利用關聯式規則解決台語文轉音系統中一詞多音之歧異

Applying Association Rules in Solving the Polysemy Problem in a Chinese to Taiwanese TTS System

林義証　Yih-Jeng Lin
建國科技大學資訊管理系
Department of Information Management
Chien-Kuo Technology University
yclin@ctu.edu.tw

余明興　Ming-Shing Yu
國立中興大學資訊科學與工程學系
Department of Computer Science and Engineering
National Chung-Hsing University
msyu@nchu.edu.tw

李尉綸　Wei-Lun Li
國立中興大學資訊科學與工程學系
Department of Computer Science and Engineering
National Chung-Hsing University
ww90053@hotmail.com

摘要

本論文提出一個利用關聯式規則解決台語文轉音系統中的一詞多音問題，台語中的一詞多音指的是一個詞在不同的場合會有不同的發音，如果不做適當處理，會造成合成之台語語音發音的錯誤，導致合成語音之瞭解度下降。有別於既有之解決一詞多音方法，關聯式規則具有推衍出較具物理意義之規則，可以模擬人們在語言使用的習慣，對於解決一詞多音有相當的助益，本文除了利用關聯式規則外，也使用了 E-Hownet 來解決訓練語料稀疏的問題。我們針對六種常見的多音詞進行實驗，分別為『上』、『下』、『不』、『你』、『我』、『他』等詞。分別得到正確率為 93.99%、94.44%、77.82%、95.11%、96.35%、95.99%。實驗結果顯示，相較於現有的實驗方法，本論文提出的方法對於上述六個多音詞皆達到更高的正確率。

Abstract

This paper proposed an approach that apply association rules to solve the polysemy problem in a Chinese to Taiwanese TTS system. In Taiwanese, one word possibly has several pronunciations. It leads to that Taiwanese TTS System can't work well if wrong pronunciation is synthesized. Thus, we need to decide which pronunciation is proper under

some condition in order to enhance the performance of a Taiwanese TTS System. The approach of association rules has the advantage of similate the usage of language by human beings. We also use the information in E-Hownet to overcome the problem of data sparsity.The experiments focus on six common used words, they are “上”(up), ”下”(down), ”不”(no), “你”, “我”, and “他”(he). Experiment results show that the proposed approach can achieve higher accuracies than the existing methods.

關鍵詞：關聯式規則、知網、一詞多音、台語文轉音系統。

Keywords: Association rules, E-Hownet, Polysemy, and Taiwanese TTS system.

一、緒論

台語，是指具有臺灣地方特色的閩南語，為臺灣第一大母語，同時在台灣也是使用人數第二多的語言。根據統計[12]，約有 73%的臺灣民眾會使用台語。隨著近年來教育部對於母語教育的倡導，台語教學也逐漸受到國人重視，因此針對於台語文轉音系統之研究也開始成為一項重要課題。

然而台語並不像中文一樣擁有正式且統一的書寫文字。因此國內台語文轉音系統的研究主要以中文文句作為輸入，之後再將中文文句轉成台語語音輸出，即中文轉台語語音合成系統(Chinese to Taiwanese TTS system) [3][6][10][11][20]。

對於一套文轉音系統來說，能合成出正確的發音是相當重要的。然而，在中文轉台語文轉音系統中存在著一詞多音現象。所謂一詞多音現象，顧名思義就是：同一個中文詞，在不同的場合中有兩種以上不同的發音。如果沒有針對一詞多音現象提供適當的辨識機制，將會造成文轉音系統誤判多音詞的正確發音，進而影響到語音合成的正確性。

一個可能的中文轉台語文轉音系統架構包含(1) 文句分析模組、(2) 音的韻律訊息處理模組、和(3) 語音處理及輸出模組等。這三個模組的功能大致如下：

1. 文句分析：

將文字對應的拼音找出來，其作法是將輸入的文字經過分析斷詞後（word segmentation）[4][7][9][11][20]，再找出中文詞對應到的台語發音並依台語的連音變調[3][6]方式決定每個音節的正確聲調。有的還有特殊符號的口語表示法或者是韻律段。

2. 音的韻律訊息：

決定每個音在發音時的參數，有音調、音量、音長、音與音之間的停頓時間、音與音之間相連音的程度等。

3. 語音的處理：

依照所給的韻律訊息，將每個音和連音做適當的處理，以得到所需要的語音。

以『我們真希望科學家能發明打下去不會痛的針』為例子，在中文轉台語語音時的文句處理過程如下：

【原始中文文句】:

我們真希望科學家能發明打下去不會痛的針。

【經過中文文句的斷詞後】:

我們/真/希望/科學家/能/發明/打/下去/不會/痛/的/針。

【轉換成台文語句後】:

阮(ghun2)/真/希望/科學家/能夠(e3-dang3)/發明/注/下去/不(bhe7)/痛/的/射。

在這個例子中，能夠發現一些文句分析模組中需要處理的多音歧異現象。首先是語意上多音歧異的問題：在原本的中文文句中的「我們」，對應到台語詞典中會發現有兩種結果，分別為「阮（ghun2）」和「咱（lan2）」。「阮（ghun2）」在語意上是不包含聽者的「我們」，相對地，「咱（lan2）」在語意上則是包含聽者的「我們」。
再來是一詞多音問題：輸入的中文文句中出現了一字詞「不」。然而，「不」的台語發音存在著/bhuaih4/、/bhe7/、/bho5/、/m7/、/but4/和/mai3/等六種。這種現象同時也會出現在其他詞上面，例如「你」、「上」、「下」和「會」等，這一類的多音歧義現象，我們稱為一詞多音。
除此之外，還有詞被分開的現象：如例句中「打下去不會痛的針」，原本中文詞「打針」對照到台語發音為「注射（/cu2 sia7/）」。但是像例句中的情況，「打針」這個詞卻被分開為「打」和「針」。如果單純地使用雙語詞典來進行中文轉台語對照的話，將會誤判為「打（/pah4/）」和「針（/ziam1/）」，而得到錯誤的台語標音。

