Neural Machine Translation for Sinhala-English Code-Mixed Text

Archchana Kugathasan[#1], Sagara Sumathipala

***Abstract*— Multilingual societies use a mix of two or more languages when communicating. It has become a famous way of communication in social media in South Asian communities. Sinhala-English Code-Mixed Texts (SCMT) are known as the most popular text representation used in Sri Lanka in the informal context such as social media chats, comments, small talks etc. The challenges in utilizing the SCMT sentences are addressed in this paper. The main focus of this study is translating code-mixed sentences written in Sinhala-English to the standard Sinhala language. Since Sinhala is a low-resource language, we were able to collect only a limited number of SCMT-Sinhala parallel sentences. Creating the parallel corpus of SCMT-Sinhala was a time-consuming and costly task. The proposed architecture of Neural Machine Translation(NMT) to translate SCMT text to Sinhala, is built with a combination of normalization pipeline, Long Short Term Memory(LSTM) units, Sequence to Sequence(Seq2Seq) and Teachers Forcing mechanism. The proposed model is evaluated against the current state-of-the-art models using the same experimental setup, which proves the Teacher Forcing Algorithm combined with Seq2Seq and Normalization improves the quality of the translation. The predicted outputs from the model are compared using the BLEU (Bilingual Evaluation Understudy) metric and our proposed model achieved a better BLEU score of 33.89 in the evaluation.**

***Keywords*— Neural Machine Translation, LSTM, Seq2Seq, Sinhala-English Code-Mixed**

I. INTRODUCTION

Code-mixing has been a practice in multilingual communities. In a given sentence, if the elements of one language such as terms, morphemes and words are mixed with the elements of another language, it is called as code-mixing. Lexicon and syntactic formulation from two different languages are combined to generate a code-mixed sentence [1]. The communities which use more than one language for communication are called multilingual communities. Most Srilankans are multilingual people who speak Sinhala-English, Tamil-English, Malay-English, etc. Several research studies have proven that multilingual communities use online social media as the chosen platform to express their opinions and feelings [2].

Posts, comments, reviews etc., are considered user-generated texts in social media. Information extraction from user-generated text has great demand when it comes to business. Analysing the sentiment, extracting the entities,

Correspondence: Archchana Kugathasan (E-mail: archchanakugathasan@gmail.com) Received: 20.12-2021 Revised:07-11-2022 Accepted: 14-11-2022

Archchana Kugathasan, from Sri Lanka Institute of Information Technology and Sagara Sumathipala from University of Moratuwa, Sri Lanka (archchanakugathasan@gmail.com, sagaras@uom.lk).

DOI: http://doi.org/10.4038/icter.v15i3.7250

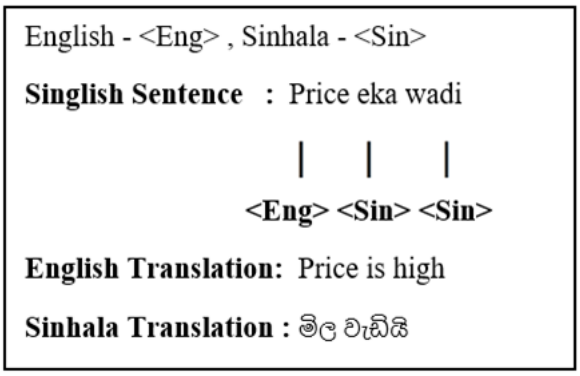

Fig.1 Example of Sinhala-English code-mixed text with language tags

identifying the user interest and providing personalized content for users has become a trending protocol followed when it comes to business marketing strategies using social media [3, 4]. Code-mixing has been identified as a barrier on utilizing user-generated texts for processing due to the mixing of languages. The need of the translation of code-mixed texts to a standard language has been a requirement for a long time. Due to the increasing amount of usage of SCMT in social media, there is a huge demand nowadays to translate SCMT into the Sinhala language. The focal point of this research study is to translate Sinhala-English Code-Mixed (SCM) sentence into a Sinhala sentence. Currently, available translation systems are not very successful in translating code-mixed texts to a standard language [5].

Code-mixed sentence of Sinhala-English has the syntax of the Sinhala language but borrow a few vocabularies from English. Figure 1 shows an example of Sinhala code-mixed text, where the word 'Price' is an English word, 'eka' and 'wadi' are transliterated Sinhala words. Transliteration is the process where a word from one language is represented using the alphabet of another language. The words 'eka' and 'wadi' are words from the Sinhala language written with the English alphabet.

Translating SCMT into Sinhala is a formidable task. The major challenge is the implementation of a Machine Translation system needs a parallel corpus [6]. This sort of dataset is typically available for standard languages, and for SCMT, there is no available data resource. Due to this issue an SCMT - Sinhala parallel corpus is built in this study. Also, this paper discusses a detailed analysis of SCMT and proposes an approach to using and adopting the prevailing models with the goal to translate SCMT to the Sinhala language. The basic architecture of the proposed model is a Neural Network model which includes the combination of normalization, Seq2Seq,

LSTM and Teacher Forcing mechanism [7]. Capability to learn temporal dependencies is very successful in LSTM [8].

The Seq2Seq model is chosen because it can map the sequence of different lengths of source and target sentences to each other [9]. The Teacher forcing mechanism is applied in the decoding phase of the Seq2Seq model to fasten the training and reduce the prediction errors. Finally, the inference model will predict the Sinhala translation for the given SCMT. BLEU evaluation metric is used to evaluate the model.

The rest of the paper is divided into the following sections: Initiated with a study on the groundwork of the research area of Normalization and Machine Translation in section II. The next section discusses code-mixing in Sri Lanka. It provides details about the challenges in SCM sentences and usage of code-mixed text in Sri Lanka. Section IV discusses the parallel corpus preparation and its features. Section V &VI includes detail such as the system architecture, model, experimental setting and the obtained result. Section VII describes the evaluation study and discusses the results, and Section VIII concludes with the conclusion.

