

**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<br>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<br>.....   | 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<br>Seq2Seq baseline model without normalization. ref and pre column refers to the<br>number of words in the reference sentence and predicted sentence, the rest of the<br>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<br>normalization.....   | 75 |
| Figure 5.5: Example of some predicted Sinhala translation and bleu score using the<br>Seq2Seq baseline model with normalization. ref and pre column refers to the number<br>of words in the reference sentence and predicted sentence, the rest of the columns<br>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<br>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<br>column refers to the number of words in the reference sentence and predicted sentence,<br>the rest of the columns shows the count of the n-gram tokens used for the calculation<br>of modified precision ..... | 72 |

## List of abbreviations

| <b>Abbreviation</b> | <b>Description</b>  |
|---------------------|---|
| SECM                | Sinhala-English Code Mixed                                  |
| LM                  | Language Model  |
| DICTIONARY          | 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**

| <b>Description</b>                       | <b>Page No</b> |
|--|----------------|
| Detailed view of model result comparison | 107            |
| Survey Questionnaire                     | 108            |
| Predicted result from the model          | 109            |

# 1. Introduction

## 1.1 Background and motivation

The mixing of linguistic elements such as terms, words, and morphemes of one language with another language is called code-mixing. In code-mixing, lexicon and syntactic formulation from different languages are mixed to generate a single sentence. This can also be described as romanization (Davies & Bentahila, 2007).

In southeast Asian contexts, bilingualism or multilingualism is a widespread trend. The potential to interact in two languages is called bilingualism and the ability to interact in more than two languages is called multilingualism. Code-mixing has been identified as a result of bilingualism and multilingualism. Researchers have different viewpoints on bilingualism. A set of researchers states that bilingualism in code-mixing is considered a skilled performance. In contrast, another group of researchers argues that bilingual people are rarely fluent in their second language. Also, the researchers state that bilingualism survives especially in communities where each language is given equal prestige (Senaratne, 2009).

Several research studies about multi-lingual communities have proven that people choose social media as the medium to express thoughts or ideas in their daily life (Chandu et al., 2017). Most people have used bilingual texts on social media. Therefore, there is a huge demand for research studies based on code-mixed text. The prime focus of this research study is based on Sinhala-English code-mixed text (SECM).

There are several questions upraised when it comes to code-mixed text.

- Does code-mixing have standard patterns?
- Is there a dominant language in the code-mixed text?
- Is code-switching a part of code-mixing?
- Does code-mixing have a grammatical structure?

These questions inspect many hidden facts that need to be clarified through more research based on code-mixed texts. In this research study, we analyse the local context of Sinhala-English code-mixed texts.

According to internet statistics from 2020, 25.9% of people are using the English language in the internet. Usage of the English language is less than one-third of the total usage. Local internet forums and social media in South East Asian countries contain texts with code-mixing. Most of the user-generated texts have been identified in the form of code-mixed text(Chandu et al., 2017).

Millions of user-generated content such as posts, comments, and reviews are daily posted on social media. Using user-generated content for personalized recommendation, sentiment analysis, entity extraction, etc., is considered one of the famous business marketing strategies in the 21<sup>st</sup> century. Due to the code-mixing, a huge amount of user-generated content lasts unprocessed for business purposes. This leads to massive information loss. So these days, there is a considerable demand to convert code-mixed text to a single language text (Dhar et al., 2018). However, due to the lack of a parallel dataset (code-mixed sentence and its target sentence), the number of researches conducted on this topic is minimal. The core target of our research study is to translate the Sinhala code-mixed text into Sinhala.

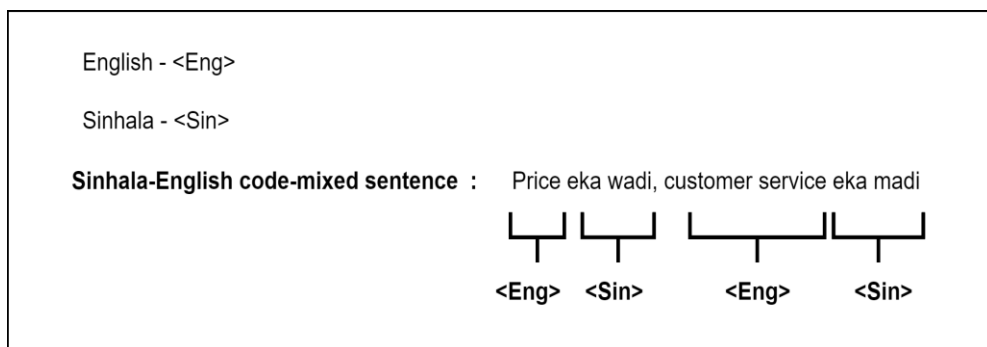


Figure 1.1 : Example of Sinhala-English code-mixed text sentence

In the research study of (Kachru, 1986), he explains the need for English in South Asian countries. Many former Anglo-American colonies have been recognized with varieties of English, which is a deviation from Standard English in the later

development world. According to Kachru's observation, in South Asia, the English language is appraised as a sign of 'achievement', 'modernization', and 'strength'. Also, his study defines code-mixing as a highlight of social, modernization, and economic status and membership in an aristocratic society. The most comprehensive code-mixing range is identified in the English language. More than the effect of colonization, code-mixing rests on several additional factors such as the amount of exposure to English, medium of instruction in school or workplace, topics and contexts of discussion, and rate of occurrence of language used in marked domains.

The code-mixed sentences are structured in the order of the native language of the writer and borrow a few vocabularies from the second language the writer is fluent in. Many countries such as India, Sri Lanka, Bangladesh, China, etc. have been identified with massive use of code-mixed texts on their local social media(Chanda et al., 2016).

Sri Lanka acknowledges Sinhala, English, and Tamil as the formal languages used for official activities. We mainly have two categories of code-mixed languages; Sinhala-English (Singlish) and Tamil-English (Tanglish). But there is no mixing between Sinhala and Tamil languages. Most Sri Lankans are bilingual. They are experts in their mother tongue and also good in English. This leads the way to the usage of code-mixed text.

Figure 1.1 shows an example of Sinhala-English code-mixed text. The words 'Price', 'customer', and 'service' are derived from English vocabulary in the sentence. The rest of the words are from the Sinhala language, written with English letters. This way of representation is called romanization or transliteration. Also, the sentence has been expressed with the Sinhala grammatical structure.

Our research study breaks the barriers to translating Sinhala-English code-mixed text to Sinhala, where we analyse the structure and pattern of code mixed text, identify the challenges in Sinhala-English code-mixed text, build a model to translate code-mixed text to a standard language and finally building a web application for the translation accessing the created model.



## 1.2 Research Problem

### 1.2.1 Challenging structure of SECM text

Singhe et al(2018), have stated that code-mixed texts diverge from the respective language mainly for 2 reasons: informal speech and transliteration. Insufficiency in the methods for formal transliteration leads the way to informal transliteration with different phonetic variations. Multi-lingual speakers often directly convert the native language scripts into Roman scripts(Singh et al., 2018). The structure of the converted text depends on the accent, region, and dialect. For example, the Sinhala word ‘කෑම’ is written in many ways such as ‘kama’, ‘kaama’, ‘kema’ etc., in Sinhala-English code-mixed text. All these words represent the single meaning ‘Food’. This is called informal transliteration. The secondly stated factor for the deviation from common words is colloquial speech. This issue mainly happens due to the usage of words with non-standard spellings. If we consider the word ‘Good’, it is written as ‘gd’, ‘gud’, ‘goood’ etc. Bi-lingual or multilingual speakers tend to make these mistakes due to the lack of knowledge in their non-native languages.

The research study of Muskyn(1984) identified 13 different patterns in code-mixing, such as heavy slang relexicalization, congruent lexicalization, borrowing, insertional code-switching, mixed pidgins, etc. We analysed our research study to specifically obtain the issues or barriers in Sinhala-English code-mixed text(Kugathasan & Sumathipala, 2020).

- **Spelling errors**

|                  |                       |
|------------------|-----------------------|
| SECM sentence    | : ‘kama vry gd’       |
| Sinhala Sentence | : ‘කෑම ගොඩාක් හොඳයි’  |
| English Sentence | : ‘Food is very good’ |

The words ‘vry’ and ‘gd’ represent the English words ‘very’ and ‘good’. It has spelling errors due to code-mixing.

- **Inconsistency in phonetic transliteration**

|                  |  |
|------------------|--|
| SECM sentence    | : ‘mama wathura bonawa’<br>‘mama vathura bonawa’<br>‘mama vathura bonawaa’ |
| Sinhala Sentence | : ‘මම චතුර බොනවා’  |
| English Sentence | : ‘I drink water’  |

The same sentence is written in different ways in SECM text. The word ‘water’ is represented as ‘vathura’, ‘wathura’ and the word ‘drinking’ is represented as ‘bonawa’, ‘bonawaa’. Same word with different transliterations. This causes an inconsistency.

- **The use of special characters and numeric characters**

|                  |                    |
|------------------|--------------------|
| SECM sentence    | : ‘4to gaththa’    |
| Sinhala Sentence | : ‘ඡායා රූප ගන්නා’ |
| English Sentence | : ‘Took photo’     |

In the SECM sentence, the word ‘4to’, absorbs the phonetic sound of the word ‘four’ and combines it with the word ‘to’, together, it represents the phonetic sound of the word photo. Likewise, there are several situations numeric and special characters are used in the SECM sentence.

- **Borrowing of words**

|                  |                        |
|------------------|------------------------|
| SECM sentence    | : ‘Service eka hondai’ |
| Sinhala Sentence | : ‘සේවාව හොඳයි’        |
| English Sentence | : ‘Service is good’    |

The sentence starts with an English ‘Service’ and suddenly switches to Sinhala transliterated words ‘eka’ and ‘hondai’. The sentence has the basic grammatical structure of Sinhala and it borrows the word service from the English language.

- **Integration of suffixes**

SECM sentence : ‘teacherla hamoma enna’  
Sinhala Sentence : ‘ගුරුවරුන් හැමෝම එන්න’  
English Sentence : ‘All the teachers are requested to come’

The word ‘teachers’ is an English word which is a singular noun and the suffix ‘la’ is in the transliterated form taken from the Sinhala language. Together the word stands for the meaning ‘teachers’ which is plural.

- **Switching for discourse marker**

SECM sentence : ‘niyama kama so ayeth kanna hithenava’  
Sinhala Sentence : ‘නියමයි කෑම ,එක නිසා ආයෙත් කන්න හිතෙනවා ’  
English Sentence : ‘Great food, so like to eat again’

In this sentence, an English discourse marker 'So' is used to join the two sentences which have Sinhala transliterated words.

### **1.2.2 Lack of parallel data for the SECM text translation**

Social media platforms in the South-East Asian context often contain code-mixed texts due to their increasing usage. We analyzed the primary data collection platforms of the code-mixed text-based research studies. According to the analysis, most of the code-mixed text data is collected from social media platforms. Figure 1.2 shows a detailed view of the data collection of several code-mixed text research studies. But only a few studies have the parallel dataset needed for the code-mixed text translation task. This is considered a barrier for low-resource languages such as Sinhala.

| Research                                   | Facebook | Twitter | SMS | Online Forums | Published datasets from online platforms | Other/ not stated |
|--|----------|---------|-----|---------------|--|-------------------|
| Bali et al (2014)                          | ✓        |         |     |               |  |                   |
| Nguyen and Doğruöz (2013)                  |          |         |     | ✓             |  |                   |
| Barman et al. (2014)                       | ✓        |         |     |               |  |                   |
| Shanmugalingam and Sumathipala et al(2018) | ✓        |         |     |               |  |                   |
| Veena, Kumar and Soman et al.(2017)        |          |         |     |               |  | ✓                 |
| Chitthranjan et al.(2014)                  |          | ✓       |     |               |  |                   |
| Mandal, Dipta and Das at el. (2018)        |          |         |     |               | ✓  |                   |
| Wong and Xia(2008)                         |          |         |     |               | ✓  |                   |
| Xue, Yin and Davison(2011)                 |          | ✓       | ✓   |               |  |                   |
| Choudhary et al. (2007)                    |          |         | ✓   |               |  |                   |
| Cook and Stevenson(2009)                   |          |         | ✓   |               | ✓  |                   |
| Sukanya et al(2015)                        |          |         |     |               |  | ✓                 |
| Han, Cook and Baldwin(2012)                |          |         |     |               |  | ✓                 |
| Mandal, t el. (2018).                      |          |         |     |               |  | ✓                 |
| Chih Yu et al(2013)                        |          |         |     |               | ✓  |                   |
| Lignos and Marcus(2013)                    |          | ✓       |     |               |  |                   |
| Volk and Clematide(2014)                   |          |         |     |               | ✓  |                   |
| Papalexakis, Nguyen and Doğruöz(2014)      |          |         |     | ✓             |  |                   |
| Kalchbrenner and Blunsom(2013)             |          |         |     |               |  | ✓                 |
| Carrera et al(2009)                        |          |         |     | ✓             |  | ✓                 |
| Dhar et al(2018)                           |          |         |     |               | ✓  |                   |
| Rijhwani et al(2016)                       |          | ✓       |     |               |  |                   |
| Masoud et al(2019)                         |          |         |     |               | ✓  |                   |

Figure 1.2:Data collection platforms of code-mixed text

### 1.2.3 Lack of successful methods for the code-mixed text translation

Numerous translation research studies have been carried out for the monolingual dataset. Also, there are multilingual Neural Machine Translation models available, where a solo model is used to convert multiple source languages to various target languages. These models motivate knowledge translation among language pairs(Lakew et al., 2018; Tan et al., 2019), zero-shot translation(direct translation among a language pair that has never been used in the training phase (Al-Shedivat & Parikh, 2019; Firat et al., 2016; Gu et al., 2019a; Johnson et al., 2017) and enhance translation of low resource language pairs (Arivazhagan et al., 2019). Rather than these benefits, multilingual NMT systems show poor performance(Firat et al., 2016; Johnson et al., 2017) and substandard translation.

Compared to the number of researches based on monolingual and multilingual translation, the number of code-mixed text translation-based researches is very few. A hybrid model, which is a combination of a Statistical model and knowledge translation approach, was introduced by Carrera et al. for code-mixed text translation(Carrera et al., 2009). Rijhwani et al. have proposed a model with word-level language identification and matrix language detection where the current monolingual translators are applied(Rijhwani et al., 2016). An augmented pipeline approach was proposed by (Singh et al., 2018) and a back translation approach was proposed by (Masoud et al., 2019) for the code-mixed text translation. But non of the mentioned approaches gave a satisfactory BLEU(Bi-Lingual Evaluation Under Study) score.

#### **1.2.4 Unavailability of tools for the Sinhala-English code-mixed text translation**

Even though some models are proposed for the code-mixed text translation, there is no tool built for the models. Therefore, in our study, we have used the implemented model and developed a translation tool for SECM to Sinhala translation.

### **1.3 Research Aim and Objectives**

This research study aims to develop a model for Sinhala-English code-mixed text to Sinhala translation.

The main aim of this project leads to the following research objectives:

1. Analyze the patterns, structure, and grammar in Sinhala-English code-mixed sentences and identify the barriers to the translation of the code-mixed text.
2. Develop a parallel corpus of SECM text collected from social media and its relevant Sinhala translation.
3. Normalization of SECM text ( Slang word normalization, spelling error corrections, and transliteration).

4. Develop a model for the SECM to Sinhala code-mixed text translation, which provides better translations with better BLEU scores.
5. Evaluate the proposed model with the state-of-the-art models
6. Build a web application using the proposed model to translate SECM sentences to Sinhala.

#### **1.4 Contribution of the Thesis**

The contribution of this research study benefits to extend the knowledge in the field of code-mixed text translation. Each contribution and the related publication are listed in the following section.

1. **An in-depth analysis of the code-mixed text :** The research systematically analyses previous studies and compares the methodologies of different tasks performed on code-mixed text. The tasks performed in code-mixed texts are language identification, normalization, code-switching, and translation. The result of this analysis revealed that there are plenty of improvements in code-mixed text-related research.

Related Publications :

Submitted and revised Journal paper -

Title - A Systematic Review of Code-Mixed Text Analysis Approaches

Journal Name – Ampersand

2. **Building the parallel corpus for SECM-Sinhala :** There are no parallel sentences available to conduct Neural Machine Translation on SECM text. In this research study, we collected 5000 SECM sentences from social media platforms and manually translated each sentence with the help of a human translator. To check the correctness of the translation, we used annotators and modified the incorrect translation according to the annotation. This is explained

in detail in Chapter 4.1. Fleiss Kappa method is used to measure the agreement between the raters, which gave a 0.88 score which shows the agreement between the raters stating the translation is valid.

3. **Identifying the challenges in SECM text and transliteration mapping:** In this part of the research, we identified the challenges which are considered the bottleneck for code-mixed text translation. Also, we proposed a mapping where each Sinhala letter is mapped to its most relevant form of transliteration.

Related Publication :

Kugathasan, A., & Sumathipala, S. (2020, March). Standardizing Sinhala code-mixed text using a dictionary-based approach. In 2020 International Conference on Image Processing and Robotics (ICIP) (pp. 1-6). IEEE.

4. **A novel approach to translating SECM sentences to Sinhala:** A neural machine translation model was implemented as a combination of the Seq2Seq model(Encoder-Decoder framework), LSTM(Long-Short Term Memory) and Teacher Forcing Algorithm. This approach solved the many identified challenges of SECM text in translation.

Related Publications :

Conference paper -

Kugathasan, A., & Sumathipala, S. (2021, September). Neural Machine Translation for Sinhala-English Code-Mixed Text. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (pp. 718-726).

Journal paper -

Kugathasan, A. and Sumathipala, S., 2022. Neural machine translation for sinhala-english code-mixed text, International Journal on Advances in ICT for Emerging Regions, Vol 15 No 3 (2022): 2022 December Issue

- 5. Implementation of a web application to translate SECM to Sinhala :** Using the implemented model, a web application is implemented. This tool can be used as a translator application where the user provides SECM text as input and the tool will translate the given input sentence as Sinhala Sentence. Also, the tool provides a feature to calculate the BLEU score based on the reference sentence.

## **1.5 Significance**

The dataset implemented for this research study can be used as an important resource in the domain of Sinhala-English Code-Mixed text translation. Also, the study points out the most frequently occurring challenges in code-mixed text. The model proposed in this thesis collectively deliver a better approach and maximize the use of available dataset to provide efficient result in the domain of code-mixed text translation.

## **1.6 Thesis Outline**

Chapter one provides an idea of what is the research is about. Chapter 2 contains an in-depth literature survey for code-mixed text analysis. Chapter 3 includes analysis of Sinhala-English code-mixed text. Chapter 4 contains the details about the proposed approach. Chapter 5 explains the prediction and evaluation of the model and Chapter 6 contains the result and discussion. Finally, Chapter 7 summarizes and explains the conclusion and the future work.



## **2. Literature Review**

The literature review section of this thesis provides a comprehensive analysis of code-mixing, a linguistic phenomenon that holds significant importance in multilingual communication. To ensure a systematic and thorough exploration of the subject, the review methodology was grounded in the identification and selection of recent and highly cited articles. By focusing on the most influential works published in the field, this review aims to offer a well-rounded understanding of the various dimensions of code-mixing. In line with this objective, the review process involved the meticulous categorization of identified articles into distinct thematic categories, such as Word-level language identification in code-mixing, identification of code-switching, normalization of code-mixing etc., allowing for a structured examination of the multifaceted aspects of code-mixing. Through this rigorous approach, the literature review endeavors to contribute to the advancement of knowledge in the field of multilingual communication and language mixing.

### **2.1 Code-mixing**

#### **2.1.1 Word-Level language identification**

Linguistic communities have shown a massive interest in building language models for code-mixed text in the previous years. According to (Bali et al., 2014), automating the process of code-mixed text analysis is considered an essential task due to the excessive usage of code-mixed text in this digital era. Furthermore, language identification of words in code-mixed is observed as an essential process in code-mixed text analysis.

According to the sentence shown in Figure 2.1, English words and Sinhala transliterated words are mixed in a single sentence. '<En>' refers to English transliterated terms, '<Sin>' refers to Sinhala transliterated terms, '<Sin> + <En>' is a combined word. Two languages are linked in one sentence. Language identification for each word in a single sentence could be considered as one of the tasks to identify whether the sentence is code-mixed if the sentence is identified with words represented in more than one language, the sentence would be marked as code-mixed.

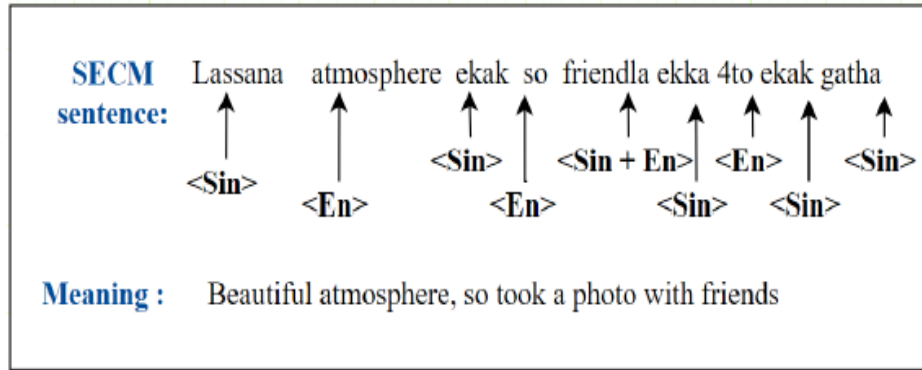


Figure 2.1: Example of Sinhala-English code-mixed text

McCallum introduced a character-based n-gram that is used with the Maximum Entropy classifier to recognize the language based on the grammatical pattern of the sentence (McCallum, 2002). (Veena et al., 2017) (2007) utilized character Conditional Random Field (CRF) (Lafferty et al., 2001), Logistic Regression (LR) (Lui & Baldwin, 2012), and Language model approaches are also used for the language identification task.

(Nguyen & Doğruöz, 2013) collected Turkish-Dutch data, which had much noise from spelling variations due to transliterations, Turkish characters which are replaced with characters from other languages, and misspelt words. Since Turkish and Dutch has so many words with common spellings, the language identification task was very challenging. Somehow the researchers managed to achieve the task by using a combined method of n-gram comparison approach along with the approach called 'langid' (Lui & Baldwin, 2012). 'langid' is an approach of Naïve Bayes with n-gram features.

A language model is implemented using a character-level n-gram approach with Witten-Bell smoothing (Chen & Goodman, 1999). A CRF-based model is proposed by (Chittaranjan et al., 2014) where the attributes such as: capitalization, contextual, character n-gram, special character, and lexicon are considered for language identification. In the research study of Veena et al. (2017), character embedding with

skip-gram architecture (Mikolov et al., 2013) has been used for the language identification of Tamil-English and Malayalam-English code-mixed text.

(Mandal et al., 2018) proposed a phonetic and character-based LSTM approach to recognize the language in the code-mixed text of Bengali-English. First, each word is encoded and to handle the phonetic encoding, two categories of phonetic libraries are created: Root Phones(RP) and Similar Phones(SP). The method variable-length sliding technique is used for the phonetic encoding process. LSTM and Recurrent Neural Network(RNN) have been used to train the RP and SP lists to recognize the language.

A research study based on language detection in Sinhala-English code-mixed text conducted by (Shanmugalingam et al., 2018), proposes an approach based on Unicode characters in roman scripts and term frequencies to identify the code-mixed text and language dictionaries to identify common words from the English language. They have used classifiers such as Support Vector Machine( SVM), Random Forest, Logistic Regression, Decision Tree, and Naive Bayes. The F-measure and the accuracy are recorded for different classifiers. For example, SVM gave the highest accuracy of 89.5% among all the other classifiers. The following Table 2.1 shows a summary of necessary research conducted on word-level language identification.

Table 2.1 : Summary of research studies on word-level language identification

| <b>Paper</b>              | <b>Dataset used</b>   | <b>Approaches</b>  | <b>Evaluation Criteria</b>  | <b>Result</b>    |
|---------------------------|---|--|-----------------------------|------------------|
| Nguyen and Doğruöz (2013) | Online forum data from Netherlands for Turkish-Dutch online community speakers. | Dictionary lookup (DICT)   | Accuracy, Precision, Recall | Accuracy = 85.8% |
|                           |   | Language model (LM)  |                             | Accuracy = 94.4% |
|                           |   | DICT+LM  |                             | Accuracy = 94.6% |
|                           |   | Logistic Regression (LR) + Assigned labels (LAB)                         |                             | Accuracy = 96.1% |
|                           |   | LR + log probability values (PROB)                                       |                             | Accuracy = 97.6% |
|                           |   | Conditional Random Fields (CRF) + Individual tokens as a feature (BASE). |                             | Accuracy = 97.5% |
|                           |   | CRF + LAB  |                             | Accuracy = 97.2% |

|   |   |  |                                   |                   |
|---|---|--|-----------------------------------|-------------------|
|   |   | CRF+PROB   |                                   | Accuracy = 97.6%  |
| Barman et al.<br>(2014)   | Facebook posts and comments collected using FaceBook graph API explorer | Baseline(Dictionary Based) Approach                                    | Accuracy,<br>Precision,<br>Recall | Accuracy=93.64%   |
|   |   | Word-level classification using SVM and without contextual information |                                   | Accuracy=95.21%   |
|   |   | Word-level classification using SVM with contextual information        |                                   | Accuracy=95.52%   |
|   |   | Sequence labelling using CRFs including contextual information         |                                   | Accuracy=95.76%   |
| Shanmugalingam<br>, Sumathipala,<br>and<br>Premachandra<br>(2018) | Tamil-English<br>Facebook posts<br>and comments                         | Support Vector Machine   | Accuracy,<br>F-Score              | Accuracy = 89.46% |
|   |   | Random Forest  |                                   | Accuracy = 86.06% |
|   |   | Logistic Regression  |                                   | Accuracy = 89.09% |

|                                      |   |   |          |                                      |
|--------------------------------------|---|---|----------|--------------------------------------|
|                                      |   | Decision Tree   |          | Accuracy = 94.5%                     |
|                                      |   | Naive Bayes   |          | Accuracy = 84.61%                    |
| Veena, M. A. Kumar, and Soman (2017) | Tamil-English and Malayalam-English Facebook posts and comments | Character embedding with skip-gram architecture, including context details for word-level language identification. Support Vector Machine(SVM) classifier for the training of the model | Accuracy | Accuracy for Malayalam-English=93%   |
|                                      |   |   |          | Accuracy for Tamil-English=95%       |
| Chittaranjan et al. (2014)           | Tweets  | CRF based approach  | Accuracy | Accuracies = 80% - 95%               |
| Mandal, S. D. Das, and D. Das (2018) | Transliterated Bengali words corpus from ICON 16 4, ICON 17     | Root phones.  | Accuracy | Accuracy of stacking method = 91.78% |
|                                      |   | Phonetic and Character encoding approaches  |          |                                      |
|                                      | English terms corpus from online resources.                     | Two deep LSTM models.<br>Ensemble models using stacking and threshold techniques  |          | Accuracy of threshold model = 92.35% |

### 2.1.2 Identification of code-switched text

Code-switching is observed as a segment of code-mixing. Changing languages among a single sentence is called code-switching. The sentence could also contain more than one grammatical structure from different languages. Figure 2.2 shows an example of a code-switched sentence. The sentence comprises both Sinhala and English words. The initial part of the sentence contains English words where English grammatical structure is used, and the latter part of the sentence contains Sinhala words where that part of the sentence has the Sinhala grammatical structure. In code-switching, the grammatical structure changes according to the language in the inter-sentential level.

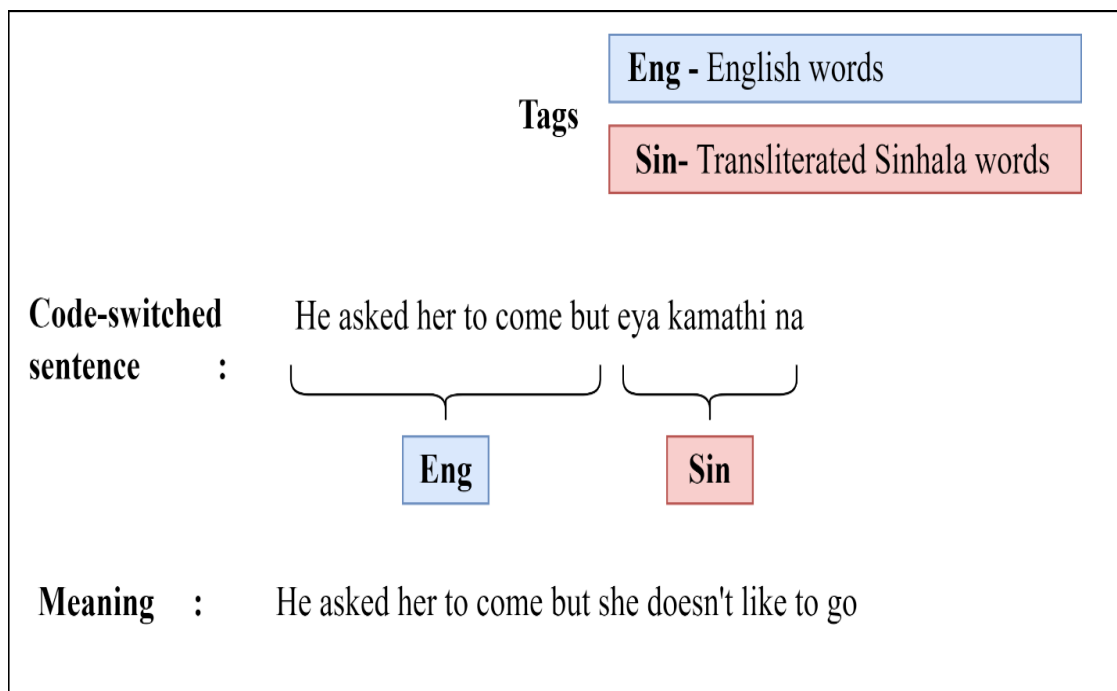


Figure 2.2: An example of a Sinhala-English code-switched sentence

(Yu et al., 2013) propose a method with two main tasks. The first task is to recognize the language of the term in a given sentence. A dataset of code-switched and non-code-switched sentences are prepared to conduct the second task, where separate language

models are built. The highest probability n-gram output obtained by the code-switching model is compared against the production obtained from non-code-switching. This technique is used to identify whether the  $n^{\text{th}}$  word in the input sentence is code-switched or not.

(Lignos & Marcus, 2013) proposed a supervised learning approach of token-level language detection to identify whether a sentence is code-mixed or not. Two monolingual corpora are used to measure the ratio probability of a word( $w$ ) in each language according to the equation shown in equation (1).

$$\frac{P(w|Spanish)}{P(w|English)} \quad (1)$$

Each word is tagged by its controlling (superior) language. For example, suppose the calculated probability is close to value 1. In that case, that word is marked as ambiguous. If the sentence contains sufficient words from each language, the sentence is labelled as a code-switched sentence.

According to (Volk & Clematide, 2014), sub-sentential tokens in the sentences containing unknown lemma and framed with quotation marks are searched and labelled. The rule is that at least two terms must be identified outside the quotation. Identified string sequence would be given as input to the Langid to identify the English word sequence and other language sequences. If the result sentence is divergent for the neighbourhood sequence, the sentence is labelled as code-switched.

Binary classification using Naïve Bayes classifier is another method used to identify the code-switched sentences by (Papalexakis et al., 2014). Suppose the  $n^{\text{th}}$  token and  $(n+1)^{\text{th}}$  token are identified as they are not from the same language the sentence is labelled as 1 and otherwise 0. A deep learning approach using Part Of Speech (POS) tagging to identify



the code-switched text is proposed by (Attia et al., 2019). According to this study, POS categories greatly assist in identifying code-switching.

Many studies have been conducted on measuring the amount or level of code-switching in a code-mixed sentence. (Barnett et al., 2000) introduced the M-Index, which is a metric used to measure the imbalance in the distribution of language in datasets. (Guzman et al., 2016) extended this idea and introduced a new metric Integration index(I-index) to identify the probability of code-switching within a dataset. This index provided a value in-between 0 to 1, where a zero I-index means there is no code-switching. In 2016, (Gambäck & Das, 2016)proposed a Code Mixing Index(CMI), which assesses the code-switching of a particular dataset based on the frequency of unique words. Table 2.2 shows the summary of research studies based on code-switching.