本文將針對中文轉台語文轉音系統中的一詞多音問題進行探討，並且利用關聯式規則(Association Rules) [1][5][8]來進行多音詞讀音預測，以期提昇對於台語一詞多音辨識的正確率。
本論文中的第一章我們會先介紹台語文轉音系統；第二章說明台語一詞多音的問題及文獻探討；第三章則是研究的方法；第四章為實驗和實驗結果；第五章為結論及未來工作。

二、台語一詞多音現象及文獻探討

2.1 台語的一詞多音

台語的一詞多音是指同一個詞在不同的場合會有不同的發音。以『我們』這個詞為例。在一篇文章中，可能會代表二種意義，一個是文章中的『我們』包含了語者和聽者;另一個是文章中的『我們』不包含聽者。而前者的『我們』在台語中的唸法為『lan2(咱)』;後者的『我們』在台語中的唸法為『ghun2(阮)』。在台語語音合成系統中一詞多音的現象很常見，也比中文更多更複雜。例如:『行』在台語語音中就有八種發音，『不』也有六種發音。如果一個台語的文轉音系統無法正確決定發音，那會導致聽者誤解語意，嚴重影響到台語語音合成的品質。『不』在台語中有六種不同的發音[11][13][15][16]，如Ex1~Ex6 所示。而這些不同的發音通常會對應的不同的意義：
(Ex1)『但是社會上的一般人並不/bo5/容易看出它的重要性。』
(Ex2)『不/m7/知浪費了多少國家資源。』

(Ex3)『讓人一時聯想不/bhe3/到他與機械的關係。』
(Ex4)『我不/but4/忍心傷害她。』
(Ex5)『不/bhuaih4/作正面肯定答覆，』
(Ex6)『我希望不/mai3/傷她的心。』
Ex7-Ex9 說明了『上』的三種發音，其中 Ex7 的『上』表示前一個(previous)的意義；Ex8 的『上』表示在上面(above)的意義；而 Ex9 的『上』則表示搭乘(take)的意義。
(Ex7) 『我**上/ding2/**半月花了好多錢去買有關台語的教科書。』
(Ex8) 『我是在這地圖**上/siang7/**的哪裡？』
(Ex9) 『我**上/jiunn7/**了公車後才發現我搭錯車了。』

表一是我們列出了一些在台語中具有一詞多音現象的詞。

表一、 台語一詞多音詞

多音字	台語發音	範　例
你	/li2/	你是誰
	/lin2/	你媽媽、你家
我	/ghua2/	我知道
	/ghun2/	我爸、我家
他	/yi7/	給他機會
	/yin7/	他老婆、他家
上	/ziunn7/	電梯上樓
	/siong7/	社會上、歷史上
	/ding2/	車上
下	/e7/	下面
	/ha7/	在下、狀況下
	/au7/	下一次
	/loh8/	下雨
大	/dua7/	下大雨
	/dai7/	大人、大學
不	/bhe7/	不會、想不到
	/bho5/	不容易
	/m7/	不知道
	/mai3/	不去
	/but4/	不歸
	/bhuaih4/	不要去
長	/diunn2/	站長、市長、學長
	/dng5/	長工、長江
	/diong2/	專長、兄長
	/ciang5/	粗長、高長(高大)
	/senn1/	長的很像
生	/cen1/	生啤酒、生魚片（不熟的意思）
	/senn1/	滋生、怕生（有生長的意思）

	/sing1/	微生物、謀生（指物品或人）
香	/hiang1/	香油、香料
	/hiong1/	香港腳、香港人、香片
	/hiunn1/	蚊香、進香、香煙
	/pang1/	香味
少	/ziou2/	不少、少不了、至少
	/ziau2/	減了什麼東西
	/siau3/	少年、少女、少林寺
天	/ten1/	天上人間
	/tinn1/	天公
	/gang1/	一天、兩天
放	/hong3/	放心、放大
	/bang3/	放走、放棄
	/kng3/	放在

2.2 文獻探討

目前針對台語一詞多音問題的相關研究中，主要以監督式學習法來達成。在監督式學習法中，訓練語料需要事先進行人工標記，其優點是擁有較高的正確率。其中較佳的研究方法包括了組合式策略（結合語言模型和決策樹）[15]、階層式方法[13]以及規則轉換學習法[16]等。

組合式策略[15]是由林金玉等人提出，組合式策略主要將單詞語言模型（Word-base Uni-gram Language Model,簡稱 WU）和決策樹中的 CHAID 演算法 (CHi-squared Automatic Interaction Detector，卡方自動互助檢視法)兩者作一個結合的組合式策略。WU 以相鄰詞為特徵判斷發音，而 CHAID 則以詞性為主要的判斷依據，也就是當訓練資料較稀疏時，組合式中的 CHAID 演算法可以決定一些正確的發音。

階層式方法[13]的概念是考慮詞的相鄰位置關係，從而統計出目標多音詞在哪種發音下比較會常出現。階層式分成四層，由條件較嚴苛的第四層(需比對一詞多音詞前後各二個詞相同)先比較，若滿足則輸出預測相對應發音，否則進入第三層(需比對詞的 trigram 相同)，若滿足則輸出預測相對應發音，不滿足則繼續往下一層比對，若都無法找到相對應的比對，則輸出訓練語料比例最高的發音。

楊文德[16]利用規則轉換學習法(Transformation based learning, TBL)解決台語的一詞多音問題，規則轉換學習法的原理則是利用以人工標記過的標準文字檔(Standard Text)和經由模型標注的答案文字檔(Answer Text)互相比對後，找出錯誤的地方，再經由轉換規則樣版(Transformation Templates)，找出轉換規則。規則轉換學習法從所有的標準文字檔與答案文字檔不同的地方，套用轉換規則樣版，取出轉換規則，接著逐一代入所有規則，記住錯誤數最少的一條，作為第一輪學習的結果，然後在引入第一條規則後，重複代入、記住最少錯誤的規則、應用規則，直到整體錯誤率不再改變為止，學習才結束。接下來，就能將測試語料套用轉換規則樣版，以進行分類作業。

