Extracting Thai Compound Nouns for Paragraph Extraction in Thai Text

Nuttida Suwanno Yoshimi Suzuki Haruaki Yamazaki
Department of Computer Science and Media Engineering
Faculty of Engineering, University of Yamanashi
4-3-11 Takeda, Kofu 400-8511

***Abstract*-In this paper, we propose a method for Thai text summarization by paragraph extraction based on the extracted Thai compound nouns and term weighting method in terms of Term Frequency Inverse Document Frequency (TF•IDF). According to the highly frequent and highly productive of Thai compound nouns in Thai text, this property shows that Thai compound nouns play the important role in summarization. The morphological analysis is used to determine Thai compound nouns and all paragraphs are ranked by summation of term weighting score. The cosine similarity between each paragraph is calculated in order to select the important paragraphs among all paragraphs. The result shows that 0.469 F-score for 45% summary of our proposed method yield the most effective approach among all experiments.**

I. INTRODUCTION

Recently, researchers are focusing their interests on the development of summarization systems using innovative approaches such as statistical approaches [1], [2], knowledge-based approaches [3] and the combination of both statistical and knowledge-based approaches. The combination of different summarization features is one of the most important subjects in recent research [4]. In general, summarization systems typically use a single word string as the unit for counting frequency [2], [5]. However, there are tasks attempting to apply NLP tools to extract multi-word phrases automatically with high accuracy and use them as the basic unit in the summarization [6]. The enhancement from such tasks showed that multi-word phrases play a major role in discriminating related phases in certain domains.

In this work, we are interested in the extracting of compound nouns and its application to text summarization. According to the highly frequent and highly productive of Thai compound nouns in Thai text, this property shows that Thai compound nouns may play an important role in the summarization. Our system is implemented by paragraph extraction to summarize Thai text because this is the practical way to extract Thai text spans from Thai documents. We use morphological analysis to determine Thai compound nouns. The important paragraphs are ranked by summation of term weighting scores in terms of TF•IDF and headline score. The cosine similarity between each paragraph is calculated to select the important paragraphs among all paragraphs. We used 50 Thai news articles as a test data. The evaluation was calculated in terms of F-score. The result shows that 0.469 F-score for 45% summary of our proposed method yield the most effective approach among all experiments.

The remainder of the paper is organized as follows: Section two details exploration of related works in Thai text summarization, then Section three illustrates the Thai compound nouns including how to extract it. Section four presents our framework and our approach for extracting important paragraphs in detail. Section five describes our experiment and evaluation. Section six presents the results and discussions and the last section includes the conclusions and future work.

II. RELATED WORK

Unlike English, Thai language has no explicit word and sentence marker. Because of this restriction of word segmentation and difficulty in Thai sentence extraction, there have been very few works on automatic Thai text summarization. In order to avoid solving the full natural language problem and to develop a summarization approach, researchers have applied statistical techniques to the summarization task for Thai. Ref. [7] had proposed a practical approach used Term Length Term Frequency (TL•TF), term weighting technique, to score words in the document as a local clustering score, and using cosine similarity between two paragraphs as a global connectivity score. However the disadvantage of this method is that best paragraphs ranked for summary selection often tend to be even longer than the average paragraph in the document. Meanwhile other researchers had proposed an approach to extract important paragraphs by using the singular value decomposition algorithm. They claimed that this method obtained a higher efficiency than it did for the previous research in the case of newspaper articles, but obtained lower efficiency in the case of essay [8]. Ref. [9] had reported a sentence extraction approach for Thai text summarization through mobile devices using TF•IDF to determine significant words. However this task has not been given a theoretical information retrieval evaluation.

