Fast Hybrid Approach for Thai News Summarization

Kietikul Jearanaitanakij[1,*], Suratan Boonpong[1], Kirttiphoom Teainnagrm[1], Thanakrit Thonglor[1],
Tiwat Kullawan[2] and Chankit Yongpiyakul[2]

[1]Department of Computer Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Latkrabang, Latkrabang, Bangkok, 10520, Thailand

[2]Dataxet Limited, Lumpini, Patumwan, Bangkok, 10330, Thailand

***Corresponding Author E-mail:** kietikul.je@kmitl.ac.th

Received: May 03, 2024; **Revised:** Jul 19, 2024; **Accepted:** Aug 09, 2024

Abstract

News summarization presents a significant challenge in Natural Language Processing (NLP). Lengthy news articles not only consume valuable time but also lead to confusion regarding key points. The ideal news summarization should swiftly produce a succinct summary while retaining the essence of the information conveyed by the news writer. While intelligent chatbots like ChatGPT and Gemini offer user-friendly text summarization, their embedded Large Language Model (LLM) cannot be downloaded for private use. Moreover, implementing them in a business process can be expensive, both in terms of pay-per-use costs and response time. The objective of this research is to develop a private Thai news summarization that effectively extracts sentences encapsulating the main idea and abstractly summarizes them. The proposed model consists of two components. The first extracts a contiguous region containing important sentences using the TextRank algorithm, while the second employs the finetuned mBART as an LLM to generate the abstractive summary from the previously extracted sentences. In other words, the proposed model extracts an important news region before passing it to mBART. This approach produces a news summary with key information and a syntactic style akin to the natural Thai language. We evaluate the summarization quality by ROUGE scores and BERTScore (precision, recall, and F1-score). The evaluation metrics Experimental results on the ThaiSum dataset show relatively high ROUGE scores and BERTScore for the proposed model compared to most of the other approaches. Furthermore, it significantly reduces the runtime, keeping it within a reasonable limit.

Keywords: News summarization, Natural language processing, TextRank, TF.IDF, mT5 model, mBART model

1. Introduction

News undoubtedly plays a crucial role in our daily lives. Reading lengthy news articles to grasp the main points can be time-consuming. Hence, there is a need for a system that can summarize news, highlighting only the key points and maintaining correct grammar. In the context of Thai news summarization, language complexity poses a significant challenge. The length of Thai phrases often complicates summarization while the subtle nuances between similar Thai words make it difficult to accurately measure the similarity of two summaries. Moreover, there is a scarcity of annotated datasets and pre-trained models for Thai, compared to more commonly spoken languages. According to a survey by Allahyari et al. [1], text summarization can be broadly divided into two strategies: extraction and abstraction.

Extractive summarization involves selecting sentences or phrases that encapsulate the main idea without altering

or rewriting the extracted text. While this method is relatively easy to implement, the syntax of its summary may lack the lexical flow or coherence found in summaries manually crafted by humans. On the other hand, the abstractive model linguistically creates logical texts to condense the original posts into a well-structured summary. However, the abstractive approach may generate a summary that includes extraneous information.

The literature reviews of text summarization are presented starting with the extractive strategy. Wong et al. [2] employed a method that extracts sentences by amalgamating various features such as extraneousness, content conveyance, events, and relevance. They found that both supervised and semi-supervised learning models yield similar performance. Qazvinian et al. [3] introduced a graph-based extractive summarization to generate a summary for a scientific document. The cited sentences are modeled as a graph, with their vertices and edges representing sentences and lexical similarities, respectively. They discovered that the citation contains vital information that is not accessible from other sections of the paper. Jain et al. [4] used a neural network to predict sentences that correspond to the summaries of the provided text. Ten features, such as sentence length, position, and cohesion, are gathered from the document and used as inputs for the neural network. Despite their method requiring numerous precalculated features, the experimental results of their strategy outperform other online summarizer tools in terms of ROUGE-1, ROUGE-2, and ROUGE-L.

The abstractive approach is another category of text summarization. Nallapati et al. [5] utilized attentional encoder-decoder recurrent neural networks to abstractly summarize a document. Their method significantly outperforms others on two datasets, and they also introduced a new dataset for the abstractive summarization task requiring multiple-sentence summaries. Tan et al. [6] introduced a graph-based attention mechanism for the seq2seq model to identify the salient information of a document. Their hierarchical decoder with a reference mechanism effectively produces abstractive summaries. When compared to the experimental results on two large datasets (CNN and DailyMail), their method yields superior ROUGE results than existing neural abstractive models. Gehrmann et al. [7] developed a bottom-up content selector for abstractive summarization. Their content selector not only selects the appropriate content but also requires a minimal number of training instances. The experimental results on CNN-DM and NYT corpora show a significant improvement over existing methods. Two convolutional-seq2seq-based models are demonstrated by Zhang et al. [8] and Hao et al. [9]. The first model generates keywords and key sentences simultaneously using a hierarchical attention mechanism, while the second architecture employs both memory and nonlocal networks to construct the feature-enhanced seq2seq model. Liu et al. [10] pretrained mBART by applying the BART model [11] to extensive monolingual corpora across multiple languages. mBART (Multilingual BART) is a versatile sequence-to-sequence model that can be fine-tuned for various tasks, including text summarization, across different language pairs. It features a transformer-based encoder-decoder architecture, pre-trained on extensive monolingual corpora using a denoising objective. This pre-training helps the model generate coherent and contextually accurate summaries. Consequently, mBART excels in multilingual tasks such as translation and summarization. Yang et al. [12] developed a generative adversarial model for abstractive summarization with multi-task constraints. Their architecture effectively generates grammatically correct and informative summaries. The experimental results on CNN/Daily Mail and Gigaword corpora indicate superior performance over baseline methods. Yang et al. [13] emulated three stages of human reading: skim, in-depth reading, and post-

editing. The post-editing stage involves a refining process to make the information in the summary clear and salient. Interestingly, the result from human evaluation shows that summaries from their architecture have better informativeness and fluency compared to baseline methods. Xue et al. [14] proposed a multilingual version of the T5 model [15] called mT5 that covers 101 languages by pretraining in a similar procedure as the original T5 but on the mC4 corpus. The mT5 model has achieved state-of-the-art performance on many cross-lingual NLP tasks, including the XTREME zero-shot classification, structure prediction, and question-answering tasks. Its encoder-decoder is fine-tuned for Thai sentence summarization, named "mt5-cpe-kmutt-thai-sentence-sum", and implemented in PyThaiNLP [16]. Ngamcharoen et al. [17] applied a bidirectional LSTM to predict keywords for use along with the original text to generate a summary abstractly. The experimental result on the ThaiSum dataset indicates that their model outperforms the traditional encoder-decoder model.

