Concatenate the Most Likelihood Substring for Generating Vietnamese Sentence Reduction

Ha Nguyen Thi Thu and Quynh Nguyen Huu

***Abstract*—Sentence Reduction is a valuable task in the framework of text summarization. In previous works, sentence is reduced by removing redudant words or phrases from original sentence and try to remain important information. In this paper we propose a new method that used Viterbi algorithm for find the most likelihood substring and then concatenate them to generate sentence reduction. Reduced sentence not only remain important information from original sentence, grammatically is ensured. The experimental results shown that, our method better than previous works and closely method that has done by human.**

***Index Terms*—Important word, Likehood substring, Sentence reduction, Vietnamese text, Virtebi algorithm.**

I. Introduction

The Goal of sentence reduction is create a system to automatically reduce the length of sentences by removing some of the words while attempting to create output sentences that:

- are grammatical
- capture the most important semantic elements of the original sentences
- make sense

There are many wide applications in sentence reduction. For example, due to time and space constraints, the generation of TV captions often requires only the most important parts of sentences to be shown on a screen (Linke-Ellis 1999; Robert-Ribes et al. 1999). A good sentence reduction module would have an impact on the task of automatic caption generation. A sentence reduction module can also be used to provide audio scanning services for the blind (Grefenstette 1998). In general, since all systems aimed at producing coherent abstracts implement manually written sets of sentence compression rules (McKeown et al. 1999; Mani, Gates, & Bloedorn 1999; Barzilay, McKeown, & Elhadad 1999), it is likely that a good sentence compression module would impact the overall quality of these systems as well. This becomes particularly important for text genres that use long sentences.

Sentence reduction is commonly expressed as a word deletion problem: given an input source sentence of words x = x1, x2, . . . , xn, the aim is to produce a target compression by removing any subset of these words (Knight & Marcu, 2002). The compression problem has been extensively studied across different modeling paradigms, both supervised and unsupervised. Typical of supervised learning model was proposed by Minh Le Nguyen, he used HMM model and lexical rule for generating sentence reduction [7] and syntax – based language model (Turner & Charniak, 2005). Apply unsupervised learning to speech summarization was proposed by Chiori Hori and Sadaoki Furui[8], in his approach, he extracted an input sentence into word sets and used dynamic programming for generating sentence reduction[9].

Our sentence reduction goal is used for automatic Vietnamese text summarization system. Because of proposed method of Vietnamese sentence reduction is supervised learning method [8]. Supervised learning need a large corpus for training and very complexity in processing while research on Vietnamese text is beginning so very difficult for building a Vietnamese text summarization corpus. Therefore, suitable conditions, we use semi – supervised learning method for sentence reduction. Additional, Vietnamese is a single syllable, also difficult to separate words, so that, we only segment sentence into two word sets, reduce complexity in word segmentation when Vietnamese word segmentation tool is less effectively [13].

The rest of paper is organized as follows: In section 2, we will introduce some related work and model of Vietnamese sentence reduction. In section 3 is presentation of our method for Vietnamese sentence reduction. Experimentals and results will show in section 4. And finally, section 5 is conclusion and future works.

II. Method of Vietnamese Sentence Reduction

A. Related Works

Knight & Marcu (2002) proposed two methods, one is the noisy channel model where the probabilities for sentence reduction (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 [7].

In the work of (Le Nguyen & Ho, 2004) two sentence reduction 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) pair, selected from a website and manually tuned to obtain the training data. Due to complexity difficulties found for the application of this big lexical ruleset, 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

Manuscript received 8 January, 2011; revised 4 April, 2011.

Ha Nguyen Thi Thu, Head of Computer Science Department, Electric Power University, Hanoi, Vietnam. (Email: hantt@epu.edu.vn).

Quynh Nguyen Huu, Dean of Information Technology Faculty, Electric Power University, Hanoi, Vietnam. (Email: quynhnh@epu.edu.vn).

applied to a specific case [8].

