Khmer News Headline Generation: A Comparative Evaluation of Extractive, Neural Network, and Transformer-Based Approaches

Bunheng Un
Applied Mathematics and Statistics
Institute of Technology of Cambodia
Phnom Penh, Cambodia
un_bunheng@gsc.itc.edu.kh

Sothea Has
Applied Mathematics and Statistics
Institute of Technology of Cambodia
Phnom Penh, Cambodia
has.sothea@itc.edu.kh

Sokkhey Phauk
Applied Mathematics and Statistics
Institute of Technology of Cambodia
Phnom Penh, Cambodia
phauk.sokkhey@itc.edu.kh

Pov Sorn
Applied Mathematics and Statistics
Institute of Technology of Cambodia
Phnom Penh, Cambodia
sorn_pov@gsc.itc.edu.kh

Hengborann Moul
Applied Mathematics and Statistics
Institute of Technology of Cambodia
Phnom Penh, Cambodia
moul_hengborann@gsc.itc.edu.kh

Phivath Sopheak
Applied Mathematics and Statistics
Institute of Technology of Cambodia
Phnom Penh, Cambodia
sopheak_phivath@gsc.itc.edu.kh

***Abstract*—Automatic headline generation remains a challenging task for low-resource languages such as Khmer due to the limited availability of annotated datasets and language-specific resources. This study investigates Khmer headline generation by comparing traditional extractive approaches, neural sequence-to-sequence, multilingual Transformer-based models. A Khmer news dataset of 65,393 article–headline pairs was collected from official Cambodian government websites and used to train and evaluate five models: Lead-1, TextRank, RNN-Attention, mT5, and mBART-50. Performance was evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore metrics. Results demonstrate that zero-shot Transformer models fail catastrophically on Khmer (mT5 ROUGE-1: 22.99; mBART-50 ROUGE-1: 12.55), while LoRA fine-tuned Transformers significantly outperform all extractive and recurrent approaches. Fine-tuned mT5 achieved the highest ROUGE-2 (65.17), ROUGE-L (66.68), and BERTScore (92.29), while fine-tuned mBART-50 obtained the highest ROUGE-1 (79.28), representing superior performance across all metrics. These results indicate that multilingual pre-trained language models are effective for low-resource headline generation and suggest that task-specific adaptation is essential to activate latent multilingual knowledge in low-resource settings. This approach can be applied to other low-resource headline generation tasks, and the benchmark is relevant for future Khmer NLP research.**

Keywords—Khmer official news, News Headline Generation; LoRA Fine-Tuning; Low-Resource Languages; Multilingual Transformers

I. Introduction

The rapid growth of digital news platforms has significantly increased the volume of information available to readers, and automatic headline generation has consequently attracted considerable attention as an important task in Natural Language Processing (NLP). Headlines play a major role in helping readers quickly identify relevant content [1], and effective headline generation offers substantial benefits for digital journalism, information retrieval, and content accessibility. Headline generation can be categorised into extractive approaches, which select important sentences from the original article (e.g., Lead-1, TextRank [2]), and abstractive approaches, which generate new text (e.g., sequence-to-sequence models with attention [3] and Transformer-based models [1]). Multilingual pre-trained models such as mT5 [4] and mBART-50 [5] are particularly promising for low-resource languages because they can transfer linguistic knowledge from high-resource languages.

Researchers have studied and improved many aspects of automatic headline generation. Mihalcea and Tarau [2] proposed TextRank for graph-based sentence ranking, Bahdanau et al. [3] introduced attention mechanisms for neural sequence-to-sequence generation, and Raffel et al. [1] presented the T5 text-to-text framework. Xue et al. [4] extended this to 101 languages with mT5, Tang et al. [5] developed mBART-50 as a multilingual denoising autoencoder, and Hu et al. [6] proposed LoRA for parameter-efficient fine-tuning. In Khmer NLP, existing work has addressed word segmentation [7], [8], sentiment analysis [9], and text classification [10], while headline generation has received limited attention [11], [12].

However, although considerable progress has been made for high-resource languages, Khmer remains a low-resource language in NLP. No study has systematically compared extractive, neural, and Transformer-based approaches for Khmer headline generation under a unified experimental setting, and the effectiveness of zero-shot versus fine-tuned multilingual Transformers for this task has not been evaluated.

Therefore, this study investigates Khmer headline generation by comparing Lead-1, TextRank, RNN-Attention, mT5-base, and mBART-50 on a dataset of 65,393 article–headline pairs using both lexical metrics (character-level ROUGE [13] and semantic metrics (BERTScore [14]). The contributions are: (1) providing a benchmark comparison of multiple headline generation approaches for Khmer, (2) assessing the effectiveness of multilingual Transformer models in a low-resource setting, and (3) contributing to the advancement of Khmer language technologies, supporting broader efforts toward digital inclusion and sustainable development.

II. Related Works

Automatic headline generation has attracted much attention in recent years as a specialised task within text summarisation, which plays a major role in enabling efficient information access from news articles. Existing approaches to this task can be broadly categorised into extractive methods, which select important words or sentences directly from the

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

source document, and abstractive methods, which generate new text that captures the semantic meaning of the original content.

