國語語音辨認中多領域語言模型之訓練、偵測與調適

Training, Detection and Adaptation of Multi-Domain Language Models for Mandarin Speech Recognition

林頌堅[1]、蔡吉龍[1]、簡立峰[2]、陳克健[2]、李琳山[1,2]
[1]國立台灣大學資訊工程學研究所
[2]中央研究院資訊科學研究所
e-mail: lsc@speech.ee.ntu.edu.tw

摘要

在本論文中，我們提出適用於不同領域間中文語言模型的自動訓練、偵測與調適方法。應用這些方法在極大詞彙國語語音辨認的語言解碼方法將可以訓練出各種不同應用領域的語言模型，為輸入的語音選擇合適應用領域的中文語言模型，對訓練語料不足的特殊領域語音辨認可以進一步提昇辨認正確率。在初步的實驗中，我們利用多領域語言模型進行語言解碼的語音辨認正確率可以比利用一般語言模型高 2~8%，從這樣的結果可以驗證語言模型調適確有其效果，並值得做進一步的研究。

一、緒論

本論文提出一系列適用於不同領域間中文語言模型(Chinese Language Models)的自動訓練(Training)、偵測(Detection)與調適(Adaptation)方法。應用這些方法在極大詞彙國語語音辨認(Mandarin Speech Recognition with Very Large Vocabulary)的語言解碼(Linguistic Decoding)方法將可以為輸入的語音選擇合適應用領域(Application Domain)的中文語言模型，進一步提昇在特定領域下的語音辨認正確率。這項結果並有助於我們瞭解在不同領域下語言的統計特性，拓展國語語音辨認在不同應用領域的適用性。

在極大詞彙國語語音辨認中，常用的語言模型是統計式的馬可夫模型(Markov Model)[1,2,3]。訓練一個可靠的統計式馬可夫語言模型需要蒐集極大量的訓練語料(Training Corpus)，但是在若干應用環境中我們無法提供大量的訓練語料[4]。一個解決的方法是利用語言模型調適的技術將目前針對極大詞彙語音辨認系統所訓練出來的語言模型加以轉換成這個領域的特定語言模型，這樣的技術便成為目前語言模型中極富挑戰性的研究[5,6]。

在本論文我們提出利用多個領域特定語言模型(Domain-Specific Language Models)來進行語言模型調適的方法，語音辨認系統會根據使用者所輸入的語音自動選擇合適的應用領域語言模型進行語言解碼。使用多領域語言模型的語音辨認系統，一來可以提昇語音辨認的正確性；二來有助於瞭解如多種資料庫存取等多領域語音辨認應用的研究。在我們的方法中，首先將蒐集到所有的訓練語料訓練一個一般語言模型(General Language Model)，希望能抽取出不分領域所共有的

語言統計特性，然後對每一個應用領域，便可以少量的領域特定語料所訓練出的語言模型組合一般語言模型，產生一個新的領域特定語言模型，以適用於新領域的語言解碼。於是，對每一個應用領域，我們都有一個最合適的領域特定語言模型來描述該領域的語言特性。在使用者輸入語音時，我們就可以利用目前輸入語音在不同領域語言模型的辨認結果，自動挑選最合適的領域特定語言模型。在以多領域語言模型進行語音辨認的實驗中，對不同領域的測試語料使用該領域特定的語言模型比只用一般語言模型，均可得到 2~8%正確率提昇，因此可以證實語言模型調適對特定領域下的語音辨認有所幫助。

在本論文所提出的多領域語言模型方法中有兩個問題亟待解決。一是對不同領域如何訓練出該領域特定的語言模型，另一個問題則是進行語音辨認時，如何偵測輸入語音的應用領域來選擇領域特定語言模型。對於多領域語言模型的訓練，我們以內插法組合一般與應用領域語言模型，也就是以不同的加權值(Weighting Value)組合兩者的詞雙連機率值(Word Bigram Probabilities)，作為新的詞雙連機率值。每當新收到一筆語料需要進行領域特定語言模型訓練時，我們就以各個領域特定語言模型來判讀這筆語料與哪一類應用領域的訓練語料較接近，判讀之後將這筆語料加入最接近的應用領域訓練語料中，訓練出新的應用領域語言模型；如果新的訓練語料與所有訓練語料都相差相當大，我們便以這筆語料新成立一類應用領域，並對該應用領域進行語言模型訓練。在本論文中，我們實驗以文字複雜度(Perplexity)[7,8]與詞雙連涵蓋率(Word Bigram Coverage Rate)作為判讀應用領域的資訊。在結合這兩種資訊後，從實驗中，我們比較這種自動判讀領域的方法，可以發現訓練語料的領域分類與由人工判讀的結果相同。

對於選擇用來進行語言解碼的語言模型，我們以領域特定語言模型對輸入語音的辨認結果來偵測這些語音最接近哪些應用領域。在使用者輸入語音到多領域的國語語音辨認系統時，語言解碼單元先用所有經調適後的領域特定語言模型對前面輸入數句語音的候選音節序列進行語言解碼，將所得到的分數作為應用領域偵測的資訊。蒐集足夠預測接下來輸入的語音可能屬於哪一個應用領域的資訊之後，便可以該應用領域的語言模型提供接下來語言解碼所需的文法限制之用。

本論文的其餘章節結構如下；第二節對語言模型調適的問題再作一個清楚的描述，並且回顧一些前人在這個研究方向的努力以及在本論文中我們所提出來的方法。第三節報告我們的多領域語言模型訓練法以及使用文字複雜度和詞雙連涵蓋率來進行訓練語料領域判讀的一些實驗結果。第四節我們介紹利用多領域語言模型進行極大詞彙語音辨認的方法和一些初步的結果。最後，我們以第五節對本論文作一個總結，並展望未來在語言模型調適的研究方向。

