A Practical Tagger for Chinese Corpora

Keh-jiann Chen, Shing-Huan Liu, Li-ping Chang
CKIP, Institute of Information Science, Academia Sinica, Taipei, Taiwan
({kchen, huan, lpchang}@iis.sinica.edu.tw)
Yeh-Hao Chin
Institute of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
(yhchin@cs.nthu.edu.tw)

Abstract

In this paper, we propose an automatic Chinese tagger having the accuracy ranged from 96% to 98% depending on the types of texts. Since large fully tagged Chinese corpus is not available, the relaxation labeling method is first adopted to select the statistically most plausible parts-of-speech for words which are categorically ambiguous. The performance of the relaxation labeling method is not satisfactory, hence we propose a hybrid approach which combines the relaxation labeling method with a rule-based method. The two methods complement each other. The accuracy of the relaxation labeling method is increased by 7%, because the statistically problematic ambiguities were resolved by the rules.

1. Introduction

In the past many automatic text taggers have been developed to assign the correct syntactic categories (i.e. parts-of-speech) to words in a large corpus. Probabilistic methods are most successful in practice [8]. In a probabilistic tagger, a probabilistic model is used to capture lexical and contextual information according to a probability distribution. In general, the parameters of the probabilistic model can be trained in two different ways:

T1. The parameters are trained by some tagged texts [7][10][12][16][18].

T2. The parameters are trained based on the hidden Markov model, which does not require any tagged corpus as the training data [2][15][17]. This type of training uses Baum-Welch algorithm [1] to compute a sequence of better and better probability estimations of parameters iteratively.

Since there is no large tagged Chinese corpus, we can not follow the type T1 training algorithm to implement a Chinese tagger. Furthermore, Chinese language has many characteristics substantially different from other languages which pose difficulties for purely probabilistic methods. Here we state some of the characteristics of Chinese language texts [4]:

C1. Chinese exhibits more free word order, which allows many possible parts-of-speech in a fixed context. In other words, the local context of a Chinese word may be relatively changeable.

C2. Chinese is a weakly marked language with little inflection. A Chinese word may play many different grammatical functions in different contexts without morphological changes. For example, a verb without inflection could be a modifier of a noun or nominalized as an argument.

The above characteristics make the probabilistic model of using co-occurrences of syntactic categories to select the most probable part-of-speech sequence very tricky. On the other hand, if we use the rule-based approach to tag a Chinese corpus, it requires a lot of context frame rules, like the TAGGIT system [13]. Therefore, we propose a hybrid method which combines a probabilistic approach with a rule-based approach. We found, as Leech [12] indicated, the rule-based approach and the probabilistic approach are complementary: "they are digging the same tunnel from opposite ends." To avoid the requirement of pretagged training corpus, we use the relaxation labeling method [20] to iteratively reestimate the probabilities of the categories for each word in the target corpus according to the contextual tags. We believe that the relaxation labeling method is simpler than the hidden Markov model with Baum-Welch algorithm. The detailed algorithm will be discussed in the next section. The result of the tagged corpus was examined by linguists to produce the most effective rules to make up the weakness of the relaxation labeling method. In actual tagging, these heuristic/linguistic rules are applied first to disambiguate the words with multiple categories or to remove impossible tags. The relaxation labeling method is applied next to make the final selection. It turns out that the accuracy of the relaxation method can be increased from 71.27% (without applying rules) to 78.44% (after applying rules) in disambiguating words with multiple categories.

2. Relaxation Labeling Method for Part-of-Speech Tagging

Relaxation labeling method has been widely used in the areas of image processing [19] and scene analysis [11][20]. The formal analysis of relaxation labeling method has been done by Rosenfeld, Hummel, and Zucker [20]. In general, relaxation labeling method consists a set of algorithms which assign labels to objects. During each assignment, relaxation technique can reduce the degree of ambiguity and select the best label among several possible choices, based on the local constraints between labels [14]. The task of assigning the right category (i.e. label) to a word (i.e. object) has similar characteristics. Therefore, the idea and principles of relaxation labeling method are suitable for solving the tagging problem.

The basic idea behind the relaxation labeling method is that the category assigned to a particular word should be agreed with the categories assigned to its neighbors. Consequently, the category assigned to the word is updated iteratively on the basis of the categories assigned to its adjacent words [11]. For example, consider the problem of part-of-speech tagging. Suppose there are two consecutive words W_1 and W_2 in a text. Word W_1 has two possible categories C_1 and C_2, and word W_2 has only one category C_3. We want to determine which category, C_1 or C_2, is more suitable for the word W_1. Based on the basic idea of relaxation labeling method, the assignment of category to word W_1 is affected by the assignment of categories to its neighboring words. Let W_2 be the only neighbor of W_1. After some statistical estimation, the results show that the category pair (C_1,C_3) has higher degree of compatibility than the pair (C_2,C_3). Through relaxation labeling process, the possibility of assigning C_1 to W_1 will be increased after each iteration, since such an assignment is agreed with the unique and fixed assignment of C_3 to W_2.

