NAMED ENTITY BOUNDARY DETECTION FOR SINHALA

Yapa Hetti Pathirannahalage Prasan Priyadarshana

208073P

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree

Master of Science (Major Component Research)

Department of Information Technology

Faculty of Information Technology

University of Moratuwa

Sri Lanka

March 2022

Declaration

I declare that this is my own research thesis, and this thesis does not incorporate without acknowledgement any material previously published submitted for a Degree or Diploma in any other 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 the acknowledgement is made in the text.

Signature: *UOM Verified Signature* Date: 18/03/2022

I have read the final thesis and it is in accordance with the approved university final thesis outline.

Signature of the supervisor: UOM Verified Signature Date: 18/03/2022

Acknowledgement

I would take this opportunity to express my sincere gratitude to everyone who guided me directly and indirectly to achieve this milestone. First and foremost, I am extremely grateful to my supervisor, Dr. Lochandaka Ranathunga for the continuous support, invaluable advice, kind guidance and patience during this whole journey. His vast experience and extensive knowledge have encouraged me throughout the time of my academic research and life. I also thank the research assistants who have been employed under the Social Media Analytics Research Group and the Accelerating Higher Education Expansion and Development (AHEAD) project, University of Moratuwa who helped me in annotating the dataset. I would also like to thank Dr. Surangika Ranathunga and all the research assistants who have been attached to the National Languages Processing Centre (NLPC), University of Moratuwa for providing me various corpus related to named entities and shaping the overall idea to identify the novelty. Finally, I would like to express my gratitude to my beloved parents, my wife and my friends for their motivation, inspiration, and continuous support throughout the difficult times.

Abstract

Named entity recognition (NER) can be introduced as a sequential categorizing task which contains a potential gravity in novel research arena. NER can be mentioned as the foundation for accomplishing most common natural language processing (NLP) tasks such as information extraction, information retrieval, semantic role labelling etc. Even though plenty of attempts have been employed on NE type detection, still there are plenty of avenues to be discovered under the NE boundary detection. Analyzing Sinhala related contents which have been published in social media can also be considered as one of the rising factors due to several human involvements in the recent past. The ultimate goal which is to obtain a constructive deep neural framework for determining named entity boundary detection has been achieved in a comprehensive manner and the model has been tested using Sinhala related statements which have been extracted through social media. Several objectives have been determined to accomplish this task considering the existing baselines. Several novelties have been identified to show off the uniqueness of this approach. Specifically, the novel concept *"Boundary Bubbles"* has been used to identify the specific entity mentions considering each head word for the identified named entities. Various experiments have been conducted based on multiple evaluation criteria and the named entity boundary detection model performs well with an average of 71% in Precision, 67% in Recall and 63% in F1 over the existing benchmarks. Hence this novel framework can be accepted as a vital solution for determining named entity boundary detection under forecasting various computational activities in social media.

Keywords:

Deep neural network, named entities, named entity boundary, named entity recognition, named entity type

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
NER	Named Entity Recognition
NE	Named Entities
NLP	Natural Language Processing
ORG, PER, LOC	Identified Some of the Named Entities
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
Bi-LSTM	Bidirectional Long Short-Term Memory
GPU	Graphics Processing Unit
GRU	Gated Recurrent Units
Bi-GRU	Bidirectional Gated Recurrent Units
BERT	Bidirectional Encoder Representations from Transformers
MRC	Machine Reading Comprehensive
BOW	Bag of Words
RNN	Recurrent Neural Network
MLP	Multi-Layer Perceptron
CRF	Conditional Random Fields
NED	Named Entity Disambiguation
IE	Information Extraction
IR	Information Retrieval