HMM-based Part-of-Speech Tagging for Chinese Corpora

Chao-Huang Chang and Cheng-Der Chen
E000/CCL, Building 11, Industrial Technology Research Institute
Chutung, Hsinchu 31015, Taiwan, R.O.C.
E-mail: changch@e0sun3.ccl.itri.org.tw

Abstract

Chinese part-of-speech tagging is more difficult than its English counterpart because it needs to be solved together with the problem of word identification. In this paper, we present our work on Chinese part-of-speech tagging based on a first-order, fully-connected hidden Markov model. Part of the 1991 United Daily corpus of approximately 10 million Chinese characters is used for training and testing. A news article is first segmented into clauses, then into words by a Viterbi-based word identification system. The (untagged) segmented corpus is then used to train the HMM for tagging using the Baum-Welch reestimation procedure. We also adopt Kupiec's concept of word equivalence classes in the tagger. Modeling higher order local constraints, a pattern-driven tag corrector is designed to postprocess the tag output of the Viterbi decoder based on trained HMM parameters. Experimental results for various testing conditions are reported: The system is able to correctly tag approximately 96% of all words in the testing data.

1 Introduction

Part-of-speech tagged corpora are very useful for natural language processing (NLP) applications such as speech recognition, text-to-speech, information retrieval, and machine translation systems. Automatic part-of-speech tagging has been intensively studied and practiced for European languages [1–4,7,8,10]. However, the technology of automatic Chinese part-of-speech tagging is still in its infancy, due to the following reasons:

1. Definition of words in Chinese is not clear; there are not breaks between two adjacent words. For example, the string 第一封信 contains four characters, but it can be divided into one, two, three, or four words by different linguists. Other difficult cases include compound words (e.g., 豬肉), split words (e.g., 洗個澡), acronyms (e.g.,原委會), and literay words.
2. Word segmentation can not be fully automatic.
3. Well-defined tag set for Chinese part-of-speech is not available.
4. A Chinese lexicon with complete parts-of-speech is hard to find.
5. Chinese part-of-speech tagging is difficult even for human, i.e., the parts-of-speech for many words are either arguable or difficult to decide.
6. Manually tagged Chinese corpora, counterparts of Brown corpus and LOB corpus in Chinese, are not available.

These intertwined problems make Chinese part-of-speech tagging an especially difficult task.

Lee and Chang Chien [5,6] used a Tri-POS Markov language model and a bootstrap training process for tagging a small Chinese corpus (1714 sentences for training and 233 sentences for testing). They reported a tagging accuracy 81.13% for all words and 87.60% for known words.

In this paper, we present our work on part-of-speech tagging a large Chinese corpus based on a hidden Markov model (HMM). This is among the first reports on automatic Chinese part-of-speech tagging in the literature [5,6].

2 The HMM-based Part-of-Speech Tagger

Kupiec [4] describes a HMM-based tagging system which can be trained with a corpus of untagged text.

There are two new features in Kupiec's tagger: (1) word equivalence classes and (2) predefined networks. Words with the same set of parts-of-speech are defined as an equivalence class. For example, "type" and "store" belong to the equivalence class *noun-or-verb*. This not only reduces the number of parameters effectively and also makes the tagging system robust. The first-order model is extended with predefined networks based on error analysis and linguistic considerations. Their experimental results show that the predefined networks reduced the overall error rate by only 0.2%. Thus, we adopt the concept of equivalence classes but consider that predefined networks are not worthwhile.

Let us briefly review the formulation of HMM for part-of-speech tagging: A first-order HMM of N states and M possible observations has three sets of parameters: state transition probability distribution A (N by N), observation probability distribution B (N by M), and initial state distribution P (N). For an observation sequence O of length T, there are algorithms, e.g., Viterbi, to uncover the hidden state sequence I. For tagging, N is the number of parts-of-speech in the language, M can be the number of words or the number of equivalence classes (as Kupiec defined). In Chinese, the number of words is more than 100,000 while the number of equivalence classes is less than 1,000. The use of equivalence classes reduces the size of B by 100 times.

The problem of tagging is: Given a word sequence (observations), find out the correct part-of-speech sequence (states).

