BURMESE (MYANMAR) – ENGLISH NAMED ENTITY TRANSLITERATION

AYE MYAT MON

UNIVERSITY OF COMPUTER STUDIES, YANGON

JUNE, 2024

Burmese (Myanmar) – English Named Entity Transliteration

Aye Myat Mon

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

June, 2024

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

…… …….

Date Aye Myat Mon

ACKNOWLEDGEMENTS

First of all, I would like to thank the Union Minister, the Ministry of Science and Technology for full facilities support during the Ph.D. Course at the University of Computer Studies, Yangon.

I also wish to offer my deep gratitude to Professor Dr. Mie Mie Khin, Rector of the University of Computer Studies, Yangon, for her contribution to the thesis and overall supporting during my thesis.

I extend my sincere gratitude to Dr. Mie Mie Thet Thwin, former Rector of the University of Computer Studies, Yangon, for giving the opportunity to do this research.

I would also like to express my special appreciation and thanks to Dr. Si Si Mar Win, Professor and Course-coordinator of the Ph.D 11th Batch, University of Computer Studies, Yangon, for the kindness, useful comments, advice and insight which are invaluable to me.

I extend my sincere gratitude to Dr. Tin Thein Thwel, Professor and former Course-coordinator of the Ph.D 11th Batch, University of Computer Studies, Yangon, for the kindness, useful comments, advice and insight which are invaluable to me.

I would like to express my deepest gratitude and appreciation to my research supervisor Dr. Khin Mar Soe, Professor, Head of Faculty of Computer Science, University of Computer Studies, Yangon, for her patient supervision, close guidance and constant encouragement throughout the term of my thesis work. Her precious suggestions and supervisions are invaluable not only for this thesis but also a long-life education.

I would like to express my deepest appreciation to the external examiner, Dr. Aung Nway Oo, Professor, Faculty of Computer Science, the University of Information Technology, for his patience in critical reading, valuable comments and suggestions in the preparation of the dissertation.

I would also like to express my humble gratitude to Dr. Win Pa Pa, Professor, Faculty of Computer Science, University of Computer Studies, Yangon, for giving her valuable advice, helpful comments and precious time to my thesis.

I would like to express my respectful gratitude to Professor Daw Aye Aye Khine, Head of Department of English for her overall supporting throughout my Ph.D. course work, doing research and for editing the drafts of the manuscript.

I would also like to take this opportunity to express my deep gratitude to my parents who supported and encouraged me not to give up my chance. Finally, I would like to thank all my friends who are also supportive and kind to me all the time.

ABSTRACT

This system presents a novel approach to Burmese (Myanmar) -English Named Entity Transliteration System leveraging Transformer models, focusing on character, sub-syllable, and syllable segmentation based on a meticulously prepared dictionary containing both foreign and native Myanmar-English entries. Transliterating named entities accurately between Myanmar and English poses significant challenges due to script differences, linguistic nuances, and varying entity structures. The proposed system addresses these challenges by incorporating advanced segmentation techniques and a comprehensive dictionary. The core of the approach lies in the segmentation of Myanmar named entities into character-level, sub-syllable, and syllable units, utilizing linguistic knowledge and domain-specific dictionaries. Linguistic rules are employed to segment Myanmar text into meaningful units, capturing the rich morphology and orthographic complexities of the Myanmar script. This segmentation process is crucial for accurately aligning Myanmar entities with their English transliterations. The system is built upon the Transformer architecture, a state-of-the-art deep learning model renowned for its sequence-to-sequence capabilities and attention mechanisms. The Transformer model is trained on a large corpus derived from our prepared Myanmar-English dictionary, learning the intricate mappings and transliteration patterns between the two languages. The performance of the system is evaluated using a benchmark dataset comprising diverse Myanmar named entities and their corresponding English transliterations. The experimental results demonstrate the efficacy of the approach, achieving superior transliteration accuracy compared to baseline methods. Extensive analyses are also conducted to investigate the impact of different segmentation strategies, dictionary sizes, and model configurations on transliteration quality. In conclusion, the Myanmar-English Named Entity Transliteration System based on character, sub-syllable, and syllable segmentation, coupled with a meticulously prepared dictionary, represents a significant advancement in cross-lingual natural language processing. The system offers a reliable and efficient solution for transliterating Myanmar named entities into English with exceptional accuracy and scalability, paving the way for enhanced multilingual communication and data interoperability.

CONTENTS

PAGE

List of Figures

List of Tables

List of Equations

CHAPTER 1

INTRODUCTION

As Natural Language Processing (NLP) continues to advance, automated transliteration of named entities has emerged as a critical component in various language-related applications. Whether it is converting text to speech or developing systems for machine translation, the ability to transliterate words from one alphabet to another is indispensable. By providing a phonetic representation of words, transliteration assists individuals in pronouncing and understanding unfamiliar terms, thus enabling effective communication across different languages and cultures.

Named entity (NE) transliteration is not new in the areas of other Asian Language processing. Special writing systems are employed in many Asian languages, leading to concentrated efforts in transliteration processing for prominent languages such as Chinese, Japanese, and Korean. However, it is imperative to conduct research on understudied languages that have limited resources. In general, the transliteration task can be seen as a simplified translation task conducted at the character or grapheme level, rather than at the word or phrase level. In the contemporary context of the Myanmar language, a significant number of borrowed words from English are found. Presently, there is an absence of a consistent standard in the Myanmar language for transcribing the rising number of borrowed words from or through English.

The complexity of human language translation in natural language processing stems from the fact that language ambiguity is influenced by the distinct features and characteristics of each language. This challenge is evident not only in Myanmar but also in other Asian languages such as Indian, Japanese, Thai, and Chinese. Word transformations in Myanmar align with those observed in other Asian languages, highlighting the shared linguistic traits and patterns among these languages. The writing system of a language also governs the types and trends of pronunciations of that language for transliteration. Myanmar is also among the languages whose writing system is different from that of English and therefore existing techniques cannot be applied for Myanmar NE transliteration without modifications. In the Myanmar script, there is a relative redundancy in its phonology inventory, allowing for the expression of phonemes in multiple ways. Furthermore, intentional use of special spellings may be employed to impart borrowed words with an exotic visual allure. Moreover, irregular transliteration can occur when the transcription adheres to the spelling

conventions of English, rather than accurately reflecting the actual pronunciation. As a result of transliteration between Myanmar and English becomes correspondingly complex.

This document presents the details of a study performed on Myanmar language to identify the problem areas of Myanmar NE transliteration and to test and evaluate the effectiveness of automatic transliteration performance using neural network-based approaches based on the prepared data.

1.1 Problem Statement

The problem statement of Myanmar-English named entity transliteration is the need for an automated system that can accurately transliterate named entities, such as names of people, places, organizations, and other proper nouns, from the Myanmar language to the English language. Transliteration involves representing words or phrases from one script or alphabet to another, while maintaining their phonetic pronunciation. Myanmar, also known as Burmese, uses a non-Latin script, making it challenging for English speakers or systems to accurately understand and pronounce Myanmar names. This poses difficulties in various applications like machine translation, speech recognition, and information retrieval, where accurate transliteration is crucial for effective communication and comprehension. The goal is to develop an efficient and reliable transliteration system that can bridge the gap between the Myanmar and English languages, enabling accurate representation and pronunciation of named entities in both languages.

1.2 Natural Language Processing

Artificial Intelligence is rapidly changing the world, affecting every aspect of our daily lives. From voice assistants using NLP and Machine Learning for making appointments in the calendar, and playing music to automatically suggesting products, so accurately that they can guess what we will need in advance. Natural Language Processing (NLP) is itself a broad field that lies under Artificial Intelligence. NLP depends upon linguistics and is responsible for making computers understand the text and spoken words the same way humans do [16].

Natural Language Processing (NLP) is an interdisciplinary field that combines computational linguistics, statistics, machine learning, and deep learning to enable computers to understand and process human language. With a history spanning several decades, NLP has emerged

as a prominent and rapidly evolving field in the world of technology. NLP-powered software applications have become ubiquitous, assisting us in various aspects of our lives. Personal assistants like Siri, Cortana, and Google Assistant employ NLP algorithms to understand and fulfill voice-based commands, while machine translation tools like Google Translator utilize NLP techniques to bridge language barriers. Grammar checking applications such as Grammarly leverage NLP to enhance writing accuracy, and autosuggestion features in search engines, Gmail, and developer's IDE rely on NLP to provide relevant suggestions and improve productivity.

The field of Natural Language Processing (NLP) is driven by the aspiration to develop computational models that accurately represent the complexities of human language. The objective is to create computer programs capable of performing a wide range of tasks involving natural language. However, effective communication through language extends beyond linguistic rules and structures. It necessitates the incorporation of common sense and world knowledge, as well as an understanding of contextual information. The ultimate ambition in NLP research is to devise models that approach human-level performance in reading, writing, listening, and speaking, enabling machines to engage in language-based tasks with a level of proficiency comparable to humans.

There are three major concerns in NLP that are described in the next three sections.

1.2.1 Natural Language Understanding

Natural Language Understanding (NLU) refers to the field of artificial intelligence (AI) that focuses on enabling machines to comprehend and interpret human language in a way that is similar to how humans understand it. NLU aims to bridge the gap between human communication and machine comprehension by enabling computers to extract meaning, infer intent, and derive insights from textual or spoken language data.

At its core, NLU involves the development of algorithms and models that can process and analyze natural language input, such as written text or spoken words, and extract relevant information from it. This includes tasks such as language parsing, part-of-speech tagging, semantic analysis, entity recognition, sentiment analysis, and more. By leveraging machine learning and statistical techniques, NLU systems can learn from large amounts of labeled data to understand the underlying structure, context, and meaning of human language.

NLU plays a crucial role in a wide range of applications and technologies. It powers virtual assistants like Siri, Alexa, and Google Assistant, enabling users to interact with these systems using natural language commands or queries. NLU is also employed in chatbots, customer support systems, sentiment analysis tools, language translation services, and information retrieval systems, among others. By understanding and interpreting human language, NLU systems enable more effective and intuitive human-computer interaction.

One of the key challenges in NLU is dealing with the inherent ambiguity and complexity of natural language. Human language is rife with nuances, idioms, context-dependent meanings, and variations across different cultures and regions. NLU systems must account for these intricacies and be able to disambiguate and interpret language accurately. This requires robust models and techniques that can handle semantic understanding, context comprehension, and even contextual disambiguation.

The advancements in deep learning, neural networks, and natural language processing have significantly contributed to the progress in NLU. Deep learning models, such as recurrent neural networks (RNNs) and transformers, have revolutionized language understanding tasks by capturing complex patterns and dependencies in textual data. By training on large-scale datasets, these models can learn to generalize and make accurate predictions, thus enhancing NLU capabilities.

1.2.2 Natural Language Generation

Natural language generation (NLG) represents the other side of the coin in NLP. NLG involves the computer's ability to generate text that exhibits the characteristics of natural language, differentiating it from the more traditional forms of computer-generated content. One of the inherent limitations of computer-generated content is its lack of fluidity, emotional depth, and human-like personality. However, NLG integrates NLP techniques to enable computers to produce text that emulates human writing. This is accomplished by identifying the central theme of a document and employing NLP to determine the most suitable approach to express the content in the user's native language. The resulting text is then generated accordingly.

1.2.3 Natural Language Acquisition

Natural Language Acquisition (NLA) refers to the process by which humans acquire and learn a natural language, such as their native language, through exposure and interaction with their environment. It is the ability to understand and produce language effortlessly, starting from infancy and continuing throughout a person's development. Natural Language Acquisition occurs through a combination of innate language learning mechanisms, exposure to linguistic input, and social interactions.

Examples of Natural Language Acquisition can be observed in children as they progress from early language development to fluency in their native language. Around 10 to 12 months, children start producing their first recognizable words. They acquire vocabulary by associating sounds with objects, actions, and concepts in their environment. For example, a child might say "mama" or "dada" to refer to their parents.

1.3 Natural Language Processing Using Statistical Approach

The successful application of statistical methods in natural language processing has been highly notable over the past two decades. The extensive accessibility of text and speech corpora has been a vital factor contributing to this success. Like other learning techniques, statistical approaches heavily depend on data, and the abundance of corpora has provided ample resources for their implementation. These methods make use of diverse mathematical techniques and leverage large text corpora to construct approximate generalized models of linguistic phenomena. These models are developed based on real-world examples found within the corpora, without the need for substantial linguistic or world knowledge.

When it comes to determining the structure of text, an NLP system can benefit from employing a Statistical NLP approach, which excels in making disambiguation decisions concerning word sense, word category, syntactic structure, and semantic scope. This approach tackles these challenges by autonomously learning lexical and structural preferences from corpora. Instead of relying solely on parsing based on syntactic categories, such as part of speech labels, the system acknowledges the wealth of information residing in the relationships between words, specifically, the tendencies of words to cluster together. This knowledge of collocation can serve as a valuable

insight into deeper semantic relationships. Notably, statistical models offer an effective solution to the issue of ambiguity: they exhibit robustness, generalize well, and handle errors and new data gracefully.

Statistical NLP models have paved the way for effective disambiguation in large-scale systems that process natural language text. These models excel in unraveling the inherent ambiguity and diverse interpretations found in language. A notable advantage of Statistical NLP is the automatic estimation of model parameters from text corpora. This automatic learning capability not only reduces the human effort involved in building NLP systems but also raises intriguing scientific questions regarding the mechanisms underlying human language acquisition [33].

1.4 Natural Language Processing Using Neural Network Approach

Natural Language Processing (NLP) using a neural network approach has revolutionized the field by leveraging the power of artificial neural networks to process and understand human language. Neural networks, particularly deep learning models, have demonstrated exceptional performance in various NLP tasks, ranging from sentiment analysis and named entity recognition to machine translation and question-answering systems.

One of the key advantages of neural network-based NLP is their ability to capture complex patterns and representations in language data. By utilizing deep neural architectures such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer models, NLP systems can effectively model the sequential and contextual nature of language, capturing dependencies and long-range dependencies within sentences and documents.

Neural network models for NLP typically involve training on large-scale annotated datasets, allowing them to learn from vast amounts of labeled examples and generalize well to unseen data. This data-driven approach enables neural networks to automatically extract meaningful features and representations from raw text, bypassing the need for explicit feature engineering.

Moreover, neural network-based NLP models can leverage pre-training techniques such as word embeddings or contextualized word representations like word2vec, GloVe, or BERT. These pre-trained representations capture semantic and syntactic information about words, enabling the

neural networks to encode rich linguistic knowledge and enhance their performance on downstream NLP tasks.

With the advancement of deep learning techniques and the availability of large-scale datasets, neural network-based NLP models have achieved state-of-the-art results in various applications. They have significantly improved the accuracy and efficiency of tasks such as text classification, sentiment analysis, natural language understanding, and language generation.

However, neural network-based NLP models also face challenges. They require substantial computational resources for training and inference due to the complexity of deep neural architectures. Additionally, they may struggle with rare or out-of-vocabulary words and can be sensitive to the quality and biases present in the training data.

Despite these challenges, the neural network approach to NLP continues to advance rapidly, with ongoing research focusing on developing more efficient architectures, better handling of rare words, addressing biases, and improving interpretability. Overall, the neural network approach has brought significant advancements to the field of NLP, pushing the boundaries of what can be achieved in understanding and processing human language.

1.5 Applications of Natural Language Processing

Natural Language Processing (NLP) has found applications in various domains, with one of the most significant being the translation of natural languages. In today's world, an abundance of new named entities emerges daily from sources such as newspapers, websites, and technical literature. However, their translations are often absent from conventional translation dictionaries. Enhancing the transliteration of named entities holds great importance for translation systems and cross-language information retrieval applications. Additionally, it facilitates the acquisition of bilingual resources from the web and aids in the extraction of translation knowledge from corpora. While named entities encompass a wide range of concepts, including people names, place names, organization names, and product names, this thesis focuses specifically on the transliteration of named entities between Myanmar and English languages. Application areas for Natural Language Processing in NE transliteration are indicated as below:

- Machine Translation

- Machine Transliteration
- Grapheme to Phoneme Conversion
- Named Entity Recognition
- Information Retrieval
- Text Mining
- Automatic text summarization
- Optical Character Recognition
- Automatic Speech Recognition
- Spelling correction/ Spell checker
- Word Segmentation
- Named entity identification and information extraction
- Text-to-speech synthesis
- Search Engine
- Question Answering etc.

Within our curated selection, we will spotlight the work that offers the most comprehensive understanding of NE transliteration principles.

1.5.1 Machine Translation

Machine Translation (MT) refers to the automated process of translating text or speech from one natural language to another using computational algorithms and techniques. It aims to bridge language barriers and facilitate communication between individuals who speak different languages. Machine translation systems utilize various approaches, ranging from rule-based methods to statistical and neural network-based models, to achieve translation.

Rule-based machine translation systems rely on linguistic rules and dictionaries to analyze the structure and meaning of the source language text and generate corresponding translations in the target language. These systems require extensive manual efforts in creating linguistic resources and rule sets, which can be time-consuming and challenging to maintain.

Statistical machine translation (SMT) approaches emerged as a significant breakthrough in the field. SMT systems rely on statistical models that learn translation patterns from parallel corpora,

which are collections of aligned texts in both the source and target languages. By analyzing these bilingual datasets, SMT models estimate the probabilities of generating target language translations given the source language input. The translation process involves selecting the most probable translation based on these probabilities. SMT systems can handle large-scale translation tasks and have shown promising results, especially when trained on vast amounts of high-quality parallel data.

More recently, neural machine translation (NMT) models have gained prominence in the field. NMT models utilize artificial neural networks, particularly sequence-to-sequence models, to directly translate text from one language to another. These models learn to encode the source language sentence into a fixed-length representation (often called an "embedding") and then decode it into the target language sentence. NMT models have shown superior performance compared to earlier approaches, capturing complex linguistic patterns and producing more fluent and accurate translations. They have benefited from advances in deep learning and have become the de facto standard in many machine translation systems.

Machine translation has both advantages and limitations. It offers a fast and cost-effective solution for translating large volumes of text, enabling communication across language barriers in various domains such as business, education, and research. However, achieving perfect translations remains a challenge. Machine translation systems often face difficulties with idiomatic expressions, context-dependent translations, rare or domain-specific vocabulary, and preserving the nuances and cultural aspects of the source text. Human post-editing is often necessary to ensure the quality and accuracy of machine-translated output, especially in professional translation settings.

Ongoing research and development in machine translation continue to advance the state of the art. This includes improving the training data quality, refining neural network architectures, incorporating advanced techniques such as attention mechanisms and transformer models, and exploring methods to enhance domain adaptation and language-specific challenges. Machine translation has made significant progress over the years and continues to play a vital role in breaking down language barriers and facilitating global communication [4] [53].

1.5.2 Machine Transliteration

Machine transliteration refers to the process of automatically converting text from one writing system or script into another. It involves mapping the characters or symbols of one script to their corresponding counterparts in another script, while attempting to preserve the pronunciation and phonetic representation of the original text.

Transliteration is often necessary when dealing with languages that use different writing systems, such as converting names or phrases from one language to another that may have different alphabets or scripts. For example, transliteration is commonly used when translating names between languages like Arabic, Chinese, Cyrillic, or Devanagari scripts and Latin script.

Machine transliteration algorithms typically rely on statistical models or rule-based approaches. Statistical models leverage large datasets of parallel texts in different scripts to learn patterns and mappings between characters. Rule-based approaches use predefined rules and linguistic knowledge to perform the transliteration.

Machine transliteration can be challenging due to the complexities and ambiguities present in languages and their writing systems. Different languages and scripts may have similar-looking characters that represent different sounds, while a single character can have multiple pronunciations or transliterations. Additionally, some languages have specific phonetic characteristics that make transliteration more difficult.

Overall, machine transliteration aims to automate the process of converting text from one script to another, making it easier to communicate and understand content across languages with different writing systems [59].

1.5.3 Grapheme to Phoneme Conversion

Grapheme-to-phoneme (G2P) conversion plays a crucial role in bridging the gap between written words and their phonetic representations. By converting the grapheme sequence, which represents the spelling of a word, into a phoneme sequence, which represents its phonetic form, G2P techniques enable the development of a phonemic lexicon in TTS and ASR systems. The accuracy of G2P conversion is paramount in achieving state-of-the-art performance in these systems, as it directly impacts their overall effectiveness. In ASR systems, where acoustic models are integral components, the pronunciation lexicons and language models rely on accurate G2P conversion.

These models are automatically constructed using large corpora. Serving as the intermediary layer between acoustic and language models, pronunciation lexicons significantly contribute to the performance of a new speech recognition task. The overall system's efficiency and accuracy heavily depend on the quality of the pronunciation component, which is directly influenced by G2P accuracy. For instance, the word 'speaker' undergoes G2P conversion resulting in the phoneme sequence 'S P IY K ER'. The overall quality of TTS systems relies heavily on the inclusion of a high-quality G2P model. If G2P conversion is inaccurate, the result is synthetic speech with unnatural pronunciation or speech that is difficult to understand.

In recent times, the utilization of neural networks has been prevalent in G2P conversion. This approach proves resilient to spelling errors and out-of-vocabulary (OOV) words, showcasing a strong ability to generalize. Additionally, it seamlessly fits into end-to-end TTS/ASR systems, which are predominantly constructed using deep neural networks. Consequently, the G2P model is trained together with other crucial components of the speech synthesizer and recognizer, enhancing the overall quality of the system [60].

1.5.4 Named Entity Recognition

Named Entity Recognition (NER) is a natural language processing (NLP) technique that involves identifying and classifying named entities in text. Named entities are specific words or phrases that represent real-world objects, such as persons, organizations, locations, dates, quantities, or monetary values. NER aims to extract and categorize these entities into predefined classes.

The process of NER typically involves analyzing a given text and locating the words or phrases that correspond to named entities. It requires understanding the context and linguistic features to accurately recognize and classify the entities. NER is an important component of various NLP applications, including information extraction, question answering, text summarization, and machine translation.

NER systems often use supervised machine learning algorithms to train models on labeled datasets. These datasets contain annotated texts where the named entities are manually labeled and assigned to specific categories. The models learn to recognize patterns and features in the text that indicate the presence of named entities. Common machine learning algorithms used for NER include conditional random fields (CRF), support vector machines (SVM), and deep learning models like recurrent neural networks (RNN) or transformer-based models like BERT.

The output of a NER system is a set of recognized named entities, typically accompanied by their corresponding entity types or labels. For example, in the sentence "Apple Inc. is planning to open a new store in New York City," the NER system would identify "Apple Inc." as an organization and "New York City" as a location.

NER is a crucial task in information extraction, as it helps in identifying and extracting relevant information from unstructured text. It aids in understanding the semantic context of the text and enables downstream applications to process and utilize the extracted entities for various purposes [35].

