SURVEY ON GENERATING SUGGESTIONS FOR ERRONEOUS PART IN A SENTENCE

VEENA S NAIR[1], AMINA BEEVI A[2]

[1]M.Tech, Computer Science & Engineering, Sree Buddha College of Engineering, Ayathil, Elavumthitta, Pathanamthitta, Kerala, India.
[2]Assistant Professor, Computer Science & Engineering, Sree Buddha College of Engineering, Ayathil, Elavumthitta, Pathanamthitta, Kerala, India

---

Abstract - *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.*

***Key Words*: Natural Language processing, Deep learning, Indian languages, RNN.**

1. INTRODUCTION

Natural language processing deals with the human language and computers interaction. Natural language processing (NLP) is either rule based or statistical based approach. Rule based means they contains a set of rules that can be already defined in the system. According to that rule they can be processed and generating results. But in the case of statistical approach there is no rules and the output that we can trained itself.

The concept of machine learning and deep learning are similar, but the difference is that, machine learning algorithms are based on rule based approach and deep learning is statistical approach. In machine learning processing can be done step by step ie., whenever the first level output is generated, then only they can be enter into the next level. It is more time consuming. The major benefit of deep learning over machine learning is fast, it takes a single step to process the input data.

The important problem of Natural Language Processing (NLP) is the modeling of the structure of the text. Beating with the NLP field feels the closest to achieving the Artificial Intelligence (AI). The task of such text is that to identify the structure. NLP can sparse much more difficult data ie., they can perform extremely complex task. Therefore the ambiguity of NLP is very high.

The disadvantage of machine learning is that the input in machine learning (ML) uses a fixed length feature vector. The significant feature of NLP is text categorization with certain applications such as web based searching, document or text classification and prediction, speech recognition, question answering, language modeling, etc.

2. DEEP LEARNING APPROACHES

The newest area of Machine Learning (ML) is Deep Learning. Most of the deep learning approaches can be performed by using neural network layers. The form of representation of data in deep learning must include input layer, output layer and a lots of hidden layers. The features from the lower layer hierarchies get to form the higher layer. The hidden layer can perform all the operations, they contains a LSTM (Long Short Term Memory) layer. This hidden layers contains all the complex layer features. They can produces semantically meaningful concepts.

In supervised learning they contains recurrent neural network (RNN), convolutional neural networks (CNN), etc. And in unsupervised learning techniques they contains sparse coding, restricted BM, etc. In the case of using deep learning approach, they use deep multi-layer belief networks, they contains a set of encoders and decoders to perform all the operations. The working principle of LSTM layers may contains four components ie., the input gate, output gate, forgot gate and a self recurrent connection. By using these gate they can add or remove the data.

A fixed length feature vector is used for many machine learning algorithms. RNN is used for efficient sentence structure representation. Therefore they can easily capture the semantics of the sentence. By using deep learning with LSTM layers they can pre-trained the word vectors and produces a short sentence [3].

2.1. Convolutional Neural Networks

In Convolutional Neural Networks (CNN) they consist of a weight and biases corresponding to each layers. Mostly they can be used for image pre-processing, POS tagging, sentiment analysis, etc. They can be also known as shift invariant or space invariant artificial neural network. In CNN they requires large amount of data and they does not works for sequential data.

2.2. Recurrent Neural Networks

Recurrent neural network (RNN) is the sub-class of artificial neural network. By using the internal memory they can process the sequence of inputs. The important thing of using RNN is that the operations can be performed may depends upon the input they received in the internal memory. The input is of two types ie., the recent past input and the present. The information can performed in RNN can be a loop. By using these inputs the loop operation can be performed in the hidden layer [4].

2.2.1. Hierarchical Recurrent Neural Networks

Hierarchical RNN are used for mainly document modeling. In the earlier stage they can be used word level and sentence level RNN's. Sentence level RNN can predict the next set of words in a document whereas in word level RNN they predicts the word sequence in the hidden layer [2].

2.2.2. Long Short Term Memory

To avoid the long short term dependency problem Long Short Term Memory (LSTM) are used. The cell state in the hidden layer, LSTM have the ability to add or remove the data. Inside the LSTM layer they contains a pointwise operation and a vector transfer. The operations can performed here and concatenate the result and copy to the next unit in the cell [5].

