JAIST Repository
https://dspace.jaist.ac.jp/

Title	A Primary Study on Summarization of Documents in Vietnamese
Author(s)	Thanh, Le Ha; Quyet, Thang Huynh; Chi, Mai Luong
Citation	
Issue Date	2005-11
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/3908
Rights	©2005 JAIST Press
Description	The original publication is available at JAIST Press http://www.jaist.ac.jp/library/jaist-press/index.html, IFSR 2005 : Proceedings of the First World Congress of the International Federation for Systems Research : The New Roles of Systems Sciences For a Knowledge-based Society : Nov. 14-17, 2118, Kobe, Japan, Symposium 5, Session 2 : Data/Text Mining from Large Databases Text Mining

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

A Primary Study on Summarization of Documents in Vietnamese

Thanh Le Ha[1] , Quyet Thang Huynh[2], Chi Mai Luong[1]

1. Institute of Information Technology, Vietnamese Academy of Science and Technology
18th Hoang Quoc Viet Str., Cau Giay Dist., Hanoi, Vietnam.
leht82@gmail.com, lcmai@ioit.ncst.ac.vn,

2. Faculty of Information Technology, Hanoi University of Technology
1st Ta Quang Buu Str., Hai Ba Trung Dist., Hanoi, Vietnam.
thanghq@it-hut.edu.vn

ABSTRACTION

There are some statistical-based sentence extraction methods applied to English documents to get the automatically summaries. In this paper, we present a Vietnamese text summarization case-study based on evaluation and extraction of highly informative sentences to abstract documents, assisting users in reducing the time required to study and grasp information in Vietnamese, particularly appropriating to news from Vietnamese sites. Our case-study combines various statistical sentences extraction methods which do not require more linguistic resources whereas provide fast approaches. From a set of sentences, we choose the most important ones depending on an input compression rates and generate a summarized document. Particularly, we use mostly Vietnamese linguistic characteristics to preprocess the source and improve the result. After using some content evaluating methods, comparing with other current approaches, our investigation shows interesting and satisfactory results. We get approximately 0.73 for the precision of traditional evaluating method, approximately 0.67 for the average of content similarities.

Keywords: Text Summarization, Sentence Extraction, Linear Combination, Statistical Methods.

1. INTRODUCTION

Language processing plays a fundamental role in information and knowledge science. With the popularity of the Internet use, various problems and needs such as Information Extraction, Text Summarization on the Web have emerged requiring new solutions in language technology. Vietnamese language technology is still in its infancy. Our researchers are carrying out various investigation, researches to deal with these problems. Among them Vietnamese Text Summarization is one of the research topics that we are focusing on in this work.

Text summarization is the process of distilling the most important information from a source(s) to produce an abridged version for a particular user and task [1].

There are many approaches and methods to attain this purpose, see Luhn[2], Edmunson[3], Jing [4], Marcu[5], Barzilay & Elhadad [6]… These approaches covered from statistical, robust to linguistic, sophisticated methods.

Luhn, Edmunson, Lin & Hovy [7] apply some sentence extraction methods, following the statistical-based approach, while Jing, Barzilay & Elhadad, Morris & Hirst [8], Mann & Thompson [9] using various NLP methods, such as Lexical Chains, Rule-based Reduction, Rhetorical Structure Theory… to get the most important, reduced sentences and output the result.

In the report of Luhn called the first text summarizer, he implemented a single measurement - the Term Frequency Method - to evaluate the importance of each sentence in the document [2]. More advanced, Edmunson using multiple measurements to get the most important sentences. There are four methods in the summarizer of Edmunson : Cue Method, Key Method, Title Method and Location Method, in which the Key Method uses the same idea as the Term Frequency Method of Luhn but using another implementation. And after assigning weights, he calculated the total weight then removed sentences which have the least total weight [3]. This approach has been used with various improved methods by Kupiec [10] or Nomoto & Matsumoto [11]..., especially for researchers who haven't completed linguistic corpora.