In this paper, we introduce a fast hybrid approach that merges both extractive and abstractive strategies. Initially, we identified all significant sentences in the news using TextRank. Sentences with a high TextRank score are typically important due to their relevance to other sentences. Subsequently, the smallest contiguous region, extending from the first sentence to a position near the last significant sentence, is extracted. These extracted sentences are fed into the abstractive summarizer, i.e., the finetuned mBART model, to generate the final summary. Our hybrid approach introduces a novel technique by leveraging TextRank to extract a crucial news region before feeding it into the mBART model. This combination aims to enhance the runtime while the correctness of the news summary is still preserved.

Our research contributes in three ways. First, the proposed model consumes significantly less runtime than other leading abstractive models. Second, the quality of news summaries it produces is promising, in terms of ROUGE, precision, recall, and F1 score, and competitive with those of other models. Lastly, it is well-suited for summarizing Thai news as it was finetuned on the ThaiSum dataset. The experimental results show a notable improvement over baseline and other models.

The remainder of this paper is structured as follows. Section 2 provides a brief explanation of the fundamental knowledge necessary to comprehend the contents of this research. The proposed model and the experimental results are discussed in Sections 3 and 4, respectively. Finally, Section 5 presents the conclusion and potential future work.

2. Fundamental Knowledge

Five fundamental concepts and the dataset related to this research are briefly described in this section.

2.1 Thai Word Segmentation

A sentence is made up of a sequence of words. Unlike many languages, the Thai language does not use spaces to separate words within a sentence. This presents a significant challenge for both novices and researchers in computational linguistics. A straightforward approach is to perform maximal matching, which involves searching for the longest prefix characters in the dictionary. In essence, the longest string of characters found in a dictionary is recognized as a single Thai word.

PyThaiNLP [16] offers three famous implementations of Thai word segmentation, namely AttaCut [18], Deepcut [19], and Newmm. The Newmm algorithm is simple and deterministic, while the other two methods segment a sentence into words based on the surrounding context. Consequently, the same string of characters may not be tokenized into the same sequence of words when it appears in different contexts.

2.2 TextRank

Mihalcea and Tarau [20] introduced a graph-based algorithm for identifying important keywords and

sentences from the source text. This algorithm considers both the local context and the global information among words/sentences throughout the text. A key advantage of this algorithm is that it doesn't require linguistic knowledge for keyword and sentence extraction. The algorithm divides the text into tokens (a token can be either a word or a sentence) and computes similarities between vector embeddings of tokens in a matrix format. From this similarity matrix, a graph is created, with vertices and edges represented by tokens and the similarities between tokens, respectively. Finally, the PageRank algorithm [21] is applied to this graph to obtain token scores for the token T_i, as shown in Eq. (1). The top N tokens based on their scores are selected to form a set of important tokens.

$$Score(T_i) = (1-d) + d * \sum_j \frac{w_{ji}}{\sum_k w_{jk}} Score(T_j), \quad (1)$$

where T_j is the neighbor of the token T_i, d is a damping factor used in the PageRank algorithm ($d=0.85$), w_{ji} is the weight between the token T_i and its predecessor T_j, and w_{jk} is the weight between the token T_j and its successor T_k. TextRank can be utilized to identify significant sentences from an article by treating sentences as tokens and forming a similarity matrix among them.

2.3 Term Frequency-Inverse Document Frequency (TF.IDF)

TF.IDF is a measure of the statistical relevance of a specific word in a document, relative to all documents in a corpus [22]. It is calculated as the product of the term frequency and the inverse document frequency. The term frequency (TF) quantifies how frequently a particular word appears in a document relative to other words. However, high-frequency words may not necessarily be important to the document as they could be stop words. This is where the inverse document frequency (IDF) comes into play. IDF is defined as the logarithm of the ratio of the corpus size to the number of documents in which a specific word appears. In other words, a word that appears infrequently in a corpus is likely to be important. The formula for calculating the TF.IDF of a given term i in document j is formally defined in Eq. (2)–(4).

$$TF.IDF_{ij} = TF_{ij} \times IDF_i, \quad (2)$$

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}, \quad (3)$$

$$IDF_i = \log_2 \frac{N}{n_i}, \quad (4)$$

where f_{ij} is the number of appearances of term i in document j, and n_i is the number of appearances of term i in a pile of N documents. Note that TF_{ij} is normalized by the maximum number of appearances of any term k in the same document.

2.4 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

Lin [23] introduced ROUGE as a metric for evaluating the quality of text summarization, including machine translation, in comparison to reference summaries. Three well-known ROUGE metrics utilized in our research are ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 and ROUGE-2 assess the overlap of unigrams and bigrams between the generated summaries and the reference summaries, respectively. ROUGE-L evaluates the overlap of the longest common subsequence of words between the generated summaries and the reference summaries. Unlike ROUGE-1 and ROUGE-2, ROUGE-L can measure the nonconsecutive matching of the word sequence, allowing it to assess similarity at the sentence-level structure.