這三種方法對於判斷一詞多音，各有千秋，如表二所示，階層式方法對於『下』有很好

的正確率；組合式策略則適用於『你』和『我』；TBL 對於『上』、『不』和『他』則表現較佳。

表二、階層式、組合式、TBL 正確率之比較

目標詞 / 實驗方法	上	下	不	你	我	他
階層式方法[13]	90.42%	94.30%	74.47%	92.12%	89.23%	92.54%
組合式策略[15]	83.88%	88.43%	69.29%	94.20%	90.59%	93.29%
TBL[16]	90.98%	90.79%	73.25%	93.78%	88.17%	94.48%
TBL（規則有排序）[16]	90.51%	91.81%	76.84%	93.78%	89.39%	94.19%

三、研究方法

3.1 關聯式規則

關聯式規則(Association Rules)，是資料探勘中的一項技術，最早是由 Agrawal & Srikant 所提出的觀念[1][5]。關聯式規則可以幫助我們瞭解資料之間所存在的相互關係。透過關聯式規則，使用者便能夠分析資料中項目與項目之間的關係，並且以容易理解的蘊涵式表達出來。

舉個較為實際的用途，像是藉由關聯式規則來分析交易記錄的資料，藉此瞭解消費者的購買行為模式[5]，例如：「如果今天一個消費者購買了『牛奶』，那麼他有多大的機率也會一起購買『麵包』？」又或者「如果消費者買了『啤酒』，那麼他還有可能順便一起買什麼商品？」。藉此掌握消費者的購買習慣，賣場就能夠制定更有效的促銷方案。

如美國著名的超級市場沃爾瑪(Wal-Mart)，他們透過多年來所記錄的商品銷售資料來進行關聯式規則分析，結果從分析出來的規則中發現兩種看似毫不相干的商品：啤酒和尿布，在星期四同時會被購買的頻率相當地高。由於從關聯式規則中發現了這個特別的現象，沃爾瑪超市嘗試著將啤酒和尿布兩種商品放置在臨近的區域，以便消費者能夠拿取。最後成功地將兩種商品的銷售量都提升上去。

關聯式規則通常以「X⇒Y」的蘊涵式來表達，其記錄著項目集之間所存在的關係。這意味者在資料庫中，若 X 項目集出現，則 Y 項目集也出現的可能性。例如當 X={Milk, Diaper}而 Y={Beer}，此時{Milk, Diaper}⇒ {Beer}表達著：若顧客買了牛奶(Milk)和尿布(Diaper)，而他們同時也會購買啤酒(Beer)的可能性。

關聯式規則在生成的同時也會一併記錄兩個參數：支持度(support)和信心度(confidence)。當規則為 X⇒Y 時，其**支持度(support)**和**信心度(confidence)**的公式分別為：

$$\text{support}(X \Rightarrow Y) = P(X \cap Y) \quad \cdots\cdots(1)$$

$$\text{confidence}(X \Rightarrow Y) = P(Y|X) \quad \cdots\cdots(2)$$

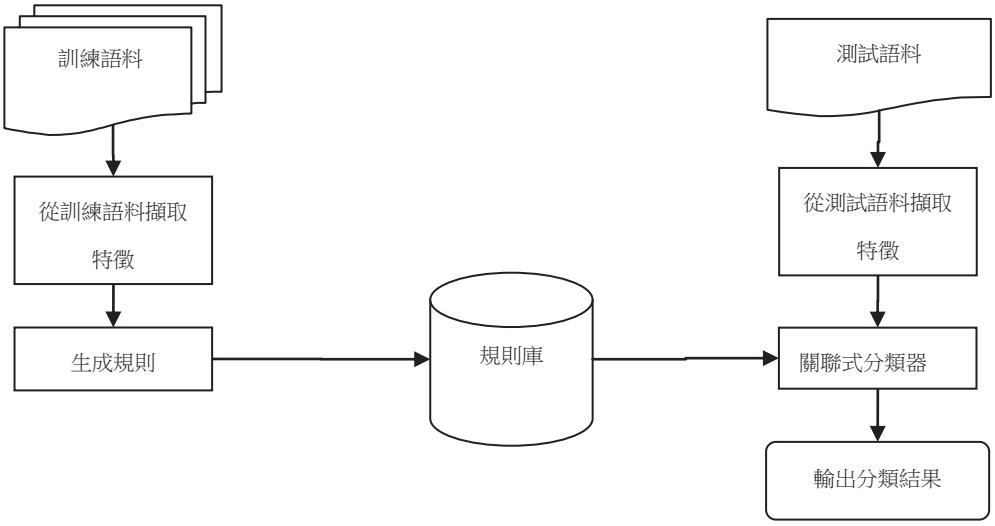
支持度代表在資料庫所有交易記錄中，項目集 X 和 Y 同時出現的機率。信心度代表著在 X 出現的前提下，Y 也出現的條件機率。利用關聯式規則來對資料進行分類預測，稱之為關聯式分類（Associative Classification，AC）[1] [5]，為監督式學習法。

在使用關聯式規則進行分類作業的時候，需要事先將關聯式規則設定排序的機制。當我們要判斷一筆資料應該屬於哪種類別的同時，往往會出現多筆規則符合這筆資料。一般

在利用關聯式規則進行關聯式分類的時候，首先以信心度高的規則優先採用，當規則之間的信心度相等的情況下，才以支持度較高的規則優先採用[5]。

3.2 利用關聯式規則解決台語一詞多音問題

圖一是本文利用關聯式規則的實驗流程，我們將語料分成訓練語料和測試語料，前者用於找出關聯式規則及相關參數；後者則用於求得外部測試之正確率。



圖一、實驗流程

3.2.1 特徵擷取

判斷一詞多音問題時，首先第一步是從語料中進行特徵擷取。在台語文轉音系統中的文句分析模組裡面，會先將中文文句作斷詞的動作。例如一段中文句子經過斷詞與詞性標記之後，從結果我們可以發現，這段中文文句中出現了三個多音詞「上」：：