1.6 Lexical Resources

Words lie at the core of language, serving as the fundamental units from which all human communication systems are constructed. Whether expressed through speech, sign language, or written text, words are the building blocks that enable us to convey thoughts, ideas, and emotions. In the realm of speech and language processing, words play a central role in diverse applications. From developing sophisticated speech recognition systems that accurately transcribe spoken words to advancing machine translation and transliteration techniques, a comprehensive understanding of words is essential. Furthermore, in the fields of psycholinguistics and generative linguistic models, lexical knowledge forms the foundation for studying and modeling the intricate processes involved in human language comprehension and production.

Knowledge of language is essential for meaningful communication through language. A vocabulary consists of the words and phrases used in a particular language. A lexicon is nothing more than a dictionary that lists all of the words of a language alphabetically. Dictionaries are storehouses of such information and therefore, they have a key role to play in NLP. A dictionary specifies the correct spelling and punctuation of the words and gives their definitions and pronunciation. A glossary is a subset of a dictionary which defines words, terms or phrases in a special field of interest. Words of a language, and the phonological, morphological, syntactic and semantic information associated with them, forms a very important part of the knowledge of language. Knowing the words is an extremely important part of knowing a language.

Many NLP applications for speech processing, transliteration, transcription and romanization tasks require the phonetic dictionary that is related to speech sounds, their production, or their

transcription in written symbols. That are corresponding to pronunciation, phonetic transcription, agreeing with pronunciation and phonetic spelling.

1.7 Motivation of the research

The motivation behind Myanmar-English Named Entity Transliteration stems from the need to bridge the linguistic gap between the Myanmar and English languages, specifically in the context of named entities. Several factors contribute to the motivation for developing a transliteration system in this domain:

Cross-Language Information Retrieval: With the increasing volume of digital content available in both Myanmar and English, there is a growing need for efficient cross-language information retrieval. However, the lack of transliteration tools poses a challenge for users who are more comfortable searching in their native language. By developing a Myanmar-English Named Entity Transliteration system, the motivation is to enable users to retrieve relevant information in their preferred language while utilizing their native language keyboard or input method.

Machine Translation and Natural Language Processing: Accurate transliteration of named entities is crucial in machine translation and natural language processing tasks. Transliteration errors can lead to incorrect translations or misinterpretations of text, impacting the overall quality and accuracy of language processing systems. The motivation behind the transliteration system is to improve the performance of these systems by providing reliable and consistent transliterations of named entities, ensuring the preservation of their original meaning and context.

Multilingual Communication: In a globalized world, effective communication between Myanmar and English-speaking individuals, organizations, and communities is increasingly important. Transliteration plays a vital role in facilitating this communication by enabling the exchange of names, organizations, and other named entities accurately and coherently across languages. The motivation for the transliteration system is to enhance multilingual communication, fostering understanding and collaboration between different language communities.

Preservation of Cultural Identity: Named entities often carry cultural significance and are integral to the identity of individuals, places, and organizations. Inaccurate or inconsistent transliteration of these entities can lead to loss of cultural identity and misrepresentation. The

motivation behind the transliteration system is to preserve the cultural and linguistic nuances embedded in named entities by providing faithful transliterations. This helps maintain the authenticity and cultural relevance of the entities in the context of language processing and translation.

Enhancing User Experience: User experience is a key factor driving the motivation for Myanmar-English Named Entity Transliteration. By providing a reliable and user-friendly transliteration system, individuals who are more comfortable with one language can seamlessly communicate with others who primarily use a different language. The transliteration system aims to enhance user experience by removing language barriers, promoting inclusivity, and facilitating effective communication and understanding across language boundaries.

Overall, the motivation for Myanmar-English Named Entity Transliteration arises from the need to overcome language barriers, improve language processing tasks, enable efficient information retrieval, foster multilingual communication, preserve cultural identity, and enhance the overall user experience in the context of named entities between the Myanmar and English languages.

1.8 Objectives of the research

The primary goal of this research is to create a comprehensive Myanmar-English named entity (NE) transliteration system that can handle all cases through exhaustive rules. However, due to limited resources such as annotated corpora, gazetteers, or well-edited dictionaries, poses a challenge. As a result, a practical approach involves employing data-driven methods to address this resource constraint. The major objectives of this research are as follows:

- To develop a western and native Myanmar-English NE transliteration dictionaries when being used by the NLP community for research purposes
- To propose a Myanmar-English automatic NE Transliteration System
- To develop the neural machine translation using transformer model for forward and backward NE transliteration
- To measure the system performance for systematic evaluation on Bilingual Evaluation Understudy (BLEU) Score and word error rate (WER)

- To investigate the effect of using units (character, sub-syllable and syllable) at different granularities in the Myanmar script for forward and backward transliteration directions

- To assist Myanmar Language applications as the front end and to develop Myanmar to English Machine Translation

1.9 Contributions of the research

- **Constructing the NE terminology dictionary for Myanmar and English:** One significant contribution is the construction of a comprehensive named entity (NE) terminology dictionary for both the Myanmar and English languages. This dictionary serves as a valuable linguistic resource that encompasses a wide range of NEs, including names of people, organizations, locations, and other entities. By creating this dictionary, the transliteration system provides a solid foundation for accurate and consistent transliteration of NEs between Myanmar and English.

- **Building a named entity transliteration module on transformer-based NN model:** The transliteration system contributes to the field by developing a dedicated named entity transliteration module using the advanced transformer-based Neural Network (NN) models. By modelling this approach, the system achieves improved accuracy and robustness in transliterating NEs between Myanmar and English, catering to the specific challenges and nuances of the two languages.

- **Designing a machine learning and deep-learning based automatic NE transliteration system:** The transliteration system's contribution extends to the design and development of an automatic NE transliteration system. Leveraging machine learning and deep learning techniques, this system learns patterns and relationships between Myanmar and English NEs, enabling it to automatically generate accurate transliterations. This significantly reduces the manual effort required for transliteration tasks and enhances the efficiency of NE processing in various applications. The system contributes to the field by conducting systematic evaluations of different transliteration units, including character, sub-syllable, and syllable levels. These evaluations utilize metrics such as BLEU Score (Bilingual Evaluation Understudy) and Word Error Rate (WER) to measure the transliteration quality

systematically. By assessing the performance of different units, the system provides insights into the most effective transliteration approaches and aids in the selection of appropriate units for specific NEs and language pairs.

- **Improving the translation quality of the existing SMT and NMT system:** Another significant contribution of the transliteration system is its impact on translation quality. By accurately transliterating NEs between Myanmar and English, the system improves the overall translation quality of existing SMT (Statistical Machine Translation) and NMT (Neural Machine Translation) systems. The incorporation of high-quality NE transliteration ensures that the translated output maintains the correct representation of NEs, preserving their semantic meaning and cultural relevance. This enhancement contributes to more accurate and contextually appropriate translations in various language translation tasks.

1.10 Organization of the research

This research is comprised of seven chapters. In the first chapter, the research provides an overview of the research problems that are the central focus of the study. The literature review conducted in Chapter 2 encompasses NE transliteration for world languages and native languages. The chapter extensively covers the statistical alignment concepts and techniques employed in Giza++, an alignment tool utilized in our work. Additionally, it classifies the research on generative transliteration into three primary categories: phonetic-based methods, spelling-based methods, and hybrid approaches. The main objective of Chapter 3 is to outline the methodology employed in constructing a comprehensive terminology dictionary for both native and foreign named entities (NE). Furthermore, the chapter presents the data statistics associated with the transliteration terminology dictionary. Within Chapter 4, a comprehensive exploration is undertaken regarding the historical background of the Myanmar Language and the intricate difficulties involved in named entity (NE) transliteration tasks. The chapter further illustrates these challenges through the inclusion of examples highlighting irregular, inconsistent, and complex NE instance pairs encountered during transliteration work. Chapter 5 provides an overview of the Myanmar-English Named Entity Transliteration System using the Transformer. Within the Transformer architecture, an essential component is the Attention function, which maps queries, key vectors, and value vectors to generate outputs, contributing significantly to improved comprehension. The pivotal discussions within

Chapter 6 delve into the experimentation phase of the Myanmar-English Named Entity Transliteration System, particularly focusing on the training and testing corpus. A thorough evaluation process is outlined, incorporating established metrics like BLEU score and WER to measure the system's transliteration accuracy. Through detailed analyses and discussions, the chapter provides valuable insights into the system's performance, shedding light on its efficacy in accurately transliterating named entities between Myanmar and English languages. Chapter 7 presents the conclusion and discussions for further research, and limitations that can be pursued in accordance with the work reported in this research.

CHAPTER 2

LITERATURE REVIEW

This chapter provides an overview of the literature relevant to the transliteration of named entities. It is divided into five main sections to comprehensively cover the topic. In Section 2.1, the theory and rationale behind Named Entity (NE) Transliteration are introduced. This section addresses the common challenges associated with designing a NE Transliteration system and presents contemporary corpora related to this research. Moving on to Section 2.2, the existing research on western NE transliteration is discussed. Section 2.3 focuses on a specialized branch of previous research, which explores local NE transliteration systems. Various approaches conducted for Myanmar NLP tasks in the context of transliteration systems are briefly discussed in this section. Section 2.4 highlights the importance of evaluating the performance of NE Transliteration systems using metrics such as BLEU (Bilingual Evaluation Understudy) Scores and WER (Word Error Rate). Finally, Section 2.5 summarizes the key findings and insights presented in this chapter.

2.1 General Process of NE Transliteration

Transliteration is the practice of converting words or written text from one writing system to another, with the aim of allowing readers to recreate the accurate spelling of transliterated words. It involves the replacement of words in the source language with phonetically or orthographically similar equivalents in the target language. One common application of transliteration is in the translation of named entities across different languages. An Automatic NE Transliteration System refers to the process of converting a named entity from one language script to another, using the appropriate characters that represent the entity in the target language. Despite the availability of bilingual lexicons, new named entities that are not included in these lexicons frequently emerge, highlighting the need for automatic NE Transliteration. This process is beneficial for various applications such as Machine Translation, Cross Language Information Retrieval, and Information Extraction, among others. Despite significant advancements in general sequence-to-sequence processing techniques for NLP tasks, the scarcity of resources remains a challenge, especially for less-studied languages.

Source and Tatget words	Letter Correspondence				Description
English to Persian					
John /dʒɒn/	J	o	h	n	*h* is a silent letter (no sound
	|	|		|	is associated to the letter)
جان /dʒɒn/	ج	ا		ن	and is not transliterated
Arabic to English					
نجيب /nædʒi:b/	ن	ج	ى	ب	short vowel /æ/ on N is
	|	|	|	|	normally not written in
Najib /nædʒi:b/	Na	j	i	b	Arabic script
English to Japanese					
Bill /bi:l/	B	i	l	l	each syllable in Japanese is a
	\	/	\	/	consonant-vowel sequence
ビル [bi-ru]		ビ		ル	
English to Hindi					
Adam /ˈædəm/	A	d	a	m	the second "a" is not
	|	|		|	transliterated in Hindi
अदम /ˈædəm/	अ	द		म	

Figure 2.1 Sample NE Transliteration Instance Pairs

2.1.1 Named Entity Transliteration Methodology

Extensive research has been conducted in the domain of machine transliteration, resulting in a variety of techniques. Upon reviewing the existing literature, these techniques are categorized into three major categories, providing a comprehensive framework for understanding the advancements in this field.

- Phoneme-based Methods
- Grapheme-based Methods
- Hybrid Methods

In each of the following three sub-sections, a comprehensive overview of the work is described using these methods.

2.1.1.1 Phoneme-based Methods

This study [21] made significant contributions to the field of Statistical Machine Transliteration by tackling the task of back-transliteration from Japanese (Katakana) to English.

Their approach involved modeling the transliteration process as a generative process, consisting of five sub-modules. The first sub-module focused on the generation of scored English phrases, while the second sub-module involved the generation of English phonemes from graphemes. Additionally, the third sub-module aimed to generate Japanese phonemes from English phonemes, and the fourth sub-module focused on generating Katakana graphemes from Japanese phonemes. Finally, the fifth sub-module addressed misspelling occurrences resulting from optical character recognition (OCR) mistakes, modeling the generation of these misspelled versions. Each of these sub-modules was modeled using specific probability distributions.

- The generation of (scored) English phrases is effectively modeled using the probability distribution P(w). This distribution allows researchers to capture the likelihood of generating English phrases, incorporating various factors such as grammar, vocabulary, and linguistic patterns. By leveraging P(w), researchers can develop robust models that accurately generate English phrases, facilitating the transliteration process with improved accuracy and fluency.
- In order to develop a model for generating of English phonemes from graphemes, researchers rely on the probability distribution P(e | w). This distribution captures the conditional probability of generating English phonemes given the corresponding graphemes in a transliteration system. By utilizing P(e | w), researchers can effectively map the written representation of English words to their corresponding phonetic representations, enabling accurate pronunciation and transcription.
- To model the generation of Japanese phonemes from English phonemes, researchers employ the probability distribution P(j | e). This distribution enables the mapping of English phonemes to their corresponding Japanese phonemes, facilitating the transliteration process. By leveraging P(j | e), researchers can generate accurate representations of Japanese phonemes based on the input English phonetic information, enhancing the quality and fidelity of the transliterated output.
- The generation of Katakana graphemes from Japanese phonemes is effectively modeled using the probability distribution P(k | j). This distribution captures the conditional probability of generating Katakana graphemes given the corresponding Japanese phonemes. By leveraging P(k | j), researchers can accurately represent Japanese phonetic information using the Katakana script, which is commonly employed for transliteration purposes. The

modeling of P(k | j) contributes to producing authentic and readable transliterations of Japanese words in the Katakana writing system.

- To account for misspellings caused by optical character recognition (OCR) during the transliteration process, researchers rely on the probability distribution P(o | k). This distribution allows for modeling the likelihood of misspelled Katakana graphemes given the correct Katakana representation. By leveraging P(o | k), researchers can address OCR errors and generate more accurate transliterations by considering the potential variations and mistakes that may occur in the Katakana writing system. The modeling of P(o | k) aids in improving the overall quality and reliability of the transliterated output.

To model each of the transliteration processes, Weighted Finite State Machine (WFSM) were employed by Knight and Graehl. For the initial task, Acceptors were used, while transducers were utilized for the subsequent tasks. The modeling of P(w) involved a straightforward unigram scoring method, where the scores of known words in the phrase were multiplied. To accomplish this, a frequency list consisting of 262,000 words from the Wall Street Journal (WSJ) corpus was utilized, providing a robust coverage of English phrases. In their approach, [44] utilized the CMU Pronunciation Dictionary to create the Transducer for P(e | w). By leveraging this resource, they could establish the mapping between English graphemes and phonemes. This allowed them to accurately generate English phonemes from the corresponding graphemes, improving the precision of the transliteration process.

In [21], for modeling P(j | e), the authors employed an Expectation-Maximization(EM) algorithm to learn the transducer. This algorithm was instrumental in generating symbol-mapping probabilities, which facilitated the mapping of English phonemes to Japanese phonemes. By utilizing the EM algorithm, they achieved a comprehensive understanding of the relationship between the phonetic systems of the two languages, enhancing the accuracy of the transliteration from English to Japanese. To address the specific challenges posed by P(k | j), they manually constructed two transducers. The first transducer was designed to merge long Japanese vowel sounds into new symbols, ensuring accurate representation of these sounds. The second transducer then mapped the Japanese phonetic sounds to their corresponding Katakana characters, preserving the integrity of the transliterated words. They considered OCR as a significant source of noise that affects the fidelity of the Katakana sequences. To account for this, they learned a transducer using an EM algorithm. This transducer was specifically trained to model the noise introduced by OCR, resulting in more

accurate and reliable Katakana sequences. Consequently, given a Katakana string o, the corresponding English word could be determined following the modeling and transduction techniques employed by in [21].

$$w = argmax_{\hat{w}} \sum_{e,j,k} P(\hat{w})\, P(e|\hat{w}) P(j|e) P(k|j) P(o|k) \qquad \text{Equation (2.1)}$$

Through the use of Bayes' rule, they reversed the cascade, enabling them to derive the English phrases from the Katakana phrases. During the inference stage, Dijkstra's shortest-path algorithm was employed. To generate a variety of transliteration options, Eppstein's k-shortest-path algorithm was also applied. In a subsequent study conducted by [44], the model was expanded to include the learning of back-transliteration from Arabic to English, resulting in several modifications. As the initial two modules solely relied on the English language, they were directly employed without any alterations. Instead of approaching the task as a two-level process involving the modeling of English phonemes to Japanese phonemes and subsequently to Katakana graphemes, these steps were integrated. In other words, the direct modeling of the conversion from English pronunciation to Arabic characters was undertaken. Considering the lack of an extensive English-Arabic dictionary during that timeframe, they chose to manually establish a small dictionary comprising only 150 words. They relied on an English pronunciation dictionary to derive the mappings from English pronunciations to Arabic words. To master the mappings from English phonemes to Arabic graphemes, they utilized an EM algorithm.

In their work on English-Chinese transliteration, [34] followed a phoneme-based methodology. They converted English phrases into phonemes, taking into consideration the monosyllabic structure of the Chinese language. To address the differences between the two languages, they applied phonological rules to modify the pronunciations. Through the use of Weighted Finite State Transducers, they established phoneme alignments between English and Chinese. They further evaluated the accuracy of the generated Chinese pronunciations by comparing them with reference data, resulting in the construction of a confusion matrix.

The authors [18] developed a transliteration system for converting English to Korean. To obtain English pronunciations, they utilized the Oxford computer-usable dictionary. Employing a probabilistic tagger, they determined the most likely Korean word based on this representation. In contrast to the standard approach of using a Markov window size of 2, they expanded the window

to a size of 4, incorporating additional contextual information for training the tagger. For the generation of k-best transliterations, they employed the Viterbi algorithm.

The previous study, [58] devised a phoneme-based English-Chinese transliteration system. They employed a deterministic approach, relying on a dictionary to convert English words into their corresponding pronunciations. To convert English phonemes into syllabic units, they employed traditional Source Channel models typically used in Statistical Machine Translation (SMT). These syllabic sequences were then transformed into pin-yin symbols and subsequently translated into character sequences. In contrast to other studies that evaluate transliteration systems based on accuracy, [58] conducted an extrinsic evaluation, specifically testing the system's performance in cross-lingual spoken document retrieval.

2.1.1.2 Grapheme-based Methods

The phoneme-based approaches to transliteration effectively capture the transliterator's functionality by incorporating the phonetic representation of words during symbol translation. However, these methods also possess significant drawbacks. As highlighted by [2], a notable limitation is the generation of pronunciations. While well-known English words can have accurate pronunciations generated, words originating from foreign languages may not always yield precise pronunciations. Additionally, the involvement of multiple intermediary steps in the phoneme-based methods introduces error propagation throughout the pipeline, adversely impacting the final outcomes. As a result, there are instances where words are transliterated solely based on their spelling in the source language. In such scenarios, a spelling-based model would prove more advantageous. Consequently, this has spurred investigations into grapheme-based methods that directly translate between scripts, bypassing the need for pronunciation modeling.

To investigate English to Korean transliterations, [20] employed a methodology similar to the aforementioned approach. They utilized a modified version of Covington's alignment algorithm to determine the alignments between English and Korean symbols. Unlike the original algorithm, which relied on match and skip operations during word traversal, resulting in one-to-one alignments, the unique nature of Korean-English transliteration necessitated many-to-many correspondences. Consequently, they introduced a bind operation to account for this requirement. Following alignment learning, they constructed a training set to predict the English symbol(s) based on the neighboring characters within a window of size 6 (3 preceding and 3 succeeding). Employing 26 decision trees,

one for each English alphabet, they could then infer the corresponding Korean symbol by traversing the English word and utilizing the relevant decision tree. The final transliterated string was obtained through the concatenation of these symbols. To prevent overfitting, they implemented post-pruning using the reduced error pruning technique. This methodology enables transliteration generation in either direction.

The study [1] also employed a grapheme-based approach to learn transliterations from English to Arabic. Initially, they utilize GIZA++ to establish the alignments between English and Arabic words. In situations where an Arabic symbol aligns with multiple English symbols, the English symbol sequence is incorporated into the English alphabet, resulting in the resegmentation of English words. Subsequently, the alignment model is retrained using the modified dataset, which includes the resegmented English words. Conditional probabilities are then computed based on the alignment counts. To infer the Arabic words, each English word undergoes the resegmentation process as described earlier, and the alignment model is utilized to determine all possible transliterations. These transliterations are scored by taking into account the product of alignment probabilities and an Arabic conditional character-level bi-gram model.

The authors [39] explored the development of an English to Hindi transliteration system using a methodology akin to the aforementioned approach. Nevertheless, they deviate from the conventional count-based approach for probability calculation and instead utilize a Conditional Random Field. For each aligned symbol pair, they generate features that encompass neighboring English alphabets within a window of 5 characters.

The paper [27] employed a grapheme-based approach to learn English-Chinese transliteration. Rather than focusing solely on one direction, they developed a joint model that enabled the simultaneous generation of source and target words. They utilized an Expectation-Maximization algorithm to learn alignments between English and Chinese words. Each aligned symbol pairing within word pairs (E,C) was treated as a transliteration unit. By constructing an n-gram transliteration model, they determined the conditional probability of a transliteration unit given the n immediate predecessor pairs. The probability of a word pair (E,C) was calculated by multiplying the probabilities of its transliteration units as estimated by this model. By marginalizing these joint probability distributions, they obtained the conditional distribution for transliteration in both directions.

The Machine Transliteration task was introduced by the Named Entities Workshop (NEWS) in 2009 [28], providing standardized datasets for various language pairs like English-Hindi, English-Hebrew, Chinese-English, Arabic-English, and more. A significant number of participating teams opted for neural networks as their preferred method to learn the target transliterations. Notably, neural networks have gained prominence in addressing the machine transliteration task, consistently outperforming phrase-based machine translation systems.

In the 2016 edition of the task, NICT's submission [29] achieved outstanding results, ranking as the best-performing team. They adopted an approach that involved training an LSTM-based RNN to encode the input sequence, generating a hidden representation that was subsequently decoded to generate the output sequence. During the decoding process, it was observed that errors tended to accumulate, resulting in a decline in the transliteration quality of the suffixes. To address this challenge, they employed target-bidirectional models that generated the target in both the left-to-right and right-to-left directions, yielding two k-best lists. An agreement model was then trained to combine these lists, and ensembles of neural networks were employed to generate transliterations by linearly interpolating probability distributions across the target vocabulary during the beam-search decoding process.

2.1.1.3 Hybrid Methods

The study [2] expanded their existing phoneme-based models by introducing a hybrid approach for Arabic-English transliterations. They developed a grapheme-based model and integrated it with the phoneme-based model to enhance the learning process. To determine the probability of an English word being the transliteration of an Arabic word, they utilized discriminative models based on both graphemes and phonemes. The scores obtained from these models were linearly interpolated to derive the final transliteration score. During the inference stage, they searched for the English word that yielded the highest score. To improve the quality of their results, they applied additional postprocessing techniques. They created a Finite State Machine specifically designed to rectify misspellings, manually assigning weights to its components due to the limited availability of misspelling training data for parameter optimization.

