CROSS LINGUISTIC NAMED ENTITY TRANSLITERATION FOR BURMESE (MYANMAR) AND ENGLISH USING TRANSFORMER MODEL

[1]Aye Myat Mon

Natural Language Processing Lab, Faculty of Computer Science
University of Computer Studies, Yangon
Myanmar
Email: ayemyatmon.ptn@ucsy.edu.mm

[2]Khin Mar Soe

Natural Language Processing Lab, Faculty of Computer Science
University of Computer Studies, Yangon
Myanmar
Email: khinmarsoe@ucsy.edu.mm

Abstract

Named Entity (NE) transliteration in Natural Language Processing entails phonetically transcribing text across diverse writing systems, playing a crucial role in applications like Entity Discovery, Information Retrieval, and Multilingual Machine Translation (MT). However, developing a reliable automatic transliteration system for English words in the Burmese (Myanmar) language proves challenging due to its complex composition system and limited data availability. To address this, our study created two bilingual NE transliteration dictionaries—native Myanmar NE dictionary and western Myanmar NE dictionary—comprising 157,105 and 129,464 name pairs, respectively. The machine assisted transliteration proficiency was evaluated using Transformer Neural Network (NN) based approach on the prepared data. Various scenarios by running different segmentation units (character, sub-syllable, and syllable) on native Myanmar-English named entity data, western Myanmar-English named entity data, and a combination of both datasets to achieve optimal results. The study evaluates experimental results using BLEU score and word error rate, demonstrating high accuracy and efficiency in transforming both Myanmar and English named entities bidirectionally.

***Keywords*: Burmese (Myanmar Language); named entity transliteration; character segmentation; sub-syllable segmentation; syllable segmentation; transformer neural network model.**

1. Introduction

With the rise of multilingualism, machine transliteration has become a vital process in Natural Language Processing. Transliteration involves transcribing words from one script to another, and it has several applications in the field. For instance, named entities like person names, organization names, and place names are usually transliterated when generating annotated parallel corpora in MT Systems. Additionally, machine transliteration can improve the performance of machine translation when a word in the source language does not have a direct equivalence in the target language. To facilitate the utility of evidence-based approaches, we manually collected two dictionaries called native and western(foreign) NE dictionaries for Burmese (Myanmar) language. This type of dictionary can solve the encountered problems and make sure to get the good performance in machine transliteration and translation systems. However, such dictionaries are extremely rare for low-resourced languages, especially for Myanmar. Motivated by this, we developed native and western NE dictionaries for My-En as one main contribution and experiments were carried out by applying Transformer NN based approach on three different segmentation units, char., sub-syl. and syl., respectively. BLEU metrics and WER were used to measure the accuracy for NE experimental results.

While Statistical Machine Translation (SMT) has been a popular form of machine translation for many years, it has many limitations. One of the main drawbacks of SMT is its dependence on the quality of the source material. This means that the accuracy of the translation is only as good as the quality of the bilingual text used for training. For languages that are less common or have fewer available resources, this can be a problem.

Neural Machine Translation, on the other hand, relies on self-learning models, making it a much more reliable solution, especially for under-resourced languages. Other advantages come in the form of speed and quality, with both increasing rate as they continue to learn.

2. Related Works

In the domain of cross-lingual information retrieval and linguistic processing, the transliteration of named entities plays a pivotal role in bridging language barriers and enhancing communication. This study delves into the specific context of Myanmar-English Named Entity Transliteration, a domain that involves the conversion of proper names and entities from Myanmar script to English script, and vice versa. As the significance of accurate transliteration continues to grow in the era of globalized information exchange, this exploration aims to survey and analyze existing works in the field. By reviewing related studies, methodologies, and advancements, we seek to gain a comprehensive understanding of the challenges and solutions in Myanmar-English Named Entity Transliteration, ultimately contributing to the enhancement of cross-language information processing and communication technologies.

The Transformer model [Vaswani *et al.* (2017)] which incorporates a self-attention mechanism, is currently the most advanced neural network architecture in the field of NLP. Over the past few years, comprehensive sequential processing techniques for NLP tasks have seen significant development. Nevertheless, resource insufficiency is still a major problem for many lesser-known languages. Numerous Asian languages utilize unique writing systems, and significant progress has been made in phonetic transcription techniques for prominent languages such as Korean, Japanese and Chinese [Merhav and Ash (2018)]. Nonetheless, further studies are necessary for understudied languages with limited resources. In the domain of NLP, transliteration is typically treated as a simplified form of translation that operates at the level of individual characters or graphemes, rather than whole words or phrases. This approach has been widely accepted and thoroughly researched, with many prior studies investigating its technical background. Neural network architectures such as the LSTM-RNN [Cho *et al.* (2014)] have emerged as particularly successful techniques for addressing a variety of NLP challenges.

The scarcity of previous research on Myanmar processing makes developing effective approaches for transliteration tasks a challenging endeavour. In a Burmese (Myanmar) name Romanization task [Ding *et al.* (2017)] found that traditional machine learning methods were more effective than NN based approaches. This underscores the importance of considering language-specific factors such as the availability and quality of training data, as well as the unique linguistic features of the language, when developing approaches for transliteration tasks. The authors [Naing *et al.* (2015)] delved the use of a rule-based transliteration strategy for post-editing machine translation from Katakana (Japanese) to Burmese, as described in their study. It focused on addressing mistranslation of Katakana out-of-vocabulary (OOV) words in Japanese to Burmese translation using statistical phrase-based machine translation (SPBMT) [Koehn *et al.* (2003a); Koehn *et al.* (2007b)]. The researchers utilized a rule-based post-editing scheme in their experiment, which involved training on 155,069 sentences from the BTEC corpus, and testing on a set of 1614 sentences. The results of the study revealed that the rule-based Katakana to Burmese translation approach outperformed the baseline PBSMT method by achieving a 19.39 BLEU score and lower out -of -vocabulary (OOV) errors by about 9.33%. This highlights the potential of rule-based approaches in improving machine translation output and reducing OOV errors.

