Pre-training LongT5 for Vietnamese Mass-Media Multi-document Summarization Task

Nicolay Rusnachenko[1][0000−0002−9750−5499], The Anh Le[2,3][0000−0003−0740−6380], and Ngoc Diep Nguyen[3]

[1] Bauman Moscow State Technical University
rusnicolay@gmail.com
[2] Vietnam Maritime University, Hai Phong, Viet Nam
anhlt@vimaru.edu.vn
[3] CyberIntellect, Moscow, Russia
diepnn83@gmail.com

Abstract Multi-document Summarization task aimed to extract the most salient information from the set of input documents. One of the main challenges that face this task is a long-term dependency problem. When we deal with texts written in Vietnamese it is also accompanied by the specific syllable-based text representation, and lack of labeled datasets. The recent advances in machine translation problem results in a significant impact on the related architecture, dubbed as *transformers*. Being pre-trained on large amounts of raw texts, transformers allows providing a deep knowledge of the texts. In this paper, we survey the findings of the language model applications for text summarization problems, including the remarkable Vietnamese text summarization models. According to the latter, we select LongT5 to pre-train and then fine-tune it for the Vietnamese Multi-document text summarization problem from scratch. We provide a result model analysis and experiments with Multi-document Vietnamese datasets, including ViMs, VMDS, and VLSP2022. We conclude that using a transformer-based model pre-trained on a large amount of unlabeled Vietnamese texts allows us to achieve promising results, with further enhancement via fine-tuning within the small amount of manually summarized texts. The pre-trained model utilized in the experiment section is published[4].

Keywords: Vietnamese Multi-document Summarization · Text Summarization · Transformers · Language Models

1 Introduction

At present, the drastically huge growth of news and event recordings becomes one of the main reasons why most of the mass-media platforms become saturated with mass-media information. Such factor becomes a crucial for manual daily news reading making the related approach unfeasible. As a task *text summarization* [12] aims to create a short version of the original texts by keeping the most

[4] https://github.com/nicolay-r/ViLongT5

concise, coherent, and salient information. Shortening the long documents by keeping the most meaningful information represents a quite consumptive task for manual execution involving the analysis and content understanding. To the best of our knowledge, such factors necessities studies in automatic text summarization approaches and systems built upon them. In terms of the result summary, such systems might be categorized as *extractive*, or *abstractive*. Summarization systems of extractive type [7] aim to rank sentences in the given text by relying on their meaning and importance, with further extraction of the high-ranked one. In turn, the abstractive type systems are focused on generative result in essay format for a given text [21,9,19].

The appearance of an attention mechanism that addresses the problem of capturing distant information in long input sequences in the Machine Translation (MT) task [2] cause a significant impact on further studies and attention implementations [28,33]. The attention mechanism represents a module in the neural network which aims to assess the importance of the given information by assigning *weights* to its components. A significant amount of investigations were directed to experiments with attention implementations as well as the integration of such modules into target-oriented machine learning models aside from the MT, including the text summarization domain. The further appearance of the *self-attention* mechanism [28] as an internal component of encoder-decoder architecture, results in a *transformer*. Transformer-based models cause a significant breakthrough in MT, resulting in further modifications [33]. The transition towards texts of a single language for transformers results in the appearance of *language models* that become recently both popular and standardized solutions in other natural language processing (NLP) domains including text summarization [9,32,15]. This paper focuses on the analysis of the recent advances of language-models to choose the promising solution for the Vietnamese Multi-document Summarization problem [27,17] of mass-media documents. It is worth noting that Multi-document Summarization faces the problem of the long contents where the importance of information might be spread in different per each document. To the best of our knowledge, we are the first who pretrain and fine-tune the Vietnamese LARGE-sized LongT5 model for Multi-document text summarization from scratch.

The remaining part of this paper is organized as follows. Section 2 provides an overview of the recent advances of transformer-based models along with their architectural updates and training techniques, with Vietnamese texts oriented models in Section 2.1 and sparse attention-based models in Section 2.2. Section 3 lists the resources that were adopted in LongT5 model pre-training and fine-tuning. The detailed description of the model pre-training process as well as the further experiments are covered in Sections 4 and 5 respectively, including comparison with the other extractive and abstractive text summarization baselines for $\text{VLSP2022}_{\text{valid}}$ and $\text{VLSP2022}_{\text{test}}$ datasets.

