主題導向之非結構化文本資訊擷取技術

劉吉軒、翁嘉緯

國立政治大學 資訊科學系

E-mail: jsliu@cs.nccu.edu.tw

Abstract. 資訊擷取(information extraction)是從自然語言文本中辨識出特定主題或事件的描述，進而萃取出相關主題或事件元素的對應資訊，如人、事、時、地、物等。因此，資訊擷取技術能依照需要的主題與事件，自動的解讀自然語言文件，將文件中的原始文字資料轉換成結構化的核心資訊。在本論文，我們提出以型態辨識的方法來處理主題導向的非結構化文本資訊擷取的問題。我們以『總政府人事任免公報』為測試對象，其精確率為98%、回收率為97%，充分印證了本資訊擷取方法處理主題導向之資訊擷取問題的可行性。

1 導論

隨著電腦技術的進步與應用的普及，網際網路與全球資訊網上大量資訊被產生、流動、與保存，這些資訊通常可以說是氾濫而雜亂無章，對我們獲取、吸收、與利用資訊的能力構成嚴肅的挑戰，其結果甚至影響我們工作與生活的產出與品質。目前在網路上獲取資訊最普遍的工具就是利用資訊檢索(information retrieval)或搜尋引擎(search engine)，以使用者提供的關鍵字或索引詞，從大量網頁、文本集合中，調出含有使用者指定的關鍵字或索引詞的子集合。然而不論資訊檢索或搜尋引擎的精準度及回召率(precision and recall)如何，使用者仍然需要自己去進行閱覽與過濾，對於資訊的解讀、吸收、與利用等加值工作仍然需要以人工的方式去完成，常常耗費大量人力與時間或因無暇兼顧而無法得到真正的資訊價值。

資訊擷取(information extraction)是從自然語言文本中辨識出特定主題或事件的描述，進而萃取出相關主題或事件元素的對應資訊，如人、事、時、地、物等[1]。因此，資訊擷取技術能依照需要的主題與事件，自動的解讀自然語言文本，將文本中的原始資料轉換成核心資訊，可供進一步的機器使用及加值處理。資訊擷取技術的研究大約從一九九零年前後開始以英文文字為主要對象，選定數個主題，如恐怖份子攻擊事件、國際企業聯盟合併等，進行先導性的技術發展。一般研究結果認為[2]，資訊擷取以事件描述型式比對(event template matching)為主，再輔以領域語言知識及推理，如字詞、句型、前後指涉分析等，可以達到百分之七十左右的正確率，其問題的困難度不如自然語言處理大，卻具有相當高的實用價值，如情報蒐集與分析等。

一般而言資訊擷取系統的建立大致上可分為兩種方式：知識工程法(knowledge engineering approach) 及自動訓練法(automatically trainable approach)[3]。知識工程法主要是透過人工的方式給定擷取規則，而給定擷取規則的人必須對處理的領域及擷取規則建立的方式有一定程度的瞭解，其處理的範圍與正確性通常取決於擷取規則的充分與適當程度。因此，知識工程法對於人工介入與人力需求的程度較高，其素質及對領域的瞭解，也會對系統表現有非常大的影響。相對而言，自動訓練法並不需要人工介入的方式來建立擷取規則，通常只要將訓練語料做適當的標註，再透過訓練演算法就可以建立擷取規則，但其擷取規則可能產生不小的錯誤率。以發展成本及可攜性而言，自動訓練法似乎是發展資訊擷取系統的一個比較好的選擇，但是當訓練語料不易取得，或是對於資訊擷取系統的正確率有較高的要求時，知識工程法可能較佔優勢。

以資訊擷取的文本對象而言，又可區分為半結構化文本與純文字文本。半結構化文本(如網頁)最主要的特點便是內容含有標籤(Tag)，提供了辨識的依據，同時資訊呈現的方式較有規則性，通常只要掌握住這些標籤與規則，就能進一步的擷取出資訊。相關的研究包括WIEN[4], SoftMealy[5], STALKER[6], IEPAD[7]等。純文字文本則不包含任何的標籤與結構，其內容完全是一長串的文字符號，在處理上無法依賴或藉助於結構特徵，而必須完全針對文字符號的組合去做資訊擷取。相關的研究包括AutoSlog[8], FASTUS[9]等。