In 2009, the Named Entities Workshop (NEWS) created a new Machine Transliteration task [Li *et al.* (2009)], which aimed to provide a standard dataset for language pairs such as English-Hindi, English-Hebrew, Chinese-English, Arabic-English, and others. The competition saw many teams participate, with most teams opting to use neural networks to learn the target transliterations. The use of neural networks proved to be more effective than phrase-based machine translation systems, resulting in superior transliteration performance. The top-performing team in the 2016 version of the task was NICT [Liu *et al*. (2016)]. They achieved this by using a LSTM-based RNN to encode input sequences to a hidden representation and decode that representation to produce the output sequence. However, decoding errors accumulate, leading to poor transliteration quality in suffixes. To solve this problem, NICT utilized target-bidirectional models, which generate two k-best lists by learning to produce the target from both left to right and right to left. They then used ensembles of these models to generate the transliterations, which are obtained by linearly interpolating probability distributions over the target vocabulary during beam-search decoding.

3. The Burmese (Myanmar) Language

The Burmese language, also known as Myanmar, is characterized by being tonal and belonging to the Burmese-Lolo branch of the Sino-Tibetan family. The language's script was influenced by the Brahmi script, which originated in India between 500 BC and 300 AD. Burmese is spoken primarily in Myanmar as it is the official language of the country. In 2007, approximately 33 million the Burmese people used Burmese as their primary language, and an additional communities of minority ethnic groups in Myanmar and surrounding nations used it as a second language. The Burmese language has a total of 12 vowels, 33 consonants, and 4

medials, which are considered the basic alphabets. This linguistic and cultural background highlights the unique characteristics and complexities of the Burmese language, which must be taken into consideration when developing language models and machine learning approaches for tasks such as named entity transliteration [Bradley (2015)].

The Burmese writing system is syllable-based, meaning that words can consist of multiple syllables and each syllable can have multiple characters. To further refine the writing system's structure, sub-syllable units can be used for specific purposes. Moreover, two example syllables are showed with character writing order numbers in Figure 1 and categories of characters in Table 1. In the Burmese Language, there are specific tasks that require the dentification of char., sub-syl. and syl. units [Mon and Soe (2020)]. This approach provides a more nuanced understanding of the language, which can be beneficial in various applications. Furthermore, the Burmese language contains a large number of English loan words, which present a unique challenge in terms of standardization of NE transliteration.

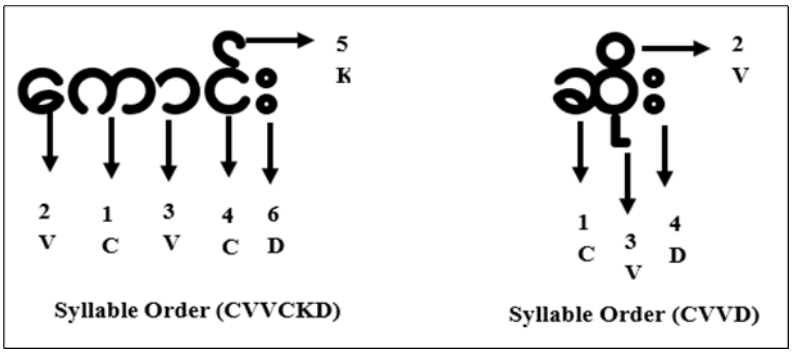

Fig. 1. Burmese Syllable Composition for "ကောင်း" ("Good" in English) and "ဆိုး" ("Bad" in English)

Notation	Description
C	ဗျည်း(Consonant) (က-အ)
M	ဗျည်းတွဲ(Medial) (ျ ြ ွ ှ)
V	သရ(Vowel) (ါ ာ ိ ီ ု ူ ေ ဲ ဳ ဴ ဵ ံ း ့ ္)
D	မှီခို့သရ(Dependent Various Sign) (း ံ ့)
K	အသတ်(Killer) (်)

Table 1. Categories of Characters

4. Difficulties

Transliterating between Myanmar and English poses a significant challenge due to two main factors. Firstly, there is a variation in the phonetic inventory of the two languages. As highlighted in a study by the author [Chang (2009)] English loanwords that are adapted into Myanmar must navigate the contrast between voiced, voiceless stops and affricates, as well as a two-way contrast with nasals and approximants. The seven-vowel system of Myanmar and its restrictions on syllable codas make it a simpler language in some respects than English. However, when transliterating from Myanmar to English, the three tones of Myanmar become redundant. In addition, the abundance of consonant clusters in English poses a challenge for transliteration to Myanmar, where such clusters are relatively limited.

One of the major challenges is the complexities of transliteration between Myanmar and English that are not only due to phonological differences but also non-phonetic orthographies. Unlike phonetic-based orthographies, both English and Myanmar use an etymologically-based orthography that can result in relative redundancy in their phonological inventory. This redundancy can lead to multiple ways of realizing phonemes, and even special spellings to give borrowed words an exotic appearance in Myanmar. Similarly, the English orthography can also cause irregular transcription, if the transliteration is based on the spelling rather than the actual pronunciation. The following sub-section analyzed the transliteration implications on Myanmar language for the letters in Wikipedia is marked as [W] and Myanmar Language Commission dictionary is marked as [M].

4.1. *Implications of Burmese (Myanmar) phonotactic*

The onset of an English syllable, which consists of one or multiple consonant letters, is a critical component of Myanmar syllable structure. In Myanmar, the initial consonant(s) of a syllable will be transcribed in the same way as English, as it is essential for Myanmar speakers to accurately identify the syllable's meaning and tone.

However, the pronunciation of the initial consonant(s) in Myanmar can vary depending on the surrounding vowels and tones, leading to some complexities in the transliteration process [Ding (2021)].

