Identification and Classification of Proper Nouns in Chinese Texts

Hsin-Hsi Chen and Jen-Chang Lee
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.
hh_chen@csie.ntu.edu.tw

Abstract

Various strategies are proposed to identify and classify three types of proper nouns in Chinese texts. Clues from character, sentence and paragraph levels are employed to resolve Chinese personal names. Character, Syllable and Frequency Conditions are presented to treat transliterated personal names. To deal with organization names, keywords, prefix, word association and parts-of-speech are applied. For fair evaluation, large scale test data are selected from six sections of a newspaper. The precision and the recall for these three types are (88.04%, 92.56%), (50.62%, 71.93%) and (61.79%, 54.50%), respectively. When the former two types are regarded as a category, the performance becomes (81.46%, 91.22%). Compared with other approaches, our approach has better performance and our classification is automatic.

1. Introduction

A Chinese sentence is composed of a string of characters without any word boundaries, so that to segment Chinese sentences is indispensable in Chinese language processing (Chen, 1990; Chen, 1994). Many word segmentation techniques (Chen & Liu, 1992; Chiang *et al.*, 1992; Sproat & Shih, 1990) have been developed. However, the resolution of unknown words, i.e., those words not in the dictionaries, form the bottleneck. Some papers (Fung & Wu, 1994; Wang *et al.*, 1994) based on Smadja's paradigm (1993) learned an aided dictionary from a corpus to reduce the possibility of unknown words. Chang *et al.* (1992) proposed a method to extract Chinese personal names from an 11,000-word corpus, and reported 91.87% precision and 80.67% recall. Wang *et al.* (1992) recognized unregistered names on the basis of titles and a surname-driven rule. Lin *et al.* (1993) presented a model to tackle a very restrictive form of unknown words. Sproat *et al.* (1994) considered Chinese personal names and transliterations of foreign words. Their performance was 61.83% precision and 80.99% recall on an 12,000-Chinese-character corpus.

This paper deals with three kinds of proper nouns - say, Chinese personal names, transliterated personal names and organization names. We not only tell if an unknown word is a proper noun, but also assign it a suitable semantic feature. In other words, '喬治布希' (George Bush) will have a feature of male transliterated personal name when it is identified. Such a rigid treatment will be helpful for further applications such as anaphora resolution (Chen, 1992), sentence alignment (Chen & Chen, 1994; Chen & Wu, 1995), *etc.* Section 2 describes the training corpora and the testing corpus we used. Sections 3, 4 and 5 propose the identification and classification methods of Chinese personal names, transliterated personal names and organization names, respectively. Section 6 presents two applications. Section 7 concludes the remarks.

2. Training Corpora and Testing Corpus

The proposed methods in this paper integrate the rule-based and the statistics-based models, so that training corpora are needed. To test the performance of language models, a good testing corpus is also necessary. This section introduces all the corpora that are used in the following sections.

NTU balanced corpus, which follows the standard of LOB corpus (Johansson, 1986), is the first training corpus. It is segmented by a word segmentation system and is checked manually. In total, this corpus has 113,647 words and 191,173 characters.

The second training corpus is extracted from three newspaper corpora (China Times, Liberty Times News and United Daily News). It is just segmented by a word segmentation system without checking manually. Although segmentation errors may exist, this corpus is 23.2 times larger than NTU balanced corpus, so that we can get more reliable word association pairs.

The third training corpus is a transliterated personal name corpus. There are 2,692 transliterated personal names, including 1,414 male's names and 1,278 female's names. Those transliterated personal names are selected from a book "English Names For You" (Huang, 1992). The last training data is a Chinese personal name corpus. It has 219,738 Chinese personal names and 661,512 characters.

Finally, the testing corpus is introduced. We randomly select six different sections from a newspaper corpus (Liberty Times News). The contents are different from the second training corpus. The following shows the statistics of the testing corpus:

(a) the political section
There are many items of news about the legislature. It includes 23,695 words and 36,059 characters.

(b) the social section
There are many items of news about police and offenders. It includes 61,846 words and 90,011 characters.