TABLE I

SURVEY RESULT - USAGE OF SINHALA-ENGLISH CODE-MIXED TEXT

Questions in Survey	Answer options	Response Percentage
Communication method often used when communicating through text in social media platforms or other online platforms?	Using Sinhala-English code-mixed text in social media	85.2%
	Using native language in social media	8.5%
	other	6.3%
What is the main reason to use Sinhala-English code-mixed text?	Using Sinhala-English code-mixed text because of easiness/flexibility with the keyboard	78.0%
	Interested in using Sinhala-English code-mixed text	12.2%
	Other	10.0%
In what kind of platforms you use Sinhala-English code-mixed text?	Social Networking sites(Facebook, Twitter, Instagram etc.)	59.80%
	Chat Applications(WhatsApp, Viber, Emo etc)	93.90%
	Community blogs	8.50%
	Discussion Forums	7.30%
	Other	1.20%

II. RELATED WORK

A. Normalization of Code-Mixed Text

The rapid growth of user-generated texts in social media allures researchers to focus on the normalization domain. Normalization of the code-mixed texts could lead the models to improve their accuracy. The first corpus for normalization was introduced by Wong and Xia et al. (2008) [10]. Source Channel Model, which finds the most suitable translation based on probability and phonetic mapping, is used to normalize the corpus text. Furthermore, this model was improved by Xue et al.(2011) as a multi-channel model that considers the phonetic factor, orthographic factor, acronym expansion, and contextual factor [11].

Two approaches were proposed by Mandal et al. (2018) [12] to convert the phonetically transliterated text to standard Roman transliteration. Sequence to Sequence (Seq2Seq) model with RNN (Recurrent Neural Network) and Long Short Term Memory (LSTM) is used in the first approach for the conversion. The second approach is based on string matching using Levenshtein Distance [13]. The first approach provided better accuracy than the second approach for the code-mixed text normalization task. Singh et al.(2018) [14] proposed a skip-gram [15] edit distance [16] method to normalize the anomalies of code-mixed text such as spelling variations and grammatical errors. Skip-gram has a similarity metric created from considering the context of a word in a given semantic space. Considering the similarity metric, the most frequently used word is used as the substitution for the variation of the same word, which normalises the data and reduces the noise.

Barik et al. (2019) [17] introduce a normalization approach with language identification with CRF (Conditional Random Field) and lexical normalization by replacing the OOV (Out Of Vocabulary) tokens with its standard tokens from the dictionary. Lourentzou et al. (2019) [18] and Dirkson et al. (2019) [19] proposed character-based and word-based normalization approaches for Out Of Vocabulary (OOV) words. Arora and Kansal (2019) [20] used a Convolutional Neural Network (CNN) model with character embedding to normalize the unstructured and noisy texts from social media. A similar approach was followed by Kayest and Jain (2019) [21] and Liu et al. (2021) [22].

B. Machine Translation

The importance of Machine Translation (MT) is increased because of the high demand for translation in overseas businesses, military services, profitable customers with the prevalence of different languages and valuable social media content for business development. Neural Machine Translation(NMT) is the currently trending domain in Machine Translation. Recurrent Neural Network [23], Seq2Seq approach [8], Attention based NMT [24] are considered trending approaches for NMT

TABLE II

CHALLENGES IDENTIFIED IN SINHALA-ENGLISH CODE-MIXED SENTENCES

Sinhala-English Code-Mixed Sentence(SCMT)	Sinhala Sentence	English Sentence	Identified Issues in SCMT
kama vry gd	කෑම ගොඩාක් හොඳයි	Food is very good	**Spelling error** - The words 'vry' and 'gd' represents the English words 'very' and 'good'.
mama wathura bonawa mama vathura bonawa mama vathura bonawaa	මම වතුර බොනවා	I drink water	**Inconsistent phonetic transliteration -** The same sentence is written in different patterns. The word 'water' is represented as 'vathura','wathura' and the word 'drinking' is represented as 'bonawa','bonawaa'.
4to gaththa	ඡායා රූප ගත්තා	Took photo	**The use of special characters and numeric characters** The word '4to', it absorbs the phonetic sound of word 'four' and combines it with the word 'to', together it represents the phonetic sound of photo.
service eka hondai	සේවාව හොඳයි	Service is good	**Borrowing of words** - The sentence starts with an English 'Service' and suddenly switches to Sinhala transliterated words 'eka' and 'hondai'.
teacherla hamoma enna	ගුරුවරුන් හැමෝම එන්න	All the teachers are requested to come	**Integration of suffixes -** the word 'teachers' is an English word which is a singular noun and the suffix 'la' is in the transliterated form taken from Sinhala. Together the word stands for the meaning 'teachers' which is plural.
niyama kama so ayeth kanna hithenava	නියම කෑම ඒ නිසා ආයෙත් කන්න හිතනවා	Great food, so like to eat again	**Switching for discourse marker -** In this sentence, an English discourse marker 'So' is used to join the two Sinhala transliterated sentences.

Many studies have been carried on translation based on monolingual datasets. Gulcehre et al.(2015) [25] present two methods, shallow and deep fusion to combine language models with Neural Machine Translation(NMT) techniques. Sennrich et al. (2016) [26] proposed two techniques to use monolingual data for translation. To fix the encoder and attention model parameters when training, the monolingual dataset is matched with dummy inputs in the first approach.

The second approach suggested is using a model trained on a parallel corpus with neural translation techniques for monolingual translation. Cheng(2019) [27] proposed a semi-supervised approach for monolingual machine translation by combining labelled and unlabelled corpus. Labelled corpus is parallel language corpus and unlabelled corpus is monolingual corpus.

There are multilingual NMT models available where a single model supports translating from multiple source languages to multiple target languages. These systems inspire knowledge translation among language pairs[28, 29], zero-shot translation(direct translation among a language pair that has never been used in the training phase) [30, 31, 32, 33] and enhance translation of low resource language pairs[34, 35]. Rather than these benefits, multilingual NMT systems show poor performance [32,34] and bad translations when accommodating many languages [36]. Zhanget al. (2020) [37] propose an improved NMT model where a normalization layer and linear transformation layers are used to overcome the representation issue of other multilingual NMT models. Also, the research study [37] addresses how the output from multilingual NMT models are affected by the unavailability of the parallel corpus. A Random Online Back Translation approach(ROBT) is proposed to overcome the issue of unseen