Our approach was difference from above related works as we focused on an extracting of compound nouns by using the natural language processing technique called Part Of Speech (POS) tagging to extract the compound nouns. Moreover, we use term weighting technique in terms of TF•IDF to score all words excluding stop words in the paragraph. The feature of headline to calculate the paragraph scores was also used. The similarity between paragraphs is measured according to cosine similarity value. The results are used to select a

0-7803-9361-9/05/$20.00 ©2005 IEEE

representative paragraph among similar paragraphs. In order to determine the appropriate threshold for cosine similarity value we also experiment using the various threshold values and various summary ratios.

III. Thai Compound Nouns

3.1. *Basic concept of compound nouns*

Similar to Chinese or Japanese, Thai text has no word boundaries. Besides, there are no explicit sentences boundaries within the document. However, there are indentations and blank lines used in the Thai writing system to indicate paragraphs. Thai words are multi-syllabic words. When stringing together, it could form a new word. Thai new words can be combined to form compound nouns, and they are invented almost daily [10]. Compound nouns are defined to take the form of a concatenated noun pair, the elements of which we will refer as N_1 and N_2 in linear order of occurrence. Examples of Thai compound nouns are "เชื้อเพลิง•ชีวภาพ" (fuel•bio = bio fuel in English), "กุ้ง•ไทย" (shrimp•Thai = Thai shrimp) and "ราคา•สินค้า" (price•product = product price)[1]. Not only are Thai compound nouns highly frequent and highly productive in Thai, but they can also be a significant word in the document. Fig. 1 shows the appearance of compound nouns in the corpus.

In Fig. 1, let ⬭ be the compound nouns, ○ and ● be the combinations of noun in compound nouns. According to Luhn's assumption, ○ and ● frequently appear in many articles more than word ⬭ which frequently appears in a particular article. Therefore, the significant value of ⬭ in the article is larger than those of ○ and ●. Due to compound nouns is the concatenation of noun pair, we also show the productive of compound nouns among common noun and proper noun by TF•IDF value of compound nouns, common noun and proper noun throughout the corpus.

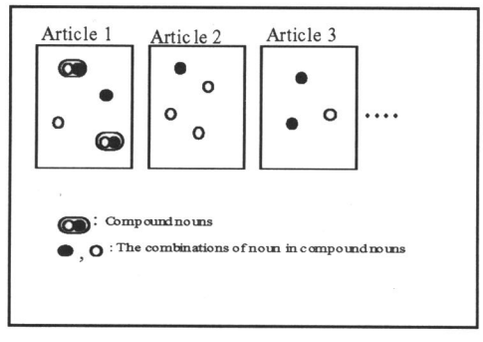

Fig. 1. The appearance of compound nouns in the corpus.

[1]With all Thai compound nouns examples, we segment the compound into its compound nouns through the use of "•" symbol. Note that no such segmentation boundary is indicated in the original Thai text.

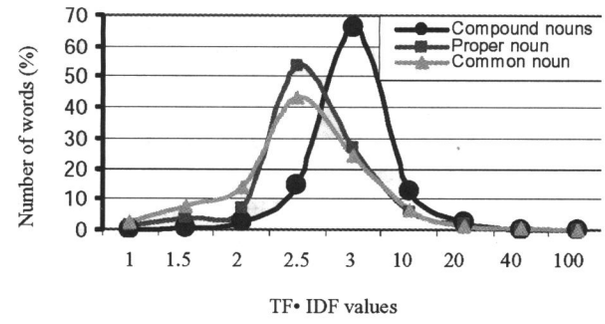

Fig. 2. The distribution of compound nouns, common noun and proper noun.

In our corpus with 417,030 words of 500 articles, we found the unduplicated of 8,901 compound nouns words, 1,718 common noun words and 5,836 proper noun words. Fig. 2 shows the comparison of distribution among compound nouns, common noun and proper noun.

According to Fig. 2, the percentage of common noun and proper noun at TF•IDF value which less than 2.5 are higher than the percentage of compound noun. On the other hand, the percentage of compound nouns was obtained higher when TF•IDF value is higher. This feature indicated that compound nouns word is more significant in Thai text than common noun and proper noun.

