Записки научных
семинаров ПОМИ
Том 529, 2023 г.

N. Rusnachenko, The Anh Le, Ngoc Diep Nguyen

PRE-TRAINING LONGT5 FOR VIETNAMESE MASS-MEDIA MULTI-DOCUMENT SUMMARIZATION

Abstract. Multi-document summarization is a task aimed to extract the most salient information from a set of input documents. One of the main challenges in this task is the 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. Recent advances in machine translation have resulted in significant growth in the use of a related architecture known as the *Transformer*. Being pretrained on large amounts of raw texts, Transformers allow to capture a deep knowledge of the texts. In this paper, we survey the findings of language model applications for text summarization problems, including important Vietnamese text summarization models. According to the latter, we select LongT5 to pretrain and then fine-tune it for the Vietnamese multi-document text summarization problem from scratch. We analyze the resulting model and experiment with multi-document Vietnamese datasets, including ViMs, VMDS, and VLSP2022. We conclude that using a Transformer-based model pretrained on a large amount of unlabeled Vietnamese texts allows us to achieve promising results, with further enhancement via fine-tuning within a small amount of manually summarized texts. The pretrained model utilized in the experiment section has been made available online at `https://github.com/nicolay-r/ViLongT5`.

§1. Introduction

At present, drastic growth of news and event recordings has become one of the main reasons why most mass media platforms have become saturated with mass media information. This factor becomes crucial for manual daily news reading, making it virtually infeasible. The *text summarization* task [12] aims to create a short version of the original texts by keeping the most concise, coherent, and salient information. Shortening long documents by keeping the most meaningful information represents a rather difficult task for manual execution, involving deep text analysis

Key words and phrases: vietnamese multi-document summarization and text summarization and Transformers and language models.

and content understanding. These factors necessitate deep research in automatic text summarization approaches and systems built upon them. In terms of the problem setting, such systems might be categorized as *extractive* or *abstractive*. Summarization systems of the extractive type [7] aim to rank sentences in the given text by relying on their meaning and importance, with further extraction of high-ranking sentences. In turn, abstractive summarization systems are focused on generating the result in an essay format for a given text [9, 19, 21].

When attention mechanisms that addressed the problem of capturing distant information in long input sequences were introduced in the machine translation (MT) task [2], they caused a significant impact on further studies and attention implementations [28, 33]. An attention mechanism is a module in the neural network which aims to assess the importance of given information by assigning *weights* to its components. A significant amount of research studies have been devoted to experiments with attention implementations as well as integration of such modules into target-oriented machine learning models aside from machine translation, including the text summarization domain.

Further appearance of the *self-attention* mechanism [28] as an internal component of the encoder-decoder architecture resulted in the *Transformer* architecture. Transformer-based models caused a significant breakthrough in MT, resulting in further modifications [33]. The transition towards texts of a single language for Transformers resulted in the introduction of *language models* that have recently become both popular and standardized solutions for other natural language processing (NLP) domains including text summarization [9, 15, 32]. In this work, we focus on the analysis of recent advances of language models to choose the most promising solution for the Vietnamese multi-document summarization problem [17, 27] of mass media documents. It is worth noting that multi-document summarization faces the problem of long contents where the importance of information might be divided unequally across the documents. 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 remainder of the paper is organized as follows. Section 2 provides an overview of recent advances in Transformer-based models along with their architectural updates and training techniques, with Vietnamese-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 pretraining and fine-tuning. Detailed descriptions of the model pretraining process as well as further experiments are covered in Sections 4 and 5 respectively, including a comparison with other extractive and abstractive text summarization baselines for the $\text{VLSP2022}_{\text{valid}}$ and $\text{VLSP2022}_{\text{test}}$ datasets.