Extractive methods identify the most salient content in a document and use it as the headline. The simplest approach, the Lead baseline, selects the opening sentence, which remains a strong baseline because journalists typically place key information at the beginning. Graph-based approaches such as TextRank [2] adapt the PageRank algorithm to rank sentences by inter-sentence similarity and select the highest-ranked one as the summary. TextRank has been shown to be competitive without labelled data and serves as a common benchmark. However, extractive methods are inherently limited to selecting existing text and cannot produce headlines that rephrase or compress the source wording.

The emergence of sequence-to-sequence (Seq2Seq) architectures with attention mechanisms [3] significantly advanced abstractive headline generation. Attention-based RNNs, particularly those using LSTM units, enabled decoders to focus on relevant parts of the source during generation and produced more natural headlines than extractive methods. However, RNN-based models suffer from sequential computation, difficulty modelling long-range dependencies, and performance degradation on lengthy documents. These limitations motivated the development of transformer-based architectures.

Transformers have become the dominant architecture for natural language generation due to their self-attention mechanism, which enables parallel processing and effective modelling of long-range dependencies. T5 [1] unified NLP tasks under a text-to-text framework, and its multilingual variant mT5 [4] extends this to over one hundred languages. BART [15] introduced denoising pretraining for seq2seq generation, and mBART-50 [5] extended this strategy to multilingual settings. Owing to their extensive pretraining and powerful contextual representations, transformer-based models are expected to outperform recurrent architectures in headline generation.

Khmer NLP research has grown considerably across foundational and applied tasks. For preprocessing, Chim et al. [7] applied deep learning to word segmentation, Taing and Li [8] proposed a bi-directional RNN segmented, and Chim et al. [16] released the largest morphologically annotated dataset for tokenization and POS tagging. For downstream tasks, Vann et al. [9] constructed a sentiment lexicon using PU learning, Kaing et al. [17] developed a BiLSTM-based sentiment analysis framework, Chhum et al. [10] demonstrated that segmentation-aware embeddings improve text classification, and Pottier et al. [18] built a Khmer NER benchmark. More recently, pretrained Khmer models have emerged: Jiang et al. [19] introduced KhmerBERT and KhmerELECTRA, and Kaing et al. [20] developed PrahokBART, a Khmer-specific seq2seq model incorporating linguistic components such as word segmentation, which outperforms multilingual baselines on generative tasks including headline generation. In summarization, Paul et al. [11] released a benchmark dataset, Navy et al. [12] evaluated extractive methods on Khmer news, and Palen-Michel and Lignos [21], [22] provided multilingual summarization resources and comparisons for less-resourced languages including Khmer.

However, although considerable progress has been made in Khmer NLP, studies on Khmer headline generation specifically remain scarce. Prior work either evaluates headline generation alongside other tasks without comparing across approach generations, or focuses on general summarisation rather than headline-specific evaluation. It remains unclear how traditional extractive methods such as Lead-1 and TextRank compare with neural abstractive models such as RNN-Attention and modern transformers such as mT5-base and mBART-50 under a unified experimental setting. There is a need for a comprehensive comparative evaluation of these approach families.

III. Methodology

This flowchart as shown in Fig. 1 illustrates the overall research methodology pipeline for Khmer news headline generation. The process begins with data collection and preprocessing, proceeds to model development comprising three approaches—extractive methods (Lead-1 and TextRank), a neural network method (RNN + Attention), and transformer-based methods (mT5 and mBART-50 with zero-shot and LoRA fine-tuning configurations), then moves through training and fine-tuning, followed by evaluation using metrics (ROUGE-1, ROUGE-2, ROUGE-L, BERTScore), and finally concludes with results and discussion.

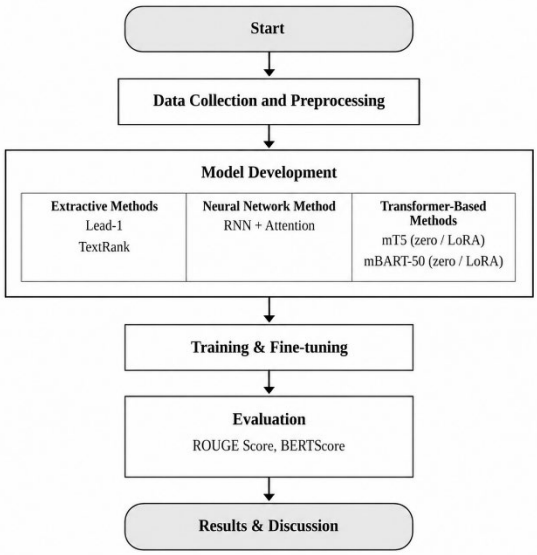


Fig. 1. Flowchart of the research for Khmer news headline generation.

A. Data sources

The dataset used in this study was collected from official news articles published by the Ministry of Information of Cambodia through its publicly accessible news. The ministry's website serves as the primary government communication channel, providing authoritative Khmer-language articles covering government activities, public policies, national events, diplomatic relations, economic development, social affairs, education, health, infrastructure, and other public-interest topics. The editorial style of the published articles follows a consistent format: each article is published with a human-written headline that concisely summaries the main content, followed by a body text that typically begins with a location and date attribution before proceeding to the substantive news content. This consistent structure ensures a reliable mapping between article bodies and their corresponding headlines, which is essential for training and evaluating automated headline generation systems.

