TEXT SUMMARIZATION ON VIETNAMESE DATASETS

Truong Dang Nghia
Information Technology University
20521658@gm.uit.edu.vn

Ngo Duc Hoc
Information Technology University
20521354@gm.uit.edu.vn

Pham Quang Huy
Information Technology University
20521412@gm.uit.edu.vn

Ngo Thi Hien Minh
Information Technology University
20521605@gm.uit.edu.vn

Abstract

In the current era, the explosion of informational text on the Internet has laid down a primitive formula for human information processing ability. To be able to grasp information quickly and easily, we need tools to help shorten text content. One of the effective solutions being researched to address the above needs is the document reunion model. In this study, we have built a data set consisting of 23,248 lines collected from Vietnamese newspaper pages. In addition, we use 3 pre-trained models, ViT5, BARTpho and mT5, applied to the above data set to build an effective automatic text summarization model to solve the raised and parallel problems. But that's the evaluation of our dataset. All three models worked effectively and produced quite good results.

I. INTRODUCTION

Today, the amount of text information is increasingly large due to the strong development of the Internet in the digital era. This makes reading comprehension and grasping information more difficult for humans. Most people are impatient with spending a lot of time reading through an article or a very long article on a particular topic and they are only interested in grasping the main information. mentioned in the text. This creates a great impetus for the natural language processing support technology industry to develop new technologies to solve difficulties in accessing information.

To help people process information quickly and effectively, many solutions in the field of natural language processing have been researched and applied. One of them is the automatic text summarization model. Instead of having to read and extract information from a long article themselves, readers can now easily access the main content in it through the automatic summarization process.

Text summarization models are divided into two types. One is in the extractive direction. Models in this direction will select important sentences in the text to return to the user, sentences that will not change anything in terms of grammar and vocabulary. The advantage of this method is to summarize important sentences without changing the sentences. However, if a text is too long and contains a lot of important information, retaining too many sentences may not be optimal. The other type is called the abtractive model. Models of this type will rely on the original text to create a summary sentence. This summary sentence can change in both grammar and vocabulary but still carries the main meaning of the text content. In this paper, we focus on the abtractive model because of its flexibility and interest.

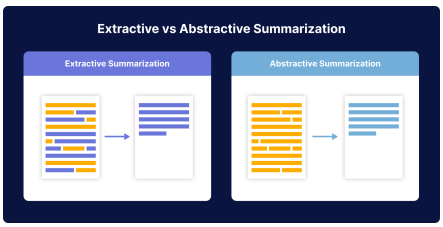

Figure 1: Two types of text summarization.

Describe the problem:

- Input: a Vietnamese text includes many sentences.
- Output : a summary sentence covers the content of the above text.

II. RELATED WORK

The summary model has been researched and applied quite commonly on English data sets. However, studies on Vietnamese data sets are more limited. Below are a few outstanding studies in building summary models for Vietnamese data sets.

- Nguyen Viet Hanh [1] in 2018 built a summary model that was trained and tested on a Vietnamese data set taken from newspaper pages. The dataset is divided into 1120 articles for the training set and 316 articles for the testing set. The author applies the Sequence-to-Sequence model [2] with 2 encoder layers and 1 decoder layer using LSTM network. In addition, to increase accuracy, the author used the Attention [3] and beam search algorithm [4] to find the best results. The results were evaluated based on the Rouge measure [5]. From the report, it shows that the model worked quite well on the Vietnamese data set.
- Lam Quang Tuong [6] and colleagues also researched an abtractive text summmarization model using the Sequence-to-Sequence model with the RNN layer. The data set used to train and evaluate the model is built

from Vietnamese newspaper pages with more than 31,000 articles. Research using Word2Vec to find the characteristics of words in the Vietnamese language. The results of the model are also very positive and have achieved the goal of solving the problem.

- Ti Hon Nguyen and Thanh Nghi Do [7] introduced a VNText dataset with more than 1 million lines for the task of summarizing text in the Vietnamese language. Then, the authors applied both extractive and abtractive summary models to test this dataset.
- Dat Quoc Nguyen [8] in 2021 introduced the BARTpho model. This is a fine-tuned model from the BART model. BARTpho is a transformer model with an encoder-decoder mechanism. This is the first large language model that works well on Vietnamese data, especially in the field of abtractive text summarization. This model is trained on a fairly large data set and has shown good results on the test data set.
- Long Phan [9] and colleagues in 2022 proposed the ViT5 model - a transformer model based on the encoder-decoder mechanism to help create summary sentences. This model is trained on the Vietnews data set of about 99,000 lines and Wikilingual data set of about 13,000 lines. The results of the model evaluated on the test set produced a quite high number and completely outperformed other large language models on the above two data sets. This model can very well solve the task of summarizing text in an abtractive direction in the Vietnamese language.

