A PRELIMINARY STUDY ON UNKNOWN WORD PROBLEM IN CHINESE WORD SEGMENTATION

Ming-Yu Lin, Tung-Hui Chiang and Keh-Yih Su

Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan 30043, R.O.C.

Abstract

Unknown word, in general, is the main factor that causes the performance of word segmentation to be unsatisfied. To recognize the words which are derived from highly productive morphemes, a set of 17 morphological rules is proposed in this paper to recognize those regular unknown words. In addition, an unknown word model is further proposed to deal with the unknown words of irregular forms such as proper name etc. With the unknown word resolution procedures, the error reduction rate of 78.34% in word and 81.87% in sentence are obtained in the task of smoothing technical manuals. To examine the procedures in more general task, a corpus of newspaper is also tested and the error reduction rate of 40.15% in word and 34.78% in sentence are observed.

1. Introduction

"Word" is the basic unit used in most Chinese information processing tasks, such as machine translation or spoken language processing. However, there is no obvious delimiter marker, except for some punctuation markers, to specify the boundaries of words. Therefore, word segmentation is essential in almost all Chinese language processing systems.

Several models for word segmentation were proposed in our previous work [Chia 92a], in which the comparisons between rule-based and statistics-based approaches were made. From that work, over 99% word segmentation accuracy rate was observed when there is not any unknown word in the corpus; while only 95-96% could be obtained in case unknown words existed. Unfortunately, in Chinese, many morphemes have high derivative abilities such that they can combine with other words or morphemes to form compounds or complex words. To enumerate all such kinds of words in the dictionary is impossible and impractical. What is more, many new words are generated every day, so it is very difficult to keep the dictionary up-to-date. Thus, the problems caused by unknown words are inevitable in processing Chinese information. Hence, how to identify unknown words is the most important issue in real Chinese language processing systems. Motivated by that, the focus is shifted to the study of unknown words in this paper.

There are two kinds of unknown words: one is *regular*, such as time, date, reduplication, etc.; while the other is *irregular*, such as proper names, compound nouns, of which the unknown words must be determined by their context instead of simple rules. Regular unknown words are likely to be predicted according to some morphological rules. However, the words of irregular forms are usually difficult to be identified from rules. They must be examined through the analysis with more high level knowledge sources, such as syntax and semantics.

In this paper, a set of 17 morphological rules are first introduced to tackle the problem caused by the regular unknown words. After applying the morphological rules, 1.81% error rate in word is observed. Compared with the error rate of 7.48% in the baseline system, it corresponds to 75.8% error reduction rate. Afterwards, an irregular unknown word model is proposed to recognize the irregular unknown words, with which 78.34% in error reduction is obtained.

This paper is organized as follows. The system architecture, the databases including the lexicon, the morphological rules, and the tasks are described in Section 2. In Section 3, the overview of the baseline models, which were derived in our previous works [Chia 92a], are given. Then, the effects of the morphological rules are investigated in Section 4. In addition, we incorporate part of speech information into the system to explore the performance both in word and lexical tag in Section 5. Furthermore, an unknown word model is proposed in Section 6 to resolve some problems caused by the unknown words in irregular forms. Finally, a summary is addressed in the last section.

2. System Architecture

The flow of the word segmentation in our system is shown in Figure 1. The system consists of four phases of processes, including the baseline word segmentation, morphological analysis, tagging, and unknown words identification. The input character string is first processed by the baseline segmentation model, in which all possible segmentation patterns are generated by looking up the dictionary and assigned the corresponding preference scores depending on the model used. Then the best N (N is set to 10 in the current implementation) word hypothesis sequences are passed to the morphological analyzer. The morphological rules are then employed to detect some particular forms of unknown words in this phase. Again, the top N candidates are output for being tagged with their lexical tags. Afterwards, the best tagged result is dispatched into the unknown word module to examine other types of unknown words. Finally, the best hypothesis is picked up as the final output.

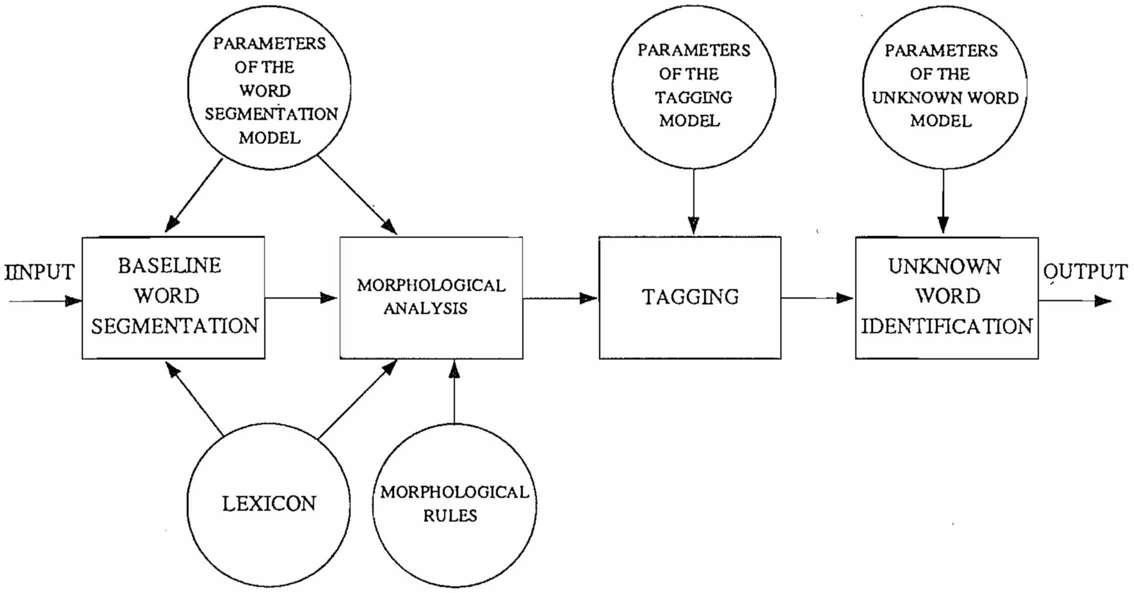

PARAMETERS OF THE WORD SEGMENTATION MODEL
PARAMETERS OF THE TAGGING MODEL
PARAMETERS OF THE UNKNOWN WORD MODEL
IINPUT
BASELINE WORD SEGMENTATION
MORPHOLOGICAL ANALYSIS
TAGGING
UNKNOWN WORD IDENTIFICATION
OUTPUT
LEXICON
MORPHOLOGICAL RULES

FIGURE 1
The block diagram of the system architecture.

