中文語料庫構建及管理系統設計

馬偉雲　謝佑明　楊昌樺*　陳克健

中央研究院資訊科學研究所

{ma, morris, kchen}@iis.sinica.edu.tw

ms8903@cis.scu.edu.tw*

摘　要

一個中文帶詞類標記的平衡語料庫，在中文自然語言的研究與應用上是不可或缺的角色，然而要構建一個數量大且高品質的語料庫往往需要投入大量的人力及時間，為了提升構建的效率以及提高語料庫管理的機能，在管理方面，我們建立了以文本為單位的資料庫系統作為語料庫的架構，並開發一管理介面。構建方面，我們設計了一套構建流程以及開發了四個子系統來幫助我們完成構建語料庫的工作。構建語料庫的第一步是語料蒐集，為此我們設計了一個語料蒐集介面，能夠蒐集網路上豐沛的電子文件資源，並在某些特定網址來源當中自動分析其文本格式。第二步是語料的斷詞及標記，我們透過未知詞擷取模組作為斷詞標記的前處理，大幅提高了斷詞標記程序的正確性，減少其後人力校正的負擔。最後一步是人工檢驗，我們設計了操作簡便的人工檢驗介面，並結合詞典與舊版本的語料庫提供使用者參考來做出正確的判斷，完成斷詞、詞類與句子的編修工作。

1. 簡介

語料庫為本（corpus-based）的研究是近年來語言學及計算語言研究的一個重要發展〔Church,Mercer93〕、〔Chen94〕、〔Huang95〕。

建立帶詞類標記的平衡語料庫是一個浩大的工程，但也是自然語言研究的基礎工程（infrastructure)。其效應可由現存語料庫，如布朗，LOB，London-Lund 等所衍生的大量研究成果得到證明。

「中央研究院平衡語料庫」簡稱「中研院平衡語料庫」(Sinica Corpus)，是世界上第一個有完整詞類標記的漢語平衡語料庫。於 1994 年公開提供給國內外學術研究使用；以期在使用過程中得到回饋，在完成目標規模前可以做必要的修正。1997 年開放的研究院語料庫 3.0 版已經達到五百萬目詞的預計規模。我們期望在 2003 年能夠達到一千萬目詞的規模。

建構一個中文的平衡帶詞類標記的語料庫，包括語料的收集、語料的整理(包含語料清潔、為語料分類、加詞類標記等等)、人工的校定。從早期的建構經驗中，由於缺乏合適的工具，我們遭遇了以下的困難：(1）早期我們以檔案的形式作為語料的最小單位，一份檔案通常包含數十篇不同的文本，文本的格式屬性以符號配合文字在文本之前表示之。這樣以檔案為單位的架構對整體語料的管理及統計相當不便，同時對人工校對的工作分配而言，也比較沒有彈性。(2)大量語料的蒐集、維護、分類、校定交由各人以檔案的方式加以處理，並無統一的處理介面，形成管理上的紊亂。(3)過去我們使用詞庫小組自行開發的系統〔Chen,Liu92〕將語料加以斷詞標記，卻發現由於文本當中未知詞的存在，使得系統的斷詞表現大幅下降，而必須倚靠事後大量的人力來加以合分詞。(4）人工校正時，由於斷詞以及詞類標記時常有歧異現象發生，校正者沒有工具立即檢驗相關的用法或範例，造成判斷上的困難，使得有時候斷詞標記的校對品質因人而異。這些問題除了造成在管理上的困難之外，同時在人工校正的過程中花費了大量的人力及時間，在斷詞標記的一致性上也不易維持。

