DIGITAL VIETNAMESE ARTICLES SUMMARIZATION USING TRANSFER LEARNING

Cao Dinh Duy Ngoc[1,1*†], Do Pham Phuc Tinh[1,1*†], Nguyen Tran Gia The[1,1*†] and Nguyen Huu Minh Tam[1,1*†]

[1*]VNUHCM, University of Information Technology, VietNam.

*Corresponding author(s). E-mail(s): 20521661@gm.uit.edu.vn; 20522020@gm.uit.edu.vn; 20521940@gm.uit.edu.vn; 20521871@gm.uit.edu.vn;

†These authors contributed equally to this work.

Abstract

Summarization is a task that helps readers to access the content without reading the entire text, which allows them to update daily information and news more thoroughly and comprehensively, all the while not being time-consuming. Recently, the task of summarizing text has been more developed than before, which has led to the emergence of more pre-trained models for the ability to summarize text with high accuracy. In this task, we collected the ViDN4ABS dataset from some digital Vietnamese newspapers, such as tuoitre.vn, thanhnien.vn, vnexpress.net, vtc.vn, after that we conducted experiments with some models that are said to be capable of abstractive summarizing texts, such as T5, ViT5, Pegasus, Prophetnet, Distilbart, BARTPho, then compared and evaluated the results between these models. Finally, we highlighted some challenges and set out the future development direction for the abstractive summarization task.

Keywords: Abstractive summarization, ViDN4ABS, mT5, ViT5, Pegasus, Prophetnet, Distilbart, BART, BARTPho

1 Introduction

The demand for information is increasing, especially in the current era of technological development. Accessing information and news is time-consuming due to the enormous amount of textual data. Extracting essential information from the text is a highly effective solution to help people quickly and comprehensively access the information. Automatic text summarization was developed to solve this problem, and people began researching it decades ago (in 1958[1]). However, until now, automatic text summarization is still a challenge. Automatic text summarization techniques can be divided into two categories: extractive summarization and abstractive summarization. Abstractive summarization is thought to be the future of text summarization; it gives coherent, more "human-like" output text. However, abstract text summarization is less developed and effective than extractive text summarization due to the limitations of representation, comprehension, and text generation methods.

In recent years, with the improvement of neural network technology, deep learning has become an effective method for many tasks, including text summarization; with the emergence of pre-trained models, abstractive text summarization tasks have significantly developed. In this project, we tested pre-trained models for abstractive summarization on the ViDN4ABS dataset that we collected from famous digital Vietnamese newspapers. Then, we compared the performance achieved by the ROUGE scale on abstractive text summarization models. The highest result was on the BARTPho-syllable model, which was 59.84% on the ROUGE-1 scale, 27.98% on the ROUGE-2 scale, and 38.41% on the ROUGE-L scale.

The remaining sections in this paper are as follows: Part 2: we carry out a literature review of previous related works. Part 3 will introduce the dataset. In part 4, we present some of the pre-trained models to use. Part 5 is the experimental setting. Part 6 is the results of this research. Finally, in part 7, we draw conclusions and some problems that we need to solve in the future.

2 Related works

2.1 Method

Recently, models for summarization tasks, especially abstractive summarization, have been developed and published significantly using different algorithms: statistics, graph-based, and transformer. Pegasus [2] and Prophetnet [3] are models explicitly developed for abstractive summarization. In addition, many models are also tuned to perform abstractive summarization tasks, such as T5, ViT5, and BARTPho. Text summarization is divided into two main domains: extractive summarization and abstractive summarization [4]. Many works have been published in extractive summarization[5]. however, research works in abstractive summarization have increased rapidly[4].

There are two approaches: supervised learning and unsupervised learning. For supervised learning, this is a common approach to text summarization. Data for this approach includes an original text and a text summarization. Humans will generate summary paragraphs based on the content of the original text. For this method, it is evaluated on popular metrics, such as ROUGE N and ROUGE L. This has many works [6]. For unsupervised learning, the output of models is based on the score of sentence relevance to the article.

