

Short Paper

A Non-Training Approach to Generating High Semantic Titles for Chinese Documents

GAI-TAI HUANG AND HSIU-HSEN YAO*

The Institute of Strategic Defense Management

National Defense Management College, NDU

Taipei, 235 Taiwan

E-mail: hgt@rs590.ndmc.edu.tw

**Department of Computer Engineering*

Yuan Ze University

Taoyuan, 320 Taiwan

E-mail: csyao@saturn.yzu.edu.tw

Due to the abundance of available data, manual title generation may become unfeasible. Traditionally, information retrieval has been applied to perform automatic title generation by exploring and searching for keywords from a document. However, titles generated through such a direct combination approach may not satisfy Chinese grammatical rules, and also may not express the semantic meaning explicitly. Thus, we propose a non-learning approach based on conceptual schema to generating titles automatically through sentence modification and recombination. Experimental results prove that our model can satisfy the automatic title generation requirements.

Keywords: relation model, conceptual schema, title generation, information retrieval, Chinese document

1. INTRODUCTION

Manual title generation is becoming unfeasible due to the abundance of available data. Traditionally, information retrieval has been applied to perform automatic title generation, where keywords are explored and then combined to form a document's title. Nevertheless, keyword combination may not satisfy Chinese grammar rules nor explicitly indicate the semantic meaning of the combined words. Since Chinese text retrieval has just been developed recently, and due to the special characteristics of the Chinese language, retrieval approaches for Chinese are quite different from approaches proposed to deal with Western languages.

This paper is organized as follows. The first section is the introduction. Previous research on title generation is introduced in section 2. Section 3 describes title generation for Chinese documents based on a conceptual schema. The design methodology and the

Received August 9, 2002; revised July 7 and November 28, 2003; accepted February 2, 2004.
Communicated by Kuo-Chin Fan.

Entity-Relation-Entity (ERE) relational model process are introduced in section 4. Experiments performed to compare title generation methods are presented in section 5. Section 6 gives the experiment results and analysis, while conclusions are drawn in section 7.

2. RELATED RESEARCH

We divide the title generation process into three phases: choosing appropriate words for the title, deciding how many title words are appropriate for a document title, and finding the correct sequence of title words that forms a readable title “sentence.”

When approaching this problem, many researchers focus on automatic summarization generation. Traditional summaries have produced using the extractive approach, which selects sentences or paragraphs from a document to create a summary [1]. The weakness of the extractive approach is its inability to take advantage of the training corpus to produce a good summary that is a very small fraction of the complete text. Sometimes, summarized sentences are too long to be used as a title. Thus, extractive summarization is not suitable for title generation. Recently, researches in this field have focused on “learning approaches” [2] that take advantage of training data to learn the correlation between document words and title words. The words that have high correlation with document content and are considered to be appropriate for a title are determined and organized into a human readable sentence. However, it is not easy to form a readable Chinese sentence due to the wide variety of Chinese sentence structures. In addition, collecting a fixed amount of training data is not practical either. Thus, we propose a non-learning approach based on a conceptual schema to generating a title automatically with prior sentence modification and recombination.

3. TITLE GENERATION BASED ON THE ERE CONCEPTUAL SCHEMA APPROACH

A new document characterization model, EAVR [3], has been proposed to solve the Chinese text retrieval problem. In the EAVR conceptual model, an information index structure that satisfies the requirements of information retrieval or information extraction is established during the context analysis stage. In addition, a new type of concept tree allows a document characteristic schema to be reorganized and reconstructed. Basically, the approach we are proposing here is based on the concepts described above. However, in order to generate a title, the document is converted into an Entity-Relational-Entity (ERE) relational concept structure, and syntactic marks together with the underlying relational model are defined. Syntactic marks are described in detail in the following.

3.1 Syntactic Marks

Syntactic marks of Chinese words generally use grammatical marks resulting from sentence structure analysis, such as nouns, verbs, adjectives, etc. One advantage of grammatical marks is their ability to show the pattern of a sentence explicitly. However, it is hard to use grammatical marks to represent the concept relations among words. That is why grammatical marks are appropriate for grammatical analysis or sentence structure

analysis, but not for semantic analysis. To overcome this drawback, we define a set of syntactic marks as follows:

- (1) Ordinary Syntactic Marks: These are mostly nouns derived through syntactic analysis and defined as entities in the ERE relational model.