(c) the entertainment section
There are many items of news about TV stars, programs, and so on. It includes 38,234 words and 55,459 characters.

(d) the international section
It contains many items of foreign news and has 19,049 words and 29,331 characters.

(e) the economic section
Many items of news about stock market, money, and so on, are recorded. It includes 39,008 words and 54,124 characters.

(f) the sports section
All items of news concern sports. It includes 36,971 words and 54,124 characters.

Every section has its own characteristics. In the political section, there are many titles. In the social section and the entertainment section, there are many Chinese personal names and organization names. In the international section, transliterated personal names are more than the other two. In the economic section, stock companies often appear. In the sports section, there are many Chinese personal names and transliterated personal names. Because the proper nouns are usually segmented into single characters, they will interfere with one another during identification and classification.

3. Chinese Personal Names

3.1 Structure of Personal Names

Chinese personal 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 one character. Theoretically, every character can be considered as names rather than a fixed set. Thus the length of Chinese personal names ranges from 2 to 6 characters.

3.2 Strategies

3.2.1 Segmentation before Identification

Input text has to be segmented roughly beforehand. This is because many characters have high probabilities to be a Chinese personal 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 personal name. On the statistical model, this type of errors is difficult to avoid. However, it is easy to capture by pre-segmentation.

3.2.2 Variation of a Character

How to calculate the score of a candidate is an important issue in this identification system. The paper (Chang *et al.*, 1992) proposes the following formula:

(1) $P(W,GN) = P(GN) * P(W|GN)$

This formula has a drawback, i.e., it does not consider the probability of a character to be the other words rather than a surname. Take the two characters '傳' and '暉' as an example. The character '傳' can form '傳播', '傳染', '傳話', and many other words. On the contrary, the character '暉' just forms a word '暉映', which is a rare word. The difference shows that the former is easier to be used as the other words than the latter. The above formula assigns the same score to '傳至' and '暉至', when '傳' and '暉' have the same frequency to be names. Intuitively, '傳至' does not look like a name, but '暉至' does. Thus '暉' should have higher score than '傳', and the variation of a character should be considered in the formula. In our model, the variation of characters is learned from NTU balanced corpus.

3.2.3 Baseline Model

Equation (2) defines the original formula. The formula used to calculate P(C*i*) is similar to Equation (1). When the variation of a character is considered, Equation (3) is formulated. The variation of a character is measured by the inverse of the frequency of the character to be the other words. Equation (4) is simplified from Equation (3).

(2) $P(C1) \times P(C2) \times P(C3)$

(3) $P(C1) \times \frac{1}{\&C1} \times P(C2) \times \frac{1}{\&C2} \times P(C3) \times \frac{1}{\&C3}$

(4) $\frac{\#C1}{\&C1} \times \frac{\#C2}{\&C2} \times \frac{\#C3}{\&C3}$

where C*i* is a character in the input sentence,
P(C*i*) is the probability of C*i* to be a surname or a name,
#C*i* is the frequency of C*i* to be a surname or a name,

&C*i* is the frequency of C*i* to contain in the other words.

For different types of surnames, different models are adopted.

(a) Single character

$$(5)\ \frac{\#C1}{\&C1}\times\frac{\#C2}{\&C2}\times\frac{\#C3}{\&C3}>Threshold1$$

$$(6)\ \frac{\#C1}{\&C1}>Threshold2$$

$$\text{and}\ \frac{\#C2}{\&C2}\times\frac{\#C3}{\&C3}>Threshold3$$

(b) Two characters

$$(7)\ \frac{\#C2}{\&C2}\times\frac{\#C3}{\&C3}>Threshold4$$

(c) Two surnames together

$$(8)\ \frac{\#C11}{\&C11}\times\frac{\#C12}{\&C12}\times\frac{\#C2}{\&C2}\times\frac{\#C3}{\&C3}>Threshold5$$