2.2 Datasets

With the intense explosion of text summarization, the number of datasets for text summarization increased rapidly, especially in the summary section. Datasets are increasingly diverse in both languages and data domains.

For Vietnamese, in recent years, there have been many datasets, such as ViMs[7], including 300 subdirectories, which are 300 news clusters, and 600 summaries. The number of articles is 1,945. VNDS[8] includes articles collected from online newspapers (tuoitre.vn, vnexpress.net, and nguoiduatin.vn). 105,418 articles in the training set, 22,642 in the validation set, and 22,644 in the test set. In VietnameseMDS[1], data is collected from the Baomoi.com website with 200 topics and 600 articles. VSoLSCSum[9], The dataset consists of 141 open-domain articles in 12 topics.

For other languages, several datasets are mainly from the news domain: Gigaword, Xsum, and Newsroom), CNN/Daily Mail, and BigPatient, WikiLingua.

3 The Dataset

3.1 Dataset Creation Process

We have created the **ViDN4ABS** dataset by collecting data from websites: thanhnien.vn, tuoitre.vn, vnexpress.net, vtc.vn in many different fields: news, education, youth, sports, health, business, economy, entertainment, science–technology, law, life, tourism, automobiles–motorcycles, culture and transportation. The digital newspapers mentioned above are reputable and top-quality in Vietnam, and the fields we focus on are also popular with people. Information is constantly updated every hour with a tremendous amount of information and massive traffic. Therefore, we decided to choose the newspaper and those fields to build the dataset for our task.

Figure 1 One article is on the thanhnien.vn in the field of youth.

For the above articles of the thanhnien.vn page (figure 1), we chose the first paragraph under the heading "Bắt bệnh cho cây" as the summary for the article, i.e., the paragraph: "Nam sinh lớp 12 ...". The remaining paragraphs after the summary

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

Table 1 An example shows the dataset.

Text: Đã có không ít tranh cãi về việc nên để Ronaldo ngôi sao 37 tuổi đang bên kia sườn dốc sự nghiệp ngồi dự bị nhưng HLV Fernando Santos vẫn trọng dụng CR7 ở vị trí xuất phát trong cả 3 trận vòng bảng. Bất ngờ xảy ra ở vòng 16 đội, Ronaldo phải ngồi ghế dự bị khi tuyển Bồ Đào Nha đấu Thụy Sĩ. Đó được xem là canh bạc mạo hiểm của HLV Fernando Santos bởi nếu sự thay đổi này không mang lại hiệu quả, ông sẽ gánh rất nhiều chỉ trích. Chưa kể những hệ lụy khó lường từ việc ngôi sao có tầm ảnh hưởng như Ronaldo phải ngồi ngoài. Thú vị thay, Goncalo Ramos, người được chọn thay thế Ronaldo đã chơi một trận đấu còn hơn cả Ronaldo thời đỉnh cao. Cú hat-trick của tiền đạo này không chỉ giúp tuyển Bồ Đào Nha thắng tưng bừng 6-1 trước Thụy Sĩ mà còn chứng minh rằng HLV Fernando Santos đã đúng. Với nhiều người màn trình diễn của tuyển Bồ Đào Nha khi không có Ronaldo chứng minh thực tế khác là đội bóng này trở nên đáng sợ hơn. Trước đó khi có Ronaldo trên sân, tuyển Bồ Đào Nha chơi gò bó, phụ thuộc khá nhiều vào ngôi sao này nên các đường tấn công thiếu biến hóa và nhịp độ chậm. Không có Ronaldo, tuyển Bồ Đào Nha có những pha tấn công biên đầy sắc bén đồng thời phản công nhanh đầy lợi hại. Vào sân từ phút 74, Ronaldo không để lại nhiều dấu ấn. Hình ảnh Ronaldo rời khỏi sân xuống đường hầm sớm báo hiệu cho tương lai không mấy sáng sủa của anh ở tuyển Bồ Đào Nha. Không thể phủ nhận tài năng cũng như những đóng góp của CR7 cho tuyển Bồ Đào Nha, nhưng phải chấp nhận thực tế "Selecao châu u" không cần phải phụ thuộc quá nhiều vào anh nữa rồi.
Summary: Quyết định để Ronaldo ngồi dự bị ở trận đấu với Thụy Sĩ được xem là canh bạc của HLV Fernando Santos nhưng ông đã chứng minh đó là quyết định đúng đắn, mang lại chiến thắng vang dội cho tuyển Bồ Đào Nha.

