Suggestion Generation for Specific Erroneous Part in a Sentence using Deep Learning

Veena S Nair[1], Amina Beevi A[2]

[1]M.Tech, [2]Assistant Professor
[1,2]Department of Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India

How to cite this paper: Veena S Nair | Amina Beevi A "Suggestion Generation for Specific Erroneous Part in a Sentence using Deep Learning" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-4, June 2019, pp.582-584, URL: https://www.ijtsrd.com/papers/ijtsrd23842.pdf

IJTSRD23842

ABSTRACT

Natural Language Processing (NLP) is the one of the major filed of Natural Language Generation (NLG). NLG can generate natural language from a machine representation. Generating suggestions for a sentence especially for Indian languages is much difficult. One of the major reason is that it is morphologically rich and the format is just reverse of English language. By using deep learning approach with the help of Long Short Term Memory (LSTM) layers we can generate a possible set of solutions for erroneous part in a sentence. To effectively generate a bunch of sentences having equivalent meaning as the original sentence using Deep Learning (DL) approach is to train a model on this task, e.g. we need thousands of examples of inputs and outputs with which to train a model.

Keywords: *Natural Language Processing (NLP), Deep Learning (DL), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM)*

INTRODUCTION

Generating suggestions for a sentence especially for Indian languages is much difficult. One of the major reason is that it is morphologically rich and the format is just reverse of English language. By using deep learning approach with the help of LSTM layers we can generate a possible set of solutions for erroneous part in a sentence. To effectively generate a bunch of sentences having equivalent meaning as the original sentence using deep learning approach is to train a model on this task, e.g. we need thousands of examples of inputs and outputs with which to train a model. The proposed system is worked based on supervised learning strategy.

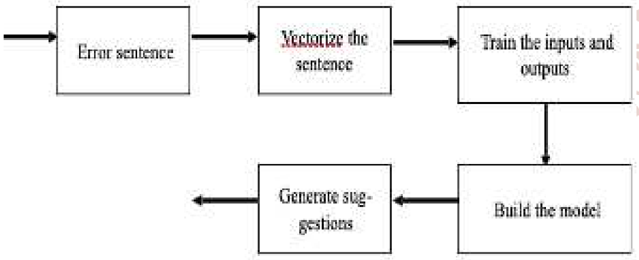


Figure1. Simple concept of suggestion generation

RELATED WORKS

Several works can be done in the field of nature language processing. Various studies and related works can be done in the case of Languages of India. Mostly the work can be done in English language, they can be explained in detail on the site [6][7]. For the training data for predictions Recurrent Neural Network (RNN) does not used exact templates. RNN with Long Short Term Memory (LSTM) are storing and accessing data. And also LSTM can predict the next level of sequence from the raw text [2]. Most difficult training task can be achieved by using deep neural network. They consist of many hidden layers, and most NLP operations can performed by using LSTM layer in the hidden layer.LSTM is one of the hidden layer of RNN [3].

In the paper [8] Adrain Sanborn and Jacek Skryzalin said that the major difficult task is to identify the sentence meaning (of any language). These task has broad applications such as image captioning, text generation, machine translation etc. In this paper [8] they can be used GloVe vector for sentence embedding and also they used pre trained word vectors for the dataset. They can do their experiments through two deep learning models such as Recurrent Neural Network (RNN) and Recursive Neural Network. Both the two network take input as tokens but they treat the input differently. Both the model has been trained, for these two models they have reached at the statement that recurrent neural network is much better than recursive neural network.

In the paper [4] Jinyue Su, Jiacheng Xu, Xipeng Qiu, Xuanjing Huang said that they proposed a sentence generator framework based on Gibbs Sampling (GS). They contains two separate model one is a language model and another one is discriminator. The language model is not aware of the constraint information and the discriminator estimates the probability of the sentences. They also proposed a candidate generator which gives more likely words at each iteration of the network. The Gibbs Sampling (GS) is purely based on Markov chain which is a statistical approach. The find out the probability of each word and the sentence is generated with maximum probability value using RNN.

A specialised generative model called conditional generative model which generate the output sample conditioned on some given input. In the paper [5] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan, they proposed an approach called TextDream which is used to search the text data on the continuous semantic space. The Text Dream is worked on some evolutionary algorithms. They can encoded the documents and decoded with Markov chain rule and check whether the solution is found or not.

METHEDOLOGY

The proposed system is generating suggestions for specific erroneous part in a Malayalam sentence with the help of Recurrent Neural Network (RNN). It is a part of Malayalam error detection and grammar correction tool of CDAC, Trivandrum. One of the most widespread forms of sequence data is text. They can be understood as either a sequence of characters or a sequence of words. Deep learning has penetrated into multiple and diverse industries, and it continues to break new ground on an almost weekly basis. Supervised learning refers to a task where we need to find a function that can map input to corresponding outputs (given a set of input-output pairs). By using deep learning with the help of tensorflow platform the input text can be processed. The input must be a text or a sentence and the goal is:

1. Collection of samples of language texts
2. Identification of errors
3. Train the network
4. Generating suggestion for errors

There are two modules in this project, data preprocessing and suggestion generation.

