# A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction

Wei-Yun Ma
Institute of Information science,
Academia Sinica
ma@iis.sinica.edu.tw

Keh-Jiann Chen
Institute of Information science,
Academia Sinica
kchen@iis.sinica.edu.tw

## Abstract

Statistical methods for extracting Chinese unknown words usually suffer a problem that superfluous character strings with strong statistical associations are extracted as well. To solve this problem, this paper proposes to use a set of general morphological rules to broaden the coverage and on the other hand, the rules are appended with different linguistic and statistical constraints to increase the precision of the representation. To disambiguate rule applications and reduce the complexity of the rule matching, a bottom-up merging algorithm for extraction is proposed, which merges possible morphemes recursively by consulting above the general rules and dynamically decides which rule should be applied first according to the priorities of the rules. Effects of different priority strategies are compared in our experiment, and experimental results show that the performance of proposed method is very promising.

## 1 Introduction and Related Work

Chinese sentences are strings of characters with no delimiters to mark word boundaries. Therefore the initial step for Chinese processing is word segmentation. However, occurrences of unknown words, which do not listed in the dictionary, degraded significantly the performances of most word segmentation methods, so unknown word extraction became a key technology for Chinese segmentation.

For unknown words with more regular morphological structures, such as personal names, morphological rules are commonly used for improving the performance by restricting the structures of extracted words (Chen et. al 1994, Sun et. al 1994, Lin et. al 1994). However, it's not possible to list morphological rules for all kinds of unknown words, especially those words with very irregular structures, which have the characteristics of variable lengths and flexible morphological structures, such as proper names, abbreviations etc. Therefore, statistical approaches usually play major roles on irregular unknown word extraction in most previous work (Sproat & Shih 1990, Chiang et. al 1992, Tung and Lee 1995, Palmer 1997, Chang et. al 1997, Sun et. al 1998, Ge et. al 1999).

For statistical methods, an important issue is how to resolve competing ambiguous extractions which might include erroneous extractions of phrases or partial phrases. They might have statistical significance in a corpus as well. Very frequently superfluous character strings with strong statistic associations are extracted. These wrong results are usually hard to be filtered out unless deep content and context analyses were performed. To solve this problem, the idea of unknown word detection procedure prior to extraction is proposed. Lin et al. (1993) adopt the following strategy: First, they decide whether there is any unknown word within a detected region with fix size in a sentence, and then they extract the unknown word from the region by a statistical method if the previous answer is "yes". A limitation of this method is that it restricts at most

one unknown word occurs in the detected region, so that it could not deal with occurrences of consecutive unknown words within a sentence. Chen & Ma (2002) adopt another strategy: After an initial segmentation process, each monosyllable is decided whether it is a common word or a morpheme of unknown word by a set of syntactic discriminators. The syntactic discriminators are a set of syntactic patterns containing monosyllabic, words which are learned from a large word segmented corpus, to discriminate between monosyllabic words and morphemes of unknown words. Then more deep analysis can be carried out at the detected unknown word morphemes to extract unknown words.

In this paper, in order to avoid extractions of superfluous character strings with high frequencies, we proposed to use a set of general rules, which is formulated as a context free grammar rules of composing detected morphemes and their adjacent tokens, to match all kinds of unknown words, for instance which includes the rule of (UW → UW UW). To avoid too much superfluous extractions caused by the over general rules, rules are appended with linguistic or statistical constraints. To disambiguate between rule applications and reduce the complexity of the rule matching, a bottom-up merging algorithm for extraction is proposed, which merges possible morphemes recursively by consulting above general rules and dynamically decides which rule should be applied first according to the priorities of the rules.

The paper is organized into 7 sections. In the next section, we provide an overview of our system. Section 3 briefly introduce unknown word detection process and makes some analysis for helping the derivation of general rules for unknown words. In section 4, we derive a set of general rules to represent all kinds of unknown words, and then modify it by appending rules constraints and priorities. In section 5, a bottom-up merging algorithm is presented for unknown word extraction. In section 6, the evaluation of extraction is presented; we also compare the performances to different priority strategies. Finally, in section 7, we make the conclusion and propose some future works.

