Computational Linguistics and Chinese Language Processing
Vol. 14, No. 3, September 2009, pp. 237-256

Modeling Taiwanese POS Tagging Using Statistical Methods and Mandarin Training Data

Un-Gian Iunn[*], Jia-hung Tai[+], Kiat-Gak Lau[#], Cheng-yan Kao[*], and Keh-jiann Chen[+]

Abstract

In this paper, we introduce a POS tagging method for Taiwan Southern Min. We use the more than 62,000 entries of the Taiwanese-Mandarin dictionary and 10 million words of Mandarin training data to tag Taiwanese. The literary written Taiwanese corpora have both Romanized script and Han-Romanization mixed script, and include prose, novels, and dramas. We follow the tagset drawn up by CKIP.

We developed a word alignment checker to assist with the word alignment for the two scripts. It searches the Taiwanese-Mandarin dictionary to find corresponding Mandarin candidate words, selects the most suitable Mandarin word using an HMM probabilistic model from the Mandarin training data, and tags the word using an MEMM classifier.

We achieve an accuracy rate of 91.6% on Taiwanese POS tagging work, and we analyze the errors. We also discover some preliminary Taiwanese training data.

Keywords: Taiwan Southern Min, POS tagging, written Taiwanese, Hidden Markov Model, Maximal Entropy Markov Model.

[*] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
E-mail: {d93001, cykao}@csie.ntu.edu.tw
[+] Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: {glaxy, kchen}@iis.sinica.edu.tw
[#] Independent scholar
E-mail: kiatgak01@gmail.com

1. Introduction

1.1 Background

There are about 46 million Southern Min speakers in the world. If we list languages by the size of their speaking population, Southern Min is ranked 21. The Southern Min speakers are mainly distributed in eight countries(Gordon, 2005). It is an important language that has received very little attention.

The percentage of Southern Min speakers in Taiwan was over 70% (Huang, 1995). Taiwan has the highest percentage of Southern Min speakers in the world. We will call this language as "Taiwanese" for simplification in this paper.

Many different types of written Taiwanese systems exist. Among these systems, the Han character script and one of the Romanized scripts (Pe̍h-ōe-jī, 白話字, *abbrev*. POJ, vernacular writing) are the most popular. Also, the mixture of the above two scripts, called the Han-Romanization mixed script (*abbrev*. as HR mixed script), has been adopted by many people (Iunn, 2009).

1.2 Motivation

In order to establish the bases of written Taiwanese processing, we have constructed some tools over the past few years, including an online Taiwanese syllable dictionary (Iunn, 2003a); an online Taiwanese-Mandarin dictionary (abbrev. OTMD) (Iunn, 2000, 2003b); a 5,800,000 syllable HR mixed script and 3,400,000 syllable POJ script Taiwanese corpus; the online Taiwanese concordancer system based on this corpus (Iunn, 2003c; Iunn & Lau, 2007); preliminary Taiwanese word frequency reports for the Taiwanese POJ and HR mixed scripts, based on the above Taiwanese corpus (Iunn, 2005); the digital archive database for written Taiwanese (*abbrev*. DADWT) literature data with POJ and HR mixed script paragraph alignment (Iunn, 2007); *etc*.

We intend to annotate the Taiwanese corpus with POS markers for more advanced applications, including Taiwanese tone sandhi TTS system improvement (Iunn *et al*. 2007), Taiwanese Treebank construction, *etc*.

1.3 Problem

The primary difficulty encountered in the POS tagging of Taiwanese corpora is the question, "What is the Taiwanese POS tagset?" To date, no standard tagset for Taiwanese has been proposed. Under the circumstances, we have temporarily employed the Chinese POS tagset established by the CKIP Group of Academia Sinica (CKIP, 1993). Unfortunately, we still encountered some problems because we did not have a Taiwanese dictionary that contained

the Mandarin POS tagset. The existing Taiwanese dictionaries merely contain basic vocabulary words, that is, nouns, verbs, adjectives, *etc*.

Moreover, there was another problem to surmount – manpower shortage. We did not have enough manpower to fully execute the POS tagging of the Taiwanese corpora.

Therefore, we proposed employing statistical procedures with the existing Mandarin resources and the OTMD to automatically complete the Taiwanese POS tagging. We used the Mandarin language model under the assumption that the word sequence in Taiwanese is similar to Mandarin.

1.4 Review

Shi (2006) translated the Mandarin sentences in the book, "Modern Chinese 800 words '現代漢語八百詞' " (by Shu-xiang Lü) into Taiwanese and Hakka to establish the T3 corpus and developed some editing tools to help in the construction of the T3 Treebank. Chou (2006) used the Brill tagger based on the HMM model to tag words in the T3 Treebank. They used a tagset size of 26, and attained tagging accuracy rates of 92.80% and 85.59% for the training and test data, respectively.

T3 Treebank has not been released publicly. Thus, we decided to use different tagsets and different tagged corpora in our experiments.

2. POS Tagging Method

Figure 1 shows our system architecture diagram.

