COMBINATION OF STATISTICAL AND LANGUAGE PROCESSING METHODS IN NEWS SUMMARIZATION:

A CASE STUDY FOR VIETNAMESE NEWS

Vo Thanh Hung, Phan Thi Tuoi, Quan Thanh Tho

Ho Chi Minh City University of Technology,

Vietnam

{vthung, tuoi, qttho}@cse.hcmut.edu.vn

Abstract: Automatic summarization refers to a technique of reducing of a text document by a computer program to create a summary that retains the most important points of the original document. Among various kinds of textual documents, news plays an important role in our life, which is always updated sequentially every day in large scale. A technique for automatic creation of new summarization would help readers to quickly capture significant information from vast amount of electronic newspapers available on the Internet nowadays. In this paper, we introduce an approach to handle this issue on Vietnamese news. By combining both statistical and linguistic methods, we are able to produce meaningful summarization of real articles collected from major newspaper channels in Vietnam.

Keywords: automatic news summarization, natural language processing, statistical methods, ontology, information retrieval.

1 INTRODUCTION

Automatic summarization is reducing a text document by a computer program to create summary capturing main meaning of the original document. In the 21st century, the electronic news plays a very important role in daily life of one's communication. With the advancement of Internet, electronic news is available in vast amount and always updated sequentially every day. Thus, a technique of automatic news summarization should be highly useful for common readers.

Thus, there is much research focusing on this domain. In general, there are two major approaches for news summarization, including single-document and multi-document techniques. Single-document summarization work on single document and get summary on it when multi-document summarization is an automatic procedure aiming at extraction of information from multiple texts written about the same topic. Each of approaches has different problems needed to solve. In this paper, we only focus on the problem for single-document using extraction-based method for handling news.

To summarize a document, typically one can use natural language processing or statistic method. TF-IDF is a basic method to solve this problem, but it is simple and did not achieve high accuracy. The graph method is another notable approach, in which every sentence is represented as a node in a graph [1], [2]. Two nodes corresponding to two semantically similar sentences will be linked by a weighted edge in the graph. After calculating all the edge weighs of the graph, a cut-up value is used to keep only edges with high weight values, based on which the final summarization is produced. In this approach, the means to calculate similar value of two sentences is very important. Thus, we will focus on this point in the next part of this paper.

Another approach is machine learning [3], which is based on some sample data to perform the training process. Then, the similarities of the sentences in the new document are calculated accordingly. After training, for new given documents, the program will perform clustering

technique to choose the most similar sentences to summarize.

Some other remarkable approaches involving statistics include neural networks [4] and fuzzy logic [5]. With the statistical approaches, the semantics of document is generally not taken into account, so this approach is independent of the language. However, the accuracy obtained by this approach is not generally high.

Some other methods use natural language to deal with this problem. WordNet was widely used to solve this problem in English [6], [7]. In this case, to find the connection between two words, WordNet provides a value to calculate. However, this approach suffered from high complexity cost which makes them run considerably slow.

In this paper, we introduce an approach for automatic summarization of Vietnamese news, which combines natural language processing and statistical methods. The rest of this paper is organized as follows. Section II presents our approach. Section III remarks some issues of implementation. Section IV discusses the experiment and the result. Finally, section V is the conclusion of the paper.

2 OUR COMBINATION APPROACH

In this part, we discuss our approach on major following steps. First, we develop ontology to store significant features of the Vietnamese language. After that, two separated statistical methods of sentences rating are presented. Finally, we describe how we combine the two methods.

2.1 Building A Simple Ontology

Ontology is a schema capturing concepts of real world in an organized and machine-readable manner. In a text document, we assume that the meaning of a paper or paragraph can be inferred from linking the semantic of its words. However, to build complete structure which captures semantic connection of all words occurring in a document is always of unacceptably high complexity cost. Thus, we only take into account the semantic connection of every consecutive pair of words in document.