An unsupervised approach was included in the work of (Turner & Charniak, 2005), where training data are automatically extracted from the Penn Treebank corpus, to fit a noisy channel model, similar to the one used by (Knight & Marcu, 2002). In the work of (Clarke & Lapata, 2006) devise a different and quite curious approach, where the sentence compression task is defined as an optimization 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 paralel corpus, it is completely knowledge driven, like a set of crafted rules and heuristics incorporated into a system to solve a certain problem[7],[8].

All these works applied for English. In Vietnamese, there are two methods for sentence reduction. And all these methods proposed Minh Le Nguyen. One of its applied HMM to Vietnamese sentence reduction and other used syntax control for reducing. And up to now hasn't significant research on Vietnamese sentence reduction [7],[8].

B. Model of sentence reduction

We use semi – supervised learning approach for generating sentence reduction and our method is carried out by 4 steps:

- ***Step 1***: Apply word segmentation VLSP tool to separate words from original sentence into 2 word sets: topic word set (is noun) and other word set (not noun).
- ***Step 2***: Important word set will be extracted according to a ratio, this ratio indicate original sentence will be reduced by information and it is calculated by number of topic word in sentence divide number of topic word in original word.

$$r = \frac{\textit{number of topic word in reduction sentence}}{\textit{number of topic word in original sentence}} \quad (1)$$

- ***Step 3***: Generating the most likelihood substring based on set of important words was extracted from step 2 and other words set by Viterbi algorithm.
- ***Step 4***: Correct grammatically in step 3 and generating reduction sentence.

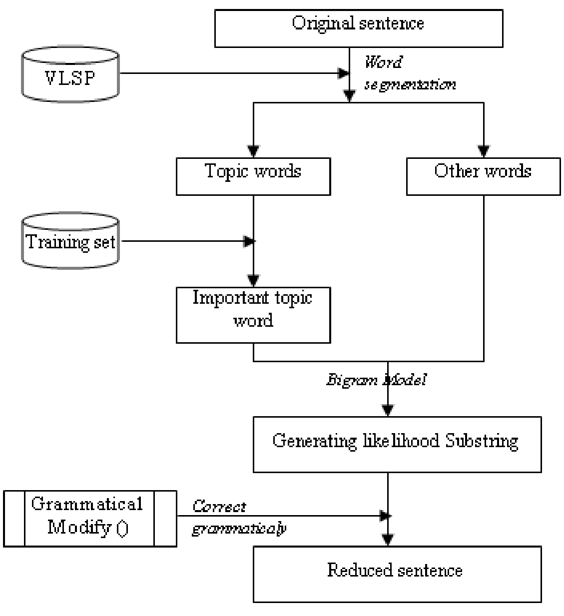

Figure 1. Model of Vietnamese Sentence Reduction.

III. Sentence Reduction by Concating the Most Likelihood Substrings

A. . Word segmentation

Vietnamese is a single syllable language so we can't separate word based on spaces. For example :an input sentence "***Iran luôn nói chương trình hạt nhân của mình là nhằm phục vụ mục đích hòa bình.". Translated to English "Iran has always said their nuclear programs is aimed at peaceful purposes***" We separate as bellow:

Iran /luôn nói /chương trình/ hạt nhân/ của /mình/ là/ nhằm/ phục vụ/ mục đích/ hòa bình.

So we have a topic word set by use a Vietnamese word segmentation tool.

Topic word set={ Iran (Iran), chương trình (programs), hạt nhân (nuclear), mục đích (purpose), hòa bình (peace) }.

B. Information Significant score

In our method, topic word is the noun expressing information. So, Information significant score only apply for words whichs is noun and calculate by the formula (2) as follow:

$$I(w_i) = \frac{N_S(w_i)}{\sum_{w_i \in S} w_i} + \frac{N_D(w_i)}{N_D} \quad (2)$$

where:

- $N_S(w_i)$ is the number of occurrence of topic word w_i in original sentence.