## 2 System Overview

The purpose to our unknown word extraction system is to online extract all types of unknown words from a Chinese text. Figure 1 illustrates the block diagram of the system proposed in this paper. Initially, the input sentence is segmented by a conventional word segmentation program. As a result, each unknown word in the sentence will be segmented into several adjacent tokens (known words or monosyllabic morphemes). At unknown word detection stage, every monosyllable is decided whether it is a word or an unknown word morpheme by a set of syntactic discriminators, which are learned from a corpus. Afterward, a bottom-up merging process applies the general rules to extract unknown word candidates. Finally, the input text is re-segmented by consulting the system dictionary and the extracted unknown word candidates to get the final segmented result.

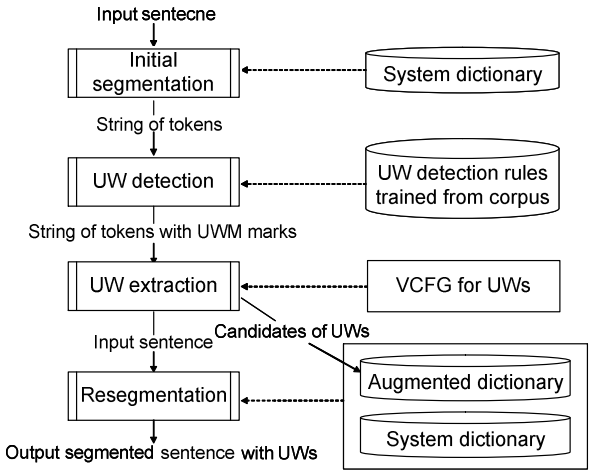



Figure 1. Flowchart of the system

(1) 若 能 提升 毛利率
if can increase gross profit rate
"if gross profit rate can be increased…"

(2) after first step word segmentation:
若 能 提升 毛 利 率
after unknown word detection:
若 能 提升 毛(?) 利(?) 率(?)
after unknown word extraction:
若 能 提升 毛利率

For example, the correct segmentation of (1) is shown, but the unknown word "毛利率" is segmented into three monosyllabic words after the

first step of word segmentation process as shown in (2). The unknown word detection process will mark the sentence as "若() 能() 提升() 毛(?) 利(?) 率(?)", where (?) denotes the detected monosyllabic unknown word morpheme and () denotes common words. During extracting process, the rule matching process focuses on the morphemes marked with (?) only and tries to combine them with left/right neighbors according to the rules for unknown words. After that, the unknown word "毛利率" is extracted. During the process, we do not need to take care of other superfluous combinations such as "若能" even though they might have strong statistical association or co-occurrence too.

## 3 Analysis of Unknown Word Detection

The unknown word detection method proposed by (Chen & Bai 1998) is applied in our system. It adopts a corpus-based learning algorithm to derive a set of syntactic discriminators, which are used to distinguish whether a monosyllable is a word or an unknown word morpheme after an initial segmentation process. If all occurrences of monosyllabic words are considered as morphemes of unknown words, the recall of the detection will be about 99%, but the precision is as low as 13.4%.

The basic idea in (Chen & Bai 1998) is that the complementary problem of unknown word detection is the problem of monosyllabic known-word detection, i.e. to remove the monosyllabic known-words as the candidates of unknown word morphemes. Chen and Bai (1998) adopt ten types of context rule patterns, as shown in table 1, to generate rule instances from a training corpus. The generated rule instances were checked for applicability and accuracy. Each rule contains a key token within curly brackets and its contextual tokens without brackets. For some rules there may be no contextual dependencies. The function of each rule means that in a sentence, if a character and its context match the key token and the contextual tokens of the rule respectively, this character is a common word (i.e. not a morpheme of unknown word).

