Exploring Extractive and Abstractive Approaches for Myanmar Language Text Summarization

Khin Nyo Nyo Theint
University of Information Technology, Yangon,
khinnyonyotheint@uit.edu.mm

Tun Ye Minn
University of Information Technology, Yangon,
tunyeminn@uit.edu.mm

Aung Khant Myat
University of Information Technology, Yangon,
aungkhantmyat@uit.edu.mm

Thinn Thinn Wai
Faculty of Computer Science
University of Information Technology, Yangon,
thinnthinnwai@uit.edu.mm

Hmway Hmway Tar
Faculty of Computer Science
University of Information Technology, Yangon,
hmwaytar@uit.edu.mm

Win Win Thant
Faculty of Computer Science
University of Information Technology, Yangon,
winwinthant@uit.edu.mm

Thet Thet Zin
Faculty of Computer Science
University of Information Technology, Yangon,
thetthetzin@uit.edu.mm

***Abstract*— Text summarization is one of the most difficult natural language processing (NLP) tasks. It is the process of creating concise, intelligent summaries from a variety of sources, including books, news articles, and social media posts. This method saves readers time in today's busy environment while giving them fast access to important information. This paper presents extractive and abstractive summarization approaches for Myanmar text. While abstractive summarizing creates the complete summary from scratch using the input text, extractive summarization chooses a subset of sentences from the original text for the summary. Abstractive summarizing is more sophisticated than extractive summarization since it involves human-like elements to create the summary from the ground up. The TextRank algorithm and cosine similarity were used for extractive summarization. The mT-5 model, a transformer-based seq2seq model, was employed to carry out abstractive summarization on the XL-Sum dataset. As a result, the extractive summarization of the proposed system yields 43.14 of the ROUGE-1 score, while the abstractive summarization contributes 36.67 of the ROUGE-1 score.**

Keywords—* *Text Summarization, Natural Language Processing, Extractive Summarization, Abstractive Summarization, Cosine Similarity, mT5

I. INTRODUCTION

Due to the rapid expansion of the internet, there is increased availability of text data online and thus automatic text summarization becomes a key issue in Natural Language Processing (NLP). The purpose of automatic text summarization is to take long texts and provide the smaller version with the most relevant information. It is especially important in fields where comprehensive analysis of documents like legal, news or academic papers is needed since relevant and brief information could be a significant factor to increase productivity and efficiency of the decision-making process.

The main goal of this paper is developing extractive and abstractive summarization methods for Myanmar news articles. For extractive summarization, it will leverage the TextRank algorithm and cosine similarity. For abstractive summarization, it will employ mT-5 transformer-based seq2seq model leveraging the XL-Sum dataset. The purpose of the paper is to generate well-structured human-like summaries that have a good coverage of information and to demonstrate the nature of extractive and abstractive summarization methods in Myanmar text summarization methodology.

In recent years, there have been many advancements in text summarization methodologies for widely spoken languages such as English, Chinese, and Japanese. But when it comes to the Myanmar language, text summarization remains very limited. Present-day research predominantly focuses on extractive methods, such as LSA-based techniques [1] and semantic approaches using Word Embedding and Graph Based Approach [2], which involve selecting important sentences from the original text. However, abstractive summarization, known for its human-like quality in generating entirely new summaries, offers a significant advantage by producing more natural and coherent text compared to extractive methods. Despite its potential, abstractive summarization techniques are yet to be developed for the Myanmar language.

This study seeks to fill the gap by introducing an abstractive summarization technique for the Myanmar language in parallel to the extractive method. Previous research has primarily utilized extractive approaches, but this work will go beyond that by incorporating the mT-5 transformer-based model to generate summaries that resemble human writing. By addressing both extractive and abstractive methods, this study will explore the qualities of Myanmar text summarization methods and differentiate itself from prior research.

II. RELATED WORKS

According to previous Myanmar text summarization research, it is pointed out that extractive summarization approaches are only developed. In this paper, it will explore both extractive and abstractive approaches of text summarization.