「我(Nh) 現在(Nd) 上(VCL) 了(Di) 公車(Na) ， 在(P) 車(Na) 上(Ncd) 發現(VE) 身(Na) 上(Ng) 沒有(VJ) 零錢(Na) 。」

一般在處理一詞多音問題時，是根據多音詞「上」前後所出現的詞、詞性以及「上」本身的詞性，作為判斷一詞多音問題用的特徵。如表三所示，多音詞「上」的樣本經過台語讀音的標記之後，即可作為生成關聯式規則時的實驗語料。

表三、 多音詞的特徵擷取（以「上」為例）

編號	讀音	特徵	前二	前一	目標詞	後一	後二	讀音
語料 1	/ziuun7/	詞	我	現在	上	了	公車	/ziuun7/
		詞性	Nh	Nd	VCL	Di	Na	
語料 2	/ding2/	詞	在	車	上	發現	身	/ding2/
		詞性	P	Na	Ncd	VE	Na	
語料 3	/siong7/	詞	發現	身	上	沒有	零錢	/siong7/
		詞性	VE	Na	Ng	VJ	Na	

3.2.2 規則生成

我們採用 Apriori 演算法來生成規則[8]。Apriori 演算法是關聯式規則探勘方法中較為著

名的演算法之一，透過逐步刪除掉支持度過低的項目集作為生成規則的機制。首先在生成規則之前，需要事先設定好最低支持度(min-support) 和最低信心度(min-confidence)。定義一個包含了 k 個項目的項目集合稱為 k-項目集(k-itemset)，並且定義符號L_k為所有高於最低支持度這個門檻值的集合，稱為大型 k-項目集(large k-itemset)，定義為L_k。Apriori 的作法是先由資料庫找出 1-itemset(僅僅包含單一項目的項目集)，再從 1-itemset 之中剔除掉支持度過低的項目集，儲存為 L_1 (large 1-itemset) 。因為 1-itemset 之中被剔除掉的項目集無法成為L_2中任何一個項目集的子集，所以接下來只需要根據L_1中記錄的項目集來生成下一階段的 2-itemset(一次記錄兩個項目的項目集)。同樣再從 2-itemset 中剔除掉支持度過低的項目集，儲存為L_2。依此類推，直到最後 n-itemset 中所有的項目集皆低於最低支持度(即代表L_n無法生成)為止。接著再根據L_1至L_{n-1}來生成規則，並且計算每條規則的信心度，保留高於或等於最低信心度參數的規則。

針對生成規則方面，我們在參數設定上作了一點修正，以便在實作上能更加適用於預測一詞多音問題。

(1) 修正支持度的計算方法：一般在生成規則時必須連帶計算出規則本身的支持度，若規則本身的支持度低於事先設定好的最低支持度，那麼這條規則將不會被採用。原本的支持度公式為：

$$\text{support}(X \Rightarrow Y) = P(X \cap Y) = \frac{|X \cap Y|}{|D|} \quad \cdots\cdots(3)$$

在解決一詞多音的實驗中，X 記錄著出現的特徵項目集，Y 記錄著預測的讀音。|X ∩ Y|代表在訓練語料中對應的特徵和讀音同時出現的次數，|D|則代表生成規則時的訓練語料筆數。以目標詞「上」的實驗語料為例子。我們從 10648 筆訓練語料中生成關聯式規則，其中一條規則為：

『業務（前一詞）⇒ /siong7/（讀音）』

根據這 10648 筆訓練語料中的統計，「上」的前一詞為『業務』且讀音為『/siong7/』的組合共出現了 5 次。如此一來，換算成支持度為： 5/10648=0.00046957。從這個例子可以發現，若依照原本的方式來計算支持度，會得到一個極小的小數。在後續的實驗中，若我們想要針對最低支持度這項參數進行更動的時候，會變得難以調整。由於我們一詞多音的實驗中採用了中文詞作為特徵，而中文詞本身容易產生資料稀疏的問題(例如出現較少見的專有名詞)。解決的方法是改回原本的出現頻率代替原本的支持度參數，出現頻率的公式如下：

$$\text{Freq}(X \Rightarrow Y) = |X \cap Y| \quad \cdots\cdots(4)$$

例如上述例子中，前一詞為『業務』且讀音為『/siong7/』的組合共出現了 5 次，則將該條規則的出現頻率記錄為 5 次，如此一來就能夠以正整數的形式進行參數設定。

(2) 規則長度的上限：規則長度，即關聯式分類中規則的基數(rule cardinality) [7]，代表著規則左式記錄的項目數量。例如一字詞「上」的實驗中，某一規則為：

「Ncd(目標詞的詞性) 社會(前一詞) => /siong7/(讀音)」

這條規則的左式是由 2 個項目所組合在一起的項目集，即規則長度為 2，同時也代表這條規則是參考了兩種特徵來判斷讀音。原本在生成規則的演算法中並沒有特別

限定規則長度，雖然長度較長的規則將會蘊藏更多的資訊量，但相對地也會拖慢生成規則和關聯式分類的整體速度。我們在實驗中加入了規則長度上限的設定，希望能夠藉此過濾掉因為長度過長而對分類沒有幫助的規則。表四為以「上」為例的生成的關聯式規則範例。

表四、 生成的關聯式規則範例（以「上」為例）

前二詞/詞性	前一詞/詞性	目標詞『上』詞性	後一詞/詞性	後二詞/詞性	讀音	信心度	頻率	支持度
-	身	-	-	Na	/siong7/	100.00%	99	0.93%
VE	-	Ng	-	-	/siong7/	100.00%	46	0.43%
Na	VH	-	-	-	/siong7/	100.00%	39	0.37%
-	車	-	-	Na	/ding2/	100.00%	21	0.20%
-	現在	-	-	-	/ziuun7/	100.00%	3	0.03%
-	身	-	-	-	/siong7/	99.35%	616	5.79%
-	車	-	-	-	/ding2/	93.98%	78	0.73%
Nd	-	-	-	-	/siong7/	90.09%	200	1.88%
在	-	-	-	-	/siong7/	80.09%	2586	24.29%
-	-	-	-	Na	/siong7/	77.95%	2574	24.17%
-	-	VCL	-	-	/ziuun7/	66.29%	474	4.45%

