ICIC Express Letters
Part B: Applications
Volume **14**, Number **11**, November **2023**

ICIC International ©2023 ISSN 2185-2766
pp. 1143–1151

EXPLORING ABSTRACTIVE SUMMARIZATION METHODS FOR VIETNAMESE USING PRE-TRAINED BERT MODELS

Duy Hieu Nguyen[1,2], Trong Nghia Hoang[3], Dien Dinh[1,2]
and Long Hong Buu Nguyen[1,2,*]

[1]Faculty of Information Technology
University of Science, Ho Chi Minh City
227 Nguyen Van Cu Street, Ward 4, District 5, Ho Chi Minh City 700000, Vietnam
19C11021@student.hcmus.edu.vn; ddien@fit.hcmus.edu.vn
[*]Corresponding author: nhblong@fit.hcmus.edu.vn

[2]Vietnam National University, Ho Chi Minh City
Linh Trung Ward, Thu Duc District, Ho Chi Minh City 700000, Vietnam

[3]School of Electrical Engineering and Computer Science
Washington State University
PO Box 642752 Pullman, WA 99164-2752, USA
trongnghia.hoang@wsu.edu

Received February 2023; accepted May 2023

Abstract. *Abstractive text summarization is a challenging task in NLP. With the innovation of the Transformer architecture in recent years, many NLP problems have been solved and achieved SOTA performance, including abstractive summarization tasks. Although there are many studies on abstractive summarization for English, there is little study for Vietnamese. In this paper, we propose a new training method for the Transformer Encoder-Decoder models using pre-trained PhoBERT and mBERT as Encoder to address the abstractive summarization task for Vietnamese. We investigated our models on two Vietnamese abstractive summarization datasets and evaluated the generated summaries using ROUGE metrics, and our methods outperformed the three strong baselines across all metrics on the Wikilingua dataset.*

Keywords: Vietnamese, Abstractive summarization, Transformer, Pre-trained BERT

1. Introduction.

Text summarization is a Natural Language Processing (NLP) task that produces a shorter version of the long input document but still retains the main idea from the input text [1]. While extractive summarization methods copy exactly most salience sentences from the input document, abstractive summarization methods use sequence-to-sequence models to generate the summarized text abstractly, similar to how humans read and summarize a document [2]. Although there are many studies in English [1-6], there is just limited study in Vietnamese text summarization and very few studies in Vietnamese abstractive summarization [7-11]. Most of them adopt the Transformer architecture and pre-trained BERT models to address these problems.

Transformer architecture was first introduced by Vaswani et al. in 2017 to replace the traditional architecture Recurrent Neural Network (RNN) in NLP. The Transformer model is comprised of 2 main components: Encoder and Decoder. The Encoder consists of several Transformer layers, including multi-head attention and a feed-forward network. The Decoder is almost the same as Encoder except for the addition of masked multi-head attention before the multi-head attention [12].

Devlin et al. adopted Transformer architecture to create BERT (Bidirectional Encoder Representations from Transformers), a language model which was pre-trained on a large

DOI: 10.24507/icicelb.14.11.1143

corpus to have better language understanding by learning the word representation from both directions. With that meaningful language representations, BERT can be finetuned in many downstream tasks to achieve state-of-the-art (SOTA) results [13].

Transformer-based Encoder-Decoder models using pre-trained BERT models have been used for abstractive summarization tasks in English and Vietnamese recently, and achieved good performance compared with other methods. Liu and Lapata employed pre-trained BERT for their abstractive model (BERTsum) to do abstractive summarization in English [3]. Nguyen et al. employed mBERT [13], PhoBERT [14], and ViBERT [15] in their Encoder-Decoder models (VieSum) to finetune Vietnamese datasets for Vietnamese abstractive summarization, and their work achieved good results in ROUGE metrics [8].

In this paper, we present another method to train Transformer Encoder-Decoder models that employ pre-trained BERT models as Encoder to address the abstractive summarization task for Vietnamese. The results show that our models outperformed the three strong baseline Encoder-Decoder methods from VieSum across all metrics on the Wikilingua [16] dataset, while still having a small under-gap compared with the baseline methods on the VietNews [17] dataset. The main contributions of this work[1] are as the following.

