A PROJECT REPORT
ON

"SENTENCE GENERATION USING LSTM NETS"

Submitted to
Information Retrieval LAB
DA-IICT

BY

KETUL SHAH 201711017

UNDER THE GUIDANCE OF
PROF. PRASENJIT MAJUMDER

Information Retrieval
DA-IICT

Near Reliance Chowkdi, DA IICT Road,
Gandhinagar, Gujarat 382007
2018

DA-IICT
GANDHINAGAR

CERTIFICATE

This is certify that the project entitled

"SENTENCE GENERATION USING LSTM NETS"

submitted by

KETUL SHAH 201711017

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the project in IR LAB at DA-IICT, Gandhinagar. This work is done during year 2018, under our guidance.

Date: 17 April 2018

(Prof. PRASENJIT MAJUMDER)
Project Guide

Acknowledgements

I am profoundly grateful to **Prof. PRASENJIT MAJUMDER** for his expert guidance and continuous encouragement with full of motivation throughout to see that this project rights its target since its commencement to its completion.

At last we must express our sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped me directly or indirectly during this course of work.

Ketul Shah

ABSTRACT

In today's era human life has been influences a lot by the modern computers. Communication between human and computer is the most crucial part. To achieve that task we need Natural Language Processing to be handy and efficient.Although this techniques can be applied to improvement of text summarizing, automated chat-bots, machine translation, text captioning and many others.

Hidden Markov Models and chains were the most used architectures for the above listed tasks in past. But as Computation power and enhancement of GPU , nowadays Neural Network architecture made breakthrough in Artificial Intelligence domain. For generation of text Recurrent Neural Networks along with its little variants like LSTM and GRUs can be used.

Project work presented in this report mainly focused on training and predicting next dialogue by training LSTM on different data-sets.Finally I am able to generate some relevant dialogues with using Ubuntu channel's IRC chat corpus in that de-velopers tend to help each other.Lastly, graph between cross entropy(Log Loss) and number of epochs plotted.

Contents

1 Introduction **2**

2 Recurrent Neural Network **3**

2.1 Basics of LSTM 3

2.2 Advantages and Limitations of LSTMs 6

2.2.1 Advantages 6

2.2.2 Limitations 6

3 Problem statement **7**

4 Project Methodology **8**

4.1 Google Colab 8

5 Text Corpus **11**

6 Experimental Results **12**

7 Conclusion and Future Scope **14**

7.1 Conclusion 14

7.2 Future Scope 14

References **14**

List of Figures

2.1 **Unrolled RNN** 3
2.2 **LSTM Memory Cell** 4
2.3 **LSTM Repeating Module** 5

4.1 **Tesla K80 GPU** 8
4.2 **Screenshot of Colab Jupyter Notebook** 9
4.3 **Developement stages** 10

5.1 **Screenshot of Ubuntu IRC** 11

6.1 **Screenshot of LSTM Training** 12
6.2 **Graph of Cross Entropy vs. Epochs** 13
6.3 **Demo of ChatBot** 13

Chapter 1

Introduction

Human like natural language can be generated using various methods including Hidden Markov model and related architectures. But because of advances in Hardware and Graphical Processing Unit(GPU) availability, Neural Network architectures really outperforms and gives us the best result till now. But for the language modeling and sentence generation we cannot use simple vanilla Neural as it is based on input and output based. As we all know the language has context hidden in it. So we need to look at the sequential architecture based models. Recurrent Neural Network is widely used for sequential modeling. It has got two variants that are LSTM(Long Short Term Memory) and GRU(Gated Recurrent Unit) which were designed for several advantages.Hidden layers in this network got the ability to remember. Popular example can be use of is/are in the sentence based on singular or plural noun.

In this project I have trained model using LSTM architecture by providing text corpus and after that trying to predict the sentences based on the seed. Seed is generally provided by the user which is generally in the form of conversation with the machine.