In our actual implementation, the Markov model is used to measure the degree of compatibility between categories. For each word in the corpus, its possible categories are initialized as an equal probability. The parameters of category bigram (or trigram) are estimated directly from the target corpus. After each relaxation labeling process, the probability of each category will be reestimated according to the compatibility with respect to its context. The bigram (or trigram) parameters have to be recalculated after each iteration. Following is an example of bigram relaxation model to tag a sequence of words in a large corpus. The general relaxation tagging algorithm can be obviously extended.

$$\begin{array}{lll} W_{i-1} & W_i & W_{i+1} \\ C_1 & C_m & C_3 \\ C_2 & C_n & C_4 \end{array}$$

Suppose we want to estimate the probabilities $P(C_m|W_{i-1}\ W_i\ W_{i+1})$ and $P(C_n|W_{i-1}\ W_i\ W_{i+1})$ by the relaxation labeling method (For the notational simplicity, here after, $P(C_i|W_{i-1}\ W_i\ W_{i+1})$ is abbreviated as $P(C_i|W_i)$)$.$ $P(C_m|W_i)$ and $P(C_n|W_i)$ are initialized as 0.5 respectively, i.e. each of the category will be assigned to an equal probability 1/x if this word has x categories. There are 4 possible category sequences (CS) formed by the immediate neighbors of the word W_i.

$CS_1 : C_1$ -- C_3
$CS_2 : C_1$ -- C_4
$CS_3 : C_2$ -- C_3
$CS_4 : C_2$ -- C_4

The question is how these 4 CS's are compatible with the categories C_m and C_n of the word W_i. The compatibility value contributed to the category C_m of word W_i from CS_1 is defined as $q_1(C_m)$:

$$q_1(C_m) = P(C_1|W_{i-1}) \times P(C_m|C_1) \times P(C_3|C_m) \times P(C_3|W_{i+1}) \qquad \text{(F1)}$$

where

$$P(C_m|C_1) =_{\text{def}} \frac{\text{Freq}(C_1\ C_m)}{\text{Freq}(C_1)} \qquad \text{(F2)}$$

$$\text{Freq}(C_1) =_{\text{def}} \sum_{\text{for all } W_j \text{ in the corpus}} P(C_1|W_j) \qquad \text{(F3)}$$

$$\text{Freq}(C_1\ C_m) =_{\text{def}} \sum_{\text{for all } W_j \text{ in the corpus}} P(C_1|W_j) \times P(C_m|W_{j+1}) \qquad \text{(F4)}$$

$P(C_3|C_m)$ has the similar definition.

The compatibilities of the category sequences CS_2, CS_3, and CS_4 for the category C_m of word W_i can also be computed similarly. The total contribution made to the category C_m of word W_i from all these 4 CS's is

$$Q(C_m) = q_1(C_m) + q_2(C_m) + q_3(C_m) + q_4(C_m) \qquad \text{(F5)}$$

By the same way, we can compute $Q(C_n)$. Then the new estimate of $P(C_m|W_i)$ is as follows

$$P(C_m|W_i) = \frac{Q(C_m)}{Q(C_m)+Q(C_n)} \tag{F6}$$

The new estimate of $P(C_n|W_i)$ can be computed similarly.

The whole corpus shall be processed from the first word to the last word in each iteration. The estimation iteratively computes a number of times and results in each iteration are accumulated until a stable value is reached. In practice, the number of iteration is not very important. As Rosenfeld and Kak [21] indicated: "we are not normally interested in reaching a limit point, but only in carrying the process through a few iterations (typically, less than ten) so as to correct initial errors and reduce initial ambiguities." In our experiment, a small portion (about 1.5%) of the input corpus is tagged manually in order to serve as the basis of comparison. During each iteration, the result of this portion, which is performed by automatic tagging, is compared with the manually tagged one, and the iterative process terminates when no significant improvement is observed. When the iterative process terminated, the category having the highest probability value for a word is selected as the right category for the word. If more than one categories have the same highest probability value, then the selection is a random one.

The above is a simple example of part-of-speech assignment by the relaxation labeling method using category bigram model. In general, the model for computing compatibility values could be varied. For instance, in the trigram model, the compatibility is computed within a window of two neighboring words at each side. Let's follow the above example. Suppose the word W_i has two categories C_m and C_n, and there are CS_1, CS_2, ..., CS_k different category sequences at its two words neighboring window. Let CS_1 be C_1 C_2 -- C_3 C_4. Then the compatibility value between CS_1 and the category C_m for the trigram model is defined as

$$q_1(C_m) = P(C_1 \,|\, W_{i-2}) \times P(C_2 \,|\, W_{i-1}) \times P(C_m \,|\, C_1\ C_2) \times P(C_3|C_2\ C_m) \times$$
$$P(C_4|C_m\ C_3) \times P(C_3|W_{i+1}) \times P(C_4|W_{i+2}) \tag{F7}$$