3.2. *Compound nouns extraction*

In order to extract Thai compound nouns, our task uses output from an available Thai word segmentation program developed by [11] as an input. This program is based on a trigram model to tag part of speech and used a learning algorithm to extract Thai words. This approach is claimed to segment Thai words correctly more than 90%. The POS of Thai words are classified into 47 types according to [12]. The input deriving from morphological analysis program consists of a stream of segmented words with POS. We then transform the sequence of nouns into compound nouns using three assigned patterns.

Pattern (1) shows compound noun (CN_1) which was extracted from the sequence of common noun (NCMN), proper noun (NPRP) and title noun (NTTL).

$$\text{Pattern (1)} \qquad CN_1 = N_1 \ldots N_n \quad , n > 1$$
$$N_n = \{\text{NCMN,NPRP,NTTL}\}$$

where n represents the number of words in a sequence order.

Pattern (2) shows compound noun (CN_2) which was extracted from the sequence of measurement classifier (CMTR) and common noun (NCMN).

$$\text{Pattern (2)} \qquad CN_2 = C_1 N_1$$
$$C_1 = \{\text{CMTR}\}$$
$$N_1 = \{\text{NCMN}\}$$

Moreover, words in Thai can be nominalized by adding a prefix such as "การ", "ความ" before that word [12]. Therefore, we also defined a sequence of nominal prefix (FIXN), common noun (NCMN) and active verb (VACT) in pattern (3) as compound nouns.

Pattern (3) $CN_3 = F_1 W_1 N_1$

$$F_1 = \{\text{FIXN}\}$$

$$W_1 = \{\text{VSTA,VACT,VATT, ADVN,XVMM,XVAM}\}$$

$$N_1 = \{\text{NCMN}\},$$

where VSTA is stative verb, VACT is active verb, VATT is attribute verb, ADVN is adverb in noun form, XVMM is pre-verb, and XVAM is pre-verb auxiliary.

The examples of compound nouns from each pattern are shown in table 1.

IV. Frame Work

4.1. Preprocessing

By using the input derived from an available morphological analysis program, we first constructed the compound nouns collocations by preprocessing with training data using the patterns as proposed in the previous section. The training data was assigned by selecting 500 business news articles from The Krungthep turakij online newspaper from year 2004 to 2005 (http://www.bangkokbiznews.com). The input data consists of headline and a body of text, with paragraph separators attached. From this step the compound nouns were extracted and each token was treated as a single word. Second, we employed a list of stop words [13] to eliminate Thai stop words. Thai stop words are unuseful words which frequently occurred in the document such as "ที่ "(at), "ใน"(in), "และ "(and), "เพื่อ "(for) etc. We also eliminated the special symbol characters from the text such as "?", "{", "}", "!" etc.

4.2. Term Frequency Inverse Document Frequency

Every word in each paragraph is then calculated by the term weighting technique. The term weighting technique has been widely investigated in Information Retrieval (IR).

TABLE 1
THE SAMPLES OF THAI COMPOUND NOUNS

Pattern	POS	Thai Compound noun	Meaning (English)
CN_1	NCMN•NCMN	ราคา•สินค้า	product price
CN_1	NCMN•NCMN	ธุรกิจ•รถยนต์	car business
CN_2	CMTR•NCMN	สวน•น้ำ	water field
CN_2	CMTR•NCMN	ร้าน•อาหาร	restaurant
CN_3	FIXN•XVMM•NCMN	ความ•ต้องการ•ธุรกิจ	business requirement
CN_3	FIXN•VACT•NCMN	การ•ประมูล•รถ	car auction

Ref. [14] cited Luhn's method that the frequency of a word occurrence in a document, as well as its relative position determines its significance in that document. More recent works have also employed Luhn's approach as a basic component for extracting relevant sentences [5] and also relevant paragraphs [7].