Chapter 2

Recurrent Neural Network

2.1 Basics of LSTM

In simple vanilla Neural Network it is assumed that input and output are independent of each other.But in this project our task is to predict next word based on the previous words supplied to machine. Hence for this kind of tasks RNN would be needed. The name recurrent suggests that the architecture has loops in it. The unrolled version of the Recurrent Neural Network is shown in Figure 2.1

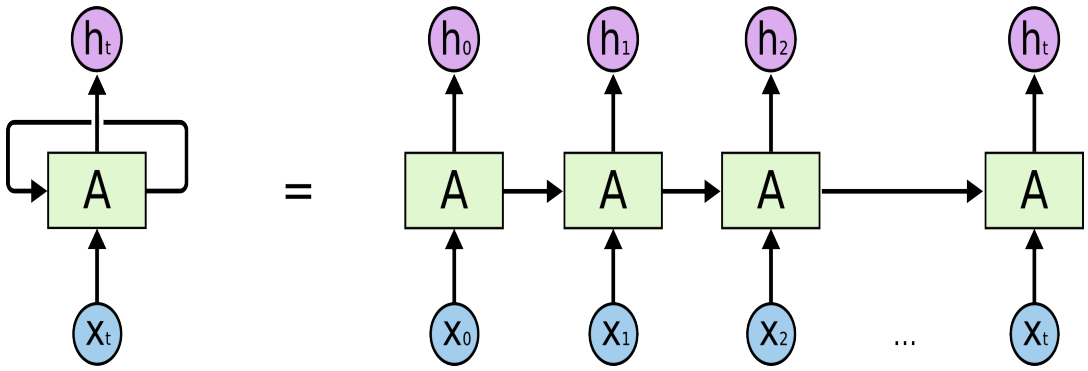

Figure 2.1: Unrolled RNN

The above figure shows RNN's chain like nature. So RNN can be used when input data is form of sequences and lists.So from the architecture we can say that every RNN cell can accept input from the previous cell or from the user input.

LSTM networks were first introduced by Hochreiter and Schmidhuber in 1997. The main purpose for the development of LSTM was to solve long range dependency between sequences of input. So essentially memory elements was added to neural network for remembering the past sequence.

Memory cells in LSTM prevents network immune to vanishing or exploding gradient problem that is generally faced by other neural networks. Memeory cell has basically four main elements which includes input , forget , output and neuron that has self-recurrent connection. Memory cell can be seen in Figure 2.2

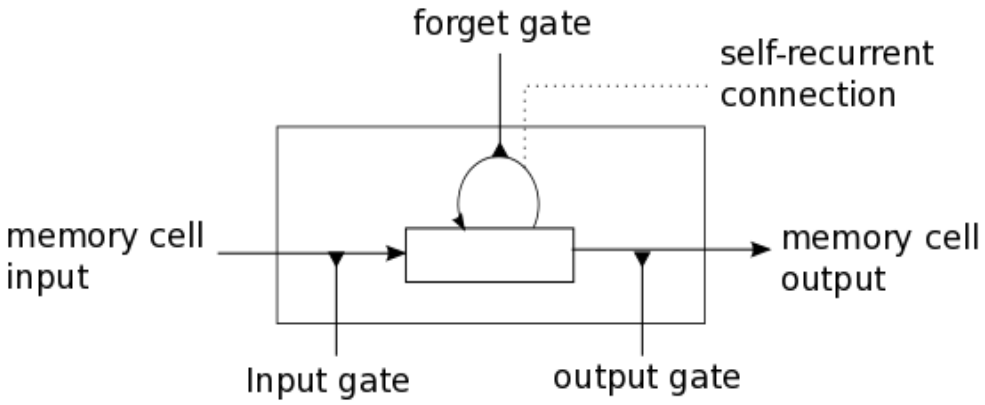

Figure 2.2: LSTM Memory Cell

The input gate can reject or allow the input signal to change the state of the memory cell. The output gate may allow or may prevent memory cell's effect on the next neuron.The need of the forget gate is to allow remembering previous state or forgetting it depending on the need.

As we have discussed in RNN about chain structure , LSTM has the same architecture except the repeating modules. The modules which are repeating contains four interacting layers instead of single layer. This can be shown in the below Figure 2.3. Yellow boxes in the figure shows neural network layers which will be used for learning purpose.

