

Tracing the Evolution of Machine Translation: A Journey through the Myanmar (Burmese)-Wa (sub-group of the Austro-Asiatic language) Corpus

Florance Yune*, Khin Mar Soe

Natural Language Processing Lab, University of Computer Studies (Lecturer), Yangon, Myanmar

ARTICLE INFO

Article history:

Received: 02 December, 2023

Revised: 06 January, 2024

Accepted: 06 January, 2024

Online: 20 January, 2024

Keywords:

Machine Translation (MT)

Low-resource Languages

Myanmar-Wa Corpus

Language Preservation

Transfer Learning

Long Short-Term Memory

Recurrent Neural Network

Pivot-Based SMT

ABSTRACT

Machine Translation (MT) has come a long way toward reducing linguistic gaps. However, its progress in efficiently handling low-resource languages—such as the Wa language in the Myanmar-Wa corpus—has not received enough attention. This study begins with a thorough investigation of the historical development of MT systems, painstakingly following their development against the complex background of the Myanmar-Wa language region. Using an interdisciplinary methodology that integrates linguistics, technology, and culture, this investigation reveals the transformative journey of Machine Translation (MT) in its pursuit of overcoming linguistic barriers. It offers a thorough study that clarifies the opportunities and limitations present in MT's progress. More broadly, by clarifying the complex relationship between technology and linguistic diversity, our work not only advances our understanding of MT's evolutionary history but also supports the conservation of endangered languages, like Wa language. The research's conclusions have implications that go beyond machine translation to the larger conversation about language preservation and how technology development coexists harmoniously. Notably, this paper is an extension of work originally presented in "2023 IEEE Conference on Computer Applications (ICCA)", acknowledging its foundation and presenting substantial advancements.

1. Introduction

The critical role played by machine translation (MT) in eliminating linguistic barriers within the context of an expanding era of globalization cannot be overstated [1]. Essential instruments for fostering international collaboration, knowledge dissemination, and cross-cultural exchanges, MT systems have evolved to facilitate communication across linguistic and cultural boundaries. This paper aims to investigate the historical development of MT systems, with a specific focus on their adaptation to the intricate linguistic environment of the Myanmar-Wa corpus. The necessity for precise and efficient language translation is underscored by the accelerated pace of globalization, characterized by the seamless flow of ideas, information, and business across international borders [2]. Propelled by advancements in artificial intelligence and computational linguistics, machine translation has emerged as a pivotal element in the global communication paradigm. It has not only surmounted the constraints of traditional translation but has also enabled the bridging of languages previously considered challenging due to limited resources, as exemplified by the Wa language within the Myanmar-Wa corpus.

The ensuing methods are structured to fulfill the objectives of this study. Commencing with a historical overview of machine translation, encompassing its development and initial challenges, the analysis subsequently delves into the distinctive attributes of the Myanmar-Wa corpus, including its cultural significance and linguistic characteristics, along with the complexities associated with data acquisition. This research article proceeds to scrutinize early attempts at applying MT to the Myanmar-Wa corpus, elucidating their deficiencies and limitations. Moreover, it investigates innovative approaches and technological advancements that have enhanced MT for this specific language combination.

The subsequent section of the study involves linguistic analysis, dissecting the unique characteristics of the Wa language and examining the metrics and procedures employed to evaluate the efficacy of the MT system. Emphasizing the pivotal role that improved MT plays in cultural preservation and the broader objective of language revitalization, this analysis focuses on the practical applications and consequences of enhanced MT for the Myanmar-Wa language pair. Finally, the system's investigation concludes with an overview of major findings, a discussion of their

*Corresponding Author: Florance Yune, floranceyune@ucsy.edu.mm

implications, and concluding reflections on the dynamic field of machine translation and its critical role in bridging linguistic gaps.

The study begins with the acknowledgment that this paper is an extension of the work originally presented in the conference paper, "Myanmar-Wa Machine Translation based on LSTM-based Deep Learning Encoder-Decoder Model" [3].

2. Related Works

The evolution of machine translation (MT) stands as an enduring quest to surmount linguistic barriers, a journey that was inaugurated by the visionary insights of Warren Weaver. In [4] the author presented groundbreaking 1949 memo, Weaver introduced the concept of computer-assisted language translation, marking the genesis of MT as a distinct area of study and laying the foundation for subsequent developments. Early MT initiatives, notably the Georgetown-IBM experiment in the 1950s, played a pivotal role in establishing the rudiments of MT technology. Despite their nascent status, these early algorithms encountered formidable challenges, often yielding imprecise and unnatural translations. The intricacies of human language, encompassing idiosyncrasies, context-dependency, and cultural nuances, presented a significant hurdle. As linguistic theories matured and computational power increased, MT systems began to exhibit promise.

A pivotal advancement in machine translation materialized with the advent of rule-based MT in the 1960s and 1970s. These systems employed grammatical analysis and language rules to generate translations, exemplified by the Systran system initially designed for diplomatic and military communication [5].

The landscape of MT underwent transformative shifts with the introduction of statistical machine translation (SMT) systems in the 1980s, as seen in IBM's *Candide* and the European Union's EUROTRA project, as presented in [6]. Utilizing statistical models and extensive multilingual corpora, SMT systems revolutionized translation paradigms, ushering in a data-driven methodology that significantly improved translation accuracy and fluency. The 21st century heralded a new era in machine translation with the rise of neural machine translation (NMT) models, including recurrent neural networks (RNNs) and transformer models like the Transformer and BERT. Fueled by deep learning, these models exhibited remarkable proficiency in recognizing idiomatic expressions, context, and intricate linguistic patterns, with further enhancements through the integration of attention mechanisms [7].

In tandem with these advancements, MT systems grapple with challenges, particularly in handling low-resource languages like Wa in the Myanmar-Wa corpus. The absence of extensive bilingual corpora for low-resource languages hampers the development of precise and contextually appropriate translations. Additional complexities arise from linguistic nuances, such as tone systems, restricted written forms, and complex morphology. The lack of established grammatical rules and linguistic resources in these languages poses a challenge to the effective generalization of machine translation systems.

In response to these challenges, the field of machine translation (MT) has evolved with innovations such as transfer learning approaches, cross-lingual pre-training, and adaptations of existing models to low-resource languages. In [8] the author introduced a notable contribution to the field in "Cross-lingual Language Model Pretraining" (*Advances in Neural Information Processing Systems*, 2019). This concept emphasizes the pretraining of

language models in a cross-lingual manner, enhancing their ability to transfer knowledge across languages. To address the issues faced by low-resource languages and enhance inclusivity, researchers are exploring novel approaches. This study traces the historical development of machine translation (MT) systems, emphasizing their adaptation to the Myanmar-Wa corpus. Through an examination of technological advancements, the study sheds light on strategies and innovations aimed at achieving efficient Wa language translation. Crucially, this work extends beyond the realm of machine translation, contributing to broader domains such as linguistics and language preservation. By scrutinizing the challenges and solutions in adapting MT to the Myanmar-Wa corpus, the study contributes valuable insights to the growing body of information on underrepresented languages. This effort plays a vital role in ensuring the survival and resurgence of languages in a globalized environment [9].