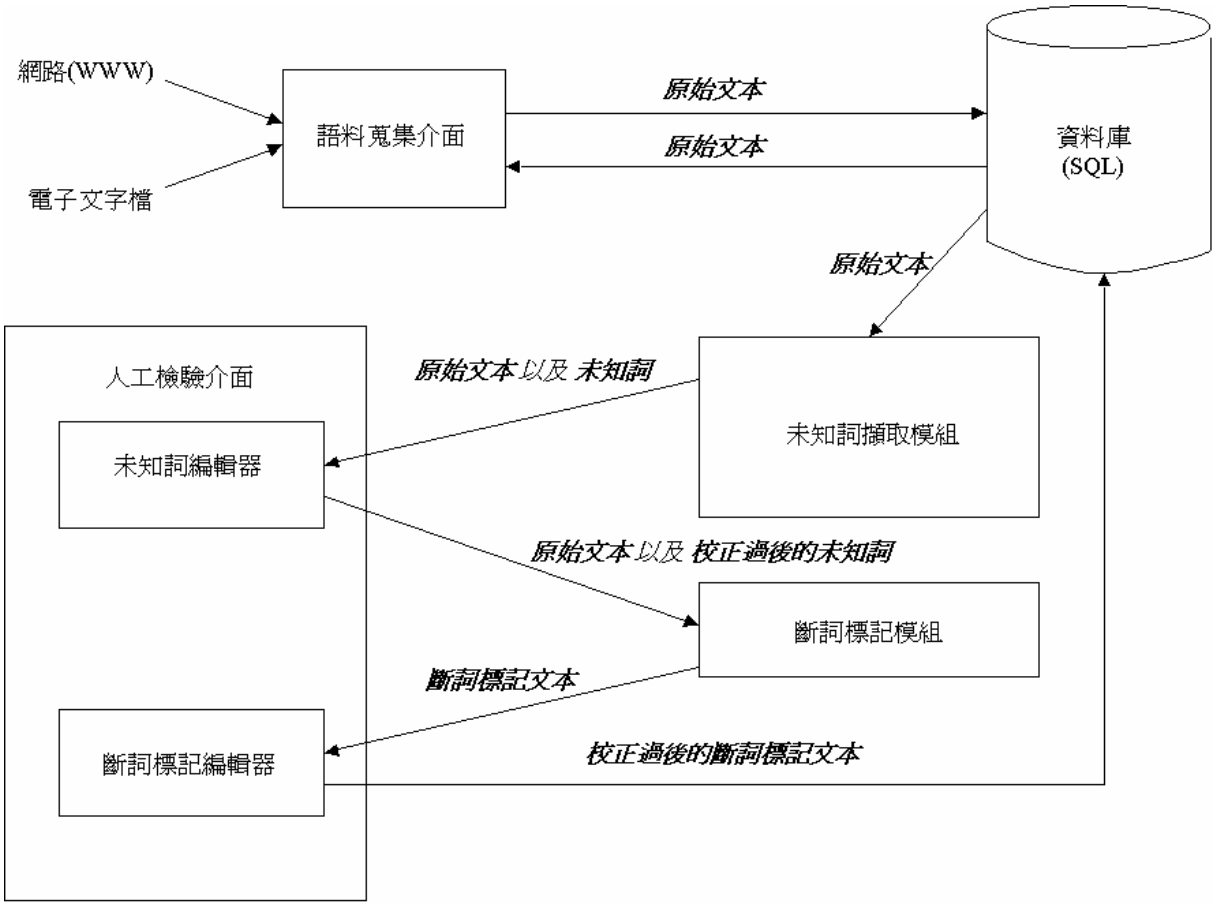
因此，我們認為要有效率的建構一個高品質的平衡語料庫，必須要訂定出一套嚴

謹的管理方式，發展出一套理想的構建流程以及開發一些合適的工具，才有可能達成。在語料的管理上，我們捨棄以往以檔案為單位的管理方式，而改用文本為單位，並採用資料庫（database）的架構儲存，一篇文本即為資料庫的一筆記錄（record），其格式屬性以此記錄的欄位（field）加以儲存。在語料的蒐集上，除了直接自相關單位取得語料之外，網路上的各類文本也是一重要來源，為此我們設計了一套語料蒐集介面，能夠將網址上的文本擷取下來，在某些特定網址上能自動分析其文體作者等格式，減輕人力的負擔。在語料的斷詞標記方面，未知詞的存在使得斷詞標記模組的錯誤率大增，因此，我們提出一改善的作法，在斷詞標記之前，先利用自行開發的未知詞擷取模組來擷取文本的未知詞，之後再加以斷詞標記，如此使得斷詞的品質大幅度的提高。最後是人工檢驗的部分，為了確保語料庫的品質，有必要將語料庫做一地毯式的人工檢驗，我們在 windows 環境下開發了一套由滑鼠操作的人工檢驗介面，除了操作簡便快速之外，最大的特色在於系統連結了詞典資料庫以及之前版本的語料庫，讓檢驗者能快速的查詢詞典中相關的資訊，也可以在先前版本的語料庫當中查詢相關的範例，來幫助檢驗者做出正確的判斷。

本文第二節將詳述建立平衡語料庫的系統設計架構及流程。第三節對系統在實務上的表現上作一討論。最後是結論。

2. 建立平衡語料庫的系統設計

本系統主要包含四個子系統，分別是語料蒐集介面、未知詞擷取模組、斷詞標記模組、以及人工檢驗介面。其中語料蒐集介面以及人工檢驗介面是在 windows 的環境，而未知詞擷取模組及斷詞標記模組是在 linux 的環境下運作。子系統間相互的溝通均是採用 client-server 的模式。系統流程圖如圖一。



圖一：系統流程圖

首先語料蒐集介面將使用者所指定的網址抓回原始文本(text)，並針對特定網址自動分析其文體作者等格式，經過使用者檢驗無誤後，將原始文本存入資料庫，準備後續斷詞及詞類標記工作，由於文本中包含各式各樣的未知詞，造成斷詞及詞類標記的困擾，因此後續處理的第一步是先經過未知詞擷取模組擷取未知詞，再將原始文本及抽取出的未知詞送交斷詞標記模組將原始文本加以斷詞標記，最後以人工逐句檢查，經過此一連串的程序以及最後的人工檢驗，得到斷詞標記後的文本，可以存入資料庫當中作為語料。

這樣的設計流程具有如下的優點：(1) 未知詞擷取模組在斷詞標記模組之前即已將未知詞標示出來，提高斷詞標記的正確率，降低人力修正的負擔。(2) 針對同一篇文本，蒐集語料和人工檢驗可以由不同的人擔任，這是因為這兩項工作所需的專業知

