使用關聯法則為主之語言模型於擷取長距離中文文字關聯性

Association Rule Based Language Models for Discovering Long Distance Dependency in Chinese

簡仁宗　陳鴻儀

國立成功大學資訊工程學系

Email：jtchien@mail.ncku.edu.tw

摘要

本論文提出一種能擷取長距離資訊的語言模型，它可以擷取多詞彙之間的關聯性，擷取的方式是使用資料探勘中十分流行的 Apriori 演算法，傳統上 n-gram 語言模型只能在 n-gram 視窗內擷取到有限距離的資訊，較長距離的資訊也就因此而流失，然而這些失去的長距離資訊對於語言模型是十分重要的，所以如何克服 n-gram 模型缺乏長距離資訊一直是非常熱門的研究課題，觸發序對就是其中一種有效的方法，其主要功能是在擷取長距離之詞序對資訊，也就是建立起詞與詞之間的關聯性，然而我們所提出的關聯法則技術能擷取多元詞組間的關聯性，可以說是進一步改良詞組數並建立更長距離資訊，而實驗結果也顯示本論文方法比起傳統觸發序對獲得較低的 perplexity，此關聯法則技術也可以有效的與其他模型調整及模型平滑化的技術結合，在語言模型的效率改善方面能有更良好的效果，最後本論文也將提出的語言模型成功的應用在語音辨識與文件分類上，並建立一套個人化之新聞瀏覽器之展示系統。

1. 簡介

拜硬體技術不斷進步的貢獻之下，，一般人會很理所當然的使用自動櫃員機提款或是利用自動空調設備來控制室內的溫度，而這都是由於電腦的自動化管理讓生活變的如此便利，正所謂“科技始終來自人性”，推動科技進步的那隻幕後

的黑手就是建立在“使人便利”的基礎之上，但是電腦自從在發明之初就存在一個與人性背道而馳的缺點，與它們的溝通需要透過一個特定的按鍵裝置，比方說要與個人電腦溝通就必須透過鍵盤或滑鼠等裝置，事實上這是使許多人對電腦望之卻步的原因，要學習如何使用鍵盤與電腦做溝通就等於是強迫人去學習一種“電腦語言”，這與“使人便利”的原則當然是互相違背的，但是反過來說如能讓電腦學習人類的語言，使電腦能更接近人類，也就能使其與人類生活的結合更加緊密，進一步如果電腦能透過語言的學習而具備了閱讀的能力，我們就可以讓電腦為我們過濾亦或分類每天所需閱讀的文件，比方說可以應用在於 e-mail 廣告過濾或是新聞文件分類等等，就可以讓電腦為我們省下更多的時間。

要克服電腦與人在語言上的鴻溝，在語言技術的領域有了聲學模型(acoustic model)與自然語言模型(natural language model)的產生，而這兩項技術的發展在國外已經行之有年，台灣自西元一九八二年起便開始有了中文聲學模型方面的研究，許多研究單位包括台清交成等大專院校，以及工研院、交通部、中研院、中華電信等都積極的投入研究的工作並且擁有了十分豐碩的研究成果，而在聲學模型已日益成熟的基礎下，自然語言模型的發展也備受矚目，誠如前文所述，語音技術發展的最終目的就是要將電腦與人類的溝通便利化，而要達到這個目的，將語音模型與自然語言模型做結合是必須的，我們的論文主要就是著墨於自然語言模型的探討，我們將會對自然語言模型中的一項十分成功且廣泛運用的技術 n-gram 語言模型做介紹，並且分析其在傳統上的缺點與改進技術，而本論文也將會針對 n-gram 模型其中一項缺點-長距離資訊的缺乏，提出一套新的改進方法，並且結合其他改進方法，進而發展出一套較有效率的 n-gram 模型，我們將會將其應用在結合聲學模型做語音辨識和文件分類的領域之上，期望對其正確率有一定幅度的改善。

而自然語言模型方面在現今有許多不同的發展，依其內容主要分為三個方向，一、根據語言學所發展出的文法(grammar)分析，二、以知識為基礎而發展的語言資料庫，三、著重於統計而發展出的 n-gram 模型。而我們主要是著墨於統計式的 n-gram 模型，在第二章中，我們將對 n-gram 模型做詳細的介紹，並對其缺點加以探討，第三章中將會介紹傳統上針對 n-gram 模型的缺點所衍生出的改進方法，並且提出一種能擷取長距離資訊的語言模型，將它應用在語音辨識或

新聞文件分類的系統上有一定幅度的幫助。

2. n-gram 語言模型簡介

目前 n-gram[11]模型的探討於各相關學術會議及期刊論文上已有相當多的文獻發表，顯示各種研究機構對此一領域的發展有相當大的期許，故投身於其中，而在各方都致力於改進 n-gram 模型之下，n-gram 模型在效能上已獲得相當不錯之成果，在本章中我們將會對 n-gram 模型的基本概念做一簡單之介紹。

2.1 n-gram 模型之應用

一般而言 n-gram 語言模型通常應用於貝式分類器(Bayes classifier)，扮演著事前機率(priori probability)或是可能性(likelihood)的角色，以語音辨識為例子而言，假設有一段聲學訊號(acoustic signal)X，我們的目標是去找尋出此訊號最有可能的對應文句(sentence) S^*，使用貝式分類架構是找出最佳事後機率的文句

$$S^* = \arg\max_S P(S \mid X) = \arg\max_S P(X \mid S) \cdot P(S) \tag{1}$$

其中 n-gram 模型 $P(S)$扮演著事前機率的角色，透過聲學模型計算可獲得一段文句的聲學模型分數 $P(X|S)$，再透過語言模型計算可獲得此文句的語言模型分數 $P(S)$，將兩機率相乘求得最佳化之文句，即為此聲學訊號最有可能之對應文句。

就文件分類的領域而言，給定一篇文件 d，目標是去找尋此篇文件所屬的類別 c (category)，假設我們總共定義了 k 個類別，並且使用這些類別所屬的文件訓練好不同類別的 n-gram 模型 L_1、L_2、……、L_k， 使用貝式分類器求得此篇文件所屬的類別 c^* 可寫成

$$c^* = \arg\max_c P(c \mid d) = \arg\max_c P(d \mid c) \cdot P(c) \tag{2}$$

在這邊 n-gram 模型 $P(d|c)$扮演的是可能性量測的角色，透過語言模型機率計算可以獲得 $P(d|c)$ 的值，而假設所有類別出現的機率是均等，只要能使 $P(d|c)$最佳化的類別語言模型，即為此文件 d 為最有可能對應之類別。