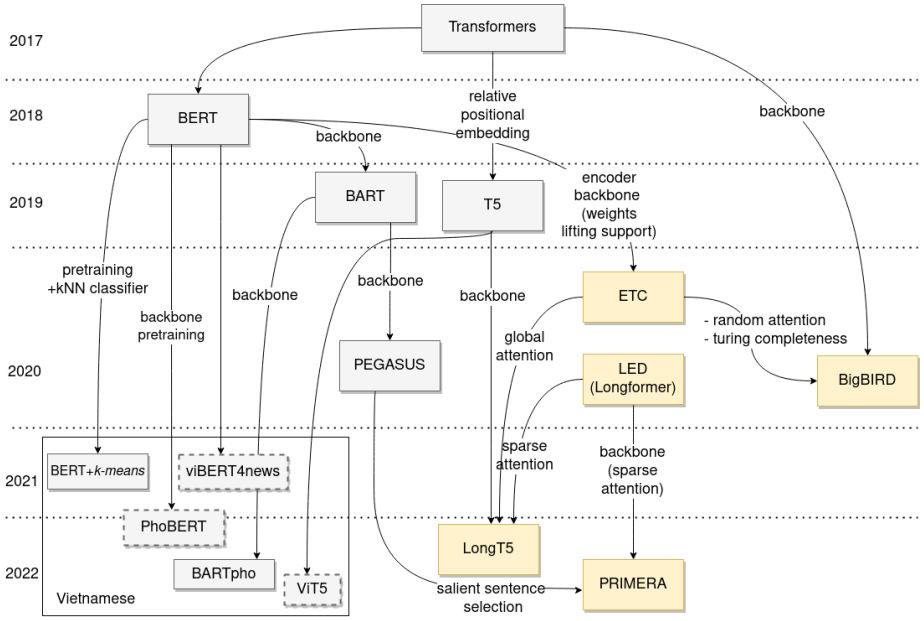

Figure 1. Tree diagram of the transformer-based models [28], placed in order of their appearance from top to bottom; *arrows* illustrate the most significant findings were found in successor models; *blocks* illustrate models with: original self-attention (gray), sparse self-attention mechanism (yellow); the highlight Vietnamese-targeted models for text summarization problems are bordered; trained/finetuned states are dotted.

2 Background

Since the text summarization problem is commonly treated as an extractive or abstractive tasks, both encoder and decoder components of the transformers could be used as *backbones*. Considering BERT architecture as a backbone, it finds its application in extractive-based text summarization problems. Due to the architecture specifics, which are considered to encode information bidirectionally, BERT could not be easily adopted for generative task format [5]. As for extractive task format, we may consider BERT as a sentence encoder, complemented with a clustering-type algorithm.

However, to address the generative limitations, in [11] the authors propose the BART framework, which represents a BERT (bidirectional transformer) complemented by an autoregressive decoder (GPT). BART proposed a denoising sequence-to-sequence framework, in which the pre-training stage includes: (1) corrupted text restoration and (2) original text reconstruction, i.e. translation. Architecturally, BART is a standard Transformer-based neural machine translation architecture [28] with the potential for customization of encoder and decoder transformer parts, including pre-training schemes modification. Being particularly effective for text generation tasks, including text summarization,

at the time of the model announcement, the authors mention a significant improvement of LARGE-sized BART over previous works on the XSum [14] dataset (Table 2).

BART has become a fundamental architecture for a variety set of text summarization oriented frameworks, such as follows. In [32], authors propose PEGASUS framework, in which the sentence-based masking strategy is based on the invented *salient sentence selection* algorithm. With the latter, authors proposed a sentence assessment metric with a limited selection of the top k-scored sentences. According to the extensive experiments on XSum and CNN/DailyMail [13] collections with LARGE-sized models (Table 2), authors illustrate that the result PEGASUS model [20] outperforms the other transformer-based solutions, such as BART, T5 [21]. Text-To-Text Transfer Transformer (T5) [21] is based on the original transformer [28] complemented by the following modifications toward layer normalization techniques and token positioning [23]. Analyzing the results of LARGE-based versions, the T5 model with the *principle sentences generation* strategy [32] in pretraining significantly outperforms the rest of the models discussed above on several common datasets (see Table 2).

2.1 Vietnamese Multi-document Summarization Models

One of the main traits of Vietnamese texts is *syllable-based* sentence segmentation – the atomic part of sentences are *syllables*. To the best of our knowledge, recent advances in Vietnamese text processing for Multi-document Summarization problems are limited by the original self-attention-based transformers application. Figure 1 illustrates the recent advances in transformer-based models for Vietnamese (bordered bottom-left corner). In this section, we overview the recent advances in *extractive* and *abstractive* text summarization approaches.

