101 年度客家委員會補助學術研究

結案報告

以資料探勘理論自動解決客語多音字岐異問題及其應用之研究

計劃主持人：黃豐隆博士 (國立聯合大學)

中華民國 101 年 11 月

本計劃承蒙客家委員會補助得以順利完成，

特此感謝!

第一章 簡 介

1.1 研究動機

電子書、語言翻譯機和語言學習系統等完善的語音合成系統不僅可以帶給我們生活極大的便利，對於弱勢文化的保存也有幫助，因為我們可以利用這些工具去學習不熟悉的語言。目前台灣客家族群約占總人口的七分之一，但是最大的通病是有的客家人本身並不太會講客家話，甚至聽不懂。方言消失即代表文化的傳承有可能會有斷層，甚至是有消失的危機，因此客語的永續傳承是相當重要的。

近年來大家越來越重視方言的問題，因此方言教學與認證的資訊越來越多，又因為網路的盛行，因此許多單位紛紛把教學資源建立成網站，但是以閩南語居多，例如互動功能的有聲詞典網站或線上語音合成系統，而客語教學以靜態的文章居多，較具有公信力的則以為主。我們的語音合成系統目的是讓使用者自行輸入中文文句，進而轉換出對應的客語詞及發音，合成出流利且自然的語音，可以讓原本不會客語的使用者做客語的語言學習。

1.2 數位學習簡介

數位學習(E-Learning)是指經由數位媒介，例如網路或多媒體來進行學習，可以是很正式的學習，例如學業上的課程，也可以是很輕鬆的學習，例如專長興趣的課程。透過網際網路，即時傳遞各種資訊和知識，E-Learning 的 E 除了代表「Electronic 電子的」，同時也包含了下列的其它意涵[26]：

1.3 研究方向

本研究的重點放在客語多音字的判斷上。從國語文句中萃取多音字特徵，使用資料探勘(Data Mining)的方法，分別以 C4.5 決策樹(Decision tree)、貝氏網路分類器、CART 決策樹以及 SVM 分類器預估多音字的類型。

第二章 多音字介紹及研究方法

2.1 客語多音字介紹

在國語中，常常會遇到一些多音字的例子，例如「隊長」的長這時發音做ㄓㄤˇ，而「長度」的長則發音為ㄔㄤˊ。而客語中同樣也會發生類似的情況，有的是國語發音和客語發音都含有多音字的現象，例如「感覺」和「睡覺」的「覺」這個字，國語發音分別為「ㄐㄩㄝˊ」和「ㄐㄧㄠˋ」，客語發音則為「gog•」和「gau」。有的則是國語發音皆一致，但是客語發音有多音字的情形，例如「屋頂」和「頂天立地」的頂這個字，國語發音皆為「ㄉㄧㄥˇ」，客語發音則分別為「dangˋ」和「dinˋ」。

2.2 監督式與非監督式方法

監督式學習是一個機器學習中的技巧，可以由訓練資料中學到或建立一個模型，並依此模型推測新的資料。訓練資料是由輸入物件和預期輸出所組成。函數的輸出可以是一個連續的值(稱為迴歸分析)，或是預測一個「分類」。

非監督式學習則不需要預先定義好的分類以及標記好的資料，不針對任何屬性學習及設計模型。「分群」就是屬於非監督式的學習，目的在於具及高相似性的資料，使得群體與群體之間產生較低的相似度。有別於監督式學習網路，無監督式學習網路在學習時並不知道其分群結果是否正確，亦即沒有告訴它何種學習是正確的，當學習完畢並經測試後，也可以將之應用到新的資料上。

資料探勘可從形式不拘的大量資料中，發掘隱藏有用的資訊，經過分析之後得到有用的知識(Knowledge)。一般而言，Data Mining包含下列五項功能：

(1)分類(classification)

(2)推估(estimation)

(3)預測(prediction)

(4)關聯分組(affinity grouping)

(5)同質分組(clustering)

本計劃用「分類」功能解決多音字問題，是屬於監督式學習的方法。「分類」是根據已知資料及其分類屬性，建立資料的分類模型，再以此分類模型對欲分類的新對象預估出類別(class)。現在有著各式各樣的分類器，各自都有強項或弱項，分類器的表現很大程度上跟要被分類的資料特性有關， 並沒有某一單一分類器可以在所有給定的問題上都表現最好。