二、語言模型調適

統計式馬可夫語言模型需要大量的語言模型作為決定模型參數的依據，但大量訓練語料的蒐集需要花費相當長久的時間與大量的人力，所以是件非常困難的工作。以往的極大詞彙的國語語音辨認的研究便是個很好的例子，在以往發展出來的實驗性國語語音辨認系統中，由於蒐集訓練語料的困難，目前能提供穩定而大量訓練語料的來源，只有中文報紙，所以在這些實驗或系統中，都以所蒐集到的新聞作為訓練語料來訓練中文語言模型，所能發展的應用也都偏向新聞聽寫的應用。但是，隨著極大詞彙語音辨認技術發展的成功，發展這項技術到其他應用

領域已是件水到渠成的工作了，比方說著名的飛行旅遊資料庫存取(Air Travel Information Service, ATIS Project)，便是美國尖端研究計畫署(ARPA)所極力推動的利用語音來存取資料庫內容應用的群體計畫[9]。在這些應用領域下的語言模型技術，自然成為一項重要的研究方向。

對於這些特殊的應用領域而言，因為輸入語音的用字、句型等等語言特性與新聞語料大不相同，顯然不能以原先利用新聞語料所訓練出來的語言模型作為語言解碼的資訊。但針對這些應用領域蒐集大量訓練語料並不是件容易的事，尤其是在一些新的應用領域，系統的雛形階段所能蒐集的語料非常的少。因此，在這種限制之下，特殊領域的語音辨認自然很難得到理想的結果。以往對於特殊應用領域的語言解碼，一個可行的方法是利用詞群語言模型(Word-Class-Based Language Models)[10,11]，這個方法針對應用領域內的用語，根據它們在這領域內的語法(Syntax)、語意(Semantics)等訊息來分群，譬如在 ATIS Project 中，我們可以將這個領域中所有可能用到的城市名稱歸為同一群，因為它們在這個領域都是指出發地和目的地，扮演同樣語法和語意的角色。對詞分群的方法可以降低語言模型的參數量，提供訊息分享(Information Sharing)的能力，因此可以增加語言解碼時的強健性(Robustness)。但是由於所能取得的訓練語料，實在相當有限，所以無法以自動分群的方法來進行語言模型訓練，需要不少以專家知識介入的地方。

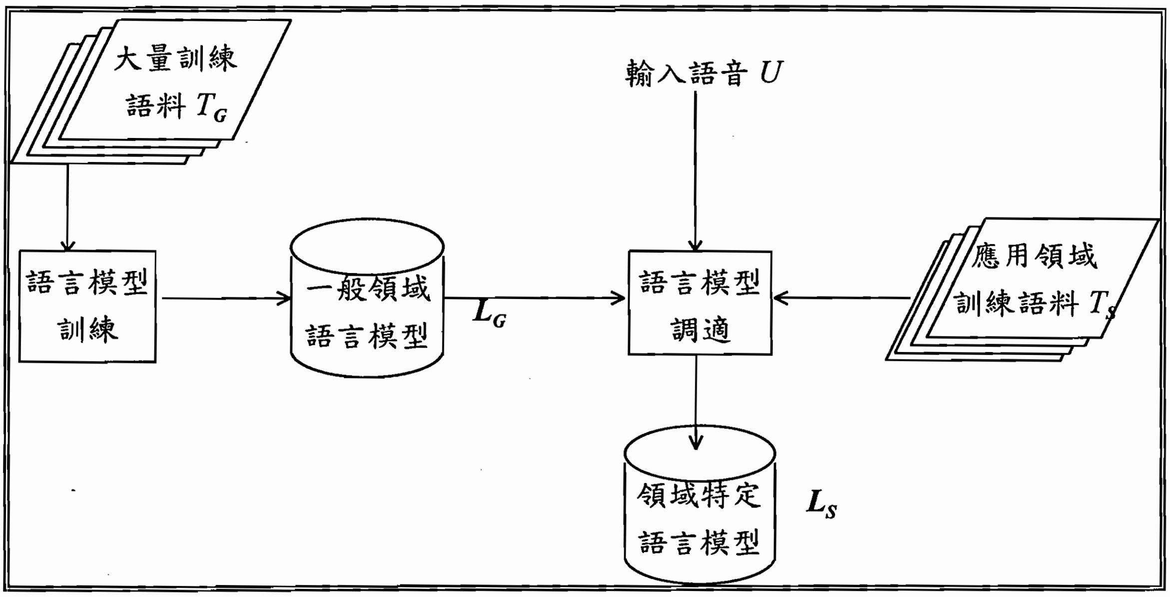

圖 1 語言模型調適示意圖

除了詞群語言模型之外，另外一種方式是以語言模型調適的方法來增加在特定領域下的語言解碼能力[12]。語言模型調適的概念是先以大量的訓練語料，訓練出一個一般語言模型 L_G，希望能從一個非常大量的訓練語料庫抽取出不分領域所共有的語言統計特性。然後當進行某應用領域 S 下的語音辨認時，對輸入的語音 U，抽取其中的語言特性資訊來對一般語言模型 L_G 進行調適，產生一個新的語言模型 L_S'，加強這個領域內特殊的用詞和語法等語言特性，以適用於新領域的語言解碼。用來對 L_G 調適的資訊可能只有 U 中的語言特性資訊，或是從少量應用領域 S 內蒐集的語料 T_S 中獲得。如圖 1 所示便是一個這樣的語言模型調適的過程。

