LatVis: Large-scale Task-specific Language Model for Low-resource Vietnamese Multi-document Summarization

LE THE ANH[*], FPT University, Viet Nam
LE HAI SON[*], Ha Noi University of Science and Technology, Viet Nam

Vietnamese Multi-document Summarization task faces three key challenges including the long input sequence problem, human-like summary generation, and the scarcity of labeled data. Thanks to Transformer-based models empowered by parallel computation architecture and attention mechanisms, the long input sequence problem is partially solved. In addition, transformer-based models trained on a massive amount of text achieve impressive results on text generation tasks, close to human-level performance. Furthermore, using the pre-training strategy using self-supervised learning objectives is a good approach to deal with the scarcity of labeled data. Based on these considerations, this paper aims to leverage a large amount of unlabeled Vietnamese text to create a large-scale pseudo-labeled multi-document dataset, and then use it to pre-train a Vietnamese task-specific language model, namely LatVis. Conducted experiments show that without any fine-tuning, our pre-trained model achieves a good performance compared with some previous models. After being fine-tuned on approximately 300 samples, our model obtained impressive Rouge Scores, 76.7%, 78.9%, and 73.9% of Rouge-1 F1, and 50.2%, 55.0%, and 46.7% of Rouge-2 F1 on VMDS, ViMS, VLSP datasets, respectively. To the best of our knowledge, this is the first public large-scale task-specific language model, specifically pre-trained for the Vietnamese multi-document summarization task, that proves to be a potential approach for natural language processing tasks in less-resourced languages.

CCS Concepts: • **Computing methodologies → Natural language generation**.

Additional Key Words and Phrases: Transformer-based Models, Task-Specific Language Model, Vietnamese Abstractive Multi-document Summarization

ACM Reference Format:
Le The Anh and Le Hai Son. 2018. LatVis: Large-scale Task-specific Language Model for Low-resource Vietnamese Multi-document Summarization. *J. ACM* 37, 4, Article 111 (August 2018), 18 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Text Summarization is defined as the task of converting a long text into a concise, coherent, and fluent summary that conveys the main meaning and salient information. This is a truly hard task, even for humans. Multi-document Text Summarization is even harder as the input is a cluster of related documents, containing up to thousands of words and overlapped contents. When working on building an supervised deep neural network for Vietnamese Multi-document Summarization, one more challenge we have to face is the lack of labeled data. Until recently, there have been only three small publicly available datasets with about 300 samples each. To deal with this challenge, some recent studies ([13], [18], [22]) leverage a large amount of unlabeled data to pre-train language

[*]Both authors contributed equally to this research. The author order is alphabetical.

Authors' addresses: Le The Anh, anhlt161@fe.edu.vn, FPT University, Can Tho, Viet Nam; Le Hai Son, haison.le001@gmail.com, Ha Noi University of Science and Technology, Ha Noi, Viet Nam.

0004-5411/2018/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

models then fine-tune them on small amount of multi-document summarization labelled data. However, these pre-trained language models have two main limitations when adapting to the Multi-document Summarization task with long and structured inputs. The first one is that they are general-purpose language models, meaning that they use general pre-training objectives to model the probability distribution over the sequence of words. For instance, BARTpho [22] is based on the BART model [6] that corrupts input text with a noising function and then learns to reconstruct the original text. The second limitation lies in their model architectures that are not suitable for working on long input sequences. We, instead, focus on building a task-specific language model using an objective function specifically designed for Multi-document Summarization. In addition, we build a large-scale database with pseudo summaries for pre-training our model.

In recent years, many pre-trained models have obtained impressive results when fine-tuning on downstream tasks. Among them, we use the PRIMERA [27] as our backbone model since it was specially designed for Multi-document Summarization. Unlike T5 [19], BART [6] which are general-purpose pre-trained language models, PRIMERA is a task-specific language one. PRIMERA leverages (1) the self-supervised objective Gap Sentence Generation proposed in [30] with Pyramid Entity strategy to select and aggregate salient information from documents and (2) Longformer architecture [1] which employs sparse local and global attentions to achieve a linear complexity with respect to the input length. The input length of Longformer can be up to 16000 tokens. These make it especially suitable for Multi-document Summarization. We build our model, namely LatVis, based on PRIMERA with three main modifications related to the word segmentation, tokenization, and named entity recognition, to make it work in Vietnamese. Following the Pyramid Entity strategy, we generate a large multi-document summarization dataset with pseudo summaries created by extracting the sentences with high Rouge Scores based on the frequency of occurrence of named entities across the documents. This dataset is then used to pre-train the LatVis model. The pre-trained LatVis is the first large-scale task-specific language model that is specially pre-trained for Vietnamese Multi-document Summarization.

2 RELATED WORK

Thanks to the advances in deep learning techniques, text summarization models have achieved impressive performance in the last decades. State-of-the-art approaches are either extractive (MatchSum [31], DiscoBERT [29], BertSumExt [8], PNBert [32]), abstractive (BRIO [10], SimCLS [9]), or hybrid (EASE [7]). However, there have been few studies on the task of Vietnamese Text Summarization. Very few of them were dedicated to Multi-document Summarization. Early studies mostly focused on Single-document Summarization using the extractive approach and mainly relied on traditional statistical methods. Thanh et al. [21] combined Word co-occurrences, TF-IDF, Position-base, Title-based, and Proper Noun-based methods to select the salient sentences, based on which the summary was generated. Phuc et al. [16] presented a graph-based system using a self-organizing map to cluster documents and extract the main idea. One of the earliest researches on Vietnamese Multi-document summarization, Ung et al. [24], introduced an extractive system consisting of three phases: pre-processing, score computing, and summarization generation. The system uses a set of features at both word and sentence levels, manually chosen to be suitable for Vietnamese news text, to compute the sentence score. These features include (1) word features (word frequency, word location), (2) sentence position, (3) time, and (4) PageRank-based sentence score.

Recent work has mainly relied on leveraging the power of transformer-based models. Quoc et al. [18] introduces extractive models that utilize variants of BERT to generate the sentence embedding. Specifically, these models concatenate all documents into one paragraph, leverage BERT to encode sentences, then use K-means clustering for ranking and selecting salient sentences to create the summary. More recently, [22] introduces BARTpho, a Vietnamese large-scale sequence-to-sequence model which is based on the BART model [6]. Two versions of BARTpho, BARTpho$_{\text{word}}$ and BARTpho$_{\text{syllable}}$, were pre-trained on huge corpora (145M automatically word-segmented sentences, 4B syllable tokens). These pre-trained models were then fine-tuned on Vietnews dataset, a Vietnamese single-document summarization, and achieved a high performance compared with some previous models. Additionally, ViT5 [15], featured a pre-trained Transformer-based encoder-decoder model specifically tailored for the Vietnamese language using T5-style self-supervised pre-training. ViT5$_{\text{large}}$ underwent evaluation on two downstream tasks, Abstractive Text Summarization, and Named Entity Recognition, achieving state-of-the-art results on Vietnamese datasets. However, our experiments with BARTPho and ViT5 on multi-document summarization datasets obtained low results. One of the main reasons for this is that BARTpho and ViT5 were not specially designed for the Multi-document Summarization task which has a long input, a cluster of related documents with a length of up to several thousands of words. Figure 1 gives an overview of some outstanding Vietnamese summarization language models including our proposed model.