- We have proposed a novel approach for finetuning the Transformer Encoder-Decoder model using PhoBERT. The result shows that RDRsegmenter is best for the Wikilingual dataset, while UITws is best for the Vietnews dataset in data preprocessing.
- We also presented the Window technique to solve the problem of losing information of the sequence after token 256th due to the max input limit of the PhoBERT model.
- Our methods achieved the best ROUGE scores compared with all three strong baselines from VieSum on the Wikilingual dataset.

This paper is organized as follows. Section 2 outlines the related work. Section 3 introduces the pre-trained BERT models that we employed in our methods. Section 4 provides overviews about the datasets, evaluation metrics, and the baseline models that we used for benchmarking our models. Section 5 describes our methods to address the abstractive summarization problem for Vietnamese. Section 6 then provides the results of our experiments compare with the strong baselines. Finally, Section 7 presents our conclusions.

2. **Related Work.** Using pre-trained BERT models for Transformer Encoder-Decoder architecture was first proposed by Liu and Lapata in [3]. In that work, a pre-trained BERT language model for English was employed as the Encoder of the customized Transformer-based Encoder-Decoder framework (BERTsum) to address both extractive and abstractive summarization tasks in English. Following [3], Nguyen et al. investigated the abstractive summarization task for Vietnamese single documents (VieSum) using the Encoder-Decoder model provided by huggingface framework [8]. Their work achieved good ROUGE scores when using mBERT, PhoBERT, and ViBERT as both Encoder and Decoder of those models [8]. Inspired by that work, we built another Transformer Encoder-Decoder model that adopted mBERT and PhoBERT as Encoder and finetuned it on the same Vietnamese benchmark datasets. Our models are mostly similar to the Transformer Encoder-Decoder models from VieSum but have 2 main differences: i) we build the model from scratch instead of using the end-to-end huggingface framework, and ii) we replace only Encoder with the pre-trained BERT model, and we do not replace the Decoder like VieSum's methods[2].

[1] We will release all preprocessed datasets and code to reproduce any step in this work to motivate further research: https://github.com/ithieund/BERTSumVN.

[2] Coding from scratch helps us modify the structures easily but the disadvantage is that we cannot replace the Decoder with a pre-trained BERT model as we cannot add cross-attention to that model.

3. **Pre-Trained BERT Models.**

This section provides background information on the pre-trained BERT models that we employed as Encoder for our Transformer-based models PhoBERT2TRANS and mBERT2TRANS.

3.1. **BERT.**

BERT stands for Bidirectional Encoder Representations from Transformers, a pre-trained language model based on Transformer architecture. BERT pre-training techniques are including Masked Language Modeling (MLM) and Next-Sentence Prediction (NSP) [13]. Our experiments employed mBERT and PhoBERT, which both support Vietnamese.

3.2. **mBERT.**

mBERT is the abbreviated name of BERT-base Multilingual Case which was released by Devlin et al. [13]. In the latest version, mBERT was trained on a multilingual dataset of over 104 top languages according to Wikipedia ranking[3]. mBERT supports 512 input tokens natively, so it is also suitable for text summarization tasks.

3.3. **PhoBERT.**

PhoBERT is the first public large-scale monolingual language model pre-trained for Vietnamese released by Nguyen et al. [14]. PhoBERT was pre-trained on a monolingual Vietnamese dataset containing 20GB of text[4]. The data was preprocessed with word and sentence segmentation using RDRsegmenter before training. The authors released 2 PhoBERT models, which differ in size: PhoBERT-base (135M parameters) and PhoBERT-large (369M parameters). This pre-trained language model outperformed all previous monolingual and multilingual methods and achieved SOTA performances on four downstream Vietnamese NLP tasks. The only problem with this model is that it supports only 256 input tokens, which are relatively small for text summarization.

4. **Datasets, Metricts, and Baselines.**

In this section, we describe the datasets, metrics we use for automatic evaluation, and the strong baseline models for benchmarking.