For instance, the rule "{Dfa} Vh" says that a character with syntactic category Dfa is a common word, if it follows a word of syntactic category Vh.

| Rule type | Example |
|---|---|
| char | {的} |
| word char | 不 {願} |
| char word | {全} 世界 |
| category | {T} |
| {category} category | {Dfa} Vh |
| category {category} | Na {Vcl} |
| char category | {就} VH |
| category char | Na {上} |
| category category char | Na Dfa {高} |
| char category category | {極} Vh T |

Table1. Rule types and Examples

The final rule set contains 45839 rules and were used to detect unknown words in the experiment. It achieves a detection rate of 96%, and a precision of 60%. Where detection rate 96% means that for 96% of unknown words in the testing data, at least one of its morphemes are detected as part of unknown word and the precision of 60% means that for 60% of detected monosyllables in the testing data, are actually morphemes. Although the precision is not high, most of over-detecting errors are "isolated", which means there are few situations that two adjacent detected monosyllabic unknown morphemes are both wrong at the mean time. These operative characteristics are very important for helping the design of general rules for unknown words later.

## 4 Rules for Unknown Words

Although morphological rules work well in regular unknown word extraction, it's difficult to induce morphological rules for irregular unknown words. In this section, we try to represent a common structure for unknown words from another point of view; an unknown word is regarded as the combination of morphemes which are consecutive morphemes/words in context after segmentation, most of which are monosyllables. We adopt context free grammar (Chomsky 1956), which is the most commonly used generative grammar for modelling constituent structures, to express our unknown word structure.

### 4.1 Rule Derivation

According to the discussion in section 3, for 96% of unknown words, at least one of its morphemes are detected as part of unknown word, which motivates us to represent the unknown word

structure with at least one detected morpheme. Taking this phenomenon into our consideration, the rules for modeling unknown words and an unknown word example are presented as follows.

| | | |
|---|---|---|
| UW | → UW UW | (1) |
| | \| ms(?) ms(?) | (2) |
| | \| ms(?) ps() | (3) |
| | \| ms(?) ms() | (4) |
| | \| ps() ms(?) | (5) |
| | \| ms() ms(?) | (6) |
| | \| ms(?) UW | (7) |
| | \| ms() UW | (8) |
| | \| ps() UW | (9) |
| | \| UW ms(?) | (10) |
| | \| UW ms() | (11) |
| | \| UW ps() | (12) |

Notes: There is one non-terminal symbol. "UW" denotes "unknown word" and is also the start symbol. There are three terminal symbols, which includes ms(?), which denotes the detected monosyllabic unknown word morpheme, ms() , which denotes the monosyllable that is not detected as the morpheme, and ps(), which denotes polysyllabic (more than one syllable) known word.

Table 2. General rules for unknown words

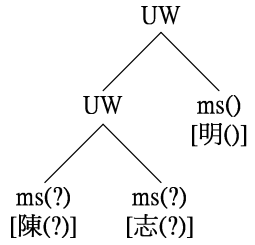



Figure 2. A possible structure for the unknown word " 陳 志 明 "(Chen Zhi Ming), which is segmented initially and detected as "陳(?) 志(?) 明()", and "明" was marked incorrectly at detection stage.

There are three kinds of commonly used measures applied to evaluate grammars: 1. generality (recall), the range of sentences the grammar analyzes correctly; 2. selectivity (precision), the range of non-sentences it identifies as problematic and 3. understandability, the simplicity of the grammar itself (Allen 1995). For generality, 96% unknown words have this kind of structure, so the grammar has high generality to generate unknown words. But for selectivity, our rules are over-generation. Many patterns accepted by the rules are not words. The main reason is that rules have to include non-detected morphemes for high generality. Therefore selectivity is sacrificed momentary. In next section, rules would be constrained by linguistic and text-based statistical constraints to compensate the selectivity of the grammar. For understandability, you can find each rule in (1)-(12) consists of just two right-hand side symbols. The reason for using this kind of presentation is that it regards the unknown word structure as a series of combinations of consecutive two morphemes, such that we could simplify the analysis of unknown word structure by only analyzing its combinations of consecutive two morphemes.

