International Journal of Innovative
Computing, Information and Control
Volume 9, Number 12, December 2013

 ISSN 1349-4198
pp. 4903–4915

A SEMI-SUPERVISED LEARNING METHOD COMBINED WITH DIMENSIONALITY REDUCTION IN VIETNAMESE TEXT SUMMARIZATION

Ha Nguyen Thi Thu and Quynh Nguyen Huu

Faculty of Information Technology
Electric Power University
235 Hoang Quoc Viet, Hanoi, Vietnam
{ hantt; quynhnh }@epu.edu.vn

Received December 2012; revised May 2013

Abstract. *The World Wide Web has brought us a vast amount of online information. When we search with a keyword, we get data feedback from many different websites and the user cannot read all the information. So, text summarization has become a hot topic, and it has attracted experts in data mining and natural language processing field. For Vietnamese, some methods of text summarization that have been proposed for English also bring some significant results. However, there still remain some difficult problems to treat with the Vietnamese language processing; typical in this are the Vietnamese text segmentation tool and text summarization corpus. In this paper, we present a Vietnamese text summarization method based on sentence extraction approach using neural network for learning combined with reducing dimensional features to overcome the cost when building term sets and reduce the computational complexity. The experimental results show that our method is really effective in reducing computational complexity, and is better than some methods that have been proposed previously.*
Keywords: Vietnamese text summarization, Sentence extraction, Neural network, Semi-supervised learning

1. **Introduction.** Text summary is one aspect of data mining that is really necessary to help readers easily find important information from a long text by shortening the length and content of the original text [1,18].

Building a successful automatic text summarization system requires a lot of different factors. First, a text summarization system must extract the most important information from original text. Second, this system must be really effective and takes a short time to show results. Third, cost to build this system is not too much [1,2,4,10,11,17,20].

Because of three problems above, researchers always try to find effective solutions to build a system of automatic text summarization. For English text, many automatic text summarization systems have been developed [1,16]. For Vietnamese text, the proposed methods are not too many. Most of its applied methods are proposed for English, so it is not high performance.

One of the reasons why applying the English text summarization method for Vietnamese is not really effective is because of the different characteristics language between English and Vietnamese. Vietnamese language is single syllable. Unlike English, words in Vietnamese text cannot be determined by whitespace [13-15,31]. Therefore, when building a text summarization system, the system is relatively complex (need integrate a word segmentation tool), system will not be effective in terms of time, cost, and accuracy of word segmentation tool does not reach 100%.

Some studies that have been proposed to construct term sets (called topic model) [27] also increase the accuracy of the system. However, the cost to construct term sets is high, needs a long time and requires experts who understand language. So it is not a good solution.

In this paper, we proposed a new method for Vietnamese text summarization using semi-supervised learning combined with feature reduction. This approach can reduce cost and complex computing, but the result is better than unsupervised learning approach.

The structure of this paper is as follows. In Section 2, we present some related works. In Section 3, we will introduce the way of feature reduction. Section 4 is methodology of sentence extraction method using neural network for training combined with dimensional feature reduction. Section 5 and Section 6 show the experimental results and conclusion.

2. **Related Works.**

Most early work on single-document summarization focused on technical documents. The first paper proposed by Luhn [22] describes research done at IBM in the 1950's. In his work, Luhn proposed that the frequency of a particular word in an article provides a useful measure of its significance. There are several key ideas put forward in this paper that have assumed importance in later work on summarization. All sentences are ranked in order of their significance factor, and the top ranking sentences are finally selected to form the auto abstract. Baxendale, also done at IBM and published in the same journal, provides early insight on a particular feature helpful in finding salient parts of documents: the sentence position. This positional feature has since been used in many complex machine learning based systems. Edmundson [11] describes a system that produces document extracts. The two features of word frequency and positional importance were incorporated from the previous two works. Two other features were used: the presence of cue words (presence of words like significant, or hardly), and the skeleton of the document (whether the sentence is a title or heading). Weights were attached to each of these features manually to score each sentence.

In the 1990's, with the advent of machine learning techniques in NLP, a series of seminal publications appeared that employed statistical techniques to produce document extracts: Bayes, neural network, hidden markov model and so on. Kupiec et al. [19] describe a method derived from Edmundson [11] that is able to learn from data. The classification function categorizes each sentence as worthy of extraction or not, using a naive-Bayes classifier. Aone et al. [1] also incorporated a naive-Bayes classifier, but with richer features. They describe a system called DimSum that made use of features like term frequency (tf) and inverse document frequency (idf) to derive signature words. The idf was computed from a large corpus of the same domain as the concerned documents. In contrast with previous approaches, that were mostly feature-based and nonsequential, Conroy et al. [18] modeled the problem of extracting a sentence from a document using a hidden Markov model (HMM). The basic motivation for using a sequential model is to account for local dependencies between sentences. Only three features were used: position of the sentence in the document (built into the state structure of the HMM), number of terms in the sentence, and likeliness of the sentence terms given the document terms.

For Vietnamese, most of methods are proposed based on the methods that applied to English and used Vietnamese word segmentation tool [31], or based on SVM model for extracting [25].

3. **Feature Reduction.**

3.1. **Text representation.**

Text representation methods describe the content or characteristics of the text. Commonly, word frequency is considered as characteristic of text.

Suppose that we have a document d, Figure 1 illustrates representing d. Here, w_{ij} is the weight of word (or term) t_{ij}, i is sentence ith in document d, j is position of word (term) in the sentence ith from left to right.

$$T = \begin{Bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \cdots & \cdots & & \cdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n} \end{Bmatrix}$$

FIGURE 1. Matrix of text representation

Example 3.1. *Have an original Vietnamese text including 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 into 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

$$T = \{t_{1,1}, t_{1,2}, \ldots, t_{1,34}\} \tag{1}$$

In the general case, the original text often includes multiple sentences, each sentence consists of many words, so if expressed in the above, the weight matrix representation of the text will have some very large, requiring more time for computation. With common Vietnamese text, first we need separate all the words (use word segmentation tool) in the text and to compute weight of each word will require more and more time treatment.

3.2. **Methodology of feature reduction.** Feature selection is one of the key topics in machine learning and other related fields. Real-life datasets are often characterized by a large number of irrelevant or redundant features that may significantly hamper model accuracy and learning speed if they are not properly excluded. Feature selection involves finding a subset of features to improve prediction accuracy or decrease the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features.

To overcome the disadvantages of large feature vector and the cost to build a large set of terms for each topic, in this paper, we propose a new method which 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 text. So, when we extract nouns from text, remarkable reduction of large feature set appears.