2.2 Named Entity Transliteration between English and Western Languages

Transliteration, as a linguistic task, revolves around converting words from a source script to a target script by considering their approximate phonetic values. The literature encompasses a plethora of transliteration processes, each employing distinct methodologies and addressing the needs of specific languages. Categorizing these studies is not a simple endeavor, given the diverse attributes they exhibit, such as transliteration direction, script variations among languages, and the different sources of information employed. Transliterating between languages with significantly distinct scripts presents inherent challenges [36].

2.2.1 English-Chinese NE Transliteration

The investigation carried out by [17] focused on English-Chinese transliteration. They evaluated their experiments using word accuracy and character accuracy metrics on a dataset comprising 46,306 English-Chinese word-pairs extracted from the LDC named entity list. The results indicated that the direct model outperformed the source-channel model in the transliteration process. Generally, phonetic-based transliteration offers a fundamental benefit of emphasizing the significance of pronunciation in the transliteration procedure. Nonetheless, the inclusion of multiple stages in the process, encompassing transformations from letter-to-sound, sound-to-letter, and occasionally sound-to-sound, amplifies the likelihood of error propagation.

The authors [30] presented a transliteration technique targeting personal names, which they referred to as semantic transliteration. By "semantic," they denoted the language of origin, gender, and given or surname attributes of the source names. Consequently, their transliteration model was structured according to the provided formula.

$$P(T|S) = \sum P(T|S, l, g)P(l, g|S) \qquad \text{Equation (2.2)}$$

In a recent study, the researchers focused on the language of origin and gender detection using sequences of four characters. They utilized three corpora consisting of Japanese, Chinese, and English names. The corpora contained 30,000 name pairs for Japanese-Chinese, 34,600 for Chinese-Chinese, and 20,600 for English-Chinese. To ensure the accuracy of their system, they removed any missing information sources from their model. To evaluate the performance of their system, the researchers employed mean reciprocal rank, word accuracy, and character accuracy. After analyzing the results, they found that the best overall accuracies achieved were 49.4% for word accuracy and

69.2% for character accuracy. These accuracies represented improvements compared to their baseline phonetic-based system. However, it should be noted that the achieved accuracies were not as high as those reported in similar studies that did not incorporate semantic information for English-Chinese transliteration. Despite this limitation, the researchers were able to demonstrate the value of considering both the language of origin and gender information in their model, showcasing the potential for further advancements in this area of research.

2.2.2 English-Korean NE Transliteration

In this paper [55], the authors focus was on the English-Korean named entity transliteration task for the NEWS 2012 dataset. Their approach begins by decomposing Korean words into individual Korean letters and then converting them into sequential Roman letters through a romanization process. Considering that a Korean word may not have a final consonant, they also create alignment results that include the null consonant in the Romanized Korean representations. Preprocessing the training data allows us to obtain alignments from English to Korean using the m2m-aligner tool. Using the alignments obtained, the authors proceed to train several transliteration models based on DirecTL-p. These models are trained using the alignments generated by the m2m-aligner. Additionally, they introduce two re-ranking methods to further refine our results. The first method is a web-based re-ranking approach that utilizes the Google search engine. They submit the English named entity and its corresponding Korean transliteration pair, generated by their model, to Google in order to obtain co-occurrence counts, which are then used to re-rank the transliteration results. The second re-ranking method is JLIS-re-ranking, which relies on three features extracted from the alignment results: the source grapheme chain feature, the target grapheme chain feature, and the syllable consistent feature. In their experimental results, they proposed method achieves promising accuracy, with a score of 0.398 in the standard run and 0.458 in the non-standard run. These results indicate that incorporating a web-based re-ranking method into the transliteration model can significantly improve the accuracy of English-Korean transliteration.

In this paper [56], the authors proposed a substring-based transliteration approach using a Conditional Random Field (CRF) model for English-Korean named entity transliteration. Their method involves aligning characters in both the source and target languages bidirectionally and grouping them into substrings to establish mappings from the source language to the target language. They formulate the transliteration task as a sequential tagging problem, where they assign tags to

the substrings in the source language corresponding to the substrings in the target language. To address this tagging problem, they employ the CRF algorithm.

For generating English substrings, they consider two types: one based on English orthography and the other based on phonemic symbols from the CMU pronouncing dictionary. Additionally, they incorporate a rule-based transliteration system based on the Korean writing method of loanwords provided by the National Institute of Korean language. Based on the evaluation results, the substring-based method utilizing English orthography outperforms other runs in terms of transliteration accuracy.

2.2.3 English-Japanese NE Transliteration

The paper [22] conducted a comprehensive study on the back-transliteration of Japanese out-of-dictionary words into English. Their approach was phoneme-based and involved four main steps. In their proposed system, they outlined the sequential steps as O->S->Is->I_T->T, where the Japanese source word O was initially transformed into an electronic representation S using optical character recognition (OCR). Next, S was converted into its phonetical presentation I_S, followed by mapping the source phonemes to target English phonemes I_T. Finally, a phoneme-to-grapheme mapping generated the target English word T. Therefore, their model was formulated as:

$$T = argmax_T \, P(T) \sum_{S,I_s,I_t} P(O|S).P(S|I_s).P(I_s|I_T).P(I_T|T) \qquad \text{Equation (2.3)}$$

Where these probabilities include P(O|S), which captures the likelihood of misspellings caused by OCR when dealing with input from katakana. Additionally, P(S|I_S) quantifies the probability of correctly pronouncing the source word, while P(I_S|I_T) handles the conversion of source sounds to their corresponding target sounds. Furthermore, P(I_T|T) represents the probability of generating the written form T based on the pronunciation in I_T, and P(T) reflects the probability of encountering a specific sequence T in the target language.

They utilized weighted finite-state transducers (WFST) and weighted finite-state acceptors (WFSA). A finite state machine (FSM) was used as a behavioral model, consisting of states, transitions, and actions. A weighted finite-state transducer, a type of FSM, incorporated parameters such as input, output, and weight for each transition. On the other hand, a weighted finite-state acceptor had a single input symbol and a weight for each transition, specifying the more probable output sequences. The authors matched this model with the transformation rules of the transliteration model, considering each transition as a transformation rule with source and target mappings and a

probability mapping to a weight. To implement the model, the authors used WFSA to represent P(T) and WFST for the remaining probabilities mentioned in Equation 2.3. They employed Dijkstra's shortest path and k-shortest paths algorithms to generate the best transliterations using WFSA. The target language model used in P(T) was a unigram model derived from the Wall Street Journal corpus, an online English name list, and an online gazetteer of place names. The English sounds inventory was obtained from the CMU pronunciation dictionary. $P(I_S|I_T)$ was calculated based on frequency information obtained from the alignment of 8,000 pairs of English and Japanese sound sequences using the estimation-maximization (EM) algorithm. In comparison to the base system, the authors' system incorporated both WFSA and WFST components. These components were automatically and manually built during the training stage and then transferred as a transliteration model to the transliteration stage.

They conducted evaluations of their automatic back-transliterator through two sets of experiments. The first experiment involved 222 katakana phrases, but no evaluation results were reported due to the difficulty in judging the task. Some of the input phrases were onomatopoetic, making them challenging to transliterate even for humans. The second experiment focused on 100 names of U.S. politicians written in katakana. In this experiment, the authors compared the performance of their system with four human transliterators who were English native speakers tasked with the same objective. The results revealed that the machine outperformed the human transliterators significantly, achieving a word accuracy of 64% compared to 24% by the humans. The reason behind the low accuracy of the human transliterators was attributed to their lack of knowledge about Japanese phonetics.

The research conducted by [7], centered around the hybridization of spelling and phonetic approaches for English-Korean and English-Japanese transliteration. During their investigation, they raised important concerns about existing hybrid models. One such concern was the omission of the interdependence between source word graphemes and phonemes, a relationship that the authors accounted for in their correspondence-based method. Another drawback of previous hybrid models was the fixed weight assignment to either the spelling or phonetics approaches, overlooking the fact that the transliteration requirements varied depending on the source word. By integrating spelling and phonetics and considering correspondence information, the researchers developed a comprehensive model to address the transliteration problem. They employed three machine learning algorithms, namely maximum entropy model, decision-tree learning, and memory-based learning,

to merge these methods into a unified framework. The transformation rules were derived from phonetics, spelling, correspondence, and a hybrid of phonetics and spelling approaches. Their evaluation results indicated improved word accuracy compared to standalone models. For English-Korean transliteration, they obtained a word accuracy of 68.4% using a dataset consisting of 7,172 word pairs, with a subset of 1,000 pairs designated for testing.

2.2.4 English-Arabic NE Transliteration

The work of [43,44] focused on proposing a method for back-transliterating Arabic out-of-dictionary words into English. In comparison to the challenges encountered in Japanese, the Arabic language presented significantly greater obstacles. These challenges stemmed from the absence of a comprehensive pronunciation dictionary that covered out-of-dictionary words from diverse origins, extending beyond English. Furthermore, the absence of short vowel notations in Arabic script and the scarcity of resources for Arabic pronunciation further complicated the task. The researchers evaluated their transliteration system using a test corpus comprising 2,800 names, which yielded a top-1 accuracy of 32.1%.

In a study conducted by [1], the focus was on English-Arabic transliteration using n-gram models. The researchers adopted a bigram model as the target language model for their transliteration approach. Training the system involved aligning word pairs from a bilingual transliteration corpus using Giza++. Transformation rules were then generated, and probabilities were assigned based on their frequencies in the corpus. To evaluate the performance of their system, they compared it to a hand-crafted model that served as a baseline. The evaluation was conducted on a corpus comprising 815-word pairs extracted from the AFP Arabic corpus. The results indicated that their system achieved a top-1 word accuracy of 69.3%, while the baseline hand-crafted system achieved an accuracy of 71.2%. Additionally, the researchers investigated the impact of transliteration in a cross-lingual information retrieval task.

The study [45] delved into the intricate process of Arabic-English transliteration, employing dynamic programming alongside substring-based transducer methodologies. What sets this research apart is its keen focus on addressing the complexities inherent in many-to-many mappings between source and target words, often overlooked in prior studies. To tackle this challenge, the researchers integrated phrase-based approaches from machine translation (MT) into their analysis. They explored two distinct methods: a monotone search utilizing a Viterbi substring decoder, and a

substring transducer. Notably, the substring transducer was highlighted for its ability to incorporate a word unigram language model, effectively filtering out low probability mappings and handling NULLs implicitly. This nuanced approach aimed to mitigate the confusion that NULLs might introduce into the transliteration process, thus enhancing overall accuracy. The evaluation of their system was meticulous, drawing from a training corpus comprising 2,844-word pairs and a test set of 300 word pairs. Additionally, a language model was independently trained on a substantial dataset consisting of 10,991 word pairs, of which 4,494 were unique. However, the reported results focused solely on seen data, indicating an overlap between the training and testing sets. While this evaluation paradigm aligns with previous studies, its adequacy for assessing the efficacy of a generative transliteration system is questioned. The primary goal of such a system lies in its ability to accurately transliterate unseen and newly emerging names, a task for which the current evaluation methodology may prove insufficient. This critique underscores the need for evaluation frameworks that capture the system's performance in real-world scenarios more effectively.

In essence, the research underscores the intricate balance between methodology and evaluation in the realm of Arabic-English transliteration. By integrating advanced techniques from machine translation and addressing the nuances of many-to-many mappings and NULL handling, the study contributes significantly to the field's evolving landscape. However, the discussion around evaluation methodologies remains pivotal, highlighting the ongoing quest for more robust assessment frameworks that align with the practical demands of generative transliteration systems.

The paper [3] introduced the ANETAC dataset, which is an English-Arabic named entity transliteration and classification dataset. The dataset, consisting of 79,924 instances, includes triplets of English named entities, their Arabic transliterations, and their corresponding classes (Person, Location, or Organization). It was created from freely available parallel translation corpora and aims to support researchers working on Arabic named entity transliteration and classification tasks. The dataset was developed as part of a previous research study and is made freely accessible in this work. The authors discuss the process of building the dataset and provide baseline results for English-to-Arabic and Arabic-to-English machine transliteration tasks. They encouraged researchers to use the dataset and strive for improved results, hoping that it will positively impact the field of Arabic-English named entity transliteration.

2.2.5 English- Bengali NE Transliteration

In their 2006 research, [15] examined a modified joint source-channel system proposed for Bengali-English transliteration. The selection of transliteration units within the source word was carried out using a regular expression based on consonants, vowels, and matra, which is a writing delimiter specific to the Bengali language. In order to handle the complexities arising from one-to-many alignments between English and Bengali, the researchers integrated hand-crafted transformation rules into their system. In cases where alignment proved challenging despite the inclusion of these rules, manual intervention during the training phase was utilized to correct errors. Upon the completion of training, the transliteration model was ready to undertake the transliteration phase. To evaluate the performance of their system, the researchers utilized a corpus consisting of 6,000 people's names. Out of this corpus, 1,200 names were set aside for testing, while the remaining 4,755 names were used for training the system. The results were highly encouraging, with their best model achieving an impressive top-1 word accuracy of 69.3% for Bengali-English transliteration and 67.9% for the back-transliteration task.

The study [32] proposed a system that aimed to convert words between two scripts used in Punjabi: Shahmukhi, which is based on the Arabic script, and Gurmukhi, which is derived from Landa, Shardha, and Takri. The transliteration system employed a set of hand-crafted transliteration rules, categorized into character mappings and dependency rules. The dependency rules served as contextual rules to handle special cases where simple character mappings failed. The system's performance was evaluated using a corpus of 45,420 words extracted from classical and modern literature, achieving an impressive average transliteration accuracy of 98.95%.

2.2.6 English-French NE Transliteration

The research conducted by [14] explored the process of generating phoneme-to-grapheme transformation rules (also referred to as letter-to-sound rules) and its challenges and applications in English and French. They highlighted the fact that the pronunciation of words in any language is influenced by various parameters, such as word position (morphophonemics) and the presence of elision or epenthesis, which can cause discrepancies between the pronunciation and the written form. Moreover, due to the diverse linguistic origins of proper names, establishing an accurate correspondence between their written representation and pronunciation poses significant difficulties, often resulting in substantial differences from the spelling. The studies have attempted to address

this issue by classifying proper names into language groups or language families, thereby improving the accuracy of systems that generate grapheme-to-phoneme rules from proper names.

2.3 Named Entity Transliteration between English and Burmese (Myanmar) Languages

The research [46] from Myanmar introduced an approach for identifying Named Entities in Myanmar using a hybrid technique. The hybrid method combined both rule-based and statistical N-gram approaches to achieve accurate results. In their methodology, they applied the statistical N-gram approach to a dataset consisting of over 10,000 person names and 350 location names. The frequencies of unigram and bi-gram syllables were pre-calculated, taking into account their positions such as position1, position2, position1-2, and more. The experiments involved analyzing a sample of 10 Myanmar text files, and the author achieved an impressive 89% accuracy in Named Entity Identification. Additionally, the named entities were categorized into three distinct classes. Following the identification process, the names were transliterated into their corresponding Myanmar phonetics using a transliteration table. This table facilitated the mapping of Myanmar syllables to their English pronunciations.

The research [8] focused on the Romanization of Burmese names, which is a crucial task in translating Burmese into languages that utilize the Latin script. Given the limited research and resources available for Burmese, they meticulously collected and manually annotated 2,335 instances of Romanization to facilitate the use of statistical methods. The annotation process involved segmenting the strings and aligning them with the corresponding Latin script. Unlike previous studies that treated syllables as indivisible units when processing Burmese, this study takes a different approach by segmenting Burmese strings into carefully designed sub-syllabic units. This segmentation enables precise and consistent alignment with the Latin script. Experimental results demonstrate that sub-syllabic units outperform syllables as more suitable units for statistical approaches in the Romanization of Burmese names

The research [31] described post-editing machine translation from Katakana(Japanese) to Burmese using rule-based transliteration. This paper presented rule-based post-editing scheme to solve translation errors for out-of vocabulary (OOV) of Katakana words that are released by Japanese to Burmese translation using PBSMT. In this experiment, 155,069 sentences (BTEC corpus) were used for training set and 1614 sentences were used for test set. The results showed that

rule-based Katakana to Burmese translation was better than the baseline (PBSMT) result in 19.39 BLEU and lower OOV errors about 9.33 percentage. By summarizing the NE transliteration between English and Burmese(Myanmar) literatures, it is observed that there is a little amount of NE tasks and lack of parallel resources.

2.4 Evaluation Metrics

The evaluation metric for Named Entity Transliteration typically involves measuring the accuracy and quality of the transliterated output. BLEU (Bilingual Evaluation Understudy) is a commonly used metric borrowed from machine translation, which measures the n-gram overlap between the transliterated output and the reference translation. It quantifies the similarity in terms of matching word sequences. One commonly used metric is the Word Error Rate (WER), which calculates the percentage of transliteration errors in the output compared to a reference. Both BLEU and WER [61] are valuable tools to assess the performance of Named Entity Transliteration systems and compare their effectiveness.

2.4.1 Bilingual Evaluation Understudy (BLEU)

Bilingual Evaluation Understudy (BLEU) is a metric used to evaluate the quality of machine-generated translations by comparing them to one or more reference translations. It uses precision and brevity penalty calculations to determine the similarity between the candidate translation and the reference translations. Here are the equations used in BLEU:

Precision calculation:

BLEU calculates precision by comparing the candidate translation (C) and the reference translation(s) (R). It counts the number of n-grams (contiguous subsequences of words) that appear in both the candidate and reference translations.

$$Precision = \frac{(Count\ of\ n{-}gram\ matches)}{(Count\ of\ candidate\ n{-}grams)} \quad \text{Equation (2.4)}$$

Brevity penalty calculation:

BLEU applies a brevity penalty to account for differences in length between the candidate and reference translations.

$$Brevity\ Penalty = \exp \frac{(1-(Reference\ Length)}{Candidate\ length} \quad \text{Equation (2.5)}$$

Combining precision and brevity penalty:

BLEU combines the precision and brevity penalty by calculating the geometric mean of the n-gram precisions.

$$BLEU\ Score = Brevity\ Penalty * (Precision\ 1 * Precision\ 2 * \ldots * PrecisionN)\text{^ } (\frac{1}{N}) \quad \text{Equation (2.6)}$$

In the equations above:

"Count of n-gram matches" represents the number of n-grams that are present in both the candidate and reference translations.
"Count of candidate n-grams" is the total number of n-grams in the candidate translation.
"Reference length" is the length (in terms of n-grams) of the reference translation(s).
"Candidate length" is the length (in terms of n-grams) of the candidate translation.
"N" is the maximum n-gram order.

2.4.1.1 Example Calculation of BLEU Score

To calculate the BLEU score for the named entity transliteration from Myanmar to English using the input named entity "အေး မြတ် မွန်" and the output named entity "Aye Myat Mon" described through the step-by-step calculations.

Step 1: Prepare the reference translation and candidate translation.
Reference translation: The correct English transliteration of the given Myanmar named entity.
Candidate translation: The predicted English transliteration generated by the transliteration system.
For example:
Reference translation: "အေး မြတ် မွန်"

Candidate translation: "Aye Myat Mon"

Step 2: Tokenize the reference and candidate translations.

Tokenization breaks the translations into individual units, typically words or characters, to facilitate comparison.

For example:

Tokenized reference translation: ["အေး", "မြတ်", "မွန်"]

Tokenized candidate translation: ["Aye", "Myat", "Mon"]

Step 3: Calculate n-gram precision.

Calculate the precision of n-grams (subsequences of length n) by comparing the candidate and reference translations.

For example:

n = 1 (unigrams)

Candidate unigrams: ["Aye", "Myat", "Mon"]

Reference unigrams: ["အေး", "မြတ်", "မွန်"]

Precision = 3/3 = 1.0

Step 4: Calculate the brevity penalty.

Determine the brevity penalty to account for differences in translation length between the candidate and reference translations.

For example:

Candidate length: 3 (number of unigrams in the candidate translation)

Closest reference length: 3 (number of unigrams in the reference translation closest in length to the candidate)

Brevity penalty = exp(1 - 1) = exp(0) = 1.0

Step 5: Calculate the BLEU score.

Combine the n-gram precision and the brevity penalty to calculate the final BLEU score.

For example:

BLEU score = brevity penalty * (n-gram precision)^(1/n)

BLEU score = 1.0 * (1.0)^(1/1) = 1.0

In this example, the calculated BLEU score for the transliteration of the given Myanmar named entity "အေး မြတ် မွန်" to English "Aye Myat Mon" is 1.0.

2.4.2 Word Error Rate (WER)

Word Error Rate (WER) is a common evaluation metric used in the field of automatic speech recognition (ASR) and natural language processing (NLP) to measure the accuracy of a system's output compared to a reference or ground truth.WER calculates the percentage of errors, specifically word-level substitutions, insertions, and deletions, made by an ASR or NLP system when transcribing or generating text. It provides a quantitative measure of the dissimilarity between the system's output and the reference text.

The formula to calculate WER involves comparing the number of errors (substitutions, insertions, and deletions) to the total number of words in the reference text:

$$WER = \frac{(S+I+D)}{N} \qquad \text{Equation (2.7)}$$

where:

S: Number of word substitutions

I: Number of word insertions

D: Number of word deletions

N: Total number of words in the reference text

The WER is typically expressed as a percentage, where a lower percentage indicates higher accuracy. For example, a WER of 10% means that, on average, 10 out of every 100 words in the system's output are incorrect compared to the reference.

WER is widely used to assess and compare the performance of ASR and NLP systems, allowing researchers and developers to measure the quality and effectiveness of different algorithms, models, or system configurations. It provides a standardized and objective measure for evaluating the accuracy and reliability of automatic transcription or text generation systems.

To calculate the Word Error Rate (WER) for named entity transliteration in Myanmar language, you would need a reference list of correct transliterations and a hypothesis list of the machine-generated transliterations. An example calculation:

Reference list (correct transliteration):

ရန်ကုန်: Yangon

Hypothesis list (machine-generated transliteration):

ရန်ကုန်: Yangon

Calculate the number of substitutions, insertions, and deletions:

Substitutions: no substitutions.

Insertions: no insertions.

Deletions: no deletions.

Calculate the total number of words in the reference list:

In this case, the total number of words in the reference list is 1.

Calculate the Word Error Rate (WER):

$$WER = \frac{(Substitutions+Insertions+Deletions)}{Total\ words\ in\ the\ reference\ list} \qquad \text{Equation (2.8)}$$

WER = (0 + 0 + 0) / 1 = 0 / 1 = 0

The Word Error Rate for this named entity transliteration task is 0, indicating that the machine-generated transliteration matches the reference transliteration perfectly.

2.5 Summary

The literature review in this chapter focuses on the previous research and transliteration of named entities. It presents five main sections covering theory, challenges, and contemporary corpora. The research conducted so far is discussed, revealing the consistent issue of poor results in low resource settings. Additionally, a specialized branch of research for the Western and Local NE transliteration systems are explored, along with the importance of employing evaluation metrics like BLEU Scores and WER.

