases. The analytical assessment shows that this method of rating pairs of phrases with an RNN encoder – decoder increases the efficiency of the translation [14]

Nguyen et. Al. devised and evaluated Pytorch on deep learning framework with the utilization of GPU. The model is an end to end [7], with no hand-crafted rules. Model is developed from a limited dataset and also provides user responses. Nonetheless, answers generated do need to be strengthened to have a constructive conversation. Bahdanau et al. (2015) recently effectively implemented such a research method on communication of neural machine by mutual learning how to manage and perceive words together. [4]. The list of Vietnam comprises of 1331 questions and responses. The conversational datasets with 300,000 are pretrained on measurements out of 1331 pairs of sentences. It obtained 887 pairs of comments and responses after checking and picking 15-word long phrases (including punctuation). 843 phrases are used in the vocabulary and 918 phrases are included in the term response. The training time of the model is 288 minutes and one second, and Google Colab (free GPU) was used [19].

3. PROPOSED ALGORITHMS

The Retrieval-based models select an answer based on the question from a set of responses. It produces no new words, so We don't have to owe grammar a look [20]. The Generative Models are really wise. They generate a response, word by word, which is dependent on the query. The answers provided are vulnerable to grammatical errors because of this. These models are challenging to practice, because they have to study the proper sentence form on their own [11]. However, when trained, the generative models outperform the retrieval-based models in terms of managing previously unseen queries and produce the user's intention of talking to a human (maybe a toddler). The proposed model of the chatbot is done using the Sequence to Sequence model with transfer learning. Here, the movie dialog is used as a training corpus [18]. This repository includes a wide selection of metadata-rich fictitious interactions taken from the scripts of raw films. It has 220,579

conversational exchanges between 10,292 pairs of movie characters and includes a total of 304,713 utterances of 9,035 characters from 617 movies while only 50,000 conversational exchanges were used for preparation. We used pre-trained word embedding from the language model such as openai gpt for transition learning, and in sequence to sequence learning model utilizing LSTM encoder and decoder shown in Figure 13, and the processes in between [25].

Line to Line Paradigm applied with RNN Predictive Machine Translation Encoder-Decoder (cho et. al.) in Studying Phrase Representations has since been the Go-To Norm for Dialog Systems and Computer Translation. It consists of two RNNs, one Encoder and one Decoder. It is constructed of two RNNs. The encoder takes a number (sentence) as input and processes at each stage one symbol (word). The job is to turn a symbol series into a vector that only covers important details in a series without missing unnecessary information [3].

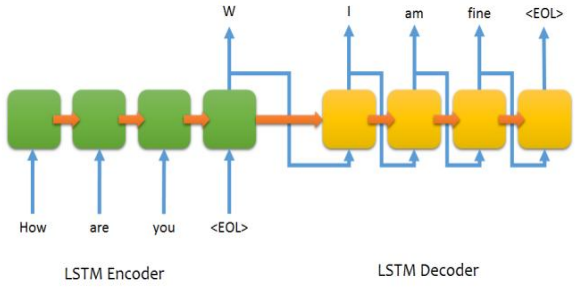

Figure 13: Sequence to sequence Model (LSTM encoder-decoder)

3.1 Working of LSTM cell

LSTM networks are extended RNNs consisting of extended memory cells known as gated cells which allow the inputs to be remembered for a long time [16]. The information in the LSTM gated cell memory may be retained or discarded and can be calculated by the weights / value allocated by the algorithm given, i.e. it learns the value of the information over time. The relation between the present layer output and next layer input shows the cell's recursive existence. This allows for knowledge to be retained in the LSTM cell from previous periods. In total there are three gates in LSTM:

- input gate
- forget gate
- output gate

Input gate decides whether to allow in fresh data or not.

Forget gate determines whether to remove unhelpful details. If the forget gate performance is 1, the knowledge is retained in the cell state and is indicated to be forgotten by closer to 0. Input gate chooses to let the forget gate input influence the input at the current time stage and it defines how much of the activation of each device is retained [16].

The proposed framework methodology consists of different steps, such as raw data selection, pre-processing of data, extraction of features and NN preparation. LSTM consists activation functions such as

- Sigmoid
- Tanh.

Sigmoid feature wants to change the correct details and only installs memory but cannot remove / forget memories on its own [13]. "Tanh" controls the values that pass across the network, which allows the cell satellite to lose information. Let it, ot, ft reflect the input – output activations, and forget gates at time phase t, respectively.