3 LATVIS MODEL

In this section, we present the LatVis model, detail its architecture and components tailored to the Vietnamese language, specifically designed for low-resource scenarios.

3.1 Model architecture

LatVis is built based on the PRIMERA model [27] which is specifically designed for the task of multi-document summarization. To tailor it for the Vietnamese language, we implement three key modifications including (1) incorporating a word segmentation model trained on Vietnamese datasets to pre-process raw text and prepare it for later processing steps before feeding it into the model, (2) training tokenizers utilizing Byte Pair Encoding (BPE) with and without segmentation for handling Vietnamese text, and (3) employing a Named Entity Recognition model trained on Vietnamese datasets for extracting named entities and selecting sentences based on their significance scores.

So far, there is no publicly available large-scale dataset for Vietnamese multi-document summarization task. We generate a large-scale labeled multi-document summarization dataset from an unlabeled dataset containing a large amount of long documents:

$$R = \{r_1, r_2, \ldots, r_n\}, \tag{1}$$

where n is the number of raw documents. Each document is transformed into a pseudo cluster, a list of related documents, by chunking each document into several parts of almost the same length:

$$C = \{c_1, c_2, \ldots, c_n\}, \tag{2}$$

here $c_i = \{d_{i_1}, d_{i_2}, \ldots, d_{i_m}\}$ is the i^{th} pseudo cluster of documents generated by chunking r_i. d_{i_j} denotes the j^{th} document in the cluster c_i. $\texttt{WordSeg}_{\texttt{VnCoreNLP}}$ [26], a Vietnamese word segmentation model, is then applied to convert each raw document d_{i_j} to the word-segemented form:

$$w_{i_j} = \texttt{WordSeg}_{\texttt{VnCoreNLP}}(d_{i_j}) \tag{3}$$

After that, the word-segmented text is fed into a Vietnamese Named Entity Recognition model to extract named entities. After considering several Vietnamese NER models such as VnCoreNLP [26], Underthesea[1], Pyvi[2], PhoNLP [14], and re-evaluating them on Vietnamese NER datasets, we choose

[1] https://github.com/undertheseanlp/underthesea

[2] https://github.com/trungtv/pyvi

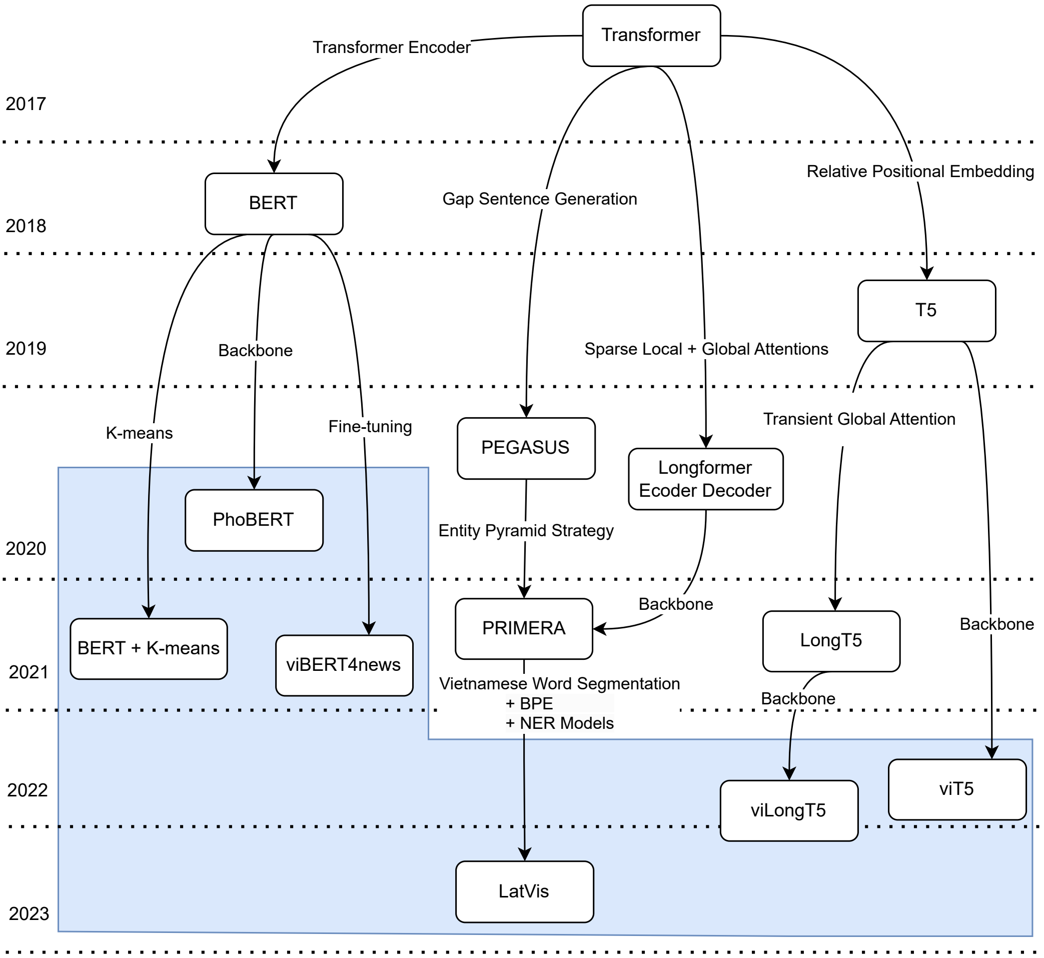

Fig. 1. Vietnamese summarization models. The models with light-blue background are the ones trained or evaluated on Vietnamese summarization tasks.

PhoNLP to extract and classify named entities into four categories including Location, Person, Organization, and Miscellaneous. Let e_{i_j} be the list of entities in the j^{th} document in the i^{th} cluster.

$$e_{i_j} = \mathsf{NER}_{\mathsf{PhoNLP}}(w_{i_j}). \tag{4}$$

The list $\{e_{i_1}, e_{i_2}, \dots, e_{i_m}\}$ is then used to compute the document frequency of each entity in the cluster c_i. After that, the document frequency of each entity is used to estimate the entity importance. The higher document frequency, the more important entity is. The list of entity with high frequencies is then used to choose the candidate sentences. Among these candidate sentences, the most representative ones are selected based on the Rouge score measuring the overlap of the sentence and the documents which the sentence doesn't appear in. The score of the k^{th} sentence in the document d_{i_j} is computed as follow.

$$Score(s_k) = \sum_{d_{i_j} \in c_i, s_k \notin di_j} Rouge(s_k, d_{i_j}). \tag{5}$$

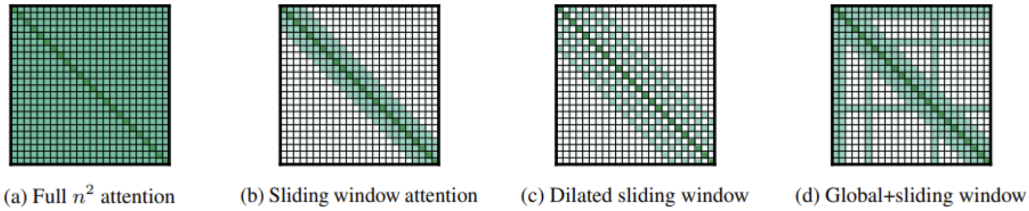
(a) Full n^2 attention
(b) Sliding window attention
(c) Dilated sliding window
(d) Global+sliding window

Fig. 2. Full self-attention pattern and the configuration of attention patterns in Longformer [1]

The top sentences with the highest scores in the i^{th} cluster are then concatenated to create a pseudo summary, ps_i (refer to the Entity Pyramid Sentence Selection algorithm in [28] for more details). So far, we generated the labeled dataset, C_l.

$$C_l = \{(c_1, ps_1), (c_2, ps_2), \ldots, (c_n, ps_n)\}. \tag{6}$$

Finally, the labelled dataset, C_l, is used to pre-train LatVis model using Vietnamese tokenizer model.