$$(9)\ \frac{\#C11}{\&C11}\times\frac{\#C12}{\&C12}>Threshold6$$

$$\text{and}\ \frac{\#C2}{\&C2}\times\frac{\#C3}{\&C3}>Threshold7$$

Because the surnames with two characters are always surnames, Model (b) neglects the score of surname part. Models (a) and (c) have two score functions. It avoids the problem of very high score of surnames. Consider the string '陳踢了王一腳，王打了陳一拳'. Because of the high scores of the characters '陳' and '王', '陳踢了', '王一腳', '王打了' and '陳一拳' may be identified according to Equation (5). Equation (6) screens out the impossible candidates. The above three models can be extended to single-character names. When a candidate cannot pass the threshold, its last character is cut off and the remaining part is tried again. The threshold is different from the original one. Thresholds are trained from Chinese personal name corpus. We calculate the score of every Chinese personal name in the corpus using the above formulas. The scores for each formula are sorted and the one which is less than 99% of the personal names is considered as a threshold for this formula. That is, 99% of the training data can pass the threshold.

3.2.4 Other Clues

Text provides many useful clues from three different levels - say, character, sentence and paragraph levels. The baseline model forms the first level, i.e., character level. The following subsections present other clues. Of these, gender is also a clue from character level; title, mutual information and punctuation marks come from sentence level; the paragraph information is recorded in cache.

3.2.4.1 Clue 1: Title The first is title. Wang *et al.* (1992) propose a model based on titles. When a title appears before (after) a candidate, it is probably a personal name. For example, '李登輝總統' and '總統李登輝'. However, there are many counterexamples, e.g., '總統向青年學子...'. Thus we cannot make sure if the characters surrounding a title form a personal name. Even so, title is still a useful clue. It can help determine the boundary of a name. In the example '李鵬常到...', '李鵬常' is identified incorrectly. When a title is included in this example, i.e., '李鵬總理常到...', the error does not occur. In summary, if a title appears, a special bonus is given to the candidate.

3.2.4.2 Clue 2: Mutual Information Chinese personal names are not always composed of single characters. For example, the name part of the sentence '陳聰明醫術非常高明' is a word. How to tell out that a word is a content word or a name is indispensable. Mutual information (Church & Hanks, 1990) provides a measure of word association. The words surrounding a word candidate are checked. When there exists a strong relationship, the word candidate has high probability to be a content word. In the example '陳家世清白，絕不會犯法...', the two words '家世' and '清白' have high mutual information, so that '陳家世' is not a personal name. Three newspaper corpora (total size is about 2.6 million words) are used to train the word association.

3.2.4.3 Clue 3: Punctuation Marks Personal names usually appear at the head or the tail of a sentence. A candidate is given an extra bonus when it is found from these two places. Candidates surrounding the caesura mark, a Chinese-specific punctuation mark, are treated in the similar way. If some words around this punctuation are personal names, the others are given bonus.

3.2.4.4 Clue 4: Gender There is a special custom in Chinese. A married woman may mark her husband's surname before her surname. That forms type 3 personal name mentioned in Section 3.1. Because a surname may be considered as a name, e.g., '安' in the personal name '林安妮' and '文' in the personal name '曾文忠', the candidates with two possible surnames do not always belong to type 3 personal name. The gender information, i.e., type 3 is always a female, helps us disambiguate the type of personal names. Some Chinese characters have high score for male and some for female. The following lists some typical examples:

male: 豪, 霸, 宏, 志, 世, 斌, 彬, 強, 正, 昌, 光

female: 佩, 月, 玉, 如, 君, 秀, 佳, 怡, 芬, 芳, 女

We count the frequencies of the characters to be male and female, and compare these two scores. If the former is larger than the latter, then it is a masculine name. Otherwise, it is a feminine name.

3.2.4.5 Clue 5: Cache A personal name may appear more than once in a paragraph. This phenomenon is useful during identification. 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:

(a) C1C2C3 and C1C2C4 are in the cache, and C1C2 is correct.
(b) C1C2C3 and C1C2C4 are in the cache, and both are correct.
(c) C1C2C3 and C1C2 are in the cache, and C1C2C3 is correct.
(d) C1C2C3 and C1C2 are in the cache, and C1C2 is correct.