For extractive text summarization, several studies of non-transformer-based approaches application [18] address such training techniques examination as: *distant supervision*, *supervised learning*. The adaptation of BERT towards the downstream tasks for texts in Vietnamese results in the appearance of PhoBERT [16], and viBERT4news[5]. In [24] authors combines Vietnamese-oriented BERT-based pretrained states and *k-means*, and the result BERT+*k-means* illustrated top results on the VMDS[6] dataset compared with prior methods. The result of the related models is illustrated in Table 1.

In the case of abstractive summarization, BARTpho [25] represents an initial study with BART-based [11] architecture application towards the domain of Vietnamese texts. The authors mentioned the importance of vocabulary by gluing syllables into complete words. Due to the latter and in terms of BERT-based approaches, the word-based model performed better than the default, syllable-based representation, and tokenization. Recently, authors of ViT5 [19] experimented with a transformer-based encoder-decoder model for the Vietnamese language, based on the T5 self-supervised pretraining. The latter illustrates the

[5] https://huggingface.co/NlpHUST/vibert4news-base-cased
[6] Dataset details in Section 3

Table 1. Results of the Vietnamese oriented text summarization models [18,24] in Rouge Scores F1 in percents for ViMs and VMDS datasets; best and second best results are bolder and underlined respectively, separately for non-transformer based models and BERT-based.

Model	ViMs R-1	ViMs R-2	VMDS R-1	VMDS R-2
LSA	62.5	36.0	62.9	37.0
LexRank	<u>69.5</u>	**46.4**	48.2	39.2
TextRank	62.8	41.6	66.2	40.8
SVR	64.5	39.7	66.9	44.3
SVMRank	63.5	41.0	<u>67.4</u>	<u>46.2</u>
MART	65.1	42.4	**70.2**	**49.6**
CNN	56.1	42.1	52.8	40.0
LSTM	**70.7**	<u>43.1</u>	52.5	39.6
XLM-R-large + *k-means*	—		<u>77.4</u>	<u>51.2</u>
PhoBERT-large + *k-means*	—		<u>77.4</u>	50.9
viBERT4news + *k-means*	—		**77.4**	**52.0**

recent advances in *abstractive text summarization* and *named entity recognition* (NER) problems [19].

2.2 Sparse Self-Attention

The main task solved by attention is the connectivity of the particular token with respect to the other mentioned in the text. However, the crucial payment of this solution lies in its computational ineffectiveness. The computational complexity of full self-attention for an input size of n is $O(n^2)$ [28]. Besides the BERT [5], such mentioned models as BART [11] and T5 [21] nested the self-attention mechanisms, and hence in practice input sequences tend to be limited by 512 tokens [31].

To address the shortcomings of the self-attention application towards longer input sequences, the series of independent works were accomplished to its *sparse* variations [3,1,31]. To manage attention behavior on that matter, in [1] authors propose Extended Transformer Construction (ETC). Alongside the other works, authors invent *relative token positioning* [3,31] as a preliminary step for attention sparsification. To distribute attention between distant tokens, authors introduce a *global-local* attention mechanism by expanding the original (local) input with *global tokens* under the following restriction: the length of the global token sequence (n_g) is expected to be significantly less than the original input sequence length (n_l). Considering the latter, authors split attention calculation into parts and prove the resulting complexity of $O(n_g^2 + n_g \cdot n_l)$ remains linearly dependent on the original input length n_l. The relative token positioning encoding as long as sparse attention implementation [31] allows to train ETC with longer input sequences and hence caused a significant impact on the result performance in question answering (QA) tasks [1]. In [3], authors propose Longformer and the

Table 2. LARGE-sized transformer-based model performances in text summarization problems; models are grouped by self-attention mechanism into original self-attention [28] (512 tokens input limit) and sparsed version (4K+ token input limit); dataset names with best results are bolded; best and second best results are highlighted in gray; according to the results, models with sparse attention tend to perform better due to the longer input sequences.