A. Data Preprocessing

The first part of the proposed system is data preparation module. For deep learning natural language project first of all we can convert the input documents as integers. Data Preprocessing is a major task in all Deep Learning applications. The work has been focused on word length three Malayalam sentences from health and newspaper domain. The input data fed into the deep neural net must be in a pattern compatible to the network. Basically data preprocessing is undergone through a series of steps like cleaning, stemming, stop word removal, etc. In the proposed system, the dataset consist of Malayalam sentences. Here the dataset contains three length word. First load the dataset using pandas. Pandas is a package used to load the documents (in any format). Then we can create some suggestions for that input sentences. Also we can labeled that input sentences and their suggestions. From that label we can find the accuracy of the network model. All the input sentences in a deep learning project can be in integer encoded because the network can only learn the datas as integers. Al the operations can be performed by using numpy module in the python.

B. Suggestion Generation

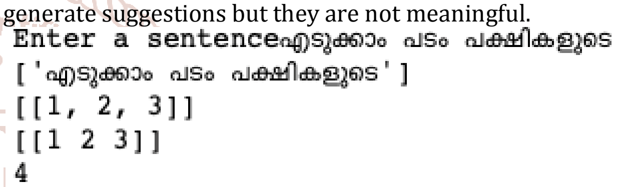
Suggestion generation is a difficult task for natural language processing. A considerable work can be done in the case of image captioning using Recurrent Neural Network (RNN). The output from the data preprocessing module can fed into suggestion generation module. Here suggestions can generated by using deep neural networks. In this project we used a sequential model with three layers such as embedding layer, flatten and dense with dropout of 0.2%. And from we can train the test set and training set, and also find the accuracy of the model. To improve the accuracy of the model we can increase the number of epochs. From that a bunch of possible solutions can be generated.

EXPERIMENTAL ANALYSIS

A. Result and Analysis

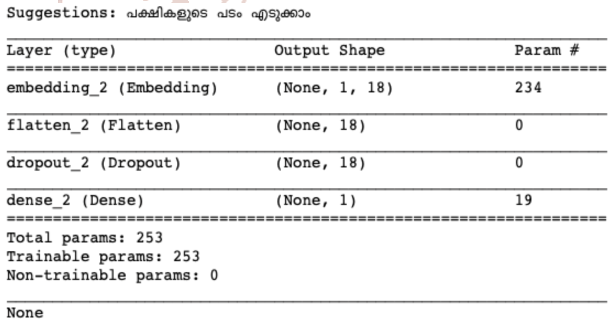
This section discusses the experimental results of the proposed system. The system that uses the operating system for macOS Mojave and python jupyter as platform. The proposed system is using original Malayalam documents given by CDAC, Trivandrum for results assessment. There are three different types of datasets are given. The proposed system is implemented using two modules ie., data preparation module and suggestion generation module. In data preparation module the input and output sentences must be encoded to integers and also removing the duplication of the similar words in the document. Natural Language Processing (NLP) is the main task of natural language generation and they can be implemented by using deep learning with the help of Recurrent Neural Network (RNN) with Long Short Term Memory layers (LSTM).

Suggestion generation for grammatically erroneous part in a sentence can be used a word length of three because this length can generate an accurate result from our network. Also late sentences can be studied by the network and also generate suggestions but they are not meaningful.



```
Enter a sentenceഎടുക്കാം പടം പക്ഷികളുടെ
['എടുക്കാം പടം പക്ഷികളുടെ']
[[1, 2, 3]]
[[1 2 3]]
4
```

Figure2. Input sentence having length three



```
Suggestions: പക്ഷികളുടെ പടം എടുക്കാം
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 1, 18)             234
_________________________________________________________________
flatten_2 (Flatten)          (None, 18)                0
_________________________________________________________________
dropout_2 (Dropout)          (None, 18)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 19
=================================================================
Total params: 253
Trainable params: 253
Non-trainable params: 0
_________________________________________________________________
None
```

Figure3. Network model summary

The proposed system shows the not exactly the suggestion generation for Malayalam grammar correction. Because the Malayalam language consist of different symbols like ൠ, എമ്മിന്റെ, etc. They cannot exactly read by the system because Malayalam is a morphologicaly rich language and no further studies can be done using deep neural network concepts. This can be overcome by using different datasets to train the network as well as also to develop a nltk tool for Malayalam language. The analysis phase includes the loss and accuracy of the training dataset. This analysis shows the word length of three as the input sentence. Here the training loss is much less than the validation loss because to increase the number of epochs on the network we can easily generate an accurate loss curve for the training data. The accuracy of the training data is high. Again to improve the accuracy of the model we can train more input sentences and inches the number of epochs of the deep neural network.

