A Novel Application of Fuzzy Set Theory and Topic Model in Sentence Extraction for Vietnamese Text

Ha Nguyen Thi Thu 1[†] and Nguyen Thien Luan2[††],

[†]Department of Computer Science, Electric Power University ,235 Hoang Quoc Viet, Ha noi, Viet Nam
††Faculty of Information Technology, Le Qui Don Technical University,100 Hoang Quoc Viet, Ha Noi, Viet Nam

Summary
Vietnamese language has common characteristics with some Asian languages such as Chinese, Japanese, Korean ... They do not define words based on spaces. In this article, we present a method that application of Fuzzy set theory and topic model to extract sentences in Vietnamese texts which have been categorized by topic. This method based on identification of important features as : length of sentence, weight of terms in sentences, position of sentences ..., then extracting important sentences according to the ratio, this ratio indicate which sentences in original text will be extracted. We also built a system based on this method and experiments have obtained good results, satisfying the given requirements.

Key words:
Vietnamese text, Sentence extraction, topic model, Fuzzy set theory

1. Introduction

" *A text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that"* (Radev et all 2002) [1]. Lin - 2000 proposed construction of an automatic text summary system including sentence- extracting process [3]. For a long time, most of automatic text summariztion systems based on sentence extraction in one or more texts and then combined them to make the summary. The earliest research on text summary used sentence extraction method based on specific features of words and phrases frequency (Luhn, 1958), position of the sentence in the text (Baxendale, 1958) and important phrases (Edmundson, 1969) [1]. Appearance frequency of a word mainly based on the tf * idf method.

Most of these methods are effectively applied for English. However, for Asian languages like Vietnamese, Chinese, Japanese or Korean with single syllable, it is very difficult to separate words. Unlike English, words of these languages can not be determined based only on a space [4]. Thus, determination of the weight of sentences based on the method of calculating the frequency of words in sentences is not appropriate . On the other hand, study on Vietnamese text summary is now something new, corpus for Vietnamese text summarization or word segmentation is still limited.

We studied Vietnamese text that categorized by topic and found that with each different theme, weight of words (hereafter called terms) would depend on the different topics. Therefore, we have built sets of terms for each topic and applied Fuzzy theory in determining degree of membership of each term in text which is easier than identifying important sentences for extracting.

Structure of this article is as follows: in section 2, we introduce structure of automatic Vietnamese text sentence extraction system. In section 3, we present methodology using topic model to build sets of terms and Fuzzy set theory to determine weight of sentence in text. Section 4 shown the result of our system and evaluation with some others method and section 5 is conclusion.

2. Modeling of sentence extraction

The method applying Fuzzy theory is quite effective in approximately solving the problems of classification, identification of texts, scripts, images, sounds, ... An Fuzzy set A on a classic set X is defined as follows:

$$\tilde{A} = \{(x, \mu_A(x)) \mid x \in \mathrm{X}\}$$

Membership function μA (x) quantifies the degree which elements x belong to basic set X. If the function gives result 0 for an element, then that element is not in the given set, the result 1 describes a full member of the set. The values in range of 0 to 1 show fuzzy elements.

