Computational Linguistics and Chinese Language Processing
vol. 3, no. 1, February 1998, pp. 27-44

Unknown Word Detection for Chinese by a Corpus-based Learning Method

Keh-Jiann Chen*, Ming-Hong Bai*

Abstract

One of the most prominent problems in computer processing of the Chinese language is identification of the words in a sentence. Since there are no blanks to mark word boundaries, identifying words is difficult because of segmentation ambiguities and occurrences of out-of-vocabulary words (i.e., unknown words). In this paper, a corpus-based learning method is proposed which derives sets of syntactic rules that are applied to distinguish monosyllabic words from monosyllabic morphemes which may be parts of unknown words or typographical errors. The corpus-based learning approach has the advantages of: 1. automatic rule learning, 2. automatic evaluation of the performance of each rule, and 3. balancing of recall and precision rates through dynamic rule set selection. The experimental results show that the rule set derived using the proposed method outperformed hand-crafted rules produced by human experts in detecting unknown words.

1. Introduction

One of the most prominent problems in computer processing of Chinese language is the identification of the words in a sentence. There are no blanks to mark word boundaries in Chinese text. As a result, identifying words is difficult because of segmentation ambiguities and occurrences of out-of-vocabulary words (i.e., unknown words). For instance, in (1), the proper name 王英雄 'Wang, Ying-Xiong' is a typical example of an unknown word, and it has ambiguous segmentation of 王 'king' 英雄 'hero'. Another example in (1) 台灣大學生 'university student in Taiwan' also has ambiguous segmentations of 台灣 'Taiwan' 大學生 'university student' , 台灣大學 'National Taiwan University' 生 'give birth to' ,and 台灣 'Taiwan' 大學 'university' 生 'give birth to' etc.:

(1) 王英雄是一個典型的台灣大學生。
'Ying-Xiong Wang is a typical university student in Taiwan.'

* Institute of Information Sicence, Academia Sinica, Taipei, Taiwan, R. O. C. E-mail: {kchen, evan}@iis.sinica.edu.tw

Most of the papers dealing with the problem of word segmentation have focused only on the resolution of ambiguous segmentation. The problem of unknown word identification is considered to be more difficult and needs to be further investigated. According to an inspection of the Sinica corpus [Chen et al., 1996], which is a balanced Chinese corpus with words segmented based on the Chinese word segmentation standard for information processing proposed by ROCLING [Huang et al., 1997], the most productive unknown words are of the following types.

1.1 Types of Unknown Words

Unknown words are defined as the words which are not in the lexicon. The following types of unknown words most frequently occur in the Sinica corpus. Table 1 shows the frequency distribution of unknown words of the most frequent 14 categories by examining 3 million-word data from the Sinica corpus.

(a) abbreviation (acronym): e.g., 中油 'China-fuel' (Nb) and 台汽 'Taiwan-bus' (Nb). (Please refer to table 1 for the meaning of each category name; for instance, Nb denotes proper names.)

It is difficult to identify abbreviations since their morphological structures are very irregular. Their affixes more or less reflect the conventions of the selection of meaning components [Huang 94]. However, the affixes of abbreviations are common words which are least informative for indicating the existence of unknown words.

(b) proper names: e.g., 陳壽 'Chen-So' (Nb), 香檳城 'Champaign-city' (Nc), and 微軟 'micro-soft' (Nb).

Proper names can be further classified into 3 sub-categories, i.e., names of people, names of place, and names of organizations. Certain key words are indicators for each different sub-category. For instance, there are about 100 common surnames which are prefix characters of Chinese personal names. The district names, such as 市 'city', 鄉 'country' etc., frequently occur as suffixes of the names of places. Identification of company names is as difficult as that of abbreviations since there is no restriction on the choice of morpheme components.

(c) derived words: e.g., 電腦化 'computer-ize' (Vh).

Derived words have affix morphemes which are strong indicators.

(d) compounds: e.g., 轉赴 'turn-go'(VCL), 獲允 'receive-permission' (VE), 搜尋法 'search-method' (Na), and 電腦桌 'computer-desk' (Na).

A compounds is a very productive type of unknown word. Nominal and verbal

compounds are easily coined by combining two words/characters. Since there are more than 5000 commonly used Chinese characters, each with idiosyncratic syntactic behavior, it is hard to derive a set of morphological rules to generate the set of Chinese compounds. To identify Chinese compounds is, thus, also difficult.

(e) numeric type compounds: e.g., 1986 年 '1986-year' (Nd), 三千 'three-thousand', and 19 巷 '19-lane' (Nc).

The characteristic of numeric compounds is that they contain numbers as major components. For instances, dates, time, phone numbers, addresses, numbers, determiner-measure compounds etc. belong to this type. Since digital numbers are the major components of unknown words of this type and their morphological structures are more regular, they can be identified using the morphological rules.