In this paper, we build a simple ontology to present connected meaning of pairs of Vietnamese words. It is produced from 625 article news about economics; resulting in around 4.2 million pairs. A pair of two words is selected from (i) one sentence; (ii) two consecutive sentences; or (iii) two sentences with a sentence between them. After that, some filters are used to get only pairs whose two words are noun, verb, and adjective. The meanings of pairs of words are reflected as a connection to each other, which is associated with a weight value ranging from 0-1. The weight value is inferred based the occurrence frequency of its two words. Generally, the higher frequency will render higher weight value.

All information about pairs and values are stored in the simple ontology. The weight value for each pair will be calculated as presented in (1) as follows.

$$WV = nP / maxW \quad (1)$$

Where:

- *nP*: number of occurrence for each pair
- *maxW*: the maximal occurrence of a word in the document

```
giá/n   đồng/nu 446 337 378
năm/n   kinh_tế/n   451 324 292
đồng/nu giá/n   446 321 301
giá/n   usd/nu  377 293 272
doanh_nghiệp/n  nhà_nước/n  760 281 312
nhà_nước/n  doanh_nghiệp/n  760 263 350
```

Figure 1. Simple ontology

Figure 1 presents an exemplar part of our ontology. One pair of words is given in one line in the following format. First, two words of the

pair are given with proper tags, followed by three numbers indicating the co-occurrence frequency of them in same sentence, two consecutive sentences and two sentences separated by one sentence in between.

This ontology is used in both methods below.

2.2 Method A: Word Connection

On the graph method [1], [2], the formula calculating the similar of two sentences is very important. Some methods analyze and compare the words (as TF-IDF), but they do not consider semantics of words in the sentences. In our approach, we calculate the similarity of two sentences, based on the semantics of words.

$$word - similarity(X, Y) = \sum_{w_1 \in X} \sum_{w_2 \in Y} similaritywords(w_1, w_2) \quad (2)$$

Where:

- X, Y are sentences
- w_1 is a word in sentence X
- w_2 is a word in sentence Y

We propose the similar of two sentences as (2).

In (2) similarity value of two sentences is total of similarity values of all pairs of words (w_1, w_2) of the two sentences X and Y respectively. The similarity value calculated of one pair (*word-similarity* (w_1, w_2)) is based on the simple ontology presented in Section II.1.

Table 1. Similar values of sentences

	S_0	S_1	S_2	S_3	S_4
S_0	X	**0.4**	**1.6**	**0.6**	**0.4**
S_1	0.4	X	**0.7**	**0.2**	**1.5**
S_2	1.6	0.7	X	**0.7**	**1.2**
S_3	0.6	0.2	0.7	X	**0.5**
S_4	0.4	1.5	1.2	0.5	X

Eventually, we obtain the graph *SG*, in which one node presents to one sentence, and each pair of nodes is linked by a weighted edge reflecting their similarity. Next step, we use a threshold to filter from SG the edges with low weight values.