We propse a model for automatic Vietnamese text sentence calculation as follows

Manuscript received August 5, 2010
Manuscript revised August 20, 2010

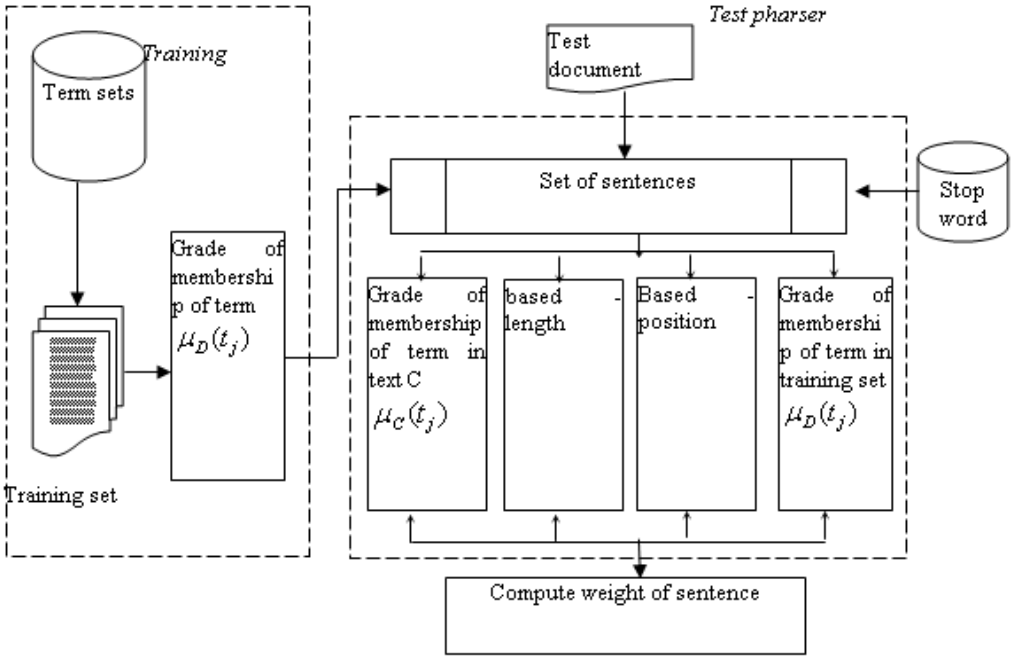

Fig 1: Calculated weight of sentence

3. Method of sentence extraction

3.1 Topic Model

Vietnamese has not a normal word segmentation, thus in this research, we have applied the Topic Model [4] to build a set of terms corresponding to the different topics. First we build sets of documents for training that has been classified by difference topic. With each topic we have a training set:

$$D = \{d_1, d_2, ..d_n\} \quad (1)$$

Where : D is the training set, collecting set of documents d_i are the same of topic.

For each topic, we build a corresponding set of terms.

$$T = \{t_1, t_2, ..t_m\} \quad (2)$$

3.2 Degree of membership of terms in training set D

Degree of membership of terms in the training set indicates importance of the term to the topic [6]. It is calculated by the ratio of total number of texts containing the term divide total texts in the training set.

$$\mu_D(t_j) = \frac{\sum_{i=1}^{n} d_i(t_j)}{n} \quad (3)$$

Where:

$\mu_D(t_j)$: is degree of membership of t_j on D

n: is number of documents in D

$$d_i(t_j) = \begin{cases} 1 \text{ if } t_j \in d_i \\ 0 \text{ if } t_j \notin d_i \end{cases}$$

3.3 Degree of membership of terms in test document.

If the degree of membership of terms in the training set D indicates importance of the term to the training set, the degree of membership of terms in test document shows the importance of the term to the test document and it is calculated by the number of appearance of term t_j over the total terms appearing in test document.

$$\mu_C(t_j) = \frac{N(t_j)}{\sum_{i=1}^{m} N(t_i)} \quad (4)$$

Where:

$\mu_C(t_j)$ indicates grade of membership of t_j in test document C.

$N(t_j)$ is number of t_j appear in C.

3.4 Calculating the importance of sentences

We propose a method for calculate weight of sentences

$$F_i = a_1 \frac{\sum \mu_D(t_j)}{\max\{\sum \mu_D(t_j)\}} + a_2 \frac{\sum_{t_j \in s_i} \mu_c(t_j)}{\max\{\sum \mu_C(t_j)\}} + a_3 \frac{length(s_i)}{length(S)} + a_4 position \quad (5)$$

Where : a_1, a_2, a_3, a_4 is the score are the linear coefficients indicate grade of membership of term t_j in training set D, grade of membership of term t_j in test document, length of s_i and position of s_i in document.

$$Position = \begin{cases} 1 \text{ if } s_i \text{ in the head of document or foot of document} \\ 0 \text{ if otherwise} \end{cases}$$

The algorithm is described as a follows:

ALGORITHM CALCULATING THE IMPORTANCE OF SENTENCES

Input C: Test Document

Output R: Set of important sentences

1. ***Initialization***

R←Ø; S←Ø;

2.Sentence segmentation

S←$s_1, s_2, .., s_k$

3. Calculating weight of each sentence in Test Document

For each s_k

For i=1 to length(s_k)

For each term in s_k calculate

$$L(s_i) \leftarrow L(s_i) + \mu_D(t_j) + \mu_C(t_j)$$

$$F(s_i) \leftarrow\!\!- L(s_i) + \frac{length(si)}{length(S)} + position$$

Fig. 2 Procedure for calculated weight of sentence

3.5 Method for extracting important sentences

After the sentences have been calculated, they will be sorted in descending order and extracted according to the ratio r which is length of set of extract sentences divided by length of set of sentences in the original text. The algorithm of sentence extraction is as follows

