

利用統計方法及中文訓練資料處理台語文詞性標記

Modeling Taiwanese POS tagging with statistical methods and Mandarin training data

楊允言¹ 戴嘉宏² 劉杰岳³ 陳克健² 高成炎¹

¹ 國立台灣大學資訊工程系

{d93001,cykao}@csie.ntu.edu.tw

² 中央研究院資訊科學研究所

{glaxy,kchen}@iis.sinica.edu.tw

³ 獨立研究學者

kiatgak@gmail.com

摘要

本文提出利用有六萬多詞條的台華辭典以及千萬詞的中文訓練資料來做台語文詞性標記的方法。台語文語料為包括全羅馬字及漢羅合用兩種書寫文本的文學資料，文類涵蓋散文、小說、劇本等，詞類集採用中央研究院詞庫小組所訂定的中文詞類集。

我們開發語詞對齊檢查程式，將兩種文本的語料逐詞對齊，透過台華辭典查詢每個語詞相對應的中文候選詞，接著利用中文訓練資料，以 HMM 機率模型挑選出最適當的中文對譯詞，再以 MEMM 分類器標記詞性。

實驗結果顯示，以此方法做台語文詞性標記，我們得到 91.49% 的正確率，並針對標記錯誤分析其原因。以此基礎，我們也得到了初步的台語文訓練語料。

Abstract

In this paper, we propose a POS tagging method using more than 60 thousand entries of Taiwanese-Mandarin translation dictionary and 10 million words of Mandarin training data to tag Taiwanese. The literary written Taiwanese corpora have both Romanization script and Han-Romanization mixed script, the genre includes prose, fiction and drama. We follow tagset drawn up by CKIP.

We develop word alignment checker to help the two scripts word alignment work, and then lookup Taiwanese-Mandarin translation dictionary to find the corresponding Mandarin

candidate words, select the most suitable Mandarin word using HMM probabilistic model from the Mandarin training data, and finally tag the word using MEMM classifier.

We achieve an accuracy rate of 91.49% on Taiwanese POS tagging work, and analysis the errors. We also get the preliminary Taiwanese training data.

關鍵詞：詞性標記，台語文，中文

Keywords: POS tagging, written Taiwanese, Mandarin

一、前言

雖然一直沒有受到足夠的重視，閩南語在全世界語言人口數有四千六百多萬，是排名第 21 位的語言，主要分佈在台灣、新加坡、馬來西亞、汶萊、中國、泰國、菲律賓及印尼等地[1]。台灣閩南語的語言使用人口，在台灣約佔 70%以上，是最主要的台灣本土語言[2]。而這個語言的名稱，至今也還是很分歧，至少有 17 種稱呼，包括台語、福建話、閩南話、福佬話、…等等[3]。本文將使用一般大眾對其的稱呼：「台語」，不打算涉入名稱的討論。

台語有漢字及羅馬字書寫兩種文字傳統，漢字書寫可追溯自 16 世紀，包括南管戲文或是天主教書籍等[4]；羅馬字書寫則可追溯自 1832 年起，包括辭典、宗教作品、報紙、啓蒙讀物、教科書、…等[5]。兩種文字各有其優缺點，漢字書寫的，較難確認其實際發音，又有訓讀字、本字、借音字、本土字等不同類別的字，專家考證的本字又各有不同[6]。但是在台灣，因為中文教育的普及，大家看到漢字書寫的台語文普遍比較不會排斥。羅馬字標注出實際的發音，語詞間以空格隔開，語詞內以連字符（hyphen）隔開每個音節，以資訊處理的角度來看，可能好用多了。

爲了建立台語計算語言學的基礎，過去幾年，我們陸續建立了台語華語對譯辭典[7]、台語文未加工語料庫等資源，以未加工語料庫爲基礎建立台語文語詞檢索系統[8]、以規則方法處理台語的變調處理問題[9]等。我們希望進一步將語料庫做更完整的標注。對於語料庫的標注，最基本且重要的，應當是詞性標記。