Model	Architectural Features	Dataset	R-1	R-2	R-L
BART [11]	Bidirectional encoder + autoregressive decoder	XSum	45.14	22.27	37.25
PEGASUS [32]	Transformer + Gap-Sentence Selection	**CNN/DailyMail**	44.17	21.47	41.11
		Multi-News	47.52	18.72	24.91
		arXiv	44.21	16.95	38.83
T5 [21]	Transformer + relative token positioning + layer norm bias and normalization changes PEGASUS pretraining strategy	CNN/DailyMail	43.41	20.99	40.77
		Multi-News	47.48	18.60	24.31
		BigPatent	67.05	52.24	58.70
		arXiv	45.86	18.40	41.62
		PubMed	48.94	22.92	45.40
LED (16K) [3]	Transformer with windowed attention	arXiv	46.63	19.62	41.83
BigBird-PEGASUS [31]	LED + sparse attention (encoder side) + random attention mask PEGASUS (PSG) pretraining strategy	arXiv	46.63	19.02	41.77
		PubMed	46.32	20.65	42.33
		BigPatent	60.64	42.46	50.01
PRIMERA [30]	Longformer, Entity Pyramid Strategy	arXiv	47.60	20.80	42.60
		Multi-News	49.90	21.10	25.90
LongT5 (4K) [9]	T5 + global-local attention from LED	CNN/DailyMail	42.49	20.51	40.18
		BigPatent	70.38	56.81	62.73
		arXiv	48.28	21.63	44.11
		PubMed	49.98	24.69	46.46

related encoder-decoder (LED), which represents a modification of the original transformer with *windowed attention* variations. Similar to implementation in ETC, the latter denotes that, for a particular token only r (radius parameter) left and right neighbored tokens are considered to attend. Figure 1 illustrates models with sparse-attention mechanisms (yellow colored). Authors experiment with LARGE sized models towards arXiv summarization dataset [4] and illustrate a better performance of LED (447M params) over PEGASUS (`4K`) and equal to BigBird (`4K`) once input size has been increased from `4K` up to `16K` tokens. [3] LED architecture caused a significant effect on models that appear further, such as PRIMERA [30] with salient sentences masking approach, LongT5 [9] described further in this section.

Alongside the findings of the ETC application, in [31] authors treat the computational problem of self-attentive mechanism connection as a *graph sparsification*. Complementing sliding window and global attention mechanisms [1] with Erdős-Rényi model [6] of independently choosing edge with fixed probability, authors aim to prove Turing Completeness of the sparse attention mechanism behind the proposed BigBird model, which is computationally linear in the number of tokens. The outcome of the latter is as follows: the more sparse the graph, the more layers are required to reach completeness. For text summarization, authors experiment with sparse attention at encoder side[7], using pre-trained

[7] Since the output is relatively short compared with the size of input

schemes from PEGASUS [32] for LARGE-sized models. The resulting model is dubbed BigBird-PEGASUS [31].

LongT5 represents a modified version of the T5 [21] which adopts the sparse attentive mechanism variations, proposed with the ETC model, including windowed attention and global-local variation, dubbed as TGlobal [9]. The latter introduces local sparsity in the attention mechanism, which allows the reduction of the quadratic cost when scaling to long inputs. Unlike T5, the modified LongT5 can handle longer input sequences before reaching the out-of-memory exceptions. It is worth noting LongT5 (4K input) reaching top results across the variety of text-generative on almost every text summarization dataset: arXiv summarization dataset, PubMed, BigPatent [22], and MediaSum [9]. As for PRIMERA (447M) model, the latter illustrates the best results in MultiNews across other models listed in Table 2 due to the specifics and news-related information utilized at the pretraining stage. Analyzing the results across multiple datasets, LongT5 illustrates the best performance across the other models discussed above. The cost of the LongT5 architectural traits is the leveraged amount of the hidden parameters. The LARGE-sized version of LongT5 [20] with a 4K input token size results in ≈780M parameters, which is almost two times larger than PRIMERA (447M) and comparable with the size of LED with 16K token input size.

3 Resources

To the best of our knowledge, there are few Vietnamese single-document summarization datasets and only three Vietnamese multi-document summarization datasets. All of them are abstractive datasets. The details of these datasets are described below, with the brief statistics described in Table 3.

NewsCorpus[8] represents a relatively large collection of 14.9M documents with unlabeled summaries crawled from about 143 Vietnamese news websites. This can be treated as a single-document summarization dataset, in which each document yields the title and sampled content.

VMDS[9] is a multi-document dataset collected from a Vietnamese online news provider baomoi.com. This dataset contains 628 documents categorized into 200 topics.

ViMs[10] represents a multi-document dataset released by Nghiem et al. [26]. This corpus was collected from different Google News domains. In total, the authors collect 1945 documents from popular news websites in Vietnam.