Instead of adopting the conventional Transformer architecture [25], the backbone of the LatVis model is based on PRIMERA [27] which leverages the Longformer encoder-decoder architecture proposed by Beltagy et al. [1]. This architecture represents a modification of the original Transformer architecture, incorporating a windowed attention mechanism (see Figure 2 for more details). One of the significant advantages of this architecture lies in its linear computational complexity relative to the input length. This efficiency enables it to process sequences of up to 16000 tokens with linear complexity and reasonable resource, such as when concatenating multiple documents. Consequently, Longformer proves to be an exceptionally suitable choice for the tasks involving lengthy documents such as multi-document summarization. The QKV (Query, Key, Value) formula is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{7}$$

here two sets of projections are used: $\mathrm{Q_s}$, $\mathrm{K_s}$, $\mathrm{V_s}$ for attention scores of sliding window attention, and $\mathrm{Q_g}$, $\mathrm{K_g}$, $\mathrm{V_g}$ for attention scores of the global attention.

In the field of summarization, task-specific pre-training objectives have shown remarkable improvements compared to using general-purpose ones. Specifically, PEGASUS [30] introduces the self-supervised Gap Sentence Generation objective function, where certain salient sentences in the input are masked, and the model is trained to generate them. PRIMERA adopts this objective function in combination with Entity Pyramid strategy [27] to identify and mask a subset of summary-like sentences. These selected sentences are replaced with special tokens, <sent mask>, in the input, and the model is then trained to generate the combined form of these sentences, creating a "pseudo-summary" as shown in figure 3. This process resembles abstractive summarization, as the model reconstructs the masked sentences using the available information from the remaining parts of the documents. Subsequently, the dataset with pseudo summaries are used to train an autoregressive language model.

3.2 Entity Pyramid Masking

To select the crucial information across documents, a novel masking strategy called the Entity Pyramid strategy [27] is introduced. The underlying idea behind this innovative masking strategy is originated from the recognition that essential information is frequently encapsulated within

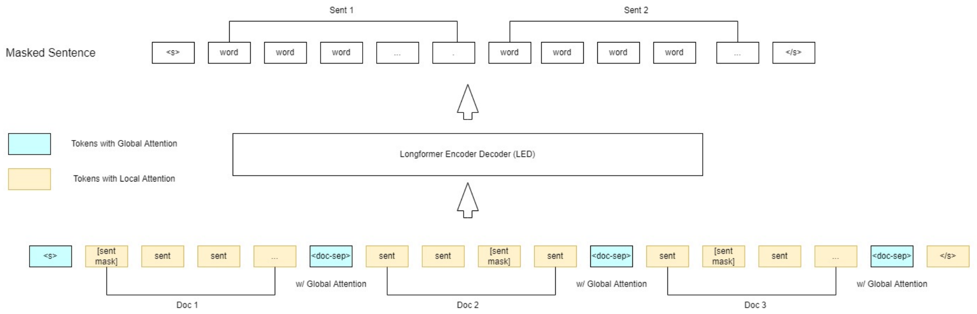

Fig. 3. Model architecture of PRIMERA. Documents are separated with <doc-sep> tokens. Selected sentences are replaced with <sent mask> tokens.

entities, and its significance increases with its appearance frequency across documents. Entity Pyramid strategy aims to select entities that better represent the entire cluster.

This approach prioritizes content overlap across multiple documents rather than exact matching between a few documents.

3.3 Vocabulary

When dealing with Vietnamese text, several tokenization techniques come into play, such as word-level, syllable-level, character-level, or sub-word level tokenization [20],[4]. However, character-level and word-level tokenization methods suffer from a limitation where tokens may lack individual meaning or struggle to handle out-of-vocabulary instances effectively. To address these shortcomings, a groundbreaking approach known as Byte Pair Encoding (BPE) was introduced by Sennrich et al. [20]. By applying BPE, most words can be represented by subwords, reducing the occurrence of the <unk> token that represents unseen words. This method has been rapidly adopted in modern NLP approaches and improved the accuracy significantly.

According to Tran et al. [22], pre-trained language models trained on word-level data can perform better than those trained on syllable-level data, especially for Vietnamese NLP tasks. Nonetheless, Long et al. [15] found that a syllable-level vocabulary can contribute a significant improvement to the model performance. As a result, our study encompasses an exploration of both word-level and syllable-level tokenization techniques, facilitating a comprehensive analysis. For the word-level vocabulary approach, we perform word segmentation using a Vietnamese word segmentation model called VnCoreNLP [26] on the pre-training dataset. We preprocessed our pre-training dataset and trained the vocabulary with a fixed size of 36K with SentencePiece [5].

3.4 Named Entity Recognition model

A Vietnamese Named Entity Recognition (NER) model is needed to effectively extract and classify general named entities from Vietnamese text. Several existing models, such as VnCoreNLP [26], Underthesea[3], Pyvi[4], and PhoNLP [14], are available. However, the absence of specific benchmark datasets for comparative evaluation presents a challenge. Consequently, our study selects the $\text{NER}_{\text{PhoNLP}}$ model, which stands as the most recent and current Vietnamese model. Leveraging the $\text{NER}_{\text{PhoNLP}}$ model, we aim to proficiently extract and classify named entities into three key categories: Location, Person, Organization. Furthermore, recognizing the significance of entities

[3] https://github.com/undertheseanlp/underthesea
[4] https://github.com/trungtv/pyvi

Dataset	Newscorpus	Newshead
Total clusters	5778893	314033
Total articles	17847516	1065571
Average length per article	154.03	2925.55
Average number of entities per article	4.87	23.67

Table 1. The statistics of Vietnamese unlabeled dataset

related to Date, Time, and Number within sentences, we also tried extracting these entities in our datasets.

4 LARGE-SCALE MULTI-DOCUMENT SUMMARIZATION DATASET GENERATION

Until recently, there was a notable absence of publicly available Vietnamese multi-document summarization datasets. As a response to this gap, our primary objective is to construct large-scale datasets specifically tailored for Vietnamese. These datasets will serve the dual purpose of facilitating pre-training of models and enabling model to adapt to low-resource scenarios as in Vietnamese, even with limited labeled data.