Table 2.2: Summary of research studies on code-switched texts

| Paper            | Dataset used   | Approaches                 | Evaluation Criteria | Result                             |
|------------------|--|----------------------------|---------------------|------------------------------------|
| Yu et al. (2013) | Mandarin-English corpus of web-based news articles,  | Language modeling approach | Accuracy            | Language model Accuracy = 79.01%   |
|                  |  | POS based model            |                     | Word based model accuracy = 41.09% |
|                  | A corpus called Sinica which is released by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) | Word based model           |                     | POS based model accuracy= 53.08%   |

|   |   |   |           |   |
|---|---|---|-----------|---|
| Lignos and Marcus(2013)                 | Tweets and Spanish corpus by MITRE for testing  | Ratio list model  | Accuracy  | Accuracy = 96.6%                            |
| Volk and Clematide(2014)                | Multilingual diachronic corpus of Swiss Alpine texts.   | Classification of word sequence approach                    | Precision | With correct language label precision = 78% |
|   |   |   |           | Unlabelled precision = 78%                  |
| Papalexakis, Nguyen, and Doǵruoz (2014) | Online discussion forum dataset obtained from the Turkish-Dutch immigrant community in the Netherlands. | Naive Bayes classifier with various combination of features |           |   |
|   |   |   |           | Language of tokens                          |

|  |  |  |                                  |  |
|--|--|--|----------------------------------|--|
|  |  | Language + previous code-switching             | Precision,<br>Recall,<br>F-score | Precision of Language + previous code-switching = 0.67           |
|  |  | Emoticons                                      |                                  | Emoticons = 0.67   |
|  |  | MWEs   |                                  | Precision of MWEs = 0.53   |
|  |  | Emoticons + MWEs                               |                                  | Precision of Emoticons + MWEs = 0.52                             |
|  |  | Language + previous code-switching + emoticons |                                  | Precision of Language + previous code-switching + emoticons=0.69 |
|  |  | Language + previous code-switching + MWEs      |                                  | Language + previous code-switching + MWEs=0.71                   |
|  |  | All  |                                  | Precision of All = 0.68  |

### 2.1.3 Normalization of code-mixed text

Normalization is the process of converting an informal text into a standard format. The massive growth of informal text in social media creates the demand for the normalization models. In Section 1.2.1 the challenges identified in code-mixed texts are stated clearly. Inconsistency in transliteration, incorrect spellings, and abbreviations in code-mixed texts are considered some of the key features that are a barrier to translating code-mixed text to a standard language.

The very first normalization model was introduced by (Wong & Xia, 2008), for a Chinese chat corpus. Phonetic mapping and SCM(Source Channel Model) are used for the normalization of the chat corpus. For the given input string, the SCM approach obtains the most suitable character string.

$$T = \{t_i\}_{j=1,2,3,..,n} \quad (2)$$

$$C = \{C_i\}_{j=1,2,3,..,n} \quad (3)$$

$$C = \underset{c}{\operatorname{argmax}} P(C|T)$$

$$C = \underset{c}{\operatorname{argmax}} P(T|C)P(C) \quad (4)$$

In equation (2), the given input string is denoted by  $T$ , and the input characters are denoted by  $t_i$ . In equation (3), the output string is denoted by  $C$ , and the output characters are denoted by  $c_i$ . Finally, Wong & Xia(2008) finds the most probable character string to the equation (4).

(Xue et al., 2011) enhanced the model of Wong & Xia(2008), where 4 factors are considered for normalization: phonetic, orthographic, contextual and acronym expansion. Hidden Markov Model approach is used by (Choudhury et al., 2007) to normalize the SMS(Short Message Text). The proposed algorithm facilitates to build the word-level Hidden Markov Model with a specific set of generalized parameters. This study was later continued by (Cook & Stevenson, 2009), to build an unsupervised noise-channel model, which increased the accuracy of normalization by 2%.

(Dutta et al., 2015), proposed an approach where the Noisy-Channel framework is used to check the misspelt words and correct them. Baye's theory is applied in the noisy channel model to identify the misspelt word, which would be replaced with the most likely word the writer expected to write. In equation (5), the cluster of all the words belonging to one language is denoted by  $V$ , which means the vocabulary. In the channel model,  $P(x/w)$  denotes the probability of how likely the word  $w$  would be accidentally written as the word  $x$ . In the language model,  $P(w)$  denotes the maximum likelihood estimate of how likely it is that  $w$  is written in the first place. The confusion matrix is used to identify the counts of the letters that are inserted/deleted/substituted by mistake.

$$\begin{aligned}
 \hat{w} &= \operatorname{argmax}_{w \in \mathcal{V}} p(w|x) \\
 &= \operatorname{argmax}_{w \in \mathcal{V}} \frac{p(w|x)p(w)}{p(x)} \\
 &= \operatorname{argmax}_{w \in \mathcal{V}} p(x|w)p(w)
 \end{aligned} \tag{5}$$

The confusion matrix is built based on the Levenstein edit distance algorithm. This matrix is categorized into insertion, deletion, substitution and transposition matrices. Wordplay, such as intentionally misspelt words to express emotions, is identified using regular expressions. The normalization here is done by replacing the repeating characters. After the replacement, if the resulting word is still not valid, it will be directed to the noisy channel model for spell checking. Also, the researchers identified and replaces the shortcuts used in words such as ‘fav’ instead of ‘favorite’ and the phonetically spelt words such as ‘4got’ used instead of ‘forgot’. The accuracy of 69.4% is gained from the overall spellchecker model.

Mandal & Nanmaran(2018) introduced a normalization approach with two main models for transliterated text. Sequence to sequence model is used to convert the user transliterated words to standard transliteration according to the ITRANS(Indian Languages Transliteration) corpus. The second model is based on the string matching algorithm, Levenshtein Distance(LD). The string matching algorithm measures the difference between the sequence of two words. The first model gave better accuracy than the second model. Another normalization approach for the anomalies in the code-mixed text is introduced by Singh et al. (2018)[7]. Different variations of the same word are captured using the skip-gram and edit distance approaches. Skip-gram considers the context and represents it in a semantic space. If the context is similar, the vector representations of variations of the same word and the consonant are also similar. Based on this, a similarity metric is defined to calculate the closest variants of a given the word. The cluster of words with similar representations are identified, and the most frequently used word is labelled as a parent. It will be used as a substitute for the words with similar representation. Substitution of the parent word representation reduces the number of words and reduces the noise.

A four-module approach is introduced by Barik et al. (2019) to normalize the code-mixed text. Tokenization and labelling are the main processes in the first module in the pipeline. Each chunk in the sentences is labelled as B (beginning), I (inside), and O (outside). The labelled data is the input to the second module, language identification, where Conditional Random Field(CRF) is used. Lexical normalization is the third module which replaces the Out Of Vocabulary(OOV) tokens with the relevant word from the dictionary. The fourth module combines the first three modules and translates the final output from the combined modules into Indonesian. Another normalization model proposed by Lourentzou et al. (2019) combines word-based and character-based models to identify the OOV by preserving contextual information. Finally, a Convolutional Neural Network model with character-level embedding is proposed by Arora & Kansal(2019) to normalize the unstructured texts and noisy sentences.



Table 2.3: Summary of research studies based on text normalization

| <b>Paper</b>                 | <b>Dataset used</b>                     | <b>Approaches</b>  | <b>Evaluation Criteria</b>             | <b>Result</b>   |
|------------------------------|---|--|--|-----------------|
| Wong and Xia(2008)           | Monolingual chat corpus for the Chinese | eXtended Source Channel Model (XSCM)                                     | F-measure                              | F1-Score = 0.88 |
| Xue, Yin, and Davison (2011) | Tweets and SMS messages                 | The multi-channel model with generic channel probabilities (MC-Generic)  | Accuracy, F-measure, Precision, Recall | Accuracy = 0.96 |
|                              |   | The multichannel model with term-dependent channel probabilities (MC-TD) |  | Accuracy = 0.96 |
|                              |   | Aspell   |  | Accuracy = 0.92 |
|                              |   | Moses  |  | Accuracy = 0.94 |
| Choudhury et al. (2007)      | Short messages over mobile phones       | Hidden Markov Model(HMM)   | Accuracy                               | Accuracy=57.7%  |

|                            |   |   |          |                   |
|----------------------------|---|---|----------|-------------------|
| Cook and Stevenson(2009)   | Short messages over mobile phones   | Unsupervised noisy channel framework                                | Accuracy | Accuracy=59.4%    |
| Dutta et al. (2015)        | Comments and Public posts written by multilingual speakers extracted from the social media  | Combination of CRF based ML approach and post-processing heuristics | Accuracy | Accuracy=69.43%   |
| Mandal and Nanmaran (2018) | Data set from Mandal, S. D. Das, and D. Das (2018), phonetically transliterated corpus, standard Roman transliterations (ITRANS) corpus | Seq2Seq approach  | Accuracy | Accuracy = 50.01% |
|                            |   | String matching using Levenshtein Distance(LD)                      |          | Accuracy = 90.27% |
|                            |   |   |          |                   |

|  |         |                            |          |                                      |
|--|---------|----------------------------|----------|--------------------------------------|
| Rajat Singh*, Nurendra Choudhary* and Manish Shrivastava(2018) | Twitter | Skip-gram & Edit Distance  | Accuracy | Accuracy on Bengali-English = 76.14% |
|  |         |                            |          | Accuracy on Hindi-English = 75.24%   |
|  |         |                            |          | Accuracy on Tamil-English = 65.97%   |
| Barik et al(2019)  | Twitter | 4 module pipeline approach | Accuracy | F1 - Score = 81.31                   |
|  |         |                            | F-Score  | Accuracy = 68.50                     |

## 2.2 Machine Translation

The significance of Machine Translation(MT) is increased due to the massive demand for translation for the purposes like military services, overseas businesses, valuable social media content from different languages, and profitable customers with a preference for different languages. Neural Machine Translation(NMT) is the currently trending domain in Machine Translation. Recurrent Neural Network (Kalchbrenner & Blunsom, 2013), Seq2Seq approach (Sutskever et al., 2014), and Attention-based NMT (Bahdanau et al., 2014) are considered trending approaches for NMT.

There are many translation-related research conducted on monolingual datasets. A combined model of shallow and fusion with NMT techniques has been introduced by (Gulcehre et al., 2015). There are two approaches proposed by (Sennrich et al., 2015) to the monolingual translation. The first approach is to match the monolingual dataset with dummy input, and the second approach is using a pre-trained model on a parallel corpus with NMT technique. A semi-supervised approach by combining the labelled and unlabeled dataset is proposed by (Cheng, 2019).

Rather than the monolingual MT models, there are multilingual NMT models, which support many languages from a single model. These models obtain the knowledge translation among language pairs(Lakew et al., 2018; Tan et al., 2019). The zero-shot translation is a direct translation among a language pair that has never been used in the training phase (Al-Shedivat & Parikh, 2019; Firat et al., 2016; Gu et al., 2019b; Johnson et al., 2017) enhances the translation of low resource language pairs.

Rather than these benefits, multilingual NMT systems show poor performance(Arivazhagan et al., 2019; Johnson et al., 2017) and bad translations when accommodating many languages. To overcome the issue of the representation of

multilingual NMT models, (Zhang et al., 2020) propose an improved approach with normalization and linear transformation layers. Furthermore, to overcome the issue of unseen training language pairs, they propose a Random Online Back Translation approach(ROBT).

If we consider the research studies related to code-mixed text translation, only a few researches have been conducted. A qualitative study on a code-switched dataset conducted by (Carrera et al., 2009) revealed that hybrid models combined with the Knowledge Translation approach (Sudsawad, 2007) and Statistical Modelling (Neale et al., 1999) achieved comparatively good translation. (Rijhwani et al., 2016)introduced a code-mixed text translation approach where the languages in a code-mixed sentence are marked as dominant(matrix language) and non-dominant(embedded language) languages. Word level language identification is the initial task in this proposed approach. After the identification of the language of each word, the dataset would be applied with a current translator to translate the data according to the users choice.

(Dhar et al., 2018) published a parallel corpus of Hindi-English code mixed with relevant English sentences. He introduced an augmentation pipeline approach. The pipeline contains the modules of language identification, dominant language identification, translation of the dominant language, and translation of the target language. Monolingual translated sentences would be the output of this augmented pipeline approach. (Masoud et al., 2019) proposed a back translation model for covert Tamil-English code-switched text. Monolingual, baseline and hybrid approaches are used to assess the system. The proposed back-translated model achieved the highest BLEU score of 25.28 for the code-switched sentences in the experiment.

### **3. Code-mixed text analysis**

#### **3.1 Background Study**

Sri Lanka acknowledges Sinhala, English and Tamil as the formal languages used for official activities. We mainly have two categories of code-mixed languages; Sinhala-English (Singlish) and Tamil-English(Tanglish). But there is no mixing between Sinhala and Tamil languages.

Kachru (1986) explains the necessity of English in the Asian continent. Many former Anglo-American colonies have been identified with varieties of English language, which is called a deviation from the standard English to the later development world. According to his observation in South Asia, the English language is considered a sign of 'modernization', 'achievement', and 'strength'. He defines code-mixing as a highlight of modernization, social and economic status and membership in an aristocracy society. The widest code-mixing range is identified in the English language. Rather than the effect of colonization, code-mixing depends on several other facets such as the medium of communication in school or workplace, amount of exposure to English, topics and contexts of discussion and rate of occurrence of language used in distinct domains.

People have massively adopted internet usage in the 21<sup>st</sup> century. Code-mixed texts are adapted to the vocabulary and grammar of languages used by the particular bilingual or multilingual user. The structure of code-mixed text depends on the individuals(Choudhary et al. 2018).

The Sinhala language has base on Brahmi script in its ornamentation of writing(Wasala, Weerasinghe, and Gamage 2006). In the latest Sinhala alphabet 41 consonants, 18 vowels and two half vowels, altogether 61 characters are there according to the Unicode standard(Punchimudiyanse and Meegama 2015). Even though there are 61 letters, the

language has only 40 different sounds represented by those letters. Singlish is a code-mixed language which is originated from the multilingual society of Sinhala-English-speaking people. Srilankan uses Singlish as one of the main communication languages in social media. It has become viral among the younger generation of the 21<sup>st</sup> century. When it comes to Sinhala codemixed text code-mixing, we identified there are several challenges in the representation of the text. Spelling errors, integration of suffixes, the use of special characters and numeric characters in the text, borrowing of words from another language, combined languages, switching of discourse markers, and inconsistent phonetic transliteration are some of the issues we have identified in Sinhala code mixed language. These challenges were described in detail in section 1.2.1.

As shown in Figure 3.1, we have conducted a survey study to collect data and understand the necessity of processing Sinhala-English code-mixed text and the usage of the code-mixed text(Kugathasan & Sumathipala, 2021). There were 82 native Sinhala-speaking citizens participated in the survey study. Most participants have stated that SECM is used as the communication medium in social media rather than the Sinhala language. Furthermore, 78% of participants have stated that usage of code-mixed text started due to the easiness/flexibility of typing using the standard keyboards. Also, 12.2% have stated they are interested in using the code-mixed text.

According to the survey study, 2.4% have stated that they started using SECM on an online platform is in between the years 1996-2000. 8.5% have stated that from 2000 to 2005, 35.4% have stated the years as 2005-2010 and the majority have said that the usage of SECM started between the years 2010-2015. Among the 82 participants, SECM text usage in social media application are 59.8%, SECM language usage in chat applications is 93.9% and community blogs, discussion forums, etc., show comparatively less usage of code-mixed text. The principal expectation of this research is to detect the usage of SECM language, and it is identified that most people use code-mixed text on online platforms.

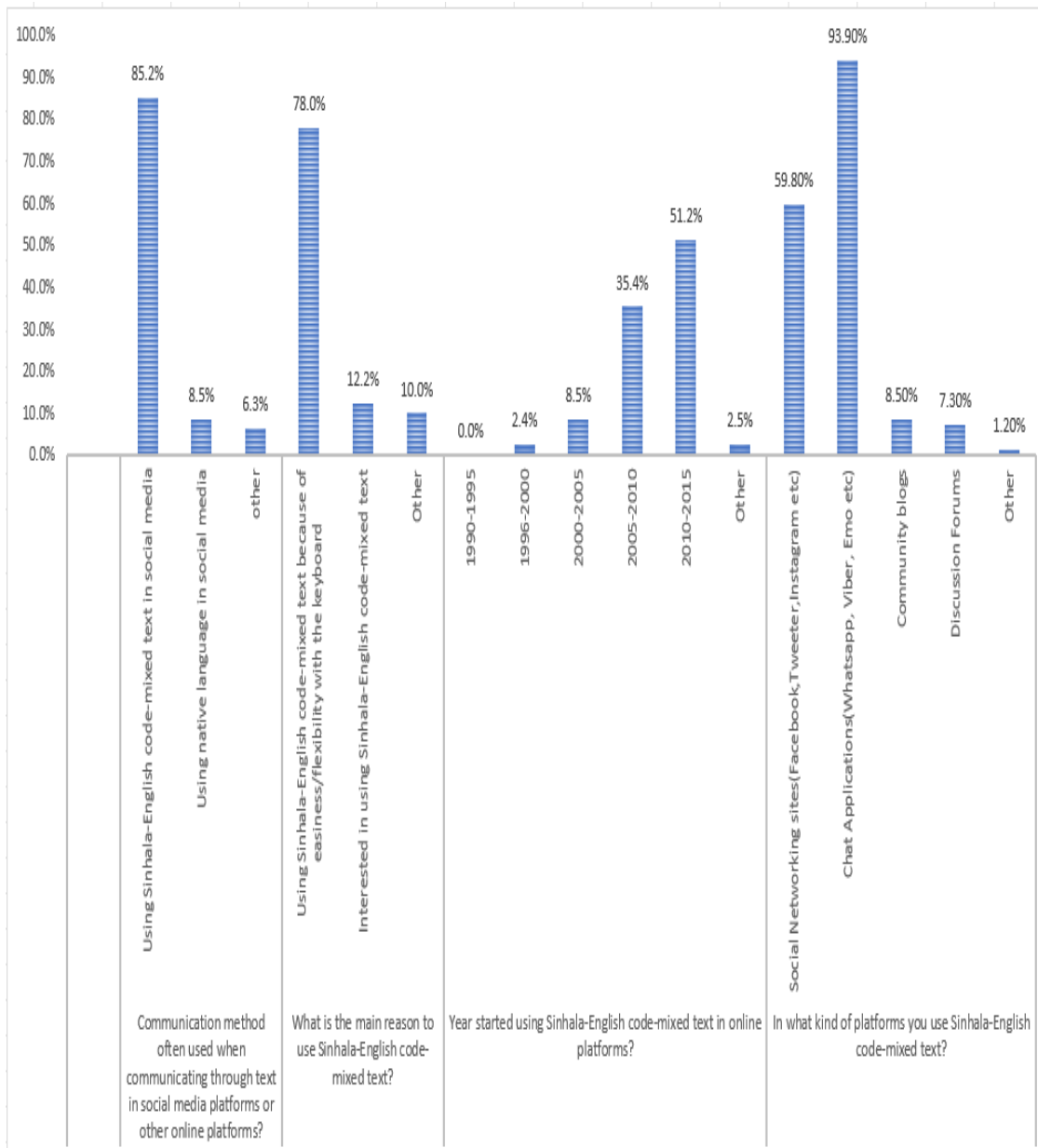


Figure 3.1: Results of the survey on Sinhala-English code-mixed text usage among Sri Lankan



### 3.2 Challenges in Sinhala-English code-mixed text

#### 3.3 Transliteration

The standard ISO15919 was published in 2001 by the International Organization for Standardization. ISO15919 is an international standard for Romanization, including Tamil and Sinhala languages. Weerasinghe et al. (2005) utilized the IPA(International Phonetic Alphabet) format to represent the Sinhala letters in their research study(Weerasinghe, Wasala, and Gamage 2005). IPA and ISO15919 represent Sinhala letters written using English alphabets, with more than 26 letters (Hettige and Karunananda 2007). A conventional tag set was proposed by Wasala Gamage, which uses 26 alphabet English letters to present the phonetical sound of Sinhala letters using the festival framework(Wasala and Gamage 2005).

Code mixing is explained as one of the approaches to write in the roman script (Davies and Bentahila, 2007). Even though the standard tag set from Hettige and Karunananda (2007) is considered the actual roman representation, the romanization in code-mixing used in multilingual societies is different from the standard romanization. Figure 3.2 and Figure 3.3 shows us how the standard romanization defined for Sinhala letters differs from the roman representation used in Singlish text.

| Basic Consonants        |    |    |     |     |    |    |    |    |      |     |     |      |    |    |    |     |    |    |    |    |    |    |    |   |   |
|-------------------------|----|----|-----|-----|----|----|----|----|------|-----|-----|------|----|----|----|-----|----|----|----|----|----|----|----|---|---|
|                         | ක  | ඝ  | ඞ   | ච   | ජ  | ඨ  | ඩ  | ඳ  | ඳ    | ඳ   | ඳ   | ඳ    | ඳ  | ඳ  | ඳ  | ඳ   | ඳ  | ඳ  | ඳ  | ඳ  | ඳ  | ඳ  | ඳ  | ඳ | ඳ |
| Phonetic Tag(ISO 15919) | k  | gh | ŋg  | c   | j  | t̪ | ɖ  | ɳ  | ɳ̪   | t   | d   | ɳ̪   | p  | b  | m  | mb  | y  | r  | l  | l  | v  | s  | h  |   |   |
| Romanized tagset        | k  | zg | zng | ch  | jh | t  | d  | zn | zndx | txh | dh  | qndh | p  | zk | m  | xmb | y  | r  | l  | zl | v  | s  | h  |   |   |
| Singlish representation | ka | ga | nga | cha | ja | ta | da | na | nda  | tha | dha | nda  | pa | ba | ma | mba | ya | ra | la | la | va | sa | ha |   |   |

Figure 3.2: Comparison between phonetic representation which are considered as the standard, romanized representation and Singlish representation for Sinhala Basic consonants

| Vowels                  |   |     |     |      |   |     |   |     |      |       |     |      |   |     |   |     |
|-------------------------|---|-----|-----|------|---|-----|---|-----|------|-------|-----|------|---|-----|---|-----|
|                         | අ | ආ   | ඇ   | ඈ    | ඉ | ඊ   | උ | ඌ   | ඍ    | ඎ     | ඏ   | ඐ    | එ | ඒ   | ඓ | ඔ   |
| Phonetic Tag(ISO 15919) | a | ā   | æ   | ǣ    | i | ī   | u | ū   | ɨ    | ĩ     | ɹ   | ṙ    | e | ē   | o | ō   |
| Romanized tagset        | a | axa | xæe | aeae | i | ixi | u | uxu | zilu | ziluu | zri | zrii | e | eze | o | oxo |
| Singlish representation | a | aa  | a   | aa   | i | ixi | u | uu  | li   | lii   | ri  | ru   | e | ee  | o | oo  |

Figure 3.3: Comparison between phonetic representation, which is considered as the standard, romanized representation and Singlish representation for Sinhala Vowels

### 3.4 Mapping of Sinhala letters to Singlish

We initiated an analysis of how each Sinhala letter is represented with Singlish text. Making the Singlish representation consistent by standardizing the Sinhala letters would reduce the unnecessary noise produced in SECM text.

Two main categories are defined for Sinhala alphabets: Vowels and Consonants. Consonants are furthermore divided into basic consonants and other consonants. Vocalic characters retrieved from Sanskrit which behave like vowels are called the ‘Basic consonants’ and ‘Other consonants’ contains the mixed alphabets which are known as ‘mifra hodiya’(Kugathanan & Sumathipala,2021). A dictionary is created where each Sinhala letter is mapped to it’s most frequently used Singlish format, as shown in Figure 3.4, Figure 3.5 and Figure 3.6.

Some letters has been mapped to more than one Singlish format according to their frequent usage. This standardization would minimize the obscurity and will reduce the noise of the Singlish sentences.

| Vowels |    |   |    |   |    |   |    |    |     |    |    |   |    |   |    |    |
|--------|----|---|----|---|----|---|----|----|-----|----|----|---|----|---|----|----|
| අ      | ආ  | ඇ | ඈ  | ඉ | ඊ  | උ | ඌ  | ඍ  | ඎ   | ඏ  | ඐ  | එ | ඒ  | ඓ | ඔ  | ඕ  |
| a      | aa | a | aa | i | ee | u | uu | li | lii | ri | ru | e | ee | o | oo | au |

Figure 3.4:Mapping of Sinhala Vowels and proposed Singlish form

| Basic Consonants |     |      |     |      |     |      |     |      |     |      |      |     |      |       |
|------------------|-----|------|-----|------|-----|------|-----|------|-----|------|------|-----|------|-------|
| ක්               | ක   | කා   | කෑ  | කා   | කී  | කී   | කු  | කූ   | කෙ  | කේ   | කේ   | කො  | කෝ   | කො    |
| k                | ka  | kaa  | ka  | kaa  | ki  | kee  | ku  | koo  | ke  | kee  | kai  | ko  | ko   | kau   |
| ජ්               | ජ   | ජා   | ජෑ  | ජා   | ජී  | ජී   | ජු  | ජූ   | ජෙ  | ජේ   | ජේ   | ජො  | ජෝ   | ජො    |
| g                | ga  | gaa  | ga  | gaa  | gi  | gee  | gu  | goo  | ge  | ge   | gai  | go  | go   | gau   |
| ඟ්               | ඟ   | ඟා   | ඟෑ  | ඟා   | ඟී  | ඟී   | ඟු  | ඟූ   | ඟෙ  | ඟේ   | ඟේ   | ඟො  | ඟෝ   | ඟො    |
| ng               | nga | ngaa | nga | ngaa | ngi | ngee | ngu | ngoo | nge | ngee | ngai | ngo | ngoo | nngau |
| ච්               | ච   | චා   | චෑ  | චා   | චී  | චී   | චු  | චූ   | චෙ  | චේ   | චේ   | චො  | චෝ   | චො    |
| ඡ්               | ඡ   | ඡා   | ඡෑ  | ඡා   | ඡී  | ඡී   | ඡු  | ඡූ   | ඡෙ  | ඡේ   | ඡේ   | ඡො  | ඡෝ   | ඡො    |
| උ                | උ   | උා   | උෑ  | උා   | උී  | උී   | උු  | උූ   | උෙ  | උේ   | උේ   | උො  | උෝ   | උො    |
| j                | ja  | jaa  | ja  | jaa  | ji  | jee  | ju  | joo  | je  | je   | jai  | jo  | jo   | jau   |
| ච්               | ච   | චා   | චෑ  | චා   | චී  | චී   | චු  | චූ   | චෙ  | චේ   | චේ   | චො  | චෝ   | චො    |
| t                | ta  | taa  | ta  | taa  | ti  | tee  | tu  | too  | te  | te   | tai  | to  | to   | tau   |
| ච්               | ච   | චා   | චෑ  | චා   | චී  | චී   | චු  | චූ   | චෙ  | චේ   | චේ   | චො  | චෝ   | චො    |
| d                | da  | daa  | da  | daa  | di  | dee  | du  | doo  | de  | de   | dai  | do  | doo  | dau   |
| ඳ්               | ඳ   | ඳා   | ඳෑ  | ඳා   | ඳී  | ඳී   | ඳු  | ඳූ   | ඳෙ  | ඳේ   | ඳේ   | ඳො  | ඳෝ   | ඳො    |
| n                | na  | naa  | na  | naa  | ni  | nee  | nu  | noo  | ne  | ne   | nai  | no  | no   | nau   |
| ච්               | ච   | චා   | චෑ  | චා   | චී  | චී   | චු  | චූ   | චෙ  | චේ   | චේ   | චො  | චෝ   | චො    |
| nd               | nda | nda  | nda | nda  | ndi | ndee | ndu | ndoo | nde | nde  | ndai | ndo | ndo  | ndau  |
| ඡ්               | ඡ   | ඡා   | ඡෑ  | ඡා   | ඡී  | ඡී   | ඡු  | ඡූ   | ඡෙ  | ඡේ   | ඡේ   | ඡො  | ඡෝ   | ඡො    |
| th               | tha | thaa | tha | thaa | thi | thee | thu | thoo | the | the  | thai | tho | tho  | thau  |
| ඳ්               | ඳ   | ඳා   | ඳෑ  | ඳා   | ඳී  | ඳී   | ඳු  | ඳූ   | ඳෙ  | ඳේ   | ඳේ   | ඳො  | ඳෝ   | ඳො    |
| dh               | dha | dhaa | dha | dhaa | dhi | dhee | dhu | dhoo | dhe | dhe  | dhai | dho | dhoo | dhau  |
| ඳ්               | ඳ   | ඳා   | ඳෑ  | ඳා   | ඳී  | ඳී   | ඳු  | ඳූ   | ඳෙ  | ඳේ   | ඳේ   | ඳො  | ඳෝ   | ඳො    |
| nd               | nda | nda  | nda | nda  | ndi | ndee | ndu | ndoo | nde | nde  | ndai | ndo | ndo  | ndau  |
| ඡ්               | ඡ   | ඡා   | ඡෑ  | ඡා   | ඡී  | ඡී   | ඡු  | ඡූ   | ඡෙ  | ඡේ   | ඡේ   | ඡො  | ඡෝ   | ඡො    |
| p                | pa  | paa  | pa  | paa  | pi  | pee  | pu  | poo  | pe  | pe   | pai  | po  | po   | pau   |
| ච්               | ච   | චා   | චෑ  | චා   | චී  | චී   | චු  | චූ   | චෙ  | චේ   | චේ   | චො  | චෝ   | චො    |
| b                | ba  | baa  | ba  | baa  | bi  | bee  | bu  | boo  | be  | be   | bai  | bo  | bo   | bau   |
| ච්               | ච   | චා   | චෑ  | චා   | චී  | චී   | චු  | චූ   | චෙ  | චේ   | චේ   | චො  | චෝ   | චො    |
| m                | ma  | maa  | ma  | maa  | mi  | mee  | mu  | moo  | me  | me   | mai  | mo  | mo   | mau   |
| ච්               | ච   | චා   | චෑ  | චා   | චී  | චී   | චු  | චූ   | චෙ  | චේ   | චේ   | චො  | චෝ   | චො    |
| mb               | mba | mbaa | mba | mbaa | mbi | mbee | mbu | mboo | mbe | mbe  | mbai | mbo | mbo  | mbau  |
| ය්               | ය   | යා   | යෑ  | යා   | යී  | යී   | යු  | යූ   | යෙ  | යේ   | යේ   | යො  | යෝ   | යො    |
| y                | ya  | yaa  | ya  | yaa  | yi  | yee  | yu  | yoo  | ye  | ye   | yai  | yo  | yo   | yau   |
| ර්               | ර   | රා   | රෑ  | රා   | රී  | රී   | රු  | රූ   | රෙ  | රේ   | රේ   | රො  | රෝ   | රො    |
| r                | ra  | raa  | ra  | raa  | ri  | ree  | ru  | roo  | re  | re   | rai  | ro  | ro   | rau   |
| ල්               | ල   | ලා   | ලෑ  | ලා   | ලී  | ලී   | ලු  | ලූ   | ලෙ  | ලේ   | ලේ   | ලො  | ලෝ   | ලො    |
| l                | la  | laa  | la  | laa  | li  | lee  | lu  | loo  | le  | le   | lai  | lo  | lo   | lau   |
| ළ්               | ළ   | ළා   | ළෑ  | ළා   | ළී  | ළී   | ළු  | ළූ   | ළෙ  | ළේ   | ළේ   | ළො  | ළෝ   | ළො    |
| l                | la  | laa  | la  | laa  | li  | lee  | lu  | loo  | le  | le   | lai  | lo  | lo   | lau   |
| ච්               | ච   | චා   | චෑ  | චා   | චී  | චී   | චු  | චූ   | චෙ  | චේ   | චේ   | චො  | චෝ   | චො    |
| v                | va  | vaa  | va  | vaa  | vi  | vee  | vu  | voo  | ve  | ve   | vai  | vo  | vo   | vau   |
| ඡ්               | ඡ   | ඡා   | ඡෑ  | ඡා   | ඡී  | ඡී   | ඡු  | ඡූ   | ඡෙ  | ඡේ   | ඡේ   | ඡො  | ඡෝ   | ඡො    |
| s                | sa  | saa  | sa  | saa  | si  | see  | su  | soo  | se  | se   | sai  | so  | so   | sau   |
| ඡ්               | ඡ   | ඡා   | ඡෑ  | ඡා   | ඡී  | ඡී   | ඡු  | ඡූ   | ඡෙ  | ඡේ   | ඡේ   | ඡො  | ඡෝ   | ඡො    |
| h                | ha  | haa  | ha  | haa  | hi  | hee  | hu  | hoo  | he  | he   | hai  | ho  | ho   | hau   |