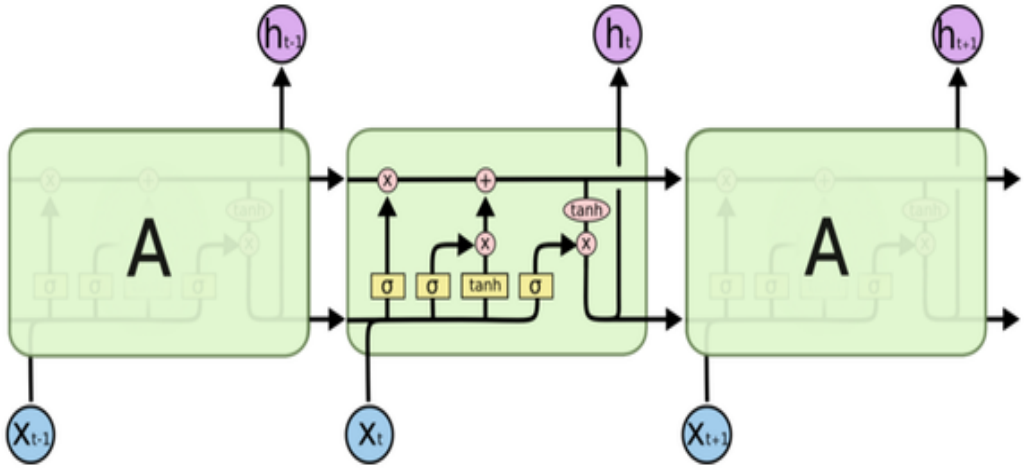

Figure 2.3: LSTM Repeating Module

As discussed earlier the forget ate decides what information should be remembered and what should be forgotten. As machine works in the binary format , for remembering the information it gives 1 as output else 0 as output for which information should be remembered. As shown in the figure 2.3 it is seen that for activation function tanh is used.

Final step would be deciding the output or prediction by the network. So it is purely based on the state of the cell in LSTM. So first cell's output is decided by the sigmoid layer and state of cell is decided based on tahh layer taht leads to value to be between -1 and 1.After than finally by multiplying this value with gate of sigmoid will lead to the output from which we can get words as output.

2.2 Advantages and Limitations of LSTMs

2.2.1 Advantages

- LSTMs can be implemented in wast areas of applications such as speech processing , Language modelling, time series prediction and also in the music composition.
- Complexity of this algorithm is O(k) for every weight and time-steps.
- Also LSTM can handle noise, distributed representatives and continuous values also.
- LSTM nets generally avoids the a long time lags because of cell's constant backpropagation.
- Another best advantage is this network can handle unlimited number of data states. This is advantage over the hidden Markov models that can work with only finite states of data.

2.2.2 Limitations

- Constant flow in memory cells happen in LSTM , so the network sees the whole inout sequece length at the same time the famous problem of finding XOR between two input sequences.
- LSTM network can't solve all the problem related to time series predictions like
- As in Figure 2.3 memory cell has additional gates in terms of input and output gate . This may result large number of weights in the whole network. So this may require complex hardware for training and testing.

Chapter 3

Problem statement

Generative model helps to reduce the laborious work of labelling the dataset.Text generation has numerous applications including machine translation , speech recognition, image captioning , helper tool for writers, chat-bots.

In this project my goal is to design interactive chat machine between human and trained model. This conversation should converge with context based sentences. In this process user should get illumination that he is talking with real human. So chat-bot interface should be modeled based on LSTM architecture with relevant text corpus.

Chapter 4

Project Methodology

In the project I have first prepossessed the english text corpus by applying regular expression to the english text. By that irrelevant symbols have been removed for the train data. Tokenized words to int using vocabulary building are supplied to machine for training the LSTM model. Because of training the weights are trained and model will be ready for predicting the words given seed to it. Generally seed is chosen from the conversation from user.

4.1 Google Colab

Colab is the cloud service provided by Google INC at free of cost. It is basically virtual machine given to user along with the Jupyter notebook interface and command line tools. Colab file can be saved in Google drive.