## 4.2 Appending Constraints

Since the general rules in table 2 have high generality and low selectivity to model unknown words, we append some constraints to restrict their applications. However, there are tradeoffs between generality and selectivity: higher selectivity usually results in lower generality. In order to keep high generality while assigning constraints, we assign different constraints on different rules according to their characteristics, such that it is only degraded generality slightly but selectivity being upgraded significantly.

The rules in table 2 are classified into two kinds: one kind is the rules which both its right-hand side symbols consist of detected morphemes, i.e, (1), (2), (7), and (10), the others are the rules that just one of its right-hand side symbols consists of detected morphemes, i.e, (3), (4), (5), (6), (8), (9), (11), and (12). The former is regarded as "strong" structure since they are considered to have more possibility to compose an unknown word or an unknown word morpheme and the latter is regarded as "weak" structure, which means they are considered to have less possibility to compose an unknown word or an unknown word morpheme. The basic idea is to assign more constraint on those rules with weak structure and less constraint on those rules with strong structure.

The constraints we applied include word length, linguistic and statistical constraints. For statistical constraints, since the target of our system is to

extract unknown words from a text, we use text-based statistical measure as the statistical constraint. It is well known that keywords often reoccur in a document (Church 2000) and very possible the keywords are also unknown words. Therefore the reoccurrence frequency within a document is adopted as the constraint. Another useful statistical phenomenon in a document is that a polysyllabic morpheme is very unlikely to be the morphemes of two different unknown words within the same text. Hence we restrict the rule with polysyllabic symbols by evaluating the conditional probability of polysyllabic symbols. In addition, syntactic constraints are also utilized here. For most of unknown word morphemes, their syntactic categories belong to "bound", "verb", "noun", and "adjective" instead of "conjunction", "preposition"…etc. So we restrict the rule with non-detected symbols by checking whether syntactic categories of its non-detected symbols belong to "bound", "verb", "noun", or "adjective". To avoid unlimited recursive rule application, the length of matched unknown word is restricted unless very strong statistical association do occur between two matched tokens. The constraints adopted so far are presented in table 3. Rules might be restricted by multi-constraints.

| | |
|---|---|
| $Freq_{docu}(LR)>=Threshold$ | (3) (4) (5) (6) (8) (9) (11) (12) |
| $P_{docu}(L\|R)=1$ | (1) (3) (7) (8) (9) (12) |
| $P_{docu}(R\|L)=1$ | (1) (5) (9) (10) (11) (12) |
| Category(L) is bound, verb, noun or adjective | (5) (6) (8) (9) |
| Category(R) is bound, verb, noun or adjective | (3) (4) (11) (12) |

Notes: L denotes left terminal of right-hand side
R denotes right terminal of right-hand side
Threshold is a function of Length(LR) and text size. The basic idea is larger amount of length(LR) or text size matches larger amount of Threshold.

Table 3. Constraints for general rules

### 4.3 Priority

To scheduling and ranking ambiguous rule matching, each step of rule matching is associated with a measure of priority which is calculated by the association strength of right-hand side symbols. In our extracting algorithm, the priority measure is used to help extracting process dynamically decide which rule should be derived first. More detail discussion about ambiguity problem and complete disambiguation process are presented in section 5.

