TEXT SUMMARIZATION FOR TAMIL ONLINE SPORTS NEWS USING NLP

Thevatheepan Priyadharshan

168290F

Degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa
Sri Lanka

April 2019

TEXT SUMMARIZATION FOR TAMIL ONLINE SPORTS NEWS USING NLP

Thevatheepan Priyadharshan

168290F

Thesis submitted in partial fulfilment of the requirement for the degree Master of Science

Department of Computational Mathematics

University of Moratuwa
Sri Lanka

April 2019

Declaration

"I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any university or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or part in print, electronic or other medium. I retain the right to use this content in whole or part in future works."

Signature: Date:

The above candidate has carried out research for the Masters dissertation under my supervision.

Signature of the supervisor: Date:

Abstract

Text summarization plays an important role in natural language understanding and information retrieval. Presently automatic text summarization getting much more attention by people because it is efficiently and effectively serving time in decision making process even in day to day life. Many approaches such as statistical based, machine learning based approaches have been presented by researchers where statistical based approaches are less semantic consideration in terms of forming summary and most machine learning approaches are language independent. Presently neural network models get more attention than the traditional approaches. There are few statistical based approaches that are presented for Tamil text summarization with less natural language processing. The primary objective of this research work is to propose a methodology to address the problem of summarization for Tamil sports news which can automatically create extractive summary for the news data with the use of Natural Language Processing (NLP) and a generic stochastic artificial neural network.

The sports news gathered from different resources has been given as input to the system where most relevant sentences will be extracted from the text and presented as an extractive summary to the input text. The input will go through six sub process such as pre-processing, feature extraction, feature vector matrix, feature enhancement, sentence extraction and summary generation. Where in the pre-processing the sentences will be initially tokenized. After this set of stop words will be removed from the tokenized output, finally named entities available within the text will be tagged such as person's name, location name, date, numeral. After pre-processing, feature extraction will be executed where features such as sentence position, sentence position related to paragraph, number of named entities, term frequency and inverse document frequency and number of numerals are employed to generate a score against each sentence available in the text.

By using these scores in feature vector matrix sub process, feature matrix will be generated for the whole text where each feature score values for each sentence available in the text is arranged in the row of the matrix. This feature vector matrix will be given as an input to the Restricted Boltzmann Machine which is embedded in the feature enhancement sub process to generate the enhanced feature matrix. After obtaining the Enhanced feature matrix in the sentence extraction process, row values are summed where it gives the summed enhanced feature values for each sentence, after this high score sentence will be extracted as the most relevant sentence to form the summary where it is considered as a sub set of sentences in the summary. And at last the summary generation process will be executed where most relevant sentence selected in the sentence extraction module will be used for cosine similarity measures with the other existing sentences in the text and another sub set of sentences will be extracted from the text to form the summary, likewise the process will be done. Finally, the sentences will be ordered as in the order in the text and extracted summary of the text will be presented. This summary generation will happen on real time by using different resources.

A comparative evaluation has been done for the text summarization systems’ result. For evaluation purpose, 30 news data set has been used, where each summary regarding to each news data set, has been evaluated by 3 Tamil speaking human assessors. Each news has been distributed among those evaluators and they have to read the news data and they have to select the sentences which will form the summary, likewise the responses for each news data set has been gathered. In the experiment, each and every summary generated by the system has been evaluated against the human generated summary and the average F-measure of this text summarization system is 76.6% which is higher than the existing approaches for the Tamil text summarization approaches.

Keywords—Extractive summarization, Natural Language Processing, Neural network, Restricted Boltzmann Machine, Feature matrix, F-measure.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my research supervisor Dr. Sagara Sumathipala for his valuable guidance extended throughout the research. This research would not have been completed to success without his immense support and guidance. Further, I would like to thank Prof. Asoka Karunananda who gave valuable guidance to the documentation of the work during the lectures. And also, I would like to thank all academic staff of the Department of Computational Mathematics who gave sufficient knowledge to complete this research.

Last but not least special thanks should be given to my family members and all of my batchmates for extending their supportive hands of friendship towards the successful drive of the research.

Table of Contents

LIST OF FIGURES

page

LIST OF ABBREVIATION

Abbreviation	Description
NLP	Natural Language Processing
RBM	Restricted Boltzmann Machine
PC	Personal Computer
Tf-IDF	Term frequency and Invers Document Frequency
DUC	Document Understanding Conferences
ROUGE	Recall Oriented Understudy for Gisting Evaluation
DNN	Deep Neural Networks
LNS	Language Neutral Syntax
SOP	Subject Object Predicate
SVM	Support Vector Machine
FFNN	Feed Forward Neural Network
MR	Mathematical Regression
PNN	Probabilistic Neural Network
DE	Differential Evolution
MRF	Markov Random Field
BM	Boltzmann Machine

LIST OF APPENDICES