B. Data Cleaning and Preprocessing

The raw news articles collected from the Ministry of Information of Cambodia underwent a series of data cleaning and preprocessing procedures to improve data quality and ensure consistency for model training. Since web-scraped data often contains formatting artifacts and duplicated content, preprocessing is a crucial step in preparing a reliable dataset for headline generation.

First, duplicate records were identified and removed to prevent repeated article–headline pairs from influencing the learning process. Articles with missing headlines, empty content fields, or incomplete text were also discarded. This step ensured that each data instance contained a valid news article and its corresponding headline. Next, HTML tags, hyperlinks, special symbols, and unnecessary formatting elements introduced during web scraping were removed from the article content. Unicode normalization was then applied to standardize Khmer character representations and eliminate inconsistencies caused by different encoding formats. Excessive whitespace, line breaks, and non-textual characters were also cleaned to improve text quality.

The cleaned articles and headlines were subsequently transformed into article–headline pairs, where the article body served as the input sequence and the corresponding headline served as the target sequence. For Transformer-based models, tokenization was performed using the model-specific tokenizer, converting text into subword units suitable for multilingual language modeling. Finally, the dataset was divided into training, validation, and testing subsets to facilitate model training, hyperparameter tuning, and performance evaluation.

These preprocessing procedures resulted in a high-quality Khmer news corpus consisting of 65,393 article–headline pairs, providing a reliable foundation for the headline generation experiments conducted in this study, as summarized in Table 1. The dataset was randomly split into training (80%), validation (10%), and testing (10%) sets at the article level.

TABLE I. SUMMARY STATISTICS OF DATASET

Summary Statistics	Article	Headline
Count	65,393	65,393
Mean	359	23
Std dev	370	8
Min	22	1
25th percentile	177	18
Median	316	22
75th percentile	463	28
Max	2,364	75

C. Model Selection

To comprehensively evaluate Khmer headline generation, this study selected five representative models spanning traditional extractive methods, deep learning approaches, and modern Transformer-based architectures.

1) Classical Methods

a) Lead-1

Lead-1 extracts the first sentence of the article as the predicted headline, exploiting the journalistic inverted pyramid structure in which key information is placed at the beginning [1]. For Khmer, sentences were segmented by splitting on the full stop character "។" (U+17D4) using a positive lookbehind regex (?<=។). The extracted sentence was truncated to 64 characters, which approximates nearly max tokens of reference headline lengths in the dataset (average 23 characters). If no sentence boundary was found, the first 64 characters of the article were used instead. This parameter-free baseline requires no training and provides a deterministic reference point.

b) TextRank

TextRank [2] ranks sentences using a graph-based adaptation of PageRank [23]. In this study, the article was first segmented into sentences on "។"; articles with fewer than two sentences fell back to Lead-1. Each sentence was represented as TF-IDF char-bigram vector, which was chosen because Khmer lacks word boundaries and character-level features avoid the need for word segmentation. A pairwise cosine similarity matrix was computed with zero diagonal, then row-normalized to form the transition matrix. PageRank iteration was run with damping factor 0.85, a maximum of 100 iterations, and convergence tolerance 10^{-4}. The top-ranked sentence (truncated to 64 characters) was returned as the headline.

2) Neural Network Model

The RNN-Attention model generates headlines abstractively using a sequence-to-sequence architecture [24] with Bahdanau attention [3]. Unlike the extractive baselines, this model can produce novel text that paraphrases the source article.

The encoder is a 2-layer bidirectional GRU [25] with 512 hidden units per direction and 128-dimensional character embeddings. Character-level tokenization was adopted to avoid dependency on Khmer word segmentation; the vocabulary was constructed from all unique characters in the training set plus four special tokens (<PAD>, <SOS>, <EOS>, <UNK>). The final forward and backward hidden states were concatenated and projected through a linear layer with tanh activation to produce the decoder initial state. Packed sequences were used to handle variable-length inputs efficiently. The attention mechanism is Bahdanau additive attention [3] with dimension 256, computing alignment scores between encoder states and the current decoder state. A source mask was applied to exclude padding positions. The decoder is a single-layer unidirectional GRU (512 units); at each step, the GRU input is the concatenation of the embedded previous token and the attention context vector. A deep output layer [26] combines the decoder state, context, and embedding to produce vocabulary logits, as illustrated in Table 2.

Training was performed using cross-entropy loss with padding tokens masked. The AdamW optimizer [27] was used with a learning rate of 3×10^{-4} and weight decay of 10^{-4}. Gradient clipping was applied at a maximum norm of 1.0 to prevent exploding gradients. Teacher forcing was annealed linearly from 0.9 to 0.1 over training to improve robustness at inference. A ReduceLROnPlateau scheduler (factor=0.5, patience=2) and early stopping (patience=5 epochs) were used to prevent overfitting. Model parameters were initialized using Xavier uniform initialization [28]. The maximum input length was set to 1,024 characters and the maximum target length to 64 characters. The batch size was 32, and training ran for a maximum of 30 epochs. At inference, beam search

with a width of 4 was used. Key hyperparameters are summarized in Table 2.