4.1.1. Simple onset

The phoneme-to-grapheme mapping is used for transliterating Myanmar, which is a language that has many consonants appearing at the beginning of syllables. The mapping overlaps with English to a large extent, and there are certain strong mappings that can be summarized in Table 2. The aspiration of obstruent is not a defining feature of the English language, which sets it apart from other languages such as Myanmar. In Myanmar, letters for non-aspirated voiceless obstruent sounds are more commonly used, and there are specific mappings between aspiration letters and certain sounds, such as <ချ-> for the sound /tʃ/ and <ရှ-> for the sound /ʃ/. The use of aspiration letters also extends to representing absent phonemes, as seen in the use of <ဖ-> for the native /p^h/ sound. The differentiation of phonemes sounds /s/ and /s^h/ are being lost in Myanmar, resulting in competition between the symbols <စ-> and <ဆ-> for the /s/ sound. However, <စ> is preferred in consonant clusters occurring at the beginning of a word, such as <စတ-> which is transcribed as /st/. Take the letter "C" as an example, in which we substitute /s/ with <ဆ-> and /k/ with <က-> in the word "CIRCUS," resulting in <ဆပ်ကပ်> [M]. Often, this mapping is done at the grapheme level, essentially leading to a literal transliteration. The conversion of <TH> to <သ> is a consistent mapping between graphemes, irrespective of the phonemes involved. For instance, consider the words LOGARITHM → လော်ဂရမ်သမ် [M] and THEORY → သီအိုရီ [M] where the <TH> is for /ð/ and /θ/ respectively.

Latin Letter	Burmese Letter
/p/	<ပ->
/t/	<တ->
/k/	<က->
/b/	<ဘ->
/d/	<ဒ->
/g/	<ဂ->
/z/	<ဇ->
/dʒ /	<ဂျ->
/m/	<မ->
/n/	<နှ->
/l/	<လ->
/j/	<ယ->
/w/	<ဝ->
/h/	<ဟ->

Table 2. The phoneme-to-grapheme mapping

Occasionally, strong grapheme-to-grapheme mappings take precedence over phoneme-to-grapheme mappings. For instance, the conversion of <J> to <ဇ_> is common in certain borrowed words, resulting in JANUARY → ဇန်နဝါရီ [M], JULY → ဇူလိုင် [M], and JUNE → ဇွန် [M]. In these cases, <J> to <ဇ_> replaces /dʒ/ with ဂျ, but this doesn't apply to examples like JOURNAL → ဂျာနယ် [M] and JURY → ဂျူရီ [M]. Additionally, etymology may play a role in cases like JESUS → ယေရှု [M], where <J> to <ယ-> and <S> to <ရှ-> may be influenced by Biblical Hebrew. The following example is about a Spanish name, JUAN→ဝှမ်[M]. Here an underlying chain of <JU>→ /xw/ → /hw/ → /ʍ/ to <ဝှ> can be considered behind the surface <JU> to <ဝှ>. Table 3 delineates specific situations in the Myanmar language concerning grapheme-to-grapheme mappings.

Grapheme	Transliteration	Example Word (English)	Transliteration (Myanmar)
<TH>	/ð/	LOGARITHM	လော်ဂရမ်သမ်[M]
<TH>	/θ/	THEORY	သီအိုရီ[M]
<J>	<ဇ_>	JANUARY	ဇန်နဝါရီ[M]
<J>	<ဇ_>	JULY	ဇူလိုင်[M]
<J>	<ဇ_>	JUNE	ဇွန်[M]
<J>	<ဂျ_>	JOURNAL	ဂျာနယ်[M]
<J>	<ဂျ_>	JURY	ဂျူရီ[M]
<J>	<ယ->	JESUS	ယေရှု[M]
<S>	<ရှ->	JESUS	ယေရှု[M]
<JU>	<ဍ>	JUAN	ဍမ်[M]

Table 3. Specific Situation in the Myanmar language related to grapheme-to-grapheme mappings

4.1.2. Onset cluster

The Myanmar language does not allow for complex onset clusters. Consonants from English onset clusters are denoted by a series of basic consonant symbols in Myanmar, with the final letter modified for the following nucleus. In clusters ending with "r" the letter "<ရ->" is used instead of "<ယ->" to avoid confusion. There are also special transliterations for clusters, such as "<CHR>" for /kr/, which is transliterated as "<ခရ->" in CHRIST→ခရစ်[M] and CHROMIUM→ခရိုမီယမ်[W]. When <CH> just stands for /k/, /k/ to <ခ> may not be triggered, CINCHONA → စင်ကိုနာ [M] nor likely be triggered in "<CHL>" for /kl/ in the following example CHLORINE→ကလိုရင်း [W]. Similarly, "<tr>" is mapped to "<ထရ->" where the aspirated "<ထ>" is used instead of the common "<တ>" for ELECTRON→ အီလက်ထရွန်[M] and TRANSISTOR →ထရန်စစ္စတာ [M] or even irregular spellings as <TR> to <တြ_> in GEOMETRY→ ဂျီသြမေတြီ [M]. The mapping for "
" is "<ဗြ->" (regularly "<ဘရ->") and can be seen in the transliteration of "Britain" as "<ဗြိတိန်>"[W]. Table 4 outlines specific English onset clusters within the Myanmar language.

