Recognizing Unregistered Names for Mandarin Word Identification

Liang-Jyh Wang, Wei-Chuan Li, and Chao-Huang Chang
Computer and Communication Research Laboratories (CCL)
Industrial Technology Research Institute (ITRI)
Hsinchu, Taiwan, R.O.C.
E-mail: changch%e0sun3.ccl.itri.org.tw@cunyvm.bitnet

Abstract

Word Identification has been an important and active issue in Chinese Natural Language Processing. In this paper, a new mechanism, based on the concept of sublanguage, is proposed for identifying unknown words, especially personal names, in Chinese newspapers. The proposed mechanism includes title-driven name recognition, adaptive dynamic word formation, identification of 2-character and 3-character Chinese names without title. We will show the experimental results for two corpora and compare them with the results by the NTHU's statistic-based system, the only system that we know has attacked the same problem. The experimental results have shown significant improvements over the WI systems without the name identification capability.

1 Introduction

Word Identification (WI, also known as Segmentation) has been an important and active issue in Chinese Natural Language Processing. Various approaches are proposed for this problem [1], such as MM (Maximum Matching) method [8], RMM (Reverse Directional Maximum Matching) method, OM (Optimum Matching) method, statistical approaches [5], and unification approaches [12]. However, there are still a number of problems to conquer towards a satisfactory WI system. Among them are a clear definition of Chinese words, an objective evaluation suite with appropriate corpora, and the processing of unknown words (such as personal names, place names, and organization names).

In this paper, we will deal with the problem of unknown words, especially personal names, although the proposed approach can be easily extended to cover place names and organization names. According to Chang, *et al.* [2], proper nouns (which compose a major part of unknown words) account for more than fifty percent of errors made by a typical system. Thus, successful processing of proper nouns is essential for a satisfactory WI system.

Almost all WI systems use a lexicon to guide the segmentation process. In fixed domains such as a classical novel or technical texts, we can put all possible words in the lexicon and avoid the unknown-word problem. However, in a dynamic domain such as newspapers, it is impossible to enumerate all possible words in advance. For example, some personal names, such as suspects or victims, often appear in only one day's news. Thus, recognition of these personal names and other unknown words is very important.

Chang, *et al.* [2] (at National Tsing-Hua University, Hsinchu, Taiwan) proposed a Multiple-Corpus approach to solve the problem. They consider the WI problem as a constraint satisfaction problem (CSP) and use a number of corpora to train their statistic-based system. The probabilities of each Chinese character as a surname, the first character and the second character in a first name are computed based on the training. Using these statistics, two-character and three-character personal names are proposed to compete with the words in the lexicon. Then, a dynamic programming technique is used to decide the most probable solution to the CSP. They reported a 90 percent average correct rate of surname-name identification. To the best of our knowledge, this is the only group that has proposed a solution to the problem.

Chang's approach is completely statistic-based and easy-to-implement. However, we argue that syntactic and semantic information must be considered in a successful WI system.

2 A Sublanguage Approach

The concept of sublanguages (i.e., languages in restricted domains) has been considered very important in natural language processing [6, 7]. A sublanguage usually has its own special syntax, semantics, and style, which are more restricted comparing with the language as a whole. In this paper, we will show how the study of a sublanguage can help identifying names and forming them in a dynamic, adaptive way.

2.1 Observation

From the United News, one of the most popular daily newspapers in Taiwan, we have acquired a newspaper corpus of more than one million characters. This corpus has been used for building our lexicon, computing statistics, and testing our WI systems for spell-checking, preprocessing for speech synthesis, and phoneme-to-word conversion.

After studying the segmentation output of the newspaper corpus, we observed that (1) unknown words are mostly personal names (translation names or otherwise), place names, and organization names in addition to those words that should have been built in the lexicon (a similar conclusion was obtained by Chang's papers); and (2) *when a personal name appears the first time, it is usually accompanied with a title (such as* taibei shizhang 台北市長 *Taipei mayor) or a role noun (such as* jizhe 記者 *reporter,* houxianren 候選人 *candidate).*

From these observations, we propose the following mechanisms to help identifying unknown words in the WI process: (1) title-driven name recognition and (2) adaptive dynamic word formation.