Category	Frequency	Meaning of Category
A	1453	/*non-predictive adjective*/
Na	34372	/*common noun*/
Nb	14813	/*proper noun*/
Nc	9688	/*location noun*/
Nd	2264	/*time noun*/
VA	6466	/*active intransitive verb*/
VC	8462	/*active transitive verb*/
VCL	811	/*active transitive verb with locative object*/
VD	448	/*ditransitive verb*/
VE	1051	/*active transitive verb with sentential object*/
VG	996	/*classificatory verb*/
VH	10492	/*stative intransitive verb*/
VHC	498	/*stative causative verb*/
VJ	1471	/*stative transitive verb*/
total:	93,285	

Table 1. *The frequency distribution of unknown words in the most frequent categories.*

From the above discussion, it is seen that identification for each different type of unknown word is difficult and might require adopting different approaches. However, the processes for detecting the occurrences of each different type of unknown word are

almost the same since they are all composed of morphemes of characters. In this paper, we focus only on the detection processes and leave the complete identification problem for future research.

1.2 Unknown Word Detection

Unknown words cause segmentation errors because out-of-vocabulary words in an input text normally are incorrectly segmented into pieces of single character word or shorter words. For instance, example (1) would be segmented into (2) after dictionary look-up and resolution of ambiguous segmentation:

(2) 王 英雄 是 一 個 典型 的 台灣 大學生。
king hero be DET CL typical DE Taiwan university-student

It is difficult to know when an unknown word is encountered since all Chinese characters can either morphemes or words and there are no blanks to mark word boundaries. Therefore, without (or even with) syntactic or semantic checking, it is difficult to tell whether a character in a particular context is a part of an unknown word or whether it stands alone as a word. As mentioned in section 1.1, compound words and proper names are the two major types of unknown words. It is not possible to list all of the compounds in the lexicon nor possible to write simple rules which can enumerate the compounds without over-generation or under-generation. Each different type of compound must be identified using either content or context dependent rules. Proper names and their abbreviations have less content regularity. Identifying them relies more on contextual information. The occurrence of typographical errors makes the problem even more complicated. There is currently no satisfactory algorithm for identifying both unknown words and typographical errors, but researchers are separately working on each different type of problem.

Chang et al.[Chang et al., 94] used statistical methods to identify personal names in Chinese text and achieved a recall rate of 80% and a precision rate of 90%. Similar experiments were reported in [Sun et al., 94]. Their recall rate was 99.77% but with a lower precision of 70.06%. Both papers default with the recognition of Chinese personal names only. Chen & Lee [Chen & Lee 94] used morphological rules and contextual information to identify the names of organizations. Since organizational names are much more irregular than personal names in Chinese, they achieved a recall rate of 54.50% and a precision rate of 61.79%. A pilot study on automatic correction of Chinese spelling errors was done by Chang [Chang 94]. He used mutual information between a character and its neighboring words to detect spelling errors and to then automatically make the necessary corrections. The error detection process achieved a recall rate of 76.64% and

a precision rate of 51.72%. Lin et al. [Lin et al., 93] did a preliminary study of the problem of unknown word identification. They used 17 morphological rules to recognize regular compounds and a statistical model to deal with irregular unknown words, such as proper names etc. With this unknown word resolution procedure, an error reduction rate of 78.34% was obtained for the word segmentation process. Since there is no standard reference data, the accuracy rates claimed in different papers vary due to different segmentation standards. In this study, we used the Sinica corpus as a standard reference data. As mentioned before, the Sinica corpus is a word-segmented corpus based on the Chinese word segmentation standard for information processing proposed by ROCLING. Therefore, it contains both known words and unknown words which are properly segmented, i.e., separated by blanks. The corpus was utilized for the purposes of training and testing. For unknown word and typographical error identification, the following two steps are proposed. The first step is to detect the existence of unknown words and typographical errors. The second step is the recognition process, which determines the type and boundaries of each unknown word and recognizes typographical errors. The reasons for separating the detection process from the recognition process are as follows:

a. For different types of unknown words and typographical errors, they may share the same detection process but have different recognition processes.
b. If the common method for spell checking is followed, an unknown word would be detected first, and a search for the best matching words performed next. Recognizing a Chinese word is somewhat different from spell checking, but they have a lot in common.
c.If the detection process performs well, the recognition process is better focused, making the total performance more efficient.

