Using Topic in Summarization for Vietnamese Paragraph

Dat Tien Dieu, Dien Dinh

Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam

Abstract—**This article delves into the realm of refining the precision of automated text summarization tasks by harnessing the underlying themes within the documents. Our training data draws upon the VNDS dataset (A_Vietnamese_Dataset_ for_Summarization), encompassing a total of 150,704 samples aggregated from diverse online news sources like vnexpress.net, tuoitre.vn, and more. These articles have been meticulously processed to ensure they align with our training objectives and criteria. This paper presents an approach to text summarization that is theme-oriented, utilizing Latent Dirichlet Allocation to delineate the document's subject matter. The data subsequently have been fed into the BERT model, which constitutes one of the subtasks within the broader domain of abstractive summarization—summarizing content based on pivotal concepts. The results attained, although modest, underscore the challenges we've confronted. Consequently, our model necessitates further development and refinement to unlock its full potential.**

Keywords—Automatic text summarization; a theme-based approach; BERTmodel; latent dirichlet allocation

I. INTRODUCTION

Text summarization is a method that allows readers to quickly grasp the essential and core contents of a document, reducing reading time while retaining important and necessary information. Text summarization is not only applied in news articles and publications but also in search results, product descriptions, technical documents, and even summarizing related articles within a given topic. One of the earliest studies in the field of automatic text summarization was conducted by Erkan and his colleagues in 2007. Since then, automatic text summarization has become a formal task and has been approached by various research groups and integrated into their language models for evaluation purposes.

In terms of the purpose of text summarization, it can be divided into theme-based summarization and generic summarization. Currently, there are two main approaches to automatic text summarization: extractive summarization, abstractive summarization, and a combination of these two methods known as hybrid summarization.

In this research, the paper focuses on two primary directions related to generating text summaries. The first direction concentrates on extracting information from the original text to create a summary. Methods in this direction emphasize selecting sentences with high coverage in the original text for inclusion in the summary. This can be achieved through graph-based methods [1] or even using neural networks to identify important sentences [2]. The second direction involves creating a summary by abstracting

meaning from the original text. In this approach, machine learning models are used to generate new sentences based on the content of the original text, resulting in a summary with a completely different grammatical structure from the original text. Additionally, there is a hybrid approach that combines extraction and abstraction, creating a summary by extracting information and summarizing a portion of the original text.

All three research directions produce automatic summaries, which are evaluated by comparing them to reference summaries created by humans, using metrics such as BLEU [3] and ROUGE [4].

Our work is based on a dataset of Vietnamese text, which is compiled by aggregating news articles from the VnExpress electronic information portal. After collecting the articles, they are preprocessed to create an experimental dataset. Each document in this dataset consists of two main components: a pre-existing summary called a reference summary and the main content of the article. Then the dataset was divided into three separate parts model training, training parameter tuning, and finally, model evaluation.

II. RELATED WORKS

A. Model ViHeartBERT (2022)

The ViHeartBERT model (2022) [5] constructs a model with applications in the medical field, designed to help both patients and doctors understand scientific literature by providing clear explanations for medical abbreviations and summarizing frequently asked questions. The dataset used for training is a specialized medical dataset called acrDrAid. This model has a BERTbase-like architecture but is trained on its own dedicated dataset.

B. Model ViT5 (2022)

The ViT5 model (2022) [6] is an encoder-decoder model based on the Transformer architecture that has been pre-trained for the Vietnamese language. Following the self-supervised T5-style pre-training approach, ViT5 was trained on a large, high-quality, and diverse dataset of Vietnamese text. They assessed the performance of ViT5 on two downstream text generation tasks: Abstractive Summarization and Named Entity Recognition (NER).

ViT5's performance against several other Transformer-based encoder-decoder models. Our experiments demonstrated that ViT5 significantly outperforms existing models and achieves the best results in Vietnamese text summarization. In the Named Entity Recognition task, ViT5 competes strongly

with the previous best results obtained from pre-trained Transformer-based models.

C. Model BARTPho (2022)

The BARTpho model (2022) [7] is one of the leading and largest single-language sequence-to-sequence models pre-trained for Vietnamese. Our BARTpho model utilizes a "large" architecture and follows the pre-training strategy of the BART long document denoising model, making it particularly suitable for natural language generation tasks.

Tests conducted on a specific task related to Vietnamese text summarization demonstrate that BARTpho, both in terms of automatic evaluation and human assessment, outperforms the strong base model mBART and enhances the current state of the art. BARTpho has been released to support future research and applications in natural language generation tasks for the Vietnamese language.