Here C*i* denotes a Chinese character. It is obvious that case (a) contradicts with case (b). Consider the string '李鵬常去打高爾夫球'. A personal name '李鵬常' is recognized. When another string '李鵬及鄧小平到廣州視察' is input, '李鵬及' and '鄧小平' are identified. Then we find the two strings '李鵬及' and '李鵬常' are similar. Here case (a) is correct. However, case (b) also appears very often in newspapers. For example, '邱永漢、邱永強兩兄弟...'. Two personal names, '邱永漢' and '邱永強' are identified. In the examples like '李煥說不清楚...' and '...李煥。', two candidates '李煥說' and '李煥' will be identified. That belongs to case (d). Consider the last examples '焦仁和表示...' and '焦仁和秘書長...'. Two candidates '焦仁' and '焦仁和' will be identified too. Now, case (c) is adopted.

In our treatment, a weight is assigned to each entry in the cache. The entry which has clear right boundary has a high weight. Title and punctuation are clues for boundary. For those similar pairs which 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.

3.3 Experiments and Discussions

Table 1 summarizes the identification results of Chinese personal names. Total Chinese personal names in each section are listed in Column 2. Column 3 shows the precision and the recall of the baseline model. The overall performance is good except for section 4 (the international section) and section 5 (the economic section). The remaining columns demonstrate the change of performance after the clues discussed in Section 3.2 are considered incrementally. If name part of a candidate is a word, word association is used to measure the relationship between the surrounding words. The increase of the precision in Column 4 verifies this idea. Theoretically, it should not decrease the recall. After checking the result, we find that some unreasonable word association comes from the training corpus. Recall that it is generated by a rough word segmentation system without manually-checking. The next clue is punctuation. The idea is that the candidates in the beginning or at the end of sentences have larger probabilities to be personal names than they are in other places. It helps some candidates with lower score to pass the threshold, but it cannot avoid the incorrect candidates to pass the threshold. Thus, the performance is dangling. Then, title is considered. The increase of the recall shows that title works well. But it decreases the precision too. From the variation of the performance, we know that cache is powerful. Both the recall and the precision increase. Finally, gender is joined. It is used when two successive characters are candidates of surnames. In other words, it focuses on type 3 personal names. Almost all type 3 personal names are identified correctly. Because this type of personal names is rare in the testing newspaper corpus, the variation is not large. Table 1 shows that our model is good except for section 4 and section 5. There are many proper nouns in the international section, and almost all of them are not included in the dictionary. All unknown words disturb one another in segmentation. For example, '立陶宛' is a country name. It is divided into three single characters by our word segmentation system. From the viewpoint of personal name identification, it is easy to regard '陶宛' as a Chinese personal name. Another source of errors is foreign names. Some of them are similar to Chinese personal names, e.g., '魏斯特' and '艾琳達'.

Table 1. Identification Results of Chinese Personal Names

	Total	Baseline Model		+ Word Association		+Punctuation		+Title		+Cache		+Gender	
	Names	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
section 1	641	90.54%	91.11%	90.78%	90.64%	89.72%	89.86%	88.84%	90.64%	91.02%	93.29%	91.32%	93.60%
section 2	1628	86.66%	93.74%	86.94%	93.67%	86.76%	93.80%	86.08%	93.86%	93.81%	93.98%	93.99%	94.16%
section 3	666	83.90%	82.13%	83.99%	79.58%	84.01%	81.23%	83.84%	82.58%	86.41%	84.99%	86.26%	84.83%
section 4	148	54.22%	91.22%	55.14%	90.54%	55.14%	90.54%	55.24%	92.57%	64.09%	95.27%	64.09%	95.27%
section 5	176	73.46%	88.07%	74.40%	87.50%	73.91%	86.93%	73.46%	88.07%	74.18%	89.77%	74.18%	89.77%
section 6	694	83.87%	93.66%	84.09%	93.66%	83.83%	94.09%	82.85%	94.67%	84.87%	95.39%	84.87%	95.39%
Total	3953	83.79%	90.99%	84.13%	90.41%	83.84%	90.67%	83.19%	91.27%	87.94%	92.46%	88.04%	92.56%