另外，文本語言也是資訊擷取技術的重要區別因素。以中文與英文來說，兩者之間最大的不同在於中文詞與詞之間並沒有明顯的界限(如英文字之間的空白)加以區隔，因此許多中文處理的第一個步驟，通常

就是利用詞典，將一個字串中的文字，比對詞典內的詞來當做斷詞的依據。不過因為字組成詞的變化程度相當大，一個句子難免會有許多種斷詞的方式，所以斷詞的錯誤率通常很高。另一個問題則是未知詞的問題，例如專有名詞，包括人名、地名、或組織名，不在詞典中的可能性非常大，而在一般句子中出現未知詞的頻率也很高，這對斷詞的正確率造成嚴重的影響。這些錯誤通常會對自然語言處理中的詞性標註、語法剖析等工作造成相當程度的困難，而使得一般以英文文本為處理對象的資訊擷取技術無法直接適用於中文文本。

隨著科技的進步與資訊數位化的趨勢，數位化之文字資料已呈指數般的大幅成長，而資訊擷取也就成為了近年來相當熱門研究領域。美國政府透過一系列的研討會(Message Understanding Conference, Text Retrieval Conference)為主軸[10]，持續推動資訊擷取研究，並規劃研究主題與實驗。這幾年，國內也有許多學者與研究人員投入資訊擷取領域，分別以網頁或其他半結構性文本為主[11][12]，及以中文純文字文本為主[13][14][15]。

我們的研究目標是發展高度實用的主題導向資訊擷取系統，以中文非結構性文本為擷取對象，並以結果的高度正確為主要考量。在本論文中，我們提出不做斷詞、不做詞性分析，而利用型態辨識的方法，搭配有限狀態自動機的運作機制，來處理中文非結構性文本資訊擷取的問題。我們的中文非結構性文本資訊擷取技術包括擷取模板的建立、多層次擷取型態的給定、及有限狀態自動機的執行工具等模組。擷取模板定義特定主題的構成語意元素，是描述主題的核心資訊，如人事異動的主題必須包括單位組織名稱、相關人員的姓名、及職位、時間等。在多層次擷取型態方面，我們處理中文語句中數個子句共用語意元素的問題。而在系統核心執行工具－有限狀態自動機的執行方面，我們利用演算法，將擷取型態轉換成相對應之有限狀態自動機，自動的進行中文語句的辨識。我們發展出一個主題導向的資訊擷取系統，並以『總政府人事任免公報』[16]為測試對象，蒐集了1981年(民國70年1月)到2003年(民國92年6月)的『總統府人事任免公報』電子檔，共1788期，約10萬個擷取目標，每一個擷取目標為一筆完整的人事異動資料。經過採樣及推測的實驗方法評估，實驗數據顯示98%的精確度與97%的回收率。

本論文以下分為四個章節，第二章描述我們研究的主題與文本特性，第三章提出型態辨識為主的中文資訊擷取方法與機制，第四章為實驗評估與結果討論，第五章為結論及未來的研究方向。

2 人事異動主題與文本

在主題導向的中文非結構性文本資訊擷取研究上，我們認為政府文本是一個相當好的試驗對象。我們從資料面、資訊與知識面、與價值面分別進行分析。首先，政府的官方文件具有下列特性：(1)有長期持續存在的主題，能提供大量的文件做為試驗資料與資料庫建置來源；(2)文件結構與主題描述方式較少變化，可以期望較高的資訊擷取正確率；(3)文件主題間存在關聯性，如人事、組織、考核、獎懲等，可以進一步的提供資訊查詢、資料探勘、與知識擷取研究；(4)政府文件可公開取得，沒有版權問題。而在價值面上，政府文件的精華資訊擷取具有極佳的先導示範作用。

我們選擇政府人事任免公報做為研究初期的試驗對象，此公報從民國三十七年起，週期性出刊(約每週一次)，記載政府各部門人事異動情形，並由總統令公告。這個特定文件只有單一主題，並且是由有限語意元素與句型組成。因此，非常適合做為我們研究起始的、典型的資料領域。政府人事任免公報的範例如下：

總統令 中華民國九十一年五月二十四日
任命鄒擅銘為國史館臺灣文獻館簡任第十職等組長。
任命楊合進為法務部簡任第十一職等權理簡任第十二職等司長。
……
任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為簡任第十一職等副處長。
總　　統　陳水扁
行政院院長　游錫堃