N : undetermined words from segmentation	O : noun (object), ex: “汽車”
NX : noun, adjective or object’s attribute, ex: “住址”	Q : measurement unit, ex: “公升”
NXX : new words resulting from segmentation	X : function word, ex: “因此”
V : quantity, numbers combined with measurement unit, ex: “一百二十公升”	T : time expression, ex: “今日”

- (2) Relation Syntactic Mark: Chinese sentences imply the semantic meaning of their characters. By taking the understanding and specific assumption about the semantic meaning of some particular characters, the relationship among words can be expressed as a relational model. Relational models that are commonly learned in Chinese sentences are described as follows and are defined as “Relations” in the ERE relational model:

A : action, ex: “發佈”, “佔”	ZOT: negation, ex: “不是”
= : synonym, ex: “澳洲” and “澳洲大陸”	P : without(no), ex: “沒有”
# : possession, ex: “含有”	-- : continuity relationship, ex: “,”
@ : belong to, ex: “均是”	* : any one(or), ex: “或”
→ : causality expression, ex: “導致”	> : part of, ex: “的”
+ : higher degree of comparison, ex: “增加”	\$: range (from to), ex: “至”
LOW : lower degree of comparison, ex: “低於”	& : and, ex: “與”
? : undetermined correlation word	< : reverse of part-of, ex: “位於”

By applying the syntactic marks defined above, we convert a document into ERE binary relationship lists, where: Set E is a set that consists of words belonging to one of {O, NX, V, NXX, N, T} syntactic marks. Set R is a set that consists of words that belong to one of {A, =, →, #, ZOT, @, >, LOW, &, <, \$, P, *, --, +, T, V} syntactic marks. The syntactic marks “V, T” may be members of set E or set R. First, we learn what the major pattern of a human-assigned title is: **ERE** list: E + R + E + ... Or **RER** list: R + E + R + ...

Thus, based on the statistical analysis described above, firstly, we convert a document content’s sentences into ERE lists. Then, we transform the automatic title generation problem into that of discovering, among the document’s ERE lists, the optimum ERE list that can serve as a representative of the document’s title. We assume that most documents are descriptive, reporting information; thus, a meaningful document context may be represented as a binary relation [4] to describe the relationship between themes. For example, the sentence “安非他命是國內危害最烈的毒品” (Amphetamine is the drug that causes the most harm domestically), after transformation into an ERE list, becomes:

Entity1	Relation	Entity2
安非他命	是(@)	國內
國內	危害(A)最烈的(>)	毒品

The ERE list of the s -th document is represented as: $D_s = (E_i, R_k, E_j)$

In the ERE Approach Model that we propose, document characteristics are first extracted and converted into the corresponding ERE lists. Afterwards, the algorithm searches for and calculates the optimum ERE lists that will serve as title candidates. Among the candidates, the best ERE list is determined as the document title. The design and algorithm steps are described in detail in the next section.

4. DESIGN AND PROCESS OF THE ERE RELATIONAL MODEL

4.1 Context Analyses and Word Segmentation

The objective of context analysis is to find and define document characteristics, where Chinese words segmentation may be involved. We will apply the approach proposed previously in [3], which uses a dictionary to accomplish segmentation and to define syntactic marks.

4.2 Conversion into an ERE Relational Model

After word segmentation is done, weight calculation of term frequency is performed to convert the document into an ERE relationship model list. A sentence is considered as a unit measurement during conversion since the meanings of words within a sentence have continuity and correlation. After function words have been removed, relevant words are converted into an ERE list according to the word order, syntactic marks and the rules defined below. Weight recalculation is performed during conversion based on term frequency (tf) multiplied by syntactic mark weight. (Words with the syntactic marks “O”, “NX”, “V”, and “A” have weights of 5, 4, 3, and 2, respectively). Conversion rules and the terminologies used inside are defined as follows:

Terminologies used in E_1 -R- E_2 list:

LE Set: Left Entity Set, the set that appears to the left of R. In other words, it is the set that represents E_1 in the E_1 -R- E_2 list.

RE Set: Right Entity Set, the set that appears to the right of R. In other words, it is the set that represents E_2 in the E_1 -R- E_2 list.

R Set: Relation word Set, the set that represents R in the E_1 -R- E_2 list.

LE Word: Entity(ies) belong(s) to E_1 before E_1 is converted into LE Set.

R Word: Relation word(s) belong(s) to R before R is converted into R Set.

RE Word: Entity(ies) belong(s) to E_2 before E_2 is converted into RE Set.