2.2 Title-driven Name Recognition

As we mentioned above, it is not plausible to put all proper names in the lexicon for a dynamic domain such as news articles. Since a new personal name usually appears with a title or a role noun, we can use the clue to design a set of word formation rules in our parsing-based WI system [11] (see the next section). Part of the set of rules in augmented CFG format are :

```
<name>   ← <title> <last> <first>
           { Build <last> <first> as a name }
<name>   ← <last> <first> <title>
           { Build <last> <first> as a name }

<title>  ← <word>
           { Test if <word> is a title }
<last>   ← <word>
           { Test if <word> is a surname }
<first>  ← <word>
           { Test if <word> is 1- or 2-char }
<first>  ← <word> <word>
           { Test if both <word> are 1-char }
```

A Chinese name usually consists of two to four characters: one- or two-character surname and one- or two-character first name. Furthermore, surnames are among a limited set. Thus, in rule 4, the augmented part is just a membership test. We can store the surname information as a feature in the lexical entries. Similarly, we have *title* and *role* features in the lexicon for rule 3. Note that in the current design, translation names of foreigners and husband surname prefixing of married women can not be correctly identified. However, this approach works for common personal names that occupy a major part of unknown words.

2.3 Adaptive Dynamic Word Formation

After a new personal name is recognized through the set of rules described above, the system will dynamically build a lexical entry for it. Thus, if the name appears in later sentences in the news article, it can be correctly identified.

In Figure 1 is an example for adaptive dynamic word formation. In the article, there are four Chinese names: ni2 shu2 yan2 倪淑媛 (4 instances), ye4 ying1 hao2 葉英豪 (1 instance), cai4 jia1 ting2 蔡佳婷 (4 instances), and wu2 xun2 long2 吳巡龍 (1 instance). In first instances, all four names come with a title: lao3shi1 老師 (teacher), ji4zhe3 記者 (reporter), er2tong2 兒童 (child), and jian3cha2guan1 檢察官 (prosecutor). Since the names are built in the lexicon dynamically, the other instances of the names can be identified with higher scores than names without title. In other words, the names with title are built with much more confidence.

2.4 Names without Title

In addition to the names with title or role, the other personal names are proposed through a surname-driven rule. In other words, when the WI system meets a surname word, a personal name proposing rule is invoked although its preference score would be much lower than regular words and names with title.

2.5 Place Names and Organization Names

The proposed mechanism can be extended to cover place names and organization names. Just like personal names appear with title, place names can be identified through the unit such as xian 縣 (county), shi 市 (city), jie 街 (street), lu 路 (road), etc. Similarly, organization names can be identified by the type such as gongsi 公司 (company), bu 部 (department or ministry), ke 課 (section), and so on. This part has not yet implemented in our system.

3 The System

Since July 1986, we have been involved in developing a series of Chinese-related NLP systems, including an English-Chinese MT system, a Japanese-Chinese MT, a Chinese Word Knowledge Base, a Chinese Parser, and a Chinese Spell-Checker. Here, we will only briefly describe the Chinese WI system as a frontend for the Chinese Parser. For more details, the reader is referred to Wang, *et al.* [11].

We consider the WI process as a parsing process with word composition grammar, instead of a CSP problem [2], a unification problem [12], or a scanning process. A set of Chinese word composition grammar rules are designed to capture the characteristics of Chinese words. The grammar representation is Augmented CFG which is also used to write the English grammar in our English-Chinese MT system. The parser we used is based on Tomita's Generalized LR Parser [10]. However, the augmented parts (tests and actions) and preference scoring module have been added.

