"Information Extraction using Convolutional Neural Networks"

Mini Project Synopsis Report

Submitted by:

Mihir (201600045)

Chandni Agarwal (201600052)

Sweta Agarwal (201600091)

Under the supervision of:

Dr. Udit Kumar Chakraborty

Associate Professor

Department of Computer Science and Engineering

Sikkim Manipal Institute of Technology

SMU SIKKIM MANIPAL UNIVERSITY

Established under Govt.of Sikkim, Act 9 of 1995, recognised under 2(f) of the UGC Act, 1956

April 2019

ABSTRACT

Information Extraction has been one of the important task in Natural Language Processing (NLP). The said task has been accomplished by using sophisticated machine learning algorithms Support Vector Machine. A large corpus of words is taken and fed into a neural network that will produce vector space uniquely and meaningfully representing each word in the transformed space. This project aims to solve the problem of Information Extraction using Convolutional Neural Networks (CNN) on pre trained word vectors that will be obtained from the previously said neural network. It is expected that acceptable results will be obtained using CNN and that will be much efficient in terms of both accuracy as well as computational cost.

CONTENT

INTRODUCTION

Convolutional Neural Networks is a burgeoning topic in the computer science research field of Deep Learning. Due to the rapid growth of data, Information Extraction has become one of the key technique for handling and organizing text data. The main idea of using CNN for Information Extraction is to reduce the computational complexities that is offered by traditional approaches. Information Extraction have found its uses in language translation, sentiment analysis, spam filtering and other NLP tasks.

Data in large unstructured corpus is to be transformed to high quality word vectors with decent dimensionalities preserving the linear regularities of the data. For this architectures like CBOW and Skip-gram can be employed using the model like word2vec. From this network resultant vectors are retrieved having retained the semantic relations between the words.

The obtained vectors will then be fed into a convolutional neural network. The Convolutional Neural Network consists of different layers like convolutional layer, pooling layer, dropout layer and fully connected layer. These Layers helps in optimizing the results by the highlighting various features that will be used for the extraction of information using kernels/filters. Additionally, for NLP we have some more ConvNets models such as CNN-rand, CNN-static, CNN-non-static, CNN-multichannel. The CNN layers extracts meaningful sub-structures that are useful for the task.

LITERATURE SURVEY

1."Convolutional neural networks for sentence classification." by Kim, Yoon.

This paper describes how a convolutional neural network is used to train pre-trained word vectors for sentence-level classification. The pre-trained word vectors were trained using continuous bag of words architecture. By little tuning of the hyperparameters, the CNN model with only one layer of convolution performed remarkably well. This model uses multiple filters to extract multiple features which increases the number of parameters. By this model it is assumed to retrieve important features from the pre-trained word vectors.

2. "Efficient estimation of word representations in vector space." by Mikolov, Tomas, et al.

This paper proposes two novel architectures that computes high quality vector representation of words from large unstructured datasets with preserved semantic relationship amongst them. The architectures introduced by this model have resulted in saving contextual similarity and thus the analogies and linear regularities between the words. However, the models thus generated does not save the statistical information and global word relationships as this model is based on the concept of local context window. By this model it is assumed to retrieve distributed vector representation of words from the large corpus.

PROBLEM DEFINITION

The efficiency of the Traditional Algorithm was measured based on the output generated from data and was not robust due to limited availability of data. But in the present era, where data is available in abundance, the computational power is limited. An architecture is required having least possible computational cost and high performance.

