

**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

<b>DECLARATION</b>	i
<b>ABSTRACT</b>	ii
<b>ACKNOWLEDGEMENT</b>	iii
<b>TABLE OF CONTENTS</b>	iv
<b>LIST OF FIGURES</b>	vi
<b>LIST OF TABLES</b>	vii
<b>LIST OF ABBREVIATIONS</b>	viii
<b>INTRODUCTION</b>	1
1.1 Problem and Motivation	1
1.2 Problem Statement	2
1.3 Overall Objective	2
1.4 Research Contribution	3
1.5 Report Structure	3
<b>LITERATURE REVIEW</b>	4
2.1 String based similarity	4
2.1.1 Character based similarity measures	5
2.1.2 Term-based Similarity Measures	7
2.2 Corpus based similarity	11
2.3 Knowledge Based Similarity	13
2.3.1 WordNet	14
2.3.2.1 Tamil WordNet and Corpus	14
2.3.2.2 Sinhala WordNet and Corpus	14
2.3.2 Semantic Similarity Measures	15
2.3.3 Semantic Relatedness Measures	16
2.5 Hybrid Similarity Measurement Techniques	16
2.4 Deep Learning Based Techniques	19
2.4.1 Convolutional Neural Networks Based Approach	20
2.4.2 Recurrent Neural Networks Based Approach	22
2.4.3. Long Short-Term (LSTM) Memory Based Approach	22
2.4.3. Gated Recurrent Units Based Approach	24
2.6 Sentence Similarity Techniques used for Tamil and Sinhala	26
2.7 Vector Representation of Words	26
2.7.1 Word Co-occurrence Matrix	26
2.7.2 Word embeddings	27
2.7.2.1 Shallow Word Embeddings	28
2.7.2.2 Contextualized Word-Embeddings	28

<b>RESEARCH METHODOLOGY</b>	31
3.1 Dataset	31
3.1.1 Dataset for Sinhala	31
3.1.2 Dataset for Tamil	33
3.2 Visualization of embedding layer	35
3.3 Preprocessing	38
3.3.1 Sentence cleanup	38
3.3.2 Stop words removal	39
3.3.3 Spelling correction	41
3.3.3.1 Traditional spell correction	41
3.3.3.2 Contextual spell correction	42
3.3.4 Tokenization	43
3.4 Architecture and Implementation	43
3.4.1 Experiment with Siamese networks	44
3.4.2 Experiment with Cosine similarity	45
3.5 Embedding Layer	45
3.6 Hidden Layer	46
3.7 Sentence similarity measurement layer	47
<b>SYSTEM EVALUATION AND RESULTS</b>	48
4.1 Baseline experiment	48
4.2 Effects of preprocessing	48
4.3.1 Effects of cleanup	49
4.3.2 Effects of stop word removal	49
4.3.3 Effects of spell correction	50
4.3 Model evaluation	52
4.4 Results summary	55
4.5 Error Analysis	59
<b>CONCLUSION</b>	62
5.1 Future works	63
<b>REFERENCES</b>	64
<b>APPENDIX-I</b>	70
<b>APPENDIX-II</b>	71

## LIST OF FIGURES

Figure 2.1:	String based similarity measurement techniques [5]	5
Figure 2.2:	Manhattan distance for similarity measures	8
Figure 2.3:	Hamming distance for similarity measures	8
Figure 2.4:	Euclidean distance for similarity measures	9
Figure 2.5:	Cosine similarity for similarity measures	9
Figure 2.6:	Jaccard similarity for similarity measures	10
Figure 2.7:	Corpus based similarity measurement techniques [5]	11
Figure 2.8:	Knowledge based similarity measurement techniques [5]	15
Figure 2.9:	Deep Learning based similarity measurement techniques	20
Figure 2.10:	A typical Siamese neural network architecture [32]	25
Figure 2.11:	CBoW, Skip-gram models architectures [64]	27
Figure 3.1:	Questionnaire for Tamil short sentence dataset collection	33
Figure 3.2:	Manual similarity score calculation method for Tamil dataset	34
Figure 3.3:	Visualization of Tamil dataset using t-SNE	36
Figure 3.4:	Visualization of Sinhala dataset using t-SNE	37
Figure 3.5:	Preprocessing steps for short sentence dataset	38
Figure 3.6:	Simplified architecture of MaLSTM model [8]	44
Figure 3.7:	System for short sentence similarity measurement [8]	45
Figure 4.1:	Model training and validation performance for Tamil dataset	53
Figure 4.2:	Model training and validation performance for Sinhala dataset	54
Figure 4.3:	Model evaluation performance for Sinhala	58
Figure 4.4:	Model evaluation performance for Tamil	58



## LIST OF TABLES

Table 2.1:	Sentence similarity for STS-2017 dataset	24
Table 3.1:	Gold Relatedness Score	32
Table 3.2:	Manually Annotated Score for Sinhala	32
Table 3.4:	Tamil - Manual similarity score calculation strategy	34
Table 3.5:	Manually annotated score for Tamil	35
Table 3.6:	List of Sinhala stop-words	39
Table 3.7:	List of Tamil stop-words	40
Table 4.1:	Baseline experiment results	48
Table 4.2:	Sentence cleanup experiment results	49
Table 4.3:	Stop words removal experiment results	50
Table 4.4:	Spelling errors or misspelling words in Tamil dataset	51
Table 4.5:	Spelling errors or misspelling words in Sinhala dataset	51
Table 4.6:	Sentence spelling correction experiment results	52
Table 4.7:	MaLSTM model parameters	53
Table 4.8:	Tamil - Short similarity score comparison	54
Table 4.9:	Sinhala - Short sentence similarity score comparison	56
Table 4.10:	Performance comparison using Pearson correlation coefficient	57
Table 4.11:	Loanwords in Tamil dataset	59
Table 4.12:	Loanwords in Sinhala dataset	59

## 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