International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume: 3 | Issue: 4 | May-Jun 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 - 6470

Phrase Structure Identification and Classification of Sentences using Deep Learning

Hashi Haris[1], Misha Ravi[2]

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

How to cite this paper: Hashi Haris | Misha Ravi "Phrase Structure Identification and Classification of Sentences 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.578-581, URL: https://www.ijtsrd.com/papers/ijtsrd23841.pdf

IJTSRD23841

ABSTRACT

Phrase structure is the arrangement of words in a specific order based on the constraints of a specified language. This arrangement is based on some phrase structure rules which are according to the productions in context free grammar. The identification of the phrase structure can be done by breaking the specified natural language sentence into its constituents that may be lexical and phrasal categories. These phrase structures can be identified using parsing of the sentences which is nothing but syntactic analysis. The proposed system deals with this problem using Deep Learning strategy. Instead of using Rule Based technique, supervised learning with sequence labelling is done using IOB labelling. This is a sequence classification problem which has been trained and modeled using RNN-LSTM. The proposed work has shown a considerable result and can be applied in many applications of NLP.

Keywords: *Deep Learning, Neural Network, Natural Language Processing, Phrase Structure, Artificial Intelligence*

INTRODUCTION

We are living in an era in which technology is flourishing at a faster pace. The world is becoming automated by the rise of Artificial Intelligence. This new era of AI has given birth to very brilliant automated devices and robots which are equivalent to human beings and possesses almost all capabilities of a human. This ability is due to the advent of deep learning technology. The Artificial Neural Network concept has been under research and study since the late 1940's. Inspired from the structure and processing of the Biological neuron, researcher studied and paved way for many applications that can be implemented using the ANNs.

Artificial Neural Network as the name suggests "neural" these are the systems inspired from human brain which intend to replicate the strategy that we humans use to learn things. Neural Networks consist of input, output and hidden layers. The hidden layers are responsible for the transformation of input data into a form that can be used by the output layer. These have become a part of Artificial Intelligence mainly due to the advent of "Back propagation". This technique allows network to adjust the nodes in the hidden layers when the predicted outcome shows deviations from the expected outcome. Another important advancement is the arrival of deep neural nets. Deep neural networks have multiple hidden layers which help the neurons to extract more features along with feature engineering in order to get the desired output. The input data is passed through multiple hidden layers and within each layer more features are extracted and the output is obtained. The deviation in the predicted output from the target output named as the error would be rectified using the backpropagation strategy until the target output is obtained or error is reduced. Then the network would perform the tasks without the human intervention. The strategies used to learn and train the network are supervised, unsupervised or reinforcement learning technique. In case of supervised learning technique, the inputs are given with certain labels and based on those labels, training occurs. In case of unsupervised learning technique, no special features except the input is provided to the network and the network learns different patterns and features on its own and produces the output. Reinforcement learning uses some rewards and penalty techniques for the network to learn.

Phrase structure is the arrangement of words in a specific order based on the constraints of a specified language. This arrangement is based on some phrase structure rules which are according to the productions in context free grammar. The identification of the phrase structure can be done by breaking the specified natural language sentence into its constituents that may be lexical and phrasal categories. These phrases are tagged using IOB sequence labelling. The phrases are classified into Noun Phrase and Verb Phrase. Many of the NLP works have been done in languages like English, Chinese, etc. But very few works have been done in Indian Languages especially South Dravidian languages like Malayalam. Due to lack of resources such languages are resource poor and corpus-based NLP tasks cannot be done. Languages like Malayalam have variations in morphology as compared to other languages like English or Hindi. The nouns have inflections due to to case, gender and number information. The verbs are inflected due to tense, aspect and mood information [1].

The state of the art is to identify the phrase structure of Malayalam language by breaking the sentence into phrases mainly Noun Phrase and Verb Phrase and tag the phrases using suitable IOB tags. Sequence labelling is the strategy applied in NLP for various applications such as named entity recognition, the proposed work is done by applying this strategy. The identification and classification have been done using deep learning. Several deep neural networks are available but for this purpose RNN (Recurrent Neural Network) is used since it is proved to be efficient in sentence structure representations. Long Short-Term Memory Units (LSTM) which is a variety of RNN is used so as to overcome some drawbacks of RNN.