This paper focuses on the unknown word detection problem only. (Note that a typographical error is considered as a special kind of unknown word.) The unknown word detection problem and the dictionary-word detection problem are complementary problems since if all known words in an input text can be detected, then the rest of the character string will be unknown words. However, this is not a simple task since there are no blanks to delimit known words from unknown words. Therefore, the word segmentation process is applied first, and known words are delimited by blanks. Since unknown words are not listed in the dictionary, they will be segmented into shorter character/word sequences after a conventional dictionary-look-up word segmentation process. Sentence (3.b) shows the result of the word segmentation process on (3.a):

(3) a. 筑波大學延請七三年諾貝爾物理學獎得主江崎出任校長，

'The University of Tsukuba invited the winner of the '73 Nobel Award in Physics, Esaki, to be the Principal.'

b. 筑 波 大學 延請 七三年 諾貝爾 物理學 獎 得主
Tsuku -ba university invite '73 Nobel physics award winner
江 崎 出任 校長，
Esa -ki be principal.

According to an examination of a group of testing data which is a part of the Sinica corpus, 4572 occurrences out of 4632 unknowns were incorrectly segmented into sequences of shorter words, and each sequence contained at least one monosyllabic word. That is, 60 of the unknown words were segmented into sequences of multi-syllabic words only. Therefore, occurrences of monosyllabic words (i.e., single character words) in the segmented input text may denote the possible existence of unknown words. This is reasonable since it is very rare for compounds or proper names to be composed of several multi-syllabic words. Therefore, the process of detecting unknown words is equivalent to making a distinction between monosyllabic words and monosyllabic morphemes which are parts of unknown words. Hence, the complementary problem of unknown word detection is the problem of monosyllabic known-word detection. If all of the occurrences of monosyllabic words are considered as possible morphemes of unknown words, the precision of prediction is very low. When the word segmentation process was applied to the testing data taken from the Sinica corpus using a conventional dictionary look-up method, 69733 occurrences of monosyllabic words were found, but only 9343 were parts of unknown words, a precision of 13.40%. In order to improve the precision, monosyllabic words, which properly fit the contextual environment, should be identified and should not be considered as possible morphemes of unknown words. In the next section, the corpus-based learning approach to identification of contextually-proper monosyllabic words is introduced. In section 3, experimental results are presented, including a performance comparison between a hand-crafted method and the proposed corpus-based learning method.

2. Corpus-based Rule Learning for Identifying Monosyllabic Words

The procedure for detecting unknown words is roughly divided into three steps: 1. word segmentation, 2. part-of-speech tagging, and 3. identification of contextually-proper monosyllabic words. The word segmentation procedure identifies words using a dictionary look-up method and resolves segmentation ambiguities by maximizing the probability of a segmented word sequence [Chiang 92, Chang 91, Sproat 96] or by using heuristic methods [Chen 92, Lee 91]. Either method can achieve very satisfactory results.

Both have an accuracy rate of over 99%. For the purpose of unknown word identification, some regular types of compounds, such as numbers, determinant-measure compounds, and reduplication, which have regular morphological structures, are also identified by means of their respective morphological rules during the word segmentation process [Chen 92, Lin 93]. The second step, part-of-speech (pos) tagging, is carried out to support step3 and the later process of unknown word identification. After pos tagging, sentence (3.b) becomes sentence (4); each word contains a unique pos:

(4) 筑 (BOUND) 波 (Nf) 大學 (Nb) 延請 (VC) 七三年 (DM) 諾貝爾 (Nb)
Tsuku -ba university invite '73 Nobel
物理學 (Na) 獎 (Na) 得主 (Na) 江 (Na) 崎 (BOUND) 出任 (VG)
physics award winnter Esa -ki be
校長 (Na) ，
principal.

Although the pos sequence may not be 100% correct, it is the most probable pos sequence in terms of pos bi-gram statistics [Liu 95]. The details of the first two steps are not the major concern of this paper. The focus here is on the step of identifying contextually-proper monosyllabic words. Hereafter, for simplicity, the term 'proper-character' will denote a contextually-proper monosyllabic word, and the term 'improper-character' will be used to denote a contextually-improper monosyllabic word which might be part of an unknown word. The way to identify proper-characters is to check the following properties:

(1) a proper-character should not be a bound-morpheme, and
(2) the context of a proper-character should be grammatical.

Hence, if a character is a bound-morpheme, it will be considered as possibly being an unknown word. However, almost any Chinese character can function either as a word or as a bound morpheme. A character's functional role is contextually dependent. Therefore, every monosyllabic word should be checked in its context for grammaticality by means of syntactic or semantic rules. For processing efficiency, such rules should be simple and have only local dependencies. It is not feasible to parse whole sentences in order to check whether or not characters are proper-characters. The task is then to derive a set of rules which can be used to check the grammaticality of characters in context. If the rules are too stringent, then too many proper-characters will be considered as improper-characters, resulting in a low precision rate. On the other hand, if the rules are too relaxed, then too many improper-characters will be considered as proper-characters, resulting in a low recall rate. Therefore, there is a tradeoff between recall and precision. In the case of unknown word detection, a higher recall rate and an acceptable precision

rate is preferred. Writing handcrafted rules is difficult because there are more than 5000 commonly used Chinese characters, and each of them may behave differently. A corpus-based learning approach is adapted to derive the set of contextual rules and to select the best set of rules by evaluating the performance of each individual rule. The approach is very similar to the error-driven learning method proposed by Brill [Brill 95]. Before the learning method is introduced, two commonly used measures for unknown word detection are defined. These two performance measures will be used throughout the paper:

Recall Rate = # of unknown word detected / total number of unknowns;

Precision Rate = # of correctly detected improper-characters / total # of guesses.