TABLE II. HYPERPARAMETERS OF RNN-ATTENTION

Component	Setting
Encoder	2-layer BiGRU, 512/dir, emb=128
Attention	Bahdanau, dim=256
Decoder	1-layer GRU, 512, deep output
Tokenizer	Character-level (from training set)
Optimizer / LR	AdamW / 3×10^{-4}
Teacher forcing	0.9 → 0.1 (linear annealing)
LR scheduler	ReduceLROnPlateau (factor=0.5, patience=2)
Early stopping	Patience 5 epochs
Max input / target	1,024 / 64 chars
Beam search	Width 4

3) Transformer-Based Models

a) mT5-base

mT5 [4] extends T5 [1] to 101 languages via span-corruption pretraining on mC4, including Khmer. The base variant (580M parameters, 12/12 encoder/decoder layers, 12 heads, dim 768) uses a SentencePiece tokenizer with 250K tokens, providing subword tokenization suited to Khmer morphology.

For zero-shot inference, articles were prefixed with "summarize: " following the T5 convention and decoded with beam search (width=4, length penalty=1.5, repetition penalty=1.2). A repetition penalty was included to mitigate repetitive loops common in low-resource language generation.

For LoRA fine-tuning, rank r=32, α=64 (α/r=2), and dropout=0.05 were applied to attention and feed-forward modules (q, k, v, o, wi_0, wi_1, wo). The relatively high rank was chosen to provide sufficient adaptation capacity for Khmer morphology. Training used AdamW [27] with lr=10^{-4}, cosine schedule, β1=0.9, β2=0.98, ε=10^{-6} (matching mT5 pretraining [4]); per-device batch 2 with gradient accumulation 4 (effective batch 8); warmup ratio 0.1; weight decay 0.01; max 10 epochs. Generation used beam search (width=4, length penalty=1.5, repetition penalty=1.2). Padding tokens in labels were masked with -100. Table 3 summarizes the configuration.

TABLE III. HYPERPARAMETERS OF MT5-BASE

Component	Setting
Base model	google/mt5-base (580M)
LoRA r / α / dropout	32 / 64 / 0.05
LoRA targets	q, k, v, o, wi_0, wi_1, wo
Task prefix	"summarize: "
Max input / target	1,024 / 64 tokens
Optimizer / LR	AdamW / 10^{-4} (cosine)
Effective batch / precision	8 / bfloat16
Early stopping	Patience 3 (ROUGE-L)
Beam / len penalty / rep penalty	4 / 1.5 / 1.2

b) mBART-50

mBART-50 [5] extends BART [15] to 50 languages via denoising pretraining. The large-50 variant (680M parameters, 12/12 layers, 16 heads, dim 1,024) uses a SentencePiece tokenizer with 250,048 tokens. Unlike mT5, mBART-50 specifies language via token identifiers rather than task prefixes; both source and target were set to "km_KH".

For zero-shot inference, the forced BOS token was set to "km_KH" to instruct the decoder to generate Khmer text. Beam search was used with width=4, max 64 tokens, length penalty=1.2, and repetition penalty=1.2.

For LoRA fine-tuning, the same rank and alpha as mT5-base were used for fair comparison: r=32, α=64, dropout=0.05, bias="none". Target modules were q_proj, k_proj, v_proj, out_proj, fc1, fc2 (naming follows the Hugging Face BART convention). A lower learning rate (3×10^{-5}) was used compared to mT5, as preliminary experiments showed that mBART-50's denoising pretraining required more conservative adaptation to avoid destabilizing the pretrained representations. The effective batch size was 32 (per-device batch 4, gradient accumulation 8). The no_repeat_ngram_size was set to 0 (disabled) because Khmer morphology involves legitimate n-gram repetitions that should not be penalized, as shown in Table 4 summarizes the configuration.

TABLE IV. HYPERPARAMETERS OF MBART50

Component	Setting
Base model	facebook/mbart-large-50 (680M)
LoRA r / α / dropout	32 / 64 / 0.05
LoRA targets	q_proj, k_proj, v_proj, out_proj, fc1, fc2
Source / Target language	km_KH / km_KH
Forced BOS token	km_KH
Max input / target	1,024 / 64 tokens
Optimizer / LR	AdamW / 3×10^{-5} (cosine)
Effective batch / precision	32 / bfloat16
Early stopping	Patience 3 (ROUGE-L)
Beam / len penalty / rep penalty	4 / 1.2 / 1.2

D. Evaluation Metrics

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) framework was adopted as the primary automatic metric due to its wide use in summarization research. This study reports the F1 scores of three ROUGE variants:

- ROUGE-1: Overlap of unigram tokens between generated and reference summaries.
- ROUGE-2: Overlap of bigram sequences, reflecting phrase-level accuracy.
- ROUGE-L: Longest Common Subsequence (LCS), measuring fluency and sentence-level structure.

ROUGE-N score is computed as:

$$ROUGE-N = \frac{\sum_{S\in\{ReferencesSummaries\}} \sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{ReferencesSummaries\}} \sum_{gram_n\in S} Count(gram_n)} \quad (1)$$

ROUGE-LCS F1 score is computed as:

$$ROUGE-L_{F1} = \frac{(1+\beta^2)\times Precision \times Recall}{Recall + \beta^2 \times Precision} \quad (2)$$

$$Precision = \frac{LCS(C,R)}{|C|} \quad (3)$$

$$Recall = \frac{LCS(C,R)}{|R|} \quad (4)$$

Where:

C : candidate (model-generated) summary
R : reference (human-written) summary

ROUGE is a commonly used baseline for summarization evaluation, but it relies on lexical overlap and cannot capture semantic similarity between generated and reference texts. To address this limitation, this study also uses BERTScore as a complementary metric.