Figure 3.5: Mapping of Sinhala Basic consonants and proposed English form

| Other consonants |      |       |      |       |      |       |      |       |      |      |       |      |      |       |
|------------------|------|-------|------|-------|------|-------|------|-------|------|------|-------|------|------|-------|
| කි               | ක    | කා    | කැ   | කෑ    | කි   | කී    | කු   | කූ    | කෙ   | කේ   | කෙ    | කො   | කෝ   | කො    |
| kh               | kha  | khaa  | kha  | khaa  | khi  | khee  | khu  | khoo  | khe  | khee | khai  | kho  | kho  | khau  |
| ඝ                | ඝ    | ඝා    | ඝැ   | ඝෑ    | ඝි   | ඝී    | ඝු   | ඝූ    | ඝෙ   | ඝේ   | ඝෙ    | ඝො   | ඝෝ   | ඝො    |
| gh               | gha  | ghaa  | gha  | ghaa  | ghi  | ghee  | ghu  | ghoo  | ghe  | ghee | ghai  | gho  | gho  | ghau  |
| ඡ                | ඡ    | ඡා    | ඡැ   | ඡෑ    | ඡි   | ඡී    | ඡු   | ඡූ    | ඡෙ   | ඡේ   | ඡෙ    | ඡො   | ඡෝ   | ඡො    |
| chh              | chha | chhaa | chha | chhaa | chhi | chhee | chhu | chhoo | chhe | chhe | chhai | chho | chho | chhau |
| කඩ               | කඩ   | කඩා   | කඩැ  | කඩෑ   | කඩි  | කඩී   | කඩු  | කඩූ   | කඩෙ  | කඩේ  | කඩෙ   | කඩො  | කඩෝ  | කඩො   |
| jh               | jha  | jhaa  | jha  | jhaa  | jhi  | jhee  | jhu  | jhoo  | jhe  | jhe  | jhai  | jho  | jho  | jhau  |
| කඳ               | කඳ   | කඳා   | කඳැ  | කඳෑ   | කඳි  | කඳී   | කඳු  | කඳූ   | කඳෙ  | කඳේ  | කඳෙ   | කඳො  | කඳෝ  | කඳො   |
| kn               | kna  | knaa  | kna  | knaa  | kni  | knee  | knu  | knoo  | kne  | knee | knai  | kno  | kno  | knau  |
| ඳ                | ඳා   | ඳෑ    | ඳැ   | ඳෑ    | ඳි   | ඳී    | ඳු   | ඳූ    | ඳෙ   | ඳේ   | ඳෙ    | ඳො   | ඳෝ   | ඳො    |
| gn               | gna  | gnaa  | gna  | gnaa  | gni  | gnee  | gnu  | gnoo  | gne  | gne  | gnai  | gno  | gno  | gnau  |
| ඵ                | ඵ    | ඵා    | ඵැ   | ඵෑ    | ඵි   | ඵී    | ඵු   | ඵූ    | ඵෙ   | ඵේ   | ඵෙ    | ඵො   | ඵෝ   | ඵො    |
| nj               | nja  | njaa  | nja  | njaa  | nji  | njee  | nju  | njoo  | nje  | nje  | njai  | njo  | njo  | njau  |
| ඨ                | ඨ    | ඨා    | ඨැ   | ඨෑ    | ඨි   | ඨී    | ඨු   | ඨූ    | ඨෙ   | ඨේ   | ඨෙ    | ඨො   | ඨෝ   | ඨො    |
| th               | tha  | thaa  | tha  | thaa  | thi  | thee  | thu  | thoo  | the  | the  | thai  | tho  | tho  | thau  |
| ඪ                | ඪ    | ඪා    | ඪැ   | ඪෑ    | ඪි   | ඪී    | ඪු   | ඪූ    | ඪෙ   | ඪේ   | ඪෙ    | ඪො   | ඪෝ   | ඪො    |
| dh               | dha  | dhaa  | dha  | dhaa  | dhi  | dhee  | dhu  | dhoo  | dhe  | dhe  | dhai  | dho  | dhoo | dhau  |
| ඵ                | ඵ    | ඵා    | ඵැ   | ඵෑ    | ඵි   | ඵී    | ඵු   | ඵූ    | ඵෙ   | ඵේ   | ඵෙ    | ඵො   | ඵෝ   | ඵො    |
| th               | tha  | thaa  | tha  | thaa  | thi  | thee  | thu  | thoo  | the  | the  | thai  | tho  | tho  | thau  |
| ධ                | ධ    | ධා    | ධා   | ධෑ    | ධි   | ධී    | ධු   | ධූ    | ධෙ   | ධේ   | ධෙ    | ධො   | ධෝ   | ධො    |
| dh               | dha  | dhaa  | dha  | dhaa  | dhi  | dhee  | dhu  | dhoo  | dhe  | dhe  | dhai  | dho  | dho  | dhau  |
| න                | න    | නා    | නැ   | නෑ    | නි   | නී    | නු   | නූ    | නෙ   | නේ   | නෙ    | නො   | නෝ   | නො    |
| n                | na   | naa   | na   | naa   | ni   | nee   | nu   | noo   | ne   | ne   | nai   | no   | no   | nau   |
| ඵ                | ඵ    | ඵා    | ඵැ   | ඵෑ    | ඵි   | ඵී    | ඵු   | ඵූ    | ඵෙ   | ඵේ   | ඵෙ    | ඵො   | ඵෝ   | ඵො    |
| ph               | pha  | phaa  | pha  | phaa  | phi  | phee  | phu  | phoo  | phe  | phe  | phai  | pho  | pho  | phau  |
| භ                | භ    | භා    | භැ   | භෑ    | භි   | භී    | භු   | භූ    | භෙ   | භේ   | භෙ    | භො   | භෝ   | භො    |
| bh               | bha  | bhaa  | bha  | bhaa  | bhi  | bhee  | bhu  | bhoo  | bhe  | bhe  | bhai  | bho  | bho  | bhau  |
| ශ                | ශ    | ශා    | ශැ   | ශෑ    | ශි   | ශී    | ශු   | ශූ    | ශෙ   | ශේ   | ශෙ    | ශො   | ශෝ   | ශො    |
| sh               | sha  | shaa  | sha  | shaa  | shi  | shee  | shu  | shoo  | she  | she  | shai  | sho  | sho  | shau  |
| ඡ                | ඡ    | ඡා    | ඡැ   | ඡෑ    | ඡි   | ඡී    | ඡු   | ඡූ    | ඡෙ   | ඡේ   | ඡෙ    | ඡො   | ඡෝ   | ඡො    |
| sh               | sha  | shaa  | sha  | shaa  | shi  | shee  | shu  | shoo  | she  | she  | shai  | sho  | sho  | shau  |

Figure 3.6: Mapping of Sinhala other consonants and proposed Singlish form

### **3.5 Summary**

In Section 3, the focus is on comprehensively analyzing code-mixed text, primarily centered on Sinhala-English (Singlish) and Tamil-English (Tanglish) language blends. The linguistic landscape of Sri Lanka is introduced, acknowledging Sinhala, English, and Tamil as official languages.

The chapter highlights the challenges associated with code-mixed text representation, ranging from spelling errors to the integration of suffixes and phonetic transliteration. The complexities of transliteration are discussed, revealing differences between standard ISO15919 and code-mixing practices. Additionally, the systematic mapping of Sinhala letters to Singlish formats is explored, aiming to standardize the representation and reduce linguistic noise in code-mixed texts.

The chapter further presents survey findings, revealing the prevalence of Sinhala-English code-mixed text usage in online platforms and social media. The motivations behind this usage, including ease of typing and flexibility offered by standard keyboards, shed light on the evolving linguistic practices in the digital age.

Overall, this section offers a comprehensive foundation for understanding the intricacies of code-mixed text analysis, ranging from linguistic challenges to transliteration methods and the implications of such practices in contemporary communication, particularly within the digital realm.

## 4. Methodology

### 4.1 Data collection and validation

Machine translation models require a remarkable number of parallel sentences to achieve a good result. Our research study needed the SECM sentences and the relevant Sinhala sentences. Since Sinhala is considered a low-resource language, no dataset is currently available for SECM-Sinhala. So initially, we used the web scraping method to collect SECM sentences from social media.

Web scraping is used to extract information from websites or web applications. Web crawling is the main feature of web scraping. Web scraping fetches webpages or the necessary text to process later. The information on the web page will be extracted and searched, parsed, and reformatted according to the need of the user. Also, the fetched data can be saved in a spreadsheet or loaded into the database. In our case, we scraped the SECM sentences from the public social media pages using SOAX (SOAX Data Collection Solution, 2022) web scraping API. Five thousand SECM sentences are extracted from social media.

As the next step in creating a parallel corpus, we manually translated the SECM sentences into Sinhala using human translators. These human translators are experts in Sinhala and English language who are Sinhala native speakers. Each translation is translated by one translator and validated using the crowd sourcing approach which is explained in the next paragraph. We provided the translator with the following guidelines.

- The translation should be grammatically correct
- The translation should be natural sounding
- Use the proposed transliteration mapping (Kugathan & Sagara,2020) for the translation

The research study used the Crowd Sourcing method to validate and correct human-translated sentences. This method is used to differentiate the excellent translation from bad ones. Our dataset is split into 15 groups where each reviewer group got approximately 300 sentences for the annotation. Each group had at least two bilingual reviewers, Sinhala native speakers and good in English.

The reviewers were commanded to check for the following factors in the translated Sinhala sentences: has the correct spelling, natural-sounding Sinhala sentence, and should have the correct grammatical structure. The reviewers were instructed to drop each sentence into one of the following categories: Fully Correct(FC) and Change Required(CR). Each record with the SECM source sentence and Sinhala target sentence had two reviewers. If a translation is annotated with both FC tags, we finalize that as the final translation.

As shown in Figure 4.1, if a translation has been annotated with one FC tag and another with CR, the reviewer who has provided the CR tag should also provide the alternative translation that he/she thinks is correct. If a translation has been labelled as CR by both the reviewers, both the reviewers should provide separate alternative translations.

After the annotation process, we filter out the translation which has been tagged with the CR tag separately and provides those translations with the proposed new translations to another human translator. The translator decides whether to consider the alternative translation or not and, if it changes what the finalized translation is.

GITHUB link for the dataset : [https://github.com/ArchchanaKugathan/SECM-to-Sinhala\\_translator-Project.git](https://github.com/ArchchanaKugathan/SECM-to-Sinhala_translator-Project.git)

| Singlish Sentence                                       | Sinhala Sentence translated by Human Translator | A1 | A2 | Alternate translation by A1 | Alternate translation by A2         | Finalized translations by A3        |
|---|---|----|----|-----------------------------|-------------------------------------|-------------------------------------|
| place eka piliwelai, clean, kama godaak rasai           | ස්ථානය පිළිවෙලට, පිරිසිදුයි, කෑම හරිම රසයි      | FC | FC | N/A                         | N/A                                 | N/A                                 |
| gaana wadi  | ගාන වැඩියි                                      | FC | FC | N/A                         | N/A                                 | N/A                                 |
| Price ekata shape wenna hoda rasata kama tika hambenawa | මිලට හරියන්න රසවත් කෑම හම්බෙනවා                 | FC | CR | N/A                         | මිලට හරියන්න හොඳ රසවත් කෑම හම්බෙනවා | මිලට හරියන්න හොඳ රසවත් කෑම හම්බෙනවා |
| kama patta, price is also reasonable                    | කෑම පව්ට, මිලද සාධාරණයි                         | CR | CR | කෑම හොඳයි, ගාන සාධාරණයි     | කෑම හොඳයි, මිලද සාධාරණයි            | කෑම හොඳයි, මිලද සාධාරණයි            |
| Service eka hodai.                                      | සේවාව හොඳයි                                     | FC | FC | N/A                         | N/A                                 | N/A                                 |
| Kama rasai specially drinks quantity ekat ok.           | කෑම රසයි, විශේෂයෙන් බීම, ප්‍රමාණය ප්‍රමාණවත්    | FC | FC | N/A                         | N/A                                 | N/A                                 |

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.

After the corrections to the parallel sentences, we randomly choose 100 sentences from the dataset to estimate the quality of the translation using Fleiss' Kappa method.

Fleiss kappa is a measure used when allocating categorical ratings to a number of items which is classified. This equation is applied to evaluate the reliability of agreement between the judgments from different people. In Cohen's Kappa, we can have only two raters, but when assessing the agreement, we can have more than two raters in Fleiss kappa. The equation calculates the degree of agreement in classification over that which would be expected by chance.

In Fleiss' Kappa  $n \Rightarrow$  number of subjects,  $m \Rightarrow$  the number of raters for each subject and  $k \Rightarrow$  number of evaluation categories. In some cases, there can be  $m$  number of raters, but not every rater needs to judge each subject. The significant task is that each subject is judged  $m$  times.



In Cohen's Kappa, for every subject  $i = 1, 2, \dots, n$  and evaluation categories  $j = 1, 2, \dots, k$ , let  $x_{ij}$  = the number of raters that assign category  $j$  to subject  $i$ . Thus,

$$0 \leq x_{ij} \leq m \quad \sum_{j=1}^k x_{ij} = m \quad \sum_{i=1}^n \sum_{j=1}^k x_{ij} = mn \quad (06)$$

The ratio of pairs of raters that agree in their evaluation on the subject  $i$  is given by the following equation.

$$p_i = \frac{\sum_{j=1}^k C(x_{ij}, 2)}{C(m, 2)} = \frac{\sum_{j=1}^k x_{ij}(x_{ij} - 1)}{m(m-1)} = \frac{\sum_{j=1}^k x_{ij}^2 - \sum_{j=1}^k x_{ij}}{m(m-1)} = \frac{\sum_{j=1}^k x_{ij}^2 - m}{m(m-1)} \quad (07)$$

Therefore the mean of the  $p_i$  is,

$$p_\alpha = \bar{p} = \frac{1}{n} \sum_{i=1}^n p_i = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^k x_{ij}^2 - m}{m(m-1)} = \frac{1}{mn(m-1)} \left[ \sum_{i=1}^n \sum_{j=1}^k x_{ij}^2 - mn \right] \quad (08)$$

The error term is defined as,

$$p_\varepsilon = \sum_{j=1}^k q_j^2 \quad (09)$$

Where,

$$q_j = \frac{1}{nm} \sum_{i=1}^n x_{ij} \quad (10)$$

So the Fleiss' Kappa is defined to be,

$$\kappa = \frac{p_a - p_s}{1 - p_s} \quad (11)$$

Kappa is defined for the  $j^{\text{th}}$  category by,

$$\kappa_j = 1 - \frac{\sum_{i=1}^n x_{ij}(m - x_{ij})}{mn(m - 1)q_j(1 - q_j)} \quad (12)$$

The formula used to calculate the standard error for  $K_j$ ,

$$s.e.(\kappa_j) = \sqrt{\frac{2}{mn(m - 1)}} \quad (13)$$

The formula used to calculate the standard error for  $K$ ,

$$s.e. = s.e.(\kappa_j) \cdot \frac{\sqrt{[\sum_{j=1}^k q_j(1 - q_j)]^2 - \sum_{j=1}^k q_j(1 - q_j)(1 - 2q_j)}}{\sum_{j=1}^k q_j(1 - q_j)} \quad (14)$$

In our studies, the agreement between the raters are used to measure the quality of the translation by obtaining Fleiss' Kappa Score. Linguistically experts in both the Sinhala and English languages are hired for the process. We obtained the judgment for each of the translations from 3 different linguists. The linguists were asked to rank the translation as 'Good' or 'Bad' according to the following criteria: Spelling errors, meaningful translation, and grammatical structure. Each translated record had 3 rating tags from 3 different raters, which belong to one of the two categories, as shown in Figure 4.2. The overall Fleiss' Kappa score we obtained for the agreement of good translation is 0.88.

| Singlish Sentence  | Sinhala Sentence translated by Human Translator                          | Rater 1 | Rater 2 | Rater 3 |
|--|--|---------|---------|---------|
| Weekday dawas wala tikak crowded. Parking adui.                          | සතියේ දවස් වල ජනාකීර්ණයි, වාහන නැවැත්වීමේ ස්ථාන අඩුයි                    | Good    | Good    | Bad     |
| Kama rasai. Service eka hodai  | කැම රසයි සේවාව නොදයි   | Good    | Good    | Good    |
| Gana hondai.   | ගාන නොදයි .  | Good    | Good    | Good    |
| Kama rasai..view eka lssnai  | කැම රසයි,දර්ශනය ලස්සනයි  | Good    | Good    | Good    |
| Mama kamatima kama tyana restaurant eka                                  | මම කැමතිම කැම තියන අවන්හල  | Good    | Good    | Good    |
| karann puluvan etena..godaak comfortable environment ekak etena tiyenva. | කැම ගොඩක් රසයි, අසට එහි බොහෝ කාලයක් ගත කළ හැකිය, ඉතා සුවසහසු සර්සරයක් ඇත | Good    | Good    | Good    |
| price eke tikak vadi.  | මිල වඩාත් වැඩියි   | Good    | Good    | Good    |
| Gana tikak wadi  | ගාන වඩාත් වැඩි   | Good    | Good    | Good    |
| Gewana gaanata paadu naha. kama rasai.                                   | ගෙවන ගානට සාමු නැහැ, කැම රසයි  | Good    | Good    | Good    |
| kama patta, price is also reasonable                                     | කැම නොදයි , මිලද සාධාරණයි  | Good    | Good    | Good    |
| hondai.Harima friendly.  | නිදහසේ ඉන්න පුළුවන්,සේවාව නොදයි,බොහෝ මිත්රශීලී                           | Good    | Bad     | Good    |
| lassana than godak thiyenawa balanna                                     | ලස්සන කැන් ගොඩක් තියෙනවා බලන්න   | Good    | Good    | Good    |
| Gampolin yannath puluwan lissanai.                                       | ගම්පොලින් යන්නත් පුළුවන් ලස්සනයි   | Good    | Good    | Good    |
| wade hodata karala deyi  | වැඩේ නොදට කරලා දෙයි  | Good    | Good    | Good    |
| supiri wadek thama akke  | සුපිරි වැඩිවෙත් තමා අක්කේ  | Good    | Good    | Good    |
| qualityata wade karala dennava   | ගුණාත්මක වැඩේ කරලා දෙනවා   | Good    | Good    | Good    |
| aulk na  | අවුලක් නෑ  | Good    | Good    | Good    |
| bari wei neee  | බැරි වෙයි නේ   | Good    | Good    | Good    |
| supiri ma sub thama  | සුපිරිම සබ් තමයි   | Good    | Good    | Good    |
| sounds quality patta..   | ශබ්දයේ ගුණාත්මකභාවය ඉතා නොදයි  | Bad     | Bad     | Good    |
| awlkma ne  | අවුලක්ම නෑ   | Good    | Good    | Good    |
| Sonicgear titan series eka patta   | සොනික්ගියර් වයිටන් ටීටන් එක නොදයි  | Good    | Bad     | Bad     |
| Crowd eka wadi. Service eka not good.                                    | සෙනඟ වැඩි ,සේවාව නොද නෑ  | Good    | Good    | Good    |
| ambience eka godaak hodai. Kama awulak na.                               | සර්සරය ගොඩාක් නොදයි, කැම අවුලක් නෑ                                       | Good    | Good    | Good    |
| Kama rasai   | කැම රසයි   | Good    | Good    | Good    |
| Kama hari ne   | කැම හරි නෑ   | Good    | Good    | Good    |
| also godaak Honda quality ekata hadala serve karanawa.                   | සේවාවේ ගුණාත්මකභාවය සුපිරි වන අතර ආහාර ඉතා නොද තත්වයේ සේවය කරනු ලැබේ     | Bad     | Good    | Bad     |
| Service eka hodai, kama rasai  | සේවාව නොදයි, කැම රසයි  | Good    | Good    | Good    |

Figure 4.2: Datasheet for Fleiss Kappa analysis

## 4.2 Text normalization

### 4.2.1 Spelling Error detection and correction

Spell checking is an essential task in our research study to reduce the unnecessary noise in the dataset. For example, the word ‘difference’ could be misspelt as ‘defference’, ‘dfference’, ‘diference’, etc. Spelling mistakes can occur due to many reasons, such as fast typing, carelessness, etc. But the main reason spelling errors occur in SECM text is that English is not the native language, and most SECM text users are not fluent in English.

So as the first initiative for the normalization in our research project, the Out Of Vocabulary(OOV) words are identified and corrected using a dictionary-based approach. Birbeck spelling error corpus. This corpus contains 6136 words, and the most frequent misspellings of those words are gathered from various sources. 36,133 misspelt words are listed in the corpus. Applying the dictionary-based approach to our SECM corpus, we corrected the spelling error and reduced the noise raised from the misspelt words in the dataset before training.

#### **4.2.2 Slang word normalization**

After the normalization of Spelling of English words, next, we focused on removing the noise created by slang words used in SECM text. Slang is described as a type of language which is too informal to be used in specific situations. Also, in some situations slang is described as a language that generally belongs to the member of a particular community in order to establish the identity of that specific community or exclude the outsider. Linguists agree that slang is a repeatedly changing linguistic occurrence. Also, some linguists argue that slang words are created as a way to define new experiences that are dominating the time.

We used the SlangNorm dictionary-based approach for Slang word normalization. The dictionary contains 5427 slang words. The dictionary includes words such as ‘3wheel’, ‘2mrw’ which would be replaced with the correct words ‘tomorrow’, ‘three wheeler’. Converting the slang into common words would reduce the unnecessary noise.

### 4.2.3 Transliteration normalization

In Sinhala-English code-mixed sentences, the same Sinhala word would be presented in many different representations, which increases the noise of the input dataset. We used the Levenshtein Edit Distance approach to normalize the transliterated words in each sentence of the corpus.

Levenshtein Edit Distance is an estimate of similarity between two strings. The least number of changes that needed to be performed to convert a string  $x$  to string  $y$  is measured by the Levenshtein Edit Distance. The number of changes can be performed by inserting, replacing, or deleting a character string from string  $x$ . If the Levenshtein Edit Distance value is smaller, it means the strings are more similar and if the value is higher, the strings are less similar.

If we consider the following example,

String  $x$  - > 'mitten'

String  $y$  - > 'fitting'

To convert the string 'mitten' into the string 'fitting' at least 3 edits are required.

1. Mitten  $\rightarrow$  Fitten (Substitution of 'M' for 'F')
2. Fitten  $\rightarrow$  Fittin (Substitution of 'i' for 'e')
3. Fittin  $\rightarrow$  Fitting (Insertion of 'g' at the end)

In this measure, 'Edit' is outlined by either insertion of a character, replacement of a character or deletion of a character.

The following is the Levenshtein Distance equation.

$$\text{lev}_{a,b}(i, j) = \begin{cases} \max(i, j) & \text{if } \min(i, j) = 0, \\ \min \begin{cases} \text{lev}_{a,b}(i-1, j) + 1 \\ \text{lev}_{a,b}(i, j-1) + 1 \\ \text{lev}_{a,b}(i-1, j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

$a \rightarrow$  string 1

$b \rightarrow$  string 2

$i \rightarrow$  the terminal character position of string 1

$j \rightarrow$  the terminal character position of string 2

In this research, Levenshtein Edit Distance approach is used to normalize the transliterations by substituting the high-frequency words with the corresponding low-frequency words based on the edit distance. A dictionary with a frequency list of the words in the corpus is maintained. So the most frequently used transliterated form of a Sinhala word would replace the other transliterated words which have minimum edit distance with the most frequently used word form for a particular word. Transliteration normalization is considered the last step of the normalization process of SECM text.

#### 4.3 Sequence to Sequence model(Encoder-Decoder framework)

Google introduces this model in 2014 with a goal of mapping a fixed length input sequence with a fixed-length output sequence even though the input and output lengths are different. For example, “Did you eat?” in English has three words as input and its output sentence in Sinhala “ඔයා කෑවද?” has two words. Regular LSTM cannot be used to map word by word when it comes to translation. LSTM is chosen because it doesn’t have information decay and the vanishing gradient problem like RNN (Van Houdt et al., 2020). Using LSTM as the basic structure with the encoder-decoder framework, we fabricate a Seq2Seq model, as shown in Figure 4.3. In this approach sequence of a source,

sentence is matched with the sequence of the target sentence(Sutskever et al.,2014). In the machine translation system source sequence would be the input and the target sequence would be the output.

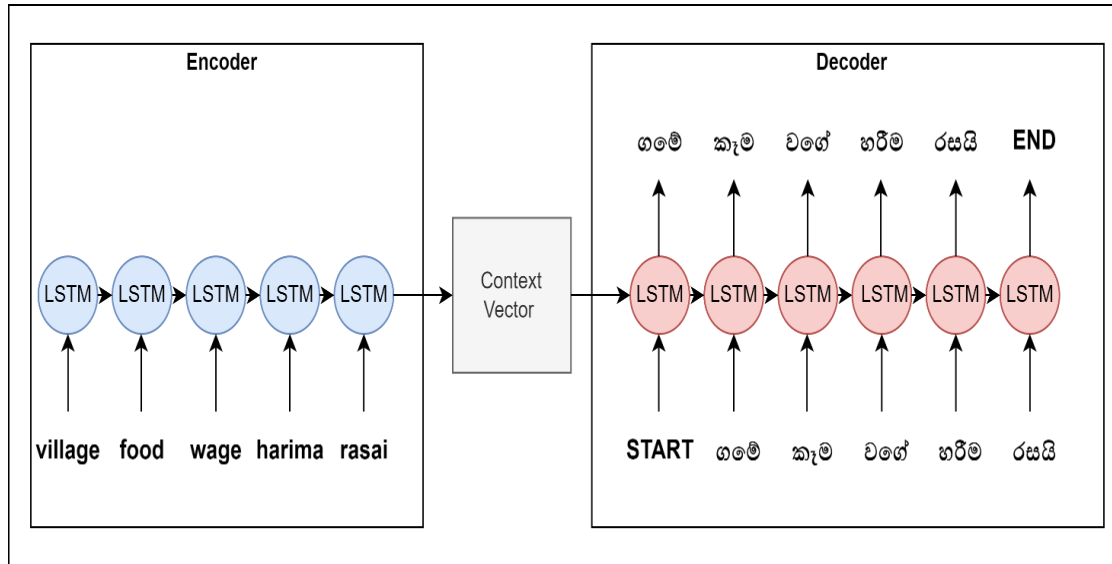


Figure 4.3: Sequence to Sequence model

Source language is read and used as the input to the encoder. A context vector which can also be called a hidden state, is created with the encoder by encoding the input data into a real-valued vector. Word by word encoder reads the input sequence. Meaning of the input sequence encoded into a single vector. Also, in each timestep LSTM units will be processing. The outputs gained from the encoder are discarded and only the hidden states have proceeded as the inputs to the decoder. One timestep from the encoder is shown in Figure 4.4.

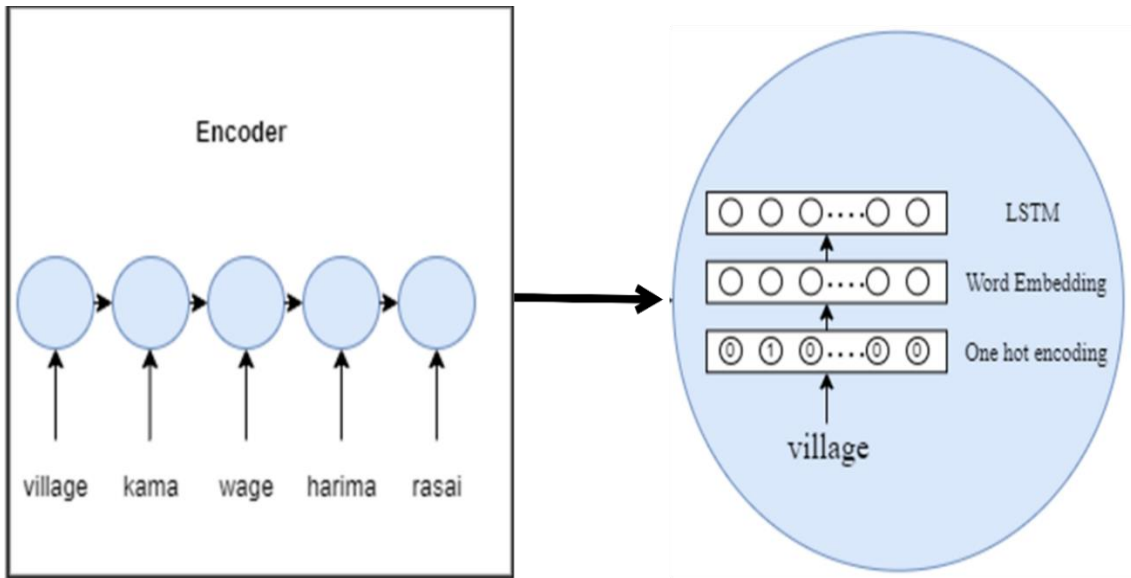


Figure 4.4: One timestep in the encoder

The decoder takes the hidden state and the *START* string as the input. Output produced by the decoder is read word by word during decoding. The teacher Forcing mechanism is used in our training part of the decoder. Section 4.6 describes the Teacher Forcing mechanism.

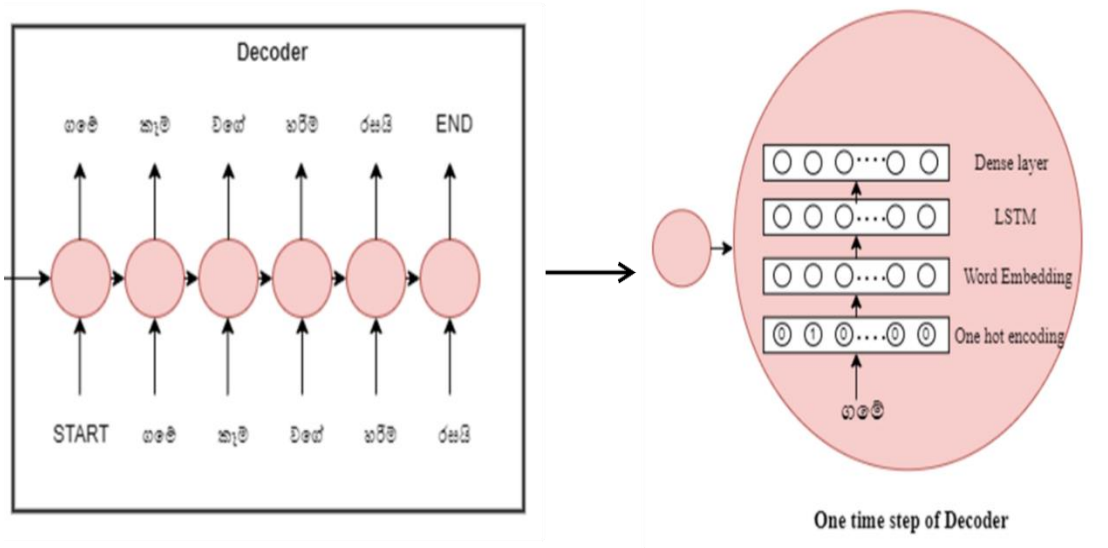


Figure 4.5: One timestep in the decoder



## 4.4 Long Short Term Memory

Recurrent Neural Networks(RNN) was the most famous approach used when it comes to Neural Machine Translation. But unfortunately, RNN has the problem of short-term memory. If we have a long sentence, RNN faces difficulty carrying the information forward through the time steps. So, when we process a lengthy sentence, RNN might leave out the important part of the sentence in the prediction. Also, when it back propagates, RNN faces the vanishing gradient issue. Gradients values are used to update the weights in the Neural Network. If the gradients get smaller values, when backpropagation happens, the gradient values will get smaller and smaller as the gradients would be multiplied in each step. If a gradient gets a minimal weight, it would not much contribute to the learning. This is called the vanishing gradient issue. The contribution of the gradient value for the learning will be stopped if the gradient value is minimal.

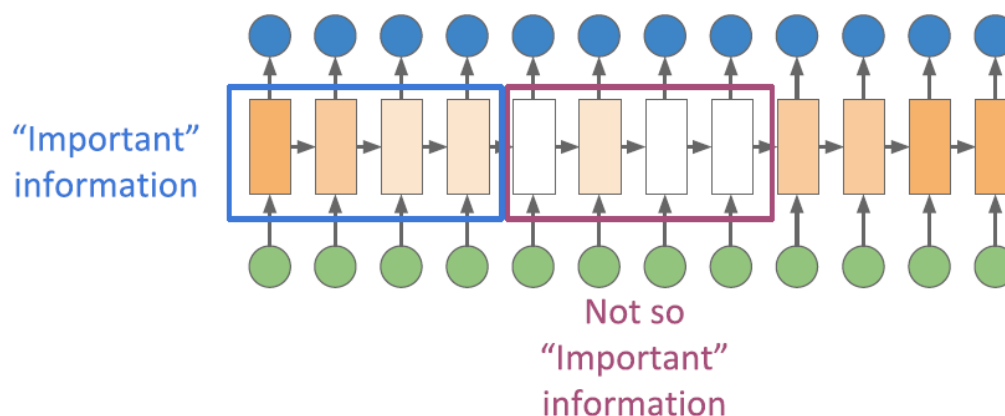


Figure 4.6 : Cell state memory maintenance

LSTM is the found solution for the vanishing gradient issue in a neural network. Gates are an internal mechanism of LSTM which regulates the flow of information. Gates studies which data sequence is the most important and which data to save in the memory or discard. By performing this task, LSTM units pass only the necessary information through the network to make predictions. LSTM and RNN have a similar flow control

where the passing of data propagates forward. The only difference is the internal operations inside each LSTM unit. Figure 4.6 shows the identification of the important information and the non-important information.

The main theme of LSTM is the cell state and the various gates. Cell states are used as memory storage which transfers only the vital information throughout the entire sequence. This clears the path where important details from earlier time steps could be passed to the later time steps reducing the vanishing gradient and short-term memory problem.

The cell state is maintained throughout the entire training, the critical information would be added, and the unnecessary details would be removed in the cell state using the gates. Gates are a neural network that decides what information should be stored in the cell state and what information should be forgotten in the training phase. The following example in Figure 4.7 explains how the cell states are maintained for a given input sentence.

In a sequence-to-sequence model, we pass the hidden state from the previous time step to the current time step. LSTM cell has three main gates: the input gate, forget gate, and output gate and the activation functions sigmoid and tanh.