智X智障兒不愼跌落漁港致死案　老師倪淑媛　提公訴啓智中心記者　葉英豪／報導
馬公市惠氏啓智中心老師倪淑媛，因疏於注意智障兒童蔡佳婷行蹤造成偷跑出校外不愼跌落漁港內致死案，昨日經澎湖地檢署偵查終結後依過失致人於死提起公訴。起訴書中指出，馬公市惠氏啓智中心特教三班老師倪淑媛（十九歲），因在四日上午欲帶兒童們參加升旗典禮時，未顧及曾有偷跑紀錄的智障兒童蔡佳婷，逕自帶其他兒童參加升旗而未託付其他老師代爲照顧，使得蔡佳婷跑出校外迷路，在當天下午不愼跌落馬公第二漁港內窒息死亡。檢察官吳巡龍在偵查中並了解，倪淑媛在案發後發現蔡佳婷失蹤後即四處尋找，而平日素行良好無前科，請院方能衡情參酌從輕量刑。

Figure 1: An Example for Adaptive Dynamic Word Formation

In the WI process, the basic unit is a character. A Chinese word is composed of one to five (may be longer) characters.

The WI system consists of a lexicon, the word composition grammar, the preference scoring module, the test functions, and the parser.

The lexicon contains a list of Chinese words (sorted by the internal code order) with the following information: the characters from which the word is composed, its frequency count, its part of speech, and some semantic features (such as title, surname, and role). The lexicon is a general purpose one; that is, it is built independent of the testing corpora. Currently, there are more than 90,000 lexical entries in the lexicon.

A rule in the word grammar consists of a context-free part and an augmented part. In addition to the unknown word identification described in the previous section, augmented parts are used for recognizing (1) replication of words; (2) numbers; (3) prefixes; (4) suffixes; and (5) the determiner measure constructions.

Since the word parser would produce two or more parses for an ambiguous sentence, a preference scoring module has been designed to choose the correct parse. Currently, the preference score is assigned based on (1) the length of the word (longer words are preferred), (2) the frequency count, and (3) semantic consideration (e.g., three-character personal names are preferred to two-character ones). The WI system is written in Common Lisp, running on a TI Micro-Explorer machine.

4 Experimental Results

Before we present the experimental results, two performance indices, *recall rate* and *precision rate*, of a WI system are defined below following Sproat and Shih [9] and Chang, *et al.* [3]. Let C be the segmentation results by the computer, H the results by the human (the correct results), and I the intersection of C and H. Then, recall rate is I divided by H, and precision rate I divided by C. For example, if there are 20 words in a sentence (i.e., H equals 20), the WI system produces 22 words for the sentence (i.e., C equals 22), and there are 18 words in common (i.e., I equals 18), the recall rate would be 0.90 and the precision rate 0.82.

To demonstrate the proposed mechanism, we have tested the WI system with two corpora: (1) ten articles from a newspaper corpus, the United Daily corpus, (2) 61 sentences from Chang et al. [4]. The first corpus is selected from the United Daily on March 8, 1991. The selection criterion is that the article does not contain any table or figure and, preferably, contains Chinese names. The second corpus is composed of difficult cases for which the NTHU WI system either can not identify the names or overgenerates some Chinese names.

In the experiment, we use four versions of the WI system to segment the ten articles. Version 1 is the WI system without name recognition capability, Version 2 the system recognizing only names with title, Version 3 the system recognizing both names with title and 3-character names, and Version 4 also recognizing 2-character names.

Recall rates (RR) and precision rates (PR) are computed automatically by comparing the segmentation output with the correct answers segmented by human. The experimental results are summarized in Table 1.

From the table, we can observe the following facts:

1. Version 2 (RR:96.17, PR:93.46) has a significant improvement over Version 1 (RR:94.77, PR:89.28). In other words, the capability for name recognition is very important in a WI system. Although Version 2 only has a limited capability (for names with title), the improvement is rather apparent. Note that in Version 2, the dynamic word formation mechanism is much more useful than in Version 3 or 4.

2. Version 3 has the best results (RR:97.51, PR:98.19) among the four versions. It is better than Version 2 for the obvious reason: the capability for identifying 3-character names without title.