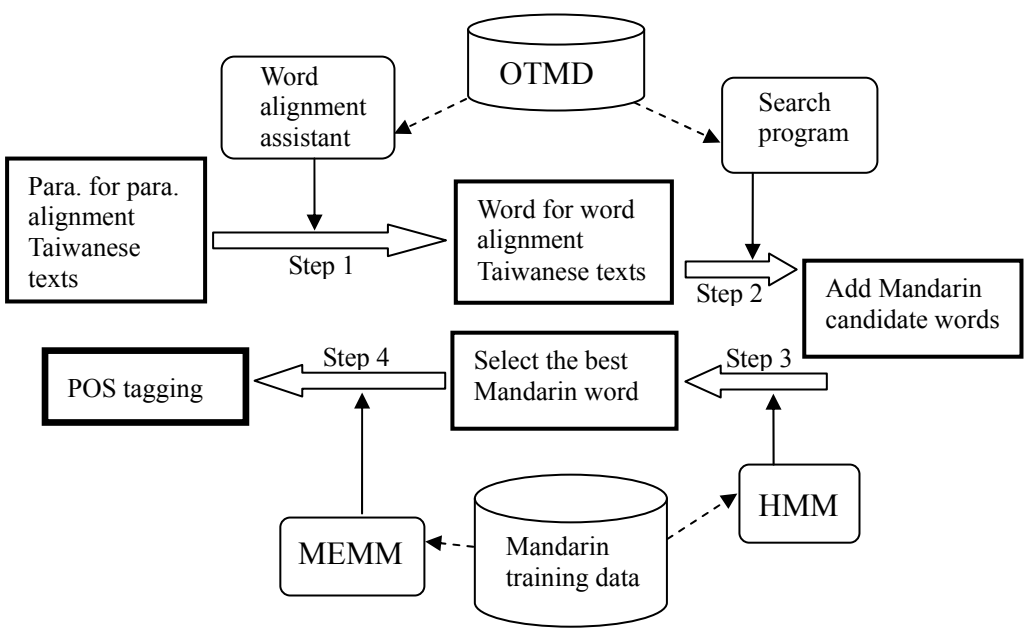

Figure 1. Taiwanese POS Tagging System Architecture Diagram

At first, the text contains both POJ and HR mixed scripts with paragraph by paragraph alignment. Step 1 converts the texts to word alignment form. Step 2 adds the Mandarin candidate words (translations). Step 3 selects the best Mandarin translation using the HMM model. Finally, we decide the POS tagging of each word using the MEMM model. The following subsection will describe this process in detail.

For example, the original texts are

"Tâi-ôan tē-it kôan ê Gio̍k-san ê hū-kūn khah kē ê só͘-chāi ... " and

"台灣 第一 懸 ê 玉山 ê 附近 較 低 ê 所在 ... "

Taiwan first high of Mt.Jade of nearby more low of place

Step 1 converts the texts to word alignment form:

"台灣/Tâi-ôan 第一/tē-it 懸/kôan ê/ê 玉山/Gio̍k-san ê/ê 附近/hū-kūn 較/khah 低/kē ê/ê 所在/só͘-chāi … "

Then, Step 2 adds the Mandarin translations:

"台灣/Tâi-ôan{台灣} 第一/tē-it{第一;絕頂} 懸/kôan{高} ê/ê{的} 玉山/Gio̍k-san{玉山}

ê/ê{的} 附近/hū-kūn{附近} 較/khah{較} 低/kē{低} ê/ê{的} 所在/só͘-chāi(去處;

地方;角頭;所在;處所;場所;間量} …"

Step 3 selects the best Mandarin translation using the HMM model (we omit the original Taiwanese texts):

"台灣 第一 高 的 玉山 的 附近 較 低 的 地方 …"

Finally, Step 4 decides the POS tagging of each word using the MEMM model:

"台灣/Tâi-ôan(Nc) 第一/tē-it(Neu) 懸/kôan(VH) ê/ê(DE) 玉山/ Gio̍k -san(Nc) ê/ê(DE)

附近/hū-kūn(Nc) 較/khah(Dfa) 低/kē(VH) ê/ê(DE) 所在/só͘-chāi(Na)… "

We will illustrate our work with Figure 1 in the following subsections.

2.1 Origin of the Corpus

The corpus we chose is part of the DADWT project achievements of the National Museum of Taiwan Literature. It contains both POJ and HR mixed scripts with paragraph by paragraph alignment, including novels, prose, dramas, and poems (Iunn, 2007).

2.2 Word by Word Alignment

First, we developed a word alignment program to aid manual processing. We arranged the word alignment of the two scripts, where the paragraphs were already aligned. This program not only collates the number of syllables in the two scripts, but it also compares and contrasts the two scripts with the entries of the OTMD. If the program does not find the two scripts within the same entry, it highlights the corresponding words to remind the user that the word may be an unknown word, an inconsistent usage of the Han character, or a typographical error.

The OTMD was announced and has been online since 2000. The main data provider is Robert L. Cheng, but many anonymous contributors also offer entries and correct the typographical errors. There are a total of more than 62,000 entries. The URL is http://iug.csie.dahan.edu. tw/q. This dictionary offers POJ, HR mixed script, and Mandarin fields, with the POJ field also offering the different accents. The pronunciation function was added in 2006, and English translation was added to more than 10,000 entries in 2007 based on Embree (1984), which contains English, Mandarin, and POJ fields.

2.3 Finding the Corresponding Mandarin Candidate Words

Next, we continued to search for the corresponding Mandarin candidate words from the POJ and HR mixed script word pairs via the OTMD. The mapping was one-to-many. In short, a Taiwanese word pair would have more than one Mandarin word counterpart. For example, "愛/ài" in Taiwanese has the meanings of "愛"'love (person),' "喜歡"'like (thing),' "要" 'want to,' "需要"'need to,' *etc.* in Mandarin. Nevertheless, we were not able to find counterparts for certain words, since they were not contained in the OTMD. We also found some that had different HR mixed script usage.

For instance, the Taiwanese word that appears as "較贏/khah-iâⁿ" 'more than' in the corpus appears as "khah 贏/khah-iâⁿ" in the dictionary. With regard to problems of this nature, we applied the following solution. If the POJ and HR mixed script word pair could not be found, we temporarily removed the HR mixed script and searched for the Mandarin word counterpart again using the POJ script. If the characters of HR mixed script were all Han characters, we regarded the Han characters as one of a Mandarin candidate word (assuming that the word is common to both Taiwanese and Mandarin).