Conversion rules:

- (1) While **LE Word** is empty, if the syntactic mark of the word received belongs to set E, then the word is included in **LE Word**.

- (2) While the syntactic mark of the first received word belongs to set R, if **LE Word** is empty, then **LE Set** is replaced into “?”. Otherwise, the entities contained in it are combined and converted into **LE Set**. The word received is then included in **R Word**.
- (3) If **LE Set** already exists, and if the syntactic mark of the first received word belongs to set E, then the relations contained in **R Word** are combined and converted into **R Set**. The received word is then included in **RE Word**.
- (4) If **LE Set**, **R Set**, and **RE Word** all exist, when punctuation is received or a sentence terminates or a word with a syntactic mark of R set is received, then entities included in **RE Word** are combined and converted into **RE Set**. Hence, a complete ERE list is generated. The received word is included in **R Word** of the next ERE list, and the current **RE Set** is inserted into **LE Set** of the next ERE list.
- (5) If both **LE Set** and **R Word** exist but **RE Word** is empty, and if a sentence terminates, then **RE Set** is replaced with “?”. Also, a complete ERE list is generated.

4.3 Title Word Candidate Generation

After a document is converted into its corresponding ERE lists, the next step is to determine the appropriate ERE list that will serve as the document title. The main idea in title generation is to simulate the major pattern of a human generated title – the ERE list (E + R + E + R + ...). First, the LE Set with the most weight in each paragraph is determined, and then the most weighted RE Set within the same sentence that follows LE Set is also determined. The next step is to find **R Set** with the mark “A” that appears first within the same sentence that follows **LE Set**. Here, for convenience, **LE Set**, **RE Set** and **R Set** are abbreviated as LE, RE, and RW, respectively. Based on these three orientations (LE, RW, RE), we can obtain two types of title patterns (LE, RE, RW) or (LE, RW, RE), depending on the word order. Thus, the title generation problem is transformed into that of constructing a meaningful (LE, RW, RE) word list. Based on Chinese document writing conventions, most authors usually state the theme and key points of a document in the first paragraph. Thus, our algorithm takes the first sentence of the first paragraph as the groundwork to deduce title word candidates. Afterwards, based on the original paragraph of these candidates and the weight of the title, we determine the title of the document.

4.4 Document Title Determination

If document analysis results in only one title set, then it is used as the title of the document. If many sets of title candidates are generated, then the most appropriate title is selected based on the following formula that chooses the most weighted title after calculation. If there are distinct candidates with similar weights, then title determination depends on the positions of the candidates. Within a document, preceding positions have higher precedence to be chosen as the title. Our evaluation approach considers the following factors as measurement metrics:

- (1) Title length: the shorter the title, the better the result.
- (2) Average weights of title words: the higher, the better.
- (3) Original positions of title words: the earlier, the better.

Thus, the evaluation formula is defined as:

$$\text{Evaluation weight} = (\text{title weight}/\text{title length}) * (\text{shortest title length}/\text{title length}) \quad (1)$$

As an example, Table 1 shows candidate titles obtained from a specific document analysis along with their corresponding metrics.

Table 1. Examples of candidate titles obtained from a specific document analysis.

No	Candidate Title	Weight	Len.	Pos.
Title1	心肌梗塞患者心肌重生採用的新療法 (Myocardial infraction sufferers adopt new treatment for myocardium regeneration)	162	16	1
Title2	心肌梗塞所使用 GCSF 是受矚目的心肌梗塞新療法 (GCSF used by myocardial infraction is an arresting new treatment)	275	23	2

After weight calculation:

$$\text{Title1} = (162/16) * (16/16) = 10.1, \text{Title2} = (275/23) * (16/23) = 8.4$$

Title1 has a higher weight and precedes Title2 in the source document; thus, Title1 = “心肌梗塞患者心肌重生採用的新療法” is chosen as the title of the document.

5. THE TITLE GENERATION COMPARISON EXPERIMENT

Title generation based on the ERE model produces high semantic titles in contrast to an approach proposed in a previous study [5], which did not. Our approach provides a mechanism that preserves the overall integrity of the title’s meaning. In addition, it is a non-learning approach, and restricting the maximum title length in advance is not required. Our approach can generate a dynamic title length based on semantic meaning and significance. Our contrastive experiment evaluated title generation based on a Chinese word retrieval approach. The experimental data and evaluation are described in detail in the next section.