paragraph will be the content section for later model training—other newspapers such as tuoitre.vn, vnexpress.net, vtc.vn we take the same as thanhnien.vn.

The collected data has some standardized spaces containing HTML tags, so we use dedicated libraries to handle this. Many articles contain non-content information, so they do not have the necessary information, which can lead to wrong predictions. Therefore, we remove articles under 300 characters and summaries under 20 characters. After preprocessing, we obtained a dataset which was ready to implement the models consisting of 37,367 rows and two columns

3.2 Dataset Overview

After creating dataset from digital newspapers: thanhnien.vn, tuoitre.vn, vnexpress.net, vtc.vn, we have 37,367 rows and two columns: text and summary.

To perform the experiments, we divided the 37,367-row dataset into three sets: 80% for the 29,893-row training set, 10% for the 3,737-row development set, and 10% for the 3,737-row test set.

In addition, we also analyzed the characteristics of the original dataset and the three sets that were divided. The maximum input length is 1024 tokens, and the summary is almost all under 64 words. In some models, because of hardware limitations, we chose the input length of 700 for their best performances.

From table 2 and Figure 2 As, the content length of the focused articles is almost less than 1024 with 91.51%, and the mode about 300-700 words, the average is 569.72 words. In the summary part, almost all summaries are less than 64 (99.98%) in length, and most are between 25-50 words, averaging 36.48 words.

Table 2 Some characteristics in the dataset.

	Entire	Training set	Validation set	test set
Number of sentences	37367	29893	3737	3737
Mean_text	569.72	570.06	566.43	570.29
Mean_summary	36.48	36.49	36.65	36.57
Text ratio <=1024	91.51%	91.48%	91.86%	91.45%
Text ratio <=700	72.64%	72.53%	72.84%	73.32%
Summary<=64 ratio	99.87%	99.87%	99.84%	99.87%

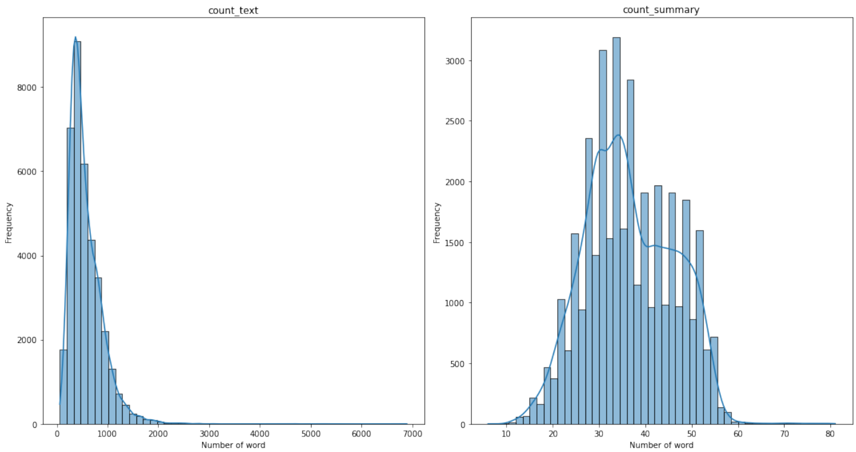

Figure 2 Frequency of text length (left) and summary length(right).

Thus, choosing 1024 words for input length and 64 words for output length will help the model learn better when the number of skips will not be too much, from which the model learns almost as many cases, and the prediction results will be more effective.