TF•IDF is a common term-weighting scheme in IR which is based on keyword frequency. The TF•IDF weight of word k in document i ($w_{i,k}$) can be calculated as follows:

$$w_{i,k} = tf_{i,k} \cdot idf_k = tf_{i,k} \cdot \log(\frac{D}{n_k}) \quad (1)$$

where $tf_{i,k}$ is the frequency of occurrence of word k in document i, n_k is the total number of documents which contains term k, and D is the total number of documents in the corpus. While the normalized paragraph p score of document i ($P_{TF\text{•}IDF}(p)$) can be calculated as follows:

$$P_{TF\text{•}IDF}(p) = \frac{\sum_{\{k\} \in P_p} w_{i,k}}{\sum n_p} \quad (2)$$

where $\{k\} \in P_p$ refers to the set of words in a paragraph p (P_p), and n_p is the total number of words in a paragraph p. The importance of a paragraph is determined by the summation of the important value of each word in the paragraph and normalized by dividing the total number of words in the paragraph excluding stop words.

4.3. Headline

It is plausible to think that terms appearing in a title or a headline are highly important. The basic idea is that the greater the number of words in a paragraph that match those in the headline, the more important the paragraph is likely to be formed. There had been some works applied this idea to calculate the score of important sentences or paragraphs to generate the summary [1], [15]. In this paper, we employed the headline score from [15]. This equation estimates the relevance between the headline and paragraph using the summation of TF•IDF value of words which both occurred in the headline and paragraph and normalized by TF•IDF value of all words in the headline. The normalized headline score of paragraph p ($P_{HL}(p)$) can be calculated as follows:

$$P_{HL}(p) = \frac{\sum_{\{k\} \in H_i \cap P_p} w_{i,k}}{\sum_{\{k\} \in H_i} w_{i,k}} \quad (3)$$

where $\{k\} \in H_i \cap P_p$ refers to the set of words which occurred in both the headline of document i (H_i) and

paragraph p (P_p), and $\{k\} \in H_i$ refers to the set of words in the headline of document i.

4.4. Paragraph score

When finishing step 4.3, the paragraph score ($P_{SCORE}(p)$) was calculated by the combination of equation (2) and (3) as follows:

$$P_{SCORE}(p) = P_{TF\text{•}IDF}(p) + P_{HL}(p). \quad (4)$$

After all paragraphs were assigned a score, the scoring paragraphs were ranked in descending order.

4.5. Paragraph extraction

We performed paragraph extraction by employing cosine similarity value in order to select the important paragraphs among similar paragraphs for the last step. The cosine similarity metric is an IR technique which determines the similarity between paragraphs by measuring the cosine of the angle between vectors of words [16]. Thus, the similarity of both paragraphs can be represented by identifying the high value of cosine similarity. In order to calculate this value, we first represented each paragraph in a document as a vector. The form of vector can be represented as follows:

$$P_i = (t_{i1}, t_{i2}, ..., t_{in}) \quad (5)$$

where n is the number of words in a document and t_{ij} is as follows:

$$t_{ij} = \begin{cases} 0 & ; t_j \text{ does not appear in } P_i \\ f(t_j) & ; t_j \text{ appears in } P_i \end{cases} \quad (6)$$

where $f(t_j)$ represents a frequency in which the word t_j appears in paragraph P_i. Given the vector representation of paragraphs $P_1, P_2, ..., P_m$ as in (5), the cosine similarity of P_i and P_j can be calculated by the inner product of their normalized vectors as follows:

$$Sim(P_i, P_j) = \frac{V(P_i) \cdot V(P_j)}{|V(P_i)||V(P_j)|} \quad (7)$$

For a set of paragraphs $P_1, P_2, ..., P_m$ of a document, we calculated the cosine similarity value of all possible pairs of paragraphs. If the cosine similarity value between P_i and P_j exceeds a predetermined threshold, then they are considered to be similar. Afterward, we selected one from a pair of paragraphs which had the higher value of paragraph score as calculated in (4). This is aimed to remove redundancy from the generated summary. Several experiments were also run in which the similarity threshold was varied in order to determine a suitable threshold value. Finally, after eliminating the similar paragraphs, a summary was generated. The summary consists of the ranked listing of top-scoring paragraphs according to the desired compression ratio, and the order in accordance with their appearance in the source text.