Furthermore, in [10], the authors introduced a novel approach to NMT by incorporating syntax-directed attention mechanisms. This research, presented at the AAAI conference on artificial intelligence, delves into the utilization of syntactic structures to enhance the attention mechanism in NMT systems. By integrating syntax-directed attention, the researchers aimed to improve the model's comprehension of sentence structures and syntactic relationships, potentially leading to more accurate and contextually relevant translations. In [10] the author found the valuable insights to the ongoing advancements in NMT, particularly in addressing syntactic complexities for improved translation quality.

3. Myanmar-Wa Corpus

The extraordinary linguistic tapestry that is the Myanmar-Wa corpus derives its significance from the merging of two separate languages, Wa and Myanmar. This corpus provides insights into the difficulties and advancements in machine translation in addition to serving as a monument to linguistic diversity [11].

The Wa language, which is spoken by the Wa people, an ethnic group that live in the border regions of China and Myanmar, is stored in the Myanmar-Wa corpus, which is located inside this complicated linguistic terrain. MT systems face significant hurdles due to the unique linguistic complexities of the Myanmar-Wa corpus, which include a limited written form and a rich tone system [12]. These difficulties include maintaining cultural context, grammatical structures, and subtle semantic differences—all essential components of accurate translation.

3.1. Myanmar Language

Spoken largely in Myanmar, the Burmese language, also called the Myanmar language, is linguistically rich and culturally significant. It has distinct linguistic characteristics that present intriguing opportunities and problems for machine translation systems. This section gives a general introduction to Burmese, emphasizing its unique vowels and consonants, and discusses its use in the larger field of machine translation. Three basic character groups—vowels, consonants, and medial—combine to form Myanmar writing, and each adds to the language's complex structure. In addition, there are 33 primary consonants in the Myanmar language, each with a unique phoneme.

The tonal and analytical characteristics of the Myanmar language contribute to its linguistic complexity [13]. Its script,

which is based on the Indian (Brahmi) prototype, has changed over the course of a millennium to reflect Myanmar's rich linguistic past. Significantly, the Myanmar language has functioned as a symbol of cultural identity within the nation as well as a means of communication. Since it serves as the foundation for translation into the Wa language, it is imperative to comprehend its subtleties in the context of machine translation.

A wide variety of vowels and consonants make up Burmese Voiced and voiceless nasals (m, n, ɲ, ŋ), stops and affricates (b, d, dʒ, ɡ, p, t, tʃ, k, ʔ), and fricatives (ð, z, θ, s, ʃ, sʰ, h) are among its consonantal inventory. Burmese also has voiceless approximants (l̥, ɱ) and voiced approximants (l, j, w). Burmese consonants are peculiar compared to those in many other languages, which presents an interesting linguistic problem for machine translation systems.

Example Consonants:

Voiced Nasals: m (မ), n(န)

Voiceless Stops: p (ပ), t (တ), k (က)

Fricatives: s (စ), h (ဟ)

Voiced Approximants: w (ဝ), j (့)

There are several vowel sounds in the Burmese vowel system, such as diphthongs and monophthongs. Along with a variety of low and mid vowels, the monophthongs include the high vowels /i/ and /u/, as well as the middle vowels /ɪ/ and /ʊ/. These vowel sounds can be combined to generate diphthongs, which give languages distinctive phonetic patterns and increased complexity.

Example Vowels:

High Vowels: /i/ (အိ), /u/ (ဥ)

Central Vowels: /ɪ/ (အိး), /ʊ/ (ဥ့)

Low Vowels: /a/ (အ), /ɛ/ (ဧ)

3.2. Wa Language

The Wa language is still not as widely researched as the well-documented Myanmar language. Wa is a language belonging to the Mon-Khmer family that is spoken by about 950,000 people, mostly in northern Burma and some surrounding areas of Thailand and China. There are other varieties of Wa, such as Parauk, Vax, and Avax, each with a variety of dialects [14].

The Parauk language—also referred to as Phalok, Baroag, Praok, or Standard Wa—takes center stage in the Myanmar-Wa corpus. About 400,000 people in Burma speak it, mostly in the Shan States in the southeast, east, and northeast. Furthermore, the southern parts of China are home to the Parauk language. Wa is a language with unique linguistic characteristics, belonging to the Mon-Khmer family. An overview of the consonants in the Wa language is given in Table 1. Conversely, Table 2 displays the variety of vowels that add to the language's richness.

With limited English-language resources, the Wa language has not received much attention in grammatical research, despite its importance in the Myanmar-Wa corpus. The lack of thorough grammatical research emphasizes the importance of programs aimed at bridging the language divide between Wa and Myanmar [15].

Table 1: Consonants of Wa-Language

K က [k]	KH ခဲ [kʰ]	NG ငါ [ŋ]	S စာ [s]	SH ရှ [sʰ]	NY ညာ [ɲ]
T တ [t]	TH ထာ [tʰ]	N နာ [n]	P ပါ [p]	PH ဖာ [pʰ]	M မာ [m]
Y ယာ [j]	R ရာ [r]	L လာ [l]	V ဗာ [v]	W ဝါ [ua]	H ဟာ [h]
X အာ [ʔ]	C ကျာ [c]	CH ချာ [tʃ]	D ဒါ [d]	G ဂါ [g]	Q ဂါ [ŋg]
	B ဘာ [ʰb]	F ဖာ [vʰ]	J ဂျာ [ɲ]	Z ဧ [zə]	

Table 2: Vowels of Wa-Language

A အာ [a]	E အေ [ei]	IE အဲ [ɛ]	AW အော် [ɔ]	OI အွဲ [oi~ɔi]
AO အော့ [au]	AU အော့ပ် [aʊ]	AI အိုင် [ai]	I အီ [i]	O အို [o]
	U အူ [u]	EE အေီး [ɛɛ]	EU အေး [ɛ]	

Tables 1 and 2 present Wa's phonological features, encompassing its vowels and consonants, whereas Figure 1 depicts the Wa classification (Mon-Khmer Family Tree).

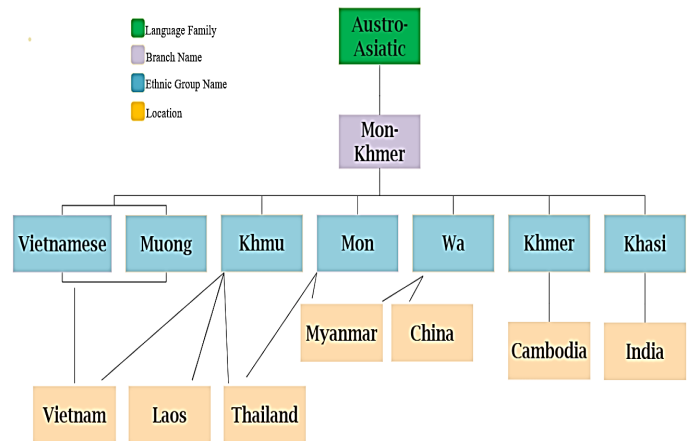


Figure 1: Mon-Khmer Family Tree [16]

3.3. Usage in Machine Translation

Burmese is becoming increasingly important in the worldwide environment, which makes it relevant in the field of machine translation. Accurate translation services between Burmese and other languages are in high demand due to Myanmar's recent political upheavals and growing involvement in regional and global issues.