4 Methodology

4.1 T5

Text-to-text Transfer Transformer (T5)[10] is a language model introduced in 2020 featuring a transformer-based architecture. T5 is an E-D Transformer with some architectural changes (applying layer normalization before one sub-block, then adding the initial input to the sub-block of the output (pre-norm)). The configuration of the T5 is quite similar to BERT[11].

In particular, all tasks (translation, classification...) with T5 have the same basic structure. The input and output of the model are in the type of a text string. We can use the same model, hyperparameters, and loss function in all tasks. In this project, we use T5 to solve the problem of summarizing text, specifically Vietnamese articles. There are five variants of T5: base, small, large, 3B, and 11B with different numbers of parameters and model sizes. Our team focuses on two of those five variations.

T5-small is a checkpoint with 60 million parameters, with only six layers of E-D. T5-small was pre-trained on the Colossal Clean Crawled Corpus (C4) dataset [12].

mT5[13] is a multilingual pre-trained text-to-text transformer model developed based on the T5 model, improved to maximize performance when used for multilingual. mT5 is pre-trained on the mC4 dataset, which includes more than 100 languages differs from the previous C4 set, which was mainly English. We focus on two variants of mT5, mT5-small with 60M parameters and mT5-base with 222M paras. mT5-base can be considered the closest model to BERT[9].

ViT5 [14] was created by a team of authors of VietAI Research and is a pre-trained Transformer-based E-D model developed based on a diverse Vietnamese dataset and meets quality standards. Other models can give good results on English datasets. However, the complexity of the Vietnamese language is why the ViT5 model was created. The ViT5 model is mainly used for two text generation tasks: Abstractive Text Summarization and Named Entity Recognition. ViT5 is also based on the E-D architecture. The ViT5 model has two variants: ViT5-base with 310M paras and ViT5-large with 860M paras.

4.2 Pegasus

Transformer models work well and are a step forward for solving NLP problems. However, abstractive summarization has yet to be developed. Therefore, Pegasus[2] was created in 2019; there are two variants, Pegasus-base with 223M paras and Pegasus-large with 568M paras, using XSum, CNN/DailyMail, WikiHow [15], and Reddit TIFU[16] datasets. This model focuses on summarizing based on the Large Transformer E-D models. In this task, we hide meaningful sentences and generate words to create a summary text. This model achieves exceptionally high results with low-resource summarization. Pegasus-large is a model designed with L = 16 layers E-D, hidden size H = 1024, feed-forward layer size F = 4096, and the number of self-attention heads A = 16. Pegasus was pre-trained with C4 and HugeNews.

4.3 Prophetnet

Prophetnet [3] is a Seq2seq pretraining model based on the Transformer E-D architecture. Compared to the original Transformer seq2seq models, there are some differences: The novel self-supervised objective called future n-gram prediction, the n-stream self-attention mechanism, the mask-based autoencoder denoising task for Seq2Seq pretraining.

Instead of optimizing one-step-ahead prediction in the traditional sequence-to-sequence model, the ProphetNet is optimized by n-step ahead prediction that predicts the subsequent n tokens simultaneously based on previous context tokens at each time step.

4.4 BART

The Bart model[17] proposed in 2019 is a monolingual model using a standard machine translation/seq2seq architecture with a bidirectional encoder (such as BERT) and a left-to-right decoder (such as GPT). BART-base consists of 6 E-D layers, while BART-large has 12. BART is about 10% more in terms of some parameters than the equivalent-sized BERT model.

BART is trained in 2 steps: Transform the text with noise functions, then learn the models to recreate the original text. Therefore, BART works exceptionally well with text generation tasks; we uses BART to summarize the text in an abstract style.

Distilbart[18]: A smaller student model is trained to imitate the actions of a larger teacher model. When performing distillation, terms are added for losses on the encoder output, the outputs of the encoder and decoder hidden layers, and the prediction layer (i.e., the soft target probabilities). DistillBartAgent[2] is used for distilBART.