BERTScore leverages contextual embeddings from pretrained transformer models to measure token-level semantic similarity using cosine similarity [14]. Unlike ROUGE, it matches words based on meaning rather than exact form, allowing it to reward semantically similar expressions even with different wording. This is particularly useful for Khmer, where multiple valid expressions can convey the same meaning.

Formally, given a candidate summary which is generated from the model $C = \{c_1, c_2, c_3 \dots, c_m\}$ and reference summary $R = \{r_1, r_2, r_3 \dots, r_n\}$, the contextual embeddings are first obtained from a pretrained model. The precision, recall, and F1 score are then computed as follows:

$$P_{BERT} = \frac{1}{|C|}\sum_{c_i \in C} \max_{r_j \in R} \cos(c_i, r_j) \quad (5)$$

$$R_{BERT} = \frac{1}{|R|}\sum_{r_j \in R} \max_{c_i \in R} \cos(c_i, r_j) \quad (6)$$

$$\mathrm{F_{BERT}} = 2 \times \frac{\mathrm{P_{BERT}} \times \mathrm{R_{BERT}}}{\mathrm{P_{BERT}} + \mathrm{R_{BERT}}} \quad (7)$$

Where $Cos(c_i, r_j)$ gives the cosine similarity between the contextual embedding of token c_i in the candidate summary and token r_j in the reference summary.

In this study, BERTScore was computed using xlm-roberta-base as backbone model with the language settings configured for Khmer (lang="km"). The choice of XLM-RoBERTa is well-suited for summarization due to its strong multilingual representation capabilities, having been pretrained on a large corpus spanning over 100 languages [29]. The F1 score of BERTScore (F_{BERT}) is reported as the primary BERTScore result, as it balances both precision and recall of semantic content between the generated and reference summaries.

IV. Results and Discussion

As described in Section 1, the aim of this study was to compare extractive, neural network model, and Transformer-based approaches for Khmer news headline generation. Five models such as Lead-1, TextRank, RNN-Attention, mT5 (LoRA), and mBART-50 (LoRA) were evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore on 3,271 test samples. Overall, the Transformer-based models achieved the highest scores, while extractive and neural models exhibited a tradeoff between surface-level overlap and semantic fidelity.

Table 1 presents the evaluation results. The Transformer-based models outperform all other approaches across every metric. mBART-50 achieves the highest ROUGE-1 (79.28), while mT5 leads on ROUGE-2 (65.17), ROUGE-L (66.68), and BERTScore (92.29). The two models differ by less than 1%, suggesting that both the text-to-text (mT5) and denoising autoencoder (mBART) pretraining objectives are equally effective when combined with LoRA fine-tuning. Among non-Transformer models, RNN-Attention achieves the highest BERTScore (88.44) while Lead-1 achieves the highest ROUGE-1 (70.29), a pattern consistent with the extractive–abstractive tradeoff [21].

TABLE V. Comparative Results of Khmer News Headline Generation Across Extractive, Neural Network, and Transformer-Based Models

Model	ROUGE -1	ROUGE -2	ROUGE -L	BERTScore
Lead-1	70.29	45.06	46.63	86.9
TextRank	68.33	37.72	40.96	85.43
RNN-Attention	65.82	48.95	52.68	88.44
mT5-base (Zero-shot)	22.99	10.11	16.4	76.9
mT5-base (Fine-tuned)	79.06	65.17	66.68	92.29
mBART-50 (Zero-shot)	12.55	7.44	9.33	74.38
mBART-50 (Fine-tuned)	79.28	64.82	66.37	92.19

Lead-1 outperforms TextRank across all metrics, with a particularly pronounced ROUGE-2 gap (45.06 vs. 37.72, a 16.3% relative difference, as shown in Table 1). This may be attributed to the editorial conventions of Khmer news writing, where the most newsworthy information is placed in the opening sentence following the inverted-pyramid structure [4]. Among abstractive models, RNN-Attention surpasses extractive methods on ROUGE-2 (48.95) and BERTScore (88.44) but underperforms Lead-1 on ROUGE-1 (65.82 vs. 70.29), suggesting that it generates semantically equivalent but lexically different output [21]. Both fine-tuned Transformers resolve this tradeoff, achieving high scores on both surface-level and semantic metrics simultaneously.

The superior performance of mT5 and mBART-50 can be attributed to their self-attention mechanisms, which capture contextual information across the entire article, and their multilingual pretraining, which enables knowledge transfer from other languages, a capability particularly valuable for low-resource Khmer. The near-parity between the two models is comparable to findings by Oguez et al. [30], who observed that different Transformer architectures converge to similar performance when fine-tuned for headline generation. However, occasional errors including information omission, word repetition, and factual inaccuracies were observed in the abstractive models, which is a common problem in neural text generation [22], ; representative output examples illustrating these failure modes are provided in the Table 6 of Appendix.