where

$$P(C_m|C_1\ C_2) =_{\text{def}} \frac{Freq(C_1\ C_2\ C_m)}{Freq(C_1\ C_2)} \tag{F8}$$

$$Freq(C_1\ C_2) =_{\text{def}} \sum_{\text{for all } W_j \text{ in the corpus}} P(C_1|W_j) \times P(C_2|W_{j+1}) \tag{F9}$$

$$Freq(C_1\ C_2\ C_m) =_{\text{def}} \sum_{\text{for all } W_j \text{ in the corpus}} P(C_1|W_j) \times P(C_2|W_{j+1}) \times P(C_m|W_{j+2}) \tag{F10}$$

$P(C_3|C_2\ C_m)$ and $P(C_4|C_m\ C_3)$ have the similar definition.

Then the total contribution made to the category C_m of the word W_i is

$$Q(C_m) = \sum_{j=1}^{k} q_j(C_m) \qquad \text{(F11)}$$

The formula for estimating $P(C_m|W_i)$ is the same as (F6), and the new estimate of $P(C_n|W_i)$ can be computed similarly. The general formulas for the relaxation labeling method can be derived by following the above example. Some experiment results of the relaxation labeling method is given in section 4.

3. Tagging Procedure for Chinese

To tag a Chinese corpus, there is an additional problem to be solved. The input corpus is a string of Chinese characters without blanks to mark words, so the first step is to identify words. The CKIP dictionary [3] is prepared to provide the words and their syntactic categories. However, no matter how large a lexicon is, many compounds and proper names may still not be included in the dictionary. Lacking word breaks and inflectional markers, it is difficult to identify those unknown words and provide their parts-of-speech. Word identification algorithms for Chinese can be found in [5][6][18], but none of them provides a satisfactory solution to the problem of unknown word identification. To solve this problem, the high frequency unknown words can be found by examining collocations in the corpus before tagging, and the rarely occurring unknown words will be fixed afterwards by human post editing.

3.1. Search for Unknown Words by Collocations

An unknown word is a word which is not included in the dictionary, and it could be segmented into two or more words after the word segmentation process by looking up the dictionary. For those high frequency compounds or proper names, the technique such as Xtract [22] can be used to find unknown words, since it satisfies the properties of collocation [22][23]. The discovered collocates will be re-examined by a linguist to identify words and to determine the syntactic categories. Those newly discovered words will serve as supplements of the lexicon.

3.2. A Reduced Part-of-Speech Tagset

Since relaxation is a probabilistic method, only categories with different contextual patterns can be discriminated. Therefore, a reduced tagset of 57 different parts-of-speech (including 9 punctuation marks) was derived from the original 178 syntactic categories from the CKIP tagset [9] as shown in Appendix 1. In the reduced tagset, the semantic criteria are not considered. In addition, some prepositions in Chinese may function as verbs if no other matrix verb exists. Such prepositions could be assigned to a special category "Pv" and leave the discrimination of preposition or verb function to parser.

3.3. The Preliminary Experiment and Derivation of Contextual Rules

As we mentioned in section 1, Chinese has many characteristics which cause difficulties for a purely probabilistic tagger. Therefore, we shall not expect that the relaxation labeling method alone can do the part-of-speech tagging job well. For example, either a verb, an adjective, or a noun could precede a head noun as a modifier of the head noun. If the modifier is a word of two categories, noun and verb, then the noun category will be always selected since it has statistical advantage. But there are many instances where verb is the correct selection. Therefore, the tagging result performed by relaxation labeling method shall be re-examined by linguists to identify contextual rules to make up the weakness of the relaxation labeling method. According to our preliminary experiment on the relaxation labeling method without applying contextual rules, poor results were produced in the following cases:

E1. Some high frequency words are easily assigned with wrong categories.

E2. For some multiple-category words, statistical computation result is usually preferred a particular category which is incorrect. Examples are those words having common noun and verb together.

For the error type E1, 21 rules are collected to take care of 21 most error-prone words. For error type E2, 6 rules are designed to eliminate the categories which are statistically preferred but incorrect. These context-dependent rules are listed in Appendix 2.

The reasons that the rules and the relaxation labeling method complement each other are

R1. The contextual rules can take care of the non-neighboring dependencies and the relaxation labeling method handles various local dependencies.

R2. The relaxation labeling method does not consider the idiosyncrasy of the individual word, but rules do. For instance, there is no pretagged corpus to know the probability distribution of categories for each individual word but it is a very useful information in the probabilistic tagging models [7][8].

3.4. Tagging Steps for Chinese Corpora

The final tagging steps we propose for Chinese corpora are the relaxation labeling method with a rule-based filtering algorithm as follows:

S1. Unknown word identification. Find the new words in the corpus and assign their syntactic categories. The new words are found by examining the collocations in the corpus. The new words will serve as supplements of the lexicon in order to improve the accuracy of word segmentation.

