SINHALA CODE-MIXED TEXT TRANSLATION USING NEURAL MACHINE TRANSLATION

Kugathasan Archchana

188069F

Masters of Philosophy

Department of Computational Mathematics

University of Moratuwa
Sri Lanka

April 2024

SINHALA CODE-MIXED TEXT TRANSLATION USING NEURAL MACHINE TRANSLATION

Kugathasan Archchana

188069F

Thesis submitted in partial fulfilment of the requirements for the
Masters of Philosophy

Department of Computational Mathematics

University of Moratuwa
Sri Lanka

April 2024

Declaration of the candidate and the supervisor

"I declare that this is my own work and this thesis does not incorporate without acknowledgement 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 acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/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)".

Signature: Date:

02.04.2024

The above candidate has carried out research for the Masters/MPhil/PhD thesis/ Dissertation under my supervision.

07/04/2024

Signature of the supervisor: Date:

Acknowledgement

This MPhil is attributed to the immense support received from many staff members of University Of Moratuwa. I would like to convey my sincere thanks to my supervisors Dr.Sagara Sumathipala(Primary Supervisor) and Dr.Thushari Silva(Associate supervisor). I am very grateful for my supervisors for their continues support provide throughout my research studies.

I would like to convey my gratitude for Dr.Sagara Sumathipala for accepting me as his student for my MPhil. He guided me with my technical skills, writing skills, presenting skills etc. Also he provided me several opportunities such as attending NLP based spring school, being an exchange student in Japan etc. He actively guided me in the planning the research, identifying the gaps, designing the methodology tasks and closely monitored my work. Dr.Sagara conducted weekly progress meetings and he provided an in-depth feedback each and every time when I show my research progress. He always brought suitable connections at the right time to guide me in a correct pathway. He was a great mentor in my MPhil journey. Also I am fortunate to get Dr.Thushari Silva as my associate supervisor as she always provided me valuable feedbacks in all of my progress presentations.

Also my heartful thanks dedicated to my loving husband Sindhujan, mother, father, sister, grandma and grandpa. They provided me there fullest support through out this MPhil journey. I wouldn’t have been completed my MPhil studies successfully without their sacrifice, devotion and prayers. Special thanks to my husband who always encouraged me and motivating me to move forward when there are obstacles.

Also I would like to thank Ms.Kanishka Silva for the tremendous guidance she provided me on completing the modules and documentations. Finally I would like to thank all my colleagues at University Of Moratuwa.

Abstract

Mixing two or more languages together in communication is called as code-mixing. In South Asian communities it has become famous due to bilingualism or multilingualism. Sinhala-English code-mixed(SECM) text is the most popular language used in Sri Lanka in casual talks such as social media comments, posts, chats, etc. On social media platforms, the contents such as posts and comments are used for personalized advertisement recommendations, post recommendations, interesting content recommendations, etc., to provide better customer service according to their interest. Due to the code-mixing nature of the language, most of the Srilankan social media content is unused for recommendation purposes. So our research study mainly focuses on translating the SECM text to the Sinhala language. Once the contents are converted to a standard language, the social media contents can be processed easily and used for the necessary purposes. In this research, we initially conduct an in-depth analysis of Sinhala-English code-mixed. Issues that are considered as barriers to translate the SECM to Sinhala are identified. Also, we conducted a thorough literature study of code-mixed text analysis. An SECM-Sinhala parallel corpus with 5000 parallel sentences are used for this research study. The approach proposed for the SECM to Sinhala translation consists of a normalization layer, Encoder-Decoder framework(Seq2Seq), LSTM and Teacher Forcing mechanism. We evaluated our proposed approach with other translation approaches proposed for code-mixed text translation, and our approach gave a significantly higher BLEU score.

Key words

Code-mixing, Bilingualism, Multilingualism, LSTM, Teacher Forcing

Table of Contents

List of figures

List of tables

List of abbreviations

Abbreviation	Description
SECM	Sinhala-English Code Mixed
LM	Language Model
DICT	Dictionary
LR	Logistic Regression
CRF	Conditional Random Field
SVM	Support Vector Machine
RP	Root phone
POS	Part Of Speech
MWE	Minimum Word Error
ITRANS	Indian Languages Transliteration
LD	Levenshtein Distance
XSCM	eXtended Source Channel Model
HMM	Hidden Markov Model
Seq2Seq	Sequence to Sequence
NMT	Neural Machine Translation
FC	Fully Correct
CR	Change Required
OOV	Out Of Vocabulary
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
BLEU	Bi-Lingual Evaluation Understudy
TER	Translation Edit Rate
GTM	General Text Matcher
METEOR	Metric for Evaluation of Translation with Explicit Ordering
WER	Word Error Rate

List of appendices