A Statistical Approach for Automatic Thai Text Summarization using Fractal Summarization

Rattasit Sukhahuta[1], Chayapol Sanpiban[1], Prasert Luekhong[2]

[1]Department of Computer Science, Faculty of Science, Chiang Mai University, Chiangmai, Thailand

rattasit.s@cmu.ac.th, kms4406015@hotmail.com

[2]College of Integrated Science and Technology, Rajamangala University of Technology Lanna, Chiang Mai, Thailand
prasert@rmutl.ac.th

Abstract

Automatic text summarization is a technique of compressing the original document into a summarized version by extracting the important concepts in term of phrases or sentences. This research focuses on a development of statistical methods using fractal summarization model by enhancing its ability to deal with the ambiguity of Thai documents. We proposed that syntactical information can be incorporated with the weight computation to enhance the accuracy and to resolve the ambiguity of the semantic meanings. Winnow algorithm is exploited to perform a sentence breaking based on a statistical model of determination between sentence breaking space (SBS) and non-sentence breaking space (NSBS) according to context around the spaces. Each sentence is computed with thematic weight given that the document in a hierarchical structure and based on the concept terms within the sentences; therefore, sentences with high thematic weight value will be extracted.

Keywords: Thai Document Summarization, Fractal Summarization, Winnow Algorithm

I. Introduction

Due to the exploration of the number of documents being stored electronically in nowadays, quick information access becomes necessary. Text summarization is a process of compressing the original text into a shortened version by capturing the important context and extracting or abstracting by paraphrasing the content into new version without losing the original content. To be able to summarize information in human-like requires an in-depth NLP analysis with deep knowledge and context understanding. In general, the summarization system is developed to extract the keywords, phrases and sentences from the document. Another approach is to deploy natural language processing techniques in the information extraction where large amount of resources and computational linguistic help to enhance the level of language understanding resulting more accurate content to be summarized. In recent surveys showed that there are few researches done on Thai document summarization. Unlike in English, there is no indicator for boundaries between words and sentences in Thai writing. Nevertheless, the paragraph can still be detected from the new line or empty spaces between the contexts. Jaruskulchai, C., and Kruengkrai, C. [3] proposed a practical approach for extracting the most relevant paragraphs from the original document to form a summary of Thai text. The idea of their approach is to exploit both local and global properties of paragraphs. The local property can be considered as clusters of significant words within each paragraph, while the global property can be thought of as relations of all paragraphs in a document. These two properties are combined for ranking and extracting summaries. The common statistical technique is based on term frequency and co-occurrence of the terms appears in the sentences to determine the importance of the terms. Suwanno, N., Suzuki, Y., and Yamazaki, H. [6] proposed a method for Thai text summarization by paragraph extraction based on the extracted Thai compound nouns and term weighting method using term frequency (TF) and inverse document frequency (IDF). Morphological analysis has been applied to determine the Thai compound nouns and all paragraphs that are ranked by the summation of term weighting score. The cosine similarity between each paragraph is calculated in order to select the important paragraphs within the document. Thangthai, A., and Jaruskulchai, C. [7] proposed the method for Thai document summarization by using an N-Gram technique for word segmentation and the summarization process that is not depending on various types of document. This methodology is based on a mathematical model that does not depend on dictionary, learning technique and training data. The important paragraph is therefore extracted using the Singular Value Decomposition algorithm.

In this paper, we propose the method of Thai document summarization using fractal summarization and Winnow algorithm. The remainder of this paper is organized as follows. Thai text analysis technique is explained in Section II. Section III describes Thai document summarization technique using fractal summarization. Section IV presents our experimental design and discussion and finally Section V presents the conclusion of this paper.

II. Thai Text Analysis

In Thai writing, there is no specific indicator for sentence boundary. The spaces are commonly used at the end of each context. They can be interpreted as the end of a sentence or phrases or clause break in a sentence. They can also be used as a separation between numerals and foreign languages. Nevertheless, incorrect spaces in the sentence can yield different meaning and sometimes misinterpretation. Figure 1 illustrates the proposed model for Thai document summarization. It basically consists of four main tasks: term weight determination, word segmentation and part-of-speech tagging, sentence boundary detection, and document summarization. We also take the syntactic categories into consideration in our approach since they help in determining the significant level of Thai words. In this research, we used 47 syntactic categories defined in the Orchid project [5] to be considered during the weight assigning process.