5.1 Experimental Data Description

- (1) **Experiment Corpus:** Our Experiment Corpus, consisting of 250 news reports and other electronic documents from various domains, including medical treatments, finance and economics, technology, and politics, were obtained randomly from the ChinaTimes [6] and IcareAsia [7] web sites in 2002.
- (2) **Dictionary:** Our dictionary consisted of about 130,000 words, some of which were obtained from a Chinese dictionary developed by Academia Sinica [8] while the rest were collected manually. We applied the approach proposed in [3] to obtain automatic labeling entity syntactic marks (noun words) and manually labeled, the syntactic marks of relation words, function words, measurement unit words, and time expression words in advance.

5.2 Evaluation

Three evaluations were performed in this study and a human-assigned title was used as a measurement standard: one to measure selection quality, another to measure the accuracy of the sequential order, and the last one to measure title length ratio.

- (1) To measure the selection quality of title words, an $F1$ metric was used (Van Rjiesbergen, 1979), defined as

$$F1 = (2 * \text{Precision} * \text{Recall}) \div (\text{Precision} + \text{Recall}), \quad (2)$$

where Precision and Recall are defined, respectively, as follows:

$$\text{Precision} = \frac{\text{The number of identical words in } T_{Auto} \text{ and } T_{human}}{\text{The number of words in } T_{Auto}} \quad (3)$$

$$\text{Recall} = \frac{\text{The number of identical words in } T_{Auto} \text{ and } T_{human}}{\text{The number of words in } T_{human}} \quad (4)$$

T_{Auto} : automatically generated title; T_{human} : human assigned title

- (2) To measure how well a generated title compared with the original human-generated title in terms of word order, we measured the number of correct title words in the hypothesis titles that were in the same order as in the reference titles in (Nye, 1984):

$$\text{Correct Order} = \frac{\text{The number of identical words in } T_{Auto} \text{ and } T_{human} \text{ with the same word order in } T_{human}}{\text{The number of identical words in } T_{Auto} \text{ and } T_{human}} \quad (5)$$

- (3) For the third metric, the length of the automatically generated title was compared to that of the human assigned title:

$$\text{Length ratio} = \frac{\text{Length of automatically generated title}}{\text{Length of human assigned title}} \quad (6)$$

5.3 Simulation of Chinese Document Title Generation

Han [5] noted that Chinese document title generation is a Component-Word based approach, Subject-word based approach, and Statistics-based approach. The first two approaches are dictionary-based and are similar to the ERE model approach proposed in this paper. One of the differences between Han's approaches and our approach is that the former use the restriction that only the first 15 most-weighted characters have an opportunity to become a title, while the latter does not. Our experiment was then divided into two parts, and a comparison was made using the dictionary-based approach and statistics

based approach, where the ERE model approach was chosen as the simulation methodology of the dictionary-based approach, while the statistics-based approach applies Han's [5] research.

6. EXPERIMENTAL RESULTS AND DISCUSSION

To illustrate the quality of the experiment results, we will first give examples of machine-generated titles, and then present the quantitative results and analysis.

6.1 Example

Original title: 微風廣場推出百貨白金卡 (Breeze Center offers department store Visa Platinum Card)

Original content: 為搶攻高消費信用卡市場，微風廣場與聯邦銀行於昨（二十三）日共同推出國內第一張百貨白金卡，持卡人除可享受百貨、超市、商品、主題餐廳九折，電影票早場優惠價及三小時免費停車外，喜愛名品的白金持卡人更可以於微風廣場一樓國際精品館選擇六期免息分期付款，同時推出一%的現金回饋讓白金卡友享受刷卡購物的樂趣。聯邦銀行表示，...。(To win over the high-consumption credit card market, Breeze Center and Union Bank offered the first department store platinum card yesterday (23) in Taiwan. Card holders, enjoying a 10% discount on general merchandise, supermarket goods, discounted prices for morning show movie tickets, and three-hour free parking. Those fond of labeled brands may choose to pay in installments of six periods with no interest in international souvenir stores on the first floor. Meanwhile, a 1% cash refund is offered to let card holders enjoy the convenience of shopping with credit cards. Union Bank's)

Automatically generated title:

Approach	Title
ERE Model Approach	微風廣場聯邦銀行推出國內第一張百貨白金卡
Statistics Based Approach	消費白金卡專屬停車優惠小時微風廣場持卡人
Select Most Weight Sentence Approach /Select First Sentence Approach	為搶攻高消費信用卡市場，微風廣場與聯邦銀行於昨（二十三）日共同推出國內第一張百貨白金卡，...，同時推出一%的現金回饋讓白金卡友享受刷卡購物的樂趣。

The results table shows that the Select Most Weight Sentence approach and the Select First Sentence approach usually generated an overly-long title, while the ERE model approach could pick suitably concise, words for the summary results to represent the document's title. It is clearly shown that the statistics-based approach may not generate a meaningful title. A detailed comparison is presented in the next section.

