Vol. XVI & Issue No. 04 April - 2023
INDUSTRIAL ENGINEERING JOURNAL

BUILDING INTUITIVE CHATBOTS FOR HUGE KNOWLEDGE BASES USING INFORMATION RETRIEVAL

Umesh D

Dr. Vikas S

Abstract

Since the beginning of artificial intelligence and machine learning, one of the primary areas of focus has been developing systems capable of comprehending human people and responding to them in the most natural manner possible. This has been one of the primary areas of focus since the beginning of artificial intelligence and machine learning. The simulation of human interaction is the intended end result of a machine-learning system with the same name as the chatbot.

Keywords – IR, LSTM, RCNN, Machine learning.

1. INTRODUCTION

Chatbots are becoming more popular as a kind of application that enables individuals to conduct simple discussions via the use of text [1]. Chatbots may be broken down into two primary categories. The first kind is known as a generative chatbot, while the second type is a retrieval-based chatbot, also known as an IR-based chatbot. The user inputs their question, and the generative chatbot provides an answer to that question automatically. The conversational training set was used as a basis for the replies that were created. The IR-based chatbots always make an effort to provide a response that is the most relevant one out of the preset replies that are included in the dataset. The Ubuntu conversation corpus was used in the development of this IR-based chatbot system.

The Ubuntu Dialogue Corpus is a very extensive database that contains one million different conversations. Context, utterance, and label are included in the dataset. The term could have a good or a negative connotation. The presence of the positive label suggests that the speech in question is suitable for the context in question, whilst the presence of the negative answer shows that the utterance in question is not suitable for the context in question. The training set consists of talks that are marked positively 50% of the time and negatively 50% of the time[3].

The pretrained Glove (Global Vectors for Word Representation) model may be used to produce the word embeddings for the provided text data. [4] When attempting to make a prediction, the word embeddings are sent via an LSTM (long short term memory) network.

The data for the test are organised into columns with one context and 10 utterances each. Only one of the 10 statements will be the ground truth speech, and the others will all be distractors. The ten-dimensional vectors will serve as the values that are expected. The significance is taken into consideration while assigning the values. The speech that is most pertinent would be given the most value among that. Recall @k is the metric that is employed, and k may be any number between one and ten.[5] k is any integer that takes the highest k utterances from 10.

2. RELATED WORK

In the area of human-machine interaction, a lot of research has been carried out throughout the course of time, and many different projects are now being worked on. Manuscript forms are used for the presentation of some of the works that are relevant to the area.Puneet Agarwal [1] in the paper highlighted the challenges they faced while building an emerging commercial chatbot. They builded a chatbot which they referred as trusted friend of every Indian youth. In his study, Ryan Lowe [2] detailed the many components that make up the Ubuntu dialogue corpus. In addition to this, they discussed two methods that are helpful in analysing the Ubuntu Dialog Corpus[6].

3. SYSTEM OVERVIEW

The flow of the paper goes in the following way:

1. **ENCODING**
2. **CREATING THE MODEL TO CLASSIFY DATA:(LSTM)**
3. **PREDICTION**
4. **EVALUATION**

Pre-processing: Stemmatization and lemmatization were applied to the data that was provided (the Ubuntu Dialog corpus). In addition, we eradicated the stop word from both the context and the utterances [7]. In a subsequent phase, we got rid of the underscores and a few words that were meaningless.

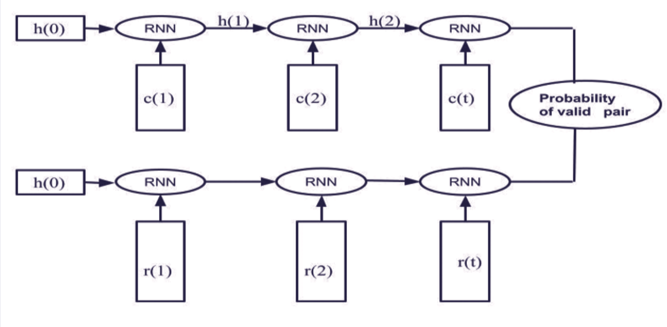
Glove Implementation: We tokenize both context and response. Here tokens represent words

- Every word has to be converted into fixed size vectors. We made use of Glove vectors to initialize the word embeddings of context and responses. In later part of project LSTM model is used for classification purpose. Because the Ubuntu dialogue corpus is such a big database, the pre-trained embedding used in Glove is scalable, making it an ideal choice for this particular project [8].