識不盡相同，因此在設計上以資料庫為中繼站將兩者分離，在人力的分配上這樣的設計提供了更大的彈性。(3) 自動化的處理提高語料處理的效率及一致性。(4) 人工檢驗介面提供完整的詞典以及舊有的標記訊息和範例，幫助使用者做判斷，提高資料的正確性。

2.1 語料蒐集介面

隨著網際網路的蓬勃發展，大量文本也以電子化的形式呈現，每個人可以輕易地閱讀這些文本，不受空間及時間的限制。從蒐集語料的觀點出發，也可以善用環境中的這項特點，設計一個使用介面，在網路環境裡獲得語料，標記屬性特徵資訊，將成果以資料庫形式儲存。

早期蒐集語料的方式是以電子文字檔為主，必須事先定義其內容格式，之後使用者將大量的語料以文本檔案的形式加以分類管理，再由程式員撰寫程式進行各項統計查詢的動作。面對數量眾多的語料，使用者必須自行紀錄工作的流程與進度，修正錯誤時也必須用傳統的方法找到該錯誤檔案再加以修正，容易造成版本不統一的問題。

為了改善這些問題，使蒐集語料的工作更有效率，我們設計的介面，必須符合下列二個要求：(1)統一版本及減少人工時程(2)利用網路改善工作流程。為了達成以上兩個要求，語料蒐集介面提供以下功能：(1)文本擷取(2)修改屬性特徵(3)自動分類資料擷取。

2.1.1 統一版本及減少人工時程

在蒐集語料的過程中，使用者除了找到文本之外，還必須判斷其屬性特徵加以人工標註，包括文類、文體、語式、主題、媒體、作者姓名、性別、國籍、母語、出版單位、出版地、出版日期、版次〔Chen94,Huang95〕。系統擷取語料時會事先自動判別文本的相關屬性特徵，當無法擷取到所有的屬性特徵時，則以視窗形式呈現的介面（如圖二所示），提供了線上修改的功能，使用者可直接在圖形介面上進行屬性特徵的

修改，避免文書編輯上的困擾，也能統一修改的格式與製作出來的版本，減少對單一文本處理的時程。

圖二：語料蒐集視窗程式介面

2.1.2 利用網路改善工作流程

目前網際網路上的電子文件普遍以 HTML 格式存在，使用者可透過本介面直接輸入目的網頁之網址，由本系統過濾掉不必要的 HTML TAG，呈現本文部分供使用者審核編輯；並由事先定義好、並撰寫於程式中的規則，判斷該文本的屬性特徵，再由使用者加以確認及修改，如此可減少繁瑣的文字檔案編輯工作；這些事先定義好的規則有賴於程式員對目的網頁的原始結構加以分析。如圖二所示之介面，使用者至中時電子報（http://news.chinatimes.com/）擷取新聞文本；程式員根據中時電子報 HTML 內文格式

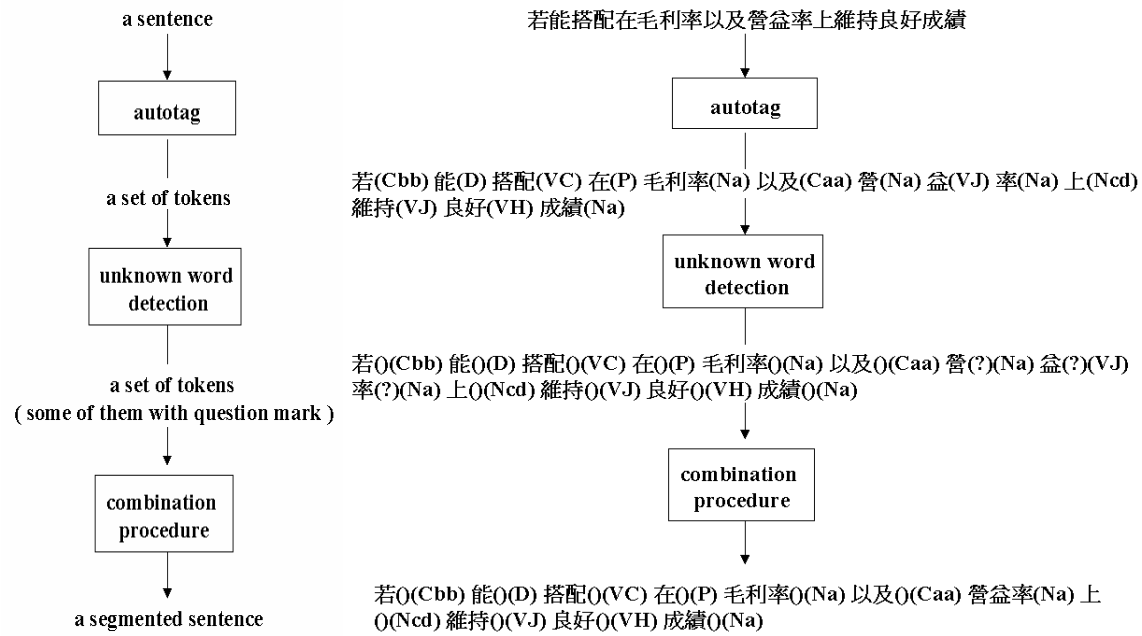
撰寫擷取的程式碼，使用者處理好語料文本後，該文本將直接透過本介面存進資料庫，讓後續的語料庫標記程式能夠直接銜接，不必再進行文字檔的轉換工作。