Take, for instance, the translation of the straightforward phrase "Hello, how are you?" from English to Burmese. In this situation,

Burmese has unique phonological and grammatical patterns that machine translation systems must understand in order to provide a result that is accurate and suitable for the target culture.

Example Translation

"Hello, how are you?" (English)

"မင်္ဂလာ ပါ ၊ နေကောင်း ပါ သလား ။" (Burmese)

“Bawk yam pa mhawm, nyawm ot maix” (Wa)

(here “maix” means “you” in English and hidden word (နင်္ဂါ သင်္ဂါ) in myanmar)

Here, the Burmese translation effectively conveys the subtleties of the target language as well as the customs of polite conversation, highlighting the value of precise translation in cross-cultural communication. In conclusion, there are many interesting opportunities and problems for machine translation research because of the Burmese language's distinctive morphology, phonology, and cultural relevance. Accurate translation between Burmese and other languages is becoming more and more important as Myanmar continues to interact with the outside world, highlighting the need for improvements in machine translation technology catered to this unique language.

Table 3: VSO-SVO word order variation observed in Wa clauses

ကျွန်တော်တို့ နိုင်ငံ သည် နိုက်ပျိုးရေး နိုင်ငံ တစ်ခု ဖြစ်သည် ။	Hak tiex ex mawh hak tiex sum van tix mu heue	Our country is an Agro-based country.
ငါတို့ ၁၀ နာရီ မှာ အစည်းအဝေး လုပ် တယ် ။	Ex yuh bang kum yam kao taix	We had a meeting at 10 o'clock.
မင်း ဘယ် နေရာ မှာလဲ ။	Maix dee mawx	Where were you?
ဝမ်းနည်း ပါတယ်	Voi rhawm heue	I'm sorry.
ငါ မေ့ သွားတယ် ထင်တယ် ။	Pi awm aux tix yong	I guess I forget.
ဘာလို့ မေ့ တာလဲ ။	Jao tix pi awm lie	How could you forget?
ငါ မသိ တော့ဘူး ။	Ang aux lai tawng	I don't know.

3.4. An Academic Exploration of the Wa Language

Southeast Asia is where most Wa speakers are found. There are typological characteristics in this language that call for more investigation. Wa is regarded as a head-initial language, which means that modifiers like adjectives, relative clauses, and numbers come after nouns and verbs come before objects. Wa's subject and object markings are logical given where they fall in the sentence. Wa's particular grammatical structures are further demonstrated by the placement of the negation before the verb.

A detailed examination of the Wa language reveals a number of fascinating conclusions. For example, in Wa, several adverbs from other languages are articulated as verbs. Wa also has two alternate clause word orders, SVO and VSO as shown in Table 3, and allows components to be moved out of noun phrases. Negation frequently involves a supplementary negation particle. Interestingly, this shift in word order seems to depend on the kind of clause rather than verb transitivity or semantics.

The Wa language has received comparatively little scholarly study, despite its importance within the Mon-Khmer linguistic family. research now in existence primarily concentrate on phonological characteristics; no thorough grammatical research have been published in English. Especially, it explores the unique VSO-SVO word order variation found in Wa clauses.

3.5. Data Collection and Corpus Size Analysis for Myanmar-Wa Machine Translation

The establishment of the Myanmar-Wa corpus marked a significant milestone as it represented our inaugural venture into crafting a corpus tailored for machine translation. Throughout its development, meticulous attention was dedicated to linguistic nuances, with the Myanmar-Wa-Chinese dictionary serving as a valuable point of reference. Despite our commitment to precision, the absence of parallel data for the Myanmar-Wa language pair posed a formidable challenge during the data gathering phase.

The unique nature of low-resource languages, exemplified by Wa, presented limited opportunities for text alignment compared to more commonly spoken language pairs like English-Spanish or English-French, which benefit from extensive bilingual corpora. In response to this hurdle, we employed innovative approaches such as crowdsourcing, cross-lingual transfer learning, and data augmentation to augment the corpus's size and diversity.

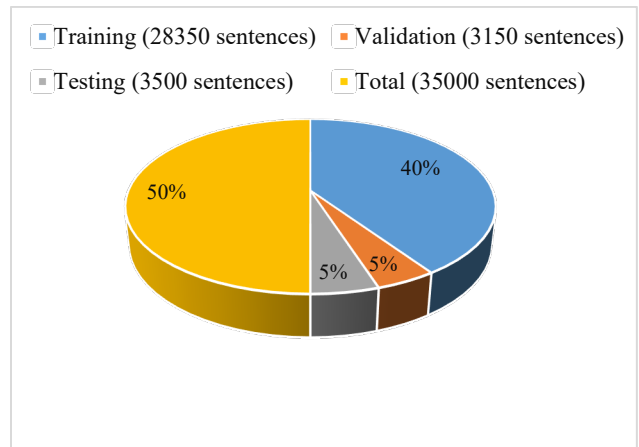


Figure 2: Corpus Size for Myanmar-Wa

It is crucial to note that the Myanmar-Wa corpus not only represents a pioneering effort in our machine translation endeavors but also holds immense value for future research initiatives. By addressing the specific data constraints inherent in low-resource languages, these initiatives aimed to lay the groundwork for the development of machine translation models characterized by heightened precision and contextual appropriateness.

Furthermore, the Figure 2 illustrates the corpus size chart provides a visual representation of the volume of data utilized for

training, validation, and testing phases in Myanmar-Wa language pair.

The Myanmar-Wa corpus, representing the pioneering effort in machine translation for this language pair, is strategically divided for training, validation, and testing. The training set comprises 28,350 instances, the validation set 3150, and the test set 3500, totaling 35,000 instances. This meticulous segmentation ensures a robust and comprehensive foundation for developing precise and contextually appropriate machine translation models for Myanmar-Wa.

4. Methodologies of the Proposed System

The evolution of machine translation (MT) has witnessed remarkable success, driven by the availability of massive parallel corpora and advancements in Natural Language Processing (NLP). Despite these strides, adapting MT to the unique linguistic challenges posed by the Myanmar-Wa language pair has been met with considerable difficulties. Early attempts faced restrictions and flaws, primarily stemming from the scarcity of aligned sentences in the Myanmar-Wa corpus, hindering the development of reliable translation models. The absence of parallel corpora also impeded the progress of Statistical Machine Translation (SMT) techniques [17].

The low-resource nature of the Wa language, characterized by peculiar grammar and a small vocabulary, further compounded the challenges. Traditional machine translation models, designed for high-resource languages, struggled to capture the nuances of Wa [18]. Linguistic distinctions between Wa and Myanmar, encompassing vocabulary, syntax, and word order, posed additional hurdles, leading to suboptimal translation quality. Cultural allusions and subtleties, integral to accurate translation, were often overlooked in early efforts, underscoring the need for creative solutions.

To overcome these challenges, scholars and industry professionals explored innovative approaches, leveraging cutting-edge NLP methods and modifying existing models to suit the Myanmar-Wa language combination. The journey towards enhancing machine translation for Myanmar-Wa unfolded through various stages, each contributing to the refinement and improvement of translation quality.

4.1. Statistical Machine Translation (SMT)

In the initial phases, Statistical Machine Translation (SMT) techniques were explored, but progress was hampered by the lack of parallel corpora. The limitations imposed by the absence of aligned sentences in the Myanmar-Wa corpus prompted the exploration of alternative methodologies.