V. Conclusion

Prior research on low-resource language headline generation has reported that extractive methods tend to outperform abstractive methods on surface-level metrics, while abstractive methods achieve higher semantic fidelity [18, 19]. However, these studies have either focused on higher-resourced languages or have not investigated whether multilingual Transformer models can resolve this tradeoff for Khmer. In this study, five headline generation approaches, namely Lead-1, TextRank, RNN-Attention, mT5 with LoRA, and mBART-50 with LoRA, were compared on a dataset of 65,393 Khmer article and headline pairs. The results demonstrate that LoRA fine-tuned Transformer models

substantially outperform all other approaches, with mT5 and mBART-50 achieving ROUGE-1 scores of approximately 79 and BERTScore above 92 (Table 1). Both fine-tuned Transformers resolve the extractive and abstractive tradeoff, achieving high surface-level overlap and semantic fidelity simultaneously.

These findings extend those of Palen-Michel and Lignos [21], confirming that the extractive and abstractive tradeoff documented in less-resourced language summarization also applies to Khmer. The tradeoff pattern is consistent with results obtained in English summarization research [22], and the near-parity between mT5 and mBART-50 is comparable to Oguez et al. [30], who observed that different Transformer architectures converge to similar performance when fine-tuned for headline generation. Our results therefore lend support to the view that multilingual pretraining provides an effective foundation for low-resource NLP, and expand prior work by demonstrating that this holds for Khmer, a language with a non-Latin script and no word boundaries.

We have demonstrated that LoRA fine-tuning effectively activates the multilingual knowledge encoded during pretraining for Khmer headline generation, regardless of whether the pretraining objective is text-to-text (mT5) or denoising autoencoder (mBART). To our knowledge, this is the first study to investigate the comparative effectiveness of these approaches for Khmer headline generation. The resolution of the extractive and abstractive tradeoff by fine-tuned Transformers is a novel finding, which provides compelling evidence that multilingual pretrained models can remove the need to choose between surface-level and semantic quality in low-resource settings. Our study provides a framework for future research to assess Transformer-based headline generation in other low-resource languages, and the character-level evaluation methodology adopted here may serve as a reference for NLP research in languages without word boundaries.

However, some limitations are worth noting. Human evaluation of fluency, grammaticality, and factual consistency was not conducted; generalizability to other domains was not investigated; and LoRA configurations were not extensively tuned. Further work is needed to incorporate human evaluation, and future work should explore domain adaptation across diverse Khmer text types and validate these results with Khmer-specific pretrained models such as PrahokBART. The approach presented in this study has potential for headline generation in other low-resource languages, and could eventually lead to practical applications in automated news production for underserved language communities.

ACKNOWLEDGMENT

I would like to sincerely thank Dr. HAS Sothea and Dr. PHAUK Sokkhey for their guidance, encouragement, and continuous support throughout this research on headline generation. Their expertise and patient supervision played a key role in shaping the direction of this study and bringing it to completion.

I am also grateful to Ministry of Information of Cambodia (information.gov.kh) for making their news articles publicly available, which served as the primary data source for this research, and to the developers of the open-source tools used in this work, including Hugging Face Transformers, ROUGE, BERTScore, and khmernltk. I further acknowledge the researchers behind the pretrained models mT5, mBART-50, and PrahokBART.

REFERENCES

[1] C. Raffel, N. Shazeer, A. Roberts, et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, no. 1, pp. 5485–5551, 2020.

[2] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in Proc. EMNLP, 2004.

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

[4] L. Xue, N. Constant, A. Roberts, et al., "mT5: A massively multilingual pre-trained text-to-text transformer," arXiv:2010.11934, 2021.

[5] Y. Tang, C. Tran, X. Li, et al., "Multilingual translation with extensible multilingual pretraining and finetuning," Trans. Assoc. Comput. Linguist., vol. 9, pp. 368–383, 2021.

[6] E. J. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-rank adaptation of large language models," in Proc. ICLR, 2022.

[7] C. Chim, R. Sagnes, T. Le, and C. Pottier, "Khmer word segmentation using deep learning," in Proc. ACOMP, 2020, pp. 140–145.

[8] N. Taing and H. Li, "Khmer word segmentation with bi-directional recurrent neural network," J. Intell. Fuzzy Syst., vol. 40, no. 4, pp. 7429–7440, 2021.

[9] R. Vann, S. Chhoeun, and S. Sambath, "KhmerNLP: A comprehensive toolkit for Khmer natural language processing," in Proc. EMNLP Syst. Demonstrations, 2022, pp. 232–240.

[10] S. Chhum, S. Heng, and N. Taing, "Khmer text classification using word embedding and neural networks," arXiv:2112.06748, 2022.

[11] S. Paul, P. Navy, and S. Chhoeun, "Benchmark dataset for automatic text summarization in Khmer," arXiv:2106.02345, 2021.

[12] P. Navy, S. Paul, and C. Pottier, "Automatic text summarization for Khmer news articles using extractive methods," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 2, pp. 45–53, 2023.

[13] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Text Summarization Branches Out, 2004, pp. 74–81.

[14] T. Zhang, V. Kishore, F. Wu, et al., "BERTScore: Evaluating text generation with BERT," arXiv:1904.09675, 2020.

[15] M. Lewis, Y. Liu, N. Goyal, et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv:1910.13461, 2020.