2.2 n-gram 模型之建立

語言模型主要的功能是在評估一段文句出現的機率，假設有一文句 S 其長度為 T 並且是由一段詞序列 $W_1W_2W_3...W_T$ 所組成，則 S 出現的機率可以寫成

$$
\begin{aligned}
P(S) &= P(W_1, W_2, ..., W_T) \\
&= P(W_1)P(W_2 \mid W_1).....P(W_T \mid W_1, W_2, ..., W_{T-1}) \\
&= \prod_{i=1}^{T} P(W_i \mid W_1 W_2, ..., W_{i-1})
\end{aligned} \tag{3}
$$

但是此種方法在計每一個詞的條件機率時都要牽涉到前面所有的詞序列，使得計算量太大而無法實現，為解決這個問題所以有 n-gram 模型的產生，在 n-gram 模型中，它是假設一個詞出現的機率只跟前面 n-1 個詞有關，因此(3)式可以近似為

$$
P(S) = P(W_1, W_2, W_3, ..., W_T) \cong \prod_{i=1}^{T} P(W_i \mid W_{i-n+1}^{i-1}) \tag{4}
$$

其中 W_{i-n+1}^{i-1} 代表 $W_{i-n+1} W_{i-n+2} ..... W_{i-1}$ 詞序列，如此一來使用 n-gram 可以大量節省計算時間與記憶體，讓實用性大為提高。而一般在建立 n-gram 機率模型 $P(W_i \mid W_{i-n+1}^{i-1})$ 最直覺的方法就是統計在詞序列 $W_{i-n+1} W_{i-n+2} ..... W_{i-1}$ 後出現 W_i 的次數再除以詞序列 $W_{i-n+1} W_{i-n+2} ..... W_{i-1}$ 在訓練文集中出現的次數，也就是

$$
P(W_i \mid W_{i-n+1}^{i-1}) = \frac{C(W_{i-n+1}^{i})}{C(W_{i-n+1}^{i-1})} = \frac{C(W_{i-n+1}^{i})}{\sum_{W_i} C(W_{i-n+1}^{i})} \tag{5}
$$

其中 $C(W_i^j)$ 代表 W_i^j 在訓練文集中出現的次數。

2.3 n-gram 模型之評估

基本上在評估一個 n-gram 模型的效果時常使用 perplexity[12]這個評估標準，而事實上它是在計算機率模型的 entropy，entropy 在訊息理論上指的是將機率 P 乘以資訊-$logP$，應用在 n-gram 模型的評估則表示為:

$$
\begin{aligned}
H_p &= -P \log P \\
&= -\lim_{Q \to \infty} \frac{1}{Q} \sum_{W_1 W_2 ... W_Q} P(W_1 W_2 ... W_Q) \log P(W_1 W_2 ... W_Q)
\end{aligned} \tag{6}
$$

其物理意義表示在計算一個 n-gram 模型的 entropy 時，必須先將詞典中的詞做組合，形成為無限長的詞序列 $W_1 W_2 ..... W_Q$，並且將所有的可能詞序列計算其機率與資訊的乘積後加總，即可得到此 n-gram 模型的 entropy。但事實上不容易實現如此複雜的計算，必須假設可以提供一段足夠長的詞序列來代表所有的詞序列組

合，這種假設在統計學上稱為此詞序列為 ergodic，故(8)式可改寫為

$$H_p = -(\frac{1}{Q})\log P(W_1W_2.....W_Q) \tag{7}$$

而 perplexity 的定義為

$$perplexity = 2^{H_p} \tag{8}$$

perplexity 代表了 n-gram 模型中的平均分支因數(average branching factor)，perplexity 越低代表 n-gram 模型在做機率評估時，所遇到的分支越少，也就是此模型的效率越好。

2.4 n-gram 模型的缺點

n-gram 模型長久已來就存在著三個重要的問題，也是研究 n-gram 模型的人一直努力的目標，我們分述如下:

1. *訓練文集與測試文集領域上之差距(domain mismatch)* :

n-gram 模型在建立時，必須要有一訓練文集來統計出此模型的機率，因此 n-gram 模型受制於它的訓練文集，當訓練文集不平均時可能會使 n-gram 模型較偏向某種領域(domain)，假設我們的訓練文集是財經類的新聞，但是此 n-gram 模型的目的是用來測試政治新聞，那麼就會造成較大的誤差，在這方面通常會使用較一般化的的平衡文集作為訓練文集來解決這個問題。但矛盾的是如果我們使用較為平衡的文集訓練出我們的 n-gram 模型，此 n-gram 模型用來測試某些特定領域的新聞是否恰當?事實上我們希望在測試政治新聞時我們的 n-gram 模型是偏向政治類的，測試財經新聞時 n-gram 模型是偏向財經類的，為了要完成這項需求，就必須對 n-gram 模型再做改進，使其具有調整之效果[3][4][5][7]。

2. *訓練文集不足(data spareness)* :

n-gram 模型在訓練時，並不能保證訓練文集能夠包含所有詞的組合，以至於所訓練出來的機率模型某些詞組相連的機率為零，或是因為訓練文集的不平衡，造成統計出來的機率模型並不夠一般化，而為了解決這個問題，就有平滑化技術的產生，在參考文獻[2]中對傳統上受歡迎的平滑化技術有詳盡的說明。

3.*長距離資訊(long distance)缺乏之問題*：

n-gram 模型在計算上的優勢是在於它使用了 n-gram 視窗(n-gram window)做為基礎，節省了大量的記憶體與運算時間，但也因為使用了這個概念使得 n-gram 模型只能擷取到視窗之內的資訊，長距離的資訊就因此而流失，而這些流失的資訊很可能會造成 n-gram 模型測試時相當程度的誤差，故如何擷取長距離的資訊一直都是 n-gram 模型中相當受到矚目的研究的課題。一般而言目前 n-gram 模型的研究均以解決此三項問題為主，本論文針對上述第三項長距離資訊的擷取提出改進方法，期望能提昇 n-gram 模型的效果。