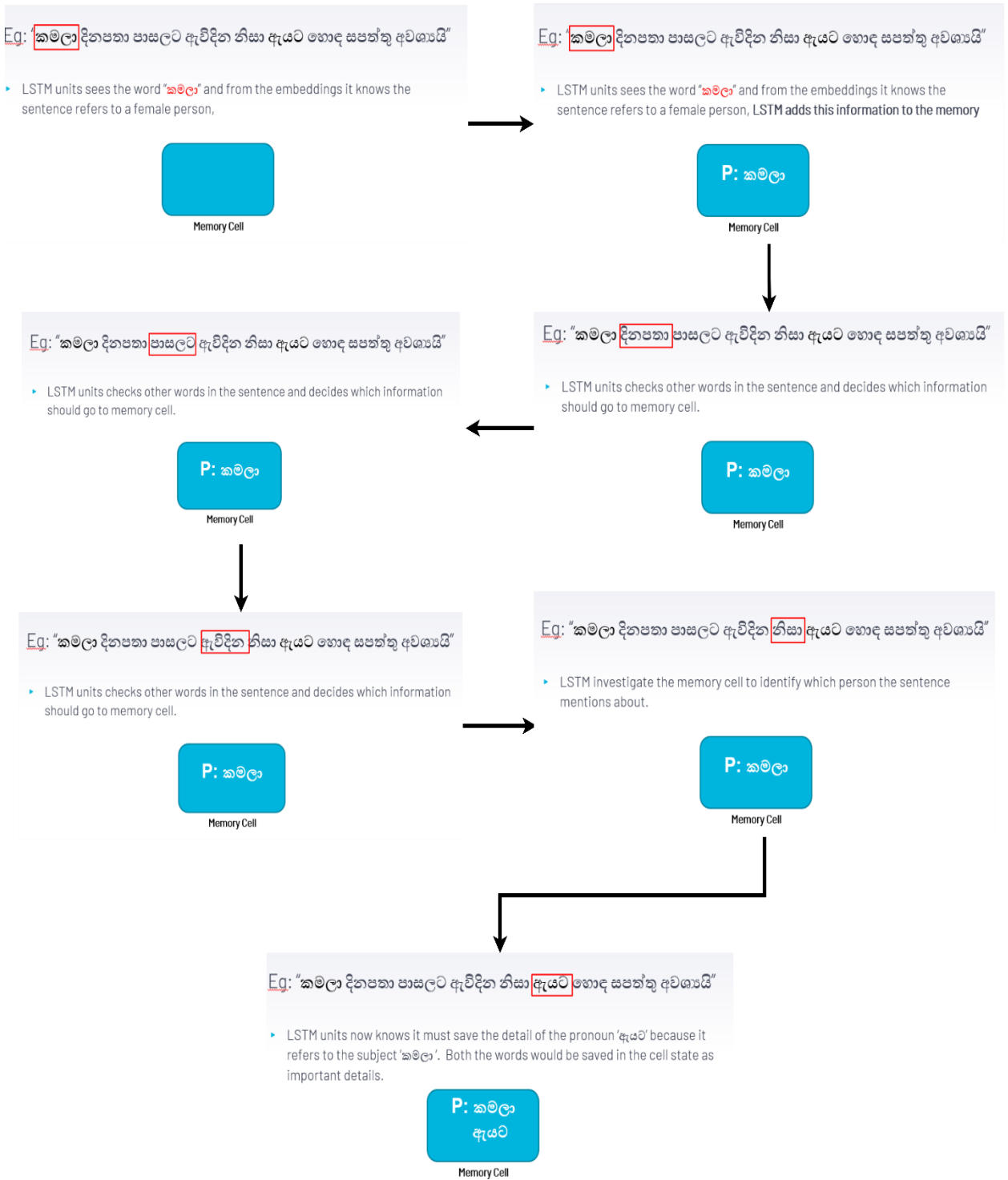


Figure 4.7 : Example of how the cell state is maintained through out a sentence.

- **Tanh**

Tanh activation function regulates the flowing values in the network, as shown in Figure 4.8. It converts the input values into a value between -1 to 1. In a neural network, whenever vectors flow through it, it would be transformed due to many mathematical operations. Some values will be exploded and become astronomical, making other values seem insignificant. The output value from tanh would stay in between the boundaries of -1 to 1.

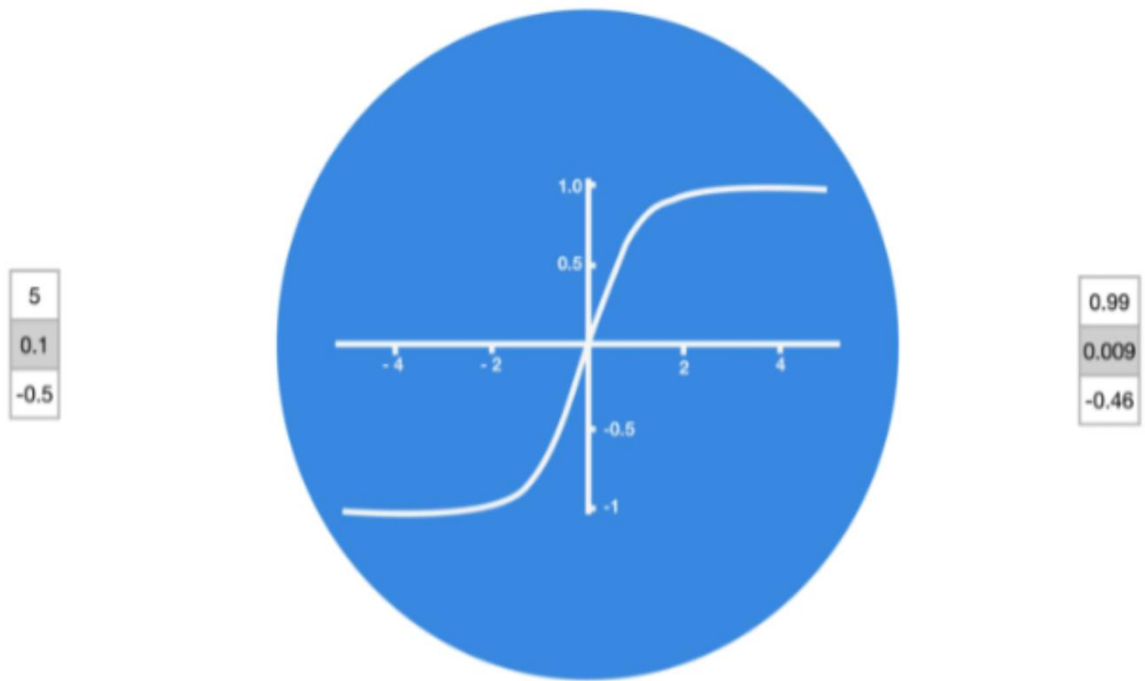


Figure 4.8: Tanh squishes values to be between -1 and 1

- **Sigmoid**

Gates have sigmoid activation functions, as shown in Figure 4.9. It is similar to tanh, the only difference would be the output values of the sigmoid function would be converted between the range of 0 to 1.

If an input value is multiplied by 0, the output value is 0, which makes the gates forget that information. However, if the input value is multiplied by 1, it gives the same value. Therefore, that information would be saved in the cell state.

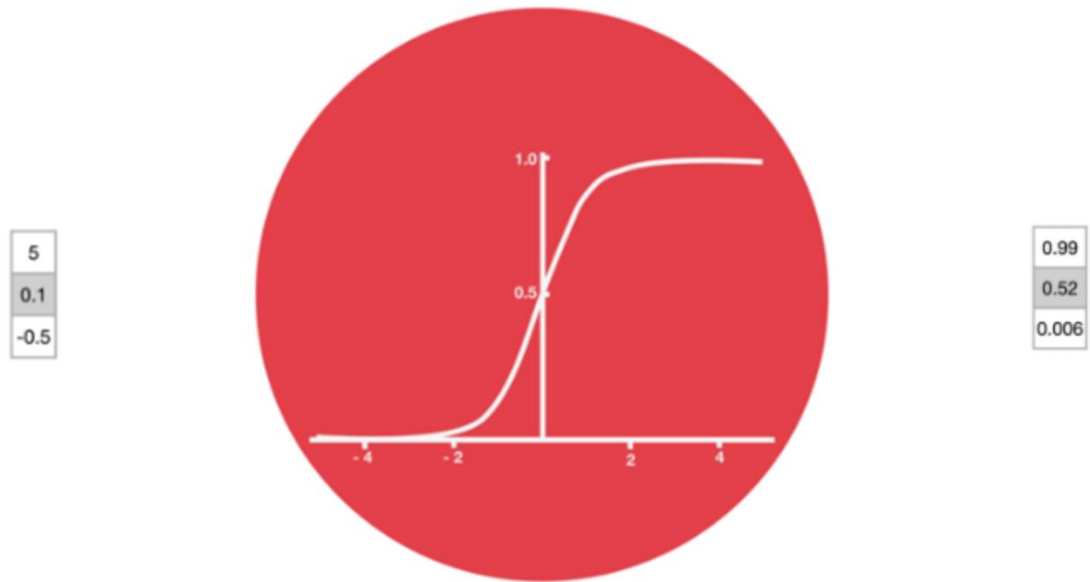


Figure 4.9: Sigmoid squishes values to be between 0 and 1

- **Forget gate**

Forget gates determine which information should be saved and which needs to be discarded. For example, information from previous hidden states passed from the previous time step and current input would be the input to the sigmoid function. The output values

would be in between the values 0 to 1. If the output value is closer to 0, then the information would be forgotten, and if the value is closer to 1, the information would be kept.

- **Input gate**

Input gates play a part of updating the cell state. The previous hidden state from the previous time step and the current input would be inputted into the sigmoid function, which decides what to keep and discard based on the output values. The hidden state and the current inputs are also inputted into the tanh function. It regulates the values between -1 to 1. Output values from tanh and sigmoid functions would be multiplied. Sigmoid will decide which information should be saved from the tanh output.

- **Cell state**

Initially, the cells state receives the pointwise multiplied from the forget gate. There is a chance of dropping values in the cell state if the value gets multiplied by values closer to 0. Next, the output from the input gate would be taken, and pointwise addition would be applied to that value, which updates the new values in the cell state. A new cell state will be created.

- **Output gate**

The final gate is the output gate which decides on what should be the next hidden state to the next timestep. Hidden states contain details from previous input, which would also be used for the next word prediction. Initially, the previous hidden state and the current input are passed into a sigmoid function. Then the newly updated cell state would be passed through tanh function. Finally, the output from sigmoid and tanh would be multiplied, and the output value will decide which information in the hidden state should be passed to the next timestep. The updated hidden state and the new cell state would be passed to the next time step. Figure 4.10 shows a detailed image of an LSTM unit.

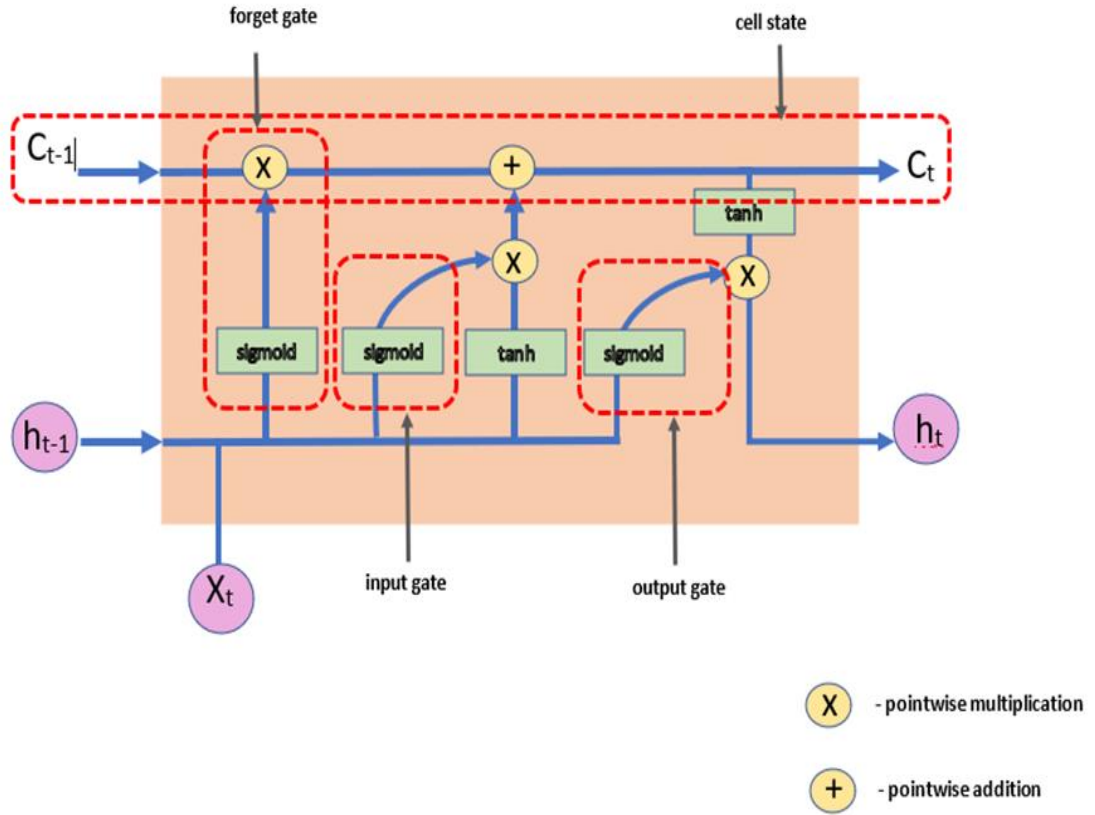


Figure 4.10: Detailed image of one LSTM unit

#### 4.5 Teacher Forcing Algorithm

Teacher forcing is a method for quickly and efficiently training recurrent neural network models that use the ground truth from a prior time step as input. In the example shown in Figure 21, if we take the left side image A, considering the timestep  $t$ , the timestep  $t$  predicts the word ‘అజ్’ as the next word in the sequence, and the predicted word will be fed as the input to  $t+1$  timestep. Likewise, the predicted word from the previous timestep will be fed as input to the next timestep. The timestep  $t$  is feeding a false prediction to the next time step  $t+1$ . If we take the whole sentence ‘అమ్మ అజ్ రజిడి’, it is not a meaningful

sentence. When we feed the wrong prediction as input to the next time step, the error gets higher and it reduces the accuracy.

In Figure 4.11, if we consider the right side image B, we use Teacher Forcing algorithm considering the timestep  $t$ , the timestep  $t$  predicts the next word as 'අස්'. But the teacher forcing algorithm does not let the network feed 'අස්' as the input to the next time step  $t+1$ . According to Teacher Forcing, the ground truth will be fed as input for each time step. The actual word that should be fed to  $t+1$  timestep is the word 'කැම'. The teacher Forcing mechanism will feed the word 'කැම', which is the ground truth, as the input to  $t+1$  timestep. Likewise, all the timesteps will be fed with the ground truth instead of the wrong predictions. Training the decoder with Teacher Forcing Algorithm will lead the model to increase its accuracy and avoid false predictions.

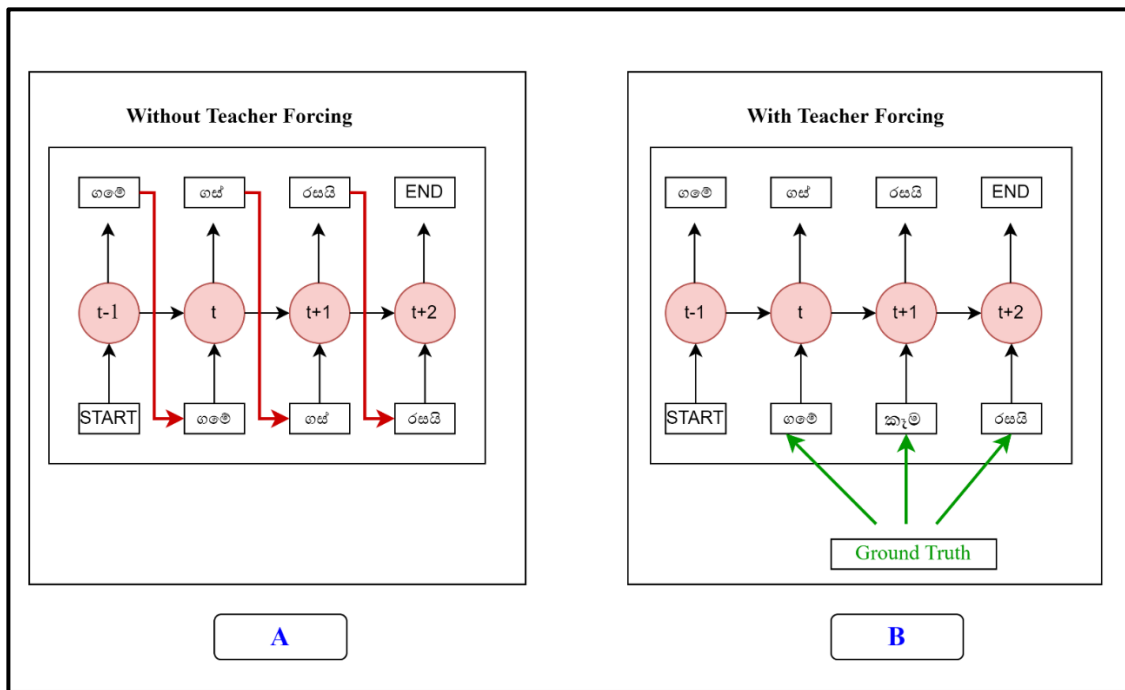


Figure 4.11 : Comparison of Teachers Forcing Vs Non-Teachers Forcing



## 4.6 Experimental Setup & Implementation of the model

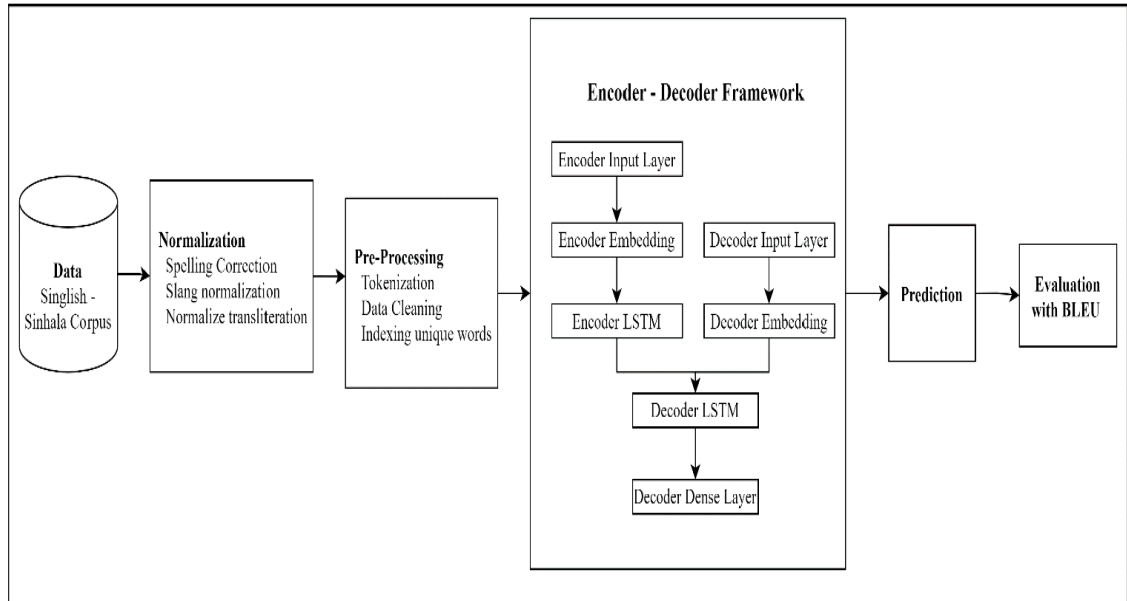


Figure 4.12 : Architecture diagram of the proposed the model

Initially, the dataset will be processed through the normalization model which we explained in Section 4.2. After the data normalization, the dataset is cleaned by converting it to lowercase, removing special characters, removing quotes, and removing unnecessary spaces. Target sentences are added with the 'START' token in the beginning of the target sentence and 'END' token is added at the completion of the target sentence. These tokens are to help the model recognize when to begin translation and end the translation. From the dataset, we identified the distinctive words in both source and target data and filtered them out. A unique number is allocated to each distinctive word identified. Separate dictionaries of word to index and vice versa is created for all distinctive words identified in source data and target data. We shuffle the data before training to lower the variance, to make sure the model overfits less and the model is more vigorous.

We allocate 70% of the dataset for training and 30% for testing. One-hot encoded data is created to train the Seq2Seq model. Encoder and decoder inputs are in the shape of a 2D array. Encoder 2D array has batch sizes of 10, maximum source sentence length is 27, and the shape of encoder input will be (10,27).

Decoder 2D array has batch sizes of 10, a maximum source sentence length of 26, and the shape of encoder input is (10,26). Decoder outputs are in the shape of 3D array with a batch size of 10, a maximum target sentence length of 26, and the number of distinct words in the target data 2233. Teachers forcing algorithm is applied in the decoder section of the sequence-to-sequence model for training. We configured the fundamental parameters like the number of training and validation samples, batch size of training data, number of epochs, and the latent dimension of the coding space of the Seq2Seq model.

The encoder and the decoder are applied with LSTM units. Input sentence from the source language is encoded using the encoder. The primary hidden layer of the encoder is the embedding layer. Large scattered vectors are transformed into a dense dimensional space in the embedding layer. Semantic relationships will be conserved even though the transformation happens. The vocabulary size and dimension of the dense embedding are the parameters passed for the embedding process. Return state of LSTM layer is set to 'true' as we want the states to be passed to the decoder. Only the actual target sequence with a hidden state and the cell state from the encoder is passed as input to the decoder. We repudiate the encoder outputs. The decoder also has embedding as its primary hidden layer. LSTM layer returns internal states and output sequence. Output sequence will be used in the training stage, and internal states are used only in the prediction phase. States passed from the encoder and the outputs given by the embedding layer in the decoder will be taken as the input of LSTM in the decoder. The dense layer is applied with the Softmax activation and decoder outputs are generated. Seq2Seq model grabs encoder and decoder input to produce decoder outputs.

The model is compiled with rmsprop optimizer. Categorical Cross Entropy is used to calculate the loss since one-hot encoded vector are created from categorical labels in our model. The GenerateBatch function is used to generate the data sets. Weights obtained from training are cached for prediction purposes.

A prediction phase in the model is built to view the translation of Singlish text to Sinhala. An unknown input sequence will be decoded to predict the output. The encoder in the prediction model encodes the input sequence into the cell and hidden states of LSTM. The decoder in the prediction model takes inputs from the encoder, such as hidden and cell states with START tag. Each timestamp in the decoder except the first timestamp is fed with the output of the previous time stamp. Decoder produces a one-hot encoded vector at each timestamp. In each timestamp, target words get appended and it repeats until it hits the word limit. The prediction phase is applied to the testing dataset.

The model is implemented using python programming language using anaconda IDE. The architecture diagram of our proposed model is shown in Figure 4.12.

#### **4.7 Implementation of the web application of SECM to Sinhala translator**

The model implementation has been performed using python with anaconda IDE. We exported the trained model and the inference model and created a web application to demonstrate the Sinhala-English code-mixed text to Sinhala translation. The website takes the SECM sentence as the input, and use our trained model in the back end to provide the translated Sinhala sentence as the output. We used python, node.js, and Flask framework. Flask is a framework written in Python. The Flask framework does not require particular tools or libraries. There is no database abstraction layer, form validation, or any other components in a flask because pre-existing third-party libraries provide common functions which make the implementation using the framework easy. Figure 4.13 shows the user interface of the SECM to Sinhala translator web app.

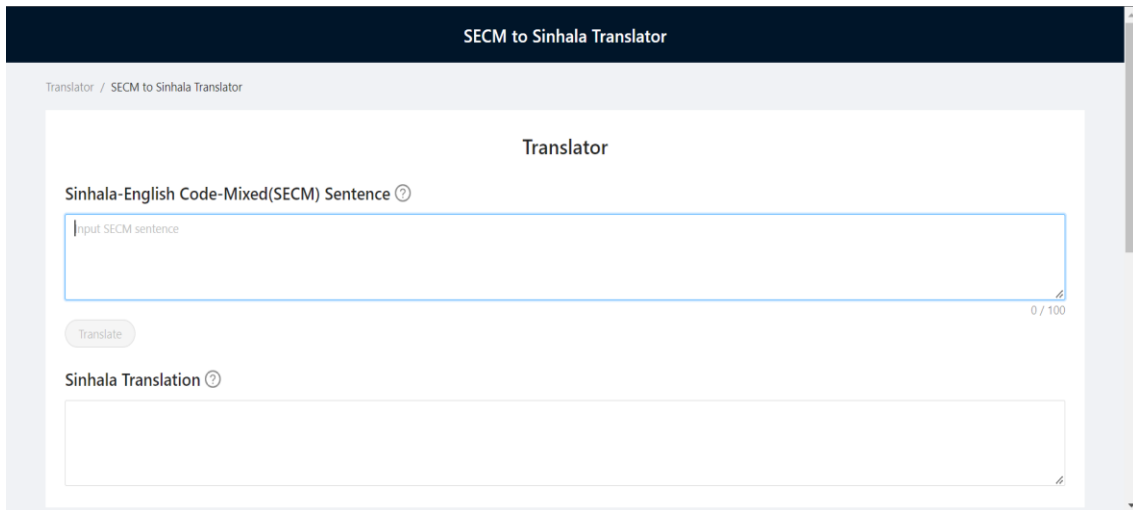


Figure 4.13: SECM to Sinhala translator web application

The web application also contains a BLEU score calculator as shown in Figure 4.14, where the predicted translation would be automatically loaded, and once we input the reference translation it calculates the BLEU score for the specific translated sentence.

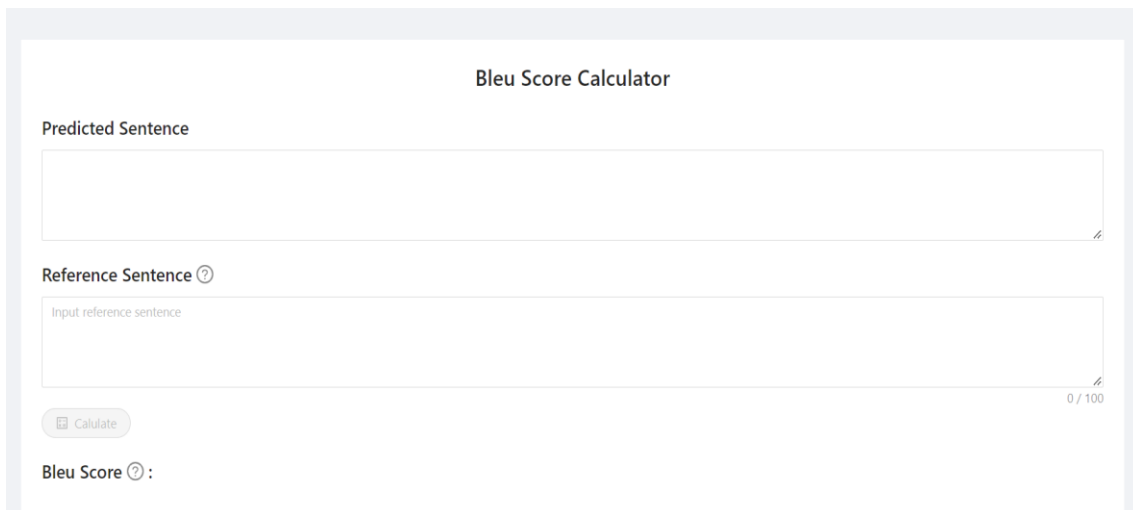


Figure 4.14: BLEU score calculator of SECM to Sinhala translator web application

## 4.8 Summary

In section 4, the methodology employed for the research is detailed, encompassing the collection, processing, and analysis of data for the code-mixed text translation model. It acknowledges the requirement for a substantial parallel dataset for effective training of machine translation models. Due to the scarcity of SECM-Sinhala datasets, web scraping was utilized to gather SECM sentences from social media sources. This involved web crawling, fetching webpages, and extracting approximately 5000 SECM sentences from public social media pages. The subsequent step involved manual translation of these sentences by human experts proficient in both Sinhala and English. The translations were then validated using a crowd sourcing approach, where survey results and raters' opinions were vital for validating translation quality.

Text normalization is addressed as a key factor to reduce noise and enhance dataset quality. Various aspects of normalization, including spelling error detection and correction, slang word normalization, and transliteration normalization, were tackled. Techniques such as dictionary-based correction, spelling error corpora utilization, and Levenshtein Edit Distance were applied to improve data quality.

The core of the research is the Sequence to Sequence (Seq2Seq) model, designed to handle variable input and output sequence lengths. This model involves an encoder-decoder architecture, with an encoder processing source language input to generate a context vector, and a decoder generating target language output based on the context vector. The approach employs Long Short Term Memory (LSTM) units to mitigate the vanishing gradient problem, enabling better information flow across sequences.

The Teacher Forcing algorithm is introduced as a crucial training technique for Seq2Seq models. It addresses the challenge of inaccurate prediction propagation by ensuring that the ground truth is used as input at each decoder time step. This mechanism enhances training efficiency and accuracy, ultimately leading to improved model predictions.

Practical implementation details of the model are provided, covering aspects such as data preprocessing, data split for training and testing, one-hot encoding, embedding layers, and model compilation using optimization techniques like rmsprop. The architecture is built around LSTM units to create a robust framework capable of effectively handling the translation task.

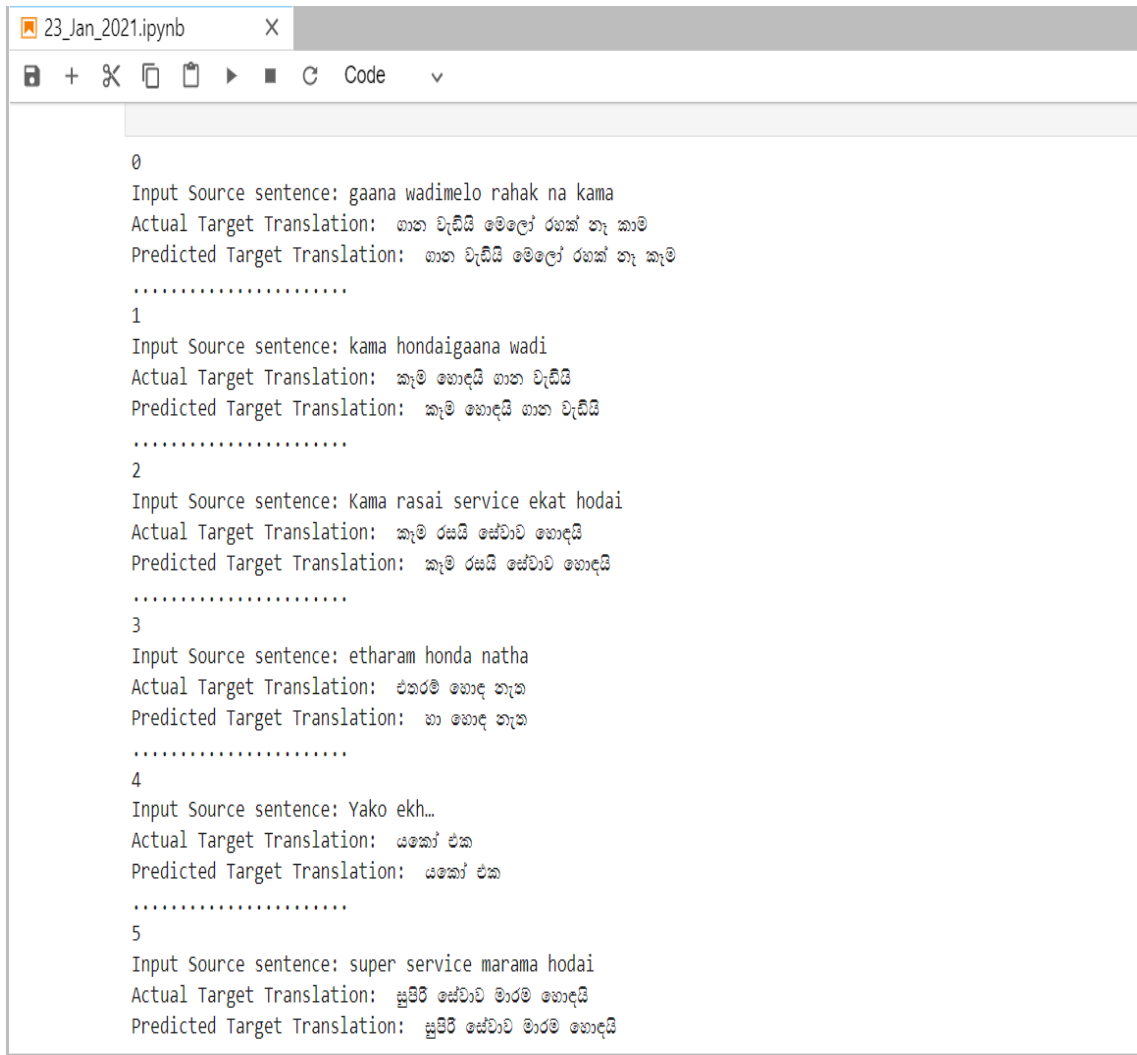
The trained model is practically applied in a web application, utilizing Python, Flask framework, and node.js to develop a user-friendly interface. Users input SECM sentences, and the application leverages the trained model to generate translated Sinhala sentences as output.

In summary, Methodology section outlines a comprehensive methodology that spans data collection and validation, text normalization, model architecture, and practical application. Each step is carefully designed to ensure the accuracy and efficacy of the translation process for code-mixed text.