In Example 3.1, we separate document d into two sets: the first set includes 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, and the size of the matrix T will be reduced; for example, with the original text in Example 3.1, instead of using the T matrix containing one row and 34 columns, we only need the matrix T' consisting of one row and 14 columns:

$$T = \{t_{1,1}, t_{1,2}, \ldots, t_{1,14}\} \tag{2}$$

So now, we reduced from 34 feature vectors to 14 feature vectors.

4. Methodology of Vietnamese Sentence Extraction Using Neural Network.

4.1. **Neural network.** A neural net is an artificial representation of the human brain that tries to simulate its learning process. Artificial neural network is an interconnected group of artificial neurons that uses a mathematical model of computational model for information processing based on a connectionist approach to computation.

A feed-forward network has a layered structure. The architecture of this class of network, besides having the input and the output layers, also has one of more intermediary layers called hidden layers. The computational unites of the hidden layer are known as hidden neurons.

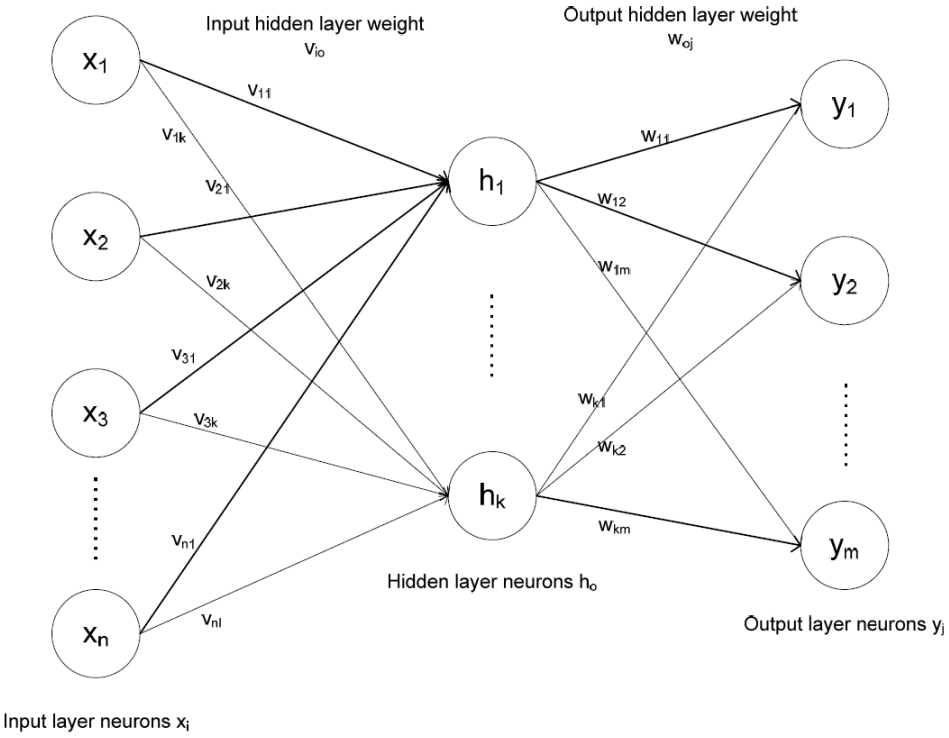

FIGURE 2. Neural network

The learning methods in neural networks are classified into three basic types:

- Supervised learning
- Unsupervised learning
- Reinforced learning

So supervised learning is meaningful: have a teacher is present during learning process and presents expected output, every input pattern is used to train the network and learning process is based on comparison between network's computed output and the correct expected output, generating "error". The "error" generated is used to change network parameters that result in improved performance. In general, supervised learning includes two phases: Forward phase and Backward phase.

- Forward: compute 'functional signal', feed forward propagation of input pattern signals through network.
- Backward: compute 'error signal', and propagate the error backwards through network starting at output units (where the error is the difference between actual and desired output values).

- Output unit:

$$w_{ij}(t+1) - w_{ij}(t) = \eta\Delta_i(t)z_j(t) = \eta(d_i(t) - y_i(t))g'(a_i(t))z_j(t) \tag{3}$$

- Hidden unit

$$v_{ij}(t+1) - v_{ij}(t) = \eta\delta_i(t)x_j(t) = \eta g'(u_i(t))x_j(t)\sum_k \Delta_k(t)w_{ki} \tag{4}$$

4.2. **Features.** Extractive summaries are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or cue words to locate the sentences to be extracted. The "most important" content is treated as the "most frequent" or the "most favorably positioned" content. Such an approach thus avoids any efforts on deep text understanding. They are conceptually simple, easy to implement. In this paper, we use sentence extraction approach for text summarization, so we need determined features, which used to calculate weight of sentence. In training D, we separate to set of sentence:

$$S = \{s_1, s_2, \ldots, s_n\}$$

Three features are used for calculating the sentence score:

- Information significant of sentence:

Information significant of sentence expresses information of sentence:

$$F_{infor}(s_k) = \sum_{i=1}^{n} I(w_i) \tag{5}$$

where:

- F_{infor}: score information of sentence s_k.
- $I(w_i)$: information significant of topic word in s_k, that is calculated by

$$I(w_i) = \frac{N_s(w_i)}{\sum\limits_{w_i \in S} w_i} + \frac{N_D(w_i)}{N_D} \tag{6}$$

where:

- $N_S(w_i)$ is number of w_i occur in sentence.
- $\sum\limits_{w_i \in S} w_i$ is the number of total of all topic words occurring in sentence.
- $N_D(w_i)$ is the number of document in training set D that has w_i.
- N_D is the total number in the training set D.

- Position of sentence:

This feature is based on the assumption that sentences occurring in initial position of both text and individual paragraphs have a higher probability of being relevant. We sign in the first sentence and the last sentence of paragraph is 1 and score is $1/i$ if other.

- Amount of information in sentence:

Amount of information in sentence can be determined by number of nouns in sentence.

$$F_{number\ of\ topic\ word}(s_k) = \frac{N(w_i)}{\max\{N(w_j)\}} \tag{7}$$

4.3. **Supervised training with neural network.** This method involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network learns the patterns inherent in sentences that should be included in the summary and those that should not be included. It uses three-layered Feed forward neural network, which has been proven to be a universal function approximator. The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. The neural network used for training is like Figure 3.

We used neural network for training; this neural network has three layers, one hidden layer. The value $[0, 1]$ of output denotes that sentence can be extracted or not. Figure 4 below is a calculating feature score algorithm.