S2. Word segmentation and initial category assignment. The target corpus is segmented into word sequences and the syntactic categories of each word are mapped into the reduced forms.

S3. Applying disambiguation rules. The context dependent rules are successively applied to determine the correct part-of-speech or to eliminate the contextually impossible categories.

S4. Applying relaxation labeling algorithm. The relaxation labeling algorithm is applied to resolve the remaining ambiguities. The relaxation process will be iterated a fixed number of times or terminated under satisfaction of certain convergent criterion.

4. Experiment Results

A large Chinese corpus from the CKIP group of Academia Sinica containing about 2 million words which includes a test data of 36436 words is segmented [5] first. The test data are tagged with their syntactic categories manually. They are only used as the reference base to check the correctness of the automatically tagged results, not for the purpose of bootstrapping. The test data contains 4 pieces of texts and are shown in Table 1. Data 1, Data 2, Data 3, and Data 4 are articles and news from Common Wealth, Liberty Time, China Time ,and ErhTung Daily, respectively.

Table 1. The Test Data

	No. sentences	No. words	No. multiple-category words	% multiple-category words
Data 1	360	2540	484	19.06%
Data 2	3272	21184	3993	18.85%
Data 3	1504	12046	1853	15.38%
Data 4	102	666	136	20.42%
total	5238	36436	6466	17.75%

As we mentioned in the previous section, the task of word segmentation is performed before the task of selecting the correct category can be done. Therefore, the correctness of word segmentation influences the accuracy of category selection. Most of the errors in word segmentation are due to unknown words. An occurrence of an unknown word implies that its segmentation and the initial categories are subject to some errors. The errors caused by word segmentation will not take into account while measuring the accuracy of tagging algorithm.

Both bigram relaxation model and trigram relaxation model are tested. The bigram model performs better than the trigram model. The result shown in the left part of Table 2 is obtained by applying bigram relaxation labeling method only. The accuracy of the relaxation algorithm is 67.84-77.07% for the multiple-category words. The total ratio of correctness is 94.73-95.63%. After examining part of the result other than the test data carefully, 21 specific rules and 6 general rules are derived as remedy for the error types E1 and E2, respectively. Then the corpus is retagged by the hybrid method. The correctness ratio of the hybrid method, compared with the relaxation labeling method, is improved from 71.27% to 80.87% for the multiple-category words. The total accuracy is 96.34-97.90%, depending on the types of texts.

Table 2. Comparison between the Relaxation and the Hybrid Method

	relaxation alone			hybrid		
	No.tagging errors	% correct tag		No.tagging errors	% correct tag	
		for all words	for multiple-category words		for all words	for multiple-category words
Data 1	111	95.63%	77.07%	58	97.72%	88.02%
Data 2	1117	94.73%	72.03%	776	96.34%	80.57%
Data 3	596	95.05%	67.84%	389	96.77%	79.01%
Data 4	34	94.89%	75.00%	14	97.90%	89.71%
total	1858	94.90%	71.27%	1237	96.61%	80.87%

Among 6466 multiple-category words in the test data, there are 3628 words whose categories are affected by the rules, but some of them are not resolved to a unique category. After applying the rules, 4897 words still remain with multiple categories which are further resolved by the relaxation algorithm. The final results show that there are 1237 errors caused by the hybrid method. Among them, 181 errors are due to the rules and 1056 errors are due to the relaxation labeling method. Table 3 shows the accurate ratio of the application of rules, and the relaxation labeling method, respectively.

Table 3. The Accurate Ratio of the Method Influenced by Rules and Relaxation

Hybrid Method	influenced by rules			influenced by relaxation		
	No. of words affected by context rules (1)	rule errors (2)	accuracy $\frac{(1)-(2)}{(1)}$	No. of multiple-category words after applying rules (3)	relaxation errors (4)	accuracy $\frac{(3)-(4)}{(3)}$
total	3628	181	95.01%	4897	1056	78.44%

The following observations are derived from this experiment:

O1. The rules not only perform well with the accuracy of 95.01% but also improve the accuracy of the relaxation labeling method from 71.27% to 78.44%. Since the rules are human adjustable and expandable, the accuracy of the rules can be easily improved by refining the underlying rules. Moreover, high accuracy of the rules indicates that more reliable contextual patterns can be generated in the input corpus. Hence the performance of the relaxation labeling method is also improved.

O2. The number of iterations needed for relaxation labeling method is reduced after applying the rules. In our experiment, it takes 3 to 5 iterations to reach a stable assignment for pure relaxation labeling method; while only 1 iteration is stable enough for the hybrid method. The rules provide more reliable contextual information, and results in the improvement on the statistically preferred categories.

O3. The total execution time of the whole input corpus required to update the probability values during each iteration for relaxation labeling method is decreased by 19.23% after applying rules under the testing environment HP 9000/835 UNIX system. The computation time is decreased since the degree of ambiguities in the input corpus is reduced by the rules.