4. MODEL ARCHITECTURE

The word embeddings obtained are fed to the LSTM (Long Short-term Memory)[10]. The architecture of LSTM model is given below.

Fig 1: Recurrent Neural Network (RNN)



RNNs are examples of types of neural networks shown in Fig 1. This RNN has a hidden state that is denoted by ht, the hidden state of the state before it is denoted by ht1, the input state is denoted by xt, the weight metric of the input neuron is denoted by W x, and the weight metric of the recurrent network is denoted by W h. The formula for representing the current state is as follows:

$$h_t = f(W_h h_{t-1} + W_x x_t) \qquad (1)$$

Long Short Term Memory networks (LSTM)

The issue of long-term dependencies in RNNs may be circumvented by the use of a specialized kind of neural network known as an LSTM model, which is what has been used here.

The word embedding produces vectors for every word in contexts and utterances. The vectors of context are fed to one neural network and the utterance are given to the other network. The hidden state gets updated every time we feed word to the RNN, and the last hidden state represents the vector of context and utterance. The learned parameters are represented by matrix M. Let the context matrix obtained be c, and let the utterance matrix obtained be uttera.

p(label=1|c,u,M)=(C + b), where b represents the bias in the analysis. (2)

In the last step of the process, the model is trained by minimising the binary cross entropy of all context-utterance pairs. The loss in binary cross entropy is denoted by the letter L and is determined using the equation that is provided below.

$$L = -\sum_{i=1}^{n} logp(label|c_i, r_i, M) \qquad (3)$$

5. RESULTS AND PERFORMANCE

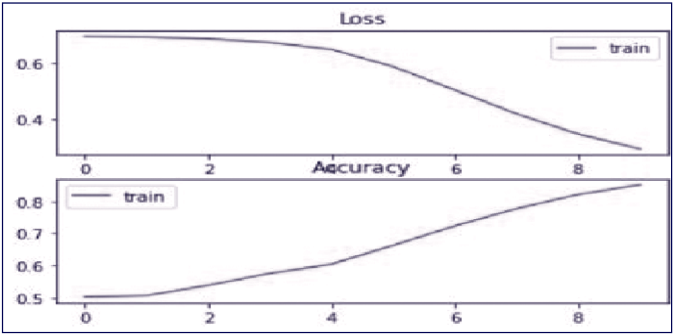
There is one context, one ground truth utterance, and nine replies spread throughout each tuple that makes up the test dataset. These are the elements that are utilised to test the model that has been suggested. While it comes to each answer, the context is taken into consideration when computing the likelihood. If the likelihood is higher, then there is a greater possibility that the reaction will be more appropriate to the situation.

Table 1: Analysis on Predicted Probability

Context	Response	Predicted Probability
what you looking for linuxuz3r? __eou__ no i mean are you looking for a spefic program? __eou__ im not sure if there is anything better then sourceforge __eou__ __eot__ no particular program, anything that interest me then contribute to the source __eou__ iwanna learn how to read code __eou__ __eot__	there is one that escapes me at the moment __eou__ most people use sourceforge __eou__	**0.7910602**
	but, from what to what? Can I hook up a formerly raid'd drive via USB/ SATA bridge like that? __eou__	**0.3378635**

The following graph Fig 2 represents the accuracy and loss of training and was produced as follows:

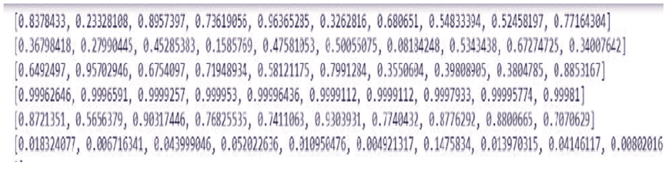
Fig 2: accuracy and loss of training Plot



	LSTM
in 10 R@7 1	0.6990099
in 10 R@8 1	0.8019802

The results of the test are summarised in the table below, with each tuple indicating the likelihood that the context was present with the replies.