There are two types of unknown words. Type one unknown words include mono-syllabic morphemes. Type two unknown words are composed with multi-syllabic words only. Only the detection of type one unknown words is considered here since type two unknown words occur very rarely as we mentioned before. An unknown word is considered successfully detected if any one of its components is detected as an improper-character. It is noted that the numerators for the recall rate and the precision rate are different since if two (or more) components of an unknown word are detected as improper-characters, it is reasonable to count only one word detection but two improper-character detections. For the corpus-based learning method, a training corpus with all the words segmented and pos-tagged is used. The monosyllabic words in the training corpus are instances of proper-characters, and the words in the training corpus which are not in the dictionary are instances of unknown words. Segmenting the unknown words using a dictionary look-up method produces instances of improper-characters. By examining the instances of proper and improper characters and their contexts, the rule patterns and their performance evaluations can be derived and can be represented as triplets (rule pattern, total # of matched instances, # of improper instances). Examples are shown in Appendix1. A contextual dependent rule may be: a uni-gram pattern, such as '{ 的 }', '{ 好 }', '{(Nh)}', '{(T)}', a bi-gram patterns, such as '{ 會 } 覺得 ', '{ 就 }(VH)', '(Na){ 上 }', '{(Dfa)}(Vh)', '(Ve){(Vj)}', or a tri-gram patterns, such as '{ 極 }(VH)(T)', '(Na)(Dfa){ 高 }', where the string in the curly brackets will match a proper-character and the other parts will match its context.

A good rule pattern has high applicability and high discrimination value (i.e., it occurs frequently and matches either proper-characters or improper-characters only, but not both). In fact, no rule has perfect discriminating ability. For instance, the rule '(Na){(Nb)}' can be applied to ' 會計 (Na) 劉 (Nb)' in (5) and ' 院士 (Na) 閻 (Nb)' in (6). The results are correct in (5) and incorrect in (6):

(5)	會計 (Na)	人員 (Na)	劉 (Nb)	小姐 (Na)
	accounting	staff	Liu	Miss.

(6)	院士 (Na)	閻 (Nb)	振興 (VC)	先生 (Na)
	academician	Yan	-strengthen	Mr.

Therefore, a greedy method is adopted in selecting the best set of unknown word detection rules. A set of rules which can identify proper-characters with high accuracy is selected. The rules with applicability greater than a threshold value are sequentially chosen according to the order of their accuracy. The rules for identifying improper-characters was not used because most improper-characters are of low frequency. Conversely, the selected rules were used as the recognition rules for proper-characters. A character matched by any one of the selected rules is considered a proper-character. Characters which are not matched by any one of the rules are considered candidates of improper-characters.

Rule selection algorithm:

1. Determine the threshold values for rule accuracy and applicability. For each rule Ri, when applied on the training corpus, the rule accuracy(Ri) = Mi / Ti, where Mi is the # of instances of matches of Ri with proper characters; Ti is the total # of matches of Ri. The rule applicability(Ri) = Ti.

2. Sequentially select the rules with the highest rule accuracy and applicability greater than the threshold value until there are no rules satisfying both threshold values.

The threshold value for rule accuracy controls the precision and recall performance of the final selected rule set. A higher accuracy requirement means fewer improper-characters will be wrongly recognized as proper-characters. Therefore, the performance of such a rule set will have a higher recall value. However, those proper-characters not matched with any rules will be mistaken as improper-characters, which lowers precision. On the other hand, if a lower accuracy threshold value is used, then most of the proper-characters will be recognized, and many of the improper-characters will also be mistakenly recognized as proper-characters, resulting in a lower recall rate and possibly a higher precision rate before reaching the maximal precision value. Therefore, if a detection rule set with a high recall rate is desired, the threshold value of rule accuracy must be set high. If precision is more important, then the threshold value must be properly lowered to an optimal point. A balance between recall and precision should be considered.

In the next section, the experimental results of different threshold values are

presented. The threshold value for rule applicability controls the number of rules to be selected and ensures that only useful rules are selected.

The selected rule type may subsume another. Shorter rule patterns are usually more general than longer ones. There are redundant rules in the initial rule selection. A further screening process is needed to remove the redundant rules. The screening process is based on the following fact: if a rule Ri is subsumed by rule Rj, then pattern of Ri is a sub-string of pattern Rj. For example the rule '{ 的 }' is more general than the rule '{ 的 } (Na)'. If the rule '{ 的 }' is selected, then the rule '{ 的 } (Na)' is redundant and can be removed from the rule set. Since a character matched by any one of the selected rules is considered a proper-character, more specific rules will be redundant and only the most general rules will remain after the screening process.