2.3 貝氏網路分類器

貝氏網路是以貝氏定理為基礎所發展出來的理論，是一種機率圖型模型，能將特定領域中的不確定性組合成模型，以供推論與診斷之用，利用事件發生的機率來推測未知資料的類別。由於機率的計算是可隨著已知樣本的增加而逐次調整的，在新樣本加入時只需局部調整某些機率值，即可得到新的分類模型，其分類模型的機動性相當高，在資料不斷增加的情況下，可以得到較好的分類效能。

藉由有向非循環圖型中得知一組隨機變數$\{X_1, X_2, \ldots X_n\}$及其 n 組條件機率分配(Conditional Probability Distributions)的性質。圖形中每個節點會包含相關連的變數的機率值，並且由父點來決定下一個子點的相關機率值，如此可以很簡單的去合併一個新的資料到貝氏網路中，並且維持貝氏網路的資料始終是最新的。

2.4 C4.5 決策樹與 CART 決策樹

決策樹是一個預測模型，用來處理類別型或連續型變數的分類預測問題，屬於監督式的學習方法，採用自頂端而下的貪婪演算法(Greedy Algorithm)。它代表的是對象屬性與對象值之間的一種映射關係，主幹是類似流程圖的樹架構，每個內部節點代表一個測試屬性，分支代表一個可能的測試輸出結果，最底層的樹葉節點則代表不同分類的類別標記。所以決策樹的主要功能是藉由分類已知的資料來建立一個樹狀結構，並從中歸納出一些規則。回到本研究，我們要分類的多音字可以套用在決策樹模型上，來幫助我們區分出該發音的類型，故可用決策樹方法分類。

CART(Classification and Regression Trees) 演算法是建構決策樹

時最常用的演算法之一，1984 年布里曼 （L. Brieman）與其同僚發表這種方法，是一種二元(binary)分割的方法。CART 以每個節點的動態臨界值做為判斷式，藉由單一輸入的變數函數，在每個節點分隔資料，並建立一個二元決策樹，根據資料的分類數及其屬性並依據 Gini 規則來決定分割條件。使用 Gini Ratio 來衡量指標，如果分散的指標程度很高，表示資料中分佈許多類別，相反的，如果指標程度越低，則代表單一類別的成員居多。一般來說，決策樹分類的正確性有賴於資料來源的多寡，若是透過龐大資料建構的決策樹，其預測和分類結果往往是符合期望的。

C4.5 則是改良自 ID3 (Iterative Dichotomizer 3, J. Ross Quinlan 於 1979 年所提出)演算法，主要是先建構一顆完整的決策樹，再針對每一個內部節點，依使用者定義的預估錯誤率(Predicted Error Rate)來做決策樹修剪的動作。而不同的節點之特徵值離散化結果是不相同的。C4.5 與 CART 之間的第一個差異是 CART 在每一個節點都呈現二分法，因此產生二元決策樹，而 C4.5 則在每一個節點產生不同數目的分支。這是因為 C4.5 對持續性變項的處理方式和 CART 相當類似，但對類別變項的處理就相當不同，根據類別值來選擇最好的屬性進行區別，屬性選擇指標使用資訊獲利（Information Gain)。下列是資訊獲利的算法。

期望訊息用於判別 D 中值組：

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (2\text{-}1)$$

其中 p_i 為任一 D 中的值組被歸類為 C_i 的機率，可由 $\frac{|C_{i,D}|}{|D|}$ 得之，在分割後，為了達到一致的判別我們需要下列的訊息：

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j) \qquad (2\text{-}2)$$

分割屬性A的資訊獲利：

$$Gain(A) = Info(D) - Info_A(D) \qquad (2\text{-}3)$$

另外一個差異則是決策樹的修剪，CART 使用決策樹的分散度為度量，來標記不同的分支樹，然後以沒有見過的預先分類好的資料（測試組）來測試這些分支樹。相反的，C4.5 並不參考其他資料，嘗試以只用訓練資料的情況下來修剪決策樹。

第三章 實驗分析與結果

我們建置了一個線上客語語音合成系統，開發環境與使用之軟體如**錯誤！找不到參照來源。**所示。

3.1 資料庫

Database Schema 如圖十一，haka2 資料表是國客語對照詞典，allspell2 資料表是客語四縣腔單字讀音表，allsealand2 資料表是客語海陸腔單字讀音表，multiwavname 資料表是四縣腔客語詞與其音檔檔名的對照表，sl_multiwavname 資料表是海陸腔客語詞與其音檔檔名的對照表。