4.1 Vietnamese multi-document summarization datasets

Currently, only three Vietnamese multi-document summarization datasets are available, posing a significant challenge for research advancement in this domain. These datasets include ViMs, VMDS, and VLSP.

The ViMs dataset [23], introduced by Nghiem et al., stands as a meticulously collected resource drawn from diverse domains, with Google News as its primary source. With a substantial count of 1,945 documents extracted from popular news websites in Vietnam, ViMs presents a robust and diverse dataset for multi-document summarization research. Notably, each cluster in the ViMs dataset contains two reference summaries, each assigned by different annotators. To select the most appropriate reference summary for each cluster, we calculate the Rouge Score of each sentence in the reference summary with respect to the entire text in the cluster and choose the sentence with the higher Rouge Score.

Another noteworthy dataset is the VMDS dataset [5], a valuable collection of multi-document data gathered from the Vietnamese newspaper website baomoi.com. It comprises 600 documents thoughtfully categorized into 200 topics, offering a diverse and informative set for analysis. Similar to ViMs, we choose the reference summary based on the sentence with the highest Rouge Score.

Furthermore, The VLSP dataset [6],which has been provided as part of a competition hosted by the Association for Vietnamese Language and Speech Processing. Encompassing Vietnamese news articles that cover a wide range of topics such as the economy, society, culture, science, and technology, the VLSP dataset serves as a valuable resource for advancing multi-document summarization techniques. In particular, The VLSP dataset includes both training and validation sets, consisting of 300 samples. Unlike the previous datasets, the VLSP dataset features only one reference summary for each cluster.

The statistic of these datasets are shown in Table 2.

[5] https://github.com/lupanh/VietnameseMDS

[6] https://vlsp.org.vn/vlsp2022/eval/abmusu

Item	ViMs	VMDS	VLSP
Number of documents	1945	628	925
Number of samples	300	200	300
Average number of documents per cluster	6.5	3.0	3.0
Average number of words per cluster	2208	1308	1853
Average number of words per summary	192	153	162

Table 2. The statistic of ViMs, VMDS, and VLSP datasets

Model	#Params	pre-trained	fine-tuned	En-Vi	Vi-En
M2M100	1.2B	-	CCMatrix+ CCAligned	35.83	31.15
Google Translate	-	-	-	39.86	35.76
Bing translator	-	-	-	40.37	35.74
Transformer-base	65M	-	PhoMT	42.12	37.19
Transformer-big	213M	-	PhoMT	42.94	37.83
mBART	448M	CC25	PhoMT	43.46	39.78
EnViT5-base	275M	CC100	MTet+ PhoMT	45.47	40.57

Table 3. Vietnamese translation model benchmark [11]

4.2 Vietnamese multi-document summarization dataset generation

In the early stages of our work, we addressed the lack of datasets problem by employing the NewsCorpus[7], a large Vietnamese single-document corpus, as a pre-training dataset. Later, we expanded our resources by leveraging an English multi-document dataset known as Newshead [3], which we translated into Vietnamese. The dataset pre-processing phase has two main steps including (1) calculating the Pyramid Rouge score, and (2) finding named entities with the extended NER labels by ultilizing the PhoNLP model.

NewsCorpus is one of the largest Vietnamese news datasets containing 14896998 documents with unlabeled summaries crawled from about 143 Vietnamese news websites. To enhance its utility, we undertook the task of transforming this dataset into a pseudo multi-document format. Each document within the dataset underwent segmentation into smaller parts, taking into consideration the length of the document. Consequently, our dataset now contains document segments ranging from 2 to 5 parts. It's worth noting that preprocessing this dataset is a time-intensive step, requiring roughly 2 to 3 weeks.

Despite the pseudo clusters in NewsCorpus being generated by chunking individual documents, which diverges from the typical definition of multi-document summarization that relies on utilizing documents from varied sources, this approach was chosen due to the lack of available data. Moreover, we believe that segmenting documents in this manner approximates the characteristics of multi-document summarization. By segmenting longer documents into smaller parts, each part can be treated as an individual document. This approach simulates the scenario where different documents contribute unique pieces of information to be put into a coherent summary. It helps in developing summarization models that can handle diverse and dispersed information, mimicking the essence of true multi-document summarization where the goal is to consolidate varied content into a unified summary.

[7] https://github.com/binhvq/news-corpus

However, one of the notable limitations of the NewsCorpus is its lack of authentic multi-document data, therefore, there is a need to create a Vietnamese multi-document dataset. Given the considerable time and effort required to construct such a dataset from scratch, we pursued a practical alternative: translating an existing English multi-document dataset. Our choice was NewsHead [3], the original pre-training dataset for PRIMERA [27]. The NewsHead dataset contains 5GB of English stories with around 300000 multi-document samples, with up to five documents per cluster. To ensure dataset quality, we took several measures:

- Comprehensive Translation: We translated not only the documents but also named entities extracted by Spacy models, ensuring that the entity recognition process during the pre-training phase would be accurate and complete.
- We choose the translation model based on a comparision among several well-known translation engines [12], including Google Translate, Bing Translator, Transformer-base, Transformer-large [25], mBART [2]) and EnViT5 [12] on PhoMT English-Vietnamese Translation Test Set [2] utilizing SacreBLEU [17] to compute the case-sensitive BLEU score in 3.

Due to constraints in GPU resources, we opted to translate the NewsHead dataset using Google Translate, a process that took 5 days to completion. The statistics of the datasets are listed in Table 1.

5 EXPERIMENTS

The primary goal of our work is to harness the potential of task-specific language models for addressing the challenges posed by low-resource languages, with a particular focus on the Vietnamese language. By leveraging and adapting a previously constructed large-scale multi-document dataset, we aim to demonstrate the effectiveness of this approach in the context of limited available data for Vietnamese. To achieve our research objectives, we have structured our experiments into two principal phases including: (i) pre-training step using our previously built large-scale datasets; (ii) fine-tuning the pre-trained models on Vietnamese low-resource context.

5.1 Pre-training phase

We present two variants of our LatVis model, each tailored to address the challenges posed by low-resource languages:

- $\text{LatVis}_{\text{NewsCorpus}}$: This variant is pre-trained on NewsCorpus dataset. In this configuration, our model undergoes pre-training utilizing 4 NVIDIA A100 40GB GPUs, requiring a convergence period of 4 days. It's noteworthy that the validation loss reaches approximately 2.16, while, in comparison, the validation loss for the English PRIMERA version [27] is approximately 1 when evaluated on an English pre-training dataset.
- $\text{LatVis}_{\text{NewsHead}}$: This variant is pre-trained with 1 NVIDIA A100 40GB GPU and takes about 10 days to converge. Remarkably, this approach achieves a more favorable validation loss of 1.26.

We pre-train our LatVis model with 496M backbone parameters, The input sequences extend to a length of 4096 tokens, while output sequences span 1024 tokens. We pre-train our model for 100K steps, with early stopping, batch size of 8, Adam optimizer with a learning rate of $3e{-}5$, with 10K warmup steps and linear decay.