Figure 4.1: Tesla K80 GPU

As shown in Figure 4.1 Google is providing developer to use Tesla K80 GPU with some limited memory in cloud. Also the provided virtual machine comes with some pre-installed libraries used for machine learning including python supported packages like tensorflow, scikit-learn ,pytocrch, numpy.

```
Downloading data from https://s3.amazonaws.com/img-datasets/mnist.npz
11493376/11490434 [==============================] - 1s 0us/step
x_train shape: (60000, 28, 28, 1)
60000 train samples
10000 test samples
Train on 60000 samples, validate on 10000 samples
Epoch 1/12
2018-01-25 23:47:16.992173: I tensorflow/stream_executor/cuda/cuda_gpu_executor.cc:892] successful NUMA node read from SysFS had negative value (-1), but th
2018-01-25 23:47:16.992422: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1030] Found device 0 with properties:
name: Tesla K80 major: 3 minor: 7 memoryClockRate(GHz): 0.8235
pciBusID: 0000:00:04.0
totalMemory: 11.17GiB freeMemory: 505.38MiB
2018-01-25 23:47:16.992455: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1120] Creating TensorFlow device (/device:GPU:0) -> (device: 0, name: Tesla K
22912/60000 [==========>...................] - ETA: 9s - loss: 0.4608 - acc: 0.855160000/60000 [==============================] - 13s 214us/step - loss: 0.2
Epoch 2/12
60000/60000 [==============================] - 11s 188us/step - loss: 0.0866 - acc: 0.9747 - val_loss: 0.0377 - val_acc: 0.9876
Epoch 3/12
58112/60000 [============================>.] - ETA: 0s - loss: 0.0656 - acc: 0.980260000/60000 [==============================] - 11s 187us/step - loss: 0.0
Epoch 4/12
60000/60000 [==============================] - 11s 188us/step - loss: 0.0533 - acc: 0.9840 - val_loss: 0.0297 - val_acc: 0.9911
Epoch 5/12
60000/60000 [==============================] - 11s 188us/step - loss: 0.0469 - acc: 0.9860 - val_loss: 0.0305 - val_acc: 0.9895
Epoch 6/12
 3584/60000 [>.............................] - ETA: 9s - loss: 0.0331 - acc: 0.990860000/60000 [==============================] - 11s 186us/step - loss: 0.0
Epoch 7/12
60000/60000 [==============================] - 11s 187us/step - loss: 0.0393 - acc: 0.9877 - val_loss: 0.0256 - val_acc: 0.9918
Epoch 8/12
52352/60000 [=========================>....] - ETA: 1s - loss: 0.0355 - acc: 0.989460000/60000 [==============================] - 11s 187us/step - loss: 0.0
Epoch 9/12
60000/60000 [==============================] - 11s 186us/step - loss: 0.0317 - acc: 0.9902 - val_loss: 0.0268 - val_acc: 0.9919
Epoch 10/12
60000/60000 [==============================] - 11s 187us/step - loss: 0.0297 - acc: 0.9915 - val_loss: 0.0275 - val_acc: 0.9922
Epoch 11/12
 2304/60000 [>.............................] - ETA: 10s - loss: 0.0325 - acc: 0.989160000/60000 [==============================] - 11s 187us/step - loss: 0.
Epoch 12/12
60000/60000 [==============================] - 11s 189us/step - loss: 0.0272 - acc: 0.9917 - val_loss: 0.0254 - val_acc: 0.9923
Test loss: 0.02544446041899282
Test accuracy: 0.9923
```

Figure 4.2: Screenshot of Colab Jupyter Notebook

For this project I have used Kears API which is written in Python language which has tensorflow library running at backend. As python is more developer friendly and widely used by the community in the world, I have also done the whole project using coding language as python using Jupyter notebook.

Figure 4.3 shows flow graph of the text generation system. So for every block different code blocks in jupyter notebook is written. Model building is essentially done using Keras API by selecting several hyper parameters.