TABLE III

SAMPLE SENTENCES FROM THE ANNOTATED CORPUS; AN1 – ANNOTATOR 1, AN2 – ANNOTATOR 2

Sinhala-English Code-Mixed Sentence	Sinhala Sentence translated by Human Translator	AN1	AN2	Alternate translation by Annotator1	Alternate translation by Annotator2	Finalized translations by the translator
gaana wadi	ගාන වැඩියි	FC	FC	N/A	N/A	N/A
Price ekata shape wenna hoda rasata kama hambenawa	මිලට හරියන්න හොඳ ටෙස්ට් කෑම හම්බෙනවා	FC	CR	N/A	මිලට හරියන්න හොඳ රසවත් කෑම හම්බෙනවා	මිලට හරියන්න හොඳ රසවත් කෑම හම්බෙනවා
calm place ekak, enjoy kranna puluwn	කාම් තැනක් , එන්ජෝයි කරන්න පුළුවන්	CR	CR	සන්සුන් තැනක් , විනෝද කරන්න පුළුවන්	සන්සුන් තැනක් , එන්ජෝයි කරන්න පුළුවන්	සන්සුන් තැනක් , විනෝද කරන්න පුළුවන්
Singappooru kola kiyalai api kiyanne me gedi hedena gahata	සිංගපූරු කෝලා කියලයි අපි කියන්නේ මේ ගෙඩි හැදෙන ගහට	FC	FC	N/A	N/A	N/A
parking loku aulak na.	පාකින් ලොකු අපහසු නැත	CR	FC	වාහන නැවැත්වීමේ ලොකු අපහසු නැත	N/A	වාහන නැවැත්වීමේ ලොකු අපහසු නැත

When it comes to code-mixed languages, the translation domain consists of only very few research. Carrera et al. (2009) [38] introduce a qualitative study on the combined code-switched corpus from social media. According to the study, hybrid models combined with Statistical Modelling [39] and the Knowledge Translation approach [40] achieved comparatively good translation. In the code-mixed machine translation model introduced by Rijhwani et al.(2016) [6], the dominant language in a sentence is called matrix language. The non-dominant language is called an embedded language. The initial task in this model is word-level language identification and matrix language detection. Then the data is applied to a current translator to translate code-mixed tweets to the language of the user's choice. An augmentation pipeline for code-mixed text machine translation is proposed by Dhar et al. (2018) [5]. They introduce a parallel corpus with code mixed Hindi-English sentences as source sentences and English sentences as target sentences. The pipeline includes language identification, matrix language identification, translation to matrix language, and translation to the target language. The final output from the model would be translated monolingual sentence. The augmentation pipeline is applied with current translation models such as Google's Neural Machine Translation System (NMTS) [41], Moses [42] and Bing Translator. Each of these models provided an improved BLEU score when the augmentation pipeline is added in the pre-processing phase. Masoud et al. (2019) [43] introduced a Back Translation model for Tamil-English code-switched text. Baseline, monolingual and hybrid approaches are used to evaluate the system. The back-translated approach gave the highest BLEU score of 25.28 for the code-switched sentences.

III. Code-Mixing in Sri Lanka

Kachru (1986) [44] explains the necessity of English in South Asia in his research study. Many former Anglo-American colonies have been identified with English language varieties, which is called a deviation from standard English to the later development world. According to his observation in South Asia, the English language is considered as a sign of 'modernization', 'achievement' and 'strength'. He defines code mixing as a highlight of modernization, social and economic status and membership in an aristocratic society. The widest code-mixing range is identified with the English language. The main reason for code-mixing in Sri Lanka occurred due to the colonization of the British.

Sri Lanka acknowledges Sinhala, English and Tamil as the formal languages used for official activities. Sri Lanka mainly has two code-mixed language categories: Sinhala- English and Tamil-English, but there is no mixing between Sinhala and Tamil languages. People have massively adopted internet usage in the 21st 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 used is depended on the individuals [45].

The Sinhala language has a base of Brahmi script in its ornamentation of writing. According to the Unicode standard, 41 consonants, 18 vowels, and 2 half vowels altogether 61 characters are there in the latest Sinhala alphabet [46]. Even though there are 61 letters, the language has only 40 different sounds represented by those letters [47]. Sinhala-English code-mixing originated from the multilingual society of Sinhala - English speaking people. Srilankans use SCM as one of the main communication languages in social media. It has become very popular among the younger generation of the 21st century.

We conducted a survey study for identifying the necessity of translation of Sinhala code-mixed text. According to a recent research study on social media usage, users aged 20-29 are 32.2% of the whole social media users[56]. To identify the extent of usage of the code-mixed text in Sri Lankan social media, we decided this specific age group would be more appropriate to collect reliable data as they are the most active age group of social media. 82 individuals participated in this survey study who are native Sinhala language speakers and aged between 20-29. According to the survey result shown in Table I, 85.2% of people have stated as using code-mixed text for writing in social media rather than the native language. Increased usage of SCMT increases the demand for processing the SCMT. The best way to use the code-mixed text is to translate the text into a standard language so the data could be easily used for Machine Learning tasks such as recommendations, sentiment analysis, entity extractions etc.

In SCMT, there are several challenges in representing the text: Spelling errors, integration of suffixes, the usage of special and numeric characters in the text, borrowing words from another language, combining languages, switching of discourse markers and inconsistent phonetic transliteration. Table II provides a detailed description of challenges in Sinhala-English code-mixed text with examples. Due to different patterns of SCMT, it is difficult to translate SCMT without a parallel corpus.

IV. Corpus Creation

Most machine translation systems need a remarkable number of parallel sentences to accomplish a good outcome. Our study required creating a parallel corpus with parallel sentences of SCMT and Sinhala text. To achieve this goal, SCM (Sinhala-English Code-Mixed) sentences were gathered from social media. 5000 SCM sentences are used to create the parallel corpus.

After the extraction process, each SCM sentence in the corpus is human translated into Sinhala sentences with the help of a human translator, who is a Sinhala native speaker. The translator followed the mapping proposed in the research study of Kugathasan and Sumathipala et al. (2020) [48] for the manual human translation process. Thus, the SCM sentence is the source sentence, and the Sinhala sentence is the target sentence.

The translated dataset is validated using the Crowd Sourcing method [49]. Using the Crowd Sourcing technique in our research aims to discriminate good translations from bad ones. We split our corpus into groups of 15 where each annotator gets approximately 300 sentences and each group had a number of 2 annotators who are Sinhala native speakers, bilingual and good in English.

The reviewers were instructed to make sure that their Sinhala translation: does not have any spelling errors, and should be grammatically correct and natural-sounding Sinhala. The annotators judge the translated Sinhala sentences into two categories. Fully Correct(FC) and Change Required(CR). If the sentence is labelled with CR, then an alternative Sinhala translation would also be provided by the same annotator. The alternative sentence provided for each SCM sentence was a more fluent and grammatically correct Sinhala sentence. When there are contradictory tags by annotators for a specific translation, only the alternative translation with the CR tag is considered. When both annotators have annotated with CR tag, the best alternate provided is selected by the human translator who worked in the initial phase of creating the corpus. Some annotated sample sentences from the corpus are shown in Table III.