In Figure 4, ***WEIGHT*** algorithm has some parameters: W' is list of sentences that include features score. I is information significant score, P is position score, M is amount of information score. ***MATCH()*** is a function that returns position of sentence s in document d.

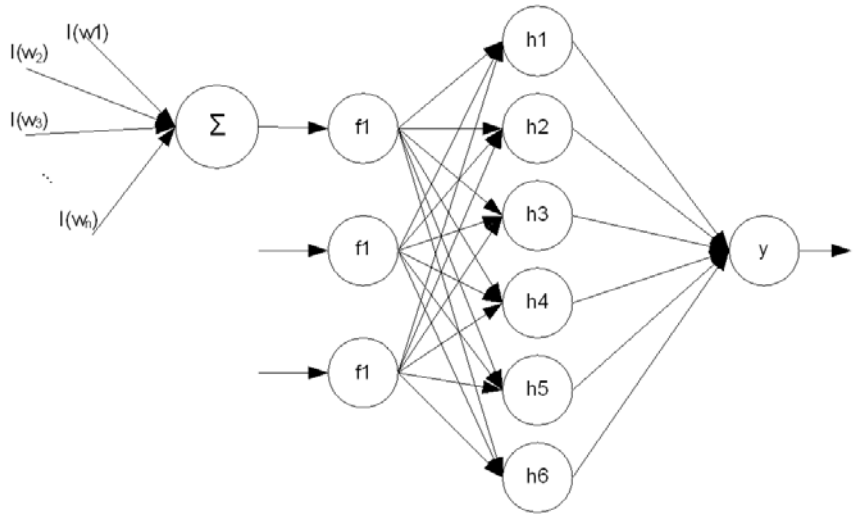

FIGURE 3. Neural network for training

FEATURE SCORE (WEIGHT)

Input
S: set of sentences; V: dictionary of nouns;
Output
W': list of sentences and value coresponding of it
1. Initalization
W'=ϕ; I=ϕ; P=ϕ; M=ϕ; x=0;y=0;
2. For i=1 to count(S) do
2.1 For k=1 to length(s_i) do
2.1.1 If (s_i[k] is noun)
2.1.1.1 x=x+I(s_i[k]);
2.1.2 I[k]←x;
2.1.3 x=0;
2.2 For k=1 to length(s_i) do
2.2.1 If (s_i[k] is noun) calculate
2.2.1.1 y++;
2.2.2 M[k] ←y;
2.2.3 y=0;
2.3 P[k]= ←***MATCH***(s_i,d);

FIGURE 4. Feature selection algorithm

SENTENCE SELECTION (SENS)
Input
C: original document; V: dictionary of nouns;
Output
C': summary;
1. Initalization
S=ϕ ; W'=ϕ ; T=ϕ ;
2. Sentence segmentation
S ← Split(C); // split C in to sentences
3. Calculating feature
For k=1 to count(S) do
W'[k] ← Weight(s[k],V);
4. Calculating score of sentence
For k=1 to count(S) do
T[k] ← Train(W'[k]);
5. decision for selection or not
6. Sorting by original document order

FIGURE 5. Sentences selection algorithm

Figure 5 is sentences selection algorithm that illustrates how to select sentences and summarizer.

This SENS algorithm has some parameters: C is a document that wants to summarize it. S is list of sentences in C, W' is weight of s_k in S and T is score of sentence s after training with neural network. ***Split*** is a procedure to split C into sentences, ***Train()*** is a function that returns score of sentence after training.

5. Experimental Results.

5.1. **Corpus.** There is not standard corpus for Vietnamese text summarization now. So, in our experiment, we built corpus manually. Documents in this corpus had been downloaded from websites's news as: http://thongtincongnghe.com, http://echip.com, http://vnexpress.net, http://vietnamnet.vn, http://tin247.com... Corpus's entitle is "information" and "technology". There are over 300 documents in it. We segmented from 300 documents into 16,117 sentences.

TABLE 1. Some documents in corpus

Document	Source	Sentences	File name
Ứng dụng Twitter trong lớp học	thongtincongnghe.com	28	18-10.txt
Hacker "sờ tới" website chính phủ Malaysia	Vietnamnet.vn	15	11-5.txt
Yahoo ra mắt công cụ tìm kiếm app cho Android	Ngoisao.net	12	12-9.txt
TQ phủ nhận điều tra chống độc quyền Microsoft	Tin247.com	21	13-8.txt
Cấu hình tối thiểu để nâng cấp lên Mac OS X Lion	Sohoa.vnexpress	18	16-3.txt
Chọn hệ điều hành của bạn	pcworld.com	69	21-10.txt
Linux ở khắp mọi nơi	Vietbao.vn	71	22-1.txt
Màn hình cảm ứng: Đằng sau những cú chạm	Pcworld	86	25-4.txt
Phanh phui bí mật thế giới ngầm hacker Việt Nam	Echip.com	7	33-4.txt
Người dùng di động quan tâm giá cả hơn sáng tạo công nghệ	baomoi.com	39	33-7.txt

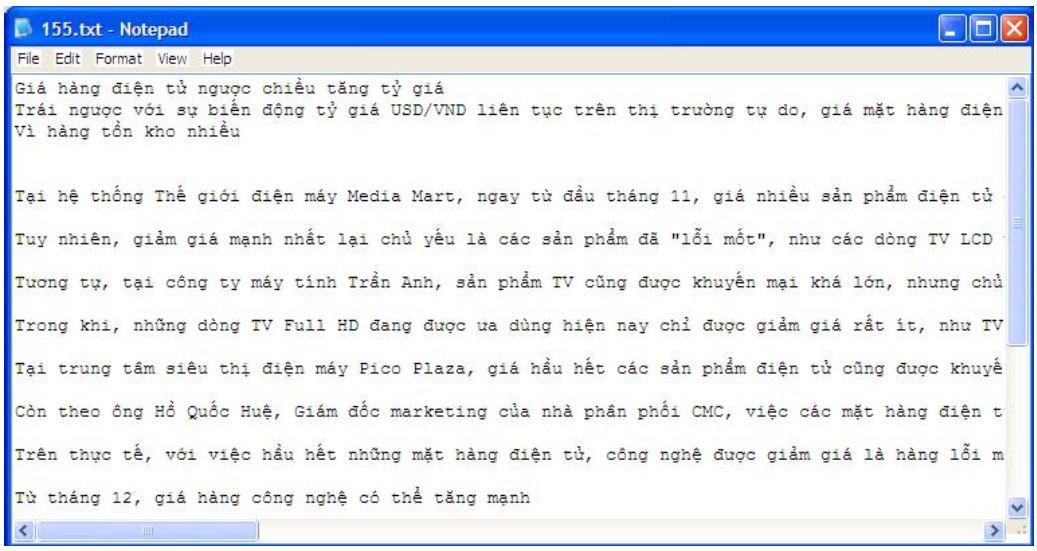

FIGURE 6. A file for training

All files downloaded from website will be saved in corpus by *.txt and preprocessed. Figure 6 illustrated a file in corpus that preprocessed.