總統令 中華民國八十五年六月十三日
經濟部政務次長楊世緘，交通部政務次長蔡兆陽另有任用；財政部政次長王政一，僑務委員會副委員長王能章、張植珊辭職已准；均應予免職。
總　　統 李登輝
行政院院長 連戰

圖一：政府人事任免公報的範例

由初步的分析可知，政府人事公報中的記載分別為有關任命或免職的命令。任命的命令是指派某人到某個職位、機關、階級等，免職的命令是免除某人現有在某機關、階級上的職位。最簡單的任命句型為：任命李大衛為行政院簡任第十三職等參事。最簡單的免職句型為：行政院簡任第十三職等參事李大衛呈請辭職；應予令免。較複雜的任命或免職句型描述多人、共用部分資訊、或個人同時有兩個以上的資訊。例如：任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為簡任第十一職等副處長。經濟部政務次長楊世緘，交通部政務次長蔡兆陽另有任用；財政部政次長王政一，僑務委員會副委員長王能章、張植珊辭職已准；均應予免職。我們進一步的分析整理發現，政府人事公報中大約有 20 幾個語意元素，30 幾種句型，部分語意元素與句型顯示於圖二。

部分語意元素代碼：

A – 任命 (appoint)	R – 階級名稱 (rank)
N – 人名 (person name)	T – 職位名稱 (title)
B – 為 (as)	Q – 免職原因 (reason of dismissal)
O – 機關名稱 (organization name)	D – 免職 (dismissal)

句型範例： ANBORT ORTNQD

圖二：政府人事主題之部分語意元素與句型

在所有語意元素中，只有部分語意元素的字詞可以被蒐集而可直接辨識，其他的語意元素則不可能掌握，如人名、機關名等。另外，相關子句間的語意元素可能被省略，需要進行分析，找出對應而補齊。我們的基本想法是以能掌握的關鍵詞部分辨識切割句子，再與已知句型進行型態比對，而依據最接近的句型，詮釋之前未辨識出的語意元素。

3 型態比對模型與機制

資訊擷取為針對特定主題或事件從文本中找到對應於相關觀念或元素的實際資料，如人、事、時、地、物等資訊。從原文到解讀出的核心資訊，需要經過字詞的辨識、語句的分析、描述方式的比對、語意關係的推理、資訊的抽取與對應等步驟。國外相關研究歸納了一套基本流程[2][3]：tokenization (word segmentation) → morphological and lexical processing (part of speech tagging, word sense tagging) → syntactic analysis (full parsing) → domain analysis (co-reference, merging partial results) 。原文先被分解成句子與字詞，並從辭典中找出各字詞的詞類與其他資訊，接著進行各種名字的辨識，包括人名、組織名、日期、幣別等。基於這兩個步驟辨識的結果，各句子被部分的分析，根據句子結構的資訊，確認各重要字詞的意義。這些辨識與分析的結果和已知的主題或事件的可能描述方式進行比對，找出最接近的模型。接著進行同指涉詞的分析，找出前後彼此對應的名詞，再進行必要的推理，確定各字詞的意義與關係。最後，總結所有的資訊，依照選定的模型，確認主題或事件中各觀念或元素適當的對應字詞。

這一套基本流程與步驟提供了研究目標的指引與可行性依據，不過由於人類的語言具有模糊、變動、文化、地域等等的特殊性，從初期的tokenization到中期的syntactic analysis及到最後階段的domain analysis都可能因為這些語言上與主題領域上的特殊性而產生一定程度的錯誤。倘若在處理的前期就產生錯誤的話，就會影響到其後處理的步驟，而這些錯誤的累積也必定會影響到最後的結果。以中文來說，由於中文詞跟詞之間並沒有明顯的界限，斷詞錯誤的情形仍很可能出現，加上中文的結構較為鬆散、多縮寫型式且詞性不易判別，所以相對於英文，其錯誤情形的累積就會更嚴重。

由於我們的目標是高度正確的、具有實用價值的資訊擷取技術，我們決定不採用斷詞及詞性標註的方式來處理原始文句，而藉由分析文本中中文語句的結構、順序及組合方式，以型態辨識的方法確認出中文語句之結構關係，再利用關鍵字詞及其特殊的型態，來推論或擷取出相關的資訊。