過去在語言模型調適上的研究

在介紹我們的研究之前，我們首先回顧一些前人在這方面所做的努力。在特定領域的語料中，我們可以觀察到有某些詞經常重複出現的出現，比方說，有關「公共安全」領域內的語料中常常出現出現像“檢查、火災、安全、樓梯、消防”這一類的詞。在參考文獻[13]中就利用這個語言現象，以一個短期記憶體(Cache Memory, Short-Term Memory)來儲存辨認過的詞，當下次預測輸入語音中是否包含有這樣的詞時，便可以得到較高的機率。

再進一步觀察，可以發現領域內的高頻詞間往往存在某些關連性(Association)，於是參考文獻[14]便提出觸發對(Trigger Pair)的觀念，其觀念就是希望從特定領域語料內抽取兩兩具有具有關連性的詞，把這樣的一對詞稱為觸發對。當輸入的語音中有觸發對的其中一個詞出現時，另外一個詞在語言模型中的機率值也跟著動態調整，來提昇語言解碼的正確性。從實驗數據顯示，經過人工選取的觸發對的確可以降低約 12%的文字複雜度。

為了使語言模型具有隨著應用領域而調整的能力，以符合特殊應用領域之特性，參考文獻[15]和[16]都以最小區辨資訊(Minimum Discrimination Information, MDI)的方法改變最小的分佈差異來調整不同領域間不同的 n 連機率。實驗結果顯示這種方法可以降低語音辨認中詞的錯誤率約 10~14%，表示這樣的方法的確有它的發展空間，而這樣的一個實驗結果也引發我們進行分析不同應用領域中語言特性的動機，於是我們便著手研究多應用領域中文語言模型調適的方法。

本論文所提出的方法

本論文所針對語言模型調適的問題，是一個藉由多領域語言模型[12]來進行調適的問題，也就是在使用者輸入語音後，語音辨認系統根據所輸入的語音選擇合適的應用領域語言模型，作為調適的語言模型，進行語言解碼。使用多領域語言模型的語音辨認系統，一來因為所使用的語言模型較一般語言模型精確，可以對各領域內不同的語言特性做較精細的描述，有助於提昇語音辨認的正確性；二來目前如多種資料庫存取等多領域語音辨認的應用已逐漸興起，以多領域語言模型整合極大詞彙語音辨認系統，有助於瞭解這方面應用的研究。

在進行語音辨認時，當使用者開始輸入語音，我們首先以少量輸入語音來預測目前輸入語音是屬於哪一種領域，比方說 S 領域，然後便以 S 的領域特定語言模型 L_S'對接下來輸入的語音進行語言解碼。領域特定語言模型 L_S'包含了不分領域共同的語言特性與這個領域內獨特的語言特性，其訓練方法如下：首先我們利用蒐集到的大量訓練語料，訓練一個包含所有不分領域的一般語言模型 L_G。L_G 具有每一個應用領域所共有的語言統計特性，以 L_G 進行語言解碼，基本上對各種領域的語音，已經可以達到還不錯的辨認效果。然後對每一個應用領域 S，以所蒐集到少量的領域特定訓練語料 T_S 來訓練領域 S 的語言模型 L_S，加強領域 S 內所特有的語言特性，再組合 L_G 與 L_S，產生新的領域特定語言模型 L_S'。於是，對每一個應用領域，我們都有一個最合適的領域特定語言模型來描述該領域的語言特性。

在本論文中，所採用的語言模型是詞雙連語言模型。以內插法(Interpolation)[17]來組合一般詞雙連語言模型 L_G 和領域 S 詞雙連語言模型 L_S 中

的詞雙連機率值，作為領域特定語言模型 L_S'內的詞雙連機率值。我們可以用式(1)的數學形式來更清楚的表達這個方法。

$$\Pr_S{}'(w_i|w_j) \overset{def}{=} (1-d)\times \Pr_G(w_i|w_j) + d\times \Pr_S(w_i|w_j) \qquad \ldots\ldots(1)$$

在這個式子中，w_i 和 w_j 代表詞典中的任意兩個詞，$\Pr_G(w_i \mid w_j)$和 $\Pr_S(w_i \mid w_j)$ 分別代表(w_j, w_i)這對詞出現在 L_G和 L_S 中的詞雙連機率值。d 是一個介於 0 和 1 之間的數值，用來調整調適詞雙連機率值 $\Pr_S'(w_i \mid w_j)$中 $\Pr_G(w_i \mid w_j)$和 $\Pr_S(w_i \mid w_j)$間的比重。一般而言，若是我們蒐集到較多的應用領域訓練語料，我們可以較信賴領域詞雙連語言模型 L_S，d 的值可以設一個較大的值；否則，我們需要仰賴一般詞雙連語言模型 L_G提供語言解碼時的資訊，d 就必須設定一個較小的值。下面，我們以一般以及領域特定詞雙連語言模型進行語言解碼的實驗，以驗證語言模型調適的效果。