This method might increase the number of the Mandarin candidate words, especially for single syllable words. For instance, the word pair "轉/chōan"'turn' appears in the text. We could not find an entry that contains both "轉" and "chōan" in the OTMD. The corresponding Mandarin translations of "chōan" in the dictionary are "扭"'twist' and "上" 'up'. We added "轉"'turn' as the supplementary Mandarin translation, but the meanings of these three words differ.

Table 1. Partial Entries of the OTMD

HR Mixed Script	POJ Script	Mandarin Translation
chōan	chōan	扭
撰	chōan	上

Note: There exists not "轉/chōan" entry in the OTMD. The Mandarin translation of "轉/chōan" will be "扭," "上" and "轉"

If the strategy was still unable to find any results, the HR mixed script was directly recognized as the Mandarin candidate word. For instance, no dictionary entry was found for the word pair appearing as "有形/iú-hêng"'tangible' in the text, neither could one be found in the search using the POJ script "iú-hêng." So, the HR mixed script "有形" was directly recognized as the Mandarin candidate word (Lau, 2007).

2.4 Selecting the Best Mandarin Translation

We employed the Hidden Markov Model and Viterbi algorithm, and we made use of the bigram word training data of the ten-million word balanced Sinica corpus of the CKIP Group of Academia Sinica to select the most appropriate corresponding Mandarin word from the Mandarin candidate words. Figure 2 is an example. The selected words are boxed and bold.

Taiwanese Word	對/ Tùi 'from'	古早/ kó͘ –chá 'ago'	以來/ í-lâi 'since'	琴/ khîm 'instrument'	有/ ū 'has'	濟濟/ chē-chē 'many'	款/ khóan 'appearance'
Corresponding Mandarin Word(s)	**從** w_{11} 對 w_{12} 對子 w_{13} 對於 w_{14}	**以前** w_{21} 古代 w_{22} 古時候 w_{23} 從前 w_{24}	**以來** w_{31}	**琴** w_{41}	**有** w_{51}	濟濟 w_{61} **很多** w_{62}	**樣子** w_{71} 樣式 w_{72} 整理 w_{73}
	$w_1 = w_{11}$	$w_2 = w_{21}$	$w_3 = w_{31}$	$w_4 = w_{41}$	$w_5 = w_{51}$	$w_6 = w_{62}$	$w_7 = w_{71}$

Figure 2. An Example of Selecting the Best Mandarin Translation

Assume that a particular sentence contains *m* words. The first word, w_1, is selected from the candidate words of $w_{11}, w_{12}, ..., w_{1n_1}$; the second word, w_2, is selected from the candidate words of $w_{21}, w_{22}, ..., w_{2n_2}$; and the m^{th} word, w_m, is selected from the candidate

words of $w_{m1}, w_{m2}, ..., w_{mn_m}$. $\hat{S} = w_1 w_2 \cdots w_m$, which is the most probable word sequence, is selected from the candidate words, such that P($\hat{S} = w_1 w_2 \cdots w_m$) is maximized.

The HMM assumes that the word w_i is only influenced by the previous word w_{i-1}, thus:

$$P(\hat{S} = w_1 w_2 \cdots w_m) \cong P(w_1) \times \prod_{i=2}^{m} P(w_i \mid w_{i-1}) \tag{1}$$

Therefore, it searches for the word sequence $\hat{S} = w_1 w_2 \cdots w_m$, which maximizes

$$\log P(w_1) + \sum_{i=2}^{m} \log P(w_i \mid w_{i-1}) \tag{2}$$

We use the Laplace smoothing method to solve the problem of $P(w_i \mid w_{i-1}) = 0$, where no bigram of $w_{i-1} w_i$ could be found in the training data in other words. It should be noted that the word string $\hat{S}$ may not be a legal Mandarin sentence.

In practice, we use the Viterbi algorithm to eliminate repeated computation and reduce the time complexity from exponential time to polynomial time. If a sentence S has *m* words, and every word has *n* candidate words, the time complexity will be $O(n^m)$. The Viterbi algorithm reduces the time complexity to $O(n^2 \times m)$ (Manning & Schütze, 1999).

2.5 Selecting the Most Appropriate POS According to the Corresponding Mandarin Word

We applied the Maximal Entropy Markov Model (MEMM) to the POS tag selection.

Manning and Schütze (1999) stated that "Maximum entropy modeling is a framework for integrating information from many heterogeneous information sources for classification. The data for a classification problem is described as a number of features. Each feature corresponds to a constraint on the model. …Choosing the maximum entropy model is motivated by the desire to preserve as much uncertainty as possible."

MEMM includes a set of possible word and tag contexts, or "histories" (*H*), and the POS tagging set (*T*):

$$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)} \tag{3}$$

where $h \in H, t \in T$, π is a normalization constant, $\{\mu, \alpha_1, ..., \alpha_k\}$ are the positive model parameters, and $\{f_1, ..., f_k\}$ stands for the features $f_j(h,t) \in \{0,1\}$. Parameter α_j corresponds to the feature f_j. The parameters $\{\mu, \alpha_1, ..., \alpha_k\}$ are then chosen to maximize the likelihood of the training data using p:

$$L(p) = \prod_{i=1}^{n} p(h_i, t_i) = \prod_{i=1}^{n} \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h_i,t_i)} \tag{4}$$

As for the POS tag t_i of the target word w_i, we selected ten features including:

(a) Words – five types of feature patterns: $w_i, w_{i-1}, w_{i-2}w_{i-1}, w_{i+1}, w_{i+1}w_{i+2}$.

(b) POS – two types of feature patterns: $t_{i-1}, t_{i-2}t_{i-1}$.

(c) Morpheme – three types of feature patterns: m_1, m_2, m_n.