Based on the promising results of language models, they have decided to train a language model based on a sequential model structure with attention mechanisms and the incorporation of document themes.

III. PROPOSAL

Following the research above, there is a need for a model that can be used to summarize a paragraph with the topic because using the topic helps to redirect the bag of words to the topic that it should be about, it helps to improve the accuracy of generation words and also the relation between generated text and original one. The attention mechanisms are really powerful but also the topic awareness, the bag of words have been narrowed down, it is easier to control the output to get better-generated text over time.

A. The Structure of a Sequential Model with Attention Mechanisms

The structure of the Sequence-to-sequence (Seq2Seq) model with an attention mechanism is depicted in Fig. 1 and serves as the foundation for developing the Pointer-generator Networks model. The model consists of:

Encoder**:** This is a bidirectional LSTM network with a single layer. It encodes the input sequence.

Decoder: This is a unidirectional LSTM network with a single layer. It decodes the encoded information from the encoder to generate the output sequence.

The document w_i is sequentially fed into the encoder to generate the encoded sequence of hidden states h_i.. At each time step t, the decoder receives the embedding of the previous word (during training, it's the word from the reference document, and during testing, it's the word selected at time step t - 1 by the decoder) and obtains the decoding state s_t.

The attention distribution is computed according to Eq. (1) and Eq. (2):

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{attn}) \quad (1)$$

$$a^t = softmax(e^t) \quad (2)$$

where, v, W_h, W_s and b_{attn} are trainable parameters. The attention distribution can be seen as a probability distribution over the vocabulary, informing the decoder about where to look to generate the next word. Subsequently, the attention distribution is used to compute the weighted sum of the hidden states from the encoder, referred to as the context vector h_t^* in Eq. (3):

$$h_t^* = \sum_i a_i^t h_i \quad (3)$$

The context vector, representing a fixed-size vocabulary, is read at the current time step, and it is combined with the decoder state s_t to generate the word probability P_{vocab} as described in Eq. (4):

$$P_{vocab} = softmax(V'(V[s_t, h_t^*] + b) + b') \quad (4)$$

where, V, V', b' and b are trainable parameters, and P_{vocab} is the probability distribution over all words in the vocabulary, providing us with the distribution of the word to be predicted as given in Eq. (5):

$$P(w) = P_{vocab}(w) \quad (5)$$

During the training process, the loss at time step t is determined by the negative log-likelihood of the word group w_t^* at that step, as defined in Eq. (6):

$$loss_t = -\log P(w_t^*) \quad (6)$$

Subsequently, the loss value for the entire input sequence is determined as per Eq. (7):

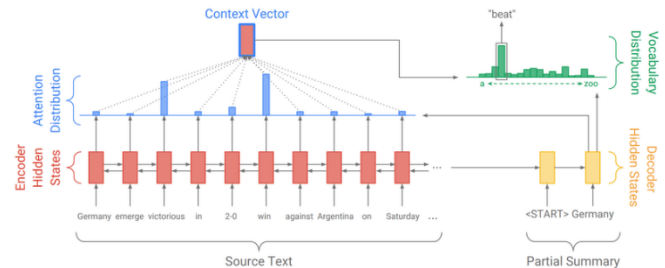$$loss = \frac{1}{T}\sum_{t=0}^{T} loss_t \quad (7)$$

Fig. 1. Structure of sequence-to-sequence attentional model.

Fig. 1 is the structure of the model that was used to feed the data to the model and get the context vector with vocabulary distribution.

B. The Generative Network Utilizes the Topic Modeling Method with Latent Dirichlet Allocation (LDA)

A model was built to demonstrate the capability of expressing the topic distribution of a text segment as well as an array representing the distribution of words related to those topics from raw data. To achieve this, using Latent Dirichlet Allocation (LDA) is an important part. This information is combined to enhance the accuracy of text summarization. There are several models which have been used to enhance the accuracy of the Natural Language Processing model, which can be referred to as "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey" of Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, Liang Zhao [8]. To learn more about how

to apply the LDA for the abstraction summarization, this paper is inspired by "LDA based topic modeling of journal abstracts" by P. Anupriya; S. Karpagavalli [13] about how to use LDA effectively in the summarization in order to get better summarization without using the same words which are in the original paragraph.

To better understand LDA, let's consider an example. having 1000 documents, each consisting of 500 words. This means it would need 1000x500 connections to understand how each word depends on each document and vice versa. Instead of having so many connections, and grouping these documents into topics, let's assume there are three topics. So, there will be 1000x3 = 3000 connections to determine which topics each document belongs to, and another 500x3 = 1500 connection to determine the word distributions that influence these 3 topics.