The similar problem occurs in the economic section. There are many company names, and some of them are similar to Chinese personal names. The company name '龍祥' is a typical example. In summary, there are three major errors. One is foreign name. They are identified as proper nouns correctly, but are assigned wrong features. About 20% of errors belong to this type. The second type of errors results from the rare surnames, which are not included in the surname table. Some rare surnames are not real surnames. They are just artists' stage names. Near 14% of errors come from this type. The other errors include place names, organization names, and so on.

4. Transliterated Personal Names

4.1 Structure of Personal Names

Compared with the identification of Chinese personal names, the identification of transliterated personal names has the following difficulties:

(a) No specific clue like surnames in Chinese personal names to trigger the identification system.

(b) No restriction on the length of a transliterated personal name. It may be composed of a single character or more, e.g., '強', '肯尼', '史帝芬', '金貝辛格' and '布魯司威利'.

(c) No large scale transliterated personal name corpus.

(d) Ambiguity in classification.

For example, '華盛頓' may denote a city or a former American president.

4.2 Strategies

4.2.1 Basic Idea

Almost all foreign names are in transliteration, not in translation. And the base of transliteration is pronunciation of foreign names. Pronunciation is composed of syllables and tones. The major difference of pronunciation between Chinese and English is syllables. The style of syllabic order is specific in transliteration. Consider an example. The transliterated personal name '史帝芬席格' has syllables 'ㄕ ㄉㄧˋ ㄈㄣ ㄒㄧˊ ㄍㄜ'. Such a syllabic order is rare in Chinese, but is not special for a transliterated string. In other words, the syllabic orders of transliterated strings and general Chinese strings are not similar. Besides, a transliterated name consists of a string of single characters after segmentation. That is, these characters cannot be put together. However, the unrestrictive length of transliterated names and homophones in Chinese result in the need of very large training corpus. The following sections show how to modify the basic idea if a large scale corpus is not available.

4.2.2 Character Condition

When a foreign name is transliterated, the selection of homophones is restrictive. Consider an example shown below:

Richard Macs　理查馬克斯　娌茶螞剋鷥

Those strings following English names have the same pronunciations. The first is usually adopted, and the second is never used. It shows that the characters used in transliteration are selected from some character set. In our model, total 483 characters are trained from our transliterated personal name corpus. They play the similar role of the surnames in the identification of Chinese personal names. If all the characters in a string belong to this set, i.e., they satisfy *character condition*, they are regarded as a candidate.

4.2.3 Syllable Condition

Because of the unrestrictive length of transliterated names, how to identify their boundary is a problem. Of course, titles and punctuation used in last section can be adopted too. But they do not always appear in the text. Thus another clue should be found. Syllable order may be a clue. Those examples like '各國', '令人' and '二發' which meet the character condition do not look like transliterated names because their pronunciations are not like foreign names. If there is a large enough transliterated name corpus, the syllable orders can be learned. However, our transliterated corpus only contain 2692 personal names. Thus only the first and the last characters are considered. For each candidate, we check the syllable of the first (the last) character. If the syllable does not belong to the training corpus, the character is deleted. The remaining characters are treated in the similar way.

4.2.4 Frequency Condition

As mentioned in Section 3.2.3, the frequency of a character to be a part of a personal name is important information. The concept may be used here. However, only large scale transliterated personal name corpus can give reliable statistical data. Based on our small training corpus, the range of the application of the information should be narrowed down. We only apply it in a candidate of length 2. This is because it is easy to satisfy the character condition for candidates of the shortest length. For each candidate which has only two characters, we compute the frequency of these two characters to see if it is larger than a threshold. If it is not, it is eliminated. The threshold is determined in the similar way as Section 3.2.3.