## 5. Performance Evaluation

### 5.1 Prediction from the model

We exported the trained model to predict the output. So 100 SECM sentences are randomly selected from the corpus and applied to our model as the input and predicted the translation of the Sinhala sentence as the output.



```
23_Jan_2021.ipynb X
+ % 🔍 📄 ▶ ■ 🔄 Code v

0
Input Source sentence: gaana wadimelo rahak na kama
Actual Target Translation: ගාන වැඩිසි මෙලෝ රහක් නෑ කාමි
Predicted Target Translation: ගාන වැඩිසි මෙලෝ රහක් නෑ කෑමි
.....

1
Input Source sentence: kama hondaigaana wadi
Actual Target Translation: කෑමි හොදයි ගාන වැඩිසි
Predicted Target Translation: කෑමි හොදයි ගාන වැඩිසි
.....

2
Input Source sentence: Kama rasai service ekat hodai
Actual Target Translation: කෑමි රසයි සේවාව හොදයි
Predicted Target Translation: කෑමි රසයි සේවාව හොදයි
.....

3
Input Source sentence: etharam honda natha
Actual Target Translation: එතරම් හොද නෑන
Predicted Target Translation: හා හොද නෑන
.....

4
Input Source sentence: Yako ekh..
Actual Target Translation: යකෝ එක
Predicted Target Translation: යකෝ එක
.....

5
Input Source sentence: super service marama hodai
Actual Target Translation: සුපිරි සේවාව මාරම හොදයි
Predicted Target Translation: සුපිරි සේවාව මාරම හොදයි
```

Figure 5.1 Prediction of Sinhala sentences for randomly given SECM sentences

The Figure 5.1 shows the prediction of Sinhala sentences for the randomly given SECM sentences input. An analysis of the predicted sentences is performed to identify whether the proposed model helped to overcome the challenges pointed out in Section 1.2.1. The following examples show some predicted Sinhala sentences from our model(Sinhala-English Code-Mixed Sentence - SECMS, Reference text- REF, Translated text - TRANS).

- **Example 1 :**

SECMS : gaana wadi  
REF : ගාන වැඩියි  
TRANS : ගණන් වැඩියි

In this sentence in example 1, even though the TRANS doesn't match the exact REF sentence, the meaning of both the sentences are the same, and the prediction is correct.

- **Example 2 :**

SECMS : place eka super clean  
REF : තැන සුපිරි පිරිසිදුයි  
TRANS : තැන සුපිරි පිරිසිදුයි

In this sentence in example 2, the SECMS contains English words such as 'place', 'super' and 'clean'. In TRANS the words are translated to Sinhala. This translation shows us that borrowing words from another language issue is sorted out with our proposed translation model.

In this sentences in example 3 and example 4,



SECMS : kaama echchara special naha

TRANS : කෑම එච්චර විශේෂ නෑහැ

SECMS : kama denna puluwan

TRANS : කෑම දෙන්න පුළුවන්

The sentences in example 3 and example 4, have the same word in two different transliterations format. But in the predicted sentence, both the words ‘kaama’ and ‘kama’ are correctly identified as one Sinhala word ‘කෑම’. The transliteration issue has also been solved with our model.

Also, the use of special characters and numeric character issues were sorted in the normalization phase with the SlangNorm dictionary.

## **5.2 Performance evaluation metrics & Algorithm**

Evaluation of machine translation models can be evaluated using several algorithms such as WER (Word Error Rate), METEOR (Metric for Evaluation of Translation with Explicit ORdering), General Text Matcher (GTM), Translation Edit Rate (TER) and CDER and BLEU (Bilingual Evaluation Understudy).

The evaluation metric selected for this study is the BLEU score. When examining a given source sentence, there can exist multiple translations that are considered perfect. These translations might vary in terms of word order or specific word choices. However, human evaluators possess the ability to discern between translations of high quality and those of lower quality.

The adoption of the BLEU metric for translation evaluation is rooted in its ability to provide a quantitative measure of translation quality. BLEU assesses the correspondence between machine-generated translations and human-generated references, offering a practical means to gauge the effectiveness of translation models. BLEU accounts for variations in word choice and word order, aligning with the nuanced nature of human language. Its correlation with human judgment, particularly in the context of distinguishing between good and poor translations, solidifies its suitability for translation assessment. By selecting BLEU as the evaluation metric, this study ensures a robust and consistent approach to appraising translation quality while maintaining alignment with human perceptual distinctions.

### **5.2.1 BLEU (BiLingual Evaluation Understudy)**

This metric measures the quality of the translation by matching the translation a professional human translation. According to a numerical metric, closeness to one or more human translations are measured. A corpus of good quality human reference translation is required if the BLEU metric is used.

BLEU score is calculated using n-gram modified precision. For each distinctive n-gram in the predicted translation, the maximum frequency count is calculated in each reference sentence. The minimum of this special count and the original count is called the clipped count. The clipped count is always lesser than the original count. When we apply the clip count instead of the original count to calculate the precision, the calculated value would be the modified precision. Modified precisions are considered a better metric compared to the precision.

$$BLEU = BP \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right) \quad (15)$$

According to the equation 15,

$p_n$  → modified precision for n-gram

$\log$  → the base of log is the natural base  $e$

$w_n$  → weight between 0 and 1 for  $\log p_n$

$BP$  → Brevity Penalty

Brevity penalty is used to penalize short machine translation.

$$BP = \begin{cases} 1, & \text{if } c > r \\ \exp \left( 1 - \frac{r}{c} \right), & \text{if } c \leq r \end{cases} \quad (16)$$

In equation (16) of Brevity Penalty,

$c$  → the number of unigram(length) in all the predicted sentences

$r$  → the best match length(closest reference sentence length to the candidate sentences) for each candidate sentence in the corpus

In most cases, the BLEU score is calculated on a corpus where there are many predicted sentences translated from the different source texts and each of the candidate sentences has several reference sentences. Then  $c$  is the total number of unigrams (length) in all the candidate sentences, and  $r$  is the total sum of the best match lengths for each candidate sentence in the dataset.

BLEU is a value between 0 to 1 because  $w_n$ ,  $p_n$ , and  $BP$  are always between 0 to 1.

$$\exp \left( \sum_{n=1}^N w_n \log p_n \right) = \prod_{n=1}^N \exp (w_n \log p_n) \quad (17)$$

$$= \prod_{n=1}^N [\exp(\log p_n)]^{w_n} \quad (18)$$

$$= \prod_{n=1}^N p_n^{w_n} \quad (19)$$

$\in [0,1]$

Usually the in-built BLEU libraries use the N count as 4 (4-gram) and  $w_n = \frac{1}{N}$ .

### 5.3 BLEU score calculation

We calculated the overall BLEU score for randomly selected 100 sentences from the corpus which we have explained in Section 5.1. We exported the randomly selected 100 SECM input sentences, the actual translation for the sentences, and the predicted translation of the sentences into separate text files. Then we applied the files with SacreBLEU(Post,2018) library to calculate the BLEU score.

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

| No | INPUT  | REFERENCE  | PREDICTION                                   | LENGTH |     | MODIFIED PRECISION |        |        |        |        |        |        |        |
|----|--|--|--|--------|-----|--------------------|--------|--------|--------|--------|--------|--------|--------|
|    |  |  |  | REF    | PRE | 1-GRAM             | 2-GRAM | 3-GRAM | 4-GRAM | 5-GRAM | 6-GRAM | 7-GRAM | 8-GRAM |
| 1  | ganan wadi   | ගන වැඩියි  | ගනන් වැඩියි                                  | 2      | 2   | 1                  | 2      | 0      | 1      | 0      | 1      | 0      | 1      |
| 2  | Budu saranai dewi pihitai  | බුදු සරණයි දෙවි පිහිටයි  | බුදු සරණයි දෙවි පිහිටයි                      | 4      | 4   | 4                  | 4      | 3      | 3      | 2      | 2      | 1      | 1      |
| 3  | place eka super clean  | තැන සුපිරි පිරිසිදුයි  | තැන සුපිරි පිරිසිදුයි                        | 3      | 3   | 3                  | 3      | 2      | 2      | 1      | 1      | 0      | 1      |
| 4  | kama raha unta gana hondatama wadi eh gaanata worth na                           | කෑම රහ උනාට ගන හොඳටම වැඩියි ඒ ගනට වටින්තේ නෑ                           | කෑම රහ උනාට ගන හොඳටම වැඩියි ඒ ගනට වටින්තේ නෑ | 10     | 10  | 10                 | 10     | 9      | 9      | 8      | 8      | 7      | 7      |
| 5  | Price eka tikak wadi Customer service eka madi Staff eka thawa improve wenna one | මිල ටිකක් වැඩියි පාරිභෝගික සේවය මදි කාර්ය මණ්ඩලය වැඩි දියුණු කළ යුතුයි | මිල ටිකක් වැඩියි හැබැයි කාර්ය මණ්ඩලය වැඩි    | 12     | 7   | 6                  | 7      | 4      | 6      | 2      | 5      | 0      | 4      |
| 6  | Meya hithan inne I phone thiyenne photo ganna witarai kiyala                     | මෙයා හිතන් ඉන්නේ අයි ෆෝන් තියෙන්නේ ෆෝටෝ ගන්න විතරයි තියන්නේ කියලා      | මෙයා තියන්නේ කියලා                           | 11     | 3   | 3                  | 3      | 1      | 2      | 0      | 1      | 0      | 1      |
| 7  | mn recommend karana thanak   | මන් නිර්දේශ කරන තැනක්  | මන් නිර්දේශ කරන තැනක්                        | 4      | 4   | 4                  | 4      | 3      | 3      | 2      | 2      | 1      | 1      |
| 8  | main road eka laga nisa noisy  | ඒරට්ත පාර ළඟ නිසා සද්ද වැඩියි  | පාර ළඟ නිසා සද්ද වැඩියි                      | 6      | 5   | 5                  | 5      | 4      | 4      | 3      | 3      | 2      | 2      |
| 9  | kaama echchara special naha  | කෑම එච්චර විශේෂ නෑහැ   | කෑම එච්චර විශේෂ නෑහැ                         | 4      | 4   | 4                  | 4      | 3      | 3      | 2      | 2      | 1      | 1      |
| 10 | kama denna puluwan   | කෑම දෙන්න පුළුවන්  | කෑම දෙන්න පුළුවන්                            | 3      | 3   | 3                  | 3      | 2      | 2      | 1      | 1      | 0      | 1      |

Initially, the number of clipped counts and the total number of the particular n-grams in the predicted sentence are extracted to calculate the modified precision as shown in Table 3. Using the retrieved values shown in Table 3, the overall BLEU score is calculated.

|                   |              |             |              |              |
|-------------------|--------------|-------------|--------------|--------------|
| N-Gram            | 1            | 2           | 3            | 4            |
| Weights(Wn)       | 0.25         | 0.25        | 0.25         | 0.25         |
| Ref Total Length  | 280          | 168         | 99           | 56           |
| Pred Total Length | 366          | 275         | 205          | 160          |
| Precesion(Pn)     | 0.76502732   | 0.61090909  | 0.48292683   | 0.35         |
| Wn*log(Pn)        | -0.066960933 | -0.12320178 | -0.181972532 | -0.262455531 |
| Brevity Penalty   | 0.595040575  |             |              |              |
| Cumulative BLEU   | 0.315462186  |             |              |              |

Figure 5.2: Calculated values to evaluate the BLEU Score

In the calculation initially, the number of unigram(length) in all the predicted sentences and the number of best match length(closest reference sentence length to the candidate sentences) for each candidate sentence in the corpus are calculated automatically. Likewise, the n-gram counts are calculated in reference and the predicted translations.

Next, the precision values will be calculated and it is converted to modified precision. Also, the brevity penalty will be calculated automatically with an in-built function. Finally, by applying the calculated values to the BLEU score equation, the score is calculated as 0.3154, as shown in Figure 5.2. Compared to state of the art for BLEU score values received for code-mixed text translation, our model gave a significantly better BLEU score.

#### 5.4 Evaluation of the proposed model with SECM dataset

This section provides the proves of why our proposed model is considered better for code-mixed text translation. We compared our model with the Baseline Seq2Seq model and the Attention model.

### 5.4.1 Experimenting with baseline Seq2Seq model

The baseline Seq2Seq model is the fundamental encoder-decoder framework with RNN as the basic unit in the network. We applied the same set of the dataset used for our study and trained with the baseline Seq2Seq model with the same set of hyperparameter settings, which is used for our proposed model. We experimented with two scenarios: the Seq2Seq baseline model without normalization and the Seq2Seq baseline model with normalization.

The model of Seq2Seq baseline model without normalization gave the training accuracy as 0.5383, training loss 1.4032, testing accuracy 0.2792, and training loss 1.76. The necessary values retrieved for the BLEU score calculation of the Seq2Seq baseline model without normalization from each sentence from the corpus are shown in Figure 5.3. The BLEU score received for the model is 0.1278, as shown in Figure 5.4.

Seq2Seq baseline model with normalization gave the training accuracy of 0.5711, training loss of 0.7753, testing accuracy of 0.2792, and testing loss of 1.75. The necessary values retrieved for the BLEU score calculation of the Seq2Seq baseline model with normalization from each sentence from the corpus are shown in Figure 5.5. The BLEU score received for the model is 0.2077 as shown in Figure 5.6.

| INPUT   | REFERENCE   | HYPOTHESIS                      | LENGTH |     | MODIFIED PRECISION |        |        |        |   |   |   |   |
|---|---|---------------------------------|--------|-----|--------------------|--------|--------|--------|---|---|---|---|
|   |   |                                 | REF    | HYP | 1-GRAM             | 2-GRAM | 3-GRAM | 4-GRAM |   |   |   |   |
| friendly environment.gedara kama wage   | මිත්රශීලී පරිසරය හොද ර කාමි වනේ                                       | සේවාව හොද කාමි වනේ              | 5      | 4   | 2                  | 4      | 1      | 3      | 0 | 2 | 0 | 1 |
| Budusaranai   | බුදු සරණයි  | බුදුසරණයි                       | 2      | 1   | 0                  | 1      | 0      | 1      | 0 | 1 | 0 | 1 |
| mu hamparama apen birthday party gannawauge partyak dilama nah                | මමයා සාමි වකාවකම අපෙන් උසන්දින සාදයක් ගන්නවායාගෙන් සාදයක් දිලුම් නැහැ | නැහැ රහයි                       | 10     | 2   | 1                  | 2      | 0      | 1      | 0 | 1 | 0 | 1 |
| kama rahaiservice eka hondaiprice adui  | කාමි රසයිසේවාව හොදයිමිල අඩුයි   | කාමි රහයි සේවාව හොදයි ගාන අඩුයි | 4      | 6   | 2                  | 6      | 0      | 5      | 0 | 4 | 0 | 3 |
| honda ekak  | හොද එකක්  | හොද එකක්                        | 2      | 2   | 2                  | 2      | 1      | 1      | 0 | 1 | 0 | 1 |
| thanks loku ammi api bellanwila park ta giya                                  | තෑහැනි ලොකු අපා අප ගොලොවල උද යානායා                                   | ලොකු කාමි ගොවාන් රසයි           | 9      | 4   | 1                  | 4      | 0      | 3      | 0 | 2 | 0 | 1 |
| Heta trip giyata kamak naa  | හෙට ගමනට යාමි කමක් නැහැ   | හෙට ගමනට යාමි කමක් නැහැ         | 5      | 5   | 5                  | 5      | 4      | 4      | 3 | 3 | 2 | 2 |
| foods nam hondai  | කාමිනම් හොදයි   | කාමිනම් හොදයි                   | 2      | 2   | 2                  | 2      | 1      | 1      | 0 | 1 | 0 | 1 |
| buffet ekanam patta   | බුලේ එකනම් සරිට   | බුලේ එකනම් විශිෂ්ටයි            | 3      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| service eka patta food uth supiri   | සේවාව දුකාමි හොදයිකාමන් විශිෂ්ටයි                                     | සේවාව සරිට කාමන් සුසිරි         | 4      | 4   | 1                  | 4      | 0      | 3      | 0 | 2 | 0 | 1 |
| Kama rasaigaana awulak na   | කාමි රසයිගාන අඩුලක් නැ  | කාමි අඩුලක් නැ                  | 4      | 3   | 3                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| Colour light eka hinda  | වර්ණ ආලේකය එක හින්දා  | එක හොදයි                        | 4      | 2   | 1                  | 2      | 0      | 1      | 0 | 1 | 0 | 1 |
| price eka godaak wadi   | මිල ගොවන් වැඩියි  | මිල ගොවන් වැඩියි                | 3      | 3   | 2                  | 3      | 0      | 2      | 0 | 1 | 0 | 1 |
| mokakda dan methana violation eka   | මොකක්ද දුන් මෙකන වයලේකන් එක   | මොකක්ද දුන් කමයි                | 5      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| Hamola ma malak wennen deyak unnama tama                                      | හාමොටම මකක් වෙන්නේ දෙයක් වුනාමි කමයි                                  | මකක් මකක් කමයි අඩුල්            | 6      | 4   | 2                  | 4      | 0      | 3      | 0 | 2 | 0 | 1 |
| supirima restaurant eka   | හොදමි අවන්හල  | හොදමි අවන්හල                    | 2      | 2   | 2                  | 2      | 1      | 1      | 0 | 1 | 0 | 1 |
| italian pizza walata hondama tanaharima rahai and spicy                       | දුකලියන් සීසා වලට හොදමි කානහරිම රසයි වගේම සැරයි                       | රහයි වගේම සීසා වලට හොදමි නැහැ   | 8      | 6   | 4                  | 6      | 2      | 5      | 1 | 4 | 0 | 3 |
| coffee eka nijama rahai   | කෝපි එක නියමි රහයි  | කෝපි එක හොදයි                   | 4      | 3   | 1                  | 3      | 0      | 2      | 0 | 1 | 0 | 1 |
| mach eka ada dinanawa wage neda lamai   | කරහය අද දිනනවා වගේ ගේද ළමයි   | එක                              | 6      | 1   | 0                  | 1      | 0      | 1      | 0 | 1 | 0 | 1 |
| seafood rice ekanam godak rasai   | වුහුදු කනාර රයිස් එකනම් ගොවාන් රසයි                                   | සීයුනි රයිස් එකනම් ගොවාන් රසයි  | 6      | 5   | 4                  | 5      | 3      | 4      | 2 | 3 | 1 | 2 |
| Oka hariyanawa dear oya eyala perma anith aya eka ni tharama dewala bedaganna | මික හරියයි වියමියා එයාට ජේනන අනෙක් අය එක්ක හිතරම් දේවල් බෙදාගන්න      | හා එක්ක එක කමයි යන්න            | 11     | 5   | 1                  | 5      | 0      | 4      | 0 | 3 | 0 | 2 |
| lhama th supiri   | හාමි සුසිරි   | හාමි සුසිරි                     | 2      | 2   | 2                  | 2      | 1      | 1      | 0 | 1 | 0 | 1 |
| drive karaddi mobile phone paavichi karaneka                                  | විනානාය ධාවනාය කිරීමේදී ජංගම දුරකථන සාමිච්චි කරන එක                   | හොදයි                           | 8      | 1   | 0                  | 1      | 0      | 1      | 0 | 1 | 0 | 1 |
| Kisima aduwak kiyanna baha price was resonable                                | කිසිදු අඩුවක් කිමට නොහැකියම්ල සාධාරණයි                                | ගොවන් හොදයි                     | 5      | 2   | 0                  | 2      | 0      | 1      | 0 | 1 | 0 | 1 |
| oyalage service eka pattama hodai   | මියාලගේ සේවාව සරිටමි හොදයි  | සේවාව සේවාව මිල වැඩියි          | 4      | 4   | 1                  | 4      | 0      | 3      | 0 | 2 | 0 | 1 |
| kama patta rahairelax karana hondama lhana                                    | කාමි දුකාමි රසයිහිදුනගේ සිරිමට හොදමි කාන                              | කාමි දුකාමි හොදයි ගාන වැඩියි    | 6      | 5   | 2                  | 5      | 1      | 4      | 0 | 3 | 0 | 2 |
| post dammata wedak ne Pata rakina mura  | පෝස් වී දැමිමට වැඩික් නැහැරටරකින වුර දේවනා                            | සුළුවන් ගාන වැඩික් නැහැ         | 9      | 4   | 1                  | 4      | 0      | 3      | 0 | 2 | 0 | 1 |
| service Honda nea   | සේවාව හොද නැහැ  | සේවාව හොදයි                     | 3      | 2   | 1                  | 2      | 0      | 1      | 0 | 1 | 0 | 1 |
| review balala giye but kamanam echchara                                       | රිටිම් බලල ගියේ එක්කාමි නම් එපිටර හොද නැහැ                            | ගොවන් වැඩියි දුකා හොදයි         | 8      | 4   | 0                  | 4      | 0      | 3      | 0 | 2 | 0 | 1 |
| meke kama maara rahai   | මෙක කාමි මාර රසයි   | මාර රසයි කාමි                   | 4      | 3   | 3                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |

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.

| N-Gram                         | 1            | 2            | 3            | 4            |
|--------------------------------|--------------|--------------|--------------|--------------|
| Weight(Wn)                     | 0.25         | 0.25         | 0.25         | 0.25         |
| Ref Total Length               | 186          | 73           | 27           | 12           |
| Pred Total Length              | 356          | 266          | 197          | 148          |
| Modified Precesion(Wn*log(Pn)) | -0.162296014 | -0.323259217 | -0.496841716 | -0.628076406 |
| Brevity Penalty                | 0.639781087  |              |              |              |
| Cumilative BLEU                | 0.127823795  |              |              |              |

Figure 5.4: BLEU score calculation values of Seq2Seq baseline model without normalization



| INPUT  | REFERENCE   | HYPOTHESIS                                       | LENGTH |     | MODIFIED PRECISION |        |        |        |   |   |   |   |
|--|---|--|--------|-----|--------------------|--------|--------|--------|---|---|---|---|
|  |   |  | REF    | HYP | 1-GRAM             | 2-GRAM | 3-GRAM | 4-GRAM |   |   |   |   |
| Service eka godaak hodai   | සේවාව ගෙවෙයක් හොඳයි   | සේවාව ගෙවෙයක් හොඳයි                              | 3      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 | 1 |
| star rating ekak denna watinava  | ස්ථාර රේටින්ග් එකක් දෙනවා වටිනාව  | ස්ථාර එකක් දෙනවා වටිනාව                          | 5      | 4   | 4                  | 4      | 2      | 3      | 1 | 2 | 0 | 1 |
| hondama view eka family eka yana hondama tanak   | හොඳම දර්ශනය එක හටුල එක යන්න හොඳම කැනක්  | හොඳම දර්ශනය එක හටුල එක යන්න හොඳම කැනක්           | 8      | 8   | 8                  | 8      | 7      | 7      | 6 | 6 | 5 | 5 |
| tikak ganan wadi eh unata kamanam hodai  | ටිකක් ගසන් වැටියි ඒ උනාව කැමතම් හොඳයි සේවාව                                       | ටිකක් ගසන් වැටියි ඒ උනාව කැමතම් හොඳයි සේවාව      | 11     | 9   | 8                  | 9      | 7      | 8      | 6 | 7 | 5 | 6 |
| Theruvan saranai   | හොඳුවන් සරණයි   | හොඳුවන් සරණයි                                    | 2      | 2   | 2                  | 2      | 1      | 1      | 0 | 1 | 0 | 1 |
| Colour light eka hinda   | වර්ණ ආලෝකය එක හින්දා  | එක එක  | 4      | 2   | 1                  | 2      | 0      | 1      | 0 | 1 | 0 | 1 |
| godaak Ganan kama Very crowded   | ගෙවෙයක් ගසන් කැමි ගෙවෙයක් සෙනඟ වැඩි   | ගෙවෙයක් ගසන් කැමි ගෙවෙයක් සෙනඟ වැඩි              | 6      | 6   | 6                  | 6      | 5      | 5      | 4 | 4 | 3 | 3 |
| Godaaak gaana wadi   | ගෙවෙයක් ගසන් වැටියි   | ගෙවෙයක් ගසන් වැටියි                              | 3      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 | 1 |
| repair aduuii pic up super   | සවස්වැටියා කිරීමේ අවස්ථාවේ සුපිරි   | එක කිරීමේ අවස්ථාවේ                               | 5      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| rice ekai chilli paste ekai patta bro  | රයිස් එකයි චිලි පේස්ට් එකයි ඉතා හොඳයි ඕ රෝ  | රයිස් රයිස් ගෙවෙයක් හොඳයි ඉතා හොඳයි              | 10     | 6   | 3                  | 6      | 1      | 5      | 0 | 4 | 0 | 3 |
| Gaana wadifood tasty   | ගසන් වැටියි කැමි රසයි   | ගසන් වැටියි කැමි රසයි                            | 4      | 4   | 4                  | 4      | 3      | 3      | 2 | 2 | 1 | 1 |
| customer service ekam hodai  | සාර්වභාග්‍යයන් සේවාවකම් හොඳයි   | සාර්වභාග්‍යයන් සේවාවකම් හොඳයි                    | 3      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 | 1 |
| Gana tikak wediSenaga  | ගසන් ටිකක් වැටියිසෙනඟ   | ගසන් ටිකක් වැටියි                                | 3      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| ub tmii ek auruddk wath dunnd  | උබ් කමයි එක් අවුරුද්දකටත් දුන්නා  | උබ් කමයි කමයි                                    | 5      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| kama hari hondaineat   | කැමි හරි හොඳයි පිළිවෙලයි  | කැමි හරි හොඳයි                                   | 4      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 | 1 |
| gedara hadapu kama wage godak rasai  | ගෙදර හදපු කැමි වගේ ගෙවෙයක් රසයි   | ගෙදර හදපු කැමි වගේ ගෙවෙයක් රසයි                  | 6      | 6   | 6                  | 6      | 5      | 5      | 4 | 4 | 3 | 3 |
| mkkda bn seen eka  | මොකක්ද බන් සීන් එක  | මොකක්ද බන් සීන්                                  | 4      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| udeta gihilla orderer karanam dawalta kannu puluwang kamath awul service ekath awul        | උදෙට ගිහිල්ලා ඔබේවර් කාරනම්දවල්ලේ කන්න පුළුවන් කැමිත් අවුල් සේවාව අවුල්           | අවුල් සේවාව අවුල් කැමිත් අවුල්                   | 10     | 5   | 4                  | 5      | 3      | 4      | 1 | 3 | 0 | 2 |
| I ordered a large portion and a regular one kisima wenasak naha boru karanne               | මින් ඕඩර් කළා ලොකු පෝරන් හා රේගුලර් එකක් කිසිම වෙනසක් නෑ වෙනරු කරන්නේ             | මින් ඕඩර් කළා ලොකු ලොකු එකක් කිසිම වෙන් නෑ ගසන්  | 13     | 10  | 7                  | 10     | 4      | 9      | 2 | 8 | 1 | 7 |
| the best roti with cheese and chicken curry godak  | හොඳම රෝට් පමන එස් හා චීන්කන් කර ගෙවෙයක් රසයි                                      | හොඳම කැපා හා කැමි චීන්කන් කර ගෙවෙයක් රසයි        | 9      | 8   | 6                  | 8      | 3      | 7      | 2 | 6 | 1 | 5 |
| place ekath super cleankamath hodai  | කැනක් සුපිරි පිරිසිදුකැමිත් හොඳයි   | කැනක් සුපිරි නෑ                                  | 4      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| Ammo oi  | අම්මේ මයි   | අම්මේ අම්මේ                                      | 2      | 2   | 1                  | 2      | 0      | 1      | 0 | 1 | 0 | 1 |
| staffsila kisima pilliwalak naservice eka godaak   | කාර්ය මණ්ඩලය කිසිම පිල්වලක් නෑ සේවාව ගෙවෙයක්                                      | කාර්ය මණ්ඩලය කිසිම නෑ සේවාව ගෙවෙයක් හොඳයි        | 10     | 9   | 6                  | 9      | 4      | 8      | 2 | 7 | 0 | 6 |
| not words to express supiri kama   | වචන නෑ කියන්න පුළුවන් කැමි  | පුළුවන් නෑ කැමි                                  | 5      | 3   | 3                  | 3      | 0      | 2      | 0 | 1 | 0 | 1 |
| service charge ekak add karana tharamata service ekak naha                                 | සේවාව ගාස්තු එකක් එකතු කරනා කරමට සේවාවක් නෑනෑ                                     | සේවාව එකක් එකක් එකතු කරන ඕනි                     | 8      | 6   | 4                  | 6      | 2      | 5      | 1 | 4 | 0 | 3 |
| Gana hondatama Wadi but food quality eka   | ගසන් හොඳම වැටියි නැටියි කැමි ගසන්මකනාවය   | ගසන් හොඳම වැටියි නැටියි කැමි වැටියි              | 7      | 6   | 5                  | 6      | 4      | 5      | 3 | 4 | 2 | 3 |
| sea food categories godaak thiyana tanak   | විදුදු ආහාර වර්ග ගෙවෙයක් නියත කැනක්   | විදුදු ආහාර ගෙවෙයක් නියත කැනක්                   | 6      | 5   | 5                  | 5      | 3      | 4      | 1 | 3 | 0 | 2 |
| kama harima nahanot hygeine  | කැමි හරිම නෑනෑ පිරිසිදු නෑ  | කැමි හරිම නෑනෑ කියලා                             | 5      | 4   | 3                  | 4      | 2      | 3      | 1 | 2 | 0 | 1 |
| I had ah tasty and spicy vegetable rice very cheap and staffs late service eka outstanding | මම කැට් රසයි හා සුර එළවලු රයිස් ගෙවෙයක් ලොකු හා කාර්ය මණ්ඩලය පරේ කයි සේවාව සුපිරි | මම කැට් රසයි හා එළවලු කැමි හා රයිස් ගෙවෙයක් රසයි | 17     | 10  | 8                  | 10     | 4      | 9      | 2 | 8 | 1 | 7 |

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

| N-Gram                | 1            | 2            | 3            | 4            |
|-----------------------|--------------|--------------|--------------|--------------|
| Weight(Wn)            | 0.25         | 0.25         | 0.25         | 0.25         |
| Ref Total Length      | 212          | 127          | 71           | 35           |
| Pred Total Length     | 317          | 288          | 214          | 166          |
| Precesion(Wn*log(Pn)) | -0.145237483 | -0.204693348 | -0.275824034 | -0.389159932 |
| Brevity Penalty       | 0.573080988  |              |              |              |
| Cumilative BLEU       | 0.207703639  |              |              |              |

Figure 5.6:BLEU score calculation values of Seq2Seq baseline model with normalization

Figure 5.7, shows the comparison in a detailed view. From the training and testing result, we can see that the model Seq2Seq baseline with the normalization layer has improved the accuracy of the model.