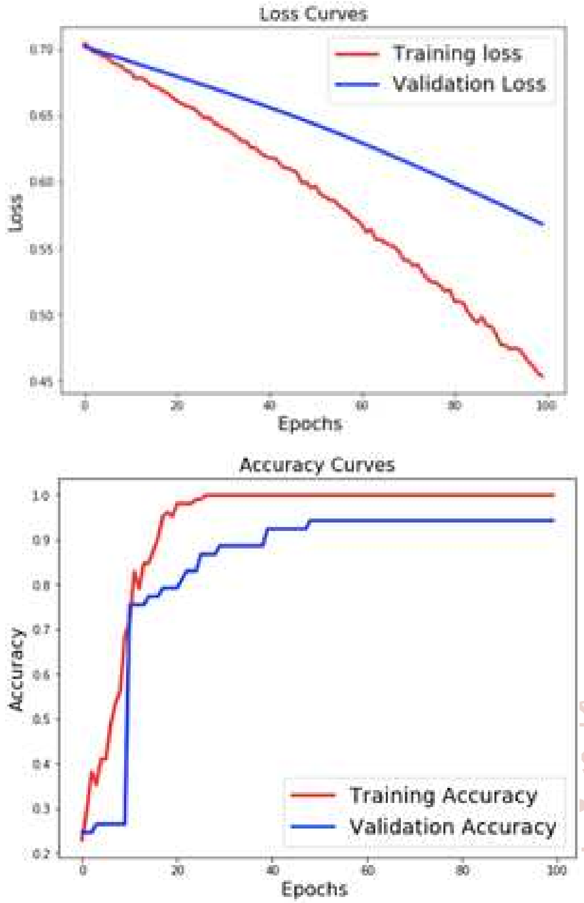


Figure4. Loss and accuracy curve for the training data

B. Performance Requirements

Performance requirements describes all the hardware specification of the system. For our project work firstly we have a Dell laptop with Intel pentium processor, but they cannot be used, because pentium processor does not contains AVX instructions. Advanced Vector eXtension (AVX) instruction is used extensions to the x86 instructions set architecture for microprocessors from Intel. For deep learning project, the minimum system requirement is it needs i3 or above processor. Mostly deep learning projects used 64 bit Ubuntu operating system. Here we used a Mac OS with 51.8 GHz Intel Core i5 processor for this project (For Mac OS only 10.12.6 Sierra or later os versions must be used). Our Mac OS contains 8GB RAM, 125GB hard drive space.

CONCLUSION

Suggestion generation is one of the task of Natural Language Processing (NLP). They can be implemented by using Keras deep learning. First the work can be done using text generation of a Malayalam document. But the disadvantage is that some of the letters in Malayalam can be repeatedly comes in the text generation. This is one of the drawback of Malayalam text generation using deep learning approach. The proposed system developed for some erroneous part in a sentence, using deep learning approach, a bunch of sentences with similar meanings has be generated. The main goal is to generate sentences for Malayalam error documents of different structures, but they need more time and need a requires more dataset. In this project only a specific structure of sentence has been used and their corresponding suggestions has been generated.

Acknowledgment

I undersigned hereby declare that the project “Suggestion Generation for Specific Errorneous part in a sentence using Deep Learning”, submitted for partial fulfillment of the requirements for the award of degree of Master of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Ms. Amina Beevi A. This project is a live re search and development project contributed from CDAC- TVM. This submission represents my ideas in my own words and where ideas or words of others have been included, I have an adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission.

References

[1] Veena S Nair, Amina Beevi A, “Survey On Generating Suggestions For Specific Erroneous Part In A Sentence”, IRJET, Volume: 05 Issue: 12 | Dec 2018.

[2] Veena S Nair, Amina Beevi A, “Malayalam Text Generation Using Deep Learning”, IJCEA, Volume XIII, Issue III, March 2019.

[3] https://towardsdatascience.com/tagged/nlp

[4] https://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras

[5] Adrian Sanborn, Jacek Skryzalin, “Deep Learning for Semantic Similarity”, Department of Computer Science Stanford University, 2015.

[6] Alex Graves. “Generating Sequences With Recurrent Neural Networks”, University of Toronto, 5 June 2014.

[7] Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria, “Recent Trends in Deep Learning Based Natural Language Processing”, 25 Nov 2018.

[8] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. “TextDream: Conditional Text Generation by Searching in the Semantic Space”, IEEE, 2018.

[9] Jinyue Su, Jiacheng Xu, Xipeng Qiu, Xuanjing Huang. “Incorporating Discriminator in Sentence Generation: A Gibbs Sampling Method”, The Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[10] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. “Sequence to Sequence Learning with Neural Networks”, Google.