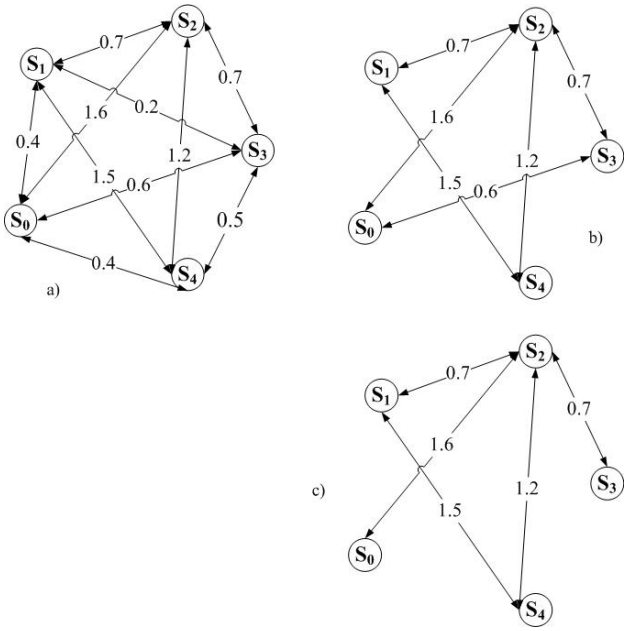

Figure 2. Graphs of five sentences before and after using the cut-up values

Table 2. The number of obtained node for each sentence

Sentences	S_0	S_1	S_2	S_3	S_4
Number of node	1	2	4	1	2

For example, in Figure 2a, we have five sentences $S = \{S_0, S_1, S_2, S_3, S_4\}$ with similarity values given as Table 1. Suppose that the threshold value is 5.5, we have the new graph as Figure 2b, and if the cut-up value is 6.5, the new graph is as shown in Figure 2c. Assume that the graph in Figure 2c is further used, the number of nodes obtained for each sentence is given in Table 2.

In this method, we assume the sentence is more important if they produce more number of connected nodes.

2.3 Method B: Noun-Phrases Connection

We also implement another method to infer the importance of a sentence based on its noun-phrases. In this method, the sentences are firstly processed to get the proper tags, e.g. noun, verb, adjective, etc. On this work, we focus only on noun, verb and adjective to form noun-phrases, because it was agreed that in 80% case, the meaning of a sentence is carried out by their noun-phrases.

In the next step, a *dependency grammar* [8] is used to parse the tag-associated words. In the parsing process, pairs of noun-phrase are determined based on verbs. We assume that a noun-phrase may have one or more nouns and other words of other kinds, e.g. adjective, article, etc. However, in this step, only nouns are stored for later used. In all pairs, subjects (agent) and objects are acquired. In some case, if the objects are not available, then only agent is stored.

Later, we represent pair as *P(agent, object)*; if object is not available then we have *P(agent, null)*. And, P_{agent}, P_{object} (or P_a, P_o) are two component of *P*(*agent*, *object*). Note that, in one sentence, there may be more than one pair like that.

Next, a *connected meaning graph* (CMG) is built. We use *CMV* and *CME* to denote the sets of nodes and arcs of the *CMG* respectively. In this graph, node is one pair with agent and/or object role above *P(agent, object)*. For 2 nodes $P_1(P_{1agent}, P_{1object})$ and $P_2(P_{2agent}, P_{2object})$, an edge is created $(P_1, P_2) \in CME$ that denotes an edge from P_1 to P_2, if and only if they have connected meaning between $P_{1object}$ and P_{2agent} or P_{1agent} and P_{2agent}. The simple ontology is reused for this step.

Table 3. Relation of sentences and noun-phrase pairs

Sentence	Noun-phrases
S_0	P_0, P_2, P_5
S_1	P_1
S_2	P_3, P_5
S_3	P_4
S_4	P_6

For example, we have five sentences $S = \{S_0, S_1, S_2, S_3, S_4\}$ and after parsing we acquire 7 pairs of noun-phrase $CMV = \{P_0, P_1, P_2, P_3, P_4, P_5, P_6\}$ in which relation of sentences and noun-phrase pairs is given as Table 3. In this case, sentence S_0 and S_2 have more than one noun-phrase pairs; another sentences have only one noun-phrase pair.

Table 4. Relation among noun-phrase pairs

	P_0	P_1	P_2	P_3	P_4	P_5	P_6
P_0	X		1	1		1	
P_1		X			1	1	
P_2			X		1		1
P_3	1			X		1	
P_4		1			X		
P_5						X	
P_6		1					X

After calculating connected meaning, we suppose that we have relation as presented in

Table 4. So, we have the final CMG as presented in Figure 3.