§2. Background

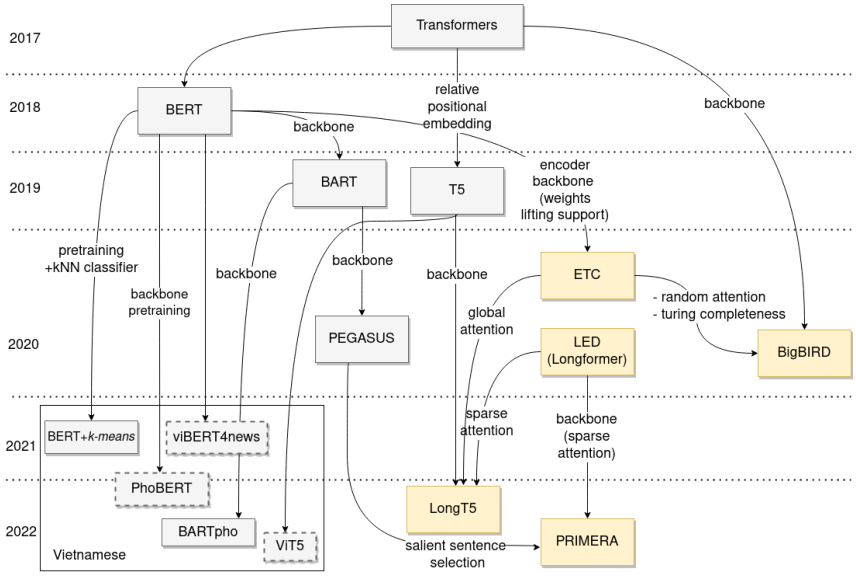

Figure 1. Tree diagram of Transformer-based models [28], placed in order of their appearance from top to bottom; *arrows* illustrate the most significant findings found in successor models; *blocks* illustrate models with: original self-attention (gray), sparse self-attention mechanism (yellow); highlighted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

Since the text summarization problem is commonly treated as either extractive or abstractive task, both encoder and decoder components of the Transformer could be used as *backbones*. We first consider the BERT

architecture as a backbone; it finds applications in extractive-based text summarization problems. Due to architectural specifics, where information is encoded bidirectionally, BERT could not be easily adopted for the generative task format [5]. As for the extractive task format, we may consider BERT as a sentence encoder, complementing it with a clustering type algorithm.

However, to address generative limitations, in [11] the authors proposed the BART framework, which represents a BERT (bidirectional Transformer) complemented by an autoregressive decoder (GPT). BART proposed a denoising sequence-to-sequence framework, where the pretraining 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 modifications of pretraining schemes. Being particularly effective for text generation tasks, including text summarization, at the time of the model announcement the authors mentioned 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 of frameworks for text summarization, including the following. In [32], the authors proposed the PEGASUS framework, where the sentence-based masking strategy was based on the invented *salient sentence selection* algorithm. With the latter, the authors proposed a sentence assessment metric with a limited selection of the top k-scored sentences. According to extensive experiments on XSum and CNN/DailyMail [13] collections with LARGE-sized models (Table 2), the authors showed that the resulting PEGASUS model [20] outperformed other Transformer-based solutions such as BART or T5 [21]. Text-To-Text Transfer Transformer (T5) [21] is based on the original transformer [28] complemented by several modifications in 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, i.e., the atomic parts of a sentence are *syllables* rather than words. To the best of our knowledge, recent advances in Vietnamese text processing for multi-document summarization problems are limited by applications of

the original self-attention-based Transformers. Figure 1 illustrates recent advances in Transformer-based models for Vietnamese (bordered bottom left corner). In this section, we survey recent advances in *extractive* and *abstractive* text summarization approaches.

For extractive text summarization, several studies of non-Transformer-based approaches [18] have addressed such training techniques as *distant supervision* and *supervised learning*. The adaptation of BERT towards downstream tasks for texts in Vietnamese have resulted in the appearance of PhoBERT [16] and viBERT4news[1]. In [24], the authors combine Vietnamese-oriented BERT-based pretrained states and *k-means*, and the resulting BERT+*k-means* showed top results on the VMDS[2] dataset compared with previous methods. Results of related models are illustrated in Table 1.

In the case of abstractive summarization, BARTpho [25] represents an initial study with BART-based [11] architecture applied to 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, the 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 recent advances in *abstractive text summarization* and *named entity recognition* (NER) [19].

2.2. Sparse Self-Attention. The main task solved by attention is the how a particular token is related to other tokens mentioned in the text. However, an important drawback 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]. Apart from BERT [5], such above-mentioned models as BART [11] and T5 [21] use nested self-attention mechanisms, and hence in practice input sequences tend to be limited by 512 tokens [31].