3. n-gram 模型改進方向

針對 n-gram 模型的問題已經有許多論文提出改善的方法，在本章中，我們將針對幾組熱門的解決方式做簡介。3.1 節是快取 n-gram 模型(cache n-gram model)與混合式 n-gram 模型(mixture n-gram model)[3]的介紹，此項技術是為了要使 n-gram 模型更符合測試文集領域所發展出來。3.2 節我們將介紹一個在平滑化技術上十分受到歡迎且有效的 Witten-Bell 平滑化技術[13]。3.3 節是對於觸發序對(Trigger pair)[8][9]的簡介，觸發序對是在擷取長距離資訊的一種有效的方法，可以用來補償 n-gram 模型長距離資訊的不足，而本論文也將提出一種改進觸發序對的方法為對照組，並在實驗中做比較研究。

3.1 快取(cache)n-gram 模型與混合式(mixture)n-gram 模型

為了要使 n-gram 模型能夠更符合測試時的領域，所以產生了模型調整的概念，它的概念是基於一篇文章或是一段文句會有一個近似的主題，比方說棒球類的新聞就比較偏向運動類的領域，與其他類別的新聞(如財經新聞)就有一段相當大的差距，而希望能在做測試時，利用文章前面出現的資訊，動態的調整我們的 n-gram 模型，使得我們的模型更能符合我們測試文集的領域，基於此種概念，就有快取模型與混合式模型的技術產生。

快取 n-gram 模型顧名思義就是相同的詞序列會在鄰近的時間點上不斷出現，比方說我們的測試文集是一篇有關金融股的新聞，也就是說此篇文件“金融股”這段詞序列會不斷出現，透過我們的詞典，會將此詞序列斷詞為“金融”與“股”

兩個詞，此時若我們在第一次測試到此詞序列時，將“金融”後面接“股”的機率提高，自然可以增強我們模型的準確性，在快取模型中會保留一塊快取記憶體，而做文件測試時，會將最近測試過的文句拿來訓練出快取 n-gram 模型 P^c 將其與原始的統計模型 P^S 做結合，我們將模型機率用(9)式表示

$$P(W_1W_2.....W_T)=\prod_{i=1}^{T+1}[(1-\mu)P^s(W_i|W_{i-n+1}^{i-1})+\mu P^c(W_i|W_{i-n+1}^{i-1})] \tag{9}$$

其中 μ 代表結合比重。

而在本論文中，我們使用的是文句階層混合式 n-gram 模型(sentence-level mixture n-gram model)，在每經過一文句後，就利用此文句所提供的資訊調整混合模型的比重參數。我們是利用奇摩網站已分類好的新聞，做為我們的分類群組。而我們會依據分群過後之文集訓練出對應於各群組之 n-gram 模型，在這邊以 P_k 表示第 k 個群組的 n-gram 模型，而在做測試時，使用權重 λ_k 將各群之模型做組合成為測試用的 n-gram 模型，也就是說假設有一文句 S 為 $W_1W_2W_3……W_T$，則此文句出現的機率為

$$P(S)=\prod_{i=1}^{T+1}P(W_i|W_{i-n+1}^{i-1})=\prod_{i=1}^{T+1}\sum_{k=1}^{m}\lambda_kP_k(W_i|W_{i-n+1}^{i-1}) \tag{10}$$

其中 m 代表混合數個數，但為此模型還須做兩點改進，第一、為了避免每個群組中的訓練文集太少，造成資料稀疏(data sparseness)，每個單一群組模型需要再結合一個一般化的模型(general model)，用以增加模型的可靠度，第二、在測試時可能會有無領域(nontopic)的文集存在，所以我們又必須將一般化模型加入，視為一個無領域的群組，在此我們將一般化模型以 P_G 表示，故上式可改寫為

$$P(S)=\sum_{k=1}^{m,G}\lambda_k\prod_{i=1}^{T+1}[\alpha_kP_k(W_i|W_{i-n+1}^{i-1})+(1-\alpha_k)P_G(W_i|W_{i-n+1}^{i-1})] \tag{11}$$

其中 α_k 為第 k 個群組模型與一般化模型的組合權重。在混合式 n-gram 模型中，有兩個權重 α_k 及 λ_k 存在，基本上混合式 n-gram 模型是依據前文來動態的調整此二權重，在初始時會使用少數保留文集估測出其初始值，測試時會在每一文句結束時再去做一次權重的調整，而調整的動作可以分別寫成(12)(13)式

$$\alpha_k^{new}=\frac{1}{\sum_{l=1}^{N_k}T_l}\sum_{l=1}^{N_k}\sum_{i=1}^{T_l}\frac{\alpha_k^{old}P_k(W_i|W_{i-n+1}^{i-1})}{\alpha_k^{old}P_k(W_i|W_{i-n+1}^{i-1})+(1-\alpha_k^{old})P_G(W_i|W_{i-n+1}^{i-1})} \tag{12}$$

其中 T_l 代表在文句 l 的詞數，N_k 表示在群組 k 的總文句數。

$$\lambda_k^{new} = \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_k^{old}P_k(W_1,...,W_{T_i})}{\sum_{j=1}^{m,G}\lambda_j^{old}P_j(W_1,...,W_{T_i})} \tag{13}$$

其中 N 代表調整的總文句數。權重的調整的主要根據測試時文件所出現的資訊，混合式 n-gram 模型會依前文在每個群組模型出現的機率為權重，動態的調整測試模型的組合權重，比方說在測試文件中不斷提到金融消息，混合式 n-gram 模型就會將模型逐步的調整到財經領域，再利用這調整過後之 n-gram 模型繼續測試後面的文句，然後再將測試而得的新資訊繼續做調整，這種遞迴式的做法是一種稱為資訊結構(Information structure)的概念。

3.2 Witten-Bell 平滑化技術

在平滑化問題上，我們引進了十分廣泛應用且受到歡迎的 Witten-Bell 平滑化技術[13]做為加強我們 n-gram 模型的基礎，平滑化技術主要建立於將 n-gram 模型中沒有訓練到的詞序列機率模型使用(n-1)-gram 模型做補償，也就是

$$P_{\text{interp}}(W_i \mid W_{i-N+1}^{i-1}) = \lambda_{W_{i-N+1}^{i-1}} P(W_i \mid W_{i-N+1}^{i-1}) + (1-\lambda_{W_{i-N+1}^{i-1}}) p_{\text{interp}}(W_i \mid W_{i-N+2}^{i-1}) \tag{14}$$