SENTENCE EXTRACTION ALGORITHM

Input :
C: Test Document
Ratio *r*
Output:
R : Set of Sentence
Begin
R=""; p=1;
// select sentence to extract
Do While
*length (R) + length(s[i_p]<= length(C) *r*
Begin
R =R + s[i_p]
t[i_p]= true;
p=p+1;
End
// arrange sentences in order
R=" "
For i=1 to n do if t[i]=true then R=R+s[i]
End

Fig. 3 Procedure for selected sentence

4. Result and evaluation

We evaluate this approach by apply it to summarize Vietnamese news and comparing with other current approaches, our research shows interesting and satisfactory results. Based on the proposed method above, we have built a system to calculate importance of sentences and automatic extraction of sentences for Vietnamese texts. We tested with four different topics: sport, technology, political, and economy. For each topic, we collected about 300 different texts for training from the Vietnamese website http://www.vnexpress.net. And here is the architecture of our automatic Vietnamese text sentence extraction.

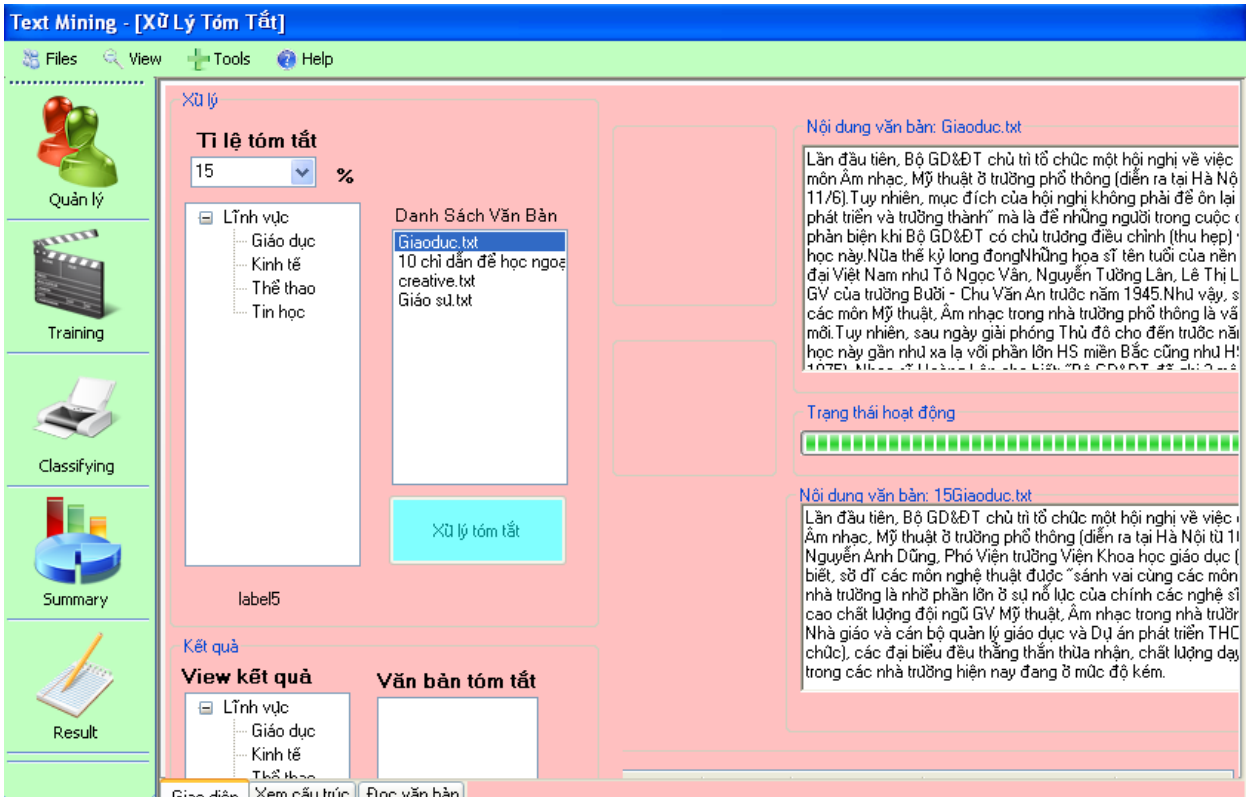

Fig. 4 Architecture for extraction

At the present, Vietnamese does not have any standard assessment method, therefore, we compare the result of extract according to the method proposed by Ha Thanh Le [Le Thanh Ha, Quyet Thang Huynh, Mai Luong , 2005] and our result. Comparison shows that our extract result is better and more stable than theirs. In addition, we also compared our result with an online summary system http://smmry.com, and result from extract by human. In evaluate our method with an online summary system. We downloaded english documents from http://news.bbc.co.uk and used for exprimentation. We translated to Vietnamese text and apply our system to extract sentences, then translated to English and compared it with result of smmry.com.