We regard the possibility of a rule application as co-occurrence and association strength of its right-hand side symbols within a text. In other words, a rule has higher priority of application while its right-hand side symbols are strongly associated with each other, or co-occur frequently in the same text. There have been many statistical measures which estimate co-occurrence and the degree of association in previous researches, such as mutual information (Church 1990, Sporat 1990), t-score (Church 1991), dice matrix (Smadja 1993, 1996). Here, we adopt four well-developed kinds of statistical measures as our priority individually: mutual information (MI), a variant of mutual information (VMI), t-score, and co-occurrence. The formulas are listed in table 4. MI mainly focuses on association strength, and VMI and t-score consider both co-occurrence and association strength. The performances of these four measures are evaluated in our experiments discussed in section 6.

| |
|---|
| $co-occurence(L,R) = f(L,R)$ |
| $MI(L,R) = \log \frac{P(L,R)}{P(L)P(R)}$ |
| $VMI(L,R) = f(L,R)MI(L,R)$ |
| $t-score(L,R) = \frac{f(L,R) - \frac{f(L)f(R)}{N}}{\sqrt{f(L,R)}}$ |

Notes: f(L,R) denotes the number of occurrences of L,R in the text; N denotes the number of occurrences of all the tokens in the text; length(*) denotes the length of *.

Table 4. Formulas of 4 kinds of priority

## 5 Unknown Word Extraction

### 5.1 Ambiguity

Even though the general rules are appended with

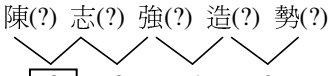
well-designed constraints, ambiguous matchings, such as, overlapping and covering, are still existing. We take the following instance to illustrate that: “拉法葉” (La Fa Yeh), a warship name, occurs frequently in the text and is segmented and detected as “拉(?) 法(?) 葉(?)”. Although “拉法葉” could be derived as an unknown word “((拉法)葉)” by rule 2 and rule 10, “拉法” and ”法葉” might be also derived as unknown words “(拉法)” and “(法葉)” individually by the rule 2. Hence there are total three possible ambiguous unknown words and only one is actually correct.

Several approaches on unsupervised segmentation of Chinese words were proposed to solve overlapping ambiguity to determine whether to group “xyz” as “xy z” or “x yz”, where x, y, and z are Chinese characters. Sproat and Shih (1990) adopt a greedy algorithm: group the pair of adjacent characters with largest mutual information greater than some threshold within a sentence, and the algorithm is applied recursively to the rest of the sentence until no character pair satisfies the threshold. Sun et al. (1998) use various association measures such as t-score besides mutual information to improve (Sproat & Shih 1990). They developed an efficient algorithm to solve overlapping character pair ambiguity.

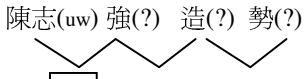
### 5.2 Bottom-up Merging Algorithm

Following the greedy strategy of (Sproat & Shih 1990), here we present an efficient bottom-up merging algorithm consulting the general rules to extract unknown words. The basic idea is that for a segmented sentence, if there are many rule-matched token pairs which also satisfy the rule constraints, the token pair with the highest rule priority within the sentence is merged first and forms a new token string. Same procedure is then applied to the updated token string recursively until no token pair satisfied the general rules. It is illustrated by the following example:

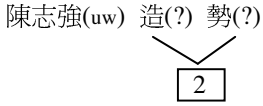
==========
System environment:
Co-occurrence priority is adopted.
Text environment:
“陳志強” (Chen Zhi Qiang), an unknown word, occurs three times.
“造勢” (take an electing activity), an unknown word, occurs two times.
“陳志強造勢” (Chen Zhi Qiang took an electing activity), a sentence, occurs one time.
Input: 陳志強造勢
After initial segmentation and detection:




priority
After first iteration:




priority
After second iteration:




priority
After third iteration:
陳志強(uw) 造勢(uw)
==========

Figure 3. Extraction process of input “陳志強造勢”.

By the general rules and greedy strategy, besides overlapping character pair ambiguity, the algorithm is able to deal with more complex overlapping and coverage ambiguity, even which result from consecutive unknown words. In finger 3, input sentence “陳志強造勢” is derived as the correct two unknown words “((陳志)強)” and “(造勢)” by rule (2), rule (10), and rule (2) in turn. “陳志強” and ”造勢” are not further merged. That is because P(造勢|陳志強)<1 violates the constraint of rule (1). Same reason explains why “陳志強” and ”造” do not satisfy rule (10) in the third iteration.