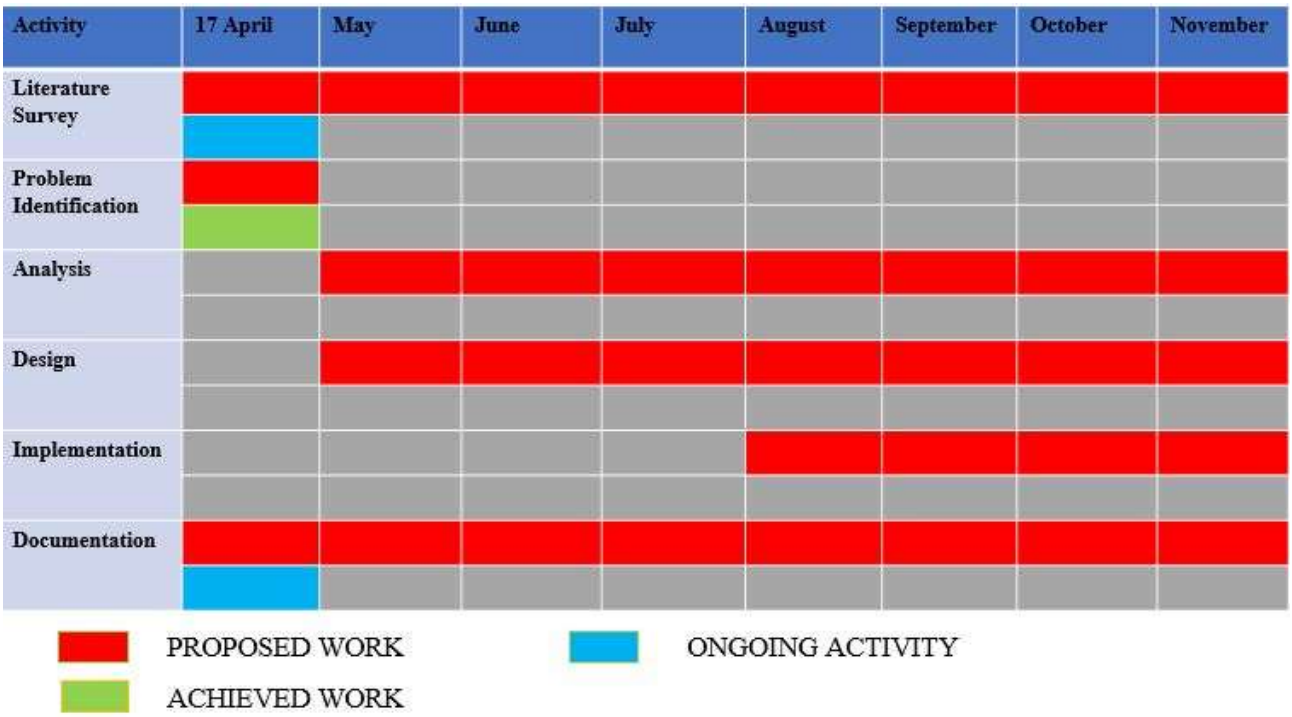
With increase in corpus size, a model is required that could embed its words saving the semantic relationships amongst them and thus can be used to extract information from the pre-trained word vectors. And all this is to be done with the aim of high efficiency both in terms of accuracy as well as computational complexity.

This project is to develop a Model for Information Extraction using Convolutional Neural Networks.

SOLUTION STRATEGY

The dataset is to be converted into a distributed vector representation of words using embedding techniques like word2vec, GloVe embedding etc. so that the resultant vectors preserve the semantic relationships between the words. The word vectors thus obtained are fed into a convolutional Neural Network. The CNN will consist of Convolutional layer, Pooling Layer, Dropout Layer, Fully Connected Layer and different activation functions. The hyperparameters of the CNN will be tuned to train the pre-trained word vectors aiming to maximize the efficiency of the model. The Desired Information will be extracted from the trained CNN and is expected to give acceptable results.

GANTT CHART


Activity
17 April
May
June
July
August
September
October
November
Literature Survey
Problem Identification
Analysis
Design
Implementation
Documentation
PROPOSED WORK
ONGOING ACTIVITY
ACHIEVED WORK

REFERENCES

[1] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

[2] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[3] https://www.davidsbatista.net/blog/2018/03/31/SentenceClassificationConvNets/