After correcting the alternatives, the corpus is updated with the corrections. We randomly choose 100 translated sentences, provided them to the linguistic experts of Sinhala language, and asked them to rank the translation good or bad, considering the following factors: spelling errors, the grammatical pattern of the sentence, and meaningful translation. In the ranking process, we gained judgments from three different linguists. Each translation has 3 rating labels from two categories. We used Fleiss' Kappa method [50, 51] to measure the reliability of the agreement between the raters while assigning a rating for the translated sentences. The Fleiss' Kappa score received for the translation of SECM to Sinhala is 0.88, which is almost near to full agreement for the translated sentences are correct.

V. System Architecture

The MT model proposed in this study is an adopted and enhanced approach to the research work of Sutskever et al. (2014) [8]. The model consists LSTM, Seq2Seq, Teachers Forcing mechanism and a normalization pipeline to translate the code-mixed text.

A. Sequence to Sequence(Seq2Seq)

Seq2Seq approach introduced by Sutskever et al.(2014) is a model with the goal of mapping the input sequence with a fixed length to an output sequence with fixed length 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. In this approach sequence of source sentences is matched with the sequence of the target sentence[20]. In this machine translation model, source sequence would be the input and target sequence would be the output. Seq2Seq model is also called as Encode-Decoder framework as shown in Figure 3.

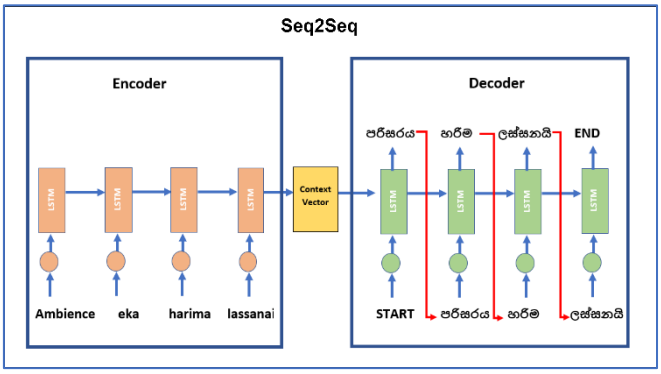

Fig. 3 Seq2Seq model

Source language is read and used as the input to the encoder. A context vector which can also be called the 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. The outputs gained from the encoder are discarded and only the hidden states have proceeded as the inputs to the decoder.

The decoder takes the hidden state and the starting string 'START' as the input. Hidden states are produced by the encoder and the input of the decoder is read word by word during decoding. In the training phase of the decoder, the Seq2Seq baseline model lets the predicted output from the previous timestamp as the input to the next timestamp in the decoder. But in our proposed approach we applied Teacher Forcing.

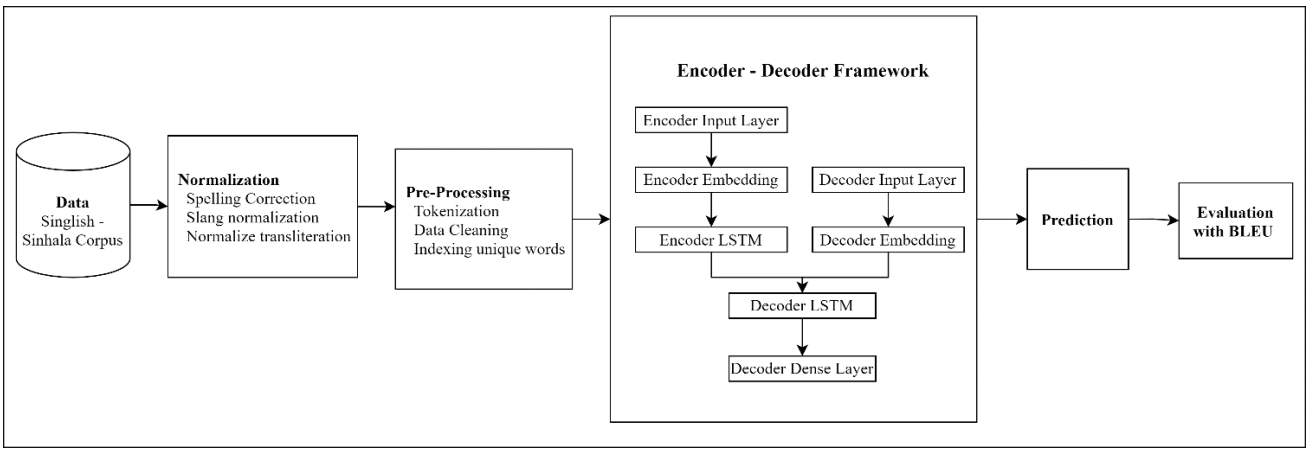

Fig. 4 System diagram of the proposed model

Source language is read and used as the input to the encoder. A context vector which can also be called as the 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.

The outputs gained from the encoder are discarded and only the hidden states have proceeded as the inputs to the decoder. Decoder takes the hidden state and the starting string 'START' as the input. Hidden states are produced by the encoder and the input of the decoder is read word by word during decoding. In the training phase of the decoder, the Seq2Seq baseline model lets the predicted output from the previous timestamp as the input to the next timestamp in the decoder. But in our proposed approach we applied Teacher Forcing Mechanism in the training phase of the decoder neglecting the predicted outputs from the timestamps.

B. *Long Short Term Memory(LSTM)*

LSTM network is chosen as the basic unit for text generation with the Seq2Seq model as shown in Figure 3. LSTM has internal technique gates that control the flow of information. Gates decides the important details to keep or forget in the cell state along the long chain of sequence. Gates learns what information is relevant and what to keep or throw away during the training. LSTM cell has three main gates, which are the input gate, forget gate and output gate as shown in Figure 5. According to the concept, when an input is given to the LSTM unit, it is converted into machine-readable vectors and these sequences of vectors would be processed one by one. In the forget gate the information from the hidden state from the previous timestep(h_{t-1}) and current input(X_i) would be passed as inputs. Forget gate has a Sigmoid activation function which turns the values between 0 to 1. If the output value from the sigmoid is closer to 0, that information will be forgotten and if it is closer to 1, it will be stored. In the input gate previous hidden state(h_{t-1}) from the previous timestep and current input(X_i) would be passed into the sigmoid function and Tanh function separately. Tanh activation function turns the values in between -1 to 1 to control the network.