2.2 未知詞擷取模組

從以往到現在處理斷詞及詞類標記的過程中，大多數的錯誤來自於未知詞的出現，例如：人名、縮寫、複合詞等在文本中出現的機率相當高，自動斷詞及標記是以詞典資料為本，因此只要有一個未知詞出現就會造成斷詞及標記的數個錯誤。

目前大多數的未知詞擷取作法與斷詞程序相連結，先將文本做斷詞，未知詞由於詞典未收錄，故會被斷成切割過短的單字或字組，之後再由統計式〔Sproat90〕、〔Smadja93,96〕、〔Wu93〕、〔Chang97〕或法則式〔Yeh91〕、〔Lin93〕的技術加以合併這些短字組成為未知詞。

本系統所採用的未知詞擷取模組的演算法大體上亦採取這樣的概念，除了與斷詞程序相連結之外，在斷詞完畢後，會再經過一未知詞偵測程序〔Chen,Bai97〕，決定那些單字是未知詞的一部份，那些是本來就能夠獨用的單字。屬於未知詞一部份的單字之後能夠啟動合併程序，有機會結合前後字組合併成未知詞。如圖三。



圖三：未知詞擷取模組流程圖及範例

當一個句子經過斷詞以及未知詞偵測程序後，未知詞會被切分為較短的成分，正常的單字詞會被上下文規律區分出來，因此我們可以區分出何者屬於正常的單字詞，何者屬於未知詞的成分，之後再利用統計及構詞律等合併程序得到未知詞，並根據未知詞的內部結構猜測其詞類〔Chen,Bai,Chen97〕。

如圖三的範例所示，當輸入的句子「若能搭配在毛利率以及營益率上維持良好成績」經過斷詞標記後得到「若（Cbb） 能（D） 搭配（VC） 在（P） 毛利率（Na） 以及（Caa）營（Na） 益（VJ） 率（Na） 上（Ncd） 維持（VJ） 良好（VH） 成績（Na）」，接下來經過未知詞偵測程序，判斷出「營（Na）」、「益（VJ）」、「率（Na）」可能是未知詞的成分（以問號標示），經過合併程序將之合併成為未知詞「營益率」，再分析「營益率」的內部結構猜測其詞類為「Na」。

在合併程序的部分，我們同時採用統計式以及法則式的技術，統計方面，不同於前人從整體語料獲得詞彙組成的統計資訊，而是從單篇文本中獲得統計資訊，這是因為我們認為大多數的未知詞在單篇文本的統計比在整體語料的統計上更有意義。

法則式的部分，我們利用構詞學與構句學的理論以及觀察到的現象，加以訂定若干合併的規則及限制，例如某些詞性標籤或是如百家姓的訊息等等都是法則所規範的對象。

2.3 斷詞標記模組

經過未知詞擷取模組之後，雖然已經可以以一個斷詞完成並標記好的句子形式呈現，然而所擷取出來的未知詞當中仍有錯誤發生的可能，因此有必要以人工校定的方式直接增刪或修改這些未知詞，再利用斷詞標記模組參考這些人工確認後的未知詞，將原始文本重新斷詞標記，得到較為正確的結果。

因此未知詞擷取模組將未知詞傳送給人工檢驗介面，如「營益率（Na）」。由人工判斷是否要修改或增刪這些未知詞。之後再將這些正確的未知詞以及原始文本傳送給斷詞標記模組。

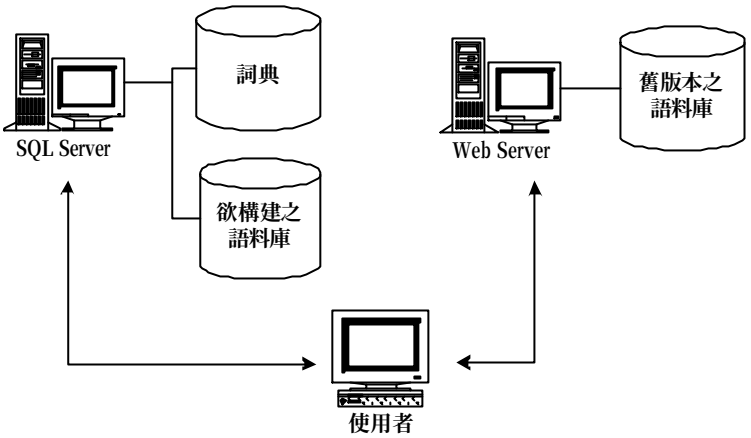
我們的斷詞標記模組是以詞典資料為本，因此當有了某一文本的未知詞資訊後，立即動態產生一部未知詞詞典的資料結構來幫助我們斷詞標記，我們可視作新增了一部未知詞的詞典和原有詞典搭配，在斷詞標記的程序中同時參考這兩部詞典，來完成這一文本的斷詞標記工作。