V. EXPERIMENT

5.1. Data

Standard data sets in Thai are not yet available for evaluating the text summarization system. Therefore we collected Thai business news of 2004 and 2005 from The Krungthep turakij online newspaper as the corpus. These news consists of 3,620 paragraphs from 500 articles. The average number of paragraphs in each article is 7.24. To observe the performance of our method, we selected 50 articles as test data.

5.2. Experimental setup

To evaluate the efficiency of proposed method, we developed four experiments.

Method A We implemented the experiment by using the method reported by [7]. This method proposed the paragraph score as local clustering score and global connectivity score. To obtain local clustering score, we used TL•TF method to indicate the significant words. Meanwhile, we computed the similarity from all paragraphs by using cosine similarity in order to get the global connectivity score. The summary was generated by extracting the n top ranked score paragraphs according to the combination scores.

Method B This experiment indicated the significant words by using the TF•IDF method and calculating the paragraph score by summation of the TF•IDF value of all words in each paragraph without considering compound nouns. To remove redundant paragraphs, this experiment also used cosine similarity value to consider similar paragraphs. Among similar paragraphs, a paragraph with highest paragraph score was selected. The paragraphs were ranked in decreasing order by the paragraph score. Finally a summary was generated by corresponding to the compression ratio.

Method C This experiment was implemented the same procedure as Method B but also using a combination of headline score to calculate the paragraph score. This method was implemented in order to evaluate the effect of the headline score.

Proposed Method The proposed method focuses especially on compound nouns by using the TF•IDF method, and headline method to score the important paragraphs as described in section three and section four. The cosine similarity was applied to eliminate redundant paragraphs.

The compression ratios of the summary from those four methods were set to 15%, 30% and 45%. To evaluate the suitable cosine similarity threshold, all experiments were tested using 0.2 – 0.7 of cosine similarity value as thresholds.

5.3. *Evaluation*

We evaluated results against the human-judged extracts according to the standard measure of precision, recall and F-score by using 50 business news articles as test data. Precision (*P*) is defined as a ratio of the number of correctly selected items to the number of all selected items. Recall (*R*) is defined as the ratio of the number of correctly selected items to the number of all items which should be selected while F-score (*F*) is the average of Precision and Recall. The evaluation can be calculated as follows:

$$P = \frac{S}{J} \tag{8}$$

$$R = \frac{S}{K} \tag{9}$$

$$F = \frac{2PR}{(R+P)} \tag{10}$$

where S represents the number of paragraphs occurring in both summaries selected by system and human judge. J represents the number of paragraphs selected by the system, and K represents the number of paragraphs selected by a human judge.

Generally, human judgment of the quality of a summary varies from person to person. Ref. [17] reported in their evaluation study that when asked people to pick the most relevant sentences in a given document, there was a little overlap of the sentences picked up by different persons. Hence it is difficult to measure the quality of a summary. In order to evaluate the effectiveness of human judgment in our work, we asked for four people as reader judges. We used F-score to evaluate among results of them and selected the best one who had the highest value to be the judge of our work.

VI. Result and Discussion

We compared summaries obtained by the previous three methods, i.e., Method A, Method B, Method C and proposed method as described in section 5.2, to illustrate the efficiency of our method. We used the F-score as a key to measure summary accuracy among all experiments. In table 2, λ refers to the cosine similarity threshold which varied from 0.2 to 0.7. Each row shows the results of F-score at 15%, 30% and 45% of summary respectively from method A, method B, method C and proposed method.