[16] C. Chim, R. Sagnes, T. Le, and C. Pottier, "Towards tokenization and part-of-speech tagging for Khmer," in Proc. PACLIC, 2021, pp. 510–519.

[17] H. Kaing, S. Ourng, and V. Phan, "A BiLSTM-based sentiment analysis scheme for Khmer NLP in time-series data," IEEE Access, vol. 12, pp. 1–12, 2024.

[18] C. Pottier, T. Le, and R. Sagnes, "Building a Khmer NER benchmark from health news data," in Proc. IALP, 2023, pp. 1–6.

[19] S. Jiang, S. Fu, N. Lin, and Y. Fu, "Pretrained models and evaluation data for the Khmer language," Tsinghua Sci. Technol., vol. 27, no. 4, pp. 709–718, 2022.

[20] H. Kaing, R. Dabre, H. Song, et al., "PrahokBART: A pre-trained sequence-to-sequence model for Khmer natural language generation," in Proc. COLING, 2025, pp. 1–14.

[21] C. Palen-Michel and C. Lignos, "LR-Sum: Summarization for less-resourced languages," in Findings EMNLP, 2023, pp. 1–15.

[22] C. Palen-Michel and C. Lignos, "Comparing approaches to automatic summarization in less-resourced languages," arXiv:2512.24410, 2025.

[23] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Comput. Netw. ISDN Syst., vol. 30, no. 1–7, pp. 107–117, 1998.

[24] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Adv. Neural Inf. Process. Syst., vol. 27, 2014.

[25] K. Cho, B. van Merriënboer, C. Gulcehre, et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv:1406.1078, 2014.

[26] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in Proc. ICML, 2013, pp. 1310–1318.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.

[28] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proc. AISTATS, 2010, pp. 249–256.

[29] A. Conneau, K. Khandelwal, N. Goyal, et al., "Unsupervised cross-lingual representation learning at scale," arXiv:1911.02116, 2020

[30] F. Oguez, C. Grozea, and C. Muller, "Towards multi-lingual headline generation for low-resource languages," in Proc. LREC, 2022, pp. 5401–5408.

APPENDIX

TABLE VI. COMPARATIVE OUTPUT EXAMPLE OF KHMER NEWS HEADLINE GENERATION ACROSS EXTRACTIVE, NEURAL NETWORK, AND TRANSFORMER-BASED MODELS