To address the shortcomings of self-attention being applied to longer input sequences, several independent works have proposed its *sparse* variations [1,3,31]. To manage attention behavior in a sparse way, the work [1] proposes the Extended Transformer Construction (ETC). In parallel with other works, the authors invented *relative token positioning* [3, 31] as a

[1]`https://huggingface.co/NlpHUST/vibert4news-base-cased`

[2]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.

	ViMs		VMDS	
Model	R-1	R-2	R-1	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**

preliminary step for attention sparsification. To distribute attention between distant tokens, the authors also introduce a *global-local* attention mechanism by expanding the original (local) input with *global tokens* under the following constraint: 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, the authors split attention calculation into parts and prove that 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 together with sparse attention [31] allows to train ETC with longer input sequences and hence causes a significant impact on the result performance in, for example, question answering (QA) [1]. In [3], the authors proposed Longformer and the related encoder-decoder architecture (LED), which represents a modification of the original Transformer with *windowed attention* variations. Similar to the implementation in ETC, windowed attention meahs that for a particular token we only consider r (radius parameter) left and right neighboring tokens as potential subjects

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]	Bidir. encoder + autoregr. 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 norm 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

for attention. Figure 1 illustrates models with sparse attention mechanisms (in yellow). The authors experiment with LARGE sized models from 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` to `16K` tokens. The LED architecture [3] caused a significant effect on models that appeared afterwards, including PRIMERA [30] with its salient sentences masking approach and LongT5 [9] considered further in this section.

Alongside the findings of ETC, in [31] the authors treat the computational problem of constructing a sparse self-attention mechanism as *graph sparsification*. Complementing sliding window and global attention mechanisms [1] with the Erdös-Rényi model [6] of independently choosing an edge with fixed probability, the 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. In particular, they show that the sparser the graph, the more layers are required to reach completeness. For text summarization, the authors experimented with sparse attention at the encoder side[3], using pretrained schemes from PEGASUS [32] for LARGE-sized models. The resulting model is known as BigBird-PEGASUS [31].

LongT5 represents a modified version of the T5 model [21] which adopts the sparse attention mechanism variations proposed in the ETC model, including windowed attention and the global-local variation known as TGlobal [9]. The latter introduces local sparsity in the attention mechanism, which allows to reduce the quadratic cost when scaling to long inputs. Unlike T5, the modified LongT5 can handle longer input sequences before failing with out-of-memory exceptions. We also note that LongT5 (4K input) achieves top results across a variety of text generative models on almost every text summarization dataset: arXiv summarization dataset, PubMed, BigPatent [22], and MediaSum [9]. As for the PRIMERA model (447M), it shows the best results in MultiNews compared to 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, we see that LongT5 has the best performance compared to other models discussed above. The cost of the LongT5 architectural traits lies in its number of 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

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

these datasets are described below, with their brief statistics described in Table 3.

NewsCorpus[4] 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[5] is a multi-document dataset collected from a Vietnamese online news provider `baomoi.com`. This dataset contains 628 documents categorized into 200 topics.

ViMs[6] 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[7] is a dataset is provided in a competition hosted by the Association for Vietnamese Language and Speech Processing. The provided data consists of Vietnamese news on various topics, including the economy, society, culture, science, and technology. Every document includes the title, anchor text and body text of individual documents, summary, and a category tag. It is divided into train ($\text{VLSP2022}_{\text{train}}$), 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 the training and validation sets in total ($\text{VLSP2022}_{\text{train+valid}}$). The compression ratio of the summaries provided per every split of the dataset amounts to 9%.

§4. Experimental Setup

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

We consider NewsCorpus dataset for the ViLongT5 pretraining. To be precise, we select the first 10^6 documents from the entire NewsCorpus.