Lexicon

The electronic dictionary used in our system is provided by Behavior Design Corporation (BDC) [BDC 92], in which there are 89,590 entries of definition. For each word, the possible lexical tags that can be attached to it are encoded in the dictionary. Currently, there are 49 different categories of tags used in the dictionary. The statistics of the dictionary are listed in Table 1.

# of characters / word	# of entries
1	1,734
2	35,492
3	19,650
4	24,054
5	6,140
6	2,020
>= 7	500
Total	89,590

TABLE 1
The statistics of the dictionary.

Morphological Rules

There are 17 morphological rules [Lin 93] available in the system which are written by linguistic experts according to a large corpus. Two of these morphological rules are only related to some particular affixes. The rest 15 rules, on the contrary, must refer to the lexical tags. All these morphological rules are listed in appendix A.

Corpus

To evaluate the performance of different segmentation models, a corpus of 9,677 sentences extracted from technical manuals are collected. This corpus is further divided into a training corpus of 7,742 sentences, i.e., 4/5 of the original set, and a testing corpus of the remaining 1,935 sentences in the following simulations. Along with the simulations performed in [Chia 92a], an *ideal* corpus is formed by extracting the sentences which contain unknown words out of the original corpus. Therefore, the original corpus is also called the *real corpus* in contrast. The effect of using the proposed models both in the ideal and the real corpora are investigated and compared in the paper. The statistics of the corpora are listed in Table 2.

	Ideal Corpus		**Real Corpus**	
	Training Set	Testing Set	Training Set	Testing Set
# of sentences	3,711	911	7,742	1,935
# of words	37,720	9,238	87,715	21,964
# of characters	62,423	15,374	148,221	37,261
Ave. # of words /sentence	10.16	10.14	11.33	11.35
Ave. # of characters /sentence	16.82	16.88	19.15	19.26

TABLE 2
The statistics of the corpora.

3. Overview of the Baseline Model

Since the baseline models have been derived in our previous work [Chia 92a], instead of repeating the detail derivations of those models, only the final forms of the computational models are listed in this paper.

Let c_1^n denote the input character sequence of n Chinese characters and $W_i = w_{i,1}, w_{i,2}, \cdots, w_{i,M_i}$ be the i-th word segmentation pattern, where M_i denotes the total number of words in W_i, the model derived in [Chia 92a] is summarized as follows:

$$\underset{w_{i,1}^{i,M_i}}{\operatorname{argmax}} \left\{ \prod_{k=1}^{M_i} P\big(w_{i,k} \mid l_{i,k-1}\big) \right\}, \tag{1}$$

where $l_{i,k-1}$ denotes the length, i.e., the number of characters, of $w_{i,k-1}$. In other words, the correlation of the word and the length of its left contextual word is considered in the model.

To compare the common rule-based approach with the baseline models, the approach using the rule that the longest word is most preferred is also implemented in this simulation. The results for these approaches both in the real and ideal corpora are shown in Table 3.

	Error rate in the training set		**Error rate in the testing set**	
Model	word (%)	sentence (%)	word (%)	sentence (%)
Max. Match	2.15	9.84	2.63	11.09
$P(w_k \mid l_{k-1})$	0.12	0.62	0.69	2.63

(a)

	Error rate in the training set		**Error rate in the testing set**	
Model	word (%)	sentence (%)	word (%)	sentence (%)
Max. Match	8.74	56.74	9.47	58.14
$P(w_k \mid l_{k-1})$	6.91	52.34	7.48	54.16

(b)

TABLE 3
The results of various word segmentation models in (a) the ideal corpus and (b) the real corpus.

Comparing the results in Table 3(a) and 3(b), it is apparent that the existence of unknown words is the main issue which causes the performance to degrade evidently. The results performed in the real corpus are 6.79% worse in word accuracy, and 51.53% degradation in sentence accuracy compared with those in the ideal case. After analyzing the errors caused by these models, two kinds of error patterns are founded. The first one is the mis-combined error, denoted by **s_ns**, such as |一 |個 |人 |→ |一 |個人 |, where two or more words which should be separated are regarded as a word. The other pattern, denoted by **ns_s**, is the over-segmentation error, where a word is mis-segmented apart into several morphemes or words, such as |轉換器 |→ |轉換 |器 |. The statistics of these two error patterns for the baseline models in the real corpus are listed in the following table.

	Errors in the training set		**Errors in the testing set**	
Models	s_ns	ns_s	s_ns	ns_s
$P(w_k \mid l_{k-1})$	260	5,904	109	1,533

TABLE 4
The statistics of the error patterns for the baseline model in the real corpus.

In Table 4, it is obvious that the error is caused mainly from over-segmentation of words. Therefore, to combine those over-segmented words into a word will improve the performance effectively. To do this, the approaches with the morphological rules and an unknown word model are introduced later in this paper.

4. The Morphological Analysis

As mentioned above, many morphemes in Chinese have high derivative capability so that they can combine with other words or morphemes to form new words, such as 總，化 . Therefore, the words formed in such a way are unable and impractical to be enumerated in the lexicon. Since the word formation processes associated with those morphemes are quite regular, they are; therefore, predictable according to a few rules. Motivated by this concern, a set of morphological rules are introduced in our system. Currently, there are 17 morphological rules in the system. They are divided into two parts according to whether part of speech is applied or not. The first part consists of two morphological rules which only relate to some particular affixes. On the other hand, the remaining 15 rules in the second part are applied with the lexical tags. Interested readers for the morphological rules are referred to appendix A or [Lin 93].

Like most rule-based approaches, the use of the morphological rules will results in the problem of redundancy, and inconsistency. Those redundant and, especially, the inconsistent rules have to be withdrawn from the rule-base to improve the performance of the system both in terms of the accuracy and efficiency. In this paper, a sequential forward selection method is used in rule ordering, which will be described in the following subsection.

4.1. Rule Ordering

To examine the effectiveness of the morphological rules, the sequential forward selection (SFS) procedure [Devi 82, Liu 93] is applied to determine the ordering of morphological rules. SFS is a simple bottom up search procedure where one rule at a time is added to the current rule set. At each stage, the rule to be included in the rule set is selected from the remaining available rules, so that the new enlarged set of the rules yields a maximum value of the criterion function used. The rule ordering procedure with the SFS is shown as follows.