5.2. **Word segmentation và labeled.** We build a dictionary of nouns and used vlsp tool for segmentation nouns. Vlsp tool is published on Internet via address: http://vlsp.vietlp. or g:8080/demo/. We selected 814 sentences in the corpus for signed label and used vlsp for segmenting nouns set. And only noun will be updated weight.

5.3. **Training.** In training phase, corpus includes 11,670 sentences, but we only used 814 sentences for labeling. Sentences will be segmented from documents and through 3 steps for training: labeled step, feature calculation step and training with neural network step.

- Labeler: Sentence labeled by manual, each sentence has been signed a number in $[0, 1]$. This processing supported with language experts. With each sentence we sign a number in $[0, 1]$. Figure 7 below illustrated signing for sentences in corpus.
- Feature calculator: 814 sentences calculated three features: information significant score, position score and amount of information score.
- Training: In training phase, first, weight of neural network has random initializer, 814 sentences with 3 features are input of neural network. Back propagation phase

ID_Textsentence	Content	Text	OutputMM
30	sau khi đã dọn dẹp xong bạn hãy vào lại mục virtual memor...	5	0.5
31	tăng ram ảo có thể giúp hệ thống của bạn tăng tốc	5	0.5
32	nếu còn dư ổ cứng thì bạn hãy đặt initial size là 500mb và m...	5	0.6
33	còn đối với những tệp tin mft - những tệp tin hệ thống chỉ có...	5	0.6
34	khi mà ổ cứng của bạn lưu càng nhiều dữ liệu thì các tệp tin...	5	0.6
35	tuy nhiên, nói như thế không có nghĩa là bạn không thể di ...	5	0.6
36	bạn vào run và gõ lệnh regedit để mở trình biên tập registry	5	0.6
37	sau khi trình biên tập registry khởi động bạn hãy tìm đến kh...	5	1
38	đóng trình biên tập registry rồi khở động lại hệ thống của bạn	5	1
39	ổ cứng của bạn sẽ thực sự chạy nhanh và ổn định hơn rất ...	5	1
40	bạn đã back-up lại dữ liệu, mã hoá và tối ưu hoá dữ liệu củ...	5	1

FIGURE 7. Sentence labeler

ID_Sentence	Infor_significant	Position	Amount_infor	OutputNN
4	0	0.25	0	0.339735209941...
5	0.813660477453...	0.2	0.083333333333...	0.433437913656...
6	0.826259946949...	0.166666666666...	0.083333333333...	0.630042016506...
7	0.828912466843...	0.142857142857...	0.166666666666...	0.639999747276...
8	0.836870026525...	0.125	0.333333333333...	0.641342580318...
9	0.836870026525...	0.111111111111...	0.416666666666...	0.763624906539...
10	0	0.1	0	0.765968441963...
11	0.854111405835...	0.090909090909...	0.25	0.649469733238...
12	0.812997347480...	0.083333333333...	0.083333333333...	0.628403306007...
13	0	0.076923076923...	0	0.336953490972...
14	0.837533156498...	0.071428571428...	0.083333333333...	0.435314506292...
15	0	0.066666666666...	0	0.669492244720...
16	0	0.0625	0	0.669340372085...
17	0.847480106100...	0.058823529411...	0.25	0.648843169212...
18	0.927718832891...	0.055555555555...	0.416666666666...	0.763591229915...
19	0	0.052631578947...	0	0.611639738082...
20	0.970159151193...	0.05	0.5	0.679388344287...
21	0.843501326259...	0.047619047619...	0.166666666666...	0.867279708385...

FIGURE 8. Update score of sentences with final weights

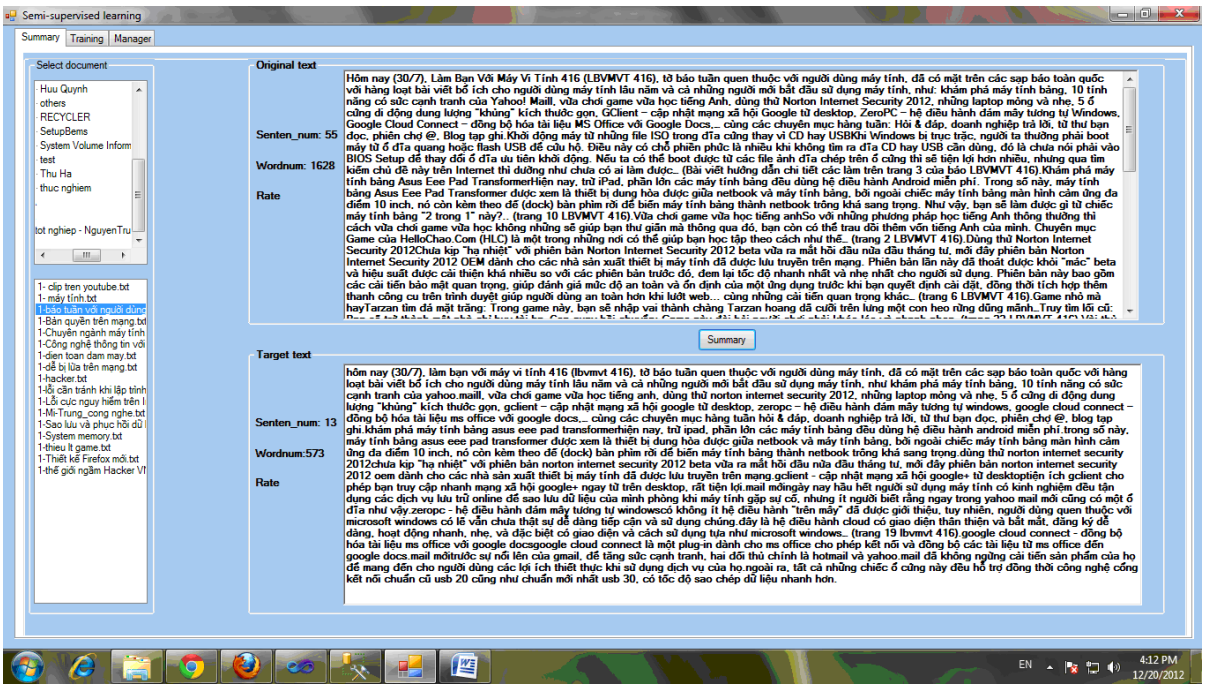

FIGURE 9. SEBSSL system

will be stopped when error ∼0.01. Figure 8 is score of sentences with final weights of neural networks.