這是一個遞迴式的定義，所有的 n-gram 模型都必須利用(n-1)-gram 模型做補償，其中 $\lambda_{W_{i-N+1}^{i-1}}$ 代表的是合併 n-gram 與 (n-1)-gram 之權重，而 Witten-Bell 平滑化技術對此一權重有一個特殊的估測方式，在這邊先對符號做以下的定義

$$N_{1+}(W_{i-n+1}^{i-1}\cdot) = |\{W_i : C(W_{i-n+1}^{i-1}W_i) > 0\}| \tag{15}$$

$N_{1+}(W_{i-n+1}^{i-1}\cdot)$ 代表在 W_{i-n+1}^{i-1} 後可接的詞數，其中下標「1+」代表是連接一個詞以上之意。權重因數定義為

$$1-\lambda_{W_{i-n+1}^{i-1}} = \frac{N_{1+}(W_{i-n+1}^{i-1}\cdot)}{N_{1+}(W_{i-n+1}^{i-1}\cdot) + \sum_{W_i} C\,(W_{i-n+1}^{i})} \tag{16}$$

即為 Witten-Bell 的 n-gram 模型建立方式，其物理意義表示在統計 W_{i-n+1}^{i-1} 出現次數時，如果 W_{i-n+1}^{i-1} 後面可接的詞數越少，我們給 $P(W_i \mid W_{i-n+1}^{i-1})$ 較大的權重，反之則使用較多的(n-1)-gram 做補償，假設在做 bigram 統計時，詞典中有一詞為「類神經」，我們發現在訓練文集中「類神經」後都接「網路」一詞，此時就不需要太多的 unigram 做補償，這是因為此名詞有獨特性，後面幾乎都接少量特定的詞，而若欲統計一詞「幾乎」後可接詞的 bigram 機率，可能會發現訓練文集中其後可接

的詞非常多，此時 unigram 的權重可以適度加大，以彌補可能較多的資訊損失，使語言模型的準確性提高。

3.3 觸發序對 (Trigger Pair) 演算法

在自然語言中，存在著許多高度關聯性的詞組，比方說“醫生、“護士”或是“陽光”、“熱”等就經常出現於於同一句子之中，但由於它們通常在句子中並不相連，所以 n-gram 模型並沒有辦法擷取到這些詞之間的關聯資訊，因此就有了觸發序對的產生，觸發序對的設計主要在於解決長距離資訊彌補 n-gram 模型的不足的問題，觸發序對由於其沒有演算法與資料結構可以快速的對資料庫做求取，故觸發序對會限制本身為“序對”、即若有一辭典 V，觸發序對會對其中所有可能的詞序對做考慮，如此一來可將促發序對的總個數控制於$|V|^2$內。

在統計觸發序對之前，我們必須訂定一個觸發序對的視窗大小，而觸發序對的選取主要是依據平均相互資訊(average mutual information) ，簡稱 AMI，它是用來評估兩個詞 W_i 和 W_j 之間的關聯性大小， AMI 以數學式表示如下

$$AMI(W_i;W_j) = P(W_i,W_j)\log\frac{P(W_i,W_j)}{P(W_i)P(W_j)} + P(W_i,\overline{W_j})\log\frac{P(W_i,\overline{W_j})}{P(W_i)P(\overline{W_j})} + P(\overline{W_i},W_j)\frac{P(\overline{W_i},W_j)}{P(\overline{W_i})P(W_j)} + P(\overline{W_i},\overline{W_j})\frac{P(\overline{W_i},\overline{W_j})}{P(\overline{W_i})P(\overline{W_j})} \tag{17}$$

其中 $P(W,W_j)$ 代表 W_i、W_j 出現在同一視窗的機率，$P(W_i,\overline{W_j})$ 代表在同一個視窗中只出現 W_i,而沒出現 W_j 的機率。透過 AMI 評估標準，我們將其選為觸發序對，以符號($W_i \rightarrow W_j$) 表示。當序對選取完畢後，必須要對每個觸發序對計算其相互資訊 MI (mutual information)，用對數表示之如下

$$MI(W_i;W_j) = \log\frac{P(W_i,W_j)}{P(W_i)P(W_j)} \tag{18}$$

如果 W_i和 W_j是相互獨立的話，則 $MI(W_i, W_j) = 0$，相互資訊反映了觸發序對中兩個詞相互間的資訊變化。而觸發序對並無法單獨使用[14]，因為它只能反映出詞與詞的資訊變化，所以我們必須將其與 unigram 做結合，如此一來所獲得的資訊

比起 n-gram 模型多了長距離的資訊，為了方便起見使用對數表示為

$$\log P(S) = \sum_{i=1}^{T} \log P(W_i) + \sum_{i=T}^{2} \sum_{j=i-1}^{\max(1,i-ws)} MI - Trigger(W_j \to W_i) \tag{19}$$

其中 $\log P(W_i)$ 即為 unigram 模型機率，ws 代表 window size，現在在我們的論文中將 window size 定為文句長度，也就是說在我們論文中的觸發序對是文句階層的觸發序對(sentence-level trigger pair)，代表我們只能擷取同一文句中的觸發序對資訊。現在我們必須將觸發序對加入 n-gram 模型之中做為長距離資訊擷取之輔助，透過線性插補(linear interpolation)的方式，我們可以一權重 a_i 將其做合併，也就是

$$\log P_{MERGED}(S) = \sum_{i=1}^{k} a_i \cdot \log P_i(S) \tag{20}$$

其中 $0 \le a_i \le 1$ 且 $\sum_{i=1}^{k} a_i = 1$，在這邊我們有兩個模型機率存在

1. $P_1(S) = P_{n\text{-}gram}(S)$ 為 n-gram 模型對文句 S 所估測出之機率。
2. $P_2(S) = P_{MI\text{-}Trigger\text{-}pair}(S)$ 為觸發序對模型對文句 S 所估測出之機率。

透過(20)式的計算，我們可以使用觸發序對計算出一段文句的機率，且此機率有長距離資訊存在，比起傳統的 n-gram 模型在資訊擷取上為佳。

4. 關聯法則與其應用

在這邊我們引入了一個在資料探勘(Data Ming)領域受到十分廣泛運用的 Apriori 演算法[1]，此演算法可以用來建立關鍵詞的關聯法則，舉例而言，假設有一組交易紀錄資料庫，此資料庫記錄著每筆交易所包含的商品，關聯法則所要擷取的就是每個商品間的相互關係，也就是說我們想知道一筆交易出現了某種商品後，還有哪些商品是可能出現在同一筆交易紀錄之中，如果說商家從關聯法則中知道顧客買了商品甲後，還有很大的機率會去買商品乙，則可將商品甲與商品乙放在附近增加顧客的方便性與商家的業績。