Assume that G1 is the original rule set and G2 is the set including the rules which are ordered through the SFS algorithm. Initially, G1 consists of all morphological rules and G2 is an empty set.

```
SFS(n rules) {
        G1= {n rules};                                          /* initialization for G1 set */
        G2= {};                                                                             /*
        initialization for G2 set */
        /* the loop of moving the best rule in G1 to G2 */
        loop( while there is any rule in G1) {
```

```
        mincost=minimum_value;                               /* initialize the variable
        for minimum cost */
        /* the loop of computing the cost of embodying each of rules in G1 to G2 */
        loop ( for each rule_i in G1) {
            cost=WordSegmentation(corpus, {rule_i}+G2);
                    /* computing the cost returned by word segmentation procedure for using
                       the new rule set which is composed of rule_i and those rules in G2 */
            /* find the rule with minimum cost */
            if (cost<mincost) then
                swap(cost,mincost);                          /* swap the minimum cost with the
                current one */
                best_rule=rule_i;                            /* current rule is assigned to be
                the best one */
            endif
        }
        move_rule(G1,G2,best_rule)                           /* move the best
        rule from G1 to G2 */
    }
}
```

ILLUSTRATION 1
The rule ordering procedure with the sequential forward selection algorithm.

Note that the cost function returned by the WordSegmentation() function is computed according to the following formula:

$$cost = w_r \times (1 - P_r) + w_p \times (1 - P_p), \tag{2}$$

where $P_r = \frac{No.\,of\,words\,identified\,correctly}{No.\,of\,words\,in\,the\,corpus}$, named as the *recall rate*, is the percentage of the words in the corpus which are identified correctly by the system; $P_r = \frac{No.\,of\,words\,identified\,correctly}{No.\,of\,words\,identified\,by\,the\,system}$, known as the *precision rate*, is the percentage of the words identified by the system being correct; w_r, w_p are defined as the weights to the error rate of recall and precision respectively and they are both defined to be *0.5* in the following test. Thus, the performance of both the recall rate and the precision rate are taken into account through the cost function defined above.

Through the SFS procedure, the results of cost versus number of rules is illustrated in Figure 2. From this figure, it is noted that the cost is decreasing as the number of rules is increasing up to 11; however, the cost remains a constant as the rule number is in the range from 11 to 16. Finally, it increases slightly when the last rule is incorporated. After checking the results, we find that it is the rule n → q + 點 (時間) which results in the increment in cost. Therefore, it should be modified or withdrawn from the rule base. In addition, those rules that rank from 11-th to 16-th are never applied in the training corpus; however, they are remained in the rule base because they may be useful in the testing set.

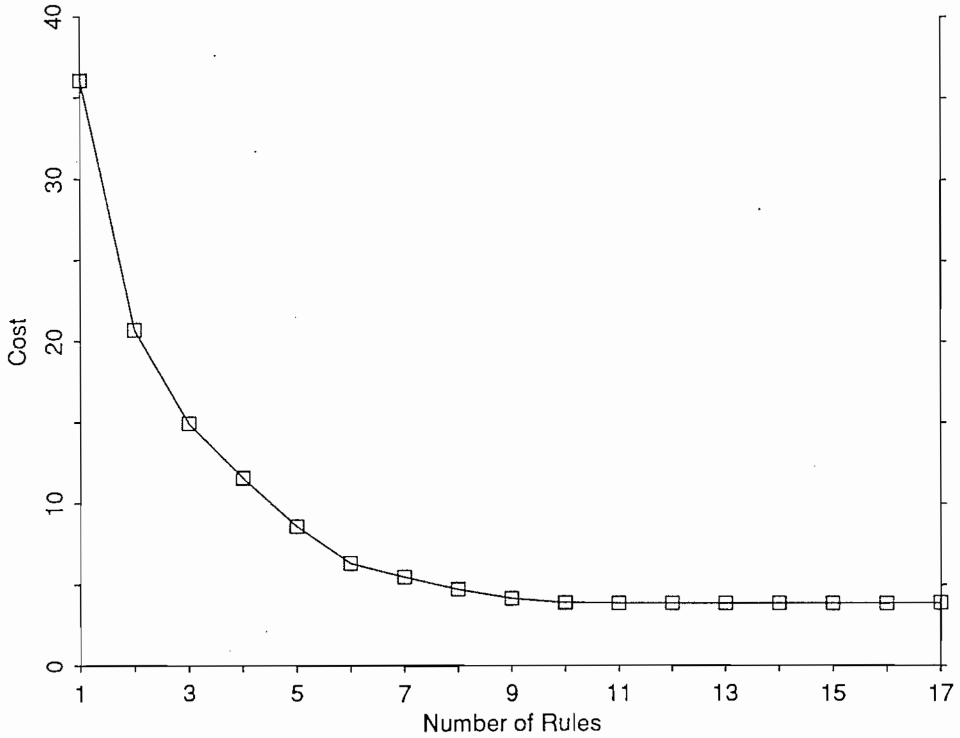


FIGURE 2
Illustration of the cost versus the number of rules through the rule ordering mechanism.

4.2. Summary of the Morphological Analysis

The results of various word segmentation models with the morphological analysis in the ideal corpus as well as the real corpus are shown in Table 5, where the values in the parentheses are the corresponding results of the baseline models.

Model	Error rate in the training set		Error rate in the testing set	
	word (%)	sentence (%)	word (%)	sentence (%)
$P(w_k \mid l_{k-1})$	0.80 (0.12)	4.04 (0.62)	1.44 (0.69)	6.26 (2.63)

(a)

Model	Error rate in the training set		Error rate in the testing set	
	word (%)	sentence (%)	word (%)	sentence (%)
$P(w_k \mid l_{k-1})$	1.32 (6.91)	8.86 (52.34)	1.81 (7.48)	10.70 (54.16)

(b)

TABLE 5
The results of the baseline word segmentation model with the morphological analysis in (a) the ideal corpus and (b) the real corpus.

It is observed that the performance in the ideal corpus degrades slightly. After applying the morphological rules, the possibility of mis-combining the words or morphemes which should be separated will inevitably increase. Therefore, the performance of the ideal corpus degrade. On the contrary, the situation of mis-combination is not so serious in the real corpus. In fact, the results are greatly improved with the morphological analysis, where it corresponds to the error reduction rate of 75.8% in word and 80.43% in sentence. The statistics of the error pattern for morphological analysis are listed in Table 6, where the corresponding results with the baseline models are tabulated in parentheses.

Models	Errors in the training set		Errors in the testing set	
	s_ns	ns_s	s_ns	ns_s
$P(w_k \mid l_{k-1})$	429 (260)	370 (5,904)	168 (109)	110 (1,533)

TABLE 6
The statistics of the error patterns in the real corpus after morphological analysis.