Our case-study follows the statistical-based sentence extraction approach, using some of above described methods and other methods that we proposed as well. These methods are all considered appropriating to linguistic characteristics of Vietnamese and do not require more Vietnamese linguistic resources. It make ours easy to implement whereas still obtain acceptable

and satisfactory result. The rest of this paper is organized as follows: in the next section a brief description of model that we use for our study is given. In Section 3 we describe structualization of documents. In Section 4, our sentence extraction methods are described. Finally evaluation and conclusions are given in Section 5 and Section 6 respectively.

2. THE MODEL

Our case-study is distinctly divided into three phases. First, in the initial phase, we structuralize the free-text document into smaller structured units i.e. paragraphs, sentences and words so we can process more easily in the following phases. Then, in the main phase, we choose important sentences which contain major information of the document by assigning weights to each of them depending on their importance. Some methods for evaluating the importance of a certain sentence are title-based, TFxIPF, position-based, proper noun-based and word co-occurrences... After choosing these important sentences, in the final phase, we re-arrange them by considering their original order and generate the result. Particularly, in all phases, we often apply linguistic characteristics of Vietnamese to our processing. And this really makes our result better.

The model of our case-study is shown in Figure 1 below.

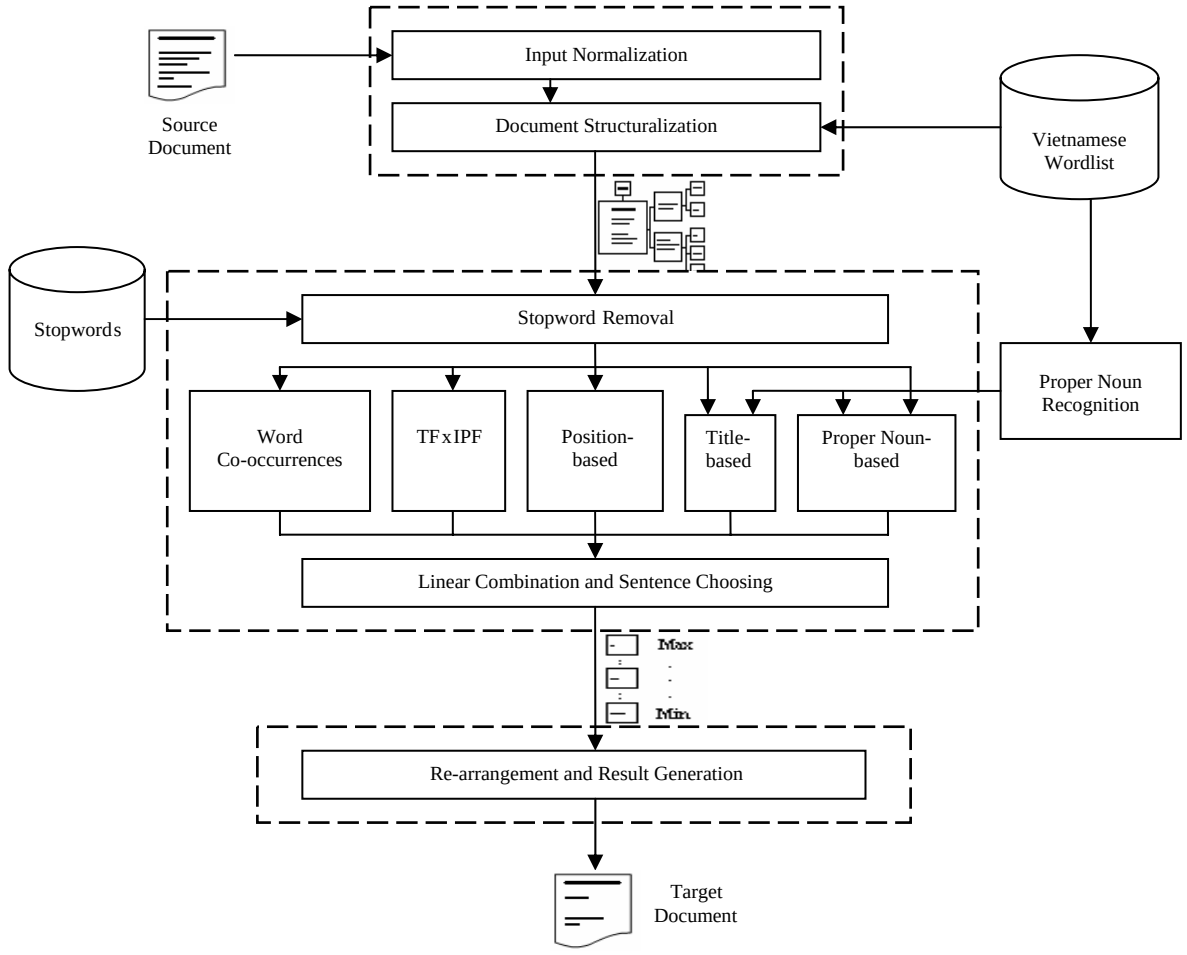

Figure 1. Model of Vietnamese Text Auto Summarizer

In this model, the input is the Vietnamese free-text document, and it will be normalized and structuralized in the initial phases. The normalization and structuralization help us to apply evaluation methods which only work with linguistic units such as words, sentences or paragraphs.

The main phase, Sentence Evaluation and Extraction, will be detailed below. These phases will help decrease a number of sentences, retain the most essential of them and remove redundant others.

Finally, in the last phase, depending on the compression rate, a certain number of sentences will be chosen and re-arranged, based on sentence clues, to generate the target - an automatically summarized document.

