

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

Keh-Jiann Chen, Ming-Hong Bai

Institute of Information Science

Academia Sinica

Taipei, Taiwan

e-mail: kchen@iis.sinica.edu.tw, evan@iis.sinica.edu.tw

Abstract

One of the most prominent problems in computer processing of 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 by 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 the 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). However most of the papers dealing with the problem of word segmentation focus their attention 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. Unknown words cause segmentation errors, because out-of-vocabulary words in an input text normally would be incorrectly segmented into pieces of single character word or shorter words. It is difficult to know when an unknown word is encountered since all Chinese characters can either be a morpheme or a word and there are no blanks to mark the 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. Compound words and proper names are the two major types of unknown words. There are many different types of compounds, such as nominal compounds, verbal compounds, determiner-measure compounds, numbers, reduplications etc. It is neither 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 by 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 etc. [Chang etc. 94] used statistical methods to identify personal names in Chinese text which achieved a recall rate of 80% and a precision rate of 90%. Similar experiments were reported in [Sun etc. 94]. Their recall rate was 99.77%, but with a lower precision of 70.06%. Both papers deal 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]. They used mutual information between a character and its neighboring words to detect spelling errors and then to automatically make the necessary corrections. The error detection process achieved a recall rate of 76.64% and a precision rate of 51.72%. Lin etc. [Lin etc. 93] made 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 claimed accuracy rates of different papers vary due to different segmentation standards. In this paper we use the Sinica corpus as a standard reference data. The Sinica corpus is a word-segmented corpus based on the Chinese word segmentation standard for information processing proposed by ROCLING [Huang etc. 96, Chen etc. 96]. Therefore it contains many occurrences of unknown words which are separated by the blanks. The corpus were utilized for the purposes of training and testing. For the 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. 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, the unknown word would be detected first, and a search for the best matching words would be 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 the typographical errors are considered as a special kind of unknown words). The problems of unknown word identification and typographical error correction will be left for future research. The unknown word detection problem and the dictionary-word detection problem are complementary problem, since if all known words in the input text can be detected, then the rest of character string would 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(1.b) shows the result of the word segmentation process on (1.a).

(1) 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，

b. 筑波大學延請七三年諾貝爾物理學獎得主江崎出任校長，

According to an examination of a testing data which is a part of Sinica corpus, there are 4572 occurrences out of 4632 unknowns which were incorrectly segmented into sequence of shorter words and each sequence contains at least one monosyllabic word. That is, 60 of the unknown words were segmented into sequences of multi-syllabic words only. Therefore, the 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 that compounds or proper names are composed by several multi-syllabic words. Therefore the processes of detecting unknown words is equivalent to making the distinction between monosyllabic words and monosyllabic morphemes which are part 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 was performed, the precision of the prediction is very low. When the word segmentation process on the Sinica corpus by a conventional dictionary look-up method, 69733 occurrences of monosyllabic words were found, but only 9343 were part of unknown words, a precision of 13.40%. In order to improve the precision, the monosyllabic words which properly fit in 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 identify contextually-proper monosyllabic words is introduced. In section 3, the experimental results are presented which includes 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, 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 94] or by heuristic methods[Chen 92, Lee 91]. Either method can achieve very satisfactory results. Both have an accuracy 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 their respective morphological rules during the word segmentation process[Chen 92, Lin 93]. The purpose of the second step, part-of-speech (pos) tagging, is for the convenience of step3 and the future process of unknown word identification. After pos tagging, sentence (1.b) becomes sentence (2); each word contains a unique pos.

(2) 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) , .