5.2 Fine-tuning phase

We explore the model's performance across three experimental settings: zero-shot, few-shot, and fully supervised to compare the performance of our model with the other Vietnamese existing models on three Vietnamese public multi-document datasets. In the last part, we have shown

Models	ViMs			VMDS			VLSP		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
BARTPho$_{\text{large}}$ (our run)	48.90	37.24	32.55	39.98	30.86	26.92	34.41	22.22	25.32
GPT-3 (our run)	**69.35**	44.24	**40.42**	**73.50**	43.35	38.24	**69.20**	41.30	**37.45**
LatVis$_{\text{Newscorpus}}$	57.58	45.51	35.64	58.40	44.60	37.13	55.00	39.74	34.33
LatVis$_{\text{Newshead}}$	68.39	**51.29**	39.16	66.28	**46.75**	**40.90**	60.64	**41.59**	37.04

Table 4. Zero-shot setting comparison on different datasets for various models. Notes: The best scores are in bold and second best scores are underlined.

Models	ViMs			VMDS			VLSP		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
BARTPho$_{\text{large}}$ (our run)	74.51	45.41	41.24	72.56	46.03	43.64	67.83	36.15	37.74
ViT5$_{\text{large}}$ (our run)	77.54	53.96	**49.57**	75.57	48.62	46.17	73.03	43.54	39.90
LatVis$_{\text{Newscorpus}}$	77.21	52.37	46.6	75.45	49.57	45.85	**74.70**	**46.78**	42.67
LatVis$_{\text{Newshead}}$	**78.98**	**55.08**	48.00	**76.76**	**50.21**	**47.39**	73.20	46.74	**42.73**

Table 5. Test result on Vietnamese Multi-document Summarization datasets. Notes: The best scores are in bold.

results to show the impact of different Vietnamese tokenizers on the text summarization model. We train and infer models with different length limits (256/ 512/ 1024) in the zero-shot few-shot setting. We use Adam as the optimizer with linear scheduled learning rate 3e−5 and batch size 8 for all the datasets. The performance of the model has been evaluated using Rouge-1, Rouge-2 and Rouge-L.

5.2.1 Zero-shot evaluation. We focus on the evaluation of abstractive summarization in a zero-shot setting, where pre-trained language models are directly utilized without any additional fine-tuning on specific datasets. To ensure the fairness among models, we set the length limit of the output at inference time to the same length. Results indicate that our model achieve substantial result compared with all the baselines across multiple datasets. Our models, which have been pre-trained on document clusters with pseudo summaries, exhibit notably enhanced performance, especially when tasked with generating longer summaries. Our model demonstrates exceptional zero-shot performance, outperforming both ViT5 [15] and BARTpho [22] model. It's worth highlighting that ViT5 lacks zero-shot inference capabilities, and the performance of BARTpho is consistently lower compared to our LatVis model because they are only trained on single-document dataset-Newscorpus. Furthermore, we conducted a comparative analysis with the widely recognized GPT-3 model to further demonstrate the effectiveness of our approach. Table 4 shows the Rouge scores for different models on various datasets with zero-shot setting.

5.2.2 Few-shot evaluation. In our few-shot evaluation phase, we undertook the fine-tuning of our LatVis model across three Vietnamese multi-document summarization datasets: ViMs, VMDS, and VLSP. This evaluation aims to examine how well the pre-trained model could adapt to new multi-document summarization datasets providing additional training samples. The results in table 5 show that our initial model achieve competitive Rouge scores when benchmarked against the most recent state-of-the-art models in text summarization, indicating its effectiveness in generating accurate abstractive summaries for Vietnamese multi-document text.

Models	Vietnews		
	R-1	R-2	R-L
BARTpho	61.14	30.31	40.15
ViT5$_{\text{base 256-length}}$	61.85	31.70	41.70
ViT5$_{\text{base 1024-length}}$	62.77	33.16	<u>42.75</u>
ViT5$_{\text{large 1024-length}}$	**63.37**	**34.24**	**43.55**
LatVis$_{\text{Newscorpus}}$	62.63	32.57	42.10
LatVis$_{\text{Newshead}}$	<u>63.01</u>	<u>33.38</u>	<u>42.74</u>

Table 6. Comparison among various models on Vietnews. Notes: The best scores are in bold and second best scores are underlined.

5.2.3 Fully-supervised evaluation. It is essential to keep in mind the small size of the datasets, which may limit the generalizability of the results. Additionally, further evaluation on larger datasets and comparisons with other state-of-the-art models would provide a more comprehensive assessment of the LatVis performance. To mitigate this issue, we have taken a more practical approach by exploring fine-tuning on a larger single-document dataset, Vietnews [8]. This dataset comprises a collection of articles gathered from diverse categories, such as "the world", "news", "law", and "business", sourced from popular Vietnamese news websites such as https://tuoitre.vn, https://vnexpress.net, and https://www.nguoiduatin.vn as well. The Vietnews dataset comprises a total of 150704 samples, split into three sets for training (70%), development (15%), and testing (15%). It is impressive to see that my model, which was specifically designed for multi-document summarization, still performs well on the single-document dataset. This indicates the adaptability and versatility of the LatVis model, showcasing its potential for handling different summarization tasks. Table 6 details the result.

5.2.4 Impact of tokenizers. We assess the performance of the LatVis model when employing different tokenization approaches for multi-document summarization. The study involved assessing three distinct tokenization approaches: PhoBERT-large [13], ViT5 [15] and our trained tokenizer. Figure 4 illustrates the performance of pre-trained LatVis model, trained on translated and preprocessed Newshead data, using various tokenization methods. The models are then finetuned and evaluated on the ViMs dataset using Rouge F1 Score. Based on the outcomes obtained with different tokenization strategies, it becomes evident that the PhoBERT tokenizer consistently outperforms the other two options, namely the 34k-vocab tokenizer and ViT5 tokenizer.

Word segmentation and BPE have been used in Vietnamese language processing, as seen in previous works. For instance, PhoBERT [13] utilizes BPE to effectively handle the rich morphology and compound words in Vietnamese. These methods are crucial for Vietnamese due to the nature of the language, which lacks explicit word boundaries. Proper segmentation is essential for accurate tokenization and, consequently, better model performance. In our work, we experimented with these established techniques and evaluated their effectiveness in the context of multi-document summarization. Our findings confirm that using a well-optimized tokenizer, like PhoBERT, which incorporates BPE and word segmentation, leads to superior performance. This result is in line with other research in the field, suggesting that while our approach may not be novel in the application of these techniques, our empirical validation contributes valuable insights into their efficacy.