In the paper "Myanmar news summarization using different word representations" by S. S. Lwin and K. T. Nwet, the authors proposed a centroid-based word embedding summarizer for Myanmar news and used different word embedding models to compare with Bag-of-words summarization. Experiments utilize the local and international news datasets; the cosine similarity score is applied to rank the sentences in terms of their similarity to the centroid vector. The results show that the default models such as FastText and BPEmb yield better scores than the BOW model, and BPEmb-based centroid summarizer performs well under specified circumstances. These studies support the hypothesis that word embedding techniques are more effective than BOW models, noting the importance of preserving syntactic and semantic dependencies to improve the quality of summarization [3]. However, it did not explore abstractive summarization techniques, leaving a gap in generating more fluent and human-like summaries.

Y. K. Thu and W. P. Pa explored text summarization methods specifically for Myanmar news articles. The authors use an encoder-decoder LSTM model to create a Recursive RNN model for headline prediction, which generates a single word forecast and iteratively called a forecasted text summarization in the form of news headlines. They gathered 5000 articles from Myanmar to train the headline prediction model and evaluate its performance relative to sequence-to-sequence models using ROUGE score values. From their results, they have ascertained that the Recursive RNN model significantly outperforms the other models, demonstrating the ability of generating concise and precise headlines for the Myanmar news articles [4]. While this study demonstrated the potential of neural network-based approaches for summarization, it remained limited to headline generation and did not explore fully-fledged abstractive summarization of longer texts.

Hasan et al. (2021) present findings on evaluating summaries in 44 languages in their paper titled "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages". They emphasized the training of the language model on fine-tunes pretrained checkpoints. The proposed system applies the XL-Sum dataset and the mT5 model for multilingual abstractive summarization [5]. However, there remains an opportunity to further enhance summarization performance by exploring monolingual models tailored specifically to Myanmar language. So, this paper utilized both models and showcases the monolingual variation of the model which outperforms the multilingual version for abstractive.

III. Background Theories

There are two main methods for summarizing text: extractive and abstractive. Abstractive summarization involves generating new sentences or phrases to convey the essence of the information, which may not be present in the original content. On the other hand, extractive summarization selects relevant sentences directly from the source text without making any changes. Past research papers have focused on assessing the importance of individual sentences and their connections, rather than exploring contextual understanding.

A. *Extractive Summarization*

One way to summarize a text document is through extractive summarization, which involves selecting key words or sentences from the original text and combining them to create a summary. This method captures the most important ideas from the original content and presents them in a new format. Terms like "selection-based summarization" or "surface-level summarization" are often used to describe this approach. An extractive summary allows readers to grasp the essential concepts and main points of a lengthy document without having to read it in its entirety. This technique can enhance productivity and save time across various fields, including the legal, medical, and business sectors.

Following a discussion of extractive summarization, an example is provided to understand more about the result of the extractive summarization system.

Input Article:

“AIIB ဘဏ် သဘောတူညီချက်ကို အောက်တိုဘာလက တရုတ် အပါအဝင် ၂၁ နိုင်ငံ လက်မှတ် ထိုးခဲ့ ယူကေ နိုင်ငံဟာ 'အာရှ အခြေခံ အဆောက်အအုံ ရင်းနှီး မြှုပ်နှံမှု ဘဏ်'(AIIB) ရဲ့ အဖွဲ့ဝင် ဖြစ်လာဖို့ လျှောက်ထားမယ့် ပထမဆုံး အနောက် အုပ်စုဝင် စီးပွားရေး အင်အားကြီး နိုင်ငံ ဖြစ်ပါတယ်။ AIIB ဘဏ်ဟာ အာရှ ဒေသက စွမ်းအင်၊ သယ်ယူ ပို့ဆောင်ရေးနဲ့ အခြေခံ အဆောက်အအုံ ဆိုင်ရာ ပရောဂျက်တွေ အတွက် ငွေကြေး ထောက်ပံ့မယ်လို့ ဆိုပါတယ်။ ဒါပေမယ့် အမေရိကန် ကတော့ ဒီဘဏ်သစ်ရဲ့ အုပ်ချုပ်ရေးပိုင်းမှာ နိုင်ငံတကာ စည်းမျဉ်းတွေနဲ့ အညီ ဆောင်ရွက် နိုင်ပါ့မလား ဆိုတဲ့ မေးခွန်းတွေ ထုတ်ခဲ့ ပါတယ်။ တရုတ် နိုင်ငံကတော့ အမေရိကန်ရဲ့ စိုးရိမ် ပူပန်မှုကို ပယ်ချ လိုက်ပြီး၊ အမေရိကန်ဟာ တရုတ် နိုင်ငံနဲ့ ပတ်သက်ရင် ကလေး ဆန်တဲ့ သံသယ လွန်မှုတွေ ထားတုန်း ဖြစ်တယ်လို့လည်း ဆင်ဟွာ သတင်း ဌာနက စွပ်စွဲ လိုက်ပါတယ်။”

