Chapter 5 POS Tagging Method

This chapter will introduce a tagging method for the Taiwanese language. We propose a POS tagging method, and use the more than 62,000 entries of the OTMD and 10 million words of Mandarin training data to tag Taiwanese. The literary written Taiwanese corpora have both POJ script and HR mixed script, and includes prose, novels, and dramas. We followed 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 OTMD to find corresponding Mandarin candidate words, selects the most suitable Mandarin word using an HMM probabilistic model from the Mandarin training data, and finally tags the word using an MEMM classifier.

5.1 Problems of POS Tagging

The primary difficulty encountered in the POS tagging of the Taiwanese corpora is the question, “What is the Taiwanese POS tagset?” To date, no standard tagging system has been established. Under the circumstances, we temporarily employed the Chinese POS tagset established by the CKIP Group of Academia Sinica (CKIP, 1993). Unfortunately, we still encountered some problems since 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.

Under the circumstances, 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.

The Mandarin language model is a ten-million word balanced corpus with about 100,000 word types and 46 tags maintained by the CKIP Group.

The statistical procedures included a Hidden Markov Model (*abbrev*. HMM) for word selection and a Maximal Entropy Markov Model (*abbrev*. MEMM) for tag selection. We assumed that the appearance of a word was only influenced by the previous word, and the HMM worked with this assumption. However, the appearance of a POS tag is not only influenced by the previous tag or word but also by other information. We selected the MEMM to predict the POS tag because the MEMM included more information that was of assistance in determining the POS tags.

5.2 POS Tagging Methods

Fig 5 - 1 shows a system architecture diagram:

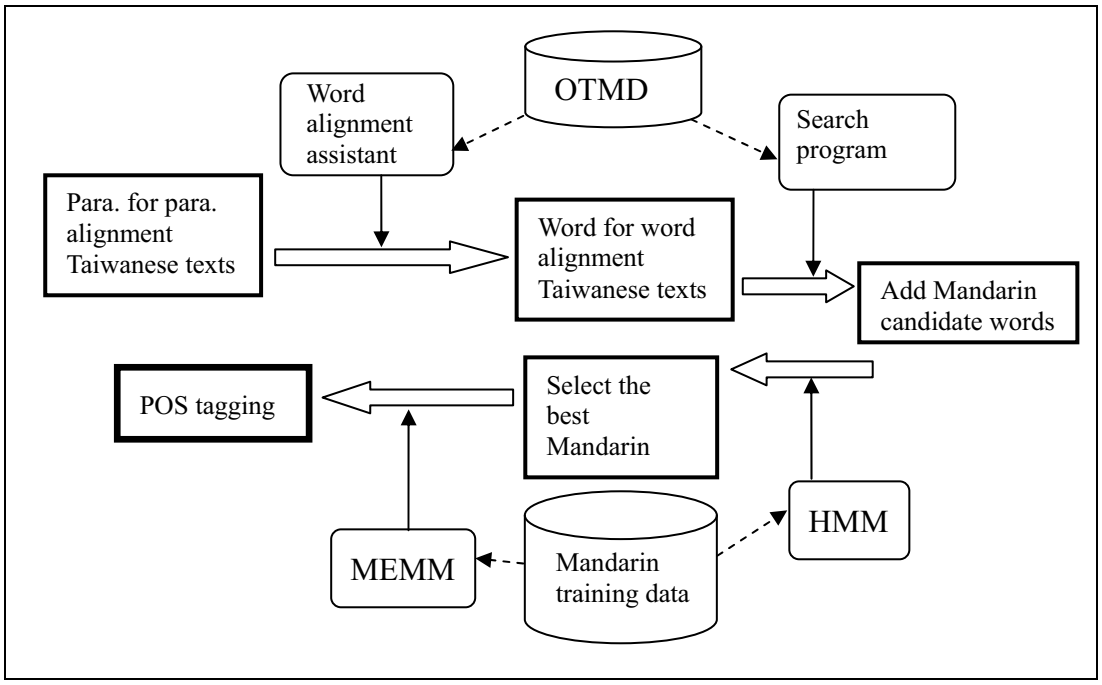

Fig 5 - 1 Taiwanese Language POS Tagging System Architecture Diagram

5.2.1 Origin of the Corpus

The corpus we chose was the result of the DADWT project of the NMTL. It contains over 2.58 million syllables in both POJ and HR mixed scripts with paragraph by paragraph alignment, including novels, prose, dramas and poems (Iunn, 2007b).