2.4 人工檢驗介面

經由前面章節的介紹，我們瞭解到如何透過網際網路收集 WWW 上的文本做為語料，再經由未知詞擷取模組及斷詞標記模組將文本加以斷詞標記。以上動作均採用自動化處理的方式，不僅有效率而且保持了斷詞標記的一致性，達到百分之九十五以上的正確結果。然而，在斷詞及詞類標記的過程中，因為有切分歧義、標記歧義與未知詞判斷的困難，所以，標記出來的詞類結果難免會有些許錯誤的地方。為了達到高品質的語料庫，人工檢驗的過程還是不能避免。因此，需要一個系統來輔助處理斷詞標記後的檢驗動作。該系統需提供簡單的介面與簡易的操作方式，來節省人力及時間。並能夠取得已有的詞典、語料標記資訊做為參考，來幫助使用者完成斷詞、詞類與句子的編修。

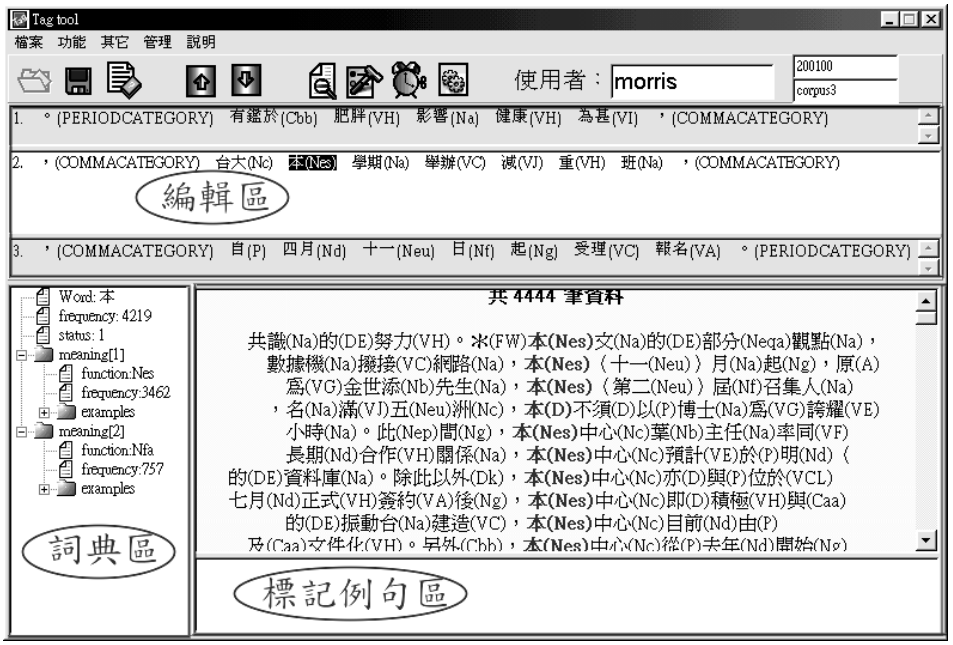
2.4.1 系統功能

本系統的開發平台是在微軟作業系統上。使用者可以在任何一台電腦上，透過區域網路的連線，連上後端語料庫及詞典的資料庫伺服器，來選取欲編輯的文本與查詢某詞的詞典資料等。該後端資料庫伺服器是以微軟 SQL Server 所建構，系統並連結以 Web Server 建構的舊版本語料庫，提供相關例句供使用者參考。如圖四所示。



圖四：系統與資料庫分佈架構

語料編輯介面規劃三個區域來表示這些訊息，分別是編輯區、字典區與標記例句區，所呈現的畫面如圖五所示。



圖五：人工檢驗畫面

現在就分別對這三個主要區域加以描述：

(1) 編輯區

該區所編輯的對象是斷詞標記後的文本，經由使用者選取文本後載入所呈現出來的。在這裡可以看到有三個小區域，分別代表上一句、編輯句與下一句，分別以不同

的顏色與字體來區分，這樣劃分的目的在提供使用者瞭解編輯過與將要編輯的句子內容，好讓使用者盡量達到詞類標記的一致性。

當使用者要進行修改時，所有的功能，都可以在主功能表中看到。我們將常用的部份以快速鍵或快速按鈕表示，以增加使用者編輯的效率。使用者可以在編輯列中利用滑鼠去選取欲處理的詞或是配合鍵盤的左右鍵去選取，然後配合快速鍵或是按下滑鼠右鍵依不同目的去選擇欲執行的項目。這些功能可以分類如下：

(a)詞的修改：包含了合分詞、改詞類、改特徵、去特徵等功能。其中，在合分詞的部份，系統會列出合分詞後的詞有哪些詞類，供使用者選擇。倘若只有一種選擇，系統會自行加入，盡量不讓使用者自行輸入，避免往後詞類標記不一致的狀況發生。在改詞類的部份，作法亦同上述的合分詞。

(b)句子修改與詞字數統計：在斷詞標記的過程中，有可能會有不正確的斷句結果產生，這時就需要合分句的功能。另外，使用者可以統計到目前為止文本當中已檢驗了多少個詞與字，系統本身也會在文本完稿之後重新編排句子序號並加以統計該文本的總詞字數，以提供使用者參考。

