Improving Quality of Vietnamese Text Summarization Based on Sentence Compression

Ha Nguyen Thi Thu
Information Technology Faculty
Electric Power University
Hanoi, Vietnam

Cuong Nguyen Ngoc
Department of Computer and Mathematics
The people's Security University
Hanoi, Vietnam

Tu Nguyen Ngoc
Information Technology Faculty
Electric Power University
Hanoi, Vietnam

Hiep Xuan Huynh
Information Technology Faculty
CanTho university
Hanoi, Vietnam

***Abstract*—Sentence compression is a valuable task in the framework of text summarization. In previous works, the sentence is reduced by removing redundant words or phrases from original sentence and tries to remain information. In this paper, we propose a new method that used Grid Model and dynamic programming to calculate n-grams for generating the best sentence compression. These reduced sentences are combined to text summarization. The experimental results showed that our method really effective and the text is grammatically, coherence and concise.**

***Keywords*—Sentence compression; topic modeling; text summarization; Grid model; n-grams; dynamic programming**

I. INTRODUCTION

Text summarization is technique allows computers automatically generated text summaries from one or more different sources. To base oneself on features of the main content and to recapitulate content from original documents that text summarization is one of the fields is interested in researchers from the 60's of the 20th century and it is still a hot topic of the forums and seminars on the current world [1].

The traditional text summarization method usually bases on extracted sentences approach [1], [9]. Summary is made up of the sentences were selected from the original. Therefore, in the meaning and content of the text summaries are usually sporadic, as a result, text summarization lack of coherent and concise. Figure 1 below illustrates the approach to extract sentences. Text summarization has one sentence with extraction rate 30%.

Vừa qua, Microsoft đã chính thức ra mắt người dùng bản cập nhật Service Pack đầu tiên dành cho Windows 7. Nhưng liệu chúng ta có nên tốn thời gian để cài đặt bản cập nhật này cho Windows không. Bạn chưa bao giờ cập nhật các bản Service Pack dành cho Windows của mình. Trước tiên, chúng ta cần lưu ý rằng Service Pack thực chất chỉ là một gói tổng hợp các bản cập nhật nhỏ lẻ đã được tung ra từ trước đó thông qua Windows Update với việc sửa chữa các lỗi và bổ sung một vài thay đổi. Microsoft tung ra Service Pack nhằm mục đích giúp những người dùng phải cài lại hệ điều hành vì một lý do nào đó có thể cập nhật Windows một cách nhanh chóng chỉ thông qua một gói tải về duy nhất.Tăng cường chất lượng âm thanh của thiết bị dùng cổng HDMI. In tài liệu XPS với nhiều khổ giấy khác nhau. Thay đổi hành động của tính năng. Hỗ trợ Advanced Vector Extensions. Hỗ trợ Advanced Format 512e cho các thiết bị lưu trữ.

Figure 1.a. Original text

Thay đổi hành động của tính năng. Hỗ trợ Advanced Vector Extensions. Hỗ trợ Advanced Format 512e cho các thiết bị lưu trữ.

Figure 1.b. Result with extraction rate 30%

Fig. 1. Extraction summary

Some other text summarization methods are the problem of natural language processing that made summary has a good linguistic score and seamlessly coherence the content of the original. One of its is a sentence compression technique [2], [3], [7]. With the compression approach, researchers focused using supervised learning techniques or using legal vocabulary or deep level language analysis techniques based on syntax tree [10]. These methods have the following characteristics:

- High cost when building the corpus for training.
- Need a long time for construction meticulously by language experts, especially construction corpus related legal vocabulary.
- Higher computational complexity.

Therefore, in this paper, we use a sentence compression method to create a text summary basing on grid model with the target:

- Use unsupervised learning to reduce costs.

- Use unsupervised learning techniques to not waste time to build corpus crafts.
- Minimize computational complexity by using dynamic programming algorithm.

The rest of the paper is organized as follows: In section 2, we will introduce some related works. In section 3 is the presentation of our method for Vietnamese feature reduction, the methodology of Vietnamese sentence compression is presented in section 4. Experiments and results will show in section 5. And finally, section 6 is a conclusion and future works.

II. Related Works

The sentence compression task is defined as the curtailment of redundant components, in sentence to produce a shorter sentence. Figure 2 below is an example

Tháng 11 ~~năm ngoái~~, Quốc hội Việt Nam thông qua dự án xây hai nhà máy điện hạt nhân ~~đầu tiên của Việt Nam~~ tại ~~tỉnh~~ Ninh Thuận.