4.1. **Datasets.**

Following [8], we investigated our abstractive models on two benchmark datasets. Table 1 shows the statistics of the raw datasets.

Table 1. Statistics factors of the two datasets

	Wikilingua			VietNews		
	Train	Val	Test	Train	Val	Test
#samples	13707	1957	3917	105418	22642	22644
#avg of words in body	519	541	519	548	549	549
#avg of words in abstract	44	45	44	38	38	38

Note: text is desegmented before counting.

4.1.1. *Wikilingua.*

Wikilingua is a large-scale benchmark dataset for Cross-Lingual Abstractive Summarization [16]. All the data were collected from wikihow.com, a website that provides instructions for any problems and subjects. Each post on that website includes a title and guiding steps to resolve a specific problem. The extracted step titles were combined to form the target summary, while the remaining texts from each step were combined to form the article body. This procedure produces article-summary pairs for all samples, which are suitable for abstractive summarization tasks. We extracted only the Vietnamese subset of this dataset to finetune our models.

[3] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[4] 19GB from the Vietnamese News corpus and 1GB from the Vietnamese Wikipedia corpus

4.1.2. *VietNews.* VietNews is the first large-scale benchmark dataset for Vietnamese single document summarization tasks [17]. The data was collected from three well-known online newspapers[5] where each article has a title, abstract, and body. All articles related to questionnaires, admissions, analytical comments, and weather forecasts were filtered out to get the final dataset with only news articles. Finally, they used NLTK and vitk tools[6] to apply sentence and word segmentation for all samples in the dataset.

4.2. **Metrics.** In this study, we used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics [18], a benchmark score for evaluating text summarization and machine translation tasks in NLP, to benchmark our methods. It measures the number of overlapping units between the generated summary and the reference summary. As many researchers do, we used ROUGE-1, ROUGE-2, and ROUGE-L for our experiments.

- ROUGE-1 and ROUGE-2: measures the overlapping unigrams and bigrams. These scores reflect how relevant the generated tokens are to the tokens in the gold label.
- ROUGE-L: measures the Longest Common Subsequences (LCS) between the two sequences. A higher score indicates greater similarity between the two sequences.

4.3. **Baselines.** To compare the output results, we borrowed the reported ROUGE scores of the three strong baselines models ViBERT2ViBERT, PhoBERT2PhoBERT, and mBERT in [8]. It is worth noting that there are two other models mBART and mT5 in that paper that achieved better ROUGE scores but those models are end-to-end Transformers, not using any pre-trained BERT model like our study. Therefore, we only compare our methods to the three strong baselines above to find out which finetune method is better.

5. **Our Methods.** Our finetuning procedure is as follows. First, we excecute data preprocessing for two datasets. Second, we conduct six training experiments with our baseline PhoBERT-based model on data segmented by current top 3 best word-segmenters RDRsegmenter [19], UETsegmenters [20], and UITws [21] to find which one is better. Finally, we finetune other PhoBERT-based versions with the data preprocessed by the best word segmenter on each dataset, and then train the mBERT-based model on two datasets to compare the performance of multilingual versus monolingual language models.

5.1. **Data preprocessing.** It was pointed out in [22] that there are a large number of duplicated samples in VietNews dataset. Therefore, we followed the practice in [22] to filter all the duplicated and overlapping samples from the raw dataset[7], and then preprocessed the cleaned data in different ways for each pre-trained BERT-based model.

The input data for mBERT is case-sensitive and does not require word segmentation. Therefore, for the VietNews dataset, we needed to convert the word-segmented text back into unsegmented text by replacing the underscore with a space.

In contrast, PhoBERT requires the input text to be word-segmented, as it was pretrained on data preprocessed with RDRsegmenter. The viWikiHow dataset is not segmented, while the VietNews dataset was pre-segmented using vitk tool by the authors. Consequently, the preprocessing procedure for each dataset is different.

- For viWikiHow dataset, we employed RDRsegmenter, UETsegmenter, and UITws to create three word-segmented versions.
- For VietNews dataset, we first applied word-desegmentation and then used RDRsegmenter, UETsegmenter, and UITws to creating three word-segmented versions.