Real Madrid are set to complete the signing of Portugal centre-back Ricardo Carvalho from Chelsea. The Spanish giants have agreed a £6.7m fee to sign the 32-year-old, who will link up once again with former Chelsea boss Jose Mourinho. "Chelsea can confirm it has agreed terms with Real Madrid for the transfer of Carvalho," said a club statement. Carvalho, who was also with Mourinho at Porto, joined Chelsea in 2004 for £19.8m and won the league three times. The Chelsea statement added: "The transfer is subject to a medical and the player agreeing personal terms. Chelsea would like to thank Riccy for his six years of service, and we wish him well in his future career." As well as winning three Premier League titles during his time at Stamford Bridge, Carvalho also won the FA Cup three times, the League Cup twice and the Community Shield twice. He becomes Real's fifth signing of the summer, following Portugal's Angel di Maria, Spanish duo Pedro Leon and Sergio Canales and Germany World Cup star Sami Khedira.

A statement on the Spanish club's website began: "Best known for his perfect positioning, his foresight and his ability to advance the ball out of the first third, Ricardo Alberto Silveira Carvalho joins Real Madrid at the age of 32 and with 19 titles to speak of." Last month, Carvalho told a Spanish newspaper it would be a dream come true to join Mourinho at the Bernabeu. "If there was a possibility to sign with Real Madrid, I would go there right now swimming or running," he said. "It would be a dream to be able to play for Madrid, which I consider to be the best club in the world, and follow the orders of the best coach in the history of football. "With Mourinho I experienced two marvellous stages at Porto and Chelsea. To have the opportunity to win another Champions League with him at Real Madrid would be tremendous." Carvalho played more than 200 games for Chelsea, scoring 11 goals. He is the fifth player to leave the Stamford Bridge club this summer after Joe Cole, Michael Ballack, Juliano Belletti and Deco.

Fig.5 Original text for experiment

From original text, we extracted 5 sentences which are the most important in it.

And follow is the result of extracting by online sumary system summry.com

SENTENCES EXTRACTED BY SUMMY.COM

The Spanish giants have agreed a £6.7m fee to sign the 32-year-old, who will link up once again with former Chelsea boss Jose Mourinho.
"Chelsea can confirm it has agreed terms with Real Madrid for the transfer of Carvalho," said a club statement. Carvalho, who was with Mourinho at Porto, joined Chelsea in 2004 for £19.

"It would be a dream to be able to play for Madrid, which I consider to be the best club in the world, and follow the orders of the best coach in the history of football.

"With Mourinho I experienced two marvellous stages at Porto and Chelsea.
Have the opportunity to win another Champions League with him at Real Madrid would be remendous." Carvalho played more than 200 games for Chelsea, scoring 11 goals.

Fig. 6 Extracted by smmry.com

And then, we use our system for extract sentence from original text.

And here is the result of extracting by our system.

SENTENCES EXTRACTED BY OUR SYSTEM

“Chelsea có thể xác nhận nó đã đồng ý các điều khoản với Real Madrid về việc chuyển nhượng của Carvalho," một câu lạc bộ tuyên bố Carvalho cùng với Mourinho tại Porto, gia nhập Chelsea vào năm 2004 với giá £ 19.8m và chiến thắng ở giải Vô địch ba lần.
Chelsea tuyên bố thêm: "chuyển giao này phải được một y tế và người chơi đồng ý các điều khoản cá nhân. Chelsea muốn cảm ơn Riccy sau sáu năm phục vụ, và chúng tôi mong muốn sự nghiệp tương lai của anh ấy tốt."
Cũng như giành ba danh hiệu Premier League trong thời gian của mình tại Stamford Bridge, Carvalho cũng đã giành FA Cup ba lần, League Cup hai lần và Community Shield hai lần.
Có cơ hội để giành chiến thắng khác với Champions League tại Real Madrid "Carvalho đã chơi hơn 200 trò chơi cho Chelsea, ghi được 11 bàn thắng..
Anh là cầu thủ thứ năm rời khỏi câu lạc bộ sân Stamford Bridge mùa hè này sau Joe Cole, Michael Ballack, Juliano Belletti và Deco.