After the graph is built, all nodes are rated. In this method, all nodes that can be visited from node $V \in CMV$ will be calculated and contributed to the rating for node V. At the end of this step, all nodes $V(agent, object) \in CMV$ have real values that indicate the relative important noun-phrase pair. For example, rating for each noun-phrase pair for CMG in Figure 3 is represented on Table 5.

Table 5. Rating for each noun-phrase pair

Noun-phrase	P_0	P_1	P_2	P_3	P_4	P_5	P_6
Rating	7	3	5	7	3	1	3

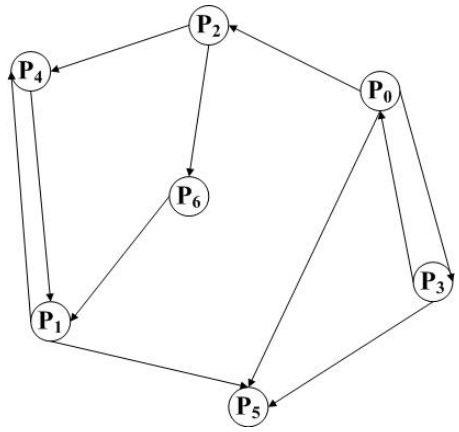

Figure 3. Connected meaning graph of noun-phrase

After all, one sentence may have one or more pairs (node in *CMG*), and then the rating of sentences is the sum of the rating of all nodes in graph. Rating for each sentence is given as Table 6 below.

Table 6. Rating for each sentence

Sentences	S_0	S_1	S_2	S_3	S_4
Number of node	13	3	8	3	3

In this example, S_0 has the highest rating.

2.4 Combination Of Two Methods To Solve Summarization Problem

To combine two methods above, we firstly normalize the rating for each sentence in both of them to the range of [0, 1]. Then, the rating for each sentence is the sum of those of it components, assuming that coefficients are of (0.5, 0.5). After all, all sentences are sorted in descending order in terms of their rating.

After normalizing the data in Table 2 and Table 6, we have two vectors whose elements are in the range of [0, 1] as follows.

$v_1 = \{0.25, 0.5, 1, 0.25, 0.5\}$

$v_2 = \{1, 0.23, 0.62, 0.23, 0.23\}$

We assume that coefficient for v_1, v_2 are 0.5, 0.5 then $v = v_1 * 0.5 + v_2 * 0.5$.

From v_1 and v_2, we have:

$v = \{0.625, 0.365, 0.808, 0.24, 0.365\}$

Then, order of sentences will be obtained as

$\{S_2, S_0, S_1, S_4, S_3\}$

For summarization, we only choose a certain number from the sorted sentences. All sentences in summarization are refined before outputting.

3 SOME ISSUES OF IMPLEMENTATION

3.1 Support Tools And Libraries

JVnTextPro [9] is used to part-of-speech (POS) tagging for Vietnamese. It is a Java open source tool, which is based on Conditional Random Fields (CRFs) and Maximum Entropy.

Another tool is GATE [10]. It is a powerful platform in natural language processing which provides a lot of tools to work with language such as tagging, parsing, etc. In this paper, it is used to parse sentences and to create the pairs of words.

3.2 Algorithms And Implementation

The structure of the system is given in Figure 4. Firstly, news articles enter to preprocessing step. In this step, the stop words are removed. The tags of words are specified by JVnTextPro. Then, a list of tagged sentences is outputted and ready to next step. The following step is the summary processing.

On processing step, we use two methods together. In the left, after building graph, *method A*, referred as *Graph Building CG*, is used, whose output is counting of sentences. *Method B*, referred as *Graph Building CG*, builds graph and gets rating for each sentence.

After that, results of two methods are combined to get a rating for each sentence and get sentences to summary.

Output refined step will refine the sentences obtained. Eventually, the final output is given.