2.1 The Part-of-Speech Tag Set

The tag set contains 46 regular tags plus 11 special tags. Regular tags include A0 (adjective), C0-C1 (conjunctions), D0-D2 (pronouns), I0 (interjection), M0 (measure), N0-N9 (nouns), P0 (preposition), R0-R6 (particles), T0 (mood), U0-U4 (numbers), V0-V4 (verbs), X0 (onomatope), Y0-Y4 (compounds), Z0-Z2 (adverbs). Special tags are for punctuations (PAR, SEN, PCT, DUN, COM, SEM, COL), unknown words (UNK), foreign words (ABC), and composed numbers (NUM, ARA). It is simplified and reorganized from the classification of Chinese Knowledge Information Processing Group (CKIP), Academia Sinica, Taipei. The original CKIP classification is a five-level system, too complicated even for human to use. Sun [12] designed a three-level tag set TUCWS of 120 tags for Chinese word segmentation. However, they tag the corpus by hand without an automatic tagger. Thus, it is difficult to decide if the set is good for automatic tagging. Other Chinese tag sets can be found in the literature: 33 tags in Su [11], 30 tags in Lee and Chang Chien [5], and 34 tags in Lee *et al.* [6]. These three tag sets are of two origins, CKIP [5] and NTHU [6, 11]. The numbers of tags in them are considered too small.

2.2 Corpus Preparation

The 1991 United Daily corpus contains more than 10 million Chinese characters, about twenty days of news articles published by United Informatics, Inc. during January through March 1991. Basically, it is a collection of articles in the form of raw text (i.e., character stream). Thus, we have to segment the character stream into a word stream before it can be used for training or testing the model. The corpus preparation process consists of the following steps:

Preprocessing Clean up inappropriate parts, such as titles, parenthesized texts, reporter information, figures, etc., in the input article. Articles mostly composed of inappropriate parts are deleted.

Clause identification Divide up the article into clauses delimited by clause-ending punctuations such as periods, commas, question marks.

Automatic word segmentation
Segment the characters in a clause into words using a dictionary-based, Viterbi decoding word identification system.

Manual correction (optional) Check the segmented text to correct segmentation errors due to unregistered words or inaccuracy of the segmentation algorithm. This step is optional but helpful especially for training.

Equivalence class look-up Words in the clause are then converted to identifiers of equivalence class (EQC-ids) via dictionary look-up.

After the above steps, an article is converted into a series of sequences of EQC-ids.

Manual tagging of the whole corpus would take several man-years. However, tagged corpus is necessary for evaluation of the model and helpful for initialization of the HMM parameters as Merialdo [8] pointed out. Thus, we also tag part of the corpus by the steps below: (1) Train the HMM using the articles to be tagged; (2) Tag the articles using the trained HMM; (3) Correct the erroneous tags by hand.

2.3 Training the Model

The untagged corpus of EQC-ids is then used for training the HMM for tagging using the Baum-Welch reestimation procedure with multiple observation sequences [9]. Before training, the model parameters, A, B, P, can be initialized with a tagged corpus.

A The tag bigrams in the tagged corpus are counted to initialize A, the state transition matrix. All counts are incremented by one then normalized.

B The EQC-id to tag correspondences are counted to set up B, the observation matrix. All possible states for an EQC are then incremented by one.

P The initial state matrix P is initialized by counting the tags of first words in the clause. All counts are incremented by one then normalized.

After training, the model parameters are adjusted to bestly predict the most probable tag sequence for the training data.

2.4 Automatic Tagging

Having the trained model parameters, we can automatically tag an unseen text based on an HMM decoding algorithm such as Viterbi's. For a given clause, the tagging process is:

Automatic word segmentation Segment the characters in the clause into words using the above-mentioned word identification system.

Equivalence class look-up Words in the clause are then converted to EQC-ids via dictionary look-up.

Viterbi decoding The sequence of EQC-ids, as observations, is then fed to the Viterbi decoder in order to find out the most probable hidden state sequence, namely, the tag sequence.