Although the pos sequence may not be 100% correct, it is the most probable pos sequence in the terms of pos bi-gram statistics[Liu 95]. The details of the first two steps is not the major concerns of this paper. The focus 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 use the term 'improper-character' to denote a contextually-improper monosyllabic word which might be part of an unknown word. The way to identify proper-characters is by checking 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 the character is a bound-morpheme, it will be considered possibly belonging to unknown word. However almost every 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 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 how 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 relax, 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 hand-crafted rules is difficult, because there are more than 5000 commonly used Chinese character 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. There are two types of unknown words. The type one unknown words contain monosyllabic morphemes. The type two unknown words are composed with multi-syllabic words only. Only the detection of the type one unknown word is considered here, since the occurrences of the type two unknown words are very rare as we mentioned before.

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

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

An unknown word is considered successfully detected, if any one of its component is detected as an improper-character. It is noticed 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 detection. 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 the instances of unknown words. Segmenting the unknown words by a dictionary look-up method produces the 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 represent as a triplet (rule pattern, # of proper instances, # of improper instances). 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)}',

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 rest 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. 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 by sequentially choosing the rules which has the highest accuracy with applicability greater than a threshold value. The selected rule set is used as the recognition rules for proper-characters. The characters without a match by any one of the rules are considered as candidates of improper-character.

Rule selection algorithm:

1. Determine the threshold values for rule accuracy and applicability.

For each rule R_i , when applied on the training corpus, the rule accuracy(R_i) = M_i / T_i , where M_i is the # of instances of matches of R_i with proper characters; T_i is the total # of matches of R_i . The rule applicability(R_i) = T_i .

2. Sequentially select the rules with the highest rule accuracy and the 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 less improper-characters would 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. However 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 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 adjusted lower to an optimal point. A balance between recall and precision should be considered.

In the next section, the experimental results on the 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 the longer rules. 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 R_i is subsumed by rule R_j , then pattern of R_i is a sub-string of pattern R_j . For example the rule '{的}' is more general than the rule '{的} (Na)'.

Screening Algorithm:

- a. Sort the rules according to their string patterns in increasing order, resulting in rules
 $R_1 \dots R_n$.
- b. For i from 1 to n ,
 if there is a j such that $j < i$, and R_j is a sub-string of R_i , then remove R_i .

3. Experimental Results

The corpus-based learning method for unknown word detection was tested on the Sinica corpus which is a balanced Chinese corpus with segmented words tagged with pos [Huang 95, Chen 96]. 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 the 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 not identified by the word segmentation program as a foreign word (for instance English,) a number, or a reduplicated compound. There were 53328 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'; '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 the bi-word compounds did not affect the results too 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. 64 times "的" occurred 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. At the initial stage, 1455633 rules were found. After eliminating the rules with frequency less than 3, 215817 rules remained. At 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 by a dictionary look-up method and 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. The characters without a match will be considered as part of an unknown word. The performance results of different rule sets are shown in Table 2 and the detail statistics are shown in Appendix 3.

The results show that there is a tradeoff between precision and recall rate, but the overall performance was much better than when hand-crafted rules written by human experts were used. The set of hand-crafted rules were written by linguists. They examined the training corpus and wrote up the rule set for proper-characters to the best of their ability. The hand-craft rules had a precision rate of 39.11% and a recall rate of 81.45% which are much lower than the rule set made by the corpus-based rule learning method. The syntactic complexity of monosyllabic words was the reason for the lower coverage of the hand-crafted rules. There were only 139 hand-crafted rules while the proposed method generated thousands of rules as shown in Table 2. The number of rule selected is increasing with respect to the decrement of the accuracy of rule selection criteria, because more rules will satisfy the lower accuracy requirement. However the number of rules after the screening process is decreasing in accordance with the decrement of the accuracy of the rule selection criteria. For instances there are 207059 number and 210552 number of rules selected respectively for the rule accuracy criterion of 98% and 95%, but after the screening process the number of rules become 70415 and 56020. The reason for this interesting fact is that to achieve a higher accuracy demands more contextual dependency rules to discriminate between proper-characters and improper-characters; on the other hand lower accuracy requirement may cause the inclusion of more shorter rules which eliminate a lot of longer rules subsumed by the shorter rules.