目前我們要做台語文語料的詞性標注，首先馬上面臨一個難題：台語的詞類集爲何？至今並沒有一套標準。在此情形下，我們暫時採用中央研究院詞庫小組所訂定的中文詞類

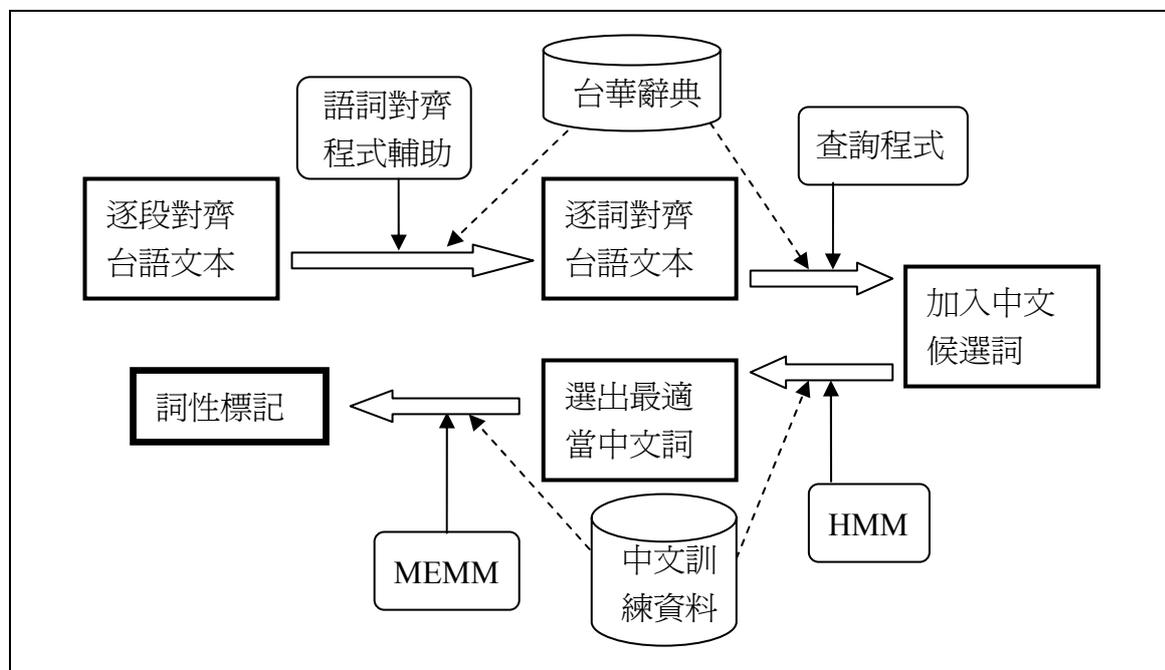
集[10]。這樣仍然會有問題，因為我們並沒有標記中文詞類集的台語辭典，現有的台語辭典，只有基本詞類如名詞、動詞、形容詞等的訊息。

另外一個問題是人力的缺乏，我們並沒有充足的人力來進行台語文語料的詞性標記。

在此情形下，本文提出利用現有的中文資源，以及台語華語對譯辭典，加上統計方法來進行台語詞性自動標記工作。

二、實驗方法

圖一顯示系統架構圖。



圖一、台語詞性標記系統架構圖

(一) 語料來源

我們所使用的語料為國家台灣文學館「台語文數位典藏資料庫（第二階段）」的計畫成果，有 258 萬多音節全羅馬字（全羅）及漢字羅馬字合用（漢羅）段落對齊的兩種文本，包括小說、散文、劇本、新詩等文類[11]。

（二）逐語詞對齊

首先，我們開發語詞對齊程式輔助人工，逐步將段落對齊的兩種文本逐語詞對齊。這支程式除了核對兩種文本的音節數之外，還將羅馬字、漢羅合用的語詞與台華辭典中的內容做比對，如果這兩個語詞沒有出現在同一詞條內，程式將會標明出來以提醒使用者，有可能是未知詞，也可能是漢字用字不一致，或是此語詞打字錯誤等原因。

台華辭典的原始資料由鄭良偉提供，楊允言在 2000 年提供線上查詢系統，詞條內容經過多次增補，目前有六萬多詞條，包括台語羅馬字、台語漢羅、華語對譯詞、英文等欄位，台語部分並附上發音功能，平均每天有 2400 多次的查詢（過去一年平均）。其中英文欄位於 2007 年新增，資料尚不完整[12]。