(c)記憶功能：在編輯的過程中，使用者會遇到某種錯誤可能一再的發生，並不希望每次都去改相同的錯誤。因些，系統提供了規則檔記錄的功能，紀錄使用者合分詞與改詞類的過程，當使用者對某詞的詞類討論出最後的定論，即可根據規則檔的記錄將目前的編輯句自動作修正。系統也提供批次更新的功能作全文修正。批次更新的處理對象是全文內容，在全文當中找出符合規則的詞並加以列出，供使用者選擇是否更新取代之。

(d)提醒的功能：目的在告知使用者目前編輯句中，有哪些詞是歧義切分詞或詞綴，以不同的方式加以顯示，讓使用者注意到在處理該詞時要特別小心。

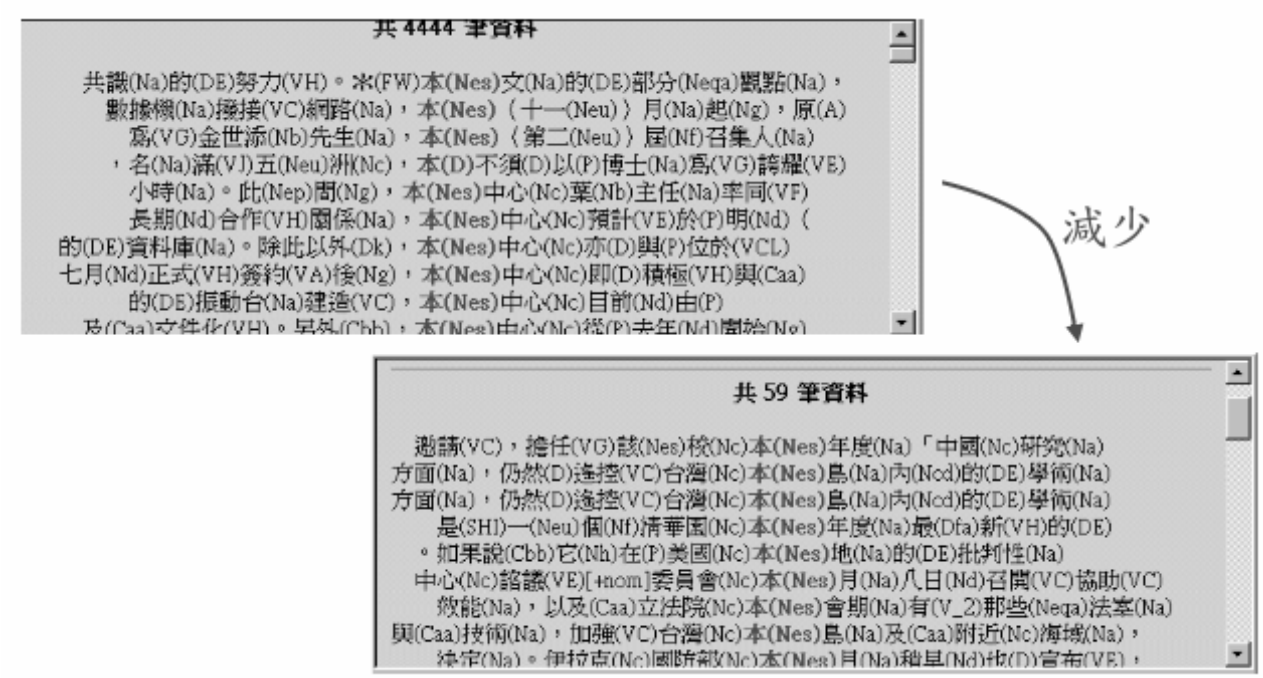
(2) 字典區

在資料庫伺服器中，我們建立了一個以詞為基礎的詞典資料庫，使用者可以輸入一個欲查詢的詞，系統會回應使用者該詞的詞類有哪些？頻率是多少？相關例句有哪

些？我們將這些結果顯示在字典區，以樹狀結構的方式呈現出來，讓使用者清楚地了解到該詞有多少詞類，以更確定詞類標記的對錯。

(3) 標記例句區

在這裡，我們整合了中央研究院開放出來供各界使用的網頁版本語料庫查詢介面(http://www.sinica.edu.tw/ftms-bin/kiwi.sh)於系統當中。該語料庫的語料平衡分配在不同的主題和語式上，總共約有五百萬目詞之多，為現代漢語無窮多的語句中一個具代表性的樣本。該網頁版的語料查詢系統，提供使用者多樣式的查詢，在顯示的結果中，包含每個文句與其斷詞標記後的結果。本系統將查詢的步驟簡化，讓使用者只要在本系統中的編輯句裡，先點選欲查詢的詞，再按下滑鼠右鍵選擇查例句，就會列出與該詞相關的句子，且每個例句都是標記過的。如此一來，使用者可以參考這樣的結果來進行詞類標記的判斷。此外，有時為了避免過多的例句產生，系統提供了前後詞類的相關限制功能，以減少輸出，讓查詢到的結果更符合我們的需要，同時節省查詢時間。如圖六所示，以「本」作為查詢的關鍵字可得到 4444 筆例句，若限制前後的詞類為「Nc」及「Na」，查詢的例句數目則大幅精減為 59 筆。