"The AIIB agreement was signed by 21 countries, including China, in October. The UK is set to become the first major Western economy to apply for membership in the Asian Infrastructure Investment Bank (AIIB). The AIIB aims to provide funding for energy, transportation, and infrastructure projects in the Asian region. However, the United States has raised questions about whether this new bank will operate in accordance with international standards in its governance. China, in response, dismissed the US's concerns, with the Xinhua news agency accusing the US of maintaining a childish level of suspicion when it comes to China."

Extractive summary:

“တရုတ် နိုင်ငံကတော့ အမေရိကန်ရဲ့ စိုးရိမ် ပူပန်မှုကို ပယ်ချ လိုက်ပြီး၊ အမေရိကန်ဟာ တရုတ် နိုင်ငံနဲ့ ပတ်သက်ရင် ကလေး ဆန်တဲ့ သံသယ လွန်မှုတွေ ထားတုန်း ဖြစ်တယ်လို့လည်း ဆင်ဟွာ သတင်း ဌာနက စွပ်စွဲ လိုက်ပါတယ်။ AIIB ဘဏ် သဘောတူညီချက်ကို အောက်တိုဘာလက တရုတ် အပါအဝင် ၂၁ နိုင်ငံ လက်မှတ် ထိုးခဲ့ ယူကေ နိုင်ငံဟာ 'အာရှ အခြေခံ အဆောက်အအုံ ရင်းနှီး မြှုပ်နှံမှု ဘဏ်'(AIIB) ရဲ့ အဖွဲ့ဝင် ဖြစ်လာဖို့ လျှောက်ထားမယ့် ပထမဆုံး အနောက် အုပ်စုဝင် စီးပွားရေး အင်အားကြီး နိုင်ငံ ဖြစ်ပါတယ်။”

“China, in response, dismissed the US's concerns, with the Xinhua news agency accusing the US of maintaining a childish level of suspicion when it comes to China. The AIIB agreement was signed by 21 countries, including China, in October. The UK is set to become the first major Western economy to apply for membership in the Asian Infrastructure Investment Bank (AIIB).”

In this case, it is seen that the outcome is the first and last sentences, which are directly selected from the original input article.

1) *Cosine Similarity*

Cosine similarity [6] is a useful metric for determining how similar two vectors are to one another in a highdimensional space. It is frequently used in natural language processing (NLP) to evaluate the semantic similarity of words, documents, and phrases. The cosine similarity approaches 1, indicating strong similarity, if the vectors are almost parallel. On the other hand, the similarity gets closer to zero if they are orthogonal. This measure is crucial to many NLP applications, including recommendation systems, summarization, and question-answering processes.

The following formula can be used to get the cosine similarity between the two vectors:

$$\mathit{cosine\ similarity} = \frac{v1 \cdot v2}{||v1||\ ||v2||} \quad (1)$$

where v1 and v2 are the vector representations of the sentences, and '·' denotes the dot product of two vectors. ||v1|| and ||v2|| are the Euclidean norms of the two vectors.

B. *Abstractive Summarization*

Abstractive text summarization involves condensing the original text by crafting new sentences that may not directly mirror the wording of the source material. Unlike extractive summarization, abstractive techniques are characterized by their complexity and flexibility, offering a more engaging and versatile approach to summarization. Consequently, there is a growing interest in exploring abstractive techniques across various languages. Based on current knowledge, there is a limited number of studies focusing on text summarization in the Myanmar language, with the majority being extractive in nature. This is partly due to the lack of proper Myanmar text datasets available for this task.