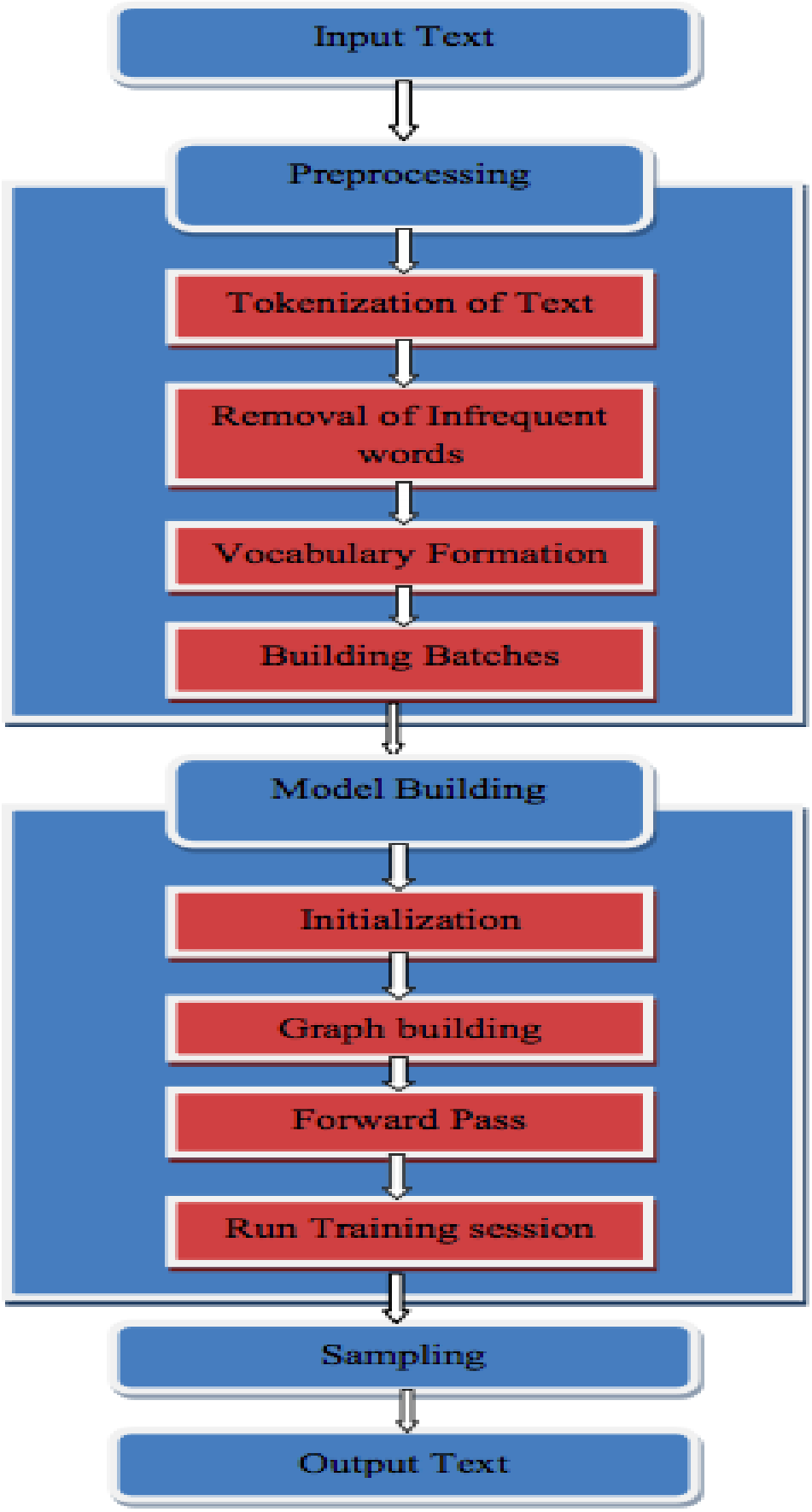

Figure 4.3: Developement stages

Chapter 5

Text Corpus

There are many available text corpus that can be used for language modeling starting from corpus of simple stories to news article and blogs. So as project goal is to design chat-bot, it would be useful if corpus of particular topic is selected and LSTM is trained using that.Here in this project I had trained LSTM using Ubuntu's IRC(Internet Relay Chat) corpus, in which generally developer help each other for any issues/bugs in the developed operating system. So for any question related to ubuntu if asked to trained machine so it can give relevant answer. Figure 5.1 shows the screenshot of IRC chat of ubuntu channel. In which it can be seen one user asks question and the community will try to converge in the direction of solution.