5.2.2 Word for Word Alignment

First, we developed a word alignment program to aid with manual processing. We arranged for the word alignment of two scripts, where the paragraphs were already aligned. This program not only collates the number of syllables in the two scripts, it also compares and contrasts the two scripts with the contents of the OTMD. If the program does not find the two words within the same entry, it highlights the corresponding words to remind the user that the

word may be an unknown word, inconsistent usage of the Han character, or typo.

The OTMD contains over 62,000 entries, including the POJ script, Taiwanese HR mixed script, Mandarin translation, and English translation (about 10,000 records). This online dictionary contains synthesized Taiwanese pronunciations, and receives an average of around 2,700 daily inquiries (counted from 1/1/2008 to 12/31/2008). The English field was added in year 2007; however, the data still needs to be completed (Iunn, 2000).

5.2.3 Searching for 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, “ài[愛]” in Taiwanese has the meanings of “愛 ‘love (person),’ ” “喜歡 ‘like (thing),’ ” “要 ‘want to,’ ” “需要 ‘need to,’ ” *etc*., in Mandarin. However, 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 word pair that appears as “較贏 [khah-iâⁿ]” 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 the HR mixed script were all Han characters, then we regarded the Han characters as a Mandarin candidate word (assuming that the word is the Taiwanese and Mandarin common words).

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

If the strategy was still unable to find any result, 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] ” 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).

5.2.4 Selecting the Best Mandarin Translation

We employed the Hidden Markov Model and Viterbi algorithm and 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.

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 $P(\hat{S} = w_1 w_2 \cdots w_m) \cong \prod_{i=1}^{m} P(w_i \mid w_{i-1})$. Therefore, it searches for the word sequence $\hat{S} = w_1 w_2 \cdots w_m$, which maximizes $\sum_{i=1}^{m} \log P(w_i \mid w_{i-1})$. In a case where $P(w_i \mid w_{i-1}) = 0$, no bigram of w_i, w_{i-1} could be found in the training data, and we backed off $P(w_i \mid w_{i-1})$ to $P(w_i)$. It should be noted that the word string $\hat{S}$ may not be a legal Mandarin sentence (Samuelsson, 2003).

In practice, we use the Viterbi algorithm to eliminate repeated computations 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).

5.2.5 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 & 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). $p(h,t)=\pi\mu\prod_{j=1}^{k}\alpha_j^{f_j(h,t)}$; 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: $L(p)=\prod_{i=1}^{n}p(h_i,t_i)=\prod_{i=1}^{n}\pi\mu\prod_{j=1}^{k}\alpha_j^{f_j(h_i,t_i)}$.

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

(a) Words: there are 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: there are two types of feature patterns: $t_{i-1}, t_{i-2}t_{i-1}$;

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

The feature patterns m_1, m_2, m_n are designated to manipulate unknown words. If w_i is an unknown word, we then 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.

The ten-million word pos tagged balanced Sinica corpus of the CKIP Group was used as the training data and the search for the most probable pos sequence was implemented using the Viterbi Algorithm to search for pos sequence *t* such that $L(p) = \prod_{i=1}^{n} p(h_i, t_i) = \prod_{i=1}^{n} \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h_i, t_i)}$ was maximized (Berger, Pietra, & Pietra, 1996; McCallum, Freitag, & Pereira, 2000; Rabiner, 1989; Ratnaparkhi, 1996; Samuelsson, 2003; Tai, 2007; Y.-f. Tsai & Chen, 2004).

5.3 Results

We used the foregoing method to perform the Taiwanese POS tagging task; however, as no standard answers were available to gauge the accuracy, we extracted partial results and checked them manually. The primary consideration of the manual checking procedure was the CWSTS of the CKIP group of Academia Sinica (CKIP, 2004). We selected seven 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 forms of prose (three), drama (one), and novels (three). We selected the first paragraph from each composition, or if the number of syllables in the first paragraph was less than 50, we selected

the second paragraph.