By this simple algorithm, unknown words with unlimited length all have possibilities to be extracted. Observing the extraction process of “陳志強”, you can find, in the extraction process, boundaries of unknown words might extend during iteration until no rule could be applied.

## 6 Experiment

In our experiments, a word is considered as an unknown word, if either it is not in the CKIP lexicon or it is not identified by the word segmentation program as foreign word (for instance English) or a number. The CKIP lexicon contains about 80,000 entries.

### 6.1 Evaluation Formulas

The extraction process is evaluated in terms of precision and recall. The target of our approach is to extract unknown words from a text, so we define "correct extractions" as unknown word types correctly extracted in the text. The precision and recall formulas are listed as follows:

$NC_i$ = number of correct extractions in document i
$NE_i$ = number of extracted unknown words in document i
$NT_i$ = number of total unknown words in document i

$$\text{Precision rate} = \frac{\sum_{i=1}^{i=150} NC_i}{\sum_{i=1}^{i=150} NE_i} \qquad \text{Recall rate} = \frac{\sum_{i=1}^{i=150} NC_i}{\sum_{i=1}^{i=150} NT_i}$$

### 6.2 Data Sets

We use the Sinica balanced corpus version 3.0 as our training set for unknown word detection, which contains 5 million segmented words tagged with pos. We randomly select 150 documents of Chinese news on the internet as our testing set. These testing data are segmented by hand according to *the segmentation standard for information processing* designed by the Academia Sinica (Huang et.al 1997). In average, each testing text contains about 300 words and 16.6 unknown word types.

### 6.3 Results

Based on the four priority measures listed in table 4, the bottom-up merging algorithm is applied. The performances are shown in table 5.

| Evaluation / Priority | Match# | Extract# | Precision | Recall |
|---|---|---|---|---|
| Co-occurrence | 1122 | 1485 | 76% | 45% |
| MI | 1112 | 1506 | 74% | 45% |
| VMI | 1125 | 1499 | 75% | 45% |
| t-score | 1125 | 1494 | 75% | 45% |

Note: There are total 2498 reference unknown word types

Table 5. Experimental results of the four different priority measures

In table 5, comparing co-occurrence and MI, we found that the performance of co-occurrence measure is better than MI on both precision and recall. The possible reason is that the characteristic of reoccurrence of unknown words is more important than morphological association of unknown words while extracting unknown words from a size-limited text. That is because sometimes different unknown words consist of the same morpheme in a document, and if we use MI as the priority, these unknown words will have low MI values of their morphemes. Even though they have higher frequency, they are still easily sacrificed when they are competed with their adjacent unknown word candidates. This explanation is also proved by the performances of VMI and t-score, which emphasize more importance on co-occurrence in their formulas, are better than the performance of MI. According to above discussions, we adopt co-occurrence as the priority decision making in our unknown word extraction system.

In our final system, we adopt morphological rules to extract regular type unknown words and the general rules to extract the remaining irregular unknown words and the total performance is a recall of 57% and a precision of 76%. An old system of using the morphological rules for names of people, compounds with prefix or suffix were tested, without using the general rules, having a recall of 25% and a precision of 80%. The general rules improve 32% of the recall and without sacrificing too much of precision.

## 7 Conclusion and Future Work

In this research, Chinese word segmentation and unknown word extraction has been integrated into a frame work. To increase the coverage of the morphological rules, we first derive a set of general rules to represent all kinds of unknown words. To avoid extracting superfluous character strings, we then append these rules with linguistic and statistical constraints. We propose an efficient bottom-up merging algorithm by consulting the general rules to extract unknown words and using priority measures to resolve the rule matching ambiguities. In the experiment, we compare effects of different priority strategies, and experimental results show that the co-occurrence measure performances best.

It is found that the performance of unknown word detection would affect the entire performance significantly. Although the performance of unknown word detection is not bad, there is still room for improvement. The possible strategies for improvement in our future work include using con-

textual semantic relations in detection, and some updated statistical methods, such as support vector machine, maximal entropy and so on, to achieve better performance of unknown word detection.