Figure 2.a. Original sentence

Tháng 11, Quốc hội Việt Nam thông qua dự án xây hai nhà máy điện hạt nhân tại Ninh Thuận.

Figure 2.b. Reduced sentence

Fig. 2. The task of sentence compression

Text summarization is based on a sentence compression approach, allows connecting multiple sentences were reduced to make a shorter document that have the meaning and grammar are accepted, to guarantee a coherent level of content and meaning.

Some studies of sentence compression showed the importance of this approach in the problem text summarization. The first people who proposed sentence compression model is Jing and McKeown in 2000, they presented one of the earliest approaches on sentence compression using machine learning and classifier based techniques. This research work was focused on removing inessential phrases in extractive summaries based on an analysis of human written abstracts. In their experiments, they used human-written abstracts, and the corpus was collected from the free daily news and headlines provided by the Benton Foundation [4].

Noisy channel is the typical of the sentence compression method, in studies of Marcu et al., they suggested two methods for sentence compression: one is the noisy channel model where the probabilities for sentence compression (P{compress|S)} 1) are estimated from a training set (Sentence, Sentencecompress) pairs, manually crafted, while considering lexical and syntactical features. The other approach learns syntactic tree rewriting rules, defined through four operators: SHIFT, REDUCE DROP and ASSIGN [9].

In the work of Le Nguyen and Ho in 2004, two sentence compression algorithms were also proposed. The first one is based on template translation learning, a method inherited from the machine translation field, which learns lexical transformation rules, by observing a set of 1500 (Sentence, Sentencereduced) pairs, selected from a website and manually tuned to obtain the training data. Due to complexity difficulties found in the application of this big lexical rule set, they proposed an improvement where a stochastic Hidden Markov Model is trained to help in the decision of which sequence of possible lexical reduction rules should be applied to a specific case [11].

Some other works used unsupervised approach. Turner and Charniak, in their work, corpus for training are automatically extracted from the Penn Treebank corpus, to fit a noisy channel model [3], similar to the one used by Knight and Marcu [8]. And Clarke and Lapata devise a different and quite curious approach, where the sentence compression task is defined as an optimal goal, from an Integer Programming problem. Several constraints are defined, according to language models, linguistic, and syntactical features. Although this is an unsupervised approach, without using any parallel corpus, it is completely knowledge driven, like a set of craft rules and heuristics incorporated into a system to solve a certain problem [2].

All these works applied to English. For Vietnamese, there are some methods for sentence compression. Minh Le Nguyen et.al and Ha Nguyen Thi Thu et. al. Minh Le Nguyen proposed two methods for sentence compression, one of its applied HMM to Vietnamese sentence compression and other used syntax control for reducing sentences [10], [11]. Ha Nguyen Thi Thu using unsupervised learning and supervised learning for creating Vietnamese text summarization based on sentence compression [5], [6].

III. Vietnamese Text Feature Reduction

A. Feature reduction problem

Considering a number of applications such as in a data processing system (the voice signal, image or pattern recognition generally) if we consider setting of features as a set of vectors of real value. Assuming that the system is only effective if the dimension vector of each individual is not too large.

The problem of dimensionality reduction occurs when data have greater dimension processing capabilities of the system [16]. Example: A face recognition/classification system based on multi-level gray image that has size mxn, corresponding to mxn dimensional vector of real value. In the experiment, an image may have m = n = 256 or 65536 dimensions. If using a multilayer perceptron network to perform the classification system. It will become difficult to build an MLP.

Therefore, the feature reduction is an important problem when we work with the data that has many features such as image, voice, text, ... The feature reduction illustrated like following figure

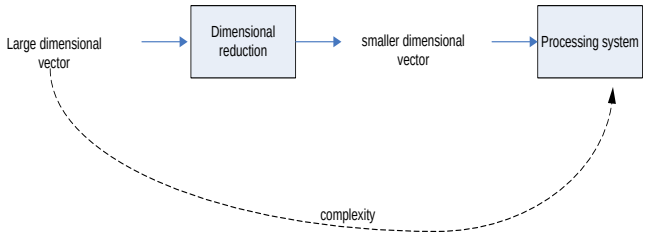

Fig. 3. Dimensional vector reduction model

B. Methodology of Vietnamese text feature reduction

Define 1: (topic word): Topic word is the nouns that have been extracted from sentences.