3.1 以型態辨識擷取中文資訊之概念形成

一般而言，語言的敘述可以視為多種特定語意元素的組合，而其組合的方式通常具有某種規則或常見的型態，所以只要能掌握到某些特定語意元素常對應之字詞的知識，再加上這些語意元素的組合方式，就可以推論出其它相關語意元素的位置，進而擷取出我們想要的資訊。舉例來說：『總統某某某先生』這個敘述可以區分成三個連續語意元素。假設我們想擷取出”某某某”這個字詞，我們只要將第一個語意元素相對應之字詞”總統”與第三個語意元素相對應之字詞”先生”當作關鍵字，利用前後關鍵字之辨識，再包夾切割出目標字詞的方式，即可擷取出第二個語意元素相對應之字詞”某某某”。

在主題導向的中文資訊擷取中，我們針對含有與主題相關資訊的敘述，將特定語意元素的組合方式稱之為擷取型態(extraction pattern)。以上述的例子而言，我們可以標示一個以三個連續之語意元素代號『TNA』組成之擷取型態，其中T(Title)表示”職稱”，N (Name)表示”姓名”，A (Appellation)表示”稱謂”。透過『TNA』的擷取型態，我們可以從許多文本中擷取出出任過特定職位的人員姓名，例如：(總統、蔣經國)，(總統、李登輝)，(總統、陳水扁)，(首相、邱吉爾)，(首相、柴契爾)等。

3.2 擷取型態中語意元素之辨識與擷取屬性

擷取型態中的語意元素組合，在字詞辨識與擷取的過程中，個別語意元素具有不同的屬性。其中最主要的差異在於直接辨識的難易，各語意元素在不同文本中出現的敘述，其對應之字詞可能變化性較大、不容易掌握，如人名、機關名；也可能變化性較小、容易掌握，如稱謂(先生、小姐等)。通常這些字詞變化性較大的語意元素，也可能會是我們的擷取目標，我們採用的基本方法為：蒐集變化性較小的語意元素所對應之字詞做為關鍵字，利用關鍵字之辨識，以前後包夾目標語意元素的方式，切割擷取出變化性較大的語意元素相對應之字詞。

我們定義了三個語意元素之辨識與擷取屬性，以因應擷取過程，針對擷取型態中個別語意元素的不同處理動作。這三個辨識與擷取屬性為: (1) **EOE**(Extraction Only Element): 屬性為EOE的語意元素表示其相對應之字詞變化多，不能掌握，如人名、機關名等。這類語意元素必須依賴前後關鍵字的辨識，進而切割擷取出相對應之字詞。(2)**ROE**(Recognition Only Element): 屬性為ROE的語意元素表示其相對應之字詞變化少，可以被蒐集而可直接辨識，但此語意元素在主題資訊中不具價值，如前述例子中的”稱謂(A)”及政府人事異動主題中的”為(B)”。其相對應之字詞是用來當做包夾切割目標語意元素字詞的關鍵字，並不需要被擷取。(3)**RTE**(Recognition exTraction Element): 屬性為RTE的語意元素表其相對應之字詞變化少，可以被蒐集而直接辨識，同時，此語意元素也是主題資訊中的重要成分。所以此語意元素相對應之字詞，既是用來當做切割前後其他目標字詞的關鍵字，也是擷取的對象本身。

基本上，由EOE、ROE及RTE這三個屬性的語意元素所組成的擷取型態，對應於較為緊密、精簡或簡短的主題描述方式。例如，擷取型態『TNA』中，語意元素『T』之屬性為RTE，對應之關鍵字為『總統、首相』等，語意元素『N』之屬性為EOE，語意元素『A』之屬性為ROE，對應之關鍵字為『先生、女士』等。在文本中出現『總統陳水扁先生』的語句，經辨識擷取後的結果將為(總統、陳水扁) 。