Table 5 - 1 and Table 5 - 2 show the test data selected for manual checking. The number of syllables, words, and incorrectly selected Mandarin words, as well as the POS tagging inaccuracy for each paragraph are noted.

Table 5 - 1 Test Data List

id	Year	Genre	Author	Article Title	No. of Syllables
1	1885	prose	Reverend Iap̍ ‘葉牧師’	Pe̍h-ōe-jī ê lī-ek ‘The Benefits of Using Pe̍h-ōe-jī, 白話字的利益’	162
2	1919	prose	H S K	Phín-hēng ê ûi-thôan ‘Inheritance of Morality, 品行的遺傳’	180
3	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
4	1950	drama	Tân Chheng-tiong ‘陳清忠’ translated	Venice ê Seng-lí-lâng ‘Venice Businessman, 威尼斯的生意人’	92
5	1890	novel	Unknown	An-lo̍k-ke ‘Safety and Happiness Street, 安樂街’	101
6	1924	novel	Lōa Jîn-seng ‘賴仁聲’	Án-niá ê Ba̍k-sái ‘Mother's Tears, 母親的眼淚’	133
7	1990	novel	Iûⁿ Ún-giân ‘楊允言’ translated	Hái-phīⁿ Sin-niû ‘Bride on The Cape, 岬角上的新娘’	94
Note: the original author of id 4 is Shakespeare, id 7 is Sòng Te̍k-lâi ‘宋澤萊’					

Table 5 - 2 Tagging Accuracy Rate for the Test Data

id	No. of Syllables	No. of Words	Errors	Tagging Errors	Accuracy Rate(%)
1	162	109	9	6	94.5
2	180	119	6	8	93.3
3	75	49	7	7	85.7
4	92	58	3	4	93.1
5	101	77	7	9	88.3
6	133	93	7	9	90.3
7	94	59	7	5	91.5
Totally	837	564	46	48	91.5

$$accuracy\ rate = (1 - \frac{\#\ tagging\ errors}{\#\ total\ words})\ 100\%$$

A total of 564 words (837 syllables) were selected, and manual checking showed that 46 words had been incorrectly selected and 48 words were found to have the wrong POS tagging, thus placing the average POS tagging accuracy rate at 91.49%. 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 Mandarin corresponding word did not always have accurate POS tagging.

Furthermore, sometimes one Taiwanese word would correspond to two Mandarin words. For instance, if the two syllables of the Taiwanese word “壁頂” were treated as only one word, the result would be “牆壁上” in Mandarin. There were also occasions wherein two Taiwanese words corresponded to only one Mandarin word counterpart. For instance, the Mandarin counterpart of the Taiwanese words “中國” and “字” was “中國字.” The former was 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%.

Examples of actual POS tagging results are shown below. These are a part of id 7. 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 in the 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 separated 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 5 - 3 Example of POS Tagging Result

Taiwanese	Mandarin Candidate Words	Selected Word	POS tagging
我[góa]	我	我	Nh
將[chiong]	將	將	D
草帽仔[chháu-bō-á]	@草帽仔	草帽仔	Na
掛[kòa]	帶;掛;戴	帶	VC
tī [tī]	在	在	P
壁頂[piah-téng]	牆壁上	牆壁上	Nc
，[,]	，	，	COMMACATEGORY
行李[hêng-lí]	行李	行李	Na
khêng[khêng]	收拾;盤點	收拾	VC