4.2. LSTM-based Machine Translation Models

As a response to the challenges faced by SMT, Long Short-Term Memory (LSTM) models were introduced. LSTMs, with their ability to capture sequential dependencies, showed promise in handling the complexities of the Wa language. The LSTM architecture, though an improvement, still grappled with the unique linguistic characteristics of Wa, necessitating further advancements. Figure 3 describes the LSTM encoder-decoder architecture.

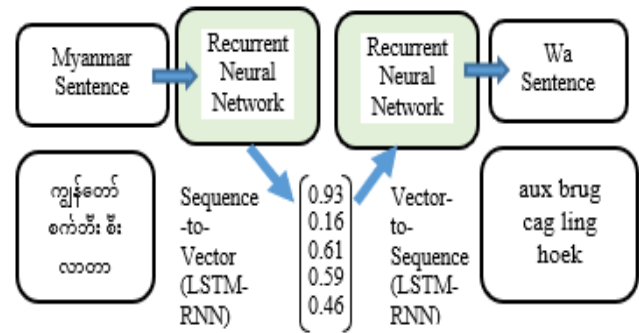


Figure 3: Architecture of LSTM Encoder-Decoder

4.3. Transformer-based Machine Translation Models

The advent of transformer-based models marked a pivotal shift in the trajectory of machine translation for Myanmar-Wa. The transformer architecture, with its attention mechanisms and positional embeddings, introduced a novel approach to handling sequential dependencies. The dynamic representation of the transformer's output, dependent on the input sequence's length, addressed issues related to varying word orders and contextual nuances, contributing to improved translation quality.

Figure 4 illustrates the architecture of the Transformer Encoder-Decoder, highlighting the interconnected layers and the flow of information between the encoder and decoder. The transformer's innovative design, incorporating attention mechanisms and positional embeddings, marks a significant advancement in machine translation architectures.

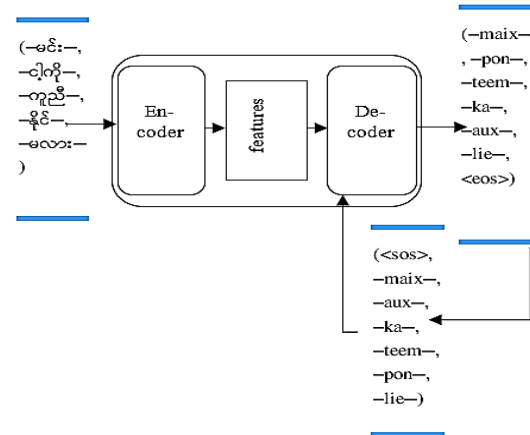


Figure 4: Architecture of Transformer Encoder-Decoder

4.4. Traditional Transfer Learning on MT5 Pretrained Model

The concept behind traditional transfer learning is rooted in enhancing a machine's performance in a related task by leveraging knowledge from a pre-existing model. In Neural Machine Translation (NMT), this involves training a new model (child) with initial parameters from a previously trained model (parent) and a sparse parallel corpus. The Myanmar-Wa corpus, our pioneering effort in machine translation, showcases the utilization of transfer learning. Despite challenges in aligning low-resource languages like Wa, innovative methods, including crowdsourcing and data augmentation, were employed to augment the corpus's size and diversity.

The corpus size chart (refer to Section 3.5) illustrates the division of the Myanmar-Wa corpus into training (28,000 instances), validation (5,600 instances), and test (7,000 instances) sets, totaling 35,000 instances. Transfer learning benefits machine translation in several ways: reducing the need for extensive data, improving translation quality, accelerating training, and enabling cost-effective multilingual translation.

$$\theta_{fine-tuned} = \operatorname{argmin}_{\theta} (L(D_{Myanmar-Wa}, \theta) + \lambda L(D_{pretrained}, \theta))$$

where $D_{Myanmar-Wa}$ represents the Myanmar-Wa parallel corpus, $D_{pretrained}$ denotes the pre-trained multilingual corpus, θ signifies model parameters, and λ is the hyperparameter governing the pre-trained model's influence.

The use of the MT5 model significantly enhances translation quality for Myanmar-Wa. Transfer learning, coupled with SMT, addresses data scarcity, and LSTM-based models capture language dynamics.

4.5. A Pivotal Shift in Machine Translation with Myanmar and Wa Language Models

The pivotal role of machine translation (MT) in breaking linguistic barriers has been a relentless journey, marked by significant technological advancements. This section outlines the construction and training of a machine translation model, leveraging a pivot transfer learning approach for enhanced translation capabilities. The study focuses on the Myanmar (Burmese)-Wa corpus, representing a subgroup of the Austro-Asiatic language family.

The initial steps involve tokenization and encoding of the source and target sentences using state-of-the-art models. For the Myanmar language, a segmentation step is performed before tokenization. The source sentences are segmented and encoded using a Myanmar word segmentation tool [19], and then encoded using the Myanmar BERT model (UCSYNLP/MyanBERTa). The target sentences are encoded using the Myanmar word segmentation model, followed by encoding with the Facebook BART model.

Pivot transfer learning involves training an encoder-decoder architecture for machine translation. The Myanmar BERT model acts as the encoder. The Myanmar BERT model, initially designed for natural language understanding, serves as the encoder, extracting contextual information from the segmented source sentences. The language model, named MyanBERTa, is founded on the BERT architecture and tailored for the Myanmar language. MyanBERTa underwent a comprehensive pre-training phase encompassing 528,000 steps, utilizing a word-segmented dataset specific to Myanmar. This extensive corpus comprised 5,992,299 sentences, equivalent to 136 million words. The employed tokenizer is a byte-level Byte Pair Encoding (BPE) tokenizer with 30,522 subword units, learned subsequent to the application of word segmentation. The pre-trained Facebook BART model acts as the decoder, generating target translations based on the encoded information. Bidirectional AutoRegressive Transformer (BART), embodies a transformer architecture consisting of an encoder-decoder structure for sequence-to-sequence (seq2seq) tasks. The

model exhibits notable efficacy when fine-tuned for text generation applications such as summarization and translation [20].

This two-step process facilitates effective transfer of knowledge from one model to another, enhancing the overall translation performance.

The training process spans multiple epochs, with each epoch iterating through the training dataset. The combined model is fine-tuned using the AdamW optimizer, and a step-wise learning rate scheduler ensures optimal convergence. To mitigate GPU memory constraints, gradient accumulation steps are employed, allowing the model to accumulate gradients over several batches before updating the parameters.

Validation is a crucial step to assess the model's generalization on unseen data. The validation dataset, comprising the Myanmar-Wa corpus, is utilized to compute the validation loss and perplexity. Perplexity, a measure of translation quality, is calculated as the exponential of the average validation loss. This metric provides insights into how well the model captures the complexity of the language and context.