實驗環境

在報告語言解碼的實驗及其結果之前，我們首先對實驗的環境參數作一介紹，爾後在本論文的實驗中，若非特別提及，則表示相同的參數。

(1)詞典：

由中研院詞庫小組提供的詞典，共 84464 目詞，取高頻之一字到四字詞共 43591 目，其詞目數分佈如表 1 所示：

詞長	一字詞	二字詞	三字詞	四字詞
詞目數	8,154 目	23,529 目	5,494 目	6,414 目

表 1 本論文所使用詞典的詞目數分佈情形

(2)語言模型訓練語料：

在語言模型訓練語料中，包括了一般與特定領域兩類。對於一般語料，我們所能蒐集到穩定與大量的中文語料是新聞語料，同時中文新聞語料包羅萬象，幾乎涵蓋了各個領域，因此我們以三家報社共九個月的新聞語料做為我們實驗中一般語言模型的訓練語料。至於特定領域語料的選取，我們從不同語料來源來蒐集，包括從現代漢語平衡語料庫所選取的哲學類(phi)和文學類(lit)相關文章、從電子佈告欄(BBS)中蒐集的棒球區(ball)和微軟視窗軟體區(win)內的討論文章、以及一些關於科技報告的文章(cs)。表 2 是這些訓練語料的大小。

領域別	一般	phi	lit	ball	win	cs
詞數	12,094,234	67,127	61,676	170,214	16,647	3,445
字數	18,384,664	100,054	87,987	226,864	23,463	5,612

表 2 本論文實驗所用各個領域訓練語料的大小

(3)測試語料：

對每一特定領域，我們從對應的領域中選取若干篇文章進行測試，但所選取的文章並不與訓練語料所選取的重複。換言之，這些語料全部都是外部測試(Outside Test)，所有的應用領域共計 20 篇測試語料。在語音輸入方面，我們以一位語者對每篇測試語料唸一遍，以金聲三號的聲學處理單元[1]將輸入的語音轉換成一序列的候選音節，將這些候選音節儲存起來，便於比較。我們

以候選音節中是否有出現正確音節的比率來衡量聲學處理單元對每一測試語料的正確性，這樣的比率我們稱為音節包含率(Inclusion Rate)。每個領域的測試語料大小及音節包含率如表 3 。

領域別	phi	lit	ball	win	cs
詞數	5,799	4,621	4,470	17,535	7,038
字數	8,407	6,269	6,011	24,271	11,223
音節包含率	99.82%	99.81%	99.75%	99.70%	99.88%

表 3 各測試文章所含有之字詞數及平均正確音節涵蓋率

實驗結果

表 4 是我們以一般及領域特定語言模型對各個領域的測試語料進行語言解碼所得到的字正確率。表中的各欄代表每一個不同領域測試語料對不同語言模型的結果，這些結果均是經過仔細調整式(1)中一般與領域語言模型的比重 d 而得到最佳辨認正確率的結果，我們也將每一應用領域的調整比重值列於最後一列中。

	phi	lit	ball	win	cs
一般 語言模型	82.52%	82.04%	81.43%	82.28%	87.04%
領域特定 語言模型	85.24%	87.03%	90.13%	87.54%	90.72%
調適比重 d	0.75	0.95	0.95	0.75	0.25

表 4 以一般及領域特定語言模型

對不同應用領域測試語料進行語言解碼所得到的結果

從上面的表中，我們可以觀察到對各個領域的測試語料，即使以一般語言模型進行語言解碼已可獲得不錯的效果，但以領域特定語言模型進行語言解碼，確實可以得到更好的效果，辨認正確率上昇的幅度從 2 到 8%。從上表中，我們也

觀察到一些有趣的現象：在這個實驗中，辨認正確率上昇幅度最高的是電子佈告欄的棒球類討論文章(ball)這個領域，經調適後，正確率可以從 81.43%上昇到 90.13%。經過仔細閱讀調適和測試語料，發現裡面內容與一般調適語料最大的不同是這類語料的文章主題(Subject)較狹隘，裡面包含了許多這領域內的術語，這些術語在調適和測試語料一再被提及，所以在經過調適之後，可以得到大幅度上昇的辨認正確性。

另外一個值得注意的現象是，科技報告文章領域(cs)的調整比重 d 遠比其他領域來的小，我們可以對照表 4 和表 2，發現這領域的語料量遠比其他領域小，因此可驗證了我們在前面的推論者，調整比重 d 與領域的語料大小有關。我們對這個現象進一步進行領域語料量與調整比重間關係的實驗，我們逐漸增加 cs 領域的語料量來觀察最佳辨認正確率與對應的比重，實驗結果列於表 5。從表 5 中，我們可以觀察到在語料量增加之下，語音辨認的正確率的確有增加的趨勢，同時，在訓練語料增加兩倍之後，調整比重已從傾向一般變成為傾向領域語言模型，可見得語料量的大小確實影響語言模型的調適與解碼效果。

語料量	5,612 字	10,768 字	17,373 字
辨認正確率	90.72%	91.31%	91.45%
調整比重 d	0.25	0.25	0.75

表 5 不同語料量大小與調整比重及辨認正確率間的關係

由上面這些實驗結果中，可以觀察對應用領域內的測試語料，以領域特定語言模型可以獲得更好的結果，同時增加語料量對語音正確率有顯著的幫助。因此如何自動判讀訓練語料的應用領域來訓練領域特定語言模型與語音辨認系統如何自動而迅速偵測使用者輸入語音的應用領域，便成為多領域語言模型的兩大問

題，在下面兩節中，我們將探討這兩個問題並試圖找出可能的解決方法。

三、多領域語言模型訓練

以多領域語言模型來進行語言解碼，我們首先要對每一個領域訓練一個領域的語言模型以及一個一般語言模型，然後以內插法組合一般與領域特定語言模型來訓練領域特定語言模型。從前一節的實驗中，我們可以瞭解到語料的增加可以進一步提昇這個領域內測試語料的辨認正確率，所以即使已經產生領域特定語言模型，若我們能蒐集到新的語料，最好能將它們歸類到最接近的應用領域，加入訓練語料中。在本節中，我們將描述多應用領域語言模型的訓練方法，尤其著重在判讀新語料的領域判讀上。