Table 5 - 3 Example of POS Tagging Result

Taiwanese	Mandarin Candidate Words	Selected Word	POS tagging
khêng[khêng]	收拾;盤點	收拾	VC
leh[leh]	咧	咧	T
，[,]	，	，	COMMACATEGORY
坐[chē]	坐	坐	VA
tòa[tòa]	住	**住	**VCL**(P)**
小店[sió-tiàm]	@小店	小店	Na
ê[ê]	的	的	DE
tha-thá-mì[tha-tha-mì]	塌塌米	塌塌米	Na
頂 kôan[téng-kôan]	上面	上面	Ncd
，[,]	，	，	COMMACATEGORY
看[khòaⁿ]	看	看	VC
窗外[thang-gōa]	@窗外	窗外	Nc
ê[ê]	的	的	DE
光景[kong-kéng]	風光;氣象;光景;風景;景氣	**景氣 **風景**	Na
，[,]	，	，	COMMACATEGORY
看[khòaⁿ]	看	看	VC
起起[khí-khí]	@起起	起起	**Nb**(VA)**
落落[lo̍h-lo̍h]	@落落	落落	VA
ê[ê]	的	的	DE
海湧[hái-éng]	海浪;海潮	海浪	Na
，[,]	，	，	COMMACATEGORY
因爲[in-ūi]	由於;因爲	因爲	Cbb
等待[tán-thāi]	留待;等待	等待	VK
朋友[pêng-iú]	友人;朋友	朋友	Na
，[,]	，	，	COMMACATEGORY
心適[sim-sek]	好玩;好玩兒;有趣;風趣;愉快;稀奇;鬧著玩	有趣	VH
心適[sim-sek]	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]	輕輕的	輕輕的	**Nb**(D)**

Table 5 - 3 Example of POS Tagging Result

Taiwanese	Mandarin Candidate Words	Selected Word	POS tagging
來[lâi]	來	來	D
點[tiám]	燃點;檢點;點;點子	點	VC
一支[chit-ki]	@一支	一支	Na
涼涼[liâng-liâng]	冷冷;涼絲絲	**冷冷 **涼涼**	VH
ê[ê]	的	的	DE
芎蕉[kin-chio]	香蕉	香蕉	Na
薰[hun]	香菸;香煙;薰	香煙	Na
。[.]	。	。	PERIODCATEGORY

5.4 Error Analysis

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

5.4.1 Incorrect Corresponding Mandarin Word Selection

An analysis of the errors made in the selection of Mandarin words or POS tags revealed that in thirteen cases the selection of inappropriate Mandarin words led to POS tagging errors. Table 5 - 4 shows the incorrect Mandarin words selected and the respective parts of speech of their tagged parts.

Table 5 - 4 The Incorrect Mandarin Words Selected and Their Respective POS

Word	Selected Mandarin Word and POS	More Appropriate Mandarin Word and POS	Remark
押/ah	強制(D)	押(VC) ‘take into custody’	
無/bô	不(D)	沒有(VJ) ‘not have’	2 times
這號/chit-hō	這樣(VH)	這種(N?) ‘this kind of’	2 times
轉/chōan	上(Ncd)	轉(Vac) ‘turn’	2 times
夭壽/iáu-siū	非常(Dfa)	早夭(VH) ‘dead early’	
價值/kè-ta̍t	值得(VH)	價值(Na) ‘value’	
活/o̍ah	生活(Na)	活(VH) ‘live’	
破相/phòa-siù^{n}	破(VHC)	殘廢(Na) ‘disabled’	
相借問 /sio-chioh-mn̄g	招呼(VC)	打招呼(VB) ‘say hello’	
著/tio̍h	就(P)	得(D) ‘need to’	

5.4.2 Absence of Appropriate Mandarin Words in the OTMD

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

Table 5 - 5 Errors Caused by the Absence of Appropriate Mandarin Words in the OTMD.

Taiwanese	Selected Mandarin by System	Appropriate Mandarin Word
tiā^{n}-tiā^{n} / tiā^{n}-tiā^{n}	常常(D) ‘often’	而已(T) ‘just’
轉 / tn̂g	調解(VC) ‘mediate’	轉(VAC) ‘turn’

5.4.3 Unknown Words from the Viewpoint of Mandarin

In the eight remaining errors, the POS tagging errors were made because the word was an unknown word. The majority of these unknown words correspond to two Mandarin words. These eight unknown words are shown in Table 5 - 6.

Table 5 - 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
食老/chia̍h-lāu	*食老 'old'	Na	V
轉了/chōan-liáu	*轉了	VH	V
法律上/hoat-lu̍t-siōng	法律上 'jural'	VC	N
非爲/hui-ûi	非爲 'infamous conduct'	A	N
窮志/kiông-chì	窮志 'exhaust the ambition'	Na	V
輕輕仔/khin-khin-á	輕輕地 'lightly'	Nb	D
生子/se^{n}-kiá^{n}	生子 'give birth to a child'	Na	V

5.4.4 Propagation Error

Four 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.