圖六：精減相關例句查詢結果

2.5　整合後的語料庫構建管理系統介面

本系統以視窗介面整合了斷詞標記模組、未知詞擷取模組與人工檢驗介面，提供使用者文本選取與查詢介面、未知詞編輯介面、人工檢驗介面與語料類別修改介面，針對這四個介面內容描述如下：

(1) 文本選取與查詢介面

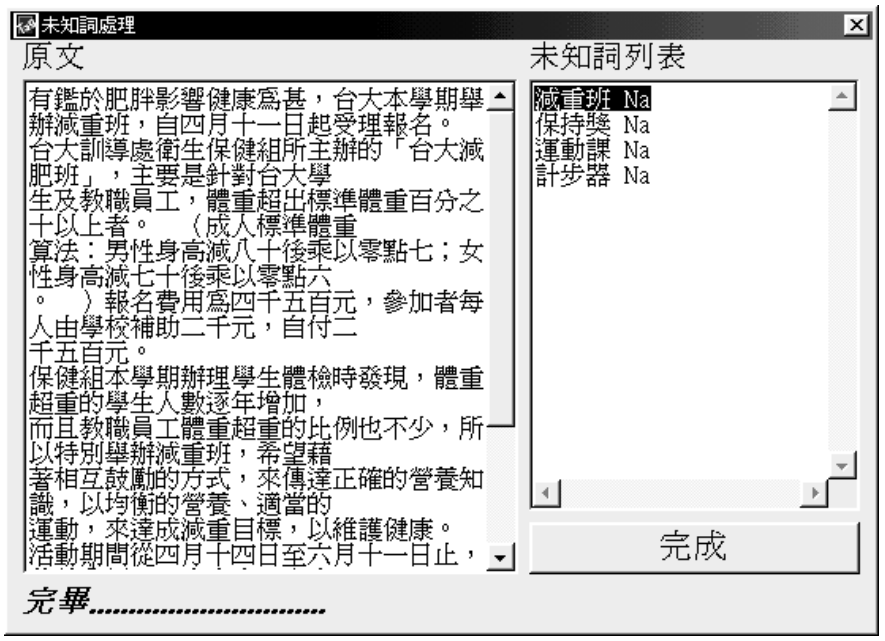
透過 2.1 節所描述的語料蒐集介面所抓來的文本，都會放在後端的語料庫伺服器中。當使用者欲檢驗文本時，可透過本介面以條件式的查詢方式從語料庫伺服器中取出。另外，系統也提供使用者以檔案的型式處理。參考畫面如下：