An example is provided to enhance the comprehension of the result derived from this method. The same input article in the extractive summarization session is tested for this method.

Abstractive summary:

"ယူကေ နိုင်ငံဟာ အာရှ ဒေသက အခြေခံ အဆောက်အအုံ ရင်းနှီး မြှုပ်နှံမှု ဘဏ်သစ် တခု ထူထောင်ဖို့ လျှောက်ထား လိုက်ပါတယ်။"

"The UK has applied to join a new infrastructure investment bank established in the Asian region."

In this case, the output is condensed into a summary that might not exactly match the original input sentences.

1) *Sequence-to-Sequence Architecture and Transformers*

Sequence-to-Sequence (Seq2Seq) [7] architectures play a pivotal role in the domain of abstractive text summarization, particularly in English contexts. Seq2Seq models— a subset of machine learning models— excel in converting input sequences into output sequences, which is essential for producing unique and cohesive text summaries. These models are composed of encoder and decoder components. The encoder processes the input sequences to produce fixed-length vector representations, which the decoder then uses to produce matching output sequences. While traditional architectures utilized RNN-based encoders, more recent developments, such as Transformers [8], have replaced them with self-attention layers, significantly enhancing training speed and parallelization. Transformer models — such as T5 [9] use multi-head self-attention layers to improve contextual understanding while maintaining the fundamental encoder-decoder architecture. T5 has proven to be effective in various downstream tasks, including text summarization.

IV. Proposed Systems

In this section, an overview of both extractive and abstractive summarizers is discussed.

A. *An overview of Extractive Summarization System*

Figure 1 illustrates the design of the proposed system architecture for Extractive Myanmar News Summarization. The proposed system has four main steps.

- Data Pre-processing
- Vectorization
- Computing Cosine Similarity Scores
- Sentence Ranking

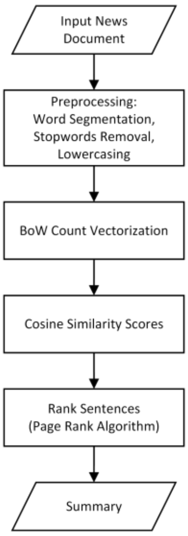

Figure 1 Proposed System Architecture for Extractive Myanmar News Summarization

1) *Data Pre-processing*

The input news document is firstly spitted into sentences. Unlike English, Myanmar text does not use white space to separate words. Moreover, the splitting of text is essential for all languages because it is the first step in linguistic processing. In this system, Myanmar news articles are segmented to words by the "Pyidaungsu" [10] Python library. Furthermore, stop words are unnecessary information and they are removed in the pre-processing step. Lastly, all the English words which are embedded within the Myanmar sentences are converted to lowercase to maintain consistency in text processing.

2) *Vectorization*

The vectorization method utilized for the Extractive Myanmar News Summarization is "Bag-of-Words (BoW) Count Vectorization" [11]. It involves counting the occurrences of each word in a document or a sentence and finally representing the text as a vector, where each component corresponds to the count of a particular word.

3) *Cosine Similarity Scores*

Cosine similarity [6] score is calculated to assess the similarity between sentences or document segments. Then, this similarity is used to select sentences for inclusion in the summary. The selection is done by measuring the word frequency vectors. This metric can be used to determine the relevance of each sentence to the overall document.

4) *Sentence Ranking*

Following the calculation of similarity scores between sentence pairs, these scores are utilized to construct a similarity matrix, which is subsequently converted into a graphical representation. Within this graph, each sentence is represented as a node, with edge weights denoting the similarity scores. The TextRank algorithm [12] assigns scores to each node based on its importance and connectivity, by iteratively updating them until convergence. Subsequently, the summary is generated by selecting the sentences with the highest TextRank scores, ensuring the inclusion of the most central and informative sentences.

B. An overview of Abstractive Summarization System

The architecture of the proposed abstractive text summarization system is shown in Figure 2. The preprocessing of the XL-Sum dataset is done only for the Myanmar language as the paper is focused solely on Myanmar language. After preprocessing the dataset, optimizing the mT5 [13] model on the prepared data, and generating summaries from the fine-tuned mT5 model is carried out.