Tanh output would be multiplied with the output from the sigmoid and the sigmoid would decide which information to keep and forget. Outputs gathered from forget gate and input gate would be utilized to upgrade the cell state. The next hidden state(h_t) would be decided by the output gate. The preceding hidden state(h_{t-1}) and the current input(X_i) passed into the sigmoid function and the newly upgraded cell state would be transited through tanh function. Sigmoid and tanh output decides the information that should be carried by the next hidden state. The upgraded new cell state (C_t) and the hidden state(h_t) would be transited to the next time step. Likewise, each unit of LSTM would run through these gates to store only the important details from the sequence.

C. *Teacher Forcing*

Using the ground truth from a prior timestamp as input for the current timestamp for quick and efficient training of Recurrent Neural Network is called as Teacher Forcing method [54]. Teacher Forcing method functions by utilizing the actual output from the previous timestamp *t* as input to the next timestep *t+1*. Figure 6 shows how the decoder of Seq2Seq model would be trained with Teacher Forcing and without Teacher Forcing. In our proposed model to translate SCMT to Sinhala, Teacher Forcing method is applied in the decoding phase.

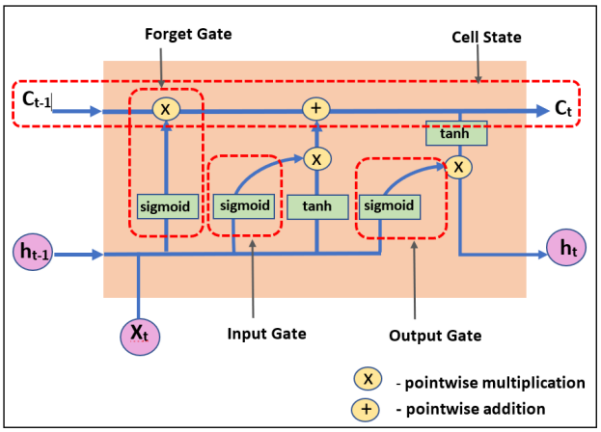

Fig. 5 Architecture inside a LSTM unit

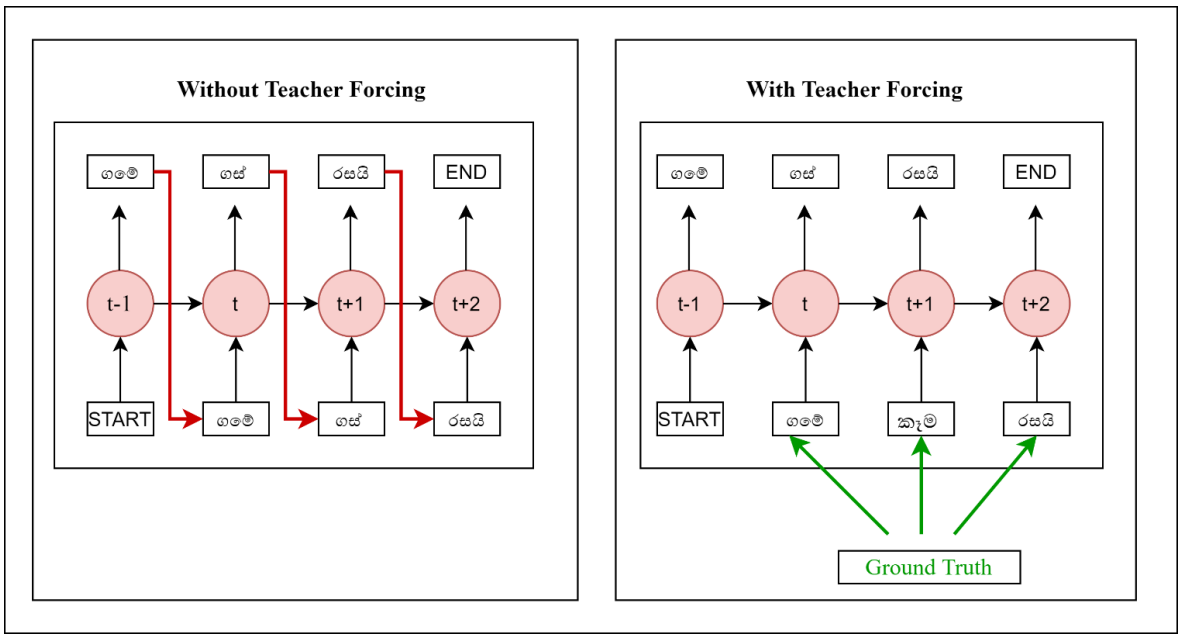

Fig 6. Example of decoder with the application of Teacher Forcing method and without Teacher Forcing method

VI. Model, Experimental Setting & Result

The initial phase of the model consists of the data pre-processing. Then, the dataset is cleaned by converting the sentences into lowercase, removing emojis, removing quotes and removing unnecessary spaces. Normalization is considered an important process when it comes to the translation of code-mixed text. Compared to monolingual sentences, code-mixed sentences have more noisy data. Dictionary-based approach and Levenshtein Edit Distance [52] based approaches are used for the normalization task in our model.

Spelling error is one of the challenges in Sinhala-English code-mixed. For example, ‘accident’ can be misspelt ‘accsident; accxident; acddent etc’. This happens mainly because most bilingual users are fluent only in their native language Sinhala and not experts in the second language English. The first step of the normalization is the out-of-vocabulary English words from the texts are normalized using the Birkbeck spelling error corpus dictionary [53], which contains 36,133 misspellings of 6,136 words gathered from various sources. Slang words in the code-mixed text were identified as another barrier to the translation of the SCM sentences. This issue is sorted using the SlangNorm dictionary, which contains 5427 slang words. For example, words such as ‘2mrw’ and ‘3wheeler’ will be replaced with the correct form ‘Tomorrow’ and ‘three wheeler’ using SlangNorm dictionary. In SCMT the same word is represented in different transliterated forms in various sentences in the corpus. Levenshtein Edit Distance approach [52] 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.

After the normalization of the sentences, target sentences are added with a ‘START’ token at the beginning of the sentence and an ‘END’ token is added at the completion of the sentence. Tokens assist the model to recognize when to begin the translation and end the translation in the decoder. The distinctive words are identified from the source and target corpus. A unique number is allocated to each distinctive word to create dictionaries of words to index and vice versa. These dictionaries are used in the embedding phase of the encoder and decoder.

In this research, a Seq2Seq model is fabricated using LSTM as the basic unit. The sequence of the source sentence is matched with the sequence of the target sentence where the source sequence would be the SCM sentence, and the target sequence would be the Sinhala sentence. 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 by LSTM units even though the transformation happens. Outputs from the encoder are repudiated and only the hidden states in the context vector are passed to the decoder.