2.5 The ThaiSum Dataset

The ThaiSum dataset [24] is a collection of 3GB of data from online news websites such as Thairath, Thai PBS, Prachatai, and The Standard, gathered between January 2014 and August 2020. It consists of 358,868 pairs of original news and its summary, all written by professional writers. The average word lengths of the news and summary are 529.5 and 37.3, respectively. The dataset covers a wide variety of news topics such as politics, business, economics, and sport.

3. Proposed Model

The proposed Thai news summarization process is depicted in **Figure 1**. Initially, we tokenize Thai news into a sequence of sentences using a Thai-Segmenter from the Python Package Index (PyPI). Subsequently, we use TextRank to measure the importance of each sentence. Sentences that rank above the average are deemed important and are selected (these are represented in red sentences in **Figure 2**).

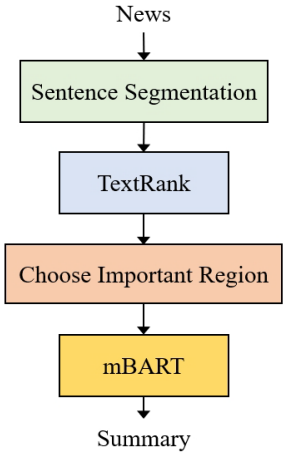

Figure 1 Proposed Thai news summarization model

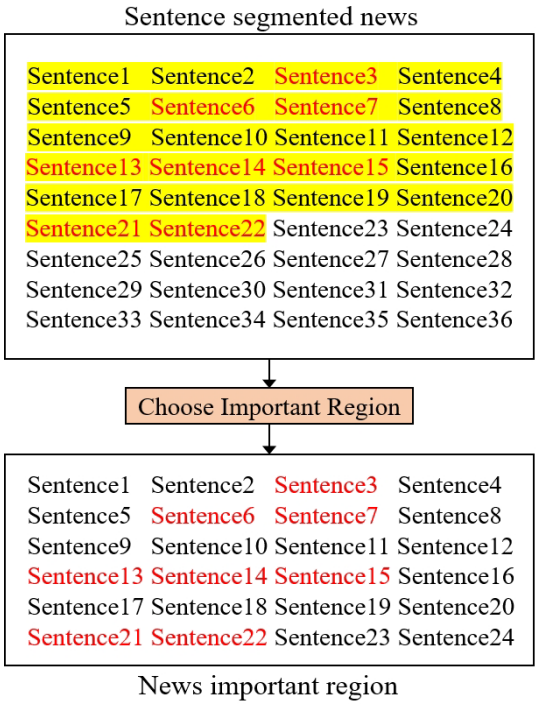

Figure 2 Selecting the important news region

According to Chumpolsathien [24], the majority of the significant sentences extracted from a document are typically found at the beginning. We select a series of sentences, from the first to the last significant sentences, as the important region (highlighted in yellow in **Figure 2**). It's worth noting that the contiguous news region may include sentences that are not important, as we aim to maintain a smooth flow of content within the segment. For completeness, a few sentences following the last significant sentence are also included in the important region, as shown at the bottom of **Figure 2**. Sentences 23 and 24 are included in the important region, while sentences 25–36 are omitted.

Finally, the important region is fed into the finetuned mBART to generate an abstractive summary. We selected mBART as our news summarizer because it was trained as a multilingual sequence-to-sequence model with a primary focus on translation tasks. Our use of the TextRank technique to reduce the news input to mBART plays a crucial role in decreasing the runtime required to generate an abstractive summary.

4. Experiments

The experimental setup and results are given in sections 4.1 and 4.2, respectively.

4.1 Experimental Setup

The experiments in this section are carried out on the ThaiSum dataset using a machine with the following specifications: CPU Core i5-13500 4.8 GHz, RAM 64GB, GPU RTX A5000 24 GB. We examine four methods (TF.IDF, TextRank, mT5, mBART) and the proposed method (a news important region from TextRank is fed to mBART). Both mBART and mT5 models are fine-tuned to the ThaiSum training and validation sets to ensure a fair comparison.

Additionally, we perform hyperparameter tuning to identify the optimal setting before fine-tuning each Large Language

Model (LLM). **Table 1** enumerates all hyperparameters for both models. **Figures 3–4** depict the loss value during the fine-tuning process of mBART and mT5, respectively. To prevent overfitting, the fine-tuning process is halted at the epoch where the training loss diverges from the validation loss. Interestingly, both mBART and mT5 stop training at the 5th epoch coincidentally.

Table 1 Hyperparameters for finetuning LLMs

Hyperparameters	mBART	mT5
learning_rate	1.00E-05	1.00E-05
weight_decay	0.061	0.044
label_smooth_factor	0.051	0.053
optim	adamw	adamw
learning_rate	1.00E-05	1.00E-05

4.2 Experimental Results

To examine the number of common words between the actual summary and the news summary produced by each model, we employ the Newmm algorithm to tokenize words in the news summary and measure the ROUGE-1 metric as shown in **Figure 5**. Both mBART and the proposed model excel over other methods by competitively generating words common to the actual summary. For the sake of clarity, mBART is an off-the-shelf LLM text summarization that is finetuned to the ThaiSum dataset while in the proposed method a news important region is extracted by TextRank and then fed to the finetuned mBART.