The algorithm to solve summarization problem is presented as follows.

```
1. Load paper to variable "paper"
2. sens = preprocessing(paper)
3. sum = summarization(sens)
4. out = refine(sum)
5. Write output (out)
```

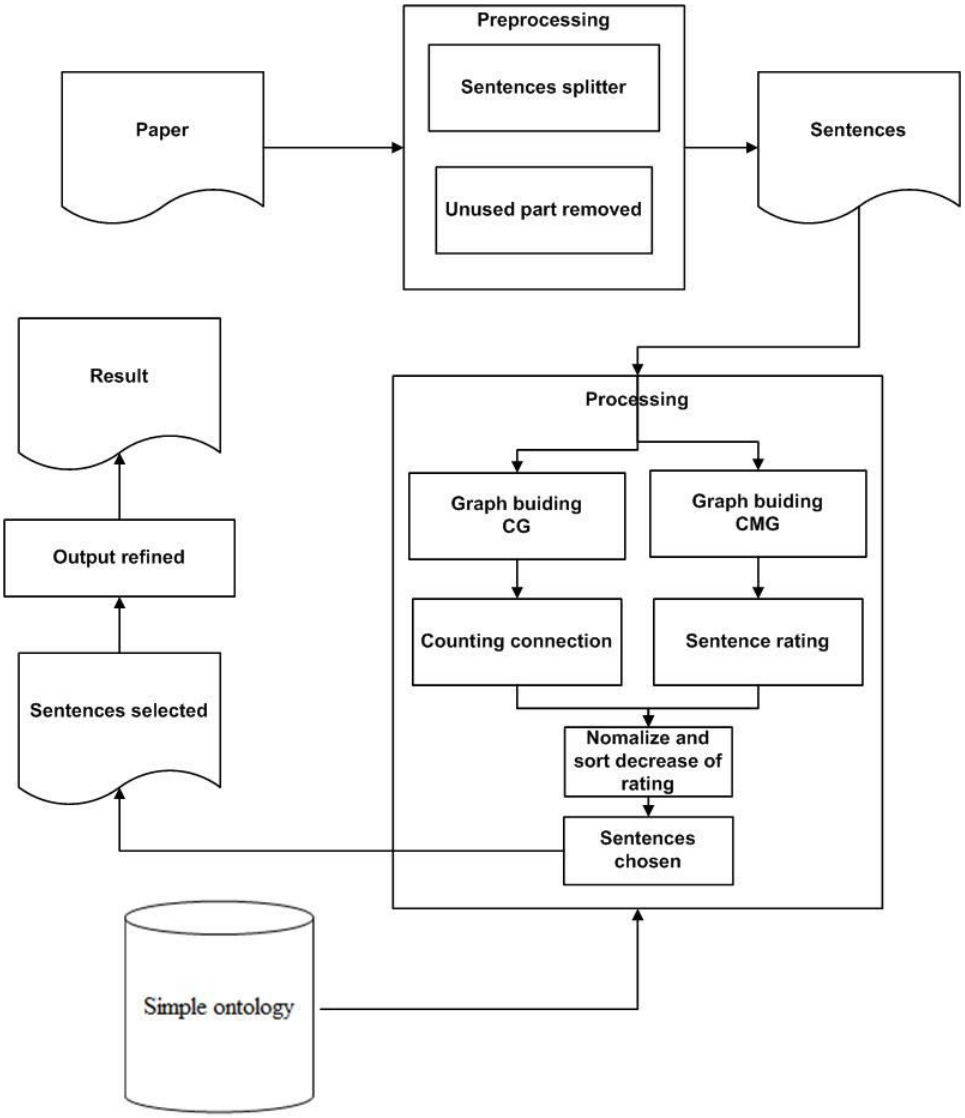

Figure 4. System structure

Preprocessing algorithm:

```
1. Get all separated sentence
2. For each sentence
      a. Remove stop word
3. Return sentences
```

Summarization algorithm:

```
1. outA = methodA(sens)
2. outB = methodB(sens)
3. normalize(outA)
4. normalize(outB)
5. tmp = combine(outA, outB)
6. sort decrease order
7. get some most sentence and refine
   to output
```

In refine step, the sentences are reordered, some words a replaced, and unnecessary words are removed. At the end, a set of sentences is given as the summary of the paper.