Fig. 2 and Fig. 3 demonstrate the differences between using an additional topic layer and not using any additional topic layer.
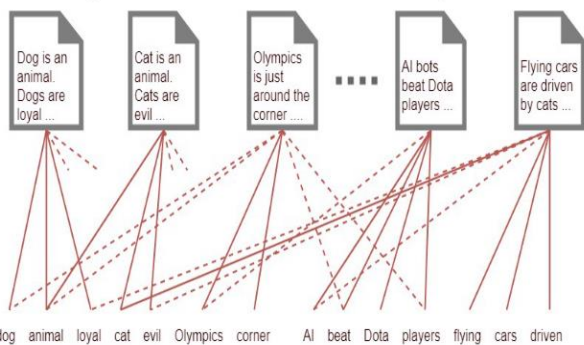
Fig. 2. Connections are necessary to determine the influence of words on documents.
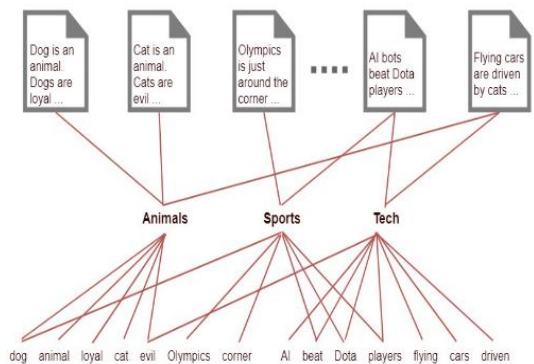
Fig. 3. Connections are necessary to determine the impact of words on the documents through topics.

Once the possible topics are identified within the documents, using CountVectorizer to train the LDA model and obtain the results as follows Fig. 4 is the next step. The distribution of words to the corresponding topic can see how the words are being distributed or which word contributes to which topic.

Fig. 4. Distribution of words to the corresponding topic.

C. Combining the Generative Network using the LDA Topic Modeling Method with the BERT Model

The Bayesian approach, which is based on the distribution of topics within the input text, can be highly effective for documents that have hidden topics.

$$decoding\ probability =$$
$$alpha * sum\_topic[\ p(word\ |\ topic)$$
$$* p(topic\ |\ paragraph)] + (1$$
$$- alpha)$$
$$* p\_transformer(word\ |\ context)$$

The combination approach of the LDA topic modeling method with the BERT model will be illustrated in the following figure:
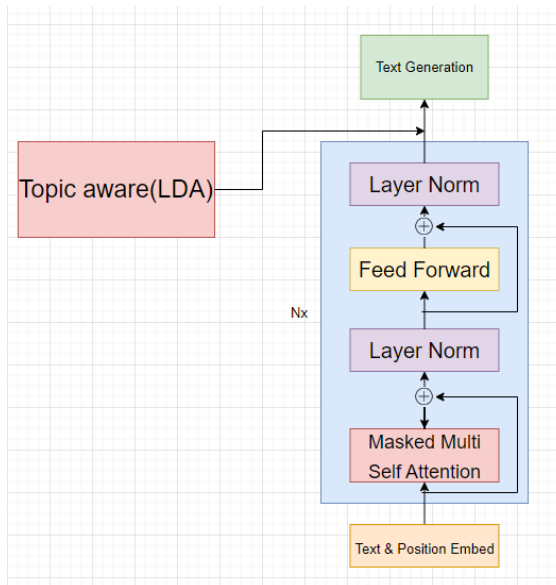

Fig. 5. The combination of BERT and LDA.

Fig. 5 demonstrates where the Topic Aware (LDA) has been added to the original BERT model and the reason for this is because this model has been designed to focus on the last layer before the model generates the words.

IV. DISCUSSION

The topic-aware approach is also potent, with several papers employing the topic as a primary feature in generating text, following the principles of "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey."[8]. This technique proves particularly effective in specific domains with extensive terminology, declarations, definitions, and so forth, such as Social Networks, Software Engineering, Crime Science, Geography, Political Science, and Medical Science. In such areas, where term ambiguity is prevalent, a model that can discern the appropriate words for generation is imperative. Conventionally, LDA has been used not only for topic detection but also to uncover hidden topics within paragraphs. When dealing with a multitude of words, identifying underlying themes can be challenging for humans, but LDA offers a viable solution. "Prediction of research trends using LDA based topic modeling" is also an inspired paper that shows there is a promising trend for using LDA-based topic modeling in the Natural Language Processing industry. The research in [10], an example of using LDA for topic modeling for new articles [12].