5.4. **Result.** Based on the proposed method, we built the program to experiment called SEBSSL system (Sentence extraction based on semi-supervised learning). Figure 9 is SEBSSL system.

It is difficult to compare our method with previous ones, because there were no widely accepted benchmarks for Vietnamese text extraction sentences. Therefore, we compare our proposed method with an algorithm suggested by Thanh et al. [31], Minh et al. [25] and baseline method gives better results than their methods.

Below is an original document. That has 57 sentences and length of it is 2035. This document is shown in Figure 10.

And Figure 11 is summary of original text. That has 18 sentences and its length is 959. Target text (summary) is in Figure 11.

Công nghệ thông tin với những tiện ích của nó đang rất được sinh viên quan tâm và ứng dụng nhằm nâng cao chất lượng học tập. Với sinh viên sư phạm thì việc ứng dụng công nghệ thông tin thể hiện rõ nhất ở việc thiết kế giáo án điện tử.Thạc sĩ Trần Thị Kim Oanh, cán bộ giảng dạy thuộc tổ phương pháp dạy học văn (Khoa Ngữ văn) cho biết: Giáo án điện tử - một hình thức soạn bài giảng trên máy vi tính - là sự kết hợp những ưu điểm đã được khẳng định của giáo án truyền thống và những tiện ích của công nghệ thông tin. Đây là phương tiện để giáo viên sử dụng một cách hiệu quả phương pháp trực quan với những ưu điểm: sinh động, cụ thể, thẩm mỹ, lôi cuốn, hấp dẫn… Theo thạc sĩ thì cách thiết kế bài giảng giáo án điện tử có rất nhiều ưu điểm như phát huy tính tích cực, chủ động, sáng tạo của học sinh; học sinh có thêm phương tiện học tập khác ngoài sách giáo khoa, bảng đen, phấn trắng. Các em sẽ được tiếp cận nội dung bài học một cách đa dạng, phong phú hơn, vận dụng được nhiều kỹ năng hơn như nghe, nhìn, đọc, nói... Chính vì những tiện ích ấy, đồng thời xuất phát từ thực tiễn việc dạy và học văn trong nhà trường hiện nay, việc hướng dẫn sinh viên sư phạm, những thầy cô giáo tương lai, thiết kế giáo án điện tử, ứng dụng công nghệ thông tin đã có sự quan tâm nhất định. Trong chương trình học tập môn phương pháp dạy học văn, sinh viên đã bước đầu được làm quen với phương pháp này.Tuy nhiên, chỉ mới có một bộ phận sinh viên trong khoa được hướng dẫn soạn bài giảng trên máy vi tính. Bạn Đinh Thị Thái Hiền (sinh viên lớp K27A) cho biết: "Hầu hết các bạn đều muốn tham gia, có bạn đã thiết kế đến 2 giáo án". Giải thích về sự hưởng ứng nhiệt tình của sinh viên, Thái Hiền cho rằng: "Sinh viên ai cũng muốn tìm tòi, khám phá, sáng tạo cái mới nhằm chuẩn bị tư thế để trở thành giáo viên tương lai. Các bạn cũng ý thức được vai trò và yêu cầu của một người giáo viên tương lai trong việc góp phần nâng cao chất lượng dạy học trong nhà trường phổ thông và hơn nữa, đó là một công việc mang đến cho họ nhiều thú vị". Những điều thú vị mà cô bạn sinh viên này "khám phá" được trong thời gian thiết kế bài giảng "Đây thôn Vĩ Dạ" của Hàn Mặc Tử chính là tiếp cận được thế giới tiện ích mênh mông của công nghệ thông tin và Internet, tha hồ lựa chọn, sáng tạo, càng làm càng mê, nâng cao khả năng sử dụng tin học, kỹ năng soạn giáo án súc tích, cô đọng và nhất là nâng cao được tri thức qua việc tìm tài liệu, hình ảnh hỗ trợ bài giảng trên Internet. Đây cũng chính là những bộc bạch của các bạn Nguyễn Thị Hạnh, Nguyễn Đình Khoa, Phạm Thành Hiệp (lớp K27B), Nguyễn Thị Thu Phương (lớp K27C) và những bạn có tham gia soạn giáo án điện tử. "Đó là cách thể hiện sự tâm huyết của mình về nghề trong tương lai, phải luôn tìm tòi sáng tạo các phương pháp giảng dạy thích hợp, hấp dẫn nhằm truyền đạt một cách đầy đủ hệ thống kiến thức đến học sinh" – bạn Minh Tân (lớp K27B) giải thích tại sao lại thích ứng dụng công nghệ thông tin vào thiết kế bài giảng.Không dừng ở việc soạn thử để học tập, một số bạn đã sử dụng giáo án này giảng dạy ở trường phổ thông trong các đợt thực tập. Đặc biệt, sinh viên Lê Ni La (lớp K26 C) đã mạnh dạn chọn đề tài "Ứng dụng công nghệ thông tin trong việc giảng dạy tác gia Hồ Chí Minh" làm luận văn tốt nghiệp đại học của mình. Nhưng con số 56 sinh viên năm học 2002 – 2003 và hơn 60 sinh viên năm học này tham gia thiết kế giáo án điện tử còn quá khiêm tốn. Trong số giáo án điện tử được thực hiện mà Thạc sĩ Trần Thị Kim Oanh hướng dẫn thì cũng còn rất nhiều thiếu sót, hạn chế. Thạc sĩ nhận xét: các giáo án soạn trên máy vi tính còn tham kiến thức, chưa hàm súc, khái quát – là những yêu cầu quan trọng của giáo án điện tử. Khả năng xử lý, các thao tác vận dụng tiện ích của công nghệ thông tin như chèn nhạc, chèn hình ảnh, tranh vẽ còn nhiều lúng túng. Rất nhiều sinh viên phải nhờ vả, thậm chí thuê người làm. Chưa đảm bảo tính khoa học, hợp lý giữa các trang (slide) nhất là phần yêu cầu cần đạt và phần củng cố dặn dò, phần phải sửa chữa, điều chỉnh nhiều nhất. Tư liệu, hình ảnh còn thiếu chọn lọc, sử dụng thiếu khoa học, đôi chỗ còn tuỳ tiện, không có tác dụng, phông nền cầu kỳ, lạm dụng yếu tố màu sắc làm giảm tính khoa học của bài giáo án điện tử. Như vậy, có thể thấy, nhu cầu tiếp cận công nghệ thông tin và ứng dụng nó vào trong việc tập soạn giáo án ở sinh viên sư phạm là đáng ghi nhận. Điều này cũng phù hợp với chủ trương, xu hướng đổi mới giảng dạy và học tập ở trường phổ thông. Không ai khác hơn sinh viên sư phạm phải là những người đi đầu trong xu hướng này. Với sinh viên các khoa tự nhiên thì có lẽ việc đó không quá khó và nhiều bỡ ngỡ như sinh viên các khoa xã hội, đặc biệt là sinh viên ngữ văn, bởi những đặc trưng rất riêng của bộ môn này. Bởi vậy, số lượng sinh viên tham gia soạn giáo án điện tử còn hạn chế.Nguyên nhân do đâu?