（三）尋找對應中文候選詞

接著，我們繼續利用台華辭典，將全羅/漢羅配對的語詞，找出其對應的中文候選詞。這是一對多的對應，亦即一個台語全羅/漢羅配對的語詞，可能有一個以上的華語對譯詞。除此外，有些語詞因為台華辭典沒有收錄而查不到，有些語詞則因漢羅的寫法不同而查不到（例如文本中出現「較贏[khah-iâⁿ]」，而辭典裡為「khah 贏[khah-iâⁿ]」）。對於此問題，我們的解決方法是：如果全羅/漢羅配對語詞查不到，暫時把漢羅拿掉，用全羅找出對應的中文候選詞，若漢羅的寫法是全漢字，也將漢羅視為中文候選詞之一（假設其為台華共通詞）；這麼做有可能讓中文候選詞的詞數增加，尤其是單音節詞（例如文本中出現「轉[chōan]」，辭典找不到此詞條，但是羅馬字為「chōan」的有兩個詞條，其中文對譯詞分別為「扭」和「上」，再加上「轉」，詞義皆不同）。

如果還是查不到，則把漢羅語詞直接當做中文候選詞（例如文本中出現「有形[iú-hêng]」，辭典中沒有這個詞條，用羅馬字「iú-hêng」查，也查不到，就直接把「有形」視為中文候選詞）[13]。

（四）挑選最適當的對應中文詞

我們採用 Markov bigram model，利用中央研究院詞庫小組千萬詞平衡語料庫的 bigram

語詞訓練資料，從中文候選詞中挑選最適當的中文詞。

假設某一句子有 m 個語詞，第一個語詞 w_1 是從 $w_{11}, w_{12}, \dots, w_{1n_1}$ 候選詞中挑出來的中文詞，第二個語詞 w_2 是從 $w_{21}, w_{22}, \dots, w_{2n_2}$ 候選詞中挑出來的中文詞，第 m 個語詞 w_m 從 $w_{m1}, w_{m2}, \dots, w_{mn_m}$ 候選詞中挑出來的中文詞。我們要從中挑選出去可能的串列 $\hat{S} = w_1 w_2 \dots w_m$ ，使得 $\prod_{i=1}^m P(w_i | w_{i-1})$ 為最大，亦即使 $\sum_{i=1}^m \log P(w_i | w_{i-1})$ 為最大。要說明的是，這個串列 \hat{S} ，可能不是合法的中文句子[14]。

$$w_i = \begin{cases} \arg \max \log_e P(w_{ij}) & i=1 \text{ or } \forall j \neg \exists w_{i-1} w_{ij} \\ \arg \max \log_e P(w_{ij} | w_{i-1}) & \text{otherwise} \end{cases}$$

在訓練資料中，若雙連詞不存在，則取最高頻的單一語詞（unigram）。

（五）根據中文詞挑選最適當的詞性

我們採用 Maximal Entropy Markov Model (MEMM) 來挑選詞性。

MEMM 包括一組包含語詞和詞性的歷程集合 H ，和詞性集合 T ， $p(h, t) = \pi \mu \prod_{j=1}^k \alpha_j^{f_j(h, t)}$ ，其中， $h \in H, t \in T$ ， π 是常數， $\{\mu, \alpha_1, \dots, \alpha_k\}$ 是大於 0 的參數， $\{f_1, \dots, f_k\}$ 是特徵 features， $f_j(h, t) \in \{0, 1\}$ ，參數 α_j 對應特徵 f_j 。

對於目標語詞 w_i 的詞性 t_i ，我們選取 10 個特徵，包括：

1. 語詞：有 $w_i, w_{i-1}, w_{i-2} w_{i-1}, w_{i+1}, w_{i+1} w_{i+2}$ 五個特徵；
2. 詞性：有 $t_{i-1}, t_{i-2} t_{i-1}$ 兩個特徵；
3. 構詞： m_1, m_2, m_n 三個特徵

m_1, m_2, m_n 是針對未知詞，如果 w_i 是未知詞，我們就對 w_i 採對大匹配來斷詞， $w_i = m_1 m_2 \dots m_n$ ，在某些情況下， $m_2 = m_3 = \dots = m_n$ 。如果 w_i 不是未知詞，則構詞的三個特徵值為設為 *null*。此外，若 w_i 為句首或句尾，某些特徵值也是 *null*，例如 $i=1$ 時， $w_{i-1}, w_{i-2} w_{i-1}, t_{i-1}, t_{i-2} t_{i-1}$ 等特徵值皆為 *null*。