The feature patterns m_1, m_2, m_n are designated to manipulate the unknown words. If w_i is an unknown word, we segment w_i with a maximal matching strategy; thus, $w_i = m_1 m_2 \cdots m_n$ and, under certain circumstances, $m_2 = m_3 = \cdots = m_n$. If w_i is a known word, the three morpheme features are set to null. Moreover, if w_i is at the beginning or end of a sentence, certain features are likewise given a null value. For instance, when *i=1*, the feature values of $w_{i-1}, w_{i-2}w_{i-1}, t_{i-1}, t_{i-2}t_{i-1}$, *etc*. are also null (Berger *et al*., 1996; McCallum *et al*., 2000; Rabiner, 1989; Ratnaparkhi, 1996; Samuelsson, 2003; Tai, 2007; Tsai & Chen, 2004).

In MEMM, the dependencies of observations are flexibly modeled whereas HMM assumes that observations are independent. We think MEMM is more suitable for the POS tagging task.

We used the "Maximum Entropy Modeling Toolkit for Python and C++" package provided by Zhang Le to implement our system (Le, 2003). The ten-million word POS tagged balanced Sinica corpus of the CKIP Group was used as the training data. Several million features were expanded from the ten features mentioned above, and the training time was about two days on Windows Server 2003 x64 SP2 with an Intel Xeon 3.2GHz processor (Quad-core), 8G DRAM.

3. Results

We used the aforementioned method to perform the Taiwanese POS tagging task; nevertheless, as no standard answers were available to gauge the accuracy rate, we extracted partial results and checked them manually. The primary consideration of the manual checking procedure was the Chinese Word Segmentation and Tagging System of the CKIP group of Academia Sinica (CKIP, 2004). We selected fourteen literary works belonging to three different eras – the Ching Dynasty, the Japanese-ruled Period, and the Post-war Era. These literary works were in the form of prose (seven), novels (five), and dramas (two). We selected the first paragraph from each composition, or, if the length (number of syllables) of the first paragraph was less than 60, we selected the second paragraph.

$$accuracy\ rate = (1 - \frac{number\ of\ tagging\ errors}{number\ of\ total\ words})\ 100\% \quad (5)$$

The test data list is shown in the Appendix. Table 2 shows the test data selected for manual checking. The number of syllables, words, and incorrectly selected Mandarin words,

as well as the POS tagging inaccuracy of each paragraph are noted.

A total of 1,038 words (1,496 syllables) were selected, and manual checking showed that 90 words had been incorrectly selected and 87 words were found to have inaccurate POS tagging, thus placing the average POS tagging accuracy rate at 91.6%. It should be noted that sometimes, even when the corresponding Mandarin word selected was inappropriate, the POS tagging result was still accurate. On the other hand, an appropriate or correct corresponding Mandarin word did not always have accurate POS tagging.

Furthermore, sometimes one Taiwanese word would correspond to two Mandarin words. For instance, while the Taiwanese word "壁頂/piah-téng" 'on the wall' is treated as only one word, the Mandarin translation "牆壁 上" should be treated as two words. There are also occasions wherein two Taiwanese words would correspond to only one Mandarin word counterpart. For instance, the Mandarin counterpart of the Taiwanese words "Tiong-kok/中國" 'Chinese' and "jī/字" 'character' was "中國字." The former is processed as an unknown word, whereas the latter, which was separated into two independent words, was processed as two words. In these types of cases, if the POS tagging was accurate, we still regarded the results as accurate. If they were to be regarded as incorrect, the average accuracy rate would drop by around 2%.

Table 2. Tagging Accuracy Rate of The Test Data

id	No. of Syllables	No. of Words	Errors	Tagging errors	Accuracy rate(%)
1	162	109	9	6	94.5
2	66	46	4	3	93.5
3	180	119	6	8	93.3
4	122	88	3	6	93.2
5	74	51	4	1	98.0
6	75	49	7	7	85.7
7	112	87	13	12	86.2
8	101	77	7	9	88.3
9	133	93	7	9	90.3
10	116	82	3	3	96.3
11	94	59	7	5	91.5
12	92	61	10	10	83.6
13	77	59	8	4	93.2
14	92	58	3	4	93.1
Totally	1,496	1,038	90	87	91.6

In Table 3 (following), examples of actual POS tagging results are shown. It is a part of id 11. In this table, the first field is the HR mixed script and POJ script (contained in brackets), and the second field is the Mandarin candidate word(s). The "@" symbol preceding the word indicates that no entry has been found for the Taiwanese word shown in the first field; hence the HR mixed script automatically served as the Mandarin candidate word. The third field contains the selected Mandarin word, and the final field contains the selected POS. All of the incorrectly selected Mandarin words or incorrectly selected POS tags are underlined and segregated by two asterisks "**" preceding the word. The correct POS tag, contained in parentheses and shown in bold type, is then added after the incorrect POS tag.

Table 3. Example of POS Tagging Result

Taiwanese	Mandarin Candidate Words	Selected Word	POS tagging
我[góa] 'I'	我	我	Nh
將[chiong] 'let'	將	將	D
草帽仔[chháu-bō-á] 'straw hat'	@草帽仔	草帽仔	Na
掛[kòa] 'hang'	帶;掛;戴	**帶 **掛**	VC
tī [tī] 'at'	在	在	P
壁頂[piah-téng] 'on the wall'	牆壁上	牆壁上	Nc
，[,]	，	，	CommaCategory
行李[hêng-lí] 'baggage'	行李	行李	Na
khêng[khêng] 'arrange'	收拾;盤點	收拾	VC
khêng[khêng] 'arrange'	收拾;盤點	收拾	VC
leh[leh] '(modal particle)'	咧	咧	T
，[,]	，	，	CommaCategory
坐[chē] 'sit'	坐	坐	VA
tòa[tòa] 'at'	住	**住	**VCL**(P)**
小店[sió-tiàm] 'store'	@小店	小店	Na
ê[ê] 'of'	的	的	DE
tha-thá-mì[tha-tha-mì] 'tatami'	塌塌米	塌塌米	Na
頂 kôan[téng-kôan] 'above'	上面	上面	Ncd
，[,]	，	，	CommaCategory
看[khòaⁿ] 'see'	看	看	VC
窗外[thang-gōa] 'outside the window'	@窗外	窗外	Nc