Screening Algorithm:

1. Sort the rules according to their string patterns in increasing order, resulting in rules R1...Rn.
2. For i from 1 to n, if there is a j such that j< i, and Rj is a sub-string of Ri, then remove Ri.

3. Experimental Results

The corpus-based learning method for unknown word detection was tested on the Sinica corpus. The Sinica corpus version 2.0 contains 3.5 million words. 3 million words were used as the training corpus and 0.15 million words for the testing corpus. The word entries in the CKIP lexicon were considered as known words. The CKIP lexicon contains about 80,000 entries of Chinese words with their syntactic categories and grammatical information [CKIP 93]. A word is considered as an unknown word if it is not in the CKIP lexicon and is not identified by the word segmentation program as a foreign word (for instance, English), a number, or a reduplicated compound. There were 93285 unknown words in the training corpus and 4632 unknown words in the testing corpus. A few bi-word compounds were deliberately ignored as unknowns, such as 分析化學 'analytical chemistry' and 技術人員 'technical member', since they are not identifiable by any algorithm which does not incorporate real world knowledge. In addition, whether these are single compounds or noun phrases made up of two words is debatable. In fact, ignoring bi-word compounds did not affect the results very much since the fact that there were only 60 such unknown words out of 4632 shows that they rarely occurred in the corpus.

The following types of rule patterns were generated from the training corpus. Each

rule contains a token within curly brackets and its contextual tokens without brackets. For some rules, there may be no contextual dependencies.

Rule type	Examples
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)

Rules of the 10 different types of patterns above were generated automatically by extracting each instance of monosyllabic words in the training corpus. Every generated rule pattern was checked for redundancy, and the frequencies of proper and improper occurrences were tallied. For instance, the pattern '{ 的 }' occurred 165980 times in the training corpus; 165916 of these were proper instances and 64 of these were improper instances (i.e., ' 的 ' occurred 64 times as part of an unknown word). Appendix 1 shows some of the rule patterns and their total occurrences counts as well as the number of improper instances. In the initial stage, 1455633 rules were found. After eliminating rules with frequency less than 3, 215817 rules remained. In the next stage, different rule selection threshold values were used to generate 10 different sets of rules. These rule sets were used to detect unknown words in the testing corpus. The testing corpus contained 152560 words. In the first step, the running text of the testing corpus was segmented into words using a dictionary look-up method which were then tagged with their part-of-speech by an automatic tagging process. Each different rule set was applied to detect the unknown words in the testing corpus. A character without a match was considered as part of an unknown word. Appendix 2 shows some examples. The performance results of different rule sets are shown in Table 2, and the detail statistics are shown in Appendix 3.

Rule selection criteria	Recall rate	Precision rate	# of rules after screening
(0) no rule applied	100%	13.40%	0
(1) rule accuracy >= 55%	63.32%	73.69%	3054
(2) rule accuracy >= 60%	63.89%	73.73%	3239
(3) rule accuracy >= 65%	64.85%	74.04%	5209
(4) rule accuracy >= 70%	68.18%	74.61%	6081
(5) rule accuracy >= 75%	73.80%	74.36%	8611
(6) rule accuracy >= 80%	77.34%	73.26%	10500
(7) rule accuracy >= 85%	81.06%	71.52%	13962
(8) rule accuracy >= 90%	87.40%	68.74%	18967
(9) rule accuracy >= 95%	93.66%	64.73%	31309
(10) rule accuracy >= 98%	96.30%	60.62%	45839

Note: all of the applicability values are set to rule frequency >= 3.

Table 2. *The experimental results of unknown word detection on the testing corpus.*

The results show that there is a tradeoff between precision and recall rate, but that the overall performance was much better than that of the handcraft rules written by human experts. They examined the training corpus and wrote up a rule set for proper-characters to the best of their ability. The handcraft rules had a precision rate of 39.11% and a recall rate of 81.45%, which are much lower than the rule set, made using the corpus-based rule learning method. The syntactic complexity of monosyllabic words was the reason for the lower coverage of the handcraft rules. Some handcraft rules are shown in Appendix 4. It is clearly shown that the handcraft rules suffer from low accuracy because a limited number of rules can be derived and the rules are usually too general to achieve high precision rates. There were only 139 hand-crafted rules while the proposed method generated thousands of rules as shown in Table 2. The number of rules selected increased with the decrement of the accuracy of the rule selection criteria because more rules satisfied the lower accuracy requirement. However, the number of rules after the screening process was lower in accordance with the decrement of the accuracy of the rule selection criteria. For instance, 207059 and 210552 rules were selected, respectively, for the rule accuracy criterion of 98% and 95%, but after the screening process, the number of rules was 45839 and 31309. The reason for this is that achieving a higher accuracy requires more contextual dependency rules to discriminate between proper-characters and improper-characters. On the other hand, a lower accuracy

requirement may cause the inclusion of many more short rules. This causes a lot of long rules to be subsumed by shorter rules eliminated during the screening process.