The result shows that the morphological analysis significantly reduces the errors caused by the over-segmentation, which is over 90%. Therefore, the performance is improved dramatically with the morphological rule approach. On the other hand, checking up the s_ns type of error in Table 6, it is observed that this approach has slight side-effect for increasing the mis-combination errors. Those mis-combinations are caused by unconditionally applying the morphological rules without regarding their contexts. For example, the mis-combination of |從 |研究到 |發展 | is caused by applying the rule " v → v(研究)+ 到 ".

To further decrease the error of mis-combination, those morphological rules should be accomplished with a context-sensitive framework, which is similar to the formulae for phrase structure rules in [Chia 92b]. It will be our future work and will not be discussed in this paper. Instead, we will pay attention to the unknown words which are formed irregularly and cannot be recognized through the morphological rules. Because lexical tags will be used as the parameters in the unknown word model, we will describe the tagging process in the next section before starting the unknown word modeling.

5. Tagging Part Of Speech

In the previous study [Chia 92a], we have shown that the incorporation of lexical information is useful in word segmentation. However, the morphological rules are applied before the tagging process. The introduction of the morphological analysis may result in changes of the formation of words or the lexical tags. Accordingly, the effect of the combination of the morphological and lexical knowledge sources is investigated in this section. To do this, we derive the word segmentation model which incorporates the lexical information as follows:

$$\widehat{W} = \underset{W_i}{\operatorname{argmax}} \sum_{T_{j,i}} P(W_i, T_{j,i} \mid c_1^n), \tag{3}$$

where $T_{j,i}$ stands for the j-th lexical sequence corresponding to the i-th word segmentation pattern W_i. To save the time for computation in the above equation, we approximate it in the following form:

$$\widehat{W} = \underset{W_i}{\operatorname{argmax}} \left\{ \max_{T_{i,j}} P(W_i, T_{i,j} \mid c_1^n) \right\}. \tag{4}$$

The term $P(W_i, T_{i,j} \mid c_1^n)$ in Eq.(4) is further derived as follows.

$$\begin{aligned} & P(W_i, T_{i,j} \mid c_1^n) \\ & = P(T_{i,j} \mid W_i, c_1^n) \times P(W_i \mid c_1^n) \\ & \approx P(T_{i,j} \mid W_i) \times P(W_i \mid c_1^n) \\ & = \frac{P(W_i \mid T_{i,j}) \times P(T_{i,j})}{P(W_i)} \times \frac{P(c_1^n \mid W_i) \times P(W_i)}{P(c_1^n)}. \end{aligned} \tag{5}$$

Note that the approximation in the above derivation is based on the fact that the lexical tags are attached only to words; therefore, it is assumed that tagging the part-of-speech is independent of the character string if the word sequence is given. In addition, since the character sequence can be determined uniquely if a word sequence is given, it causes that $P(c_1^n \mid W_i) = 1$ holds for all word segmentation patterns. Besides, the term $P(c_1^n)$is the same constant to each segmentation ambiguity and it does not affect the result in Eq.(4) if being neglected. Hence, the criterion in Eq.(4) is rewritten in the following form:

$$\widehat{W} = \underset{W_i}{\operatorname{argmax}} \left\{ \max_{T_{i,j}} P(W_i \mid T_{i,j}) \times P(T_{i,j}) \right\}. \tag{6}$$

Concerning the i-th word segmentation pattern $W_i = w_{i,1}, w_{i,2}, \cdots, w_{i,M_i}$ of M_i words, let $t_{i,j,k}$ denote the part-of-speech that is attached to $w_{i,k}$ in the j-th lexical sequence $T_{i,j} = t_{i,j,1}, t_{i,j,2}, \cdots, t_{i,j,M_i}$. To make the computation in Eq.(6) feasible, $P(W_i \mid T_{i,j})$ and $P(T_{i,j})$ are approximated as follows.

$$\begin{aligned} P(W_i \mid T_{i,j}) &= P\left(w_{i,1}^{i,M_i} \mid t_{i,j,1}^{i,j,M_i}\right) \\ &\approx \prod_{k=1}^{M_i} P_i(w_k \mid t_{j,k}). \end{aligned} \tag{7}$$

$$
\begin{aligned}
P(T_{i,j}) &= P_i\left(t_{j,1}^{j,M_i}\right) \\
&\approx \prod_{k=1}^{M_i} P_i\left(t_{j,k} \mid t_{j,k-1}\right).
\end{aligned}
\tag{8}
$$

It is noted that $P_i(\cdot)$, which denotes the probability function that relates to the i-th word segmentation pattern, is introduced in the above equations to prevent from notational confusion. Therefore, the word segmentation model with lexical knowledge incorporated is represented as the following formula:

$$
\underset{w_{i,1}^{i,M_i}}{\arg\max}\left\{\max_{t_{i,j,1}^{i,j,M_i}}\left[\prod_{k=1}^{M_i} P\left(w_{i,k} \mid t_{i,j,k}\right) \times P\left(t_{i,j,k} \mid t_{i,j,k-1}\right)\right]\right\}. \tag{9}
$$

The results with the morphological and lexical analysis in the ideal corpus as well as the real corpus are shown in Table 7, where the values in the parentheses are the results with the model $P(w_k \mid l_{k-1})$ before incorporating the lexical information.

	Error rate in the training set		Error rate in the testing set	
Model	word (%)	sentence (%)	word (%)	sentence (%)
$P(w_k \mid t_k) \times P(t_k \mid t_{k-1})$	0.69 (0.80)	3.48 (4.04)	1.45 (1.44)	7.14 (6.26)

(a)

	Error rate in the training set		Error rate in the testing set	
Model	word (%)	sentence (%)	word (%)	sentence (%)
$P(w_k \mid t_k) \times P(t_k \mid t_{k-1})$	1.24 (1.32)	8.58 (8.86)	1.74 (1.81)	11.16 (10.70)

(b)

TABLE 7
The word and sentence error rate of various word segmentation models with the morphological analysis in (a) the ideal corpus and (b) the real corpus.

From Table 7, it is observed that the results are improved slightly, except for the testing set in the ideal corpus. However, the improvement is not significant enough to show the superiority to incorporate the lexical information. Since the morphological rules are applied in a context-free manner, the errors of mis-combination resulting from the morphological analysis cannot be recovered even with a tagger. Besides, by using this tagging model, the number of parameters is much larger than those of the baseline models so that the over-tuning phenomena is more apparent. Hence, the results in the training set can be improved more significant than those in the testing set.

