COMPUTER PROCESSING OF ORIENTAL LANGUAGES, VOL. 11, NO. 4, 1998

Named Entity Extraction for Information Retrieval[1]

HSIN–HSI CHEN, YUNG–WEI DING, and SHIH–CHUNG TSAI

Abstract: Name extraction is indispensable for both natural language understanding and information retrieval. However, proper names are major unknown words in natural language texts, and unknown word identification is still a challenge problem in natural language processing. This paper deals with identification of person names, organization names and location names from Chinese texts. Different types of information from different levels of text are employed, including character conditions, statistic information, titles, punctuation marks, organization and location keywords, speech–act and locative verbs, cache and n–gram model. We also clarify which strategies can be used in which cases, i.e., queries and/or documents. In our experiments, the recall rates and the precision rates for the extraction of person names, organization names, and location names under MET data are (87.33%, 82.33%), (76.67%, 79.33%) and (77.00%, 82.00%), respectively.

Keywords: Chinese language processing, Information retrieval, N–gram model, Named entity extraction, Word segmentation.

1. Introduction

People, affairs, time, places and things are five basic entities in a document. When we catch the fundamental entities, we can understand a document to some degree. These entities are also the targets that users are interested in. That is, users often issue queries to retrieve such kinds of entities in information retrieval systems. Thompson and Dozier [1] reported an experiment over periods of several days in 1995. It showed 67.8%, 83.4%, and 38.8% of queries to Wall Street Journal, Los Angeles Times, and Washington Post, respectively, involve name searching. Besides name searching, name identification has many applications. Chen and Wu [2] consider person names as one of cues in sentence alignment. Chen and Lee [3] show its application to anaphora resolution. Chen and Bian [4] propose a method to construct white pages for Internet/Intranet users automatically. They extract information from World Wide Web documents, including proper nouns, E–mail addresses and home page URLs, and find the relationship among these data.

Name extraction is indispensable for both natural language understanding and information retrieval. However, proper names are major unknown words in natural language texts. Chen, He and Xu [5] examined TREC–5 Chinese collection and found that there were 287 university and college names, and 627 company names. Only 21 out of 287 names and 14 out of 627 are included in their dictionary. Unknown word identification is a challenge problem in natural language processing. Many papers [6–8] touch on this problem. In a famous message understanding system evaluation and message understanding conference (MUC), which is sponsored by Tipster Text Program of DARPA, named entity, which covers named organizations, people, and locations, along with date/time expressions and monetary and percentage expressions, is one of tasks for evaluating technologies. In MUC–6 named entity task, the systems developed by SRA [9] and BBN [10] on the person name recognition portion have very high recall and precision scores (over 94%). From 1996, MUC extends the named entity

[1] The authors are with Department of Computer Science and Information Engineering, National Taiwan University, Taipei, TAIWAN, R.O.C.

extraction from English to Multilingual Entity Task (MET). MET–2 focuses on Chinese and Japanese.

Name extraction in Chinese documents is more difficult than that in English documents. This is because there are not any word boundaries in Chinese sentences. Because of the segmentation problem in Chinese, character–based approach [11] is usually adopted in Chinese information retrieval. Kwok [12] even claims that word segmentation may not be a pre–requisite for Chinese. However, we can find segmentation is needed in cross–language information retrieval [13]. When we issue a Chinese query to retrieve documents in another language, we have to introduce query translation. Besides, segmentation result is an important cue to improve precision when used in information filtering. In these ways, segmentation is required.

Segmentation problem makes name extraction in Chinese texts more difficult. Chang, et al. [14] proposed a method to extract Chinese person names from an 11,000–word corpus, and reported 91.87% precision and 80.67% recall. Wang, et al. [15] recognized unregistered names on the basis of titles and a surname–driven rule. Sproat, et al. [16] considered Chinese person names and transliterations of foreign words. Their performance was 61.83% precision and 80.99% recall on a 12,000–Chinese–character corpus. Chen and Lee [3, 17] propose various strategies to identify and classify Chinese person names, transliterated names, and organization names in Chinese texts. The precision and the recall for these three types of proper names are (88.04%, 92.56%), (50.62%, 71.93%) and (61.79%, 54.50%), respectively. In the above approaches, Chen and Lee [3] deal with more types of proper names than others, but all of these papers do not clarify the differences of name extraction in queries and documents.