4.1 Apriori 演算法

假設我們有一組新聞文件資料庫 D，裡面包含了$|D|$篇文章，每篇文章均是辭典 $L = \{w_1, w_2, ..., w_n\}$ 的子集合，用上面的例子解釋就是各種商品的集合之意，而關聯法則以 $X \Rightarrow Y$ 的型式表示，其中 X、Y 均是 L 的子集合(subset)且互

相獨立，如果在所有包含 X 的文章中有 $c\%$ 同時也包含了 Y，則我們可以稱關聯法則 $X \Rightarrow Y$ 存在於資料庫 D 中的信賴度(confidence)為 c，此外若有 s% 的文章同時包含 X 與 Y，則我們可稱關聯法則 $X \Rightarrow Y$ 以支持度(support) s 存在於資料庫 D 中，換句話說，信賴度是一種量測關聯法則強弱的標準，而支持度則是表示統計上出現的頻率，事實上我們實作時會訂定信賴度與支持度的門檻，我們擷取出來之關聯法則的信賴度與支持度均必須大於此門檻。

以下即為擷取關聯法則的演算法流程，是以資料探勘中的 Apriori 演算法做為基礎所改寫而成若我們以簡單的例子說明之，假設我們共有三詞，分別以 *a*、*b*、*c* 代表，Apriori 演算法就是在找尋此三詞的關聯性，它的概念就是先將這些詞兩兩為一組建立序對集合*(a,b)*，*(a,c)*，*(b,c)*，並且對資料庫搜尋每一序對，是否同時出現在於同一文句中，假設只有*(a,b)*，*(b,c)*序對符合這項條件，則將*(a,c)*刪除，此時我們建立*(a,b)*，*(b,c)*的關聯法則，此關聯法則的層級(step) 為二 ，不過我們必須計算其信賴度與支援度，例如我們可以計算同一篇文章出現詞 *a* 且出現詞 *b* 的機率，即為其信賴度。而我們會再將剩餘下之序對*(a,b)*，*(b,c)*做結合成為*(a,b,c)* 並搜尋訓練文集中*(a,b,c)*會不會同時出現於一文句中，如果沒有則刪除之，有則可以計算給定任二詞，出現第三詞的機率，此時稱關聯法則的層級為三，找尋出的關聯法則之層級是依其詞數而定，層級越大，代表其關聯法則中的字數越多。上述只是概念性的做法，Apriori 演算法事實上會做一些節省時間的動作，而其中最重要的部分就是建立雜湊樹(hash tree)的資料結構以節省更多的運算時間。

Apriori 演算法

1) L_1 = { words | counts is large than support threshold };
2) for (k = 2 ; $L_{k-1} \neq \varnothing$; k++) do begin
3) C_k = unification (L_{k-1}) ; // produce new candidate sets
4) for all sentence t do begin
5) C_t = subset(C_k, t) ; // candidates contained in sentence t
6) for all candidate $c \in C_t$ do
7) Increment count of candidate c ;
8) end
9) L_k = { $c \in C_k$ | count of candidate is large than support threshold }
10) end
11) Answer = $\bigcup_k L_k$;

表一、 關聯法則演算法

使用我們改變後的 Apriori 演算法，我們可以獲得詞與詞之間的關聯法則，而關聯法則的形式如下

$WordSeq \Rightarrow Y$　confidence = $c\%$　suppport = $s\%$ 代表出現詞序列 *WordSeq* 的文章中有 c%的機率會出現 Y，而有 s% 的文章同時包含了 *WordSeq* 與 Y。以下為利用西元二千零一年十二月二十八號到西元二千零一年十二月三十一號期間的政治新聞所擷取出的兩條關聯法則範例，第一條關聯法則的層級為二，第二條的層級則為三

(1)小三通　⇒　大陸

confidence = 90%　support = 6.25%

(2)大陸 小三通 ⇒　兩岸

confidence = 100%　support = 2.25%

以上例而言，出現"小三通"後，新聞語料有 90 % 的機率也會出現 "大陸" 這個詞，這個關聯法則佔了總文句 6.25%，如果"大陸"與"小三通"同時出現後，新聞語料會有 100 %的機率也會出現"兩岸"，而此關聯法則佔了總文句 2.25%。

4.2 關聯法則為主之 n-gram 模型機率計算

為了使關聯法則更能反映語言模型的特性，我們將關聯法則做一稍微的變動，我們將其使用相互資訊的形式表示，就有如觸發序對一般對任一關聯法則 *WordSeq* ⇒ B，我們計算其相互資訊並用對數表示之

$$MI - Association(WordSeq; B) = \log \frac{P(WordSeq, B)}{P(WordSeq)P(B)} \tag{21}$$

如使用關聯法則為長距離資訊的輔助的 n-gram 模型機率，為了方便以對數表示為

$$\log P(S) = \sum_{i=1}^{T} \log P(W_i) + \sum_{i=1}^{T} MI - Association(W_1 W_2 ..... W_{i-1} \Rightarrow W_i) \tag{22}$$

其中 $P(W_i)$代表 unigram 模型之機率。$MI\text{-}Association(W_1 W_2 ... W_{i-1} \Rightarrow W_i)$ 代表使用詞序列 $W_1 W_2 ... W_{i-1}$ 所找出之最大層級之關聯法則 *WordSeq* ⇒ W_i 的相互資訊，如同觸發序對一般 ws 代表 window size，在這邊我們也將 window size 定為文句長度，也就是說我們的關聯法則是文句階層的關聯法則(sentence-level

association rule)。如同觸發序對一般，我們要將關聯法則與傳統 n-gram 模型做結合，如同(20)式，此時

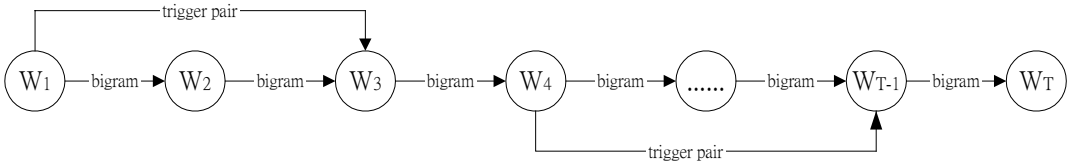
1. $P_1(S) = P_{n\text{-}gram}(S)$ 為 n-gram 模型對文句 S 所估測出之機率。