CHAPTER 3

MYANMAR-ENGLISH NAMED ENTITY TRANSLITERATION TERMINOLOGY DICTIONARY

The chapter delves into the intricacies of creating a large Named Entity terminology dictionary, offering a detailed description of the methods employed and the steps taken throughout the construction process.

3.1 Corpus

In the field of linguistics, a corpus (plural corpora) plays a crucial role as a large and organized collection of texts, typically stored and processed electronically. These corpora serve as valuable resources for conducting statistical analyses, hypothesis testing, and validating linguistic rules within a specific domain. Researchers rely on corpora to examine occurrences, study language patterns, and gain insights into various linguistic phenomena. Whether it is a monolingual corpus containing texts in a single language or a multilingual corpus encompassing text data from multiple languages, corpora provide a fundamental foundation for numerous NLP research endeavors.

3.2 Lexicon

In the field of Natural Language Processing (NLP), a lexicon refers to a collection of words or terms along with their associated information, such as part-of-speech tags, semantic labels, pronunciation, and other relevant linguistic features. It can be thought of as a vocabulary or dictionary specific to a particular language or domain.

A lexicon serves as a foundational resource for various NLP tasks, including text processing, information retrieval, sentiment analysis, machine translation, and named entity recognition, among others. Lexicons can be created manually by linguists and lexicographers or generated automatically using computational methods. Lexicons often contain additional metadata about words, such as their frequency of occurrence, collocations (words that tend to appear together), synonyms, antonyms, and other semantic relations. Some lexicons also incorporate sentiment or emotional information

associated with words, helping sentiment analysis systems determine the sentiment polarity of a given text.

Lexicons can vary in size and scope depending on their intended application. For example, a general-purpose lexicon covers a wide range of words from a language, while a domain-specific lexicon focuses on vocabulary related to a specific field like medicine, finance, or sports. Overall, lexicons are fundamental resources in NLP that provide linguistic information about words, enabling machines to understand and process natural language more effectively.

3.3 Dictionary

A dictionary refers to a mapping or lookup table that associates each unique word in a corpus or dataset with a unique index or identifier. The purpose of this mapping is to provide a convenient and efficient way to represent and manipulate textual data during various NLP tasks. A dictionary would be used to map each unique word in the training data to a specific index or identifier. This allows for the representation of text data in a numerical format that can be easily processed by machine learning algorithms or other NLP models.

The NE Transliteration terminology dictionary is a specific type of dictionary. It is designed to handle named entity (NE) transliterations, which involve converting names from one script or language to another based on their pronunciation rather than their meaning. In this case, the dictionary contains a list of candidate names transliterations along with their corresponding pronunciations or phonetic representations.

The construction of a Native and Western Burmese (Myanmar)-English NE terminology dictionary mentioned in the facts is described as a valuable contribution to the development of a NE transliteration system. This dictionary would contain entries that associate Burmese (Myanmar) names with their English transliterations or vice versa, providing a resource for accurately converting names between the two languages.

Overall, a dictionary in NLP serves as a mapping between words and their representations, facilitating various language processing tasks such as information retrieval, machine translation, and transliteration. It enables efficient manipulation and analysis of textual data in a computational setting.

3.4 Importance of Dictionary

The dictionary plays a crucial role in NE transliteration, specifically in the context of converting names from one script or language to another based on their pronunciation. Here are the key reasons why a dictionary is important in NE transliteration:

Mapping Pronunciations: A dictionary provides a mapping between names in one language or script and their corresponding transliterations in another language or script based on their pronunciation. It captures the phonetic information necessary for accurately converting names between different writing systems.

Consistency and Accuracy: By using a dictionary, NE transliteration systems can ensure consistency and accuracy in converting names. The dictionary serves as a reference for selecting the most appropriate transliteration for a given name, reducing ambiguity and potential errors.

Candidate Name Transliterations: The dictionary contains a list of candidate transliterations for each name, considering variations in pronunciation or spelling. It allows the system to consider multiple possibilities and choose the most suitable transliteration based on specific transliteration rules or linguistic patterns.

Development and Evaluation: The dictionary serves as a valuable resource for developing and evaluating NE transliteration systems. It provides a foundation for training transliteration models, testing their performance, and assessing the quality of transliterations generated by the system.

Linguistic and Cultural Considerations: The dictionary incorporates linguistic and cultural knowledge related to names and their transliterations. It captures specific conventions, phonetic rules, and language-specific patterns, enabling the system to produce accurate and culturally appropriate transliterations.

The dictionary plays an important role in NE transliteration by providing the necessary mapping, candidate transliterations, and linguistic knowledge to accurately convert names between different scripts or languages. It ensures consistency, accuracy, and cultural sensitivity, contributing to the overall effectiveness of NE transliteration systems.

3.5 Myanmar-English Named Entity Terminology Dictionary Construction.

To collect the transliteration instances, the researcher began with two Myanmar-English parallel NE instance pair, and then moved to resources on the internet to enlarge the scale of the data. Specifically, we used the ALT corpus [9,10,11,40], UCSY corpus [47,48] and Wikipedia data for western Myanmar-English language pair. The ALT corpus consists of twenty thousand parallel sentences from news articles and the UCSY corpus contains two hundred thousand parallel sentences collected from different domains, including local news articles and textbooks.

Word alignment is the natural language processing task of identifying translational relationships among the words and multi-word units in parallel corpora. Automatic word alignment in bilingual or multilingual parallel corpora has been a challenging issue for natural language processing. Giza++ is the automatic word alignment tool. It is very common and convenient to use the word alignment generated from GIZA++ for most statistical machine translation systems. We used the GIZA++ toolkit [9,37,38] to obtain the raw alignments between the source and target language, based on which the transliteration instances were filtered from ALT (Asian Language Treebank) and UCSY parallel corpus. Moreover, we further collected NE instances of places, organizations, and person names from the internet sources. The dictionary was encoded in Unicode format.

Moreover, there is no freely available for Myanmar native name transliterations. Thus, we developed a parallel data for transcription for Myanmar native names on the Myanmar matriculation exam results and Wikipedia data. We consider all possible transliterations are especially focus on a real-world transliteration of Myanmar people (i.e., local transliteration) [62]. To the best of our knowledge, this is the first large transliteration dictionary for Myanmar language. It can be applied not only in Myanmar NET system but also in other Myanmar NLP research areas.

For English transliteration systems, the freely available resources are accessed on the web. However, any large pronunciation dictionary for Myanmar language to transliterate named entities is not found in the web. Therefore, the first large amount of pronunciation dictionary for Myanmar language has been built for applying in Myanmar NET system. The detailed process of building a large Myanmar NE terminology dictionary will be described in the above paragraphs and data sample described in the following Table 3.1 and Table 3.2.

Table 3.1 Sample Myanmar-English NE Transliteration Instance Pairs for Western Script

Type of NE	**Named Entity in Myanmar**	**Named Entity in English**
Person Name	ဘာရက် ဟူစိန် အိုဘားမား	Barack Hussein Obama
Person Name	မစ်ရှဲလ် အိုဘားမား	Michelle Obama
Person Name	ဆာရှာ အိုဘားမား	Sasha Obama
Person Name	မာလီယာ အိုဘားမား	Malia Obama
Person Name	ဒေါ်နယ်လ် ထရန့်	Donald Trump
Place Name	ယူနိုက်တက် စတိတ်	United States
Place Name	ဆန် ဖရန်စစ္စကို	San Francisco
Place Name	ကယ်လီဖိုးနီးယား	California
Place Name	အမေရိကန်	American
Place Name	နယူးယောက်	New York
Organization Name	ဝါရှင်တန် ဒီစီ	Washington DC
Organization Name	အက်ပဲ	Apple
Organization Name	ဂူဂယ်	Google
Organization Name	ယူကျူ(ဘ်)	YouTube
Organization Name	ကုမ္ပဏီ	Company

Table 3.2 Sample Myanmar-English NE Transliteration Instance Pairs for Native Script

Type of NE	Named Entity in Myanmar	Named Entity in English
Person Name	အေးမြတ်မွန်	Aye Myat Mon
Person Name	ဆုမြတ်မို	Su Myat Mo
Person Name	ဆုမြတ်မွန်	Hsu Myat Mon
Person Name	ဝင်းလဲ့လဲ့ဖြူ	Win Lei Lei Phyu
Person Name	ဆန်းဆုဆုရည်	Sann Su Su Yee
Place Name	မြန်မာ	Myanmar
Place Name	ရန်ကုန်	Yangon
Place Name	ပုဂံ	Bagan
Place Name	ပုသိမ်	Pathein
Place Name	ထားဝယ်	Dawei
Organization Name	ဘီဘီစီ	BBC
Organization Name	မဇ္ဈိမ	Mizzima
Organization Name	စကိုင်း နက်	Sky Net
Organization Name	အမ်အာတီဗီ	MRTV
Organization Name	ဖော်အဲဗား	Forever

3.5.1 Data Statistics

Detailed data statistics for the transliteration dictionary are presented in Table 3.3 and Table 3.4. These tables provide comprehensive insights into the characteristics and composition of the dictionary. They likely include information such as the number of entries or words in the dictionary, the distribution of transliteration variants, statistical measures such as frequency or occurrence rates

of different entities, and any other relevant metrics that highlight the overall data quality and diversity.

Table 3.3 Data statistics for Western names

Data Set	**Number of Entities**
Training	127464
Development	2000
Test	2000

Table 3.4 Data statistics for Native names

Data Set	**Number of Entities**
Training	155105
Development	2000
Test	2000

3.6 Summary

The availability of the large Myanmar-English NE Transliteration dictionary marks a significant breakthrough in Myanmar transliteration research. By capturing the complexities of Myanmar names and their conversion to English, this dictionary opens new avenues for accurate and effective transliteration systems. Through the utilization of NN based transformer models, the quality of the data within the dictionary was evaluated and validated. This pioneering effort in Myanmar transliteration research not only provides valuable insights for future advancements but also showcases the importance of well-constructed transliteration dictionaries in facilitating successful cross-language conversions.

CHAPTER 4

BURMESE (MYANMAR) CHARACTER WRITING SYSTEM AND TRANSLITERATION ISSUES

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 of 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 [63]. 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. 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 4.1 and categories of characters in Table 4.1. In the Burmese Language, there are specific tasks that require the dentification of char., sub-syl.[12] and syl. units [36,54]. 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.

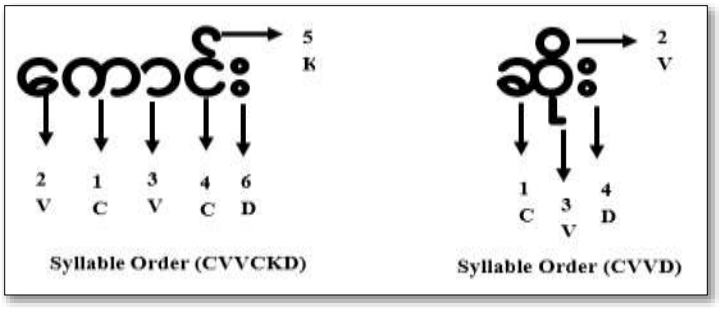

Figure 4.1 The structure of Burmese (Myanmar) Syllable

Table. 4.1 Categories of Characters

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

4.1 Myanmar Syllable Composition

The basic structure of Myanmar syllable composition follows a consistent pattern. There is a breakdown of the components in a Myanmar syllable.

Initial consonant (Onset): This is the first sound or consonant in a syllable. It can be a single consonant character or a consonant cluster (two or more consonants combined). Not all syllables have an initial consonant. If there is no initial consonant, the syllable begins with the medial vowel.

Medial vowel (Nucleus): This is the vowel sound that appears after the initial consonant, or at the beginning of the syllable if there is no initial consonant. Myanmar has a set of eight vowel characters, and each vowel character represents a specific vowel sound. The medial vowel is an essential component of every syllable.

Final consonant (Coda): This is the last sound or consonant in a syllable, appearing after the medial vowel. Not all syllables have a final consonant. If there is no final consonant, the syllable ends with the medial vowel. The final consonant can be a single consonant character or a consonant

cluster. The possible combinations of final consonants are more limited compared to the combinations of initial consonants.

Tone mark (Tone): Myanmar is a tonal language, meaning that the pitch or tone of a syllable can change the meaning of a word. Some syllables may have a tone mark, which is a diacritic symbol that indicates the specific tone associated with the syllable. There are four main tones in Myanmar: high, low, creaky, and stopped. The tone mark is placed on top of the syllable or attached to the final consonant.

The Burmese language employs an abugida writing system, where consonant letters are utilized to represent syllables that contain an implicit inherent vowel. Diacritics are used to modify the consonant letters, allowing for the creation of consonant clusters, changes in tones, and alteration of the inherent vowel. To visually represent a Burmese syllable, a numerical system is employed, indicating the character order within the composition. Consonant letters are denoted by 1 and 4, while diacritics are represented by 2, 3, 5, and 6. The combination of 2 and 3 with 1 forms consonant clusters. Tone marks are indicated by 5, and 6 functions as the virama, which depresses the inherent vowel of 4 to form the syllable's coda. In general, Burmese syllables consist of multiple characters, and their identification follows specific rules, including the attachment of diacritics and consonant letters with a virama to the modified letter [8].

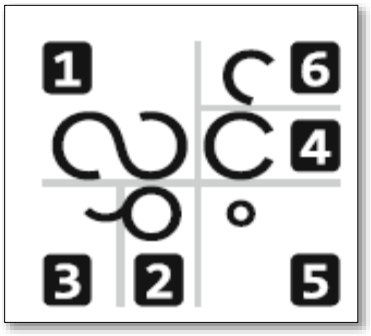

Figure 4.2 Myanmar Syllable Composition

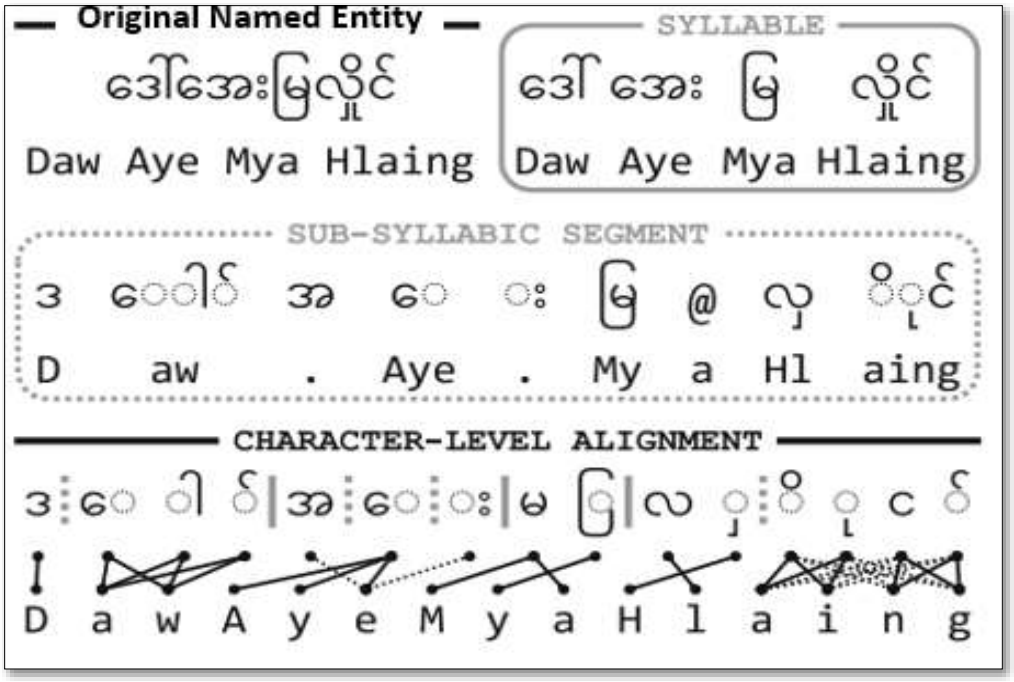

Figure 4.3 Named Entity Segmentation Process using Char., Sub-Syl. and Syl. Units

The figure 4.3 shows the NE transforming process from syllables to characters. In Burmese writing, sub-syllable segmentation is a meticulous process that involves analyzing the alignment of characters at both the individual and unit levels. The upper-left corner presents a raw transliteration instance, while the upper-right corner displays the corresponding syllables. To indicate inserted or silent segments, the sub-syllabic segmentation employs the symbol @/. The alignment of characters at the individual level is complex, as depicted by the dash lines that demonstrate the influence of surrounding Burmese characters on the spelling. In the character-based analysis, solid vertical bars are used to signify syllable boundaries, while dashed vertical bars represent sub-syllabic unit boundaries.

4.2 Challenges

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 [6], 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 analysed the transliteration implications on Myanmar language for the letters in Wikipedia and Myanmar Language Commission dictionary.

4.2.1 Analysis for Burmese Transliteration on Phonotactic Issues

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 [13].

4.2.1.1 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 4.2. 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 /ph/ sound. The differentiation of phonemes sounds /s/ and /sh/ 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>. 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 → လော်ဂရစ်သမ် and THEORY → သီအိုရီ where the <TH> is for /ð/ and /θ/ respectively.

Table 4.2 The phoneme-to-grapheme mapping

Latin Letters	Burmese Letters
/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->

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, JULY → ဇူလိုင်, and JUNE → ဇွန်. In these cases, <J> to <ဇ_> replaces /dʒ/ with ဂျ, but this doesn't apply to examples like JOURNAL → ဂျာနယ် and JURY → ဂျူရီ. Additionally, etymology may play a role in cases like JESUS → ယေရှု, where <J> to <ယ-> and <S> to <ရှ-> may be influenced by Biblical Hebrew. The following example is about a Spanish name, JUAN→ဝှမ်. Here an underlying chain of <JU>→ /xw/ → /hw/ → /ʍ/ to <ဝှ> can be considered behind the surface <JU> to <ဝှ>.

Table 4.3. Specific Situation in the Myanmar language related to grapheme-to-grapheme mappings

Example Word (English)	Specific Graphemes	Transliteration Letters of Graphemes	Transliteration (Myanmar)
LOGARITHM	<TH>	<သ_>	လော်ဂရမ်သမ်
THEORY	<TH>	<သ_>	သီအိုရီ
JANUARY	<J>	<ဇ_>	ဇန်နဝါရီ
JULY	<J>	<ဇ_>	ဇူလိုင်
JUNE	<J>	<ဇ_>	ဇွန်
JOURNAL	<J>	<ဂျ_>	ဂျာနယ်
JURY	<J>	<ဂျ_>	ဂျူရီ

4.2.1.2 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 and CHROMIUM→ခရိုမီယမ်. When <CH> just stands for /k/, /k/ to <ခ> may not be triggered, CINCHONA → စင်ကိုနာ nor likely be triggered in "<CHL>" for /kl/ in the following example CHLORINE→ကလိုရင်း. Similarly, "<tr>" is mapped to "<ထရ->" where the aspirated "<ထ>" is used instead of the common "<တ>" for ELECTRON→ အီလက်ထရွန် and TRANSISTOR →ထရန်စစ္စတာ or even irregular spellings as <TR> to <တြ_> in GEOMETRY→ ဂျီသြမေတြီ. The mapping for "
" is "<ဗြ->" (regularly "<ဘရ->") and can be seen in the transliteration of "Britain" as "<ဗြိတိန်>".

Table 4.4. Specific English onset clusters in the Myanmar language

Example Word (English)	Specific Grapheme	Transliteration Letters of Graphemes	Transliteration (Myanmar)
CHRIST	<CHR> (/kr/)	<ခရ_>	ခရစ်
CHROMIUM	<CHR> (/kr/)	<ခရ_>	ခရိုမီယမ်
CHLORINE	<CHL> (/kl/)	<က_>	ကလိုရင်း
CINCHONA	<CH> (/k/)	<က_>	စင်ကိုနာ
ELECTRON	<TR> (/tr/)	<ထရ_>	အီလက်ထရွန်
TRANSISTOR	<TR> (/tr/)	<ထရ_>	ထရန်စစ္စတာ

GEOMETRY	<TR> (/tr/)	<တြ_>	ဂျီသြမေတြီ
BRITAIN	 (/br/)	<ဗြ_>	ဗြိတိန်

4.2.1.3 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. If a hiatus begins with /i/, it is customary to use <ယ-> (or <ရ->) rather than <အ> in words such as UNION→ ယူနီယန်. Take note that /ju/ is altered to <ယူ> at the word's start to prevent the combination of <အျ>. In the word MERCURY → မာကျူရီ, it is common to use <+ျူ> for the sound /ju/ after a general onset. In LOUISIANA → လူဝီစီယားနား, at times, <o_> is added after /u/. Some stable borrowed words in Myanmar use independent vowel letters at the beginning, such as APRIL → ဧပြီ and AUGUST → ဩဂုတ်. For triphthongs, syllables may be re-segmented with semi-vowel insertion, as seen in POWER → ပါဝါ , where /aʊr/ is analyzed as /a.wṛ/. In the word WIRE → ဝိုင်ယာ, /aɪr/ is analyzed as /aɪ.ṛ / with nasalization and the addition of <ယ>, as there is no standalone /aɪ/ rhyme in Myanmar. The Table 4.5 summarizes how English words are transliterated into Myanmar according to the described conventions.

Table 4.5. Specific situations in the Myanmar language related to sound clusters and vowels

Example Word (English)	Specific Graphemes	Transliteration Letters of Graphemes	Transliteration (Myanmar)
IODINE	<I>	<အ->	အိုင်အိုဒင်း
UNION	<U>	<ယ-> (or <ရ->)	ယူနီယန်
MERCURY	<U>	<ယူ>	မာကျူရီ
LOUISIANA	<U>	<+ျူ>	လူဝီစီယားနား

APRIL	<A>	Independent Vowel Letters	ဧပြီ
AUGUST	<A>	Independent Vowel Letters	ဩဂုတ်
POWER	<OR> /a.wṛ /	Semi-vowel Insertion	ပါဝါ
WIRE	<IR> /aɪr/	<ယ>	ဝိုင်ယာ

4.2.2 Analysis for Native Burmese Transliteration Issues

Analysis for Native Burmese Transliteration issues refers to the examination and study of challenges and problems encountered when transliterating native Burmese words or text into another writing system or script. Transliteration involves representing the sounds and characters of one language using the symbols and characters of another language.

When it comes to Burmese transliteration, there are several issues that can arise due to the unique characteristics of the Burmese language and writing system. Some common challenges include:

Non-standardized transliteration: Burmese transliteration lacks a standardized system, leading to variations in representing Burmese sounds and characters in different contexts or by different individuals. This can result in inconsistencies and difficulties in accurately transliterating Burmese words.

Tonal representation: Burmese is a tonal language, meaning that the pitch or tone of a syllable can change the meaning of a word. Transliterating tonal information into non-tonal writing systems can be challenging, as there is no direct one-to-one correspondence between Burmese tones and the characters of other scripts.

Unique Burmese characters: Burmese script has its own set of characters that may not have direct equivalents in other scripts. Transliterating these characters accurately can be problematic, as finding suitable substitutes in another writing system may not always capture the exact pronunciation or meaning.