4 EXPERIMENT AND RESULT

4.1 Data Test

The data are used to test the program including 105 papers about economic from news online papers such as Tuoi Tre [11], VnExpress [12] and VnEconomy [13].

Table 7. Number of paper from sources

Source	Number of paper
Tuoi Tre	13
Thanh Nien	9
VnEconomy	83

Table 7 presents number of paper for each source. We can see that, most of paper is from VnEconomy news paper; because our domain is about economy. Some papers are very short and some other is of longer length. In the average, each paper has 26.78 sentences. The largest number of sentences in one paper is 390 and the smallest is 3.

4.2 A Case Study

In this part of paper, we present a case study of our work.

A paper is given below whose sentence IDs are denoted as $S = \{S_0, S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9\}$:

S_0 – "*Hãng đánh giá tín nhiệm Standard & Poor's Ratings Services (S&P) cho rằng, áp lực thanh khoản sau vụ bắt giữ "bầu" Kiên chỉ giới hạn ở ngân hàng ACB*".

***(English quick translation:** Standard & Poor's Ratings Services (S&P) stated that the financial pressure from the arresting of Mr.Kien was only limited in ACB bank.)*

S_1 – "*Cũng theo S&P, khó có khả năng xảy ra tình huống tương tự tại các ngân hàng khác của Việt Nam*".

***(English quick translation:** Also from S&P, the chance of similar events in other Vietnam bank was rather low.)*

S_2 – "*Trong một tuyên bố phát đi vào đi vào ngày hôm nay (29/8), S&P nhận định, việc Ngân hàng Nhà nước đảm bảo hỗ trợ thanh khoản cho Ngân hàng Thương mại Cổ phần Á Châu (ACB) sau vụ bắt giữ ông Nguyễn Đức Kiên, nguyên Phó chủ tịch ACB, thường gọi là "bầu" Kiên, và ông Lý Xuân Hải, nguyên Tổng giám đốc ACB, có thể giảm thiểu nguy cơ lan rộng rủi ro trong hệ thống ngân hàng Việt Nam"*.

***(English quick translation:** From an announcement made on today, S&P stated that the fact that the State Bank ensured the payment capability of ACB after the Mr.Kien arresting will reduce risk of Vietnam bank system .)*

S_3 – "*Cũng trong tuyên bố này, S&P đánh giá cao những chuyển biến vĩ mô tích cực của kinh tế Việt Nam*".

***(English quick translation:** Also on that statement, S&P praised the movement on the top of Vietnam economic structure.)*

S_4 - "*Sau mấy năm tăng trưởng tín dụng ở mức cao, những mất cân đối trong nền kinh tế đã*

giảm bớt nhờ chính sách bình ổn của Chính phủ từ năm 2011".

***(English quick translation:** After few years of fast credit growing, the unbalance in economic was reduced due to some policies made by the Government since 2011.)*

S_5 – *"Sự giảm tốc cần thiết trong tăng trưởng vốn vay, sự lập lại ổn định đối với giá cả của các tài sản, và sự xuống thang của lạm phát đã đạt được nhờ những chính sách này", S&P viết".*

***(English quick translation:** The necessary reduction of investment loan rate, the re-establishment of price and property and the downgrading of inflation had been achieved thanks to those policies, said S&P)*

S_6 – *"Hãng đánh giá tín nhiệm này cũng ghi nhận rằng, gần đây, các nhà hoạch định chính sách của Việt Nam đã bắt đầu thể hiện ý định giải quyết những vấn đề dài hạn trong hệ thống ngân hang".*

***(English quick translation:** This firm also noted that recently the policy maker of Vietnam began showing their determination in solving long-term problems of banking system.)*