ê[ê] ‘of’	的	的	DE
光景[kong-kéng] ‘scenery’	風光;氣象;光景;風景;景氣	**景氣 **風景**	Na
，[,]	，	，	CommaCategory
看[khòaⁿ] ‘see’	看	看	VC
起起[khí-khí] ‘up’	@起起	起起	**Nb**(VA)**
落落[lo̍h-lo̍h] ‘down’	@落落	落落	VA
ê[ê] ‘of’	的	的	DE
海湧[hái-éng] ‘wave’	海浪 ;海潮	海浪	Na
，[,]	，	，	CommaCategory
因爲[in-ūi] ‘because’	由於 ;因爲	因爲	Cbb
等待[tán-thāi] ‘wait’	留待 ;等待	等待	VK
朋友[pêng-iú] ‘friend’	友人 ;朋友	朋友	Na
，[,]	，	，	CommaCategory
心適[sim-sek] ‘pleasant’	好玩 ;好玩兒 ;有趣 ;風趣 ;愉快 ;稀奇 ;鬧著玩	有趣	VH
心適[sim-sek] ‘pleasant’	好玩 ;好玩兒 ;有趣 ;風趣 ;愉快 ;稀奇 ;鬧著玩	有趣	VH
，[,]	，	，	CommaCategory
輕輕仔[khin-khin-á] ‘lightly’	輕輕的	輕輕的	**Nb**(D)**
來[lâi] ‘toward’	來	來	D
點[tiám] ‘light’	燃點;檢點;點;點子	點	VC
一支[chi̍t-ki] ‘a’	@一支	一支	Na
涼涼[liâng-liâng] ‘cool’	冷冷;涼絲絲	**冷冷 **涼涼**	VH
ê[ê] ‘of’	的	的	DE
芎蕉[kin-chio] ‘banana’	香蕉	香蕉	Na
薰[hun] ‘tobacco’	香菸;香煙;薰	香煙	Na
。[.]	。	。	PeriodCategory

4. Error Analysis

This section discusses how a more thorough check was performed to analyze the error conditions.

4.1 Selection of Inappropriate Mandarin Word

An analysis of the errors made in the selection of Mandarin words or POS tags revealed that the selection of inappropriate Mandarin words led to POS tagging errors in 25 cases. Table 4 shows the incorrect Mandarin words selected and their respective POS.

Table 4. The Selected Incorrect Mandarin Words and Their Respective POS

Word	Selected Mandarin word and POS	More appropriate Mandarin word and POS	Remark
押/ah	強制(D) ‘compel’	押(VC) ‘take into custody’	
bat/bat	知道(VK) ‘know’	曾(D) ‘ever’	
無/bô	不(D) ‘not’	沒有(VJ) ‘not have’	2 times
chham/chham	和(P) ‘and’	摻(VC) ‘accompany’	
進前 /chìn-chêng	之前(Ng) ‘before’	向 前(P Nes) ‘forward’	
這號/chit-hō	這樣(VH) ‘such’	這種(Nep Nf) ‘this kind of’	2 times
轉/chōan	上(Ncd) ‘above’	轉(Vac) ‘turn’	2 times
外/gōa	外(Ng) ‘outside’	開外(Neqa) ‘more’	2 times
夭壽/iáu-siū	非常(Dfa) ‘very’	早夭(VH) ‘dead early’	
加/ke	上(Ncd) ‘above’	多(Dfa) ‘more’	
價值/kè-ta̍t	值得(VH) ‘worthy’	價值(Na) ‘value’	
腳/kha	個(DE) ‘(a numerary adjunct)’	下(Ncd) ‘under’	
黃 hóaⁿ/n̂g-hóaⁿ	罕(D) ‘rarely’	淺黃(A) ‘light yellow’	
倚/óa	依(P) ‘in accordance with’	靠(VJ) ‘lean against’	
活/o̍ah	生活(Na) ‘life’	活(VH) ‘live’	
破相/phòa-siùⁿ	破(VHC) ‘break’	殘廢(Na) ‘disabled’	
細漢/sè-hàn	小時候(Nd) ‘in one's childhood’	年幼(VH) ‘young’	
相借問 /sio-chio̍h-mn̄g	招呼(VC) ‘greet’	打招呼(VB) ‘say hello’	
搭/tah	地方(Na) ‘location’	搭(VC) ‘construct’	
tio̍h/tio̍h	就(P) ‘(an auxiliary confirming and stressing the verb following)’	著(VCL) ‘come into contact with’	
著/tio̍h	就(P) ‘(an auxiliary confirming and stressing the verb following)’	得(D) ‘need to’	

4.2 Absence of Appropriate Mandarin Translation in OTMD

There were fourteen errors made in inappropriate Mandarin word selection due to the absence of an appropriate Mandarin word in the OTMD. This also led to errors in the POS tagging. The discovery indicates the necessity of expanding the entries of the OTMD. Table 5 tabulates these errors.