四、實驗結果

4.1 實驗設定

4.1.1 實驗語料描述

我們採用中央研究院平衡語料庫 3.0（簡稱 ASBC 3.0）[14]作為實驗語料的來源。中研院平衡語料庫是世界上第一個擁有完整詞類標記的漢語平衡語料庫，裡面收錄的文句包含了斷詞以及詞性標記，很適合作為我們一詞多音研究中的實驗語料。我們抽取出一些常見的多音詞作為實驗語料，並經由人工方式將正確發音標出，其中包括『上』、『下』、『不』、『你』、『我』、『他』等六個多音詞。表五為每種多音詞的語料筆數和可能會出現的讀音標記。

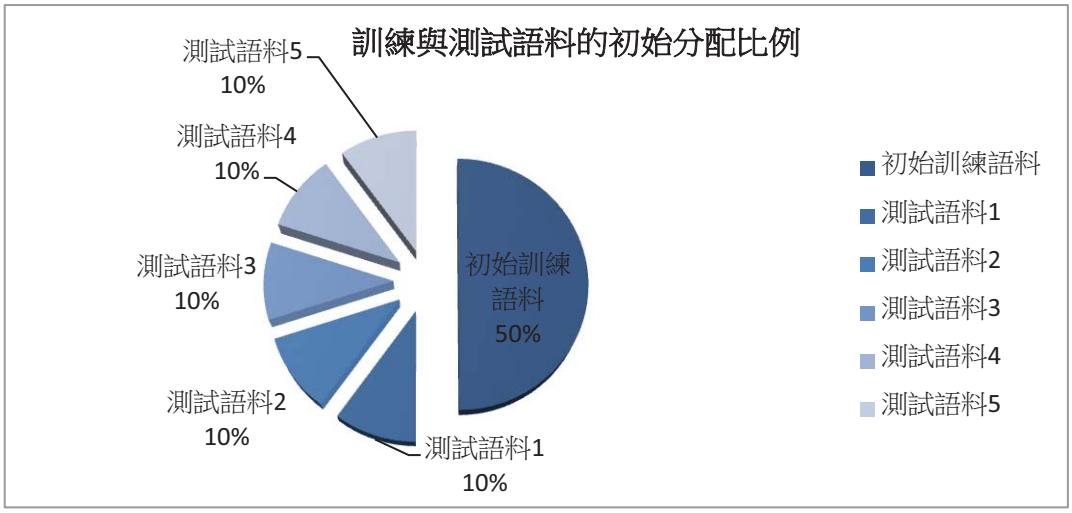
表五、 語料筆數和各種讀音標記

多音詞	台語讀音標記種類	讀音種類數	總筆數
上	/siong7/、/ding2/、/ziuun7/	3	21,294
下	/ha7/、/au7/、/loh8/、/e7/	4	6,840
不	/bho5/、/bhuaih4/、/bhe7/、/m7/、/but4/、/mai3/	6	38,688
你	/li2/、/lin2/	2	2,229
我	/ghua2/、/ghun2/	2	5,627
他	/yi1/、/yin1/	2	3,708

4.1.2 訓練與測試語料分配比例

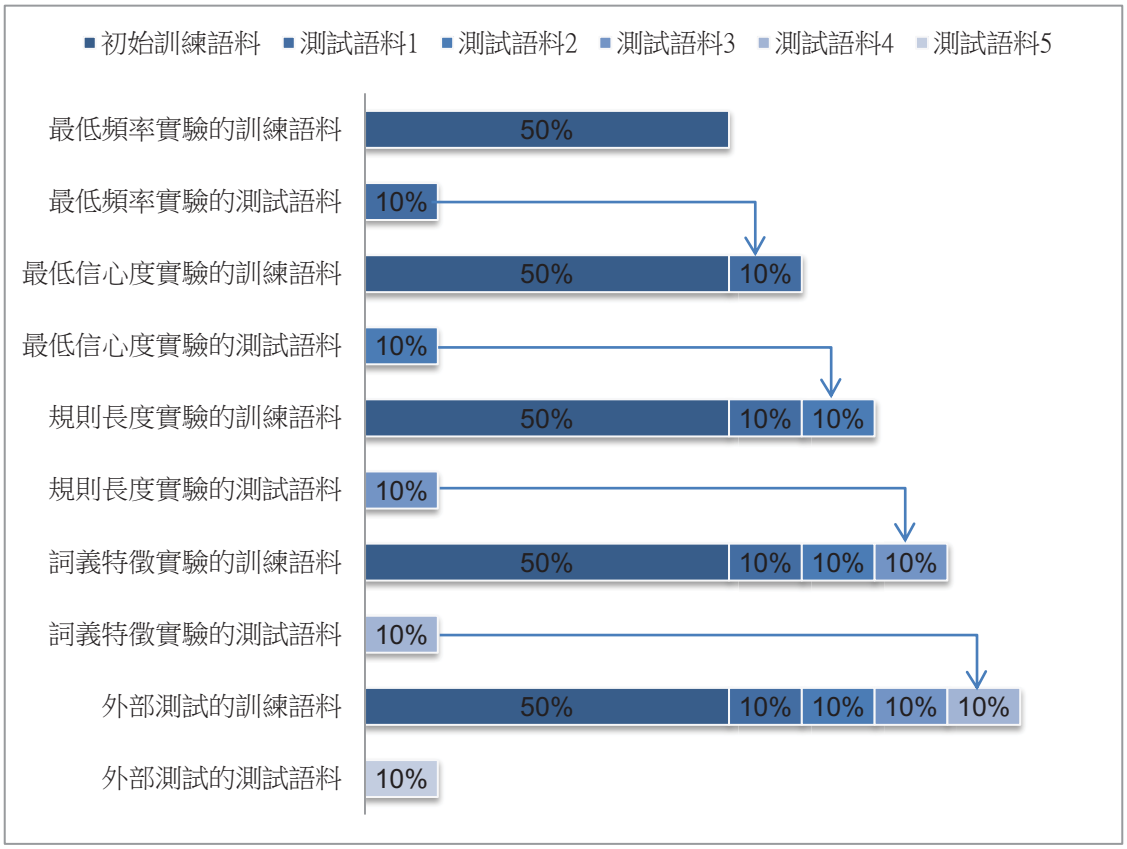
我們採用的訓練、測試語料的分配比例如圖二的說明。一開始，先從實驗語料中隨機選取出 50%的實驗語料作為初始訓練語料。剩餘的 50%實驗語料再平均分成五等份作為測