4. Conclusion and Future Research

The corpus-based learning approach proved to be an effective and easy method for finding unknown word detection rules. The advantages of using a corpus-based method are as follows:

a. The syntactic patterns of proper-characters are complicated and numerous. It is hard to hand-code each different pattern, yet most high frequency patterns are extractable from the corpus.

b. The corpus provides standard reference data not only for rule generation, but also for rule evaluation. The hand-craft rules can also be evaluated automatically and incorporated into the final detection rule set if the rule has a high accuracy rate.

c. It is easy to control the balance between the precision and the recall of the detection algorithm since we know the performance of each detection rule based on the training corpus.

Different types of unknown words have different levels of difficulty in identification. The detection of compounds is the most difficult because some of their morphological structures are similar to common syntactic structures. The detection of proper names and typographical errors is believed to be easier because of their irregular syntactic patterns. The results with respect to different types of syntactic categories were checked. Appendix 3 shows that the recall rates of proper names (i.e., category Nb) were less affected by the higher precision requirement. There was no data for typographical errors, but the detection of typographical errors is believed to be similar to the detection of proper names; that is, a higher precision can be achieved without sacrificing the recall rate. If parallel corpora with and without typographical errors are available, the corpus-based rule learning method can also be applied to the detection of typographical errors in Chinese.

After the unknown word detection process, an identification algorithm will be required to find the exact boundaries and the part-of-speech of each unknown word. This will require future research. Different types of rules will be required in identifying different compounds and proper names. The corpus can still play an essential role in the generation of rules and in their evaluation.

Acknowledgments

The authors wish to thank Dr. Charles Lee and the anonymous reviewers for their useful comments on this paper.

References

Brill, Eric, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," *Computational Linguistics* 21.4 (1995), pp.543-566.

Chang, J. S., C. D. Chen, & S. D. Chen, "Word Segmentation through Constraint Satisfaction and Statistical Optimization," *Proceedings of ROCLING IV* (1991), pp. 147-165.

Chang, C. H., "A Pilot Study on Automatic Chinese Spelling Error Correction" *Communication of COLIPS*, 4.2 (1994), pp.143-149.

Chang J. S.,S.D. Chen, S. J. Ker, Y. Chen, & J. Liu,1994 "A Multiple-Corpus Approach to Recognition of Proper Names in Chinese Texts", *Computer Processing of Chinese and Oriental Languages*, 8.1(1994), pp.75-85.

Chen, H.H., & J.C. Lee, "The Identification of Organization Names in Chinese Texts", *Communication of COLIPS*, 4.2(1994), pp. 131-142.

Chen, K.J., C.R. Huang, L. P. Chang & H.L. Hsu, "SINICA CORPUS: Design Methodology for Balanced Corpora," *Proceedings of PACLIC 11th Conference* (1996), pp.167-176.

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

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

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

Huang, C.R., K.J. Chen, & Li-Li Chang, "Segmentation Standard for Chinese Natural Language Processing," *International Journal of Computational Linguistics and Chinese Language Processing* 2.2 (1997), pp. 74-62.

Lee,H.J. & C.L. Yeh, "Rule-based Word Identification for Mandarin Chinese Sentences- A Unification Approach," *Computer Processing of Chinese and Oriental Languages*, 5.1 (1991), pp. 97-118.

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

Liu S. H., K. J. Chen, L.P. Chang, & Y.H. Chin, "Automatic Part-of-Speech Tagging for Chinese Corpora," *Computer Processing of Chinese and Oriental Languages*, 9.1(1995), pp.31-48.

Sproat, R., C. Shih, W. Gale, & N. Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational Linguistics*, 22.3 (1996), pp.377-404.

Sun, M. S., C.N. Huang, H.Y. Gao, & Jie Fang, "Identifying Chinese Names in Unrestricted Texts", *Communication of COLIPS*, 4.2 (1994), pp. 113-122.