Model	Output Example
Article	ក្នុងខណៈដឹកនាំគណៈប្រតិភូជាន់ខ្ពស់នៃរាជរដ្ឋាភិបាលកម្ពុជាបំពេញទស្សនកិច្ចផ្លូវការនៅលើទឹកដីនៃសាធារណរដ្ឋប្រជាធិបតេយ្យប្រជាមានិតឡាវនោះ នារសៀលថ្ងៃទី១៣ ខែកុម្ភៈ ឆ្នាំ២០២៣ សម្តេចអគ្គមហាសេនាបតីតេជោ ហ៊ុន សែន នាយករដ្ឋមន្ត្រី នៃព្រះរាជាណាចក្រកម្ពុជាបានអញ្ជើញចូលជួបសម្តែងការគួរសម ជាមួយឯកឧត្តមបណ្ឌិត Xaysomphone Phomvihane ប្រធានរដ្ឋសភា នៃសាធារណរដ្ឋប្រជាធិបតេយ្យប្រជាមានិតឡាវ។ នេះបើតាមការឱ្យដឹងរបស់អ្នកយកព័ត៌មានរបស់ទូរទស្សន៍ជាតិកម្ពុជា។ក្នុងជំនួបនោះ ឯកឧត្តមបណ្ឌិតប្រធានសភាឡាវ បានមានប្រសាសន៍សម្តែងនូវការស្វាគមន៍យ៉ាងកក់ក្តៅបំផុត ជូនចំពោះសម្តេចតេជោនាយករដ្ឋមន្ត្រី និងគណៈប្រតិភូជាន់ខ្ពស់រាជរដ្ឋាភិបាលកម្ពុជា ដែលបានអញ្ជើញបំពេញទស្សនកិច្ចផ្លូវការ នៅសាធារណរដ្ឋប្រជាធិបតេយ្យប្រជាមានិតឡាវ។ ឯកឧត្តមប្រធានរដ្ឋសភា បានចាត់ទុកថា ដំណើរបំពេញទស្សនកិច្ចផ្លូវការរបស់សម្តេចតេជោទៅកាន់ប្រទេសឡាវនៅពេលនេះ នឹងជួយជំរុញនិងពង្រីកបន្ថែមទៀតនូវទំនាក់ទំនងកិច្ចសហប្រតិបត្តិការគ្រប់ជ្រុងជ្រោយ ឱ្យកាន់តែរឹងមាំទ្វេឡើងថែមទៀត ក្នុងនាមជាប្រទេសភូមិឆងរបងជាមួយ និងជាប្រទេសជន្លងជាមួយគ្នា ដែលមានប្រវត្តិដ៏យូរលង់ណាស់មកហើយនោះ។ឯកឧត្តមប្រធានរដ្ឋសភា បានផ្ញើសាកសួរសុខទុក្ខជូនចំពោះ សម្តេចអគ្គមហាពញាចក្រី ហេង សំរិន ប្រធានរដ្ឋសភា សម្តេចវិបុលកត្តិ សាយ ឈុំ ប្រធានព្រឹទ្ធសភា នៃព្រះរាជាណាចក្រកម្ពុជា ។សម្តេចតេជោនាយករដ្ឋមន្ត្រីមានប្រសាសន៍ ថ្លែងអំណរគុណ ជូនចំពោះឯកឧត្តមប្រធានសភាឡាវ ដែលបានទទួលជួបធ្វើកិច្ចសន្ទនាជាមួយសម្តេច។ សម្តេចតេជោនាយករដ្ឋមន្ត្រីមានប្រសាសន៍ចាត់ទុកថា ដំណើរទស្សនកិច្ចរបស់សម្តេចទៅកាន់ប្រទេសឡាវនៅពេលនេះ នឹងជួយជំរុញកិច្ចសហប្រតិបត្តិការលើគ្រប់វិស័យ នៃប្រទេសទាំងពីរឱ្យកាន់តែ រឹងមាំទ្វេឡើងថែមទៀត។ សម្តេចតេជោ បានមានប្រសាសន៍អរគុណចំពោះស្ថាប័ននិតិប្រតិបត្តិឡាវ ដែលបានគាំទ្រដល់រដ្ឋាភិបាលឡាវ ក្នុងការផ្តល់សច្ចាប័នផ្លូវច្បាប់នានា ក្នុងការរួមចំណែកជួយដល់ការស្តារនិងអភិវឌ្ឍន៍សេដ្ឋកិច្ចរបស់ប្រទេសកម្ពុជា។បន្ទាប់មកនៅរាត្រីថ្ងៃដដែលនោះដែរ សម្តេចអគ្គមហាសេនាបតីតេជោ ហ៊ុន សែន នាយករដ្ឋមន្ត្រីនៃព្រះរាជាណាចក្រកម្ពុជា បានអនុញ្ញាតឱ្យប្រតិភូសមាគមមិត្តភាព ឡាវ-កម្ពុជា ចូលជួបសម្តែងការគួរសម។ សម្តេចតេជោនាយករដ្ឋមន្ត្រីមានប្រសាសន៍ថា កម្ពុជាមានសេចក្តីសោមនស្សរីករាយចំពោះភាពរីកចម្រើន នៃកិច្ចសហប្រតិបត្តិការរវាងសមាគមមិត្តភាព ឡាវ-កម្ពុជា ដែលបានលះបង់នូវពេលវេលាដ៏មានតម្លៃ ក្នុងការជួយគ្នាទៅវិញទៅមក ពិសេសទំនាក់ទំនងរវាងប្រជាជន និងប្រជាជន នៃប្រទេសទាំងពីរ។ សម្តេចតេជោនាយករដ្ឋមន្ត្រីបានលើកឡើងថា សមាគមមិត្តភាព ឡាវ-កម្ពុជា និងសមាគមមិត្តភាពកម្ពុជា-ឡាវ បានធ្វើកិច្ចការជាច្រើនលើផ្នែកមនុស្សធម៌ និងការជួយគ្នាទៅវិញទៅមក។ **Source:** J. Narı and K. Chandan, "Cambodia-Laos leadership express satisfaction with strong friendly relations and good cooperation between the two countries," *Ministry of Information, Kingdom of Cambodia*, Feb. 14, 2023. [Online]. Available: https://www.information.gov.kh/articles/97452
Reference	សម្តេចតេជោនាយករដ្ឋមន្ត្រី ហ៊ុន សែន អញ្ជើញជួបសម្តែងការគួរសម ជាមួយប្រធានសភានៃសាធារណរដ្ឋប្រជាធិបតេយ្យប្រជាមានិតឡាវ
Lead-1	ក្នុងខណៈដឹកនាំគណៈប្រតិភូជាន់ខ្ពស់នៃរាជរដ្ឋាភិបាលកម្ពុជាបំពេញទស្សនកិច្ចផ្លូវការនៅលើទឹកដីនៃសាធារណរដ្ឋប្រជាធិបតេយ្យប្រជាមាន
TextRank	ក្នុងជំនួបនោះ ឯកឧត្តមបណ្ឌិតប្រធានសភាឡាវ បានមានប្រសាសន៍សម្តែងនូវការស្វាគមន៍យ៉ាងកក់ក្តៅបំផុត ជូនចំពោះសម្តេចតេជោ នាយករដ្ឋមន្
RNN-Attention	សម្តេចតេជោ ហ៊ុន សែន អញ្ជើញចូលជួបសម្តែងការគួរសមជាមួយប្រធានរដ្ឋសភា
mT5-base (Fine-tuned)	សម្តេចតេជោ ហ៊ុន សែន អញ្ជើញចូលជួបសម្តែងការគួរសម ជាមួយប្រធានរដ្ឋសភា នៃសាធារណរដ្ឋប្រជាធិបតេយ្យប្រជាមានិតឡាវ
mBART-50 (Fine-tuned)	សម្តេចតេជោ ហ៊ុន សែន អញ្ជើញចូលជួបសម្តែងការគួរសម ជាមួយប្រធានរដ្ឋសភា នៃសាធារណរដ្ឋប្រជាធិបតេយ្យប្រជាមានិតឡាវ