MONOLINGUAL SENTENCE SIMILARITY MEASUREMENT USING SIAMESE NEURAL NETWORKS FOR SINHALA AND TAMIL LANGUAGES

Nilaxan Satkunanantham

179337B

M.Sc. in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

MONOLINGUAL SENTENCE SIMILARITY MEASUREMENT USING SIAMESE NEURAL NETWORKS FOR SINHALA AND TAMIL LANGUAGES

Nilaxan Satkunanantham

179337B

This dissertation submitted in partial fulfillment of the requirements for the Degree of MSc in Computer Science Specializing in Data Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

DECLARATION

I declare that this is my own work, and this dissertation does not incorporate without acknowledgment any material previously 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 acknowledgment is made in the text.

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

UOM Verified Signature

Signature: .... Date: 2021-05-31 ....

Name: Nilaxan Satkunanantham

The supervisor/s should certify the thesis/dissertation with the following declaration.

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for the MSc PG Diploma Project.

Signature of the supervisor: .... Date: ....

Name: Dr. Surangika Ranathunga

ABSTRACT

Sentence similarity plays a key role in text-processing related research such as plagiarism checking and paraphrasing. So far, only conventional unsupervised sentence similarity techniques such as string-based, corpus-based, knowledge-based, and hybrid approaches have been used to measure sentence similarity for Tamil and Sinhala languages. In this research, we introduce a Deep Learning methodology to measure sentence similarity for these two languages, which makes use of Siamese Recurrent Neural Networks techniques together with a word-embedding model as the input representation. This approach achieved a 3.07% higher Pearson correlation coefficient for the Tamil dataset of 2500 sentence pairs and a 3.61% higher Pearson correlation coefficient for the Sinhala dataset of 5000 sentence pairs. Both these results outperform that of the conventional unsupervised sentence similarity techniques applied on the same datasets.

Keywords - Sentence-similarity, Sinhala, Tamil, Siamese neural network, LSTM, deep-learning, fastText, natural language processing

ACKNOWLEDGEMENT

I would like to express profound gratitude to my advisor, Dr. Surangika Ranathunga, for her invaluable support by providing relevant knowledge, materials, advice, supervision, and useful suggestions throughout this research work. Her expertise and continuous guidance enabled me to complete my work successfully.

I am grateful for the support and advice given by Dr. Charith Chitraranjan, by encouraging the continuation of this research. Further, I would like to thank all my colleagues for their help in finding relevant research material, sharing knowledge and experience, and for their encouragement.

I am as ever, especially indebted to my parents for their love and support throughout my life. I also wish to thank my loving wife, who supported me throughout my work. Also, I wish to express my gratitude to all my colleagues at Sysco LABS, for the support given to me to manage my MSc research work.

Finally, I would also like to expand my deepest gratitude to all those who have directly and indirectly guided us in completing this research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	**Description**
RNN	Recurrent Neural Networks
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
BOW	Bag of Words
CBOW	Continuous Bag-of-Words
POS	Part of Speech
IR	Information Retrieval
Q&A	Question and Answer
VSM	Vector Space Model
MaLSTM	Manhattan LSTM
LCS	Longest Common SubString
STS	Semantic Text Similarity
SVD	Singular Value Decomposition
HAL	Hyperspace Analogue to Language
GLSA	Generalized Latent Semantic Analysis
ESA	Explicit Semantic Analysis
CL-ESA	Cross-Language Explicit Semantic Analysis
PMI-IR	Pointwise Mutual Information - Information Retrieval
SCO-PMI	Second-order Co-Occurrence Pointwise Mutual Information
NGD	Normalized Google Distance
DISCO	DIStributionally similar words using CO-occurrences
Bi-LSTM	Bidirectional LSTM
GRU	Gated Recurring Units
Bi-GRU	Bidirectional GRU
STS	Semantic Text Similarity
RDF	Resource Description Framework
ERCNN	Enhanced Recurrent Convolutional Neural Networks
CARNN	Context Aligned RNN
SA-BiLSTM	Self-Attention based BiLSTM
NNLM	Feedforward Neural Net Language Model
CoVe	Contextual Word Vectors
BERT	Bidirectional Encoder Representations from Transformers
ELMo	Embeddings from Language Model
ULMFiT	Universal Language Model Fine-tuning for Text Classification
CVT	Cross-View Training
t-SNE	t-Distributed Stochastic Neighbouring Embedding