5. Conclusion

The relaxation labeling method looks at the context of two (or a few) concatenating words. Therefore, it fails to account for (1) relationship beyond that context and (2) lexical preference for each individual word.

The first disadvantage is due to the feature of probabilistic method. The second disadvantage is due to the lack of pretagged corpus to provide the probability distribution of categories of each individual word. The rule-based approaches can overcome these disadvantages to some extent by considering long distance dependencies and the idiosyncrasies of each individual word. However it is very tedious to consider every fine-grained difference by rules, so the relaxation labeling method takes care most of the local relations, and the rules patch the rest, including non-local constraints.

If there is no pretagged training corpus, the relaxation labeling method is a good tagging method for the following advantages:

A1. It does not need pretagged training data.

A2. Bootstrapping and rule-based methods can be incorporated easily.

A3. It is simpler and more flexible than the hidden Markov model with Baum -Welch algorithm.

The hybrid method is considered as a one-shot method. Once we have a tagged corpus by this method, the well-known probabilistic tagging algorithms without iterations, such as the trigram model in Church [7], can replace the relaxation labeling method to speed up the tagging process. However the rules are retained to remedy the deficiency of the probabilistic method as mentioned above.

Acknowledgments

The research of this paper was partially supported by NSC grant NSC 84-0301-H-001-008.

References

[1] Baum, L. E., "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process," *Inequalities*, 3, 1972.

[2] Chang. C. H., and Chen. C. D., "HMM-based Part-of-Speech Tagging for Chinese Corpora," in *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Ohio State University, 1993.

[3] Chen, K. J., and CKIP, "The Chinese Knowledge and Information Project and Chinese Electronic Dictionary," in *Proceedings of the Joint Chinese-Japan Symposium on Information Processing*, 1991 (in Chinese).

[4] Chen, K. J., "Design Concepts for Chinese Parsers," in *Proceedings of the 3rd International Conference on Chinese Information Processing*, 1992.

[5] Chen, K. J., and Liu, S. H., "Word Identification for Mandarin Chinese Sentences, " in *Proceedings of COLING-92'*, 1992.

[6] Chiang, T. H., Chang, J. S., Lin, M. Y. and Su, K. Y., "Statistical Models for Word Segmentation and Unknown Word Resolution," in *Proceedings of ROCLING V*, 1992.

[7] Church, K. W., "A Stochastic Parts Program and Noun Phrase for Unrestricted Text," in *Proceedings of the 2nd Conference on Applied Natural Language Processing (ACL)*, 1988.

[8] Church, K. W., and Mercer, R. L. "Introduction to the Special Issue on Computational Linguistics Using Large Corpora," *Computational Linguistics*, vol. 19, no. 1, 1993.

[9] CKIP, "Analysis of Syntactic Categories for Chinese," Technical Report #93-05, Institute of Information Science, Taipei, 1993 (in Chinese).

[10] DeRose, S. J., "Grammatical Category Disambiguation by Statistical Optimization," *Computational Linguistics*, vol. 14, no. 1, 1988.

[11] Eklundh, J. O., and Rosenfeld, A., "Some Relaxation Experiments Using Triples of Pixels," *IEEE Transactions on System, Man, and Cybernetics*, vol. SMC-10, no. 3, March 1980.

[12] Garside, R., Leech, G., and Sampson G., *The Computational Analysis of English: a Corpus-Based Approach*, Longman, 1987.

[13] Greene, B. B., and Rubin, G. M., *Automated Grammatical Tagging of English*, Department of Linguistics, Brown University, Providence, Rhode Island, 1971.

[14] Hummel, R. A., and Zucker S. W., "On the Foundation of Relaxation Labeling Process," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 3, May 1983.

[15] Kupiec, J., "Robust Part-of-Speech Tagging Using a Hidden Markov Model," *Computer Speech and Language*, vol. 6, 1992.

[16] Lin, Y. C., Chiang, T. H., and Su, K. Y., "Discrimination Oriented Probabilistic Tagging," in *Proceedings of ROCLING V*, 1992.

[17] Merialdo, B., "Tagging Text with a Probabilistic Model," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1991.

[18] Pong, T. Y., and Chang, J. S., "A Study of Word Segmentation and Tagging for Chinese," in *Proceedings of ROCLING VI*, 1993 (in Chinese).

[19] Rosenfeld, A., "Iterative Methods in Image Analysis," *Pattern Recognition*, vol. 10, 1978.

[20] Rosenfeld, A., Hummel, R. A., and Zucker, S. W., "Scene Labeling by Relaxation Operations," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 6, June 1976.

[21] Rosenfeld, A., and Kak, A. C., *Digital Picture Processing*, 2nd Edition, vol. 2, Academic Presses, 1982.

[22] Smadja, F., "Retrieving Collocations from Text: Xtract," *Computational Linguistics*, vol. 19, no. 1, 1993.