[5] tuoitre.vn, vnexpress.net, and nguoiduatin.vn

[6] https://www.nltk.org and https://github.com/phuonglh/vn.vitk

[7] We remove all duplicates in each *train*, *val*, and *test* sets first. After that, we filter the overlapping samples between *train*, *val*, and *test* sets to get the final dataset without duplication.

The preprocessed datasets are named viWikiHow-Abs-Sum and VietNews-Abs-Sum[8] to avoid confusion with the original datasets. Table 2 shows the statistics of those datasets.

Table 2. Statistic factor of the two processed datasets

	viWikiHow-Abs-Sum			VietNews-Abs-Sum		
	Train	Val	Test	Train	Val	Test
#samples	13707	1957	3917	99134	22184	22498
#avg of words in body	519	541	519	543	548	548
#avg of words in abstract	44	45	44	37	38	38

Note: text is desegmented before counting.

5.2. **Finding the best word segmenter.** PhoBERT requires data to be word-segmented before finetuning, but what word-segmentation method is better is the big question. Therefore, before training PhoBERT-based models with different settings, we have conducted 3 training experiments on each dataset with our baseline model PhoBERT2TRANS. PhoBERT2TRANS is a Transformer Encoder-Decoder model where Encoder is replaced by a pre-trained model PhoBERT-base, while Decoder is a vanilla Transformer decoder with 8 layers, attention dimension is 64 and other parameters (vocab size, hidden size) are the same as the encoder. For each dataset, we trained the model on the *train* set and then evaluated it on the *val* set to get the *eval loss* at the end of each epoch.

For optimization, we employed AdamW with constant learning rate = 5e-5 and cross-entropy loss with label smoothing = 0.1. We finetuned the model with a batch size of 32 in max 100 epochs. The gradient was accumulated in 2 steps for viWikiHow-Abs-Sum dataset and 10 steps for VietNews-Abs-Sum dataset[9] before updating the model weights through backpropagation. To avoid overfitting, we applied early stopping [23] during training phase with delta = 0 and patience = 5, which means if the model has no improvement in the *eval loss* for 5 consecutive checkpoints, training will be terminated.

During the prediction phase on *test* set, we employed Beam Search with beam size = 3. To penalize finished hypotheses that do not have the expected sequence length, we implemented the Min Length Penalty, following the methodology outlined in [3]. We also employed Length Normalization [24] with alpha = 0.6 to normalize the score between the long and short hypotheses. In this experiment, we set the expected output length to 20. The results in Table 3 show that RDRsegmenter is best on viWikiHow-Abs-Sum, while UITws is better on VietNews-Abs-Sum dataset when finetuning PhoBERT-based model and decoding with expected min output length = 20.

Table 3. ROUGE scores of our baseline model PhoBERT2TRANS after finetuning with data preprocessed by different word segmenters

	viWikiHow-Abs-Sum			VietNews-Abs-Sum		
	R1	R2	RL	R1	R2	RL
PhoBERT2TRANS + RDRsegmenter	**49.86**	**21.07**	**32.95**	<u>56.82</u>	<u>25.75</u>	**36.84**
PhoBERT2TRANS + UETsegmenter	49.48	20.91	32.85	56.8	25.6	36.77
PhoBERT2TRANS + UITws	<u>49.73</u>	<u>20.94</u>	<u>32.9</u>	**56.89**	**25.8**	<u>36.83</u>

Note: R1, R2, and RL denote ROUGE-1, ROUGE-2, and ROUGE-L scores.

The best scores are bolded numbers and the second-best scores are underlined.

[8] We release the preprocessed datasets at https://huggingface.co/datasets/ithieund/viWikiHow-Abs-Sum and https://huggingface.co/datasets/ithieund/VietNews-Abs-Sum.

[9] It is due to the fact that the number of samples in dataset VietNews is much bigger than in Wikilingua.