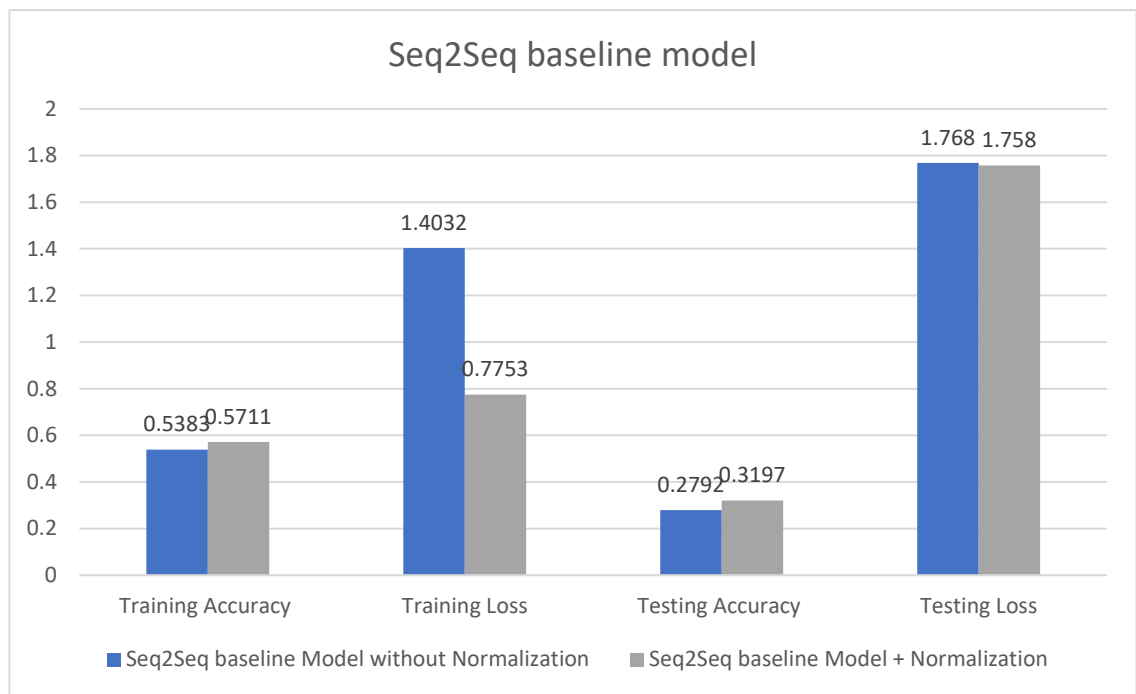


Figure 5.7: Seq2Seq baseline model result

#### 5.4.2 Experimenting with baseline Seq2Seq Attention model

In this experiment, we again used the same dataset and we used to train our model, to Seq2Seq with Attention model. This attention technique allowed for a considerable improvement in machine translation systems by focusing on the relevant parts of the input sequence where attention is given. Here also, we experimented with the normalization layer and without the normalization layer. The Seq2Seq Attention model without normalization gave the training accuracy as 0.7055, training loss 0.303, testing accuracy 0.303, and training loss 1.15. We calculated the BLEU score for randomly selected 100

sentences from the corpus. The necessary values retrieved for the BLEU score calculation of the Seq2Seq Attention-based model without normalization from each sentence from the corpus are shown in Figure 5.8. We calculated the BLEU score received for the model is 0.2895, as shown in Figure 5.9.

| INPUT  | REFERENCE   | HYPOTHESIS  | LENGTH |     | MODIFIED PRECISION |        |        |        |   |   |   |   |
|--|---|---|--------|-----|--------------------|--------|--------|--------|---|---|---|---|
|  |   |   | REF    | HYP | 1-GRAM             | 2-GRAM | 3-GRAM | 4-GRAM |   |   |   |   |
| Service eka godaak hodai   | සේවාව ගොඩනැගීම  | සේවාව ගොඩනැගීම                                      | 3      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 | 1 |
| star rating ekak denna watinava  | ස්ථාර අවදානමක් ඇති බවට පත්වීම   | ස්ථාර අවදානමක් ඇති බවට පත්වීම                       | 5      | 4   | 4                  | 4      | 2      | 3      | 1 | 2 | 0 | 1 |
| price high   | මිල වැඩියි  | මිල වැඩියි  | 2      | 2   | 2                  | 2      | 1      | 1      | 0 | 1 | 0 | 1 |
| Theruwana saranai  | නොදැන සරණයි   | නොදැන සරණයි   | 2      | 2   | 2                  | 2      | 1      | 1      | 0 | 1 | 0 | 1 |
| Colour light eka hinda   | වර්ණ සාලකයක් එක හින්දා  | එක එක   | 4      | 2   | 1                  | 2      | 0      | 1      | 0 | 1 | 0 | 1 |
| mutton biriyani eka echcharama special naha  | මිටිමස්ස් කිරිගස්ස් එක එපිටුම විශේෂ කැහැ                                | මිටිමස්ස් කිරිගස්ස් එක විශේෂ කැහැ                   | 6      | 5   | 5                  | 5      | 3      | 4      | 1 | 3 | 0 | 2 |
| Godaak gaana wadi  | ගොඩනැගීම ගානා වැඩියි  | ගොඩනැගීම ගානා වැඩියි                                | 3      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 | 1 |
| repear aduiii pic up super   | අළුත්වැඩියා කිරීම ආදිය සුපිරි   | එක කිරීම ආදිය                                       | 5      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| rice ekai chilli paste ekai patta bro  | රයිස් එකයි චිලි පේස්ට් එකයි ඉතා හොඳයි ඕ රස්                             | රයිස් රයිස් ගොඩනැගීම ඉතා හොඳයි                      | 10     | 6   | 3                  | 6      | 1      | 5      | 0 | 4 | 0 | 3 |
| customer service ekanm hodai   | සාර්වභෝග්‍යයන් සේවාවලට හොඳයි  | සාර්වභෝග්‍යයන් සේවාවලට හොඳයි                        | 3      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 | 1 |
| Gana tikak wedi Senaga   | ගානා විකුණ වැඩියි සෙනගා   | ගානා විකුණ වැඩියි                                   | 4      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| ub tmii ek auruddk wath dunnd  | උඩු කමයි එක් අඩුරුදුකමක් දැනෙන  | උඩු කමයි කමයි                                       | 5      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| kama hari hondaineat   | කැමි හරි හොඳයි පිළිවෙලයි  | කැමි හරි හොඳයි                                      | 4      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 | 1 |
| mkkda bn seen eka  | මමකක්ද බිත්තියක් එක   | මමකක්ද බිත්තියක්                                    | 4      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| udeta gihilla oredar karanam dawalta kannu puluwang kamath awul service ekath awul | උදේ ගිහිල්ලා බිවේන කාරනම්දවල්ට කන්න පුළුවන් කැමිත් ආදිය සේවාව ආදිය      | ආදිය සේවාව ආදිය කැමිත් ආදිය                         | 10     | 5   | 4                  | 5      | 3      | 4      | 1 | 3 | 0 | 2 |
| I ordered a large portion and a regular one kisima wasak naha boru karanne         | මින් ඕවර් කළා ලොකු පෝර්ශන් හා රේගුලර් එකක් කිසිම වෙනසක් නැ වෙරා කැරන්තේ | මින් ඕවර් කළා ලොකු ලොකු එකක් කිසිම මින් නැ ගානා     | 13     | 10  | 7                  | 10     | 4      | 9      | 2 | 8 | 1 | 7 |
| the best roti with cheese and chicken curry godak rasai                            | හොඳම රොට් සමඟ චීස් හා චික්කන් කැර ගොඩනැගීම රසයි                         | හොඳම කැනා හා කැමි චික්කන් කැර ගොඩනැගීම රසයි         | 9      | 8   | 6                  | 8      | 3      | 7      | 2 | 6 | 1 | 5 |
| place ekath super cleankamath hodai  | ආහල් සුපිරි පිරිසිදුකමක් හොඳයි  | ආහල් සුපිරි නැ                                      | 4      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 | 1 |
| staffslla kisima pilwalak naservice eka godaak parakui ganath wadi                 | කාර්ය මණ්ඩලය කිසිම පිලිවෙලක් නැ සේවාව ගොඩනැගීම සරත්කුඩු ආහල් වැඩියි     | කාර්ය මණ්ඩලය කිසිම නැ සේවාව ගොඩනැගීම හොඳයි සේවාව ගො | 10     | 9   | 6                  | 9      | 4      | 8      | 2 | 7 | 0 | 6 |
| service charge ekak add karana tharamata service ekak naha                         | සේවාව ගාස්තු එකක් එකතු කරන තරම්ම සේවාවක් නැහැ                           | සේවාව එකක් එකක් එකතු කරන ඕනි                        | 8      | 6   | 4                  | 6      | 2      | 5      | 1 | 4 | 0 | 3 |
| Gana hondatama Wadi but food quality eka hondai                                    | ගානා හොඳම වැඩියි නැබැයි කැමි ගුණාත්මකභාවය හොඳයි                         | ගානා හොඳම වැඩියි නැබැයි කැමි වැඩියි                 | 7      | 6   | 5                  | 6      | 4      | 5      | 3 | 4 | 2 | 3 |
| kama harima naharot hygiene  | කැමි හරිම නැහැ පිරිසිදුකම නැ  | කැමි හරිම නැහැ කියලා                                | 5      | 4   | 3                  | 4      | 2      | 3      | 1 | 2 | 0 | 1 |

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.

Seq2Seq attention model with normalization gave the training accuracy of 0.7022, training loss of 0.5023, testing accuracy of 0.3105, and testing loss of 1.0522. The necessary values retrieved for the BLEU score calculation of the Seq2Seq Attention-based model with

normalization are shown in Figure 5.10. The BLEU score received for the model is 0.3154 as shown in Figure 5.11.

|                                   |              |              |              |             |
|-----------------------------------|--------------|--------------|--------------|-------------|
| N-Gram                            | 1            | 2            | 3            | 4           |
| Weight(Wn)                        | 0.25         | 0.25         | 0.25         | 0.25        |
| Ref Total Length                  | 311          | 176          | 94           | 45          |
| Pred Total Length                 | 430          | 337          | 258          | 197         |
| Modified<br>Precesion(Wn*log(Pn)) | -0.080998074 | -0.162399734 | -0.252416201 | -0.36913531 |
| Brevity Penalty                   | 0.68768892   |              |              |             |
| Cumilative BLEU                   | 0.289567188  |              |              |             |

Figure 5.9:BLEU score calculation values of Seq2Seq Attention model without normalization

| INPUT   | REFERENCE  | HYPOTHESIS  | LENGTH |     | MODIFIED PRECISION |        |        |        |   |   |   |
|---|--|---|--------|-----|--------------------|--------|--------|--------|---|---|---|
|   |  |   | REF    | HYP | 1-GRAM             | 2-GRAM | 3-GRAM | 4-GRAM |   |   |   |
| star rating ekak denna watinava   | ස්ටාර් රේටින්ග් එකක් දෙනවා වැඩිනවා   | ස්ටාර් එකක් දෙනවා වැඩිනවා                           | 5      | 4   | 4                  | 2      | 3      | 1      | 2 | 0 | 1 |
| hondama view eka family eka yana hondama tanak  | හොඳම දර්ශනය එක පවුල එක යන්න හොඳම කුසක්   | හොඳම දර්ශනය එක පවුල එක යන්න හොඳම                    | 8      | 8   | 8                  | 7      | 7      | 6      | 6 | 5 | 5 |
| Colour light eka hinda  | වර්ණ ආලෝකය එක හින්දා   | එක එක   | 4      | 2   | 1                  | 2      | 0      | 1      | 0 | 1 | 0 |
| godaak Ganan kama Very crowded  | ගොඩාක් ගසන් කැමි ගොඩාක් සෙනග වැඩි  | ගොඩාක් ගසන් කැමි ගොඩාක් සෙනග වැඩි                   | 6      | 6   | 6                  | 6      | 5      | 5      | 4 | 4 | 3 |
| mutton biriyani eka echcharama special naha   | මිම්බන් බිරියාහි එක එච්චරම් විශේෂ න්‍යාය   | මිම්බන් බිරියාහි එක විශේෂ න්‍යාය                    | 6      | 5   | 5                  | 3      | 4      | 1      | 3 | 0 | 2 |
| repear aduilli pic up super   | අළුත්වැඩියා කිරීම අඩුයි මිකස් සුපිරි   | එක කිරීම අඩුයි                                      | 5      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 |
| rice ekai chilli paste ekai patta bro   | රයිස් එකයි චීලි පේස්ට් එකයි ඉසා හොඳයි ඕ රෝස්   | රයිස් රයිස් ගොඩාක් හොඳයි ඉසා හොඳයි                  | 10     | 6   | 3                  | 6      | 1      | 5      | 0 | 4 | 0 |
| Gaana wadifood tasty  | ගාන වැඩියි කැමි රසයි   | ගාන වැඩියි කැමි රසයි                                | 4      | 4   | 4                  | 4      | 3      | 3      | 2 | 2 | 1 |
| customer service ekanm hodai  | සාරිසේවකයන් සේවාවනම් හොඳයි   | සාරිසේවකයන් සේවාවනම් හොඳයි                          | 3      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 |
| sudu adagena kalu awidin puduma lassana<br>kathawakhamogema acting supireeeidiri<br>anagathe gana positive hithanna pulwan dan jaya | සුදු ඇඳගෙන කළු ඇඳිනි සුදුම් ලූස්සන කකාවක්<br>කුමාරසේම් රඟකැමි සුපිරි ඉදිරි දුනාගසන් ගැන ධනාත්මක<br>හිතවත් සුළුබන් දැන්ජය ඓවා මිත්තොටම් | කැමි වගේ  | 19     | 2   | 0                  | 2      | 0      | 1      | 0 | 1 | 0 |
| Gana tikak wediSenaga   | ගාන විකස් වැඩියිසෙනග   | ගාන විකස් වැඩි                                      | 3      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 |
| ub tmii ek auruddk wath dunnd   | එම් කමයි එක් අවුරුද්දකවත් දුන්නා   | එම් කමයි කමයි                                       | 5      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 |
| kama hari hondaineat  | කැමි හරි හොඳයි මිළිවෙලයි   | කැමි හරි හොඳයි                                      | 4      | 3   | 3                  | 3      | 2      | 2      | 1 | 1 | 0 |
| mkkda bn seen eka   | මොකක්ද බන් සීන් එක   | මොකක්ද බන් සීන්                                     | 4      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 |
| udeta gihilla oredder karanam dawalta kanna<br>puluwang kamath awul service ekath awul  | උදේට ගිහිල්ලා මීච්චි කරනාමිදුවට් කන්න සුළුවන්<br>කැමි අවුල් සේවාව අඩුල්  | අඩුල් සේවාව අඩුල් කැමි අඩුල්                        | 10     | 5   | 4                  | 5      | 3      | 4      | 1 | 3 | 0 |
| I ordered a large portion and a regular one kisma<br>wenasak naha boru karanne  | මින් මිච්චි කලා ලොකු සෝශන් හා ඓවුලු එකක් කිසිම මින් නා<br>වෙනසක් නා ඓවුරු කරන්නේ   | මින් මිච්චි කලා ලොකු ලොකු එකක් කිසිම මින් නා<br>ගාන | 13     | 10  | 7                  | 10     | 4      | 9      | 2 | 8 | 1 |
| the best roti with cheese and chicken curry godak<br>rasai  | හොඳම රෝටි සමඟ චීස් හා චීන්කන් කර ගොඩාක් රසයි   | හොඳම හාන හා කැමි චීන්කන් කර ගොඩාක් රසයි             | 9      | 8   | 6                  | 8      | 3      | 7      | 2 | 6 | 1 |
| place ekath super cleankamath hodai   | භාගන් සුපිරි පිරිසිදුකැමිත් හොඳයි  | භාගන් සුපිරි නා                                     | 4      | 3   | 2                  | 3      | 1      | 2      | 0 | 1 | 0 |
| Ammo oi   | අම්මෝ මයි  | අම්මෝ අම්මෝ   | 2      | 2   | 1                  | 2      | 0      | 1      | 0 | 1 | 0 |
| not words to express supiri kama  | වචන නා කියන්න සුපිරි කැමි  | සුපිරි නා කැමි                                      | 5      | 3   | 3                  | 3      | 0      | 2      | 0 | 1 | 0 |
| service charge ekak add karana tharamata service<br>ekak naha   | සේවාව ගාස්තු එකක් එකතු කරමිට සේවාවක් නානා  | සේවාව එකක් එකක් එකතු කරන මිනි                       | 8      | 6   | 4                  | 6      | 2      | 5      | 1 | 4 | 0 |
| Gana hondatama Wadi but food quality eka hondai   | ගාන හොඳම වැඩියි හැබැයි කැමි ගුණාත්මකභාවය<br>හොඳයි  | ගාන හොඳම වැඩියි හැබැයි කැමි වැඩියි                  | 7      | 6   | 5                  | 6      | 4      | 5      | 3 | 4 | 2 |
| kama harima nahanot hygeine   | කැමි හරිම නානා පිරිසිදුන් නා   | කැමි හරිම නානා කියලා                                | 5      | 4   | 3                  | 4      | 2      | 3      | 1 | 2 | 0 |
| crowd wadikamanam prashnayak na   | සෙනග වැඩියි කැමිමච් ජ රස්තියක් නා  | සෙනග වැඩියි කැමිමච් කමයි                            | 7      | 4   | 3                  | 4      | 2      | 3      | 1 | 2 | 0 |
| kama rahaigodaak varieties thiyana  | කැමි රසයි ගොඩාක් වර්ග කියනවා   | කැමි රසයි ගොඩාක් වර්ග කියනවා                        | 5      | 5   | 5                  | 5      | 4      | 4      | 3 | 3 | 2 |
| Service eka supiri Kama rasa depend wenawa dina<br>gedara uyapu kaama wage  | සේවාව සුපිරි කැමි රස විසෙන්වී වෙනවා දින අනුවෙන්<br>ගෙදර උපසු කැමි වගේ  | සේවාව සුපිරි කැමි වෙනවා<br>ගෙදර හදසු කැමි වගේ       | 8      | 4   | 4                  | 4      | 2      | 3      | 1 | 2 | 0 |
| Oscillate kiwwe machn   | මිස්මිලේම් කීම්මි මවන්   | නානා  | 3      | 1   | 0                  | 1      | 0      | 1      | 0 | 1 | 0 |

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.

|  |             |              |              |              |
|--|-------------|--------------|--------------|--------------|
| N-Gram   | 1           | 2            | 3            | 4            |
| Weight(Wn)                                     | 0.25        | 0.25         | 0.25         | 0.25         |
| Ref Total Length                               | 293         | 170          | 99           | 54           |
| Pred Total Length                              | 386         | 300          | 228          | 174          |
| Modified<br>Precesion( $W_n \cdot \log(P_n)$ ) | -0.06891619 | -0.141996009 | -0.208556445 | -0.292517813 |
| Brevity Penalty                                | 0.64136546  |              |              |              |
| Cumilative BLEU                                | 0.314697851 |              |              |              |

Figure 5.11: BLEU score calculation values of Seq2Seq attention model with normalization

Figure 5.12, shows the comparison of the accuracies and loss in a detailed view.

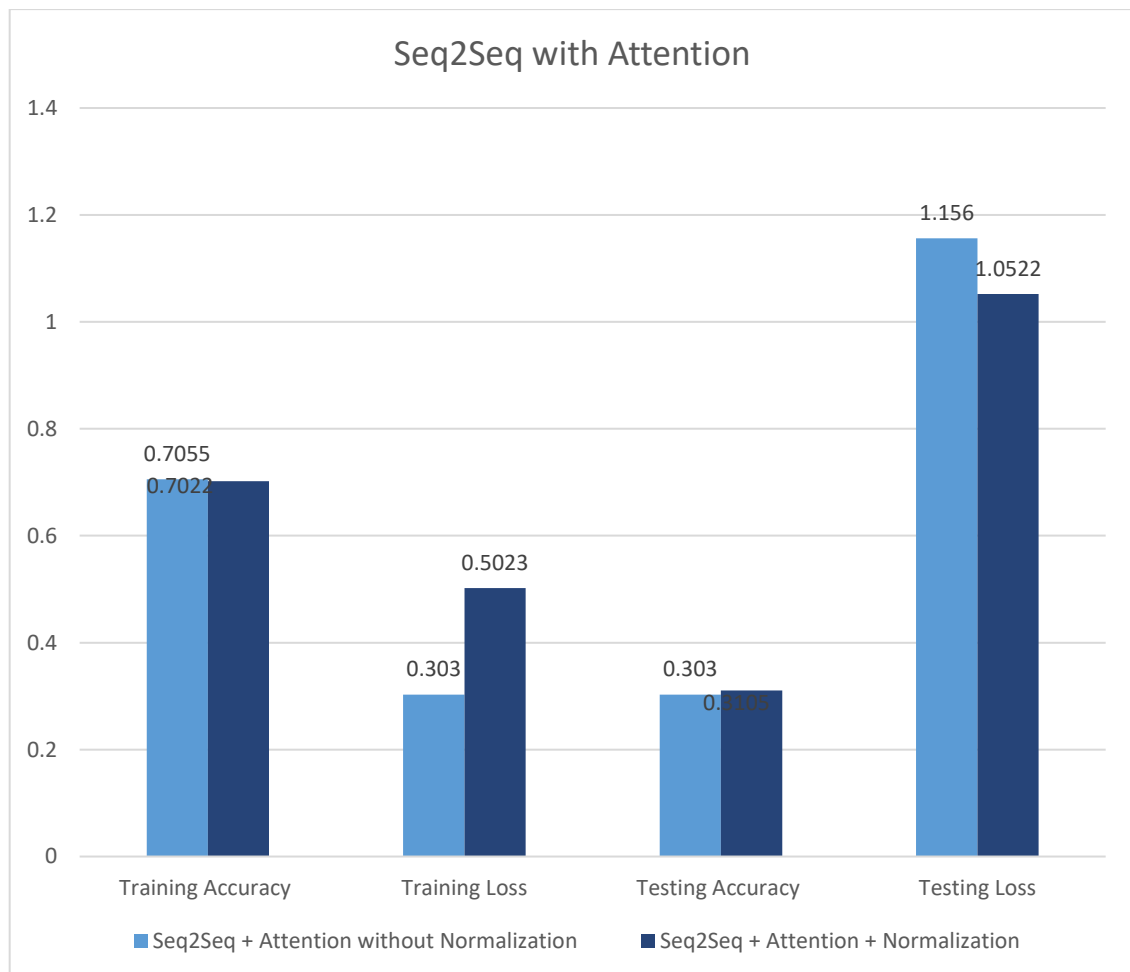


Figure 5.12: Seq2Seq with Attention model result

### 5.4.3 Experimenting with the proposed model

Our proposed approach is a Seq2Seq model with Teacher Forcing Algorithm. Also, the basic units of our neural network use the LSTM as the basic unit, which prevents the vanishing gradient issue. The model is evaluated with the same scenarios which we used with other models: with the normalization layer and without the normalization layer.

The Seq2Seq Teacher Forcing model without normalization gave the training accuracy as 0.7142, training loss 0.5095, testing accuracy 0.3717, and testing loss 0.3872. The BLEU score received for the model is 0.3154. Seq2Seq Attention model with normalization gave the training accuracy of 0.70157, training loss of 0.5095, testing accuracy of 0.3787, and testing loss of 0.3866. The BLEU score received for the model is 0.3389, as shown in Section 5.3. Figure 5.13, shows the result.

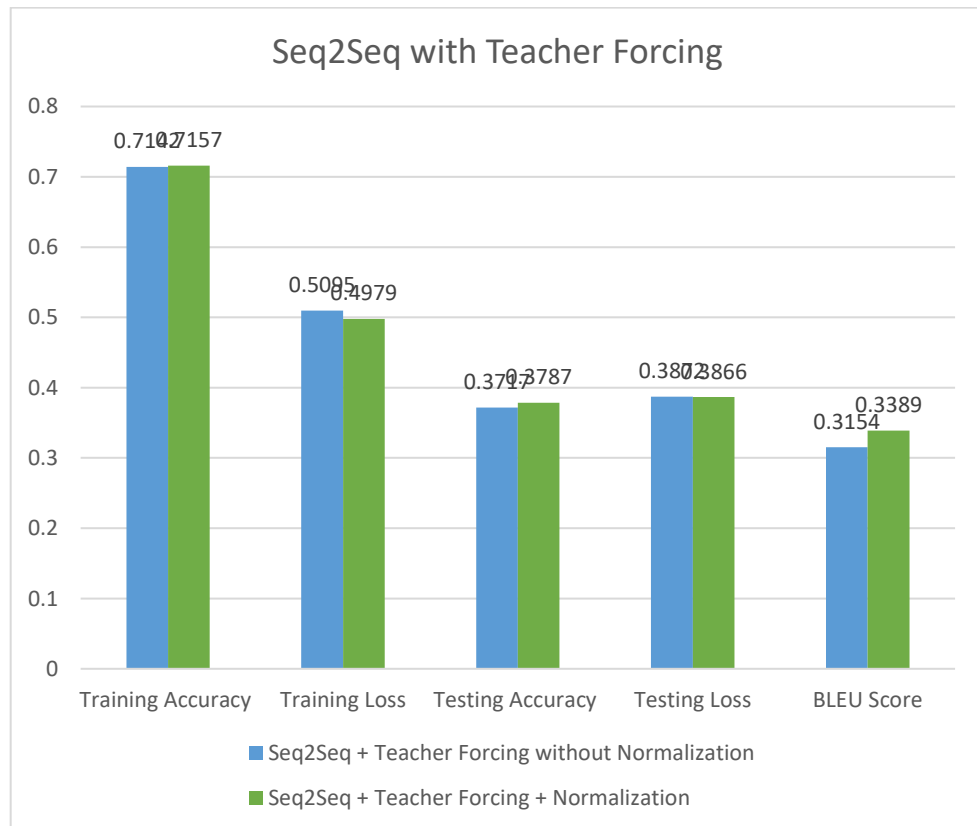


Figure 5.13: Experiment values of Seq2Seq with Teacher Forcing model

## 5.5 Evaluation of the proposed model with Hindi-English code-mixed dataset

To check the efficiency of our proposed model, we downloaded a publicly available parallel corpus of Hinglish-English code-mixed sentences(Srivasta & Singh,2020). Hinglish(Hindi-English code-mixed text) is the source language and English is the target language.

Fifty sentences from the corpus were sampled for the evaluation study of different code-mixed translation systems. Bing Translator(BT) provided a BLEU score of 0.139, Google Translate(GT) provided a BLEU score of 0.14, and the combined approach of Augmented Pipeline(AP) and Google Translate provided a BLEU1 score of 0.153. Our model gave the BLEU1 score as 0.293. As we can see in Figure 5.14, our model offers a significantly higher BLEU score for the code-mixed text from another language pair.

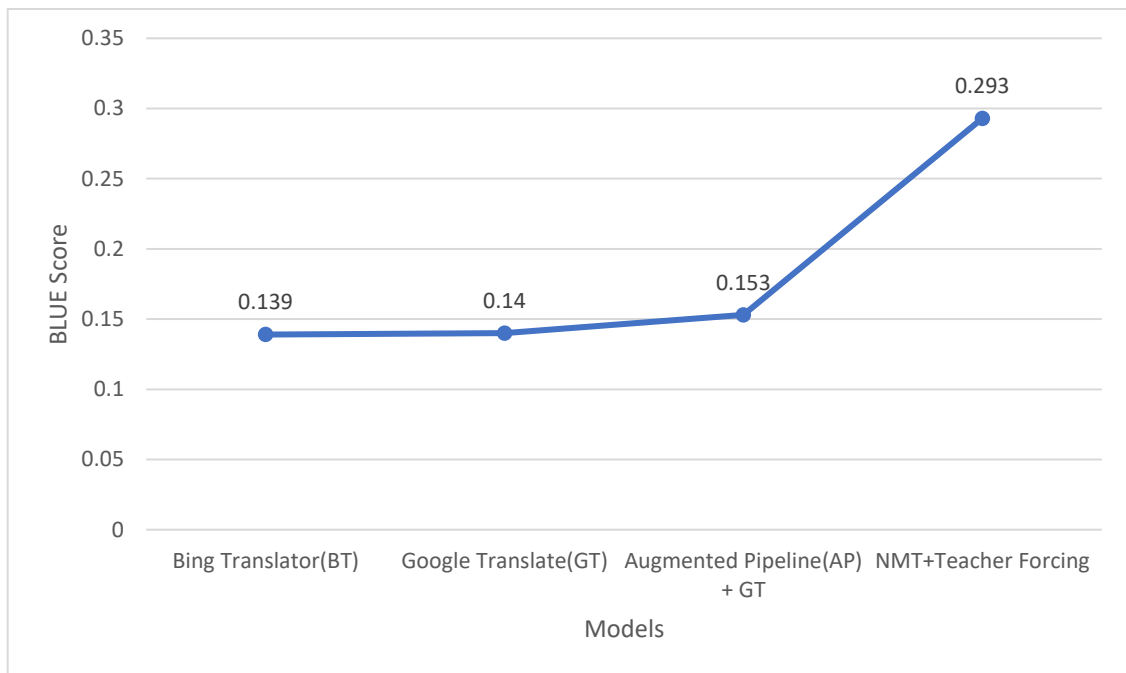


Figure 5.14: BLEU Score comparison for Hindi-English code-mixed translation

## 5.6 Summary

This section delves into the performance evaluation of the proposed model for code-mixed text translation, focusing on the prediction from the model and the metrics used for evaluation.

This section starts by providing details on exporting the trained model to predict the output. Randomly selected SECM sentences are input into the model, and the predicted translations in Sinhala are generated as output. An analysis is carried out to determine if the model overcomes the challenges highlighted earlier. Several examples are presented, showcasing the model's performance in handling various aspects of code-mixed text translation, such as meaning preservation, transliteration, and handling special characters.

The chosen evaluation metric is the BLEU score, a well-established method for assessing translation quality. BLEU measures the correspondence between machine-generated translations and human references, providing a quantitative measure of translation quality. The chapter explains how BLEU is calculated using modified precision, weight factors, and brevity penalty. The formula and calculations for BLEU are detailed, demonstrating its relevance in evaluating translation models. The calculated overall BLEU score for the model is reported as 0.3154, indicating its translation quality. This score is compared to the state of the art for BLEU score values received for code-mixed text translation, emphasizing the significant improvement offered by the proposed model.

The section also involves the evaluation of the model with the SECM dataset. A comparison is drawn with the baseline Seq2Seq model and the Seq2Seq Attention model. Different scenarios, with and without normalization layers, are experimented with. The BLEU scores for each model variant are presented, showcasing the superiority of the proposed Seq2Seq Teacher Forcing model in terms of translation quality.



Furthermore, the proposed model's efficacy is demonstrated using a Hindi-English code-mixed dataset. Comparative BLEU scores with other translation systems, such as Bing Translator and Google Translate, highlight the robust performance of the proposed model across different language pairs.

In summary, this section provides an in-depth evaluation of the proposed code-mixed text translation model's performance, backed by detailed explanations of prediction, BLEU metric, and comparative analyses with other models. The results substantiate the model's effectiveness in addressing the challenges of code-mixed text translation.

## **6. Result & Discussion**

The main aim of this research study is to translate Sinhala-English code-mixed text to the Sinhala language. The evaluation of the proposed model has been conducted considering two aspects: how does our proposed model perform compared to the current methods available for the code-mixed text translation, and what is the performance of our model with the code-mixed text from another language.

The current models which have performed code-mixed text translation Seq2Seq baseline model, and the Seq2Seq model with attention. We furthermore experimented with the model with and without normalization to check whether normalization has an effect on the models. The following Table 6.1 summarizes the results of our experimental study. It shows the training and testing accuracies and loss and also shows the BLEU score values obtained for each model.

Table 6.1: Comparison of models and the result

| Model   | Training Accuracy | Training Loss | Testing Accuracy | Testing Loss | Precision         |                   |                   |                   | Brevity Penalty (BP) | BLEU Score    |
|---|-------------------|---------------|------------------|--------------|-------------------|-------------------|-------------------|-------------------|----------------------|---------------|
|   |                   |               |                  |              | 1-gram            | 2-gram            | 3-gram            | 4-gram            |                      |               |
|   |                   |               |                  |              | $W_1 = 0.25$      | $W_2 = 0.25$      | $W_3 = 0.25$      | $W_4 = 0.25$      |                      |               |
|   |                   |               |                  |              | $W_1 * \log(P_1)$ | $W_2 * \log(P_2)$ | $W_3 * \log(P_3)$ | $W_4 * \log(P_4)$ |                      |               |
| Seq2Seq Baseline Model without Normalization    | 53.83             | 1.4032        | 27.92            | 1.76         | -0.16229          | -0.323259         | -0.496841         | -0.628076         | 0.6397               | <b>0.1278</b> |
| Seq2Seq Baseline Model + Normalization          | 57.11             | 0.7753        | 31.97            | 1.75         | -0.145237         | -0.204693         | -0.275824         | -0.389159         | 0.573                | <b>0.2077</b> |
| Seq2Seq + Attention without Normalization       | 70.55             | 0.303         | 30.3             | 1.15         | -0.080998         | -0.162399         | -0.252416         | -0.369135         | 0.6876               | <b>0.2895</b> |
| Seq2Seq + Attention + Normalization             | 70.22             | 0.5023        | 31.05            | 1.05         | -0.0689162        | -0.141996         | -0.208556         | -0.292517         | 0.6413               | <b>0.3146</b> |
| Seq2Seq + Teacher Forcing without Normalization | 71.42             | 0.5095        | 37.17            | 0.38         | -0.066960         | -0.1232           | -0.181972         | -0.262455         | 0.595                | <b>0.3154</b> |
| Seq2Seq + Teacher Forcing + Normalization       | 71.57             | 0.4979        | 37.87            | 0.38         | -0.06046          | 0.1232717         | -0.189274         | -0.251089         | 0.6326               | <b>0.3389</b> |

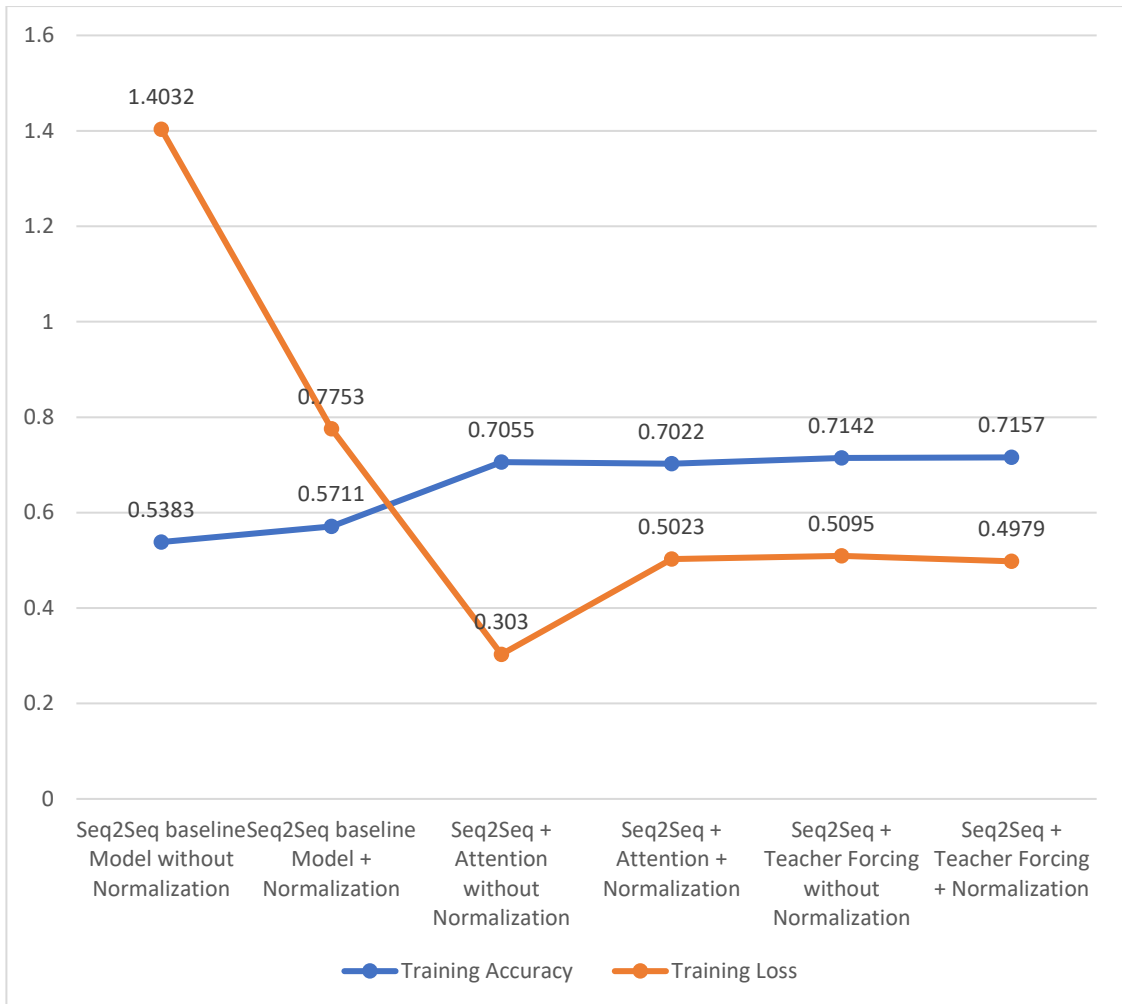


Figure 6.1: Training accuracy and loss of experimented models

If we compare the testing accuracies and loss among the models which are shown in Figure 6.2, our proposed model provides a better score for test accuracy and test loss. Also, it can be seen that the Seq2Seq baseline model, Seq2Seq attention model, and Seq2Seq Teacher Forcing models which have the normalization module performs better than the models without the normalization. This result points out that the normalization of the input dataset could increase the performance of any model.