[23] Wang, M. C., Chen, K. J., and Huang, C. R., "The Identification and Classification of Unknown Words in Chinese - A N-grams-based Approach," in *Proceedings of the Joint Conference of the 8th ACLIC/the 2nd PacFoCoL*, 1994.

Appendix 1. The Reduced Tagset

A	non-predicate adjective
b	bound morpheme
Caa	coordinate conjunction
Cab	listing conjunction
Cba	conjunction occurring at the end of a sentence
Cbb	subordinate conjunction following the subject of a sentence
Cbc	subordinate conjunction occurring at the initial position of a sentence
D	general adverb
Da	quantity adverb
De	relative marker
Dfa	degree adverb preceding the stative verb
Dfb	degree adverb following the stative verb
Di	aspect
Dk	adverb always occurring at the initial position of a sentence
DM	determiner-measure word
I	interjection
Na	common noun
Nb	proper noun

Nc	place noun
Nd	time noun
Ne	determiner
Nf	classifier
Ng	localizer
Nh	pronoun
P	preposition
Pa	preposition always taking a temporal or locative argument
Pb	preposition that also has a verb tag and the verb tag is a high frequency one
Pc	preposition often occurring at the initial position of a sentence
Pv	coverb
S	archaic saying
SHI	special tag for the word "是"
Str	bound character string
T	particle
TZAI	special tag for the word "在"
VA	intransitive active verb
VB	pseudo-transitive active verb
VC	transitive active verb
VC1	transitive verb taking a locative argument
VD	ditransitive verb
VE	active verb with sentential object
VF	active verb with verbal object
VG	classificatory verb
VH	intransitive stative verb
VI	pseudo-transitive stative verb
VJ	transitive stative verb
VK	stative verb with sentential object
VL	stative verb with verbal object
VR	verb-resultative verb

Appendix 2. Contextual Rules

Some specific notations in the following rules should be explained:

N1. "A << B" indicates that A immediately precedes B.

N2. "A < B" indicates that A locates in front of B.

N3. "#" is a beginning or ending marker.

N4. "{A,B}" indicates "either A or B".

N5. "(A)" indicates that A is optional.

N6. "A-B" indicates A but not including B.

N7. Each condition of the rule is associated with a marker like "→ A" indicating that the word should be tagged as category "A".

The context rules are listed as below.

SR1. specific rule for the word "下去"
C1. {V*,PV} (<< {了,不}) << "下去" → Di
C2. otherwise → VA

```
SR2.  specific rule for the word "中"
      C1. "中" (<< PAUSECATEGORY) << {b,Nc,Nb} → Nc
          (NOTE : PAUSECATEGORY indicates the punctuation mark "、")
      C2. otherwise → Ng
SR3.  specific rule for the word "對"
      C1. "對" << (<< {De,T}) << # → VH
      C2. otherwise → P
SR4.  specific rule for the word "當中"
      C1. -的 << V* << "當中" << # → Di
          (NOTE : "-的" indicates any word other than "的")
      C2. # << "當中" → Nc
      C3. otherwise → Ng
SR5.  specific rule for the word "為"
      C1. "為" << 了          → PV
          "為" << TZAI
      C2. V* << "為"                  → VG
          "為" (<< {DM,Ne}) << #
          {化,以} < "為"
          的 << "為"
      C3. "為" < 所 << V* << #        → P
          (NOTE : the word "所" can not immediately follow "為")
          "為" < W(所*) << #
          (NOTE : "W(所*)" indicates the complex words prefixed the word "所")
      C4. otherwise → discard the category "P" (i.e. retain categories "PV" and
          "VG") for future processing
SR6.  specific rule for the word "開始"
      C1. {從,自從,自,於} < -D << "開始" << {#,T,P*}                → Ng
          (NOTE: "-D" indicates any category other than "D")
          # << {DM[Nfg],N*[+temporal_relation],Ne}+ << "開始"
          (NOTE 1: "DM[Nfg]" indicates the head of the DM word is the category
                   "Nfg".
           NOTE 2: "N*[+temporal_relation]" indicates the nouns which have the
                   feature "+temporal_relation".
           NOTE 3: "+" indicates that the categories "DM" "N*", or "Ne" can occur
                   not only once.)
      C2. "開始" (<< T) << {#,的}   → VH
          "的" << "開始" << #
          "一" << "開始" << D
      C3. otherwise → VL
SR7.  specific rule for the word "過"
      C1. "過" << VH (monosyllabic) (<< #) → Dfa
      C2. {V*,PV} << "過" < {-N[+temporal_relation],-DM[Nfg]} → Di
          (NOTE : The partial condition "過" < {-N[+temporal_relation],
                  -DM[Nfg]} means that it will be satisfied if and only if the
                  categories following "過" is not "N[+temporal_relation]" and
```

```
                 not "DM[Nfg]".)
       C3. "過" (<< T) << # → VH
       C4. otherwise → VJ
SR8.   specific rule for the word "有"
       C1. "有" << DM[Nfg]                    → Pb
           "有" << DM << [+temporal_relation]
       C2. otherwise → VJ
SR9.   specific rule for the word "沒有" and "有沒有"
       C1. {沒有,有沒有} << {N*-Ng-Ne,DM,Dfa}       → VJ
           {沒有,有沒有} << -VE < {之,的}
           {沒有,有沒有} (<< T) << #
           {沒有,有沒有} (<< D) << VH << Na
           Dfa << {沒有,有沒有}
           {沒有,有沒有} << Da << 的
           {沒有,有沒有} << -VE < {A,VA,VH,VF} << 的
       C2. otherwise → D
SR10.  specific rule for the word "上"
       C1. monosyllabic V << "上"              → the category of "V"
           monosyllabic V << 不 << "上"
           monosyllabic V << 得 << "上"
           (NOTE : These are morphological rules. We merge "上", "不上", "得上
                   " with the preceding monosyllabic verb to form a compound,
                   and then assign the category of the preceding verb to the
                   compound.)
       C2. "上" << Nd → Ne
       C3. "上" << {了,過}          → VC1
           {D*-Di,C*} << "上"
       C4. otherwise → Ng
SR11.  specific rule for the word "到"
       C1. {DM,Ne} << "到" << {DM,Ne}         → Caa
       C2. "到" < {為止,底,止,時,末}           → Pa
           {由,從} < "到"
           "到" << DM << [+temporal_relation]
           "到" << {DM[Nfg],[+temporal_relation]}
       C3. {VA,VB,VC,VD,VE,VH,VI,VJ} << "到" → the category of "V"
           (NOTE : This is a morphological rule. We merge "到" with the
                   preceding verb to form a compound, and then assign the
                   category of the preceding verb to the compound.)
       C4. otherwise → VC1
SR12.  specific rule for the word "不到"
       C1. {PV,V*} << "不到" → category of "V"
           (NOTE : This is a morphological rule. We merge "不到" with the
                   preceding word to form a compound, and then assign the
                   category of the preceding word to the compound.)
       C2. "不到" << Nc → VC1
```