2.2.3. Bidirectional Recurrent Neural Network

Bidirectional recurrent neural network (BRNN) which connects two hidden layers with opposite directions for the same output field. The output of a hidden state can be forward to the next layer and also they can backward the output itself to that corresponding state. By using BRNN, they can generate the missing words and to generate a meaningful sentence ie., they can give a significant impact for the generation of words [4].

2.2.4. Deep Recurrent Neural Network

Here they consist of two hidden layers. So they are also known as deep two layer network. A single hidden layer does not capture the raw data so the additional layer in deep RNN can capture unusual semantic pattern in the corpus. Compared with the vanilla recurrent neural network, the model has just one hidden layer network [4].

3. CLASSIFICATION OF INDIAN LANGUAGES

The languages of India can be mainly classified into two major categories such as Indo-Aryan and Dravidian languages. Languages such as Malayalam, Tamil, Telungu are belongings to Dravidian family and other languages like Gujarati, Bengali, Marati, etc. are belonging to Indo-Aryan languages.

3.1. Research Problem

In the case of Indian language, they are morphologically very rich. Our research area focus on Malayalam language. Malayalam is originated from Sanskrit and Tamil language and the pattern is like SOV (Subject-Object-Verb) format. There are various studies held upon the word error correction but in the case of generating sentence for a corpus is not done.

Sentence ordering is a difficult task in NLP, basically in the case of Indian languages. Normally they can be used Hidden Markov based models (HMM) to model the document structure. They can be used a probabilistic approach for the word alignment. Unlike all the previous approaches, they do not use any features that can be predict the next set of sentences or words in a corpus [6].

3.2. N-Gram Approach

For predicting the next set of items in a document or any other corpuses n-gram technique will be used. N-gram means generating the combination of adjacent n words in a sentence. If the count of n is two then it will takes two words or letters for processing, called bigram. If we taken the count as three then it will be called trigram approach. By using n-gram technique we can generate a possible set of solutions or suggestions for the input text [7].

4. CONCLUSION

Since there is no study can be done in the case of Indian languages. So that it is much more difficult to perform the task. Our future work is to study all the Indian languages with its grammar and generating possible suggestions for the input text. For numerical computation using data flow graph Tensor flow is used. It is an open source software. By using Tensor flow with python the future work can be done.

REFERENCES

[1] Veena S Nair, Amina Beevi A, “Survey on Generating Suggestions For Erroneous Part In A Sentence”.

[2] Lajanugen Logeswaran, Honglak Lee, Dragomir Radev, “Sentence Ordering and Coherence Modeling using Recurrent Neural Networks”, Department of Computer Science & Engineering, University of Michigan and Yale University, 2017.

[3] Abdalraouf Hassan, Ausif Mahmood, “Deep Learning for Sentence Classification”, Department of Computer Science & Engineering, University of Bridgeport, CT, 06604, USA.

[4] Arathi Mani, “Solving Text Imputation Using Recurrent Neural Networks”, Department of Computer Science, Stanford University, Stanford, CA 94305.

[5] Gene Lewis, "Sentence Correction using Recurrent Neural Networks", Department of Computer Science, Stanford University, Stanford, CA 94305.

[6] Stephan Vogel, Hermann Ney, Christoph Tillmann, "HMM-Based Word Alignment in Statistical Translation", Lehrstuhl ffir Informatik V, RWTH Aachen, D-52056 Aachen, Germany.

[7] Ratnasingam Sakuntharaj, Sinnathamby Mahesan, "Use of a Novel Hash-Table for Speeding-up Suggestions for Misspelt Tamil Words", IEEE, 2017.

BIOGRAPHIES

Veena S Nair, received the Bachelor's Degree in Computer Science and Engineering from Mahatma Gandhi University, Kerala, India in 2016. She is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Kerala, India. Her research area of interest includes the field of cyber security, machine learning, and intelligent systems.

Amina Beevi A, She is an Assistant Professor in the Department of Computer Science and Engineering, Sree Buddha College Of Engineering, Kerala. Her research area of interest includes the field of data mining, algorithm analysis.