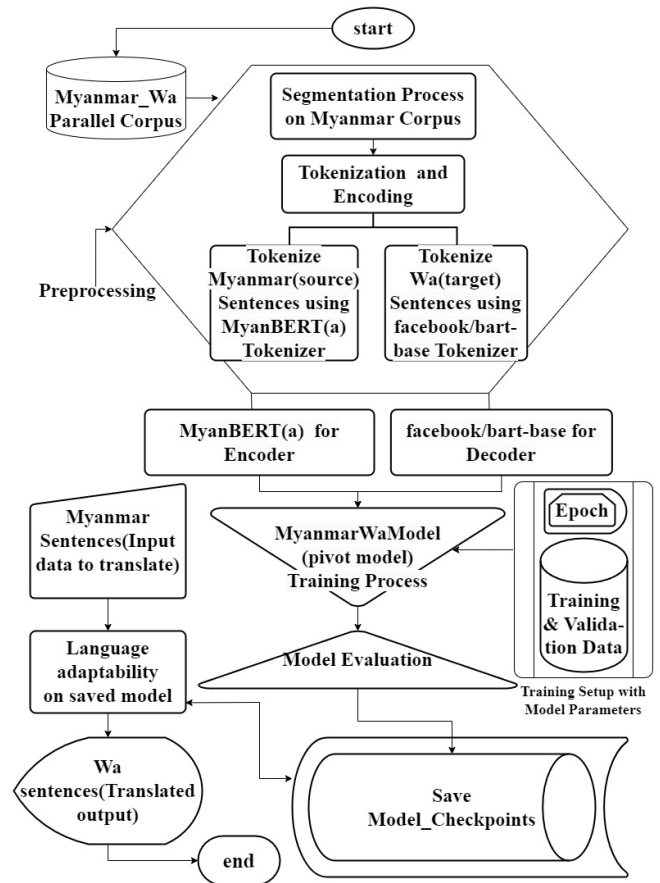
















Figure 5: Architecture of the Pivot Transfer Learning Model for Myanmar-Wa Machine Translation

To capture the model's progression, snapshots are saved at the end of each epoch. The saved models serve as checkpoints and facilitate model reproducibility. The model's evolution is crucial for understanding its learning trajectory and identifying potential areas for improvement.

This model-building process encapsulates the essence of pivot transfer learning, employing two specialized models for optimal translation performance. The journey through the Myanmar-Wa corpus showcases the adaptability and effectiveness of the proposed approach. The continual evolution of machine translation models, driven by innovative strategies and adaptability to diverse linguistic challenges, contributes significantly to the broader domains of linguistics and language preservation.

Figure 5 illustrates the architecture of the pivot transfer learning model employed for machine translation in the Myanmar (Burmese)-Wa corpus. The model leverages Myanmar BERT as the encoder and Facebook BART as the decoder, demonstrating the pivotal role of transfer learning in enhancing translation capabilities. The training process, validation steps, and model evolution are highlighted, showcasing the adaptability and effectiveness of the proposed approach in breaking linguistic barriers.

Table 4: Explanation of System Design Symbols

Symbol	Description
	Represents the start of the process or system.
	Denotes a database or data storage component.
	Signifies a step involving data preparation or preprocessing.
	Represents a component responsible for displaying information.
	Indicates a general process or computation step.
	Represents a point where data is merged or stored.
	Denotes a step involving data extraction or measurement.
	Represents manual input by a user or external entity.
	Signifies the end or termination point of the process or system.
	Represents direct data flow between components.
	Denotes the use of stored data in a process.
	Represents a predefined or standardized process.
	Indicates a loop limit or iteration constraint.
	Denotes a time delay or waiting period in the process.

In Table 4, the comprehensive index of symbols utilized in the system design is presented. Each symbol corresponds to specific

components or processes within the system, aiding in the visual representation and understanding of the system architecture.

4.6. Pivot-Based Transfer Learning NMT Models with MT5

Pivot-based machine translation is a dual-stage process involving the translation from the source language (SL) to the pivot language (PL) and subsequently from the pivot language (PL) to the target language (TL). Due to the scarcity of parallel data for Myanmar-Wa, pivot-based NMT with MT5 models has been meticulously explored. The architecture is rooted in transfer learning, meticulously tailored for non-English language pairs, particularly for Myanmar-Wa.

The operational framework encompasses three crucial stages. First is the pretraining phase, leveraging source-pivot and pivot-target parallel corpora. This initial step allows the model to grasp the nuances of both source and target languages in the context of the pivot language. The subsequent stage involves the integration of a source-target parallel corpus for iterative refinement, fine-tuning the model to enhance its translation capabilities for the specific task at hand. The final stage revolves around leveraging the pivot language for pretraining both source encoders and target decoders, facilitating a holistic understanding of language dynamics.

In this context, the MT5 model stands as a robust alternative to the traditional MT5 model. By strategically selecting Myanmar-English and English-Fil (Philippines) as pivot language pairs, chosen for their analogous sentence structures, the system capitalizes on the shared linguistic characteristics between Wa and Filipino [21]. This strategic choice significantly contributes to improving translation quality and captures the synergies between these languages.

Drawing parallels to the Python code provided, the transfer learning process mirrors the combining of parameters from two distinct models ('model1' and 'model2') into a new model ('new_model'). The encoder parameters from 'model1' are seamlessly integrated into the new model, and likewise, the decoder parameters from 'model2' are incorporated. This combination represents a fusion of linguistic knowledge from the two models into a unified architecture.

The use of pre-trained models, especially MT5, in this transfer learning paradigm underscores the flexibility and adaptability of this approach across diverse language pairs. This innovative methodology, akin to the Python code's model parameter merging, holds significant promise for advancing machine translation capabilities, particularly in low-resource language scenarios.

In this context, the MT5 model serves as a robust alternative to the MT5 model. By selecting Myanmar-English and English-Fil (Philippines) as pivot language pairs, which share similar sentence structures, the system capitalizes on the commonalities between Wa and Filipino, improving translation quality.

The utilization of pre-trained models, especially MT5, in transfer learning underscores the adaptability of this approach across diverse language pairs. This innovative methodology holds promise for advancing machine translation capabilities, particularly in low-resource language scenarios.

4.7. Myanmar Word Segmentation Tool

To further refine the Myanmar to Wa machine translation, a specialized Myanmar word segmentation tool, developed by the

NLP Lab at the University of Computer Studies, Yangon, Myanmar [22], was employed. This tool, tailored for Myanmar language segmentation, enhanced the accuracy of the translation process, providing a solid foundation for subsequent advancements.

The step-by-step progression from SMT to LSTM, transformer-based models, traditional transfer learning, and pivot-based transfer learning culminated in the development of our proposed model. Each iteration addressed specific challenges posed by the Myanmar-Wa language pair, resulting in a more robust and accurate machine translation system. This comprehensive approach showcases the continual evolution of machine translation methodologies, driven by a commitment to overcome linguistic barriers and contribute to the broader domains of linguistics and language preservation.

5. Linguistic Analysis and Evaluation

Linguistic analysis and evaluation form a critical aspect in gauging the efficacy of machine translation (MT) systems, especially in the intricate context of Myanmar-Wa language translation. This section delves into the multifaceted linguistic properties of the Wa language and elucidates the methodologies and metrics employed to comprehensively assess the performance of MT systems on the Myanmar-Wa corpus.