BARTPho[19]: These are a large number of pre-trained Sequence-to-Sequence models used for English, as opposed to Vietnamese. As a result, BARTPho was developed. BARTPho has two versions: BARTPho-word with 420M paras and BARTPho-syllable - the first public large-scale monolingual sequence-to-sequence models, pre-trained for Vietnamese with 396M paras. BARTPho has 12 E-D layers.

BART[17] is a multilingual BART model, but compared to BARTPho, the model dedicated to Vietnamese, we found that the performance of BARTPho is outstanding.

5 Experimental setting

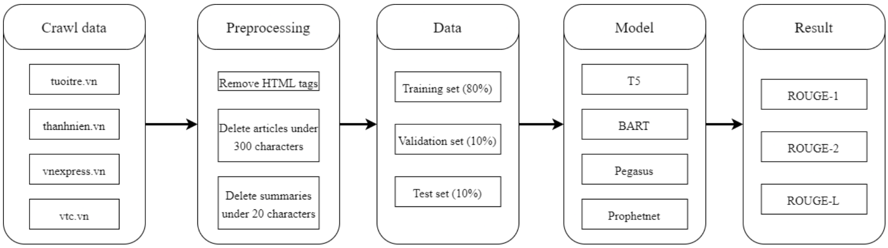

Figure 3 Experimental process

In this part, we do it through five steps. First, we collect data from digital newspapers in a variety of fields. Then, processing the raw data because there are some articles that are not useful to the model, which can cause interference that makes the model inferior. After processing the collected raw data, we obtained a dataset of 37,367 rows and two columns, divided into three sets: training set, validation set, and test set in the ratio of 8:1:1. Once we have the datasets, we install models with different versions for the abstractive summarization tasks. Our hyperparameter is presented in table 3:

[2] https://github.com/facebookresearch/ParlAI/blob/main/projects/anti_scaling/README.md

Table 3 Hyperparameters Experimental Setting

Model	Max input length	Max output length	Batch size	Learning rate	epoch
T5-small	1024	62	32	1.00E-05	3
mT5-small	1024	62	32	5.00E-05	3
mT5-base	700	64	8	5.00E-05	1
mT5-base	700	64	32	5.00E-05	3
ViT5-base	1024	64	4	1.00E-04	3
ViT5-base	1024	64	4	1.00E-04	5
ViT5-large	1024	64	2	1,00E-04	3
Pegasus-large	700	64	32	1.00E-04	3
Prophetnet-base	1024	62	8	1.00E-04	3
Distilbart	1024	64	48	5.00E-05	3
BARTPho-syllable	1024	64	2	1.00E-04	3
BARTPho-syllable	1024	64	2	1.00E-04	5
BARTPho-word	1024	64	2	1.00E-04	3

6 Experimental results

6.1 Metrics

In our experiment, we use the F-measure of the ROUGE score to evaluate the model performance. **ROUGE**[20] is an automatic summarization evaluation benchmarking metric that compares the machine-generated summary to the reference summary to determine the quality of the produced summary (ideal or human-written). The number of words that overlap between the reference and the computer-generated summaries is used to calculate ROUGE scores. More specifically, we use ROUGE-N (ROUGE1, ROUGE-2) and ROUGE-L to evaluate the model performance.

ROUGE-N: It denotes the overlapping of n-grams between the system-generated summary and the reference summary.

$$ROUGE-n=\frac{\sum_{S\in RS}\sum_{\mathrm{gram}_n\in S}\mathrm{Count}_{\mathrm{match}}\left(\mathrm{gram}_n\right)}{\sum_{S\in RS}\sum_{\mathrm{gram}_n\in S}\mathrm{Count}(\mathrm{gram}_n)}.$$

Where:

- RS is a set of reference summaries
- n stands for the length of the n-gram, gram_n
- Countmatch(gram) is the maximum number of n-grams co-occurring in a generated summary and a set of reference summaries.