Table 5. Errors Caused by Absence of Appropriate Mandarin Word Option in OTMD

Taiwanese	Selected Mandarin by System	Appropriate Mandarin Word	Remark
chak/chak	促(VF) 'urge'	擠(VC) 'crowd'	
chūn/chūn	絞(VC) 'twist'	陣(Nf) '(a numerary adjunct)'	2 times
kah/kah	和(Caa) 'and'	得(DE) 'a particle used after a verb'	3 times
leh/leh	咧(T) '(modal particle)'	在(P) 'doing'	3 times
煞/soah	結束(VHC) 'finish'	卻(D) 'but'	
teh/teh	在(P) '(an indicator or location)'	著(Di) '(an adverbial particle)'	
頂/téng	頂(VC) 'lift'	上(Nes) '(the first half part)'	
tiāⁿ-tiāⁿ / tiāⁿ-tiāⁿ	常常(D) 'often'	而已(T) 'just'	
轉 / tńg	調解(VC) 'mediate'	轉(VAC) 'turn'	

4.3 Unknown Words from the Viewpoint of Mandarin

Ten of the POS tagging errors were made because the word was an unknown word. Parts of these unknown words correspond to two Mandarin words. These unknown words are tabulated in Table 6.

Table 6. Unknown Words from The Viewpoint of Mandarin

Taiwanese Word	Corresponding Mandarin Word	Selected POS by System	Correct POS
bē 會/bē-ē	不會 'be unable to'	Nb	D
廟埕/biō-tiâⁿ	廟前院 'temple square'	Na	Nc (Na Nc)
食老/chiȧh-lāu	年老 'old'	Na	VH
轉了/chōan-liáu	轉 後 'after turning'	VH	VC Ng
牛摃灣/Gû-taⁿ-oan	牛摃灣 '(a place name)'	VA	Nc
法律上/hoat-lu̍t-siōng	法律上 'jural'	VC	N (Na Ncd)
非爲/hui-ûi	非爲 'infamous conduct'	A	N (A Na)
窮志/kiông-chì	窮志 'exhaust the ambition'	Na	V (VH Na)
輕輕仔/khin-khin-á	輕輕地 'lightly'	Nb	D (VH DE)
生子/seⁿ-kiáⁿ	生孩子 'give birth to a child'	Na	VA (VH Na)

4.4 Propagation Error

Five of the POS tagging errors were probably due to the occurrence of a previous POS tagging error. These are categorized as propagation errors and include one unknown word.

4.5 Other Cases

The personal name "天賜" of "天賜 ah/Thian-sù ah" (not an unknown word) which has been tagged as "A" with the suffix "ah" tagged as "T" or "Di" (which appeared twice in all; once, the selected Mandarin word was "啊" and in other instance it was "了").

The Taiwanese word "對/tùi" under general circumstances is synonymous with the Mandarin word "從"'from.' This word appeared ten times in the test data. The system selected the Mandarin word "對"'for' eight times and the word "從" twice for its counterpart. Nevertheless, under both circumstances, the POS tag of the word was always "P"; thus, the different word choice did not affect the accuracy of the POS tagging.

There were also 30 errors made that leave us unable to clearly explain the reasons. Table 7 lists some examples.

Table 7. Example of Some POS Tagging Errors

Left Context	Word and POS	Correct POS	Right Context	id
	lūn 'discuss' (Na)	VE	thȧk'read'(VC) pe̍h-ōe-jī 'vernacular writing'(Na) khah-iâⁿ'better than'(VJ) …	1
chò 'do'(VC) thâi-lâng 'kill someone'(VA)	hōan 'criminal' (VC)	Na	siū 'be subjected to'(P) sí-hêng 'death penalty'(Na) ê'of'(DE) 7(Neu) lâng 'people'(Na)	2
lâng'people'(Na)	ch̍it-ē 'once' (Nd)	D	chia̍h-lāu 'old'(VH)	3
tùi 'from' (P)	khí-thâu 'beginning'(VH)	Nv	chiū'then'(D) chin'very'(Dfa) tāng 'waver'(VAC)	4
Má-lī 'a person name'(Nb) ê'of'(DE) lāu-pē'father'(Na)	sí 'dead' (Dfb)	VH	ê'of'(DE) sî'time'(Na)	10
khòaⁿ 'look at' (VC)	khí-khí'up'(Nb)	VA	lo̍h-lo̍h'down'(VA) ê'of'(DE) hái-éng 'tide'(Na)	11
n̂g-hóaⁿ 'turned yellow'(VH) àm-tām'dim'(VH) ê'of'(DE) lō̤-teng'streetlamp'(Na)	chhiō 'shine'(D)	VC	lóng'always'(D) bē'not'(D) hn̄g'far'(VH)	12

4.6 Summary of Error Conditions

A summary of the causes of the errors made during the POS tagging and their frequency percentages is tabulated in Table 8.

Table 8. The Reason of POS Tagging Errors

Reason	Count	Percentage(%)	Remark
Selection of Inappropriate Mandarin Word	25	28.7	
Absence of Appropriate Mandarin Word	14	16.1	
Unknown Word	10	11.5	
Personal Name	4	4.6	
Propagation Error	4	4.6	Includes an unknown word
Totally	57	65.5	After discounting the repeat count

5. Discussion

5.1 Is Improvement Possible?

The ideal situation would be to resolve the foregoing errors and use this method to conduct the Taiwanese POS tagging to achieve an accuracy rate of 97.1%. Nevertheless, there is an apparent difficulty in the realization of this goal.

There are differences between the Taiwanese word order and the Mandarin word order; thus, the selection of the incorrect Mandarin word, and consequently incorrect POS tagging, occurred with high probability. The absence of appropriate Mandarin translation was the second leading cause of the POS tagging errors.

The unknown word problem was also a cause of POS tagging errors. From the Mandarin perspective, these words are not actually unknown words; this problem mostly resulted from the fact that translations between different languages are not one-to-one mappings. Another significant factor involves the use of hyphens in the POJ script, as their usage has not yet been standardized. It is probable that due to the use of Han characters, word boundaries are relatively vague in the different languages of the Chinese language family.