Identification of proper names in queries is different from that in large–scale texts. The major difference is that query is always short. Thus its context is much shorter than full texts and some technologies involving larger contexts are useless. This paper will introduce language models to extract person names, organization names and location names. Besides, we will tell which methods can be employed to which cases, i.e., queries and/or documents. Sections 2, 3, 4 present our language models for different types of proper names. Section 5 discusses the applicable domains. MET data is used to measure the performance of each strategy, and discuss the applicability. Section 6 concludes the remarks.

2. Named People Extraction

Person names can denote Chinese or foreigners. The formulation rules for these two types of names are totally different. We often transliterate a foreign name into Chinese. Thus the language models are different for the extraction of Chinese person names and transliterated names.

2.1 Chinese person names

Chinese person names are composed of surnames and names. Most Chinese surnames are single character and some rare ones are two characters. The following shows three different types:

(a) Single character like ‘趙’, ‘錢’, ‘孫’ and ‘李’.

(b) Two characters like '歐陽' and '上官'.

(c) Two surnames together like '蔣宋'.

Most names are two characters and some rare ones are single characters. Theoretically, every character may be used to form a name rather than a fixed set. Thus the length of Chinese person names ranges from 2 to 6 characters.

In our model, input text is segmented roughly beforehand. This is because many characters have high probabilities to be a Chinese person name without pre–segmentation. Consider the example '蘇聯與南韓達成 ...'. The character '韓' has a high score to be a surname. In this aspect, '達成' is easy to be a name. If the input text is not segmented beforehand, it is easy to regard '韓達成' as a Chinese person name. On the statistical model, this type of errors is difficult to avoid. However, it is easy to capture by pre–segmentation.

Three kinds of recognition strategies shown below are adopted:

(1) name–formulation statistics

(2) context cues, e.g., titles, positions, speech–act verbs, and so on

(3) cache

Name–formulation statistics form the baseline model. It proposes possible candidates. The context cues add extra scores to the candidates. Cache records the occurrences of all the possible candidates in a paragraph. If a candidate appears more than once, it has high tendency to be a person name. The following illustrates each strategy in details.

Name–formulation statistics are trained from a person name corpus in Taiwan [18]. It contains near one million Chinese person names, including 489,305 men and 509,110 women. Each contains surname, name and sex. During training, we divide the corpus into two partitions according to sex of persons. In our method, we postulate that the formulation of names is different for male and female. At first, we get 598 surnames from this person name corpus, and then compute the probabilities of these characters to be surnames. Of these, surnames of very low frequency like "是", "那", etc., are removed from this set to avoid too much false alarms. Only 541 surnames are left, and are used to trigger the person name identification system. Next, the probability of a Chinese character to be the first character (the second character) of a name is computed for male and female, separately.

The following models are adopted to select the possible candidates. We consider the above three types of surnames.

Model 1. Single character

(i) P(C_1)*P(C_2)*P(C_3) using the training table for male > Threshold$_1$
and
P(C_2)*P(C_3) using the training table for male > Threshold$_2$, or

(ii) P(C_1)*P(C_2)*P(C_3) using the training table for female > Threshold$_3$ and
P(C_2)*P(C_3) using the training table for female > Threshold$_4$

Model 2. Two characters

(i) P(C_2)*P(C_3) using the training table for male > Threshold$_2$, or

(ii) P(C_2)*P(C_3) using the training table for female > Threshold$_4$

Model 3. Two surnames together

P(C_{12})*P(C_2)*P(C_3) using the training table for female > Threshold$_3$,
P(C_2)*P(C_3) using the training table for female > Threshold$_4$ and
P(C_{12})*P(C_2)*P(C_3) using the training table for female >
P(C_{12})*P(C_2)*P(C_3) using the training table for male

where C_1, C_2 and C_3 are a continuous sequence of characters in a sentence, and they denote surname and names, respectively,
C_{11} and C_{12} denote the first and the second surnames, and
P(C_i) is the probability of C_i to be a surname or a name.

For different types of surnames, different models are adopted. Because the surnames with two characters are always surnames, Model 2 neglects the score of surname part. Both Models 1 and 3 consider the score of surnames. We compute the probabilities using the training tables for male and female, respectively. In Models (1) and (2), either score of male name or score of female name must be greater than thresholds. In Model (3), the person names must denote a female. In this case, the probability to be female must be greater than the probability to be male. The above three models can be extended to the single–character names. When a candidate cannot pass the thresholds, its last character is cut off and the remaining string is tried again. Thresholds are trained from our person name corpus. We let 99% of the training data pass the thresholds. The thresholds for male and female are computed separately.