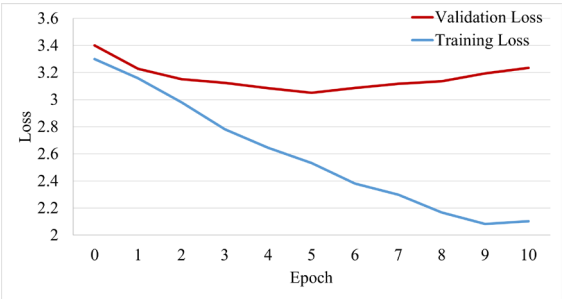

Figure 3 Loss during finetuning of mBART

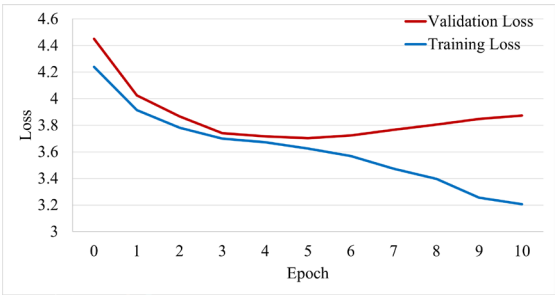

Figure 4 Loss during finetuning of mT5

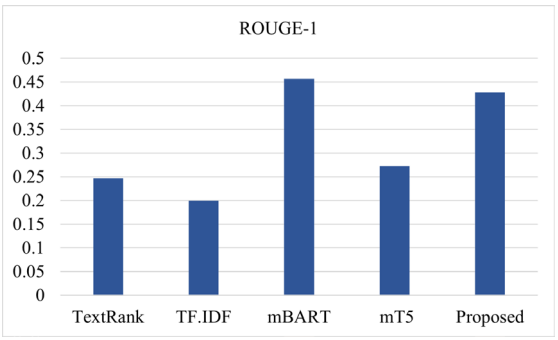

Figure 5 ROUGE-1 comparison

A high ROUGE-1 score alone is insufficient to indicate a good summary, as the common words may not be in the correct sequence. For instance, “I ate an apple” and “An apple ate I” have the same ROUGE-1 score, but they convey different meanings. Therefore, we require ROUGE-2 to measure the word order in the news summary.

As shown in **Figure 6**, the ROUGE-2 scores for all models are approximately one-third lower than the ROUGE-1 scores. We observed that many of the news summaries from TextRank and TF.IDF overlook the secondary theme sentences present in the ground truth, producing only one or two main sentences as a summary. As a result, the ROUGE-2 values for TextRank and TF.IDF are low and unacceptable. In contrast, the ROUGE-2 scores for LLM-based models are acceptable and significantly higher than those of both TextRank and TF.IDF methods.

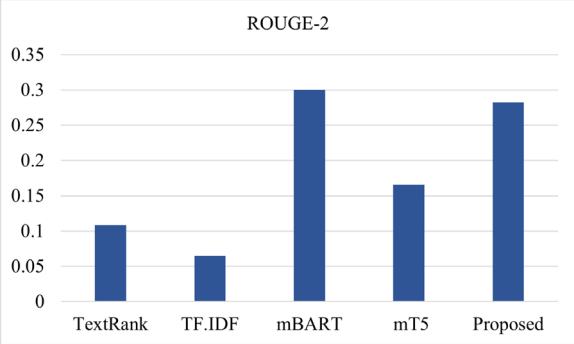

Figure 6 ROUGE-2 comparison

However, ROUGE-2 might not be sufficient to assess the quality of a news summary as it strictly evaluates the bigram order of the text. A sentence with a slightly different word order, but similar sentence structure, might convey the same meaning. For example, "You are beautiful" and "You are so beautiful" have different ROUGE-2 scores but convey a similar meaning.

To capture the similarity in sentence structure, we examine ROUGE-L, which measures the longest common subsequence between the ground truth and the news summary. **Figure 7** shows that both mBART and the proposed model continue to outperform other models, as their ROUGE-L scores are significantly high. Therefore, we will only consider these two models, i.e., mBART and the proposed model, for further experiments to identify the best method for Thai news summarization.

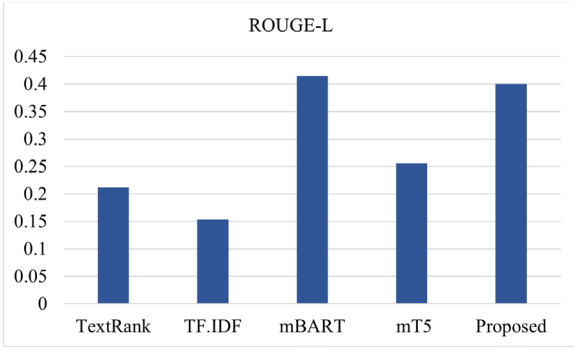

Figure 7 ROUGE-L comparison

It appears that mBART is on par with the proposed model in all ROUGE scores. The subsequent experiment aims to differentiate them. We collected 1000 longer external Thai news articles from the internet and used the cutting-edge ChatGPT to generate the reference summaries. The selection criterion for these news articles is that their length must be significantly greater than that of articles in ThaiSum. On average, the word count in an article is twice that of ThaiSum. News articles equally cover a wide range of categories, e.g., politics, economics, business, technology, health, sports, environment, education, science, law, and entertainment. The experimental outcomes for mBART and the proposed model are reported as follows.

Figure 8 presents a comparison of precision, recall, and F1 measures calculated by BERTScore, a metric that computes a similarity score for each word in the news summary against each word in the ground truth text. These three metrics are more delicate than ROUGE scores since they not only measure the exact match but also the similarity between words. Both the proposed model and mBART demonstrate competitive performance.

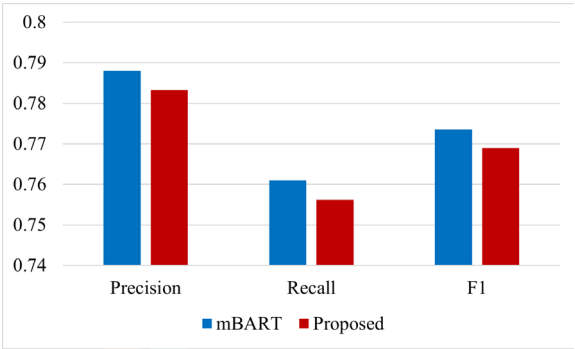

Figure 8 Precision, recall, and F1 comparison

The final comparison is particularly crucial. **Figure 9** presents a comparison of the ROUGE scores and average runtime between the proposed method and mBART. As shown in **Table 2**, while the proposed method's ROUGE scores are approximately 3.4% lower than mBART's, its average runtime is significantly 32% less. This tradeoff suggests that although the proposed method occasionally removes potential sentences from the original news (resulting

in slightly lower ROUGE scores than those of mBART), its substantial runtime improvement makes it a viable choice.