訓練資料為詞庫小組一千萬詞平衡語料，並以 Viterbi 演算法實作[14-18]。

三、實驗結果

我們利用上述方法執行台語文詞性標記的工作，不過因為沒有標準答案可以檢查正確率，我們只好抽取部分資料，以人工檢查結果，人工檢查時，主要參考中央研究院詞庫小組的中文斷詞系統。我們選取七篇文章，時間涵蓋清國、日治及終戰後三個不同的時代，文類包括散文（三篇）、劇本（一篇）及小說（三篇），每篇文章挑選第一段，若第一段文字太少，則挑選第二段。表一為測試資料，為挑選出來做人工檢視的，並列出每個段落的音節數、語詞數、挑錯語詞數、詞性標記錯誤數，以及詞性標記正確率。

表一、測試資料及其詞性標記正確率

id	年	文類	作者	題目	音節數	語詞數	語詞錯誤	標記錯誤	正確率
1	1885	散文	葉牧師	Pèh-ōe-jī ê lī-ek(白話字的利益)	162	109	9	6	94.50%
2	1919	散文	H S K	Phín-hēng ê ûi-thôan(品行的遺傳)	180	119	6	8	93.28%
3	1990	散文	陳義仁	Lāu-lâng ê kè-tat(老人的價值)	75	49	7	7	85.71%
4	1950	劇本	陳清忠譯	Venice ê Seng-lí-lâng(威尼斯的生意人)	92	58	3	4	93.10%
5	1890	小說	佚名	An-lòk-ke(安樂街)	101	77	7	9	88.31%
6	1924	小說	賴仁聲	Án-niá ê Bák-sái(母親的眼淚)	133	93	7	9	90.32%
7	1990	小說	楊允言譯	Hái-phī Sin-niû(岬角上的新娘)	94	59	7	5	91.53%
					837	564	46	48	91.49%

說明：id 4 原著者為莎士比亞，id 7 原著者為宋澤萊

$$\text{正確率} = \left(1 - \frac{\text{標記錯誤數}}{\text{語詞數}} \right) \times 100\%$$

所選取的資料，總共有 564 個語詞（837 個音節），經過人工檢查，有 46 個語詞挑選錯誤、48 個語詞的詞性標記錯誤，詞性標記的平均正確率為 91.49%。要說明的是，有時所挑選出來的中文詞雖然是錯誤的，但是詞性標記結果可能仍然是正確的，另一方面，中文詞選對，未必詞性標記的結果是正確的。

此外，有時一個台語詞，對應到兩個中文詞，如台語的「壁頂」對應中文的「牆壁上」；

也有兩個台語詞對應一個中文詞的，如台語的「中國 字」對應到中文的「中國字」。前者會當成未知詞來處理，而後者，因為拆開的兩部分也都是詞，所以就當成兩個詞處理。這個部分，如果詞性標記是正確的，我們仍將結果視為正確。如果這部分視為錯誤，則平均的正確率會下降 2%左右。

以下為實際的詞性標記結果舉例，此為 id 7 的一部份。其中，第一個欄位是漢羅台語文，第二欄為台語羅馬字（以中括弧刮起來），第三欄為中文候選詞（以大括弧刮起來，前面加‘@’符號者，表示第一欄和第二欄的台語文並未出現在辭典的同一詞條中，直接以漢羅台語文當成中文候選詞），第四欄為挑選出的中文詞（以角括弧刮起來），最後一欄為所挑選的詞性。我們將挑選的中文詞錯誤或是挑選的詞性標記錯誤者加上底線，之前並加上‘**’，錯誤的詞性之後，加上正確的詞性標記（粗體，以小括弧刮起來）。

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-teng2]{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)
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[khòa ⁿ]{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>(Na)
， [,]<,>(COMMACATEGORY)