4.3 Experiments and Discussions

The identification system scans a segmented sentence from left to right. It finds the character string that meets the character condition, syllable condition and frequency condition. Table 2 shows

Table 2. Identification Results of Transliterated Personal Names

	Total Names	System	Correct	Error	Lose	Precision	Recall
Section 1	52	64	34	30	18	53.13%	65.38%
Section 2	9	88	6	82	3	6.82%	66.67%
Section 3	238	300	180	120	58	60.00%	75.63%
Section 4	301	301	230	71	71	76.41%	76.41%
Section 5	34	152	26	126	8	17.11%	76.47%
Section 6	214	300	134	166	80	44.67%	62.62%
Total	848	1205	610	595	238	50.62%	71.93%

the precision and the recall are both good for sections 3 and 4, i.e., the entertainment and the international sections. However, sections 2 and 5 (the social and the economic sections) have bad precision. The average recall tells us that the trigger to the identification system is useful. The reasons why the recall is not good enough are: some transliterated personal names (e.g., '魏斯特' and '艾琳達') look like Chinese personal names, and the identification of Chinese personal names is done before that of transliterated personal names. Although they are correctly identified as personal names, they are assigned wrong features. Similarly, transliterated nouns like popular brands of automobiles ('飛雅特' and '雪佛蘭'), Chinese proper nouns ('利多', '連拉' and '華隆') and Chinese personal names ('朱士列') look like transliterated personal names. That decreases the precision. Besides these types of nouns, boundary errors affect the precision too. For telling out the error rates from classification, we made another experiment. If the identified results are not classified, the average precision is 81.46% and the average recall is 91.22%.

5. Organization Names

5.1 Structures of Organization Names

Structures of organization names are more complex than those of personal names. Some organization names are composed of proper nouns and content words. For example, '台北市政府' is made up of the place name '台北市' and the content word '政府'. A personal name can also be combined a content word to form an organization name, e.g., '黃東和內科診所'. Some organization names look like personal names, e.g., '龍祥'. Some organization names are composed of several related words . For example, '海峽兩岸交流基金會' contains four words '海峽', '兩岸', '交流' and '基金會'. Several single-character words can also form an organization name, e.g., '開康證券'. Some organization names have nested structures. Consider the string: '美國眾議院 外交委員會 亞太小組'. The group '亞太小組' is a part of the committee '外交委員會', and the committee itself is a part of '美國眾議院'. Such complex structures make identification of organization names very difficult.

Basically, a complete organization name can be divided into two parts: name and keyword. In the example '台北市政府', '台北市' is a name, and '政府' is a keyword. Many words can serve as names, but only some fixed words can be regarded as keywords. Thus, keyword is an important clue to identify the organizations. However, there are still several difficult problems. First, keyword is usually a common content word. It is not easy to tell out a keyword and a content word. Second, a keyword may appear in the abbreviated form. For example, '投顧' is an incomplete keyword of '投資顧問公司'. Third, the keyword may be omitted completely. For example, '宏碁' (Acer). The following shows two rough classifications, and discusses their features.

(1) Complete organization names

(a) Structure: This type of organization names is usually composed of proper nouns and keywords.

(b) Length: Some organization names are very long, so it is hard to decide their length. Fortunately, only some keyword like '聯盟會', '協會', '基金會', '組織', and so on, have this problem.

(c) Ambiguity: Some organization names with keywords are still ambiguous. For example, '天下雜誌' and '聯合報'. They usually denote reading matters, but not organizations. However, if they are used in some contexts, e.g., "天下雜誌 總經理" and "聯合報 董事長", they should be interpreted as organizations.

(2) Incomplete organization names

(a) Structure: These organization names often omit their keywords.

(b) Ambiguity: The abbreviated organization names may be ambiguous. For example, '兄弟', '熱火', '金磚' and '公牛' are famous sport teams in Taiwan or in U.S.A., however, they are also general content words.