The decoder also has embedding as its primary hidden layer. Hidden states passed from the encoder and the outputs given by the embedding layer in the decoder will be taken as the input of LSTM layer in the decoder. Teachers Forcing mechanism is applied in the training part of the decoder. Decoder pursues to implement a word at *t+1* timestamp, considering the actual output at *t* timestamp, not the predicted output. This lets the model learn from the actual values rather than wrongly predicted values. LSTM layer in the decoder returns internal states and output sequences. Internal states are stored and used in the prediction phase. The dense layer is applied with the Softmax activation, and decoder outputs are generated.

The data is shuffled 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. Encoder and decoder inputs are in the shape of a

TABLE IV

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	
1	ganan wadi	ගාන වැඩියි	ගණන් 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 ගාන හොඳටම වැඩියි ඒ ගානට වටින්නේ නෑ	කෑම රහ උනාට ගාන හොඳටම වැඩියි ඒ ගානට වටින්නේ 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 මදියි 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 ෆොටෝ ගන්න විතරයි කියන්නේ කියලා	මෙයා කියන්නේ 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	ප්‍රධාන 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 නැහැ	කෑම එච්චර විශේෂ නැහැ	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

2D array. The encoder 2D array has batch sizes of 10, the maximum source sentence length is 27, and the shape of the encoder input will be (10,27). The decoder 2D array has batch sizes of 10, a maximum source sentence length of 26 and the shape of the encoder input is (10,26). Decoder outputs are in the shape of a 3D array with a batch size of 10, the maximum target sentence length 26. NumPy, Pandas, TensorFlow, Sacrebleu are some important libraries used to build the model in the technological point of view.

After the training phase of the model, to produce the translation outputs, a prediction phase is implemented. In the prediction phase, an input sequence from the corpus(SCM sentence) will be provided to predict the Sinhala translation. This phase contains an encoder-decoder framework without Teacher Forcing mechanism, where the predicted output from the previous timestamp *t* would be fed for the current timestamp *t+1* instead of the actual output. Figure 4 shows the system architecture of the proposed model.

VII. EVALUATION & DISCUSSION

We evaluated the performance of our system by comparing our model with the most commonly used translation models. We applied our dataset to the Seq2Seq Baseline model [8] and the Attention model [24] with the same experimental setting. Each model was trained with the normalization pipeline and without the normalization pipeline. After training the models, we evaluated the translation outputs using BLEU [55] metric.

$$BLEU = BP.\exp\left(\sum_{n=1}^{N} W_n log P_n\right) \quad (1)$$

$$BP = \begin{cases} 1 & if\ c > r \\ \exp\left(1 - \frac{r}{c}\right) & if\ c \le r \end{cases} \quad (2)$$

In the BLEU score equation (1), *BP* is the Brevity Penalty, *N* is the number of n-grams(1-gram,2-gram,3-gram,4-gram), W_n is the weight for each modified precision, P_n is the modified precision [55]. P_n for each n-gram up to 4-gram is calculated based on the clipped count and the total number of the particular n-gram in the predicted sentence [55]. When the n-gram order is greater than the length of the reference sentence, to avoid the zero division error the total number of n-gram values is set to 1.

The Brevity Penalty(BP) depends on the values of *c*, the count of unigrams in all the predicted sentences and *r* is the most probable matching length of sentence in the corpus. Hundred Sinhala code-mixed sentences are selected from the corpus. Its relevant translation of Sinhala sentences is predicted using our proposed model. Initially, the number of clipped counts [55] and the total number of the particular n-gram in the predicted sentence are extracted to calculate the modified precision as shown in Table IV. Then the overall BLEU score is calculated for those hundred sentences. Finally, the same evaluation approach with the same experimental setting as explained in Section IV, is applied with the Seq2Seq Baseline model and Attention models with and without the normalization task. A summary of the comparison among the models is shown in Table V.

TABLE V

COMPARISON OF RESULTS RECEIVED FROM DIFFERENT MODELS

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	**12.78**
Seq2Seq Baseline Model + Normalization	57.11	0.7753	31.97	1.75	-0.145237	-0.204693	-0.275824	-0.389159	0.573	**20.77**
Seq2Seq + Attention without Normalization	70.55	0.303	30.3	1.15	-0.080998	-0.162399	-0.252416	-0.369135	0.6876	**28.95**
Seq2Seq + Attention + Normalization	70.22	0.5023	31.05	1.05	-0.0689162	-0.141996	-0.208556	-0.292517	0.6413	**31.46**
Seq2Seq + Teacher Forcing without Normalization	71.42	0.5095	37.17	0.38	-0.066960	-0.1232	-0.181972	-0.262455	0.595	**31.54**
Seq2Seq + Teacher Forcing + Normalization	71.57	0.4979	37.87	0.38	-0.06046	-0.1232717	-0.189274	-0.251089	0.6326	**33.89**

Seq2Seq Baseline model with normalization and without normalization showed the lowest performance and achieved the lowest BLEU score compared to the other two models. Among the Attention and Teacher Forcing models, the best BLEU score is 33.89 received by Teacher Forcing Algorithm, proving the proposed model comparatively works well with Sinhala-English Code-Mixed text. Also, the comparison study with and without the normalization task demonstrated that the models performed well and provided a better BLEU score when the normalization pipeline is applied to each of the models. Not only the BLEU scores, but the proposed model also achieved comparatively fair values for training and testing accuracies and loss as shown in Figure 8.

An analysis of the predicted sentences is performed to identify whether the proposed model helped to overcome the challenges pointed out in Table II.

If we take the sample sentence (1) shown in Table IV, (Code-mixed text - CMT, Reference text- REF, Translated text - TRANS):

CMT : gaana wadi (1)
REF : ගාන වැඩියි
TRANS : ගණන් වැඩියි

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

In the sample sentence (3) shown in Table IV,

CMT : place eka super clean (3)
REF : තැන සුපිරි පිරිසිදුයි
TRANS : තැන සුපිරි පිරිසිදුයි

In sentence (3), the CMT sentence 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 the sample sentence (9),(10) shown in Table IV,