Example 1: *table, human, computer,…* is the topic words

For the Vietnamese text, some text processing problem often uses word segmentation tool for separate words in text. In previous works, we proposed a method for feature reduction that is published in [5], [6]. Documents can reduce complexity computing of large feature set by using a word, segmentation tool for separating word into two word sets: nouns set (called topic word) and other words set. In any text, nouns contain information of the text. So, when we extract nouns from text, a remarkable reduction of large feature set.

Example 2: Have an original Vietnamese text include 34 words.

"*Các nhà nghiên cứu thuộc trường Đại học Michigan vừa tạo ra một nguyên mẫu đầu tiên cho hệ thống tính toán quy mô nhỏ, có thể chứa dữ liệu một tuần khi tích hợp chúng vào trong những bộ phận rất nhỏ như mắt người.*"

Translate in to English:

"Researchers at the University of Michigan have created a first prototype system for small-scale computing, which can contain data for a week while integrating them into very small parts as the human eyes."

Like this document, we must calculate weight for 34 words. And representation matrix with 1 row and 34 columns like below

$$T = \{t_{1,1}, t_{1,2}, ..., t_{1,34}\} \quad (1)$$

In Example 2, we separate document d into two sets, the first set include noun and the second set is remain of words.

Noun set T' = {*nhà, nghiên_cứu, trường, đại_học, Michigan, nguyên_mẫu, hệ_thống, quy_mô, dữ_liệu, tuần, chúng, bộ_phận, mắt, người*}.

Other set O'= {*Các, thuộc, vừa, tạo, ra, một, đầu_tiên, cho, tính_toán, nhỏ, có_thể, chứa, một, khi, tích_hợp, vào, trong, những, rất, như* }.

Use text separation technique in two sets, the size of the matrix T will be reduced, for example, with the original text in example 1, instead of using the T matrix contains one row and 34 columns, we only need the matrix T' consists of one row and 14 columns:

$$T = \{t_{1,1}, t_{1,2}, ..., t_{1,14}\} \quad (2)$$

IV. Using Grid Model for Vietnamese Text Summarization

Methods Vietnamese sentence compression based on unsupervised learning techniques with grid model combined with dynamic programming to choose the best sentence shortened. The calculation is based on the set of noun in a sentence that limit the loss of information in a sentence.

To calculate the probability of a sequence P $(w_1 , w_2 , .. , w_n)$. Use chain rule of probability:

$$P(X_1...X_n) = P(X_1)P(X_x \mid X_1)P(X_3 \mid X_1^2)...P(X_n \mid X_1^{n-1}) = \prod_{k=1}^{n} P(X_n \mid X_1^{k-1}) \quad (3)$$

Apply chain rule for the words, to receive:

$$P(w_n^n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1}) = \prod_{k=1}^{n} P(w_n \mid w_1^{k-1}) \quad (4)$$

With N - grams, the conditional probability approximation of the next word in the sequence is

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-N+1}^{n-1}) \quad (5)$$

In this paper, we use bi-gram to reduce complexity in calculation, therefore, when using the bi-gram model to predict the conditional probability of the next word can use approximately formula as follows:

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-1}) \quad (6)$$

Define 2. (Word substring): *Word substring is initialized by a topic word and stop with a topic word, no any topic word between its.*

Example 3: In his memory, **Flowers** bloom along the **river.**

Define 3. (T**he most Likelihood word substring**) *is the word substring in which, every word has maximum bi-gram.*

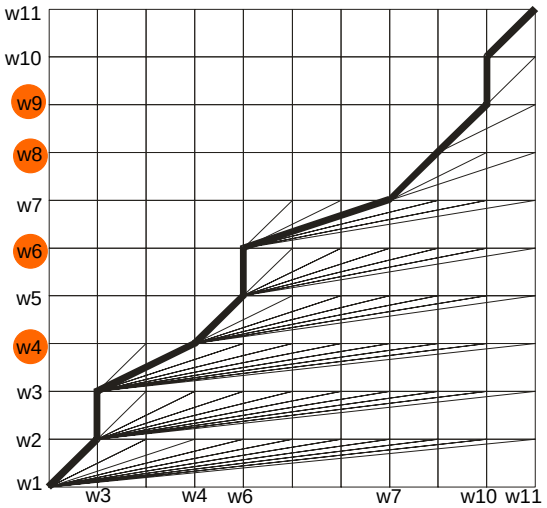

Fig. 4. Grid model for a 11 words

Suppose there are 11 word in the sentence W , W is represented by :

$$W = \{w_1, w_{2,}..., w_{11}\} \quad (7)$$

Using word identification and word segmentation tools into two sets of a separate word, inside the collective noun includes words w_4, w_6, w_8, w_9. The grid model with this sentence is created as Figure 4 below