我們可以用圖 2 中的演算法來表示我們的多領域語言模型訓練方法。在初始的時候，在步驟 I-1 到 I-3，我們首先訓練每一個應用領域的詞雙連語言模型 $\boldsymbol{L_S}$ 與一般的詞雙連語言模型 $\boldsymbol{L_G}$，並以內插法組合一般與領域特定的詞雙連機率值，作為新的領域特定詞雙連機率值，所以在步驟 I-3 完成後我們對每一個應用領域有一個領域特定語言模型 $\boldsymbol{L_S}$'，已經可以進行多領域的語音辨認應用。而在系統已經使用之後，每當新收到一筆語料 T，這筆語料可以加入目前已有的領域，或另外成立一類新的應用領域。步驟 II-1 中首先以目前的領域語言模型來判讀 T 與哪一類應用領域的訓練語料較接近，判讀之後便可將將這筆語料加入最接近的應用領域 $\boldsymbol{S}$ 訓練新的領域特定語言模型 $\boldsymbol{L_S}$'(步驟 II-2)；如果新的訓練語料 T 與所有訓練語料都相差相當大，我們便以這筆語料新成立一類應用領域 $\boldsymbol{S}$，再訓練屬於該應用領域的詞雙連語言模型 $\boldsymbol{L_S}$'(步驟 II-3)。

```
I.初始:
    I-1. L_G = TrainingModel(T_G)
        // 以極大量語料 T_G 訓練一般領域語言模型
    I-2. Λ = {}
    I-3. for any domain S
            L_S = TrainingModel(T_S),   Λ = Λ ∪{L_S}
            // 對應用領域 S，訓練領域語言模型 L_S
            L_S' = Interpolation(L_G,L_S)
            // 組合產生領域特定語言模型
II.當新語料 T 進入:
    II-1. S = DomainDetermination(T,Λ)
        // 判讀新進入的語料 T，屬於哪一個領域
    II-2. if (S != NULL)
        // 找到合適的應用領域
            T_S += T, L_S = TrainingModel(T_S), L_S' = Interpolation(L_G,L_S)
            // 將 T 加入應用領域 S 中，產生領域特定語言模型
    II-3. else
        // 沒找到合適的應用領域
            T_S = T, L_S = TrainingModel(T_S), Λ = Λ ∪{L_S}
            L_S' = Interpolation(L_G,L_S)
            // 以 T 成立新應用領域 S，產生領域特定語言模型
```

圖 3 多領域語言模型訓練與調適的演算法

訓練語料所屬領域特定語言模型的判讀

判讀新語料屬於哪一個應用領域，事實上是一種文件分類(Document Classification)的問題。在過去的資訊檢索(Information Retrieval)研究中，已經提出相當多自動化的文件分類方法[18,19,20]，這些方法可以歸納成三個步驟。

一、選定文件的集合與分類的類別，並選定文件的全部或者具有較多資訊的部份(比方說，題目或摘要等等)作為訓練或判讀的資料，稱為文件的簡述(Profile)，

簡述可以是為是該文件的代用品(Representation)。

二、根據訂定的篩檢規則(如：詞類和頻率等等)，從簡述中挑選出關鍵詞(Keywords)來。

三、可以用向量模式(Vector Space Model)或機率模式(Probability Model)來估測文件與類別間的相關程度，以進行類別的判讀。

由此可知，這些方法多以關鍵詞為分類的主要資訊，換言之，這樣的方法只考慮關鍵詞本身的局部性資訊。在本論文中，我們嘗試利用詞雙連語言模型中包含的詞與詞間的關連性(Association)來判讀語料所屬的應用領域，希望能夠捕捉到更進一步的資訊，得到更好的判讀結果。下面簡介本論文所用的兩類詞關連性資訊，文字複雜度(Perplexity)和詞雙連涵蓋率(Word Bigram Coverage)，同時報告以各種領域特定語言模型量測在第二節實驗中所用各種領域測試語料的文字版本(Text Transcription)做為測試語料所得到的文字複雜度和詞雙連涵蓋率來探討這兩類資訊做為領域判讀的可能性。

1. 文字複雜度

在本論文中，我們以詞雙連語言模型對測試語料的文字複雜度(Perplexity) [7,8]的大小來衡量領域特定語言模型與新的訓練語料間的關連性。在定義文字複雜度之前，我們首先定義詞群語言模型在長度為 N 測試語料 W 上的熵值(Entropy)H_p 為

$$H_p = -\frac{1}{\sum_{i=1}^{N}|w_i|}\sum_{i=1}^{N}\log \Pr(w_i|w_{i-1}) \qquad ......(2)$$

這裡，w_i 代表測試語料 W 中出現的詞，$|w_i|$則是代表詞 w_i 的字數。H_p 可以作為評估不同語言模型對測試語料預測能力的強弱。H_p 值愈大，代表這樣的語言模型對此篇測試語料的預測能力愈強，反之亦然。為了方便表示出每個字平均後接字的數目，接著我們定義文字複雜度為

$$PP = 2^{Hp} \qquad ......(3)$$

平均文字複雜度的物理意義可認為是用語言模型來預測測試文章平均每個字後面可以接字的數目。所以觀察語言模型的文字複雜度可以判斷用來訓練語言模型的語料與測試語料字詞間的關連性。舉例而言，如果一個語言模型 L 對某一測試語料 W 有較小的文字複雜度，也就是說，利用 L 來預測 W 中任意給定的字，平均每個字的後接字數都很少，這樣意謂著 L 的訓練語料與 W 的字詞間有很大的關連性。反之則 L 的訓練語料與 W 可能沒有關連性。所以我們可以用訓練好的領域特定語言模型來判讀新的訓練語料是否與這個領域的其他訓練語料有關連。我們可以對每一個領域特定語言模型來量測新的訓練語料的文字複雜度，將新的訓練語料加入那些得到較小文字複雜度的語言模型所屬的領域中。