CMT : kaama echchara special naha (9)
TRANS : කෑම එච්චර විශේෂ නැහැ

CMT : kama denna puluwan (10)
TRANS : කෑම දෙන්න පුළුවන්

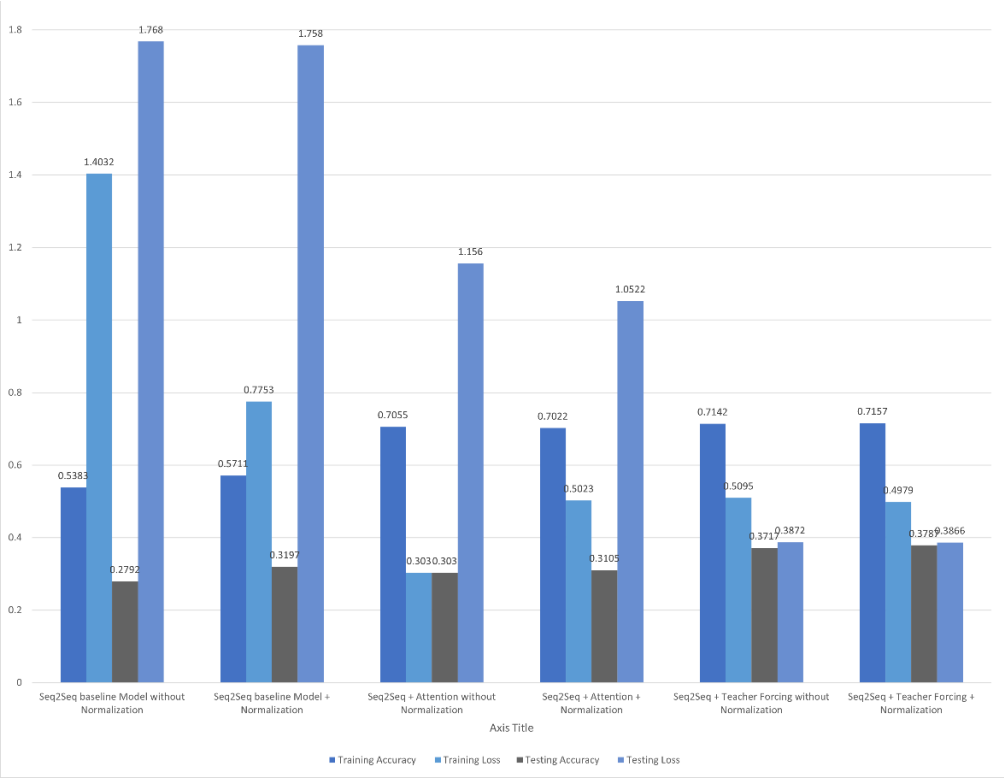

Fig 8. Experimented models accuracies, loss & relevant BLEU scores

The sentences (9) and (10) 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. The use of special characters and numeric character issues were sorted in the normalization phase with the SlangNorm dictionary.

VIII. Conclusion

The main goal of this research is to utilize the user-generated Sinha-English code-mixed sentences and convert the sentences into a standard language, so the code-mixed texts can also be used for several research and business purposes.

From analyzing the challenges in SCMT text, we pointed out the key issues that have been a barrier to processing the Sinhala-English code-mixed text. Creating a dataset for this research study was a challenging task due to the unavailability of current resources. The dataset created in the study was created following several processes such as manual translation with a human translator, crowdsourcing to annotate the dataset to check whether the human-translated sentences are correct and rating the translation with linguistic experts to analyze the Fleiss’ Kappa score. The received score of 0.88 shows almost full agreement with the translation. The corpus created in this study using proper rules and regulations could promote research based on the Sinhala code-mixed domain. The proposed approach, which is a combination of the Seq2Seq model with the LSTM unit and the Teachers Forcing mechanism gives a comparatively higher BLEU score of 33.89 for code-mixed text translation compared to the other models. Moreover, the evaluation study proves that most of the challenges identified in SCM sentences can be solved using our proposed model. But somehow, a few of the challenges such as integration of suffixes, and change of discourse marker remain unsolved.

This research study can be considered an initiative for Sinhala-English code-mixed text translation. As the future work of this study, we are planning to solve the rest of the challenges which we were not able to solve with the current proposed model. Furthermore, we would like to extend the corpus to focus on other tasks of code-mixing such as sentiment analysis, language identification, entity extraction etc.

References

[1] E. E. Davies and A. Bentahila, “Contact linguistics: Bilingual encounters and grammatical outcomes,” 2007.

[2] K. R. Chandu, M. Chinnakotla, A. W. Black, and M. Shrivastava, “Webshodh: A code mixed factoid question answering system for web,” in International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, 2017, pp. 104– 111.

[3] M. Yang, Y. Ren, and G. Adomavicius, “Understand- ing user-generated content and customer engagement on facebook business pages,” Information Systems Research, vol. 30, no. 3, pp. 839–855, 2019.

[4] E. Qualman, Socialnomics: How social media trans- forms the way we live and do business. John Wiley & Sons, 2012.

[5] M. Dhar, V. Kumar, and M. Shrivastava, “Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach,” in Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 131–140. [Online]. Available: https://www.aclweb.org/anthology/W18- 3817

[6] S. Rijhwani, R. Sequiera, M. C. Choudhury, and K. Bali, “Translating codemixed tweets: A language detection based system,” in 3rd Workshop on Indian Language Data Resource and Evaluation-WILDRE- 3, 2016, pp. 81–82.

[7] P. Goyal, S. Pandey, and K. Jain, “Deep learning for natural language processing,” New York: Apress, 2018.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” arXiv preprint arXiv:1409.3215, 2014.

[9] K. Cho,B.van Merriënboer, C.Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://www.aclweb.org/anthology/D14-1179

[10] K.-F. Wong and Y. Xia, "Normalization of chinese chat language," Language Resources and Evaluation, vol. 42, no. 2, pp. 219–242, 2008.

[11] Z. Xue, D. Yin, and B. D. Davison, "Normalizing microtext," in Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence. Citeseer, 2011.

[12] S. Mandal, S. D. Das, and D. Das, "Language identification of bengali-english code-mixed data using character & phonetic based lstm models," arXiv preprint arXiv:1803.03859, 2018.

[13] L. Yujian and L. Bo, "A normalized levenshtein dis- tance metric," IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pp. 1091– 1095, 2007.

[14] R. Singh, N. Choudhary, and M. Shrivastava, "Auto- matic normalization of word variations in code-mixed social media text," arXiv preprint arXiv:1804.00804, 2018.

[15] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling." in LREC, vol. 6. Citeseer, 2006, pp. 1222–1225.

[16] [16]E. S. Ristad and P. N. Yianilos, "Learning string-edit distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 5, pp. 522–532, 1998.

[17] A. M. Barik, R. Mahendra, and M. Adriani, "Nor- malization of indonesian-english code-mixed twitter data," in Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), 2019, pp. 417– 424.

