

**SINHALA CODE-MIXED TEXT TRANSLATION USING
NEURAL MACHINE TRANSLATION**

Kugathan 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**

Kugathan 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.

Signature of the supervisor:

07/04/2024

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 heartfelt 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

Declaration of the candidate and the supervisor	i
Acknowledgement.....	ii
Abstract	iii
List of figures	vi
List of tables.....	ix
List of abbreviations.....	x
List of appendices	xi
1. Introduction	1
1.1 Background and motivation	1
1.2 Research Problem.....	4
1.2.1 Challenging structure of SECM text.....	4
1.2.2 Lack of parallel data for the SECM text translation	6
1.2.3 Lack of successful methods for the code-mixed text translation	7
1.2.4 Unavailability of tools for the Sinhala-English code-mixed text translation.....	8
1.3 Research Aim and Objectives	8
1.4 Contribution of the Thesis	9
1.5 Significance	11
1.6 Thesis Outline.....	11
2. Literature Review.....	12
2.1 Code-mixing.....	12
2.1.1 Word-Level language identification	12
2.1.2 Identification of code-switched text.....	18
2.1.3 Normalization of code-mixed text	24
2.2 Machine Translation.....	31
3. Code-mixed text analysis	33
3.1 Background Study	33
3.2 Challenges in Sinhala-English code-mixed text.....	36
3.3 Transliteration	36
3.4 Mapping of Sinhala letters to Singlish	37
3.5 Summary	40
4. Methodology	41
4.1 Data collection and validation	41
4.2 Text normalization	46

4.2.1	Spelling Error detection and correction	46
4.2.2	Slang word normalization	47
4.2.3	Transliteration normalization	48
4.3	Sequence to Sequence model(Encoder-Decoder framework).....	49
4.4	Long Short Term Memory	52
4.5	Teacher Forcing Algorithm	58
4.6	Experimental Setup & Implementation of the model.....	60
4.7	Implementation of the web application of SECM to Sinhala translator.....	62
4.8	Summary	64
5.	Performance Evaluation	66
5.1	Prediction from the model	66
5.2	Performance evaluation metrics & Algorithm	68
5.2.1	BLEU (BiLingual Evaluation Understudy)	69
5.3	BLEU score calculation.....	71
5.4	Evaluation of the proposed model with SECM dataset.....	73
5.4.1	Experimenting with baseline Seq2Seq model.....	74
5.4.2	Experimenting with baseline Seq2Seq Attention model.....	77
5.4.3	Experimenting with the proposed model	81
5.5	Evaluation of the proposed model with Hindi-English code-mixed dataset	82
5.6	Summary	83
6.	Result & Discussion.....	84
6.1	Summary	95
7.	Conclusion and Future Work	97
7.1	Summary of achievements	97
7.2	Limitations.....	99
7.3	Future Works.....	100
8.	Publications	101
	References	102
	Appendix - A : Detailed view of model result comparison	107
	Appendix - B : Survey questionnaire	108
	Appendix C : Predicted result from the model.....	109

List of figures

Figure 1.1 : Example of Sinhala-English code-mixed text sentence	2
Figure 1.2:Data collection platforms of code-mixed text	7
Figure 2.1: Example of Sinhala-English code-mixed text	13
Figure 2.2: An example of a Sinhala-English code-switched sentence	18
Figure 3.1: Results of the survey on Sinhala-English code-mixed text usage among Srilankan	35
Figure 3.2: Comparison between phonetic representation which are considered as the standard, romanized representation and Singlish representation for Sinhala Basic consonants	36
Figure 3.3: Comparison between phonetic representation, which is considered as the standard, romanized representation and Singlish representation for Sinhala Vowels	37
Figure 3.4:Mapping of Sinhala Vowels and proposed Singlish form.....	37
Figure 3.5: Mapping of Sinhala Basic consonants and proposed Singlish form	38
Figure 3.6: Mapping of Sinhala other consonants and proposed Singlish form	39
Figure 4.1:Sample sentences from corpus which is annotated using crowd sourcing approach, A1 -> Annotator 1, A2 -> Annotator2, A3 -> Annotator3; FC -> Fully Correct, CR -> Change Required, N/A - Not Applicable fields show no changes needed.	43
Figure 4.2:Datasheet for Fleiss Kappa analysis	46
Figure 4.3: Sequence to Sequence model	50
Figure 4.4: One timestep in the encoder	51
Figure 4.5: One timestep in the encoder	51
Figure 4.6 : Cell state memory maintenance.....	52