5.3. **Final settings.** In our baseline model PhoBERT2TRANS, we employed a pre-trained model PhoBERT-base as the Encoder, which supports only 256 input tokens. This limit caused the model to discard any information from tokens beyond the 256th position and thus affect the quality of the model output. To deal with this problem, we implemented a Window technique with window size = 256 that slides the encoder along both halves of the input sequence. This technique allowed the model to capture contextual information from both halves. The resulting context vectors from the first and second halves were concatenated into a single context vector (512 × hidden size), which contained the complete information from 512 tokens. The new context vector was then fed into the Decoder to produce the output. We conducted our Window experiment with PhoBERT2TRANS on two versions: with PhoBERT-base (135M parameters) and PhoBERT-large (369M parameters) as the Encoder. The result was reported in Table 4.

Table 4. Effect of Window technique on PhoBERT-based abstractive models

Method	viWikiHow-Abs-Sum			VietNews-Abs-Sum		
	R1	R2	RL	R1	R2	RL
PhoBERT2TRANS	49.86	21.07	32.95	56.89	25.8	36.84
PhoBERT2TRANS + Window	**51.14**	**22.27**	**33.34**	**57.26**	**26.17**	**37.09**
PhoBERTLarge2TRANS + Window	44.69	15.35	28.72	52.11	16.77	31.25

Note: R1, R2, and RL denote ROUGE-1, ROUGE-2, and ROUGE-L scores.

The best scores are bolded numbers and the second-best scores are underlined.

Besides PhoBERT-based models, we conducted another experiment with the same settings but employed a pre-trained model mBERT as the encoder (named mBERT2TRANS) to compare performance of the multilingual versus monolingual language model in transfer learning for Vietnamese abstractive summarization task.

During the prediction phase, we decoded each model three times which correspond to the expected min output length of 20, 30, and 40 thanks to the min length penalty technique. After that, we calculated ROUGE scores and reported the results in Table 5.

Table 5. Effect of min length penalty technique on our PhoBERT2TRANS + Window and mBERT2TRANS models

Method	viWikiHow-Abs-Sum			VietNews-Abs-Sum		
	R1	R2	RL	R1	R2	RL
PhoBERT2TRANS + Window (minL20)	51.14	22.27	33.34	57.26	**26.17**	**37.09**
PhoBERT2TRANS + Window (minL30)	53.36	22.74	33.62	**58.76**	26.16	36.71
PhoBERT2TRANS + Window (minL40)	53.87	**22.78**	33.16	58.04	25.52	35.4
mBERT2TRANS (minL20)	51.01	21.28	33.53	56.02	24.81	36
mBERT2TRANS (minL30)	52.85	21.68	33.98	57.25	25.08	36.19
mBERT2TRANS (minL40)	**55.02**	22.13	**34.21**	58.71	25.24	35.99

Note: R1, R2, and RL denote ROUGE-1, ROUGE-2, and ROUGE-L scores.

The best scores are bolded numbers and the second-best scores are underlined.

6. **Results.** Table 4, and Table 5 report the ROUGE-1, ROUGE-2, and ROUGE-L scores on the two benchmark datasets of our methods.

The results in Table 4 show that the Windows technique we applied to our model PhoBERT2TRANS allowed it to comprehend the full context of 512 tokens, that is why it gets the best ROUGE score on both datasets. While we predicted that employing a larger PhoBERT model would result in improved performance, it produced the lowest score, suggesting that using a more extensive encoder may cause the model to overfit and perform poorly at the decoding phase.

Table 5 reports the effect of the min length penalty technique on the prediction phase of our best-performing monolingual PhoBERT-based abstractive model compared with the multilingual mBERT-based abstractive model. The results show that setting min output length = 40 yields the best results on viWikiHow-Abs-Sum dataset, while min output length = 20 and 30 get better results on VietNews-Abs-Sum. PhoBERT-based models get better performance than mBERT-based ones on VietNew-Abs-Sum dataset. That can be explained as PhoBERT was pre-trained on a large dataset that contain 19GB of Vietnamese news articles, while mBERT was all pre-trained on Wikipedia articles.

Table 6 reports our models' performance in ROUGE-1, ROUGE-2, and ROUGE-L compared with the three strong baseline models from [8] on the viWikiHow-Abs-Sum dataset. Our methods outperformed all three strong baselines from 1.72% to 2.21% across all metrics.