表 6 是我們以各種領域特定語言模型量測不同領域測試語料所得到的文字複雜度。表 6 中，由左而右的各欄分別是 phi 、 lit 、 ball 、 win 和 cs 各領域的測試語料在各種領域特定語言模型下所得到的文字複雜度，而每一列分別代表各種領域特定語言模型對不同應用領域測試語料的文字複雜度。

	phi	lit	ball	win	cs
L_{phi}	**352**	765	1067	765	521
L_{lit}	554	**301**	1174	781	587
L_{ball}	601	838	**37**	604	534
L_{win}	876	1239	1349	**402**	473
L_{cs}	1249	2221	2376	1812	**362**

表 6 以各種領域特定語言模型對不同應用領域測試語料的文字複雜度

2. 詞雙連涵蓋率

我們定義詞雙連涵蓋率如下：對一篇具有 m 對詞雙連對的語料 W,$B = \{b_1, b_2, \ldots b_i, \ldots b_m\}$，代表語料中所有詞雙連對所成的集合，其中 b_i 代表測試文章中第 i 對詞雙連對。 M 則表示這個領域中所有訓練語料內不同詞雙連對所成的集合。則詞雙連涵蓋率可定義如下:

$$\text{詞雙連涵蓋率} = 100\% \times \frac{\sum_{i=1}^{m} \chi_i}{|B|} \qquad \chi_i = \begin{cases} 1 & \text{if } b_i \in M \\ 0 & \text{otherwise} \end{cases} \qquad \text{......(4)}$$

由上面的定義，我們可以知道對一新語料 W 而言，詞雙連涵蓋率代表的是這個語料中的詞雙連對出現在應用領域訓練語料的比率。詞雙連對所代表的則是詞的前後相連關係，因此詞雙連涵蓋比率愈高，顯然新語料 W 中詞的前後關連性與這個應用領域愈像。所以，詞雙連涵蓋率可以用來作為新語料屬於哪一個應用領域的判讀標準。

表 7 是我們以各種領域特定語言模型量測不同領域測試語料所得到的詞雙連涵蓋率。表 7 中，由左而右的各欄分別是 phi 、 lit 、 ball 、 win 和 cs 各領域的測試語料在各種領域特定語言模型下所得到的詞雙連涵蓋率，而每一列分別代表各種領域特定語言模型對不同應用領域測試語料的詞雙連涵蓋率。

	phi	lit	ball	win	cs
L_{phi}	**37.28%**	31.39%	23.45%	28.62%	12.42%
L_{lit}	25.95%	**52.24%**	20.77%	27.89%	8.97%
L_{ball}	24.68%	29.28%	**99.70%**	35.55%	11.96%
L_{win}	13.73%	17.48%	15.93%	**40.28%**	12.94%
L_{cs}	5.02%	3.30%	3.30%	4.71%	**17.88%**

表 7 以各種領域特定語言模型對不同應用領域測試語料的詞雙連涵蓋率

應用領域判讀的討論

從表 6 和表 7 之中，我們可以觀察到每一應用領域測試語料在對應的領域特定語言模型下所得到的文字複雜度都比其他領域特定語言模型所得到者來得低；詞雙連涵蓋率中也有類似的現象發生，每一應用領域測試語料在對應的領域特定語言模型下的詞雙連涵蓋率都比其他領域特定語言模型來得高。我們可以驗證前面所推論者，新進入的語料在領域特定語言模型下得到較低的文字複雜度和較高的詞雙連涵蓋率，語料與這個應用領域間應存有某種關連性。這樣的結果證實我們可以用文字複雜度和詞雙連涵蓋率來作為應用領域判讀的資訊。

從以上兩表中，我們也可以觀察到一些有趣的現象。首先我們可以看到 ball 領域的測試語料對 ball 的領域特定語言模型 L_{ball} 有非常低的文字複雜度與非常高的詞雙連涵蓋率，這可以進一步驗證我們在前一節中所觀察的現象，我們所收集到的電子佈告欄的棒球討論區語料裡存在許多的術語，這些術語一再出現的結果造成 L_{ball} 對這領域的測試語料有非常低的文字複雜度與非常高的詞雙連涵蓋率。

另一個觀察是 win 和 cs 兩個領域都是有關電腦技術方面的領域，在這些領域中是主題最接近的兩個領域。這樣的關係對文字複雜度和詞雙連涵蓋率的量測

有什麼影響呢？以 win 的領域特定語言模型 L_{win} 來測量 cs 領域的測試語料得到在所有領域特定語言模型僅次於本身領域特定語言模型 L_{cs} 的效果，這再次的證實了我們以文字複雜度和詞雙連涵蓋率作為應用領域判讀資訊的可行性。反過來，雖然以 cs 的領域特定語言模型 L_{cs} 來測量 win 領域的測試語料得不到這樣好的效果，但造成這種現象的主要原因是由於 cs 領域的訓練語料量並不充足，所以我們無法得到非常可靠的語言模型估計。

在組合兩種資訊之後，我們以測試語料來進行應用領域判讀，在總共 20 篇的測試語料中，所有由自動判讀的結果都與原先人工分類者相同。雖然實驗的測試語料太少了，所以可以得到這麼好的結果。但是以文字複雜度與詞雙連涵蓋率作為應用領域判讀的優點也可略見其端倪。