- $\sum_{w_i \in S} w_i$ is the total number of topic word w_i in original sentence.

- $N_D(w_i)$ is the number of documents in training set that occur topic word w_i .

- N_D is the total number of documents in training set *D*.

We don't use tf x idf method (term frequency–inverse document frequency) for calculating information weight of word. Because, in our method, document sets for training was classified by topic, so use tf x idf is not appropriate.

C. Extracting important topic words

Ratio r of a sentence indicate how much amount of information will be extracted, and it is calculated by number of topic words in reduced sentence divide by number of topic words in original sentence. This ratio is only used for extracting a set of important topic word from topic word set that has high score *I(wi)*.

Suppose ratio r=50% with the above sentence, the number of topic word will be extracted from original sentence is a set of 3 words.

D. Methodolody of sentence reduction

This section describes Sentence Reduction Algorithm based find substrings that is the most likelihood. In the first step, the original sentence is segmented into two sets that are T and O. A set of words, T, that consists of nouns and a set of words, O, which consists of the rest of sentence. In the second step, the significance score of each word in set T is computed. In the next step, the important word set (called R) will be extracted from T according the ratio r. The set of all words in set R are then used to generate the reduced sentence by dynamic programming to find the most likelihood substrings. At last, the method uses the syntax tree model to grammatically correct the sentence that is generated from the set R.

Algorithm 3.4.1 Extracting important words from original sentence

Input: *T , O, r*
Output: *T'*
Begin
1. For *i:=1* to *count(T)* do

$$I(w_i) \leftarrow \frac{N_S(w_i)}{\sum_{w_i \in S} w_i} + \frac{N_D(w_i)}{N_D};$$

2. *Sort I(wi)* in *descending* order;
3. While (# topic word in *T'< r* * # topic word in *T*)
T'← w_i;
End;

In the above algorithm, *T* is topic word set, *O* is other word set, *r* is reduction ratio, *T'* is important word set, w_i is topic word.

Reduced sentence contains set of important topic word that was extracted from previous step. Then, we used n-grams model and Viterbi algorithm for better determining sentence reduction. Viterbi algorithm is practiced with each pair nouns (w_i,w_j) in original sentence that was removed unimportant words from step 3, between us can be verb, adverb, adjective etc other than nouns. N-grams and Viterbi is used for calculating which subsequence is most likelihood .

$$S = \sum \textit{the most likehood subsequen ce of each noun pair}\,(w_i, w_j) \qquad (3)$$

For example "*Có thể coi HVA là một "tổ chức" hacker hoạt động có tôn chỉ mục đích rõ ràng*." In English "***HVA that can be considered as a hacker organization does everything according to a clear purpose.***". We have some noun pairs (HVA (HVA) , tổ chức (organization)), (tổ chức (organization) , hacker(hacker)), (hacker (hacker), tôn chỉ), (tôn chỉ, mục đích (purpose)).

Algorithm 3.4.2 Generating sentence reduction

Input: *T',O*
Output: *S*
1. Determining the next word of w_i
For each topic word pair (w_i,w_j)
If *j>i+1* then
For *k:=i+1* to *j* do
{ *S(i,k)*←max(*N-grams(wi, vk)*);
backpoint←agrmax *(S(i,k))*; }
2. Determining the most likelihood substring (w_i,w_j)
For *m:=k* to *j-1* do
For *l:=k+1* to *j* do
{ *S(k,l)*←max (backpoint + *N-grams*(k,l);
backpoint ← agrmax (backpoint +*N-grams(k,l)*; }
3.Concating the most likelihood substring for generating reduction sentence

Example 3.1. An input sentence "Trong tháng 4, Thủ tướng Việt Nam Nguyễn Tấn Dũng sẽ tham dự Hội nghị Thượng đỉnh về an toàn hạt nhân do Tổng thống Mỹ Barack Obama chủ trì tại Washington.**"**. Translated to English ***"In April , Vietnam's Prime Minister Nguyen Tan Dung will attend the Summit on Nuclear nuclear that is held by the U.S. President Barack Obama in Washington"***.