To couple the tagger into the system is, however, essential because the lexical information is required in the following unknown word model. To examine the effectiveness of the tagger, the error of the tagging process is also listed in Table 8. Note that in this paper, a correct tagging to a word is defined when both the word segmentation and the lexical tag are correct simultaneously.

Model	**Error rate in the training set**		**Error rate in the testing set**	
	tag (%)	sentence (%)	tag (%)	sentence (%)
$P(w_k \mid t_k) \times P(t_k \mid t_{k-1})$	7.56	46.86	8.92	51.70

(a)

Model	**Error rate in the training set**		**Error rate in the testing set**	
	tag (%)	sentence (%)	tag (%)	sentence (%)
$P(w_k \mid t_k) \times P(t_k \mid t_{k-1})$	7.77	51.52	9.50	58.40

(b)

TABLE 8

The tag error of the morphological analysis in (a) the ideal corpus and (b) the real corpus.

6. Unknown Word Modeling

The unknown words in the corpus can be categorized into the following classes.

1. The words should be contained in the dictionary, such as 若要，不僅，爭議，驚魂未定 . For the corpus of technical manuals, there are 263 words in the training set and 72 words in the testing set of this class; while in the newspaper corpus, 141 and 40 words in the training set and the testing set are categorized to this class respectively.
2. The words should be combined through the morphological rules, such as 牛肝，牛心 . For the technical manuals, there are 35 words in the training set and 7 words in the testing set of this class. In the newspaper corpus, 12 and 2 words in the training set and the testing set belong to this class respectively.
3. Abbreviations, such as 國大，立委，台獨 . No word in the corpus of technical manuals are classified to this class. However, for the newspaper corpus, there are 6 words in the training set and 2 words in the testing set of this class.
4. Proper nouns, biographical names, and geographical names, such as 德國聯邦共合國，大牛，胡適. There are 2 words in the training set and none in the testing set of this class. As to the newspaper corpus, 39 and 6 words in the training set and testing set belong to this class respectively.
5. Others: this class includes the words of typographical errors in the corpus, such as 吩附（咐） and missing lexical tags in the dictionary, such as 閣 . In addition, several words in the dictionary are in conflict with the principals of word formation

announced by Computational Linguistics Society R.O.C., which should be withdrawn from the dictionary.

Due to the incompleteness of the dictionary and the morphological rules, the words in class 1 and 2 are regarded as unknown words. They should be restored by renewing the dictionary or modifying the morphological rules. In this paper, the words in class 3 and 4 are what we are really interested in. Nevertheless, the words of class 1 and 2 will always appear unless a dictionary in unlimited size is available.

Since the morphological rules are written for detecting unknown words which are formed regularly, they cannot identify those words which are neither formed regularly nor able to be enumerated entirely in the dictionary. Therefore, a statistical model is further proposed in this paper to tackle the problems caused by these kinds of unknown words. In viewing the word segmentation results, several unknown words are segmented into a series of separate characters, such as| 國 | 大 | 黨 | 部 | 透露 | 層 | 峰 | 消息 |. In the current task, over 72% of the irregular type of unknown words belong to this case. Therefore, we will attack this kind of error in this paper. To deal with this kind of unknown words, only the region in which all words are of single character is considered to have the possibility of possessing a unknown word in our model. It means that the unknown words of length over than 2, such as 曼德拉 and 德國聯邦共合國 , are not taken into account currently.

To consider the region of interest R_u as shown below, which is composed of N_u separate characters, $w_1, w_2, \cdots, w_{N_u}$, with their corresponding tags $t_1, t_1, \cdots, t_{N_u}$, the contexts associated with this region are w_b, w_e, and their tags are t_b, t_e respectively. Here, we further assume that there is only one unknown word resides in the suspected region in the unknown word model. Accordingly, two decisions relating to this suspected region have to be made by the unknown word fixing strategy: (1) to decide whether there is any unknown word in R_u; (2) to determine the way of combination of the unknown word if the previous answer is "yes."

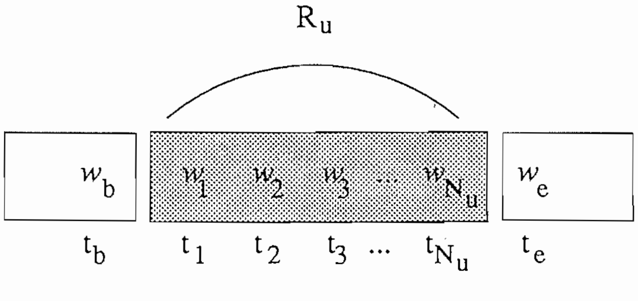


FIGURE 3
The suspected unknown word region.

To answer the first question, a likelihood ratio γ is defined as follows:

$$\gamma = \frac{P\Big(E_{uw} = 1 \mid (w_b, t_b), \Big(w_1^{N_u}, t_1^{N_u}\Big), (w_e, t_e)\Big)}{P\Big(E_{uw} = 0 \mid (w_b, t_b), \Big(w_1^{N_u}, t_1^{N_u}\Big), (w_e, t_e)\Big)}, \qquad (10)$$

where E_{uw} is an indicator; $E_{uw} = 1$ denotes the existence of unknown words, otherwise $E_{uw} = 0$. The number of parameters associated with the Eq.(10) are too many to be handled in practice. Hence, it is approximated as the following equation:

$$\gamma = \frac{P\left(E_{uw} = 1 \mid t_b, t_1^{N_u}, t_e\right)}{P\left(E_{uw} = 0 \mid t_b, t_1^{N_u}, t_e\right)} \approx \frac{\left[\prod_{i=0}^{N_u+1} P(t_i \mid t_{i-1}, E_{uw} = 1)\right] \times P(t_e, E_{uw} = 1)}{\left[\prod_{i=1}^{N_u+1} P(t_i \mid t_{i-1}, E_{uw} = 0)\right] \times P(t_e, E_{uw} = 0)}; \qquad (where\ t_0 = t_b,\ t_{N_u+1} = t_e). \tag{11}$$

Currently, it is regarded that there is an unknown word in the region if $\gamma > 1$; otherwise, the suspected region is considered without any unknown word.

If the suspected region is considered with an unknown word, each possible way of combination associated with the unknown word shown below is given a preference score according to a scoring function. To clearly describe this function, we take the second case (UW 2.2) for example. The score of the case (UW 2.2), where the unknown word is combined by w_2 and w_3, is defined as follows:

$$P\left(UW = w_2 w_3 \mid (w_b.t_b), \left(w_1^{N_u}, t_1^{N_u}\right), (w_e, t_e), E_{uw} = 1\right), \tag{12}$$

where $UW = w_2 w_3$ represents the event that the unknown word is composed of w_2 and w_3.