ROUGE-L: It denotes the Longest Common Subsequence (LCS) matching between the reference summary and system-generated summary.

$$\begin{cases} P_{LCS}(R,S)=\dfrac{\mathrm{LCS}(R,S)}{\mathsf{S}} \\ R_{LCS}(R,S)=\dfrac{LCS(R,S)}{\mathsf{R}} \\ R_{LCS}(R,S)=\dfrac{\left(1+\beta^2\right)P_{LCS}(R,S)R_{LCS}(R,S)}{\beta^2P_{LCS}(R,S)+R_{LCS}(R,S)} \end{cases}$$

Where:

- R and S are the machine-generated summary length and the reference summary length, respectively
- LCS(R, S) is the LCS between R and S
- $P_{LCS}(R,S)$ is the precision of LCS(R, S)
- $R_{LCS}(R,S)$ is the coverage of LCS(R, S)
- $\beta = P_{LCS}(R,S)/R_{LCS}(R,S)$.

6.2 Results

After training the models with the input, and output maximum lengths and the parameters of the models are shown in table 3, we obtain the best results with each model presented in table 4 as follows:

Table 4 The best results of models

Model	Validation set			Test set		
	ROUGE 1	ROUGE 2	ROUGE L	ROUGE 1	ROUGE 2	ROUGE L
T5-small	29.61	11.11	22.54	29.58	11.03	22.46
mT5-small	51.00	19.59	32.04	51.01	19.54	31.97
mT5-base	55.27	23.21	35.66	54.84	22.77	35.40
ViT5-base	59.03	27.96	38.67	58.91	27.74	38.60
ViT5-large	59.33	28.25	38.86	59.14	27.93	38.62
Pegasus-large	26.11	11.25	20.54	25.66	10.98	20.12
Prophetnet-base	28.70	14.37	21.93	28.88	14.13	21.85
Distilbart	53.22	20.08	32.86	53.06	19.56	32.65
BARTPho-syllable	59.88	28.35	38.64	59.84	27.98	38.41
BARTPho-word	57.79	25.45	36.74	47.98	13.97	28.90

According to the evaluation of the Validation set in table 8, it can be seen that the BARTPho-syllable model has the highest results with 59.88% on the ROUGE-1 score, 28.35% on the ROUGE-2 score, and 38.64% on the ROUGE-L score.

Similar to the evaluation on the Test set in table 8, the BARTPho-syllable model continues to have the best results, scoring 59.84% on ROUGE-1, 27.98% on ROUGE-2, and 38.41% on ROUGE- L.

It can be concluded that the BARTPho model has the best performance of the models we tested for the task of abstractive text summarization because it is the model especially built for Vietnamese. However, it is not too far off from the best results from the other models.

From the results in table 4, we can know that the language model based on pre-training has achieved good results. This is expected because these pre-trained models are trained on a large-scale external corpus to capture deeper semantic information of natural language. And right now, pretraining-based models nearly completely dominate the list of different NLP jobs.

Table 5 Wrong sample using BARTPho

Actual Summary: Quyết định để Ronaldo ngồi dự bị ở trận đấu với Thụy Sĩ được xem là canh bạc của HLV Fernando Santos nhưng ông đã chứng minh đó là quyết định đúng đắn, mang lại chiến thắng vang dội cho tuyển Bồ Đào Nha
Predict summary: Không có Cristiano Ronaldo, tuyển Bồ Đào Nha chơi kém cỏi và phản công nhanh hơn hẳn.

Machine-generated summaries are often different from reference summaries, the machine-generated summaries of models tend to be very meaningful and relevant, so a phenomenon not captured matching evaluation metrics such as Rouge.

On the other hand, the model sometimes 'misinterprets' the semantics of the text and generates a summary with a comical interpretation as shown by examples in table 1 and table 5. Models often do not yet know the information about phrases to complete the end of a new sentence. Capturing the 'meaning' of complex sentences remains a weakness of these models.

