



Rule-Based Translation of Sinhala Slang and Colloquial Expressions into English for Enhanced Cross-Cultural Communication

I. S. Gallage

(Reg. No.:MS23017474)

A THESIS

SUBMITTED TO

SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2024

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Dr. Dilshan De Silva

Approved for MSc. Research Project:


MSc. Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign:

Gallage Imasha Sandali

Date:13/12/2024.....

ABSTRACT

Rule-Based Translation of Sinhala Slang and Colloquial Expressions into English for Enhanced Cross-Cultural Communication

Gallage Imasha Sandali

MSc. in Information Technology

Supervisor: Dr. Dilshan De Silva

December 2024

Begin This research paper presents the development and evaluation of a hybrid translation system designed to translate Sinhala slang and colloquial expressions into English. Recognizing the challenges of accurately conveying informal language in cross-cultural communication, the study addresses the need for precise and contextually relevant translations. Sinhala slang and colloquialisms often contain nuanced meanings, cultural references, and idiomatic expressions that are challenging to capture with conventional machine translation approaches. To address these complexities, the proposed system combines linguistic rules, pattern recognition algorithms, and context-based translation models, specifically tailored for Sinhala slang and colloquial styles. A novel component of this research is the combination of rule-based matching (for known slang) with unsupervised learning using Word2Vec embeddings and K-Means clustering for new slang detection. This hybrid approach enhances the system's ability to differentiate formal from informal language by leveraging predefined rules for familiar slang while dynamically identifying emerging slang patterns through clustering. This integration enables more targeted translation processing by identifying slang, which is subsequently handled by rule-based frameworks developed through data collection and analysis of Sinhala slang expressions. The system's effectiveness is rigorously evaluated using metrics such as accuracy, precision, recall, F1-score, and Bilingual Evaluation Understudy score, confirming its utility in promoting cross-cultural understanding through culturally sensitive translations. The outcomes of this research advance rule-based and unsupervised learning-supported machine translation techniques for informal language, fostering communication across diverse linguistic and cultural contexts.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my supervisor, Dr. Dilshan De Silva, for his invaluable guidance, steadfast support, and encouragement throughout this research journey. I am deeply thankful for his patience and for always being available to discuss ideas, provide direction, and offer insightful feedback. His mentorship has not only shaped my academic progress but has also inspired me to persevere through challenges, motivating me to continuously improve and refine my work.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS	v
List of Figures	vi
List of Tables	vii
Chapter 1 Introduction	1
1.1 Background	1
1.2 Research Problem	2
1.3 Objectives	3
Chapter 2 Literature Review	5
2.1 Review of Existing Literature	5
2.2 Gap Identification	15
2.3 Conceptual Framework	17
Chapter 3 Methodology	24
3.1 Research Design	24
3.2 Sampling Strategy	26
3.2.1 Dataset Characteristics	26
3.2.2 Data Preprocessing	26
3.3 System Development	29
3.3.1 Rule Implementation	29
3.3.2 Integration and Testing	33
3.3.3 Evaluation of the System	34
3.3.4 System Requirements	38
Chapter 4 Development of the Web Application	40
4.1 Overview of Web Application Development	40
4.2 Design and Architecture	40
4.2.1 Error Correction	44
4.2.2 Security Considerations	45
4.2.3 System Evaluation	46
Chapter 5 Results	50
Chapter 6 Conclusion	60
Chapter 7 Discussion	63
References	66

List of Figures

Figure 3-1 System Flow	32
Figure 3-2 Slang detection performance	36
Figure 3-3 Translation quality	36
Figure 5-1 Home screen of the application	50
Figure 5-2 Slang detection example 1	51
Figure 5-3 Slang detection example 2	51
Figure 5-4 Translation example 1	52
Figure 5-5 Translation example 2	52
Figure 5-6 Translation results from the proposed system	53
Figure 5-7 Translation results from Google Translate	53
Figure 5-8 Slang dictionary	54
Figure 5-9 performance in detecting Sinhala slang	55
Figure 5-10 BLEU score value	55

List of Tables

Table 6-I Future enhancements to the system	62
---	----