四、使用多領域語言模型之語言解碼

在我們的多領域語言模型之語言解碼方法，一開始我們以所有應用領域的領域特定語言模型進行語言解碼，逐漸排除不可能的應用領域，一直到最後只剩少數一、二個可能的領域。選取應用領域所依據的資訊是以對應的領域特定語言模型對輸入語音進行語言解碼所得到的分數。圖 3 是我們的多領域語言模型語言解碼方法。在使用者輸入語音到多領域的國語語音辨認系統時，前面幾句我們先以所有領域特定語言模型來偵測這些語音最接近於哪些應用領域(步驟 I-2)。語言解碼單元在接受聲學處理單元比對出的候選音節序列 Σ 後，在步驟 II-2 時，以每一個領域特定語言模型對 C 進行語言解碼，從中搜尋出最有可能的字串。在進行完所有領域的語言解碼之後，步驟 II-3 裡，選取其中解碼分數最高的字串 C_{S^*}，

將這條字串輸出(步驟 II-4)，同時記錄下這條字串所屬的領域 S^*(步驟 II-5)。在蒐集足夠預測接下來輸入語音的可能應用領域資訊後，便可以目前可能應用領域的領域特定語言模型提供接下來語言解碼所需的文法限制之用(步驟 III-1)，完成語言模型調適。

```
I. 初始：
    I-1. Λ = {}, Λ' = {}
        // 將解碼集合 Λ 和修正領域集合 Λ'設為空集合
    I-2. for all L_S'
            Λ = Λ ∪ {L_S'}
            // 將系統內所有領域的調適後領域特定語言模型 L_S'加入 Λ 中
II. 使用者輸入語音 U
    II-1. Σ = AcousticMatching(U)
        // 聲學處理單元將 U 辨認為候選音節序列 Σ
    II-2. for all L_S' in Λ
            (C_S,P_S)= LinguisticDecoding(Σ,L_S')
            // 以調適後領域特定語言模型 L_S'對 Σ 進行語言解碼，得到字串
              C_S與解碼分數 P_S
    II-3. select S* , such that P_S* ≧ P_S for all L_S' in Λ
        // 找出分數最高的領域 S*
    II-4. output C_S*
        // 輸出由 L_S*'解碼出的字串 C_S*
    II-5. Λ' = Λ' ∪ {L_S*'}
        // 更新修正領域集合 Λ'
III. 經過數句之後：
    III-1. Λ = Λ'
        // 以修正領域集合 Λ'更新解碼集合 Λ，調適語言模型
```

圖 3 多領域語言模型語言解碼的演算法

多領域語言模型語言解碼的實驗

這個實驗中，我們以前面實驗所用的五個應用領域來進行多領域語言模型的

語言解碼，討論語音所屬應用領域的偵測與比較一般語言模型與應用多語言模型語言解碼得到的辨認正確性。首先來看看輸入語音的句數與應用領域偵測正確性的關係，我們以本節中所提出的利用多領域語言模型的語言解碼方法，對 20 篇測試語料進行語言解碼，記錄下每一次輸入語句後所得到的前三個偵測到的應用領域，加以統計，圖 4 所示是以輸入句數為橫座標，應用領域偵測的正確率為縱座標所得到的關係圖。從圖 4 可以看出前三個偵測的結果就已經相當不錯。在本圖中，我們可以同時觀察到輸入句數與應用領域偵測的正確率有正面的影響，也就是說，輸入的句數愈多，可以偵測到愈正確的應用領域。這個結果證實了本節所提出的利用多領域語言模型的語言解碼方法具有語言模型調適的能力。

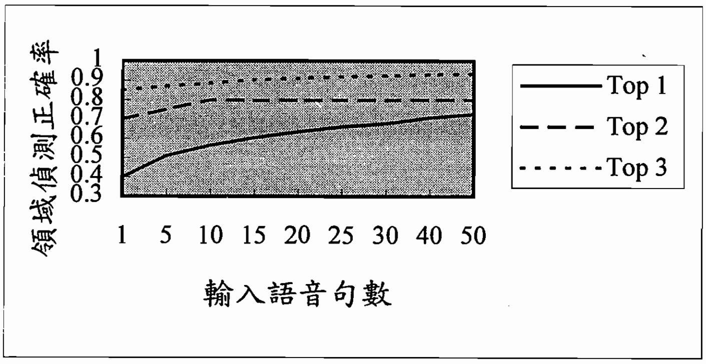

圖 4 輸入語音句數與應用領域偵測正確性的關係圖

接下來我們來看看實際應用這個多領域語言模型的語言解碼方法所得到的語音辨認效果。表 8 是對各種應用領域的測試語料所得到的字正確率，對照於在第二節中的表 4 中以一般語言模型進行語言解碼所得到的字正確率，可以看出這個多領域語言模型的語言解碼方法確實有助於辨認正確率的提昇。此外，我們可以發現在某些訓練語料相當缺乏的領域中，使用多領域語言模型甚至可以得到比已知應用領域時直接採用領域特定語言模型還要好的辨認結果，因為我們可以利