Table 2: The table containing the top k recalls for the model's results.



```
[0.8378433, 0.23328108, 0.8957397, 0.73619056, 0.96365285, 0.3262816, 0.680651, 0.54833394, 0.52458197, 0.77164304]
[0.36798418, 0.27990445, 0.45285383, 0.1585769, 0.47581053, 0.50055075, 0.08184248, 0.5343438, 0.67274725, 0.34007642]
[0.6492497, 0.95702946, 0.6754097, 0.71948934, 0.58121175, 0.7991284, 0.3550604, 0.39808905, 0.3804785, 0.8853167]
[0.99962646, 0.9996591, 0.9999257, 0.999953, 0.99996436, 0.9999112, 0.9999112, 0.9997933, 0.99995774, 0.99981]
[0.8721351, 0.5656379, 0.90317446, 0.76825535, 0.7411063, 0.9303931, 0.7740432, 0.8776292, 0.8800665, 0.7070629]
[0.018324077, 0.006716341, 0.043999046, 0.052022636, 0.010950476, 0.004921317, 0.1475834, 0.013970315, 0.04146117, 0.00802016
```

6. CONCLUSION

An algorithm that provides an explanation of the procedure as well as the findings that were reached is described in this work. It has come to our attention that the model that was given and The use of methodology has been shown to be an effective strategy for resolving the issue at hand, and it is possible that its use may contribute to the overall improvement of the matter in question.

7. FUTURE DEVELOPMENT

We were only able to utilise a portion of the data for training since our computing resources were restricted. We would want to apply the training on the whole one million data set in order to create a more accurate training model.

8. REFERENCES

1. *An Information Retrieval-based Approach for Building Intuitive Chatbots for Large Knowledge Bases -Andreas Lommatzsch and Jonas Katins TU Berlin, DAI-Labor, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany*
2. *The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems -Ryan Lowe,Nissan Pow* , Iulian V. Serban† and Joelle Pineau*
3. *A. Bordes, J. Weston, and N. Usunier. Open question answering with weakly supervised embedding models. In MLKDD, pages 165– 180. Springer, 2014. [4] J. Pennington, R. Socher, and C.D. Manning. GloVe: Global Vectors for Word Representation. In EMNLP, 2014*
4. *D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014.*
5. *Nikita Hatwar "AI BASED CHATBOT" International Journal of Emerging Trends in Engineering and Basic Sciences (IJEEBS) Volume 3, Issue 2 (March-April 2016).*
6. *Pratik Salve "College Enquiry Chat Bot" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 3.*
7. *Aniket Dole "Intelligent Chat Bot for Banking System" International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 4 Issue IV, April 2016.*
8. *BayuSetiaji" Chatbot Using A Knowledge in Database" International Conference on Intelligent System, Modling and Simulation 2016.*
9. *I. Aaltonen, A. Arvola, P. Heikkil¨a, and H. Lammi. Hello pepper, may i tickle you?: Children's and adults' responses to an entertainment robot at a shopping mall. In Procs. of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI '17, pages 53–54, New York, USA, 2017. ACM.*

10 *T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol. Rasa: Open source language understanding and dialogue management. arXiv preprint arXiv:1712.05181, 2017.*

11. *A. Mishra and S. K. Jain. A survey on question answering systems with classification. J. King Saud Univ. Comput. Inf. Sci., 28(3):345–361, July 2016.*
12. *K. Toutanova, V. Lin, W.-t. Yih, H. Poon, and C. Quirk. Compositional learning of embeddings for relation paths in knowledge base and text. In Procs. of the 54th Meeting of the Assoc for Comp. Linguistics, volume 1, pages 1434–1444, 2016.*

AUTHORS

Mr. Umesh D, Department of Computer Science, Smt. Indira Gandhi Govt. First Grade Women's College Sagar Karnataka India.
Email: Umesh.sirmv@gmail.com

Dr. Vikas S, Assistant Professors, Department of Computer Science and Engineering, Visvesvaraya Technological University, CPGS Mysuru, Jnana Sangama, VTU Main Road, Machhe, Belagavi – 590 018, (Karnataka) India.
Email: Vikas.smg@gmail.com