They are all good resources for learning about text summarization to apply to this study.

III. DATASET

A. General process.

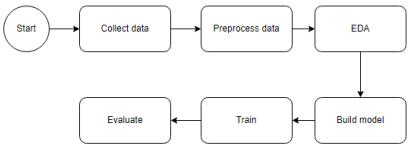

Figure 2: General process.

First of all, we collect and preprocess the data. This stage is quite important to prepare data and create a good quality training source for the learning model.

Next, we analyze and explore (EDA) the entire data set, the goal is to find important information in the data set to establish appropriate inputs for the model.

Finally, there is the process of model building, training and evaluation. At this stage, we will set up the model's layers and based on the process analyzed above to set up hyperparameters for the model. After training, we will use appropriate metrics to evaluate how well the model works.

The dataset we built is a collection of texts, and it is an unstructured data, which is why it is difficult to be approach with models without applying preprocessing methods. So in the next step, we preprocessed texts in corpus. It includes converting all letters of comments to lowercase situation, removing teencode and stopwords, replacing acronyms with complete words and correct spelling mistakes to reduce data dimentions, which can improve computational time. The more cleaning data, the less dimensions and noise.

B. Feature of Vietnamese

In Vietnamese, words have two types: simple words and compound words. A single word consists of only one word, for example the words “di”, “ban”, “toi”. Meanwhile, unlike English, Vietnamese also has compound words created by combining two or more words that have a relationship in meaning. For example, "hoc sinh" is a compound word consisting of two simple words: "hoc" and "sinh". Therefore, when working with Vietnamese, word tokenizing is an important task in the pre-processing step because it will represent words more transparently.

In Vietnamese, a word can represent many meanings, depending on the specific context. For example, the word "di" in the sentence "Ong ay di roi" can mean that the subject has left the current location or it can also mean "chet", depending on the case we will consider the corresponding meaning. In addition, due to regional differences, Vietnamese will have its own slang and the way words are used to express a phenomenon will also be very different.

The Vietnamese language has three main types of words: nouns, verbs and adjectives. Nouns refer to things and phenomena in nature such as personal names and titles, for example "Huy", "toi", "ban", "cho", "meo". Verbs are words that describe the behavior of things and phenomena such as "an", "ngu", "chay". Meanwhile, adjectives refer to the properties of things and phenomena such as "dep", "xau", "met". There are also adverbs to supplement the meaning of the sentence.

Sentences are made up of two components: subject and predicate. The subject describes the subject who will perform the action in the sentence, while the predicate describes the action the subject will perform. Sentences can be simple sentences or compound sentences. Simple sentences include only one subject and predicate, while compound sentences include many subjects and predicates connected by conjunctions. Vietnamese uses the same SVO grammatical structure as English. However, the difference is that in English, adjectives come before nouns, and in Vietnamese, it's the opposite. In English, verbs are divided according to tense, which is not the case in Vietnamese.

C. Collecting and preprocessing dataset

The data set collected from Bao dan tri [10] and vnexpress [11] includes 23,248 lines. Both of them are reliable sources for news in Vietnam. Each line is represented by 4 columns: topic, title, text and summary. Below is the codebook that describes the information and meaning of the columns:

No.	Name	Description	Range
1	topic	descibe a topic of news	16 topic
2	title	header of news	text
3	text	full content of news	text
4	summary	brief content of news	text

Table 1: The codebook of dataset

The data set includes a total of 16 topics: welfare, real estate, technology, unions, education, entertainment, business, philanthropy, law, health, world, sports, jobs, culture, cars and society. Each topic collects about 1500 news. Below is an image of the distribution of the number of articles by topic.

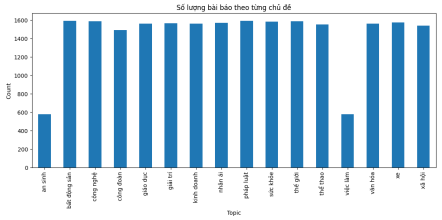

Figure 3: The distribution of topic dataset.

The figure above shows that this dataset is fairly balanced at each label. Therefore, we can expect the model to be able to learn vocabulary evenly across topics without being biased towards a particular topic. Next, once we have collected the data, we move on to preprocessing.

Figure 4: First 5 rows of dataset.

On the entire dataset, we will filter out null rows and then delete these rows, because during the data collection process, we noticed that there are many articles that only have images without text, so after collected, these data lines have null values. In addition, articles are continuously updated, so when crawling data, many articles appear more than once, so we also delete these rows. Then we conducted a survey and found that the majority of articles had more than about 500 words, so we decided to delete lines with a word count of less than 200 words in the text column to ensure that when building the model, we would Do not appear too many zero values. For each data value in the two columns text and summary, we will initially standardize punctuation, specifically semicolons, exclamation marks, question marks, and punctuation marks. three dots will change to a dot. The remaining marks will be deleted. Next, we will remove extra spaces in sentences and special characters such as &$%#. Finally, we remove additional information in parentheses, such as the source information of a certain definition in the sentence. After processing, we only retain 2 columns, text and summary, from the original 4 columns. To summarize, the shape of the dataset after preprocessing is 20792 rows and 2 columns.