English Cluster	Example Word (English)	Transliteration (Myanmar)
/kr/	CHRIST	ခရစ်[M]
/kl/	CHLORINE	ကလိုရင်း[W]
/tr/	TRANSISTOR	ထရန်စစ္စတာ[M]
/br/	BRITAIN	ဗြိတိန်[W]

Table 4. Specific English onset clusters in the Myanmar language

4.1.3. Null onset and hiatus

In situations where the onset of a sound cluster is missing, it is a common convention to use <အ-> as a placeholder, either at the start of a word or within a word hiatus. This is exemplified in the word IODINE → အိုင်အိုဒင်း [M]. If a hiatus begins with /i/, it is customary to use <ယ-> (or <ရ->) rather than <အ> in words such as UNION→ ယူနီယန် [M]. Take note that /ju/ is altered to <ယူ> at the word's start to prevent the combination of <အျ>. In the word MERCURY → မာကျူရီ [M], it is common to use <+ျူ> for the sound /ju/

after a general onset. In LOUISIANA → လူဝီစီယားနားး [W], at times, <o_> is added after /u/. Some stable borrowed words in Myanmar use independent vowel letters at the beginning, such as APRIL → ဧပြီ [M] and AUGUST → သြဂုတ် [M]. For triphthongs, syllables may be re-segmented with semi-vowel insertion, as seen in POWER → ပါဝါ [M], where /aʊr/ is analyzed as /a.wṛ/. In the word WIRE → ဝိုင်ယာ [M], /aɪr/ is analyzed as /aɪ.ṛ / with nasalization and the addition of <ယ>, as there is no standalone /aɪ/ rhyme in Myanmar. Table 5 illustrates specific situations in the Myanmar language pertaining to sound clusters and vowels.

Situation	Placeholder	Example Word (English)	Transliteration (Myanmar)
Missing Onset	<အ->	IODINE	အိုင်အိုဒင်း[M]
/i/ Hiatus	<ယ-> (or <ရ->)	UNION	ယူနီယန်[M]
/ju/ After Onset	<+ျူ>	MERCURY	မာကျူရီ[M]
<o_> is added after /u	<+ဝီ>	LOUISIANA	လူဝီစီယားနားး[W]
Stable Borrowed Words	Independent Vowel Letters	APRIL	ဧပြီ[M]
Stable Borrowed Words	Independent Vowel Letters	AUGUST	သြဂုတ် [M]
Triphthongs	Semi-vowel Insertion	POWER	ပါဝါ[M]
Nasalization and Addition	<ယ>	WIRE	ဝိုင်ယာ[M]

Table 5. Specific situations in the Myanmar language related to sound clusters and vowels

5. Dictionary and Methodology

This study presents an in-depth exploration of the construction of Named Entity (NE) terminology dictionaries for the Myanmar language. It discusses the various steps involved in this process, including the use of two large corpus of text, manual annotation of NEs, and machine learning techniques for NE transliteration. Additionally, the experiment is carried out with Transformer NN model to My-En NE transliteration, comparing the effectiveness of various segmentation units. Our research aims to provide a valuable resource for NLP applications, including NE Transliteration and Entity Linking, and to contribute to the development of NLP tools for Myanmar language.

5.1. *Building parallel NE data*

To collect the western (foreign) NE, the ALT corpus [Ding *et al.* (2018a, 2019b and 2020c)], UCSY corpus [ShweSin *et al.* (2018)] and Wikipedia data are all valuable resources for NLP researchers studying machine transliteration between different languages. The ALT corpus, which consists of 20,000 parallel sentences from Wiki news articles, was one of the key sources of NE instances for the study of Myanmar-English transliteration. To supplement the ALT corpus, the UCSY corpus, a collection of 200,000 parallel sentences taken from local textbooks and Myanmar news articles is used. Through the use of the GIZA++ toolkit [Och and Ney (2003)] and [Och (2003)] we were able to filter out the relevant transliteration instances from the raw alignments between the Myanmar and English languages. By leveraging the Wikipedia data, we were able to gather additional NE instances that were relevant to the study of My-En transliteration. The combination of the ALT corpus, UCSY corpus, and Wikipedia data provided a comprehensive set of NE instances for the NLP researchers to use in their study of machine transliteration between Myanmar and English.

Native Myanmar NE dataset focuses on providing real-world transliterations of Myanmar people, organization, and place names. This local transliteration is essential because it ensures that the names are transcribed accurately and are easily recognizable by the Myanmar people. All possible transliterations are considered, taking into account the different dialects and variations of Myanmar names are taken from Myanmar

matriculation exam results[a] and local news resources. It is aimed to access the potency of our dataset and therefore segregated it into three categories, namely training, validation and testing. Table 6 provides the statistics on each part of the dataset. To provide a clearer picture of how our dataset works, this paper included examples of the different units used in the native and western Myanmar transcription through English of person name, place name and organization name noted as (Org.) with various examples in Table 7 and 8. These examples demonstrate the variety of units used in Myanmar transliteration, including the use of character(char.), sub-syllable(sub-syl). [Ding *et al.* (2017)] and syllable(syl.) units [Thu *et al.* (2013)]. The NE data were encoded with Unicode format [Hosken and TunTunLwin (2012)] specifically because it is a widely used standard for character encoding.

	#Mix My-En Data	#Western My-En Data	#Native My-En Data
Train	282,569	125,464	153,105
Validation	2,000	2,000	2,000
Test	2,000	2,000	2,000
Total	**286,569**	**129,464**	**157,105**

Table 6. Data Statistics for Mix, Western and Native My-En NE instance pairs

Categories	Seg. Units	Western NE (My)	Western NE (En)	Translation (My-En)
Person	char.	ခ ရ စ ် စ တ ီ း နာ း	C h r i s t i n a	ခရစ်စတီးနား - Christina
	sub-syl.	ခ @ ရ စ် စ @ တ ီ း နာ း	C h r i s t i n a	ခရစ်စတီးနား - Christina
	syl.	ခ ရစ် စ တီး နား	C h r i s t i n a	ခရစ်စတီးနား - Christina
Place	char.	က ယ ် လ ီ ဖ ို့ း နီ း ယ ား	C a l i f o r n i a	ကယ်လီဖို့နီးယား - California
	sub-syl.	က ယ် လ ီ ဖ ို့ း နီ း ယ ား	C a l i f o r n i a	ကယ်လီဖို့နီးယား - California
	syl.	ကယ် လီ ဖို့း နီး ယား	C a l i f o r n i a	ကယ်လီဖို့နီးယား - California
Org.	char.	က ေ ာ ် ပ ရ ေ း ရ ှ င ် း	C o r p o r a t i o n	ကော်ပရေးရှင်း- Corporation
	sub-syl.	က ေ ာ် ပ @ ရ ေ း ရှ င် း	C o r p o r a t i o n	ကော်ပရေးရှင်း- Corporation
	syl.	ကော် ပ ရေး ရှင်း	C o r p o r a t i o n	ကော်ပရေးရှင်း- Corporation