3. STRUCTURALIZATION

For the purpose of documents processing, the free-text input will be preprocessed and modeled into the *Document* object. This object includes an array of *Paragraph* objects which represent document paragraphs. The *Paragraph* object, in turn, has an array of *Sentence* objects which correlate one-to-one with sentences. A *Sentence* object includes the vector of terms which it contains. Besides that, each *Document*, *Paragraph* and *Sentence* object has some additional information such as the position of the paragraph in the document, the uniqueness of the sentence in the paragraphs, the number of words in the sentences and the correlative string content of the objects… In general, with this structuralization, we can work with these objects as the actual linguistic units.

The most important and difficult work in this phase is the process of Word Segmentation because Vietnamese is a monosyllable language like Japanese, Chinese or Korean. Unlike English, there are no word boundaries in Vietnamese. Although it's no problem when Vietnamese people speak or write, the computer can't understand what a word is. We can't use the blank spaces to determine the word boundaries like in English. To solve this problem, Japanese, Chinese and Korean researchers proposed some methods with high accuracy. In Vietnam, There are also some complex Word Segmentation methods such as those of Huyen el al [12], Ha Le An [13]… Especially, there is a method of Dinh Dien which achieved approximately 97% of accuracy [14]. But these methods require more lexical resources as well as corpora. Moreover, they are time consuming and not suitable for our approach. Here we use the method of Max Length Word Matching to achieve the word segmentation. Based on a Vietnamese wordlist, we find the longest string which matches with a Vietnamese word and produce the array of terms. After that, we use the vector space model to vectorize the array into the vector of indexes of term. Although it's a very simple way to segment words, we choose it because it's easy for implementing, fast for processing and have reasonable accuracy for the following phases.

Here is the *Max length matching algorithm* applied to Vietnamese Word Segmentation :

Input : String *s* which we want to segment.
dic which contains list of Vietnamese words.
Output : The corresponded array of Vietnamese words.