Xt = (x1, x2, x3, ..., xn) (1)

ht is the hidden state

ht = (h1, h2, h3, …., hn) (2)

And ct is memory cell that represents

ct = (c1,c2,c3, ... cn) (3)

The current cell output ht is as follows ht = ot tanh (ct)

- New memory (ct)
- current output (ht) next step of time and it repeats

3.2 Word Embeddings

Term Embedding is a technique to achieve dense term representation in a small vector space. Increasing term, represented by a fixed-length vector, is a space point which can be seen. This technique captures semantic connections between the words. There are some fascinating properties of the term vectors. Word Embedding is usually performed in the network's first layer: Embedding layer, in which a word is mapped (index to word) from a vocabulary to a dense vector of the given scale. In the seq2seq model, the embedding layer weights are trained in accordance with the model's other parameters [1] shown in Figure 14.

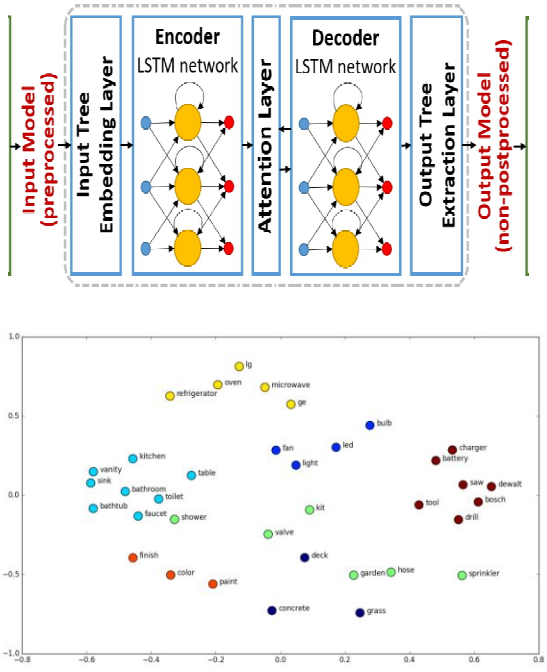

Figure 14: Word Embedding graph

4. RESULTS AND DISCUSSIONS

4.1 LSTM Encoder Decoder Architecture

The results are for random inputs from the test data.
RMSE for training dataset = 6.42
RMSE for testing dataset =18.24
The model has shown to respond well for all answers with RMSE below 20 for both train and test data while it was above 50 for ANN. Compared to generic answers a normal ANN gives, the answers given by this encoder decoder structure using LSTM gave a better dynamic answers, albeit , still meaningless [24]. The grammar is more accurate compared to basic generative models when manually evaluated shown in Figure 15.

While the testing and training loss explains how well a model learns compared to another model, when the task is about conversational agents these losses are stripped of their meaning [17]. The main problems with generative type conversational agents occur once they are outside the testing and training phases. These phases only evaluate a model based on how much its replicating the output. A chatterbot is good only if it works the same way for messages it has never seen before. This is extremely hard if a model does not "think" before it answers shown in Figure 16. In our case, the LSTM, although it beats other generative models, it still doesn't have any mechanism to think [23].

Figure 15. Result Screen 1

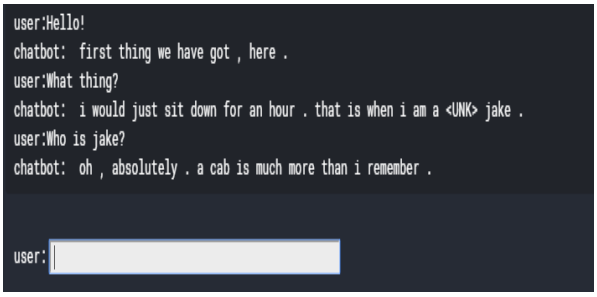

Figure 16. Result Screen when input is taken from the user

5. CONCLUSION

LSTM based encoder decoder structure has shown to be more robust, cleaner and faster than a ANN model. This model also shows promise as a generative based model which prior generative approaches were not able to show. The sentence structure is more dynamic albeit still mostly meaningless as it is free from context when building sentences. But the biggest advantage shown by this model is the relative ease of building. All this model needs to work well is lots of data unlike older approaches which require constant human intervention to work properly. This clean approach is desirable and favourable for wide usages of chatterbots.