Consonant clusters and complex orthography: Burmese has complex orthographic rules and allows for the formation of consonant clusters. Transliterating these clusters and handling complex spelling patterns can be difficult, as some writing systems may not have provisions to represent these intricacies.

Vowel representation: Burmese vowels can have various positions and combinations within syllables. Transliterating these vowel structures and accurately representing vowel sounds in another script can be a challenge.

To address these transliteration issues, it is crucial to develop standardized guidelines and transliteration systems that take into account the unique features of the Burmese language. This can help ensure consistency, accuracy, and clarity in transliterating Burmese words and texts for different purposes such as academic research, language learning, or communication in multilingual contexts.

There are many inconsistent and irregular spellings between Myanmar and English. These spellings are used as naming entities of honorific and can be seen in the following table [64].

Table 4.6 Inconsistent Spellings in Myanmar Native Names

Honorific	**Burmese**	**Usage**
Shin	ရှင်/သျှင်	Used by monks and noble men and women (Archaic; Shin Arahan, Shin Ye Htut, Yawei Shin Htwe)
Bo, Bygyoke (or) Bo	ဗိုလ်(or)ဗိုလ်ချုပ်	Used for military officers(e.g., Bogyoke Aung San)
	ဘို	Used as part of given names (e.g., Bo Phyu)
Daw	ဒေါ် (or) တော်	Used for mature women or women in a senior position (e.g., Daw Aung San Suu Kyi) (or) (e.g., Shwe Nan Taw)
Du Wa, Duwun	ဒူးဝါး, ဓူဝံ	Used for Kachin Chiefs.
	(ဒူ (or) ဓူ)	Used as part of given names.

Mai(or)Me	မယ်	Used by some young women in lieu of မ but exceedingly rare. (e.g., Mai Khway)
Maung (abbr. Mg)	မောင်	Sometimes used as part of given names
Nai (or) Naing	နိုင်	Used by Mon men; equivalent to U (e.g., Naing Shwe Kyin), from Mon.
Nant	နမ့် (or) နန့်	Used by Karen (especially in West Pwo Karen) women (e.g., Nant Thinzar Win)
Sai	စိုင်း (or) (ဆိုင်း)	Used by Shan men (e.g., Sai Htee Saing), from Shan.
Salai (or) (Salaing)	ဆလိုင်း	Used by Chin Men
Sao	စဝ် (or) စပ်	Used by Shan royalty စပ်ဖွေသိုက် (or) စဝ်ဖွေသိုက် (e.g., Sao Shwe Thaik) from Shan
Sa	စ (or) သ	Used by Karen Men (especially in West Pwo Karen) e.g., Sa Nay Lin Kyaw (စနေလင်းကျော်) e.g., သီရိလင်္ကာ (Srilanka)
Sawbwa (or) Sawbwar	စော်ဘွား	Burmese approximation of Shan(Saopha), used as a suffix for Shan Chiefs e.g., Naungshwe Sawbwa Sao Shwe Thaik)
Saya(or) Sayar	ဆရာ	Used for males of senior rank or age
Sayadaw (or) Sayartaw	ဆရာတော်	Used for senior monks (e.g., Sayadaw U Pandita)
Sayama (or) Sayarma	ဆရာမ	Used for females of senior rank or age

Tekkatho (or) University	တက္ကသိုလ်	Used by writers (Archaic; e.g., Tekkatho Phone Naing)
Thakin (or) Thakhin (or) Master	သခင်	Used by the members of Dobama Asiayone “the Thakins”(Archaic; e.g., Thakin Kodaw Hmaing)
Theippan (or) Science	သိပ္ပံ	Used by writers (Archaic; e.g., Theippan Maung Wa)
U or Oo	ဦး	Used for mature men or men in a senior position and monks (e.g., U Thant, U Ottama)
	ဦး	Used as part of given name (e.g., Hay Man Oo)

4.3 Summary

In summary, the Burmese language, also known as Myanmar language, has its own distinct structure and is written using the Burmese script. Transliteration from Burmese to other languages presents several challenges due to differences in script, tonal distinctions, unique sounds, and linguistic features. These challenges include mapping the Burmese characters to Latin characters, accurately representing tonal nuances, dealing with unique sounds, and capturing linguistic features. As a result, different transliteration systems and variations exist. To ensure accurate representation and effective communication, it is crucial to understand the specific context and purpose of transliteration when dealing with Burmese. Despite the challenges, proper attention and consideration can facilitate successful transliteration and bridge the gap between Burmese and other languages.

CHAPTER 5

TRANSFORMER BASED BURMESE(MYANMAR)-ENGLISH NAMED ENTITY TRANSLITERATION

Named entity transliteration involves converting named entities from one language script to another accurately, crucial for tasks like cross-language information retrieval (CLIR) and machine translation (MT). While a straightforward method would be mapping each character from the source language to its common counterpart in the target language, the complexities of language, including ambiguous spellings and pronunciations, necessitate data-driven approaches. Most transliteration systems rely on contextual information to disambiguate and ensure accuracy, acknowledging the inherent challenges posed by linguistic variations and inconsistencies across languages [15,19,48].

While transliteration has been a long-studied problem, some important aspects received little attention. There is not clear guidance that addresses a number of common design considerations faced when building a robust multilingual transliteration system, such as data representation and the huge gap in results depending on the language pairs and transliteration direction [41]. Like many other NLP fields recently, neural transliteration systems have gained popularity. The Transformer method uses a simple neural network architecture based solely on attention mechanisms. That motivated us to learn if it can produce strong results on transliteration as it did on translation.

Recent work in Neural Machine Translation (NMT) has proposed a number of approaches to use neural networks in variable-length sequence-to-sequence tasks such as transliteration. The encoder-decoder architecture [42] is a recurrent neural network setup with two parts. An encoder is fed input tokens one at a time and encodes them into a hidden state vector. At the end of the input sequence, an end-of-sentence token is fed to signify the end of the encoding phase. Next, the hidden state output of the encoder is fed into the decoder. The decoder emits tokens and updated hidden states, which are recursively fed into itself, until there are no more output tokens to produce. An additional mechanism, attention allows the decoder to focus on different parts of the input sequence and capture long-range dependencies. More recently, the Transformer [57] model was proposed, which avoids the need for sequential processing, relying only on self-attention. A benefit of this approach is there is no information bottleneck in the encoded hidden state vector as in the Encoder-

Decoder approach. Additionally, because there is no longer a sequential recurrent network, model training can be better parallelized, decreasing model training time.

This chapter delves into the intricate workflow of developing a Transformer-based Burmese (Myanmar) – English Named Entity (NE) transliteration system. Harnessing the power of state-of-the-art deep learning architectures like Transformers, this system aims to accurately transliterate Burmese named entities into English equivalents.

5.1 Burmese (Myanmar) – English Named Entity Transliteration System

The figure represents the system flow of a sophisticated Myanmar-English Named Entity Transliteration System. This system is designed to accurately transliterate Myanmar named entities into their corresponding English counterparts using advanced Natural Language Processing (NLP) techniques, particularly leveraging the Transformer model architecture. The system begins with the collection of a diverse dataset containing pairs of Myanmar named entities and their English transliterations. This dataset undergoes thorough preprocessing, including cleaning, tokenization into sub word units, and segmentation into training, validation, and test sets. Tokenization plays a crucial role in breaking down the input Myanmar text and output English transliterations into meaningful units, which are then mapped to numerical indices using vocabulary mappings. These indices are embedded into continuous vector representations through embedding layers in the Transformer model.

The heart of the system lies in the Transformer model architecture, meticulously configured with encoder and decoder layers, multi-head self-attention mechanisms, position-wise feedforward networks, and other components to capture intricate linguistic patterns and dependencies. During training, the model learns to predict English transliterations given input Myanmar named entities using teacher forcing, where the actual target tokens are fed into the decoder. Optimization algorithms like Adam fine-tune the model parameters to minimize loss and enhance transliteration accuracy. The trained model undergoes evaluation on a validation set, where metrics such as edit distance and character-level accuracy assess its performance. Fine-tuning may occur based on evaluation results to optimize transliteration quality.

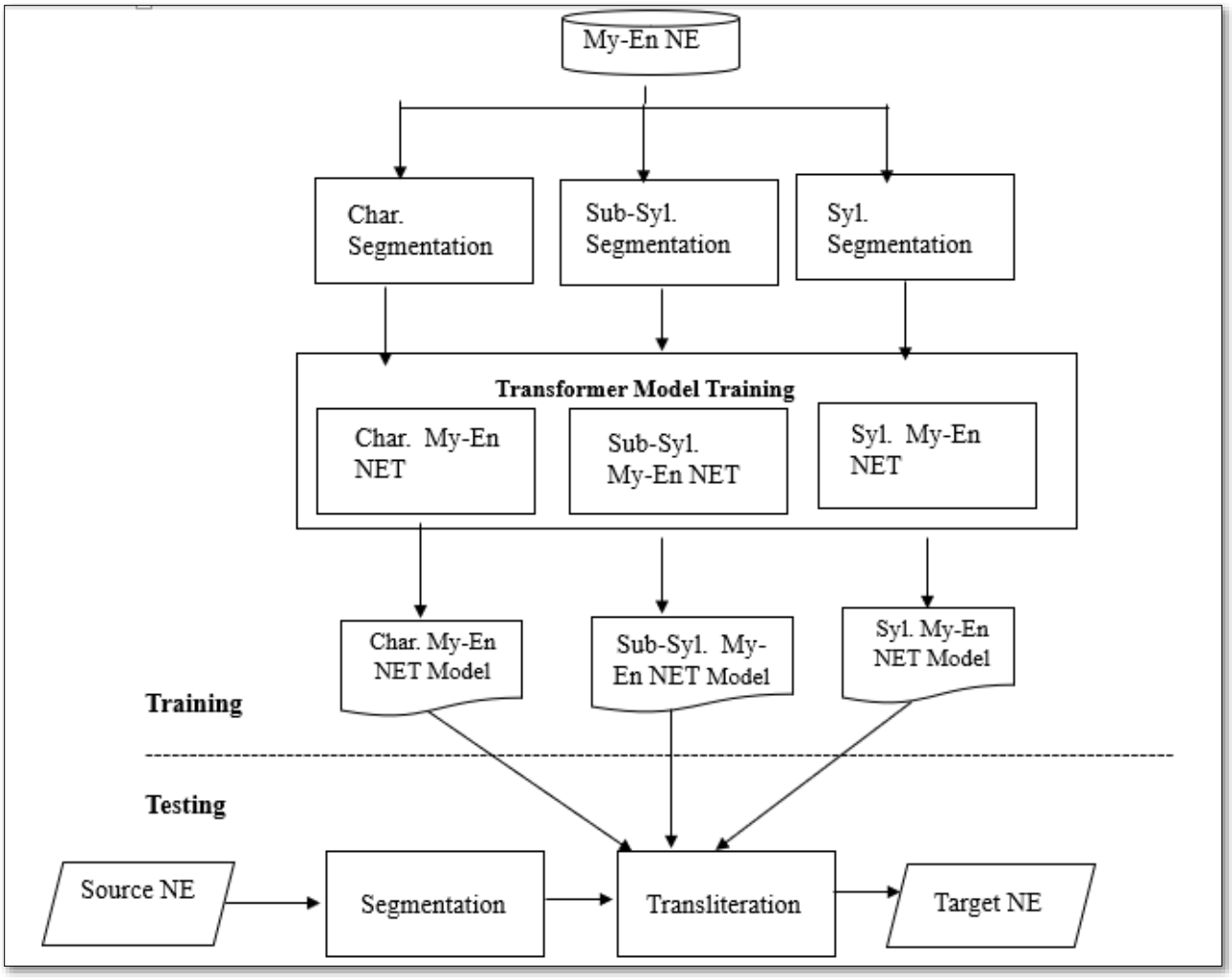

Figure 5.1 (Burmese) Myanmar-English Named Entity Transliteration System

5.2 Transformer Model Architecture

The Transformer model architecture is a groundbreaking neural network architecture introduced in the paper "Attention Is All You Need" by Vaswani et al. in 2017 [57]. It has revolutionized the field of natural language processing (NLP) and is widely used in tasks such as machine translation, text generation, and named entity recognition. Below is an overview of the key components of the Transformer model architecture:

- **Encoder-Decoder Structure**

The Transformer model consists of an encoder and a decoder, each composed of multiple identical layers. This architecture is similar to sequence-to-sequence models but without recurrent neural networks (RNNs) or convolutional layers.

- **Encoder Layers**

Each encoder layer in the Transformer consists of two main components: Self-Attention Mechanism: Computes attention scores between all positions in the input sequence to capture dependencies within the sequence. It allows each word to attend to other words, learning contextual representations. Feedforward Neural Network: Applies a pointwise fully connected feedforward network to each position independently and identically. This network introduces non-linearity and helps capture complex patterns in the data.

- **Decoder Layers**

Similar to encoder layers, each decoder layer in the Transformer comprises two main components: Self-Attention Mechanism (Decoder Self-Attention): Computes attention scores between positions in the decoder input sequence. It prevents the model from attending to future tokens during training by masking out future positions. Encoder-Decoder Attention Mechanism: Computes attention scores between the decoder input and the encoder outputs. This mechanism helps the decoder focus on relevant parts of the input sequence for generating output.

- **Attention Mechanism**

The attention mechanism in the Transformer is key to its success. It allows the model to focus on different parts of the input sequence (or encoder outputs during decoding) depending on the relevance to the current position. The attention scores are computed using scaled dot-product attention, which balances the importance of different positions.

- **Positional Encoding**

Since the Transformer model does not inherently understand the order of tokens in a sequence like RNNs or CNNs, positional encoding is added to the input embeddings. Positional encoding provides information about token positions, enabling the model to learn sequential relationships.

- **Multi-Head Attention**

To capture different types of information and enhance model performance, the attention mechanism in the Transformer is multi-headed. This means that it computes multiple sets of attention scores in parallel (heads), and the results are concatenated and linearly transformed to obtain the final attention output.

- **Layer Normalization and Residual Connections**

Layer normalization is applied after each sub-layer (self-attention and feedforward network) in both the encoder and decoder layers. Residual connections are also used, where the input to each sub-layer is added to its output before normalization. These techniques aid in stabilizing training and mitigating the vanishing gradient problem.

- **Output Layer**

The output layer of the decoder generates the final predictions (e.g., token probabilities in language modeling or softmax scores in machine translation) based on the decoder's contextual representations and attention mechanisms. Overall, the Transformer model architecture has proven to be highly effective in capturing long-range dependencies, facilitating parallel computation, and achieving state-of-the-art results in various NLP tasks. Its modular and attention-based design has inspired numerous subsequent architectures and advancements in the field.

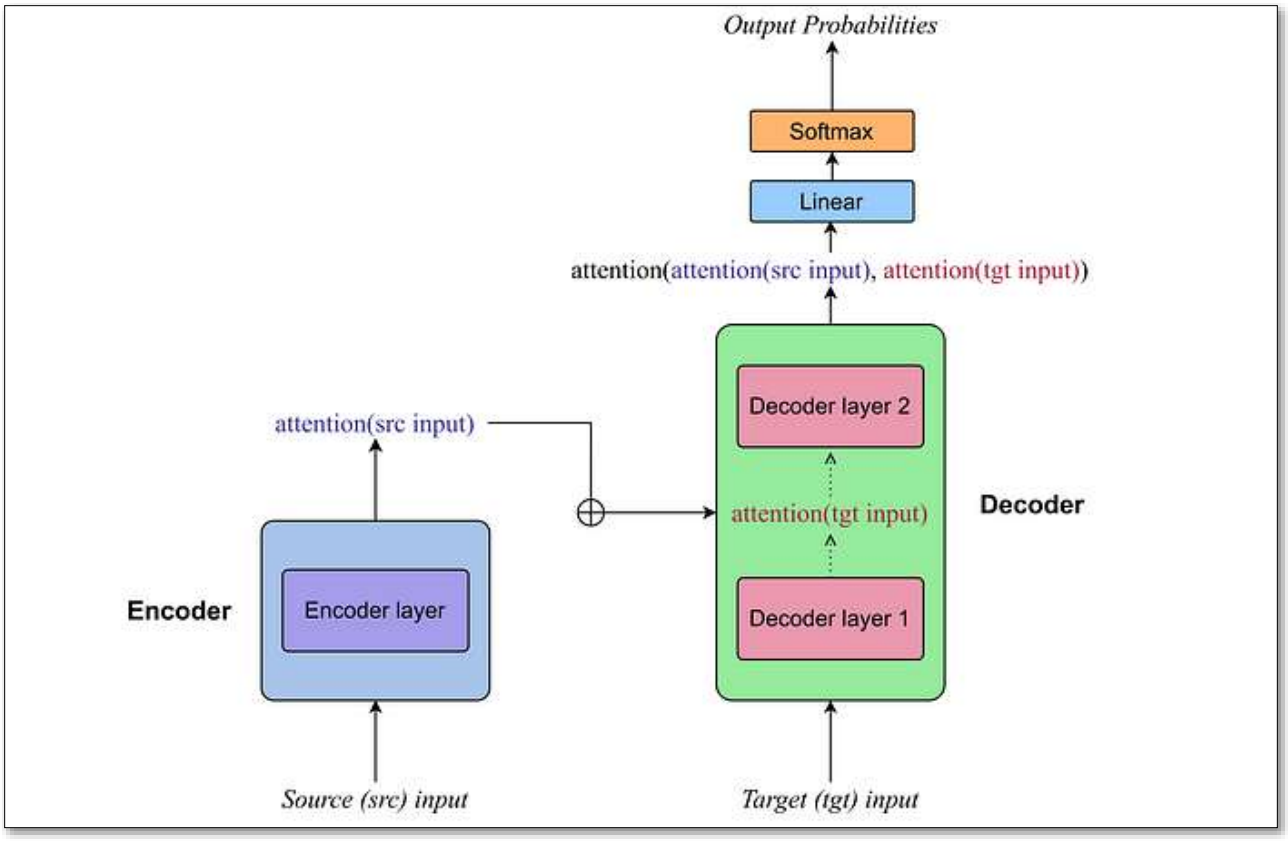

Figure 5.2 Overview of the Transformer Model Architecture

5.3 Transformer based Myanmar-English Named Entity Transliteration System

The Encoder-Decoder workflow lays the foundation for the Myanmar-English Named Entity (NE) Transliteration system, leveraging the Transformer architecture's power. The Encoder component processes the input Myanmar text, capturing intricate linguistic nuances using self-attention mechanisms and multi-head attention layers. This encoding phase embeds the input sequence into a continuous vector space, preserving essential semantic and syntactic information. On the other hand, the Decoder component generates the corresponding English transliteration by attending to the encoded Myanmar representation and iteratively predicting the output tokens. Together, this Encoder-Decoder framework forms the backbone of the NE Transliteration system, enabling accurate and contextually relevant conversions between Myanmar and English named entities.

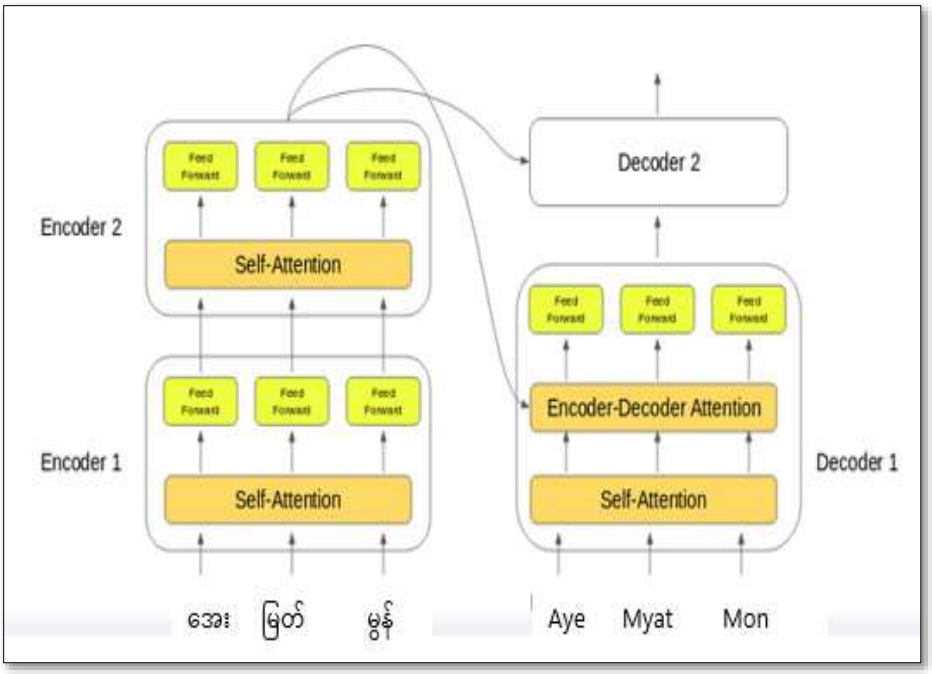

Figure 5.3 Encoder-Decoder Structure of the Transformer Model

In the Myanmar-English Named Entity Transliteration system, all input and output tokens undergo a crucial transformation into vectors through learned embeddings before being fed into the Encoder and Decoder components. The process of converting tokens to embeddings involves mapping each token to a dense vector representation learned during the training phase. This embedding step is essential as it allows the model to capture semantic and syntactic similarities between tokens, enabling effective information processing.

In the context of the Myanmar-English Named Entity Transliteration system, Positional Encoding enhances the model's understanding of token positions within input and output sequences. Each token's embedding is augmented with positional encoding vectors, which encode the token's position in the sequence relative to others. This encoding is essential because tokens with similar semantic meanings but different positions should be treated differently by the model to preserve contextual coherence and syntactic accuracy. Where PE is positional encoding, pos is the position and i is the dimension.

$$\boldsymbol{PE}_{(pos,2i)} = \mathbf{sin}\left(\frac{pos}{\mathbf{10000}^{2i/dmodel}}\right) \qquad \text{Equation (5.1)}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/dmodel}}\right)$$ Equation (5.2)

5.3.1 Self Attention in Transformer

Understanding the meaning and context of words in a sentence is facilitated through self-attention, or intra-attention, a mechanism that computes a representation of a sequence by relating different positions within it. Self-attention layers, which connect all positions in a sequence through a fixed number of operations, are faster than recurrent layers and enable a better understanding of word meaning and context. In the Transformer architecture, the Attention function maps a query and a set of key and value vectors to generate an output, enhancing the comprehension of words' meanings and contexts. The use of vectors for query, key, and value, along with Scaled Dot-Product Attention, enables the calculation of attention weights for each word in a sentence, contributing to a weighted sum of values for a final score. Self-attention, or intra-attention, serves as a crucial mechanism for understanding words in a sentence by relating different positions within the sequence and computing a comprehensive representation, ultimately aiding in grasping meaning and context effectively.