3.3 模板建立

資訊擷取技術針對文本中特定主題的資訊，進行抽取與對應，而這些抽取出來的文字，必須能完整的對應於主題所需的各部分資訊。倘若只是一味的利用擷取型態來辨識中文語句，而忽略了主題資訊的完整性，那麼擷取出來的結果就會零散而不建全。因此，擷取模板(extraction template)的建立就有其必要性。擷取模板為一組特定語意元素的欄位集合，必須根據不同的擷取主題加以定義。以政府人事異動的主題而言，我們定義的擷取模板包括姓名、組織單位、職位、職等、異動種類、異動原因、異動時間等語意元素，一個擷取目標為文本中有關某一個人員的異動情形，由一個完整的擷取模板來描述。在政府人事公報文本中，我們發現不同擷取目標共用部分語意元素的情形，這是中文前後文句中常見的省略、沿用、及總結等的慣例用法。因此，對於部分擷取目標而言，直接的資訊擷取只能得到不完整的主題資訊。藉由資訊模板的定義與規範，相鄰擷取目標的部分資訊將能互相參照補齊，而建立完整的主題資訊。為了完成這樣一個動作，在建立擷取模板時，就必需針對每一語意元素的主題資訊構成屬性加以定義，主要分成三種：(1)required, 表示此語意元素是必須存在的；(2)optional, 表示此語意元素可能存在，但也可以不存在；(3)context-dependent, 表示此語意元素可出現在前後相關的中文語句中。

茲以『任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員。』為例說明，我們可以觀察到以下特點：(1)人員為人事異動主題中擷取目標的主體，一個擷取目標至少必須具有人員姓名與職稱的存在；(2)語句『黃崇烈為專員』含有一個擷取目標，但除了人員姓名與職位兩個語意元素外，異動種類、組織單位、職等等語意元素並沒有出現在前述的語句中，而是必須由前一個句子中的部分語意元素沿用；(3)職等的語意元素並不是擷取目標的必要元素，有些擷取目標有，但也有擷取目標不具備。因此，在人事異動的主題中，人員姓名及職稱的主題資訊構成屬性為”required”，表示這兩個語意元素必須在擷取目標所在之文句中擷取。組織單位、異動種類與異動原因的主題資訊構成屬性為”context-dependent”，表示這三個語意元素可出現在前後相關的語句中。職等的主題資訊構成屬性為”optional”，表示這個語意元素可能存在，也可能不存在。

3.4 多層次擷取型態

以資訊擷取之觀點而言，主題相關敘述語句中之各語意元素，除了具有連續性的關係之外，更有著前後相關子句『語意元素共用』的關係。而這共用性的關係包羅甚廣，可能為共用的時間、共用的單位、共用的幣值等等。舉例來說，中文語句『總統候選人民進黨陳水扁先生、國民黨連戰先生、親民黨宋楚瑜先生』中，『民進黨陳水扁先生』、『國民黨連戰先生』、『親民黨宋楚瑜先生』均共用了外層字詞『總統候選人』。為了處理『外層語意元素共用』的問題而完整的擷取所有相關資訊，我們設計了多層次的擷取型態，其辨識過程將是從外層共用的部份先行處理，然後再依序從其內層進行下一步的擷取動作。繼續以前述語句『總統候選人民進黨陳水扁先生、國民黨連戰先生、親民黨宋楚瑜先生。』為例，我們可以建立多層次擷取型態為『C{PNA}』來處理此種類型之語句，其中語意元素C為參與活動人員身分，屬性為RTE，並建立其相對應之關鍵字『總統候選人』；”{PNA}”表示可重複出現之內層擷取型態，語意元素P為組織名稱，屬性為RTE，並建立其相對應之關鍵字『民進黨』、『國民黨』及『親民黨』；語意元素 N為人名，屬性為EOE；語意元素A為稱謂，屬性為ROE，並建立其相對應之關鍵字『先生』及『女士』、『小姐』)。因此，經由多層次擷取型態 『C{PNA}』辯識後的擷取結果將為(總統候選人、民進黨、陳水扁)、(總統候選人、國民黨、連戰)及(總統候選人、親民黨、宋楚瑜) 。

3.5 有限狀態自動機

我們以有限狀態自動(finite state automata)做為擷取型態的運作機制。有限狀態自動機的主體在於各狀態間之transition function。我們將擷取型態中各語意元素所代表的辨識資訊轉換成各狀態之transition function。轉換的演算法如圖三所示，其中State的初始值為1 (初始狀態)，『Pattern.length()』可以計算擷取型態中語意元素的總個數，『Pattern.char(I)』表示擷取型態中第I個語意元素，『- Pattern.char(I) 』表示除了『擷取型態中第I個語意元素』的其他語意元素，『＊』表示任何的語意元素，Final_State表最終狀態(接受狀態)，而0-state表拒絕狀態(sink state)。