Procedure :

1. { Declaration }
   t*ermList* is an array of string;
   lmax is an positive integer number;
   tempQueue is a queue of string;

2. {load all dic's Vietnamese words }
   termList := getWords(dic);
   {get the length of the longest word in dic }
   lmax = getLengthOfLongestWord(dic);

3. tempString := getSubstring(s,1,lmax);

4. tempString := trimToEvenGram(tempString);
{ Grams here we mean strings delimitated by spaces }

5. If tempString in termList then
   Add(tempString into tempQueue);
   s = s – tempString;
   Else
   s = trimTheMostLeftGram(s);

6. If s Is Not Empty Then Goto 3;

7. Return toArray(tempQueue);

Figure 2. The Max Length Matching algorithm

Some works remaining in this phase are Sentence Segmentation and Notation Disambiguation. Sentence Segmentation can be carried out by using sentence delimitators. Some sentence delimitators such as point, semicolon and three-dots. We have to distinguish sentence points and the decimal point or the point lie in internet addresses as well as email addresses. This work can be attained by the Notation Disambiguation module.

At the same time, other properties of *Document* object would be set. We note that Title is also a sentence and it's modeled by a special *Sentence* object. In addition, some position and association clues of each sentence are also gathered for the purpose of generating final result.

4. SENTENCE EXTRACTION METHODS

In the main phase, we implement some suitable methods to extract essential sentences. For the purpose of quantifying the importance of each sentence, we assign weights for them depending on certain methods.

4.1. Title-based method

The idea of this method is the title of document should be chosen for the summary. Furthermore, some appropriate words (terms) belong to the title can be used for evaluating other sentences in the document. So, we first determine the title, choose it to be the title of the summary and extract terms from it. These terms are called Title Word. Then we count the number of Title Word in each sentence and assign Title Weight for them. The more Title Words they contain, the higher Title Weight they have.

4.2. Position-based method

In some types of document such as scientific documents or news, the first and the last sentences in a paragraph are more important and contain more the percentages of document meaning than the others. So we can assign greater position weight for the first and the last sentences of all paragraphs. This idea is also true with the position of each paragraph in the document.

4.3. Proper noun-based method

Proper noun-based method is similar to the Title-based. Instead of extracting Title Words which are presented in a Vietnamese wordlist, we extract the proper nouns in the title, use them for assigning the Proper Noun Weights to the sentences. However, the way to determine proper nouns is clearly different from the process of Word Segmentation. This is a complex problem and nowadays we haven't found the perfect solution for this yet. Here we use some heuristics to determine proper nouns based on Vietnamese proper noun's characteristics. Just simple but we get the satisfactory results.

4.4. Word Co-occurrences method

The idea of this method is what paragraph that has more correlation with others will be important. The correlation is evaluated by the number of common terms between paragraphs. Alternatively, we can use some correlating formulae such as cosine or dice, too. Based on them, we assign the same Correlation Weight for all sentences which belong to the same paragraph and different Correlation Weight for the others.

Assume that a document have n paragraph P_1, P_2, ..., P_n. We determine the correlation of P_i with P_j by using the Cosine :

$$PC_{ij} = \cos(P_i, P_j) = \frac{\sum p_{ik} . p_{jk}}{\sqrt{\sum (p_{ik})^2} . \sqrt{\sum (p_{jk})^2}} \quad \text{i,j} = \overline{1,n} \quad (1)$$

PC_{ij} will be stored as the (i,j) element of an $n \times n$ 2-dimension array.

The Correlation Weight of paragraph P_i will be calculated as following :

$$CP(P_i) = \sum_{j=1}^{n} PC_{ij}, \; j = \overline{1,n} \quad (2)$$

Based on them, we assign the same Correlation Weight for all sentences which belong to the same paragraph and different Correlation Weight for the others.

4.5. TFxIPF (Term Frequency times Inverse Paragraph Frequency)

This method originates from the well-known TFxIDF Estimate. TFxIDF is used to determine specific terms in a certain paragraph. One terms is called a specific term of a paragraph if it occurs more in the paragraph and occurs less in other paragraphs of the document.