看[khòan]{看}<看>(VC)
 起起[khí-khí]{@起起}<起起>(**Nb)(VA)
 落落[lòh-lòh]{@落落}<落落>(VA)
 ê[ê]{的}<的>(DE)
 海湧[hái-éng]{海浪;海潮}<海浪>(Na)
 , [,]<,>(COMMACATEGORY)
 因為[in-ūi]{由於;因為}<因為>(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)
 來[lái]{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[liàng-liàng]{冷冷;涼絲絲}<冷冷>(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)

四、分析

我們針對選錯中文詞或詞性標記錯誤之處，做更詳細的檢視，發現其中有 13 處是因為選錯中文詞導致詞性標記錯誤。表二列出選錯的中文詞及其標記的詞性。

表二、系統選錯的中文詞

台語詞	所選的中文詞及詞性	較適當的中文對譯及詞性	說明
押/ah	強制(D)	押(VC)	
無/bô	不(D)	沒有(VJ)	2 次
這號/chit-hō	這樣(VH)	這種(N?)	2 次
轉/chōan	上(Ncd)	轉(Vac)	2 次
夭壽/iáu-siū	非常(Dfa)	早夭(VH)	
價值/kè-tát	值得(VH)	價值(Na)	
活/óah	生活(Na)	活(VH)	
破相/phòh-siū ⁿ	破(VHC)	殘廢(Na)	
相借問/sio-chioh-m̄ng	招呼(VC)	打招呼(VB)	
著/tiòh	就(P)	得(D)	

有兩處選錯中文詞是因為台華辭典中沒有正確中文詞的選項，也導致詞性標記錯誤。這表示台華辭典還需要繼續增補。表三列出這兩個語詞。

表三、台華辭典缺中文對譯導致選錯的詞

台語詞	系統所選的中文詞	較正確的中文詞
tiā ⁿ -tiā ⁿ / tiā ⁿ -tiā ⁿ	常常(D)	而已(T)
轉 / tng	調解(VC)	轉(VAC)

另外，還有八處錯誤是由於未知詞的詞性標記錯誤。這些未知詞大部分是兩個中文詞。表四列出此八個未知詞。

表四、中文未知詞

台語詞	中文	系統所選的詞性	正確的詞性
bē 會/bē-ē	不會	Nb	D
食老/chiah-lāu	*食老	Na	V?
轉了/chōan-liáu	*轉了	VH	V?
法律上/hoat-lùt-siōng	法律上	VC	N?
非為/hui-ûi	非為	A	N?
窮志/kiōng-chì	窮志	Na	V?
輕輕仔/khin-kin-á	輕輕的	Nb	D?
生子/se ⁿ -kiá ⁿ	生子	Na	V?

還有四個詞性標記錯誤的地方，可能是因為之前一個詞性標記錯誤而受到影響的，屬於傳播錯誤(propagation error)，包括一個未知詞。

人名的部分，「天賜 ah/Thian-sù ah」的「天賜」（不是未知詞）被標記為「A」，後綴的「ah」被標記為「T」或「Di」（共出現兩次，一次選「啊」另外一次選「了」）。

另外有一個台語詞「對/tùi」，在大部分語境下，其中文對譯為「從」。這個語詞在測試資料中共出現九次，不過系統有七次挑選出的中文詞是「對」，只有兩次挑選「從」。但是因為詞性標記都是「P」，對詞性標記正確率沒有影響。

其它的錯誤共有 18 處，暫時沒有辦法明確分析出其詞性標記錯誤的原因。

總結我們所分析的詞性標記錯誤的原因及其比例，列於表五。

表五、詞性標記錯誤分析

錯誤原因	次數	比例	說明
選錯中文詞	13	27.08%	
沒有正確的中文詞可選	2	4.17%	
未知詞	8	16.67%	
人名	4	8.33%	
傳播錯誤	4	8.33%	包括一未知詞
總計	30	62.50%	扣除重複算的

最理想的情形，如果上述錯誤都得到解決，以此方法做台語詞性標記，將可以達到 96.81% 的正確率，但是顯然有極大的困難。

台語的詞序與中文的詞序畢竟有差異，選錯中文詞導致詞性標記錯誤是比例最高的。而沒有正確中文詞可選的問題，可以透過增補台華對譯辭典的詞條獲得解決，但是正確率僅提升不到 5%。