2. $P_2(S) = P_{MI\text{-}Association}(S)$ 為關聯法則模型對文句 S 所估測出之機率。

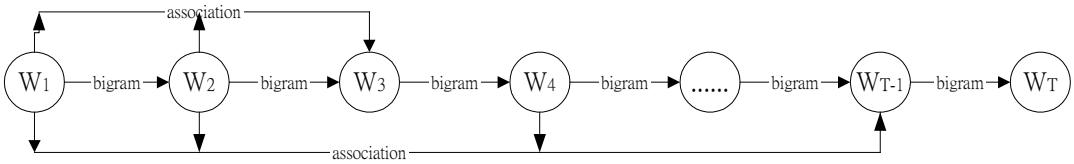
4.3 關聯法則與觸發序對之比較

使用關聯法則做資訊擷取與觸發序對最大的不同在於我們透過關聯法則可以獲得多元詞組(multi-word)之間的關聯性，而觸發序對只能擷取詞與詞之間(word pair)的關聯性，互相比較之下我們的方法是較強健的，圖一為關聯法則與觸發序對之示意圖，圖中箭頭表示關聯性。由圖中我們可以清楚的看出傳統的 bigram 模型只能由前面所出現的詞來對目前所出現的詞做機率評估，即 n-gram 模型文句間的關聯性是循序的，且受制於 n-gram 視窗之大小，而觸發序對則可跳脫此關聯性，只要是同一段文句中所出現的詞都可以有相互間的關聯性存在，不過觸發序對模型的限制在於只能擷取詞與詞之間的關係，而我們所提出之關聯

bigram



bigram 與 Trigger pair



bigram 與 Association rule

圖一、 bigram 模型、Trigger pair 與 Association rule 之比較

法則序對則可以將關聯性擴大，變成多元詞組間的相互關係，可以說是觸發序對的延伸研究。

5. 實驗

5.1 實驗資料庫

為了將本論文方法時實現在中文系統中，首先必須製作了一套詞典，詞典中主要部分是 CKIP 中文詞庫[15]，它主要是利用國語日報辭典中約四萬目詞的原始資料加以分類，並且附加部分的語法及語意訊息在其中，本論文只使用到詞出現的頻率，並無使用到語法與語意資訊，取出其中一、二、三、四詞的部分作為我們的基本辭典，並不定期由人工更新我們這部詞典的新詞。

另外我們準備了兩組基本的實驗資料庫，第一是 CKIP 平衡語料庫，這是一個十分一般化的語料庫，有近二千萬字之現代漢語語料及超過五百萬字之古代漢語語料。另外我們在西元二千零一年四月十日至四月十六號擷取民視即時新聞、中央社、中時電子報、電子新聞網、聯合新聞網、ETtoday 與鉅亨網等網站的新聞文件共 3118 篇，包含有科技、社會、休閒、國際、體育、影視、政治及財經等八大類別，依日期將其區分為訓練文集(四月十日到四月十四日共 2234 篇)和測試文集(四月十五、十六日共 884 篇)。並且我們以 bigram 模型來驗證本研究方法。

5.2 不同語言模型之實驗結果

我們會對平滑化技術做評估，將傳統 n-gram 模型與加入 Witten-Bell 平滑化技術的 n-gram 模型做比較，在這邊我們使用 CKIP 平衡語料庫加上新聞訓練文集做訓練，並對測試文集做評估，表二的第一列為傳統 n-gram 模型所得之結果，第二列則為加入 Witten-Bell 改進後之結果。另外我們會在本小節中對混合式模型的效能做評估，在混合式模型方面，我們使用 CKIP 平衡語料做為一般化模型之訓練文集，並經由新聞訓練文集分類好的八個群組訓練群組模型，對測試文集做測試，所測得之 perplexity 如表二第三列所示。接下來是對觸發序對與原始模型的效能做比較，在這邊我們使用 CKIP 平衡語料庫加上新聞訓練文集做訓練，

並對測試文集做評估，所得之結果列於表二的第四列。

Bigram(Baseline)	258.8
Bigram + Witten-Bell	193.5
Bigram + Mixture n-gram	201.5
Bigram + MI-Trigger pair	237.5

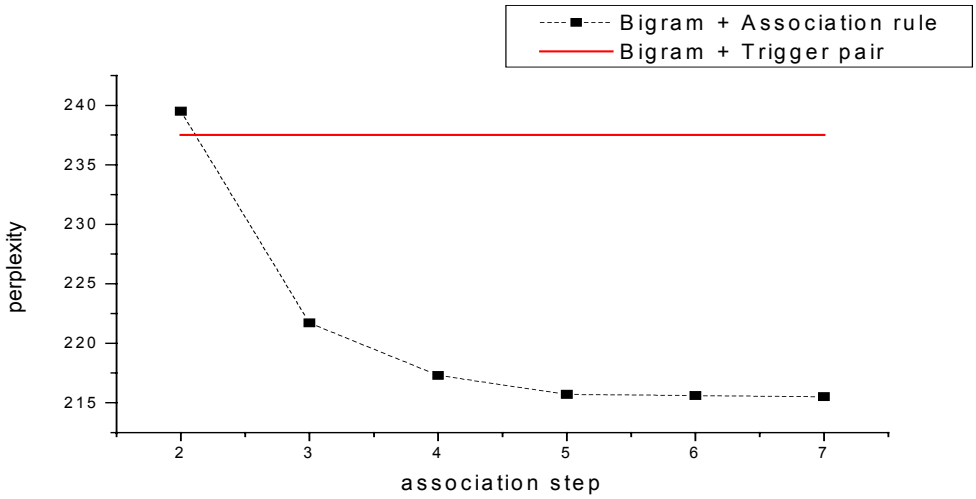
表二、不同語言模型改進技術之 perplexity 比較

5.3 關聯法則之模型效能

在本實驗中，一開始我們將會測試關聯法則依最大層級數對於 perplexity 之影響，在這邊我們使用 CKIP 平衡語料庫與新聞訓練文集做訓練，並且對測試文集做測試，表三即為所得之結果，圖二為其與觸發序對的圖形化表示，可以看出隨著最大層級的增加，perplexity 有一定程度的下降，由此項觀察可以證明我們所提出的相互資訊為基礎的關聯法則對於 n-gram 模型的改進有相當程度的改善，並且層級數較高的情形下，模型的效能優於觸發序對，而我們也發現最大層