TRANSLATE TO ENGLISH

"Chelsea can confirm it has agreed terms with Real Madrid for the transfer of Carvalho," said a club statement Carvalho, who was also with Mourinho at Porto, joined Chelsea in 2004 for £19.8m and won the league three times.

The Chelsea statement added: "The transfer is subject to a medical and the player agreeing personal terms. Chelsea would like to thank Riccy for his six years of service, and we wish him well in his future career."

As well as winning three Premier League titles during his time at Stamford Bridge, Carvalho also won the FA Cup three times, the League Cup twice and the Community Shield twice.

Have the opportunity to win another Champions League with him at Real Madrid would be remendous." Carvalho played more than 200 games for Chelsea, scoring 11 goals.

He is the fifth player to leave the Stamford Bridge club this summer after Joe Cole, Michael Ballack, Juliano Belletti and Deco.

Fig. 7 Extracted by our system

Pression is the traditional assessment method is given by:

$$Pression = \frac{correct}{correct + wrong}$$

In which: *correct* the number of sentences extracted by both human and system. *wrong* is the number of sentences extracted by the system but not by human.

Table 1: Traditon of evaluation

	Compression rate			
Method	80%	60%	40%	20%
Smmy.com	0.761	0.783	0.697	0.61
Ours	0.875	0.802	0.67	0.64
Human	0.91	0.85	0.863	0.82
Ha Le Thanh	0.62	0.754	0.698	0.543

5. CONCLUSION

This article represents the application of Fuzzy theory and Topic Model to calculate the importance of sentences in the extract process - one of the processes in automatic Vietnamese text summary system. Based on the extracted sentences, we further reduced such sentences to make the summary more compact in terms of space, more concise in terms of content and meanings. As Vietnamese is much similar to the Chinese (over 80% of Vietnamese are borrowed from Chinese), Japanese, Korean, this method once well-applied to Vietnamese, will definitely be able to apply to Chinese, Japanese and Korean.

Acknowledgments

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 Huu Quynh, Dr Nguyen Van Vinh, Dr Nguyen Phuong Thai for their great help in building the experimental summarizing application.

References

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

[2] 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.

[3] 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.

[4] Mark Steyvers and Tom Griffiths. "Probabilistic Topic Models".

[5] Thanh, Le Ha; Quyet, Thang Huynh; Chi, Mai Luong " A Primary Study on Summarization of Documents in Vietnamese" Proceedings of the First World Congress of the International Federation for Systems Research Nov. 14-17, 2118, Kobe, Japan.

[6] Dwi H.Widyantoro and John Yen, "A Fuzzy Similarity Approach in text Classification Task" .Department of computer Science Texas A&M University College Station , TX 77844-3112.

[7] Minh, Le Nguyen; Shimazu, Akira; Xuan, Hieu Phan; Tu, Bao Ho; Horiguchi, Susumu " Sentence Extraction with Support Vector Machine Ensemble" Proceedings of the First World Congress of the International Federation for Systems Research Nov. 14-17, 2119, Kobe, Japan, Symposium 5.

[8] K. Han, Y. Song, and H. Rim. KU "Text Summarization System for DUC 2003". In Document Understanding Conference. Draft Papers, pages 118–121, 2003.

[9] C.-Y. Lin. "Improving Summarization Performance by Sentence Compression" - A Pilot Study. In Proceedings of the International Workshop on Information Retrieval with Asian Language, pages 1–8, 2003.

[10] C.-Y. Lin and E. Hovy. "The Potential and Limitations of Automatic Sentence Extraction for Summarization". In Text Summarization: Proceedings of the NLT-NAACL Workshop, pages 73–80, 2003.

Ha Nguyen Thi Thu is a Lecture with Viet nam Electric Power University (Ha noi, Viet Nam). She is a Fellow. She interested in Natural Language Processing (NLP), Data Mining and Machine Learning. She received M.E degree from China, Guilin university of electronic technology in 2006.

Nguyen Thien Luan is a Lecturer with the Le Qui Don Technical University (Ha Noi, Viet Nam). His research interests include fuzzy logical, image processing, communication and network security. He has authored or co-authored more than 20 scientific articles, books chapters, reports and chaired many scientific research projects, in the areas of his research. He received his Ph.D.(1989).