Besides the baseline model, titles, positions and special verbs are important local cues. When a title such as '總統' (President) appears before (after) a string, it is probably a person name. There are 476 titles in our database. Person names usually appear at the head or the tail of a sentence. Persons may be accompanied with speech–act verbs like "發言", "說", "提出", etc. For these cases, extra scores are added to help strings pass the thresholds.

Finally, we present a global cue. A person name may appear more than once in a paragraph. We use cache to store the identified candidates and reset cache when next paragraph is considered. There are four cases shown below when cache is used:

(1) $C_1C_2C_3$ and $C_1C_2C_4$ are in the cache, and C_1C_2 is correct.

(2) $C_1C_2C_3$ and $C_1C_2C_4$ are in the cache, and both are correct.

(3) $C_1C_2C_3$ and C_1C_2 are in the cache, and $C_1C_2C_3$ is correct.

(4) $C_1C_2C_3$ and C_1C_2 are in the cache, and C_1C_2 is correct.

Cases (1) and (2) (cases (3) and (4)) are contradictory. In our treatment, a weight is assigned to each entry in the cache. The entry that has clear right boundary has a high weight. Titles, positions, and special verbs are cues for boundary. For those similar pairs that have different weights, the entry having high weight is selected. If both have high weights, both are chosen.

When both have low weights, the score of the second character of a name part is critical. It determines if the character is kept or deleted.

2.2 Transliterated person names

Transliterated person names denote foreigners. Compared with Chinese person names, the length of transliterated names is not restricted to 2 to 6 characters. The following strategies are adopted to recognize transliterated names:

(1) character condition

Two special character sets are retrieved from Hornby [19] and Huang [20]. The first character of transliterated names must belong to a 280–character set, and the remaining characters must appear in a 411–character set. Some of them are listed below:

First character: ‘丁’, ‘于’, ‘丹’, ‘井’, ‘內’, ‘切’, ‘厄’, ‘夫’, ‘孔’, ‘尤’, ‘’, etc.
Other character: ‘丁’, ‘上’, ‘于’, ‘士’, ‘大’, ‘山’, ‘丰’, ‘丹’, ‘井’, ‘什’, ‘今’, etc.

The character condition is a loose restriction. The string that satisfies the character condition may denote a location, a building, an address, etc. It should be employed with other cues (refer to (2)–(4)).

(2) titles

Titles used in Chinese person names are also applicable to transliterated person names. Thus, 聖約翰 港 will not be recognized as a transliterated person name. There are 440 titles in the current version. For example, ‘一兵’, ‘二兵’, ‘下士’, ‘下官’, ‘上兵’, ‘上尉’, ‘上將’, etc.

(3) name introducers

Some words like “叫”, “叫作”, “叫做”, “名叫”, and “尊稱” can introduce transliterated names when they are used at the first time.

(4) special verbs

Persons always appear with some special verbs like “發表”, “暗示”, and so on. Thus the same set of verbs used in Chinese person names are also used for transliterated person names.

Besides the above strategies, a complete transliterated person name is composed of first name, middle name and last name. For example, 阿卜杜勒・巴塞特・阿里・賽格拉西. The first, the middle and the last names are connected by a dot.

Cache mechanism is also helpful in the identification of transliterated names. A candidate that satisfies the character condition and one of the cues will be placed in the cache. At the second time, the cues may disappear, but we can recover the transliterated person name by checking cache. The following shows an example:

... 米切爾法官 ..., ... 米切爾因為 ...

Title does not show up, when the name is mentioned at the second time.

2.3 Performance evaluation

Table 1. The performance of named people extraction.

	POS	ACT	COR	INC	MIS	SPU	REC	PRE
met1(F)	683	770	618	11	54	141	90	80
met2(T)	327	348	291	1	35	56	89	84
met2(D)	142	159	132	2	8	25	93	83

We test our system with three sets of MET data. It includes MET–1 formal run (100 documents and 124,186 bytes), MET–2 training (104 documents and 84,976 bytes) and MET–2 dry run (26 documents and 32,772 bytes). The material is selected from various journalistic sources such as newspaper text from Renmin Ribao (People's Daily), radio scripts from China Radio International, and newswire text from Xinhua newswire service. Table 1 shows the performance of named people extraction. POS denotes total number (COR+INC+MIS) of proper names. ACT denotes the number (COR+INC+SUP) of names that our system generates. COR denotes the number of correct names. INC denotes the number of incorrect names. MIS denotes the number of missed proper names. SUP denotes the number of spurious names that our system generates. REC denotes recall rate (COR/POS), and PRE denotes precision rate (COR/ACT). In our experiments on named people extraction, the average recall rate is 87.33% and the average precision rate is 82.33%.