5.2 Hyphen Problems, Distinction between Taiwanese and Mandarin

In Taiwanese, some words take on the POJ script, thus, the use of the hyphen. Used one way, they separate the syllables of words, making it possible for a syllable to correspond to a Han character; used another way, they serve as word separators. Each syllable in a hyphenated word represents a unigram, and a space separates each word. Unfortunately, no original word

boundaries of Han character writing can be found to correspond to the hyphenated word.

In addition, Taiwanese has around 3,000 legal syllables, whereas Mandarin has around 1,200 legal syllables (Chan, 2008). Because of this, it may be said that the Taiwanese language has more single-syllable words. Nevertheless, as a single-syllable word may have several corresponding Han characters, the use of two-syllable or multi-syllable words resolves most of the problems.

For instance, if the Taiwanese word "這個"'this one' is written as "chit ê" (no hyphen used), the syllable "chit" may be made to correspond to several Mandarin words, such as "這" 'this,' "職"'job,' "質" 'quality,' "織," 'knit,' *etc*. The syllable "ê" may also be made to correspond to several Mandarin words, such as "的" 'of,' "個" '(a numerary adjunct),' "鞋" 'shoe,' *etc*. If the word is written as "chit-ê" (hyphenated), it definitely corresponds to "這個" in HR script. Hence, under the POJ script, the writer may tend to use a hyphen to link a single-syllable word to another single-syllable word if these two single-syllable words may likely form one composite word or one phrase. Present practices show that the word "這個" may appear hyphenated or in a separated syllable form, thus creating inconsistencies.

As the use of hyphenated words creates the problem of one Taiwanese word corresponding to two Mandarin words, if the original text is not revised and the Mandarin corresponding word is manifested as an unknown word, it may be possible to just remove the hyphen and try again. This method may reduce the chance of POS tagging errors due to the unknown word factor.

5.3 The Distinction between Different Eras or Different Genres

We investigated whether texts of a different era or a different literary genre would affect the accuracy rate of the POS tagging. Table 10 shows the POS tagging accuracy rates for texts of three types of literary genres and Table 11 shows the POS tagging accuracy rates for texts of literary works belonging to three different periods or eras. Table 9 shows that the POS tagging accuracy rate for novel materials is comparably lower than other genres; whereas Table 10 indicates that the POS tagging accuracy rate for the materials written in the Post-war era are comparably lower than the other periods investigated. Basically, there are no significant differences among three genres or three eras as a whole.

Table 9. Tagging Accuracy Rate for Different Genres

Genre	No. of Words	No. of Tagging Errors	Accuracy Rate (%)
Prose	549	43	92.2
Novel	372	36	90.3
Drama	117	8	93.2

Table 10. Tagging Accuracy Rate for Different Eras

Era	No. of Words	No. of Tagging Errors	Accuracy Rate (%)
Ching Dynasty	232	18	92.2
Japanese-ruled	359	27	92.5
Post-war	447	42	90.6

After deliberation, we found that the individual writing style of authors is actually the dominant factor of the POS tagging accuracy. From Table 2, the individual POS tagging accuracy varies from 83.6% to 98.0%.

6. Conclusion and Future Works

We proposed a Taiwanese POS tagging method using a statistical method and Mandarin training data, and we achieved an accuracy rate of 91.6%. Due to the lack of Taiwanese training data, we sought the help of Mandarin.

This strategy could also be applied to other languages that lack resources. We think that this is a very important idea. It is preferable to select an intermediate language close to the target language from the viewpoint of the language family.

We also developed an online Taiwanese word segmentation and POS tagging system for people who are interested in this topic. Users can input Taiwanese text and get the POS tagging results. It is somewhat difficult for a user to prepare both POJ and HR mixed scripts; therefore, we also provide the functions in the absence of one of these two scripts (Iunn, *et. al*., 2007). This, however, will decrease the accuracy rate.

If we can construct a Taiwanese-Mandarin parallel corpus, we can use other methods like the Coerced Markov Models proposed by Fung and Wu (1995) to accomplish the Taiwanese POS tagging task.

We hope that we can proceed to the construction of Taiwanese Treebank.

Acknowledgments

This research was supported in part by National Science Council of Taiwan, under the contract number NSC 95-2221-E-122 -006. Thanks to the National Museum of Taiwanese Literature of Taiwan for providing plentiful written Taiwanese literature data. We also thank the anonymous reviewers for their constructive opinions.

Reference

Berger, A. L., Pietra, S. A. D., & Pietra, V. J. D. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1), 39-71.

Chan, K. I. (2008). *Comparison with the Usage of Academic and Non-academic Taiwanese Words*. Master thesis, National Taitung University.

Chou, S. Y. (2006). *T3 Taiwanese Treebank and Brill Part-of-Speech Tagger*. Master thesis, Hsin-chu: National Tsing Hua University.

CKIP. (1993). *Analysis of Chinese Part-of-speech.* The Association for Computational Linguistics and Chinese Language.

CKIP. (2004). *Chinese Word Segmentation and Tagging System*, (Retrieved 2009/4/10) http://ckipsvr.iis.sinica.edu.tw/.

Embree, B. L. M. (1984). *A dictionary of Southern Min* '台英辭典,' Taipei: Taipei Language Institute.

Fung, P., & Wu, D. K. (1995). Coerced Markov Models for cross-lingual lexical tag relations. in the *6th International Conference on Theoretical and Methodological Issues in Machine Translation*, 1, 1995, Leuven, Belgium, 240-255.

Gordon, R. G. Jr. ed. (2005). *Ethnologue: Languages of the world* (15th ed.). Dallas:SIL International.