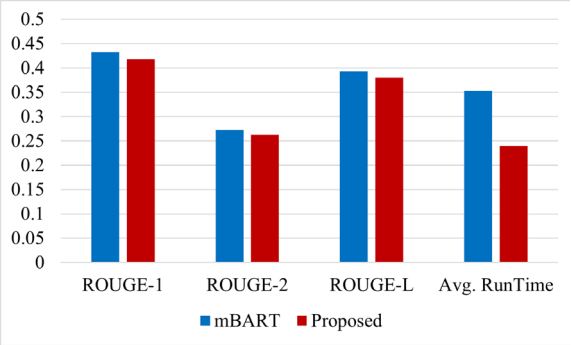

Figure 9 ROUGE scores and runtime comparison

Table 2 Percentage of ROUGE scores and runtime improvements

Metric	Percentage of improvement (%)
ROUGE-1	-3.35
ROUGE-2	-3.70
ROUGE-L	-3.25
Runtime	32.10

Consequently, the proposed model can compete with the mBART regarding the quality of news summary and surpasses mBART in terms of runtime. This makes the proposed model ideal for summarizing lengthy news at a promising speed. This improvement highlights the influence of the important region extracted by TextRank. If the region size is too small, the resulting summary might miss key points from the original news. Conversely, if the region's size is too large, it could include unnecessary information, leading to longer runtimes and an overly detailed summary.

To visualize the news summaries produced by all models, we display the original news in **Figure 10** and the corresponding summaries in **Figure 11.** The proposed model and mBART create identical summaries, encompassing approximately the first 3 sentences of the original news. This aligns with the ThaiSum report [24] stating that the main idea of articles is typically contained within the first three to four sentences. This is a common news presentation format that most writers follow, where the first paragraph serves as a summary of all the facts.

Upon a quick look, mT5 appears to generate a summary similar to the previous two models. However, it struggles with summarizing news that includes numbers, as indicated by the red underlined text. This misinterpretation of numerical data could be due to the original mT5 not receiving sufficient numerical data samples before finetuning with the ThaiSum dataset, leading to difficulties in comprehending the context and significance of numbers in a text. Consequently, the ROUGE scores of mT5 are significantly lower, about 30–40%, compared to the proposed model and mBART.

ปลัด มท. เผยยอดลงทะเบียนหนี้นอกระบบครบ 60 วัน มูลหนี้ 9,240 ล้านบาท ประชาชนลงทะเบียนแล้วกว่า 1.36 แสนราย ไกล่เกลี่ยสำเร็จแล้ว 9,066 ราย มูลหนี้ลดลงกว่า 606 ล้านบาท วันที่ 29 ม.ค. 2567 นายสุทธิพงษ์ จุลเจริญ ปลัดกระทรวงมหาดไทย เปิดเผยถึงการลงทะเบียนแก้ไขปัญหาหนี้นอกระบบ เป็นวันที่ 60 นับตั้งแต่เปิดลงทะเบียนเมื่อวันที่ 1 ธันวาคม 2566 เป็นต้นมา โดยจากข้อมูลของสำนักการสอบสวนและนิติการ กรมการปกครอง เมื่อเวลา 15.00 น. มีมูลหนี้รวม 9,240.091 ล้านบาท ประชาชนลงทะเบียนแล้ว 136,002 ราย เป็นการลงทะเบียนผ่านระบบออนไลน์ 115,169 ราย และการลงทะเบียน ณ ศูนย์อำนวยการแก้ไขหนี้นอกระบบ 20,833 ราย รวมจำนวนเจ้าหนี้ 103,441 ราย มีพื้นที่/จังหวัดที่มีผู้ลงทะเบียนมากที่สุด 5 ลำดับแรก 1. กรุงเทพมหานคร ยังคงมากที่สุด มีผู้ลงทะเบียน 11,074 ราย เจ้าหนี้ 7,523 ราย มูลหนี้ 822.216 ล้านบาท 2. จังหวัดนครศรีธรรมราช มีผู้ลงทะเบียน 5,557 ราย เจ้าหนี้ 5,048 ราย มูลหนี้ 375.714 ล้านบาท 3. จังหวัดสงขลา มีผู้ลงทะเบียน 5,049 ราย เจ้าหนี้ 3,930 ราย มูลหนี้ 330.967 ล้านบาท 4. จังหวัดนครราชสีมา มีผู้ลงทะเบียน 4,826 ราย เจ้าหนี้ 3,523 ราย มูลหนี้ 385.373 ล้านบาท 5. จังหวัดขอนแก่น มีผู้ลงทะเบียน 3,579 ราย เจ้าหนี้ 2,951 ราย มูลหนี้ 304.731 ล้านบาท ขณะที่จังหวัดที่มีผู้ลงทะเบียนน้อยที่สุด 5 ลำดับแรก ได้แก่ 1. จังหวัดแม่ฮ่องสอน มีผู้ลงทะเบียน 219 ราย เจ้าหนี้ 229 ราย มูลหนี้ 12.712 ล้านบาท 2. จังหวัดระนอง มีผู้ลงทะเบียน 304 ราย เจ้าหนี้ 218 ราย มูลหนี้ 20.785 ล้านบาท 3. จังหวัดสมุทรสงคราม มีผู้ลงทะเบียน 358 ราย เจ้าหนี้ 278 ราย มูลหนี้ 13.077 ล้านบาท 4. จังหวัดตราด มีผู้ลงทะเบียน 435 ราย เจ้าหนี้ 326 ราย มูลหนี้ 17.974 ล้านบาท และ 5. จังหวัดสิงห์บุรี มีผู้ลงทะเบียน 467 ราย เจ้าหนี้ 343 ราย มูลหนี้ 23.594 ล้านบาท