圖七：文本選取與查詢介面

(2) 未知詞編輯介面

在文本確定之後，倘若該文本已有斷詞標記過的資料，會直接取出進行上次未完的編輯動作。反之，系統會先將此未標記過的原始文本進行未知詞擷取，使用者可針

對擷取出來的未知詞進行新增、刪除及修改等編輯動作，如圖八所示。再將原始文本及編輯過後的未知詞送到斷詞標記模組加以斷詞標記。



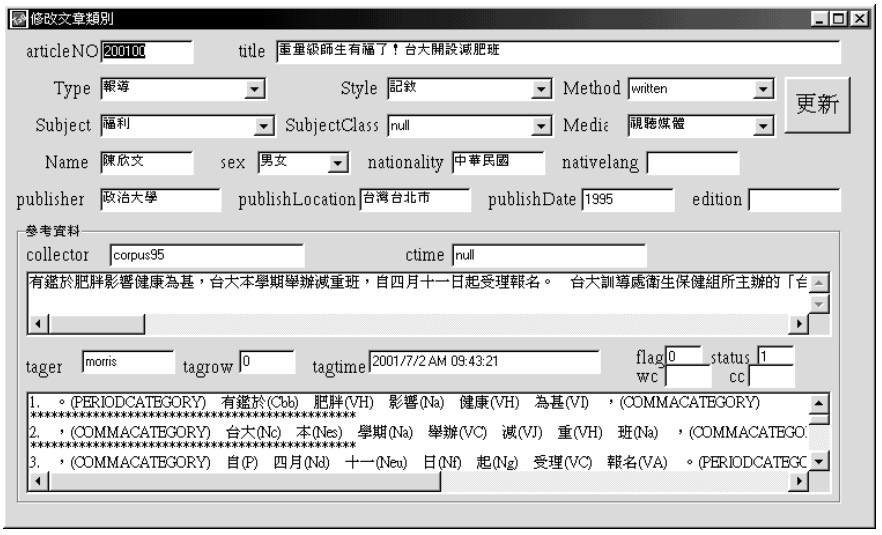
圖八：未知詞編輯介面

(3) 人工檢驗介面

請參照 2.4 節中的說明。

(4) 語料類別修改介面

當使用者發現文本的類別說明不太正確或有需要補充的時候，可以透過該介面進行修改。如圖九所示。



圖九：語料類別修改介面

2.5.1 系統輔助功能

本系統中尚有一些其它輔助功能的設計，分別說明如下：

(a) Help 功能：提供一些參考資料，如分詞標準，新修訂的詞類，與以前做過的疑難雜症的判例等，供使用者參考，目的在讓使用者於詞類標記時有個統一的標準。

(b) 提醒儲存功能：當使用者換句編輯或離開系統時，若所編修的文本有易動，系統會提醒使用者要做儲存的動作。

(c) 復原修改的功能：目前提供單次復原修改。

3. 系統成效

前述各個系統均已開發完成，並實際運作近半年的時間。在語料庫的維護與管理上，由於採用文本為單位的資料庫，因此管理維護上更簡易也更有系統；語料蒐集的部分，過去完全由人工來標示文本的格式、特徵，平均一篇文本約花五分鐘左右的時間，現在由於許多網址的 HTML 格式是固定的格式，可由語料蒐集介面自動標示文本的格式、特徵，人工只做必要的修改，故大幅縮短所需時間，平均在二十秒之內可校對完一篇原始文本的格式特徵。語料的斷詞標記部分，過去由於未知詞的存在，斷詞標記的效果大打折扣，使得人工校正時通常要花大量的時間將未知詞的部分做合分詞的修正，平均一篇六百字的文本需時四十分鐘。現在經過未知詞擷取模組的前處理，使得斷詞標記的表現有著顯著的改善，特別是有許多文本重複著大量的未知詞，例如新聞文本。斷詞標記的改善使得之後的人工校正更加的容易且快速，平均一篇六百字的文本縮短其校正時間為十五分鐘。

4. 結論

本論文提出一套構建及管理語料庫的系統設計，可以大幅縮短構建時程以及減少所花費的大量人力資源，經過實際運作近半年的時間，這樣的構建方式在語料蒐集、語料整理、人工校對等方面在成效上跟過去相比均有相當的改善。

本系統的後續研究方向主要有三：(1) 語料蒐集的格式判斷：目前我們只針對特定的網址來源，對固定的 HTML 格式抽取出文本的特徵，並無法自動的分析判斷所有網址的 HTML 格式。(2) 未知詞擷取模組的持續改善：未知詞擷取模組針對新聞類目前已達九成左右的正確率，然而對於其他文類或是長度過長的文本，由於未知詞在這些文本的統計訊息較不明顯，擷取出來的效果仍有持續改善的空間。(3) 將系統應用到古文或其他少數方言的語料庫構建。

參考文獻

Church, K. W. and R. L. Mercer, "Introduction to the Special Issue on Computational Linguisitcs Using Large Corpora," Computational Linguistics, Vol.19, No.1, 1993, pp.1-24.

Chen, Keh-jiann, Shing-huan Liu, Li-ping Chang and Yeh-Hao Chin, "A Practical Tagger for Chinese Corpora," Proceedings of ROCLING VII, 1994, pp.111-126.

Hsu, Hui-li and Chu-Ren Huang, "Design Criteria for a Balanced Modern Chinese Corpus," Proceedings of ICCPOL'95, Hawaii.

Smadja, Frank, "Retrieving Collocations from Text: Xtract," Computational Linguistics, Vol.19, No. 1, 1993, pp. 143-177.

Smadja, Frank, McKeown, K.r. and Hatzivasiloglou, V. "Translating Collocations for Bilingual Lexicons: A Statistical Approach," Computational Linguistics, Vol. 22, No.1,1996

Chang, Jing Shin, "Automatic Lexicon Acquisition and Precision-Recall Maximization for Untagged Text Corpora," National Tsing-Hua Universith, P.h.d. thesis, 1997

Wu, M. W. and Su, K. Y. "Corpus-based Automatic Compound Extraction with Mutual Information and Relative Frequency Count," Proceedings of ROCLING VI, Nantou, Taiwan, ROC, Sep. 1993 pp.207-216

Sproat, Richard and Shin, Chilin "A Statistical Method For Finding Word Boundaries In Chinese Text," Computer Processing of Chinese & Oriental Language, Vol. 4, No. 4, March 1990.

Yeh, Ching-Long and Lee, His-Jian, "Rule-Based Word Identification for Mandarin Chinese Sentences – A Unification Approach," Computer Processing of Chinese & Oriental Languages, Vol. 5, No.2, March 1991.

Lin, M.Y., Chang, T. H. and Su, K. Y., "A preliminary study on unknown word problem in Chinese word segmentation," Proceedings of 1993 R.O.C. Computational Linguistics Conference, Taiwan, 1993, pp.119-137.

Chen, Keh Jiann, Bai, Ming Hong, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," Proceedings of ROCLING X, Taipei, Taiwan, ROC, 1997, pp.159-174.

Chen, Keh Jiann and Liu Shing Huan, "Word Identification for Mandarin Chinese Sentences," Proceedings of COLING-92, vol. I, 1992, pp. 101-107.

Chen, C.J., M.H. Bai, & K.J, Chen, 1997, "Category Guessing for Chinese Unknown Words," Proceedings of 4th Natural Language Processing Pacific Rim Symposium(NLPRS' 9 7) pp.35-40.

詞庫小組, "**中文詞類分析（三版）**," CKIP Technical Report no.93-05.

詞庫小組, "**中央研究院平衡語料庫的內容與說明（修訂版）**," CKIP Technical Report no.95-02/98-04.