Đem băn khoăn này hỏi một số sinh viên không tham gia thiết kế giáo án điện tử, chúng tôi ghi nhận, phần lớn ý kiến đều khẳng định không biết sử dụng máy vi tính thành thục, thậm chí có bạn còn thừa nhận chưa thể tự đánh máy một văn bản. Tuy có được học tin học nhưng thời gian học không nhiều, quá trình thực hành cũng hạn chế, hầu hết chưa được học sử dụng phần mềm Power Point - một công cụ chủ yếu để thiết kế giáo án điện tử - nên không dám đăng ký thực hiện. Đây là một điểm yếu của sinh viên Việt Nam hiện nay, ngoài điểm yếu về ngoại ngữ. Còn với sinh viên đăng ký thực hiện thì lại vướng vào vấn đề khác, khá tế nhị - vấn đề tài chính" – bạn Thái Hiền "than". Phải lên mạng tìm hình ảnh, chọn nhạc, âm thanh, phim…, những bạn có máy tính còn đỡ, còn không phải đi thuê, tốn tiền nhiều lắm... Ngoài ra, nó cũng cần nhiều công sức, thời gian. Sinh viên Trịnh Thị Minh Hương (K27 B) kể: "Tụi mình nhiều lúc đã dở khóc dở cười vì cái chuyện ghi bài giảng đã thiết kế vào đĩa CD. Ngồi cả ngày ở dịch vụ cố gắng làm xong thì mới phát hiện ở đó không ghi được CD. Còn chỗ ghi CD thì không cho thuê máy". Và hậu quả là đủ giáo án được soạn rất công phu, đẹp mắt nhưng vẫn không đủ thời gian để nộp. Đó là trường hợp của các bạn Võ Văn Khôi, Tăng Thị Tuyết Mai (lớp K27C) và nhiều bạn khác. Cũng có nhiều bạn lúc đầu đăng ký nhưng khi thực hiện không kham được đã bỏ cuộc. Nhìn ở góc độ khác, Thạc sĩ Trần Thị Kim Oanh cho biết: Để soạn giáo án điện tử thì sinh viên phải có hiểu biết cơ bản về máy vi tính, biết sử dụng những tiện ích như Power Point, Herosoft… Có thể nói, không có hiểu biết về công nghệ thông tin, không sử dụng được máy vi tính, không thể soạn giảng trên CD hoặc VCD. Có nhiều lý do sinh viên chưa tích cực ứng dụng công nghệ thông tin, song có lẽ, điều kiện khách quan là các em chưa được đào tạo một cách hệ thống, cơ bản về máy tính trong quá trình học đại học. Như thế có thể thấy mức độ phổ cập tin học cho sinh viên hiện nay quá chậm và chưa hiệu quả. Đáng lẽ, họ phải được trang bị kiến thức tin học cơ bản từ khi học phổ thông thì nhiều sinh viên (nhất là sinh viên các tỉnh vùng sâu, vùng xa) mới biết lơ mơ, thậm chí mù tịt. Công tác này gần đây có được quan tâm hơn nhưng xem ra còn hạn chế. Mặt khác, quan niệm của cán bộ giảng dạy về việc ứng dụng công nghệ thông tin vào dạy học vẫn chưa thống nhất.Giải pháp nào cho thực trạng trên? Với sinh viên thì cần nhất là tinh thần tự học tập, tự đào tạo, khao khát chiếm lĩnh tri thức là cách tốt nhất để tự nâng cao trình độ chuyên môn và kỹ năng nghiệp vụ của mình. "Sinh viên cần quan tâm nghiêm túc hơn, đúng đắn hơn về việc rèn luyện năng lực thiết kế bài giảng (nhất là đối với giáo án truyền thống), vận dụng những điều đã học vào thực tế, nhanh chóng nắm lấy công nghệ thông tin với những tiện ích của nó để tích cực hoá, sinh động hoá việc dạy văn bởi vì đó là con đường tốt nhất để thực hiện phương pháp trực quan trong xu hướng đổi mới dạy học hiện nay" – Thạc sĩ Trần Thị Kim Oanh nhấn mạnh. Thạc sĩ cũng đề nghị nhanh chóng tăng cường chương trình đào tạo tin học, đưa tin học vào học đường, sớm giảng dạy cho sinh viên ứng dụng tin học, mỗi khoa nên có phòng nghe nhìn chuyên dụng. Sinh viên Lê Minh Tân (lớp K27B) đề xuất với Đoàn – Hội trường là nên mở các buổi chuyên đề về tin học để phổ cập kiến thức căn bản về phần mềm Power Point cho sinh viên để giáo án điện tử nhận được sự quan tâm của nhiều sinh viên và nhất là giúp sinh viên có cơ hội tiếp cận với công nghệ thông tin ngay trước khi họ rời trường. Thật ra, trung tâm tin học của trường đã tổ chức nhiều lớp chuyên đề như thế và có chính sách miễn giảm học phí cho sinh viên nhưng chưa thấm vào đâu so với nhu cầu của sinh viên và học phí, thời gian cũng chưa phải lúc nào cũng thuận lợi cho sinh viên. Với tư cách là Chủ tịch Hội Sinh viên Trường Đại học Sư phạm, Đinh Thị Thái Hiền cho biết Đoàn trường và Hội sinh viên đã, đang và sẽ hỗ trợ sinh viên trong việc tiếp cận và ứng dụng công nghệ thông tin vào việc học tập, thiết kế thử giáo án. Cụ thể, sẽ phối hợp tổ chức chuyên đề phổ cập tin học cho sinh viên, giới thiệu các lớp học tin học miễn phí hoặc học phí thấp cho sinh viên, phối hợp với các câu lạc bộ học thuật các khoa giới thiệu các nguồn tài liệu, các địa chỉ website trên Internet phục vụ cho việc ứng dụng công nghệ thông tin vào học tập, nhất là trong việc soạn giáo án điện tử. "Nhưng quan trọng nhất là sự quyết tâm và ý thức trách nhiệm với nghề, với xã hội của mỗi sinh viên" – Thái Hiền gửi gắm.

