An Effective Deep Learning Approach for Extractive Text Summarization

Minh-Tuan Luu
PhD. Student, School of Information and Communication Technology,
Hanoi University of Science and Technology, No.1 Dai Co Viet street, Hai Ba Trung district, Ha noi, Vietnam
Lecturer, School of Information Technology and Digital Economics, National Economics University,
No.207 Giai Phong street, Hai Ba Trung district, Hanoi, Vietnam
tuanlm@neu.edu.vn

Thanh-Huong Le*
Lecturer, School of Information and Communication Technology, Hanoi University of Science and Technology,
No.1 Dai Co Viet street, Hai Ba Trung district, Hanoi, Vietnam
huonglt@soict.hust.edu.vn (Corresponding author)

Minh-Tan Hoang
Student, School of Information and Communication Technology, Hanoi University of Science and Technology,
No.1 Dai Co Viet street, Hai Ba Trung district, Hanoi, Vietnam
tan.hm1211@gmail.com

Abstract - Nowadays, most research on extractive text summarization uses deep learning approaches as they provide better performances than the others. However, a difficulty in these approaches is the shortage of a large dataset for training summarization systems. To deal with this problem, we take advantage of contextualized word embeddings from pre-trained BERT models to produce sentence embedding vectors. These vectors are then used as the input of a Multi-Layer Perceptron classifier for sentence selection. The outputs of the Multi-Layer Perceptron classifier are processed by a Maximal Marginal Relevance algorithm to remove redundant sentences. Finally, the selected sentences are rearranged using information about sentence position in the original document to create a summary. Our proposed system is evaluated by using both English and Vietnamese datasets. Experimental results show that our system achieves promising results comparing to existing researches in this field.

Keywords: Extractive summarization; deep learning; BERT; RoBERTa; PhoBERT, sentence position.

1. Introduction

Text summarization is an important task in natural language processing. It can be divided into two categories: extractive summarization and abstractive summarization. Extractive summarization concentrates on selecting important sentences from the input text to put in the summary. Meanwhile, abstractive summarization aims to generate a concise summary for the input text by paraphrasing the main content of the original document. In this paper, we focus on extractive summarization for English and Vietnamese text.

Recently, neural network-based approaches have been applied for extractive summarization and provide better results than other approaches. In neural text summarization, a good representation of the input document is an important factor for evaluating sentences. Several machine learning and deep learning approaches have been investigated to solve this problem. However, these techniques often require a large training dataset and suffer from a high training cost. Some approaches solve this problem by using pre-trained word embeddings such as word2vec [Mikolov *et al.* (2013)] and Glove [Pennington *et al.* (2014)]. These word embedding vectors are context-independent, which may result in incorrect meaning in some cases.

Researches on generating context-based representing models show that using pre-trained sentence embeddings [Conneau *et al.* (2017)] provides better performance than using word embeddings in natural language processing tasks. Cera *et al.* (2018) propose two pre-trained Universal Sentence Encoders (USEs) basing on Transformer (USE_T) [Vaswani *et al.* (2017)] and Deep Averaging Network (USE_D) [Lyyer *et al.* (2015)] for the English language. Yang *et al.* (2019) expand the USE pre-trained models for 16 languages. These models were pre-trained on a large unlabeled text to generate sentence embeddings. Devlin *et al.* (2019) propose Bidirectional Encoder Representations from Transformers (BERT), a new language representation model that is designed to pre-train deep bidirectional representations from the unlabeled text. BERT is the best semantic embeddings until now, as it has been used in many state-of-the-art models in natural language processing tasks.

This paper aims at developing an extractive summarization system that can be applied for both English and Vietnamese language. The summarization system is constructed as a classification model, in which sentences in the summary have the label 1, and 0 otherwise. We integrate two pre-trained BERT models to generate document embeddings, including RoBERTa [Liu *et al.* (2019)] for English and PhoBERT [Nguyen and Nguyen (2020)] for Vietnamese. The classification task is performed by a Multi-Layer Perceptron (MLP). Maximal Marginal Relevance (MMR) is used to remove sentences with overlapping information. Finally, the summary is generated by rearranging output sentences of MMR, using information about sentence positions in the original document. Our proposed summarization model is evaluated with both English and Vietnamese languages, using CNN and Baomoi datasets, respectively. Experimental results show that our system achieves better results compared to existing researches using the same dataset.

Our main contributions can be summarized as follows:

- Applying the pre-trained BERT to represent the sentence embedding vectors, in order to have a better understanding of the input text;
- Proposing a Multi-Layer Perceptron to classifying sentences to be included in the summary;
- Integrating a Maximal Marginal Relevance to remove redundant information;
- Experimenting with both English and Vietnamese datasets to prove the generality of the proposed method.

The rest of this paper is organized as follows. Section 2 discusses related works in extractive text summarization. Our proposed text summarization is introduced in Section 3. Our experiments are described in Section 4. Finally, Section 5 concludes the paper and proposes future works for our research.

2. Related works

Researches on extractive summarization can be divided into three main categories: (i) unsupervised techniques; (ii) traditional machine learning-based techniques; and (iii) deep learning-based techniques. Unsupervised techniques are dominated early researches on extractive summarization, as text summarization corpus are rare at that time. The unsupervised techniques rely on selecting sentences with top-ranking scores. The scores are computed based on surface features such as time sequence [Wasson (1998)], term frequency [Luhn (1958)], TF*IDF [Erkan *et al.* (2004)], sentence length [Cao *et al.* (2015a)], sentence position [Ren *et al.* (2017)]. Graph-based methods are also widely used to score sentences [Mihalcea *et al.* (2004); Choi *et al.* (2011)]. These methods represent each input document as a graph, in which each sentence is represented as a vertex; two relevant sentences are connected by an edge. The sentence importance is evaluated by using a ranking algorithm on this graph. Maximal Marginal Relevance method [Carbonell and Goldstein (1998)] is proposed to remove redundant sentences from the summary.

Machine learning techniques have been used to receive a better evaluation of sentence importance when some text summarization datasets are available (e.g., [Kupiec *et al.* (1995); Wong *et al.* (2008)]). Kupiec *et al.* (1995) consider extractive text summarization as a text classification task, in which the sentences appearing in the summary will have the label **1**, and **0** otherwise. A Naive Bayes classifier is used to train their summarization system. Wong *et al.* (2008) investigate four types of sentence features (i.e., surface, content, relevance, and event features) and incorporate them into supervised and semi-supervised approaches for training the text summarization system. Probabilistic Support Vector Machine is used in their approach for supervised learning, while the co-training of Probabilistic Support Vector Machine and Naïve Bayesian Classifier is used for semi-supervised learning.

Recently, deep learning techniques have been successfully applied to text summarization and get better results comparing to traditional approaches. Zhang *et al.* (2016) extract salient sentences for the summary by using Convolutional Neural Networks (CNN). Nallapati *et al.* (2017) treat extractive summarization as a sequential labeling task. Sentences of the input document are encoded and then classified into two classes: selected or not selected. These systems compute selection probability for each sentence, then generate a summary based on these probabilities until reaching the summary limit. Zhou *et al.* (2018)develop an end-to-end neural network for text summarization by jointly learning to score and selecting sentences. To optimize the ROUGE evaluation metric, several approaches train their neural summarization models by using a reinforcement learning objective (e.g., [Narayan *et al.* (2018); Wu and Hu (2018)]).

Wu and Hu (2018) propose a neural coherence model to capture the cross-sentence semantic and syntactic coherence patterns, using a reinforcement learning mechanism. The system's reward is computed by evaluating the system output using the Rouge measures. Zhang *et al.* (2018) propose a latent variable extractive summarization model that uses directly human summarization and a sentence compression model to generate the summary. In this approach, sentences are considered as latent variables. Sentences with activated variables are used to generate the summary. This technique solves the problem of depending on sentence-level labels, which is often used in extractive summarization systems.

Jadhav and Rajan (2018) develop a neural sequence-to-sequence model for extractive summarization that models the interaction of keywords and salient sentences using a two-level pointer network-based architecture. The text summarization system in [Kamal *et al.* (2018)] uses a hierarchical structured self-attention mechanism to capture the hierarchical structure of the document and to create the sentence and document embeddings. The attention mechanism provides an extra source of information to guide the summary extraction. The model computes the probabilities of the sentence-summary membership basing on several features such as information content, salience, novelty, and positional representation.

Zhang *et al.* (2019) propose a document encoding model named HIerachical Bidirectional Encoder Representations from Transformers (HIBERT) and pre-train it using unlabeled data. This model provides promising results when applying it to the text summarization task.

Most of the researches on Vietnamese text summarization relies on sentence features such as TF*IDF, sentence position, etc… to compute the score of the sentences (e.g., [Ha *et al.* (2005); Dinh and Nguyen (2012)]). Discourse structure has been used in [Nguyen and Le (2008)] to generate text summarization. Some researches use machine learning algorithms such as Nguyen *et al.* (2005), Nguyen *et al.* (2012), Nguyen *et al.* (2012). Salient sentences are extracted by using Support Vector Machine [Nguyen *et al.* (2005)], by applying genetic algorithm [Nguyen *et al.* (2012)], and taking advantage of the semi-supervised algorithm [Nguyen *et al.* (2012)]. Lam *et al.* (2017) construct a Sequence to Sequence with Attention model and a beam search to generate the final summary. Words of the input document are embedded as word vectors before using them as input of the text summarization model. The model is trained by a self-collected news dataset with 31,429 articles, in which the abstract of each news is used as its summary.

Most of the above approaches have not been applied in an efficient way to represent the semantic structure of the input document, which leads to the redundancy of information in the summary. In this paper, we propose a text summarization model to deal with the problem mentioned above. Our model is represented in the next section.

3. Our proposed model

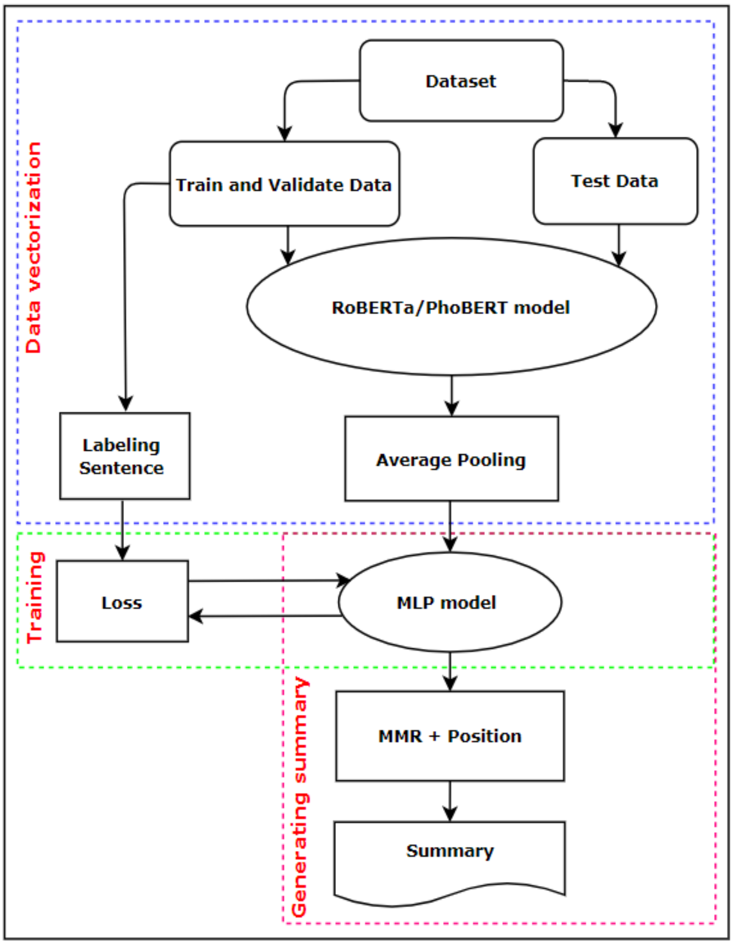

Fig. 1. Our proposed extractive single-document summarization system

Our proposed extractive single-document summarization system includes three main modules: (i) Data vectorization, (ii) Training, and (iii) Generating summary. These modules are presented in Fig. 1 above.

3.1. Data vectorization

This module encodes sentences of the input document as embedding vectors using an optimized pre-trained BERT model. The RoBERTa model [Liu *et al.* (2019)] is used in our experiment for English text, and the PhoBERT model [Nguyen and Nguyen (2020)] is used for Vietnamese text. First, the BERT model creates an index vector for each token from the input sentence. These indexed vectors are combined to obtain token embeddings of the respective sentences. The token embeddings of each sentence are processed by the Average Pooling operation to create the sentence embedding, which is used as the input for the MLP model. The following section will introduce RoBERTa and PhoBERT models in detail.

3.1.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) [Devlin *et al.* (2019)] is a new language representation model, which is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers. The BERT model is based on Transformer architecture [Vaswani *et al.* (2017)]. This architecture includes L layers, with each block containing A self-attention heads and H hidden dimensions. Sentences are used as the input of the model. The BERT model was implemented with two steps, pre-training and fine-tuning. In the pre-training step, the model was trained on an unlabeled dataset with 16GB of uncompressed text in total. The unlabeled text was taken from BookCorpus [Zhu *et al.* (2015)] and English Wikipedia. In the fine-tuning step, the model was initiated with pre-trained parameters and fine-tuned parameters using labeled data from downstream tasks. The BERT implementation includes three models: BERT$_{\mathrm{BASE}}$ (12-layer, 768-hidden, 12-heads, 110M parameters), BERT$_{\mathrm{LARGE}}$ (24-layer, 1024-hidden, 16-heads, 340M parameters), and BERT$_{\mathrm{BASE}}$ multilingual models.

3.1.2. RoBERTa

RoBERTa is proposed by Liu *et al.* (2019). It is developed to optimize the pre-trained BERT model [Devlin *et al.* (2019)]. The RoBERTa model uses the same pre-trained BERT model architecture. The main differences of RoBERTa comparing to BERT are:

- The RoBERTa model was trained on a bigger dataset of 160GB documents. This dataset is a combination of five datasets: BOOKCORPUS [Zhu *et al.* (2015)], English Wikipedia, CC-News [Nagel (2015)], OpenWebtext [Gokaslan *et al.* (2019)], and Stories [Trinh and Le (2018)]). The RoBERTa took a longer time than the BERT to train the model with 500K steps.
- The training method of RoBERTa is different from BERT. The RoBERTa model eliminated the Next Sentence Prediction (NSP) task from its training process. Instead, it was trained using Dynamic Masking so that the masked tokens would be generated when a sentence was included in the model. The model was trained with a larger batch size so it prevented better noise during the training process. The maximum length of a sentence vector is 512. The RoBERTa model was trained using the BERT$_{\mathrm{LARGE}}$ model (L = 24, H = 1024, A = 16, 355M parameters).

In our proposed model, we use the RoBERTa model that has the sentence vector with the maximum length of 256, and the batch size of 256. During the training process, we freeze the RoBERTa model and fine-tune it on the CNN dataset.

3.1.3. PhoBERT model

Nguyen and Nguyen (2020) developed two pre-trained language models, PhoBERT$_{\mathrm{BASE}}$ and PhoBERT$_{\mathrm{LARGE}}$ for Vietnamese, using BERT$_{\mathrm{BASE}}$ and BERT$_{\mathrm{LARGE}}$ architectures, respectively. PhoBERT model was trained by the same method as RoBERTa [Liu *et al.* (2019)], using a dataset of 20GB uncompressed document, which is a combination of Vietnamese Wikipedia corpus (~1GB) and Vietnamese news corpus (~19GB) (https://github.com/binhvq/news-corpus). These Vietnamese datasets were preprocessed by word segmentation before tokenizing using the BPE algorithm [Sennrich *et al.* (2016)]. The maximum length of a sentence vector after word segmentation was 256, less than that of the RoBERTa model. Since PhoBERT models were trained on the Vietnamese dataset, they provided good results in many Vietnamese natural processing tasks.

In our proposed model, we use PhoBERT$_{\mathrm{BASE}}$ model that has also the sentence vector with a maximum length of 256 and the batch size of 256. During the training process, we freeze the PhoBERT$_{\mathrm{BASE}}$ model and fine-tune it on the Baomoi dataset, too.

3.2. Training

The purpose of the training task is to learn a classifier that computes the selection probability of input sentences to be included in the summary. We perform this task by implementing a Multi-Layer Perceptron (MLP) network model using the Back Propagation algorithm. Our proposed MLP model (Fig. 2) consists of a 768-dimension input layer to adapt with the output dimension of RoBERTa model and PhoBERT model, a hidden layer with 256 neurons using the ReLU activation function, an output layer with 2 neurons using the **softmax** activation function. The model was implemented and trained by the AdamW optimizer [Loshchilov and Hutter (2019)].

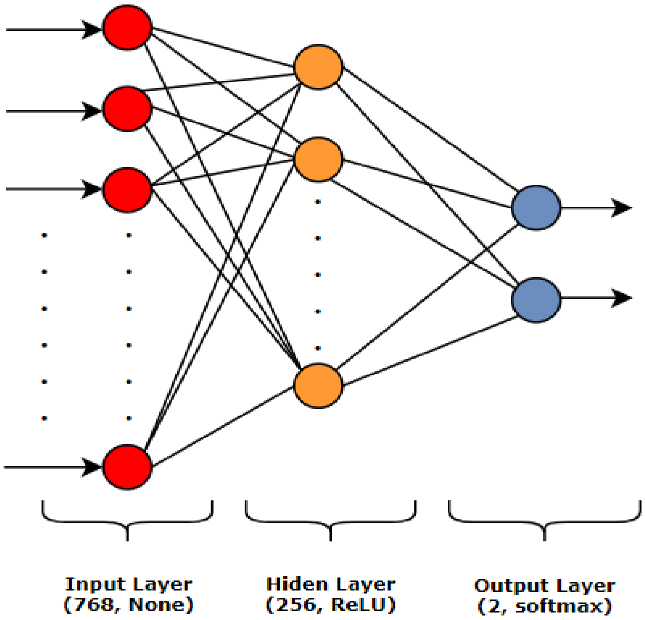

Fig. 2. The architecture of our proposed MLP network model

3.3. Generating summary

Sentences from the input document are selected to include in the summary by the order of descending probability until reaching the summary length. To prevent overlap content among sentences in the summary, we apply the MMR method [Carbonell and Goldstein (1998)] to measure the similarity among sentences.

The MMR method was originally proposed to solve Information Retrieval (IR) problem to measure the relevance between the user query Q and sentences in the document. The MMR is calculated by the formula:

$$MMR \overset{def}{=} Arg \max_{D_i \in C \setminus S} \left[\lambda \left(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right) \right] \tag{1}$$

In which:

- C is the set of sentences from the input documents;
- S is the set of existing sentences in the summary;
- Sim_1 is the *similarity* between the considering sentence D_i and the query Q;
- Sim_2 is the *similarity* between the considering sentence D_i and the existing sentences in the summary D_j (Sim_2 can be equal to Sim_1);
- λ is a *parameter* ($\lambda \in [0;1]$).

The parameter value λ is chosen depending on each problem. If it is necessary to return information around the query, the parameter λ is adjusted with a smaller value. If the result needs to be diverse, the parameter λ is adjusted with a greater value. A high MMR means the considered item is both relevant to the query and contains minimal similarity to previously selected items.

To apply the MMR method to the task of document summarization, we redefine the formula to calculate the MMR measure as follows:

$$MMR \overset{def}{=} Arg \max_{D_i \in C \setminus \{S, Q\}} \left[\lambda \left(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right) \right] \tag{2}$$

where:

- C is the set of candidate sentences for the summary;
- Q is a sentence in the set C that is best described the main idea of the input document;
- S is the set of the sentences that are already included in the summary;
- Sim_1, Sim_2 are the similarities between the two sentences u and v, being calculated by the formula:

$$Sim_1(u,v) = Sim_2(u,v) = \frac{\sum_{w \in v} tf_{w,u} tf_{w,v} (idf_w)^2}{\sqrt{\sum_{w \in u} (tf_{w,u} idf_w)^2}} \quad (3)$$

where $tf_{w,u}$ is the term frequency of the word w in the sentence u; idf_w is the importance of the word w; and λ is the chosen parameter.

Applying the MMR method to the task of document summarization. For documents in the form of news, sentences at the beginning of a document often contain more important information than the others. To take advantage of sentence positions and selection probabilities, we integrate this information to the MMR measure by replacing $Sim_1(D_i, Q)$ by *probability*position* in the formula (2) as follows:

$$MMR \overset{def}{=} Arg \max_{D_i \in C \setminus \{S,Q\}} \left[\lambda \left(probability * position - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right) \right] \quad (4)$$

In which:

- *position* is the sentence position;
- *probability* is the selection probability of the sentence.
- The main point of applying the MMR method is to eliminate redundant information in the summary. To do that, three steps needed to be carried out are:
- Determine the main topics of the input documents;
- Find sentences relevant to the main topics;
- Eliminate redundant sentences whose similarity with existing sentences in the summary is larger than a certain threshold.

4. Experiment and Evaluation

4.1. Datasets

We experimented with our proposed system using two datasets: the CNN dataset for English and the Baomoi dataset for Vietnamese. The purpose of using the CNN dataset is to compare results with state-of-the-art works in extractive summarization. Experiments with the Baomoi dataset is to evaluate our proposed system with another language (Vietnamese), aiming at proving the generality of our system.

The CNN/Daily Mail dataset [Hermann *et al.* (2015)] includes 312,085 articles with 92,579 articles from the CNN dataset and 219,506 articles from the Daily Mail. The summary of each article is the highlight sentences written by the article's author. We used the method of Hermann *et al.* (2015) to divide the CNN dataset into the training, validation, and testing datasets, which include 90,266; 1,220; and 1,093 documents, respectively. Since each summary in the CNN dataset contains 3 sentences in average, we also chose 3 sentences to include the summary that was generated by the system. Statistic information about the CNN/Daily Mail dataset is shown in Table 1 below.

Table 1. The statistics of the CNN/Daily Mail dataset

	CNN			**Daily Mail**		
	Train	***Valid***	***Test***	***Train***	***Valid***	***Test***
# months	95	1	1	56	1	1
# documents	90,266	1,220	1,093	196,961	12,148	10,397
#queries	380,298	3,924	3,198	879,450	64,835	53,182
Avg # tokens	762	763	716	813	774	780
Vocab size	118,497			208,045		

Since there is no available Vietnamese text summarization corpus that is shared among the research community, we use a corpus named ‘Baomoi’. The corpus was created by gathering articles from a Vietnamese online newspaper (http://baomoi.com). Each article consists of three parts: headline, abstract, and article. The abstract is more likely the key information of the article than its summary. However, since we cannot find any better source, the Baomoi dataset is still our best choice to be used as the summarization corpus at the moment. We take the article part and the abstract part to serve as the original document and its summary. The average length of the original document and its summary are 503 words and 45 words, respectively. The final dataset consists of 1,000,847 news articles, in which 900,847 samples are used for training, 50,000 samples for validation, and 50,000 ones for testing. The abstract section of a Baomoi article has approximately 2 sentences in average. These sections were used for training the system and evaluating the system’s accuracy.

To create the training dataset, we assigned a sentence label with **1** if that sentence was in the summary.

4.2. Preprocessing

Firstly, we extracted the content and the abstract of each article and segmented these texts into sentences. **StanfordNLP** and **VnCoreNLP** libraries were used to do the segmentation task for English and Vietnamese text, respectively. To label these sentences, we compared them with sentences in the abstract basing on the maximum total of the Rouge-2 and Rouge-L measures using the **Rouge-score 0.0.4** library. Next, these sentences were tokenized to create index vectors for these tokens. These index vectors were used as the input for the RoBERTa/PhoBERT model to obtain the token embedding vectors. Finally, the token embeddings of each sentence were processed using the avgPooling1d function in the **PyTorch** library to generate a 768-dimension sentence embedding vector which would be used as the input for the MLP model.

4.3. Experimental design

First of all, we implemented some basic methods that were good at extractive single-document summarization for both the CNN and Baomoi datasets. We used the rouge-score 0.0.4 library to evaluate the summary’s quality of the models. Table 2 presents the experimental results on both CNN and Baomoi datasets that we implemented.

Table 2. Experimental results of some basic methods. Marking with ‘*’ denotes the systems being reimplemented by us

Methods	**CNN**			**Baomoi**		
	Rouge-1	***Rouge-2***	***Rouge-L***	***Rouge-1***	***Rouge-2***	***Rouge-L***
LexRank [Erkan and Radev (2004)]*	22.9	6.6	17.2	38.5	17.0	28.9
TextRank [Mihalcea and Tarau (2004)]*	26.0	7.3	19.2	44.7	19.2	32.9
LEAD [Wasson (1998)]*	29.0	10.7	19.3	46.5	20.3	30.8

Next, we implemented summarization models using some modern pre-trained models, including the USE_T model, the RoBERTa model for English, and the PhoBERT model for Vietnamese. The USE_T model was chosen because it had provided better results than the others in [Cera *et al.* (2018); Yang *et al.* (2019)]. We carried out 4 model scenarios, in which scenarios 1 and 2 were only for the CNN dataset since the USE_T model does not support Vietnamese. Scenarios 3 and 4 were experimented on both the CNN and Baomoi datasets to choose the most efficient model for our proposed summarization system. We used the TensorFlow library to inherit the pre-trained USE_T model that was stored on **Tensorflow-hub**, the **Transformers** library to inherit the RoBERTa model, the **PhoBERT with Transformers** library to inherit the PhoBERT model, and the PyTorch library to develop Multi-Layer Perceptron classification model. In our experiments, our models were trained on Google Colab with the following configuration: GPU V100, 12GB RAM. The MLP model was trained with an initial learning rate of 2.10^{-3} by 6 epochs for the CNN dataset, 7 epochs for the Baomoi dataset. After each epoch, the learning rate would be automatically reduced by 10% using the Scheduling mechanism in the PyTorch library until the last epoch. The experimental model scenarios are shown below.

- Scenario 1 (USE_T+MLP). The system used the USE_T model in combination with the MLP network to train the model for calculating the sentential selection probability.
- Scenario 2 (USE_T+MLP+MMR+Position). The system used the model in scenario 1 in combination with the MMR measure and the sentence position to select sentences putting on the summary.
- Scenario 3 (RoBERTa/PhoBERT+MLP). The system used the RoBERTa model (for CNN dataset) and PhoBERT model (for Baomoi dataset) in combination with the MLP network to train the model for calculating the sentential selection probability.

- Scenario 4 (RoBERTa/PhoBERT+MLP+MMR+Position). The system used the model in scenario 3 combining with the MMR measure and the sentence position to select sentences putting on the summary.

With scenarios 1 and 2, our model was trained on the CNN dataset by 6 epochs with a batch size of 50 and a training time of approximately 6 hours. With scenarios 3 and 4, our model was trained by 6 epochs with a batch size of 256 and a training time of approximately 8 hours for the CNN dataset. For the Baomoi dataset, our system was trained by 7 epochs with a batch size of 256 and a training time of approximately 48 hours. The experimental results are shown in Table 3 below.

Table 3. Experimental results with the MODEL scenarios. Marking with '-' denote that we did not reimplement on the corresponding dataset

Scenarios	**CNN**			**Baomoi**		
	Rouge-1	***Rouge-2***	***Rouge-L***	***Rouge-1***	***Rouge-2***	***Rouge-L***
Scenario 1	28.9	10.3	19.3	-	-	-
Scenario 2	30.1	11.5	20.1	-	-	-
Scenario 3	31.36	11.69	28.22	52.509	24.695	37.794
Scenario 4	32.18	12.31	28.87	52.511	24.696	37.796

The experimental results show that, although the text summarization system using the USE_T model and the MLP model in scenario 1 did not eliminate the redundant information, it still gave better results than some basic methods such as LexRank, TextRank, and LEAD on the same CNN dataset (in Table 2). The model that combines the MMR measure and the sentence position to eliminate the redundant information in scenario 2 gave better results than the model in scenario 1. Although the model in scenario 3 did not eliminate the redundant information yet, it still provided much better results than the models in scenarios 1 and 2. These results proved that our proposed summarization system using RoBERTa/PhoBERT model was more effective than the one using the USE_T model on the same CNN dataset. Besides, the model in scenario 3 gave better results than some basic methods on the same Baomoi dataset. Our model in scenario 4, which had removed the redundant information from the summary, clearly provided better results than the one in scenario 3 on both CNN and Baomoi datasets. Comparing with the model's results of scenario 3, the F1-scores of R-1, R-2, and R-L of the model in scenario 4 increased 0.82%, 0.62%, and 0.65% on the CNN dataset; and 0.002%, 0.001%, and 0.002% on the Baomoi dataset, respectively. On the Baomoi dataset, the experimental results of the model in scenario 4 increased a few percentage amounts compared to that in scenario 3 since the summary length is small (two sentences). However, these results showed that the model in scenario 4 is the best. Because of that, we selected this model to develop our proposed summarization system. Table 4 shows an example of the system output with our best model on the CNN dataset.

Table 4. An example of the CNN dataset

Human
Beloved children's performer Lois Lilienstein has died . She was a member of CBC and Nickelodeon TV stars Sharon , Lois and Bram . CNN independently confirmed with Sharon and Bram's manager that Lilienstein passed away at 78 of a rare cancer.
Our system
Lois Lilienstein, co-star of "Sharon, Lois & Bram's Elephant Show" - the Canadian preschool show that ran on Nickelodeon during the early 1990s - has died, aged 78. Her son, David Lilienstein, told CBC News that his mother died in Toronto on Wednesday night from a rare form of cancer first diagnosed last October. "She knew it was happening, she was at peace with it, and she died very peacefully and not in pain," he told the Canadian broadcaster.

Table 5 shows an example of the system output with our best model on the Baomoi dataset.

Table 5. An example of the Baomoi dataset

Human

Bước sang ngày làm việc thứ hai, vòng đàm phán hạt nhân giữa Iran và Nhóm P 5+1 (gồm năm nước uỷ viên thường trực Hội đồng Bảo an Liên hợp quốc là Nga, Mỹ, Trung Quốc, Anh và Pháp với Đức) tại Vienne (Áo) đã nhận được nhiều đánh giá tích cực, trái ngược với những ý kiến không mấy lạc quan đưa ra trước đó.

Our system

Theo thoả thuận tạm thời ký kết hồi tháng 11/2013, Iran đồng ý giảm hoặc đóng băng một số hoạt động hạt nhân trong sáu tháng để đổi lại việc phương Tây giảm bớt trừng phạt hiện nay và không áp đặt thêm trừng phạt mới. Đàm phán giữa Iran với Nhóm P 5+1 từ ngày 18-20/2 tại Vienne là nỗ lực của sáu cường quốc trong việc tìm kiếm một thoả thuận toàn diện cuối cùng liên quan tới chương trình hạt nhân gây tranh cãi của Tehran trước thời điểm thoả thuận sơ bộ hết hiệu lực vào tháng Bảy tới.

The experimental results on both CNN and Baomoi datasets have shown that our proposed summarization system provides promising results for both Vietnamese and English text summarization tasks.

4.4. Evaluation and Discussion

The experimental results of our proposed model and some other models are presented in Table 6.

Table 6. Comparison and evaluation results of the methods. Marking with '*' denotes the systems being reimplemented by us, marking with '-' denotes that we did not reimplement on the corresponding dataset

Methods	**CNN**			**Baomoi**		
	Rouge-1	***Rouge-2***	***Rouge-L***	***Rouge-1***	***Rouge-2***	***Rouge-L***
LexRank [Erkan and Radev (2004)]*	22.9	6.6	17.2	38.5	17.0	28.9
TextRank [Mihalcea and Tarau (2004)]*	26.0	7.3	19.2	44.7	19.2	32.9
LEAD [Wasson (1998)]*	29.0	10.7	19.3	46.5	20.3	30.8
Cheng and Lapata (2016) [Narayan *et al.* (2018)]	28.4	10.0	25.0	-	-	-
REFRESH [Narayan *et al.* (2018)]	30.4	11.7	26.9	-	-	-
USE_T+MLP (our)	28.9	10.3	19.3	-	-	-
USE_T+MLP+MMR+Position (our)	30.1	11.5	20.1	-	-	-
RoBERTa/PhoBERT+MLP (our)	31.36	11.69	28.22	52.509	24.695	37.794
RoBERTa/PhoBERT+MLP+MMR+ Position (our model)	**32.18**	**12.31**	**28.87**	**52.511**	**24.696**	**37.796**

Table 6 shows the sentence position and the MMR probability play an important role in the text summarization systems. The results in Table 6 also show that our proposed document summarization system has given significantly better results than the systems that we have experimented with and other modern systems that were published on two CNN and Baomoi datasets, respectively. These results prove that the text summarization system using the optimized pre-trained BERT models, MLP, MMR, and the sentence position has achieved good efficiency for extractive single-document summarization problems for both English and Vietnamese languages.

5. Conclusion and Future Work

In this paper, we have proposed an efficient extractive text summarization system using the RoBERTa model [Liu *et al.* (2019)] for English and the PhoBERT model [Nguyen and Nguyen (2020)] for Vietnamese; the MLP model for evaluating sentence selection; the MMR measure to eliminate redundant information and to generate the document's summary. The experimental results on both CNN and Baomoi datasets shown that our proposed model is significantly better than other modern systems. These results demonstrate that our system is efficient for both English and Vietnamese languages. In the future, we will investigate other modern models for capturing the sentential meaning in a document such as Generative Pre-Training model (GPT) [Radford *et al.* (2018)] to continue improving the quality of the system's summary.

Acknowledgments

This work is supported by Hanoi University of Science and Technology (HUST) under grant project number T2020-PC-208.

References

[1] Cao, Ziqiang; Wei, Furu; Dong, Li; Li, Sujian; Zhou, Ming (2015a): Ranking with recursive neural networks and its application to multi-document summarization. In AAAI, pages 2153–2159.

[2] Cao, Ziqiang; Wei, Furu; Li, Sujian; Li, Wenjie; Zhou, Ming; Houfeng, WANG (2015b): Learning summary prior representation for extractive summarization. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), volume 2, pages 829–833.

[3] Carbonell, Jaime; Goldstein, Jade (1998): The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. Research and Development in Information Retrieval.

[4] Cera, Daniel; Yanga, Yinfei; Kong, Sheng-yi; Hua, Nan; Limtiaco, Nicole (2018): Universal Sentence Encoder.

[5] Cheng, Jianpeng; Lapata, Mirella (2016): Neural summarization by extracting sentences and words. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 484–494, Berlin, Germany. Association for Computational Linguistics.

[6] Choi, Yong Suk (2011): Tree pattern expression for extracting information from syntactically parsed text corpora," In: Data Mining and Knowledge Discovery 1–21.

[7] Conneau, Alexis; Kiela, Douwe; Schwenk, Holger; Barrault, Loic; Bordes, Antoine (2017): Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.

[8] Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K: (2019): BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.

[9] Dinh, Quang-Truong; Nguyen, Quang-Dung (2012): Mot giai phap tom tat van ban tieng Viet tu dong. Hoi thao quoc gia lan thu XV: Mot so van de chon loc cua Cong nghe thong tin va truyen thong. Ha Noi, Viet Nam, 03-04/12/2012.

[10] Erkan, Gunes; Radev, Dragomir R.: LexRank (2004): Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22 (2004) 457-479.

[11] Gokaslan, Aaron; Cohen, Vanya (2019): Openwebtext corpus. http://web.archive.org/ save/http://Skylion007.github.io/ OpenWebTextCorpus.

[12] Ha, Thanh Le; Huynh, Quyet Thang; Luong, Chi Mai (2005): A Primary Study on Summarization of Documents in Vietnamese. Proceeding of the First International Congress of the International Federation for Systems Research, Kobe, Japan, Nov 15- 17, pp.234-239.

[13] Hermann, Karl Moritz; Kocisk, Tomas Y.; Grefenstette, Edward; Espeholt, Lasse; Kay, Will; Suleyman, Mustafa; Blunsom, Phil (2015): Teaching machines to read and comprehend," In Advances in Neural Information Processing Systems 28. pages 1693–1701.

[14] Jadhav, Aishwarya; Rajan, Vaibhav (2018): Extractive Summarization with SWAP-NET: Sentences and Words from Alternating Pointer Networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 142–151 Melbourne, Australia, July 15 - 20.

[15] Kamal, Al-Sabahi; Zhang, Zuping; Mohammed, Nadher (2018): A Hierarchical Structured Self-Attentive Model for Extractive Document Summarization (HSSAS). IEEE Access, Volume 6. pp. 24205-24212.

[16] Kupiec, Julian; Pedersen, Jan; Chen, Francine (1995): A trainable document summarizer. Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pages 68–73. ACM.

[17] Lam, Quang-Tuong; Pham, The-Phi; Do, Duc-Hao (2017): Tom tat van ban tieng Viet tu dong voi mo hinh sequence-to-sequence. Tap chi khoa hoc Truong Dai hoc Can Tho. So chuyen de: Cong nghe thong tin, pp. 125-132.

[18] Liu, Yinhan; Ott, Myle; Goyal, Naman; Du, Jingfei; Joshi, Mandar; Chen, Danqi; Levy, Omer; Lewis, Mike; Zettlemoyer, Luke; Stoyanov, Veselin (2019): RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692v1 [cs.CL].

[19] Loshchilov, Ilya; Hutter, Frank (2019): Decoupled Weight Decay Regularization. arXiv:1711.05101v3 [cs.LG].

[20] Luhn, Hans Peter (1958): The automatic creation of literature abstracts. IBM Journal of research and development, 2(2):159–165.

[21] Lyyer, Mohit; Manjunatha, Varun; Boyd-Graber, Jordan; Daume, Hal (2015): Deep unordered composition arivals syntactic methods for text classification. In Proceedings of ACL/IJCNLP.

[22] Mihalcea, Rada; Tarau, Paul (2004): TextRank: Bringing Order into Texts. pp.1364-1368.

[23] Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg; Dean, Jeffrey (2013): Distributed representations of words and phrases and their compositionality. In Proceedings of NIPS'13.

[24] Nagel, Sebastian (2016): Cc-news. http: //web.archive.org/save/http: //commoncrawl.org/2016/10/newsdataset-available.

[25] Nallapati, Ramesh; Zhai, Feifei; Zhou, Bowen (2017): Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In AAAI, pages 3075–3081.

[26] Narayan, Shashi; Cohen, Shay B.; Lapata, Mirella (2018): Ranking Sentences for Extractive Summarization with Reinforcement Learning. Proceedings of NAACL-HLT 2018, pages 1747–1759 New Orleans, Louisiana, June 1 - 6.

[27] Nguyen, Dat Quoc; Nguyen, Anh Tuan (2020): PhoBERT: Pre-trained language models for Vietnamese. arXiv:2003.00744v3 [cs.CL].

[28] Nguyen, M.L.; Akira, Shimazu; Phan, Xuan-Hieu; Ho, Tu-Bao; Susumu, Horiguchi (2005): Sentence Extraction with Support Vector Machine Ensemble. Proceedings of the First World Congress of the International Federation for Systems Research: The New Roles of Systems Sciences For a Knowledge-based Society 2005.

[29] Nguyen, Quang Uy; Pham, Tuan Anh; Truong, Cong Doan; Nguyen, Xuan Hoai (2012): A Study on the Use of Genetic Programming for Automatic Text Summarization,", KSE, 2012 Fourth International Conference on Knowledge and Systems Engineering, pp.93-98.

[30] Nguyen, Thi Thu Ha (2012): Phat trien mot so thuat toan tom tat van ban tieng Viet su dung phuong phap hoc ban giam sat. Hoc vien ky thuat quan su.

[31] Nguyen, Trong-Phuc; Le, Thanh-Huong (2008). Tom tat van ban tieng Viet su dung cau truc dien ngon. Hoi thao ICT.rda 2008.

[32] Pennington, Jeffrey; Socher, Richard; Manning, Christopher D. (2014): Glove: Global vectors for word representation. Proceeding of EMNLP.

[33] Radford, Alec; Narasimhan, Karthik; Salimans, Tim; Sutskever, Ilya (2018): Improving Language Understanding by Generative Pre-Training. url: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[34] Ren, Pengjie; Chen, Zhumin; Ren, Zhaochun; Wei, Furu; Ma, Jun; Rijke, Maarten de. (2017): Leveraging contextual sentence relations for extractive summarization using a neural attention model. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 95–104, New York, NY, USA. ACM.

[35] Sennrich, Rico; Haddow, Barry; Birch, Alexandra (2016): Neural machine translation of rare words with subword units. In Association for Computational Linguistics (ACL), pages 1715–1725.

[36] Trinh, Trieu H.; Le, Quoc V. (2018): A simple method for commonsense reasoning. arXiv preprint arXiv:1806.02847.

[37] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz; Polosukhin, Illia (2017): Attention is all you need.

[38] Wasson, Mark (1998): Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. Proceedings of the 17th international conference on Computational linguistics-Volume 2.

[39] Wong, Kam-Fai; Wu, Mingli; Li, Wenjie (2008): Extractive Summarization Using Supervised and Semi-Supervised Learning. Proceedings of the 22nd International Conference on Computational Linguistics, pages 985–992.

[40] Wu, Yuxiang; Hu, Baotian (2018): Learning to Extract Coherent Summary via Deep Reinforcement Learning. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), pp. 5602-5609.

[41] Yang, Yinfei; Cer, Daniel; Ahmad, Amin; Guo, Mandy; Law, Jax; Constant, Noah; Abrego, Gustavo Hernandez; Yuan, Steve; Tar, Chris; Sung, Yun-Hsuan; Strope, Brian; Kurzweil, Ray (2019): Multilingual Universal Sentence Encoder for Semantic Retrieval. arXiv:1907.04307v1 [cs.CL].

[42] Zhang, Xingxing; Lapata, Mirella; Wei, Furu; Zhou, Ming (2018): Neural Latent Extractive Document Summarization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 779–784 Brussels, Belgium, October 31 - November 4.

[43] Zhang, Xingxing; Wei, Furu; Zhou, Ming (2019): HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. arXiv:1905.06566v1 [cs.CL].

[44] Zhang, Yong; Meng, Joo Er.; Pratama, Mahardhika (2016): Extractive Document Summarization Based on Convolutional Neural Networks'', IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, pp. 918-922.

[45] Zhou, Qingyu; Yang, Nan; Wei, Furu; Huang, Shaohan; Zhou, Ming; Zhao, Tiejun (2018): Neural Document Summarization by Jointly Learning to Score and Select Sentences. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 654–663 Melbourne, Australia, July 15 - 20.

[46] Zhu, Yukun; Kiros, Ryan; Zemel, Richard; Salakhutdinov, Ruslan; Urtasun, Raquel; Torralba, Antonio; Fidler, Sanja (2015): Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. arXiv preprint arXiv:1506.06724.

Authors Profile

Minh-Tuan Luu, He received her Engineer degree on Information Technology from Hanoi University of Science and Technology (HUST), Vietnam in 1998; Getting master degree on Information Technology from Hanoi University of Science and Technology, Vietnam in 2002. He has been working as a lecturer of School of Information Technology and Digital Economics, National Economics University since 2004 and he has been a PhD student in specialization in Information System at Hanoi University of Science and Technology, Vietnam since 2017. His main research topics include artificial intelligence, text summarization, discourse structure, syntactic parsing, question answering, information extraction, plagiarism detecting, and recommendation systems.

Thanh-Huong Le, She received her Bachelor degree on Informatics from Hanoi University of Science and Technology, Vietnam in 1997; Master degree on Robotics from Free University of Brussels (VUB), Belgium in 2001; and PhD degree on Natural Language Processing in 2004 from Middlesex University, United Kingdom. She has been working as a lecturer of School of Information and Communication Technology, Hanoi University of Science and Technology since 1998. Her main research topics include discourse structure, syntactic parsing, text summarization, question answering, information extraction, plagiarism detecting, and recommendation systems.

Minh-Tan Hoang, He is studying for a Bachelor of Computer Science at Hanoi University of Science and Technology (HUST), Vietnam. He is a final-year student of School of Information and Communication Technology. He will get Bachelor degree on Computer Science from Hanoi University of Science and Technology in June, 2021. His main research topics include reinforcement learning, syntactic parsing, text summarization, question answering, information extraction, and recommendation systems.