Figure 10 Original News#1

The other two models, TextRank and TF.IDF, are based on the extractive approach. TextRank provides the first 9 sentences of the articles as a summary. However, some of these sentences delve into the finer details of the news and should not be considered as the main points. TF.IDF, on the other hand, generates a poor summary by focusing on the minor elements of the news. Furthermore, it includes fragments that do not form a complete sentence

Proposed model / mBART
ปลัด มท. เผยยอดลงทะเบียนหนี้นอกระบบครบ 60 วัน มูลหนี้ 9,240 ล้านบาท ประชาชนลงทะเบียนแล้วกว่า 1.36 แสนราย ไกล่เกลี่ยสำเร็จแล้ว 9,066 ราย มูลหนี้ลดลงกว่า 606 ล้านบาท

mT5
ปลัด มท. เผยยอดลงทะเบียนหนี้นอกระบบครบ 60 วัน มูลหนี้ 9,400 ล้านบาท ประชาชนลงทะเบียนแล้วกว่า 100,000 ราย ไกล่เกลี่ยสำเร็จแล้ว 9,.75 ราย มูลหนี้ลดลงกว่า 234 ล้านบาท

TextRank
ปลัด มท. เผยยอดลงทะเบียนหนี้นอกระบบครบ 60 วัน มูลหนี้ 9,240ล้านบาท ประชาชนลงทะเบียนแล้วกว่า 1.36 แสนราย ไกล่เกลี่ยสำเร็จแล้ว 9,066 ราย มูลหนี้ลดลงกว่า 606ล้านบาท วันที่ 29 ม.ค. 2567 นายสุทธิพงษ์ จุลเจริญ ปลัดกระทรวงมหาดไทย เปิดเผยถึงการลงทะเบียนแก้ไขปัญหาหนี้นอกระบบ เป็นวันที่ 60 นับตั้งแต่เปิดลงทะเบียนเมื่อวันที่ 1 ธันวาคม 2566 เป็นต้นมา โดยจากข้อมูลของสำนักการสอบสวนและนิติการ กรมการปกครอง เมื่อเวลา 15.00 น. มีมูลหนี้รวม 9,240.091ล้านบาท ประชาชนลงทะเบียนแล้ว 136,002 ราย เป็นการลงทะเบียนผ่านระบบออนไลน์115,169 ราย และการลงทะเบียน ณ ศูนย์อำนวยการแก้ไขหนี้นอกระบบ 20,833

TF.IDF
วัน มูลหนี้ 9,240 ล้านบาท ประชาชนลงทะเบียนแล้วกว่า 1.36 แสนราย ไกล่เกลี่ยสำเร็จแล้ว 9,066 วันที่ 29 ม.ค. 2567 นายสุทธิพงษ์ จุลเจริญ ปลัดกระทรวงมหาดไทย เปิดเผยถึงการลงทะเบียนแก้ไขปัญหาหนี้นอกระบบ เป็นวันที่ 60 นับตั้งแต่เปิดลงทะเบียนเมื่อวันที่ 1 ประชาชนลงทะเบียนแล้ว 136,002 ราย เป็นการลงทะเบียนผ่านระบบออนไลน์ 115,169 ราย รวมจำนวนเจ้าหนี้ 103,441 2. จังหวัดนครศรีธรรมราช 4. จังหวัดนครราชสีมา ลำดับแรก ได้แก่ 1. จังหวัดแม่ฮ่องสอน 3. จังหวัดสมุทรสงคราม และ 5. จังหวัดสิงห์บุรี

Figure 11 News#1 Summaries

To further visualize the performance of each summarization model, we provide an additional news example in **Figure 12**, along with its summaries from various models in **Figure 13**. Both the proposed method and mBART yield the same summary, which is made up of several sentences from the original news. These sentences encapsulate the details in the subsequent sentences.

ชาวสวนยิ้ม ส่งออกทุเรียนสดผ่านแดน ปี66 พุ่ง 44.1% จีนครองแชมป์ยอดสั่งสูงสุดกว่า 9 หมื่นล้านบาท คต.เร่งกระตุ้นค้าชายแดน-ผ่านแดนผ่านยุทธศาสตร์การค้า วันนี้ (29 ม.ค.2567) นายรณรงค์ พูลพิพัฒน์ อธิบดีกรมการค้าต่างประเทศ (คต.) กล่าวว่า แม้ว่าปี 2566 การค้าชายแดนกับประเทศเพื่อนบ้านของไทยชะลอตัวจากสถานการณ์ภายในประเทศที่เปราะบาง เช่น ลาว ที่อัตราเงินเฟ้อสูงและค่าเงินกีบยังคงอ่อนค่าอย่างต่อเนื่อง สถานการณ์การสู้รบในเมียนมา ความต้องการบริโภคในประเทศของกัมพูชาที่ยังไม่ฟื้นตัวอย่างเต็มที่ แต่การค้าผ่านแดนของไทยกลับมาขยายตัวอีกครั้ง ทั้งมูลค่าการค้ารวมและการส่งออก โดยเฉพาะการค้าผ่านแดนไปจีน คิดเป็นสัดส่วน 52% ของการค้าผ่านแดนไทย มีมูลค่าการค้ารวม 423,116 ล้านบาท ในขณะที่การค้าชายแดนกับประเทศเพื่อนบ้าน 4 ประเทศ มีมูลค่าการค้ารวม 929,730 ล้านบาท แบ่งเป็นการส่งออก มูลค่า 580,100 ล้านบาท การนำเข้า มูลค่า 349,630 ล้านบาท และไทยได้ดุลการค้า รวม 230,471 ล้านบาท โดยมาเลเซียเป็นประเทศคู่ค้าชายแดนอันดับ 1 ของไทย มีมูลค่าการค้ารวม 287,155 ล้านบาท รองลงมา ลาว 260,512 ล้านบาท เมียนมา 220,327 ล้านบาท และกัมพูชา 161,736 ล้านบาท สินค้าส่งออกชายแดนสำคัญในปี 2566 ได้แก่ น้ำมันดีเซล 40,143 ล้านบาท น้ำมันสำเร็จรูปอื่น ๆ 19,672 ล้านบาท และเครื่องคอมพิวเตอร์และอุปกรณ์ อื่น ๆ 15,120 ล้านบาท ด่านศุลกากรที่มีมูลค่าการค้าชายแดนสูงสุด ได้แก่ ด่านศุลกากรสะเดา มูลค่า 226,254 ล้านบาท รองลงมาได้แก่ ด่านศุลกากรแม่สอด และด่านศุลกากรอรัญประเทศ มีมูลค่าการค้า 106,835 ล้านบาท และ 97,185 ล้านบาท ตามลำดับ