未知詞導致的詞性標記錯誤，佔第二高的比例，以中文的角度看，這些不是真正的未知詞，大都是因為兩個不同語言的語詞對譯未必是一對一的緣故；另外一個重要的原因是，台語羅馬字在連字符(hyphen)的使用上也尚未標準化，漢語系各語言，可能是使用漢字的關係，語詞的界線相對不明確，台語使用羅馬字書寫，利用連字符，一方面斷開一語詞的各音節，讓一個音節仍可連結一個漢字，另一方面，則又負擔的分詞的功能，有連字符連接的音節，代表同一語詞，語詞間有空白分隔開；但是分詞的部分，原來的漢字書寫無可與之對應。

一個音節可以再細分為聲母、韻母、聲調三部分，由這三部分所組成的音節，台語約有 3,000 個音節，華語有 1,200 個左右，在此前提下，台語有較多的單音節詞。但是，一個單音節可能對應到好幾個不同的漢字，雙音節或以上的語詞則解決大部分的問題。例如台語的「這個」，若寫成「chit ê」(沒有連字符)，看到「chit」，可對應的漢字包括「這、職、質、織、...」等，看到「ê」，可對應的漢字包括「的、個、鞋、...」等，如果寫成「chit-ê」(有連字符)時，通常閱讀者可直接對應到「這個」，因此在台語的羅馬字書寫，書寫者可能會傾向把單音節詞和另一個單音節詞用連字符連起來，如果這兩個單音節詞能形成複合詞或詞組。現實的情形是，在語料中，「這個」有加連字符，有的沒有加，存在著不一致的現象。

因為連字符導致一個台語詞對應兩個中文詞的問題，如果不修改文本，當中文對譯詞是未知詞時，也許可考慮再回頭拆掉連字符試試看。這樣做，也許可以降低因為未知詞導致詞性標記錯誤的機會。

至於不同時代及不同文類的文本，是否在詞性標記的正確率有所差異？根據表一的資料，表六列出三種不同文類文本的詞性標記正確率，表七列出三個不同時代文本的詞性標記正確率。由表六看來，小說類文本的詞性標記正確率較低，表七則顯示，不同時代文本的詞性標記正確率，並沒有顯著的差異。不過，因為資料量少，此分析結果還需進一步驗證。

表六、不同文類文本詞性標記正確率比較

文類	語詞數	標記錯誤	正確率
散文	277	21	92.42%
劇本	58	4	93.10%
小說	229	23	89.96%

表七、不同時代文本詞性標記正確率比較

年代	語詞數	標記錯誤	正確率
清國	186	15	91.94%
日治	212	17	91.98%
戰後	166	16	90.36%

五、未來方向

在缺乏台語文訓練資料的情形下，我們繞了一圈，利用台語華語對譯以及中文的訓練資料，讓台語文詞性標記達到 91.49%的正確率。這兩百多萬音節的台語文語料的詞性標記結果雖然沒有完全正確，但是應該足以提供做為台語文語詞及詞性的訓練資料，可供進一步研究使用。

如果這份台語文的訓練資料是可用的，未來我們希望能透過比較中文和台語文的雙連語詞或雙連詞性，進一步分析中文和台語文的異同處。

我們並開發台語文斷詞、詞性標示系統系統[19]供大眾使用，不過一般人要同時準備台語文全羅馬字及漢羅合用兩種文本有點困難，因此我們還提供只用台語羅馬字或只用漢羅合用（包括完全使用漢字）的台語詞性標記，做法上與上述的相同，只是在查閱台華

辭典時少核對一個欄位，這會造成中文候選詞增加，有可能造成詞性標記錯誤的機會。其結果如何，有待進一步分析。

另外，漢羅合用台語文是比較容易取得的文本，需先經過斷詞才能進行後續的詞性標記，因此我們也將斷詞系統整合在此線上系統中。

致謝

本研究得到國科會計畫「台語文語法結構樹建置(1/3)」NSC 95-2221-E-122 -006 經費補助，特此致謝。此外，也感謝三位匿名審查者所提供的建設性意見，讓本文得以更為周延。