In VieSum, the three strong baselines are trained on the raw dataset VietNews without duplicate removal. Therefore, to ensure a fair comparison with these baselines, we

Table 6. Our models' benchmark compares with the strong baselines from VieSum on viWikiHow-Abs-Sum dataset

Method	viWikiHow-Abs-Sum		
	R1	R2	RL
Baselines (from VieSum)			
ViBERT2ViBERT	53.08	20.18	31.79
PhoBERT2PhoBERT	50.4	19.88	32.49
mBERT2mBERT	52.82	20.57	31.55
Ours			
PhoBERT2TRANS + Window (minL20)	51.14	22.27	33.34
PhoBERT2TRANS + Window (minL30)	53.36	22.74	33.62
PhoBERT2TRANS + Window (minL40)	53.87	**22.78**	33.16
mBERT2TRANS (minL20)	51.01	21.28	33.53
mBERT2TRANS (minL30)	52.85	21.68	33.98
mBERT2TRANS (minL40)	**55.02**	22.13	**34.21**

Note: R1, R2, and RL denote ROUGE-1, ROUGE-2, and ROUGE-L scores.
The best scores are bolded numbers and the second-best scores are underlined.

Table 7. Our models' benchmark compares with the strong baselines from VieSum on VietNews dataset (without duplicates removal)

Method	VietNews		
	R1	R2	RL
Baselines (from VieSum)			
ViBERT2ViBERT	59.75	27.29	36.79
PhoBERT2PhoBERT	**60.37**	**29.12**	**39.44**
mBERT2mBERT	59.67	27.36	36.73
Ours			
PhoBERT2TRANS + Window (minL20)	57.95	26.65	37.47
PhoBERT2TRANS + Window (minL30)	59.03	26.52	36.97
PhoBERT2TRANS + Window (minL40)	58.21	25.84	35.67
mBERT2TRANS (minL20)	57.2	25.67	36.68
mBERT2TRANS (minL30)	58.02	25.83	36.76
mBERT2TRANS (minL40)	59.18	25.86	36.45

Note: R1, R2, and RL denote ROUGE-1, ROUGE-2, and ROUGE-L scores.
The best scores are bolded numbers and the second-best scores are underlined.

conducted an additional round of training and evaluation on the raw dataset VietNews to get ROUGE scores, which are reported in Table 7. The result shows that our methods still have a small under-gap compared with the strong baseline models.

7. **Conclusion.** In this paper, we proposed a novel method for finetuning pre-trained BERT models for Vietnamese abstractive summarization. It demonstrated that our methods outperformed three strong baselines from VieSum across all metrics on viWikiHow-Abs-Sum dataset. In contrast to [8], our methods follow the Transformer Encoder-Decoder architecture and are totally implemented from scratch, which can be modified without any limitation to optimize the training and prediction phases. Moving forward, we plan to apply coverage mechanisms and topic-based techniques to obtaining better model output.

Acknowledgment. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] A. M. Rush, S. Chopra and J. Weston, A neural attention model for abstractive sentence summarization, *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp.379-389, 2015.

[2] C. Li, F. Liu, F. Weng and Y. Liu, Document summarization via guided sentence compression, *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, pp.490-500, 2013.

[3] Y. Liu and M. Lapata, Text summarization with pretrained encoders, *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp.3730-3740, 2019.

[4] H.-Q. Le, Q.-A. Nguyen, Q.-H. Duong, M.-Q. Nguyen, H.-S. Nguyen, T. D. Thanh, H.-Y. T. Vuong and T. M. Nguyen, UETfishes at MEDIQA 2021: Standing-on-the-shoulders-of-giants model for abstractive multi-answer summarization, *Proc. of the 20th Workshop on Biomedical Language Processing*, pp.328-335, 2021.

[5] D.-C. Can, Q.-A. Nguyen, Q.-H. Duong, M.-Q. Nguyen, H.-S. Nguyen, L. N. T. Ngoc, Q.-T. Ha and M.-V. Tran, UETrice at MEDIQA 2021: A prosper-thy-neighbour extractive multi-document summarization model, *Proc. of the 20th Workshop on Biomedical Language Processing*, pp.311-319, 2021.