[8] https://github.com/ThanhChinhBK/vietnews

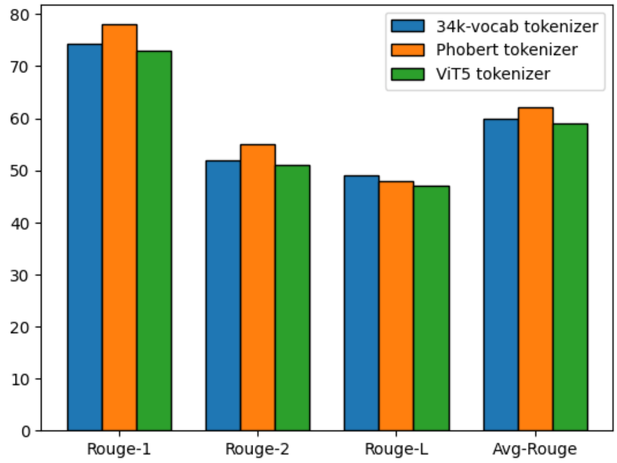

Fig. 4. LatVis model performance with different tokenizers fine-tuned and evaluated on the ViMs dataset

6 CONCLUSION

To sum up, we introduce LatVis, the first large-scale pre-trained task-specific language model which is specially pre-trained for the low-resource Vietnamese multi-document summarization task. By using Pyramid Entity strategy [27] to rank sentence importance, we generated the first large-scale Vietnamese dataset[9] that is useful in building multi-document summarization models. The pre-trained LatVis models achieve a comparable performance without any fine-tuning. After being fully fine-tuned on multi-document datasets, LatVis models obtained impressive performance on all given datasets. Despite the challenges posed by the scarcity of data and the potential for overfitting and the resource limitations pose additional substantial challenges, our approach to addressing these obstacles underscores the dedication to producing meaningful outcomes. By exploring diverse datasets and fine-tuning strategies, we have made a small contribution to the field of Vietnamese multi-document summarization, laying the foundation for future advancements.

Table 7 below provides some sample summaries generated from BARTPho, ViT5, GPT-3 and our LatVis. As detailed, LatVis and GPT-3 tends not to generate duplicate information and can handle long summaries much better than BARTPho or ViT5. We also translated the generated text to English for better understanding. The duplicated information in generated text is hightlighted in blue color.

Model	Summary

[9] the dataset with pseudo summaries

BARTPho	Ngày 1-6 , ông Lê Văn Khoa , Phó Chủ tịch UBND TP. HCM , cho biết sẽ xây dựng nút giao thông mới nhằm xoá điểm kẹt xe lên xuống cảng Cát Lái , nhằm xoá điểm kẹt xe đi qua ngày càng tăng mạnh (18.000-20.000 lượt xe / ngày) và khiến cho khu vực qua nút cũng như khu vực luôn rơi vào ùn tắc . Tại nút giao này , sẽ có hai làn xe ô tô vành đai 2 nhằm giảm tình trạng kẹt xe cục bộ , nhằm xoá điểm kẹt xe lên xuống cảng Cát Lái , tạo thuận lợi cho xe đi về giữa cầu Phú Mỹ và đường Nguyễn Thị Định . Vành đai 2 sẽ xây dựng cầu vượt thẳng qua nút giao theo hướng đường Vành đai 2 rộng 17 m , dài 316 m cho bốn làn xe ô tô (hướng từ cầu Phú Mỹ thẳng qua cầu Kỳ Hà 3 và ngược lại) . Cũng trong giai đoạn đầu xây dựng , trong giai đoạn đầu xây dựng cầu vượt thẳng qua nút giao theo hướng đường Vành đai 2 rộng 17 m , dài 316 m cho bốn làn xe ô tô (hướng từ cầu Phú Mỹ thẳng qua cầu Kỳ Hà 3 và ngược lại) ; theo hướng từ cầu Kỳ Hà 3 sẽ xây dựng hầm chui dưới nút giao cắt rẽ về đường Nguyễn Thị Định để xuống cảng Cát Lái . Cùng với đó , thành phố sẽ xây dựng các công trình giao thông nhằm giảm lưu lượng xe qua nút giao thông trên đường Nguyễn Thị Định , Nguyễn Thị Định để giảm bớt áp lực
	On June 1st, Mr. Le Van Khoa, Vice Chairman of the People's Committee of Ho Chi Minh City, announced plans to construct a new traffic interchange in order to alleviate traffic congestion at the entrance and exit of Cat Lai Port, in order to alleviate traffic congestion (18,000-20,000 vehicles per day) and prevent the area around the intersection from constantly experiencing gridlock. At this intersection, there will be two lanes of the Beltway 2 for the purpose of reducing local traffic congestion, in order to alleviate traffic congestion to and from Cat Lai Port, facilitating traffic between Phu My Bridge and Nguyen Thi Dinh Street. Beltway 2 will construct a straight overpass bridge at the intersection in the direction of the road Beltway 2, 17 meters wide, 316 meters long with four lanes for cars (from Phu My Bridge straight through Ky Ha 3 Bridge and vice versa). Also, in the initial construction phase, in the initial construction phase, the straight overpass bridge at the intersection in the direction of the road Beltway 2, 17 meters wide, 316 meters long with four lanes for cars (from Phu My Bridge straight through Ky Ha 3 Bridge and vice versa) will be built; in the direction from Ky Ha 3 Bridge, an underground tunnel will be constructed under the intersection to turn onto Nguyen Thi Dinh Road to reach Cat Lai Port. Along with that, the city will construct transportation infrastructure projects to reduce the traffic flow at the intersection on Nguyen Thi Dinh Road, Nguyen Thi Dinh Road in order to alleviate congestion.
ViT5	Ngày 3/6 , UBND TP HCM khởi công xây dựng Dự án cầu vượt , hầm chui vòng xoay Mỹ Thuỷ tại phường Cát Lái , quận 2 . Theo ông Lê Văn Khoa , việc triển khai dự án sẽ phải thu hồi , giải toả hơn 5,76 ha đất , làm ảnh hưởng đến cuộc sống của gần 150 hộ dân thuộc phường Cát Lái và Thạnh Mỹ Lợi . Theo ông , việc triển khai dự án sẽ phải thu hồi , giải toả hơn 5,76 ha đất , làm ảnh hưởng đến cuộc sống của gần 150 hộ dân thuộc phường Cát Lái . Dự án cầu vượt Mỹ Thuỷ được chia làm hai giai đoạn đầu tư : Giai đoạn 1 và giai đoạn 3 xây dựng cầu vượt , hầm chui dưới nút giao rẽ trái về đường Nguyễn Thị Định , cầu Kỳ Hà 3 sẽ hình thành vòng xoay như hiện nay . Trên mặt nền hiện hữu , từ các hướng đường Nguyễn Thị Định và Vành đai 2 sẽ xây dựng bốn nhánh rẽ phải : Từ đường Nguyễn Thị Định đi đường Vành đai 2 , nối cầu Phú Mỹ qua khu công nghệ cao ; nút giao thông Phú Hữu dẫn lên cao tốc Long Thành - Dầu Giây . Xây dựng cầu vượt Mỹ Thuỷ là công trình trọng điểm nhằm giảm ùn tắc giao thông tại cửa ngõ vào cảng Cát Lái , tạo thuận lợi cho xe đi về giữa cầu Phú Mỹ và đường Vành đai Đông (Vành đai 2) , hướng về đường cao tốc Long Thành - Dầu Giây . Xây dựng đường chui dưới lòng đất tại nút giao sẽ hình thành vòng xoay như hiện nay .