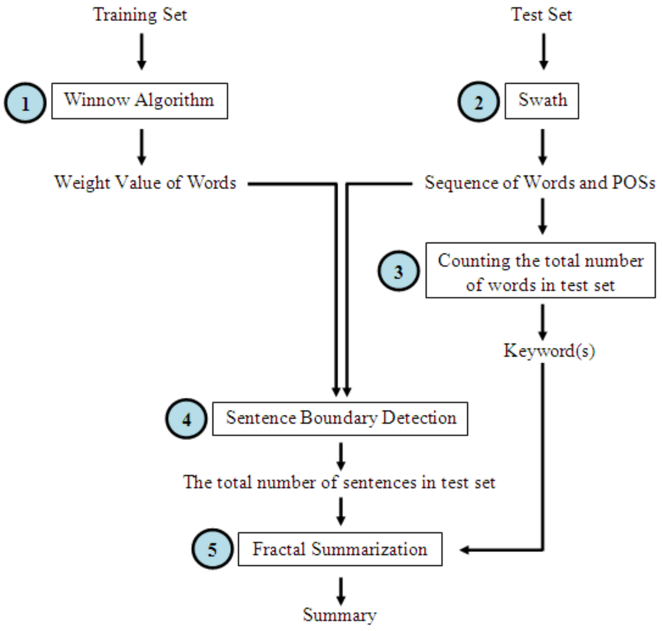
Training Set
Test Set
1
Winnow Algorithm
2
Swath
Weight Value of Words
Sequence of Words and POSs
3
Counting the total number of words in test set
Keyword(s)
4
Sentence Boundary Detection
The total number of sentences in test set
5
Fractal Summarization
Summary

Figure 1 Overview of the proposed method

A. *Thai sentence boundary detection*

During the preprocessing steps, the Thai document first needs to be segmented and tagged with Orchid part-of-speech tags. Then each paragraph in the document is broken up into sentences using Winnow algorithm [2]. Winnow is a neuron-like network where several nodes are connected to a target node as shown in Figure 2. Each node is called specialist that looks at a particular value of an attribute of the target concept and vote for a value concept based on its specialty; i.e. a value of the attribute it examines. The Winnow algorithm updates the weight of each specialist based on the vote of that specialist. Initially, the weight of each specialist is set to 1. In the case that Winnow algorithm predicts incorrectly, the weight of the specialist that predicts incorrectly is halved and the weight of the specialist that predicts correctly is multiplied by 3/2. In case that the Winnow algorithm predicts correctly, the weight of the specialist that predicts incorrectly is 1/2. [1, 2]

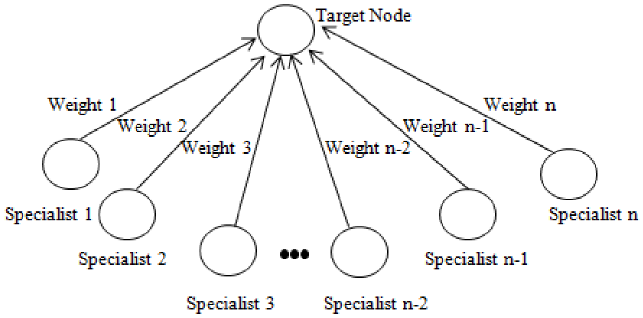

Figure 2 Example of Winnow network

B. *Finding weight value of each word in training set*

In this technique the spaces within the document will be classified into two types: non-sentence breaking space (NSBS) or sentence breaking space (SBS). Paragraph can be separated into sentences by sentence-breaking space in between. This algorithm considers five consequent words before and after a space each time contained in ORCHID corpus [5] for determining the space as whether a true sentence breaking space or not. The weight value of each word in the Orchid training set. There are two types of data set. First is sentence breaking space data set which contains 10 words around sentence breaking space and second is non-sentence breaking space data set which contains 10 words around non-sentence breaking space as shown in the following sentences:

#1
ใน/*RPRE*ปลายปี/*NCMN<space>/PUNC2529/NCNM*
<space>/PUNC ***;Non Sentence Breaking Space***
กระทรวงวิทยาศาสตร์/*NPRP*
<space>/PUNC; Non Sentence Breaking Space
โดย/*RPRE*มติ/*NCMN*คณะ/*NCMN*รัฐมนตรี/*NCMN*
ได้/*XVAM*จัดตั้ง/*VACT*ศูนย์/*NPRP*ขึ้น/*XVAE*
// ; indication for Sentence Breaking Space
#2
เพื่อ/*JSBR*พัฒนา/*VACT*เทคโนโลยี/*NCMN*
อิเล็กทรอนิกส์/*NCMN*ใน/*RPRE*ประเทศ/*NCMN*

In the example of training sentences, *<space>/PUNC* appearing within the sentences will be treated as a non-breaking space where the sentence ending marker '//' will be used to indicate a breaking spaces. In this research, there are 7,000 data sets generated from ORCHID corpus [5] for training set and 4,500 data sets consisted of 10 words around sentence breaking space and 2,500 data sets consist of 10 words around non-sentence breaking space. The training set is trained by Winnow to find weight value of each word. The weight of the word which appeared around sentence breaking space is initialized to 1.The weight of the word which appeared around non-sentence breaking space is initialized to -1. Winnow algorithm updates the weight of each word in each data set based on the prediction algorithm. If the Winnow algorithm predicts incorrectly, the weight of the word that predicts incorrectly is halved and the weight of the word that predicts correctly is multiplied by 3/2. If the Winnow algorithm predicts correctly, the weight of the word that predicts incorrectly is 1/2.

C. Finding sequence and part of speech of each word in test set

The Annual Report 2007 of NECTEC [5] is manually organized into a hierarchical structure as Title, Section and Sub-Section. The test set is segmented and tagged with part-of-speech by Swath [1,4] to find the sequence and part of speech of each word in test set. This process is considering the words that have the most occurrences. The number of word occurring in test set for finding keyword(s) of test set according to the assumption that "Keyword is the words that occur in test set more than to 20% of the total number of words in the test set".

D. Finding the boundary of sentences

We used weight value of each word in training set to find the boundary of sentence in test set by considering words around each space to determine that the space whether it is a sentence breaking space or non-sentence breaking space. If the total weight value of words around space is more than or equal to total number of words around the space, this space is determined as a sentence breaking space *(SBS)*. If the total weight value of words around space is less than the total number of words around space, this space is determined as a non-sentence breaking space *(NSBS)*. After that, we count the total number of sentences in the test set as in the following example:

สืบเนื่อง@VSTA มา@XVAE โดย@JSBR ตลอด@ADVN ***<SP1>***
ทั้ง@DDBQ นี้@DDAC ***<SP2>****โดย@FIXแบ่ง@VACTเป็น@VSTA*
กิจกรรม@NCMN ต่างๆ@NPRP ***<SP3>****คือ@VSTAการ@FIXN*
ดำเนิน@VACT โครงการวิจัย@NCMN ร่วม@ADVNกัน@ADVN
<SP4>*การ@FIXN แลกเปลี่ยน@VACT และ@JCRGถ่ายทอด@VACT*
เทคโนโลยี@NCMN ระหว่าง@RPRE กัน@ADVN ***<SP5>****รวม@VACT*
ทั้ง@DDBQ การ@FIXN เข้า@VACTร่วม@ADVN แสดง@VACT
ข้อคิดเห็น@NCMN

Base on the example Thai text that has been segmented and tagged with Orchid part-of-speech. There are five spaces specified with <SP>. The main objective is to identify whether these spaces are a space that breaks up the sentence (sentence boundary indicator) or a non-breaking space. By considering the weight summation of all the words around each space, if the weight is greater than or equal to the number of words around the space, the space will be considered as a sentence breaking space or else it will be treated as a non-sentence breaking space. According to the example, we can decide that <SP1> is the sentence breaking space and the spaces <SP2>-<SP5> are the regular spaces that break the context or phrases.