Figure 12 Original News#2

Proposed model / mBART
ชาวสวนยิ้ม ส่งออกทุเรียนสดผ่านแดน ปี66 พุ่ง 44.1% จีนครองแชมป์ยอดสั่งสูงสุด กว่า 9 หมื่นล้านบาท คต.เร่งกระตุ้นค้าชายแดน-ผ่านแดนผ่านยุทธศาสตร์การค้า

mT5
ชาวสวนยิ้ม ส่งออกทุเรียนสดผ่านแดน ปี66 พุ่ง 44.1% จีนครองแชมป์ยอดสั่งสูงสุด กว่า 9 หมื่นล้านบาท คต.เร่งกระตุ้นค้าชายแดน-ผ่านแดนผ่านยุทธศาสตร์การค้า วันนี้ (29 ม.ค.63)

TextRank
ชาวสวนยิ้ม ส่งออกทุเรียนสดผ่านแดน ปี66 พุ่ง 44.1% จีนครองแชมป์ยอดสั่งสูงสุดกว่า 9 หมื่นล้านบาท คต.เร่งกระตุ้นค้าชายแดน-ผ่านแดนผ่านยุทธศาสตร์การค้า วันนี้ (29 ม.ค.2567) นายรณรงค์ พูลพิพัฒน์ อธิบดีกรมการค้าต่างประเทศ (คต.) กล่าวว่า แม้ว่าปี 2566 การค้าชายแดนกับประเทศเพื่อนบ้านของไทยชะลอตัวจากสถานการณ์ภายในประเทศที่เปราะบาง เช่น ลาว ที่อัตราเงินเฟ้อสูงและค่าเงินกีบยังคงอ่อนค่าอย่างต่อเนื่อง สถานการณ์การสู้รบในเมียนมา ความต้องการบริโภคในประเทศ ของกัมพูชาที่ยังไม่ฟื้นตัวอย่างเต็มที่ แต่การค้าผ่านแดนของไทยกลับมาขยายตัวอีกครั้งทั้งมูลค่าการค้ารวมและการส่งออก โดยเฉพาะการค้าผ่านแดนไปจีน คิดเป็นสัดส่วน 52% ของการค้าผ่านแดนไทย มีมูลค่าการค้ารวม 423,116ล้านบาท

TF.IDF
สถานการณ์การสู้รบในเมียนมา โดยเฉพาะการค้าผ่านแดนไปจีนมีมูลค่าการค้ารวม 423,116 ล้านบาท ล้านบาท ด่านศุลกากรมุกดาหาร นครพนม 142,567ล้านบาท ในขณะที่การค้าชายแดนกับประเทศเพื่อนบ้าน 4 ล้านบาท แบ่งเป็นการส่งออก มูลค่า 580,100 ล้านบาท และไทยได้ดุลการค้า รวม 230,471ล้านบาท ล้านบาท เมียนมา 220,327ล้านบาท และกัมพูชา 161,736ล้านบาท ได้แก่ น้ำมันดีเซล 40,143ล้านบาท น้ำมันสำเร็จรูปอื่น ๆ 19,672ล้านบาท และเครื่องคอมพิวเตอร์และอุปกรณ์ อื่น ๆ 15,120ล้านบาท ด่านศุลกากรสะเดา มูลค่า 226,254ล้านบาท และด่านศุลกากรอรัญประเทศ และ 97,185ล้านบาท มีมูลค่าการค้า

Figure 13 News#2 Summaries

The mT5 summary is akin to those of the proposed model and mBART, except it includes a date fragment from the next sentence. In contrast, TextRank and TF.IDF produce extensive texts, many of which do not convey the main concept of the original news. TextRank tends to select sentences, some overly detailed, from the initial part of the article, while TF.IDF extracts fragment sentences from the entire document.

The final evaluation assesses human judgment on the satisfaction of news summaries generated by various models. **Table 3** presents the average satisfaction scores, ranging from 1 (lowest) to 5 (highest), based on evaluations from 10 volunteers who each reviewed the same set of 100 news summaries. These 100 samples are randomly picked from 1000 external news articles in the previous experiment. The results in **Table 3** align with the ROUGE scores, precision, recall, and F1-score comparisons from earlier experiments. Both the proposed method and mBART achieve similarly high scores, while TextRank and TF.IDF receive lower scores due to their lengthy summaries containing unnecessary information.

Table 3 Average satisfaction scores from volunteers

Method	Average satisfaction score
Proposed	4.71
mBART	4.83
mT5	3.15
TextRank	2.25
TF.IDF	1.87

5. Conclusion

In this paper, we introduce a fast hybrid Thai news summarization model capable of extracting significant sentences from the original article and abstractly summarizing these sentences into a coherent Thai paragraph, as suggested by the fine-tuned Large Language Model (LLM). Experimental results on the ThaiSum dataset indicate that our approach outperforms existing models in terms of Rouge-1, Rouge-2, and Rouge-L scores. Our research contributes in three ways. First, our proposed model consumes significantly less runtime compared to the state-of-the-art abstractive mBART model. Second, the news summaries it generates are of high quality and can compete with those produced by state-of-the-art LLMs. Lastly, our model is particularly effective for summarizing Thai news, as it has been fine-tuned on the ThaiSum dataset. Furthermore, our proposed model can handle lengthy Thai news by eliminating unnecessary sentences before using the LLM to summarize it abstractly. Our technique can be applied to other LLMs with text summarization services to reduce the runtime. It is worth noting that the training data's bias may cause ethical implications. When the data is biased towards specific perspectives, the resulting summaries might mirror these biases, potentially causing a skewed portrayal of the news. One limitation of our approach is that it may not reduce the LLM workload when important sentences are uniformly distributed throughout an article, as the important region will encompass all sentences in the original news. A potential solution could be to summarize all key points within the article by selecting multiple important regions, each representing one key point. A possible future work is an adaptation of the proposed method to other languages by replacing the Thai sentence segmenter with an appropriate algorithm and finetuning the mBART model using a news summarization dataset in the target language.