A. RELATED WORKS

1. Convolutional Neural Networks

Convolutional neural networks have been used in many NLP tasks so far such as the pos tagging, chunking, sentence modeling, etc. CNNs are designed in such a way that can capture more important features from sentences. Works have been reported in literature on sentence classification. Such a work done by Kim [2] on sentence classification tasks like question type classification, sentiment, subjectivity classification. But this vanilla network had difficulty in modeling long distance dependencies. This issue was solved by designing a dynamic convolutional neural network. CNNs require large training data in developing semantic models with contextual window. This becomes an issue when data scarcity occurs. Another issue with CNNs is they find difficulty in preserving sequential data.

2. Sentence ordering and coherence modeling using RNN

Modeling the structure of coherent texts is a key NLP problem. The task of coherently organizing a given set of sentences has been commonly used to build and evaluate models that understand such structure. An end-to-end unsupervised deep learning approach based on the set-to-sequence framework is proposed to address this problem [3]. RNNs are now the dominant approach to sequence learning and mapping problems. An RNN-based approach to the sentence ordering problem which exploits the set-to-sequence framework of Vinyals, Bengio, and Kudlur (2015) is proposed. The model is based on the read, process and write framework. The model is comprised of a sentence encoder RNN, an encoder RNN and a decoder RNN.

3. Deep learning for sentence classification

Neural network-based methods have obtained great progress on a variety of natural language processing tasks. The primary role of the neural models is to represent the variable-length text as a fixed-length vector [4]. These models generally consist of a projection layer that maps words, sub-word units or n-grams to vector representations (often trained beforehand with unsupervised methods), and then combine them with the different architectures of neural networks. Recurrent neural networks (RNN) are one of the most popular architectures used in NLP problems because their recurrent structure is very suitable to process the variable-length text. A recurrent neural network (RNN) [Elman, 1990] is able to process a sequence of arbitrary length by recursively applying a transition function to its internal hidden state vector h_t of the input sequence. The activation of the hidden state h_t at time-step t is computed as a function f of the current input symbol x_t and the previous hidden state h_{t-1}. A simple strategy for modeling sequence is to map the input sequence to a fixed-sized vector using one RNN, and then to feed the vector to a softmax layer for classification or other tasks.

B. PROPOSED SYSTEM

The proposed system titled "Phrase Structure Identification and Classification of sentences using Deep Learning" for Indian languages identifies the phrases within a Malayalam sentence and classifies them into Noun and Verb Phrase. Phrase Structure Grammar is an inevitable part of a language. It helps to correctly frame a sentence in an order based on the phrase structure rules. The proposed system has a modern solution to the identification and classification of phrases without the help of rules and no linguistic knowledge is necessary which is possible through the most trending hot technology named "Deep Learning". The relevance of phrase structure identification is checking the grammar of a natural language becomes more accurate and efficient if the phrases are identified correctly and it helps to track the error within phrase level and it's easy to rectify those error phrases. Instead of taking into consideration the entire document or sentence, phrase level identification helps in easy error tracking and rectification of grammar.

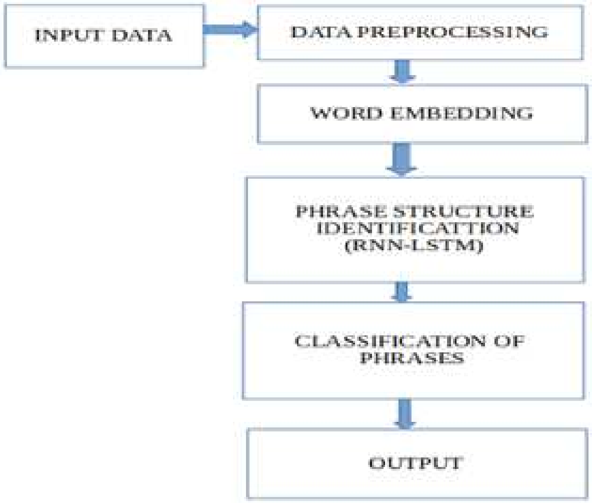

INPUT DATA
DATA PREPROCESSING
WORD EMBEDDING
PHRASE STRUCTURE IDENTIFICATTION (RNN-LSTM)
CLASSIFICATION OF PHRASES
OUTPUT

Figure B.1: Architecture of Proposed System

Phrase Structure identification and classification has been done by the combination of two modules:

1. Data Preprocessing

The proposed system can be generalized into a Text Classification problem. Text Classification is a common task in NLP which would transform a sequence of text with indefinite length into a category of text. Data Preprocessing is a major task in all Deep Learning applications. 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 consists of Malayalam sentences. The preprocessing done here is splitting the sentences into phrases. Each sentence has a sentenced and its constituent phrases also have the same sentence id. The dataset is represented using pandas data frame. The work has been focused on word length three Malayalam sentences from health and newspaper domain.