6.2 Analysis of Generated Titles

Some experimental results are listed below (Table 2) to show the differences be-

tween the results obtained using the ERE model approach and those obtained using the statistics-based approach. It is obvious that the statistics-based approach usually does not generate meaningful titles. It is also evident that titles generated by the ERE model approach are quite meaningful. Moreover, the meanings of the titles generated by the ERE model approach are almost the same as those of human-assigned titles for the corresponding documents.

Table 2. Comparison table of automatically generated titles.

Original Title	ERE Title	Statistics Title
倫敦證交所可能進行合併 (London Stock Exchange(LSE) may proceed a merge)	市場週一傳出倫敦證券交易所有可能與其他交易所合併 (Market spreads that London Stock Exchange (LSE) may merge with other Stock Exchanges)	合併有可能 (Merger, possible)
陳總統會晤馬紹爾總統 (President Chen meets Marshall Islands' President)	陳總統在總統府與諾特總統進行首次晤談 (President Chen and President Note proceed with the first interview)	總統諾特諾特總統陳總統中華民國訪問伉儷勳章馬國 (President Note, Note President, President Chen, ROC, interview, a couple, medal of honor, Marshall)
香港精神病患去年急升 15% (Hong Kong psychiatric patients increase rapidly up to 15% last year)	香港醫院顯示精神科專科門診前年上升一成五 (Hong Kong hospital showed an increase of 15% in outpatient services in the psychiatry department last year)	精神精神病忽視求診人病患社會精神病患首次求診首次求診人香港 (Spirit, psychotic, ignore, diagnose, patient, patient, society, psychotic, first diagnosis, first patient, Hong Kong)

6.3 Analysis of Experimental Results

- (1) Comparison of average length of title words: First, we analyzed the title lengths. The lengths of human-assigned titles were about 14-15 letters, and the ERE model approach produced title lengths of 26-27 letters, while the statistics-based approach generated titles with average lengths of 28-29 letters, twice those of the original titles, as shown in Fig. 1 below.
- (2) Comparison of average F1 value, Precision value, and Recall value: The ERE model approach had an average precision value of about 34%, while the statistics-based approach's was just 27%. The ERE model approach and the statistics-based approach had average recall rates of 58% and 41%, respectively. Thus, the resulting F1 metrics for both the ERE model approach and statistics-based approach were 41% and 28% respectively, as shown in Fig. 2 below.
- (3) Comparison of the correct order of title words: The ERE model approach achieved an average correct order value of 95%, while the statistics-based approach achieved only 80%. (Fig. 3)

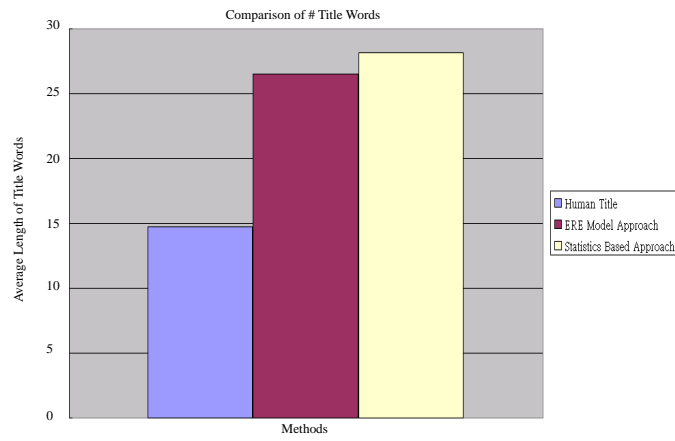


Fig 1. Comparison of the average lengths.