試語料，分別給予編號為測試語料 1 號至 5 號。



圖二、實驗語料切割

圖三則說明了每份訓練與測試語料的使用時機。在最初的實驗中，我們會先利用初始訓練語料生成關聯式規則，並且使用 1 號測試語料來評估實驗的效能。而後續的第二個實驗中，再將 1 號測試語料併入原本的訓練語料中，成為新的訓練語料（初始訓練語料加上 1 號測試語料），並且使用 2 號測試語料來評估實驗的效能，以此類推。而最後的外部測試實驗中，將會使用 5 號測試語料來評估實驗效能。



圖三、實驗語料分配比例

4.2 參數設定之實驗

在生成關聯式規則之前需要事先設定最低支持度和最低信心度兩種相關的參數。除此之外，我們在實驗中加入了規則長度上限的設定，希望能夠藉此過濾掉因為長度雖長但對於分類卻沒有幫助的規則。當一條規則低於設定的最低支持度或最低信心度，或者規則長度超過了設定的上限，這條規則將不會被記錄到規則庫中。理論上，即使沒有刻意設定上述參數，仍然有可能生成出關聯式規則來進行分類作業。但是這麼作會直接導致生成出來的規則數量過於龐大，進而拖慢分類的速度，而且不見得能夠達到最佳的預測正確率。

反過來說，若參數設定太過於嚴格（例如：最低支持度或最低信心度的設定過高、又或者規則長度上限設定過短），如此一來可以提供資訊的規則又會減少，進而造成分類的正確率下降。

接下來的實驗中，我們將逐步探討最低支持度、最低信心度和規則的長度上限應該設定為多少較佳。此外，為了因應台語一詞多音問題中使用了中文詞作為特徵的情況，最低支持度在計算方式上面需要額外作修正，以下章節將會特別說明。

4.2.2 最低出現頻率的設定實驗

我們分別將每個多音詞透過內部測試來探討規則的最低出現頻率應該為多少較佳。我們分別將最低出現頻率設定從 1 次到 10 次進行分類作業。在這個實驗中，我們採用初始訓練語料(佔整體實驗語料的 50%)來生成規則，並且採用 1 號測試語料來評估分類的正確率。

表六、 最低出現頻率參數的實驗結果

多音詞 / 最低出現頻率	上	下	不	你	我	他
1	91.88%	92.55%	76.69%	94.25%	92.36%	94.13%
2	92.25%	92.70%	77.80%	94.69%	92.71%	94.67%
3	92.35%	92.26%	77.15%	95.13%	93.58%	94.93%
4	92.21%	92.26%	76.51%	93.81%	93.40%	94.67%
5	92.21%	91.82%	76.17%	93.36%	92.71%	94.67%
6	92.02%	91.68%	76.01%	93.36%	92.88%	94.67%
7	91.69%	91.53%	75.76%	93.36%	92.88%	94.67%
8	91.55%	91.24%	75.39%	93.36%	92.88%	94.67%
9	91.55%	90.95%	75.21%	93.36%	92.88%	93.07%
10	91.55%	90.80%	75.08%	93.36%	92.88%	93.07%
最低出現頻率的最佳設定	3	2	2	3	3	3
訓練語料筆數	10648	3420	19344	1114	2813	1854
最低支持度	0.028%	0.058%	0.010%	0.269%	0.107%	0.162%

從表六的實驗結果可以發現，當每個多音詞的最低出現頻率設定在 2~3 之間時，能夠讓

得到較佳的正確率。值得一提的是，實驗中的訓練語料筆數在不同多音詞之間存在著一定的差距。

4.2.3 最低信心度設定實驗

接下來的實驗裡面，我們分別將每個目標詞透過內部測試來探討規則的最小信心度的設定為多少最為適合。我們將最低信心度設定從 0%到 95%，每次提昇 5%，分別各作一次分類作業。我們將初始訓練語料和 1 號測試語料合併成為新的訓練語料(佔整體實驗語料的 60%)，作為這次實驗的訓練語料。在這個實驗中，採用 2 號測試語料來評估分類的正確率。我們從實驗結果中挑選出較佳的最低信心度設定，以正確率較高者優先，若相同則取設定值較高者，如此一來能夠再減少一部分不必要的規則。結果如表七。

表七、最低信心度參數的實驗結果

目標詞 / 最低信心度	上	下	不	你	我	他
0%~45%	92.11%	93.27%	77.62%	94.25%	94.44%	94.13%
50%	92.11%	93.27%	77.64%	94.25%	94.44%	94.13%
55%	92.11%	93.27%	77.64%	94.25%	94.44%	94.13%
60%	92.11%	93.27%	77.64%	94.25%	94.44%	94.13%
65%	92.11%	93.27%	77.64%	94.25%	94.44%	94.13%
70%	92.11%	93.42%	77.54%	94.25%	94.44%	94.13%
75%	92.11%	93.42%	77.28%	94.25%	94.44%	94.13%
80%	92.11%	93.42%	76.97%	94.25%	94.44%	94.13%
85%	92.11%	93.42%	76.14%	94.25%	94.44%	94.13%
90%	92.11%	93.27%	75.14%	94.25%	94.44%	94.13%
95%	91.97%	93.13%	73.84%	94.25%	94.44%	94.13%
最佳的**最低信心度**設定	90%	85%	65%	95%	95%	95%
多音詞的門檻值	81.28%	69.56%	41.37%	87.80%	80.15%	87.85%