Table 7. Sample NE segmented and Translation of Western My-En Data

Categories	Seg. Units	Native NE(My)	Native NE (En)	Translation (My-En)
Person	char.	မ ေ ာ ငဲ ထ ိ န ် လ ငဲ း	M g H t e i n L i n	မောင်ထိန်လင်း- Mg Htein Lin
	sub-syl.	မ ေ ာင် ထ ိန် လ ငဲ း	M g H t e i n L i n	မောင်ထိန်လင်း- Mg Htein Lin
	syl.	မောင် ထိန် လင်း	M g H t e i n L i n	မောင်ထိန်လင်း- Mg Htein Lin
Place	char.	ဂ ေ ါ ် ရ န ် ဂ ျ ီ က ျ ွ န ် း	G a w y a n g y i K y u n	ဂေါ်ရန်ဂျီကျွန်း- Gawyangyi Kyun
	sub-syl.	ဂ ေ ါ် ရ န် ဂ ျ ီ က ျ ွ န် း	G a w y a n g y i K y u n	ဂေါ်ရန်ဂျီကျွန်း- Gawyangyi Kyun
	syl.	ဂေါ် ရန် ဂျီ ကျွန်း	G a w y a n g y i K y u n	ဂေါ်ရန်ဂျီကျွန်း- Gawyangyi Kyun
Org.	char.	အ ို ့ း ရ ှ င ် း စ ူ ပ ါ စ ငဲ တ ာ	O c e a n S u p e r C e n t e r	အိုးရှင်းစူပါစင်တာ-Ocean Super Center
	sub-syl.	အ ို့ း ရှ ငဲ း စ ူ ပ ါ စ ငဲ တ ာ	O c e a n S u p e r C e n t e r	အိုးရှင်းစူပါစင်တာ-Ocean Super Center
	syl.	အိုး ရှင်း စူ ပါ စင် တာ	O c e a n S u p e r C e n t e r	အိုးရှင်းစူပါစင်တာ-Ocean Super Center

Table 8. Sample NE segmented and Translation of Native My-En Data

5.2. *Transformer neural network model*

One of the key advantages of the Transformer model is its aptitude to tackle long sequences of data. With RNN models [Su and C-C. Jay Kuo (2019a); Su and C-C. Jay Kuo (2022b)], [Su *et al.* (2018)] longer sequences

[a] http://myanmar.results.news/

can become computationally expensive and require a lot of memory. However, the Transformer's self-attention mechanism means that it can process long sequences more efficiently. This has made the Transformer an ideal choice for tasks such as language modelling, where long input sequences are common. Automated translation is a complex task that involves converting a sentence from one language to another. In the case of Myanmar-English transliteration, the task requires us to find a sequence output in English that has the same meaning as the input sentence in Myanmar. This process involves not only understanding the meaning of each word but also understanding the grammatical structure and cultural nuances of both languages.

The Transformer neural network architecture represents a significant breakthrough in machine learning technology. By incorporating the principles of attention and self-attention in a stack of encoders and decoders, this architecture is capable of processing massive amounts of data with exceptional speed and precision. The Transformer architecture is made up of a self-attention layer and a feedforward neural network in each encoder, and a self-attention layer, a decoder attention layer and a feedforward neural network in each decoder. In this architecture, the input data is processed by a series of encoders and decoders, with each encoder using self-attention and a feed-forward neural network to process the data. The final encoder sends the information to the decoders for further processing. Figure 2 shows the Transformer architecture for My-En NE transliteration for syllable segmentation unit.

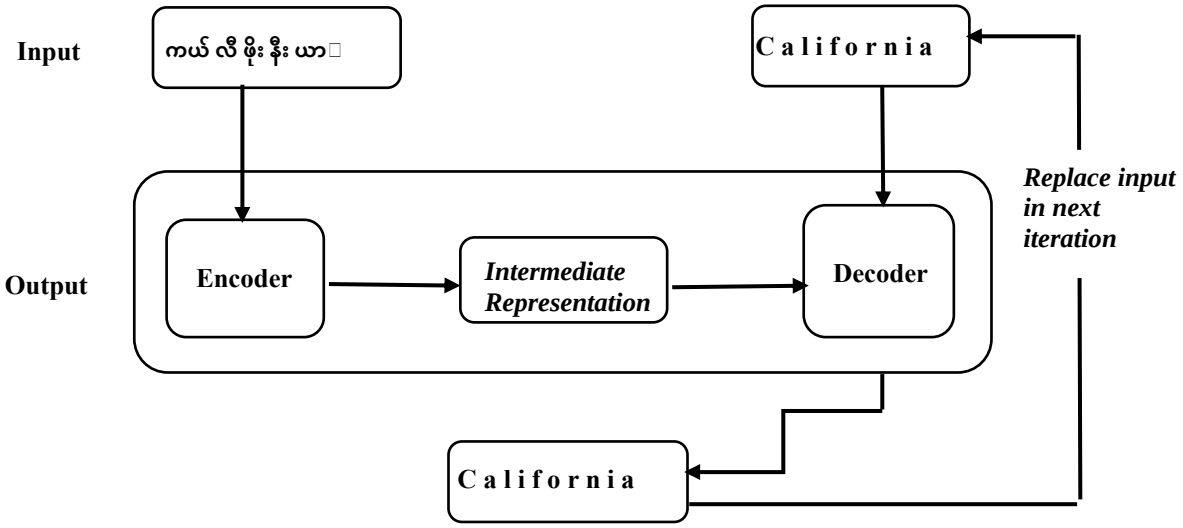

Fig. 2. Transformer Architecture for Burmese (Myanmar) to English NE Transliteration for syllable unit