Pattern-driven Tag Correction First-order models are not enough to describe local constraints for predicting part-of-speech tags. Higher-order models have much more parameters to estimate and need a lot more training data and resources (memory, CPU time). Kupiec [4] proposed using networks to model higher-order context based on error analysis and linguistic considerations. However, using networks is considered not elegant and had only very limited success. We use a simple pattern-driven tag corrector to postprocess the tag output: The EQC-id sequence is matched against predefined patterns; when a match is found, the corresponding tag corrections are made. These patterns are designed according to analysis of error patterns.

2.5 The Dictionary

The general dictionary has some 80,000 lexical entries each of which contains the Chinese characters and its EQC-id. The original dictionary is a collaborated work of CCL/ITRI with Academia Sinica, Taipei: ITRI collected the words, their pronunciations and word frequencies, while Academia Sinica provided syntactic and semantic markers. For our purpose, only the words and their syntactic information (parts-of-speech) are useful. As mentioned, we restructured the general dictionary based on our newly designed compact tag set. For purpose of comparison, we also constructed a closed dictionary in which the words and their tags in the training and testing corpora are collected.

2.6 An Example

In the following, we use a real-world example to illustrate the tagging process.

- A News Paragraph

  今年地價稅自十一月十六日開徵以來，每天仍有一百通以上電話，向台北市稅捐處查詢如何申請以自用地宅用地稅率千分之二核課，

- Clause Identification

  1. 今年地價稅自十一月十六日開徵以來，
  2. 每天仍有一百通以上電話，

3. 向台北市稅捐處查詢如何申請以自用地宅用地稅率千分之二核課,

- Word Segmentation

1. 今年 地價稅 自 十一 月 十六 日 開徵 以來 ,
2. 每 天 仍 有 一百 通 以上 電話 ,
3. 向 台北 市 稅捐處 查詢 如何 申請 以 自用地宅 用地 稅率 千 分之 二 核課 ,

- EQC-ids

1. 123 0 120 2 134 2 135 115 128 8
2. 124 135 112 15 2 234 152 116 8
3. 45 121 42 0 133 158 133 20 0 116 116 269 314 130 0 8

- Equivalence Classes

1. N1 UNK P0 NUM M0 NUM M0N3 V0 N8 COM
2. D1 M0N3 Z0 P0V2 NUM A0M0V0 D0N8 N0 COM
3. P0V0 N3 M0N4 UNK N0V0 D2Z0 N0V0 C1N3P0 UNK N0 N0 D1N5U0 U2 U0 UNK COM

- Tagging Results

1. N1 UNK P0 NUM M0 NUM M0 V0 N8 COM
2. D1 M0 Z0 V2 NUM M0 N8 N0 COM
3. P0 N3 N4 UNK V0 D2 N0 P0 UNK N0 N0 U0 U2 U0 UNK COM

- Correct Tags

1. N1 N0 P0 NUM M0 NUM M0 V0 N8 COM
2. D1 M0 Z0 V2 NUM M0 N8 N0 COM
3. P0 N3 N4 N2 V0 D2 V0 P0 N0 N0 N0 U0 U2 U0 V0 COM

3 Experimental Results

The whole tagging system, including word segmentation module, equivalence class mapper, HMM trainer, and Viterbi decoder, is implemented in C on a Sun Sparcstation.

A tagged corpus, called **corpus1**, was prepared through the steps described in the subsection **Corpus Preparation**. The corpus is composed of 1,418 clauses or 12,284 word tokens. A larger corpus, called **corpus3**, contains 3,784 clauses. **corpus3** is segmented but untagged, useful only for training.

There are totally 338 word equivalence classes: Each of the 100 most frequently used ambiguous words is assigned a unique EQC-id; the rest 238 EQC-ids are assigned to sets of words with the same set of possible tags.

3.1 Inside Test, Uniformly Initialized, General Dictionary

Condition	*#Words*	*#Hits*	*Accuracy*
All	12,284	10,610	86.37%
Known	11,389	10,610	93.16%
Ambiguous	3,906	3,135	80.26%

Table 1: Accuracy Rates (inside, uniform, general)

Table 1 shows the experimental results for an inside test on **corpus1**. The 80,000-word general dictionary was used and the model parameters are uniformly initialized, i.e., the tags in the corpus are not used to initialize the parameters.