S_7 – *"Điều này được thể hiện qua "các sáng kiến hợp nhất và tăng cường sức mạnh cho ngành ngân hàng, công nhận rõ ràng hơn những vấn đề về chất lượng tài sản, và sự tăng cường giám sát", tuyên bố của S&P viết".*

***(English quick translation:** It was reflected by the "ideas of unification and strengthening of the banks, clear recognition of issues on property and increasing monitoring", said S&P.)*

S_8 – *"S&P cho rằng, việc Chính phủ Việt Nam nới lỏng chính sách tiền tệ thời gian qua là phản ứng trước lạm phát giảm, áp lực lãi suất cao đối người vay vốn, và sự giảm tốc tăng trưởng".*

***(English quick translation:** S&P stated that the loose of control from Vietnam government of financial policy is was to react to decreasing inflation rate, high interest rate for the loaner and reduced speed of growing.)*

S_9 – *"Đồng thời, S&P khuyến nghị: "Quá trình phục hồi niềm tin vào hệ thống ngân hàng và chính sách tiền tệ của Việt Nam đang ở giai đoạn đầu và cần có sự quản lý thận trọng".*

***(English quick translation:** At the same time, S&P suggested that: "The process of belief recovery on bank system and financial policy of Vietnam is just of initial stage and should be closely monitored).*

A result of rating for sentences is:

i) By method A:

To use reasonable cut-up, the relation of sentences is given on Table 8. And the result is: *{1, 6, 7, 7, 7, 2, 7, 3, 8, 6}*

Table 8. Relation of sentences

	S_0	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
S_0	X		1							
S_1		X	1	1	1		1		1	1
S_2	1	1	X	1	1		1		1	1
S_3		1	1	X	1		1	1	1	1
S_4		1	1	1	X	1	1		1	1
S_5					1	X			1	
S_6		1	1	1	1		X	1	1	1
S_7				1			1	X	1	
S_8		1	1	1	1	1	1	1	X	1
S_9		1	1	1	1		1		1	X

ii) By method B:

Number of noun-phrase: 35

Number of arc in CMG: 332

Number of the noun-phrase on each sentence is given on Table 9.

Table 9. Number of noun-phrase on each sentence

Sentence	S_0	S_1	S_2	S_3	S_4
Number of noun-phrase	4	0	7	0	2
Sentence	S_5	S_6	S_7	S_8	S_9
Number of noun-phrase	2	7	8	1	4

The rating for sentences is *{82, 0, 110, 0, 36, 21, 146, 178, 19, 76}*

By normalizing and combination of them we have:

v_1 *= {0.13, 0.75, 0.88, 0.88, 0.88, 0.25, 0.88, 0.38, 1, 0.75}*

v_2 *= {0.46, 0, 0.62, 0, 0.2, 0.12, 0.82, 1, 0.11, 0.43}*

v = {0.295, 0.375, 0.75, 0.44, 0.54, 0.185, 0.85, 0.69, 0.555, 0.59}

The order of these sentences is:

{S_6, S_2, S_7, S_9, S_8, S_4, S_3, S_1, S_0, S_5}

Then, if one sentence is used for the summary then we can get *{S_6}*; if two sentences are used, then we have *{S_6, S_2}*; and so on. For example, 2-sentence summarization of the above news is as follows. The boldfaced texts indicate refinements made to make the news more semantically meaningful.