FIGURE 10. Original text

công nghệ thông tin với những tiện ích của nó đang rất được sinh viên quan tâm và ứng dụng nhằm nâng cao chất lượng học tập.với sinh viên sư phạm thì việc ứng dụng công nghệ thông tin thể hiện rõ nhất ở việc thiết kế giáo án điện tử.chính vì những tiện ích ấy, đồng thời xuất phát từ thực tiễn việc dạy và học văn trong nhà trường hiện nay, việc hướng dẫn sinh viên sư phạm, những thầy cô giáo tương lai, thiết kế giáo án điện tử, ứng dụng công nghệ thông tin đã có sự quan tâm nhất định.những điều thú vị mà cô bạn sinh viên này "khám phá" được trong thời gian thiết kế bài giảng "đây thôn vĩ dạ" của hàn mặc tử chính là tiếp cận được thế giới tiện ích mênh mông của công nghệ thông tin và internet, tha hồ lựa chọn, sáng tạo, càng làm càng mê, nâng cao khả năng sử dụng tin học, kỹ năng soạn giáo án súc tích, cô đọng và nhất là nâng cao được tri thức qua việc tìm tài liệu, hình ảnh hỗ trợ bài giảng trên internet."đó là cách thể hiện sự tâm huyết của mình về nghề trong tương lai, phải luôn tìm tòi sáng tạo các phương pháp giảng dạy thích hợp, hấp dẫn nhằm truyền đạt một cách đầy đủ hệ thống kiến thức đến học sinh" – bạn minh tân (lớp k27b) giải thích tại sao lại thích ứng dụng công nghệ thông tin vào thiết kế bài giảng.đặc biệt, sinh viên lê ni la (lớp k26 c) đã mạnh dạn chọn đề tài "ứng dụng công nghệ thông tin trong việc giảng dạy tác gia hồ chí minh" làm luận văn tốt nghiệp đại học của mình.với sinh viên các khoa tự nhiên thì có lẽ việc đó không quá khó và nhiều bỡ ngỡ như sinh viên các khoa xã hội, đặc biệt là sinh viên ngữ văn, bởi những đặc trưng rất riêng của bộ môn này.đem băn khoăn này hỏi một số sinh viên không tham gia thiết kế giáo án điện tử, chúng tôi ghi nhận, phần lớn ý kiến đều khẳng định không biết sử dụng máy vi tính thành thục, thậm chí có bạn còn thừa nhận chưa thể tự đánh máy một văn bản.tuy có được học tin học nhưng thời gian học không nhiều, quá trình thực hành cũng hạn chế, hầu hết chưa được học sử dụng phần mềm power point - một công cụ chủ yếu để thiết kế giáo án điện tử - nên không dám đăng ký thực hiện.có nhiều lý do sinh viên chưa tích cực ứng dụng công nghệ thông tin, song có lẽ, điều kiện khách quan là các em chưa được đào tạo một cách hệ thống, cơ bản về máy tính trong quá trình học đại học.đáng lẽ, họ phải được trang bị kiến thức tin học cơ bản từ khi học phổ thông thì nhiều sinh viên (nhất là sinh viên các tỉnh vùng sâu, vùng xa) mới biết lơ mơ, thậm chí mù tịt.với sinh viên thì cần nhất là tinh thần tự học tập, tự đào tạo, khao khát chiếm lĩnh tri thức là cách tốt nhất để tự nâng cao trình độ chuyên môn và kỹ năng nghiệp vụ của mình."sinh viên cần quan tâm nghiêm túc hơn, đúng đắn hơn về việc rèn luyện năng lực thiết kế bài giảng (nhất là đối với giáo án truyền thống), vận dụng những điều đã học vào thực tế, nhanh chóng nắm lấy công nghệ thông tin với những tiện ích của nó để tích cực hoá, sinh động hoá việc dạy văn bởi vì đó là con đường tốt nhất để thực hiện phương pháp trực quan trong xu hướng đổi mới dạy học hiện nay" – thạc sĩ trần thị kim oanh nhấn mạnh.thạc sĩ cũng đề nghị nhanh chóng tăng cường chương trình đào tạo tin học, đưa tin học vào học đường, sớm giảng dạy cho sinh viên ứng dụng tin học, mỗi khoa nên có phòng nghe nhìn chuyên dụng.sinh viên lê minh tân (lớp k27b) đề xuất với đoàn – hội trường là nên mở các buổi chuyên đề về tin học để phổ cập kiến thức căn bản về phần mềm power point cho sinh viên để giáo án điện tử nhận được sự quan tâm của nhiều sinh viên và nhất là giúp sinh viên có cơ hội tiếp cận với công nghệ thông tin ngay trước khi họ rời trường.thật ra, trung tâm tin học của trường đã tổ chức nhiều lớp chuyên đề như thế và có chính sách miễn giảm học phí cho sinh viên nhưng chưa thấm vào đâu so với nhu cầu của sinh viên và học phí, thời gian cũng chưa phải lúc nào cũng thuận lợi cho sinh viên.với tư cách là chủ tịch hội sinh viên trường đại học sư phạm, đinh thị thái hiền cho biết đoàn trường và hội sinh viên đã, đang và sẽ hỗ trợ sinh viên trong việc tiếp cận và ứng dụng công nghệ thông tin vào việc học tập, thiết kế thử giáo án.cụ thể, sẽ phối hợp tổ chức chuyên đề phổ cập tin học cho sinh viên, giới thiệu các lớp học tin học miễn phí hoặc học phí thấp cho sinh viên, phối hợp với các câu lạc bộ học thuật các khoa giới thiệu các nguồn tài liệu, các địa chỉ website trên internet phục vụ cho việc ứng dụng công nghệ thông tin vào học tập, nhất là trong việc soạn giáo án điện tử.

FIGURE 11. Target document

TABLE 2. Experimental results

Method	Rate			
	80%	60%	40%	20%
Ours	0.875	0.82	0.76	–
Thanh et al.	0.62	0.754	0.698	0.543
Baseline	0.91	0.9	0.842	0.63
Minh et al.	0.83	0.71	0.73	–

5.5. **Evaluation.** At present, Vietnamese does not have any standard assessment method; therefore, we compare the result of our method with the results of extract according to the method proposed by Thanh et al. [31], Minh et al. [25] and baseline method.