Association Step	2	3	4	5	6	7
Bigram +Association	239.5	221.7	217.3	215.7	215.6	215.5

表三、依關聯法則最大層級不同所測得之 perplexity 比較



圖二、觸發序對與關聯法則不同層級所得之 perplexity 比較

級到達五就發生了飽和狀態，在此之後隨著層級數增加，perplexity 並不會有顯著的改善，故在本論文之後的實驗我們都將關聯法則的最大層級定為五，用以節省記憶體與運算時間。我們會將所有提到的技術相互結合做比較，在這邊使用為 CKIP 平衡語料庫加上新聞訓練文集做訓練，而後對測試文集做 perplexity 之評估比較，結果如表四所示，在這邊可以發現我們的方法可以有效的結合平滑化技術與混合式模型，對於 n-gram 模型的改進可以更進一步。

Bigram (Baseline) + Witten-Bell	193.5
Bigram + Mixture n-gram + Witten-Bell	178.3
Bigram + Witten-Bell+MI-Trigger Pair	168.8
Bigram + Mixture n-gram + Witten-Bell+ MI-Trigger pair	160.4
Bigram + Witten-Bell+ MI-Association	148.2
Bigram +Mixture n-gram + Witten-Bell+ MI-Association	135.8

表四、不同結合技術之 perplexity 比較

5.3 語音辨識之實驗

我們將 n-gram 語言模型與語音辨識的工作做結合，語音辨識是以隱藏式馬可夫模型(HMM)為基礎，特徵參數為二十六階語音特徵參數，由 12 階的 MFCC、12 階的 delta MFCC、delta log energy 與 delta delta log energy 所組成，語音訊號的取樣頻率為 8kHz，解析度為 16 bits，音框大小為 256 點(23.22ms)，音框位移大小為 85 點(7.74ms)。所使用的語音資料庫為 Mandrain Across Taiwan(MAT) 所提供的 MAT-160。測試語料由不確定人數之男性及女性所錄音之國語獨立詞與文句透過電話錄音共 500 句，供做便是測試用。表五為使用上述語料所做出之辨識結果，第一列(Baseline)為單純使用音節模型辨識技術所得之結果，第二列(Bigram)為音節模型辨識分數再加上語言模型辨識分數所得之結果，第三列(Bigram + MI-Trigger pair)則是先對語言模型使用觸發序對改進後所得之分數，再與音節模型分數合併所得之結果，第四列(Bigram + MI-Association) 則是先對語言模型使用關聯法則補償技術後所得之分數，再與音節模型分數合併所

得之辨識率，上述語言模型與聲學模型分數之合併比重均為 1:1，並且語言模型事先都經過 Witten-Bell 平滑化技術解決其平滑化問題。

Baseline	51.33
Bigram	51.86
Bigram + MI-Trigger pair	52.31
Bigram + MI-Association	52.92

表五、不同語言模型所測得之音節正確率(%)

5.4 文件分類之實驗

在本實驗中，我們將透過所提出的改進方法對文件分類的工作做模擬，在這邊我們使用新聞訓練文集對八種不同的領域分別訓練出 n-gram 模型，成為原始模型，另外加入觸發序對與關聯法則成為改進後之模型，由此二模型為基礎進行文件分類模擬工作之正確率比較，表六為所得之結果，由表中可以觀察到我們改進過後的模型在文件分類的工作上比起傳統之 n-gram 模型在分類的正確率上有小幅度的改進，而我們認為改進幅度並不如預期的主要原因是由於我們所選取的新聞文件有未確定性(ambiguous)的問題，有些新聞文件在網頁上是分類於政治領域，但實際上若將其分類於財金領域也未嘗不可，這些文件造成我們在模擬分類的實驗時錯誤率的增加。

	科技	社會	休閒	國際	體育	影視	政治	財經	平均
Bigram(Baseline)	64.8	77.1	69.8	72.6	84.6	75.4	86.9	72.1	75.4
MI-Trigger pair	66.6	78.3	69.6	70.9	85.2	76.0	86.9	74.8	76.0
MI-Association	66.8	78.3	70.8	71.9	85.2	75.8	88.6	75.1	76.6

表六、不同語言模型所做之文件分類正確率(%)

5.5. 個人化新聞文件瀏覽器

在論文最後，我們將以自然語言模型為發展基礎，透過模型機率的評估，對

於網路新聞文件做線上分類，並依據個人閱讀的習慣，建立出一套符合個人需求之新聞文件瀏覽器，期望能藉由這套系統增加一般人在閱讀新聞文件上的效率，圖三即為此瀏覽器在視窗上的執行時的畫面，藉由畫面中的“文件更新”按鈕，此瀏覽器會透過網路獲得最新之新聞文件，而“啟動學習”的按鈕則會將閱讀過的文件資訊加入語言模型之中，最後由語言模型的預測分數將新聞文件做排序，使用者可能較有興趣的新聞會優先列在標題欄內，假設使用者是一位籃球迷對於 NBA 的相關報導十分的關心，在六月十四日 NBA 總冠軍賽正打的火熱時在運動類新聞選取了幾篇相關的文件，如圖三所示，當此使用者在六月十五日瀏覽運動類新聞時，系統會優先將 NBA 相關的文件列於標題欄，如圖四所示。