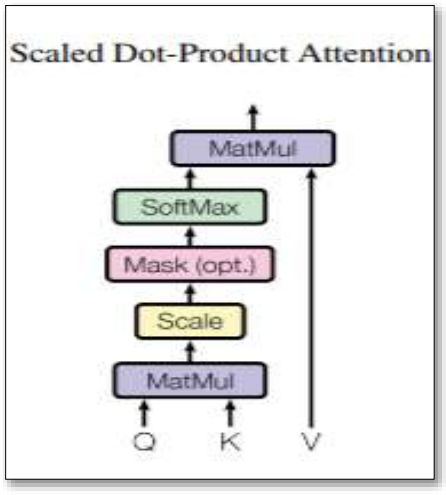

Figure5.4 Scaled Dot-Product Attention

Calculating self-attention is a fundamental operation within the Transformer architecture, pivotal for capturing contextual dependencies and semantic relationships within a sequence. In the context of the Myanmar-English Named Entity Transliteration system, self-attention mechanisms play a crucial role in both the Encoder and Decoder components. The process of calculating self-attention involves three key steps: computing attention scores, applying softmax normalization, and generating weighted context vectors. In the Transformer architecture, the encoder's input vectors are transformed into three distinct vectors: the query vector, key vector, and value vector, each playing a crucial role in self-attention mechanisms.

Step1: Dot-Product

The process of determining how much focus should be placed on other words in an input sentence involves taking the dot product of the query and key for each word. This dot product, also known as Dot-Product, is instrumental in understanding attention mechanisms within neural networks. It helps in allocating attention weights to different words based on their relevance and importance in the context of the sentence. By calculating the dot product for each word, the model can effectively decide how much attention to give to individual words, thereby enhancing the overall understanding and processing of textual data within the neural network framework.

Table 5.1 Dot-Product Calculation

Words	Query	Key	Value	Matmul
ဆွစ်	Q1	K1	V1	Q1.K1
ဇာ		K2	V2	Q1.K2
လန်		K3	V3	Q1.K3

Step 2: Scale the Dot Product

To scale the Dot-Product in the context where the dimension of the key vector is 64, a common practice involves dividing the Dot-Product by the square root of the dimension. In this case, as the dimension is 64, the Dot-Product would be divided by 8 to achieve scaling. This scaling process is crucial in neural network architectures, particularly in attention mechanisms, as it helps manage the magnitude of attention weights and ensures that they are appropriately adjusted relative to the dimensionality of the key vectors. By dividing the Dot-Product in this manner, the model can

maintain stability and efficiency in its attention computations, contributing to improved performance and robustness in handling complex input data.

Table 5.2 Scale the Dot Product Calculation

Words	Query	Key	Value	Matmul
ဆွမ်	Q1	K1	V1	Q1.K1
စာ		K2	V2	Q1.K2
လန်		K3	V3	Q1.K3

Step 3: Apply Softmax to normalize the scaled values

Softmax normalization is a key step in many machine learning algorithms, especially in the context of attention mechanisms. After scaling the Dot-Product by dividing it by the square root of the dimension of the key vector (e.g., dividing by 8 when the dimension is 64), the resulting values are passed through a Softmax function. This Softmax operation transforms the scaled values into a probability distribution where all values are positive and collectively sum up to 1. This normalization process is critical as it ensures that the attention weights assigned to different words or elements in the input sequence are interpretable as probabilities, representing the relative importance or relevance of each element in the context of the overall sequence. Softmax normalization facilitates clearer and more intuitive understanding of attention weights, aiding in the effective management of attention mechanisms within neural network architectures.

Table 5.3 Apply Softmax to normalize the scaled values

Words	Query	Key	Value	Matmul	Scale($^1/\sqrt{d_k}$)	SoftMax
ဆွမ်	Q1	K1	V1	Q1.K1	Q1.K1/8	X_{11}
စာ		K2	V2	Q1.K2	Q1.K2/8	X_{12}
လန်		K3	V3	Q1.K3	Q1.K3/8	X_{13}

Step 4: Calculate the weighted sum of the values

After normalizing the scores using techniques like Softmax, the next step in many neural network architectures involves applying a Dot-Product operation between these normalized scores

and the corresponding value vectors. This Dot-Product computation essentially combines the importance weights assigned to different elements in the input sequence with their respective values. By calculating this Dot-Product for each element and summing them up, we effectively aggregate the weighted information from the input sequence. This process is crucial in attention mechanisms as it allows the model to focus on relevant parts of the input by assigning appropriate weights and integrating the relevant information to generate meaningful outputs.

Table 5.4 Calculate the weighted sum of the values

Words	Query	Key	Value	Matmul	Scale($^1/\sqrt{d_k}$)	SoftMax	Matmul	Sum
ဆွစ်	Q1	K1	V1	Q1.K1	Q1.K1/8	X_{11}	X_{11}.V1	Z_1
ဇာ		K2	V2	Q1.K2	Q1.K2/8	X_{12}	X_{12}.V2	
လန်		K3	V3	Q1.K3	Q1.K3/8	X_{13}	X_{13}.V3	

Calculating Self Attention

The complete equation for Self-Attention

$$Attention\ (Q,K,V) = softmax\ \left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad \text{Equation (5.3)}$$

These steps are repeated for every word in the sentence.

Table 5.5 Calculating Self Attention

Step2Words	Query	Key	Value	Matmul	Scale($^1/\sqrt{d_k}$)	SoftMax	Matmul	Sum
ဆွစ်	Q1	K1	V1	Q1.K1	Q1.K1/8	X_{11}	X_{11}.V1	Z_1
ဇာ		K2	V2	Q1.K2	Q1.K2/8	X_{12}	X_{12}.V2	
လန်		K3	V3	Q1.K3	Q1.K3/8	X_{13}	X_{13}.V3	
Words	Query	Key	Value	Matmul	Scale($^1/\sqrt{d_k}$)	SoftMax	Matmul	Sum
ဆွစ်		K1	V1	Q2.K1	Q2.K1/8	X_{21}	X_{21}.V1	
ဇာ	Q2	K2	V2	Q2.K2	Q2.K2/8	X_{22}	X_{22}.V2	Z_2
လန်		K3	V3	Q2.K3	Q2.K3/8	X_{23}	X_{23}.V3	
Step2Words	Query	Key	Value	Matmul	Scale($^1/\sqrt{d_k}$)	SoftMax	Matmul	Sum
ဆွစ်		K1	V1	Q3.K1	Q3.K1/8	X_{31}	X_{31}.V1	
ဇာ		K2	V2	Q3.K2	Q3.K2/8	X_{32}	X_{32}.V2	
လန်	Q3	K3	V3	Q3.K3	Q3.K3/8	X_{33}	X_{33}.V3	Z_3

5.4 Summary

The summary of the Transformer-based Named Entity Transliteration System encapsulates its end-to-end workflow and exceptional model performance. Beginning with comprehensive data collection and preprocessing techniques, the system curate's high-quality datasets essential for robust transliteration models. The integration of Transformer architecture, featuring multi-head attention mechanisms and positional encodings, enables the system to capture nuanced linguistic patterns and generate accurate transliterations. Through rigorous model training and validation, the system attains superior performance metrics, demonstrating its efficacy in real-world applications requiring precise cross-lingual entity transliteration.

CHAPTER 6

EXPERIMENTAL RESULTS

This chapter delves into the intricate workings and detailed results of the Myanmar-English Named Entity Transliteration System. This system stands at the intersection of linguistic nuance, computational algorithms, and cross-cultural understanding, aiming to bridge the gap between Myanmar's rich linguistic heritage and the global reach of the English language.

6.1 Experimental Setting

The Transformer-based named entity transliteration approach underwent rigorous evaluation in the context of the Myanmar-English language pair, encompassing transliteration in both directions. The evaluation aimed to assess the system's performance in accurately mapping named entities between Myanmar and English, considering the linguistic nuances and orthographic differences inherent in these languages.

6.1.1 Preprocessing for Named Entity Transliteration System

One of the primary goals of preprocessing is to enhance data quality by removing noise, standardizing text formats, and addressing script-specific variations. In the context of named entity transliteration, preprocessing plays a pivotal role in ensuring accurate and reliable conversion of named entities between different scripts or languages. Preprocessing encompasses a series of tasks aimed at cleaning, normalizing, and preparing the input data for effective transliteration processing.

6.1.1.1 Character based Named Entity Segmentation

The character dataset consists of Myanmar-English named entities sourced from various official documents and online sources. The dataset comprises native Myanmar named entities and foreign English named entities, covering entity types such as person names, location names, and organization names. An analysis of character distribution reveals that Myanmar named entities predominantly consist of characters from the Myanmar script, while English named entities primarily use Latin characters with occasional diacritics.

Prior to segmentation, the researcher conducted data cleaning to remove duplicate entries and standardize entity formats. Normalization is also performed by converting all text to lowercase and removing diacritics from Myanmar script characters. Tokenization was carried out at the character level, breaking down each named entity into individual characters. For Myanmar text, the researcher utilized the following character-level segmentation strategy to handle the inherent character structure of the language, while English text was tokenized into alphabetic characters.

```
#!/usr/bin/env python

# -*- coding: utf-8 -*-

import re, codecs

fo = open("input.txt", "r+", encoding='utf-8')

fw = open("output.txt", "w", encoding='utf-8')

for line in fo.readlines():

line = re.findall('.', line)

print(line, file=fw)

'''for character in line:

print(character,file=fw)'''
```

Table 6.1 Western Data Sample for Character NET Pairs

Categories	**Seg. Units**	**Western NE (My)**	**Western NE (En)**	**Usage**
Person	char.	ခ ရ စ ် စ တ ီ း နာ း	C h r i s t i n a	The usage of the name "Christina" is primarily as a given name for girls. A personal name, identifying an individual person
Place	char.	က ယ ် လ ီ ဖ ို ု း နာ ီ း ယ ာ း	C a l i f o r n i a	“California” is the name of a state in the United States, located on the West Coast. A place name, identifying an individual place
Org.	char.	က ေ ာ ် ပ ရ ေ း ရ ှ င ် း	C o r p o r a t i o n	“Corporation” is often used more broadly to refer to large businesses or companies

Table 6.2 Native Data Sample for Character NET Pairs

Categories	**Seg. Units**	**Native NE(My)**	**Native NE (En)**	**Usage**
Person	char.	မ ေ ာ င ် ထ ိ န ် လ ငွ ် း	M g H t e i n L i n	The usage of an individual's name “Mg Htein Lin” is context-specific and can depend on factors such as their profession, achievements,

				or public activities in Myanmar native names.
Place	char.	ဂေါ်ရန်ဂျီကျွန်း	G a w y a n g y i K y u n	The island name “Gawyangyi Kyun” is commonly used in the local community or region.
Org.	char.	အိုးရှင်းစူပါစင်တာ	O c e a n S u p e r C e n t e r	The name "Ocean Super Center" suggests a business in the retail sector, possibly a supermarket or a large store that offers a wide range of products, including groceries, household items, and possibly other goods.

6.1.1.2 Sub-Syllable based Named Entity Segmentation

In languages like Myanmar with complex syllabic structures, sub-syllable segmentation plays a vital role in accurately transliterating named entities. Sub-syllable segmentation involves breaking down named entities into smaller phonetic or graphemic units, enhancing the granularity of transliteration processing according to [8].

Table 6.3 Western Data Sample for Sub-Syllable NET Pairs

Categories	**Seg. Units**	**Western NE (My)**	**Western NE (En)**	**Usage**
Person	sub-syl.	ခ@ရစ်စ@တီးနား	C h r i s t i n a	The usage of the name "Christina" is primarily as a given name for girls.

				A personal name, identifying an individual person
Place	sub-syl.	က ယ် လ ီ ဖ ိ ု း နှ ိ း ယ ာ း	C a l i f o r n i a	“California” is the name of a state in the United States, located on the West Coast. A place name, identifying an individual place
Org.	sub-syl.	က ေ ာ ် ပ @ ရ ေ း ရှ ် း	C o r p o r a t i o n	“Corporation” is often used more broadly to refer to large businesses or companies

Table 6.4 Native Data Sample for Sub-Syllable NET Pairs

Categories	**Seg. Units**	**Native NE(My)**	**Native NE (En)**	**Usage**
Person	sub-syl.	မ ေ ာ င် ထ ိ န် လ ် း	M g H t e i n L i n	The usage of an individual's name “Mg Htein Lin” is context-specific and can depend on factors such as their profession, achievements, or public activities in Myanmar native names.
Place	sub-syl.	ဂ ေ ါ ် ရ န် ဂျ ီ ကျ ွန် း	G a w y a n g y i K y u n	The island name “Gawyangyi Kyun” is commonly used in the local community or region.

Org.	sub-syl.	အ ိ ု း ရွ င် း စ ူ ပ ါ စ ် တ ာ	O c e a n S u p e r C e n t e r	The name "Ocean Super Center" suggests a business in the retail sector, possibly a supermarket or a large store that offers a wide range of products, including groceries, household items, and possibly other goods.

6.1.1.3 Syllable based Named Entity Segmentation

Syllable segmentation is a fundamental aspect of named entity transliteration for the Myanmar language, ensuring accurate conversion of Myanmar script into English or other languages. Syllable-based segmentation divides named entities into syllabic units, reflecting the phonetic and orthographic structure of Myanmar script according to [54].

Table 6.5 Western Data Sample for Syllable NET Pairs

Categories	**Seg. Units**	**Western NE (My)**	**Western NE (En)**	**Usage**
Person	syl.	ခ ရ စ် စ တီး နား	C h r i s t i n a	The usage of the name "Christina" is primarily as a given name for girls. A personal name, identifying an individual person
Place	syl.	ကယ် လီ ဖိုး နီး ယား	C a l i f o r n i a	“California” is the name of a state in the United States, located on the West Coast. A place name, identifying an individual place

Org.	syl.	ကော် ပ ရေး ရှင်း	C o r p o r a t i o n	“Corporation” is often used more broadly to refer to large businesses or companies

Table 6.6 Native Data Sample for Syllable NET Pairs

Categories	**Seg. Units**	**Native NE(My)**	**Native NE (En)**	**Usage**
Person	syl.	မောင် ထိန် လင်း	M g H t e i n L i n	The usage of an individual's name “Mg Htein Lin” is context-specific and can depend on factors such as their profession, achievements, or public activities in Myanmar native names.
Place	syl.	ဂေါ် ရန် ဂျီ ကျွန်း	G a w y a n g y i K y u n	The island name “Gawyangyi Kyun” is commonly used in the local community or region.
Org.	syl.	အိုး ရှင်း စူ ပါ စင် တာ	O c e a n S u p e r C e n t e r	The name "Ocean Super Center" suggests a business in the retail sector, possibly a supermarket or a large store that offers a wide range of products, including groceries, household items, and possibly other goods.

6.2 Models and Parameter Setting

One of the key advantages of the Transformer model [57] is its aptitude to tackle long sequences of data. With RNN models [49,50,51], longer sequences 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. 6.1 shows the Transformer architecture for My-En NE transliteration for syllable segmentation unit.

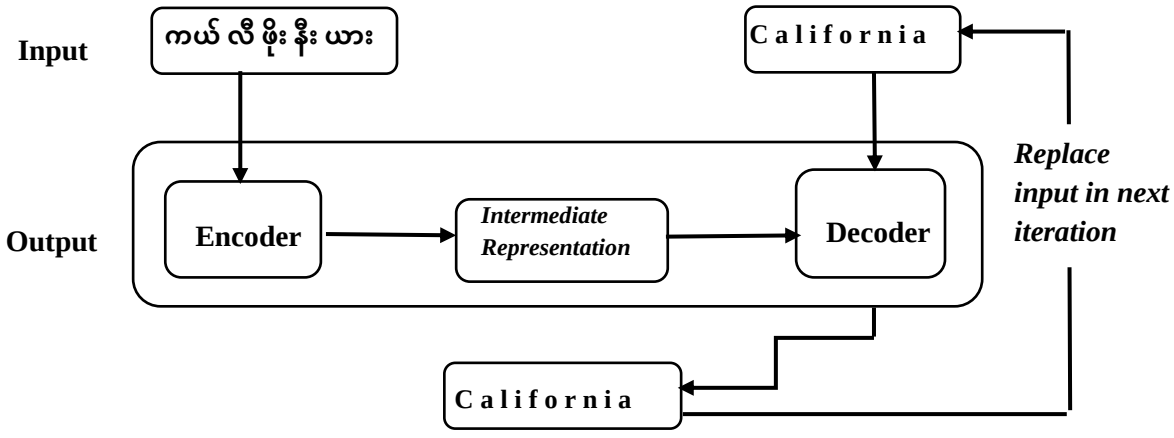

Figure. 6.1 Transformer Architecture for Burmese (Myanmar) to English NE Transliteration for syllable unit

The study on machine translation involved training a Transformer model using the OpenNMT toolkit [25,26]. 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 6.1. 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 [52], 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 [57], 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 Colab 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 6.7. Analyzing Table 6.7 provides insights into the key parameters that influenced the outcomes of the Transformer experiment.

Table 6.7 Hypher Parameter Settings for 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

6.3 Myanmar - English Named Entity Transliteration System Results and Details

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 [23] and WER [24] 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 6.8,6.9 and 6.10 which shows the overall system performance with regard to BLEU and WER.

Table 6.8 System Evaluation Results for Mix Data in term of 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