用其他主題較為接近的應用領域所訓練出來的語言模型來彌補某些沒有出現在這個領域訓練語料中的語言特性。

測試語料的領域	phi	lit	ball	win	cs
語音辨認字正確率	84.43%	85.68%	88.70%	87.13%	91.13%

表 8 利用多領域語言模型的語言模型調適法進行語音辨認得到的字正確率

五、結論

在本論文中，我們提出適用於不同領域間中文語言模型的自動訓練、偵測與調適方法。在多領域語言模型的訓練上，我們根據文字複雜度與詞雙連涵蓋率來判讀新訓練語料的應用領域，然後對每一個應用領域利用領域內的訓練語料訓練一個語言模型，再以內插法組合這個語言模型與由不分領域的大量語料訓練的一般語言模型，得到這個應用領用的領域特定語言模型。在進行語音辨認時，系統自動偵測輸入語音選擇最適合的領域特定語言模型，作為調適的語言模型，來提供語言解碼所需的文法限制。我們以利用領域特定語言模型進行語言解碼得到的分數作為偵測應用領域所需的資訊。應用這些方法在極大詞彙國語語音辨認的語言解碼方法將可以為輸入的語音選擇合適應用領域的中文語言模型，對訓練語料不足的特殊領域語音辨認可以進一步提昇辨認正確率。

目前我們利用多領域語言模型作為語言模型已經得到一些初步的結果，值得我們繼續發展這類技術的研究。在未來的方向上，我們希望能夠繼續提昇應用領域偵測的準確性，同時在多領域的應用環境中，需要更精簡的語言模型，結合前人在語言模型調適中所發展出來的技術，諸如：短期記憶體、觸發對和最小區辨資訊等等，希望能夠提供更有效的語言模型調適技術，找出更符合在應用領域中

的語言特性。另外，利用文字複雜度和詞雙連涵蓋率等詞關連性資訊可以得到相當令人滿意的應用領域判讀結果，我們希望擴展這類的研究到文件分類的技術上。

參考文獻

[1] H-W. Wang, et. al., "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data," in Proceedings of ICASSP'95, pp. 61-64, Detroit, USA, 1995.

[2] Y-J. Yang, S-C. Lin, L-F. Chien, K-J. Chen, and L-S. Lee, "An Intelligent and Efficient Word-Class-Based Chinese Language Model for Mandarin Speech Recognition with Very Large Vocabulary," Proc. ICSLP'94, pp. 1371-1374, Yokohama, Japan, 1994.

[3] Y-C. Chang, S-C. Lin, L-F. Chien, K-J. Chen, and L-S. Lee, "Methodology, Implementation and Application of Word-Class Based Language Model in Mandarin Speech Recognition," Proc. ROCLING VII, pp. 17-34, Hsinchu, ROC, 1994. (in Chinese)

[4] S-C. Lin, L-F. Chien, K-J. Chen, and L-S. Lee, "Unconstrained Speech Retrieval for Chinese Document Databases with Very Large Vocabulary and Unlimited Domains," Proc. EUROSPEECH'95, pp. 1203-1206, Madrid, Spain, 1995.

[5] C-H. Lee, "Stochastic Modeling in Spoken Dialogue System Design," Speech Communication, Vol. 15, pp. 311-322, 1994.

[6] R. Cole, et. al., "The Challenge of Spoken Language Systems: Research Directions for the Nineties," IEEE Trans. Speech and Audio Processing, Vol. 3, No. 1, pp.1-20, 1995.

[7] F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," in Proceedings of IEEE, Vol. 73, No. 11, pp. 1616-1624, Nov. 1985.

[8] K.F. Lee, *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*, Ph.D. Dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, Apr. 1988.

[9] Speech and Natural Language: Proceedings of the ARPA Workshop, Morgan Kaufmann Publishers, CA, USA, 1994.

[10] B.Suhm and A. Waibel, "Towards Better Language Models for Spontaneous Speech," in Proceedings of ICSLP'94, Vol. II, pp. 831-834, Yokohama, Japan,

1994.

[11] M. McCandless and J. Glass, "Empirical Acquisition of Language Models for Speech Recognition," in Proceedings of ICSLP'94, Vol. II, pp. 835-838, Yokohama, Japan, 1994.

[12] S. Matsunaga, T. Yamada, and K. Shikano, "Task Adaptation in Stochastic Language Models for Continuous Speech Recognition," in Proceedings of ICASSP'92, Vol. I, pp. 165-168, San Francisco, California, USA, 1992.

[13] R. Kuhn and R. D. Mori, "A Cache-Based Natural Language Model for Speech Recognition," IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. PAMI-12, No. 6, pp. 570-583, Jun. 1990.

[14] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-Based Language Models: A Maximum Entropy Approach," in Proceedings of ICSSP'93, Vol. II, pp. 45-48, Adelaide, South Australia, 1993.

[15] S. D. Pietra, V. D. Pietra, R. L. Mercer, S. Roukos, "Adaptive Language Modeling Using Minimum Discriminant Estimation," in Proceedings of ICASSP'92, Vol. I, pp. 633-636, San Francisco, California, USA, 1992.

[16] R. Rosenfeld, "A Hybrid Approach to Adaptive Statistical Language Modeling," in Proceedings of Human Language Technology Workshop, pp. 76-81, 1994.

[17] P.S. Rao, M. D. Monkowski, and S. Roukos, "Language Model Adaptation via Minimum Discrimination Information," in Proceedings of ICASSP'95, Vol. I, pp. 161-165, Detroit, Michigan, USA, 1995.

[18] R. R. Larson, "Experiments in Automatic Library of Congress Classification," JASIS, Vol. 43, No. 2, pp. 130-149, 1992.

[19] D. D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," in Proceedings of SIGIR'92, pp. 37-50, Copenhagen, Denmark, 1992.

[20] P. S. Jacobs, "Using Statistical Methods to Improve Knowledge-Based News Categorization," IEEE Expert, Vol. 8, No. 2, pp.13-23, Apr. 1993.