Precision is the traditional assessment method given by:

$$Precision = \frac{correct}{correct + wrong} \tag{8}$$

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

6. **Conclusion.** Vietnamese natural language processing is one of the current hot topics in Vietnam. Currently, experts still try to find methods to solve problems related to natural language processing Vietnamese as Vietnamese text classification, Vietnamese text clustering, Vietnameset text summarization.

In this paper, we have proposed method to Vietnamese text summarization using neural networks; in addition, we also implement specific methods of dimensional reduction to reduce processing complexity, more accurate and effective than previous methods.

Acknowledgment. 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] C. Aone, M. E. Okurowski, J. Gorlinsky and B. Larsen, A trainable summarizer with knowledge acquired from robust NLP techniques, in *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury (eds.), Cambridge, MA, MIT Press, 1999.

[2] K. Knight and D. Marcu, Summarization beyond sentence extraction: A probabilistic approach to sentence compression, *Artif. Intell.*, vol.139, no.1, pp.91-107, 2002.

[3] M. Collins, *Head-Driven Statistical Model for Natural Language Parsing*, Ph.D. Thesis, Univ. of Pennsylvania, 1999.

[4] A. B. Goldberg, *New Directions in Semi-Supervised Learning*, Ph.D. Thesis, 2010.

[5] C. Nobata et al., A summarization system with categorization of document sets, *Proc. of the 3rd NTCIRT*, 2003.

[6] C. Y. Lin and E. H. Hovy, Automatic evaluation of summaries using n-gram co-occurrence statistics, *Proc. of Human Language Technology Conference*, Edmonton, Canada, 2003.

[7] D. Jurafsky and J. H. M. Daniel, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2008.

[8] D. Hakkani-Tür, Statistical sentence extraction for information distillation, *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol.4, pp.IV1-IV4, 2007.

[9] D. Das and A. F. T. Martins, *A Survey on Automatic Text Summarization*, 2007.

[10] *Document Understanding Conferences*, http://duc.nist.gov/.

[11] H. P. Edmundson, New methods in automatic extracting, *Journal of the Association for Computing Machinery*, pp.264-285, 1969.

[12] E. Hovy and C.-Y. Lin, Manual and automatic evaluation of summaries, *Proc. of the ACL-02 Workshop on Automatic Summarization*, vol.4, pp.45-51, 2002.

[13] N. T. T. Ha and N. T. Luan, Implement some features for better determining weight of sentence in Vietnamese text, *International Journal of Computer Science & Knowledge Engineering*, vol.4, no.2, pp.131-134, 2010.

[14] N. T. T. Ha and N. T. 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*, pp.41-46, 2010.

[15] N. T. T. Ha and N. H. Quynh, A new method for Vietnamese sentence extraction based on important information of topic word and linguistic score, *Proc. of IEEE on Multimedia and Computational Intelligence*, pp.567-570, 2010.

[16] E. Hovy and C. Lin, Automated text summarization in summarist, *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, pp.18-24, 1997.

[17] E. Hovy and C. Lin, Automated text summarization and the summarist system, *TIPSTER, Proc. of a Workshop on Held at Baltimore*, Maryland, pp.197-214, 1998.

[18] J. M. Conroy, J. D. Schlesinger and D. P. O'Leary, *Using HMM and Logistic Regression to Generate Extract Summaries for Duc*, DUC, 2001.

[19] J. Kupiec, J. Pedersen and F. Chen, A trainable document summarizer, *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.68-73, 1995.

[20] K. Jezek and J. Steinberger, *Automatic Text Summarization (The State of the Art 2007 and New Challenges)*, Znalosti, 2008.

[21] K. M. Svore, L. Vanderwende and C. J. C. Burges, Enhancing single-document summarization by combining RankNet and third-party sources, *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.448-457, 2007.

[22] H. P. Luhn, The automatic creation of literature abstracts, in *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury (eds.), Cambridge, MA, The MIT Press, 1999.

[23] M. Osborne, Using maximum entropy for sentence extraction, *Proc. of ACL Workshop on Automatic Summarization*, Philadelphia, Pennsylavania, USA, 2002.

[24] M. L. Nguyen, *Statistical Machine Learning Approach to Cross Language Text Summarization*, Ph.D. Thesis, Japan Advance Institute of Science and Technology, 2004.

[25] L. N. Minh, A. Shimazu, H. P. Xuan, B. H. Tu and S. Horiguchi, Sentence extraction with support vector machine ensemble, *Proc. of the 1st World Congress of the International Federation for Systems Research: The New Roles of Systems Sciences for a Knowledge-Based Society*, 2005.

[26] M. Hirohata, Y. Shinnaka, K. Iwano and S. Furui, Sentence extraction-based presentation summarization techniques and evaluation metrics, *ICASSP*, pp.1065-1068, 2005.

[27] M. Andrews and G. Vigliocco, The hidden Markov topic model: A probabilistic model of semantic representation, *Topics in Cognitive Science*, vol.2, pp.101-113, 2010.

[28] N. J. Nilson, *Introduction to Machine Learning*, Draft Book.

[29] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation computational linguistics (ACL), *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, pp.311-318, 2002.

[30] T. Hirao, H. Isozaki, E. Maeda and Y. Matsumoto, Extracting important sentences with support vector machines, *Proc. of COLING*, pp.342-348, 2002.

[31] L. H. Thanh, T. H. Quyet and M. L. Chi, A primary study on summarization of documents in Vietnamese, *Proc. of the 1st World Congress of the International Federation for Systems Research: The New Roles of Systems Sciences for a Knowledge-Based Society*, 2005.

[32] T. M. Mitchell, *Machine Learning*, Mc GrawHill, http://www.cs.cmu.edu/∼tom/mlbook.html.

[33] V. Qazvinian, L. S. Hassanabadi and R. Halavati, Summarising text with a genetic algorithm-based sentence extraction, *Int. J. Knowledge Management Studies*, vol.2, no.4, pp.426-444, 2008.

[34] X. Zhu, *Semi-Supervised Learning with Graphs*, Ph.D. Thesis, 2005.

[35] Y. Fujii, N. Kitaoka and S. Nakagawa, Automatic extraction of cue phrases for important sentences in lecture speech and automatic lecture speech summarization, *INTERSPEECH*, Antwerp, Belgium, pp.2801-2804, 2007.