C4. otherwise → VJ

SR13. specific rule for the word "起來"
C1. {看,乍看,吃,喝,做,聽,嚐,閱讀,聞,說} << "起來" → D
(NOTE : This is a morphological rule. We merge "起來" with the preceding word. The newly merged word shall be assigned to the category "D".)
C2. V* (<<{了,不}) << "起來" → Di
C3. otherwise → VA

SR14. specific rule for the word "用"
C1. {VJ,Ne,之} << "用" (<< {De,T}) << #→ Na
一點 << "用" << {也,都}
C2. otherwise → PV

SR15. specific rule for the word "要"
C1. {就,將} << "要" → D
{比,比起,較} < "要"
C2. "要" << De → VC
"要" << {DM,N*-Ne-Ng} << #
C3. "要" << {DM,N*-Ng} → VE
C4. otherwise → D

SR16. specific rule for the word "連"
C1. "連" < {都,也,亦,皆} → Cab
C2. "連" << {V*,PV} → D
C3. {DM,Ne} << "連" → Na
以 << "連" << 為
C4. otherwise → Nb

SR17. specific rule for the word "一點"
C1. "一點" << {都,也} → Da
C2. VH (<< 了) << "一點" << # → Dfb
{VH,VI,VJ,VK,VL} << "一點"
C3. otherwise → Nf

SR18. specific rule for the word "起"
C1. {V*,PV} << "起" → Di
C2. {C*,D*-Di} << "起" → VC
C3. otherwise → Ng

SR19. specific rule for the word "去"
C1. "去" << {V*,PV} → D
C2. otherwise → VC1

SR20. specific rule for the word "可能"
C1. {De,DM,VJ} << "可能" → Na
C2. Dfa << "可能" → VH
"可能" << {的,#}
C3. otherwise → D

SR21. specific rule for the word "來"
C1. 起 < "來" (not including "起來") → T
C2. {DM[Nfg],N*[+temporal_relation]} << "來" → Ng

C3. "來" << {V*,PV} → D

C4. otherwise → VA

GR1. general rule #1

If a monosyllabic word has multiple categories, and one of the categories is "Nf", then the category "Nf" shall be discarded.

GR2. general rule #2

If the current word has categories "Na" and "V*" simultaneously, and the condition "current word << {P*,Di,TZAI,DM[Nfi],Dfb}" is satisfied, then the category "Na" shall be discarded.

GR3. general rule #3

If the current word has categories "Na" and "V*" simultaneously, and the condition "D*-Di (<< V*) << current word" is satisfied, then the category "Na" shall be discarded.

GR4. general rule #4

If the current word has categories "Ne" and "VH" simultaneously,

C1. current word (<< {Dfb,T}) << #→ VH

C2. otherwise

case 1: If the word has categories "Ne" and "VH" only, then we assign "Ne" to the word.

case 2: If the word has categories other than "Ne" and "VH", then all of the categories are retained for future processing.