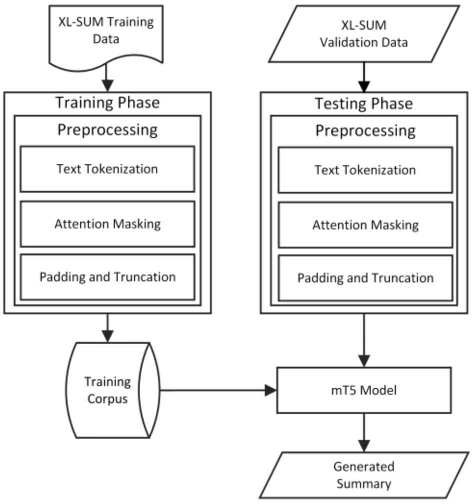

Figure 2 Proposed System Architecture for Abstractive Myanmar News Summarization

1) Dataset

The dataset used for abstractive summarization is XL-Sum [5] (released by Google Research), which was crawled from the BBC website. This large-scale dataset includes 1M professionally annotated article-summary pairs from 45 different languages and targeted at multilingual summarization. The XL-Sum is the first publicly available summarization dataset for various languages, fostering research opportunities, particularly for low-resource languages. For Myanmar language, the dataset is splitted into Train/Dev/Test sets, with 5761 articles in the training set and 719 each in the respective development and test sets. Key data fields include 'id', 'URL', 'title', 'summary', and 'text'.

2) mT5

The mT5 model is selected for fine-tuning in abstractive text summarization due to its unique characteristics compared to other pre-trained language models. The T5, mentioned in the background theories section, offers a unified Seq2Seq framework addressing a wide range of NLP challenges. Its extension, mT5 inherits all the capabilities of T5 and was trained on an expanded version of the C4 dataset, which encompasses over 10,000 web page contents across 101 languages, including Myanmar. When compared to alternative multilingual models like multilingual BERT and XLM-R, the mT5 consistently demonstrates superior performance across various tasks, particularly excelling in text summarization.

3) Tokenization

Data is tokenized by using the 250k wordpiece [14] vocabulary provided by the mT5 checkpoint. The methodology behind the mT5 tokenizer involves a comprehensive process to prepare text data for the mT5 model, particularly when applied to languages like Myanmar.

4) Fine-Tuning Configuration

To Fine-tune the mT5 models on the Xl-Sum Myanmar dataset introduced in the previous section, the Adafactor [15] optimizer is used with 90 warm-up steps with a batch size of 10 training epochs. The learning rate used for fine-tuning mT5 is $5e-4$.

V. Evaluation And Results

For evaluating the performance of the two architectures introduced in this paper, test data from XL-Sum Dataset is utilized and different ROUGE scores are computed for each method as an evaluation.

A. Evaluation

One of the metrics for evaluation of text summarization is recall-oriented understudy for gisting evaluation (ROUGE) [16]. It works by comparing human summary (one or several reference summaries) and system summary based on n-grams. The ROUGE is a popular method for calculating the efficiency of auto-generated summaries. Precision in ROUGE means that how much of the system summary was relevant. Recall in ROUGE simply means how much of the reference summary is the system summary recovering or capturing. The F1-SCORE in ROUGE means a balance between recall and precision, giving a single value to gauge overall performance. The Precision, Recall, and F1-score in ROUGE are determined as follows:

$$Recall = \frac{number\ of\ overlapping\ words}{total\ words\ in\ reference\ summary} \quad (2)$$

$$Precision = \frac{number\ of\ overlapping\ words}{total\ words\ in\ system\ summary} \quad (3)$$

$$F1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (4)$$

A ROUGE-1 refers to the overlap of unigrams between the system summary and the reference summary. A ROUGE-2 refers to the overlap of bigrams between the system and reference summaries. The overlap of unigrams, bigrams, trigrams, and higher order n-grams is measured by ROUGE-N. It relies on recall-based measures and compares n-grams between the system and reference summaries.

A ROUGE-L operates on the principle of finding the longest common subsequence (LCS) between two text sequences. It offers the advantage of not requiring consecutive matches but rather assessing in-sequence matches, capturing sentence-level word order. The ROUGE-Lsum first divides summaries into sentences, then assesses ROUGE-L scores for each sentence individually. It subsequently aggregates these scores and computes the average to determine the overall ROUGE-L Score.