Appendix 1. Samples of rule patterns.

rule	frequency	error	accuracy
{的}	165980	64	99.71 %
{是}	41089	78	98.10 %
{也}	16066	11	99.31 %
{她}	6185	4	99.35 %
{這}	5046	1	99.80 %
{或}	4582	3	99.34 %
{該}	2302	2	99.13 %
{(T)}	177641	177	99.00 %
{(Nh)}	73034	344	99.53 %
{(Caa)}	46659	392	99.16 %
{(SHI)}	41089	78	99.81%
{(Dfa)}(VH)	11037	39	99.65 %
{(Nh)}(Na)	6640	62	99.07 %
{(P)}(Nh)	6247	52	99.17 %
{(Nep)}(Na)	4030	26	99.35 %
(Na){(VCL)}	8062	299	96.30 %
(VC){(Di)}	4155	76	98.18 %
(VE){(VJ)}	1884	46	97.56 %
(VJ){(VJ)}	1489	53	96.44 %
(VJ){(Dfa)}	1004	5	99.50 %
{與}(Na)	3933	6	99.85 %
{及}(Na)	2831	18	99.36 %
{在}(VC)	2451	2	99.92 %
(VH){地}	1787	14	99.22 %
(VC){者}	1731	1	99.94 %
(Na){很}	1172	0	100 %
{再}(VC)(Na)	221	0	100 %
{令}(Na)(VH)	200	0	100 %
{各}(Na)(Na)	190	3	98.42 %
{極}(VH)(T)	187	1	99.47 %
(Na)(Dfa){高}	263	0	100 %
(Na)(VH){地}	248	1	99.60 %
(Na)(Na){時}	231	2	99.14 %
(T)(Na){則}	174	0	100 %
{會}覺得	139	1	99.28 %
{才}知道	124	0	100 %
{拿}著	121	0	100 %
{迄}今	117	0	100 %
的{話}	1406	2	99.86 %
並{非}	319	0	100 %

Appendix 2. Samples of testing results.

The first line contains the original text. The second line shows the result of word segmentation and pos tagging. The third line is the result of unknown word detection, where improper-characters are marked with '(?)'.

```
**********
有的時候我想吃點美國菜，
有的 (Nepa) 時候 (Na) 我 (Nh) 想 (VE) 吃 (V) 點 (Na) 美國 (Nc) 菜 (Na),
有的 ()(Nepa) 時候 ()(Na) 我 ()(Nh) 想 ()(VE) 吃 ()(V) 點 ()(Na) 美國 ()(Nc) 菜 (?)(Na) ,
**********
微軟過去兩年也推出了近百種新產品，
微 (D) 軟 (VH) 過去 (Nd) 兩年 (DM) 也 (D) 推出 (VC) 了 (VJ) 近百種 (DM) 新 (VH) 產品 (Na),
微 ()(D) 軟 (?)(VH) 過去 ()(Nd) 兩年 ()(DM) 也 ()(D) 推出 ()(VC) 了 ()(VJ) 近百種 ()(DM) 新 ()(VH) 產品 ()(Na),
**********
即使營收和獲利成長開始減慢，
即使 (Cbb) 營收 (Na) 和 (Caa) 獲利 (VH) 成長 (VH) 開始 (VL) 減 (VJ) 慢 (VH) ，
即使 ()(Cbb) 營收 ()(Na) 和 ()(Caa) 獲利 ()(VH) 成長 ()(VH) 開始 ()(VL) 減 (?)(VJ) 慢 ()(VH) ，
**********
一九九四將是日本教育的改革年，
一九九四 (DM) 將 (D) 是 (SHI) 日本 (Nc) 教育 (VC) 的 (T) 改革 (VC) 年 (Nf) ，
一九九四 ()(DM) 將 ()(D) 是 ()(SHI) 日本 ()(Nc) 教育 ()(VC) 的 ()(T) 改革 ()(VC) 年 (?)(Nf) ，
**********
日本可能出現第一個個人主義世代。
日本 (Nc) 可能 (D) 出現 (VH) 第一個 (DM) 個 (Nf) 人 (Na) 主義 (Na) 世代 (Na) 。
日本 ()(Nc) 可能 ()(D) 出現 ()(VH) 第一個 ()(DM) 個 (?)(Nf) 人 (?)(Na) 主義 ()(Na) 世代 ()(Na) 。
**********
筑波大學延請七三年諾貝爾物理學獎得主江崎出任校長，
筑 (BOUND) 波 (Nf) 大學 (Nb) 延請 (VC) 七三年 (DM) 諾貝爾 (Nb) 物理學 (Na) 獎 (Na) 得主 (Na) 江 (Na) 崎 (BOUND) 出任 (VG) 校長 (Na) ，
筑 (?)(BOUND) 波 (?)(Nf) 大學 ()(Nb) 延請 ()(VC) 七三年 ()(DM) 諾貝爾 ()(Nb) 物理學 ()(Na) 獎 (?)(Na) 得主 ()(Na) 江 (?)(Na) 崎 (?)(BOUND) 出任 ()(VG) 校長 ()(Na) ，
**********
就連整個體系中最官僚的教育當局——日本文部省，
就 (Da) 連 (D) 整個 (DM) 體系 (Na) 中 (Ng) 最 (Dfa) 官僚 (Na) 的 (T) 教育 (VC) 當局 (Na) — (BOUND) 日本 (Nc) 文 (BOUND) 部 (Nc) 省 (Nc) ，
就 ()(Da) 連 ()(D) 整個 ()(DM) 體系 ()(Na) 中 ()(Ng) 最 ()(Dfa) 官僚 ()(Na) 的 ()(T) 教育 ()(VC) 當局 ()(Na) — ()(BOUND) 日本 ()(Nc) 文 (?)(BOUND) 部 (?)(Nc) 省 (?)(Nc) ，
**********
也在調整一向溫吞的改革步伐。
也 (D) 在 (VCL) 調整 (VC) 一向 (D) 溫 (VHC) 吞 (VC) 的 (T) 改革 (VC) 步伐 (Na) 。
也 ()(D) 在 ()(VCL) 調整 ()(VC) 一向 ()(D) 溫 (?)(VHC) 吞 (?)(VC) 的 ()(T) 改革 ()(VC) 步伐 ()(Na) 。
**********
業者可以更準確地捕捉各個特定人口群，
業者 (Na) 可以 (D) 更 (D) 準確 (VH) 地 (Na) 捕捉 (VC) 各個 (DM) 特定 (A) 人口 (Na) 群 (Nf) ，
業者 ()(Na) 可以 ()(D) 更 ()(D) 準確 ()(VH) 地 ()(Na) 捕捉 ()(VC) 各個 ()(DM) 特定 ()(A) 人口 ()(Na) 群 (?)(Nf) ，
**********
```