Additionally, the BERT model has demonstrated its position as a state-of-the-art (SOTA) innovation in the Natural Language Processing industry. The introduction of the attention mechanism has revolutionized the landscape of methods, paving the way for a more accurate and efficient approach to extracting meaning from text using computers. It stands as the initial and rational choice for fine-tuning in various applications, including our case. Therefore, our ultimate goal is to effectively combine these two methods.

There is a limit characteristic of the BERT model is about the length of input, it is just 256. Because of that, the BERT multilingual model has been used to replace the original BERT model. According to "How Multilingual is Multilingual BERT" by Telmo Pires, Eva Schlinger, and Dan Garrette [9]. Additionally, following the "Pre-training of Deep Bidirectional Transformers for Language Understanding" [11], it is obvious using pre-trained models which have the transformer architecture will be more efficient than other architectures.

V. EXPERIMENTS

A. Datasets

This study was experimentally evaluated on a dataset comprising 830,643 articles collected from the Vnexpress website. Each article includes a title, author's summary, and the body of the article. Of this dataset, 80% was used for training, and 10% was used for experimental evaluation. On average, each article contains about 20 sentences, with each sentence consisting of approximately 25 words or fewer.

Table I provides a detailed description of the preprocessed dataset. The data was divided into three sets: Training (Train), Validation (Validation), and Test (Test) in a respective ratio of 70:15:15. The average number of sentences in the summary section across the sets is 1.22 sentences, with approximately 28 words per summary sentence. Meanwhile, the average number of sentences in the text to be summarized is around 17 sentences, with approximately 418 words.

TABLE I. THE DATASET INFORMATION

	Training Set	Validation Set	Test Set
Numbers of sample	105,418	22,642	22,644
#avg number of sentences in abstract	1.22	1.22	1.22
#avg number of words in abstract	28.48	28.54	28.59
#avg number of sentences in body	17.72	17.81	17.72
#avg number of word in body	418.37	419.66	418.74

B. Settings

The experiments have been done with the generative network model using a pointer on a computer equipped with a GPU, 12.79 GB of RAM, and a 16GB GPU, running Tensorflow three library. The training parameters were set as follows:

- Number of hidden layers: 4
- Learning rate: 0.001
- Embedding word length: 128
- Beam size: 4
- Number of heads: 8
- Input sentence length: 512
- Output sentence length: 200

C. Results

Beginning the experimental phase with the LDA training model (a model that determines document topics based on the LDA algorithm). Since the number of topics is not predetermined, attempting to determine the number of topics through several methods is a crucial aspect.

First, the K-means clustering method has been used to determine the number of topics and assess how dividing topics influences the model's generative capability. After performing this, the result is obtained in the following chart:
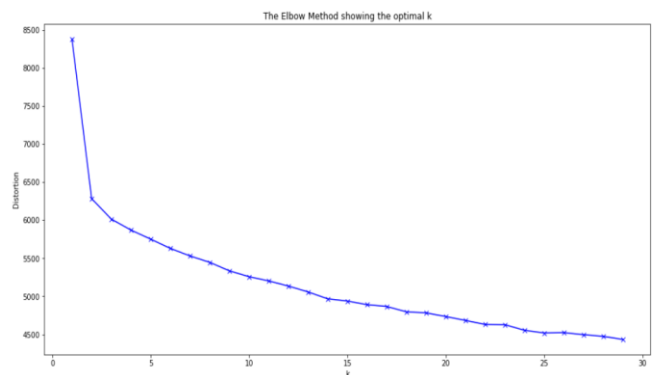

Fig. 6. The distribution of the number of topics when inputting.

Fig. 6 demonstrates the best number of topics that should be divided paragraphs into. Below Fig. 7 shows the distribution of topics in paragraphs, it can be seen obviously the paragraphs have the same topics tend to gather together.
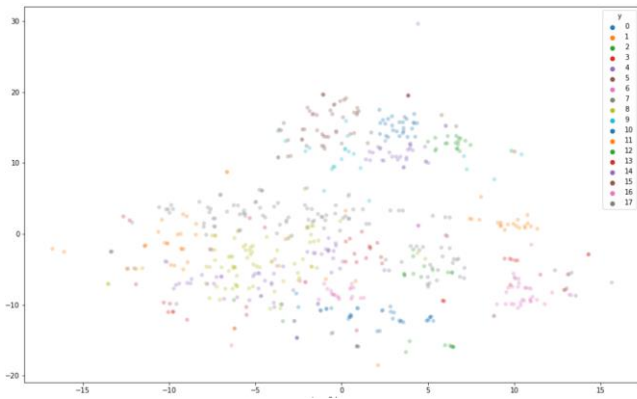
Fig. 7.   Distribution and clustering of topics in paragraph.