VLSP2022[11] is a dataset is provided in a competition hosted by the Association for Vietnamese Language and Speech Processing. The provided data is Vietnamese news on various topics, including the economy, society, culture, science, and technology. Every document includes: title, anchor text and body text of single documents, summary, category tag. It is divided into train ($\text{VLSP2022}_{\text{train}}$)

[8] https://github.com/binhvq/news-corpus
[9] https://github.com/lupanh/VietnameseMDS
[10] https://github.com/CLC-HCMUS/ViMs-Dataset
[11] https://vlsp.org.vn/vlsp2022/eval/abmusu

Table 3. Statistics of the Vietnamese datasets utilized for the model training and evaluation; NewsCorpus dataset represents only raw clustered documents without summaries.

Dataset	#doc	#samples	#docs per cluster	#words per document	#words per summary
NewsCorpus	14 896 998	–	–	–	–
VMDS	628	300	3.00	1308.00	153.00
ViMs	1 945	300	6.50	2208.00	192.00
$\text{VLSP2022}_{\text{train}}$	621	200	3.11	1925.75	168.48
$\text{VLSP2022}_{\text{valid}}$	304	100	3.04	1815.41	167.68
$\text{VLSP2022}_{\text{train+valid}}$	925	300	3.00	1853.00	162.00
$\text{VLSP2022}_{\text{test}}$	914	300	3.05	1762.40	153.05

validation ($\text{VLSP2022}_{\text{valid}}$) and test datasets ($\text{VLSP2022}_{\text{test}}$). The datasets contain several document clusters. Each cluster has 3-5 documents that illustrate the same topic. There are only 300 samples in training and validation sets in total ($\text{VLSP2022}_{\text{train+valid}}$). The compression ratio of the summaries provided per every split of the dataset represents 9%.

4 Experiential Setup

We experiment with $\text{LongT5}_{\text{LARGE}}$-TGlobal (2K/512) – is a case insensitive LongT5 version with Transient Global Attention mechanism with 2048/512 input/output tokens respectively, and size of the original T5_{LARGE} [21]. We refer to this model as ViLongT5 in further. Next, we provide the details of input data preparation and organization of the pre-training using Vietnamese datasets described in Section 3.

We consider NewsCorpus dataset for the ViLongT5 pretraining. Precisely speaking, we select the first 10^6 documents from the whole NewsCorpus. Due to the specifics of this dataset, which consists of raw documents only (Table 3), additional post-processing was applied toward document clustering and summary generation for the composed clusters. We perform the artificial transformations of the documents to the multi-document by interpreting every document as a *cluster* – a list of paragraphs, where every paragraph is considered as a sub-document of the original content. For the preliminary document summarization, we consider the *principle sentence generation* strategy from PEGASUS [32] by relying on the results of the extensive experiments [9]. For each document we select the five most salient sentences by `pyramid-rouge` [30] score. To emphasize a separation between documents in a cluster, we consider an auxiliary document separation token $\langle \textit{doc-sep} \rangle$. To emphasize the end of each sentence and the whole input sequence, we adopt $\langle \textit{sent-sep} \rangle$ and $\langle \textit{eos} \rangle$ auxiliary tokens respectively.

By default, the core LongT5 [9] is designed for the "Sentence Piece" based tokenization model [10]. To meet these requirements, we then compose case-

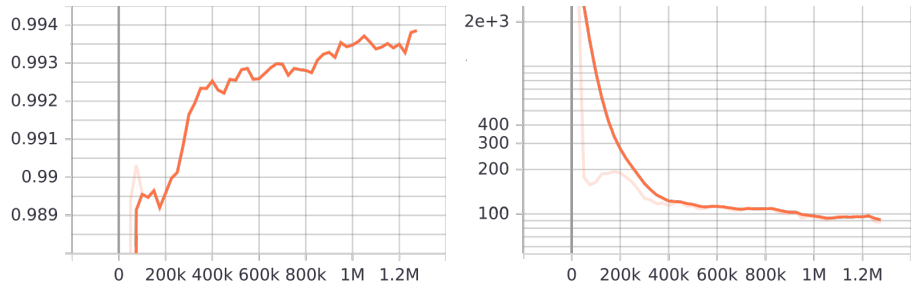

Figure 2. *Accuracy* (left) and *Loss* (right) parameter dynamics during the LongT5$_{\text{LARGE}}$-TGlobal (2K/512) pretraining stage over NewsCorpus dataset documents (details in Section 4); Y-axis corresponds to logarithmic-scaled values; X-axis represents the number of steps passed from `0` to `1.275M`, where each step involves feed-forward and back-propagation over a single batch.