[4] `https://github.com/binhvq/news-corpus`
[5] `https://github.com/lupanh/VietnameseMDS`
[6] `https://github.com/CLC-HCMUS/ViMs-Dataset`
[7] `https://vlsp.org.vn/vlsp2022/eval/abmusu`

Table 3. Statistics of Vietnamese datasets utilized for 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

Due to the specifics of this dataset, which consists of raw documents only (Table 3), additional postprocessing was applied towards document clustering and summary generation for the composed clusters. We perform artificial transformations of the documents into multi-document data by interpreting every document as a *cluster*, i.e., a list of paragraphs where every paragraph is considered as a subdocument of the original document. For preliminary document summarization, we consider the *principle sentence generation* strategy from PEGASUS [32] by relying on the results of previous extensive experiments [9]. For each document, we select the five most salient sentences by their `pyramid-rouge` [30] score. To emphasize the separation between documents in a cluster, we consider an auxiliary document separation token $\langle doc - sep \rangle$. To emphasize the end of every sentence and the entire input sequence, we adopt $\langle sent - sep \rangle$ and $\langle eos \rangle$ 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 a case-insensitive Vietnamese-oriented `SentencePiece` model[8]. To prepare 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 Vietnamese texts, all syllables were merged into words with an auxiliary "_" (underscore) character. We apply stemming and lowercasing. In terms of the stemming operation, all syllables of a

[8]We adopt the native Google SentencePiece library: `https://github.com/google/sentencepiece`.

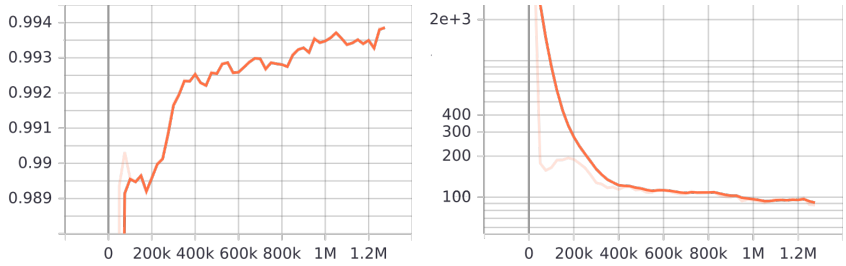

Figure 2. *Accuracy* (left) and *Loss* (right) parameter dynamics during the $\text{LongT5}_{\text{LARGE}}$-TGlobal (2K/512) pretraining stage over NewsCorpus dataset documents (details in Section 4); Y-axis corresponds to logarithmic-scaled values; X-axis represents the number of steps passed from `0`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

word are concatenated with the underscore character. For this operation, we used the `VnCoreNLP` [29] library[9]. The size of the resulting vocabulary was established at `32K` tokens.

We consider the original implementation of the LongT5 model architecture provided by the `flaxformer`[10] library. For ViLongT5 pretraining, the default configuration and hyperparameters setup were used as shown in [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 set to be 8. The latter results in the average training speed of ≈4 samples per second.

§5. Analysis and Discussion of the Results

The pretraining statistics of such parameters as *accuracy* and *loss* are illustrated in Figure 2. We terminate the pretraining process once it has reached `1.275M` steps over NewsCorpus documents, where each step includes forward propagation and backpropagation over a single input batch.

According to Figure 2 (left), once ViLongT5 has reached ≈`100K` pretraining steps it has a relatively high training accuracy of `0.989`, with

[9]We used the `wseg` annotation type.

[10]`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	—

further accuracy values increasing up to `0.994`. In terms of the *loss* parameter, we have found a significant decrease within the first ≈`600K` steps, flattening out once we get closer to `1.275M` which finally leads us to the termination of the pretraining process.

We use the checkpoint of the model pretrained with `1.275M` steps to continue fine-tuning with an additional `10K` steps on small Vietnamese multi-document summarization datasets, which we divide into the train, validation, and test sets in the proportion of 8:1:1. Considering the results of behavioral aspects mentioned above, we provide postprocessing involving output trimming by keeping only information until the first $\langle eos \rangle$ appearing in the output[11]. 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;

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

(3) $\text{VLSP2022}_{\text{valid}}$ and $\text{VLSP2022}_{\text{test}}$ according to the corresponding competitions[12].