5.2 Strategies

This section introduces some strategies used in the identification. Keyword is a good indicator for an identification system. It plays the similar role of surnames. Keyword shows not only the possibility

Table 3. Identification Results of Organization Names

	Total Names	System	Correct	Error	Lose	Precision	Recall
Section 1	596	512	394	118	202	76.95%	66.11%
Section 2	650	749	414	335	236	55.27%	63.69%
Section 3	703	601	391	210	312	65.06%	55.62%
Section 4	207	207	153	54	54	73.91%	73.91%
Section 5	347	366	150	216	197	40.98%	43.23%
Section 6	1064	711	442	269	622	62.17%	41.54%
Total	3567	3146	1944	1202	1623	61.79%	54.50%

of an occurrence of an organization name, but also its right boundary. For each sentence, we scan it from left to right to find keywords. Because keyword is a general content word, we need other strategies to tell out its exact meaning. These strategies also have the capabilities to detect the left boundary if there is an organization name.

Prefix is a good marker for possible left boundary. For example, '國立' (National), '省立' (Provincial), '私立' (Private), and so on. The name part of an organization may be formed by single characters or words. These two cases are discussed as follows.

(a) single characters

After segmentation, there may be a sequence of single characters preceding a possible keyword. The character may exist independently. That is, it is a single-character word. In this case, the content word is not a keyword, so that no organization name is found. If these characters cannot exist independently, they form the name part of an organization. The left boundary of the organization is determined by the following rule:

We insert a single character to the name part until a word is met.

(b) word(s)

Here, a word is composed of at least two characters. If the word preceding the possible keyword is a place name or a personal name, then the word forms the name part of an organization. Otherwise, we use word association model to determine the left boundary. The postulation is: the words to compose a name part usually have strong relationships. The mutual information mentioned in Section 3.2.4.2 is also used to measure the relationship of two words.

Part of speech is useful to determine the left boundary of an organization. The categories of verbs are very typical. The name part of an organization cannot extend beyond a transitive verb. If a transitive verb precedes a possible keyword, then no organization name is found. Numeral and classifier are also helpful. For example, '公司' (company) in '三家公司...' (three companies ...) is not a keyword due to the critical parts of speech. Because a tagger is not involved before identification, the part of speech of a word is determined wholly by lexical probability.

5.3 Experiments and Discussions

Table 3 shows the precision and the recall for every section. Section 4 (The International Section) has better precision and recall than other files. Most errors result from organization names without keywords, e.g., '金匯通', '復華', '大公投顧', '兄弟', and so on. Even keywords appear, e.g., '上市公司' and '各國大學', there may not always exist organization names. Besides error candidates and organization names without keywords, error left boundary is also a problem. Consider the examples: '不爲國安局' and '長老教會'. In the first, '不爲' should not be included; and in the second, a word '基督' is lost.

6. Applications

The semantic classification of proper nouns is useful in many applications. Here, anaphora resolution and sentence alignment are presented. In general, pronoun often refers to the nearest proper noun (Chen, 1992). But it is not always true. The following shows a counter example:

"問華德教授，他說那是正常的師生戀，既然雙方都是獨身男女，總不會不准談戀愛吧。至於後來趙靜雯去了那裡，爲甚麼失蹤，他一概不知，並輕描淡寫的說：「也許加拿大不適合她，跑回臺灣去了。」

The first pronoun '他' (he) refers to the personal name '華德'. It is a normal example. The second pronoun '他' (he) refers to the same person, but the nearest personal name is '趙靜雯' rather than '華德'. If we know the gender of every personal name, then it is easy to tell out which person is referred. In the above example, the gender of the Chinese pronouns '他' (he) and '她' (she) is masculine and feminine, respectively; the persons '華德' and '趙靜雯' are male and female, respectively. Therefore, the correct referential relationships can be well-established. In the experiment of the gender assignment, 3/4 of Chinese personal name corpus is regarded as training data, and the remaining 1/4 is for testing. The correct rate is 89%. Sentence alignment (Chen & Chen, 1994) is important in

setup of a bilingual corpus. Personal name is one of important clues. Its use in aligning English-Chinese text is shown in the paper (Chen & Wu, 1995).