Table I shows the topic words t_i and its significance score.

TABLE I. SIGNIFICANT SCORE OF TOPIC WORD.

Topic word t_i	**Significant score $I(t_i)$**
Tháng	0.21
thủ tướng	0.6
việt nam	0.35
nguyễn tấn dũng	0.2
hội nghị	0.433
thượng đỉnh	0.4
hạt nhân	0.68
tổng thống	0.27
mỹ	0.39
barack obama	0.5
Washington	0.1

Set of important word when ratio r =60%
T'={ Tháng, thủ tướng, Việt nam, hạt nhân, tổng thống, Mỹ, Washington}

Figure 3 indicates apply n-grams and viterbi for determining the most likelihood substring between a pair of noun. Text in the circle is nouns, the blue path is the most likelihood sequence for sentence reduction corresponding with relativity ratio r=60%.

Table II shows the reduction results with 40%, 60% and 80% reduction ratios.

TABLE II REDUCED SENTENCES WITH VARIOUS RATIOS.

Ratio	**Sentence Reduction**
80%	Trong tháng 4, thủ tướng Việt nam tham dự hội nghị về an toàn hạt nhân do tổng thống Mỹ Barack Obama chủ trì tại Washington *In April, Vietnam's Prime Minister attend the Summit on Nuclear security that is held by the U.S. President Barack Obama in Washington.*
60%	Trong tháng 4, thủ tướng Việt nam tham dự an toàn hạt nhân do tổng thống Mỹ chủ trì tại Washington. *In April, Vietnam's Prime Minister attend the Nuclear security that is held by the U.S. President in Washington.*

IV. EXPERIMENTAL RESULTS

Our experiment used the corpus of 100 Vietnamese text. We collected from Vietnamese Vnexpress online newspaper (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 were then used to generate reduction sentence for our reduction algorithm SRBLS.

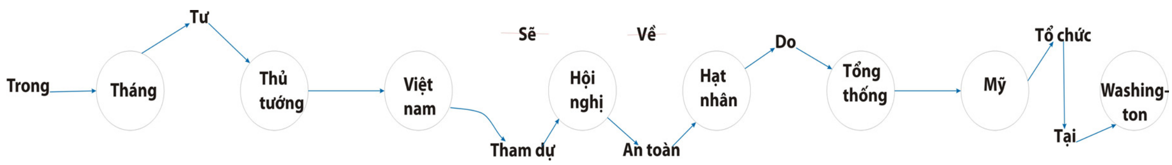

Figure 2. The most Likelihood substring with relativity ratio =60% .

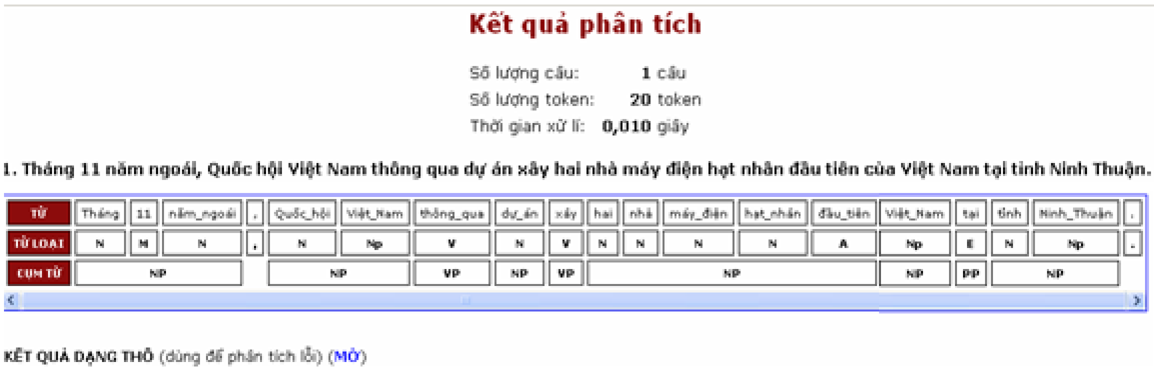

Figure 3. VLSP Word segmentation tool.