3. Although Version 4 has one more function, identification of 2-character names without title, than Version 3, the result (RR:96.32, PR:97.51) is slightly worse than Version 3. This is mainly

set	#words	Version 1		Version 2		Version 3		Version 4	
		RR1	PR1	RR2	PR2	RR3	PR3	RR4	PR4
x1	317	97.79	95.38	98.42	96.59	97.48	96.26	95.58	94.39
x5	46	89.13	82.00	91.30	95.45	91.30	95.45	91.30	95.45
x6	168	93.45	85.33	99.40	98.82	99.40	98.82	98.21	98.21
x7	415	95.19	91.45	95.19	91.45	96.14	97.32	93.98	95.63
x8	373	98.66	96.59	99.20	97.63	98.12	98.92	97.32	98.64
x17	279	93.17	84.92	93.88	87.88	98.57	98.57	94.62	97.46
x25	343	92.13	84.72	92.13	84.72	97.95	98.82	97.37	98.52
x26	260	98.85	96.62	99.62	98.85	100.00	100.00	99.23	99.61
x27	311	91.64	80.97	92.60	83.24	96.14	97.71	94.53	96.71
x38	216	97.70	94.80	100.00	100.00	100.00	100.00	99.23	99.62
Total	2,728	94.77	89.28	96.17	93.46	97.51	98.19	96.32	97.51

Table 1: Experimental results for the first corpus

because the gain (recognition of 2-character names) is less than the loss (misintepreting 2 single-character words as a 2-character name).

4. We will analyze the imperfections by the WI system in a subsection after the comparison with NTHU's system.

Comparison with NTHU's System

In Chang, *et al.* [4], which we will call NTHU's system, they reported a 95 percent precision rate and a recall rate greater than 95 percent, and listed 5 samples (A-samples) the name in which their system can identify correctly, 34 examples (B-samples) for which the names are missed, and 22 examples (C-samples) for which Chinese names are over-generated. Among them, we found 3 A-samples, 6 B-samples, and 3 C-samples contain personal names with title. Since NTHU's system is completely statistic-based, it can not make use of the title information. On the other hand, our sublanguage-based system would process these samples correctly.

These 61 examples are fed to our WI system for comparison of the name recognition algorithms. The following results are for reference only, since the comparison is rather unfair (the examples are mostly the cases their system can not recognize correctly).

1. For the 5 A-samples, our system can recognize four of them. The only A-sample it failed to identify is: huang2 rong2 you2 you2 de0 dao4 黃蓉幽幽的道 . Our segmentation result is huang2-rong2-you2 you2 de0 dao4, while the correct result is huang2-rong2 you2-you2 de0 dao4. The reason is (1) our lexicon does not have the adverb you2-you2, and (2) we prefer 3-character names over 2-character ones. Note that NTHU's system can process all 5 cases successfully.

2. For the 34 B-samples, our system can identify 25 of them correctly. That is, there are 9 B-samples the names in which both our system and NTHU's system can not identify. We will discuss the reasons why these cases can not be recognized in the next subsection.

3. For the 22 C-samples for which NTHU's system overgenerates personal names, our system has processed 16 of them correctly. We will discuss the reasons in the next section why our system also overgenerates personal names for the other 6 C-samples.

4. For these 61 samples, our system can process 45 of them correctly.

Some Imperfections

There are still some problems remained unsolved in our WI system. Some are problems for WI systems in general. The others are specific to name recognition systems only.

1. Two-character names are difficult to recognize, especially when followed by a single-character word. For example, in yi1 jing4 gang1 ba3 fa3 bao3 qu3 chu1 易靜剛把法寶取出 , yi1-jing4 is a 2-character name. However, our WI system produces a 3-character name yi1-jing4-gang1, since gang1 (just) is a single character word. Although human usually can identify the names correctly by context, our WI system proposed the 3-character names understandably.

2. The name of a married woman is usually prefixed with her husband's surname. Thus, a 3-character name would become 4-character, i.e., husband's surname, father's surname, and a 2-character given name, e.g., xu3 lin2 yan2 mei2 許林艷梅 . Currently, this kind of names cannot be identified correctly, although a word-grammar rule can be easily added.

3. Some single-character surnames, such as lian2 年 (year), tang1 湯 (soup), ceng2 曾 (once), and huang2 黃 (yellow), are common single-character words. Thus, the name recognition algorithm sometimes overgenerates a personal name by

combining one such word with two following characters.