6. Experimental setting

The study on machine translation involved training a Transformer model using the OpenNMT toolkit [Klein *et al*. (2017a); Klein *et al*. (2020b)]. In order to ensure reproducibility of the results and to provide transparency regarding the experimental setup, the hyper-parameters employed in the experiments have been expressed in the Table 9. These hyper-parameters were chosen after conducting a thorough literature review. The performance of the model was evaluated using a variety of metrics, which are discussed in detail in the subsequent paragraphs.

In the realm of contemporary NMT systems, the Transformer neural network architecture is the most effective choice due to its superior performance in terms of quality and efficiency. This architecture has been shown to outperform other mainstream NMT architectures [Su *et al.* (2019)], including deep RNN and CNN. Traditional sequence-to-sequence models have long relied on recurrent networks such as LSTM or GRU. However, the Transformer neural network architecture, as outlined in [Vaswani *et al.* (2017)], eliminates the need for such networks. This leads to better parallelization during model training, ultimately reducing the amount of time required for training. It enables to streamline our training process and achieve optimal results for the Transformer model using OpenNMT toolkit. By leveraging the powerful features of this toolkit, it also enables to fine-tune our system and optimize its performance. The consistent use of hyper-parameters on OpenNMT helped us to ensure that our model was trained to meet the specific needs and objectives. For deep learning experiment, we utilized Google Colaboratory with a single GPU to train a Transformer neural network on our prepared dataset. In the conducted Transformer experiment, specific hypher-parameter settings were crucial for the study's methodology, and these settings are detailed in Table 9. Analyzing Table 9 provides insights into the key parameters that influenced the outcomes of the Transformer experiment.

#Parameter	#Setting
Model Architecture	Transformer
Number of Layers	6
RNN Size	512
Word Vector Size	512
Transformer Feed-Forward Size	2048
Multi-Head Attention Heads	8
Encoder Type	Transformer
Decoder Type	Transformer
Position Encoding	Not Used
Training Steps	50,000
Maximum Generator Batches	2
Dropout	0.1
Batch Type	Tokens
Normalization	Tokens
Batch Size	1,024
Gradient Accumulation Batches	Every 2 Batches
Optimizer	Adam
Beta 2	0.998
Decay Method	Noam
Warm-up Steps	8,000
Learning Rate	2
Maximum Gradient Norm	0
Parameter Initialization Glorot	Not Used
Label Smoothing	0.1
Validation Frequency	Every 10,000 Steps
Checkpoint Saving Frequency	Every 10,000 Steps
World Size	1
GPU Rank	Single GPU Rank

Table 9. Hypher parameter settings for Transformer experiment

7. Experimental Results

Overall investigations are evaluated on three kinds of our prepared data: 286,569 mixing native and western My-En NE instance pairs,129,464 western My-En NE instance pairs and 157,105 native My-En NE pairs with three segmentation units using Transformer neural network models in order to determine their effectiveness in accurately transliterating between two languages. Based on the evaluation, the most significant result for system performance was the BLEU score [Kumai *et al.* (2008)] and WER [Klakow and Peters (2002)] for the Mix character units on En-My and native sub-syllable units on My-En transformer systems. It was found that the Mix character system achieved a BLEU score of 72, while the native sub-syllable-based system attained a BLEU score of 71. These results indicate that both systems performed reasonably well in generating accurate translations. Furthermore, both systems had low word error rate (WER) which further supports their effectiveness. Overall, their evaluation results suggest that the Mix-character performance-based segmentation units and native sub-syllable-based segmentation units for Transformer NN models are the most effective in terms of system. The evaluation results are presented in detail in Table 10,11 and 12 which shows the overall system performance with regard to BLEU and WER.

#Data	#Seg. Units	#En-My		#My-En	
		BLEU	**WER**	**BLEU**	**WER**
Mix Data	Char.	**72**	**0.22**	58	0.22
	Sub-Syl.	63	0.25	56	0.24
	Syl.	45	0.67	50	0.32

Table 10. System evaluation results for Mix data in term of BLEU and WER

#Data	#Seg. Units	#En-My		#My-En	
		BLEU	**WER**	**BLEU**	**WER**
Western Data	Char.	54	0.34	65	0.21
	Sub-Syl.	54	0.32	66	0.19
	Syl.	44	0.47	66	0.22

Table 11. System evaluation results for Western data in term of BLEU and WER

#Data	#Seg. Units	#En-My		#My-En	
		BLEU	**WER**	**BLEU**	**WER**
Native Data	Char.	58	0.55	52	0.40
	Sub-Syl.	56	0.23	**71**	**0.18**
	Syl.	50	0.39	52	0.22

Table 12. System evaluation results for Native data in term of BLEU and WER

7.1. *Findings and discussions*

The case study being conducted on transliteration instances were instrumental in helping us identify the key challenges and limitations of our model. The results of our investigations are presented in Table 13, which compare the outputs of different data sets for the same transliteration occurrences. These case studies highlight the need for a more nuanced and context-sensitive approach to cross-lingual text conversion. Transliterating borrowed English words into Myanmar is a difficult task due to the fact that the transcription may contain inaccurately spelled Myanmar words. Table 13 illustrates this point with the example of the pair "Cardiff" and "ကားဒစ်ဖ်." In Myanmar native language, it is not permissible to use <စ်ဖ်> to represent <iff>, making it difficult for syllable-based processing systems to handle such exceptional structures. Another NE instance also provides a challenging example of the pair "Djokovic" and "ဂျိုကိုဗစ်," where all existing systems failed to provide accurate results for both the En→My and My→En directions. The spelling of <Djo> caused difficulty in En→My processing. While in My→En processing, all systems indicated the more usual spelling of “Jokovic” instead of "Djokovic." The En→My processing was hampered by <Djo> as it caused transcription difficulties where <d> had to be transcribed separately as <ဒ…>. These challenges can be solved by doing further investigations to develop more accurate transliteration systems.