It's difficult to compare our method with previous ones, because there were not widely accepted benchmarks for Vietnamese text reduction sentence. Therefore, we compare our proposed method with manual sentence reduction generated by humans, called Human, and sentence reduction method using syntax control, called Syn.con, proposed by M.L. Nguyen and S. Horiguchi [9]. Figure 4 shows two examples of our reduction methods in testing on the Vietnamese language. Each reduction example is attached to an English translation. The reduction results of ours in the all examples are close to human reduction.

Example 1	
Original	Một quan chức cao cấp của Mỹ cho biết Hoa Kỳ và Nga "đang đạt được tiến bộ tốt" trong thỏa thuận giảm vũ khí nguyên tử. ***A senior U.S. official said the United States and Russia are "good progress" in the agreement to reduce nuclear weapons.***
Our method (with ratio 60%)	Mỹ cho biết Hoa Kỳ và Nga đang thỏa thuận giảm vũ khí ***The US said The United States and Russia agree to reduce weapons.***
Human	Mỹ cho biết Hoa Kỳ và Nga đang thỏa thuận giảm vũ khí nguyên tử. ***The US said The United States and Russia agree to reduce nuclear weapons.***
Example 2	
Original	Iran luôn nói chương trình hạt nhân của mình là nhằm phục vụ mục đích hòa bình. ***Iran has always said their nuclear programs is aimed at peaceful purposes***
Our method (with ratio r= 70%)	Iran nói chương trình hạt nhân của mình là phục vụ hòa bình. ***Iran said their nuclear program is aimed at peaceful.***
Human	Iran nói chương trình hạt nhân là phục vụ hòa bình. ***Iran said nuclear program is aimed at peaceful.***
Example 3	

Original	Có thể coi HVA là một "tổ chức" hacker hoạt động có tôn chỉ mục đích rõ ràng. ***HVA that can be considered as a hacker organization does everything according to a clear purpose.***
Our method (with ratio r= 70%)	có thể coi HVA là một tổ chức hacker có mục đích rõ ràng. ***HVA that can be considered as a hacker organization does according to a clear purpose.***
Human	HVA là một tổ chức hacker hoạt động có mục đích. ***HVA is a hacker organization does according to purpose.***

Figure 4. Some examples for sentence reduction.

In this experiment, we use the evaluation way as Knight and Marcu [7]. Table III shows the sentence reduction results that are carried out by our method, Human and Syn.con for Vietnamese text.

TABLE III EXPERIMENTAL RESULTS.

Method	Compression	Grammatically	Information
Baseline	x	X	X
Our method	65.25	7.2 ± 1.3	6.1 ± 1.2
Human	61.2209	8.333333	6.34524
Syn.con	67	6.5 ± 1.7	6 ± 1.1

Table III 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. Table III also shows the word significance weight in the fourth column, which indicates the number of important words of original sentence that occur in reduced sentence.

V. Conclusion

Reseach on Vietnamese text is beginning, so we applied semi-supervised learning approach to Vietnamese sentence reduction while Vietnamese text hasn't got a fully corpus for text summarization. In our method, we reduced complex word segmentation by segment input sentence into two word sets. Then, we used Viterbi algorithm to generate sentence reduction. Sentence that was generated by our method is correct in grammar, has high in linguistic, good readable and understandable. Our experimental results on a corpus of 5000 sentences of Vietnamese text shows that the proposed sentence reduction method achieved.acceptable results compared to human reduction.

In the future work, we will make abstracts from Vietnamese text consisting of multiple sentences.

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 for their great help in building the experimental summarizing application.