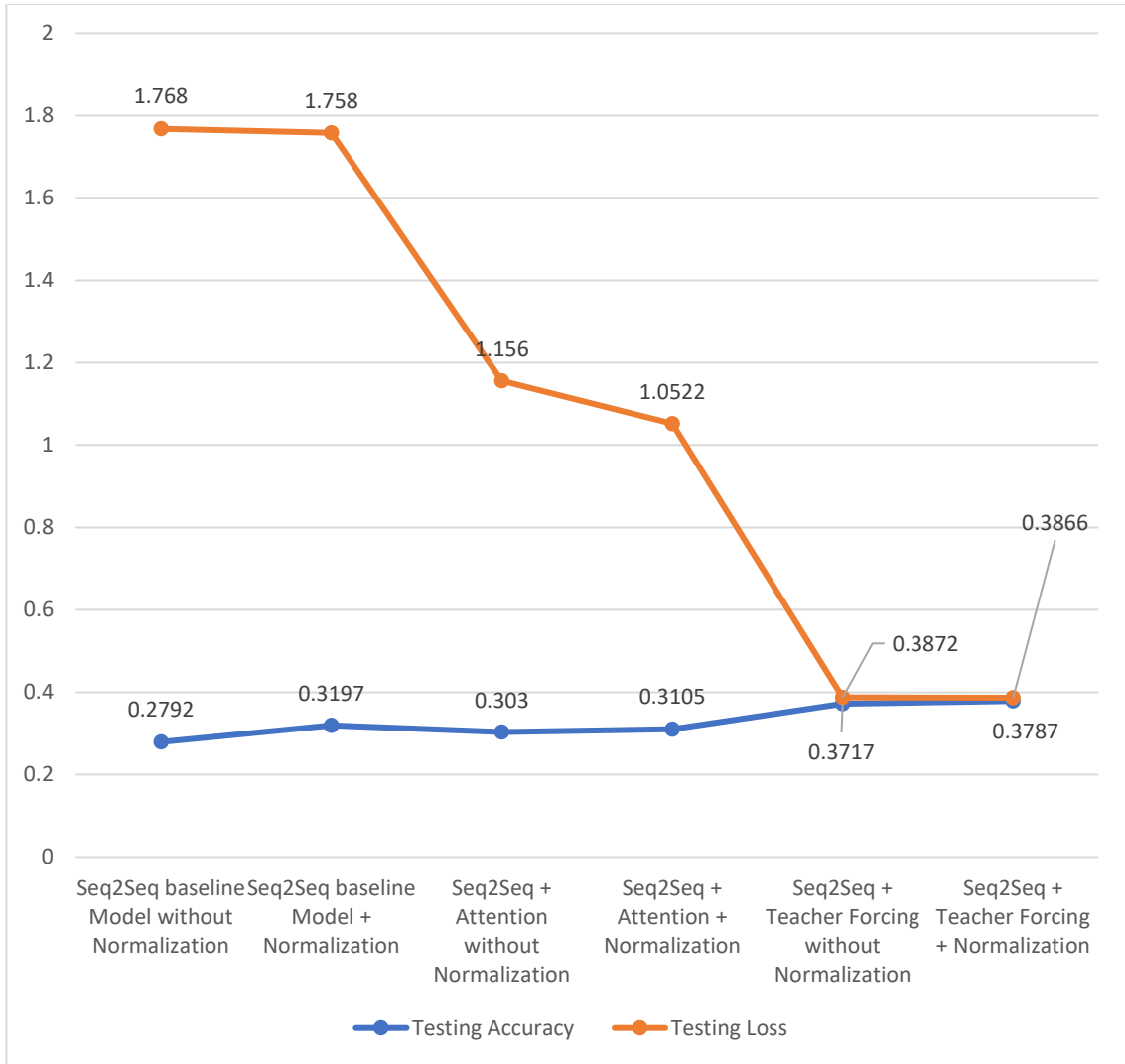


Figure 6.2: Testing accuracy and loss of experimented models

After training the models, we exported the trained models and implemented the inference model. The inference model predicted the Sinhala output sentences for the randomly inputted SECM sentences. Finally, we stored the predicted sentences from each model separately and calculated the BLEU scores as explained in Chapter 5.

The following Figure 6.3 shows the BLEU score values of each of the experimented models. The BLEU score calculation proved that our proposed approach gives a

significantly better BLEU score for Sinhala-English code-mixed text translation compared to the other models.

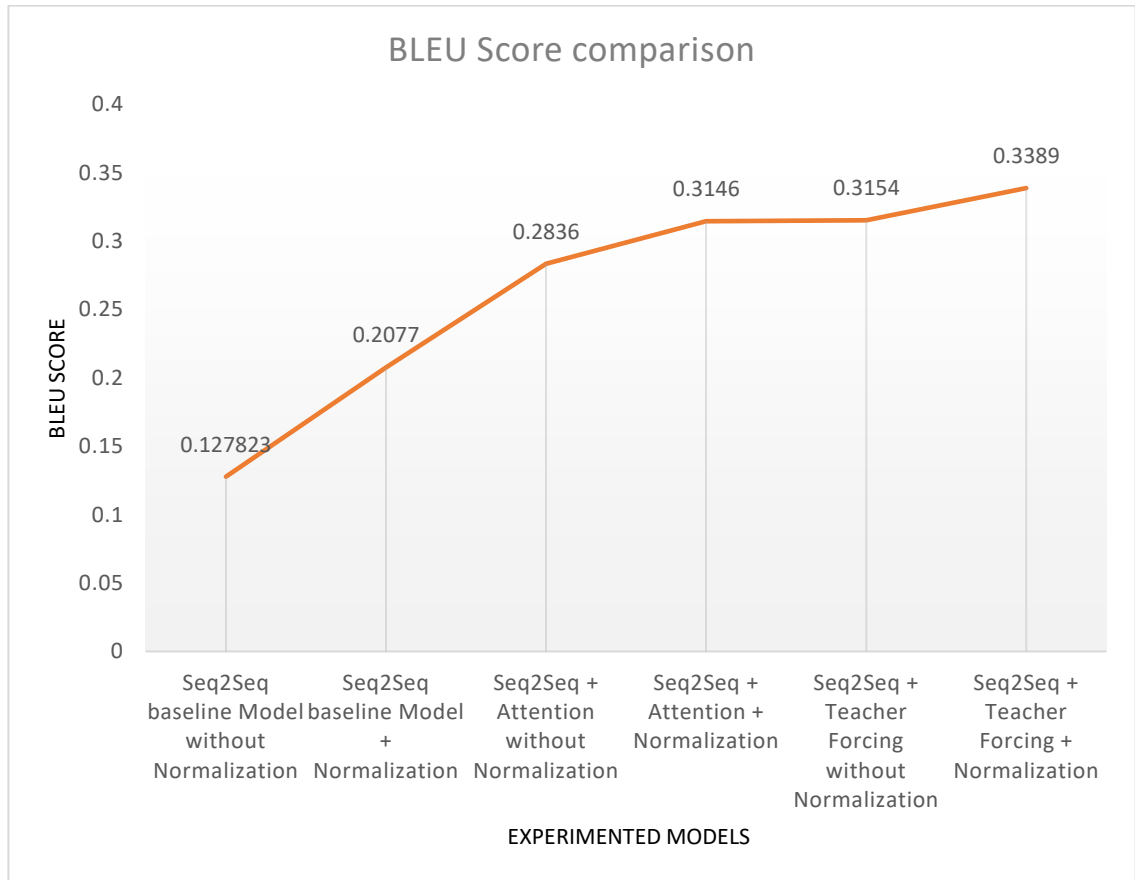


Figure 6.3: Experimented models & BLEU scores

Furthermore, as we explained in Section 5.5, we followed the research study of (Srivasta & Singh,2020), where Hindi-English code-mixed text translation has been conducted. They have applied the randomly chosen 100 sentences from their dataset and applied it to the Bing Translator(BT), Google Translate(GT), and Augmented pipeline with GT models. We tested the same dataset with the same hyper parameter setting and we calculated the BLEU score for the Hinglish-English dataset. The BLEU score received for our model of Seq2Seq with Teacher Forcing is significantly higher than the BLEU scores

received for the other models: Bing Translator(BT), Google Translate(GT) and Augmented pipeline with GT.

Section 1.2.1 points out the challenges that are considered the barrier for Sinhala-English code-mixed text translation. Spelling errors, inconsistent phonetic transliteration, the use of special characters and numeric characters, borrowing of words, integration of suffixes, and switching of discourse markers. Even though our model couldn't provide the solution for all the issues, but most of the challenges can be considered solved with our proposed model.

If the same word has a different transliteration, the normalization module in the model analyses the similar representation of the same word using the Levenshtein edit distant approach and chooses the most frequently used form of representation as to the standard, and converts the specific word to its standard form. This sorts out the issue of transliteration of translation.

The issue of spelling errors and the use of special characters and numeric character issues are sorted using the dictionary-based approaches. The issues of borrowing of words and use of discourse marker issues have been automatically solved by the Seq2Seq approach with the teacher forcing algorithm because the Teacher Forcing algorithm uses the ground truth as the input in each timestep in the decoder. This makes the model identify and learn the correct Sinhala words from the training phase of the model. In the prediction phase, the model provides the output according to what it has learned from the training phase. Section 5.1 provides examples for the outputs which proves that the model has sorted the mentioned issues.

Exporting the implemented model of Seq2Seq with LSTM unit and Teacher Forcing mechanism, we created an SECM to Sinhala translator web application as described in

Section 4.7. The following Figure 6.4 to Figure 6.8 shows the input, output, and the calculated BLEU score from the translator web application.

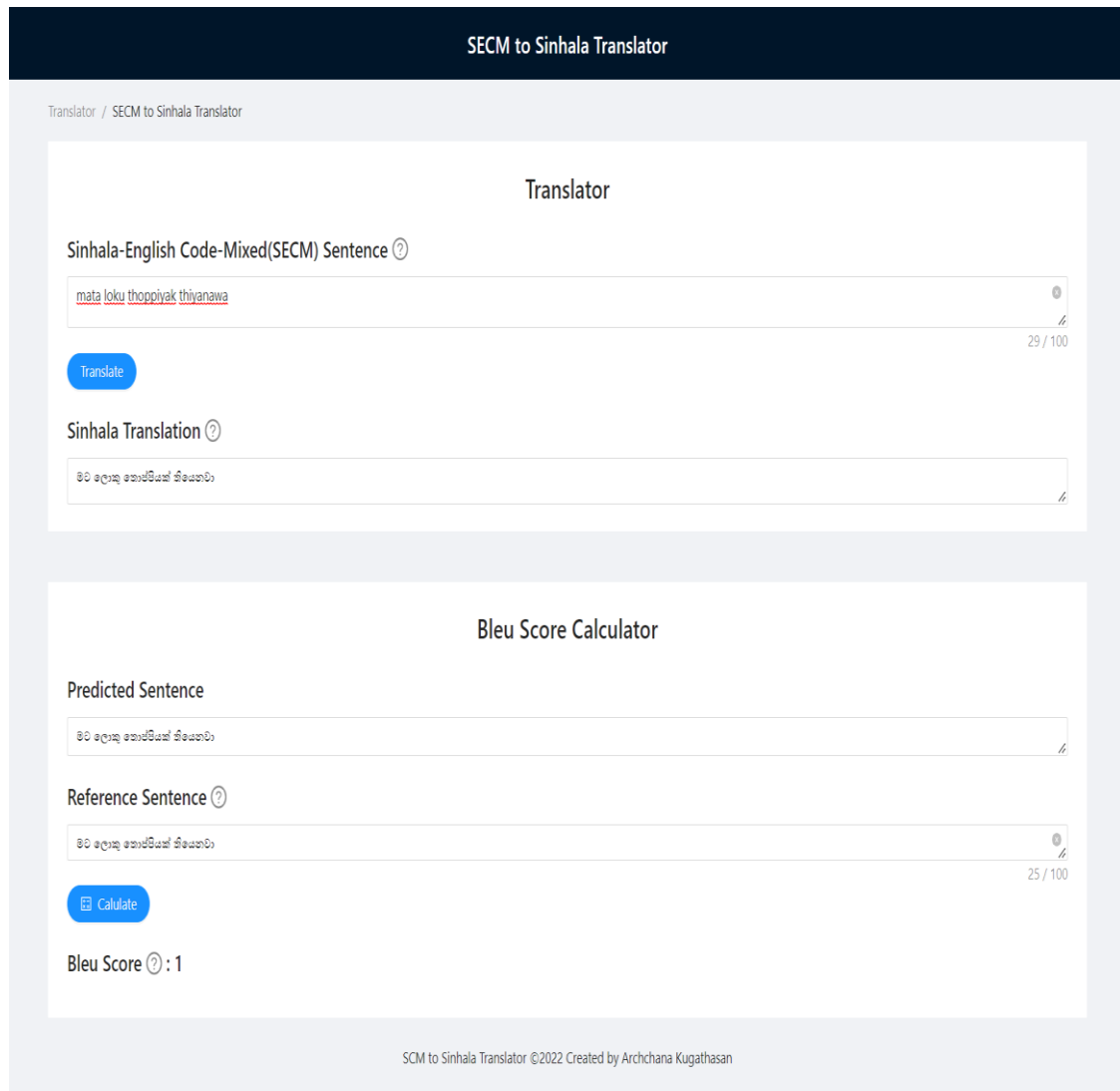


Figure 6.4: Example 1 from the web application SECM to Sinhala translator

In the example shown in Figure 6.4, the given input sentence is ‘*mata loku thoppiyak thivanawa*’, and the predicted Sinhala sentence is ‘මට ලොකු තොප්පියක් තියෙනවා’. The BLEU score calculator gave the calculated BLEU score value as 1, because the predicted translation matches the reference translation 100%.

**SECM to Sinhala Translator**

Translator / SECM to Sinhala Translator

### Translator

Sinhala-English Code-Mixed(SECM) Sentence ?

mata problem godaak 20 / 100

[Translate](#)

Sinhala Translation ?

මම ගැටලුව ගොඩක්

---

### Bleu Score Calculator

Predicted Sentence

මම ගැටලුව ගොඩක්

Reference Sentence ?

මම ගැටලු ගොඩක් 14 / 100

[Calculate](#)

Bleu Score ? : 1.384292958842266e-231

SCM to Sinhala Translator ©2022 Created by Archchana Kugathasan

Figure 6.5: Example 2 from the web application SECM to Sinhala translator

In the example shown in Figure 6.5, the given input sentence is ‘*mata problem godaak*’. The predicted Sinhala sentence is ‘මම ගැටලුව ගොඩක්’ but the expected actual translation is ‘මම ගැටලු ගොඩක්’. Even though the sentence doesn’t provide 100% perfect



translation, the translator has identified the English word ‘problem’ and has matched it to the correct Sinhala word ‘ගැටලුව’ This proves that the ‘Borrowing of words from another language’ issue has been sorted by our proposed translator.

Figure 6.6: Example 3 from the web application SECM to Sinhala translator

In the example shown in Figure 6.6, the given input sentence is ‘*mama fried rice kawa*’, and the predicted Sinhala sentence is ‘මම ආරයිච් රයිස් සි කව්’ but the expected actual

translation is ‘මම ඒරයිච් රයිස් කෑවා’. Even though the sentence doesn’t provide 100% perfect translation, the translated sentence provides a basic idea about the sentence for the reader.

SECM to Sinhala Translator

Translator / SECM to Sinhala Translator

Translator

Sinhala-English Code-Mixed(SECM) Sentence ?

heta mata moon balanna puluwan 30 / 100

Translate

Sinhala Translation ?

හෙට මට සඳ දැක ගත හැකිය

Bleu Score Calculator

Predicted Sentence

හෙට මට සඳ දැක ගත හැකිය

Reference Sentence ?

හෙට මට සඳ දැක ගත හැකිය 22 / 100

Calculate

Bleu Score ? : 1

SCM to Sinhala Translator ©2022 Created by Archchana Kugathasan

Figure 6.7: Example 4 from the web application SECM to Sinhala translator

In the example shown in Figure 6.7, the given input sentence is ‘heta mata moon balanna puluwan’, the predicted Sinhala sentence is ‘හෙට මට සඳ දැක ගත හැකිය’. In this example also, the model has identified the English word ‘moon’ and has translated it correctly

according to the context without any grammatical errors. The BLEU score calculator gave the calculated BLEU score value as 1, because the predicted translation matches the reference translation 100%.

The screenshot displays the 'SECM to Sinhala Translator' web application. It is divided into two main sections: 'Translator' and 'Bleu Score Calculator'.

**Translator Section:**

- Header:** SECM to Sinhala Translator
- Breadcrumb:** Translator / SECM to Sinhala Translator
- Section Title:** Translator
- Input:** Sinhala-English Code-Mixed(SECM) Sentence (?). The input text is 'issara thiyana putuwe waadiwenna' (32 / 100 characters).
- Action:** A blue 'Translate' button.
- Output:** Sinhala Translation (?). The output text is 'ඉස්සරා තියනා පුටුවේ වාඩි වෙන්න'.

**Bleu Score Calculator Section:**

- Section Title:** Bleu Score Calculator
- Input 1:** Predicted Sentence. The text is 'ඉස්සරා තියනා පුටුවේ වාඩි වෙන්න'.
- Input 2:** Reference Sentence (?). The text is 'ඉස්සරා තියනා පුටුවේ වාඩි වෙන්න' (30 / 100 characters).
- Action:** A blue 'Calculate' button.
- Output:** Bleu Score (?): 1.0032743411283238e-231

**Footer:** SCM to Sinhala Translator ©2022 Created by Archchana Kugathasan

Figure 6.8: Example 5 from the web application SECM to Sinhala translator

In the example shown in Figure 6.8, the given input sentence is ‘*issara thiyana putuwe waadiwenna*’, and the predicted Sinhala sentence is ‘අතීතයේ පුටුවට වාඩි වන්න’, but the expected translation is ‘ඉස්සරහ නියෙන පුටුවේ වාඩි වෙන්න’. Even though the translation is not fully correct, the Singlish word ‘*issara*’ has been identified with the meaning of ‘*ancient*’ and presented with the word ‘අතීතයේ’. The meaning is correct but the problem is the meaning of the word ‘*issara*’ in this context doesn’t match. In this context, the correct word should provide the meaning as ‘front’.

## 6.1 Summary

This section a comprehensive overview of the research study's outcomes and discussions, focusing on the proposed model's performance in translating Sinhala-English code-mixed text. The chapter evaluates the model's effectiveness against existing methods, examines its performance across different languages, and discusses the challenges it successfully addresses.

The central objective of the research is to translate Sinhala-English code-mixed text into Sinhala. The evaluation is conducted from two perspectives: comparing the proposed model's performance against existing code-mixed text translation methods and assessing its effectiveness with code-mixed text from another language.

The existing models for code-mixed text translation, including the Seq2Seq baseline model and the Seq2Seq model with attention, are evaluated. The effects of normalization on these models are also explored. A summary of the results, including training and testing accuracies, losses, and BLEU scores, is presented in Table 6.1. Notably, the proposed model exhibits greater test accuracy and loss compared to other models, highlighting its effectiveness.

The model's inference capability is demonstrated by predicting Sinhala output sentences for randomly inputted SECM sentences. The calculated BLEU scores further validate the

proposed model's efficiency, with Figure 6.3 illustrating its significantly better performance compared to other models.

Additionally, the chapter draws parallels with a Hindi-English code-mixed dataset's translation performance. The proposed model's BLEU score outperforms other models like Bing Translator and Google Translate for the same dataset.

The research study addresses several challenges outlined in Section 1.2.1 that hinder Sinhala-English code-mixed text translation. The normalization module tackles issues related to spelling errors, special characters, and numeric characters. The model also effectively handles transliteration variations by utilizing the Levenshtein edit distance approach. The Seq2Seq approach with the teacher forcing algorithm helps overcome borrowing of words and discourse marker issues. The section emphasizes that while not all challenges are entirely resolved, significant progress has been made.

The section concludes with an illustration of the proposed model's performance using the SECM to Sinhala translator web application. Various examples demonstrate the model's capabilities in handling different types of code-mixed sentences, indicating its versatility and adaptability. While some translations may not be entirely accurate, the model maintains context and provides meaningful output, effectively addressing language mixing challenges.

## **7. Conclusion and Future Work**

### **7.1 Summary of achievements**

The main objective of this research study is to build a model to translate Sinhala-English code-mixed sentences. The reason to select this research topic is the need to translate the SECM text, which is the most frequently used language in social media communication, and the unavailability of a translator tool.

As the initial step of this research study, we conducted an in-depth literature study on the research based on code-mixed text. We were able to identify the research gap clearly from the in-depth study. Next, analysis of Sinhala-English code-mixed text is conducted. First, we analyzed the history of Sinhala-English code-mixed text usage and how it started with colonization in the past. Then we conducted a survey to prove the usage of SECM texts in current social media communications. According to the survey results, the most used language in social media communication in Sri Lanka is identified as SECM and it proves the need for a translator for SECM to Sinhala, otherwise, most of the social media text will be left unprocessed and unused.

Next, we focused on analyzing the challenges in the SECM sentences: inconsistent phonetic transliteration, Spelling errors, the use of special characters and numeric characters, borrowing of words, integration of suffixes, and switching of discourse markers. Then we analyzed the Sinhala transliterated format, how it differs from the standard transliteration, and the actual transliteration(Singlish) used in society. We proposed mapping for Sinhala letters(Vowels, basic consonants, and other consonants) with its relevant Singlish mapping according to the frequency of usage.

One of the most important parts of the research is corpus creation. MT systems need a remarkable amount of parallel sentences. Since Sinhala is considered a low-resource

language, we couldn't find an already available corpus. Preparation of the parallel corpus was a very challenging task due to a lack of resources. The SECM sentences are web scrapped from social media sites. We hired a human translator to translate the SECM sentence to Sinhala, which led to parallel corpus creation. Since we can't rely only on the translator, we validated the dataset using the crowd-sourcing approach. Our corpus contained 5000 parallel sentences. We divided the data into 15 groups and conducted the annotation. Each sentence in the corpus was rechecked at least by two people and labelled whether the translation from the human translator is fully correct(FC) or a change is required(CR). The translations labelled as the change required were again checked and changed by the human translator. To measure the correctness of the human-translated sentences, we randomly selected 100 parallel sentences from the prepared corpus and asked some linguistic experts to rate the translation 'good' or 'bad'. Then we calculated the Fleiss Kappa score of 0.88, which states almost the full agreement between the raters. Finally, the parallel corpus was created which can be considered a remarkable achievement with the limited resources.

We developed a normalization module for our proposed system. The normalization module contains the spelling error identification and correction using the dictionary-based approach, slang word normalization using the dictionary-based approach, and transliteration normalization using the Levenshtein edit distant algorithm. According to the experiment results explained in Section 5.4, its seen that when the machine translation models are applied with the normalization module, the performance of the model shows a significant improvement in the accuracies. Furthermore, improvement in the accuracies led to achieving better BLEU scores which can be considered a significant achievement in the proposed approach.

Our proposed approach contains the method of Seq2Seq approach with LSTM basic units and the Teacher Forcing algorithm applied in the decoder phase. In the experimental studies, our proposed approach showed a better performance compared to the current

state-of-the-art models. It provided better accuracy, loss, and BLEU score values. Most of the challenges considered barriers to the translation of SECM to Sinhala translation have been broken from our proposed approach. As explained in Section 5.1 and Section 6, the challenges of borrowing words from the English language, transliteration issues, numerical and special character representation, and spelling error issues are sorted from our proposed model.

Finally, to make our proposed translation model (Seq2Seq with LSTM units and Teacher forcing mechanism) be used easily, we built a web application from our proposed model. We used our trained model in the back end, implemented a function to predict the Sinhala translation for the given SECM sentence, and connected it with the web UI. So anyone could easily use our proposed SECM to Sinhala translator through the web application. Also, it is more understandable for non-expert users to use the model in a web application form.

## **7.2 Limitations**

The main limitation is our research study is the resources for the dataset. We went through a huge process to prepare the parallel corpus of SECM–Sinhala because both are low-resource languages. It was challenging to find a human translator who is both an expert in SECM and Sinhala. After finding a human translator, it took a long time to get the SECM sentences translated.

After the translation, we were in need of validating the translation. So we choose the crowdsourcing approach. Getting the dataset annotated from the crowdsourcing method was a challenging task because the annotators were busy and provided half work done, so we had to send frequent reminders to get the annotation done. Also, to rate the translation, we needed linguistical experts, which was challenging as most of them were busy. This whole task of preparing the dataset was very time-consuming.



Also, there were limitations in the reference materials for the literature survey study. We couldn't access some journal papers, books, or articles from some website as they were not free to download.

Also, the computer used for the implementation of the model doesn't have a GPU. So the training of a model once normally took around a minimum of 72 hours. Furthermore, the tasks such as experimenting with different hyperparameter setting and training with different models for comparison studies was challenging with a low-performance computer. In some situations, the computer stopped working due to a heavy load in the memory and we had to run the whole training of the models from the beginning.

### **7.3 Future Works**

We have described many challenges in SECM sentences in Section 1.2.1. Most of the challenges are sorted from our proposed model. However, still, there are challenges such as Integration of suffixes, Switching of discourse marker which has to be solved. Also, by improving the number of parallel sentences in our corpus we could be able to improve the accuracy. We will make the corpus with the 5000 parallel SECM-Sinhala sentences publicly available and let it improve further. Also, we are planning to get datasets from other languages with code-mixing issues. Therefore, we will try to enhance or modify our approach specifically to the structure of other code-mixed text languages.

## 8. Publications

| Title, Journal or Conference  | Status         | Date       |
|---|----------------|------------|
| <p><b>Conference -</b><br/>Kugathasan, A., &amp; Sumathipala, S. (2020, March). Standardizing sinhala code-mixed text using dictionary based approach. In <i>2020 International Conference on Image Processing and Robotics (ICIP)</i> (pp. 1-6). IEEE.</p>                                 | Published      | March/2020 |
| <p><b>Conference -</b><br/>Kugathasan, A., &amp; Sumathipala, S. (2021, September). Neural Machine Translation for Sinhala-English Code-Mixed Text. In <i>Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)</i> (pp. 718-726).</p> | Published      | Sep/2021   |
| <p><b>Journal -</b><br/>Kugathasan, A. and Sumathipala, S., 2022. Neural machine translation for sinhala-english code-mixed text, <i>International Journal on Advances in ICT for Emerging Regions</i>, Vol 15 No 3 (2022): 2022 December Issue</p>   | Published      | Feb/2023   |
| <p><b>Journal -</b><br/>A Systematic Review of Code-Mixed Text Analysis Approaches. <i>Ampersand, Interdisciplinary Journal of Language Sciences and Bilingualism</i>.</p>  | Major revision | Nov/2023   |

## References

- Al-Shedivat, M., & Parikh, A. P. (2019). Consistency by agreement in zero-shot neural machine translation. *ArXiv Preprint ArXiv:1904.02338*.
- Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., & Cherry, C. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv Preprint ArXiv:1907.05019*.
- Attia, M., Samih, Y., Elkahky, A., Mubarak, H., Abdelali, A., & Darwish, K. (2019). *POS tagging for improving code-switching identification in Arabic*. 18–29.
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ArXiv Preprint ArXiv:1409.0473*.
- Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). “i am borrowing ya mixing?” an analysis of english-hindi code mixing in facebook. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 116–126.
- Barnett, R., Codó, E., Eppler, E., Forcadell, M., Gardner-Chloros, P., Van Hout, R., Moyer, M., Torras, M. C., Turell, M. T., & Sebba, M. (2000). The LIDES Coding Manual: A Document for Preparing and Analyzing Language Interaction Data Version 1.1—July 1999. *International Journal of Bilingualism*, 4(2), 131–271.
- Carrera, J., Beregovaya, O., & Yanishevsky, A. (2009). Machine translation for cross-language social media. *PROMT Americas Inc*.
- Chanda, A., Das, D., & Mazumdar, C. (2016). Unraveling the English-Bengali code-mixing phenomenon. *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, 80–89.
- Chandu, K. R., Chinnakotla, M., Black, A. W., & Shrivastava, M. (2017). WebShodh: A Code Mixed Factoid Question Answering System for Web. In G. J. F. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeriot, T. Mandl, L. Cappellato, & N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (Vol. 10456, pp. 104–111). Springer International Publishing. [https://doi.org/10.1007/978-3-319-65813-1\\_9](https://doi.org/10.1007/978-3-319-65813-1_9)
- Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4), 359–394.
- Cheng, Y. (2019). Semi-supervised learning for neural machine translation. In *Joint training for neural machine translation* (pp. 25–40). Springer.

- Chittaranjan, G., Vyas, Y., Bali, K., & Choudhury, M. (2014). *Word-level language identification using crf: Code-switching shared task report of msr india system*. 73–79.
- Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., & Basu, A. (2007). Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJ DAR)*, 10(3), 157–174.
- Cook, P., & Stevenson, S. (2009). *An unsupervised model for text message normalization*. 71–78.
- Davies, E. E., & Bentahila, A. (2007). CAROL MYERS-SCOTTON, Contact linguistics: Bilingual encounters and grammatical outcomes. *Language in Society*, 36(03). <https://doi.org/10.1017/S0047404507070285>
- Dhar, M., Kumar, V., & Shrivastava, M. (2018). Enabling Code-Mixed Translation: Parallel Corpus Creation and MT Augmentation Approach. *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, 131–140. <https://aclanthology.org/W18-3817>
- Dutta, S., Saha, T., Banerjee, S., & Naskar, S. K. (2015). Text normalization in code-mixed social media text. *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, 378–382. <https://doi.org/10.1109/ReTIS.2015.7232908>
- Firat, O., Sankaran, B., Al-Onaizan, Y., Vural, F. T. Y., & Cho, K. (2016). Zero-resource translation with multi-lingual neural machine translation. *ArXiv Preprint ArXiv:1606.04164*.
- Gambäck, B., & Das, A. (2016). *Comparing the level of code-switching in corpora*. 1850–1855.
- Gu, J., Wang, Y., Cho, K., & Li, V. O. (2019a). Improved zero-shot neural machine translation via ignoring spurious correlations. *ArXiv Preprint ArXiv:1906.01181*.
- Gu, J., Wang, Y., Cho, K., & Li, V. O. (2019b). Improved zero-shot neural machine translation via ignoring spurious correlations. *ArXiv Preprint ArXiv:1906.01181*.
- Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., & Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *ArXiv Preprint ArXiv:1503.03535*.
- Guzman, G. A., Serigos, J., Bullock, B., & Toribio, A. J. (2016). *Simple tools for exploring variation in code-switching for linguists*. 12–20.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., & Corrado, G. (2017). Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5, 339–351.

- Kachru, B. B. (1986). The power and politics of English. *World Englishes*, 5(2–3), 121–140. <https://doi.org/10.1111/j.1467-971X.1986.tb00720.x>
- Kalchbrenner, N., & Blunsom, P. (2013). *Recurrent continuous translation models*. 1700–1709.
- Kugathasan, A., & Sumathipala, S. (2020). Standardizing Sinhala Code-Mixed Text using Dictionary based Approach. *2020 International Conference on Image Processing and Robotics (ICIP)*, 1–6. <https://doi.org/10.1109/ICIP48927.2020.9367353>
- Lafferty, J., McCallum, A., & Pereira, F. C. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*.
- Lakew, S. M., Cettolo, M., & Federico, M. (2018). A comparison of transformer and recurrent neural networks on multilingual neural machine translation. *ArXiv Preprint ArXiv:1806.06957*.
- Lignos, C., & Marcus, M. (2013). *Toward web-scale analysis of codeswitching*. 90.
- Lui, M., & Baldwin, T. (2012). *Langid. Py: An off-the-shelf language identification tool*. 25–30.
- Mandal, S., Das, S. D., & Das, D. (2018). *Language Identification of Bengali-English Code-Mixed data using Character & Phonetic based LSTM Models* (arXiv:1803.03859). arXiv. <http://arxiv.org/abs/1803.03859>
- Masoud, M., Torregrosa, D., Buitelaar, P., & Arčan, M. (2019). *Back-translation approach for code-switching machine translation: A case study*. 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science.
- Neale, M. C., Boker, S. M., Xie, G., & Maes, H. M. (1999). *Statistical modeling*. Richmond, VA: Department of Psychiatry, Virginia Commonwealth University.
- Nguyen, D., & Doğruöz, A. S. (2013). *Word level language identification in online multilingual communication*. 857–862.
- Papalexakis, E., Nguyen, D., & Doğruöz, A. S. (2014). *Predicting code-switching in multilingual communication for immigrant communities*. 42–50.
- Rijhwani, S., Sequiera, R., Choudhury, M. C., & Bali, K. (2016). *Translating codemixed tweets: A language detection based system*. 81–82.
- Senaratne, C. D. (2009). *Sinhala-English code-mixing in Sri Lanka: A sociolinguistic study*. LOT.