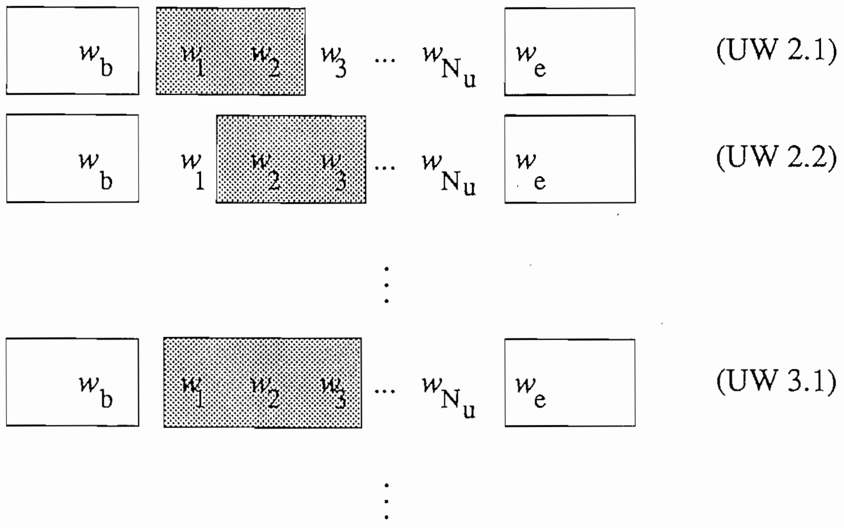


FIGURE 4
The possible types of combination with an unknown word existing in suspected region; the shaded regions indicate the possible positions and formations of the unknown word.

Again, Eq.(12) is too complex to make the computation feasible. In implementation, it is simplified as the following formula:

$$P\left(LT = t_1, UT = (t_2, t_3), RT = t_4, L_{uw} = 2 \mid t_b, t_1^{N_u}, t_e, E_{uw} = 1\right), \tag{13}$$

where L_{uw} is the random variable expressing the length (number of words to be combined) of the unknown word. In such a way, this model will only take into consideration the information related to the length of the unknown word, the lexical tags of its elementary word and its left and right contexts. Furthermore, Eq.(13) is derived as follows:

$$P\Big(LT = t_1, UT = (t_2, t_3), RT = t_4, L_{uw} = 2 \mid t_b, t_1^{N_u}, t_e, E_{uw} = 1\Big)$$

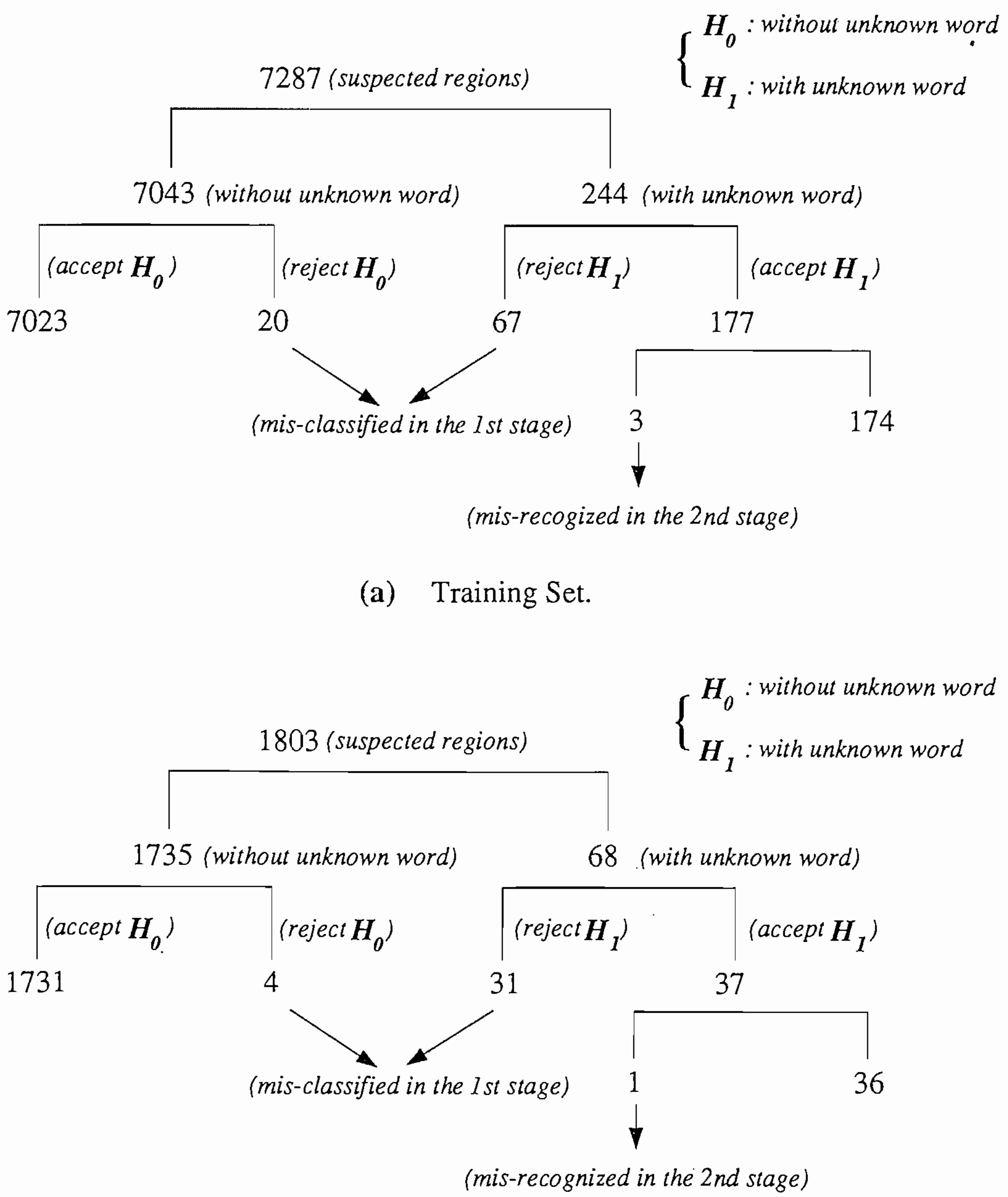
$$\begin{aligned} &= P\Big(RT = t_4 \mid UT = (t_2, t_3), LT = t_1, L_{uw} = 2, t_b, t_1^{N_u}, t_e, E_{uw} = 1\Big) \\ &\quad \times P\Big(UT = (t_2, t_3) \mid LT = t_1, L_{uw} = 2, t_b, t_1^{N_u}, t_e, E_{uw} = 1\Big) \\ &\quad \times P\Big(LT = t_1 \mid L_{uw} = 2, t_b, t_1^{N_u}, t_e, E_{uw} = 1\Big) \\ &\quad \times P\Big(L_{uw} = 2 \mid t_b, t_1^{N_u}, t_e, E_{uw} = 1\Big) \end{aligned} \tag{14}$$