	On June 3rd, the People's Committee of Ho Chi Minh City commenced the construction of the My Thuy roundabout overpass and tunnel project in Cat Lai Ward, District 2. According to Mr. Le Van Khoa, the implementation of the project will require the reclamation and resettlement of more than 5.76 hectares of land, affecting the lives of nearly 150 households in Cat Lai Ward and Thanh My Loi. According to Mr. Le Van Khoa, the implementation of the project will require the reclamation and resettlement of more than 5.76 hectares of land, affecting the lives of nearly 150 households in Cat Lai Ward. The My Thuy Overpass project is divided into two investment phases: Phase 1 and Phase 3 for the construction of the overpass and the tunnel under the left-turn junction to Nguyen Thi Dinh Street. Ky Ha 3 Bridge will form a roundabout as it is currently. On the existing ground, from the directions of Nguyen Thi Dinh Street and Belt Road 2, four right-turn branches will be constructed: From Nguyen Thi Dinh Street to Belt Road 2, connecting to Phu My Bridge through the high-tech industrial zone; the Phu Huu interchange leading to the Long Thanh - Dau Giay expressway. The construction of the My Thuy Overpass is a key project aimed at reducing traffic congestion at the entrance to Cat Lai Port, providing convenience for vehicles traveling between Phu My Bridge and the Eastern Belt Road (Belt Road 2), heading towards the Long Thanh - Dau Giay expressway. The construction of the underground tunnel at the interchange will form a roundabout as it is currently.
GPT-3	UBND TP.HCM đã khởi công xây dựng dự án nút giao thông Mỹ Thuỷ tại phường Cát Lái, quận 2. Dự án này nhằm giảm ùn tắc, tai nạn giao thông tại cửa ngõ vào cảng Cát Lái. Nút giao Mỹ Thuỷ là điểm giao của các tuyến quan trọng như đường Nguyễn Thị Định - Đồng Văn Cống, đường Vành đai 2 và nút giao thông Phú Hữu. Hiện tại, lượng xe lưu thông qua nút giao này đang gia tăng đáng kể, ảnh hưởng đến việc vận chuyển hàng hoá vào cảng Cát Lái. Vì vậy, dự án này sẽ tăng khả năng thông xe, giảm ùn tắc và đáp ứng tốt nhu cầu vận chuyển hàng hoá vào cảng Cát Lái và kết nối đường Vành đai 2. Trên hành lang của nút giao sẽ có hầm chui và cầu vượt để tách dòng xe, đảm bảo an toàn giao thông. Dự án sẽ được triển khai theo hai giai đoạn, với tổng mức đầu tư khoảng 2.337 tỷ đồng. Giai đoạn 1 dự kiến hoàn thành vào quý I/2018, và giai đoạn 2 được dự kiến hoàn thành sau đó.
	The People's Committee of Ho Chi Minh City has commenced the construction of the My Thuy traffic interchange project in Cat Lai Ward, District 2. This project aims to alleviate congestion and traffic accidents at the entrance to Cat Lai Port. The My Thuy interchange serves as a crucial intersection for important routes such as Nguyen Thi Dinh - Dong Van Cong Road, Ring Road 2, and the Phu Huu traffic interchange. Currently, the volume of vehicles passing through this interchange is significantly increasing, affecting the transportation of goods to Cat Lai Port. Therefore, this project will enhance traffic flow, reduce congestion, and effectively meet the demand for transporting goods to Cat Lai Port and connecting to Ring Road 2. The interchange will include tunnels and overpasses to separate traffic streams and ensure road safety. The project will be implemented in two phases, with a total investment of approximately 2,337 billion Vietnamese Dong. Phase 1 is expected to be completed in the first quarter of 2018, followed by Phase 2.

LatVis	Sáng 3/6 , UBND TP HCM khởi công xây dựng Dự án cầu vượt , hầm chui vòng xoay Mỹ Thuỷ tại phường Cát Lái , quận 2 . Nút giao thông Mỹ Thuỷ là điểm giao của các tuyến quan trọng là đường Nguyễn Thị Định - Đồng Văn Cống dẫn vào cảng Cát Lái ; đường Vành đai 2 nối cầu Phú Mỹ qua khu công nghệ cao ; nút giao thông Phú Hữu dẫn lên cao tốc Long Thành - Dầu Giây . Theo ông Lê Văn Khoa , Phó Chủ tịch UBND TP , nút giao bằng hiện hữu không còn đáp ứng nổi lượng xe đi qua ngày càng tăng mạnh (18.000-20.000 lượt xe / ngày) và khiến cho khu vực qua nút cũng như xung quanh luôn rơi vào ùn tắc . Dự án sẽ được thi công hai giai đoạn . Giai đoạn 1 với số vốn khoảng 840 tỷ đồng , giai đoạn 2 (hoàn chỉnh) khoảng 1.400 tỷ đồng , trong đó chi phí xây dựng 1.000 tỷ , giải phóng mặt bằng gần 400 tỷ . Dự kiến cuối năm 2018 , toàn bộ dự án sẽ khánh thành , đưa vào sử dụng . Việc xây dựng nút giao mới khác nhằm xoá điểm kẹt xe lên xuống cảng biển Cát Lái , tạo thuận lợi cho xe đi về giữa cầu Phú Mỹ và đường vành đai Đông (Vành đai 2) , hướng về đường cao tốc TP. HCM - Long Thành - Dầu Giây .
	On June 3rd, the People's Committee of Ho Chi Minh City started the construction of the My Thuy Roundabout Overpass and Tunnel Project in Cat Lai Ward, District 2. The My Thuy traffic interchange is a crucial junction for several important routes, including Nguyen Thi Dinh Street, Dong Van Cong Street leading to Cat Lai Port, Belt Road 2 connecting Phu My Bridge to the high-tech industrial zone, and the Phu Huu traffic interchange leading to the Long Thanh - Dau Giay Expressway. According to Mr. Le Van Khoa, Deputy Chairman of the People's Committee of Ho Chi Minh City, the existing interchange can no longer handle the increasing volume of traffic, which has reached 18,000-20,000 vehicles per day, leading to constant congestion in the area and its surroundings. The project will be executed in two phases. Phase 1 will require approximately 840 billion VND in funding, while Phase 2 (completion) will cost around 1,400 billion VND, with construction costs accounting for 1,000 billion VND and land clearance expenses of nearly 400 billion VND. It is expected that by the end of 2018, the entire project will be completed and put into operation. The construction of the new interchange aims to alleviate traffic congestion to and from Cat Lai Port, facilitating the movement of vehicles between Phu My Bridge and the Eastern Belt Road (Belt Road 2), as well as towards the Ho Chi Minh City - Long Thanh - Dau Giay Expressway.
Gold summary	Theo ông Lê Văn Khoa, Phó Chủ tịch UBND TP, nút giao bằng hiện hữu không còn đáp ứng nổi lượng xe đi qua ngày càng tăng mạnh (18.000-20.000 lượt xe / ngày) và khiến cho khu vực qua nút cũng như xung quanh luôn rơi vào ùn tắc. Trong giai đoạn I, chủ đầu tư sẽ đầu tư các công đoạn như : Xây dựng 1/2 chính nằm trên Vành đai 2 vượt qua đường Nguyễn Thị Định và rạch Mỹ Thuỷ ; Xây dựng một hầm chui cho các phương tiện giao thông rẽ trái từ đường Vành đai 2 vào đường Nguyễn Thị Định đến cảng Cát Lái ; Xây dựng 1/2 cầu Kỳ Hà 3 ; Xây dựng 1 đường nhánh rẽ phải cho các cklưu thông từ cảng Cát Lái vào Vành đai 2 ; Xây dựng các đường chui dành riêng cho xe hai bánh dưới các cầu Kỳ Hà 2, Kỳ Hà 3, Mỹ Thuỷ cơ bản Cải tạo vòng xoay hiện hữu. Việc xây dựng nút giao này nhằm kéo giảm tình trạng ùn tắc giao thông, TNGT, phục vụ phát triển kinh tế vùng trọng điểm phía Nam ". Theo ông Khoa, lễ khởi công hôm nay mới là bước khởi đầu của quá trình xây dựng dự án. Do đó lãnh đạo thành phố đề nghị các sở ngành, đơn vị liên quan cần tập trung nỗ lực thực hiện các công việc theo chức năng, nhiệm vụ của mình để công trình đúng tiến độ, chất lượng.