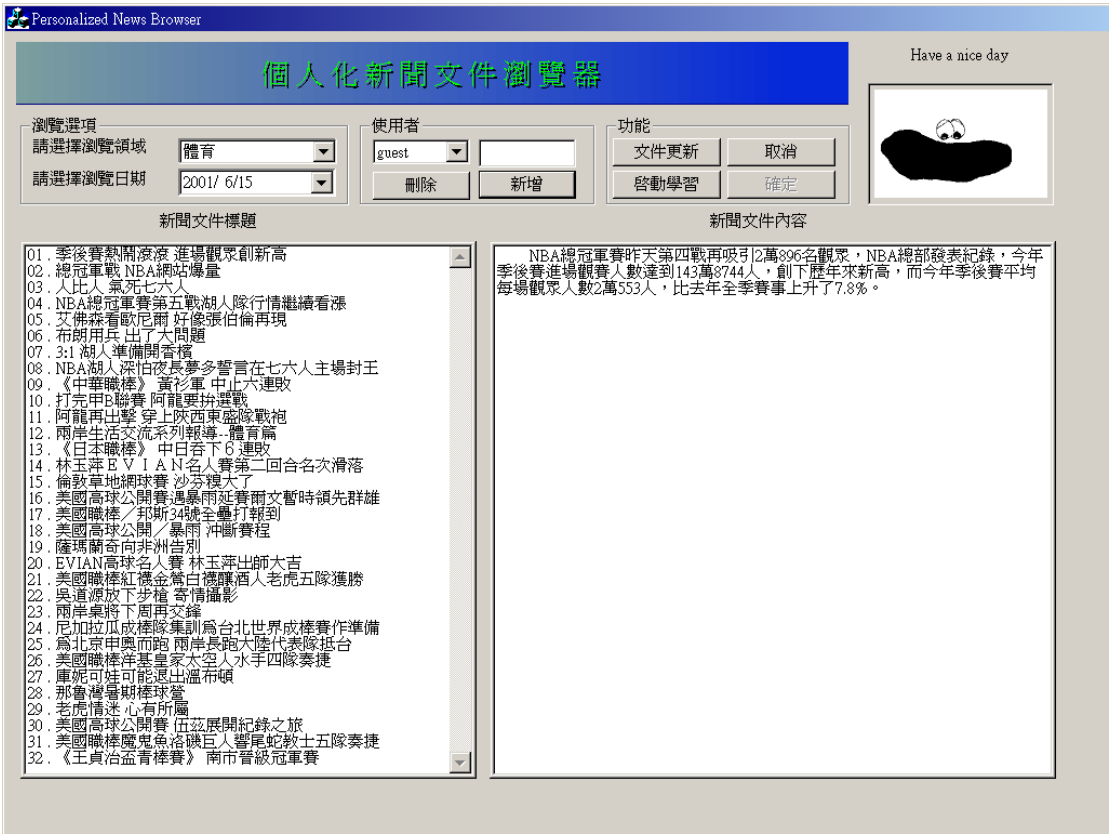
6. 結論及未來研方向

在本論文中我們對於傳統的 n-gram 模型做了完整的介紹，從模型的建立與評估到模型缺點的探討，都有一套完整的說明，而我們也針對每項 n-gram 模型的缺點介紹了近幾年來一些較為受歡迎的解決方式，包括了混合式 n-gram 模型、Witten-Bell 平滑化技術與觸發序對等，而在我們論文中的實驗，也證明了這些方法對於改進 n-gram 模型是十分有效的。

另外我們也在論文中提出了一個關聯法則的技術，此方法是透過一個在資料探勘上十分受歡迎的 Apriori 演算法，利用文句結構的特性，使用文句前面所提供的資訊來建立文句中前後文的關聯法則，將其用於 n-gram 模型的改善上得可以得到不錯的效果，在實驗方面本方法可以有效降低 perplexity，證明我們的改進過後之 n-gram 模型比起傳統的模型效能為佳，最後我們將其應用在語音辨識與文件分類的工作上在正確率上也有一定幅度的改善。

未來在 n-gram 模型上的研究應不僅止於解決傳統 n-gram 模型上的缺陷，傳統上的 n-gram 模型是從統計的概念發展而出，嚴格來講並不是一個完整的自然語言，只能說是其中的一個重要的部分，未來必須有效的結合語言文法與知識背景的語言學才能算是真正的語言模型，而如何將三者融合[10]是一個十分困難

圖三、個人化新聞文件瀏覽器展示介面(2001/6/14)



圖四、個人化新聞文件瀏覽器展示介面(2001/6/15)

的問題，因為 n-gram 模型是機率模型，語言文法則是一個人工定的條例，並不是一組機率模型，有人透過剖析(Parsing)將其機率分析而出，這是一門重要且艱深的學問，而知識背景的語言學又更複雜了，只有人工定的分數沒有機率模型，要將其與 n-gram 模型結合則又必須花費更多的功夫，但是唯有克服這個困難，才能夠大幅度的提昇語言模型的效率。

參考文獻

[1] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, Proceedings of the 20th VLDB Conference, Santiago-Chile, pp.487-499, 1994。

[2] S. F. Chen and J. Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling”, Computer Speech and Language , vol.13, 359-394 , 1999。

[3] P. R. Clarkson and A. J. Robinson, “Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache”, Proc. of ICASSP, pp.799-802 , 1997。

[4] R. Iyer and M. Ostendorf, “Relevance weighting for combining multi-domain data for n-gram language modeling”, Computer Speech and Language, vol.13, pp.267-282, 1999。

[5] R. M. Iyer and M. Ostendorf, “Modeling long distance dependence in language : Topic Mixtures Versus dynamic cache models”, IEEE Transaction on speech and audio processing , vol.7 , January 1999。

[6] F. Jelinek and R. L. Mercer, “Interpolation estimation of Markov source parameters from sparse data”, Proceedings of the workshop on pattern recognition in Practice, North-Holland, Amsterdam, The Netherlands, pp.381-397, May 1980。

[7] D. Klakow, “Selecting Articles from the Language Model Training Corpus”, Proc of ICASSP, pp.1695 –1698, 2000。

[8] R. Rosenfeld, “A maximum entropy approach to adaptive statistical language model”, Computer Speech and Language, vol 10 , pp.187-228 , 1996.

[9] R. Lau, R. Rosenfeld, and S. Roukos, “Trigger-based language models: A

maximum entropy approach” , in Proc. Int. Conf. Acoustics, Speech, Signal Processing, vol. II, pp. 45–48. , 1993

[10] M. Meteer and J. R. Rohlicek, “Statistical language modeling combining N -gram and context free grammars” , in Proc. Int. Conf. Acoustics, Speech, Signal Processing, vol. II, pp. 37–40 , 1993.

[11] C. D. Manning, H. Schutze, “Foundations of statistical natural language processing”, Massachusetts Institute of Technology pp.315-407, 1999。

[12] L. Rabiner and B.H. Juang, “Funadamental of Speech Recognition”, Prentice Hall, pp.321-387, 1993。

[13] I. H. Witten and T. C. Bell, “The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression.”, IEEE Transactions on Information Theory , vol.37, pp.1085-1094, 1991。

[14] G. D. Zhou and K. T. Lua, “Interpolation of n-gram and mutual-information based trigger pair language models for Mandarin speech recognition”, Computer Speech and Language, vol. 13, pp.125-141, 1999。

[15] CKIP, http://godel.iis.sinica.edu.tw, 中央究院資訊科學研究所詞庫小組。