Here, one term isn't considered in a document but in a paragraph and its TFIPF Weight is calculated as below :

$$w_i = tf \times ipf \; = tf \times \log \frac{N}{n_i} \quad (3)$$

Where *tf* is the times which term *i* occurs in the paragraph, *N* is the total number of paragraphs in the document and n_i is the number of paragraphs that contain term *i.*

Then all terms are calculated the TFIDF Weight and sorted in the descending order of this estimate. A preordained percentage of terms which have the highest TFIDF Weight will be used to evaluate all sentences of document in the way that similar to Title or Proper noun-based methods.

4.6. Linear Combination.

To combination all above sentence extracting methods, we use a linear formula to calculate the final weight of each sentence :

$$\mathbf{W} = a.\mathbf{W_{TB}} + b.\mathbf{W_{PS}} + c.\mathbf{W_{PN}} + d.\mathbf{W_{CO}} + e.\mathbf{W_{TFIDF}} \quad (4)$$

Where :

- W_{TB}, W_{PS}, W_{PN}, W_{CO}, W_{TFIDF} in this order are Title Weight, Position Weight, Proper Noun Weight, Correlation Weight and TFIPF Weight.
- *a, b, c, d, e* are the linear coefficients.

As we know, all above methods can't be used for all types of document and each method has different effect. For example, when we use this approach to summarize a story, the position-based method will be unsuitable while the most efficient method here is the TFIPF. In this case, the coefficient of position-based is set to zero and the coefficient of TFIPF is set to the highest. Other example, if we can't determine the title, the coefficient of Title method is certainly set to zero.

So, these linear coefficients express "the contribution" of each method. They are manually or machine learning-based refined by monitoring the result. Moreover, setting the coefficients and watching the corresponding result will help us to determine which methods should be used which others. That will refine the later results.

5. EVALUATION

We evaluate this approach by apply it to summarize Vietnamese news and short scientific documents. After using some content evaluating methods [15,16], comparing with other current approaches, our research shows interesting and satisfactory results.

In the document set including news of several subjects we collected from the Vietnamese Vnexpress online newspaper (http://vnexpress.net/Vietnam/Home), we randomly choose 55 documents being input for our application and for Vietnamese linguisticians to make comparison and evaluation. The testing was carried out at five compression rates : 10%, 20%, 30%, 40%, 50%.

In this case-study, we use two summarization evaluation methods: the traditional method with the precision and the content similarities-based method.

5.1. Precision

Precision is measured by bringing into comparison each pair of summarized results of the same document between standard method and two methods which are examined now – basic method and our method.

$$\text{Precision} : P = \frac{A}{A+B} \quad (5)$$

Where A is the number of sentences chosen by both 2 methods, B is the number of sentences chosen by the standard method (summarized by linguistic experts) and not chosen by the treated method.

From that, the precision of the whole method is averaged out for every document group which has the same compression rate. Following is the result:

	Compression rate				
Method	**10%**	**20%**	**30%**	**40%**	**50%**
Baseline	**0.783**	**0.302**	**0.219**	**-**	**-**
Ours	**0.863**	**0.543**	**0.754**	**0.698**	**0.601**

Figure 3. The traditional evaluating method

5.2. Content Similarity

The formula for defining the content similarity between set to evaluate and set needs to be evaluated is:

$$Sim(S_i, J_i) = cosine(S_i, J_i) = \frac{\sum_{j=1}^{n} Si_j \cdot Ji_j}{\sqrt{\sum_{j=1}^{n} (Si_j)^2} \cdot \sqrt{\sum_{j=1}^{n} (Ji_j)^2}} \quad (6)$$

Where m is the document number of set to evaluate J. S is the result document generated by our application.

