IR Based Chatbot

Snigdha, Ramya B Y

Dr. Maunendra Sankar Desarkar(Assistant Professor), Suvodip Dey

cs20mtech11010@iith.ac.in, cs20mtech11008@iith.ac.in

Computer Science and Engineering, Indian Institute of Technology,Hyderabad

Abstract: Building the machines for understanding and replying to human beings in a natural way possible has been a field of interest from the long period in artificial intelligence and machine learning. The chatbot is one of the implementations in machine learning with the same objective.

Background: Various methods or techniques are used in past time for building an effective AI based chatbot. In past various techniques like TF IDF, Word2Vec, RNN and LSTM have been used for the same

***Aim*:** The main objective of this project is to design IR based chatbot. Ubuntu Dialog Corpus is used as the dataset for the same.

***Technique/method*:** For encoding the context and the utterances glove pre trained embedding[3] have been used. Later, Long Short Term Memory has been used as model for classification.

***Derived results*:**

***Conclusions*:** The chatbot designed is successful in producing effective results.

1. INTRODUCTION

These days chatbots[1] are one of the widely used applications which helps people to have easy conversations in the form of text.There are basically two types of chatbots.One is generative chatbot and the other is retrieval based chatbot which is also called IR based chatbot. The generative chatbots generates the response for the query given by the user.The generated responses are based on the conversational training set.The IR based chatbots tries to give the most appropriate response from the predefined responses present in the dataset.To implement this IR based chatbot system we used Ubuntu dialogue corpus.

The ubuntu dialogue corpus is a very huge dataset with one million dialogues. The dataset has context,utterance and label. The label can either be positive or negative.The positive label indicates that given utterance is appropriate for the given context and the negative response indicates that the utterance is not appropriate for the context. The training set has 50% positive labelled dialogues and 50% negative labelled dialogues.

The word embeddings for the given text data can be obtained from the pretrained Glove[4](Global Vectors for Word Representation) model.The word embeddings are fed to LSTM(Long short term memory) network to get the prediction.

The test data consists of columns of one context and ten utterances.Among the ten utterances only one will be the ground truth utterance and rest all are distractors.The predicted values will be the vectors of size ten.The values are assigned based on the relevance.Most relevant utterance would get highest value among that.The metric used is recall@k where k can be any number between one to ten .k is any number which takes topmost k utterances from ten.

2. LITERATURE SURVEY

Over a period of time a lot of research has been done and many work are carried in the field of human machine interaction. Some works related to the field are presented in form of manuscript.

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.

Ryan Lowe[2] in his paper described the various aspects of Ubuntu dialog corpus. Also they mentioned about two techniques useful in analysing Ubuntu Dialog Corpus.

3. SYSTEM OVERVIEW

The flow of the project goes in the following way:

1.ENCODING

2. CREATING THE MODEL TO CLASSIFY DATA:(LSTM)

3. PREDICTION

4. EVALUATION

Pre processing: The given data(Ubuntu Dialog corpus) was stemmed and lemmatized.
Further we removed the stopword from context and utterances. In the later step we removed underscores and few meaningless words.

Glove Implementation: Glove pre trained embedding is scalable for the large corpus, suitable for this project as ubuntu dialog corpus is a huge corpus.

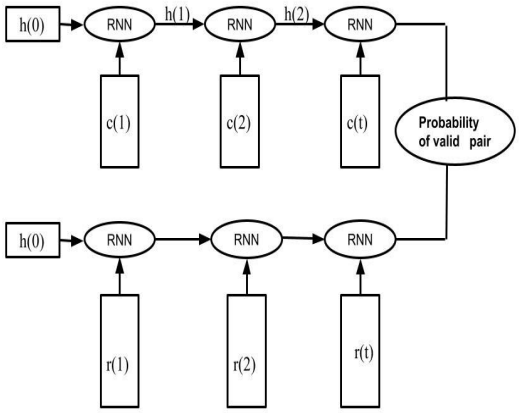
We tokenize both context and response .Here tokens represents words

- 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.

5. MODEL ARCHITECTURE

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



Recurrent Neural Network(RNN)

RNN's are the types of neural networks.This RNN forms hidden state represented by h_t ,the hidden state of previous state is represented by h_{t-1} . The input state is represented by x_t ,the weight metric of input neuron be W_x and the weight metric of the recurrent network be W_h .The formula for representing current state is given as

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

Long Short Term Memory networks(LSTM)

The model used here is a special type of neural network called LSTM which help to avoid the problem of long-term dependencies in RNNs.

The obtained word embeddings are fed to the above neural network.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 .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 utterance matrix be u.The predicted value obtained is c^t =Mr.The similarity between predicted and actual values can be found by concatenating the predicted matrix and context matrix.Let the concatenated matrix be C.Then apply sigmoid function to get the probability between the valid pair.

p(label=1|c,u,M)= $\sigma(C + b)$ where b is the bias.

Finally the model is trained by minimizing the binary cross entropy of all context ,utterance pairs.Let the binary cross entropy loss be represented by L and is calculated by the below equation.

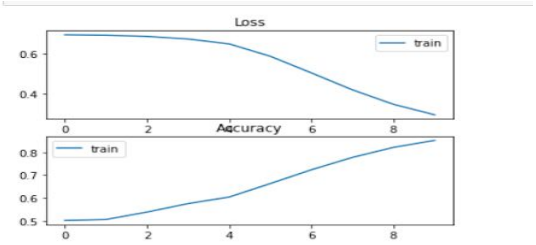
$$L = -\sum_{i=1}^{n} logp(label|c_i, r_i, M)$$

6. RESULTS AND PERFORMANCE

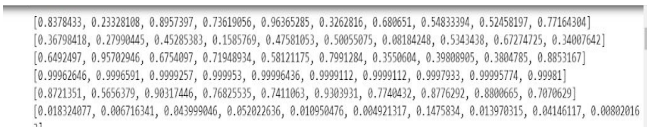
Over each tuple of test dataset, there is a context and there is on ground truth utterance and 9 responses which is used for testing of the model proposed. For each response, the probability is calculated w.r.t the context. Higher the probability, the more chances of being a better response to the context.

Context	Response	Predicted Probablity
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__ i wanna 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

For training accuracy and loss, the graph has been plotted as below:



Below is the snapshot of the test results, where each tuple represents the probability of the context with responses.



Below is the table for top k recalls for the model result.

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

7. CONCLUSION

In this paper, an algorithm explaining the method and the concluded results are presented. It has been observed that the presented model and the methodology has proved to a significant and approach to overcome the situation and can be utilized for the betterment of the observed problem.

8. FUTURE DEVELOPMENT

Due to the limited computational capabilities we used only part of data for training.We would like to apply on whole one million data to get the more accuracy for training model.

9. REFERENCES

[1] Research paper:**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

[5] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014.