(SUM-VN): *"S&P ghi nhận rằng, gần đây, các nhà hoạch định chính sách của Việt Nam đã bắt đầu thể hiện ý định giải quyết những vấn đề dài hạn trong hệ thống ngân hang. Trong một tuyên bố phát đi vào đi vào ngày hôm nay (29/8), hãng đánh giá tín nhiệm này nhận định, việc Ngân hàng Nhà nước đảm bảo hỗ trợ thanh khoản cho Ngân hàng Thương mại Cổ phần Á Châu (ACB) sau vụ bắt giữ ông Nguyễn Đức Kiên, nguyên Phó chủ tịch ACB, thường gọi là "bầu" Kiên, và ông Lý Xuân Hải, nguyên Tổng giám đốc ACB, có thể giảm thiểu nguy cơ lan rộng rủi ro trong hệ thống ngân hàng Việt Nam".*

In English:

(SUM-EN): *S&P noted that recently the policy maker of Vietnam began showing their determination in solving long-term problems of banking system. From an announcement made on today, this firm stated that the fact that the State Bank ensured the payment capability of ACB after the Mr.Kien arresting will reduce risk of Vietnam bank system.)*

4.3 Result

After processing of our program for each paper, some sentences are chosen and given as a summary of this paper. In the other way, we invite some persons to read the papers and choose some "important" sentences as summary. Finally, results from program are given and compared to results from the invited persons.

Table 10. Testing of the result for 105 papers

	Number of paper	Rate (%)
Acceptable	74	70.48
Wrong	31	29.52
Total	105	100

Table 10 gives the accurate of the method for over 105 economic papers. To evaluate an accurate of our method, we rate it by the precision value as

Precision = 74 / (74 + 31) = 70.48%

5 CONCLUSION

In this paper, we propose new method that includes natural language processing and statistic method to solve summarization problem. Our approach gives a result with acceptable quality. In addition, our proposed approach is language-independent thus it can be easily adapted for other languages.

References

[1] Dipanjan Das and André F.T. Martins, "A Survey on Automatic Text Summarization," Literature Survey for the Language and Statistics II course at CMU.,

2007.

[2] Gunes Erkan and Dragomir R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," *Journal of Articial Intelligence Research 22*, pp. 457-479, 2004.

[3] Wesley T Chuang and Jihoon Yang, "Extracting sentence segments for text summarization: a machine learning approach," in *Proceeding SIGIR '00 Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, USA, pp. 152 – 159.

[4] Khosrow Kaikhah, "Automatic text summarization with neural networks," in *Second IEEE International Conference On Intelligent Systems*, 2004, pp. 40-44.

[5] Mandar Mitra, Amit Singhal, and Chris Buckley, "Automatic Text Summarization by Paragraph Extraction," in *In Proceedings of the Workshop on Intelligent Scalable Summarization at the ACL/EACL Conference*, Madrid, Spain, 1997, pp. 39–46.

[6] Regina Barzilay and Michael Elhadad, "Using Lexical Chains for Text Summarization," in *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, 1997, pp. 10-17.

[7] Rui Pedro Chaves, "WordNet and Automated Text Summarization," in *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, NLPRS, ,* Tokyo, Japan, 2001.

[8] Joakim Nivre, "Dependency Grammar and Dependency Parsing," Växjö University: School of Mathematics and Systems, 2005.

[9] Cam-Tu Nguyen, Xuan-Hieu Phan, and Thu-Trang Nguyen. (2010) JVnTextPro. [Online]. http://jvntextpro.sourceforge.net/

[10] The University of Sheffield. (1995) GATE. [Online]. http://gate.ac.uk/

[11] Tuoi Tre Online. [Online]. http://tuoitre.vn/

[12] VnExpress. (1997) [Online]. http://vnexpress.net

[13] VnEconomy. [Online]. http://vneconomy.vn/

[14] R. M. Aliguliyev, "Automatic Document Summarization by Sentence Extraction," *Computational Technologies*, vol. 12, no. 5, pp. 5-15, 2007.

[15] Berlin Chen, Yao-Ming Yeh, Yao-Min Huang, and Yi-Ting, "Chinese Spoken Document Summarization Using Probabilistic Latent Topical Information," in *ICASSP 2006*, 2006, pp. 969-972.

[16] (2003) Thanh Nien Online. [Online]. http://www.thanhnien.com.vn/pages/default.aspx

[17] Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258-268, 2010.