NOTE : General rule #4 must be applied before general rule #5 because rule #4 is more specific.

GR5. general rule #5

If the current word has multiple categories, one of categories is "Ne" and the other categories do not contain "VH",

C1. current word (<< {VH,A}) << {Na,Nc,Nd} → Ne
current word (<< V*) << De << {Na,Nc,Nd}

C2. otherwise → discard the category "Ne" for future processing

GR6. general rule #6

If the current word has multiple categories, and one of the categories is "Dfa",

C1. current word (<< 不) << {VH,VI,VJ,VK,VL} → Dfa

C2. otherwise → discard the category "Dfa" for future processing

Appendix 3. Sample Results

In the following sample results, the probability value for each category is indicated inside the parenthesis, and the tagging errors are indicated as "**".

今天(Nd 1.0) 強盜(Na 1.0) 橫行(VA 1.0)，如果(Cbb 1.0) 民眾(Na 1.0) 都(Da 1.0) 能(D 1.0) 扮演(VC 1.0) 官兵(Na 1.0) 的(De 1.0) 角色(Na 1.0)，成爲(VG 1.0) 民眾(Na 1.0) 抓(VC 1.0) 強盜(Na 1.0)，那(Ne 1.0) 就(Da 0.38, D 0.57, P 0.05) 可以(D 0.79, VH 0.21) 天下太平(VH 1.0) 了(Di, 0.09, T 0.91)。過去(Nd 0.98, VC1 0.02) 一年(DM 1.0) 來(T 0.0, Ng 1.0, D 0.0, VA 0.0)，由於(Cbb 1.0) 政治(Na 1.0) 與(Caa 0.63, P 0.37) 經濟(VH 0.25, Na 0.75) 方面(Nf 0.04, Na 0.96) 的(De 1.0) 變局(Na 1.0)，社會(Na 1.0) 上(Ng 1.0, Ne 0.0, VC1 0.0) 出現(VH 1.0) 很多(Ne 1.0, VH 0.0) 失序(VH 1.0) 的(De 1.0) 現象(Na 1.0)，使(VL 1.0) 關心(VK 1.0) 國家(Na 1.0) 大事(Na 1.0) 的(De 1.0) 朋友(Na 1.0) 憂心忡忡(VH 1.0)。知識份子(Na

1.0) 論(VC 0.68, P 0.32) 天下(Nc 1.0) 大事(Na 1.0) 都(Da 1.0) 從(Pa 1.0) 理(VC 0.39, VJ 0.05, Na 0.56) 字(Na 1.0) 上(Ng 1.0, Ne 0.0, VC1 0.0) 著手(VF 1.0)，所以(Cbc 1.0) 談到(VE 1.0) 要 (D 1.0, VC 0.0, VE 0.0) 解決(VC 1.0) 問題(Na 1.0)，必然(D 1.0) 歸結到(VJ 1.0) 制定(VC 1.0) 更(Dfa 1.0) 多 (VH 0.98, Ne ,0.01, Da 0.01) 法律(Na 1.0) 上(Ng 1.0, Ne 0.0, VC1 0.0)。可是(Cbc 0.86, D 0.14) 我(Nh 1.0) 看(VC 0.66, VE 0.34)** 整個(DM 1.0) 社會(Na 1.0) 秩序(Na 1.0) 的(De 1.0) 大(VH 1.0) 問題(Na 1.0)，實際(VH 1.0) 可以(D 0.72, VH 0.28) 用(PV 1.0, Na 0.0) 簡單(VH 1.0) 的(De 1.0) 官兵(Na 1.0) 與(Caa 0.63, P 0.37) 強盜(Na 1.0) 的(De 1.0) 觀念(Na 1.0) 來(T 0.0, Ng 0.0, D 1.0, VA 0.0) 說明(VE 1.0)。我們(Nh 1.0) 都(Da 1.0) 知道(VK 1.0) 官兵(Na 1.0) 抓(VC 1.0) 強盜(Na 1.0) 的(De 1.0) 邏輯(Na 1.0)，官兵(Na 1.0) 代表(VK 0.04, Na 0.96)** 好人(Na 1.0)，強盜(Na 1.0) 代表(VK 0.04, Na 0.96)** 壞人(Na 1.0)，我們(Nh 1.0) 都(Da 1.0) 知道(VK 1.0) 壞人(Na 1.0) 怕(VK 1.0) 好人(Na 1.0)，所以(Cbc 1.0) 社會(Na 1.0) 上(Ng 1.0, Ne 0.0, VC1 0.0) 雖然(Cbb 1.0) 有(Pb 0.0, VJ 1.0) 不少(VH 1.0) 壞人(Na 1.0)，我們(Nh 1.0) 仍(D 1.0) 可以(D 0.73, VH 0.27) 安心(VH 1.0) 過日子(VA 1.0)。