In the figure 4 illustrated a original sentence, that has 11 words. In which, w_4, w_6, w_8, w_9 are all the topic words (nouns). We have 3 word substrings. Reduced sentence can be generated by these steps:

- Step 1: Let start with the first word substring $w_1..w_4$: Initialize with w_1, w_1 has 3 ways: $w_1 \rightarrow w_2$, $w_1 \rightarrow w_3$, $w_1 \rightarrow w_4$.
- Step 2: Calculate probability of these ways from w_1. $S_{12}=w_1 \rightarrow w_2$, $S_{13}=w_1 \rightarrow w_3$, $S_{14}=w_1 \rightarrow w_4$. After that, chose the most likelihood probability.
- Step 3: Track the point that has the most likelihood probability. If not the end of word substring, Loop step2: continue to new way from track point to another word in word substring.
- Step 4: Continue with the next word substring
- Step 5: Reduced sentence is the projection of words on the horizontal line.

SENTENCE COMPRESSION BASED ON GRID MODEL ALGORITHM

Input
W: original sentence;
Output
S: reduced sentence;
Initialization
$i \leftarrow 0; f \leftarrow 0; j \leftarrow 0; S \leftarrow \emptyset; N \leftarrow \emptyset; O \leftarrow \emptyset;$
1. Separate sentence W to two word sets.
For *i=1* **to** ***length(W)*** **do**
If *w(i)* is noun **then**
$N \leftarrow w(i);$
Else
$O \leftarrow w(i);$
2. **Calculate initial probability**
For each *w(i)* in sentence W
If *Pr(w(i)|start) > 0* **then**
f=i; // start of reduced sentence.
$S = S \cup w(f)$
3. **Representation w(i) on the Grid**
4. **Generate sentence compression**
Loop
While not end sentence W **do**
For *i=f* **to** ***length(W)*** **do**
If *w(i)* is noun **then break**;
For *j= f* **to** *i* **do**
K=argmax(Pr(w(j)|w(f))
f= j;
$S = S \cup w(j);$
End Loop

Fig. 5. Sentence compression based on Grid model

Like figure 4, reduced sentence contain these words: w_1, w_3, w_4, w_6, w_7, w_8, w_9, w_{10}, w_{11}. Figure 5. is the algorithm of Vietnamese sentence Compression using grid model.

V. EXPERIMENTAL

A. Corpus

Our experiment used the corpus of 100 Vietnamese text. We collected from Vnexpress online news (http://VnExpress.Net). We then used the VLSP word segmentation tool (http://vlsp.vietlp.org:8080/demo/?page=seg_pos_chunk) to segment Vietnamese text into words. After correcting them manually, we obtained more than 200,000 words, which was used in previous works [5].

B. Building text summarization system

This system has been built based on our proposed and use for automatic testing and experimental. Here is the interface of system.

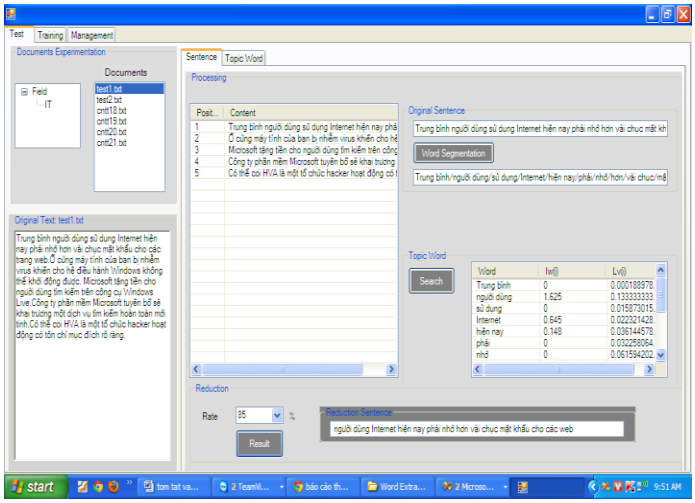

Fig. 6. Text summarization system

It's difficult to compare our method with previous ones, because there were no widely accepted benchmarks for Vietnamese text reduction sentence. Therefore, we compare our proposed method with manual sentence compression generated by humans, called Human, sentence compression method using syntax control, called Syn.con, proposed by M.L. Nguyen [10] and another our method based on determining likelihood substring (Called DLSS) [5].