3. Named Organization Extraction

The structure of organization names is more complex than that of person names. Basically, a complete organization name can be divided into two parts, i.e., name and keyword. The following specifies the rules we adopted to formulate its structure.

OrganizationName → OrganizationName OrganizationNameKeyword

e.g., 聯合國 部隊

OrganizationName → CountryName OrganizationNameKeyword

e.g., 美國 大使館

OrganizationName → PersonName OrganizationNameKeyword

e.g., 羅慧夫 基金會

OrganizationName → CountryName OrganizationName

e.g., 美國 國防部

OrganizationName → LocationName OrgnizationName

e.g., 伊利諾州 州府

OrganizationName → CountryName {D|DD} OrganizationNameKeyword
where D is a content word.

e.g., 中國 國際 廣播電台

OrganizationName → PersonName {D|D} OrganizationNameKeyword

e.g., 羅慧夫 文教 基金會

OrganizationName → LocationName {D|D} OrganizationNameKeyword

e.g., 台北 國際 廣播電台

The first five rules show organization names, country names, person names and location names can be placed into the name part of organization names. Person names can be found

Table 2. The performance of named organization extraction.

	POS	ACT	COR	INC	MIS	SPU	REC	PRE
met1(F)	705	608	496	7	202	105	70	82
met2(T)	268	270	206	5	57	59	77	76
met2(D)	80	83	66	2	12	15	83	80

by the approaches specified in the last section. Location names will be touched on in the next section. Transliterated names may appear in the name part. We use the same character sets mentioned in the last section. If a sequence of characters meet the character condition, the sequence and the keyword form an organization name. In current version, we collect 776 organization names and 1059 organization name keywords. Some of keywords are shown below:

‘人民醫院’, ‘人事部’, ‘人壽’, ‘三溫暖’, ‘土地局’, ‘士校’, ‘大使館’, ‘大飯店’, etc.

The last three rules are very special. They specify that common content words may be inserted in between the name part and the keyword part. For example, ‘陶聲洋防癌基金會’ (S.Y.Dao Memorial Fund). A content word ‘防癌’ appears between a person name ‘陶聲洋’ and an organization name keyword ‘基金會’. In current version, at most two content words are allowed.

As mentioned, transliterated person names and location names in the above rules still have to satisfy the character condition. However, the character set is trained from transliterated person name corpus. It may not be suitable for location names. Consider an example “帕鬆錯湖旅遊度假村”. “帕鬆錯湖”, which is a lake in China, is not a transliterated name. The characters “鬆” and “錯” do not belong to the character set. Here, we utilize the feature of multiple occurrences of organization names in a document and propose n–gram model to deal with this problem. Although cache mechanism and n–gram use the same feature, i.e., multiple occurrences, their concepts are totally different. For organization names, we are not sure when a pattern should be put into cache because its left boundary is hard to decide. In our n–gram model, we select those patterns that meet the following criteria:

(1) It must consist of a name and an organization name keyword.

(2) Its length must be greater than 2 words.

(3) It does not cross sentence boundary and any punctuation marks.

(4) It must occur at lease two times.

The same MET data are used to test the performance of extraction of organization names. Table 2 lists the experiment results. The average recall rates and the average precision rates for the identification of organization names are 76.67% and 79.33%, respectively.

4. Named Location Extraction

The structure of location names is similar to that of organization names. A complete location name is composed of name part and keyword part. We use the following rule to formulate this structure.

Table 3. The performance of named location extraction.

	POS	ACT	COR	INC	MIS	SPU	REC	PRE
met1(F)	1664	1594	1251	11	402	332	75	78
met2(T)	908	819	697	28	183	94	77	85
met2(D)	170	162	135	8	27	19	79	83

LocationName → PersonName LocationNameKeyword
LocationName → LocationName LocationNameKeyword

Currently, we have 45 location keywords. The following shows some examples:

‘山’, ‘中心’, ‘公路’, ‘以北’, ‘以西’, ‘以東’, ‘以南’, ‘半島’, ‘半球’, ‘市’, ‘市中心’, etc.