3.2 文句分析

本系統支援的輸入為國語中文文句。假設使用者輸入一句中文文句「中興大學建國百年畢業典禮即將於六月舉行」，這個文句先送入中文斷詞器做斷詞與詞性標記，斷完詞的結果為:「中興大學 建國 百年 畢業典禮 即將 於 六月 舉行」，共有 8 個中文詞，詞性串列為：「Nc,Nc,Nd,Na,D,P,Nd,Vc」。詞與詞性的對照如(E3)：

中興大學 建國 百年 畢業典禮 即將 於 六月 舉行 （E1）

Nc Nc Nd Na D P Nd Vc

來把斷出的中文詞各自到國客語對照詞典查出對應的客語詞並做連音變調，像這句中的「即將」所對應的客語詞「臨將」，此時系

統會列出所有找到的客語詞。若在國客語對照詞典中找不到詞，可能的原因有：

(1) 國語斷詞的輸出詞太長，客語詞典中只有此詞的部分資訊。舉個簡單的例子，本系統的中文斷詞器會把「中興大學」當成一個詞輸出，但是在客語詞典中，「中興」與「大學」卻是兩個詞，造成找不到正確的對應客語詞，且無法使用客語詞的音檔。當這種情況發生時，我們把長度三個字以上的詞取出，使用國客語對照詞典中的國語詞欄位，用長詞優先斷詞，把此詞拆成較短的詞，即可把中文詞「中興大學 」轉成對應的客語詞「中興」「大學」。

(2) 客語詞典中沒有此國語詞的資訊。此時只能把此國語詞拆成單字，去查找單字拼音表，找出單字的讀音。

(3)上述過程可以簡化為**錯誤! 找不到參照來源。**十二的流程圖。找出客語詞之後，把客語詞列出，如：「中興大學 建國 百年 畢業典禮 臨將 於 六月 舉行」，並且同時列出這些客語詞的客語讀音，其中「中興大學」的「中」已經變調過了，如(E4)：

中興大學 建國 百年 畢業典禮 臨將 於

zungˇ hin↗ tai hog gien gued bag↖ ngienˇ bid↖ ngiab dien↖ li↗ limˇ ziong↗ iˇ

六月 舉行

liug↖ ngied gi↖ hi (E2)

3.3 實驗方法

進行多音字實驗前需要標記訓練語料，我們所使用的語料是 Chinese GigaWord 語料庫[33]，來源是由許多個單位收集新聞語料而來，並且由 Linguistic Data Consortium 發行，授權讓人取得使用。目前我們所使用的版本為 2007 年所發行的第三版，其新聞語料主要有以下三個來源：Central News Agency of Taiwan、 Xinhua News Agency of Beijing 和 Zaobao Newspaper of Singapore。

我們所使用的語料要以繁體中文為主，所以選擇從台灣中央通訊社(Central News Agency)所收集而來的語料。在此份語料中，含有約五千多萬行句子，包含詞性標記的檔案大小約 6GB。語料庫中的文

章都是斷好詞，並且標上詞性標記的語料，其詞性標記是由台灣中央研究院所標記。

而我們也有採用中文結構樹資料庫(TreeBank)的資訊，其是由 86 年起由中央研究院詞庫小組(CKIP)從中央研究院現代漢語平衡語料庫(Sinica Corpus)中抽取句子，經由電腦自動剖析成結構樹，再加以人工修正、檢驗後所得的成果。目前版本是 3.0 版，包含了 6 個檔案，61,087 個中文樹圖，361,834 個詞。中文句子的語法結構表達採取中心語主導原則(Head-Driven Principle)，剖析中文句子時，詞組類型由中心語決定，並且參照中心語和其他成分所記載的語法和語意訊息，表達出句子中詞和詞之間的語法結構和語意角色關係。

表一：多音字的類型、發音及語料數目

多音字	四縣腔發音類型	海陸腔發音類型
行	hang∨、hong∨、hen	hang、hong、hen＋
調	tiau∨、tiau、diau	tiau、tiau＋、diau∨
重	cung、ciung∨、cung↙	chung＋、chung、chung↘
差	ca↙、cai↙	ca↘、cai↘
易	id˚、i	rhid˚、rhi∨
口	kieu↘、heu↘	kieu↙、heu↙

	訓練語料	測試語料	總數
行	4961 筆	895 筆	5856 筆
調	2426 筆	389 筆	2815 筆
重	4336 筆	383 筆	4719 筆
差	730 筆	186 筆	916 筆
易	546 筆	237 筆	783 筆
口	3201 筆	387 筆	3588 筆

從多音字的目標詞的詞性組合開始實驗，且慢慢增加特徵數，從前一詞、後一詞與目標詞的詞性組合，前二詞、後二詞與目標詞的詞性組合，前三詞、後三詞與目標詞的詞長詞性組合…等。總共實驗了