REFERENCES

1. Su. Ming, Wu. Chung, Huang Kun, H. Qian and Hsin Wang, "A Chatbot Using LSTM-based Multi-layer Embedding for Elderly Care", IEEE Conference on Orange Technologies, 2017. https://doi.org/10.1109/ICOT.2017.8336091
2. Yi Sanghyun and Kyomin Jung, "A Chatbot by Combining Finite State Machine, Information Retrieval, and Bot-Initiative Strategy", Semantic Scholar, 2017.
3. Kyunghyun Cho, B. Dzmitry, B. Fethi, Yoshua Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", [online: https://arxiv.org/abs/1406.1078]
4. Trang Nguyen and Maxim Shcherbakov, "A Neural Network Based Vietnamese Chatbot", IEEE Int. Conf. on System Modeling & Advancement in Research Trends, 2018. https://doi.org/10.1109/SYSMART.2018.8746962
5. "Overview of deep learning", 31st youth academic annual conference of Chinese association of automation, Wuhan, China, Nov 2016.
6. Alex Sherstinsky (9 Aug 2018). "Fundamentals of Recurrent Neural Network and Long Short-Term Memory", [Online]. Available: https://arxiv.org/abs /1808. 03314
7. "Chatbot Using Gated End-to-End Memory Networks", International Research Journal of Engineering and Technology (IRJET), 2018, India.
8. Vaswani et. al. "Attention is all you need.", [online: https: //arxiv.org/abs/1706.03762]

9. Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, Chunfang Liu,"A survey on Deep Transfer Learning", IEEE Int. Conf. on Artificial Neural Networks(ICANN 2018).
10. Bayu Setiaji and Feery Wahyu Wibowo, "Chatbot Using a Knowledge in Database: Human-to-Machine Conversation Modeling", IEEE Int. Conf. Intelligent Systems, Modelling and Simulation, 2016. https://doi.org/10.1109/ISMS.2016.53
11. AM Rahman, Abdullah Al Mamun, Alma Islam,"Programming challenges of chatbot: Current and future prospective", IEEE Conf. on Humanitarian Technology, 2017.
12. Hiremath et. al., "Chatbot for education system", International Journal of Advance Research, Ideas And Innovations In Technology, vol 4, Issue 3
13. Kolla, B.P., Raman, A.R." Data Engineered Content Extraction Studies for Indian Web Pages", Advances in Intelligent Systems and Computing, vol 711, 2019. https://doi.org/10.1007/978-981-10-8055-5_45
14. Prakash, K.B., Rangaswamy, M.A.D., Raman, A.R. "Statistical interpretation for mining hybrid regional web documents", Communications in Computer and Information Science, vol 292 CCIS, 2012. https://doi.org/10.1007/978-3-642-31686-9_58
15. Prakash, K.B., Rajaraman, A." Mining of Bilingual Indian Web Documents", Procedia Computer Science, vol 89, 2016
16. Ismail, M., Prakash, K.B., Rao, M.N, "Collaborative filtering-based recommendation of online social voting", International Journal of Engineering and Technology(UAE), vol 7, issue 3, 2018. https://doi.org/10.14419/ijet.v7i3.11630
17. Prakash, K.B, "Content extraction studies using total distance algorithm", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, 2017
18. Prakash, K.B., Rajaraman, A., Lakshmi, M, "Complexities in developing multilingual on-line courses in the Indian context", Proceedings of the 2017 International Conference On Big Data Analytics and Computational Intelligence, ICBDACI 2017, 2017
19. Prakash, K.B., Kumar, K.S., Rao, S.U.M, "Content extraction issues in online web education", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, 2017 https://doi.org/10.1109/ICATCCT.2016.7912086
20. Prakash, K.B., Rajaraman, A., Perumal, T., Kolla, P. "Foundations to frontiers of big data analytics" , Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016, 2016
21. Prakash, K.B., Ananthan, T.V., Rajavarman, V.N, "Neural network framework for multilingual web documents", Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014, 2014 https://doi.org/10.1109/IC3I.2014.7019797
22. Prakash, K.B., Rangaswamy, M.A.D., Raman, A.R, Attribute based content mining for regional web documents, IET Seminar Digest, vol 2013, issue 8, 2013
23. Ratnam, D.V., Gutta, R.C., Kumar, M.R., (...), Pavan, P.S., Santhosh, C., " Effect of high-k dielectric materials on mobility of electrons", International Journal of Emerging Trends in Engineering Research8(2), 2020. https://doi.org/10.30534/ijeter/2020/12822020
24. Suresh, B., Nikhilesh, T., Abhishek, T., (...), Chandra Sekhar Yadav, G.V.P., Ghali, V.S, "Qualitative subsurface analysis in quadratic frequency modulated thermal wave imaging", International Journal of Emerging Trends in Engineering Research, 8(1), 2020. https://doi.org/10.30534/ijeter/2020/06812020
25. Suresh, B., Sathvik, K.N., Imran, S., (...), Vijayalakshmi, A., Ghali, V.S, "Diagnosing osteoporosis through numerical simulation of bone sample by non-stationary thermal wave imaging", International Journal of Emerging Trends in Engineering Research, 8(3), 20202. https://doi.org/10.30534/ijeter/2020/27832020