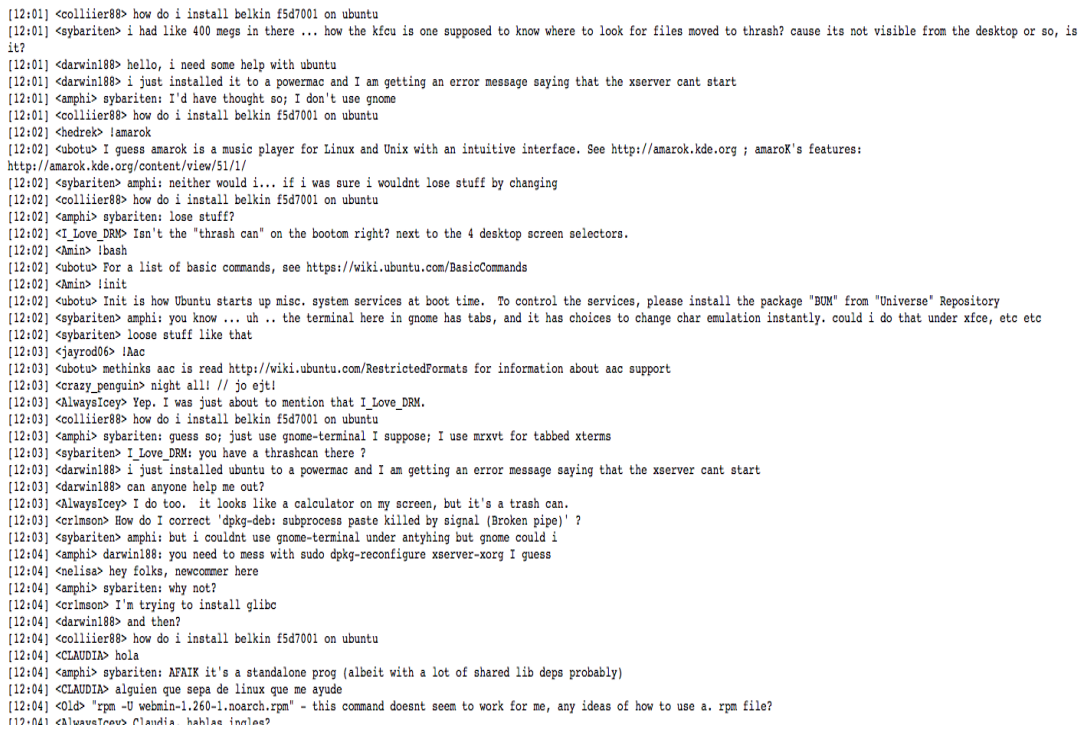
[12:01] <colliier88> how do i install belkin f5d7001 on ubuntu
[12:01] <sybariten> i had like 400 megs in there ... how the kfcu is one supposed to know where to look for files moved to thrash? cause its not visible from the desktop or so, is it?
[12:01] <darwin188> hello, i need some help with ubuntu
[12:01] <darwin188> i just installed it to a powermac and I am getting an error message saying that the xserver cant start
[12:01] <amphi> sybariten: I'd have thought so; I don't use gnome
[12:01] <colliier88> how do i install belkin f5d7001 on ubuntu
[12:02] <hedrek> !amarok
[12:02] <ubotu> I guess amarok is a music player for Linux and Unix with an intuitive interface. See http://amarok.kde.org ; amaroK's features: http://amarok.kde.org/content/view/51/1/
[12:02] <sybariten> amphi: neither would i... if i was sure i wouldnt lose stuff by changing
[12:02] <colliier88> how do i install belkin f5d7001 on ubuntu
[12:02] <amphi> sybariten: lose stuff?
[12:02] <I_Love_DRM> Isn't the "thrash can" on the bootom right? next to the 4 desktop screen selectors.
[12:02] <Amin> !bash
[12:02] <ubotu> For a list of basic commands, see https://wiki.ubuntu.com/BasicCommands
[12:02] <Amin> !init
[12:02] <ubotu> Init is how Ubuntu starts up misc. system services at boot time. To control the services, please install the package "BUM" from "Universe" Repository
[12:02] <sybariten> amphi: you know ... uh .. the terminal here in gnome has tabs, and it has choices to change char emulation instantly. could i do that under xfce, etc etc
[12:02] <sybariten> loose stuff like that
[12:03] <jayrod06> !Aac
[12:03] <ubotu> methinks aac is read http://wiki.ubuntu.com/RestrictedFormats for information about aac support
[12:03] <crazy_penguin> night all! // jo ejt!
[12:03] <AlwaysIcey> Yep. I was just about to mention that I_Love_DRM.
[12:03] <colliier88> how do i install belkin f5d7001 on ubuntu
[12:03] <amphi> sybariten: guess so; just use gnome-terminal I suppose; I use mrxvt for tabbed xterms
[12:03] <sybariten> I_Love_DRM: you have a thrashcan there ?
[12:03] <darwin188> i just installed ubuntu to a powermac and I am getting an error message saying that the xserver cant start
[12:03] <darwin188> can anyone help me out?
[12:03] <AlwaysIcey> I do too. it looks like a calculator on my screen, but it's a trash can.
[12:03] <crlmson> How do I correct 'dpkg-deb: subprocess paste killed by signal (Broken pipe)' ?
[12:03] <sybariten> amphi: but i couldnt use gnome-terminal under antyhing but gnome could i
[12:04] <amphi> darwin188: you need to mess with sudo dpkg-reconfigure xserver-xorg I guess
[12:04] <nelisa> hey folks, newcommer here
[12:04] <amphi> sybariten: why not?
[12:04] <crlmson> I'm trying to install glibc
[12:04] <darwin188> and then?
[12:04] <colliier88> how do i install belkin f5d7001 on ubuntu
[12:04] <CLAUDIA> hola
[12:04] <amphi> sybariten: AFAIK it's a standalone prog (albeit with a lot of shared lib deps probably)
[12:04] <CLAUDIA> alguien que sepa de linux que me ayude
[12:04] <Old> "rpm -U webmin-1.260-1.noarch.rpm" - this command doesnt seem to work for me, any ideas of how to use a. rpm file?
[12:04] <AlwaysIcey> Claudia, hablas ingles?