The result is:

	Compression rate				
Document	**10%**	**20%**	**30%**	**40%**	**50%**
1	**1**	**0.12**	**0.16**	**0.12**	**0.11**
2	**1**	**0.12**	**0.10**	**0.08**	**0.07**
3	**0.30**	**0.13**	**0.10**	**0.08**	**0.07**
4	**1**	**0.11**	**0.12**	**0.08**	**0.09**
5	**0.13**	**0.10**	**0.13**	**0.06**	**0.09**
6	**1**	**0.12**	**0.14**	**0.07**	**0.07**
7	**1**	**0.12**	**0.14**	**0.07**	**0.06**

Figure 4. The content similarities evaluating method

CONCLUSION

This article presents the combining of various methods in features extraction, applied to automatic summarization of Vietnamese documents. The summarization application is used in the processes of searching and summarizing news sites, allowing users to reduce required reading time. Through experimental results, this application shows relatively high precision in summarizing documents. However, in order to make it easier to read and understand, we will improve methods to compress sentences, to display documents focus on the features of Vietnamese language, and develop an effective applied model, well combining various methods to improve the compatibility among various writting.

ACKNOWLEDGEMENT

We would like to thank the linguistic experts of Viet Nam Linguistic Institute, Dr. Ho Ngoc Duc, Ms. Nguyen Hoang Anh for their great help in building the experimental summarizing application.

We would also like to thank our colleagues in the Department of Pattern Recognition and Knowledge Engineering, Institute of Information Technology, Vietnamese Academy of Science and Technology, for their continuing supports during completion of this work.

REFERENCES

[1]. Mani & Maybury. 2001. *Automatic Summarization*, ACL 2001. Slices.

[2]. Luhn, H.P. 1958. *The automatic creation of literature abstracts*. IBM Journal of Research and Development 2, p.159 - 165.

[3]. Emunson, H.P. 1969. *New Methods in Automatic Extracting*, Journal of Association for Computing Machinery 16(2), p.264-285.

[4]. Jing, H. 2000. *Sentence Reduction for Automatic Text Summarization*, Proceedings of the 6th Conference on Applied Natural Language Processing.

[5]. Marcu, D. 1997. *From discourse structures to text summaries*. Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, p.82 - 88.

[6]. Barzilay, R., & Elhadad, M. 1997. *Using lexixal chains for text summarization*. Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), p.10-17.

[7]. Lin, C., & Hovy, E.H. 1997. *Identifying topics by position*. Proceedings of the Applied Natural Language Processing Conference (ANLP-97), p.183–290.

[8]. Morris, J., & Hirst, G. 1991. *Lexical cohesion computed by thesaural relations as an indicator of the structure of text*. Computational Linguistics 17, p.21–48.

[9]. Mann & Thompson. 1988. *Rhetorical Structure Theory: towards a functional theory of text organization*. Text, volume 8, p. 243-281.

[10]. Kupiec et al. 1995. *A trainable document summarizer*. Proceedings of the 18th annual ACM SIGIR Conference on Research and Development in Information Retrieval, p. 68–73.

[11]. Nomoto & Matsumoto. 2001. *A new approach to unsupervised text summarization*. Proceedings of the 24th annual ACM SIGIR Conference on Research and Development in Information Retrieval, p. 24 - 36.

[12]. Huyen N.T. M., Luong V.X., Phuong L.H. 2003. *A case study of the probabilistic tagger QTAG for Tagging Vietnamese Texts*. Proceedings of ICT.rda'03, Hanoi, Vietnam.

[13]. Le An Ha. 2003. *A method for word segmentation in Vietnamese*. Proceedings of Corpus Linguistics 2003, Lancaster, UK, March. (http://clg.wlv.ac.uk/papers/Ha-CL-03.pdf).

[14]. D.Dien, H. Kiem, and N.V. Toan. 2001. *VietnameseWord Segmentation*. Proceedings of NLPRS'01 (The 6th Natural Language Processing Pacific Rim Symposium), Tokyo, Japan, p. 749-756.

[15]. Radev et al. 2003. *Evaluation challenges in large-scale document summarization*. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan. Association for Computational Linguistics, p.375-382.

[16]. Udo Hahn , Inderjeet Mani. 2000. *The Challenges of Automatic Summarization*, Computer, v.33 n.11, p.29-36,.