[6] X.-D. Doan, L.-M. Nguyen and K.-H. N. Bui, Multi graph neural network for extractive long document summarization, *Proc. of the 29th International Conference on Computational Linguistics*, Gyeongju, Korea, pp.5870-5875, 2022.

[7] T.-T. Nguyen, H.-H. Nguyen and K.-H. Nguyen, A study on Seq2seq for sentence compression in Vietnamese, *Proc. of the 34th Pacific Asia Conference on Language, Information and Computation*, Hanoi, Vietnam, pp.488-495, 2020.

[8] H. Nguyen, L. Phan, J. Anibal, A. Peltekian and H. Tran, VieSum: How robust are transformer-based models on Vietnamese summarization?, *arXiv Pre-Print*, https://doi.org/10.48550/arXiv.2110.04257, 2021.

[9] L. Phan, H. Tran, H. Nguyen and T. H. Trinh, ViT5: Pretrained text-to-text Transformer for Vietnamese language generation, *Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, Hybrid: Seattle, Washington + Online, pp.136-142, 2022.

[10] H. Q. To, K. V. Nguyen, N. L.-T. Nguyen and A. G.-T. Nguyen, Monolingual vs multilingual BERTology for Vietnamese extractive multi-document summarization, *Proc. of the 35th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, pp.692-699, 2021.

[11] N. Ti-Hon and D. Thanh-Nghi, Text summarization on large-scale Vietnamese datasets, *Journal of Information and Communication Convergence Engineering*, Korea Institute of Information and Communication Engineering, pp.309-316, 2022.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Proc. of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Red Hook, NY, USA, pp.6000-6010, 2017.

[13] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proc. 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, MN, USA, pp.4171-4186, 2019.

[14] Q. Nguyen and A. T. Nguyen, PhoBERT: Pre-trained language models for Vietnamese, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp.1037-1042, 2020.

[15] T. V. Bui, T. O. Tran and P. Le-Hong, Improving sequence tagging for Vietnamese text using transformer-based neural models, *Proc. of the 34th Pacific Asia Conference on Language, Information and Computation*, Hanoi, Vietnam, pp.13-20, 2020.

[16] F. Ladhak, E. Durmus, C. Cardie and K. McKeown, WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp.4034-4048, 2020.

[17] V.-H. Nguyen, T.-C. Nguyen, M.-T. Nguyen and N. X. Hoai, VNDS: A Vietnamese dataset for summarization, *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pp.375-380, doi: 10.1109/NICS48868.2019.9023886, 2019.

[18] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, *Text Summarization Branches Out*, Barcelona, Spain, pp.74-81, 2004.

[19] D. Q. Nguyen, D. Q. Nguyen, T. Vu, M. Dras and M. Johnson, A fast and accurate Vietnamese word segmenter, *Proc. of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, ISBN: 979-10-95546-00-9, 2018.

[20] T.-P. Nguyen and A.-C. Le, A hybrid approach to Vietnamese word segmentation, *2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, Hanoi, Vietnam, pp.114-119, doi: 10.1109/RIVF.2016.7800279, 2016.

[21] D.-V. Nguyen, D. V. Thin, K. V. Nguyen and N. L.-T. Nguyen, Vietnamese word segmentation with SVM: Ambiguity reduction and suffix capture, in *Part of the Communications in Computer and Information Science Book Series (CCIS, Volume 1215)*, doi: 10.1007/978-981-15-6168-9_33, 2020.

[22] N. L. Tran, D. Le and D. Q. Nguyen, BARTpho: Pre-trained sequence-to-sequence models for Vietnamese, *Proc. of Interspeech 2022*, pp.1751-1755, doi: 10.21437/Interspeech.2022-10177, 2022.

[23] L. Prechelt, Early stopping – But when?, in *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*, G. Montavon, G. B. Orr and K. R. Müller (eds.), vol.7700, Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-35289-8_5, 2012.

[24] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey et al., Google's neural machine translation system: Bridging the gap between human and machine translation, *arXiv Pre-Print*, https://doi.org/10.48550/arXiv.1609.08144, 2016.