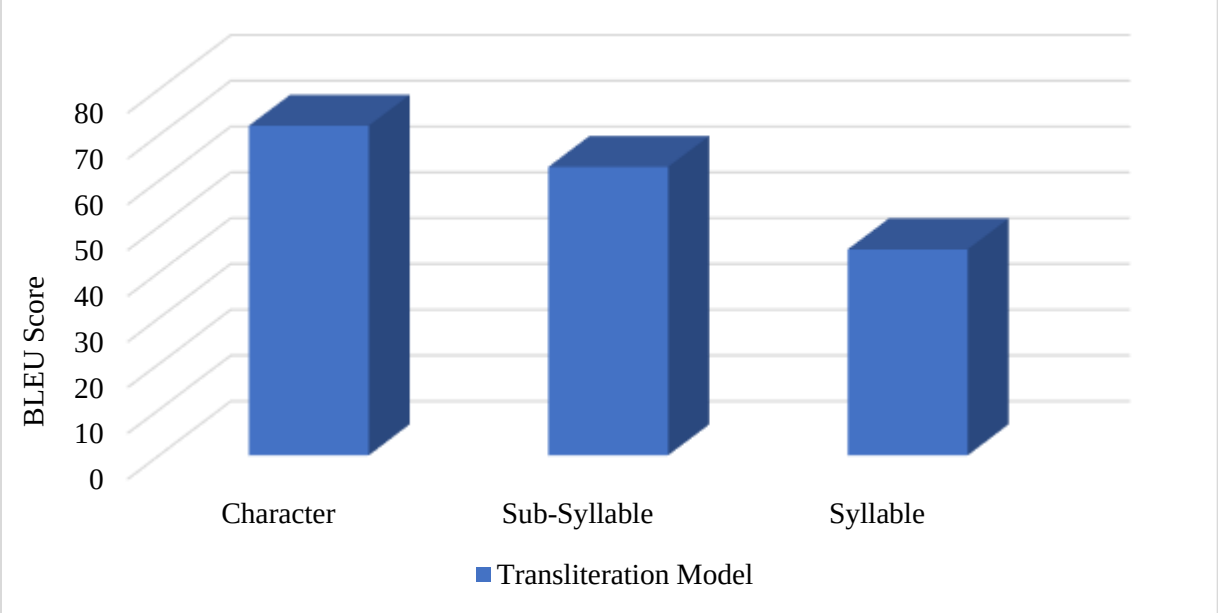

Figure6.2 The Evaluation Results of En-My NET in terms of BLEU for Mix Data

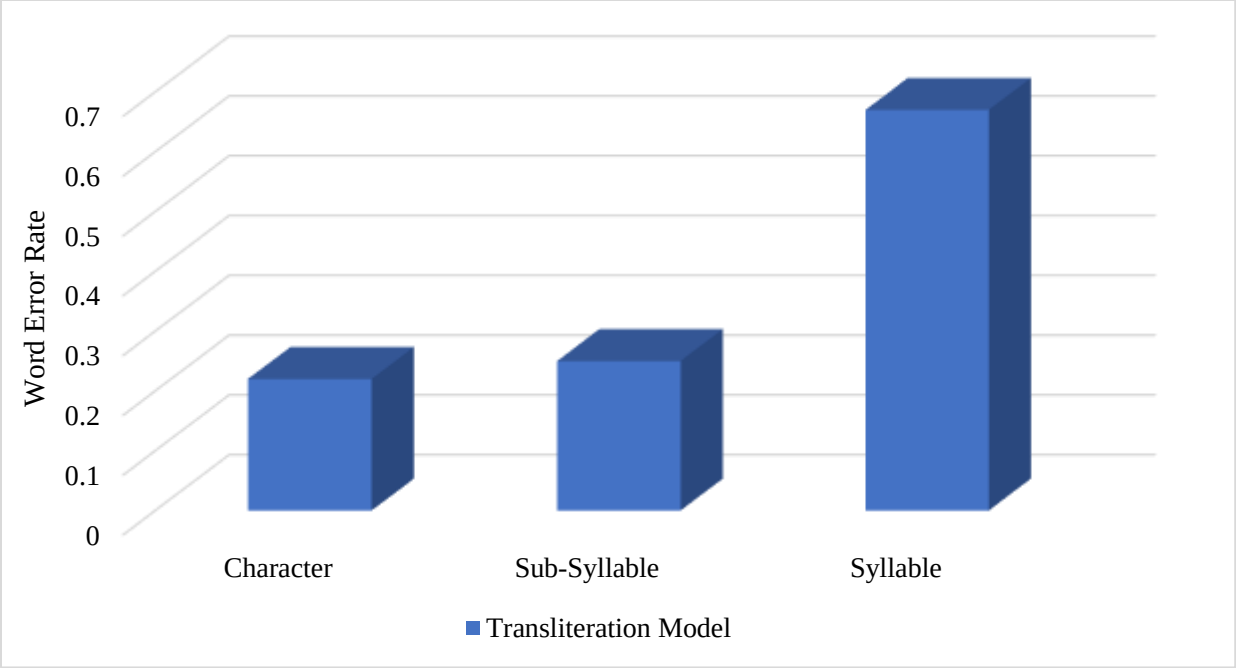

Figure6.3 The Evaluation Results of En-My NET in terms of WER for Mix Data

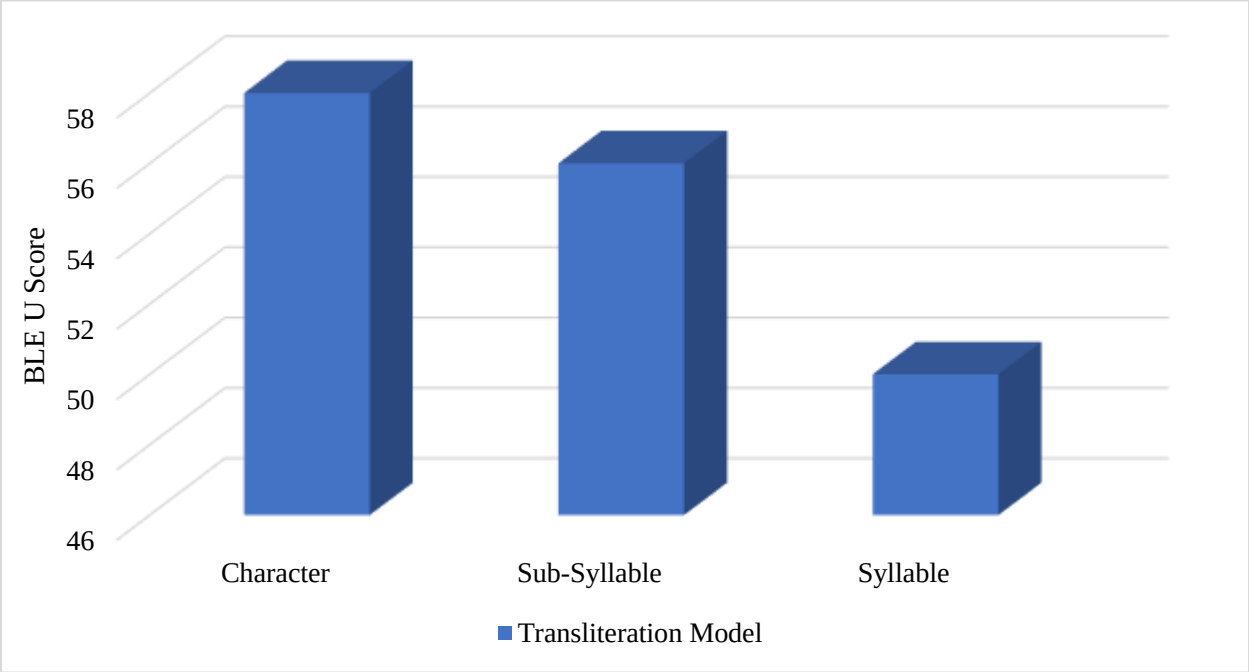

Figure6.4 The Evaluation Results of My-En NET in terms of BLEU for Mix Data

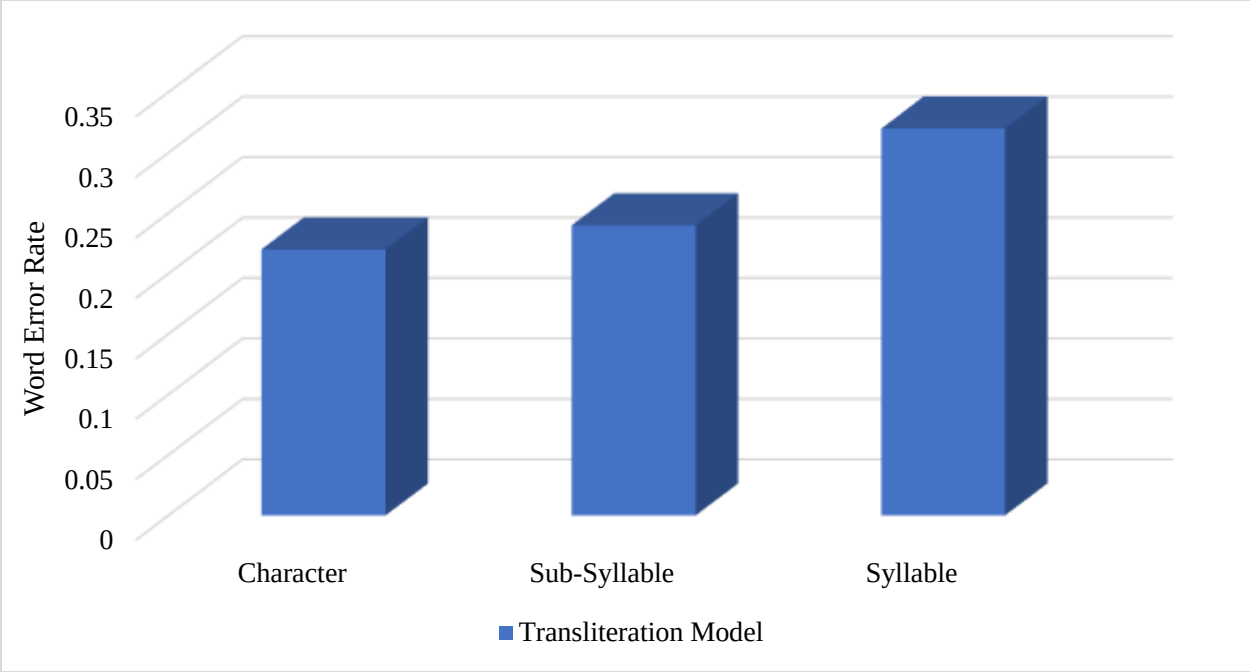

Figure6.5 The Evaluation Results of My-En NET in terms of WER for Mix Data

For En-My and My-En NET, the system performs best at the character level with the highest BLEU score with 72 and lowest WER with 0.22, indicating a high level of accuracy in character-level transliteration. However, at the syllable level, the performance drops significantly, with a lower BLEU score and higher WER, suggesting challenges in accurately transliterating syllables. While character-level transliteration demonstrates high accuracy, efforts should be made to balance accuracy with computational efficiency, especially when dealing with larger datasets or real-time transliteration requirements on mixing dataset.

Table 6.9 System Evaluation Results for Western Data in term of BLEU and WER

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

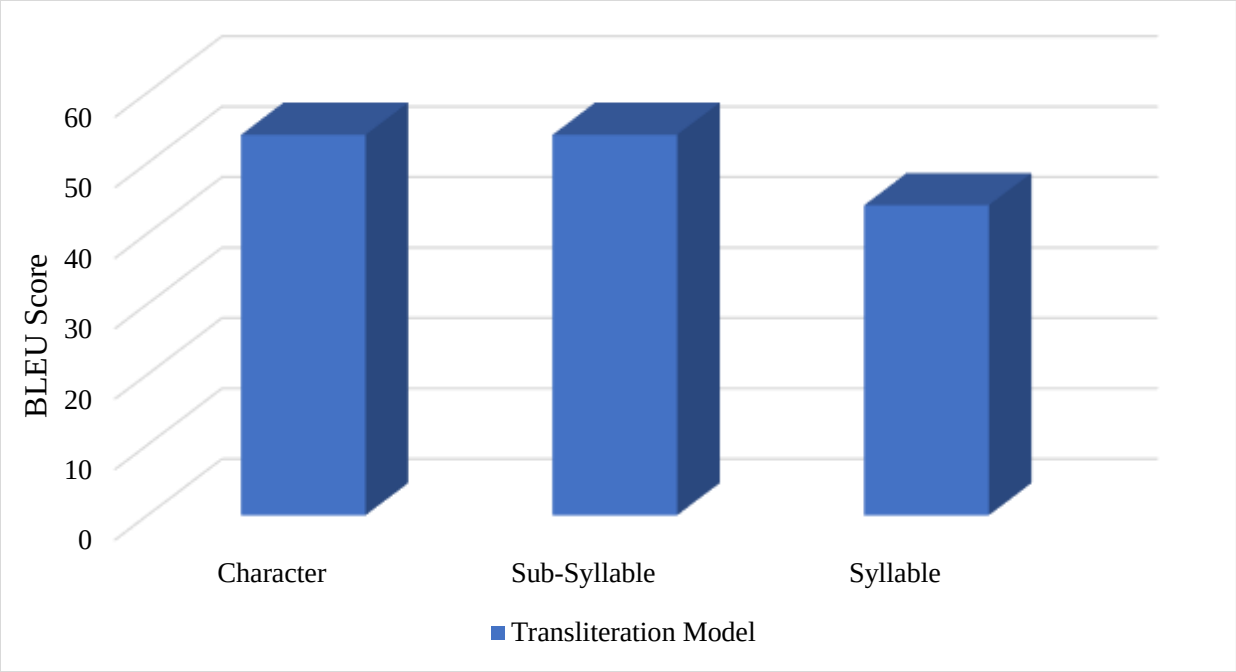

Figure6.6 The Evaluation Results of En-My NET in terms of BLEU for Western Data

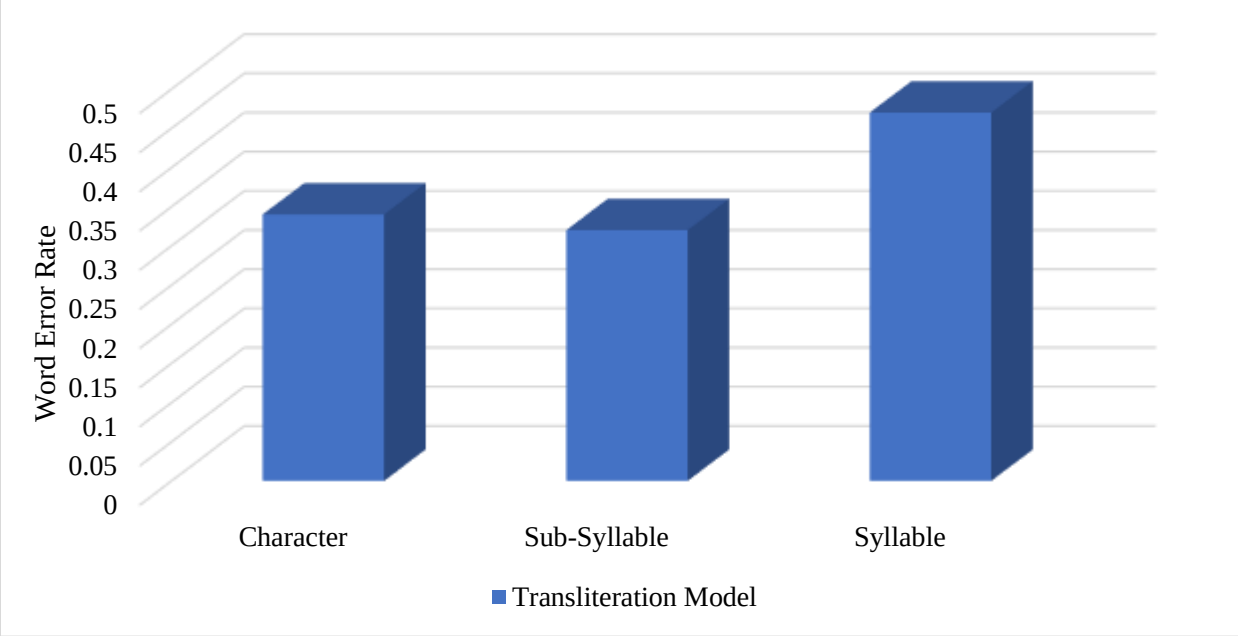

Figure6.7 The Evaluation Results of En-My NET in terms of WER for Western Data

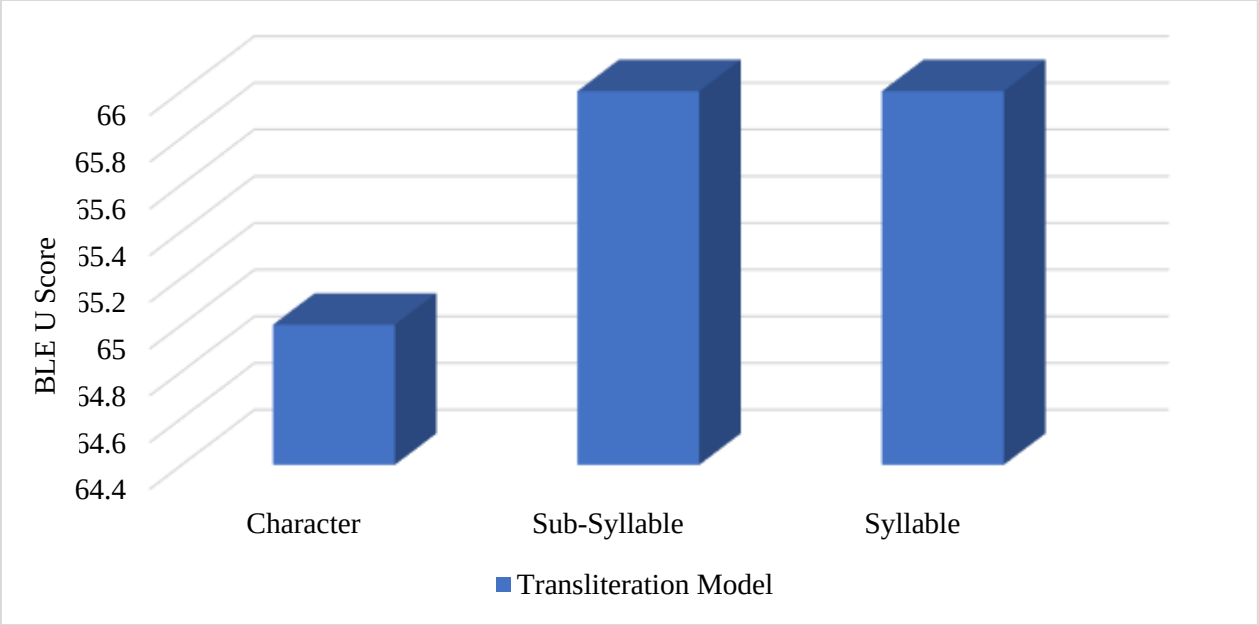

Figure 6.8 The Evaluation Results of My-En NET in terms of BLEU for Western Data

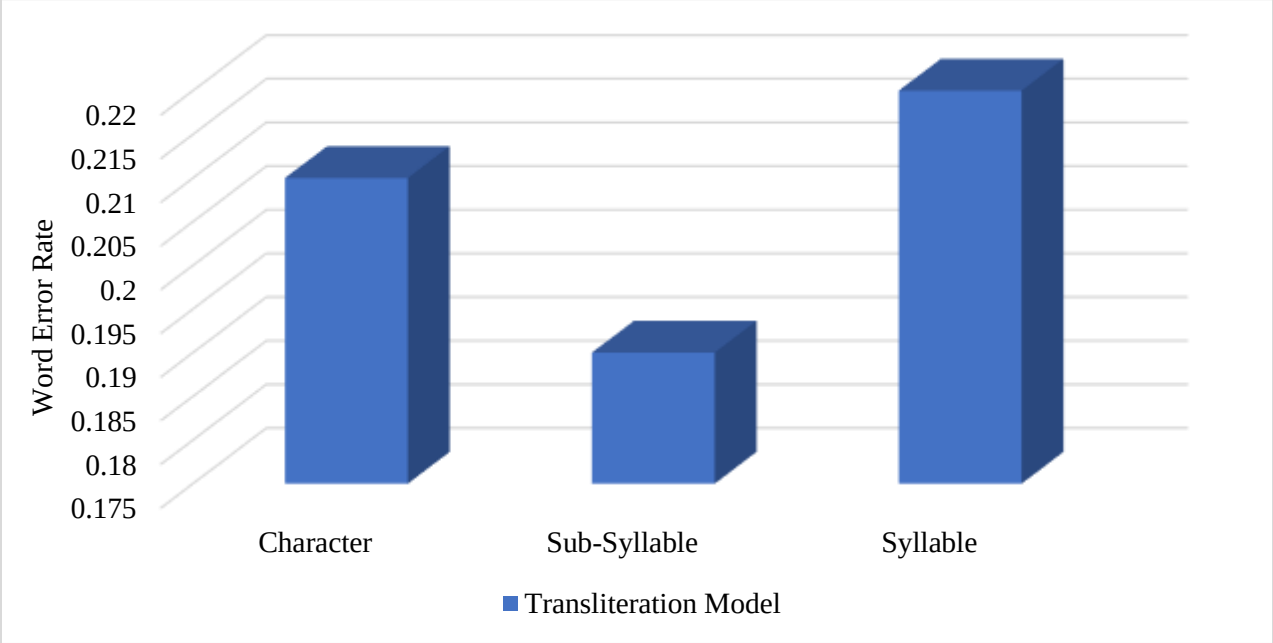

Figure 6.9 The Evaluation Results of My-En NET in terms of WER for Western Data

Using foreign data, particularly with sub-syllable segmentation, yielded competitive results. For En-My (English to Myanmar) transliteration, sub-syllable segmentation achieved a BLEU score of 54 with a low Word Error Rate (WER) of 0.32. Similarly, for My-En (Myanmar to English) transliteration, sub-syllable segmentation achieved a BLEU score of 66 with a WER of 0.19, indicating effective transliteration accuracy. In contrast, character-level segmentation and syllable

segmentation showed lower BLEU scores and slightly higher WER, suggesting that leveraging foreign data with sub-syllable segmentation can enhance transliteration performance.

Table 6.10 System Evaluation Results for Native 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