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *ArXiv Preprint ArXiv:1511.06709*.

Shanmugalingam, K., Sumathipala, S., & Premachandra, C. (2018). *Word level language identification of code mixing text in social media using nlp*. 1–5.

Singh, R., Choudhary, N., & Shrivastava, M. (2018). *Automatic Normalization of Word Variations in Code-Mixed Social Media Text*.

<https://doi.org/10.48550/ARXIV.1804.00804>

Sudsawad, P. (2007). *Knowledge translation: Introduction to models, strategies and measures*. Southwest Educational Development Laboratory

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27.

Tan, X., Chen, J., He, D., Xia, Y., Qin, T., & Liu, T.-Y. (2019). Multilingual neural machine translation with language clustering. *ArXiv Preprint ArXiv:1908.09324*.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8), 5929–5955. <https://doi.org/10.1007/s10462-020-09838-1>

Veena, P. V., Kumar, M. A., & Soman, K. P. (2017). An effective way of word-level language identification for code-mixed facebook comments using word-embedding via character-embedding. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1552–1556. <https://doi.org/10.1109/ICACCI.2017.8126062>

Volk, M., & Clematide, S. (2014). *Detecting code-switching in a multilingual alpine heritage corpus*. 24–33.

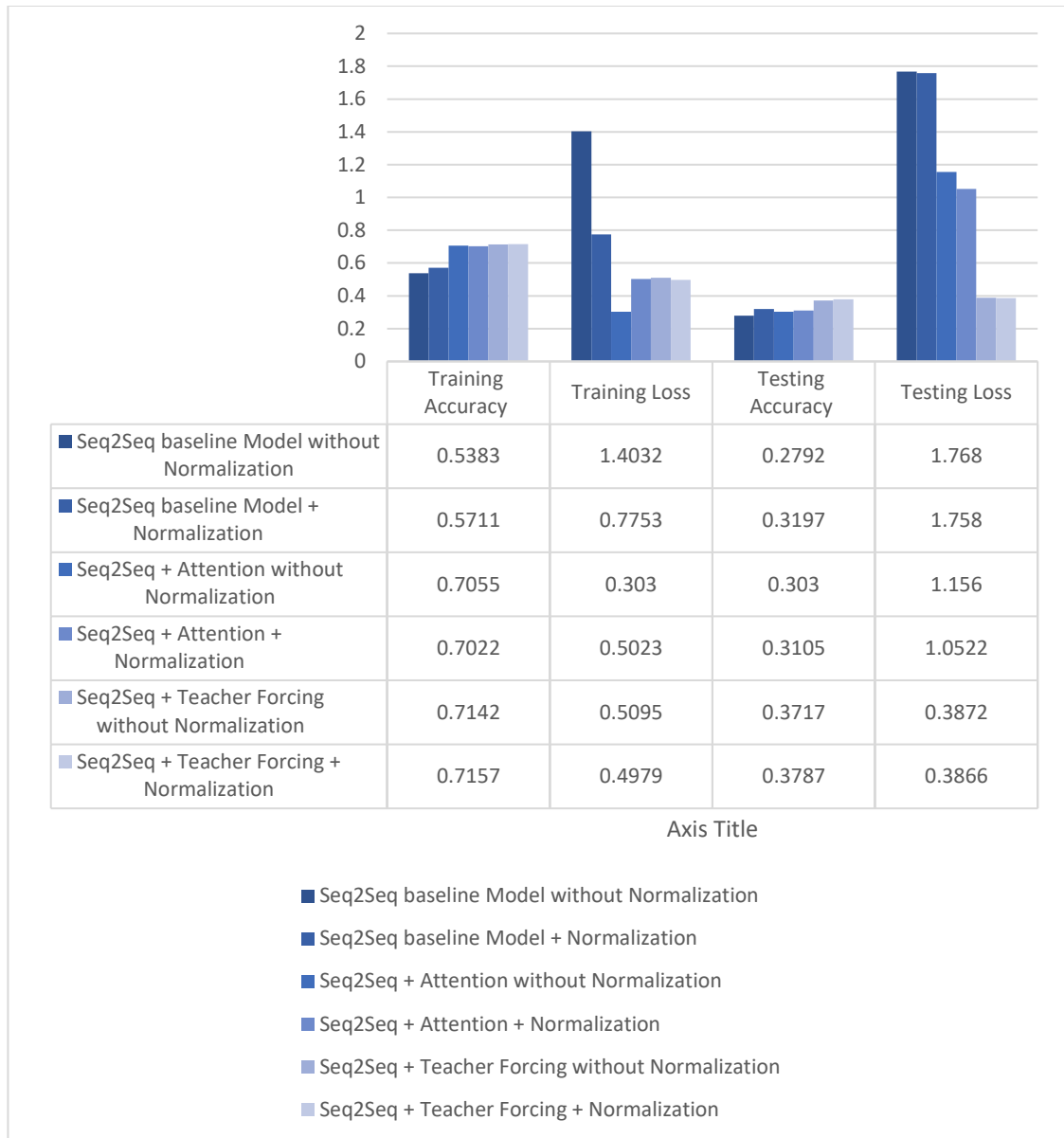
Wong, K.-F., & Xia, Y. (2008). Normalization of Chinese chat language. *Language Resources and Evaluation*, 42(2), 219–242.

Xue, Z., Yin, D., & Davison, B. D. (2011). *Normalizing microtext*. Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence.

Yu, L.-C., He, W.-C., Chien, W.-N., & Tseng, Y.-H. (2013). Identification of code-switched sentences and words using language modeling approaches. *Mathematical Problems in Engineering*, 2013.

Zhang, B., Williams, P., Titov, I., & Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. *ArXiv Preprint ArXiv:2004.11867*.

## Appendix - A : Detailed view of model result comparison





# Appendix - B : Survey questionnaire

Singlish usage among public ☆

Questions Responses 17 Settings

### Singlish code-mixed text analysis

This survey is taken in need of collecting some data regarding Singlish code-mixed text for my MPhil research study.

Please refer the following example how a Sinhala sentence is represented in Singlish

Sinhala - මොකක්ද?  
Singlish - Oya kawada?

Your name? \*

Short answer text

Which type of communication method you often use when communicating through text in social media? \*

Sinhala

Singlish

Other...

If you often use Singlish text, what is the main reason you use Singlish text? \*

You like Singlish.

You like to type in Sinhala language but you use Singlish because it is easy to type

Other...

If we can build a converter which can directly transform Singlish to Sinhala would you like to use that? \*

Yes

No

Which time period did you start using Singlish? \*

1990-1995

1996-2000

2000-2005

2005-2010

2010-2015

Other...

In what kind of platforms you use Singlish? \*

Social Networking sites(Facebook,Tweeter,Instagram etc)

Chat Applications(Whatsapp, Viber, Emo etc)

Community blogs

Discussion Forums

Other...

## Appendix C : Predicted result from the model

```
23_Jan_2021.ipynb X
Code v
.....
6
Input Source sentence: Gana hondai
Actual Target Translation: ගාන හොඳයි
Predicted Target Translation: ගාන හොඳයි
.....
7
Input Source sentence: rice eka wadakma naha
Actual Target Translation: රයිස් එක වැඩිකම නැහැ
Predicted Target Translation: රයිස් එක වැඩිකම නැහැ
.....
8
Input Source sentence: ow
Actual Target Translation: ඔව්
Predicted Target Translation: ඔව්
.....
9
Input Source sentence: indian kama kiyala apiwa rawatanawa
Actual Target Translation: ඉන්දියානු කෑම කියලා අපිට රවට්ටනවා
Predicted Target Translation: කියලා කෑම කියලා
.....
10
Input Source sentence: kama hodaipirisidui
Actual Target Translation: කෑම හොඳයිපිරිසිදුයි
Predicted Target Translation: කෑම
.....
11
Input Source sentence: restaurant crew ekka marama hodai kamath patta
Actual Target Translation: අවන්හල කාර්ය මණ්ඩලය එක්ක මාරම හොඳයි කෑමත් පව්ට
Predicted Target Translation: අවන්හල කාර්ය මණ්ඩලය එක්ක මාරම හොඳයි කෑමත් පව්ට
.....
```

```
23_Jan_2021.ipynb X py36
.....
Predicted Target Translation: කෑම
.....
11
Input Source sentence: restaurant crew ekka marama hodai kamath patta
Actual Target Translation: අවන්හල කාර්ය මණ්ඩලය එක්ක මාරම හොඳයි කෑමත් පව්ට
Predicted Target Translation: අවන්හල කාර්ය මණ්ඩලය එක්ක මාරම හොඳයි කෑමත් පව්ට
.....
12
Input Source sentence: Mama wathura bibi hitiye me post eka baladdi mata wathurath pita ugure giyane
Actual Target Translation: මම බීබී හිටියේ මේ පොස්ට් එක බලද්දී මට වතුර පිට උතුර ගියානේ
Predicted Target Translation: මම මේ කරනවා
.....
13
Input Source sentence: location eka hodai
Actual Target Translation: තැන හොඳයි
Predicted Target Translation: තැන හොඳයි
.....
14
Input Source sentence: Budusaranai
Actual Target Translation: බුදු සරණයි
Predicted Target Translation: බුදු සරණයි
.....
15
Input Source sentence: Sinhala minissu depethata bedala tambiyek janadipathi kenek undata thereitamanta salli hambenawa nam Ratat
a hena gahawath kamak ne kiyala hitha minissuGotabaya hondamai kiyanne nechanda padhanamak thiyena sinhala nayakayek enne echchra
i
Actual Target Translation: සිංහල මිනිස්සු දෙවැන්නට බේදලා නමිනියෙක් ජනාතිපති කෙනෙක් වුනාම තේරෙයි නමිනිට සල්ලි හම්බෙනවා නම් රටට හෙහා හහවත් කම
ක් නෑ කියලා හිතන මිනිස්සුහොටාරිය හොඳමයි කියන්නේ නේ වන්ද පදමක් තියෙන්නේ සිංහල නායකයෙක් ජනතේන් එව්වරයි
Predicted Target Translation: හොඳයි
```

```

23_Jan_2021.ipynb X
Code py36
Predicted Target Translation: හොඳයි
.....
16
Input Source sentence: tuk tuk eke awe hoyaganna amaru una Kamai service ekai patta
Actual Target Translation: ටුක් ටුක් එකේ ආවේ හොයාගන්න ඉවාරු උනා කැමියි හේවාටයි ඉනා හොඳයි
Predicted Target Translation:
.....
17
Input Source sentence: Ado parnse inn paradesak karaya tho gonek unata umbe kiyanna EPA
Actual Target Translation: අඩුවන් සරණයි නව නව ශක්තිය ධාරිතය ලැබෙවා
Predicted Target Translation: මම
.....
18
Input Source sentence: kama rahaikama genna paraku wenawa
Actual Target Translation: කැමි රහයි කැමි හේනන පරක්ක වෙනවා
Predicted Target Translation: කැමි රහයි කැමි එක වෙනවා
.....
19
Input Source sentence: Theruwan saranaithawa thawt shakthiya dhahirryaa labeewa
Actual Target Translation: හෙරුවන් සරණයි නව නව ශක්තිය ධාරිතය ලැබෙවා
Predicted Target Translation: හෙරුවන් සරණයි
.....
20
Input Source sentence: addicted to your food patta rasai
Actual Target Translation: කැමිටි ලැබ්බෙහි වෙලා ඉනා රසයි
Predicted Target Translation: කැමිටි කැමිටි වෙලා ඉනා රසයි
.....
21
Input Source sentence: small family ekak run karanne harima pirusudui
Actual Target Translation: හොඳි පවුල එකක් රන් කරන්නේ හරිම පිරිසිදුයි
Predicted Target Translation: හොඳි පවුල එකක් කරන්නේ හරිම නමයි

```

```

23_Jan_2021.ipynb X
Code py36
22
Input Source sentence: rest karala family eka kala ena honda thanak
Actual Target Translation: හන්සි වෙලා පවුලේ කමියන් එක්ක කාලා එන්න හොඳ කුනක්
Predicted Target Translation: පවුල එක යන්න හොඳ කුනක්
.....
23
Input Source sentence: oppu kranna kiyapan
Actual Target Translation: ඔස්සු කරන්න කියපන්
Predicted Target Translation: කරන්න හොඳ කුනක්
.....
24
Input Source sentence: fight karanna one na as deka loku karala baluwama athi
Actual Target Translation: රන්වු වෙනන ඕනි නෑ ලැස් දෙක ලොකු කරලා බැලුවනම් ලැබී
Predicted Target Translation:
.....
25
Input Source sentence: pasta ekanam supiri
Actual Target Translation: පාස්තා එකනම් සුපිරි
Predicted Target Translation: එකනම් සුපිරි
.....
26
Input Source sentence: eya nm ona kenek ekka fight karaibaya wenna deyak na
Actual Target Translation: එය නම් ඕනා කෙහෙක් එක්ක රන්වු වෙයි බිය වෙනන දෙයක් නෑ
Predicted Target Translation: ඕනා ඕනා එකනම් කන්නා පුළුවන්
.....
27
Input Source sentence: wahane drive krna galapena
Actual Target Translation: වහනේ එලවන්න හැලපෙන
Predicted Target Translation: යන්න හොඳ නෑ
.....
28

```

```

23_Jan_2021.ipynb X
Code py36
28
Input Source sentence: balalama ganna tibbanam hondai
Actual Target Translation: බලලම ගන්න නීතිවනම් හොඳයි
Predicted Target Translation: ගන්න ගන්න යන්න තැනක්
.....
29
Input Source sentence: kama rasai bada pirena kawa
Actual Target Translation: කෑම රසයි බඩ පිරෙන්න කැවෑ
Predicted Target Translation: කෑම රසයි බඩ පිරෙන්න කැවෑ
.....
30
Input Source sentence: Budusaranai puta
Actual Target Translation: බුදු සරණයි පුතා
Predicted Target Translation: බුදු සරණයි
.....
31
Input Source sentence: lassana atmosphere ekakkama rahai
Actual Target Translation: ලස්සන පරිසරයක් කෑම රහයි
Predicted Target Translation: ලස්සන පරිසරයක් කෑම රහයි
.....
32
Input Source sentence: price wedikama raha na
Actual Target Translation: ගාන වැඩියිකෑම රහ නෑ
Predicted Target Translation: ගාන හොඳවම වැඩියි
.....
33
Input Source sentence: Aye waliyakd manda
Actual Target Translation: ආයේ වලිකක්ද මන්දා
Predicted Target Translation: ආයේ ආයේ යන්න මිනි
.....
34

```

```

23_Jan_2021.ipynb X
Code py36
.....
34
Input Source sentence: Ponnoyoth nilamelata adinawane akai
Actual Target Translation: පොන්නමයන් නිලමම දේවනවා එකයි
Predicted Target Translation:
.....
35
Input Source sentence: supiri wada meyage ganath adui
Actual Target Translation: සුපිරි වැඩියා මෙයාගේ ගැනක් අඩුයි
Predicted Target Translation: සුපිරි හරිම හොඳයි
.....
36
Input Source sentence: good but kama samahara welawata honda na
Actual Target Translation: හොඳ හැබෑයි කෑම සමහර මෙලාවට හොඳ නෑ
Predicted Target Translation: හොඳ හැබෑයි කෑම සමහර සමහර හොඳ නෑ
.....
37
Input Source sentence: BUDU SARSMAI DEVI PIHITAI
Actual Target Translation: බුදු සරණයි දෙවි පිහිටයි
Predicted Target Translation: බුදු සරණයි දෙවි පිහිටයි
.....
38
Input Source sentence: kma denna puluwang
Actual Target Translation: කෑම දෙනන පුළුවන්
Predicted Target Translation: කෑම දෙනන පුළුවන්
.....
39
Input Source sentence: mama order karapu nathi dewaluth bill ekata add vela
Actual Target Translation: මම ඇණවුම් නොකළ දේවලුන් බිලට එකතු වෙලා
Predicted Target Translation: මම ඇණවුම් නෑ

```

```
23_Jan_2021.ipynb X
+ 🔍 📄 ▶ ■ 🔄 Code v py36 ○
39
Input Source sentence: mama order karapu nathi dewaluth bill ekata add vela
Actual Target Translation: මම ඇණවුම් තොරතුරු දෙවෙනස් කිරීමට එකතුවෙලා
Predicted Target Translation: මම ඇණවුම් නෑ
.....
40
Input Source sentence: kama echchra special naha
Actual Target Translation: කෑම එවීමේ විශේෂ නෑහැ
Predicted Target Translation: කෑම එවීමේ විශේෂ නෑහැ
.....
41
Input Source sentence: udama kiyala thiyena eka kiyopan booruwa
Actual Target Translation: උඩම කියල නිසා එක කියපත් බුරුවා
Predicted Target Translation: හාන වැඩියි කියලා
.....
42
Input Source sentence: Senga wadi nisa nidahasa tikak adui
Actual Target Translation: සෙනහ වැඩියි නිසා නිදහසටිකක් අඩුයි
Predicted Target Translation: සෙනහ වැඩියි නිසා නිසා
.....
43
Input Source sentence: Meya hithan inne I phone thiyenne photo ganna witarai kiyala
Actual Target Translation: මේයා හිතත් ඉන්නේ අයි මොත් නිසෙන්නේ මොවො හත්තා විතරයි කියන්නේ කියලා
Predicted Target Translation:
.....
44
Input Source sentence: Umba koheda ude hawasa yanne
Actual Target Translation: උඹි මතුවෙල උදේ හවස යන්නේ
Predicted Target Translation: යන්නා ඕනි නෑ
.....
45
```

```
23_Jan_2021.ipynb X
+ 🔍 📄 ▶ ■ 🔄 Code v py36 ○
Predicted Target Translation: හොඳම කෑම
.....
45
Input Source sentence: rasama pisa kawe methanin
Actual Target Translation: මේමි රසවත්ම පීසා කෑම
Predicted Target Translation: නෑ පීසා මොඳයි
.....
46
Input Source sentence: high ganan
Actual Target Translation: මොඩාක් ගණන්
Predicted Target Translation: මොඩාක් ගණන්
.....
47
Input Source sentence: Mokakda bn me siddiya
Actual Target Translation: මොකක්ද බන් මේ සිද්දිය
Predicted Target Translation: මොකක්ද බන් මේ
.....
48
Input Source sentence: dawalta senaga godaak wadi
Actual Target Translation: දවල්ට සෙනහ මොඩාක් වැඩියි
Predicted Target Translation: කමයි සෙනහ මොඩාක් වැඩියි
.....
49
Input Source sentence: hondama kama
Actual Target Translation: හොඳම කෑම
Predicted Target Translation: හොඳම කෑම
.....
50
Input Source sentence: gully smell inside the resturant pirisudu madi
Actual Target Translation: ගලී ගඟක් එතරා අවන්හල ඇතුළේ පිරිසිදු මදි
Predicted Target Translation: අවන්හල එක කමයි අවන්හල එක සුපිටි
.....
Mode: Command Ln 1, Col 1 23_Jan_2021.ipynb
```

```
23_Jan_2021.ipynb X
+ 🔍 📄 ▶ ⏪ ⏩ Code py36
51
Input Source sentence: order karala godak wela balagena inna oni
Actual Target Translation: ඔබට කරලා ගොඩනැක් වෙලාව බලාගෙන ඉන්න ඔබ්
Predicted Target Translation: ඔබට කරලා ගොඩනැක් වෙලාව ඉන්න ඔබ්
.....
52
Input Source sentence: Ai wenida padiya
Actual Target Translation: අයි වෙනිදා පඩිය
Predicted Target Translation:
.....
53
Input Source sentence: Gana tikak wadi
Actual Target Translation: ගාන ටිකක් වැඩි
Predicted Target Translation: ගාන ටිකක් වැඩි
.....
54
Input Source sentence: Khmd genna gnne
Actual Target Translation: කොහොමද ගන්නා ගන්නේ
Predicted Target Translation: කොහොමද කොහොමද
.....
55
Input Source sentence: siraaa neh
Actual Target Translation: සිරා නේ
Predicted Target Translation: සිරා නේ
.....
56
Input Source sentence: balan next episode eka enakam
Actual Target Translation: බලන් ඊළඟ කතාවය එක එකකන්
Predicted Target Translation: හොඳ නෑ කෑමින් රසයි
.....
57
Mode: Command Ln 1, Col 1 23_Jan_2021.ipynb
```

```
23_Jan_2021.ipynb X
+ 🔍 📄 ▶ ⏪ ⏩ Code py36
.....
57
Input Source sentence: ona tharam kanna puluwang price ekath shape
Actual Target Translation: ඕන කරම් කන්න පුළුවන්ග් ගානක් ගේජ්
Predicted Target Translation: ඕන කරම් කන්න පුළුවන්ග් ගානක් ගේජ්
.....
58
Input Source sentence: kathawa lassanai
Actual Target Translation: කතාව ලස්සනයි
Predicted Target Translation: කතාව ලස්සනයි
.....
59
Input Source sentence: supiri porak
Actual Target Translation: සුපිරි පොරක්
Predicted Target Translation: සුපිරි
.....
60
Input Source sentence: lunch ganna yanna late unama mehema thamai
Actual Target Translation: දවල් කෑම් ගන්නා යන්නා පරක්කුටුණාම් මේහෙම් නම්යි
Predicted Target Translation: අපි කෑම් ගන්නා යන්නා හොඳයි
.....
61
Input Source sentence: Gaana wadi
Actual Target Translation: ගාල වැඩියි
Predicted Target Translation: ගාන වැඩියි
.....
62
Input Source sentence: mkkda bn seen eka
Actual Target Translation: මොකක්ද බන් සීන් එක
Predicted Target Translation: මොකක්ද බන්
Mode: Command Ln 1, Col 1 23_Jan_2021.ipynb
```

```

23_Jan_2021.ipynb X
+ 🔍 📄 ▶ ⏪ ⏩ Code py36
Actual Target Translation: මොකක්ද හන් කන් ජක
Predicted Target Translation: මොකක්ද හන්
.....
63
Input Source sentence: main road eka laga nisa noisy
Actual Target Translation: ජරධනා හර ලහ නියා සර්ද වැඩියි
Predicted Target Translation:
.....
64
Input Source sentence: I ordered a large portion and a regular one kisima wenasak naha boru karanne
Actual Target Translation: මන් ඕඩර් කලා ලොකු පෝහන් හා රේගුලර් එකක් කිසිම වෙනසක් නෑ බොරු කරන්නේ
Predicted Target Translation: මන් ඕඩර් ලොකු ලොකු ලොකු හා කිසිම එකක්
.....
65
Input Source sentence: hari shok thanagodaak lassanai
Actual Target Translation: හරි හෝක් තුනක්වොඩාක් ලස්සනයි
Predicted Target Translation: හරි හරි ලස්සනයි
.....
66
Input Source sentence: mama oyalage salad ekata marama kamathiy good luck
Actual Target Translation: මම ඔයාලගේ සැලඬි එකට මාරම කැමතියි සුඹ සැකුම
Predicted Target Translation: මම මන් මාරම හොඳයි
.....
67
Input Source sentence: Marama lassanai
Actual Target Translation: මාරම ලස්සනයි
Predicted Target Translation: මාරම ලස්සනයි
.....
68
Input Source sentence: godaak Ganan kama Very crowded
Actual Target Translation: හොඩාක් ගණන් කෑම හොඩාක් සෙසාන වැඩි
.....
Mode: Command Ln 1, Col 1 23_Jan_2021.ipynb

```

```

23_Jan_2021.ipynb X
+ 🔍 📄 ▶ ⏪ ⏩ Code py36
.....
68
Input Source sentence: godaak Ganan kama Very crowded
Actual Target Translation: හොඩාක් ගණන් කෑම හොඩාක් සෙසාන වැඩි
Predicted Target Translation: හොඩාක් ගණන් කෑම හොඩාක් සෙසාන වැඩි
.....
69
Input Source sentence: place ekath super cleankamath hodai
Actual Target Translation: තුනක් සුපිරි පිරිසිදුකෑමක් හොඳයි
Predicted Target Translation: සුපිරි තුන
.....
70
Input Source sentence: Mama mage hubby ta wenna habbu kaale anayak wela madie kiyalamai hithenne oyage post eka dakkama aparaaadee
Actual Target Translation: මම මගේ හබ්බ් වෙන්න හබ්බු කාලේ දැහැන් වෙලා මිදි කියලමයි හිතන්නේ ඕයාගේ පොස්ට් එක දැක්කම් අපරාදේ
Predicted Target Translation: මම
.....
71
Input Source sentence: kama raha unta gana hondatama wadieh gaanata worth na
Actual Target Translation: කෑම රහ උනාට හානා හොඳම වැඩියි ඒ හානාට වරින්ගේ නෑ
Predicted Target Translation: කෑම රහ උනාට හානා හොඳම වැඩියි ඒ හානාට වරින්ගේ නෑ
.....
72
Input Source sentence: sialli walata ne hati tata wenumen
Actual Target Translation: සල්ලි වලට හේ හැටි වෙනුවෙන්
Predicted Target Translation: සල්ලි වලට හේ
.....
73
Input Source sentence: biriyani eke price eka wadi
Actual Target Translation: බිරියානි එකේ ගාන වැඩියි
Predicted Target Translation: බිරියානි එකේ ගාන වැඩියි
.....

```

```

23_Jan_2021.ipynb X
Code py36
74
Input Source sentence: price eka tikak wadi kama eka good
Actual Target Translation: ගාන එකක් වැඩියි කැමි එක හොඳ
Predicted Target Translation: ගාන එකක් වැඩියි කැමි එක හොඳ
.....
75
Input Source sentence: prawns witharak rahai
Actual Target Translation: ඉස්සෝ එතරක් රහයි
Predicted Target Translation: ඉස්සෝ ඉස්සෝ
.....
76
Input Source sentence: Price eka tikak wadiCustomer service eka madi Staff eka thawa improve wenna one
Actual Target Translation: මිල එකක් වැඩියි පාරිභෝගික සේවය මදි කාර්ය මණ්ඩලය වැඩි දියුණු කළ යුතුයි
Predicted Target Translation: මිල එකක් වැඩියි හැඟිලි කාර්ය මණ්ඩලය වැඩි
.....
77
Input Source sentence: Oyage kathawa koitaram lissanada kiyala kiyanna mata wachana naaoya aaima kathaawagenakal maga balagena hit iye Shashi
Actual Target Translation: ගියාගේ කතාව කතාබහරම් ලස්සනද කියලා කියන්න මම වචන හැ
Predicted Target Translation: කතාව කතාව කියලා කියන්න මම කියන්න රහයි
.....
78
Input Source sentence: Colour light eka hinda
Actual Target Translation: වර්ණ ආලෝකය එක චිත්තු
Predicted Target Translation: එක එක හොඳයි
.....
79
Input Source sentence: hodama company
Actual Target Translation: හොඳම කොම්පනිය
Predicted Target Translation: හොඳම
.....
Mode: Command Ln 1, Col 1 23_Jan_2021.ipynb

```

```

23_Jan_2021.ipynb X
Code py36
Predicted Target Translation: හොඳම
.....
80
Input Source sentence: ganan wadi
Actual Target Translation: ගාන වැඩියි
Predicted Target Translation: ගණන් වැඩියි
.....
81
Input Source sentence: spicy chiness food walata super resturant ekak
Actual Target Translation: සැර වයිනිස් කෑම වලට සුපිරි අවන්හල එකක්
Predicted Target Translation: වයිනිස් කෑම වලට සුපිරි අවන්හල එකක්
.....
82
Input Source sentence: menu card eke thiyena godak items naha
Actual Target Translation: මෙහිලා කාඩ් එකක් එකක් නියත ගොඩාක් දේවල් හැහැ
Predicted Target Translation: එකක් එකක් එකක් නියත ගොඩාක් හැහැ
.....
83
Input Source sentence: clean na kanna hithannagananam adui
Actual Target Translation: පිරිසිදු හැ කන්න භික්කන් ගණන්හම් අඩුයි
Predicted Target Translation: පිරිසිදු හැ කන්න හැ
.....
84
Input Source sentence: pattama chiness resturant eka
Actual Target Translation: පට්ටම් වයිනිස් අවන්හල එක
Predicted Target Translation: වයිනිස් අවන්හල එක
.....
85
Input Source sentence: aduma tarame ilaga akawath dipan
Actual Target Translation: අඩුම කරමි ඵලක එකවත් දීපන්
Predicted Target Translation: අන්තර්ගතය
.....
Mode: Command Ln 1, Col 1 23_Jan_2021.ipynb

```



```

23_Jan_2021.ipynb X
Code py36
Predicted Target Translation: යන්න හොඳ නෑ
.....
86
Input Source sentence: Great place for dinner and fresh juicesprice ekath resonable
Actual Target Translation: හොඳම තැන රූ කෑමට හා නැවුම් ජූස්වහාත් ටිසනාකිල්
Predicted Target Translation: හොඳම තැන හා සුපිරි
.....
87
Input Source sentence: owa biwama kata kaaranawa thibaha wedi wenawa
Actual Target Translation: ඕවා බීච්චම් කට කාරනාවා තිබහ වැඩි වෙතවා
Predicted Target Translation: නෑ
.....
88
Input Source sentence: Ape weerayata bubusaranai
Actual Target Translation: අපේ වීරයාට බුදු සරණයි
Predicted Target Translation: අපේ බුදු සරණයි
.....
89
Input Source sentence: nikan nikan hode mathak une ehma nee
Actual Target Translation: නිකන් නිකන් හොඳ මනක් උනේ එහෙමි නේ
Predicted Target Translation: හොඳ නෑ කෑම
.....
90
Input Source sentence: Budu saranai dewi pihitai
Actual Target Translation: බුදු සරණයි දෙවි පිහිටයි
Predicted Target Translation: බුදු සරණයි දෙවි පිහිටයි
.....
91
Input Source sentence: mn recommend karana thanak
Actual Target Translation: මන් නිර්දේශ කරන නැනක්
Predicted Target Translation: මන් නිර්දේශ කරන නැනක්
.....
Mode: Command Ln 1, Col 1 23_Jan_2021.ipynb

```

```

23_Jan_2021.ipynb X
Code py36
.....
91
Input Source sentence: mn recommend karana thanak
Actual Target Translation: මන් නිර්දේශ කරන නැනක්
Predicted Target Translation: මන් නිර්දේශ කරන නැනක්
.....
92
Input Source sentence: Ammo oi
Actual Target Translation: අම්මෝ ඔයි
Predicted Target Translation: අම්මෝ එක
.....
93
Input Source sentence: I orderd a cup of tea cup eka hariyata wash karalath naha
Actual Target Translation: මම නේ කන්ජපයක් ඕඩ් එක හරියට හෝදලා කරලන් නෑ
Predicted Target Translation: මම ඕඩ්
.....
94
Input Source sentence: kama hondaigana shape
Actual Target Translation: කෑමි හොඳයි හාන අේජ්
Predicted Target Translation: කෑමි හොඳයි හාන අේජ්
.....
95
Input Source sentence: place eka super clean
Actual Target Translation: තැන සුපිරි පිරිසිදුයි
Predicted Target Translation: තැන සුපිරි පිරිසිදුයි
.....
96
Input Source sentence: order karala godak wela wait karanna oni
Actual Target Translation: ඕඩ් කරලා හොඩාක් වෙලාවේ ඉන්න ඕනි
Predicted Target Translation: ඕඩ් කරලා හොඩාක් වෙලාවේ ඉන්න ඕනි
.....

```

```
23_Jan_2021.ipynb X
Code py36
.....
95
Input Source sentence: place eka super clean
Actual Target Translation: තැන සුපිරි පිරිසිදුයි
Predicted Target Translation: තැන සුපිරි පිරිසිදුයි
.....
96
Input Source sentence: order karala godak wela wait karanna oni
Actual Target Translation: ඕඩර් කරලා ගොඩාක් වෙලාව ඉන්න ඕනි
Predicted Target Translation: ඕඩර් කරලා ගොඩාක් වෙලාව ඉන්න ඕනි
.....
97
Input Source sentence: melo rahak nafamily ekan yana ba
Actual Target Translation: මෙලෝ රහක් හැ පවුල එක්කත් යන්න බෑ
Predicted Target Translation: මෙලෝ රහක් හැ පවුල බෑ
.....
98
Input Source sentence: Lasana kathawakMan Kalpana kale film ekak unanam kauda main character karanna hoda kiyala
Actual Target Translation: ලස්සන කතාවක් මත කල්පනා කලේ සිල්ම් එකක් උනානම් ආකෘද් ප්රධාන චරිතය කරන්නා හොඳ කියලා
Predicted Target Translation: ලස්සන අවිනාල එක
.....
99
Input Source sentence: Kama rasaiservice hodai
Actual Target Translation: කෑම් රසයිසේවාට් හොඳයි
Predicted Target Translation: කෑම් රසයි සේවාට් හොඳයි
.....
100
Input Source sentence: wadanam supiri
Actual Target Translation: වැඩනම් සුපිරි
Predicted Target Translation:
```