## References


[1] Chen, H.H., & J.C. Lee, 1994,"The Identification of Organization Names in Chinese Texts", *Communication of COLIPS*, Vol.4 No. 2, 131-142.

[2] Sun, M. S., C.N. Huang, H.Y. Gao, & Jie Fang, 1994, "Identifying Chinese Names in Unrestricted Texts", *Communication of COLIPS*, Vol.4 No. 2, 113-122

[3] Lin, M. Y., T. H. Chiang, & K. Y. Su, 1993," A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," Proceedings of ROCLING VI, pp. 119-137

[4] Richard Sproat and Chilin Shih, "A Statistical Method for Finding Word Boundaries in Chinese Text," *Computer Processing of Chinese and Oriental Languages,* 4, 336-351, 1990

[5] Sun, Maosong, Dayang Shen, and Benjamin K. Tsou. 1998. Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data. In Proceedings of COLING-ACL '98,pages 1265-1271

[6] Ge, Xianping, Wanda Pratt, and Padhraic Smyth. 1999. Discovering Chinese Words from Unsegmented Text. In SIGIR '99, pages 271-272

[7] Palmer, David. 1997. A Trainable Rule-based Algorithm for Word Segmentation. In Proceedings of the Association for Computational Linguistics

[8] Chiang, T. H., M. Y. Lin, & K. Y. Su, 1992," Statistical Models for Word Segmentation and Unknown Word Resolution," Proceedings of ROCLING V, pp. 121-146

[9] Chang, Jing-Shin and Keh-Yih Su, 1997a. "An Unsupervised Iterative Method for Chinese New Lexicon Extraction", to appear in International Journal of Computational Linguistics & Chinese Language Processing, 1997

[10] C.H. Tung and H. J. Lee , "Identification of unknown words from corpus," *International Journal of Computer Processing of Chinese and Oriental Languages,* Vol. 8, Supplement, pp. 131-146, 1995

[11] Chen, K.J. & Wei-Yun Ma, 2002. Unknown Word Extraction for Chinese Documents. In Proceedings of COLING 2002, pages 169-175

[12] Chen, K.J. & Ming-Hong Bai, 1998, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," international Journal of Computational linguistics and Chinese Language Processing, Vol.3, #1, pp.27-44

[13] Church, Kenneth W., 2000, "Empirical Estimates of Adaptation: The Chance of Two Noriegas is Closer to p/2 than p*p", Proceedings of Coling 2000, pp.180-186.]

[14] Allen James 1995 Natural Language understandding. Second Edition, page 44

[15] Chen, K.J. & S.H. Liu, 1992,"Word Identification for Mandarin Chinese Sentences," *Proceedings of 14th Coling*, pp. 101-107

[16] Huang, C. R. Et al.,1995,"The Introduction of Sinica Corpus," *Proceedings of ROCLING VIII*, pp. 81-89.

[17] Huang, C.R., K.J. Chen, & Li-Li Chang, 1997, "Segmentation Standard for Chinese Natural Language Processing," International Journal of Computational Linguistics and Chinese Language Processing, Accepted.

[18] Chomsky, N. 1956 Three models for the description of language. IRE Transactions on Information Theory, 2, 113-124

[19] Church, K. and Hanks, P., "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, Vol.16, March. 1990, pp.22-29

[20] Smadja, Frank, "Retrieving Collocations from Text: Xtract," Computational Linguistics, Vol. 19 , No. 1, 1993, pp.143-177

[21] Smadja, Frank, McKeown, K.R. and Hatzivasiloglou, V. "Translating Collocations for Bilingual Lexicons," Computational Linguistics, Vol. 22, No.1, 1996

[22] Church, K, W. Gale, P. Hanks, and D. Hindle. 1991 "Using Statistics in Lexical Analysis," in Zernik (ed.) Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, pp. 115-164, Lawrence Erlbaum Associates Publishers