insensitive Vietnamese language-oriented `SentencePiece` model[12]. To compose this model, we consider original documents from all datasets mentioned in Section 3, with NewsCorpus limited by the first 10^6 documents. Due to the specifics of the Vietnamese texts, all syllables were merged into words with the auxiliary "_" (underscore) character. We apply the stemming and lowercasing. In terms of stemming operation, all the syllables of the related word were concatenated with the underscore character. For such operation, the `VnCoreNLP` [29] library[14] was considered. The size of the result vocabulary was established as `32K` tokens.

We consider the original implementation of the LongT5 model architecture provided by `flaxformer`[15] library. For the ViLongT5 pretraining, the default configuration and hyperparameters setup was considered [21]. The whole process lasts 3.7 days and is performed on 2 × NVIDIA A100 GPUs (40GB each). For such parameters, the maximum possible batch size of the model was considered to be set as 8. The latter results in the average training speed of ≈4 samples per second.

5 Result Analysis And Discussion

The pre-training statistics of such parameters as *accuracy* and *loss* are illustrated in Figure 2. We terminate the pre-training process once it reached `1.275M` steps over NewsCorpus documents, where each step includes a feedforward and back-propagation of a whole input batch.

According to the Figure 2 (left), once ViLongT5 reached ≈`100K` pre-training steps, it illustrates a relatively high training accuracy of `0.989`, with further parameter value increment up to `0.994`. In terms of the *loss* variation, it is possible to investigate a significant decrease within the first ≈`600K` steps and reach the

[12] We adopt the native Google SentencePiece library[13]
[14] We `wseg` annotation type
[15] https://github.com/google/flaxformer

Table 4. Results of the baseline models in comparison with pretrained and fine-tuned ViLongT5; «∗» corresponds to the preliminary state finetuned with 5K steps only and excluding $\text{VLSP2022}_{\text{valid}}$ dataset; models ranked by R-2 measure results.

Model	Rank	Dataset	Rouge Scores (F1) R-1	R-2	R-L	AVG. R
ViLongT5	—	$\text{VLSP2022}_{\text{train+valid}}$	62.00	39.20	38.30	46.50
ViLongT5	—	VVV_{test}	62.90	39.60	37.20	46.50
ViLongT5	—	$\text{VVV}_{\text{valid}}$	52.90	33.20	33.30	39.80
$\text{hybrid}_{\text{the\_coach team}}$	#1	$\text{VLSP2022}_{\text{valid}}$	51.68	31.50	48.93	—
$\text{LexRank+MMR}_{\text{baseline}}$	#8	$\text{VLSP2022}_{\text{valid}}$	48.36	26.50	44.21	—
$\text{rule}_{\text{baseline}}$	#10	$\text{VLSP2022}_{\text{valid}}$	46.40	25.82	42.84	—
ViLongT5∗	#13	$\text{VLSP2022}_{\text{valid}}$	45.70	24.83	42.85	—
$\text{anchor}_{\text{baseline}}$	#19	$\text{VLSP2022}_{\text{valid}}$	43.81	19.31	39.28	—
$\text{ViT5}_{\text{abstractive-baseline}}$	#20	$\text{VLSP2022}_{\text{valid}}$	31.29	30.77	27.97	—
$\text{hybrid}_{\text{the\_coach team}}$	#1	$\text{VLSP2022}_{\text{test}}$	49.62	29.37	47.01	—
$\text{LexRank+MMR}_{\text{baseline}}$	#6	$\text{VLSP2022}_{\text{test}}$	47.72	26.25	43.39	—
$\text{rule}_{\text{baseline}}$	#7	$\text{VLSP2022}_{\text{test}}$	46.27	26.11	42.73	—
ViLongT5	#10	$\text{VLSP2022}_{\text{test}}$	45.16	24.48	42.08	—
$\text{anchor}_{\text{baseline}}$	#19	$\text{VLSP2022}_{\text{test}}$	43.21	18.86	38.69	—
$\text{ViT5}_{\text{abstractive-baseline}}$	#20	$\text{VLSP2022}_{\text{test}}$	32.26	14.97	28.95	—

flat once getting closer to `1.275M` which finally leads us to the termination of the pretraining process.