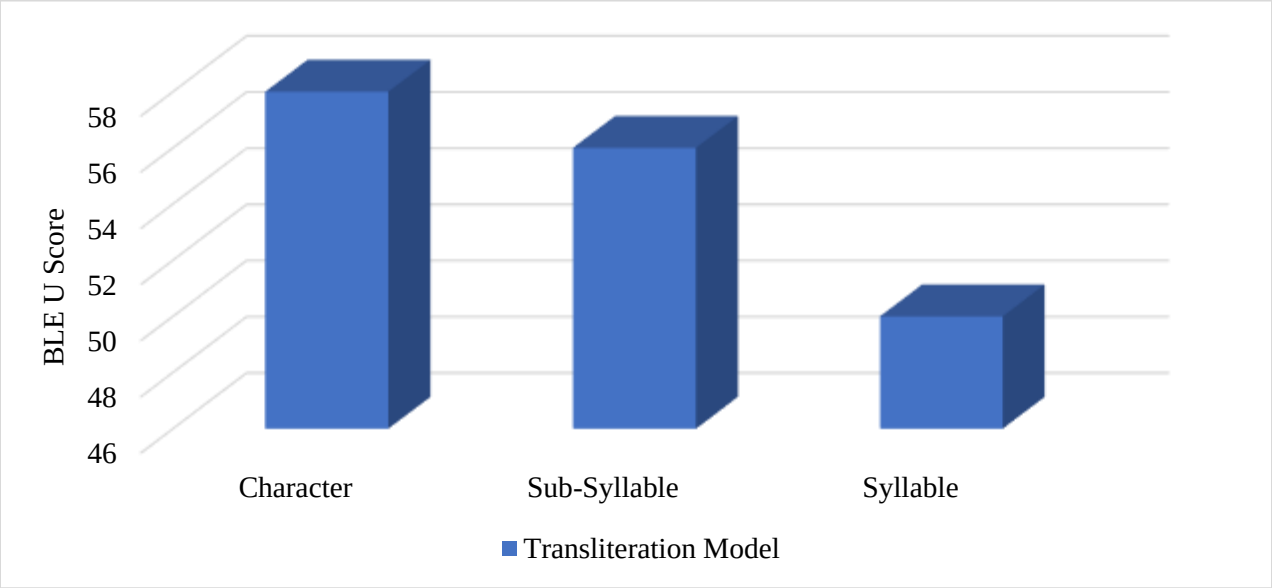

Figure6.10 The Evaluation Results of En-My NET in terms of BLEU for Native Data

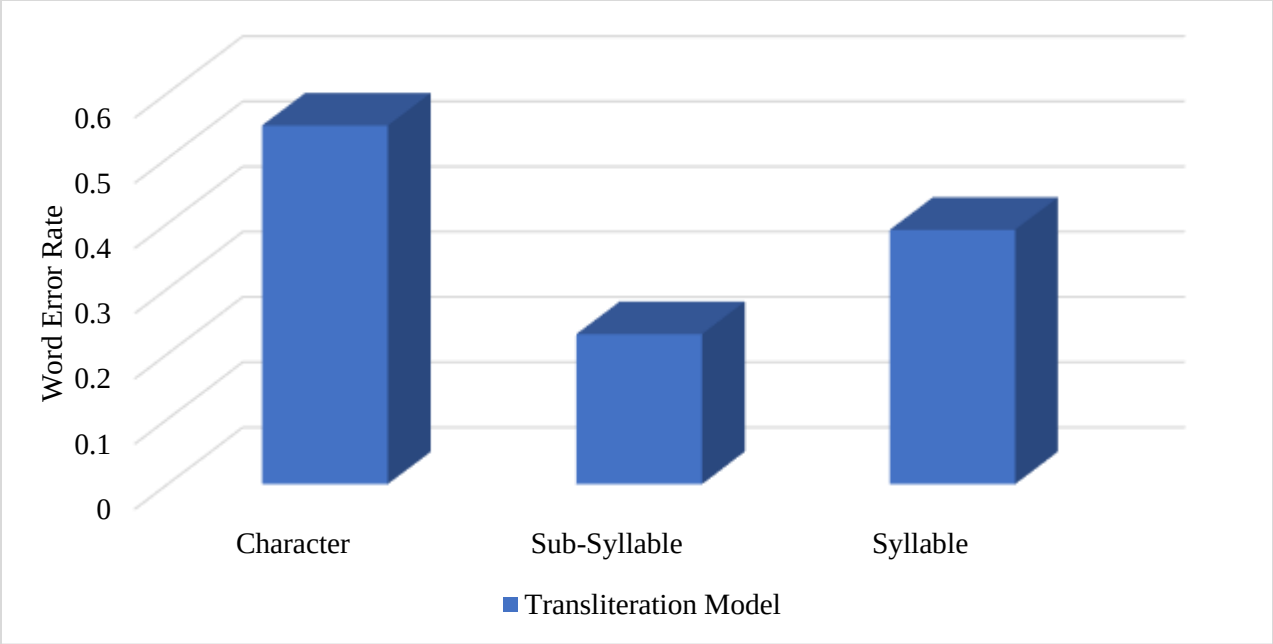

Figure6.11 The Evaluation Results of En-My NET in terms of WER for Native Data

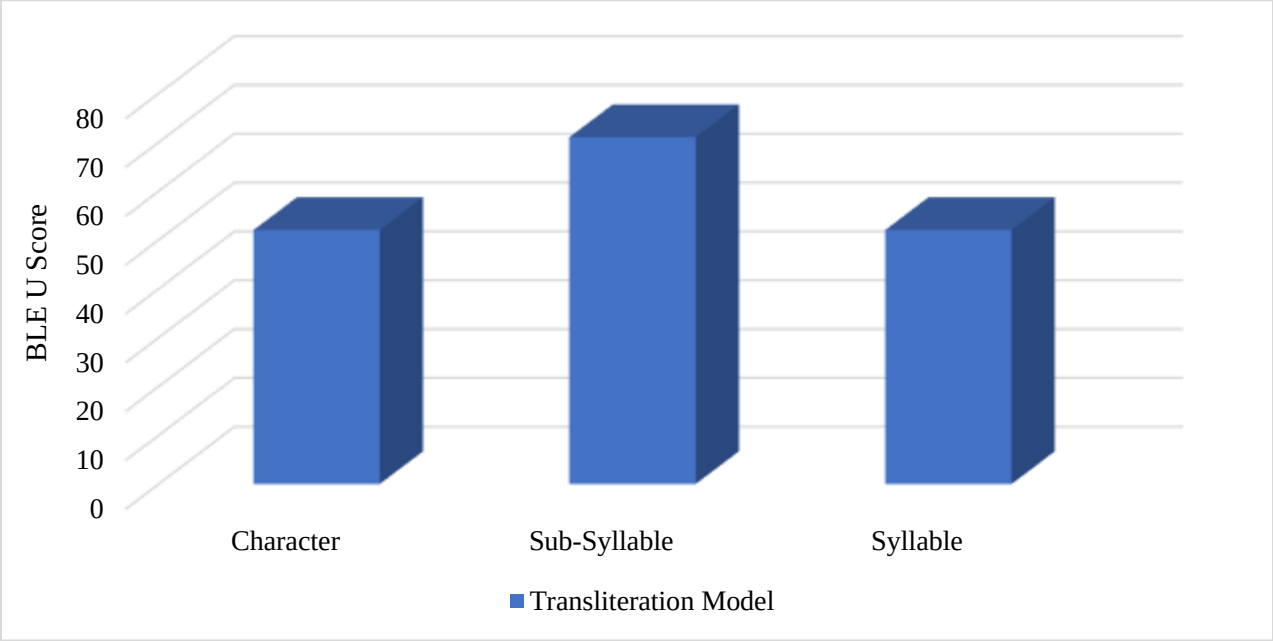

Figure6.12 The Evaluation Results of My-En NET in terms of BLEU for Native Data

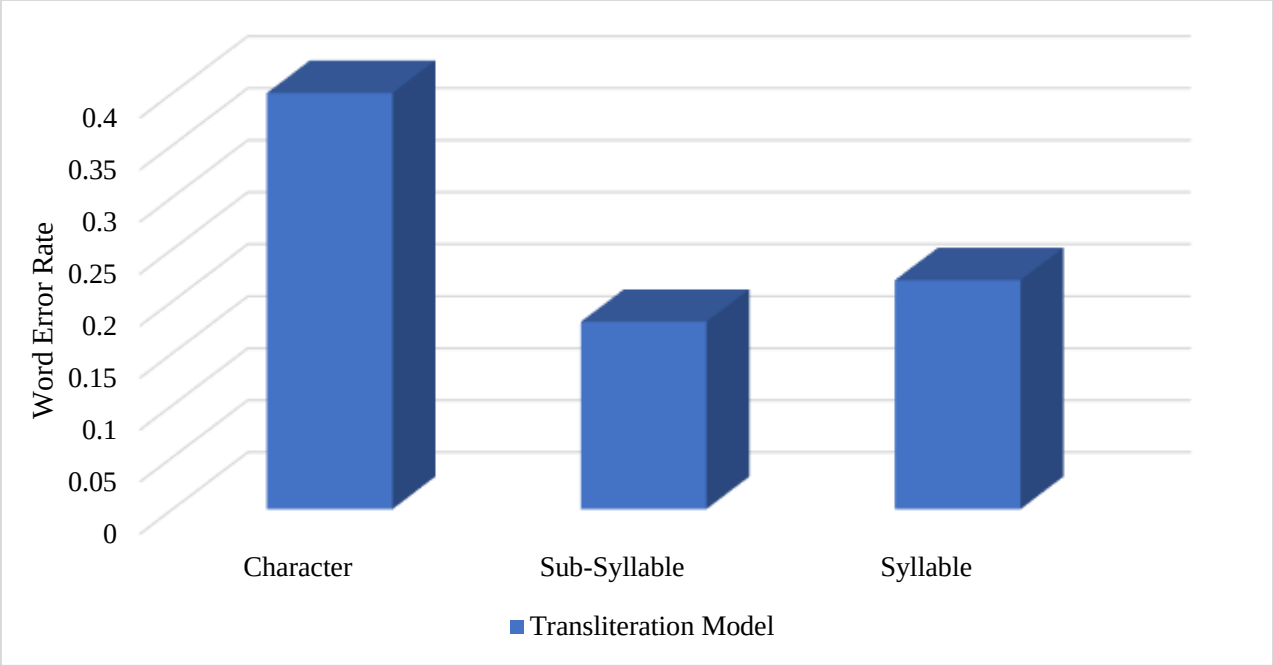

Figure6.13 The Evaluation Results of My-En NET in terms of WER for Native Data

The experimental results demonstrate the impact of data sources and segmentation units on Myanmar-English Named Entity Transliteration. Utilizing native data, particularly with sub-syllable segmentation, led to significant improvements in transliteration quality. For En-My (English to Myanmar) transliteration, sub-syllable segmentation achieved a BLEU score of 56 with a low Word Error Rate (WER) of 0.23. Similarly, for My-En (Myanmar to English) transliteration, sub-syllable segmentation achieved the highest BLEU score of 71 with a WER of 0.18, indicating accurate and effective transliteration. In contrast, character-level segmentation and syllable segmentation yielded lower BLEU scores and higher WER, underscoring the importance of leveraging native data and employing appropriate segmentation techniques for optimal transliteration performance.

6.4 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 6.11, 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 6.11 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.

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 6.11, 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.

Table 6.11 Findings and Discussions on Some Hypotheses Results

Data	**Seg. Units**	**Reference (My)**	**Reference (En)**	**En→My (Hypotheses)**	**My→En (Hypotheses)**
Mix	char.	ကားဒစ်ဖ်	Cardiff	ကာဒီဖ်	Car**difft**
	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

Foreign	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**

6.5 Summary

In this chapter on Myanmar-English Named Entity Transliteration, Experiments are conducted to evaluate the impact of different factors on transliteration performance. The experimental setting involved using various datasets, segmentation units, and preprocessing tools. For preprocessing, the experiments focused on three segmentation units: character-level segmentation, sub-syllable segmentation, and syllable segmentation to identify and segment entities for transliteration. The experimental results revealed interesting insights into the transliteration process. When using character-level segmentation, a BLEU score of 72 is achieved with a Word Error Rate (WER) of 0.22 for En-My (English to Myanmar) transliteration on mixing data, while

My-En (Myanmar to English) transliteration yielded a BLEU score of 71 with a WER of 0.18 on sub-syllable segmentation of the native dataset.

CHAPTER 7

CONCLUSION AND FUTURE WORK

Overall, this chapter serves as a comprehensive overview of the research journey, highlighting both the achievements and the areas for improvement in Myanmar-English named entity transliteration. By addressing the identified limitations and exploring new research directions, the field can continue to advance towards more robust and effective translation systems, benefiting communication and knowledge exchange across languages and cultures.

Myanmar, characterized as a low-resourced language, faces significant challenges in the development of language processing tools such as parallel dictionaries and transliteration systems. The absence of freely available Myanmar-English parallel dictionaries hinders the advancement of named entity transliteration, which remains in its nascent stages, requiring substantial improvements in accuracy and efficiency. To address this gap, the initial focus of this paper was on collecting parallel named entity pairs to lay the groundwork for developing robust transliteration models. The creation of a large-scale Myanmar-English parallel terminology dictionary marks a crucial step forward in supporting natural language processing tasks specific to the Myanmar language. The intricacies of human language transliteration, particularly in the realm of natural language processing, stem from the unique features and nuances inherent in each language. This challenge is not exclusive to Myanmar but resonates across various Asian languages like Indian, Japanese, Thai, and Chinese. Myanmar, with its distinctive writing system and phonology, presents a complex transliteration landscape, necessitating specialized techniques tailored to its linguistic characteristics. Unlike languages with similar writing systems to English, Myanmar's phonological redundancy and stylistic variations pose additional complexities, influencing the accuracy and consistency of transliteration efforts. The complexity of Myanmar-English transliteration extends beyond phonological considerations to encompass orthographic and stylistic variations inherent in the Myanmar script. This divergence from English conventions requires customized transliteration approaches that go beyond straightforward phonetic mappings. The intentional use of unique spellings and borrowed word adaptations adds layers of intricacy to the transliteration process, contributing to irregularities and challenges in achieving accurate and intuitive transliteration results. As such, the development of effective Myanmar-English transliteration models demands a deep

understanding of linguistic nuances, orthographic intricacies, and the interplay between phonetic representations and visual aesthetics in both languages.

In this study, three NET models are trained using Transformer architecture for transliteration tasks. The first model is character-based, the second is Myanmar sub-syllable-based, and the third is Myanmar syllable-based (baseline). These models were trained on a mixture of native data, foreign data, and a combination of both. Our aim was to evaluate their performance in transliterating English to Myanmar (En-My) and Myanmar to English (My-En). The experimental results revealed interesting findings. The BLEU (Bilingual Evaluation Understudy) and WER (Word Error Rate) metrics showed significant improvements for the Myanmar character-based NET model in the En-My transliteration task. This indicates that the character-level approach better captures the nuances and phonetic details required for accurate transliteration from English to Myanmar script. On the other hand, the Myanmar sub-syllable-based NET model exhibited superior performance in the My-En transliteration task, as evidenced by higher BLEU and lower WER scores compared to the other models. These results underscore the importance of considering linguistic characteristics and data representation methods when designing transliteration models. While character-level models excel in certain tasks, sub-syllable-based approaches may be more suitable for different language pairs and directions of transliteration, highlighting the need for tailored solutions in multilingual NLP tasks.

7.1 Advantages of the System

There are several advantages to having a Myanmar English Named Entity Transliteration System. Firstly, it facilitates seamless communication between Myanmar speakers and English speakers by accurately converting names and entities from one language to another. This is particularly useful in cross-cultural interactions, business transactions, and academic exchanges.
Secondly, such a system enhances information retrieval and search functionalities by ensuring that names and entities are correctly transliterated and indexed. This improves the accuracy of search results and makes information more accessible across languages. Additionally, a Myanmar English Named Entity Transliteration System promotes cultural understanding and appreciation by preserving the authenticity and pronunciation of Myanmar names and entities in English texts. It helps maintain the integrity of cultural identities in global contexts. Overall, the system streamlines

communication, improves information retrieval, and fosters cultural preservation and understanding, making it a valuable tool in various domains.

7.2 Limitations of the System

One of the primary limitations is the unavailability of training data for Pali words. Named entities can be diverse and context-dependent, making it challenging to curate a comprehensive dataset that covers all possible variations and entities accurately. Limited data could lead to suboptimal performance, especially for less common or specialized entities. Transliterating Myanmar named entities into English (and vice versa) can be inherently ambiguous due to variations in pronunciation, spelling, and structure. For example, the same Myanmar entity could be transliterated into multiple English variations based on context, dialect, or personal preference. Handling such variability effectively remains a challenge. Moreover, Transformer models rely on pre-defined vocabularies, which may not cover all named entities encountered in real-world scenarios. Out-of-vocabulary entities can lead to transliteration errors or inaccurate mappings, especially for rare or newly coined terms.

7.3 Future Works

Future enhancements in the Myanmar-English Named Entity Transliteration System using Transformer models can focus on several key areas to improve accuracy, coverage, and usability. One avenue for advancement is the exploration of advanced data augmentation techniques. By incorporating diverse datasets from social media, domain-specific texts, and parallel corpora, the system can expand its vocabulary and better handle a wide range of named entities, including rare or specialized terms. Another promising direction is the adoption of hybrid segmentation strategies. Combining character-level, sub-syllable, and syllable-level segmentation dynamically based on contextual cues and linguistic patterns can enhance the system's adaptability and transliteration accuracy. Adaptive segmentation approaches would enable the system to effectively handle complex named entities with varying linguistic structures and contextual nuances.

Lastly, the development of domain-specific models tailored to specialized industries or domains can further enhance transliteration accuracy and relevance. Domain adaptation techniques, specialized vocabularies, and domain-specific training data can ensure that the system accurately

transliterates domain-specific named entities, thereby catering to the diverse needs of users across different sectors. By pursuing these future directions, the Myanmar-English Named Entity Transliteration System can evolve into a more robust, accurate, and user-centric solution, meeting the demands of real-world transliteration tasks effectively.

Author's Publications

P [1] **Aye Myat Mon**, Khin Mar Soe. "Clustering Analogous Words in Myanmar Language Using Word Embedding Model". Proceedings of the 17th International Conference on Computer Applications (ICCA 2019), pages 154-159,Yangon, Myanmar on February 27-28, 2019.

P [2] **Aye Myat Mon**, Chenchen Ding, Hour Kaing, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. "A Myanmar (Burmese)-English Named Entity Transliteration Dictionary". Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 2973–2976,Marseille,France, 11–16 May 2020.

P [3] **Aye Myat Mon**, and Khin Mar Soe. "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 Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 38-43. IEEE, 2020.

P [4] **Aye Myat Mon** and Khin Mar Soe. "Neural Named Entity Transliteration for Myanmar to English Language Pair." The 13th International Conference on Future Computer and Communication, (ICFCC 2021), Singapore (pp. 70-74), WCSE2021.

P [5] Benjamin Marie, Hour Kaing, **Aye Myat Mon**, Chenchen Ding, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita: "Supervised and Unsupervised Machine Translation for Myanmar-English and Khmer-English". WAT@EMNLP-IJCNLP 2019: 68-75

P [6] **Aye Myat Mon** and Khin Mar Soe. "CROSS LINGUISTIC NAMED ENTITY TRANSLITERATION FOR BURMESE (MYANMAR) AND ENGLISH USING TRANSFORMER MODEL" Indian Journal of Computer Science and Engineering, VOLUME 14 ISSUE 1 Jan-Feb 2024 (pp. 120-131).

P [7] **Aye Myat Mon**, Mya Than Hnin, and Su Su Win "A Case Study on Myanmar-English Named Entity Dictionary Enhancement" Journal of Information Technology, Research and Innovation (JITRI) 2023, Vol-3, Issue-2 (pp. 70-77)

Bibliography

[1] AbdulJaleel, N., & Larkey, L. S. (2003, November). Statistical transliteration for English-Arabic cross language information retrieval. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 139-146).

[2] Al-Onaizan, Y., & Knight, K. (2002, July). Translating named entities using monolingual and bilingual resources. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 400-408).

[3] Ameur, M. S. H., Meziane, F., & Guessoum, A. (2019). ANETAC: Arabic named entity transliteration and classification dataset. arXiv preprint arXiv:1907.03110.

[4] Arnold, D., Balkan, L., Humphreys, R. L., Meijer, S., & Sadler, L. Machine translation: An introductory guide.

[5] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014 Sep 1.

[6] Chang, C. (2003). "High-interest loans": The phonology of English loanword adaptation in Burmese.

[7] Choi, K., Isahara, H., & Oh, J. (2011). A comparison of different machine transliteration models. arXiv e-prints, arXiv-1110.

[8] Ding, C., Pa, W. P., Utiyama, M., & 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.

[9] Ding, C., Utiyama, M., and Sumita, Eiichiro. (2018). NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. ACM TALLIP, Vol. 18, Issue 2, Article No. 17.

[10] 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: Syllablebased Tokenization and Part-of-Speech Tagging. ACM TALLIP, Vol. 19, Issue 1, Article No. 5.

[11] 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 TALLIP, Vol.19, Issue 3, Article No. 40.

[12] Ding, C., Pa, W. P., Utiyama, M., and Sumita, E. (2017).Burmese (Myanmar) name romanization: A subsyllabic segmentation scheme for statistical solutions.In Proc. of PACLIC, pp. 191—202.

[13] Ding, Chenchen. "Transliteration of Foreign Words in Burmese." arXiv preprint arXiv:2110.03163 (2021).

[14] Divay, M., & Vitale, A. J. (1997). Algorithms for grapheme-phoneme translation for English and French: Applications for database searches and speech synthesis. Computational linguistics, 23(4), 495-523.

[15] Ekbal, A., Naskar, S. K., & Bandyopadhyay, S. (2006, July). A modified joint source-channel model for transliteration. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions (pp. 191-198).

[16] Eisenstein, Jacob. Introduction to natural language processing. MIT press, 2019.

[17] Gao, W., Wong, K. F., & Lam, W. (2005). Improving transliteration with precise alignment of phoneme chunks and using contextual features. In Information Retrieval Technology: Asia Information Retrieval Symposium, AIRS 2004, Beijing, China, October 18-20, 2004. Revised Selected Papers 1 (pp. 106-117). Springer Berlin Heidelberg.

[18] Jung, S. Y., Hong, S., & Paek, E. (2000). An english to korean transliteration model of extended markov window. In COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics.

[19] Jiang R, Banchs RE, Li H. Evaluating and combining name entity recognition systems. InProceedings of the Sixth Named Entity Workshop 2016 Aug (pp. 21-27).

[20] Kang, B. J., & Choi, K. S. (2000, October). English-Korean automatic transliteration/backtransliteration system and character alignment. In Proc. ACL (pp. 17-18).

[21] Knight, K., & Graehl, J. (1997). Machine transliteration. arXiv preprint cmp-lg/9704003.

[22] Knight, K. and Graehl, J. (1998), “Machine transliteration”, in proceedings of the 35th annual meetings of the Association for Computational Linguistics, Madrin, Spain, pp. 128-135.

[23] Kumai H, Sagawa H, Morimoto Y. “NTCIR-7 Patent Translation Experiments at Hitachi”. InNTCIR 2008.

[24] Klakow D, Peters J. “Testing the correlation of word error rate and perplexity”. Speech Communication. 2002 Sep 1;38(1-2):19-28.

[25] 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).

[26] 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.

[27] Li, H., Zhang, M., & Su, J. (2004, July). A joint source-channel model for machine transliteration. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04) (pp. 159-166).

[28] Li, H., Kumaran, A., Pervouchine, V., & 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).

[29] Liu, L., Finch, A., Utiyama, M., & Sumita, E. (2020). Agreement on target-bidirectional recurrent neural networks for sequence-to-sequence learning. Journal of Artificial Intelligence Research, 67, 581-606.

[30] Li, H., Sim, K. C., Kuo, J. S., & Dong, M. (2007, June). Semantic transliteration of personal names. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 120-127).

[31] Naing, H. M. S., Thu, Y. K., Pa, W. P., Kato, H., Finch, A., Sumita, E., & 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).

[32] Malik, M. A. (2006, July). Punjabi machine transliteration. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 1137-1144).

[33] Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT press.

[34] Meng, H. M., Lo, W. K., Chen, B., & Tang, K. (2001, December). Generating phonetic cognates to handle named entities in English-Chinese cross-language spoken document retrieval. In IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01. (pp. 311-314). IEEE.

[35] Mo, H. M., & Soe, K. M. (2019). Syllable-Based Neural Named Entity Recognition for Myanmar Language. arXiv preprint arXiv:1903.04739.

[36] Mon, A. M., & 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.

[37] Och, F. J., and Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1), 19-51.

[38] Och, F. J. (2003). Minimum error rate training in statistical machine translation. In Proc. of ACL Vol. 1, pp.160—167.

[39] Pingali, P., Ganesh, S., Yella, S., & Varma, V. (2008). Statistical transliteration for cross language information retrieval using HMM alignment model and CRF. In Proceedings of the 2nd workshop on cross lingual information access (CLIA) addressing the information need of multilingual societies.

[40] Riza, H., Purwoadi, M., Gunarso, Uliniansyah, T., Aw AiTi, Aljunied, S. M., Luong C. M., Vu T. T., Nguyen P.T., Chea, V., Sun, R., Sam, S., Seng, S., Soe, K. M.,Nwet, K. T., Utiyama, M., and Ding, C. (2016) Introduction of the Asian Language Treebank. In Proc. of OCOCOSDA, pp. 1—6.

[41] Rosca M, Breuel T. Sequence-to-sequence neural network models for transliteration. arXiv preprint arXiv:1610.09565. 2016 Oct 29.

[42] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Advances in neural information processing systems. 2014;27.

[43] Stalls, B. G., & Knight, K. (1998). Translating names and technical terms in Arabic text. In Computational Approaches to Semitic Languages.

[44] Stalls, B. and Knight, K. (1998), “Translating Names and Technical Terms in Arabic Text”, in proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages, Montreal, Canada, pp. 34-41.

[45] Sherif, T., & Kondrak, G. (2007, June). Bootstrapping a stochastic transducer for Arabic-English transliteration extraction. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 864-871).

[46] Swe, T. T., & Htay, H. H. A Hybrid Method for Myanmar Named Entity Identification and Transliteration to English (Doctoral dissertation, MERAL Portal).

[47] Sin, Y. M. S., Soe, K. M., and Htwe, K. Y. (2018). Large scale Myanmar to English neural machine translation system. In Proc. of GCCE, pp. 464—465.

[48] Shao Y, Hardmeier C, Nivre J. Multilingual named entity recognition using hybrid neural networks. InThe sixth Swedish language technology conference (SLTC) 2016.

[49] Su, Yuanhang, and C-C. Jay Kuo. "On extended long short-term memory and dependent bidirectional recurrent neural network." Neurocomputing 356 (2019): 151-161.

[50] 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).

[51] Su, Yuanhang, Yuzhong Huang, and C-C. Jay Kuo. "Dependent bidirectional RNN with extended-long short-term memory." (2018).

[52] 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.

[53] Thurmair, G. (2005). Hybrid architectures for machine translation systems. Language resources and evaluation, 39, 91-108.

[54] 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).

[55] Wu, C. K., Wang, Y. C., & Tsai, R. T. H. (2012, July). English-Korean named entity transliteration using substring alignment and re-ranking methods. In Proceedings of the 4th Named Entity Workshop (NEWS) 2012 (pp. 57-60).

[56] Wang, Y. C., & Tsai, R. T. H. (2011, November). English-korean named entity transliteration using statistical substring-based and rule-based approaches. In Proceedings of the 3rd Named Entities Workshop (NEWS 2011) (pp. 32-35).

[57] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.

[58] Virga, P., & Khudanpur, S. (2003, July). Transliteration of proper names in cross-lingual information retrieval. In Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition (pp. 57-64).

[59] Yadav, M., Kumar, I., & Kumar, A. (2023, March). Different Models of Transliteration-A Comprehensive Review. In 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA) (pp. 356-363). IEEE.

[60] Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. (2019). Grapheme-to-phoneme conversion with convolutional neural networks. Applied Sciences, 9(6), 1143.

[61] https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b

[62] https://2020.myanmarexam.org/

[63] https://en.wikipedia.org › wiki › Burmese_language

[64] Wikipedia:Naming conventions (Burmese)

List of Acronyms

NLP	Natural Language Processing
NE	Named Entity
IDE	Integrated Development Environment
NLU	Natural Language Understanding
NLG	Natural Language Generation
NLA	Natural Language Acquisition
NMT	Neural Machine Translation
DMNT	Deep Neural Machine Translation
SMT	Statistical Machine Translation
MT	Machine Translation
RTL	Right to Left
LTR	Left to Right
G2P	Grapheme to Phoneme
TTS	Text to Speech
ASR	Automatic Speech Recognition
OOV	Out of Vocabulary
NER	Named Entity Recognition
ML	Machine Learning
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BLEU	Bilingual Evaluation Understudy
WER	Word Error Rate
OCR	Optical Character Recognition
WFSM	Weighted Finite State Machine
WSJ	Wall Street Journal

CMU	Carnegie Mellon University
EM	Expectation Maximization
NEWs	Named Entities Workshop
NICT	National Institute of Information and Communications Techonology
LDC	Linguistic Data Consortium
CRF	Conditional Random Field
WFSA	Weighted Finite State Acceptors
FSM	Finite State Machine
ALT	Asia Language Treebank
CLIR	Cross Language Information Retrieval