7. Concluding Remarks

This paper proposes various strategies to identify and classify Chinese proper nouns. The performance evaluation criterion is very strict. Not only are the proper nouns identified, but also suitable features are assigned. The performance (precision, recall) for the identification of Chinese personal names, transliterated personal names and organization names is (88.04%, 92.56%), (50.62%, 71.93%) and (61.79%, 54.50%), respectively. When the criterion is loosed a little, i.e., Chinese personal names and transliterated personal names are regarded as a category, the performance is (81.46%, 91.22%). Compared with the approaches (Sproat *et al.*, 1994; Fung & Wu, 1994; Wang *et al.*, 1994), we deal with more types of proper nouns and we have better performance.

Some difficult problems should be tackled in the future. Foreign proper nouns may be transformed in part by transliteration and translation. The example "George Town" is transformed into '喬治城'. The character '城' (town) results in translation and '喬治' (George) comes from transliteration. This problem is interesting and worthy of resolving. The performance of identification of organization names is not good enough, especially for those organization names without keywords. It should be investigated further.

Acknowledgments

The research was supported in part by National Science Council, Taipei, Taiwan, R.O.C. under contract NSC83-0408-E002-019. We are also thankful for the anonymous referees' comments.

References

Chang, J.S.; *et al.* (1992) "Large-Corpus-Based Methods for Chinese Personal Name Recognition," *Journal of Chinese Information Processing*, Vol. 6, No. 3, pp. 7-15.

Chen, H.H. (1990) "A Logic-Based Government-Binding Parser," *Proceedings of 13th COLING*, Vol. 2, pp. 48-53.

Chen, H.H. (1992) "The Transfer of Anaphors in Translation," *Literal and Linguistic Computing*, Vol. 7, No. 4, pp. 231-238.

Chen, H.H. (1994) "The Contextual Analysis of Chinese Sentences with Punctuation Marks," *Literal and Linguistic Computing*, Vol. 9, No. 4, pp. 281-289.

Chen, K.H. and Chen, H.H. (1994) "A Part-of-Speech-Based Alignment Algorithm," *Proceedings of 15th COLING*, pp. 166-171.

Chen, K.J. and Liu, S.H. (1992) "Word Identification for Mandarin Chinese Sentences," *Proceedings of 14th COLING*, pp. 101-107.

Chen, H.H. and Wu, Y.Y. (1995) "Aligning Parallel Chinese Texts Using Multiple Clues," *Proceedings of 2nd PACLING*, pp. 29-48.

Chiang, T.H., *et al.* (1992) "Statistical Models for Word Segmentation and Unknown Word Resolution," *Proceedings of 5th ROCLING*, pp. 121-146.

Fung, P. and Wu, D. (1994) "Statistical Augmentation of a Chinese Machine-Readable Dictionary," *Proceedings of 2nd WVLC*, pp. 69-85.

Johansson, S. (1986) *The Tagged LOB Corpus: User's Manual*, Norwegian Computing Centre for the Humanities, Bergen.

Huang, Y.J. (1992) *English Names for You*, Learning Publish Company, Taiwan.

Lin, M.Y.; Chiang, T.H. and Su, K.Y. (1993) "A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," *Proceedings of 6th ROCLING*, Taiwan, pp. 119-141.

Smadja, F. (1993) "Retrieving Collations from Text: Xtract," *Computational Linguistics*, Vol. 19, No. 1, pp. 143-177.

Sproat, R. and Shih, C. (1990) "A Statistical Method for Finding Word Boundaries in Chinese Text," *Computer Processing of Chinese and Oriental Languages*, Vol. 4, No. 4, pp. 316-351.

Sproat, R.; *et al.* (1994) "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Proceedings of 32nd Annual Meeting of ACL*, New Mexico, pp. 66-73.

Wang, L.J.; Li, W.C. and Chang, C.H. (1992) "Recognizing Unregistered Names for Mandarin Word Identification," *Proceedings of 14th COLING*, Nantes, pp. 1239-1243.

Wang, M.C.; Chen, K.J. and Huang, C.R. (1994) " The Identification and Classification of Unknown Words in Chinese: A N-Gram Approach," *Proceedings of PACFocol 2*, pp. 17-31.