After preprocessing, we divide the data set into 3 parts: train, validation and test with a ratio of 8: 1: 1 respectively.

Finally, there is the EDA (Exploratory Data Analyst) process. This is the process that takes place after the preprocessing stage, the purpose of which is to find some useful features of the data. Here, we will probe 4 parameters: number of longest and shortest words in a text sentence, number of longest and shortest words in a summary. Below are the survey results:

- Maximum of word text : 9647
- Maximum of word summary: 140
- Minimum of word text : 48
- Minimum of word summary : 9
- 80% text length < 811
- 80% summary length < 47

IV. BUILDING MODEL

The input to the problem is a long paragraph and our goal is a short paragraph that covers the entire content of the input text. With problems of this type, we can use the Sequence-to-Sequence (Seq2Seq) model to solve. The Seq2Seq model includes an encoder layer to encode the context vector input character string and a decoder layer that will receive the context vector from the encoder, combined with the label character string to create the corresponding output string. However, this model still has the disadvantage of not handling the problem of long-range dependencies well, that is, the model will not work well if the input sentence is too long and due to the sequentiality of the model architecture, it is difficult to run parallel.

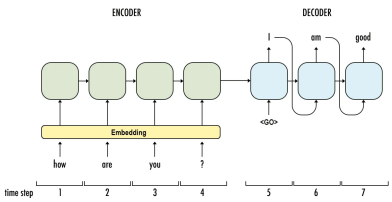

Figure 5: Sequence-to-Sequence Architect.

To solve the problem of the Seq2Seq model, the Transformer architecture was born. The Transformer model is built with the purpose of processing a long piece of text as input. Since its launch until now, this model has achieved great breakthroughs, contributing to the development of the field of natural language processing. Many famous large language models have been born based on Transformer architecture such as BERT, GPT, T5, BART.

In this study, we will use 3 pre-trained models: ViT5, BARTpho and mT5 to run tests on the collected data set. Below are the parameters we chose to train the models:

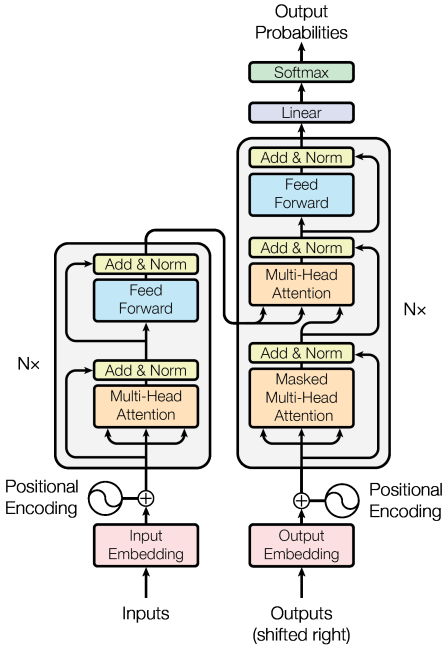

Figure 6: Sequence-to-Sequence Architect.

- epoch = 3
- per_device_train_batch_size = 2
- per_device_test_batch_size = 2
- learning_rate = 1e-4

V. RESULT AND EVALUATION

A. Result

In text generation tasks, there are two popular metrics to evaluate models: BLEU [12] and Rouge. Particularly in the field of text summarization, the Rouge measure is more commonly used. This measurement has two commonly used forms: Rouge-N and Rouge-L.

Rouge-N: based on the n-gram overlap between the summary sentence generated by the model (prediction) and the original summary sentence (reference) to evaluate the results. The two commonly used forms are Rouge-1 and Rouge-2. There are also 3 evaluation cases: precision, recall and f1. Precision is based on the number of overlapping n-grams between the prediction sentence and the reference sentence divided by the total number of n-grams of the prediction. Recall is similar to precision but we will divide it by the total number of n-grams of the reference. F1 is a combination of precision and recall. Below is the formula for the Rouge-N measure:

$$Rouge\text{-}N(recall) = \frac{ngram\,pred. \bigcap ngram\,ref.}{\#ngram\,in\,ref.}$$

$$Rouge\text{-}N(Precision) = \frac{ngram\,pred. \bigcap ngram\,ref.}{\#ngram\,in\,ref.}$$

$$Rouge\text{-}N(F1) = 2 * \frac{recall * precision}{recall + precision}$$

Rouge-L: based on the overlap of the longest sequence (LCS - Longest Common Subsequence) between the prediction sentence and the reference sentence to calculate the results. This measure also has 3 cases precision, recall and f1. For example, we have the sentence:

- Reference: "The quick brown fox jumps over the lazy dog."
- Prediction: "A quick brown fox jumps over a lazy dog."

LCS in the above case would be "quick brown fox jumps over lazy dog". The calculation formula for each case is:

$$Rouge\text{-}L(recall) = \frac{LCS(pred., ref.)}{\#words\,in\,ref.}$$

$$Rouge\text{-}L(precision) = \frac{LCS(pred., ref.)}{\#words\,in\,pred.}$$

$$Rouge\text{-}L(F1) = 2 * \frac{recall * precision}{recall + precision}$$

To evaluate the model results, we use 3 measures Rouge-1, Rouge-2 and Rouge-L in case f1. Below is the table of results of the models:

Model	Rouge-1	Rouge-2	Rouge-L
ViT5	0.587	0.282	0.389
BARTpho	0.575	0.258	0.370
mT5	0.507	0.193	0.336

Table 2: The result of models

As we have seen, all three models produce quite good results. Among them, ViT5 is the model with the highest index. This is because this model is trained on the same type of problem (text summarization) and the same data domain as our study (the data source is collected from Vietnamese newspapers). On the contrary, mT5 is a multilingual model, meaning it does not focus on any specific language, so the results of this model are lower than the other two models that focus on the Vietnamese language.

B. Error Analyst

First, we calculate the overlap of words between the summary sentence – summary column and the article content – text column. We calculated based on 2-grams and found that the majority of data lines had quite high results, which showed that the summary statements were basically correct. However, there are still cases where the overlap is low and even the result is zero, which makes it difficult for the model to generate a high-quality summary for the original content.

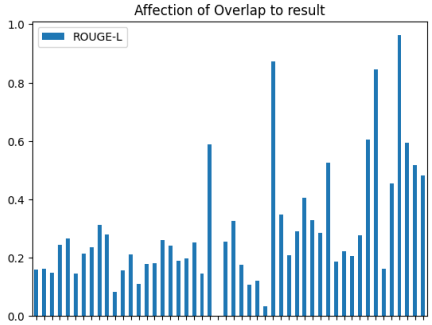

Figure 7: Affection of overlap to result

Next, to verify our thinking, we measured the influence of overlap on the results. The figure 6 shows the correlation

between overlap and model results. As we can see, the higher the overlap, the better the model performance. Therefore, to ensure model quality, what needs to be done is to improve the quality of summary sentences in lines with low overlap and remove noisy data lines, for example lines whose content is completely in English.

VI. CONCLUSION AND DEVELOPMENT

Text summarization is a necessary task in the current period when the amount of information expressed in words is increasingly created. There have been many studies on building automatic summary models, but most of them are still mainly in English. In this study, we focus on building an effective summary model in the Vietnamese language. We have built a Vietnamese data set of 23,248 articles collected on Vietnamese news websites. In addition, we also use 3 models to run experiments and evaluate on this dataset.

However, due to time and infrastructure constraints, we still only train models with quite small parameter sets. This may not demonstrate the model's reliability. In addition, the data set still has noise that affects the model. In the future, we will focus on improving data quality and running the model with a large enough parameter set to provide more accurate results. Furthermore, we plan to build a Seq2Seq model running on this dataset to compare the results with Transformer models, clarifying the advantages and disadvantages of the models.

REFERENCES

[1] Nguyen Viet Hanh, *"Nghien cuu tom tat van ban tu dong va ung dung"*, 2018.

[2] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, *"Sequence to Sequence Learning with Neural Networks"*, 2014.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, *"Attention Is All You Need"*, 2017.

[4] Markus Freitag, Yaser Al-Onaizan, *"Beam Search Strategies for Neural Machine Translation"*, 2017.

[5] Chin-Yew Lin, *"ROUGE: A Package for Automatic Evaluation of Summaries"*, Information Sciences, Institute University of Southern California.

[6] Lam Quang Tuong , Pham The Phi, Do Duc Hao, *"Tom tat van ban tieng Viet tu dong voi mo hinh Sequence-to-Sequence"*, 2017.

[7] Ti Hon Nguyen, Thanh Nghi Do, *"Text Summarization on Large-scale Vietnamese Datasets"*, 2022.

[8] Nguyen Luong Tran, Duong Minh Le, Dat Quoc Nguyen, *"BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese"*, 2021.

[9] Long Phan, Hieu Tran, Hieu Nguyen, Trieu H. Trinh, *"ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation"*, 2022.

[10] Bao dan tri, *https://dantri.com.vn/*.

[11] vnexpress, *https://vnexpress.net/*.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, *"BLEU: a Method for Automatic Evaluation of Machine Translation"*, 20.