III. Factal Document Summarization

Thai document summarization using fractal summarization is mainly based on statistical approach to automatically generating summaries of Thai text document. Our approach basically consists of four main tasks: (i) Thai text pre-processing as described in section II. (ii) determination of the term weight based on several features: thematic feature, location feature and heading feature (iii) thematic weight of the sentence is determined, and finally (iv) sentence summary generation. In fractal summarization, the text document will be organized as a hierarchical structure. For instance, the document will be organized into chapter, section, sub-section and so on. Hench, each part contains heading, title, section, subsection and so on. Each level of document will be treated as a *node and child-node* respectively. The source document is partitioned into range-blocks and transformed into fractal tree according to the document structure. Each range-block is represented by a node in the fractal tree illustrating in Figure 3.

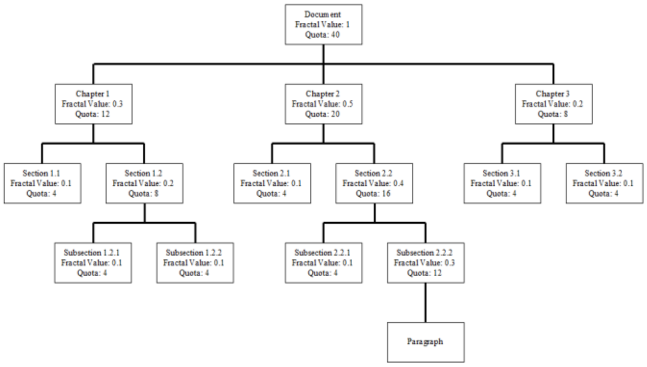

Figure 3 Example of Fractal Tree

The computed weight value and salient features of each sentence are depending on its location within the tree-like structure. The number of sentences of the child-nodes will be propagated to grandchild-nodes according to the fractal value of grandchild-nodes. If the number of sentences of the child-node exceeds the threshold value, the system will process its child-nodes iteratively until the number of sentence is less than the threshold value. The fractal value of each node is calculated as the sum of sentence weights of the sentences under the range-block. The document with deep level structure will eventually have an impact the time of the weight computation. The user may choose the compression ratio to specify the ratio of sentences to be extracted as the summary. The total number of sentence quotas of the summary can be calculated accordingly and it will be propagated to the child-nodes directly proportional to their fractal values.

A. *Thematic Feature*

Thematic feature is a feature that being considered to determine the significant level of the sentences that represent the main context of the document. It also indicates that the sentences contain important keywords or key phrases. In the fractal summarization, the thematic features will be computed as thematic weight using *tf-idf* scores that should be proportional to the term frequency within a range-block, but inversely proportional to the frequency of range-blocks containing the term, that is

$$w_{ir} = tf_{ir} \times \log_2 \left[\frac{N' \times |t_i|}{n'}\right] \qquad (1)$$

Where

tf_{ir} the frequency of term t_i in range-block r

N' the number of range-blocks in the document

n' is the number of range-blocks where term t_i occurs, and $|t_i|$ is the length of the term t_i

The fractal sentence's score based on thematic feature for sentence $k_{(sk)}$ in range-block r is computed as follows:

$$FSS_{T(k,r)} = \sum_{t_i \in s_k} w_{ir} \tag{2}$$

$$RBSS_{T(r)} = \sum_{k \in r} FSS_{T(k,r)} \tag{3}$$

B. *Location Feature*

Location Weight is the weight of the sentence by considering the location where the sentence appears in the document. The location weight of a range-block is $1/p$, where p is the shortest distance of the range-block to the first or last range-block under same parent range-block. The fractal sentence score based on location feature of any sentences in range-block r is computed by:

$$FSS_{L(k,r)} = \frac{1}{\prod_{\substack{y \in path\ from\ k \\ to\ z(excluding\ r)}} \begin{bmatrix} \min\big(d(y, first\ child\ of\ y's\ parent)\big), \\ d(y, last\ child\ of\ y's\ parent)) \end{bmatrix}} \tag{4}$$

$$RBSSL_{(r)} = \frac{1}{\min(d(r,first),d(r,last))} \tag{5}$$

where $first$ is the first range-block in the same level of *r*; *last* is the last range-block in the same level or *r; d*(*r*, *first*) and *d*(*r*, *last*) are distances between range-block *r* and the first range-block, also the distance between range-block *r* and the last range-block, respectively.

C. *Heading Feature*

The sentence scores based on the heading feature of a sentence is dynamic and depends on which document level we are considering in the document. The fractal sentence score is based on the heading feature for sentence $k_{(sk)}$ in the range-block r, $FSSH_{(k,r)}$ is:

$$FSSH_{\mathrm{H}(k,r)} = \sum_{\substack{y \in path \\ from\ root\ to\ r}} \frac{\sum_{t_i \in s_k \cap\ heading(y)} w_{iv}}{\prod_{q \in path\ from\ y\ to\ r} m_q} \tag{6}$$

$$RBSS_{H(r)} = \sum_{k \in r} FSSH_{\mathrm{H}(k,r)} \tag{7}$$

where w_{iy} is the $tf - idf$ score of term i in the range-block y, and m_q is the number of children of range-block q. The $RBSS_{(r)}$ is then computed as the sum of the normalized values of $RBSS_{T(r)}$, $RBSS_{L(r)}$, and $BSSR_{H(r)}$.The individual feature score of a range-block is divided by the maximal feature score of all sibling nodes of the range-block; hence, the feature scores are normalized such that the maximum score of each feature is 1 [8]. The details of the fractal summarization algorithm are performed in following steps:

```
(1) Selects the compression ratio to specify the ratio of the extracted sentences
(2) Selects the threshold value to specify the maximum number of sentences extracted from each node
(3) Calculates the total sentence quota of the summary
(4) Partition the document into range-blocks according to the document structure
(5) Transforms the document into a fractal tree
(6) Set the current node to the root node of the fractal tree and initialize its fractal value to 1
(7) Repeat.
  (7.1) For each child-node under the current
  node, calculate the fractal value of a
    child-node.
```

$$Fv(child\_of\_x) = Fv(x)C\left[\frac{RBSS(child\_of\_x)}{\sum_{y\ \in\ children\ of\ x} RBSS_{(y)}}\right]$$

where *Fv*(*x*) is the fractal value of range-block *x*
RBSS(*y*) is the range-block significance score of range-block *y*
y is any child of *x*
C is a constant,0 < C ≤ 1
D is the fractal dimension,0 <*D* ≤ 1

```
      (7.2) Allocate quota to child-nodes in
      proportion to fractal values
  (7.3) For each child-node
  If the quota is less than the threshold
     select the sentences in the range-block by
     extracting the sentences
  Else
     set the current node to the child-node
(8) Until all the child-nodes under the current node are processed
```

The $RBSS$ of range-block r, $\mathrm{RBSS}_{(r)}$, is computed as a summation of the fractal sentence scores based on the thematic, location, and heading features for all sentences within range-block r. The details of the fractal sentence scores based on thematic weight, location weight, and heading features as described.

IV. Experiment and Discussion

In this experiment, we used NECTEC Annual Report 2007 as a data test set. The data set is divided into 122 topics where each topic contains about 30-50 sentences in average. We evaluated the summary results by first determined the sentence boundary of each input document. Then fractal summarization is applied to generate a summary from the test set by selecting sentences with the top i sentence score in each section. The parameter i is a mean value showing the sentence's quota of that section. For example, if the sentence quota of that section is 5, we will select the sentences in this section with the top five sentences score for generating the summary. The experiment results generated from test set for 4 experiments which have different compression ratio and threshold value shown as follows:

Experiment	Compression Ratio	Threshold
I	20%	3
II	20%	5
III	25%	3
IV	25%	5

To evaluate the implemented system, the correct answers which are the captured sentences will be extracted manually for each document by ten participants who have knowledge about the content of our test corpus. The performance of the summarization system is measured based on precision value that is the ratio of sentences being accepted by the users over the total number of sentences being extracted by the summarization engine.

$$Precision = \frac{\begin{array}{c} the\ number\ of\ sentences\ extracted\ by \\ users\ as\ correct\ results \end{array}}{total\ number\ of\ sentences}$$

Table 1. Summarization the Average precision Evaluation

	Exp.I	Exp.II	Exp.III	Exp.IV
1	73.85	76.19	61.84	62.96
2	76.92	84.13	63.16	64.20
3	78.46	71.43	65.79	62.96
4	73.85	73.02	61.84	61.73
5	83.08	79.37	65.79	66.67
6	81.54	77.78	67.11	66.67
7	70.77	68.25	63.16	59.26
8	64.62	68.25	55.26	55.56
9	76.92	73.02	64.47	65.43
10	70.77	69.84	63.16	58.02
Avg	**75.08**	**74.13**	**63.16**	**62.35**

In Table 1, the experiment results showed that with 25% compression ratio and threshold = 3 yields the best result. In this experiment, the extracted sentences are based on the sentences computed boundary by Winnow. Thus Thai sentences containing context fragment mostly are long sentences with incomplete content information. This information illustrates that the better results do not always come from higher number of compression ratio or threshold. Our experiment results showed the overall performance of the sentence extraction based on an average precision is 75.08%. More interestingly, the proportion of text summarization performance is higher than 60%, which implies that the results extracted automatically by the summarization are close to human choices.

V. Conclusion

With the exploration of the documents being stored and published electronically, there must be a simply and quicker ways to access these information. Information summarization is a technique of extracting specific content and to discover most of the important issues of information and to present the user with a shorter version to the original document. In this paper, we proposed a method that employed the statistical techniques of Winnow to detect the boundary of Thai sentences and using fractal summarization that explores term relationships within the document given that the sentences are organized in a hierarchical structure, where fractal values are computed considering several features including thematic, heading, and location. The proposed method used syntactical information to resolve the ambiguity of Thai document structure. The problem of Thai sentence's structure ambiguity is thereby resolved by determining the sentence boundary and sentence breaking spaces using Winnow algorithm. The fractal summarization has also proven to be a feasible method for improving the performance of Thai text summarization applications.

References

[1] Charoenpornsawat, P. Feature-based Thai Word Segmentation. M.Eng. Thesis, Chulalongkorn University, 1998.

[2] Charoenpornsawat, P., and Sornlertlamvanich, V. Automatic Sentence Break Disambiguation for Thai, *Proceedings of the 19th International Conference on Computer Processing of Oriental Languages*, 2001.

[3] Jaruskulchai, C., and Kruengkrai, C.A practical text summarizer by paragraph extraction for Thai, *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, 2003.

[4] Meknavin, S., Charoenpornsawat, P., and Kijsirikul, B. Feature-based Thai Word Segmentation, *Proceedings of the Natural Language Processing Pacific Rim Symposium*, 1997.

[5] Sornlertlamavanich, V., Charoenporn, T., Isahara, H. ORCHID: Thai Part-Of-Speech Tagged Corpus, Technical Report, Linguistics and Knowledge Science Laboratory, NECTEC, Bangkok, 1997.

[6] Suwanno, N., Suzuki, Y., and Yamazaki, H.Extracting Thai Compound Nouns for Paragraph Extraction in Thai Text, *Proceedings of the IEEE Natural Language Processing and Knowledge Engineering*, 2005.

[7] Thangthai, A., and Jaruskulchai, C.
Text Summarization using Singular Value Decomposition for Thai, *Proceedings of the 7th National Computer Science and Engineering Conference*, 2003.

[8] Yang, C. C., and Wang, F. L., Hierarchical Summarization of Large Documents, *Journal of the American Society for Information Science and Technology*, **59(6)**, 887-902, 2008.