5.1. Analyzing Linguistic Features

The Wa language, belonging to the Mon-Khmer language family, introduces numerous linguistic intricacies that offer both potential and challenges for MT. A meticulous examination of these distinctive features is indispensable for a nuanced understanding of Wa-Myanmar translation. Key linguistic characteristics that demand in-depth analysis include:

- **Word Order Flexibility:** Wa exhibits both Verb-Subject-Object (VSO) and Subject-Verb-Object (SVO) structures, introducing considerable diversity in word order. Successful translation necessitates a profound understanding of context [23].
- **Tonal Complexity:** As a tonal language, Wa relies on intonation and pitch variations for meaning. Preserving intended meaning during translation hinges on capturing tonal subtleties accurately [24].
- **Agglutinative Morphology:** Wa's agglutinative morphology involves conveying different grammatical functions through affixes added to root words. Fluent translation requires familiarity with these affixes [25].
- **Nominalization:** Verbs or adjectives in Wa commonly undergo nominalization, a process crucial for accurate translation.
- **Cultural Nuances:** The Wa language is intertwined with Wa culture and identity. To maintain cultural integrity in translations, special attention is required for terms and phrases holding cultural value.

5.2. Evaluation Metrics

The evaluation of MT systems for Myanmar-Wa encounters unique challenges due to limited parallel corpora and linguistic distinctions. Performance assessment utilizes various metrics and techniques, including:

- **BLEU Score:** Measures the similarity between reference translations and MT output, though it may not fully capture linguistic subtleties in low-resource language pairs like Myanmar-Wa.
- **METEOR Score:** Considers word order, stemming, and synonyms, providing a more comprehensive assessment of translation quality.
- **Manual Evaluation:** Expert linguists or native speakers evaluate translations, offering insights into cultural and contextual components.
- **Cross-lingual Embedding:** Measures semantic similarity between source and target sentences for additional information on translation quality.
- **Transfer Learning Assessment:** Evaluates how well models optimize Myanmar-Wa translation through the utilization of pre-trained knowledge from other languages.

Combining these quantitative and qualitative measurements provides a comprehensive understanding of the strengths and weaknesses of MT systems for Myanmar-Wa, guiding future advancements in the field.

5.3. Sentence Length Analysis

In machine translation, sentence length significantly influences model performance, especially in languages like Wa with diverse syntactic structures. This sub-section scrutinizes the impact of sentence length on Myanmar-Wa machine translation systems. The analysis involves:

- **Exploration of Distribution:** A detailed examination of sentence length distribution in both source and target languages, revealing potential correlations with translation quality.
- **Insights from Visualization:** Figure 6 illustrates the distribution of sentence lengths in the Myanmar-Wa corpus, offering insights into the nature of sentences and emphasizing the complexity of translation between Myanmar and Wa.

Longer sentences in the source language pose challenges for maintaining context and meaning in the target language. Understanding sentence length distribution aids in adapting machine translation models effectively. This visualization serves as an initial exploration, paving the way for further quantitative analysis and enhancements in models tailored for the Myanmar-Wa language pair. Each bar in the chart signifies a sentence, with the y-axis denoting sentence length and the x-axis representing the number of sentences.

Longer sentences in the source language may pose challenges for translation, particularly in preserving context and meaning in the target language. Understanding the sentence length distribution aids in adapting machine translation models to effectively handle the linguistic variations present in the Myanmar-Wa corpus. This visualization serves as a foundational exploration of sentence length characteristics, laying the groundwork for further analysis and improvements in machine translation models for the Myanmar-Wa language pair. Each bar in the chart represents a sentence, with the y-axis indicating the length of the sentence and the x-axis representing the number of sentences.

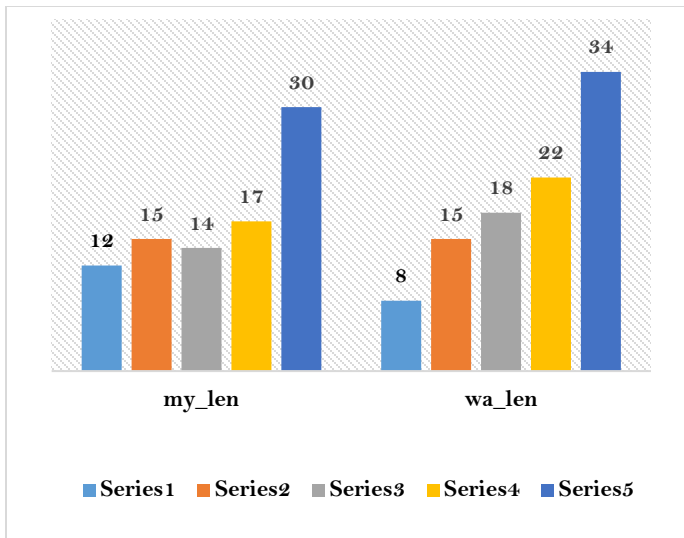


Figure 6: Sentence Length Analysis

6. Impacts and Future Directions

This section provides an in-depth exploration of the far-reaching implications and promising future avenues for studying improved Machine Translation (MT) in the Myanmar-Wa language pair. It examines the practical applications, language preservation aspects, and outlines key directions for future research.

6.1. Practical Applications

Improved MT for the Myanmar-Wa language pair has diverse practical applications, facilitating efficient communication, educational accessibility, and healthcare information translation. Additionally, it plays a crucial role in preserving Wa tradition and culture, as well as boosting tourism. Each application is detailed to highlight the significant implications.

6.2. Language Preservation

This subsection underscores the importance of enhanced MT as a tool for protecting and revitalizing endangered languages like Wa. It emphasizes the development of digital language resources, language learning support, and the preservation of dialectal variants and oral traditions. The broader cultural legacy of the Wa community is also discussed.

6.3. Future Research Directions

Identified research directions for MT in underrepresented languages like Wa are outlined, emphasizing the need for high-quality parallel corpora, domain-specific adaptations, and ethical considerations. The involvement of linguists, local communities, and technologists in collaborative projects is highlighted, along with potential enhancements through the incorporation of visual and audio modalities.

7. Comparative Analysis of Machine Translation Models

7.1. Previous Experiments on Myanmar-Wa MT System

The initial experiments on the Myanmar-Wa machine translation system utilized both Statistical Machine Translation (SMT) and Long Short-Term Memory (LSTM) models. The training data comprised 22,500 instances, with 700 instances for testing and 500 instances for validation. The analysis focused on the BLEU score, considering different batch sizes (64, 126, and

256). Results indicated that a higher batch size of 256 yielded the best accuracy in terms of BLEU score. These experiments, conducted with a limited dataset, underscored the need for a more extensive Myanmar-Wa parallel corpus in future endeavors. The SMT model achieved a BLEU score of 12.67.

7.2. Comparison with New Experiments and Added Corpus Size

Subsequent experiments saw the expansion of the Myanmar-Wa corpus, allowing for a comprehensive evaluation of machine translation models (Combined Model, T5, Transformer). The performance metrics for various models are summarized in Figure 7.