According to table 2, by comparing proposed method with Method A, the results of the proposed method were superior to Method A for all cosine similarity thresholds. This suggested that compound nouns and terms appearing in a headline were more significant in paragraph score calculation than term length. The best results of proposed method and Method A were obtained when cosine similarity threshold was 0.3 at 45% of summary, whereas gradually decreased results of both methods were obtained when cosine similarity threshold was increased. The decreasing of F-score was possibly associated with irrelevant paragraphs being included in the summary, therefore the matching paragraphs between the system summaries and human summaries were poor. The worst results of our proposed method, Method B and Method C were obtained when cosine similarity threshold was 0.2 at 15% summary. This may possibly caused by dissimilar paragraphs being assigned to be considered and the number of selected paragraphs in a summary tending to be low. Therefore less matching paragraphs between the systems and human judges were selected and caused the least precision values of proposed method, Method B and Method C. When we considered the F-scores between Method C and proposed method among all summary ratios, F-score obtained from proposed method is higher than it is from Method C at the 0.2, 0.3, 0.4 and 0.5 of cosine similarity threshold.

Fig. 3 shows the F-score curves for all methods at different cosine similarity thresholds. From Fig. 3, we can see the curve of our method resides over other methods. It was indicated that our method yielded the most effective F-score among all experiments at every cosine similarity threshold and this shows that compound nouns were influenced to weight the terms in the document.

TABLE 2
THE F-SCORE VALUES

λ	F-Score											
	Method A			Method B			Method C			Proposed method		
	15%	*30%*	*45%*	*15%*	*30%*	*45%*	*15%*	*30%*	*45%*	*15%*	*30%*	*45%*
0.2	0.130	0.324	**0.382**	0.114	0.229	**0.264**	0.220	0.260	**0.301**	0.246	0.389	**0.400**
0.3	0.149	0.362	**0.389**	0.134	0.318	**0.335**	0.267	0.325	**0.361**	0.283	0.461	**0.469**
0.4	0.148	0.362	**0.381**	0.152	0.366	**0.369**	0.262	0.410	**0.425**	0.271	0.451	**0.464**
0.5	0.130	0.345	**0.374**	0.148	0.408	**0.386**	0.240	0.434	**0.438**	0.262	0.437	**0.442**
0.6	0.129	0.292	**0.355**	0.134	0.401	**0.400**	0.242	0.437	**0.438**	0.254	0.424	**0.435**
0.7	0.112	0.309	**0.369**	0.132	0.394	**0.386**	0.234	0.432	**0.406**	0.251	0.426	**0.401**

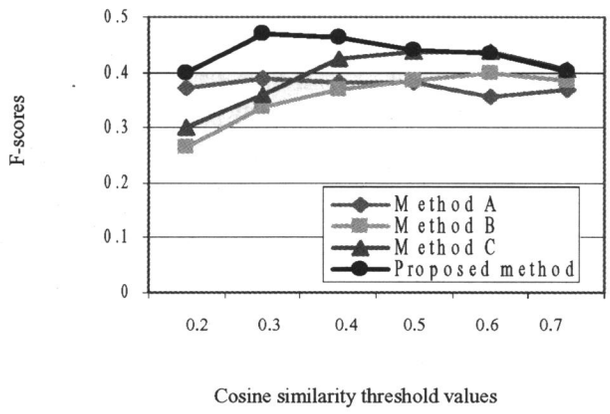

Fig. 3. F-score values at several of cosine similarity thresholds at 45% summary.

VII. Conclusion and Future Work

In this paper, we proposed a Thai text summarization system that extracts paragraphs from a document. The focus of this work is to apply POS to extract Thai compound nouns, a term weighting method in terms of TF•IDF, and headline score. According to the evaluation, our method showed the effectiveness of Thai compound nouns for extracting paragraphs and generating Thai summary.