References

[1] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut, "Text Summarization Techniques: A Brief Survey," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 397–405, 2017, doi: 10.14569/IJACSA.2017.081052.

[2] K. F. Wong, M. Wu and W. Li, "Extractive summarization using supervised and semi-supervised learning," in *Proc.*

22nd International Conference on Computational Linguistics, Manchester, UK, Aug. 18–22, 2008, pp. 985–992, doi: 10.3115/1599081.1599205.

[3] V. Qazvinian, D. R. Radev, S. M. Mohammad, B. Dorr, D. M. Zajic, M. Whidby and T. Moon, "Generating Extractive Summaries of Scientific Paradigms," *Journal of Artificial Intelligence Research*, vol. 46, pp. 165–201, 2013, doi: 10.1613/jair.3732.

[4] A. Jain, D. Bhatia and M. K. Thakur, "Extractive Text Summarization Using Word Vector Embedding," in *2017 International Conference on Machine Learning and Data Science (MLDS)*, Noida, India, Dec. 14–15, 2017, pp. 51–55, doi: 10.1109/MLDS.2017.12.

[5] R. Nallapati, B. Zhou, C. Santos, Ç. Gulçehre and B. Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond," in *Proc. 20th SIGNLL Conference on Computational Natural Language Learning*, Berlin, Germany, Aug. 11–12, 2016, pp. 280–290, doi: 10.18653/V1/K16-1028.

[6] J. Tan, X. Wan and J. Xiao, "Abstractive Document Summarization with a Graph-Based Attentional Neural Model," in *Proc. the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, Jul. 30–4, 2017, pp. 1171–1181, doi: 10.18653/v1/P17-1108.

[7] S. Gehrmann, Y. Deng and A. Rush, "Bottom-Up Abstractive Summarization," in *Proc. 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Oct. 31–4, 2018, pp. 4098–4109, doi: 10.18653/v1/D18-1443.

[8] Y. Zhang, D. Li, Y. Wang, Y. Fang and W. Xiao, "Abstract Text Summarization with a Convolutional Seq2seq Model," *applied sciences*, vol. 9, no. 8, 2019, Art. no. 1665, doi: 10.3390/app9081665.

[9] Z. Hao, J. Ji, T. Xie and B. Xue, "Abstractive Summarization Model with a Feature-Enhanced Seq2Seq Structure," in *2020 5th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, Singapore, Jul. 17–19, 2020, pp. 163–167, doi: 10.1109/ACIRS49895.2020.9162627.

[10] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020, doi: 10.1162/tacl_a_00343.

[11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703.pdf

[12] M. Yang, X. Wang, Y. Lu, J. Lv., Y. Shen and C. Li, "Plausibility promoting generative adversarial network for abstractive text summarization with multi-task constraint," *Information Sciences*, vol. 521, pp. 46–61, 2020, doi: 10.1016/j.ins.2020.02.040.

[13] M. Yang, C. Li, Y. Shen, Q. Wu, Z. Zhao and X. Chen, "Hierarchical Human-Like Deep Neural Networks for Abstractive Text Summarization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2744–2757, 2021, doi: 10.1109/TNNLS.2020.3008037.

[14] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua and C. Raffel, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Jun. 6–11, 2021, pp. 483–498.

[Online]. Available: https://aclanthology.org/2021.naacl-main.41.pdf.

[15] 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.

[16] W. Phatthiyaphaibun, K. Chaovavanich, C. Polpanumas, A. Suriyawongkul, L. Lowphansirikul and P. Chormai, "PyThaiNLP: Thai Natural Language Processing in Python," in *2023 Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, Singapore, Dec. 6, 2023, pp. 25–36, doi: 10.18653/v1/2023.nlposs-1.4.

[17] P. Ngamcharoen, N. Sanglerdsinlapachai and P. Vejjanugraha, "Automatic Thai Text Summarization Using Keyword-Based Abstractive Method," in *2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Chiang Mai, Thailand, Nov. 5–7, 2022, pp. 1–5, doi: 10.1109/iSAI-NLP56921.2022.9960265.

[18] P. Chormai, P. Prasertsom and A. Rutherford, "AttaCut: a fast and accurate neural Thai word segmenter," *arXiv*, vol. abs/1911.07056, pp. 1–13, 2019, doi: 10.48550/arXiv.1911.07056

[19] *DeepCut: A Thai word tokenization library using Deep Neural Network*, Zenodo, Sep. 23, 2019, doi: 10.5281/zenodo.3457707.

[20] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," in *Proc. 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, Jul. 25–26, 2004, pp. 404–411.

[21] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998, doi: 10.1016/S0169-7552(98)00110-X.

[22] J. Leskovec, A. Rajaraman and J. D. Ullman, "Data mining," in *Mining of Massive Datasets*, 2nd ed. Cambridge, UK: Cambridge University Press, 2014, ch. 1, sec. 1.3.1, pp. 8–9.

[23] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proc. Workshop on Text Summarization Branches Out*, Barcelona, Spain, Jul. 25–26, 2004, pp. 74–81.

[24] N. Chumpolsathien, "Using Knowledge Distillation from Keyword Extraction to Improve the Informativeness of Neural Cross-lingual Summarization," M.S. thesis, Beijing Institute of Technology, Beijing, China, 2020.