To confirm the meaningfulness and accuracy of this representation, T-SNE (t-Distributed Stochastic Neighbor Embedding) examined the distribution and clustering of the data. Fig. 8 demonstrates distribution topics over each paragraph, it can be referred to the contribution of those topics to the big topics in each paragraph.

Fig. 8.   Distribution of topics to each paragraph in the dataset.

With the results above, it shows that the distribution of topics across documents is uneven. This is because different words in different texts have varying meanings.

After completing the LDA experiments, the next experiments are combining LDA with the BERT model. Initially, the BERT model was trained from scratch, meaning there is no pre-trained model used, to assess its effectiveness and training capabilities without a suitable pre-trained model. However, the results were not very good, achieving ROUGE 1, ROUGE 2, and ROUGE L scores of 5.4%, 4.7%, and 5.6%, respectively. Therefore, training the BERT model from scratch using the above dataset did not yield the desired results. Consequently, to continue with better performance, it is urging to use a pretrained model to reduce training time and improve the accuracy of predicting the next word during text summarization. This decision led to significant improvements compared to training BERT from scratch. However, due to BERT's limitation in input text length, that will be better when transitioning to another pretrained model called BERT Multilingual. This model has a longer input text length capability than BERT and offers better embedding support. Table II is the result that model achieved in training multiple times.

TABLE II.   THE RESULT OF THE MODEL WITH THE ROUGE 1, ROUGE 2 VÀ ROUGE L

	ROUGE-1	ROUGE-2	ROUGE-L
Pre-trained BERT and LDA	0.102	0.081	0.113

VI. CONCLUTIONS AND DEVELOPMENT DIRECTIONS

A. Conclusions

Currently, with the experiment with the BERT model, it seems that the combination of LDA and the BERT model has not achieved the desired effectiveness. The ROUGE-1, ROUGE-2, and ROUGE-L scores have not shown significant improvements, and the generated words do not appear to be highly related to the input text's topic, and the grammar is not as polished as desired, despite the good training results of the LDA model.

Given that the LDA model's results are promising, it is a great idea to consider exploring other language models with the ability to incorporate new features into the model, such as BART. BART is a powerful language model that can be fine-tuned for various natural language processing tasks, including text summarization. It has demonstrated strong performance in abstractive summarization tasks and might provide improvements, especially when combined with LDA. Further experimentation and fine-tuning of this combined approach could yield better results.

B. Development Directions

Plan for the next step is to continue development based on models that achieve high ROUGE-1, ROUGE-2, and ROUGE-L scores in combination with the LDA model. To diversify the number of topics and ensure equal topic coverage, training an LDA model using deep learning methods is needed. This will enable us to create a model that can predict the topics of text effectively and seamlessly integrate it into state-of-the-art (SOTA) natural language processing models. This approach should help improve the quality and relevance of generated summaries in various applications.

REFERENCES

[1] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, vol. 22, p. 457-479, Dec 2004.

[2] R. Nallapati, B. Zhou, C. N. dos santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," 2016.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, (Philadelphia, Pennsylvania, USA), pp. 311-318, Association for Computational Linguistics, July 2002.

[4] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Text Summarization Branches Out, (Barcelona, Spain), pp. 74-81, Association for Computational Linguistics, July 2004.

[5] Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui, and Steven Quoc Hung Truong. 2022. ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 328–337, Marseille, France. European Language Resources Association.A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," 2015.

[6] Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5:

Pretrained Text-to-Text Transformer for Vietnamese Language Generation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 136–142, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.Z. Cao, W. Li, S. Li, F. Wei, and Y. Li, "Attsum: Joint learning of focusing and summarization with neural attention," 2016.

[7] Nguyen Luong Tran, Duong Minh Le, Dat Quoc Nguyen. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese 2022.

[8] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, Liang Zhao . Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Dec 2018.

[9] Telmo Pires, Eva Schlinger, and Dan Garrette. How Multilingual is Multilingual BERT?. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy. Association for Computational Linguistics. 2019.

[10] Rahul Kumar Gupta, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, Thoudam Doren Singh , Prediction of research trends using LDA based topic modeling. June 2022.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. May 2019

[12] Mykyta Kretinin; Giang Nguyen Topic Modeling on News Articles using Latent Dirichlet Allocation,p. 1543-9259, Oct 2022

[13] P. Anupriya; S. Karpagavalli. LDA based topic modeling of journal abstracts, Nov 2015