Data	**Seg. Units**	**Reference (My)**	**Reference (En)**	**En→My (Hypotheses)**	**My→En (Hypotheses)**
Mix	char.	ကားဒစ်ဖ်	Cardiff	ကာဒီဖ်	Car**dift**
	sub-syl.	ကားဒစ်ဖ်	Cardiff	ကာဒီအက်ဖ်	Car**dif**
	syl.	ကားဒစ်ဖ်	Cardiff	ကာဒီအက်ဖ်အက်ဖ်	Car**ဒစ်ဖ်**
	char.	ခိုင်ဇင်သန့်	Khaing Zin Thant	ခိုင်ဇင်သန့်	**Khaing** Zin Thant
	sub-syl.	ခိုင်ဇင်သန့်	Khing Zin Thant	ခိုင်ဇင်သန့်	**Khing** Zin Thant
	syl.	ခိုင်ဇင်သန့်	Khine Zin Thant	ခိုင်ဇင်သန့်	**Khine** Zin Thant
Western	char.	ကားဒစ်ဖ်	Cardiff	ကာဒီဖ်	**Kadif**
	sub-syl.	ကားဒစ်ဖ်	Cardiff	ကားဒစ်ဖ်	**Kadif**
	syl.	ကားဒစ်ဖ်	Cardiff	ကာဒစ်	**Kadif**
	char.	ဂျိုကိုဗစ်	Djokovic	ဒီဂျိုကိုဗစ်	**Jo**kovic
	sub-syl.	ဂျိုကိုဗစ်	Djokovic	ဒီဂျိုကိုဗစ်	**Jo**kovic
	syl.	ဂျိုကိုဗစ်	Djokovic	ဒီဂျိုးကိုဗစ်	**Jo**kovic
Native	char.	ခင်လပြည့်ဝင်း	Khin La Pyae Win	ခင်လပြည့်ဝင်း	Khin La Pyae **Win**
	sub-syl.	ခင်လပြည့်ဝင်း	Khin La Pyae Wynn	ခင်လပြည့်ဝင်း	Khin La Pyae **Wynn**
	syl.	ခင်လပြည့်ဝင်း	Khin La Pyae Winn	ခင်လပြည့်ဝင်း	Khin La Pyae **Winn**
	char.	စန္ဒာထွန်း	Sandar Htun	စန္ဒာထွန်း	Sandar **Htun**
	sub-syl.	စန္ဒာထွန်း	Sandar Htoon	စန္ဒာထွန်း	Sandar **Htoon**
	syl.	စန္ဒာထွန်း	Sandar Tun	စန္ဒာထွန်း	Sandar **Tun**

Table 13. Findings and discussions on some hypotheses results

For native Myanmar name entities, our transliteration model has achieved impressive results when dealing with one-to-many associations for Myanmar native names. For instance, consider names like ခင်လပြည့်ဝင်း (Khin La Pyae Win) and စန္ဒာထွန်း (Sandar Htun). These names illustrate a unique aspect of the Myanmar language: a single Myanmar syllable can yield multiple possible transliterations. For example, the syllable "Win" can be transliterated as "Winn" or "Wynn," while "Tun" can be represented as "Htun" or "Htoon." Our model's ability to handle such variations is a testament to its versatility and its effectiveness in preserving the

richness and diversity of the Myanmar language. Based on the results presented in Table 13, it can be concluded that utilizing character or sub-syllable units in Myanmar is preferable to using syllables for transliteration. The reason is that Myanmar syllables have limited processing capability when dealing with exceptional structures, whereas sub-syllables and characters provide more flexibility.

8. Conclusion and Future Works

Based on our findings, an internal dictionary of Myanmar-English NE terminology that incorporates transliterations for both local and western languages have been developed. The experiments involved utilizing Transformer-based NN techniques to assess their efficacy on the data we prepared. Despite the limited scope of the NE corpus utilized, our neural network models were able to achieve satisfactory outcomes for transliteration duties. It is anticipated that as more data is gathered and further testing is conducted, neural network-based transliteration models will become increasingly relevant in this domain. This investigation represents the initial exploration of utilizing neural networks for Myanmar NE transliteration. In the future, developing an interactive and user-friendly transliteration tool or application that incorporates the Transformer model can facilitate real-world applications and assist users in seamlessly transcribing Myanmar named entities into English, potentially aiding in fields like language translation, information retrieval, and natural language processing.

Acknowledgments

I would like to express my appreciation to my dedicated supervisor for their continuous guidance, motivation, and unwavering support throughout the course of this study. Their extensive knowledge and innovative thinking have been indispensable pillars of support throughout my work.