In future research we intend to extend the next part of our system to translate Thai to another language using the results of text summarization based on our model.

References

[1] M. Banko, V. O. Mittal, M. J. Witbrock, "Headline Generation Based on Statistical Translation", *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong, 2000.

[2] H. Jing, "Sentence reduction for automatic text summarization", *In Proceedings of The sixth conference on Applied natural language processing*. New York, USA, pp. 310-315, 2000.

[3] Y. Seki, "Sentence extraction by tf/idf and position weighting from newspaper articles", *In Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text summarization and Question Answering*,Tokyo, Japan, 2003.

[4] C. Aone, M. E. Okurowski and J. Gorlinsky, " Trainable, Scalable Summarization Using Robust NLP and Machine Learning". *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Canada, pp. 62-66, 1998.

[5] J. Fukumoto, "Text summarization based on itemized sentences and similar parts detection between documents", *In Proceedings of the Third NTCIR Workshop on research in Information Retrieval, Automatic Text summarization and Question Answering*. Tokyo, Japan, 2003.

[6] E. Tzoukermann, J. L. Klavans and C. Jacquemin, "Effective use of natural language processing techniques for automatic conflation of multi-word terms: The Role of Derivational Morphology, Part of Speech Tagging and Shallow Parsing". *In Proceedings of the 20th Annual International ACM SIGIR, USA,* pp. 148-155, 1997.

[7] C. Jaruskulchai and C. Kruengkrai, "A practical text summarizer by paragraph extraction for Thai", *In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, IRAL-2003*, Sapporo, Japan, 2003.

[8] C. Kruengkrai and C. Jaruskulchai, "Generic text summarization using local and global properties of sentences". *In Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence, WI-2003*, Halifax, Canada, 2003.

[9] C. Jaruskulchai, A. Khanthong and W. Tantiprasongchai, "A framework for delivery of Thai content through mobile devices", *In Proceedings of the Closing Gaps in the Digital Divide Regional Conference on Digital GSM*. Bangkok, Thailand, 2003.

[10] A. Kawtrakul, M. Suktarachan, P. Varasai and H. Chanlekha, "A State of the Art of Thai Language Resources and Thai Language Behavior Analysis and Modeling", *Coling 2002 Post-conference workshops: the 3rd Workshop on Asian Language Resources and International Standardization*, Taipei, Taiwan, 2002.

[11] S. Meknavin, P. Charoenpornsawat, and B. Kijsirikul, "Feature-based Thai word segmentation", *In Proceedings of Natural Language Processing Pacific Rim Symposium '97* , Phuket, Thailand, pp 41-46, 1997.

[12] V. Sornlertlamvanich, N. Takahashi and H. Isahara, "Building a Thai Part-Of-Speech Tagged Corpus (ORCHID)". *The Journal of the Acoustical Society of Japan*, vol. 20, pp. 189-140, 1999.

[13] C. Jaruskulchai, "An automatic indexing for Thai text retrieval", *Ph.D. Thesis*, George Washington University, USA, 1988.

[14] S. Gerard and B. Christopher. "Term–weighting approaches in automatic retrieval", *Information Procession & Management*, vol. 24, pp. 513-523, 1988.

[15] C. Nobata, S. Sekine, K. Uchimoto and H. Isahara, "A summarization system with categorization of document sets", *In Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text summarization and Question Answering*, Tokyo, Japan, 2003.

[16] F. Fukumoto, Y. Suzuki and J. Fukumoto, "An automatic extraction of paragraphs based on the degree of context dependency", *In Proceeding of the Fifth conference on Applied natural language processing*. Washington. DC, USA, pp. 291-298, 1997.

[17] J. Goldstein, V. Mittal, J. Carbonell and J. Callan, "Creating and evaluating multi-document sentence extract summaries". *In Proceedings of the International Conference on Information and Knowledge Management, CIKM*, Washington. DC, USA, 2000.