以下特徵：目標詞與後一詞可以生成的詞性、目標詞與前一詞可以生成的詞性、目標詞與前一詞跟後一詞可以生成的詞性、目標詞前第一詞詞性、目標詞前第二詞詞性、目標詞前第三詞詞性、目標詞後第一詞詞性、目標詞後第二詞詞性、目標詞後第三詞詞性、目標詞詞性。以例子「我 今天 早上 去 學校 差點 遲到」解說，在「差點」這詞裡面含有多音字「差」:

我　今天　早上　去　學校　差點　遲到　(E4)

Nh　Nd　Nd　VCL　Nc　D　VH

以目標詞「差點」為例，記錄下需要的特徵，需注意的是當遇到句首沒有詞的時候，則詞性標記為「Begin」;當遇到句尾沒有詞的時候，則詞性標記為「End」:

該目標詞前第一詞詞性： Nc
該目標詞前第二詞詞性： VCL
該目標詞前第三詞詞性：Nd
該目標詞後第一詞詞性：VH
該目標詞後第二詞詞性：END
該目標詞後第三詞詞性：END
目標詞與前一詞可以生成的語法類別(取 TreeBank 頻率最高者)：NULL
目標詞與後一詞可以生成的語法類別(取 TreeBank 頻率最高者)：VP
目標詞與前一詞跟後一詞可以生成的語法類別(取 TreeBank 頻率最高者)：NULL
目標詞詞性：D

3.4 實驗的結果

我們依序將特徵從目標詞性開始，慢慢增加特徵的數目，內部測試、外部測試、訓練語料與測試語料一起測試(全部當訓練)、訓練語料與測試語料一起測試(取 66%當訓練語料，其餘當測試語料)這四種測試是最終的實驗結果，結果發現，不同多音字分類器正確率最高的特徵組合幾乎皆是採用 C4.5 演算法,且選取特徵為「前三後三詞性、目標詞與後一詞可以生成的語法類別、目標詞與前一詞可以生成的語法類別、目標詞與前一詞跟後一詞可以生成的語法類別、目標詞性」

的時候；只有在「行」這個字時，是 CART 演算法的正確率較佳，不過跟 C4.5 演算法的正確率也相去不遠。

在修改特徵組合的過程中，可以看出目標詞性，對於正確率有相當大的影響，而再加上目標詞前後的詞性以及和前後詞可以生成的詞性，能使預測的正確率上升。整體而言，預測多音字的正確率都有80%以上。但是要注意的是，如果當某多音字的某個發音，可以選出的訓練語料太少時，將可能因為資料稀疏，使得訓練出來的分類器無法有效分類未知的語料，導致分類器的正確率降低。

本計劃已分析這六個客語多音字實驗的數據，惟本結案報告只列出多音字「**行**」與「調」二字之實驗數據。其中表格上層的 I. 內部測試、II. 外部測試、III. 訓練語料與測試語料一起測試(全部當訓練）、IV. 訓練語料與測試語料一起測試(取 66%當訓練語料，其餘當測試語料)這四個實驗使用特徵為「目標詞前第一詞詞性、目標詞前第二詞詞性、目標詞前第三詞詞性、目標詞後第一詞詞性、目標詞後第二詞詞性、目標詞後第三詞詞性、目標詞與後一詞可以生成的語法類別、目標詞與前一詞可以生成的語法類別、目標詞與前一詞跟後一詞可以生成的語法類別、目標詞性」。

表二：C4.5 決策樹於前三後三詞性、A、B、C、目標詞性，多音字「**行**」的各發音正確率

	Precision	**Recall**	**F-Measure**
hangˇ	0.92	0.958	0.939
hen	0.452	0.322	0.376
hongˇ	0.922	0.847	0.882
加權平均	0.9	0.906	0.902

(A:目標詞與後一詞可以生成的語法類別 B:目標詞與前一詞可以生成的語法類別 C:目標詞與前一詞跟後一詞可以生成的語法類別)

表三：CART 決策樹於前二後二詞性、目標詞性，多音字「行」的 confusion matrix

實際 \ 預測	hangˇ	Hen	hongˇ
hangˇ	1407	23	44
Hen	55	32	0
hongˇ	50	4	376

(A:目標詞與後一詞可以生成的語法類別 B:目標詞與前一詞可以生成的語法類別 C:目標詞與前一詞跟後一詞可以生成的語法類別)