Figure 5.1: Screenshot of Ubuntu IRC

Chapter 6

Experimental Results

In Figure 6.1 training of LSTM network can be seen. It can be observed that in starting phase of epochs, the cross entropy loss of the model is very high and thus it is predicting the words very poorly. And yes it can be compared to the child learning English language.

```
Epoch 1/1
12586/12586 [==============================] - 10s 770us/step - loss: 5.5506
Loss improved from 6.557857509639103 to 5.550561289178838. Saving weights.

-> seed: "and then whatever you like , gimp for example
 i purchased some tv shows" ...

 to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to to

---------- epoch 3/50 ----------
Epoch 1/1
 6016/12586 [=============>................] - ETA: 5s - loss: 5.015612586/12586 [==============================] - 10s 770us/step - loss: 4.9969
Loss improved from 5.550561289178838 to 4.996913215531731. Saving weights.

-> seed: "the easiest thing to do is a cd to it s installed directory , and" ...

 i have to get the same
 i m not
 i m not
 i m not
 i m not
 i m not
 i m not
 i m not
 i m not
 i m not
 i m not
 i m not

---------- epoch 4/50 ----------
Epoch 1/1
12586/12586 [==============================] - 10s 774us/step - loss: 4.5223
Loss improved from 4.996913215531731 to 4.522315947027857. Saving weights.

-> seed: "it to a powermac and i am getting an error message saying that the xserver" ...


 i do it
 what s it
 i dont it in the same
 i do i install it
 it s, it s, it s, it s, it s, it s, it s, it s, it s, it
```

Figure 6.1: Screenshot of LSTM Training

Figure 6.2 shows the graph between cross entropy loss versus number of epochs used for training of the network.