$$\begin{aligned} &\approx P(RT = t_4 \mid UT = (t_2, t_3), L_{uw} = 2, E_{uw} = 1) \\ &\quad \times P(UT = (t_2, t_3) \mid LT = t_1, L_{uw} = 2, E_{uw} = 1) \\ &\quad \times P(LT = t_1 \mid L_{uw} = 2, E_{uw} = 1) \\ &\quad \times P(L_{uw} = 2 \mid E_{uw} = 1) \end{aligned}$$

In a similar way, the scores corresponding to the other types of the unknown word in Figure 4 can be computed by analogy.

Experimental Results and Discussions

In the training corpus, there are 336 irregular unknown words, in which there are 247 double-character words, 69 tri-character words, and the rest 20 words are composed of over 3 characters. That is, at most 247 unknown words can be possibly identified through the model described above, for only the case of double-character words being considered in the simulation. Meanwhile, there are 89 irregular unknown words in the testing set, where there are 71 double-character words. The examination of the unknown word model is illustrated as follows.


H_0 : without unknown word
H_1 : with unknown word
7287 (suspected regions)
7043 (without unknown word)
244 (with unknown word)
(accept H_0)
(reject H_0)
(reject H_1)
(accept H_1)
7023
20
67
177
(mis-classified in the 1st stage)
3
174
(mis-recogized in the 2nd stage)
H_0 : without unknown word
H_1 : with unknown word
1803 (suspected regions)
1735 (without unknown word)
68 (with unknown word)
(accept H_0)
(reject H_0)
(reject H_1)
(accept H_1)
1731
4
31
37
(mis-classified in the 1st stage)
1
36
(mis-recognized in the 2nd stage)

(a) Training Set.

(b) Testing Set.

ILLUSTRATION 2
The illustration of the error types in the unknown word modeling.

In the above illustration, the null hypothesis H_0 is defined as follows:
H_0: There is no unknown word in the suspected region.
H_1: There is at least one unknown word in the suspected region.
Note that there are 247 double-character words in the training set, but only 244 single word regions containing unknown words, it implies that at most three of the suspected region include more than one unknown word. Nevertheless, it is reasonable for us to assume that there is only one unknown word in a single word region.

From the above illustration, it is observed that 87 errors arise from the first stage by using Eq.(11) to inspect the suspected region; that is, 1.19% (87/7,287) error rate is for the first stage. Aside from that, there are additional 3 identification errors imposed using Eq.(14) in the second stage; it means that there is 1.69% (3/177) error rate introduced by the second stage. Thus, there are total 90 errors, which correspond to 1.23% (90/7,287) error rate, resulting from using the current model to identify the unknown words in the training set.

With the taxonomy described above, the unknown words of class 3 and class 4 are what you are really interested in. But there are only 2 words belonging to class 4, i.e., geographical names, none for class 3 in the training set; what is more, none of unknown words in the testing set belongs to class 3 or class 4. In view of the recognition of unknown words in the training set, 174 of the total 247 unknown words, i.e., 70.44%, are identified correctly. However, the rest 70 ones are missed, and another 20 mis-combined errors are imposed through the unknown word model. In the testing set, 36 of the 71 unknown words are recognized correctly; it corresponds to 50.7% recognition rate. According to the above analysis, it is apparent that the errors are mainly introduced from the first stage. Therefore, to improve the performance of the model in the future, Eq.(11) should be modified.

The progressive results of the unknown word recognition procedure are summarized in Table 9. Compared with the baseline model, the error reduction rate of 78.34% in word and 81.87% in sentence are obtained with the unknown word recognition procedure.

		Error rate in the testing set	
	Computational Model	word (%)	sentence (%)
BS	$P(w_k \mid l_{k-1})$	7.48	44.16
BS+MA	$P(w_k \mid l_{k-1})$	1.81	10.70
BS+MA+TG	$P(w_k \mid t_k) \times P(t_k \mid t_{k-1})$	1.74	11.16
BS+MA+TG+UW	unknown word model	1.62	10.70

Note: BS: (baseline); MA: (morphological analysis); TG: (tagging); UW: (unknown word model).

TABLE 9
The progressive results in our approaches on unknown word recognition.

To examine our approaches in a more general task, we also test a corpus of newspaper (Free Times), which consists of 400 training sentences and 100 testing sentences; the results are shown in Table 10. From this table, the error reduction rate of 40.15% in word and 34.78% in sentence can be observed.

		Error rate in the testing set	
	Computational Model	word (%)	sentence (%)
BS	$P(w_k \mid l_{k-1})$	19.00	69.0
BS+MA	$P(w_k \mid l_{k-1})$	13.06	50.0
BS+MA+TG	$P(w_k \mid t_k) \times P(t_k \mid t_{k-1})$	12.21	52.0
BS+MA+TG+UW	unknown word model	11.37	45.0

Note: BS: (baseline); MA: (morphological analysis); TG: (tagging); UW: (unknown word model).

TABLE 10
The results on newspaper task.

Looking into the errors in more detail, 3 unknown words of class 3 and 22 ones of class 4 are identified correctly in the training set, where originally, there are 6 and 39 unknown words of class 3 and class 4 respectively. It means that 55.56% unknown words in these two classes are recovered. Actually, the 17 mis-recognized class 4 unknown words are all caused by the missing of the first stage. Hence, how to select more discriminative features in the first stage is a key issue to improve the model in our next work. On the other hand, 1 of 2 unknown words for class 3, and 5 of 6 unknown words for class 4 are recognized correctly in the testing set; it corresponds to 75% recognition rate for these two classes. Both these two errors are tri-character words that are not considered in the current models. Although the promising results have shown the superiority of the resolution procedure, the model proposed in this paper, however, only tackles a very restrictive form of unknown words. We will extend and modify the model to more general cases in the future.