表四：CART 決策樹於前二後二詞性、目標詞性，多音字「行」的各發音正確率

	Precision	**Recall**	**F-Measure**
hangˇ	0.931	0.955	0.942
Hen	0.542	0.368	0.438
hongˇ	0.895	0.874	0.885
加權平均	0.906	0.912	0.908

(A:目標詞與後一詞可以生成的語法類別 B:目標詞與前一詞可以生成的語法類別 C:目標詞與前一詞跟後一詞可以生成的語法類別)。

表五：「調」的各種特徵組合的正確率(門檻值 60%)

	C4.5	貝氏網路	CART	SVM
I.內部測試	86.8096	83.4295	88.8293	99.3405
II.外部測試	86.6075	86.6945	86.3233	73.1439
III.訓練語料與測試語料一起測試(全部當訓練)	88.881	85.7889	90.9414	99.2895
IV.訓練語料與測試語料一起測試(取 66%當訓練語料，其餘當測試語料)	86.2069	85.5062	85.8934	70.8464
特徵類型改變比較				
A+B+C、前三後三詞性(套用 III)	79.5737	74.1030	85.8615	98.0462
A+B+C、前三後三詞性(套用 IV)	72.5183	72.9363	75.4441	69.5925
C、前三後三詞性、目標詞性(套用 III)	88.8810	86.2877	91.0835	99.2895
C、前三後三詞性、目標詞性(套用 IV)	86.2069	84.117	84.6395	70.4284
前三後三詞性、目標詞性(套用 III)	88.3481	85.6838	91.4032	99.2540
前三後三詞性、目標詞性(套用 IV)	86.1024	83.6991	85.0575	70.3239
前二後二詞性、目標詞性(套用 III)	87.389	84.7247	91.4032	96.9499
前二後二詞性、目標詞性(套用 IV)	86.1024	83.908	86.2069	73.9812
前一後一詞性、目標詞性(套用 III)	87.9218	84.6181	88.8009	90.4085
前一後一詞性、目標詞性(套用 IV)	86.1024	84.117	85.7889	82.1317
目標詞性(套用 III)	77.6199	77.6199	77.6199	77.6199
目標詞性(套用 IV)	78.0564	78.2654	78.2654	78.0564

(A:目標詞與後一詞可以生成的語法類別 B:目標詞與前一詞可以生成的語法類別 C:目標詞與前一詞跟後一詞可以生成的語法類別 門檻值:在取出的語料中發音最多個數佔總發音個數之百分比)(單位：%)

第四章 結論

本研究的重點係在客語多音字的語義岐異上面，我們從國語文句中萃取多音字特徵，使用資料探勘(Data Mining)的方法，分別以 C4.5 決策樹(Decision tree)、貝氏網路分類器、CART 決策樹以及 SVM 分類器預估多音字的類型。

本研究速角步探討網際網路上建置客語語音合成系統，我們的系統包含有三個模組：文句分析、律韻訊息分析與語音合成，並建立相關之語料庫與語音庫。我們錄製客語單字之基本語音合成單元與不同時長之靜音檔，配合韻律參數作為合成之用。為改善所錄製合成單元之能量差異，對所有單元進行能量正規化處理；並處理有多音字的客語詞，接著找出客語詞對應的音檔，並且加入適當的停頓，以提升輸出的語音品質。

語音合成中的串接法合出的客語連續語音，會在遇到入聲字時產生發音急促的現象，本研究將入聲字音檔的長度補足至特定長度，以解決此問題。此外當遇到含有「行」、「調」、「重」、「差」、「易」、「口」這幾個多音字的詞且詞典無此詞的音檔時，則經由多音字分類器模組，根據輸入文句的特徵，選出一個此時發音機率最高的發音，而回傳給系統的語音合成模組。(註:本結案報告只列出「行」與「調」字之實驗結果)

系統由 11 位熟悉客語之人士進行 10 分等級之聽測實驗(MOS)，進行合成語音品質之自然度分析。聽測實驗之統計結果顯示，合成語音的自然度分數已達 7.9 分，證明合成語音具有自然度。

目前，我們實作完成之客語語音合成系統已可供線上使用，透過使用者自行輸入中文文句，合成出對應客語語音，包含多音字之岐異在內。使用者只要連上網路，就可線上立即聽取並學習，每一詞彙均標示出拼音，以便於學習。我們所建置的系統符合數位學習的精神，達到幫助使用者學習客語的目標，期望能對推廣客語貢獻一份心力。

最後，本計劃承蒙客家委員會補助得以順利完成，特此感謝!