We use checkpoint of model pre-trained with `1.275M` steps to continue fine-tuning with extra `10K` steps on small Vietnamese multi-document summarization datasets, which we divide into the train, validation, and test sets with the proportion of 8:1:1. Considering the results of behavioral aspects mentioned above, we provide post-processing involving output trimming by keeping only information until the first appeared $\langle eos \rangle$ in output[16]. Table 4 illustrates the obtained results for:

1. $\text{VLSP2022}_{\text{train+valid}}$;
2. $\text{VLSP2022}_{\text{train+valid}}$+ViMs+VMDS (test/valid) or VVV in short;
3. $\text{VLSP2022}_{\text{valid}}$ and $\text{VLSP2022}_{\text{test}}$ according to the related competitions[17].

In terms of the $\text{VLSP2022}_{\text{test}}$ assessment, the proposed ViLongT5 in 13th out of 20 participants on $\text{VLSP2022}_{\text{valid}}$[18] and 10th place on $\text{VLSP2022}_{\text{test}}$. Models ranked by R2-F1 measure results. Table 4 lists the results of other baselines as well as the top submissions for comparison ($\text{hybrid}_{\text{the\_coach team}}$). First it is worth to mention that abstractive approaches with generative texts are tend to

[16] Summaries provided by ViLongT5 model might include multiple entries of $\langle eos \rangle$ token

[17] https://aihub.ml/competitions/341

[18] Preliminary version of the "ViLongT5∗" was used, for which the $\text{VLSP2022}_{\text{valid}}$ dataset has been excluded from fine-tuning

perform worse than generative in terms of the result assessment systems. Analyzing the baseline results of the purely extractive and abstractive approaches it is possible to investigate the large gap in the obtained results and importance of the originally salient sentences in the result summary, especially with long-common-sequence assessment (R-L). The hybrid approach ($\text{hybrid}_{\text{the\_coach team}}$) text summarization approach illustrates the highest result. Application of the LexRank [7] + MMR [8] correspond to extractive baseline approach ranked by #8 and #6 in $\text{VLSP2022}_{\text{valid}}$ and $\text{VLSP2022}_{\text{test}}$ respectively. The latter outperforms the results of our model by $\approx$ 5.7% (R-1), 7% (R-2), and 3% (R-L) respectively. In that sense, the application of $\text{hybrid}_{\text{the\_coach team}}$ performs better by 15% (R-1), 23% (R-2) and 12.5% (R-L). Our assumption on relatively large percentage increase of R-2 is due to the relatively low results across all the VLSP2022 models listed in Table 4. Results of the $\text{rule}_{\text{baseline}}$ correspond to the case of selecting first and last sentence for each cluster of the documents, ranked with #10 and #7 in $\text{VLSP2022}_{\text{valid}}$ and $\text{VLSP2022}_{\text{test}}$ respectively. The $\text{anchor}_{\text{baseline}}$ is a result of the input duplication, results in #19 rank. Model ViT5 has been adopted as a zero-shot abstractive baseline, and ranked #20.

6 Conclusion

The recent appearance of language models significantly addresses the long-range information memorizing is what becomes a result of the vast amount of further studies, focused on context length increment. In this work, we survey the transformers and their variations and evolution toward the internal self-attention mechanism implementations. The main highlights that overcome the main problem of self-attention with its computational complexity were shown. Considering highlights and the lack of their recent application findings for the Vietnamese language, we adopt and experiment with one of the promising models (LongT5) for the abstractive multi-document text summarization in mass-media texts. One of the largest and publicly available NewsCorpus of raw texts has been adopted for the initial pre-training. We experiment with the finetuned version and due to the pre-train specifics investigate with the summaries representing the retelling of the most salient sentences.

References

1. Ainslie, J., Ontanon, S., Alberti, C., Cvicek, V., Fisher, Z., Pham, P., Ravula, A., Sanghai, S., Wang, Q., Yang, L.: ETC: Encoding long and structured inputs in transformers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 268–284. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.19, https://aclanthology.org/2020.emnlp-main.19
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. ArXiv **abs/2004.05150** (2020)

4. Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W., Goharian, N.: A discourse-aware attention model for abstractive summarization of long documents. arXiv preprint arXiv:1804.05685 (2018)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
6. Erdős, P., Rényi, A., et al.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5**(1), 17–60 (1960)
7. Erkan, G., Radev, D.R.: LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research **22**, 457–479 (12 2004). https://doi.org/10.1613/jair.1523,
8. Goldstein, J., Carbonell, J.: Summarization: (1) using MMR for diversity- based reranking and (2) evaluating summaries. In: TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998. pp. 181–195. Association for Computational Linguistics, Baltimore, Maryland, USA (Oct 1998). https://doi.org/10.3115/1119089.1119120, https://aclanthology.org/X98-1025
9. Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.H., Yang, Y.: LongT5: Efficient text-to-text transformer for long sequences. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 724–736. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.findings-naacl.55, https://aclanthology.org/2022.findings-naacl.55
10. Kudo, T., Richardson, J.: 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. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2012, https://aclanthology.org/D18-2012
11. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: 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. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.703, https://aclanthology.org/2020.acl-main.703
12. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (apr 1958). https://doi.org/10.1147/rd.22.0159, https://doi.org/10.1147/rd.22.0159
13. Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., Xiang, B.: Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. pp. 280–290. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/K16-1028, https://aclanthology.org/K16-1028
14. Narayan, S., Cohen, S.B., Lapata, M.: Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In: Proceed-

ings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1797–1807. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1206, https://aclanthology.org/D18-1206

15. Nenkova, A., Passonneau, R.: Evaluating content selection in summarization: The pyramid method. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. pp. 145–152. Association for Computational Linguistics, Boston, Massachusetts, USA (5 2004), https://aclanthology.org/N04-1019
16. Nguyen, D.Q., Tuan Nguyen, A.: PhoBERT: Pre-trained language models for Vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1037–1042. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.92, https://aclanthology.org/2020.findings-emnlp.92
17. Nguyen, M.T., Nguyen, H.D., Nguyen, T.H.N., Nguyen, V.H.: Towards state-of-the-art baselines for vietnamese multi-document summarization. In: 2018 10th International Conference on Knowledge and Systems Engineering (KSE). pp. 85–90 (2018). https://doi.org/10.1109/KSE.2018.8573420
18. Nguyen, M.T., Nguyen, H.D., Nguyen, T.H.N., Nguyen, V.H.: Towards state-of-the-art baselines for vietnamese multi-document summarization. In: 2018 10th International Conference on Knowledge and Systems Engineering (KSE). pp. 85–90 (2018). https://doi.org/10.1109/KSE.2018.8573420
19. Phan, L., Tran, H., Nguyen, H., Trinh, T.H.: 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. pp. 136–142. Association for Computational Linguistics (2022), https://aclanthology.org/2022.naacl-srw.18
20. Phang, J., Zhao, Y., Liu, P.J.: Investigating efficiently extending transformers for long input summarization. arXiv preprint arXiv:2208.04347 (2022)
21. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
22. Sharma, E., Li, C., Wang, L.: BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2204–2213. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1212, https://aclanthology.org/P19-1212
23. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 464–468. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2074, https://aclanthology.org/N18-2074
24. To, H.Q., Nguyen, K.V., Nguyen, N.L.T., Nguyen, A.G.T.: Monolingual vs multilingual BERTology for Vietnamese extractive multi-document summarization. In: Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation. pp. 692–699. Association for Computational Lingustics, Shanghai, China (11 2021), https://aclanthology.org/2021.paclic-1.73

25. Tran, N.L., Le, D.M., Nguyen, D.Q.: Bartpho: Pre-trained sequence-to-sequence models for vietnamese. In: Proceedings of the 23rd Annual Conference of the International Speech Communication Association (2022)
26. Tran, N.T., Nghiem, M.Q., Nguyen, N.T., Nguyen, N.L.T., Van Chi, N., Dinh, D.: Vims: a high-quality vietnamese dataset for abstractive multi-document summarization. Language Resources and Evaluation **54**(4), 893—920 (2020)
27. Ung, V.G., Luong, A.V., Tran, N.T., Nghiem, M.Q.: Combination of features for vietnamese news multi-document summarization. In: 2015 Seventh International Conference on Knowledge and Systems Engineering (KSE). pp. 186–191. IEEE (2015)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
29. Vu, T., Nguyen, D.Q., Nguyen, D.Q., Dras, M., Johnson, M.: VnCoreNLP: A Vietnamese natural language processing toolkit. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 56–60. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-5012, https://aclanthology.org/N18-5012
30. Xiao, W., Beltagy, I., Carenini, G., Cohan, A.: 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). pp. 5245–5263. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.360, https://aclanthology.org/2022.acl-long.360
31. Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al.: Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems **33** (2020)
32. Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org (2020)
33. Zheng, Z., Yue, X., Huang, S., Chen, J., Birch, A.: Towards making the most of context in neural machine translation. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. IJCAI'20 (2021)