C. Results

In this experiment, we use the evaluation way as Knight and Marcu [9]. Table I shows the sentence compression results that are carried out by our method, Human and Syn.con for Vietnamese text.

TABLE I. EXPERIMENTAL RESULTS

Method	Compression	Grammatically	Information
Baseline	x	X	X
DLSS	63.26	6.83 ± 1.3	6.78 ± 1.2
Our method	78.1	8.2	8.8
Human	61.2209	8.333333	8.34524
Syn.con	67	6.5 ± 1.7	6 ± 1.1

Table 1 shows compression ratios in the second column, which indicates that the lower the compression ratio the shorter the reduced sentence. The Grammaticality in the third column, which indicates the appropriateness of reduced sentence in term of grammatical.

VI. CONCLUSION

Many Vietnamese text summarization researches were published that showed the importance of Vietnamese information processing problem today. In this paper, the text summarization method based on the sentence compression approach to the target reduce time when applying unsupervised learning method and to not waste cost to build corpus crafts and to reduce computational complexity by using dynamic programming algorithm. The experimental results illustrate our approach is satisfactory requirement of the text summary and can be applied to a number of different languages .

ACKNOWLEDGEMENTS

We would like to thank the experts of University of Engineering and technology, Vietnam of University, and Japan Advanced Institute of Science and Technology , Dr Nguyen Le Minh, Dr Nguyen Van Vinh**,** Dr Nguyen Phuong Thai, Vu Xuan Luong for their great help in building the experimental summarizing application and Vietnamese dictionary corpus.

REFERENCES

[1] Ani Nenkova and Kathleen McKeown, Automatic Summarization, Foundations and Trends in Information Retrieval, Vol. 5, No. 2–3 , p 103–233, 2011.

[2] Clarke, J., & Lapata, M.. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 377–384, 2006.

[3] Clarke, J., & Lapata, M., Global inference for sentence compression: An integer linear programming approach. Journal of Artificial Intelligence Research, 31, 399–429, 2008.

[4] Hongyan Jing and Kathleen R. McKeown. Cut and paste based text summarization, In Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2000), pages 178–185, 2000.

[5] Ha Nguyen Thi Thu, Quynh Nguyen Huu "Method of Sentence reduction in Vietnamese Text Based on Determining Likelihood Substring". International Conference on intelligen Network and Computing, November, 2010.

[6] Ha Nguyen Thi Thu and An Nguyen Nhat, A Method for Generating Vietnamese Text Sentence Reduction Based on Bayesian Network, International Journal of Innovative Computing, Information and Control, pp 407-416, Vol 11, No2. , 2015.

[7] Hal Daume´ III and Daniel Marcu, A Noisy-Channel Model for Document Compression, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 449-456.

[8] Turner, J., & Charniak, E. Supervised and unsupervised learning for sentence compression. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 290–297, 2005.

[9] Knight, K., & Marcu, D. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. Artificial Intelligence, 139 (1), 91–107, 2002.

[10] M.L. Nguyen and S. Horiguchi, "A Sentence Reduction Using Syntax Control", Proc. Of 6th Information Retrieval with Asian Language, pp. 139-146, 2003.

[11] M.L. Nguyen and S. Horiguchi, "Example-Based Sentence Reduction Using the Hidden Markov Model" ACM Transactions on Asian Language Information Processing, Vol. 3, No. 2, pp146-158, 2004.

[12] Maria Soledad Pera and Yiu-Khai Ng, A Naïve Bayes Classifier for web document summaries created by using word similarity and significant factors, International Journal on Artificial Intelligence Tools, Vol. 19, No. 4, pp. 465–486, 2010.

[13] Trevor Cohn, Mirella Lapata, Sentence Compression as Tree Transduction, Journal of Artificial Intelligence Research 34, pp. 637-674,2009.

[14] Youngjoong Ko, Jinwoo Park, Jungyun Seo, Improving text categorization using the importance of sentences, Information Processing and Management 40, pp, 65–79, 2004.

[15] Ziegler, C. and M. Skubacz, 2007. Content extraction from news pages using particle swarm optimization on linguistic and structural features. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Nov. 2-5, IEEE Computer Society Washington, DC., USA., pp: 242-249.

[16] Galavotti et al., 2000] Galavotti, L. –Sebastiani, F. –Simi, M.: Feature selection and negative evidence in automated text categorization. Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, ECDL-00, Lisbon, 2000