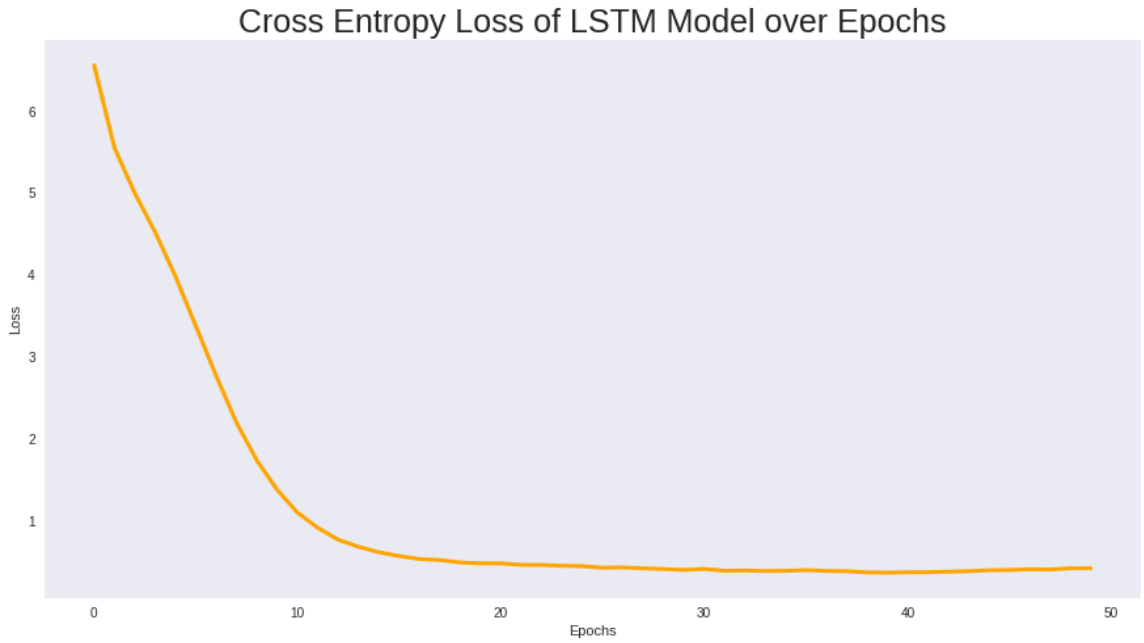

Figure 6.2: Graph of Cross Entropy vs. Epochs

Figure 6.4 shows the final demonstration of conversation between two machines where initial seed is provided by user. The output generated by first block is supplied to another as seed.

```
In [104]: chatbot("is there a way to lower my screen resolution, Ubuntu loaded tops and is hard on my eyes",15)

          User 0 is saying :
          is there a way to lower my screen resolution, Ubuntu loaded tops and is hard on my eyes

          User 1 is saying :
           now on boot an i tried source to work they me get pissed a dell as if as ve can

          User 0 is saying :
           now on boot an i tried source that in knorrie
           keithsr with that. if you need to run

          User 1 is saying :
           now on a an.. might
           darwin you. think this on for linux, you just says

          User 0 is saying :
           now on boot an i m source
           adrums so.
           you can t even up the on in

          User 1 is saying :
           there. then the...
           you can help me getting ndiswrapper is.. in you if

          User 0 is saying :
           on on help sata how to do then
           amphi ok i want to give it i have a raid

          User 1 is saying :
           now on boot an i tried file it in if root now darwin. try are
           experiencing
           keithsr

          User 0 is saying :
           now on
           an can i an server;
```

Figure 6.3: Demo of ChatBot

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

The project presents text generation system using LSTM network.In this project I have used the corpus of ubuntu IRC chats that can be used for training LSTM model. In this I have tried training model using different hyper-parameters like epochs and learning rate. Although , the model is not giving the perfect output in terms of sentence, it is able to generate relevant word based on the seed provided. One of the reason is the less vocabulary size provided. As GPU access is provided by Google colab training part becomes really easy and time efficient.

Developed code can be found at the link :- https://goo.gl/QVJ62M

7.2 Future Scope

- Proper text corpus selection can be done by further analysis including pre-processing.
- Online training based Model which learns with every conversation with user.
- Reinforcement learning can be added to make more accurate output.

References

[1] *"Training deep and recurrent networks with hessian-free optimization." In Neural networks*; Martens J. and Sutskever I.

[2] *"Generating text with recurrent neural networks., "* ; Sutskever I., Martens J. and Hinton G.E

[3] *" Writing Stories with Help from Recurrent Neural Networks.,"* ; Roemmele M. and Intelligence D.D.N

[4] *"Text generation from keywords.,"* ;Uchimoto K., Isahara H., and Sekine S.