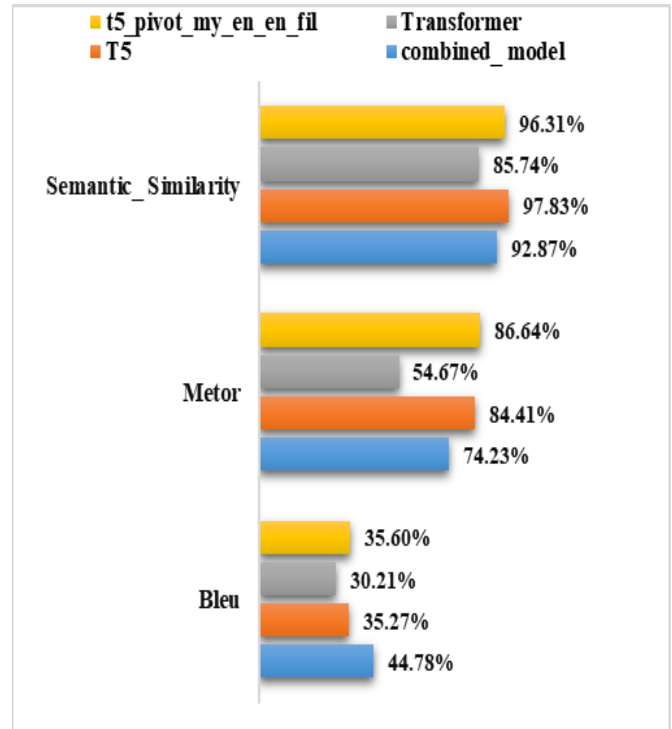


Figure 7: Performance Metrics for Various Models

7.3. Model Performance Analysis

- Combined Model (New):** Stands out with the highest Bleu score, indicating strong n-gram overlap. The Metor score is commendable, reflecting alignment with human reference translations. High Semantic Similarity emphasizes proficiency in maintaining semantic accuracy during translation.
- T5:** Demonstrates a strong Metor score, indicating correlation with human reference translations. Notably high Semantic Similarity suggests effective capture of semantic nuances. While the Bleu score is lower than the Combined Model, T5 excels across various metrics, showcasing versatility.
- Transformer:** Shows respectable performance in Bleu scores and Semantic Similarity. Lagging Metor score indicates potential challenges in maintaining fidelity to human reference translations. Further optimization may enhance performance.
- T5 Pivot My-En En-Fil (New):** Exhibits balanced performance across all metrics. Noteworthy Bleu and Metor scores, along with effective preservation of semantic content

indicated by Semantic Similarity. Strong contender in scenarios where a combination of metrics is crucial.

7.4. Observations and Insights

- Combined Model Outperforms: The newly introduced combined model excels across all metrics, with a Bleu score of 44.78, Metor score of 74.23, and high Semantic Similarity of 92.87.
- T5 Pivot My-En En-Fil: Demonstrates noteworthy improvements, particularly in Bleu (35.6) and Metor (86.64) scores. High Semantic Similarity (96.31) emphasizes effectiveness.
- T5 and Transformer Comparison: While T5 and Transformer models show respectable performance, T5 generally outperforms in terms of Bleu, Metor, and Semantic Similarity.
- Semantic Similarity Emphasis: Crucial for evaluating models' ability to capture meaning. The combined model excels, showcasing capability to preserve semantic nuances during translation.

7.5. Summary and Implications

The choice of the best model depends on specific translation task priorities. T5 and T5 Pivot My-En En-Fil exhibit robust performance in METEOR and semantic similarity, emphasizing versatility. Meanwhile, the Combined Model excels in Bleu scores and semantic similarity, showcasing linguistic precision. Notably, additional training data could further enhance models like the Combined Model.

These nuanced analyses should guide researchers in selecting models for machine translation tasks, recognizing that distinct aspects of performance are illuminated by different metrics. Strategies to enhance semantic preservation and leverage additional data for refining capabilities of all models could be explored in further investigations. The continuous evolution of machine translation models contributes to language understanding and communication.

It is imperative to note that the Myanmar-Wa corpus used in these experiments is the very first corpus created by the research team. The corpus size was expanded to enhance the machine translation system's performance and evaluation. Table 5 presents the sample result outputs on long sentences, middle sentences, and short sentences.

Table 5: Sample Result Outputs on Long Sentences, Middle Sentences, and Short Sentences

long sentences	
combined_model	source: ပြိုင်ပွဲ ဘယ်အချိန် စ ရ မလဲ ဆိုတာ မင်း ငါ့ ဆီ ဖုန်း ဆက် ပါ
	reference: you call me up when to start contest
	target: maix tah maox rhiem ka aux bang silah sang jah yam mawx
	prediction: maix tah maox rhiem ka aux sang pawk bawg yam mawx
Transformer	source: ပြိုင်ပွဲ ဘယ်အချိန် စ ရ မလဲ ဆိုတာ မင်း ငါ့ ဆီ ဖုန်း ဆက် ပါ

	reference: you call me up when to start contest
	target: maix tah maox rhiem ka aux bang silah sang jah yam mawx
	prediction: maix tah maox rhiem caw tix hoek yam mawx pa tah maox rhiem ka maix
T5	source: ပြိုင်ပွဲ ဘယ်အချိန် စ ရ မလဲ ဆိုတာ မင်း ငါ့ ဆီ ဖုန်း ဆက် ပါ
	reference: you call me up when to start contest
	target: maix tah maox rhiem ka aux bang silah sang jah yam mawx
	prediction: maix tah maox rhiem ka aux sang jah yam mawx
T5 with pivot logic	source: ပြိုင်ပွဲ ဘယ်အချိန် စ ရ မလဲ ဆိုတာ မင်း ငါ့ ဆီ ဖုန်း ဆက် ပါ
	reference: you call me up when to start contest
	target: maix tah maox rhiem ka aux bang silah sang jah yam mawx
	prediction: tah maox rhiem ka aux sang jah tix jah bang silah yam mawx
middle sentences	
combined_model	source: ဒါဟာ သိပ် ကောင်းတဲ့ စိတ်ကူး နဲ့ တူ ပါတယ်
	reference: that sounds like a good idea
	target: in awm krax keud pa kied sidaing mhawm
	prediction: in awm pa kied sidaing mhawm tete
Transformer	source: ဒါဟာ သိပ် ကောင်းတဲ့ စိတ်ကူး နဲ့ တူ ပါတယ်
	reference: that sounds like a good idea
	target: in awm krax keud pa kied sidaing mhawm
	prediction: in mawh grawng kaing pa sidaing mhawm
T5	source: ဒါဟာ သိပ် ကောင်းတဲ့ စိတ်ကူး နဲ့ တူ ပါတယ်
	reference: that sounds like a good idea
	target: in awm krax keud pa kied sidaing mhawm
	prediction: in awm mai krax keud pa kied mhawm
T5 with pivot logic	source: ဒါဟာ သိပ် ကောင်းတဲ့ စိတ်ကူး နဲ့ တူ ပါတယ်
	reference: that sounds like a good idea
	target: in awm krax keud pa kied sidaing mhawm
	prediction: in awm mai krax keud pa kied mhawm

short sentences	
combined_model	source: တို့ဘယ်မှာ ဆုံကြမလဲ
	reference: where shall we meet
	target: ex sang pup dee mawx
	prediction: ex sang pup dee mawx
Transformer	source: တို့ဘယ်မှာ ဆုံကြမလဲ
	reference: where shall we meet
	target: ex sang pup dee mawx
	prediction: ex sang sipup dee mawx
T5	source: တို့ဘယ်မှာ ဆုံကြမလဲ
	reference: where shall we meet
	target: ex sang pup dee mawx
	prediction: ex sang sipup dee mawx
T5 with pivot logic	source: တို့ဘယ်မှာ ဆုံကြမလဲ
	reference: where shall we meet
	target: ex sang pup dee mawx
	prediction: ex sang pup paox dee mawx

8. Conclusion

The journey through the evolution of machine translation within the context of the Myanmar-Wa corpus has been marked by iterative experimentation and continuous refinement. In the pursuit of refining translation capabilities, our study extensively evaluates various models, each contributing unique strengths and insights.