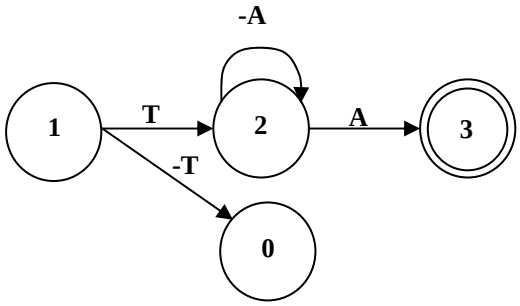
```
0   Begin
1     State = 1
2     For I = 1 to Pattern.length() do
3     If(attribute of Pattern.char(I) is not EOE ) then
4       Generate state transition from “State to (State+1)”,
5       Add transition input as “Pattern.char(I)”
6       If (attribute of Pattern.char(I-1) is not EOE then
7          Generate state transition from “State to 0-state”,
8          Add transition input as “ - Pattern.char(I) ”
9
10      State = State + 1
11    Else
12      Generate state transition from “State to State”,
13      If ( “ -Pattern.char(I+1) ” does not exist)
14         Add transition input as “＊”
15      Else
16         Add transition input as “-Pattern.char(I+1)”
17  End
```

圖三：擷取型態之有限狀態自動機轉換演算法法)

此演算法的基本目標為根據擷取型態及其中各語意元素之辨識與擷取屬性，自動建立一個對應此擷取型態之有限狀態自動機及其中各狀態間的狀態轉移。此擷取型態之辨識與擷取就由其對應之有限狀態自動機執行，而產生適當之輸出結果，包括成功擷取個別語意元素對應之字詞(進入最終狀態及狀態轉移過程之辨識紀錄) 或型態不符合(無法進入最終狀態)。以擷取型態『TNA』為例，語意元素T的屬性為RTE，依據圖三演算法的第4至第5行，將會建立一個從1號狀態至2號狀態的transition，而其transition input為T；而依據演算法的第7至第8行，系統將會建立一個從1號狀態至0號狀態(0-state)的transition，其transition input為-T。而語意元素N的屬性為EOE，依據演算法的第12行，將會建立一個從2號狀態至2號狀態的transition，接著依據演算法的第16行，其transition input為-A。語意元素A的屬性為ROE，依據演算法的

第4至第5行，將會建立一個從2號狀態至3號狀態的transition，其transition input為A。其中最終狀態(Final_State)為3號狀態。圖四為依照此演算法對擷取型態『TNA』所自動建立之狀態轉換表及其相對應之狀態轉換圖。

	Transition Input			
State	**T**	**-T**	**A**	**-A**
1	**2**	**0**		
2			**3**	**2**