The accuracy rate for all words is 86.37% (1,674 errors out of 12,284 words). Excluding unknown words (words not in the dictionary), the accuracy rate is 93.16% (779 errors). In other words, approximately half of the errors can be attributed to unknown words. If we only consider ambiguous (multi-POS) words, the accuracy is 80.26% (771 errors). We can also observe that only about 35% of the words are ambiguous. (The difference between the latter two numbers of error is due to special usage of some registered words, e.g., 天天 'everyday' is Z0 (adverb) in the dictionary but is used as a company name N2 in 天天百貨 'Everyday Department Store'.)

3.2 Inside Test, Initialized with Tagged Text, General Dictionary

Tagged texts are useful for initializing the model parameters before training. Table 2 shows that the accuracy for ambiguous words was improved by about three percent (from 80.26% to 83.21%). The accuracy

Condition	*#Words*	*#Hits*	*Accuracy*
All	12,284	10,725	87.31%
Known	11,389	10,725	94.17%
Ambiguous	3,906	3,250	83.21%

Table 2: Accuracy Rates (inside, initialized, general)

rate for known words was also improved to more than 94 percent.

3.3 Inside Test, Closed Dictionary

Condition	*#Words*	*#Hits*	*Accuracy*
All	12,284	11,895	96.83%
Known	12,284	11,895	96.83%
Ambiguous	2,432	2,043	84.00%

Table 3: Accuracy Rates (inside, closed)

All words and their used tags in corpus1 are collected to form an ideal dictionary, so-called closed dictionary, for tagging the corpus. The HMM-based tagger is able to correctly tag 96.83% of all words or 84.00% of ambiguous words (Table 3). The accuracy rate is comparable to that of Kupiec's HMM-based English tagger for the well-known Brown corpus.

3.4 Outside Test, General Dictionary

Train	*Test*	*All*	*Known*	*Ambiguous*
800	618	85.80%	92.37%	78.16%
1,000	418	86.58%	92.83%	79.95%
1,200	218	86.90%	92.16%	79.40%
3,784	1,418	85.14%	91.83%	76.40%

Table 4: Accuracy Rates (outside, general)

Table 4 shows the results for outside tests. The corpus is divided into two parts: one for training, the other for testing. The first two columns (Train and Test) are the numbers of clauses (not words) used for training and testing, respectively. The accuracy rates are not as good as those for inside tests: degraded by about 2 percent for known words, by 5 percent for ambiguous words. In general, the system is able to tag approximately 80 percent of ambiguous words correctly.

In the last row, corpus3 (3,784 clauses, 35,849 words, translated AP news) was used for training while corpus1 (1,418 clauses, 12,284 words, domestic news) for testing. Due to difference of text type, accuracy rates are degraded by about 3 percent for ambiguous words. However, the system is still able to assign correct tags to 91.83 percent of all words. This shows the robustness of the model, due to the concept of equivalence classes.

3.5 Outside Test, Closed Dictionary

Train	*Test*	*All*	*Known*	*Ambiguous*
800	618	96.01%	96.01%	80.24%
1,000	418	96.20%	96.20%	82.27%
1,200	218	95.41%	95.41%	79.91%

Table 5: Accuracy Rates (outside, closed)

Table 5 summarizes the results for outside tests on closed dictionary. Approximately 96% of all words and 80% of ambiguous words are tagged correctly.