7. Summary

Since we have shown in our previous work that the existence of unknown words is the main factor that causes the performance of word segmentation task to be unsatisfied, we, therefore, shift the focus to this issue in the paper. Unknown words are generally formed in terms of regular or irregular ways. First, in this paper, a set of 17 morphological rules are applied to recognize those regular unknown words. In addition, an unknown word model is further proposed to deal with the unknown words of irregular forms. With the unknown word resolution procedures, the error reduction rate of 78.34% in word and 81.87% in sentence are obtained in a task of technical manuals. To examine the procedures in more general task, a corpus of newspaper is also tested and the error reduction rate of 40.15% in word and 34.78% in sentence are observed.

Acknowledgement

The authors would like to express our gratitude to Miss Shu-Jun Ke (Behavior Design Corporation) for her efforts in preparing the morphological rules and corpus. Her works

provide useful training and testing materials for verifying the proposed models. Special thanks are given to the Behavior Design Corporation for providing the Chinese-English Electronic Dictionary to this research project.

References

[BDC 92] Behavior Design Corporation, "The BDC Chinese-English Electronic Dictionary: Version 2," 1992.

[Chia 92a] Chiang Tung-Hui, Ming-Yu Lin and Keh-Yih Su, "Statistical Models for Word Segmentation and Unkonwn Word Resolution," *Proceedings of 1992 R.O.C. Computational Linguistics Conference (ROCLING V)*, pp. 121-146, Taipei, Taiwan, 1992.

[Chia 92b] Chiang Tung-Hui, Yi-Chung Lin and Keh-Yih Su, "Syntactic Ambiguity Resolution Using A Discrimination and Robustness Oriented Adaptive Learning Algorithm," *Proceedings of International Conference on Computational Linguistics (COLING '92)*, pp. 352-358, Nates, France, Jul. 23-28, 1992.

[Devi 82] Devijver, P. A. and J. Kittler, "Pattern Recognition: A Statistical Approach," *Prentice-Hall International Inc.*, 1982.

[Lin 93] Lin Ming-Yu, *A Study in Chinese Word Segmentation, master thesis*, National Tsing Hua University, Taiwan, R.O.C., 1993.

[Liu 93] Liu Yuan-Ling, Shih-ping Wang and Keh-Yih Su, "Corpus-based Automatic Rule Selection in Designing a Grammar Checker," to appear in *1993 R.O.C. Computational Linguistics Conference (ROCLING VI).*

Appendix A Morphological Rules

(1) 附著在後一詞項的詞綴：共計3個。

1 總：「總 工程師」

2 主：「主 電腦板」

3 副：「副 處長」

(2) 與前一詞項組合的詞綴：共計13個。

1 氏：「劉 氏」

2 師：「工程 師」

3 們：「同學 們」

4 族：「安公子 族」

5 器：「轉換 器」

6 員：「操作 員」

7 碼：「字元 碼」

8 鍵：「說明 鍵」

9 式：「階層 式」

10 極了：「美 極了」

11 多了：「好 多了」

12 的很：「好 的很」

13 得很：「快樂 得很」

(3) 限定詞與數量詞的形成

例：一個個，一條條

規則表示法：n → q+cl+cl

a → q+cl+cl

(n：名詞；a：形容詞；q：數詞；cl：量詞)

規則說明：數詞+量詞+量詞可形成一個名詞或形容詞。

(4) 日期與時間

例；八十二年，五月，十八日，六點，廿七分，卅九秒

規則表示法：n → q+ 年

n → q+ 月

n → q+ 日

n → q+ 點

n → q+ 分

n → q+ 秒

(5) 名詞的前綴（簡稱 npfx）

例：「非　產品」，「全　比例尺」

規則表示法：n → npfx+n

這一類的詞有：全、初、反、非、老，共計5個。

(6) 名詞的後綴（簡稱 nsfx）

例：「教育　性」，「叉　狀」

規則表示法：n → n+nsfx

這一類的詞有：性、家、狀，共計3個。

(7) 動詞的後綴（簡稱 vsfx）

例：「情緒　化」，「綠　化」

規則表示法：v → n+vsfx（v：動詞）

這一類的詞有：化，共計1個。

(8) 結果詞綴（簡稱 rsfx）

例：「跑　得」，「哭　得」

規則表示法：v → v+rsfx

這一類的詞有：得，共計1個。

（9）趨向標誌（簡稱vdir）

例：「動　起來」，「哭　出來」，「吞　下去」，「拿　出」，「丟　下」

規則表示法：v→v＋vdir

這一類的詞有：上、下、回、出、進、過、起、來、去、上去、下去、下來、上來、回去、回來、過去、過來、進去、進來、出去、出來、起來，共計22個。

（10）動相標誌（簡稱phase）

例：「喝　光」，「用　完」，「吃　掉」，「寫　好」，「叫　住」，「用　作」，「飛　向」，「走　向」

規則表示法：v→v＋phase

這一類的詞有：住、入、到、完、掉、好、過、光、開、做、作、成、為、向，共計14個。

（11）～（13）為動詞的特殊句型

（11）特殊動詞（一）

例：「丟丟　看」，「吃吃　看」，「寫寫　看」

規則表示法：v→v＋v＋看

說明：出現在「看」前的兩個動詞必須完全一樣。

（12）特殊動詞（二）

例：「高高興興」，「歡歡喜喜」，「漂漂亮亮」，「迷迷糊糊」

規則表示法：v→va＋va＋vb＋vb

說明：va表動詞的前一成份，而vb表動詞的後一成分。

(13) 特殊動詞（三）

例：「打打　球」，「跑跑　步」，「寫寫　字」，「高高」，「瘦瘦」，「慢慢」，「黑黑」

規則表示法：v →v+v+n

v →v+v

說明：兩個動詞必須完全一樣。

(14) 動詞前綴（簡稱vpfx）

例：「已　定義」，「反　革命」，「未　成年」

規則表示法：v →vpfx+v

這一類的詞有：反、未、已，共計3個。

(15) 形容詞前綴（簡稱apfx）

例：「非　人」，「非　人性」

規則表示法：a →apfx+n

這一類的詞有：非，共計1個。

(16) 副詞後綴（簡稱advsfx）

例：「快樂　地」，「勉強　地」

規則表示法：adv →adv+advsfx

這一類的詞有：地，共計1個。

(17) 動詞中綴（簡稱vifx）不予合併

例：「快樂　不　快樂」，「同　不　同意」

規則表示法：v+vifx+v →v+vifx+v

這一類的詞有：不，共計1個。