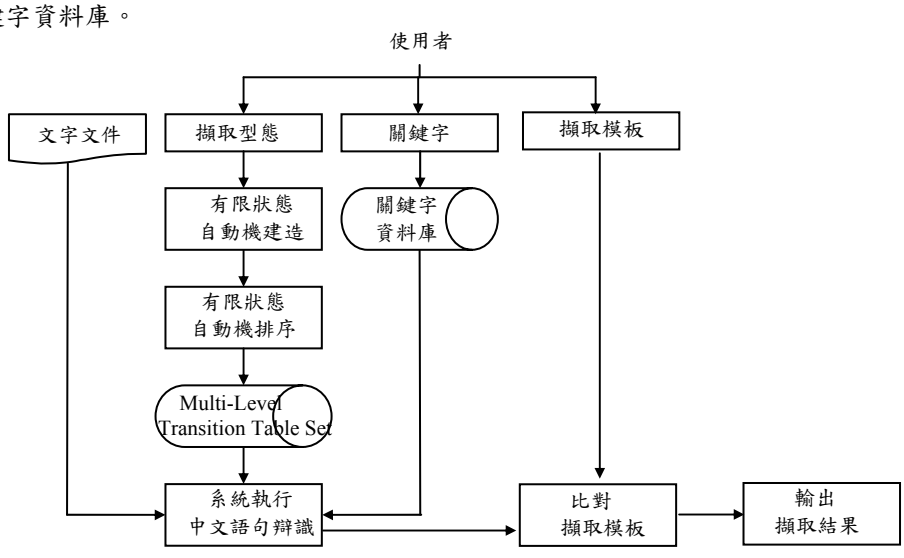


圖四：擷取型態『TNA』之狀態轉換表及圖

此有限狀態自動機的運作過程如下，由1號狀態起始，開始針對目標語句進行辨識。若辨識到語意元素T所對應的關鍵字，則進入2號狀態。反之，若在1號狀態時遇到除了語意元素T外的任何其他字詞，此有限狀態自動機將會進入0號狀態。在2號狀態時，若辨識到語意元素A所對應的關鍵字，則進入3號狀態。在由2號狀態進入3號狀態以前，所遇到的除了語意元素A外的任何其他字詞，將會使此有限狀態自動機以迴圈的方式停留在2號狀態，而這一個迴圈的動作則對應於屬性為EOE的語意元素的字詞蒐集擷取。以擷取型態『TNA』來說，此一迴圈的動作動應於語意元素N對應字詞之擷取。

3.6 系統架構

我們發展的型態比對模型與機制包括語意元素屬性的訂定、關鍵字的蒐集、擷取模板、多層次擷取型態及有限狀態自動機的運作對應等，針對含有特定主題的大量中文文本，進行個別擷取目標主題相關資訊的完整萃取，希望能達成高度正確的資訊擷取，進而匯集成具有實用價值的特定主題結構性資料庫。我們將這些模組整合成一個系統架構(圖五)，其運作流程大致分為三個階段，第一階段為由使用者根據其擷取需求，建立主題領域所需的各項資訊，包括擷取模板的定義、擷取型態的建立、屬於ROE及RTE的語意元素所對應的關鍵字的蒐集與輸入。此階段為擷取主題的訂定與辨識資訊的設置。在第二階段中，由系統將使用者給予之各種擷取型態自動轉換成相對應之有限狀態自動機，接著再透過排序演算法以狀態數目的多少，建立個別有限狀態自動機被執行比對之優先次序。通常，短的擷取型態較籠統，長的擷取型態則較明確。所以，應該以較長的擷取型態優先比對，以避免產生辨識上之錯誤或模糊。排序後之有限狀態自動機被存至Multi-Level Transition Table Set資料庫，而使用者所建立之關鍵字及其對應之語意元素則存至關鍵字資料庫。



圖五：系統架構與執行流程

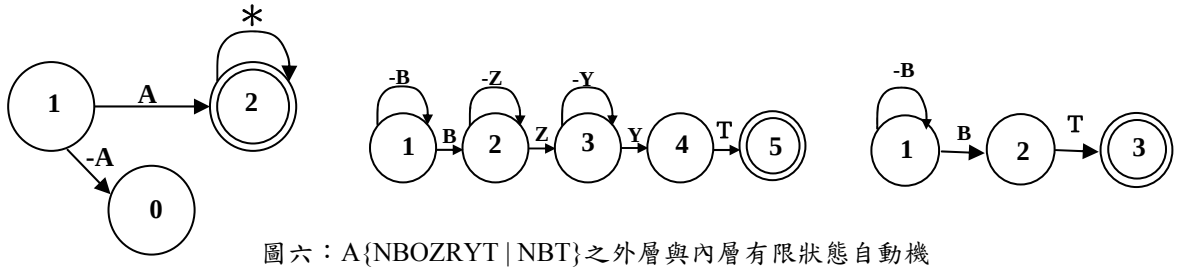
第三階段為系統執行辨識與擷取階段，系統先將輸入文本以句號及逗號切割成個別的句子，再針對每一個句子進行型態辨識。系統從Multi-Level Transition Table Set資料庫中讀取一個Transition Table以執行一個擷取型態之有限狀態自動機，由外而內進行比對與擷取的動作。有限狀態自動機在執行時，系統會將目前遇到的中文字詞與關鍵字資料庫裡的關鍵字詞進行比對(採長詞優先的方式)，經比對確認的語意元素則輸入到有限狀態機中進行狀態的轉移。當有限狀態自動機停留在Final_State，則表示目前處理的中文語句符合此有限狀態自動機所表示的擷取型態。反之，若此中文語句未解譯完就進入0-state或者此中文語句解譯完後最後的狀態不是停留在Final_State，就表示目前比對的型態並不吻合，系統會再讀取下一個擷取型態的有限狀態機繼續比對辨識。