There are 16,442 built–in location names in current versions. For the treatment of location names without keywords, we also introduce some locative verbs like ‘來自’, ‘前往’, and so on. The objects following this kind of verbs may be location names. For example, in the string “飛往聖路易斯”, “聖路易斯” will be identified. Cache is also useful. For example, assume ‘巴塞隆納市’ is recognized as a location name and placed in cache. When ‘巴塞隆納’ appears, it will be identified as a location name even if the location name keyword is omitted. N–gram model is also employed to recover those names that do not meet the character condition.

Table 3 lists the performance evaluation of language models for identification of location names. The average recall rates and the average precision rates are 77.00% and 82.00%, respectively. The performance of extraction of organization and location names is very similar, and is a little worse than that of person names. The major reason is the former two depend on keyword sets very much. The training data for organization and location names are much smaller than that for person names.

5. Name Extraction Strategies and IR

Proper names are plausible index terms, but proper names are infrequent words relative to other content words in corpora. In information retrieval, most frequent and less frequent words may be regarded as unimportant words and be neglected. Previous sections propose methods to extract the interesting terms. We know there are significant differences between document and query representations. Documents provide longer context, while queries usually have shorter context. The strategies suitable for documents are not always applicable to queries. We clarify the possible strategies into four different levels of text: document, paragraph, sentence and character levels. The contexts of these four levels range from the longest to the shortest. The following shows what strategies can be applied at each level.

(a) character level: baseline model, character condition

(b) sentence level: titles, keywords, punctuation marks, speech–act and locative verbs

(c) paragraph level: cache

(d) document level: n–gram model

Chinese character is a basic unit for composing a proper name. We collect a specific set for Chinese surname, and two character sets for transliterated names. A possible candidate must belong to the specific set. This character condition is independent of context. Because it is easy to produce many false alarms, context information from sentence, paragraph and document levels is employed to eliminate impossible cases.

Then we discuss information from sentence level. Several keywords can introduce proper names. They may be nouns or verbs. Titles, organization and location keywords are such kinds of nouns. Speech–act and locative verbs also have this function, but they are less used in queries than titles, organization and location keywords.

Information from paragraph and document levels cannot be employed to query processing. At first, we discuss the model at document level. N–gram model is usually applied to a large document collection to extract the repeating patterns. These patterns may be new words or just insignificant patterns. Filtering rules may be added or human may be involved to eliminate insignificant patterns. The papers [21, 22] based on Smadja's paradigm [23] learned an aided dictionary from a corpus to reduce the possibility of unknown words. Because proper names are some sort of unknown words, n–gram model is also useful. Surname, length of patterns (e.g., 2–4 words in Chinese person names), organization keywords and location keywords are plausible cues to filter out the useless patterns.

Cache is a mechanism from paragraph level. Similar to the n–gram model, it employs the repeating occurrences of patterns to extract proper names. The major differences are the criteria of candidates and the scope of context. The entities that are entered into cache must meet some conditions. Chinese person names must pass the examination of baseline models. Titles and positions give them extra weights. Transliterated names must belong to the specific character sets. Speech–act and locative verbs may introduce proper names. Organization and location keywords are overt marks. These conditions are regarded as pre–filter in cache approach and post–filter in n–gram approach. Besides this difference, the entities in the cache are cleared when a new paragraph starts. Thus, the context is narrower in cache approach.

6. Concluding Remarks

This paper presents several strategies to extract person names, organization names and location names. Because of the short queries, only information from character and sentence levels can be employed to query processing. When all the information is used, the recall rates and the precision rates for the extraction of person names, organization names, and location names under MET data are (87.33%, 82.33%), (76.67%, 79.33%) and (77.00%, 82.00%), respectively. Surnames, organization and location keywords are the most important cues. When they are absent from the context, it is not easy to capture the proper names.

References

[1] P. Thompson and C. Dozier, "Name searching and information retrieval," in *Proceedings of Second Conference on Empirical Methods in Natural Language Processing,* Providence, Rhode Island, 1997, **pp.** .

[2] H. H. Chen and Y. Y. Wu, “Aligning parallel Chinese–English texts using multiple clues,” in *Proceedings of 2nd Pacific Association for Computational Linguistic Conference*, Queensland, Australia, 1995, pp. 39–48.