Next a representation of the input text is created so that it could be fed into the deep learning model. The phrases in the text are represented as vectors. The phrases are mapped into a dictionary with a sequence of numbers and these numbers are the vector representation of the phrases. These phrases are then padded. Since it is classification model the phrases are tagged into Noun and Verb phrase. These tag sequences are represented as dictionary with sequence of numbers and then padded. To train the model the tag sequences or the labels are represented in a categorical format. The Noun phrase and Verb phrase are labelled as NP and VP. Sequence labelling is the strategy applied with IOB labels I-refers to inside of a chunk, O- outside of a chunk, B- beginning of a chunk. The proposed model is trained using supervised learning and it's a sequence or text classification task so the labels are already tagged and the dataset is prepared in a compatible manner to the DNN model.

2. Phrase Structure Identification and Classification Training and Testing

The data that is given into the model has been preprocessed, so the next task in a Deep Learning application is building up a model. Recurrent Neural Networks are known to handle sequential data and has been widely used and recommended in literature for text classification tasks. Recurrent Neural Networks are nothing but a chunk of neural network into which input data is given and it outputs a hidden layer. This single chunk of neural network represented as a loop forms a Recurrent Neural Network as the name suggests, it is recurrent in nature. But RNN's have a drawback of vanishing and exploding gradient problem. The neural networks work based on a Backpropagation algorithm. This is done mainly to backtrack the network and rectify the error. At the time of backtracking the network, some gradient values get vanished or some of them increase exponentially which would lead to fault in prediction of the output. In theory RNN's can handle long term dependencies but in practice it is found that RNN's lack that ability. RNN's can predict the previous or the next information but it cannot accurately predict the word or information that are at a larger distance from the current input word, this is the long-term dependency problem. To overcome these kinds of issues persisting in RNN's a variant of RNN named Long Short-Term Memory Units (LSTM) are used.

C. BIDIRECTIONAL LSTM'S

Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems. In problems where all timesteps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. The first on the input sequence as is and the second on a reversed copy of the input sequence. This can provide additional context to the network and result in faster and even fuller learning on the problem. LSTM in its core, preserves information from inputs that has already passed through it using the hidden state. Unidirectional LSTM only preserves information of the past because the only inputs it has seen are from the past. Using bidirectional will run the inputs in two ways, one from past to future and one from future to past and what differs this approach from unidirectional is that, in the LSTM that runs backwards we can preserve information from the future and using the two hidden states combined we are able in any point in time to preserve information from **both past and future** .BiLSTMs show very good results as they can understand context better.

D. EXPERIMENTAL ANALYSIS

The Phrase Structure Classification model is defined then the model is compiled and fit the model with all hyperparameters. The first layer is an Embedding layer to which the input is mapped. A functional Keras API is used to build the model which uses Bidirectional LSTM with 100 memory units and recurrent dropout of 0.1. The activation function used is softmax since it is suitable for binary classification problems. The optimizer used is adam optimizer which is widely used with binary cross entropy as loss function. The model is compiled using the above hyperparameters and it is fitted with batch size of 32 and 100 epochs. Once the model is trained, the model is undergone testing by giving as input test data which are also phrases and the model would classify the phrases by predicting them as Noun Phrase and Verb Phrase.

For development of the proposed system following are the requirements:

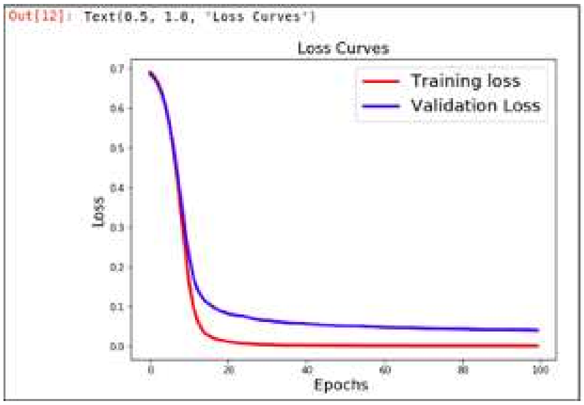
TensorFlow is an open-source software library used to develop and deploy machine learning applications. Keras is an open source neural network library written in Python. It is designed for faster implementation of deep neural networks. It is user friendly, modular and extensible. Jupyter notebook is an interactive development environment to designed to support interactive data science and scientific computing across all programming languages. Pandas is a software library designed for the Python programming language for data manipulation and analysis. NumPy is a Python programming language library that deals with multi-dimensional arrays, matrices and various mathematical functions for operating on these arrays and matrices. The performance required for the proposed system are: Processor – Intel core i3 or higher RAM – 4GB or higher, Speed – 1.80 Ghz or higher, OS – Ubuntu 16 or higher, Mac OS, OS type-64-bit, Programming language – Python 3.6, Disk space – 491.2 GB or higher.

E. Results

This section discusses the experimental results of the proposed system. First of all, the proposed system uses original Malayalam documents given by CDAC, Trivandrum for results assessment. The proposed system is composed of two modules. The first module named Data Preprocessing module deals with converting Malayalam sentences into phrases and labelling them with IOB tags. For word embedding skip gram and cbow models were also implemented. For the proposed system the embedding is done by just encoding the input phrases into some integer vectors. The second module which is the training and testing phase of the proposed model named Phrase structure identification and classification training and testing deals with the development of the model which provided an accuracy of 100% for 210 sentence phrases with validation accuracy of 100% with training loss of 1.1286e-04 and validation loss of 0.0395.

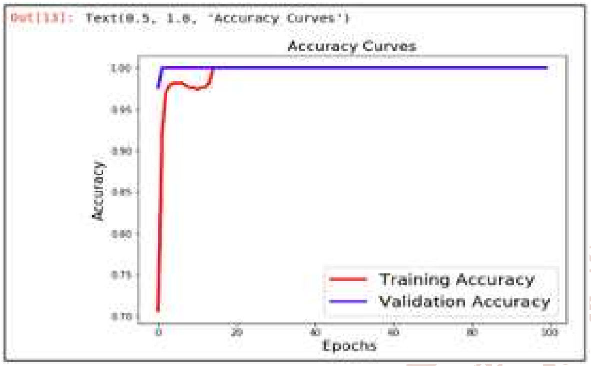
F. Analysis

The proposed system has been modeled as a Sequence Classification Problem using a sequence labelling strategy with RNN-Bidirectional LSTM. The following graphs show the loss and accuracy curves.



```
Out[12]: Text(0.5, 1.0, 'Loss Curves')
```

Figure D.b.1: Loss Curve



```
Out[13]: Text(0.5, 1.0, 'Accuracy Curves')
```

Figure D.b.2: Accuracy Curve

G. CONCLUSION & FUTURE WORKS

Deep learning has been widely used in many NLP tasks as it needs only little engineering by hand. The proposed system is basically a sequence classification problem under NLP which has been solved by applying supervised learning strategy. RNN with LSTM has been used to train the model. The proposed problem has been done only for 200 sentences which is just a miniature version of the problem and it can be further improved with the help of more data. The Sequence IOB labelling technique with Bidirectional lstms have given a considerable accuracy with small loss percentage. The proposed system can be further improved and used in the development of many NLP tools. This strategy can be applied for other languages also. But still there are various areas in which deep learning is still in its childhood stage as in case of processing with unlabeled data. But with the developing researches it is expected that deep learning would become more enhanced and can be applied in different areas.

H. ACKNOWLEDGMENT

The proposed work is an R&D project of CDAC-TVM. The work was done under supervision of Knowledge Resource Centre of CDAC-TVM. A sincere gratitude to CDAC-TVM for providing an opportunity to work on their project, giving permission to access their resources and guiding for the successful completion of this work.

I. REFERENCES

[1] Latha R Nair and David peter S "Language Parsing and Syntax of Malayalam Language" 2nd International Symposium on Computer, Communication, Control and Automation (3CA 2013)

[2] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[3] Lajanugen Logeswaran, Honglak Lee, Dragomir Radev "Sentence Ordering and Coherence Modeling using Recurrent Neural Networks"

[4] Pengfei Liu Xipeng Qiu and Xuanjing Huang "Recurrent Neural Network for Text Classification with Multi-Task Learning "Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence