Journal of Computer Science and Cybernetics, V.37, N.2 (2021), 123–143
DOI 10.15625/1813-9663/37/2/15980

A HYBRID MODEL USING THE PRE-TRAINED BERT AND DEEP NEURAL NETWORKS WITH RICH FEATURE FOR EXTRACTIVE TEXT SUMMARIZATION

TUAN LUU MINH[1,2], HUONG LE THANH[1,*], TAN HOANG MINH[1]

[1] *Hanoi University of Science and Technology, Hanoi, Vietnam*
[2] *National Economics University, Hanoi, Vietnam*

Abstract. Deep neural networks have been applied successfully to extractive text summarization tasks with the accompany of large training datasets. However, when the training dataset is not large enough, these models reveal certain limitations that affect the quality of the system's summary. In this paper, we propose an extractive summarization system based on a Convolutional Neural Network, an Encoder-Decoder, and a Fully Connected network for sentence selection. The pretrained BERT multilingual model is used to generate embedding vectors from the input text. These vectors are combined with TF-IDF values to produce the input of the text summarization system. Redundant sentences from the output summary are eliminated by the Maximal Marginal Relevance method. Our system 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 works using the same dataset. It confirms that our approach can be effectively applied to summarize both English and Vietnamese languages.

Keywords. Extractive Summarization; BERT multilingual; CNN; Encoder-Decoder; TF-IDF feature

1. INTRODUCTION

The strong development of the Internet has made the volume of data in general and text data in particular increasing rapidly, causing challenges for searching and capturing information. One solution to this problem is text summarization. This is the process of extracting important and valuable information from one or more documents to generate a concise summary. The summary needs to ensure the accuracy of the content and the meaning of the source text [6, 14].

The representation of the input document is an important factor to extract sentence features in a neural network based on document summarization systems. Several machine learning and deep learning approaches have solved this problem by using pretrained word embeddings such as word2vec [27] and Glove [33]. However, these word embedding vectors

*Corresponding author.
E-mail addresses: tuanlm@neu.edu.vn (T. L. Minh); huonglt@soict.hust.edu.vn (H. L. Thanh); tan.hm1211@gmail.com (T. H. Minh).

are context-independent, which may result in incorrect meaning in some cases. Recent works on generating context-based representing models have shown that using pretrained sentence embedding provides better performance than using word embeddings in natural language processing tasks [10]. Cera et al. [8] proposed two pretrained Universal Sentence Encoder (USE) basing on Transformer [39] and Deep Averaging Network [17] for the English language. Yang et al. [42] expanded the pretrained USE models for 16 languages without Vietnamese. These models were pretrained on a large unlabeled text to generate sentence embedding.

With the introduction of the pretrained BERT (Bidirectional Encoder Representations from Transformers) [11], the obtained results in the natural language processing tasks have made significant progress. The strength of the pre-trained BERT model is the ability to learn bidirectional context with the attention mechanism from the transformer that allows capturing the important factors of the input document. The application of the BERT model to the semantic information representation of the input document has also improved the accuracy of text summarization systems (e.g., [23, 44, 46]).

This paper aims at developing an extractive summarization system that can be applied for both English and Vietnamese languages. The summarization system is constructed as a classification model, in which sentences in the summary have the label 1, and 0 otherwise. We use pre-trained BERT multilingual (M-BERT) model [34] that supports multi-languages including Vietnamese to generate document embeddings. The sentence classification task is performed by using a CNN, an Encoder-Decoder, an FC layer combining with the TF-IDF feature as input for the model. Maximal Marginal Relevance (MMR) is used to eliminate redundant information to generate the summary. 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 research using the same dataset. The standard ROUGE measures [4] including $F1-$Score measures on Rouge-1 $(R-1)$, Rouge-2 $(R-2)$, and Rouge-L $(R-L)$ are used to evaluate the quality of extractive text summarization systems.

Our main contributions can be summarized as follows:

- Proposing an effective extractive text summarization system that can be applied for both English and Vietnamese languages;
- Applying the pretrained BERT multilingual model to represent the token embedding vectors of the sentences, in order to have a better understanding of the input text;
- Proposing the CNN and fine-tuning with a novel activation function to extract sentence features. Subsequently, proposing the k-max pooling layer to obtain the sentence vectors;
- Proposing Encoder-Decoder model with a bidirectional Long Short Term Memory for both the encoder and the decoder to associate the context of the considered sentence in the text;
- Combining TF-IDF feature to the model and reducing the dimension of the TF-IDF vector by using the Fully Connected layer without activation function;

The rest of this paper is organized as follows. The related works in extractive text summarization are discussed in Section 2. Section 3 presents some background of our proposed approach. The proposed text summarization system is introduced in Section 4. Our experiments are described in Section 5. Finally, Section 6 concludes the paper and proposes future works for our research.

2. RELATED WORKS

Recently, deep learning techniques have been successfully applied to text summarization and get better results comparing to other approaches. Zhang et al. [45] extracted salient sentences for the summary by using Convolutional Neural Networks (CNN). Nallapati et al. [30] treated extractive summarization as a sequential labeling task. In this approach, sentences of the input document were encoded and then classified into two classes: selected or not selected. This system computed selection probability for each sentence, then generated a summary basing on these probabilities until reaching the summary limit. Zhou et al. [47] developed 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 trained their neural summarization systems by using a reinforcement learning objective. Wu et al. [41] proposed a neural coherence system to capture the cross-sentence semantic and syntactic coherence patterns, using a reinforcement learning mechanism. The system's reward was computed by evaluating the system's output using the Rouge measure. Zhang et al. [43] proposed a latent variable extractive summarization system that uses direct human summarization and a sentence compression system to generate the summary. In this approach, sentences were considered as latent variables. Sentences with activated variables were used to generate the summary. This technique solved the problem of depending on sentence-level labels, which was often used in extractive summarization systems. Jadhav and Rajan [18] developed a neural sequence-to-sequence model for extractive summarization that modeled the interaction of key words and salient sentences using a two-level pointer network-based architecture. Al-Sabahi et al. [5] proposed a summarization system using a hierarchical structured self-attention mechanism to capture the hierarchical structure of the document and to create the sentence and document embedding. The attention mechanism provided an extra source of information to guide the summarization extraction. The system computed the probabilities of sentence-summary membership basing on several features such as information content, salience, novelty, and positional representation. Zhang et al. [44] proposed a document encoding system named HIerachical Bidirectional Encoder Representations from Transformers and pretrained it using unlabeled data. This system provided promising results when applying to the summarization task. These systems often required a large training dataset and a high training cost in general.

There are only a few works on Vietnamese text summarization. However, most of them rely on sentence features such as sentence position, TF-IDF, title, similarity, etc... to compute sentence scores (e.g., [37]). Machine learning and deep learning algorithms were used in several works such as [21, 28, 38, 15]. Salient sentences were extracted by using Support Vector Machines in [28], applying a genetic algorithm in [38]), and taking advantage of a semi-supervised algorithm in [15]. Lam et al. [21] constructed a Sequence to Sequence with Attention model and beam search to generate the summary, the words of the input document were embedded as word vectors to use as input of the summarization system. The system was trained by a self-collected news dataset with $31,429$ articles, in which the abstract of each news was used as its summary.

However, most of the above approaches have not applied an efficient way to represent the semantic structure of the input document, which causes redundancies in the summaries. To deal with this problem, we propose a text summarization system that exploits the representing power of deep neural networks. Specifically, we treat the text summarization task as a classification problem, using a pretrained BERT model to encode input text.

3. BACKGROUND

This section briefly introduces background information of main components of the proposed summarization system, including: (i) BERT [11], (ii) BERT multilingual [34], (iii) CNN [20], (iv) LSTM cell and biRNN [13], and (v) MMR method [7].

3.1. BERT model

BERT model [11] is designed to pretrain deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers is a new language representation model. The main characteristics of the BERT model are summarized below.

Input representation. The input of the BERT model can be considered as a concatenation operator of two segments (each segment is often a sentence) denoted as $x_1, x_2, \ldots, x_M$; $y_1, y_2, \ldots, y_N$. These two segments are represented as a single input sequence with special tokens that consists of *[CLS]*,*[SEP]*, *[EOS]* to distinguish them as *[CLS]*, $x_1, x_2, \ldots, x_M$, *[SEP]*, $y_1, y_2, \ldots, y_N$, *[EOS]* where:

- *CLS* is the token that was added to the beginning of the first segment;
- *SEP* is the token that was added to the end of the first segment;
- *EOS* is the token that was added to the end of the sequence (at the end of the second segment);
- Satisfying the condition $M + N < T$ (where T is the maximum length of the sequence during training. BERT is pretrained on a large unlabeled text corpus that is a combination of BooksCorpus ($800M$ words) [48] and English Wikipedia ($2,500M$ words), which totals $\sim 16GB$ of uncompressed text and fine-tuned using labeled data from the downstream tasks.

BERT architecture. BERT model uses Transformer architecture [39] has L layers, each block of the BERT model uses A self-attention heads and hidden dimension H, and takes the input as a set of sentences of the input text. There are many BERT models, consist of $BERT_{BASE}$ ($12-$layer, $768-$hidden, $12-$heads, $110M$ parameters), $BERT_{LARGE}$ ($24-$layer, $1,024-$hidden, $16-$heads, $340M$ parameters), and BERT-Base, Multilingual models support multiple languages for natural language processing (NLP) tasks.

BERT implementation. BERT model uses two pretraining and fine-tuning tasks to generate state-of-the-art models for the NLP tasks.

(i) *Pretraining BERT.* BERT is pretrained with two actions. First action, Masked Language Model (MLM), a set of random tokens of the input sequence is selected and replaced with a special token *[MASK]*. The MLM is a cross-entropy loss on predicting the masked tokens. BERT model chooses 15% of the input tokens for replacement. For these selected tokens, 80% tokens were replaced with *[MASK]*, 10% tokens were unchanged, and 10% tokens were replaced with the randomly selected vocabulary token. The second action, Next Sentence Prediction (NSP) is a binary classification loss that predicts the next two segments of the original text. The *'Position'* labels are generated by taking the next segment in the original text. The *'Negative'* labels are generated by pairing segments from the text in the dataset. For example, we need to pretrain the BERT model using a text dataset of 100, 000 segments, thus we have 50, 000 train samples (pairs of segments) as train data, with 50% of the pair of segments, the second segment will be the next segment of the first segment, these labels are denoted as *'Position'*. With 50% of the other pair of segments, the second segment will be a random segment from the dataset, these labels are denoted as *'Negative'*.

(ii) *Fine-tuning.* Depending on each specific downstream task, BERT model will be fine-tuned with its training dataset to generate a more suitable semantic model for that downstream task. The model uses two fine-tuning strategies [1, 11]:

   (a) Reduce the learning rate to increase the model's accuracy;

   (b) Freeze some previous layers to freeze the learned weights from pretrained models (the model only updates weights on the higher layers) to increase the speed of the model's train.

3.2. BERT multilingual

The pretrained BERT models [11] only support monolingual English processing tasks in NLP. Basing on the pretrained BERT models, Pires et al. [34] has developed BERT multilingual (M-BERT) models to support multi-languages including Vietnamese. M-BERT models have the same architecture as the corresponding pretrained BERT models. There are two available M-BERT models based on BERT-Base models to be BERT-Base, Multilingual Cased (104 languages, 12$-$layer, 768$-$hidden, 12$-$heads, 110M parameters) and BERT-Base, Multilingual Uncased (102 languages, 12$-$layer, 768$-$hidden, 12$-$heads, 110M parameters). M-BERT models are trained on Wikipedia pages of 104 languages with the corresponding vocabulary instead of only being trained on monolingual English datasets with English vocabulary like the BERT models.

3.3. CNN

The CNN model architecture for sentence-level classification tasks is described by Kim et al. (2014) [20]. The CNN model often uses multiple filters to obtain multiple features. These features form the penultimate layer and are fed forwards to an FC layer with the softmax activation function that its output is the probability distribution of the labels. Extracting a feature from a filter is described below.

Let $x_i \in \mathbb{R}^m$ be a $m-$dimensional word vector corresponding to the i^{th} word in the sentence. A sentence of length n (padded if necessary) is represented as

$$x_{1\to n} = x_1 \oplus x_2 \oplus x_3 \oplus \cdots \oplus x_n, \tag{1}$$

where, $\oplus$ is a concatenation operation. In general, suppose that $x_{i\to i+j}$ denote the concatenation of words $x_i, x_{i+1}, \ldots, x_{i+j}$. The convolution operation between the filter $w \in \mathbb{R}^{hm}$ and a window of h words to generate a new feature. The c_i feature is produced from a window of words $x_{i\to i+h-1}$ to present as

$$c_i = f(w.x_{i\to i+h-1} + b), \tag{2}$$

where, $b \in \mathbb{R}$ is the bias term and f is the nonlinear function. This filter is applied to every possible window of words in the sentence $\{x_{1\to h}, x_{2\to h+1}, \ldots, x_{n-h+1\to n}\}$ to produce a *feature map*

$$c = [c_1, c_2, c_3, \ldots, c_{n-h+1}], \tag{3}$$

where, $c \in \mathbb{R}^{n-h+1}$. The max-overtime pooling operation [9] is applied on the feature map to take the maximum value $\hat{c} = \max\{c\}$ as the feature corresponding to each filter.

We use the CNN model architecture in [20] and fine-tune it to apply our proposed text summarization model by applying k-max pooling operation [19] on each feature map instead of the max-overtime pooling operation [9] to select the k maximum values as the feature corresponding to each filter.

3.4. LSTM cell and biRNN

LSTM cell. LSTM (Long Short Term Memory) cell is described by Graves et al. [13]. Let an input sequence $x = (x_1, \ldots, x_T)$. At step t (with $t = 1, 2, \ldots, T$), the Recurrent Neural Network (RNN) computes the hidden vector sequence $h = (h_1, \ldots, h_T)$ and the output vector sequence $y = (y_1, \ldots, y_T)$ by the following formulas

$$\begin{aligned} h_t &= H(W_{xh}x_t + W_{hh}h_{t-1} + b_h), & (4)\\ y_t &= W_{hy}h_t + b_y, & (5)\end{aligned}$$

where, W terms are weight matrices (e.g., W_{xh} is input hidden weight matrix); the b terms are bias vectors (e.g., b_h is hidden bias vector), and H is the hidden layer function. Figure 1a) illustrates an LSTM cell. H is performed by the following summary function:

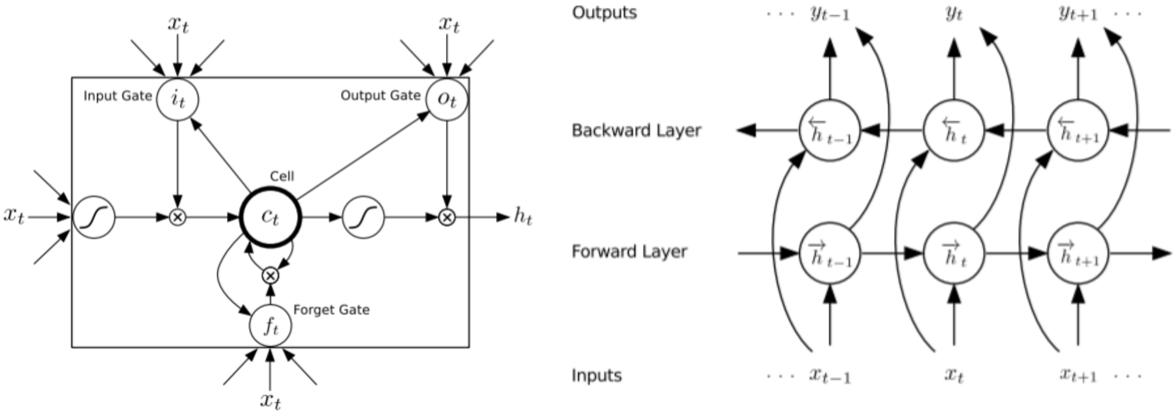
$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), & (6)\\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), & (7)\\ c_t &= f_t c_{t-1} + i_t \tanh((W_{xc}x_t + W_{hc}h_{t-1} + b_c)), & (8)\\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), & (9)\\ h_t &= o_t \tanh(c_t), & (10)\end{aligned}$$

where, σ is the sigmoid logistic function, and i, f, o, and c are the input gate, forget gate, output gate, and the cell activation vectors, and they are the same size as the hidden vector h.

biRNN. biRNN (Bidirectional RNN) processes the data in both directions with two separate hidden layers, which are then fed forwards to the same output layer. Figure 1b presented a biRNN, the biRNN computes the forward hidden sequence $\overrightarrow{h_t}$ (with $t = 1, 2, \ldots, T-1, T$), the backward hidden sequence $\overleftarrow{h_t}$ (with $t = T, T-1, \ldots, 1$), the output sequence y_t at step t, and updating the output layer:

$$\begin{aligned} \overrightarrow{h_t} &= H(W_{x\overrightarrow{h}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}), & (11)\\ \overleftarrow{h}_t &= H(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}), & (12)\\ y_t &= H(W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y). & (13)\end{aligned}$$

We will construct an Encoder-Decoder architecture by using biLSTMs for both the encoder and the decoder for our proposed system.



(a) Long Short-term Memory Cell

(b) Bidirectional RNN

Figure 1. Long Short Term Memory cell and Bidirectional RNN [13]

3.5. MMR method

The originally MMR method [7] was proposed to measure the relevance between the user query Q and sentences in the document in the Information Retrieval (IR) problem. The formula calculating MMR is

$$\text{MMR} \overset{def}{=} \arg \max_{D_i \in C \setminus S} \left[\lambda \left(Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right) \right], \tag{14}$$

where,

- 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 selected 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. The formula calculating the MMR measure was redefined to apply the document summarization task as follows

$$\text{MMR} \overset{def}{=} \arg \max_{D_i \in C \setminus \{S, Q\}} \left[\lambda \left(Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right) \right], \tag{15}$$

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

$$Sim_1(u,v) = Sim_2(u,v) = \frac{\sum_{w\in v} tf_{w,u} tf_{w,v} (idf_w)^2}{\sqrt{\sum_{w\in u} (tf_{w,u} idf_w)^2}}, \tag{16}$$

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.

The main point of applying the MMR method is to eliminate redundant information in the summary. We carried out three steps as below:

- 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. OUR PROPOSED SYSTEM

Our proposed text summarization system is shown in Figure 2, which consists of three main modules: (i) Token embedding, (ii) Sentence Classification, and (iii) Summarization.

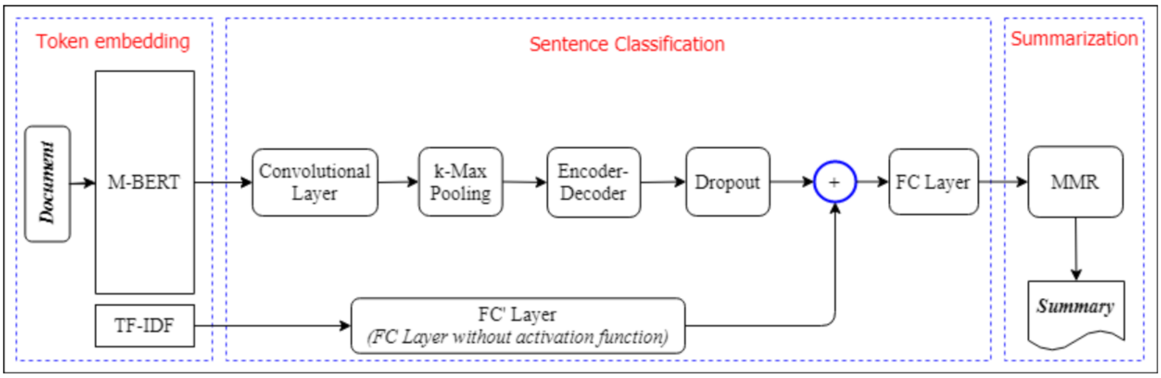


Figure 2. Our proposed text summarization system

4.1. Token embedding

The first step of the token embedding process is to split the input text into sentences. Then, the first 32 sentences from each document are taken to represent that document. The first 64 tokens of each sentence are taken to represent that sentence (padding if necessary). These sentences are processed by an M-BERT model to obtain their token embeddings. The M-BERT model is fine−tuned by summarization datasets during our experiments to get the best representation of the input text.

The token embedding vectors are used as the input of the Convolutional layer in our CNN model to obtain the 400−dimension sentence vectors.

4.2. Sentence classification

Sentence classification is the major component of the system. The purpose of this task is to compute the selection probability of the input sentences to be included in the summary. To perform this task, the sentential embedding vectors are processed by a CNN, an encoder–decoder model, and a dropout layer to extract vectors of sentence features. The $TF - IDFs$ of each sentence are put into the fully connected layer without an activation function (denoted as FC' layer in Figure 2, and then are combined with vectors of sentence features in another fully connected layer to produce the selection probability of each sentence. The components of our proposed system are described in detail below.

Convolutional layer and k-max Pooling layer. We use the CNN architecture [20] described in Section 2.3 above and fine-tune it in our experiments. The input of the Convolutional layer is a tensor to be sharpened $(n, 1, D * L, H)$. Here, n is the batch size; D is the number of sentences in the document; L is the sentence length; H is the dimensions of the three last hidden layers of M-BERT. In our experiment, we choose $n = 32$, $D = 32$, $L = 64$, $H = 3 * 768$. To reduce the computational complexity of the model, we do not use $l2$ constraint since it does not affect the system's accuracy. The Convolutional layer uses a novel activation function called **mish**[3], proposed by Misra and Landskape [29], to improve the efficiency of the neural network architecture. This function is calculated by the following formula

$$f(x) = x \tanh(softplus(x)) = x \tanh(\ln(1 + e^x)), \tag{17}$$

where $softplus(x)) = \ln(1 + e^x)$.

Next, we feed forward the $k-$max pooling layer by applying the k-max pooling operation [19] instead of using the max-overtime pooling operation [9]. The $k-$max pooling operation in formula (18) is applied to each feature map to select k maximum values

$$k_\ell = \max\left(k_{top}, \lceil \frac{L - \ell}{L} s \rceil\right), \tag{18}$$

where, k_l is a function involving the sentence length and the network depth; l is the number of the current convolutional layer; L is the total number of convolutional layers in the network; k_{top} is the fixed pooling parameter for the topmost convolutional layer; s is the sentence length. Figure 3 illustrates the Convolutional layer architecture with $k-$max pooling $(k = 2)$ applying to our proposed system.

Encoder-Decoder. Our Encoder-Decoder model is a bidirectional Long Short Term Memory ($biLSTM$) for both the encoder and the decoder. Each biLSTM has $1,024$ hidden states (equals to $2 * 512$ hidden states) to associate the context of the considered sentence in the text. Inputs of this model are the output sentence vectors $(s_1, s_2, \ldots, s_m)$ of the $k-$max pooling layer, whereas its outputs are sentence vectors $(s'_1, s'_2, \ldots, s'_m)$ of $1,024$ dimensions. Our proposed Encoder-Decoder architecture is illustrated in Figure 4 below.

Dropout layer. The FC layer is prone to be overfitted. To prevent this problem, the output sentence vectors of the Encoder-Decoder are selected by a dropout layer with a dropout rate of 0.2.

[3] Available at https://github.com/digantamisra98/Mish

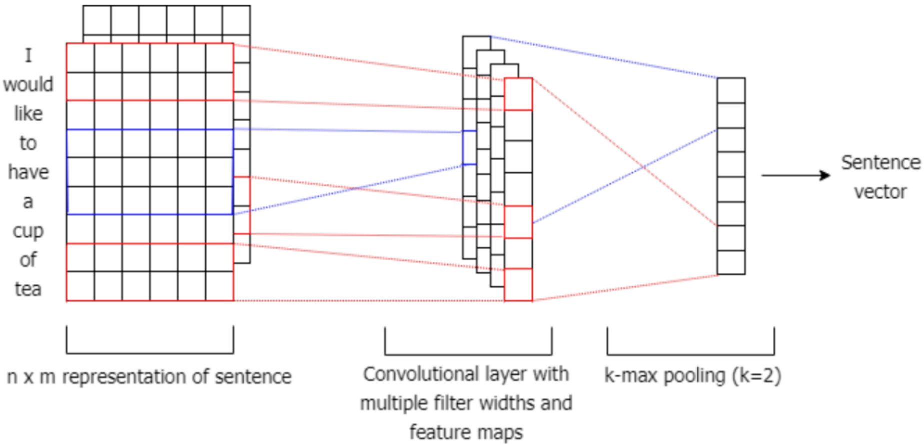

I
would
like
to
have
a
cup
of
tea
Sentence vector
n x m representation of sentence
Convolutional layer with multiple filter widths and feature maps
k-max pooling (k=2)

Figure 3. The Convolutional layer architecture with $k-$max pooling ($k = 2$) applying to our proposed system

TF-IDF feature and FC' layer. Viet et al. [35] have pointed out the efficiency of the TF-IDF feature in text summarization systems. Because of that, we also use the TF-IDF as a feature in our system. Since the TF-IDF vector is large (equal to the size of the dictionary), we use the Fully Connected layer without an activation function (denoted as FC' layer in Figure 2) to reduce the dimension of the TF-IDF vector to reduce the system's complexity. The dictionary is limited to $40,000$ words with the highest term frequency, so the dimension of the TF-IDF vector is also $40,000$. The dimensions of the input and output vectors of the FC' layer are $40,000$ and 128, respectively.

Concatenating operation. The output vector of the FC' layer is concatenated with the output vector of the dropout layer by the concatenating operation (denote as $\oplus$ to obtain a vector of $1,152$ dimensions, which is used as the input vector of the fully connected layer.

Fully Connected (FC) layer. The FC layer converts a $1,152-$dimension input vector to a $2-$dimension output vector, using a softmax activation function. This output vector contains the probability to select a sentence to put in the summary.

4.3. Summarization

Sentences from the input document are sorted in descending order by the selection probabilities. These sentences are selected to include in the summary until reaching the summary length. To eliminate redundancy, we apply the MMR method in [25] to measure the similarity among sentences and eliminate redundant sentences whose similarity with existing sentences in the summary is larger than a certain threshold.

5. EVALUATION

5.1. Datasets

To evaluate the proposed system, we carried out experiments with two different languages: English and Vietnamese. The English datasets consist of DUC 2001 [2], DUC 2002 [3], and

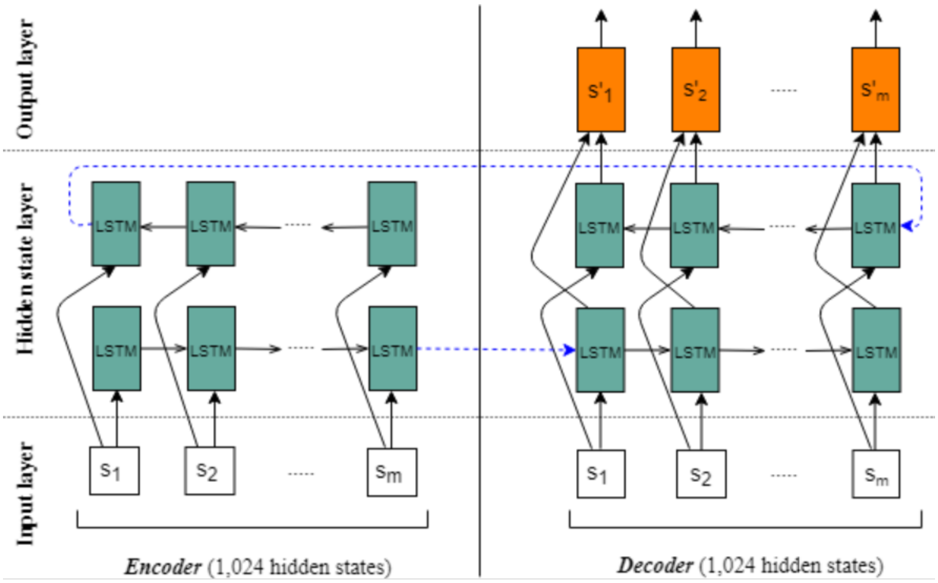


Figure 4. The architecture of our proposed Encoder-Decoder model

CNN [16]. Brief information of the DUC 2001 and DUC 2002 datasets is summarized in Table 1.

Table 1. The statistical information on the DUC 2001 and DUC 2002 datasets

Dataset	**#Doc.**	**#Sen.(Avg)**	**#Sum.(Avg)**
DUC 2001	486	40	2.17
DUC 2002	471	28	2.04

The purpose of using the CNN dataset is to compare our system with the state-of-the-art ones in extractive summarization, as the CNN dataset is usually used in evaluating summarization systems. Our experiments with the Vietnamese dataset (Baomoi) aim at evaluating our system with another language and guaranteeing the generality of our system.

The CNN/Daily Mail dataset consists of 312, 085 articles, in which 92, 579 articles from the CNN dataset, and the highlights written by the article's authors. These highlights have three sentences on average. We split the CNN dataset into training/validation/test datasets with 90, 266/1, 220/1, 093 documents, respectively. Table 2 describes brief statistic information of the CNN/Daily Mail dataset.

Table 2. The brief statistics on the CNN/Daily Mail dataset

	CNN			**Baomoi**		
	Train	**Valid**	**Test**	**Train**	**Valid**	**Test**
#documents	90,266	1,220	1,093	196,961	12,148	10,397
Avg#tokens	762	763	716	813	774	780
Vocab size		118,497			208,045	

Since there is no available Vietnamese summarization dataset, we use a dataset called 'Baomoi', which was created by gathering articles from a Vietnamese online newspaper (http://baomoi.com). Each article consists of three parts: headline, abstract, and article's content. The abstract consists of 2 sentences with 45 words on average, which represents the main information of the article. Since we cannot find any better source, the Baomoi dataset is our best choice to be used as the summarization corpus at the moment. The average length of the article's content is 503 words. The final dataset consists of $1,000,847$ news articles, being splitted into training/validation/test datasets with $900,847/50,000/50,000$ documents, respectively.

5.2. Preprocessing

First, we separated the content and the summary from each article in the CNN and Baomoi datasets. The **StanfordCoreNLP**[4] and **VnCoreNLP**[5] libraries were used for splitting English and Vietnamese text into sentences, respectively. Next, the **rouge-score 0.0.4**[6] library was used to label sentences from each dataset basing on the maximum total of $R-2$ and $R-L$.

5.3. Experiments

To choose the best model for our proposed extractive summarization system, we carried out experiments with four scenarios, using the CNN dataset:

- Scenario 1 ($M-BERT+CNN+FC+TF-IDF$): The system uses the M-BERT model combined with the CNN layer, the FC layer, and the TF-IDF feature.
- Scenario 2 ($M-BERT+CNN+Encoder-Decoder+FC+TF-IDF$): The system uses the model in scenario 1, combined with the $Encoder-Decoder$ model to associate the context of the sentences in the document. This scenario is used to evaluate the efficiency of using the $Encoder-Decoder$ to the summary system.
- Scenario 3 ($M-BERT+CNN+FC+TF-IDF+MMR$): The system uses the model in scenario 1, combined with the MMR method to eliminate overlapping information in the summary. This scenario aims at evaluating the efficiency of applying the MMR method to the system.
- Scenario 4 ($M-BERT+CNN+Encoder-Decoder+FC+TF-IDF+MMR$): The system uses the model in Scenario 2, combining with the MMR method to eliminate overlap information in the summary.

Our system was trained on Google Colab using GPU $V100$, RAM $25Gb$. Our experiments used the initial learning rate of 2.10^{-3} over 10 epochs, the batch size of 32 and the train time of approximately 29 hours for the CNN dataset and 63 hours for the Baomoi dataset. After each epoch, the learning rate was automatically reduced by 10% using the **scheduling** mechanism in the **PyTorch**[7] library until reaching the end of the last epoch.

The system performance was evaluated by $R-1$, $R-2$, and $R-L$ scores, computed by using the rouge−score 0.0.4 library.

Experimental results with the above−mentioned four scenarios are shown in Table 3 below.

Scenario 2 gains higher results than Scenario 1, with the support of the encoder-decoder. The values of $R-1$, $R-2$, and $R-L$ in Scenario 2 increase 0.33%, 0.68%, and 0.16%

[4] Available at http://stanfordnlp.github.io/CoreNLP/
[5] Available at http://github.com/vncorenlp/VnCoreNLP/
[6] Available at http://github.com/google-research/google-research/tree/master/rouge/
[7] Available at https://github.com/pytorch/pytorch/

Table 3. Experimental results with the four scenarios

Scenarios	**CNN**		
	R-1	**R-2**	**R-L**
Scenario 1 (M-BERT+CNN+FC+TF-IDF)	31.62	12.01	28.57
Scenario 2 (M-BERT+CNN+Encoder-Decoder+FC+TF-IDF)	31.95	12.69	28.76
Scenario 3 (M-BERT+CNN+FC+TF-IDF+MMR)	32.54	12.60	29.52
Scenario 4 (M-BERT+CNN+Encoder-Decoder+FC+TF-IDF+MMR)	32.64	13.04	29.53

respectively, comparing to Scenario 1. Scenario 3, which is the combination of the MMR method into Scenario 1 to eliminate redundant information in the summary, provides better results than Scenario 1. Table 4 shows that Scenario 4 is the best result among the four scenarios. It proves the importance of the TF-IDF feature, the $Encoder - Decoder$, and the MMR method in the text summarization task.

To exam the generality of our approach, we also carried out experiments with a Vietnamese dataset (Baomoi), using our best model (Scenario 4). This experiment obtained 54.67%, 25.26%, and 37.48% for the $R-1$, $R-2$, and $R-L$ scores, respectively.

Table 4 shows an example of our system's output on the CNN dataset. The source text of this summary is in Appendix 1.

Table 4. An example of the system's output on the CNN dataset

Human

"The Cold War has been over for a long time," President Obama says. The thaw in ties has dominated discussion at the Summit of the Americas in Panama. The top leaders from the United States and Cuba haven't met for substantive talks in more than 50 years.

Our system's summary

Panama City, Panama(CNN)Ending a decades-long standstill in U.S.-Cuba relations, President Barack Obama met for an hour Saturday with his Cuban counterpart Raul Castro, the first time the two nations' top leaders have sat down for substantive talks in more than 50 years. The meeting in a small conference room on the sidelines of the Summit of the Americas came as the two countries work to end the Cold War enmity that had led to a total freeze of diplomatic ties. "The Cold War has been over for a long time," Obama said during opening remarks at the summit Saturday.

Table 5 shows an example of our system's output on the Baomoi dataset. The source text of this summary is in Appendix 2.

Table 5. An example of the system's output on the Baomoi dataset

Human

Đang chạy, chiếc xe con bất ngờ leo lên dải phân cách làn đường ở đầu cầu Bến Thuỷ. Vụ tai nạn khiến giao thông ùn tắc khoảng 1 giờ đồng hồ.

Our system's summary

Theo đó, vào khoảng thời gian trên, khi chiếc xe con mang BKS: 37A-048.45 đang điều khiển xe lưu thông theo hướng Bắc - Nam bất ngờ lao lên dải phân cách làn đường ô tô và xe máy. Các nhân viên của trạm soát vé cầu Bến Thuỷ đã phải ra điều tiết giao thông, tránh ùn tắc nghiêm trọng trên tuyến đường.

5.4. Evaluation and comparison

To evaluate our system, we reimplemented some basic methods to be used as the baseline systems, including LexRank [12], LEAD [40], and TextRank [26]. These methods were chosen because they provided good results in extractive summarization and easy to implement. These systems were experimented with the CNN and Baomoi datasets. In addition, we compared our system with the state-of-the-art models in [31] which used deep learning and reinforcement learning techniques. Two models in [31] that were compared with our system are REFRESH and Cheng and Lapata's one. Since the models in [31] are hard to reimplement, we reused the results reported in that paper for the CNN dataset. Table 6 shows the performance comparison between our system and the systems mentioned above, using the CNN and Baomoi datasets.

Table 6. Comparison and evaluation results of the methods. Marks with '*' denotes the systems that we re-implement, marks with '-' denotes that the systems do not implement on the corresponding dataset (Note that our model is M-BERT+CNN+Encoder-Decoder +FC+TF-IDF+MMR).

Methods	**CNN**			**Baomoi**		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
LexRank [12]*	22.9	6.6	17.2	38.5	17.0	28.9
LEAD [40]*	29.0	10.7	19.3	46.5	20.3	30.8
TextRank [26]*	26.0	7.3	19.2	44.7	19.2	32.9
Cheng and Lapata [31]	28.4	10.0	25.0	-	-	-
REFRESH [31]	30.4	11.7	26.9	-	-	-
Our model	**32.67**	**13.04**	**29.53**	**54.67**	**25.26**	**37.48**

Tables 6 indicates that our proposed system is more efficient than LexRank, TextRank, LEAD, and modern models in [31]. The nearest model to ours is REFRESH, which also uses the convolutional layer and the max pooling layer to encode input sentences. Our model is different from the REFRESH in several points. The M-BERT is applied to our system to obtain token embeddings of the input sentences, which are used as the input for the convolutional layer and k-max pooling layer. We believe that this is a powerful way to represent the input text. The biLSTM which is used in our encoder-decoder provides more information than the LSTM using in the REFRESH. Experimental results in Table 6 have shown that integrating the TF-IDF information and the MMR technique into the system providing better performance than the REFRESH.

There are only a few works on Vietnamese text summarization. Most of them used small datasets with different characteristics than ours to evaluate the system. Therefore, we cannot compare our models with such systems. Nguyen et al. [38] used a set of 30 articles collecting from Vietnamese newspapers with the summaries being generated manually by selecting salient sentences and compressing them. Truong and Nguyen [37] used two datasets, one with only five paragraphs, and another with 25 articles. Nguyen [15] constructed a dataset of 300 articles, with the summary being generated manually by human. The author used different measures than us to evaluate the result.

The research whose evaluation dataset is quite similar to us is Lam et al. [21]. This dataset was collected automatically from newspapers with $31,429$ articles for training, $5,893$

articles for validating, and 1,964 articles for testing. The evaluation results on the test set are 39%, 31.4%, 36.7% F_1−score for $R-1$, $R-2$, and $R-L$, respectively. Our F_1−scores of $R-1$ and $R-L$ are higher than that of the system in [21], but the $R-2$ score is a little lower. However, since the dataset in [21] is smaller than ours, the above comparison only has a relative meaning.

We also reimplemented the method in [22] and compared it with our system using the DUC 2001/DUC 2002 dataset. The system in [22] uses word embeddings to represent text and CNN to extract summarizing sentences. Since the approach in [22] is not suitable for the Vietnamese language, we did not compare this approach with ours using the Baomoi dataset. Experimental results with the DUC 2001 as the training dataset and the DUC 2002 as the test one are shown in Table 7. Since we only find one research [22] using the same dataset, we compare our system with that research.

Table 7. Experimental results on the DUC 2001 and DUC 2002 datasets. Marks with '*' denotes the system re-implemented by us. Marks with '-' denotes that the system was not implemented on the corresponding dataset.

Methods	**DUC 2001/DUC 2002**		
	R-1	**R-2**	**R-L**
Laugier et al.[22]	42.48	16.96	-
Laugier et al.[22]*	41.83	16.78	-
M-BERT+CNN+Encoder-Decoder+FC+TF-IDF+MMR	**48.29**	**23.40**	**43.80**

Table 7 points out that our proposed system providing significant improvement comparing to the system in [22]. Experiments above have proved that our proposed system can provide promising results with both English and Vietnamese languages, comparing to existing research in extractive text summarization.

6. CONCLUSION AND FUTURE WORKS

The paper has proposed an efficient extractive single-document summarization system using the M-BERT model to embed the input document, combining with the TF-IDF feature as input for the model that uses the CNN (with k-max pooling), the Encoder-Decoder, the FC layer, and the MMR method to generate a summary. Experiments with the proposed system on the DUC 2001, DUC 2002, CNN datasets (for English), and the Baomoi dataset (for Vietnamese) provided better results than existing methods that were published. These results show that our proposed text summarization system has been highly effective for both English and Vietnamese text. In the proposed system, we used the M-BERT model to embed the input document that was limited in length (since the M-BERT model limits the length of the input document). In the future, we will apply the optimized pre-trained BERT models, such as RoBERTa [24], PhoBERT [32] to embed the input document, or the Generative Pre-Training (GPT)[36] to further improve the quality of the system's summary. A method to replace the biLSTM in the encoder-decoder by a transformer is also considered in our future work. In addition, we will investigate a method to apply the proposed single-document summarization model to develop a multi-document summarization system for both English and Vietnamese languages in our next research.

APPENDIXES

Appendix 1: A source text in the CNN dataset

Panama City, Panama (CNN) Ending a decade - long standstill in U.S. - Cuba relations, President Barack Obama met for an hour Saturday with his Cuban counterpart Raul Castro, the first time the two nations' top leaders have sat down for substantive talks in more than 50 years. The meeting in a small conference room on the sidelines of the Summit of the Americas came as the two countries work to end the Cold War enmity that had led to a total freeze of diplomatic ties. And while both leaders proclaimed progress had been made, a key stumbling block – Cuba's place on the U.S. list of countries that sponsor terror – remained unresolved. "This is obviously an historic meeting, "Obama said at the beginning of his session with Castro, claiming that decades of strain had done little to benefit either Cubans or citizens of the United States. "It was time for us to try something new," he said. "We are now in a position to move on a path toward the future. "Castro, who earlier in the day said he trusted Obama, acknowledged there would be difficult stumbling blocks as his nation works to repair ties with the United States. But he said those differences could be surmounted. "We are willing to discuss everything, but we need to be patient, very patient," Castro said. "We might disagree on something today on which we could agree tomorrow. "Speaking to reporters after his session with Castro, Obama said the meeting was "candid and fruitful" and could prove to be a "turning point" in his push to defrost ties with Cuba. But he said he hadn't yet decided whether to remove Cuba's designation as a state sponsor of terror, an outcome that had previously been expected during the summit. The State Department provided Obama with a review of the terror status this week. "I want to make sure I have a chance to read it, study it before we announce publicly what the policy outcome is going to be," Obama said. "But in terms of the overall direction of Cuba policy, I think there is a strong majority both in the United States and in Cuba that says our ability to engage, to open up commerce and travel and people to people exchanges is ultimately going to be good for Cuban people. "On Friday night, Obama and Castro greeted each other courteously amid an explosion of camera flashes, shaking hands before dining at the inaugural session of the conference. The two sat at the same table but not directly next to one another. Before Obama arrived in Panama on Wednesday, he spoke with Castro by phone, laying the groundwork for what will become a new era of relations between the neighboring countries. "The Cold War has been over for a long time, "Obama said during opening remarks at the summit Saturday. "I'm not interested in having battles, frankly, that began before I was born. "That exhortation, however, seemed to be lost on Castro himself, who expanded what was meant to be a six - minute speech into a 50 - minute address lecturing leaders on Cuba's revolution and giving a litany of perceived grievances to Cuba over the past 50 years. But he distinguished Obama from past American presidents, saying he respected Obama's move toward reconciliation. "In my opinion, President Obama in an honest man, "Castro said through an interpreter. "I admire him, and I think his behavior has a lot to do with his humble background. "U.S. administration official said Castro's long list of grievances was expected, despite the move toward diplomatic ties. "(What's) unique and new is what he said about the president, "the official said of Castro's praise for Obama. Obama announced in December that he was seeking to renew diplomatic relations with Cuba after half a century of strife, including eventually opening embassies in Washington and Havana. Obama set to test engagement doctrine with Cuba in Panama.

His meeting with Castro on Saturday isn't being billed as a formal bilateral session, but Obama's aides are still characterizing the event as the highest - level engagement with the Cuban government since then - Vice President Richard Nixon met with Fidel Castro in 1959. "We're in new territory here, "Ben Rhodes, Obama's deputy national security adviser, said Friday. "The reason we're here is because the President strongly believes that an approach that was focused entirely on isolation, focused entirely on seeking to cut off the Cuban people from the United States of America had failed. "The overtures to Cuba have not been universally popular in the United States; some lawmakers were irate that Obama was seeking to engage what they regard as a corrupt government. "A recommendation to remove Cuba from the list of State Sponsors of Terrorism would represent another significant misstep in a misguided policy, "Sen. Bob Menendez, a Democrat who used to the chair the Foreign Relations Committee, wrote in a statement last week. In Latin America, however, Obama was receiving a warm welcome after announcing he was seeking to engage Havana in talks over reopening embassies and removing barriers to commerce and travel. 9 things you wanted to ask about the Cuban embargo He noted to applause during a session Friday that this was the first summit with Cuba in attendance. And he's cast the decision to reopen the U.S. relationship with Cuba as beneficial to the entire hemisphere, which has also embraced his immigration executive action. But even as Obama landed in Panama, the longstanding gulfs between the two countries ' governments were on display. Dissidents opposed to Castro's regime were violently accosted this week by supporters of the Cuban government, a scuffle the White House said was unacceptable. "As we move toward the process of normalization, we'll have our differences, government to government, with Cuba on many issues – just as we differ at times with other nations within the Americas, just as we differ with our closest allies," Obama said at a meeting of civil society leaders Friday. "There's nothing wrong with that." "But I 'm here to say that when we do speak out, we're going to do so because the United States of America does believe, and will always stand for, a certain set of universal values, "he said. The long history between the U.S. and Cuba Obama closed out his time in Panama with a news conference where he covered topics ranging from Hillary Clinton's expected presidential announcement to his framework deal with Iran on its nuclear program. The President had pointed criticism for Sen. John McCain, R - Arizona. Earlier this week, McCain accused Secretary of State John Kerry of intentionally mischaracterizing what the sides had agreed to in the Iran nuclear deal. "John Kerry is delusional, "McCain said on the Hugh Hewitt show, a conservative talk radio program, adding that the view from the Supreme Leader of Iran of the provisions agreed to" is probably right, " rather than what the United States maintains are the agreed provisions. While discussing the Iran agreement Saturday, Obama brought up those remarks without being asked. "When I hear someone like Sen. McCain recently suggest that our secretary of state, John Kerry, who served in the United States Senate, (is) a Vietnam veteran, who's provided exemplary service to this nation, is somehow less trustworthy of the interpretation of what's in a political agreement than the Supreme Leader of Iran, that's an indication of the degree to which partisanship has crossed all boundaries," he said at the news conference. After the President's remarks, McCain tweeted "So Pres. Obama goes to # Panama, meets with Castro and attacks me - I 'm sure Raul is pleased." As for his 2008 Democratic rival, Obama said, "If she decides to run, if she makes an announcement, she's going to have some strong messages to deliver," he said.

Appendix 2: A source text in the Baomoi dataset

Vụ tai nạn nói trên xảy ra vào khoảng 23h15 ngày 19/1, tại cầu Bến Thuỷ, trên Quốc lộ 1A. Theo đó, vào khoảng thời gian trên, khi chiếc xe con mang BKS: 37 A-048. 45 đang điều khiển xe lưu thông theo hướng Bắc - Nam bất ngờ lao lên dải phân cách làn đường ô tô và xe máy. Hiện trường vụ tai nạn. Sau cú đâm mạnh, đầu chiếc xe con bị hư hỏng, toàn bộ chiếc xe con bị nằm gác trên dải phân cách đường. Những người ngồi trên chiếc xe con này đã không có ai bị thương. Do buồn ngủ, thay vì đi sang làn đường cho ô tô, chiếc xe con này chọn con đường riêng bằng cách leo lên dãi phân cách. Ngay sau khi vụ tai nạn xảy ra, chủ xe đã tiến hành gọi cứu hộ giao thông đến để giải cứu chiếc xe này. Tuy nhiên, phải mất gần giờ đồng hồ sau, chiếc xe này mới tách ra khỏi dải phân cách. Sau khi "lỡ" leo lên dải phân cách, đầu của chiếc xe đã bị hư hỏng nặng. Vụ tai nạn xảy ra ngay trên đầu cầu nên đã khiến giao thông qua đây bị hỗn loạn. Các nhân viên của trạm soát vé cầu Bến Thuỷ đã phải ra điều tiết giao thông, tránh ùn tắc nghiêm trọng trên tuyến đường. Sau khi bị tai nạn, chủ xe đã gọi điện nhờ xe cứu hộ đến để giải quyết vụ việc.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their useful comments and valuable suggestions, which help us to improve and clarify this manuscript.

This research is supported by Hanoi University of Science and Technology (HUST) under the grant project No. T2020-PC-208.

REFERENCES

[1] "https://viblo.asia/p/fine-tuning-pretrained-model-trong-pytorch-va-ap-dung-vao-visual-saliency-prediction-4p856ly1zy3."

[2] "https://www-nlpir.nist.gov/projects/duc/guidelines/2001.html."

[3] "https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html."

[4] "What is rouge and how it works for evaluation of summarization tasks?" *url:https://rxnlp.com/how-rouge-works-for-evaluation-of-summarizationtasks/#.XOO5Z8j7TIW*, 2019.

[5] K. Al-Sabahi, Z. Zuping, and M. Nadher, "A hierarchical structured self-attentive model for extractive document summarization (hssas)," *IEEE Access*, vol. 6, pp. 24 205–24 212, 2018.

[6] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: A brief survey," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, no. 10, 2017.

[7] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 335–336.

[8] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[10] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 670–680.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

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

[13] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference On Acoustics, Speech And Signal Processing*. IEEE, 2013, pp. 6645–6649.

[14] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258–268, 2010.

[15] N. T. T. Ha, "Developing some vietnamese text summarization algorithms using semi-supervised learning method," *Military Technical Academy*, 2012.

[16] K. M. Hermann, T. Kočiskỳ, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, 2015, pp. 1693–1701.

[17] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd Annual Meeting of the Association For Computational Linguistics And the 7th International Joint Conference on Natural Language Processing (volume 1: Long papers)*, 2015, pp. 1681–1691.

[18] A. Jadhav and V. Rajan, "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 (volume 1: Long papers)*, 2018, pp. 142–151.

[19] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics., 2014.

[20] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1746–1751.

[21] Q. T. Lam, T. P. Pham, and D. H. Do, "Vietnamese text summarization with sequence-to-sequence," *Scientific Journal of Can Tho University. Special topic: Information Technology*, pp. 125–132, 2017 (Vietnamese).

[22] L. Laugier, E. Thompson, and A. Vlissidis, "Extractive document summarization using convolutional neural networks - reimplementation," *Available: https://www.semanticscholar.org/paper/Extractive-Document-Summarization-Using-Neural-Laugier/ed0f189bbbbccceefb41ccb36e1c5b62bc36d2fb*, 2018.

[23] Y. Liu, “Fine-tune bert for extractive summarization,” *arXiv preprint arXiv:1903.10318*, 2019.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[25] H. C. Manh, H. Le Thanh, and T. L. Minh, “Extractive multi-document summarization using k-means, centroid-based method, mmr, and sentence position,” in *Proceedings of the Tenth International Symposium on Information and Communication Technology*, 2019, pp. 29–35.

[26] R. Mihalcea and P. Tarau, “Textrank: Bringing order into text,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 404–411.

[27] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, 2013, pp. 3111–3119.

[28] L. N. Minh, A. Shimazu, H. P. Xuan, B. H. Tu, and S. Horiguchi, “Sentence extraction with support vector machine ensemble,” in *Proceedings of the First World Congress of the International Federation for Systems Research: The New Roles of Systems Sciences for a Knowledge-based Society*. JAIST Press, 2005.

[29] D. Misra and Landskape, “Mish: A self regularized non-monotonic activation function,” *arXiv preprint arXiv:1908.08681*, 2020.

[30] R. Nallapati, F. Zhai, and B. Zhou, “Summarunner: A recurrent neural network based sequence model for extractive summarization of documents,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[31] S. Narayan, S. B. Cohen, and M. Lapata, “Ranking sentences for extractive summarization with reinforcement learning,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2018, pp. 1747–1759.

[32] D. Q. Nguyen and A. T. Nguyen, “Phobert: Pre-trained language models for vietnamese,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 1037–1042.

[33] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[34] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual bert?” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4996–5001.

[35] V. N. Quoc, H. Le Thanh, and T. L. Minh, “Abstractive text summarization using lstms with rich features,” in *International Conference of the Pacific Association for Computational Linguistics*. Springer, 2020, pp. 28–40.

[36] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” *https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_ understanding_paper.pdf*, 2018.

[37] N. Q. D. Truong, Quoc Dinh, “A method to summarize vietnamese text automatically,” in *Proceeding of the 15th National Conference Some Selected Issues of Information and Communication Technology*, 2012.

[38] N. Q. Uy, P. T. Anh, T. C. Doan, and N. X. Hoai, “A study on the use of genetic programming for automatic text summarization,” in *2012 Fourth International Conference on Knowledge and Systems Engineering.* IEEE, 2012, pp. 93–98.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.

[40] M. Wasson, “Using leading text for news summaries: Evaluation results and implications for commercial summarization applications,” in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, 1998, pp. 1364–1368.

[41] Y. Wu and B. Hu, “Learning to extract coherent summary via deep reinforcement learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[42] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-h. Sung *et al.*, “Multilingual universal sentence encoder for semantic retrieval,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 87–94.

[43] X. Zhang, M. Lapata, F. Wei, and M. Zhou, “Neural latent extractive document summarization,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 779–784.

[44] X. Zhang, F. Wei, and M. Zhou, “Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5059–5069.

[45] Y. Zhang, J. E. Meng, and M. Pratama, “Extractive document summarization based on convolutional neural networks,” in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society.* IEEE, 2016, pp. 918–922.

[46] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X.-J. Huang, “Extractive summarization as text matching,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6197–6208.

[47] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, “Neural document summarization by jointly learning to score and select sentences,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 654–663.

[48] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 19–27.

Received on April 04, 2021
Accepted on May 06, 2021