REFERENCES

[1] Ha Nguyen Thi Thu, Quynh Nguyen Huu, Cuong Do Duc, "A novel important word based sentence reduction method for Vietnamese text", Proc. of IEEE on Intellectual Technology in Industrial Practice, pp 401-405, China – Changsha September 2010.

[2] Ha Nguyen Thi Thu, Nguyen Thien Luan "A Novel Application of Fuzzy Set Theory and Topic Model in Sentence Extraction for Vietnamese Text", International Journal of Computer Science and Network Security, Vol. 10 No. 8 pp. 41-46, 2010.

[3] Ha Nguyen Thi Thu, Quynh Nguyen Huu "A New method for Vietnamese text Sentence Extraction based on important information of topic word and linguistic score", Proc. of IEEE on Multimedia and Computational Intelligence, China – Wuhan September 2010.

[4] Ha Nguyen Thi Thu, Quynh Nguyen Huu "Method of Sentence Reduction in Vietnamese Text Based on Determining Likelihood Substring". IEEE- International Conference on intelligen Network and Computing, (Accepted) , November, 2010.

[5] Ha. Nguyen Thi Thu, Luan Nguyen Thien, Implement some features for better determining weight of sentence in vietnamese text, International Journal of Artificial Intelligence and Computational Research (IJAICR), (Accepted), December, 2010.

[6] Ha Nguyen Thi Thu, Quynh Nguyen Huu "A Semi – Supervised Learning Approach for Generating Vietnamese Sentence Redution", Proc. of IEEE on 2010 International Conference on Computer and SoftwareModeling (ICCSM 2010) Manila, Philippines. December 4-5, 2010. (Acepted)

[7] Trevor Cohn, Mirella Lapata "Sentence Compression as Tree Transduction" Journal of Artificial Intelligence Research 34 (2009) 637-674

[8] 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, June 2004, Pages 146-158.

[9] Chiori Hori and Sadaoki Furui, "Speech Summarization: An Approach through Word Extraction and a Method for Evaluation" IEICE Trans. INF. SYST,.Vol . E87-D, No.1 January 2004. pp15- 25

[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] KNIGHT, K. AND MARCU, D. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artif. Intell. 139, 1 , 91-107, 2002.

[12] Nguyen, M.L.; Shimazu, A.; Horiguchi, S.; Ho, B.T.; Fukushi, M. (2004). Probabilistic Sentence Reduction Using Support Vector Machines. In the Proceedings of the 20th international conference on Computational Linguistics.

[13] Dipanjan Das and Andre F.T. Martins (2007). A Survey on Automatic Text Summarization

[14] Chin-Yew Lin and Eduard Hovy "The Potential and Limitations of Automatic Sentence Extraction for Summarization". In Proceedings of the HLT-NAACL 2003 Workshop on Automatic Summarization, May 30 to June 1, 2003, Edmonton,Canada.

[15] 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.

[16] Thanh Le Ha, Quyet Thang Huynh, Chi Mai Luong, "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, 2005.

Ha Nguyen Thi Thu received the B.S. degree in computer science and MSc degree in computer science from Guilin University of Electronic Technology in 2003 and 2006, respectively. She is currently a PhD student at Le Qui Don technical University, Viet Nam. MSc Ha has been working for Information Technology Faculty at Electric Power University, Hanoi, Vietnam, from 2007. She is the author or co-author of many international journal articles and international conference papers. Her research topics include Natural Language Processing (NLP), Data Mining and Machine Learning.

Quynh Nguyen Huu received the B. S. degree in informatics, M.S. and Ph. D. degrees in computer science, all from Hanoi University of Viet Nam, in 1998, 2004 and 2010, respectively. Dr. Quynh has been working for Information Technology Faculty at Electric Power University from 1999. His current research interests include content based image retrieval, intelligent image processing, GIS, image features extracting, and multimedia systems, Image Database Management Systems, Natural Language processing. Dr. Quynh has published over 15 international journal articles and international conference papers.