4. Some surnames are rather unusual, such as lian 蓮 (lotus), ping2 萍 (duckweed), and que4 卻 (but). This would make the names not recognizable. There is a tradeoff between a complete surname list and a minimal common surname list. On the one end, a complete surname list would help name recognition but it helps overgeneration as well. On the other end, a minimal list would limit the overgeneration while missing some would-be names.

5. Some single-character words are very difficult to identify when they can be grouped as two-character words with the characters in the neighbour. A famous example is ba3 shou3 把手 (a handle). The problem is very difficult to solve for any WI systems.

6. Even when the title information is used, overgeneration of personal names is still hard to avoid. In the following is one of such examples:

   - yao1 qing3 tai2 bei3 di4 fang1 fa3 yuan4 zhang1 lu3 xue2 jian3 cha3 guan1 tan2 yao4 wu4 lan4 yong4 wen4 ti2.

   邀請臺北地方法院張履學檢察官談藥物濫用問題

   Both the correct name zhang1-lu3-xue2 張履學 and an overgenerated name tan2-yao4-wu4 談藥物 are produced by our system. A fine adjustment of the scoring function should be able to overcome this problem. However, there are so many similar problems such that it would be a real problem when we develop a full-scale system.

7. In Version 4 of our system, 2-character names without title are recognized in addition to those of Version 3, i.e., names with title and 3-character names without title. However, both the recall rate and precision rate of Version 4 are lower than those of Version 3. The major reason is that too many 2-character names are generated.

5 Conclusion

In this paper, we have proposed a new mechanism for identifying unknown words, especially personal names, in Chinese newspapers. The proposed mechanism includes title-driven name recognition, adaptive dynamic word formation, identification of 2-character and 3-character Chinese names without title. We have also shown the experimental results for two corpora and have compared them with the results by the NTHU's WI system.

Although there are still some problems remained unsolved (as discussed above), the experimental results have shown significant improvements over the WI systems without the name identification capability.

Acknowledgement

This paper is a partial result of the project No. 33H3100 conducted by the Industrial Technology Research Institute, Taiwan, under the sponsorship of the Ministry of Economic Affairs, R.O.C.

References

[1] ACCC. *The Status and Progress of Chinese Language Processing Technology.* Association for Common Chinese Code, International, Beijing, China, 1991.

[2] J.-S. Chang, S.-D. Chen, Y. Chen, J. S. Liu, and S.-J. Ker. A Multiple-corpus Approach to Identification of Chinese Surname-names. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pages 87–91, 1991.

[3] J.-S. Chang, C.-D. Chen and S.-D. Chang. Chinese word segmentation through constraint satisfaction and statistical optimization. In *Proc. of ROCLING IV*, pages 147–165, 1991.

[4] 張俊盛，陳舜德，鄭縈，劉顯仲，柯淑津．多語料庫的中文姓名辨識，１９９１年全國計算語言學聯合學術會議，杭州，１９９１．

[5] C. K. Fan and W. H. Tsai. Automatic word identification in Chinese sentences by the relaxation technique. In *Proc. of National Computer Symposium*, pages 423–431, Taipei, Taiwan, 1987.

[6] R. Grishman and R. Kittredge, editors. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[7] R. Kittredge and J. Lehrberger, editors. *Sublanguage: Studies of language in restricted domains.* Walter de Gruyter, Berlin, 1982.

[8] N. Liang. On the automatic segmentation of Chinese words and related theory. In *Proc. of the 1987 International Conference on Chinese information processing*, pages 454–459, Beijing, 1987.

[9] R. Sproat and C. Shih. A statistic method for finding word boundaries in Chinese text. *Computer Processing of Chinese & Oriental Languages*, 4(4):336–351, March, 1990.

[10] M. Tomita. *Efficient Parsing for Natural Language.* Kluwer Academic Publishers, 1986.

[11] L.-J. Wang, T. Pei, W.-C. Li, and L.-C. Huang. A parsing method for identifying words in Mandarin Chinese. In *Proc. of IJCAI-91*, pages 1018–1023, 1991.

[12] C.-L. Yeh and H.-J. Lee. Unification-based word identification for Mandarin Chinese sentences. *Proc. of 1988 ICCPCOL*, pages 27-32, Toronto, Canada, 1988.