	According to Mr. Le Van Khoa, Vice Chairman of the People's Committee of the city, the existing intersection can no longer accommodate the increasing volume of traffic passing through (18,000-20,000 vehicles per day), causing constant congestion in the area and its surroundings. In Phase I, the investor will invest in stages such as: constructing half of the main road overpassing Belt Road 2 over Nguyen Thi Dinh Street and My Thuy Canal; building an underpass for left-turning traffic from Belt Road 2 to Nguyen Thi Dinh Street to Cat Lai Port; constructing half of Kỳ Hà 3 Bridge; building a branch road for traffic from Cat Lai Port to Belt Road 2; constructing dedicated lanes for two-wheeled vehicles under Kỳ Hà 2, Kỳ Hà 3, and basic renovation of the existing roundabout. The construction of this intersection aims to alleviate traffic congestion, reduce traffic accidents, and serve the economic development of the key southern region. According to Mr. Khoa, today's groundbreaking ceremony marks the beginning of the project's construction process. Therefore, the city's leadership recommends that relevant departments and units focus on their respective functions and tasks to ensure that the project progresses on schedule and meets the required quality standards.

Table 7. Gold and generated summaries on the ViMs dataset. The text in blue indicates the repeated information.

One of the most significant challenges we encountered when applying the PRIMERA model is the limited availability of Vietnamese datasets. The pre-training dataset plays a crucial role in constructing effective models for Vietnamese text summarization. Its quality directly impacts the performance of our model. However, the existing large datasets primarily cater to single-document tasks or translated datasets, which may not be an optimal fit for our multi-document summarization task. In the long run, it becomes imperative to gather a substantial news dataset comprising clusters of Vietnamese documents with shared topics, aligning better with the objectives of pre-training.

Additionally, further exploration and experimentation are essential concerning Vietnamese Named Entity Recognition (NER) and Tokenizer models. Furthermore, we are currently researching and developing methods to compute sentence scores or determine important sentences instead of methods used in PRIMERA paper. This approach involves considering various information aspects to score the key sentences. Investigating the results of training these models from scratch on Vietnamese multi-document news datasets can provide valuable insights into enhancing performance.

REFERENCES

[1] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150* (2020).

[2] Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. 2021. PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation. arXiv:2110.12199 [cs.CL]

[3] Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating Representative Headlines for News Stories. In *Proc. of the the Web Conf. 2020*.

[4] Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. arXiv:1804.10959 [cs.CL]

[5] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, 66–71. https://doi.org/10.18653/v1/D18-2012

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[7] Haoran Li, Arash Einolghozati, Srinivasan Iyer, Bhargavi Paranjape, Yashar Mehdad, Sonal Gupta, and Marjan Ghazvininejad. 2021. EASE: Extractive-Abstractive Summarization with Explanations. In *Proceedings of the Third Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, 85–95.

[8] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Online, 3730–3740.

[9] Yixin Liu and Pengfei Liu. 2021. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. In *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)*. Association for Computational Linguistics, 1065–1072.

[10] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing Order to Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2890–290.

[11] Chinh Ngo, Trieu H. Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang Luong. 2022. MTet: Multi-domain Translation for English and Vietnamese. https://doi.org/10.48550/ARXIV.2210.05610

[12] Chinh Ngo, Trieu H. Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang Luong. 2022. MTet: Multi-domain Translation for English and Vietnamese. arXiv:2210.05610 [cs.CL]

[13] Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. *CoRR* abs/2003.00744 (2020). arXiv:2003.00744 https://arxiv.org/abs/2003.00744

[14] Linh The Nguyen and Dat Quoc Nguyen. 2021. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. *CoRR* abs/2101.01476 (2021). arXiv:2101.01476 https://arxiv.org/abs/2101.01476

[15] Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Association for Computational Linguistics, 136–142. https://aclanthology.org/2022.naacl-srw.18

[16] Do Phuc and Mai Xuan Hung. 2008. Using SOM based graph clustering for extracting main ideas from documents. In *2008 IEEE International Conference on Research, Innovation and Vision for the Future in Computing and Communication Technologies*. 209–214. https://doi.org/10.1109/RIVF.2008.4586357

[17] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. arXiv:1804.08771 [cs.CL]

[18] To Huy Quoc, Kiet Van Nguyen, Ngan Luu-Thuy, and Nguyen Anh Gia-Tuan. 2021. Monolingual vs multilingual BERTology for Vietnamese extractive multi-document summarization. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Association for Computational Lingustics, Shanghai, China, 692–699. https://aclanthology.org/2021.paclic-1.73

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[20] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. arXiv:1508.07909 [cs.CL]

[21] Le Ha Thanh, Thang Huynh Quyet, and Mai Luong Chi. 2005. A Primary Study on Summarization of Documents in Vietnamese. In *IFSR 2005: Proceedings of the First World Congress of the International Federation for Systems*. JAIST Press.

[22] Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*.

[23] Nhi Thao Tran, Minh Quoc Nghiem, Nhung TH Nguyen, Ngan Luu Thuy Nguyen, Nam Van Chi, and Dien Dinh. 2020. ViMs: a high-quality Vietnamese dataset for abstractive multi-document summarization. *Language Resources and Evaluation* 54, 4 (2020), 893–920.

[24] Van-Giau Ung, An-Vinh Luong, Nhi-Thao Tran, and Minh-Quoc Nghiem. 2015. Combination of features for vietnamese news multi-document summarization. In *Proceedings of The Seventh International Conference on Knowledge and Systems Engineering (KSE)*. 186–191.

[25] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *ArXiv* abs/1706.03762 (2017).

[26] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. *CoRR* abs/1801.01331 (2018). arXiv:1801.01331 http://arxiv.org/abs/1801.01331

[27] Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 5245–5263.

[28] Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 5245–5263. https://doi.org/10.18653/v1/2022.acl-long.360

[29] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-Aware Neural Extractive Text Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5021–5031. https://doi.org/10.18653/v1/2020.acl-main.451

[30] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv:1912.08777 [cs.CL]

[31] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive Summarization as Text Matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6197–6208. https://doi.org/10.18653/v1/2020.acl-main.552

[32] Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for Effective Neural Extractive Summarization: What Works and What's Next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1049–1058. https://doi.org/10.18653/v1/P19-1100