[3] H. H. Chen and J. C. Lee, “Identification and classification of proper nouns in Chinese texts,” in *Proceedings of 16th International Conference on Computational Linguistics*, **location**, 1996, pp. 222–229.

[4] H. H. Chen and G. W. Bian, “Proper name extraction from Web pages for finding people in Internet,” in *Proceedings of ROCLING X*, Taipei, Taiwan, 1997, pp. 143–158.

[5] A. Chen, J. He, and L. Xu, “Chinese text retrieval without using a dictionary,” in *Proceedings of 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, **location**, 1997, pp. 42–49.

[6] I. Mani, et al., “Identifying unknown proper names in newswire text,” in *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, Columbus, Ohio, 1993, pp. 44–54.

[7] D. McDonald, “Internal and external evidence in the identification and semantic categorization of proper names,” in *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, Columbus, Ohio, 1993, pp. 32–43.

[8] W. Paik, et al., “Categorizing and standardizing proper nouns for efficient information retrieval,” in *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, Columbus, Ohio, 1993, pp. 154–160.

[9] G. R. Krupka, “SRA: Description of the SRA system as used for MUC–6,” in *Proceedings of Sixth Message Understanding Conference*, **location**, 1995, pp. 221–235.

[10] R. Weischedel, “BBN: Description of the PLUM system as used for MUC–6,” in *Proceedings of Sixth Message Understanding Conference*, **location**, 1995, pp. 55–69.

[11] L. F. Chien, “Fast and quasi–natural language search for gigabytes of Chinese texts,” in *Proceedings of 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, **location**, 1995, pp. 112–120.

[12] K. L. Kwok, “Comparing representations in Chinese information retrieval,” in *Proceedings of 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, **location**, 1997, pp. 34–41.

[13] H. H. Chen, S. J. Huang, Y. W. Ding, and S. C. Tsai, “Proper name translation in cross–language information retrieval,” in *Proceedings of 17th International Conference on Computational Linguistics*, **location**, 1998, **pp.** .

[14] J. S. Chang, et al., “Large–corpus–based methods for Chinese personal name recognition,” *Journal of Chinese Information Processing*, Vol. 6, No. 3, 1992, pp. 7–15.

[15] L. J. Wang, W. C. Li, and C. H. Chang, “Recognizing unregistered names for Mandarin word identification,” in *Proceedings of 14th International Conference on Computational Linguistics*, Nantes, 1992, pp. 1239–1243.

[16] R. Sproat, et al., “A stochastic finite–state word–segmentation algorithm for Chinese,” in *Proceedings of 32nd Annual Meeting of ACL*, New Mexico, 1994, pp. 66–73.

[17] H. H. Chen and J. C. Lee, “The identification of organization names in Chinese texts,” *Communication of Chinese and Oriental Languages Information Processing Society*, Vol. 4, No. 2, 1994, pp. 131–142.

[18] ROCLING, *ROCLING Text Corpus Exchange*, ROC Computational Linguistics Society, **location**, 1993.

[19] A. S. Hornby, *Oxford Advanced Learner's Dictionary of Current English*, **puiblisher**, **location**, 1984.

[20] Y. J. Huang, *English Names for You*, Learning Publishing Company, Taiwan, 1992.

[21] P. Fung and D. Wu, “Statistical augmentation of a Chinese machine–readable dictionary,” in *Proceedings of 2nd Workshop on Very Large Corpora*, **location**, 1994, pp. 69–85.

[22] M. C. Wang, K. J. Chen, and C. R. Huang, “The identification and classification of unknown words in Chinese: A N–Gram approach,” in *Proceedings of PAcFocol 2*, **location**, 1994, pp. 17–31.

[23] F. Smadja, “Retrieving collations from text: Xtract,” *Computational Linguistics*, Vol. 19, No. 1, 1993, pp. 143–177.

Hsin–Hsi Chen was born in Chiayi, Taiwan, R.O.C., on September 23, 1957. He received the B.S. and the M.S. degrees in computer science and information engineering in 1981 and 1983, respectively, and the Ph.D. degree in electric engineering in 1988, all from National Taiwan University, Taipei, Taiwan, R.O.C.

Since August 1995, he has been a Professor in Department of Computer Science and Information Engineering, National Taiwan University. His research interests are computational linguistics, Chinese language processing, information retrieval and extraction, Internet and database design.

He is member of ROCLING and ACL. He is a member of governing board of ROCLING and an editorial board of Communications of COLIPS.