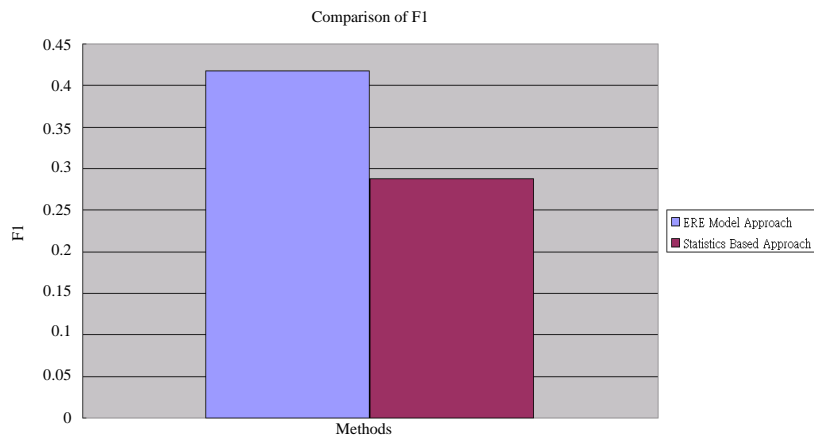


Fig. 2. Comparison of F1 values of title words.

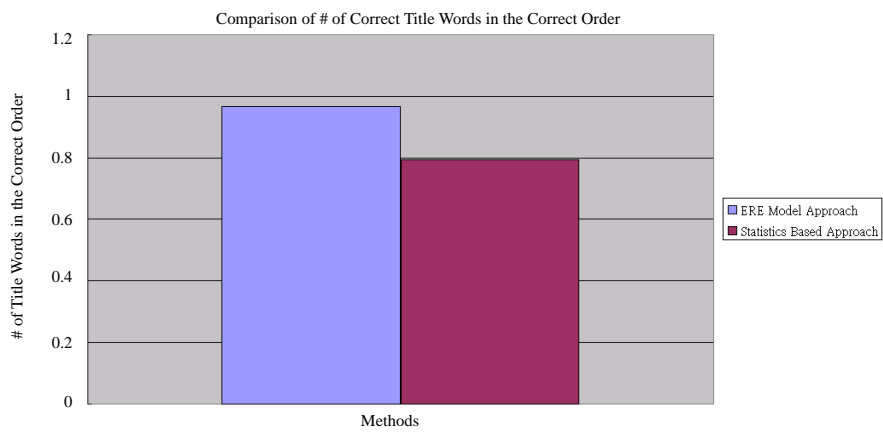


Fig. 3. Comparison of correct order rates of title words.

Comparing the two approaches (the ERE model approach and the statistics-based approach), although both generate longer title lengths than those produced by humans, we can see from the experimental results that the ERE model approach has better performance. That is, the ERE model approach generates titles with semantic meanings that are nearly the same as those of human assigned titles.

7. CONCLUSIONS

While automatic title generation for Chinese documents remains a non-semantic based approach, the ERE model approach we have proposed in this paper can generate a title that is close to a human-generated title and is highly semantic. Although this approach requires the use of a dictionary to complete the syntactic marking process, it provides better semantic results compared with previously proposed approaches.

REFERENCES

1. K. McKeown and R. Radev; "Generating summaries of multiple news articles," *18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995, pp. 1-9.
2. R. Jin and A. G. Hauptmann, "Title generation for spoken broadcast news using a training corpus," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Vol. 2, 2000, pp. 680-683.
3. G. T. Huang and H. H. Yao, "Chinese text information extraction based on conceptual schema," *International Conference on Chinese Computing-eLearning*, 2001, pp. 242.
4. J. C. Wan and W. L. Yao, "Chinese syntactic and semantic analysis based on binary relations," *Communications of COLIPS*, Vol. 8, 1998, pp. 31-42.
5. K. S. Han, Y. C. Wang, and F. F. Wu, "Research on extracting subject from Chinese text," in *Proceedings of the 5th International Workshop Information Retrieval with Asian Languages*, 2000, pp. 211-212.
6. <http://news.chinatimes.com/>.
7. <http://www.icareasia.com/>.
8. http://godel.iis.sinica.edu.tw/CKIP/r_content.htm.

Gai-Tai Huang (黃志泰) was born in 1964. He received the M.S. degree in Information Management from National Defense Management College, Taiwan, in 1992 and Ph.D. degree in Computer Engineering from Yuan Ze University, Taiwan, in 2003. Currently he is a strategic studies instructor in National Defense Management College. His research interests are information extraction, web mining and electronic commerce.

Hsiu-Hsen Yao (姚修慎) received the M.S. degree in Computer Engineering from National Taiwan University, Taiwan, in 1985 and Ph.D. degree in Computer Engineering from Case Western Reserve University, U.S.A., in 1989. Currently he is a Professor in Yuan Ze University. His research interests are communication, information theory and electronic commerce.