Conflicts of Interest

The authors state no conflict of interest

References

[1] Bradley D. “Burmic languages in Myanmar”. Continuum of richness of languages and dialects in Myanmar. 2015:167-90.
[2] Chang, C.B., 2009. English loanword adaptation in Burmese. Journal of the Southeast Asian Linguistics Society, 1, pp.77-94.
[3] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
[4] Ding, C., Pa, W.P., Utiyama, M. and Sumita, E., 2018. Burmese (Myanmar) name romanization: A sub-syllabic segmentation scheme for statistical solutions. In Computational Linguistics: 15th International Conference of the Pacific Association for Computational Linguistics, PACLING 2017, Yangon, Myanmar, August 16–18, 2017, Revised Selected Papers 15 (pp. 191-202). Springer Singapore.
[5] Ding C. Transliteration of Foreign Words in Burmese. arXiv preprint arXiv:2110.03163. 2021 Oct 7.
[6] Ding, C., Utiyama, M. and Sumita, E., 2018. NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(2), pp.1-18.
[7] Ding, C., Aye, H.T.Z., Pa, W.P., Nwet, K.T., Soe, K.M., Utiyama, M. and Sumita, E., 2019. Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(1), pp.1-34.
[8] Ding, C., Yee, S.S.S., Pa, W.P., Soe, K.M., Utiyama, M. and Sumita, E., 2020. A Burmese (Myanmar) treebank: Guideline and analysis. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(3), pp.1-13.
[9] Hosken M, TunTunLwin M. “Representing Myanmar in Unicode”. Unicode Technical Note. 2012;13:1-67.
[10] Koehn, P., Och, F.J. and Marcu, D., 2003. Statistical phrase-based translation. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (pp. 127-133).
[11] Koehn, P., Federico, M., Shen, W., Bertoldi, N., Bojar, O., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Zens, R. and Constantin, A., 2007, August. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. In CLSP Summer Workshop Final Report WS-2006, Johns Hopkins University.
[12] Klein G, Hernandez F, Nguyen V, Senellart J. The OpenNMT neural machine translation toolkit: 2020 edition. InProceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track) 2020 Oct (pp. 102-109).
[13] Klein G, Kim Y, Deng Y, Senellart J, Rush AM. “Opennmt: Open-source toolkit for neural machine translation”. arXiv preprint arXiv:1701.02810. 2017 Jan 10.
[14] Kumai H, Sagawa H, Morimoto Y. “NTCIR-7 Patent Translation Experiments at Hitachi”. InNTCIR 2008.
[15] Klakow D, Peters J. “Testing the correlation of word error rate and perplexity”. Speech Communication. 2002 Sep 1;38(1-2):19-28.
[16] Li, H., Kumaran, A., Pervouchine, V. and Zhang, M., 2009, August. Report of NEWS 2009 machine transliteration shared task. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009) (pp. 1-18).
[17] Liu, L., Utiyama, M., Finch, A. and Sumita, E., 2016, June. Agreement on target-bidirectional neural machine translation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 411-416).
[18] Merhav Y, Ash S. Design challenges in named entity transliteration. arXiv preprint arXiv:1808.02563. 2018 Aug 7.
[19] Mon, A.M. and Soe, K.M., 2020, November. Phrase-based named entity transliteration on Myanmar-English terminology dictionary. In 2020 23rd conference of the oriental COCOSDA international committee for the co-ordination and standardisation of speech databases and assessment techniques (O-COCOSDA) (pp. 38-43). IEEE.

[20] Naing, H.M.S., Thu, Y.K., Pa, W.P., Kato, H., Finch, A., Sumita, E. and Hori, C., 2015. Rule Based Katakana to Myanmar Transliteration for Post-editing Machine Translation. In Proceedings of the Annual Conference of the Language Processing Society of Japan (pp. 257-260).
[21] Och, F.J. and Ney, H., 2003. A systematic comparison of various statistical alignment models. Computational linguistics, 29(1), pp.19-51.
[22] Och, F.J., 2003, July. Minimum error rate training in statistical machine translation. In Proceedings of the 41st annual meeting of the Association for Computational Linguistics (pp. 160-167).
[23] ShweSin, Y.M., Soe, K.M. and Htwe, K.Y., 2018, October. Large scale Myanmar to English neural machine translation system. In 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE) (pp. 464-465). IEEE.
[24] Su, Yuanhang, and C-C. Jay Kuo. "On extended long short-term memory and dependent bidirectional recurrent neural network." Neurocomputing 356 (2019): 151-161.
[25] Su, Yuanhang, and C-C. Jay Kuo. "Recurrent neural networks and their memory behavior: a survey." APSIPA Transactions on Signal and Information Processing 11, no. 1 (2022).
[26] Su, Yuanhang, Yuzhong Huang, and C-C. Jay Kuo. "Dependent bidirectional RNN with extended-long short-term memory." (2018).
[27] Su, Yuanhang, Kai Fan, Nguyen Bach, C-C. Jay Kuo, and Fei Huang. "Unsupervised multi-modal neural machine translation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10482-10491. 2019.
[28] Thu, Y. K., Finch, A., Sagisaka, Y., & Sumita, E. (2013). “A study of myanmar word segmentation schemes for statistical machine translation” (Doctoral dissertation, MERAL Portal).
[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Authors Profile

In 2012, **Aye Myat Mon** earned her Master of Computer Science (M.C.Sc.) degree with credits from the University of Computer Studies, Pathein, Myanmar. She is currently pursuing her Ph.D. at the University of Computer Studies, Yangon, Myanmar, and is an active member of the Natural Language Processing and Speech Processing Lab at UCSY. Her research interests encompass Natural Language Processing, Machine Learning, and Deep Learning. Aye Myat Mon has also been involved in the Asia Language Treebank (ALT) projects, a collaborative endeavor between the National Institute of Information and Communications Technology (NICT), Japan, and UCSY. In addition, she completed an internship at the Advanced Translation Technology Laboratory (ASTREC), Universal Communication Research Institute, NICT, Kyoto, Japan, from April 2019 to March 2020.

Dr. Khin Mar Soe completed her Ph.D. in Information Technology in the year 2005. Currently, she holds the position of a professor and serves as the Head of the Natural Language Processing Lab within the Faculty of Computer Science at the University of Computer Studies, Yangon. She actively supervises Master's theses and Ph.D. research in the field of Natural Language Processing. Additionally, Dr. Khin Mar Soe has contributed to projects such as ASEAN MT, a machine translation initiative for South East Asian languages, and the Asia Language Treebank (ALT) project, which involved research collaboration between NICT, Japan, and UCSY.