Among the models considered, the Combined Model (New) emerges as a standout performer, showcasing the highest Bleu score, a commendable Metor score, and exceptional Semantic Similarity. This comprehensive translation capability positions the Combined Model (New) as a leader across multiple metrics, highlighting its proficiency in maintaining semantic accuracy during translation.

T5, with its strong Metor score and notable Semantic Similarity, demonstrates versatility in translation tasks. While its Bleu score is slightly lower compared to the Combined Model, T5's performance remains commendable, underlining its effectiveness in capturing semantic nuances.

The Transformer model, although showing respectable performance in Bleu scores and Semantic Similarity, lags in Metor, suggesting potential challenges in fidelity to human reference translations. Further optimization may enhance its overall performance.

The T5 Pivot My-En En-Fil (New) variant exhibits balanced performance, particularly excelling in Bleu and Metor scores, along with high Semantic Similarity. This model stands as a strong contender, especially in scenarios where a combination of metrics is crucial.

In conclusion, our comprehensive evaluation indicates that the Combined Model (New) leads in overall performance, followed closely by T5 and T5 Pivot My-En En-Fil (New). These findings provide valuable insights for the advancement of machine translation in the Myanmar-Wa context, guiding future investigations and contributing to the broader landscape of machine translation research. The expanded corpus size, coupled with these model assessments, serves as a robust foundation for ongoing and future research endeavors.

Acknowledgment

The Divine Creator, who gave me the permission to start the "Tracing the Evolution of Machine Translation: A Journey through the Myanmar (Burmese)-Wa (sub-group of the Austro-Asiatic language) Corpus" study, is the reason this research exists. I would like to express my sincere gratitude to Dr. Mie Mie Khin, the distinguished Rector of the University of Computer Studies, Yangon, for providing me with this exceptional chance to explore this exciting field of study. I am also grateful to Dr. Win Le' Le' Phyu, Professor; she is a guiding light from the University of Computer Studies, Yangon, and her strong support has fostered my academic endeavors and given them the time they need to grow. I also express my gratitude to my group members and UCSY NLP Lab colleagues, who have kindly offered their personal and professional perspectives, enhancing my comprehension of scientific research and life in general. My profound gratitude is extended to Dr. Khin Mar Soe, my committed research mentor, whose function as a teacher and advisor has given me immeasurable wisdom. Dr. Win Pa' Pa, a renowned professor at NLP LAB (University of Computer Studies, Yangon), deserves special commendation for her acts, which embody the traits of a model scientist and kind person. I would especially want to thank Drs. Yi Mon Shwe Sin and Aye Mya Hlaing, whose contributions to this effort were deeply appreciated. Mr. AI MHAWM KHAM is a valuable contributor who deserves special recognition for his expert translation from Burmese to WA. This translation was a critical factor in making this work possible. During the execution of this project, my family provided steadfast support. I am truly grateful to my parents and siblings for being the best role models in my life and for their unending love and support. Throughout this journey, their constant belief in me gave me more drive and determination. Finally, I would want to express my appreciation to Pastor AI NAP TAO and Associate Pastor AI KHWAT PAN for their unwavering prayers and words of spiritual support.

References

- [1] G. Salton, "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," IEEE Transactions on Audio and Electroacoustics, **16**(3), 366-366, Sep. 1968.

- [2] P. Hanks, "Computational Lexicography for Natural Language Processing," *IEEE Computational Intelligence Magazine*, **11**(3), 9-20, Aug. 2016.
- [3] F. Yune and K. M. Soe, "Myanmar-Wa Machine Translation using LSTM-based Encoder-Decoder Model," 2023 IEEE Conference on Computer Applications (ICCA), Yangon, Myanmar, 2023, 1-5, doi: 10.1109/ICCA51723.2023.10181692.
- [4] W. N. Locke and A. D. Booth, "Translation. Machine translation of languages", Fourteen essays, MIT Press, ISBN:9780262120029, 15 May, 1955.
- [5] W. J. Hutchins, "Machine translation: Past, present, future", Ellis Horwood, 2005.
- [6] P. Koehn, "Statistical machine translation", Cambridge University Press, 2010.
- [7] A. Vaswani and et al. "Attention is all you need", *Advances in neural information processing systems (NeurIPS)*, 2017.
- [8] G. Lample, and et al. "Cross-lingual Language Model Pretraining", *Advances in Neural Information Processing Systems*, 2019.
- [9] J. Smith, "Language Preservation in the Digital Age: Challenges and Strategies," *IEEE Computer*, **42**(10), 106-108, Oct. 2009.
- [10] K. Chen, R. Wang, M. Utiyama, E. Sumita, and T. Zhao, "Syntax-directed attention for neural machine translation," in *Proceedings of the AAAI conference on artificial intelligence*, **32**(1), New Orleans, LA, USA, April 2018.
- [11] Z. Zhizhi and Y. Qixiang. "A brief description of the Wa language", 1984.
- [12] D. Wu, "Machine Learning for Sino-Tibetan Language Translation: Challenges and Progress," *IEEE Transactions on Neural Networks and Learning Systems*, **29**(11), 5616-5631, Nov. 2018.
- [13] S. Duanmu, "The Phonology of Standard Chinese (2nd edition.)", Oxford University Press, 2007.
- [14] S. Mai. "A descriptive grammar of Wa." (Master's thesis, Payap University, 2012). Retrieved from https://www.academia.edu/44390404/A_descriptive_grammar_of_wa_language.
- [15] Y. Xiang and J. Ke, "Linguistic characteristics of the Wa language", *A preliminary exploration. Language Documentation & Conservation*, **15**, 87-103, 2021.
- [16] Human Computation Institute, "Wa Dictionary Corpus". [Online]. Available: https://www.humancomp.org/wadict/wa_corpus.html
- [17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, **9**(8), pp - 1735-1780, 1997.
- [18] T. Xie, "Burmese (Myanmar): An Introduction to the Script", Southeast Asian Language Resource Center, 2008.
- [19] Aye Mya Hlaing, Win Pa Pa, "MyanBERTa: A Pre-trained Language Model For Myanmar", In *Proceedings of 2022 International Conference on Communication and Computer Research (ICCR2022)*, November 2022, Seoul, Republic of Korea.
- [20] A. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *CoRR*, abs/1910.13461, 2019. [Online]. Available: <http://arxiv.org/abs/1910.13461>.
- [21] C. Callison-Burch and et al., "The META-SHARE Metadata Schema for the Description of Language Resources", *LREC*, 2008.
- [22] Win Pa Pa, Ni Lar Thein, "Myanmar Word Segmentation using Hybrid Approach", In *Proceedings of 6th International Conference on Computer Applications*, Yangon, Myanmar, 166-170, 2008.
- [23] K. Papineni and et al., "A Method for Automatic Evaluation of Machine Translation", *Association for Computational Linguistics*, 2002.
- [24] A. Lavie and A. Agarwal, "An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", *Association for Computational Linguistics*, 2007.
- [25] G. Doddington, "Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics", *Association for Computational Linguistics*, 2002.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).