Appendix 3. The detailed performance results for the different rule sets.

The first column shows the categories of unknown words.

The second column is the number of occurrences of the unknown words in the category shown in column one.

The third column is the recall rates of the unknown words detected under different rule sets.

Category	# of Unknown Words	Frequency > 2 Accuracy >= 55%	60%	70%	80%	90%	95%	98%
A	63	66.67%	66.67%	66.67%	74.60%	79.37%	87.30%	96.83%
Na	1396	75.07%	76.29%	79.87%	85.24%	92.12%	95.85%	97.13%
Nb	1511	87.16%	87.56%	90.47%	95.90%	98.28%	99.47%	99.60%
Nc	424	67.92%	67.92%	74.76%	75.94%	89.86%	91.04%	95.52%
Nd	24	16.67%	16.67%	25.00%	37.50%	50.00%	79.17%	83.33%
Nh	62	4.84%	4.89%	35.48%	75.81%	88.71%	90.32%	93.55%
VA	151	31.79%	32.45%	34.44%	54.30%	69.54%	83.44%	86.76%
VB	25	20.00%	20.00%	24.00%	40.00%	64.00%	84.00%	84.00%
VC	439	14.58%	14.58%	20.05%	41.91%	73.13%	89.29%	94.99%
VCL	63	14.29%	14.29%	15.87%	36.51%	79.37%	90.48%	96.83%
VD	48	2.08%	2.08%	8.33%	56.25%	77.08%	89.58%	93.75%
VE	70	4.29%	4.29%	4.29%	12.86%	24.29%	78.57%	88.57%
VG	69	7.25%	7.25%	10.15%	21.74%	40.58%	69.57%	86.96%
VH	137	22.65%	24.09%	35.77%	60.58%	73.72%	84.67%	89.78%
VHC	23	91.30%	91.30%	91.30%	95.65%	95.65%	95.65%	95.65%
VJ	67	8.96%	8.96%	11.94%	25.37%	44.78%	67.16%	83.58%
Total:	4572							
Recall:		63.32%	63.89%	68.18%	77.34%	87.40%	93.66%	96.30%
Precision:		73.70%	73.73%	74.61%	73.27%	68.74%	64.73%	60.63%

Appendix 4. Some examples of the handcraft rules.

The items in the curly brackets match a proper-character and the items in the round brackets match its context according to their linear order. The symbol ',' in the rules denotes an 'or' relation.

1. (之 , 的){ Na, Nc }

2. (Di){ Na, Nc }(DE)

3. (之){ VH }

4. (P, Da, Dk, D, Neqa, DM){ VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, SHI }

5. { Da, Dk, D, Neqa }(VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK,

VL, V_2, SHI)

6. (Dfa){ VH, VI, VJ, VK, VL }

7. { Dfa }(VH, VI, VJ, VK, VL, VE, D)

8. (VH, VI, VJ, VK, VL){ Dfb }

9. { VA, VCL }(在 , 至 , 自 , 離 , 經)

10. (VA , VCL){ 在 , 至 , 自 , 離 , 經 }

11. { Na, Nc, Ncd, Nd, Neu, Nes, Nep, Neqa, Neqb, Nf, Ng, Nh, VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, SHI, DM }(Ng, Ncd)

12. (Na, Nc, Ncd, Nd, Neu, Nes, Nep, Neqa, Neqb, Nf, Ng, Nh, VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, DM) { Ng , Ncd }