One significant limitation of the ROUGE score is its inability to recognize different words that convey the same meaning. This is because ROUGE primarily focuses on matching syntax rather than understanding semantics. As a result, if two sequences express the same idea using different words, they may receive a low ROUGE score. Moreover, the current evaluation criteria heavily rely on reference summaries to judge the quality of generated summaries, which may not fully capture the essence of the summarization process. Despite these drawbacks, ROUGE provides a standardized and established metric for comparing automatically generated summaries to human-produced reference summaries, making it a common benchmark for evaluating summarization systems.

B. *Results*

In the results section, as for extractive summarization, the generated summary is compared with the original summary from the Xl-Sum dataset. As for abstractive summarization, each summary from fine-tuned monolingual mT5 and multilingual mT5 models are compared. The comparison result is showcased with different ROUGE scores as an evaluation.

1) Performance Evaluation of Extractive Myanmar News Summarization

For evaluating performance, the Extractive Myanmar News Summarization is applied to a set of 3,000 Myanmar news articles sourced from the XL-Sum Dataset, which contains annotated reference summaries. The ROUGE-1, ROUGE-2, and ROUGE-L scores of the system are shown in Table 1.

The extractive summarization system demonstrated a commendable performance with moderate Precision, Recall, and F-1 Score values across various ROUGE metrics. Notably, the system achieved significant scores for ROUGE-1, ROUGE-2, and ROUGE-L, indicating its effectiveness in capturing the essence of the original text. To further elevate its capabilities, integrating contextual embeddings or advanced deep learning models could enhance the system's ability to capture the nuanced context of the text.

Table 1 Experimental Results for Extractive Summarization

	Precision	**Recall**	**F1-Score**
ROUGE-1	39.73	51.24	43.14
ROUGE-2	15.91	23.85	18.1
ROUGE-L	27.77	35.72	30.04

2) Performance Comparison of Multilingual Model and the proposed Monolingual Model of Abstractive Summarization on XL-Sum Dataset

The comparison for Abstractive Summarization involves the implemented proposed Monolingual Model, which is fine-tuned solely on the Myanmar dataset from XL-Sum Dataset, and a Multilingual Model fine-tuned with all the 44 languages of XL-Sum Dataset as the baseline. The resulting ROUGE-1, ROUGE-2, and ROUGE-L scores of the systems are discussed in Table 2.

Table 2 Accuracies of multilingual model and monolingual model with XL-Sum Dataset

	Multilingual Model	**Monolingual Model**
ROUGE-1	15.96	36.67
ROUGE-2	5.15	16.78
ROUGE-L	14.18	29.98

In comparing the performance of the multilingual and monolingual models based on the ROUGE metrics, it is evident that the monolingual model outperforms the multilingual model. Specifically, the monolingual model achieves higher ROUGE scores, indicating its superior ability to capture the essence of the original text and generate more accurate and comprehensive summaries. This underscores the importance of language-specific fine-tuning for achieving optimal performance in text summarization tasks.

VI. Conclusion

This paper addresses the limited development in Myanmar text summarization by exploring both extractive and abstractive approaches. The previous researches mainly concentrated on extractive methods for text summarization with word embedding and headline generation through neural networks. However, no research has been done regarding the abstractive summarization by using pretrained models for Myanmar language. To address this problem, this paper used cosine similarity and TextRank for extractive summarization, and the multilingual mT5 for abstractive summarization with Myanmar category XL-Sum dataset to describe the nature of both methods.

The results of the proposed systems are quite satisfactory in terms of providing summaries for Myanmar news articles. Since there is a scarcity of research in the abstractive summarization for Myanmar, the proposed systems could not be compared to any earlier work and can now serve as a baseline for any future works in this field. Moreover, with this paper, one can find the differences in nature of how both methods generate different summaries even though same dataset is used.

Based on the findings of this study, one limitation was the size of the dataset used. Due to the lack of dataset availability for Myanmar text summarization, further studies with a larger sample size and a broader variety of text types are suggested.