7 Conclusion and Future Works

7.1 Conclusion

In this paper, we have implemented and compared the models used for abstractive text summarization. We provide a comprehensive overview of currently available abstractive text summarization models. Additionally, we have built the dataset ViDN4ABS about the domain of electronic newspaper data. Then, give the results of analyzing the performance of different models on the dataset ViDN4ABS. Finally, we report the results of the performance analysis of various models on the ViDN4ABS dataset. The results indicate that BARTpho models[**?**] consistently give the best results in abstract summarization of digital newspapers. Experimental results show that our approach has promise in solving the abstractive summarization task.

7.2 Future Works

On the contrary, there is still room for future work. Since understanding a document depends on its structure. So, we will propose methods to enhance the completeness of the meaning of sentences generated in the summary text and investigate strategies for effectively combining sentences to provide a better summary for long texts. In addition, we also suggest a new dataset of domain-specific digital newspapers improve performance. As part of our future work, we plan to focus on this dataset and provide new models while enhancing current ones for abstract summaries of the new domain.

References

[1] Kumar, A., Uttam, A.K.: Information extraction and sentence ordering in multi-document summarization using preference learning. In: 2022 International Conference on Edge Computing and Applications (ICECAA), pp. 1146–1149 (2022). IEEE

[2] Zhang, J., Zhao, Y., Saleh, M., Liu, P.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: International Conference on Machine Learning, pp. 11328–11339 (2020). PMLR

[3] Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., Zhou, M.: Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. arXiv preprint arXiv:2001.04063 (2020)

[4] Alomari, A., Idris, N., Sabri, A.Q.M., Alsmadi, I.: Deep reinforcement and transfer learning for abstractive text summarization: A review. Computer Speech & Language **71**, 101276 (2022)

[5] To, H.Q., Van Nguyen, K., Nguyen, N.L.-T., Nguyen, A.G.-T.: Monolingual versus multilingual bertology for vietnamese extractive multi-document summarization. arXiv preprint arXiv:2108.13741 (2021)

[6] Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al.: Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023 (2016)

[7] 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)

[8] Nguyen, V.-H., Nguyen, T.-C., Nguyen, M.-T., Hoai, N.X.: Vnds: A vietnamese dataset for summarization. In: 2019 6th NAFOSTED Conference on Information and Computer Science (NICS), pp. 375–380 (2019). IEEE

[9] Nguyen, M.-T., Lai, D.V., Do, P.-K., Tran, D.-V., Le Nguyen, M.: Vsolscsum: Building a vietnamese sentence-comment dataset for social context summarization. In: Proceedings of the 12th Workshop on Asian Language Resources (ALR12), pp. 38–48 (2016)

[10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., *et al.*: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)

[11] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[12] Dodge, J., Sap, M., Marasovic, A., Agnew, W., Ilharco, G., Groeneveld, D., Gardner, M.: Documenting the english colossal clean crawled corpus (2021)

[13] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)

[14] Phan, L., Tran, H., Nguyen, H., Trinh, T.H.: Vit5: Pretrained text-to-text transformer for vietnamese language generation. arXiv preprint arXiv:2205.06457 (2022)

[15] Koupaee, M., Wang, W.Y.: Wikihow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305 (2018)

[16] Kim, B., Kim, H., Kim, G.: Abstractive summarization of reddit posts with multi-level memory networks. arXiv preprint arXiv:1811.00783 (2018)

[17] 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. arXiv preprint arXiv:1910.13461 (2019)

[18] Li, Z., Wang, Z., Tan, M., Nallapati, R., Bhatia, P., Arnold, A., Xiang, B.,

Roth, D.: Dq-bart: Efficient sequence-to-sequence model via joint distillation and quantization. arXiv preprint arXiv:2203.11239 (2022)

[19] Tran, N.L., Le, D.M., Nguyen, D.Q.: Bartpho: Pre-trained sequence-to-sequence models for vietnamese. arXiv preprint arXiv:2109.09701 (2021)

[20] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)