Huang, S. F.(1995). *Language, Society and Ethnicity* (2nd ed.). Taipei: Crane.

Iunn, U. G. (2000). *Online Taiwanese-Mandarin Dictionary*. (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/q/q.asp.

Iunn, U. G. (2003a). *Online Taiwanese Syllable Dictionary.* (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/TG/jitian/.

Iunn, U. G. (2003b). Survey of the Online Taiwanese-Mandarin Dictionary-- Discussion of Building Technique and its Utilization. in the *Proceedings of 3rd International Conference on Internet Chinese Education*, 2003b, Overseas Chinese Affairs Commission, 132-141.

Iunn, U. G. (2003c). *Online Taiwanese Concordancer System.* (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/TG/concordance/.

Iunn, U. G. (2005). *Taiwanese Corpus Collection and Corpus Based Syllable / Word Frequency Counts for Written Taiwanese*. (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/guliau-supin.asp.

Iunn, U. G. (2007). New Manifestation of the Taiwanese vernacular literature -- Introduction to Digital Archive for Written Taiwanese. *National Museum of Taiwanese Literature Communication*, 15, 42-44.

Iunn, U. G. (2009). *Processing Techniques for Written Taiwanese-- Tone Sandhi and POS Tagging*. PhD thesis, National Taiwan University.

Iunn, U. G., & Lau, K. G. (2007). Introduction to online Taiwanese Dictionaries and Corpora. *Language, Society and Culture Series 2: Multiculturalism Thinking of the Language Policy*, 2007, Institute of Linguistics of Academia Sinica, 311-328.

Iunn, U. G., Lau, K. G., Tan-Tenn, H. G., Lee, S. A., & Kao, C. Y. (2007). Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods. *International*

Journal of Computational Linguistics and Chinese Language Processing, 12(4), 349-370.

Iunn, U. G., Lau, K. G., & Tai, C. H. (2007). *Online Taiwanese Word Segmentation and POS Tagging System*, (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/TGB/tagging/tagging.asp.

Lau, K. G. (2007). *Finding Mandarin Candidate Words by POJ script and Han-Romanization mixed script word pair*, (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/nmtl/dadwt/pos_tagging/clhl_hoagi_hausoansu.asp.

Le, Z. (2003). Maximum Entropy Modeling Toolkit for Python and C++. (Retrieved 2009/7/10) http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, Cambridge: MIT Press.

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. in the *Proceedings of 17th International Conference on Machine Learning*, Stanford University, 591-598.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. in the *Proceedings of the IEEE*, 77, 257-286.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania, 133-142.

Samuelsson, C. (2003). Statistical methods. in the *Oxford Handbook of Computational Linguistics*, Oxford University Press, 358-375.

Shi, D. M. (2006). *T3 Taiwanese Treebank and Brill Parser*, Master thesis, Hsin-chu: National Tsing Hua University.

Tai, C. H. (2007). *Word and POS tagging selection for Taiwanese Language*, (Retrieved 2009/4/10) http://140.109.19.105/.

Tsai, Y. F., & Chen, K. J. (2004). Reliable and Cost-Effective Pos-Tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(1), 2004, 83-96.

Appendix

Test Data List.

id	Year	Genre	Author	Article title	No. of Syllables
1	1885	prose	Reverend Ia̍p ‘葉牧師’	Pe̍h-ōe-jī ê lī-ek ‘The Benefits of Using Pe̍h-ōe-jī, 白話字的利益’	162
2	1893	prose	Reverend Kam ‘甘牧師’	Chhi^{n}-mî o̍h ‘Blind Study, 青瞑學’	66
3	1919	prose	H S K	Phín-hēng ê ûi-thôan ‘Inheritance of Morality, 品行的遺傳’	180
4	1935	prose	Ong Chong-têng ‘汪宗程’	Chín-chai kì ‘Earthquake Disaster Record, 震災記’	122
5	1954	prose	Ô· Bûn-tî ‘胡文池’	Tōa-soa^{n} chhiù^{n}-koa ‘A High Mountains sing, 大山唱歌’	74
6	1990	prose	Tân Gī-jîn ‘陳義仁’	Lāu-lâng ê kè-ta̍t ‘The Value of The Elderly People, 老人的價值’	75
7	2000	prose	Tân Bêng-jîn ‘陳明仁’	Sûn-chêng Ông Pó-chhoan ‘Pure Love Ông Pó-chhoan, 純情王寶釧’	112
8	1890	novel	Unknown	An-lo̍k-ke ‘Safety and Happiness Street, 安樂街’	101
9	1924	novel	Lōa Jîn-seng ‘賴仁聲’	Án-niá ê Ba̍k-sái ‘Mother's Tears, 母親的眼淚’	133
10	1955	novel	N̂g Hôai-un ‘黃懷恩’	Chháu-tui téng ê bîn-bāng ‘Dreams on the Grass Stack, 草堆上的夢’	116
11	1990	novel	Iû^{n} Ún-giân ‘楊允言’ translated	Hái-phī^{n} Sin-niû ‘Bride on The Cape, 岬角上的新娘’	94
12	2006	novel	Lâu Sêng-hiân ‘劉承賢’	Chia̍h-chōe ‘Plead Guilty, 伏罪’	92
13	1924	drama	Lîm Bō·-seng ‘林茂生’	Hì-chhut: Lō·-tek kái kàu ‘ Drama: Ruth Reformed Church, 戲齣:路得改教’	77
14	1950	drama	Tân Chheng-tiong ‘陳清忠’ translated	Venice ê Seng-lí-lâng ‘Venice Businessman, 威尼斯的生意人’	92
Note: the original author of id 11 is Sòng Te̍k-lâi ‘宋澤萊,’ id 14 is Shakespeare					