There are some further exciting potentials for future research, based on the findings of this study. Due to the weakness discovered in the ROUGE measure, it would be highly significant to construct assessment criteria that encompass more than simply the reference description.

Appendix

Table 3 Summaries of news data generated by proposed systems

Input Article: IS ပိုင် ရေနံတွင်းတွေကို ဗြိတိသျှ တိုက်လေယာဉ် ၄ စီးက တိုက်ခိုက်ခဲ့ ဗုံးကြဲ တိုက်ခိုက်ဖို့ အတွက် ယူကေ အမတ်တွေ အတည်ပြုပြီး

မကြာမီမှာပဲ ဆိုက်ပရပ်စ်မှာ ရှိတဲ့ Akrotiri အက်ရော့တီရီ လေတပ်စခန်းက တိုနေးဒိုး တိုက်လေယာဉ် ၄ စီး တိုက်ခိုက်မှုမှာ ပါဝင်ခဲ့ပါတယ်။ "Shortly after UK lawmakers approved airstrikes, four British Tornado fighter jets from the Akrotiri airbase in Cyprus targeted ISIS-controlled oil fields in a bombing raid." ဆီးရီးယား အရှေ့ပိုင်းက IS တွေ ထိန်းချုပ်ထားတဲ့ ရေနံတွင်းတွေကို တိုက်ခိုက် ထိမှန်ခဲ့တယ်လို့ ကာကွယ်ရေး ဝန်ကြီးက ပြောပါတယ်။ "The defense minister stated that airstrikes successfully hit oil fields controlled by ISIS in eastern Syria." ဝန်ကြီးချုပ် ဒေးဗစ် ကင်မရွန်း ကတော့ IS အပေါ် တိုက်ခိုက်မှုတွေဟာ အချိန် ကြာမြင့်မှာ ဖြစ်ပြီး ပုံမှန် ပြုလုပ်သွားဖို့ လိုတယ်လို့ သတိပေး ပြောဆို ခဲ့ပါတယ်။ "Prime Minister David Cameron warned that the attacks on ISIS will be prolonged and must be carried out regularly." IS အပေါ် လေကြောင်း တိုက်ဖို့အတွက် ယူကေ ပါလီမန်မှာ ဗုဒ္ဓဟူးနေ့က ဆွေးနွေးငြင်းခုံပြီး မဲခွဲ ဆုံးဖြတ်ရာမှာ ထောက်ခံမဲ ၃၉၇ မဲ၊ ကန့်ကွက်မဲ ၂၂၃ မဲနဲ့ အတည်ပြုခဲ့တာ ဖြစ်ပါတယ်။ "The UK Parliament debated and voted on Wednesday regarding airstrikes on IS, and the decision was approved with 397 votes in favor and 223 votes against." အက်ရော့တီရီ လေတပ်စခန်းမှာ ရှိနှင့်ပြီးဖြစ်တဲ့ တိုက်လေယာဉ် ၈ စီး အပြင် နောက်ထပ် ၈ စီး ထပ်ပို့လိုက်တယ် လို့လည်း ယူကေ ကာကွယ်ရေး ဝန်ကြီး မိုက်ကယ် ဖာလန်က ပြောပါတယ်။ "UK Defense Secretary Michael Fallon said that in addition to the eight fighter jets already stationed at the Akrotiri airbase, another eight have been sent."
Human Summary: ဗြိတိသျှ တော်ဝင် လေတပ်ရဲ့ တိုနေဒိုး တိုက်လေယာဉ်တွေက ဆီးရီးယားမှာ ရှိတဲ့ IS အစ္စလာမ္မစ် နိုင်ငံအဖွဲ့ အပေါ် လေကြောင်း တိုက်ခိုက်မှုတွေ လုပ်ခဲ့တယ်လို့ ယူကေ ကာကွယ်ရေး ဝန်ကြီးဌာနက အတည်ပြု ပြောဆိုခဲ့ပါတယ်။ "The UK Ministry of Defense confirmed that Royal Air Force Tornado fighter jets carried out airstrikes on IS targets in Syria."
Extractive Summary: IS ပိုင် ရေနံတွင်းတွေကို ဗြိတိသျှ တိုက်လေယာဉ် ၄ စီးက တိုက်ခိုက်ခဲ့ ဗုံးကြဲ တိုက်ခိုက်ဖို့ အတွက် ယူကေ အမတ်တွေ အတည်ပြုပြီး မကြာမီမှာပဲ ဆိုက်ပရပ်စ်မှာ ရှိတဲ့ Akrotiri အက်ရော့တီရီ လေတပ်စခန်းက တိုနေးဒိုး တိုက်လေယာဉ် ၄ စီး တိုက်ခိုက်မှုမှာ ပါဝင်ခဲ့ပါတယ်။ "Shortly after UK lawmakers approved airstrikes, four British Tornado fighter jets from the Akrotiri airbase in Cyprus targeted ISIS-controlled oil fields in a bombing raid." ဆီးရီးယား အရှေ့ပိုင်းက IS တွေ ထိန်းချုပ်ထားတဲ့ ရေနံတွင်းတွေကို တိုက်ခိုက် ထိမှန်ခဲ့တယ်လို့ ကာကွယ်ရေး ဝန်ကြီးက ပြောပါတယ်။ "The defense minister stated that airstrikes successfully hit oil fields controlled by ISIS in eastern Syria."
Abstractive Summary: ဆီးရီးယား အရှေ့ပိုင်းမှာ IS အစ္စလာမ္မစ် နိုင်ငံအဖွဲ့ အပေါ် လေကြောင်း တိုက်ခိုက်မှုတွေ ဆက်လက် ဖြစ်ပွားနေတယ်လို့ ယူကေ ကာကွယ်ရေး ဝန်ကြီး မိုက်ကယ် ဖာလန်က ပြောပါတယ်။ "UK Defense Secretary Michael Fallon stated that airstrikes against IS in eastern Syria are continuing."