參考文獻

- [1] Gordon, Raymond G. Jr., Ed., *Ethnologue : Languages of the world*. 15th ed. SIL International, 2005. [Online] Available: <http://www.ethnologue.com/> [Accessed: Jun. 30, 2008].
- [2] 黃宣範, *語言，社會與族群意識*, 台北：文鶴, 1993.
- [3] 李勤岸 洪惟仁, “沒有名字的語言？「台灣話」、「閩南話」還是 Hòh-ló 話？”, *台灣文學館通訊* 15 期, 台南：國家台灣文學館, pp36-41, May. 2007.
- [4] 吳守禮, *閩台方言研究集 1*. 台北：南天, 1995.
- [5] 張裕宏, *白話字基本論：白話文對應&相關的議題淺說*, 台北：文鶴, 2001.
- [6] H.K. Tiunn, “Writing in Two Scripts : A Case Study of Digraphia in Taiwanese,” *Written Language and Literacy*, vol. 1, no. 2, pp. 223-231, 1998.
- [7] 楊允言, “台文華文線上辭典建置技術及使用情形探討”, in *2003 第三屆全球華文網路教育國際學術研討會*, 2003, pp. 132-141.
- [8] 楊允言 and 劉杰岳, “台語文線頂辭典 kap 語料庫簡介”, in *語言、社會與文化系列叢書之二 語言政策的多元文化思考*, 鄭錦全等 Ed. 台北：中央研究院語言學研

究所, 2007, pp. 132-141.

- [9] U.G. Iunn, K.G. Lau, H.G. Tan-Tenn, S.A. Lee and C.Y. Kao, "Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 12, no. 4, Dec. 2007, pp. 349-370.
- [10] 詞庫小組, "中文詞類分析," 詞庫小組, 台北, 台灣, Tech. Rep. no.93-05, 1993.
- [11] 楊允言, "台語白話文學ê全新表現——台語文數位典藏資料庫簡介," *台灣文學館通訊* 15 期, 台南: 國家台灣文學館, pp42-44, May. 2007. [Online] Available: <http://iug.csie.dahan.edu.tw/iug/Ungian/Chokphin/Phenglun/DADWT/dadwt.asp> [Accessed: Jun. 30, 2008].
- [12] 楊允言, "台語文/華文辭典," Dec. 2000. [Online]. Available: <http://iug.csie.dahan.edu.tw/q/q.asp> [Accessed Jun. 30, 2008].
- [13] 劉杰岳, "全羅漢羅對照文本找華語候選詞" Aug. 2007. [Online]. Available: http://iug.csie.dahan.edu.tw/nmtl/dadwt/pos_tagging/clhl_hoagi_hausoansu.asp [Accessed Jun. 30, 2008].
- [14] C. Samuelsson, "Statistical methods," in *the Oxford Handbook of Computational Linguistics*, R. Mitkov, Ed. New York: Oxford Univ. Press, 2003, pp. 358-375.
- [15] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania, 1996, pp. 133-142. [Online] Available: <http://acl.ldc.upenn.edu/W/W96/W96-0213.pdf> [Accessed: Jun. 30, 2008].
- [16] A. McCallum, D. Freitag and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," in *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford Univ., Jun. 2000, pp. 591-598. [Online] Available: <http://www.cs.umass.edu/~mccallum/papers/memm-icml2000.ps> [Accessed: Jun. 30, 2008]
- [17] Y.F. Tsai and K.J. Chen, "Reliable and Cost-Effective Pos-Tagging," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 9, no. 1, Feb. 2004, pp. 83-96. [Online] Available: <http://rocling.iis.sinica.edu.tw/CKIP/paper/>

Reliable_and_Cost-Effective_PoS-Tagging.pdf [Accessed: Jun. 30, 2008]

[18]戴嘉宏,“台語選詞跟詞性” Jun. 2007. [Online]. Available: <http://140.109.19.105/>
[Accessed Jun. 30, 2008].

[19]楊允言,劉杰岳,戴嘉宏,“台語文斷詞、詞性標示系統” Aug. 2007. [Online]. Available:
<http://iug.csie.dahan.edu.tw/TGB/tagging/> [Accessed Jun. 30, 2008].