Rule selection criteria	Recall rate	Precision rate	# of rules after screening
(0) no rule applied	100%	13.40%	0

(1)	rule accuracy $\geq 55\%$	63.32%	73.69%	12996
(2)	rule accuracy $\geq 60\%$	63.89%	73.73%	15250
(3)	rule accuracy $\geq 65\%$	64.85%	74.04%	17875
(4)	rule accuracy $\geq 70\%$	68.18%	74.61%	18559
(5)	rule accuracy $\geq 75\%$	73.80%	74.36%	20191
(6)	rule accuracy $\geq 80\%$	77.34%	73.26%	23047
(7)	rule accuracy $\geq 85\%$	81.06%	71.52%	30097
(8)	rule accuracy $\geq 90\%$	87.40%	68.74%	36563
(9)	rule accuracy $\geq 95\%$	93.66%	64.73%	56020
(10)	rule accuracy $\geq 98\%$	96.30%	60.62%	70415

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

4. Conclusion and Future Research

The corpus-based learning approach proved to be an effective and easy method of finding the 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 patterns, yet most high frequency patterns are extractable from the corpus.
- b. The corpus provides a standard reference data not only for rule generation but also for rule evaluation. The hand-craft rules can also be evaluated automatically and be 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 difficulties in identifying them. 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 are 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), is less affected by the higher precision requirement. there was no data for typos, but the detection of typos is believed to similar to the detection of proper names; that is, a higher precision can be achieved without sacrificing the recall rate. If a parallel corpora with and without typos is available, the corpus-based rule learning method could 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 the rules and 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, 1995," Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," *Computational Linguistics* Vol.21, No. 4, pp.543-566.
- Chang, J. S., C. D. Chen, & S. D. Chen, 1991," Word Segmentation through Constraint Satisfaction and Statistical Optimization," Proceedings of ROCLING IV, pp. 147-165.
- Chang, C. H., 1994,"A Pilot Study on Automatic Chinese Spelling Error Correction" *Communication of COLIPS*, Vol.4 No. 2, 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*, Vol. 8, No. 1, 75-85.
- Chen, H.H., & J.C. Lee, 1994,"The Identification of Organization Names in Chinese Texts", *Communication of COLIPS*, Vol.4 No. 2, 131-142.

- Chen, K.J., C.R. Huang, L. P. Chang & H.L. Hsu, 1996, "SINICA CORPUS: Design Methodology for Balanced Corpora," *Proceedings of PACLIC 11th Conference*, pp.167-176.
- Chen, K.J. & S.H. Liu, 1992, "Word Identification for Mandarin Chinese Sentences," *Proceedings of 14th Coling*, pp. 101-107.
- 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.
- Huang, C. R. Et al.,1995,"The Introduction of Sinica Corpus," *Proceedings of ROCLING VIII*, pp. 81-89.
- Lee,H.J. & C.L. Yeh, 1991, "Rule-based Word Identification for Mandarin Chinese Sentences- A Unification Approach," *Computer Processing of Chinese and Oriental Languages*, Vol. 5, No. 1, 97-118.
- 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.
- Liu S. H., K. J. Chen, L.P. Chang, & Y.H. Chin, 1995, "Automatic Part-of-Speech Tagging for Chinese Corpora," *Computer Processing of Chinese and Oriental Languages*, Vol. 9, No. 1, 31-48.
- Sproat, R., C. Shih, W. Gale, & N. Chang,1994, "A Statistical Finite-State Word-Segmentation Algorithm for Chinese," *Proceedings of 32nd ACL Conference*.
- 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.

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

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 such that the improper-characters are marked with '(?)'.

有的時候我想吃點美國菜。

有的(Neqa) 時候(Na) 我(Nh) 想(VE) 吃(V) 點(Na) 美國(Nc) 菜(Na),
有的(Neqa) 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 detail performance results of 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 with the category shown in column one.

The third column is the number of unknown words detected.

The last column is the recall rate. 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%