5.4.5 Other Cases

The personal name "天賜" of "天賜 ah/Thian-sù ah" (not an unknown word)

was tagged as “A,” with the suffix “ah” tagged as “T” or “Di” (this appeared twice in all; once, the selected Mandarin word was “啊” and in the other instance it was “了”).

The Taiwanese word “對/tùi” under general circumstances is synonymous with the Mandarin word “從.” This word appeared nine times in the test data. The system selected the Mandarin word “對” seven times and the word “從” twice for its counterpart. However, 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 18 other errors made, mainly due to our inability to clearly analyze the proper POS tags for the words at the time.

5.4.6 Summary of Error Conditions

A summary of the causes of the errors made during the POS tagging and their frequency percentages is given in Table 5 - 7.

Table 5 - 7 The Reasons for the POS Tagging Errors

Reason	Count	Percentage(%)	Remark
Selection of inappropriate Mandarin word	13	27.1	
Absence of appropriate Mandarin word	2	4.2	
Unknown word	8	16.7	
Personal name	4	8.3	
Propagation error	4	8.3	Includes an unknown word
Totally	30	62.5	After discounting the repeat count

5.5 Discussion

5.5.1 Is Improvement Possible ?

The most ideal situation would be to resolve the foregoing errors and then use this method to conduct the Taiwanese POS tagging to achieve an accuracy rate of 96.8%. However, 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 an incorrect Mandarin word, and consequently incorrect POS tagging, occurred with high probability. Although it is possible to add new entries to the OTMD to resolve the problem of unavailable appropriate Mandarin word choices, the accuracy rate could only be raised by about 5%.

The unknown word problem was the second leading 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.

5.5.2 Hyphen Problems, Distinction between Taiwanese and Mandarin

In Taiwanese, some words take on the POJ script and, thus, the use of the hyphen. On one hand, they are used to separate the syllables of words, making it possible for a syllable to correspond to a Han character; on the other hand, 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 (K.-i. Chan, 2008). Because of this, it may be said that the Taiwanese language has more single-syllable words. However, 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” is written as “chit ê” (no hyphen used), the syllable “chit” may be made to correspond to several Mandarin words, such as “這,” “職,” “質,” “織,” *etc*. The syllable “ê” may also be made to correspond to several Mandarin words, such as “的,” “個,” “鞋,” *etc*. If the word is written as “chit-ê” (hyphenated), it is usually directly read as “這個.” 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 the presence of inconsistencies.

Since 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.

5.5.3 The Distinction between Different Eras or Different Genres

For questions about whether texts of a different era or a different literary genre would affect the accuracy rate of the POS tagging, please refer to the data shown in the following tables. Table 5 - 8 shows the POS tagging accuracy rates for the texts of three types of literary genres and Table 5 - 9 shows the POS tagging accuracy rates for the texts of literary works belonging to three different periods or eras. Table 5 - 8 shows that the POS tagging accuracy rate for novel materials are comparably lower; whereas Table 5 - 9 indicates that no significant difference may be noted in the POS tagging accuracy rates for the literary works of different periods. However, due to the limited amount of data available, further empirical studies are necessary to attest to the foregoing analysis findings.

Table 5 - 8 Tagging Accuracy Rates for Different Genres

Genre	No. of Words	No. of Tagging errors	Accuracy rate(%)
prose	277	21	92.4
drama	58	4	93.1
novel	229	23	90.0

Table 5 - 9 Tagging Accuracy Rates for Different Eras

Era	No. of Words	No. of Tagging errors	Accuracy rate (%)
Ching Dynasty	186	15	91.9
Japanese-ruled	212	17	92.0
Post-war	166	16	90.4

5.6 Summary

We proposed a Taiwanese POS tagging method using a statistical method and Mandarin training data, and achieved an accuracy rate of 91.5%. 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 thought that this was 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 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 (Lau & Iunn, 2007). However, that will

decrease the accuracy rate.

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

A more suitable tagset for Taiwanese and an electronic word dictionary based on the Taiwanese word segmentation standard are necessary for advanced searches. We hope that we can proceed to the construction of the Taiwanese Treebank.