4.2.4 規則長度上限設定之實驗結果

當我們在生成規則時，可以選擇只將指定長度以下的規則記錄到規則庫中。例如當我們設定長度上限為 2 時，生成的規則最多只能一次參考兩個特徵來判斷讀音。如果生成的規則長度越長，所能夠提供的資訊量會越多，但相對也會導致生成的規則數量劇增。我們將初始訓練語料、1 號測試語料、2 號測試語料合併成為新的訓練語料(佔整體實驗語料的 70%)，作為這次實驗的訓練語料。在這個實驗中，採用 3 號測試語料來評估分類的正確率。表八是長度上限最佳設定的實驗結果。

表八、規則長度上限的實驗結果

目標詞 / 長度上限	上	下	不	你	我	他
1	91.27%	93.57%	76.73%	92.44%	94.79%	93.87%
2	92.49%	95.47%	77.07%	92.44%	94.62%	93.60%
3	92.30%	94.88%	76.68%	93.78%	94.27%	93.60%
4	92.30%	94.88%	76.65%	93.78%	94.27%	93.60%
5 以上	92.30%	94.88%	76.65%	93.78%	94.27%	93.60%
長度上限的最佳設定	2	2	2	3	1	1

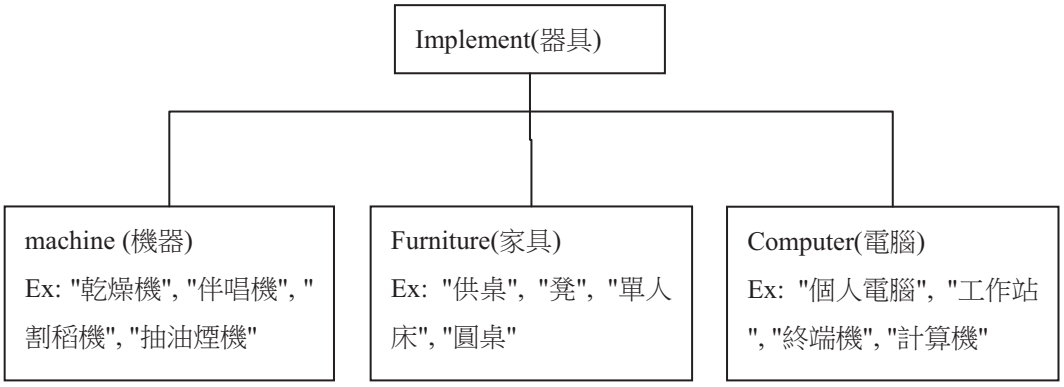
4.3 追加詞義特徵

由於中文詞的數量龐大，會有資料稀疏導致有些情形無法被訓練到的問題，例如『上』的一字詞多音實驗語料中，其中一句經過斷詞與詞性標記後的語料為：

廟(Na)/裡(Ncd)/供桌(Na)/上(Ncd)/的(DE)/水果(Na)/容易(VH)/順手牽羊(VA)

原本『上』的台語讀音應該為/ding2/，分類器卻誤將這筆語料分類為/siong7/。這是由於『上』的前一詞出現『供桌』的情況，在訓練語料中相當罕見，這將會導致沒有規則支持『上』應該念/ding2/。在人工標記的時候，標記者之所以能夠判斷此時的『上』應該念/ding2/，是因為標記者知道『供桌』其實意義上接近『桌子』。當『桌子』或『桌』等詞彙出現在『上』的前一詞時，『上』的讀音應該標記成/ding2/。為了解決這個問題，我們將使用廣義知網中文詞知識庫(E-HowNet) [17] [18] [19]來查詢對應的詞義，並且追加至特徵中。

我們根據輸入的中文詞和詞性，透過中文詞庫中的八萬目詞來查詢其對應到的詞義。如此一來，原本詞頻較低的中文詞將有機會被納入一個出現頻率較高的新特徵裡面。如果仍然因為詞義的頻率過低而導致規則庫中找不到對應的詞義，能夠再根據 E- HowNet 的樹狀結構來取得上一層的詞義，如圖四所示。



圖四、 E-Hownet 詞義樹範例（以『furniture|家具』為例子）

以上述『而且廟裡供桌上的水果容易順手牽羊』句子為例，我們將追加詞義作為新的特徵如表九所示：

表九、追加詞義作為新特徵的範例

特徵位置	前二詞	前一詞	目標詞	後一詞	後二詞
詞(WORD)	裡	供桌	上	的	水果
詞性(POS)	Ncd	Na	Ncd	DE	Na
詞義(Word Sense)	Internal	Furniture	Upper	Relation	Fruit

原本『供桌』等詞彙存在著出現頻率過低的問題，但是透過詞義 Furniture（家具）作為參考之後，就能夠得知『供桌』在詞義上也是一種家具。同樣地，當我們要判斷『供桌/上』、『凳/上』、『單人床/上』等情況的時候，分類器也能根據 Furniture（家具）這個特徵來判斷『上』的讀音。

我們將初始訓練語料、1 號至 3 號測試語料合併成為新的訓練語料(佔整體實驗語料的 80%)，作為這次實驗的訓練語料。在這個實驗中，採用 4 號測試語料來評估分類的正確率。從表十的實驗結果顯示，當我們添加詞義作為特徵之後，能夠再提升所有多音詞的分類正確率。

表十、追加詞義作為特徵的實驗結果

目標詞 / 使用特徵	上	下	不	你	我	他
詞&詞性	93.14%	93.42%	77.23%	93.33%	95.13%	94.12%
詞&詞性&詞義	93.38%	94.30%	77.33%	95.11%	95.83%	94.65%