References

[1] S. S. Lwin and K. T. Nwet, "Single-Document Myanmar Text Summarization using Latent Semantic Analysis(LSA)," *Sixteenth International Conferences on Computer Applications(ICCA 2018)*, 2018.

[2] S. A. T. Oo, H. Hun, L. M. Htut, N. Lynn, S. K. H. Myo and K. K. Min, "Semantic Text Summarization for Myanmar Language using Word Embedding and Graph Based Approach," *2023 Seminar on Information Computing and Processing (ICP)*, pp. 186-191, 2023.

[3] S. S. Lwin and K. T. Nwet, "Myanmar News Summarization Using Different Word Representations," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 2285-2292, 2021.

[4] Y. Thu and W. P. Pa, "Generating Myanmar News Headlines using Recursive Neural Network," *18th International Conference on Computer Applications 2020*, pp. 199-205, 2020.

[5] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y. F. Li, Y. B. Kang, M. S. Rahman and R. Shahriyar, "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, p. 4693–4703, 2021.

[6] Wikipedia, "Cosine similarity - Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Cosine_similarity?form=MG0AV3. [Accessed 16 October 2024].

[7] I. Sutskever, O. Vinyals and Q. V. Le, " Sequence to Sequence Learning with Neural Networks," *Advances in Neural Information Processing Systems*, vol. 4, 2014.

[8] A. Vaswani, N. Shazeer, N. Parmar and J. Uszkoreit, "Attention Is All You Need," *Advances in Neural Information Processing Systems, 2017*, 2017.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1-67, 2020.

[10] K. H. San, "Pydiaungsu Library," 2020. [Online]. Available: https://github.com/kaunghtetsan275/pyidaungsu.

[11] V. Zhou, "A Simple Explanation of the Bag-of-Words Model," 2019. [Online]. Available: https://victorzhou.com/blog/bag-of-words/.

[12] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004.

[13] L. Xue, N. Constant, . A. Roberts, M. Kale, R. A. Rfou, A. Siddhant, . A. Barua and C. Raffel, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 483–498, 2021.

[14] X. Song, "A Fast WordPiece Tokenization System," Google Research Blog, 10 December 2021. [Online]. Available: https://research.google/blog/a-fast-wordpiece-tokenization-system/.

[15] N. Shazeer and M. Stern, "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost," *arXiv preprint*, 2018.

[16] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004.