In terms of the $\text{VLSP2022}_{\text{test}}$ assessment, the proposed ViLongT5 model placed 13th out of 20 participants on $\text{VLSP2022}_{\text{valid}}$[13] and 10th place on $\text{VLSP2022}_{\text{test}}$. Models were ranked by the 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, we note that abstractive approaches with generative texts tend to perform worse than generative in terms of the scores assigned by result assessment systems. Analyzing the baseline results of purely extractive and abstractive approaches, we can see a large gap in the obtained results and importance of the salient sentences originally appearing in the result summary, especially with long-common-sequence assessment (R-L). The hybrid approach ($\text{hybrid}_{\text{the\_coach team}}$) to text summarization leads to the highest results. Application of the LexRank [7] + MMR [8] corresponds to an extractive baseline approach ranked at #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 a relatively large increase in the results in terms of R-2 is due to the relatively low results across all VLSP2022 models listed in Table 4. Results of the $\text{rule}_{\text{baseline}}$ correspond to the case of selecting the first and last sentence for every cluster of documents, and it ranks at #10 and #7 in $\text{VLSP2022}_{\text{valid}}$ and $\text{VLSP2022}_{\text{test}}$ respectively. The $\text{anchor}_{\text{baseline}}$ denotes simple input duplication, and it ranks at #19. The ViT5 model has been adopted as a zero-shot abstractive baseline, and ranked #20.

§6. Conclusion

The recent introduction of large language models significantly alleviates the problem of long-range information memorization, especially as a result of numerous further studies focused on increasing their context lengths. In this work, we survey the recently proposed Transformers and their variations and evolution in the internal self-attention mechanisms. We have shown the main highlights that overcome the primary problem of self-attention with its computational complexity. Considering the highlights

[12] `https://aihub.ml/competitions/341`

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

and the lack of their recent applications for the Vietnamese language, we adopt and experiment with one of the most promising models (LongT5) for 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 pretraining. We experiment with the fine-tuned version and, due to the pretraining specifics, investigate the summaries produced by combining the most salient sentences.

References

1. J. Ainslie, S. Ontanon, C. Alberti, V. Cvicek, Z. Fisher, P. Pham, A. Ravula, S. Sanghai, Q. Wang, L. Yang, *ETC: Encoding long and structured inputs in transformers.* — Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Online), Association for Computational Linguistics, November 2020, pp. 268–284.
2. D. Bahdanau, K. Cho, Y. Bengio, *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473 (2014).
3. I. Beltagy, M. E. Peters, A. Cohan, *Longformer: The long-document transformer*, ArXiv **abs/2004.05150** (2020).
4. A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian, *A discourse-aware attention model for abstractive summarization of long documents*, arXiv preprint arXiv:1804.05685 (2018).
5. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding.* — 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) (Minneapolis, Minnesota), Association for Computational Linguistics, June 2019, pp. 4171–4186.
6. P. Erdős, A. Rényi, et al., *On the evolution of random graphs.* — Publ. Math. Inst. Hung. Acad. Sci **5**, No. 1 (1960), 17–60.
7. G. Erkan, D. R. Radev, *LexRank: Graph-based lexical centrality as salience in text summarization.* — J. Artificial Intelligence Research **22** (2004), 457–479.
8. J. Goldstein, J. Carbonell, *Summarization: (1) using MMR for diversity- based reranking and (2) evaluating summaries*, TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998 (Baltimore, Maryland, USA), Association for Computational Linguistics, October 1998, pp. 181–195.
9. M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, Y. Yang, *LongT5: Efficient text-to-text transformer for long sequences*, Findings of the Association for Computational Linguistics: NAACL 2022 (Seattle, United States), Association for Computational Linguistics, July 2022, pp. 724–736.

10. T. Kudo, J. Richardson, *SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (Brussels, Belgium), Association for Computational Linguistics, November 2018, pp. 66–71.
11. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online), Association for Computational Linguistics, July 2020, pp. 7871–7880.
12. H. P. Luhn, *The automatic creation of literature abstracts.* — IBM J. Res. Dev. **2**, No. 2 (1958), 159–165.
13. R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, B. Xiang, *Abstractive text summarization using sequence-to-sequence RNNs and beyond*, Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (Berlin, Germany), Association for Computational Linguistics, August 2016, pp. 280–290.
14. S. Narayan, S. B. Cohen, M. Lapata, *Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Brussels, Belgium), Association for Computational Linguistics, October-November 2018, pp. 1797–1807.
15. A. Nenkova, R. Passonneau, *Evaluating content selection in summarization: The pyramid method*, Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004 (Boston, Massachusetts, USA), Association for Computational Linguistics, 5 2004, pp. 145–152.
16. D. Q. Nguyen, A. T. Nguyen, *PhoBERT: Pre-trained language models for Vietnamese*, Findings of the Association for Computational Linguistics: EMNLP 2020 (Online), Association for Computational Linguistics, November 2020, pp. 1037–1042.
17. M.-T. Nguyen, H.-D. Nguyen, T.-H.-N. Nguyen, V.-H. Nguyen, *Towards state-of-the-art baselines for vietnamese multi-document summarization*, 2018 10th International Conference on Knowledge and Systems Engineering (KSE), 2018, pp. 85–90.
18. M.-T. Nguyen, H.-D. Nguyen, T.-H.-N. Nguyen, V.-H. Nguyen, *Towards state-of-the-art baselines for vietnamese multi-document summarization*, 2018 10th International Conference on Knowledge and Systems Engineering (KSE), 2018, pp. 85–90.
19. L. Phan, H. Tran, H. Nguyen, T. H. Trinh, *ViT5: Pretrained text-to-text transformer for Vietnamese language generation*, 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, 2022, pp. 136–142.
20. J. Phang, Y. Zhao, P. J. Liu, *Investigating efficiently extending transformers for long input summarization*, arXiv preprint arXiv:2208.04347 (2022).

21. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, *Exploring the limits of transfer learning with a unified text-to-text transformer.* — J. Machine Learning Research **21**, No. 140 (2020), 1–67.
22. E. Sharma, C. Li, L. Wang, *BIGPATENT: A large-scale dataset for abstractive and coherent summarization*, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Florence, Italy), Association for Computational Linguistics, July 2019, pp. 2204–2213.
23. P. Shaw, J. Uszkoreit, A. Vaswani, *Self-attention with relative position representations*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2 (Short Papers) (New Orleans, Louisiana), Association for Computational Linguistics, June 2018, pp. 464–468.
24. H. Q. To, K. Van Nguyen, N. L.-T. Nguyen, A. G.-T. Nguyen, *Monolingual vs multilingual BERTology for Vietnamese extractive multi-document summarization*, Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (Shanghai, China), Association for Computational Lingustics, 11 2021, pp. 692–699.
25. N. L. Tran, D. M. Le, D. Q. Nguyen, *Bartpho: Pre-trained sequence-to-sequence models for vietnamese*, Proceedings of the 23rd Annual Conference of the International Speech Communication Association, 2022.
26. N. T. Tran, M. Q. Nghiem, N. TH Nguyen, N. L. T. Nguyen, N. V. Chi, D. Dinh, *Vims: a high-quality vietnamese dataset for abstractive multi-document summarization.* — Language Resources and Evaluation **54**, No. 4 (2020), 893–920.
27. V.-G. Ung, A.-V. Luong, N.-T. Tran, M.-Q. Nghiem, *Combination of features for vietnamese news multi-document summarization*, 2015 Seventh International Conference on Knowledge and Systems Engineering (KSE), IEEE, 2015, pp. 186–191.
28. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, Proceedings of the 31st International Conference on Neural Information Processing Systems (Red Hook, NY, USA), NIPS'17, Curran Associates Inc., 2017, p. 6000–6010.
29. T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, M. Johnson, *VnCoreNLP: A Vietnamese natural language processing toolkit*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (New Orleans, Louisiana), Association for Computational Linguistics, June 2018, pp. 56–60.
30. W. Xiao, I. Beltagy, G. Carenini, A. Cohan, *PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization*, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Dublin, Ireland), Association for Computational Linguistics, May 2022, pp. 5245–5263.
31. M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., *Big bird: Transformers for longer sequences*, Advances in Neural Information Processing Systems **33** (2020).
32. J. Zhang, Y. Zhao, M. Saleh, P. J. Liu, *Pegasus: Pre-training with extracted gap-sentences for abstractive summarization*, Proceedings of the 37th International Conference on Machine Learning, ICML'20, JMLR.org, 2020.

33. Z. Zheng, X. Yue, S. Huang, J. Chen, A. Birch, *Towards making the most of context in neural machine translation*, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20, 2021.

Bauman Moscow State
Technical University
E-mail: rusnicolay@gmail.com

FPT University, Can Tho, Viet Nam;
CyberIntellect, Moscow, Russia
E-mail: anhlt161@fe.edu.vn

CyberIntellect, Moscow, Russia
E-mail: diepnn83@gmail.com

Поступило 6 сентября 2023 г.