4 Error Analysis

4.1 Confusion Matrix

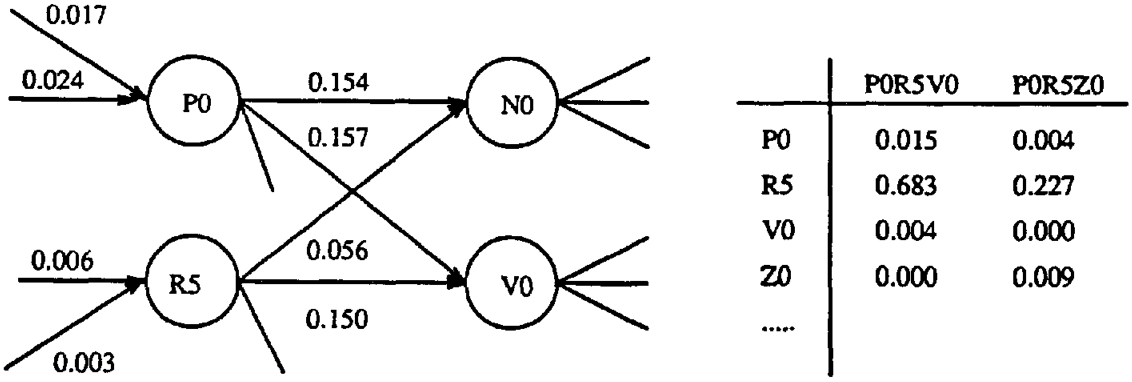
Table 6 shows part of the confusion matrix for the test described in subsection 3.2; only the confusing parts-of-speech are shown.

The ANVZ problem: Due to lack of inflections in Chinese, a Chinese word can have many different parts-of-speech, yet only one form. It is sometimes very difficult even for human to identify the correct tag. For example, Chinese does not have -ing ending for nominalization of verbs, -ly for adverbs, -tion for verbal nouns, -en for past participles. Thus, a word such as 分散 can be a verb (V0) 'distribute', a noun (N0) 'distribution', an adjective (A0) 'distributive', 'distributing' or 'distributed', and an adverb (Z0) 'distributively' in different contexts. Nouns and verbs are especially hard to distinguish. That is why the V0-N0 (180), N0-V0 (47) confusions are common.

The RP problem: Open classes, such as nouns and verbs, have large population, while closed classed, such as prepositions and particles have small

	A0	C0	C1	N0	P0	R0	R5	V0	V4	Z0	others	rate
A0	63	0	0	**27**	0	0	0	**42**	0	3	14	42.3%
C0	0	106	0	0	8	0	0	0	0	1	0	92.2%
C1	0	14	19	0	5	0	0	3	0	2	0	44.2%
N0	0	0	0	281	0	0	0	**47**	0	1	10	81.2%
P0	0	8	1	0	195	0	**154**	**25**	**26**	1	6	46.8%
R0	0	0	0	0	0	452	0	0	0	0	0	100.0%
R5	0	0	0	0	0	0	20	0	0	12	0	62.5%
V0	1	0	8	**180**	10	0	0	461	4	6	14	67.4%
V4	0	0	0	0	0	0	0	3	27	0	3	81.8%
Z0	13	11	5	3	0	0	3	5	0	222	8	82.2%
T0	0	0	0	0	0	10	0	0	0	0	8	5.6%

Table 6: Part of Confusion Matrix



	P0R5V0	P0R5Z0
P0	0.015	0.004
R5	0.683	0.227
V0	0.004	0.000
Z0	0.000	0.009
.....		

A: state transition probabilities

B: observation probabilities

Figure 1: The RP Problem

population. In general, this is not a problem for tagging. However, in our tag set, R5 (aspect prefix) has only three members 在 (P0 R5 V0), 將, and 正. The former two words are also common prepositions (P0). From the experiments, we observed that while 在 is a preposition in most instances, it is always tagged as R5 (aspect). After studying the trained model parameters A, B, P, we found (Figure 1) that R5 was assigned large probabilities in B matrix (0.683 for 在 , 0.227 for 將) since R5 has only three words while P0 was assigned much smaller probabilities (Due to the probabilistic characteristic, sum of the observation probabilities for a state, such as P0, R5, must be one.) In addition, R5 and P0 have not significant difference in the incoming or outcoming entries of A matrix because of the characteristic of unsupervised learning: all instances of 在 are considered as possible candidates for R5. We consider this as a weakness of HMM for tagging.

4.2 Error Patterns

Tagging errors usually occur in clusters; that is, an error may cause further mistagging of its neighbors if they are also ambiguous. Common patterns of mistagging include V0-V0 (as N0-N0), Z0-V0 (as A0-N0), V0-N0 (as C1-Z2), V0-P0 (as N0-R5), P0-N0 (as R5-V0), P0-N1 (as R5-V4), and N0-V0-N0 (as U1-C1-Z2). They can be classified into three types:

ANVZ type These error patterns are due to the above-mentioned ANVZ problem. This type of error is reasonable.