Figure 4.7 : Example of how the cell state is maintained thorough out a sentence. . .	54
Figure 4.8: Tanh squishes values to be between -1 and 1	55
Figure 4.9: Sigmoid squishes values to be between 0 and 1	56
Figure 4.10: Detailed image of one LSTM unit	58
Figure 4.11 : Comparison of Teachers Forcing Vs Non-Teachers Forcing.....	59
Figure 4.12 : Architecture diagram of the proposed the model	60
Figure 4.13: SECM to Sinhala translator web application.....	63
Figure 4.14: BLEU score calculator of SECM to Sinhala translator web application	63
Figure 5.1 Prediction of Sinhala sentences for randomly given SECM sentences	66
Figure 5.2: Calculated values to evaluate the BLEU Score	73
Figure 5.3: Example of some predicted Sinhala translation and bleu score using the Seq2Seq baseline model without normalization. ref and pre column refers to the number of words in the reference sentence and predicted sentence, the rest of the columns shows the count of the n-gram tokens used for the calculation.....	75
Figure 5.4: BLEU score calculation values of Seq2Seq baseline model without normalization.....	75
Figure 5.5: Example of some predicted Sinhala translation and bleu score using the Seq2Seq baseline model with normalization. ref and pre column refers to the number of words in the reference sentence and predicted sentence, the rest of the columns shows the count of the n-gram tokens used for the calculation	76
Figure 5.6:BLEU score calculation values of Seq2Seq baseline model with normalization.....	76
Figure 5.7: Seq2Seq baseline model result	77

Figure 5.8: Example of some predicted Sinhala translation and bleu score using the Seq2Seq + Attention model without normalization. ref and pre column refers to the number of words in the reference sentence and predicted sentence, the rest of the columns shows the count of the n-gram tokens used for the calculation.....	78
Figure 5.9: BLEU score calculation values of Seq2Seq Attention model without normalization.....	79
Figure 5.10: Example of some predicted Sinhala translation and bleu score using the Seq2Seq model without normalization. ref and pre column refers to the number of words in the reference sentence and predicted sentence, the rest of the columns shows the count of the n-gram tokens used for the calculation.	79
Figure 5.11: BLEU score calculation values of Seq2Seq attention model with normalization.....	80
Figure 5.12: Seq2Seq with Attention model result	80
Figure 5.13: Experiment values of Seq2Seq with Teacher Forcing model.....	81
Figure 5.14: BLEU Score comparison for Hindi-English code-mixed translation....	82
Figure 6.1: Training accuracy and loss of experimented models.....	86
Figure 6.2: Testing accuracy and loss of experimented models	87
Figure 6.3: Experimented models & BLEU scores.....	88
Figure 6.4: Example 1 from the web application SECM to Sinhala translator.....	90
Figure 6.5: Example 2 from the web application SECM to Sinhala translator.....	91
Figure 6.6: Example 3 from the web application SECM to Sinhala translator.....	92
Figure 6.7: Example 4 from the web application SECM to Sinhala translator.....	93
Figure 6.8: Example 5 from the web application SECM to Sinhala translator.....	94

List of tables

Table 2.1 : Summary of research studies on word level language identification	15
Table 2.2:Summary of research studies on code-switched texts.....	21
Table 2.3:Summary of research studies based on text normalization	28
Table 5.1 : Example of some predicted Sinhala translation and bleu score. ref and pre column refers to the number of words in the reference sentence and predicted sentence, the rest of the columns shows the count of the n-gram tokens used for the calculation of modified precision	72

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

Description	Page No
Detailed view of model result comparison	107
Survey Questionnaire	108
Predicted result from the model	109