[18] I. Lourentzou, K. Manghnani, and C. Zhai, "Adapting sequence to sequence models for text normalization in social media," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, 2019, pp. 335–345.

[19] A. Dirkson, S. Verberne, A. Sarker, and W. Kraaij, "Data-driven lexical normalization for medical social media," Multimodal Technologies and Interaction, vol. 3, no. 3, p. 60, 2019.

[20] M. Arora and V. Kansal, "Character level embedding with deep convolutional neural network for text nor- malization of unstructured data for twitter sentiment analysis," Social Network Analysis and Mining, vol. 9, no. 1, pp. 1–14, 2019.

[21] M. Kayest and S. K. Jain, "An incremental learning approach for the text categorization using hybrid optimization," International Journal of Intelligent Computing and Cybernetics, 2019.

[22] J. Liu, S. Zheng, G. Xu, and M. Lin, "Cross- domain sentiment aware word embeddings for review sentiment analysis," International Journal of Machine Learning and Cybernetics, vol. 12, no. 2, pp. 343–354, 2021.

[23] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1700–1709. [Online]. Available: https://www.aclweb.org/anthology/D13-1176

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[25] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," arXiv preprint arXiv:1503.03535, 2015.

[26] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96. [Online]. Available: https://www.aclweb.org/anthology/P16-1009

[27] Y. Cheng, "Semi-supervised learning for neural ma- chine translation," in Joint training for neural machine translation. Springer, 2019, pp. 25–40.

[28] S. M. Lakew, M. Cettolo, and M. Federico, "A com- parison of transformer and recurrent neural networks on multilingual neural machine translation," arXiv preprint arXiv:1806.06957, 2018.

[29] X. Tan, J. Chen, D. He, Y. Xia, T. Qin, and T.-Y. Liu, "Multilingual neural machine translation with lan- guage clustering," arXiv preprint arXiv:1908.09324, 2019.

[30] M. Al-Shedivat and A. Parikh, "Consistency by agreement in zero-shot neural machine translation," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1184–1197. [Online]. Available: https://www.aclweb.org/anthology/N19- 1121

[31] J. Gu, Y. Wang, K. Cho, and V. O. Li, "Improved zero-shot neural machine translation via ignoring spurious correlations," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1258–1268. [Online]. Available: https://www.aclweb.org/anthology/P19-1121

[32] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y.Wu, Z.Chen, N.Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," Transactions of the Association for Computational Linguistics, vol. 5, pp. 339–351, 2017. [Online]. Available: https://www.aclweb.org/anthology/Q17- 1024

[33] O. Firat, B. Sankaran, Y. Al-onaizan, F. T. Yarman Vural, and K. Cho, "Zero-resource transla- tion with multi-lingual neural machine translation," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 268–277. [Online]. Available: https://www.aclweb.org/anthology/D16-1026

[34] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry et al., "Massively multilingual neural machine translation in the wild: Findings and challenges," arXiv preprint arXiv:1907.05019, 2019.

[35] T.-L. Ha, J. Niehues, and A. Waibel, "Toward multi- lingual neural machine translation with universal en- coder and decoder," arXiv preprint arXiv:1611.04798, 2016.

[36] K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, "Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021.

[37] B. Zhang, P. Williams, I. Titov, and R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, Jul. 2020, pp. 1628–1639. [Online]. Avail- able: https://www.aclweb.org/anthology/2020.acl- main.148

[38] J. Carrera, O. Beregovaya, and A. Yanishevsky, "Ma- chine translation for cross-language social media," PROMT Americas Inc, 2009.

[39] M. C. Neale, S. M. Boker, G. Xie, and H. M. Maes, "Statistical modeling," Richmond, VA: Department of Psychiatry, Virginia Commonwealth University, 1999.

[40] P. Sudsawad, Knowledge translation: introduction to models, strategies and measures. Southwest Educa- tional Development Laboratory, National Center for the …, 2007.

[41] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey et al., "Google's neural machine translation system: Bridging the gap between hu- man and machine translation," arXiv preprint arXiv:1609.08144, 2016.

[42] P. Koehn, H. Hoang, A. Birch, C. Callison- Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 177–180. [Online]. Available: https://www.aclweb.org/anthology/P07-2045

[43] M. Masoud, D. Torregrosa, P. Buitelaar, and M. Arčan, "Back-translation approach for code- switching machine translation: A case study," in 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science. AICS2019, 2019.

[44] B. B. Kachru, "The power and politics of english," World Englishes, vol. 5, no. 2-3, pp. 121–140, 1986.

[45] N. Choudhary, R. Singh, I. Bindlish, and M. Shrivas- tava, "Sentiment analysis of code-mixed languages leveraging resource rich languages," arXiv preprint arXiv:1804.00806, 2018.

[46] M. Punchimudiyanse and R. Meegama, "Unicode sinhala and phonetic english bi-directional conversion for sinhala speech recognizer." IEEE International Conference on Industrial and Information Systems 2015, 2015.

[47] A. M. Gunasekara, A comprehensive grammar of the Sinhalese language. Asian Educational Services, 1999.

[48] A. Kugathasan and S. Sumathipala, "Standardizing sinhala code-mixed text using dictionary based ap- proach," in 2020 International Conference on Image Processing and Robotics (ICIP). IEEE, 2020, pp. 1–6.

[49] E. Estellés-Arolas and F. González-Ladrón-de Gue- vara, “Towards an integrated crowdsourcing defini- tion,” Journal of Information science, vol. 38, no. 2, pp. 189–200, 2012.
[50] T. R. Nichols, P. M. Wisner, G. Cripe, and L. Gu- labchand, “Putting the kappa statistic to use,” The Quality Assurance Journal, vol. 13, no. 3-4, pp. 57– 61, 2010.
[51] J. J. Randolph, “Online kappa calculator,” Retrieved October, vol. 20, p. 2011, 2008.
[52] G. Navarro, “A guided tour to approximate string matching,” ACM computing surveys (CSUR), vol. 33, no. 1, pp. 31–88, 2001.
[53] “Birkbeck spelling error corpus / roger mitton,” oxford Text Archive. [Online]. Available: http://hdl.handle.net/20.500.12024/0643
[54] I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning (adaptive computation and machine learning series),” p. 372, 2016.
[55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meet- ing of the Association for Computational Linguistics, 2002, pp. 311–318.
[56] K. Simon, “DIGITAL 2022: GLOBAL OVERVIEW REPORT,” Jan. 26, 2022. https://datareportal.com/reports/digital-2022-global-overview-report