RP type Those error patterns involving R5 are due to the RP problem. The type of error should be eliminated by model improvement or postprocessing.

idiomatic type Some idiomatic expressions are composed of highly ambiguous words. For example, in "以 ... 為 準", all the three words 以 (C1 N3 P0), 為 (C1 P0 V0), 準 (A0 N0 Z2), are 3-way ambiguous words. That is why the V0-N0 sequence is frequently mistagged as C1-Z2.

If we consider the mistagging of unknown words, more long tagging error clusters would appear. Actually, an unknown word not only causes mistagging of the word itself but also affects the tagging of its neighbors.

4.3 Without Equivalence Classes

Train	*Test*	*w/o EQC*	*EQC*
800	618	77.02%	80.24%
1,000	418	76.90%	82.27%
1,200	218	77.68%	79.91%
1,418	inside	83.80%	84.00%

Table 7: Accuracy Rates (closed, ambiguous words only)

To verify feasibility of the concept of equivalence classes, we implemented a version of the HMM tagger considering each word as a unique observation (without EQC). Table 7 compares the results for inside/outside tests on closed dictionary. To our surprise, the concept of equivalence classes not only has the advantages of saving space/time and making the tagger robust but also achieve higher tagging accuracy, especially in case of outside tests. This might be due to insufficient training data for the much larger number of parameters to estimate. Nevertheless, it also proves that the concept is valid and useful.

5 Concluding Remarks

We have presented our initial effort for Chinese part-of-speech tagging using a first-order fully-connected hidden Markov model and Kupiec's concept of equivalence classes. The experimental results show that the tagging model is promising. We have also discussed our observations on some imperfections of the current model. In the near future, we will (1) use the whole UD corpus to further validate and verify the system, (2) try to implement a second-order HMM, (3) attempt to solve part of the unknown word tagging problem, (4) attempt to solve part of the compound word problem, (5) use heuristic rules for postprocessing the tagging output, (6) perform word identification and part-of-speech tagging concurrently, and (7) integrate the tagging HMM with the linguistic decoder of a Chinese speech recognition system.

Acknowledgements

This paper is a partial result of the project no. 37H2100 conducted by the ITRI under sponsorship of the Minister of Economic Affairs, R.O.C.

References

[1] K. Church. A stochastic parts program and noun phrase parser for unresticted text. In *Proc. of ICASSP-89*, pages 695–698, Glasgow, Scotland, 1989.

[2] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proc. of the Third Conference on Applied Natural Language Processing*, Trento, Italy, April 1992.

[3] S. DeRose. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14:31–39, 1988.

[4] J. Kupiec. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, 6:225–242, 1992.

[5] H.-J. Lee and C.-H. Chang Chien. A Markov language model in handwritten Chinese text recognition. In *Proc. of Workshop on Corpus-based Researches and Techniques for Natural Language Processing*, Taipei, Taiwan, September 1992.

[6] H.-J. Lee, C.-H. Dung, F.-M. Lai, and C.-H. Chang Chien. Applications of Markov language models. In *Proc. of Workshop on Advanced Information Systems*, Hsinchu, Taiwan, May 1993.

[7] Y.-C. Lin, T.-H. Chiang, and K.-Y. Su. Discrimination oriented probabilistic tagging. In *Proc. of ROCLING V*, pages 87–96, Taipei, 1992.

[8] B. Merialdo. Tagging text with a probabilistic model. In *Proc. of ICASSP-91*, pages 809–812, Toronto, 1991.

[9] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[10] B. Santorini. Part-of-speech tagging guidelines for the Penn Treebank project. University of Pennsylvania, Pennsylvania, March 1991.

[11] K.-Y. Su, Y.-L. Hsu, and C. Saillard. Constructing a phrase structure grammar by incorporating linguistic knowledge and statistical log-likelihood ratio. In *Proc. of ROCLING IV*, pages 257–275, Pingtung, Taiwan, 1991.

[12] M.S. Sun, T.B.Y. Lai, S.C. Lun, and C.F. Sun. The design of a tagset for Chinese word segmentation. In *First International Conference on Chinese Linguistics*, Singapore, June 1992.