4.4 外部測試與各種實驗方法比較

我們將上述實驗中所得到的最佳參數設定套用到最後的外部測試實驗中，並且以中文詞、詞性、詞義作為特徵來生成關聯式規則。我們將初始訓練語料、1 號至 4 號測試語料合併成為新的訓練語料(佔整體實驗語料的 90%)，作為外部測試的訓練語料。在最後的外部測試實驗中，採用 5 號測試語料來評估分類的正確率。表十一為本文提出之方法及參數設定和文獻上之方法正確率的比較，實驗結果顯示利用關聯式規則搭配多種參數之設定及加入詞義資訊，有助於大大提升一詞多音的預測正確率，表十一也說明本文提出的方法優於其他文獻上的方法。

表十一、各種方法外部測試正確率之比較

目標詞 / 實驗方法	上	下	不	你	我	他
階層式方法[13]	90.42%	94.30%	74.47%	92.12%	89.23%	92.54%
組合式策略[15]	83.88%	88.43%	69.29%	94.20%	90.59%	93.29%
TBL[16]	90.98%	90.79%	73.25%	93.78%	88.17%	94.48%
TBL（規則有排序）[16]	90.51%	91.81%	76.84%	93.78%	89.39%	94.19%
本論文方法	93.99%	94.44%	77.82%	95.11%	96.35%	95.99%

五、結論

本論文利用關聯式規則來預測台語文轉音系統中一詞多音現象，並針對關聯式分類實驗進行參數上的調整，同時藉由 E-Hownet 來追加詞義特徵，以避免中文詞作為特徵時容易出現的資料稀疏問題。

從實驗結果能夠說明，在常見的『上』、『下』、『不』、『你』、『我』、『他』等多音詞，我們提出的方法擁有更好的預測正確率。儘管如此，將關聯式規則應用在一詞多音預測上仍然有許多進步的空間。尤其是面對一個多音詞能夠通用於兩種讀音的情況，在現有的實驗結果中，依然容易出現誤判的情況。未來能夠針對現有的台語一詞多音預測模組的系統架構作出這方面的修正，使訓練模型能夠處理讀音通用的情況。

參考文獻

[1] B. Liu, W. Hsu and Y. Ma, "Integrating classification and association rule mining",

Knowledge Discovery and Data Mining, pp. 86, 80, 1998.

[2] Brill, E. "Automatic grammar induction and parsing free text: A transformation-based approach." *In Proceedings of the 31st Meeting of the Association for Computational Linguistics*. Columbus, Ohio, USA, pp. 259-265, 1993.

[3] C. H. Hwang, "Text to Pronunciation Conversion in Taiwanese", *Master thesis, Institute of Statistics, National Tsing Hua University*, 1996.

[4] C. Shih and R. Sproat, "Issues in Text-to-Speech Conversion for Mandarin", *Computational Linguistics and Chinese Language Processing*, Vol., 1, No. 1, pp. 37-86, 1996.

[5] Fadi Thabtah, "A review of associative classification mining." , *Knowledge Engineering Review*, Vol. 22, pp. 37-65, 2007.

[6] J. Y. Huang, "Implementation of Tone Sandhi Rules and Tagger for Taiwanese TTS", *Master thesis, Department of Communication Engineering, National Chiao Tung University*, 2001.

[7] M. S. Yu, T. Y. Chang, T. H. Hsu, and Y. H. Tsai, "A Mandarin Text-to-Speech System Using Prosodic Hierarchy and a Large Number of Words", *Proceedings of the 17th Conference on Computational Linguistics and Speech Processing, (ROCLING XVII)*, pp. 183-202, 2005.

[8] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules." Presented at *Proc. 20th Int. Conf. Very Large Data Bases*, VLDB, 1994.

[9] S. H. Chen, S. H. Hwang, and Y. R. Wang, "A Mandarin Text-to-Speech System", *Computational Linguistics and Chinese Language Processing*, Vol. 1, No. 1, pp. 87-100, 1996.

[10] Y. J. Lin, M. S. Yu, and C. J. Huang, "The Polysemy Problems, An Important Issue in a Chinese to Taiwanese TTS System", *Proceedings of the 2008 International Congress on Image and Signal Processing* , Paper number P1234, May 28-30, Sanya, China, 2008.

[11] Y. J. Lin, M. S. Yu, and C. Y. Lin, "Using Chi-Square Automatic Interaction Detector to Solve the Polysemy Problems in a Chinese to Taiwanese TTS System" , in *Proceeding of The 8th International Conference on Intelligent Systems Design and Applications* (*ISDA* 2008), pp.362-367, 2008.

[12] Y. J. Lin, W. S. Ji, M. S. Yu, and S. D. Lee, "Some Important Issues on Text Analysis in a Chinese to Taiwanese TTS System", *Proceedings of the 9th IEEE International Workshop on Cellular Neural Networks and their Applications* , 2005.

[13] Y. J. Lin, M. S. Yu, C. Y. Lin, and Y. C. Lin, "A Layered Approach to the Polysemy Problem in a Chinese to Taiwanese TTS System, " *Journal of Information Science and Engineering*, Vol. 26, No. 5, 2010.

[14] 中央研究院平衡語料庫 http://db1x.sinica.edu.tw/cgi-bin/kiwi/mkiwi/mkiwi.sh

[15] 林金玉，"中文轉台語文轉音系統中一詞多音之預測"，國立中興大學資訊科學與工

程學系碩士論文，2008。

[16] 楊文德，”利用規則轉換學習方法解決台語一詞多音之歧義”，國立中興大學資訊科學與工程學系，2012。

[17] 董振東、董強，知網(HowNet)，http://www.keenage.com/

[18] 廣義知網知識本體架構(Extended-HowNet Ontology) http://ckip.iis.sinica.edu.tw/taxonomy/

[19] 廣義知網知識本體架構線上瀏覽系統 http://ehownet.iis.sinica.edu.tw/

[20] 鍾祥睿，”台語 TTS 系統之改進”，國立交通大學電信工程系所碩士論文，2002。