我們以中文語句『任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員。』為執行範例，說明系統運作過程。在第一階段，使用者所建立的主題領域資訊中與此範例有關的部分顯示於表一。在系統模組中，這些資訊分別存於擷取型態、擷取模板、關鍵字等不同模組，為了呈現上的方便，我們將之顯示於一個表中(表一)。

表一：由使用者建立之部分主題領域資訊

多層次擷取型態	A{NBOZRYT | NBT}	
語意元素	屬性	對應關鍵字
A (異動種類)	RTE	任命、特任、特派、派
N (人員姓名)	EOE	N/A
B (身分賦予)	ROE	為
O(組織單位)	EOE	N/A
Z(職等種類)	ROE	簡任第
R(職等等級)	EOE	N/A
Y(職等稱謂)	ROE	職等
T(職位)	RTE	處長、專員、部長、…

在第二階段，系統建立多層次擷取型態所對應的有限狀態自動機。表一所顯示的擷取型態將會產生一個處理外層中文語句的有限狀態自動機及兩個處理內層中文語句的有限狀態自動機(圖六)。接著在第三階段，當此一擷取型態被選取與語句『任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員』進行比對時，關鍵字"任命"會先被比對確認，其相對應之語意元素A則輸入至有限狀態自動機進行狀態的轉移，此時有限狀態自動機將會由1號狀態進入2號狀態。由於2號狀態為最終狀態且有擷取的動作，再加上狀態轉移的語意元素為＊(任何的語意元素)，所以最後的擷取結果將為"任命"及"鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員"，其中擷取結果"任命"將會被暫存起來，而"鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員"將繼續交由內層之有限狀態自動機進行辨識。



圖六：A{NBOZRYT | NBT}之外層與內層有限狀態自動機

接著目前之語句將被切割成兩個子句，其中"鍾萬梅為行政院客家委員會簡任第十二職等處長"經內層有限狀態自動機辨識，同時比對擷取模板後所得到擷取結果如下：任命(A)、鍾萬梅(N)、行政院客家委員會(O)、十二(R)、處長(T)。語句"黃崇烈為專員"則被另一個內層有限狀態自動機所辨識，經比對擷取模板後所得到擷取結果如下：任命(A)、黃崇烈(N)、行政院客家委員會(O)、專員(T)。其中異動種類與組織單位兩個語意元素在擷取模板中所定義的主題資訊構成屬性為"context-dependent"，而職等之語意元素的屬性為"optional" 。因此，以"鍾萬梅"及"黃崇烈"為主體的兩個擷取目標將在這兩個語意元素上，相互參照內外層擷取結果，補上擷取模板所定義的語意元素資訊，既"鍾萬梅"的擷取目標補上外層的異動種類資訊"任命"，而"黃崇烈" 的擷取目標補上外層的異動種類資訊"任命"及前一個擷取目標("鍾萬梅")的組織單位資訊" 行政院客家委員會"，但職等之語意元素之資訊仍然維持空白。這些語意元素之資訊填補動作符合原文之含意。

4 實驗評估

為了有效驗證此一主題導向資訊擷取系統，我們大量蒐集與轉換『總政府人事任免公報』，建立了從1981年(民國70年1月)到2003年(民國92年6月)的『總統府人事任免公報』電子檔，共1788期的實驗文本資料。每一期的人事任免公報為以總統令形式發布的政府各部門人事訊息，長短不一。內容主要為以句號區隔的人事異動命令，每一道人事命令有時只針對一個人，有時則可能牽涉到一、二十人。由於人事異動主要是描述人員職務工作的變更，所以，一個擷取目標就是以每一個人員為主體的相關異動資訊。本研究的實驗文本資料中，約有10萬個擷取目標。而人事命令中有許多精簡、共用、省略的情形，系統必須為每一個擷取目標建立完整的異動資料，包括『人員姓名』、『異動種類(就任/免職)』、『組織單位』、『職等』、『職稱』等。

在系統運作的第一階段，由系統發展者扮演使用者的角色，定義擷取模板，建立了約30種擷取型態、約130個關鍵字。接著，系統經過第二階段的有限狀態自動機建置，於第三階段，對實驗文本資料一一進行辨識擷取。最後，產生約10萬個擷取目標異動資料的輸出。

4.1 評估方式

資訊擷取結果的正確性的檢驗，唯有以人工方式，一一核對每筆擷取資料與原來的相關人事命令是否吻合，同時，也必須確定擷取模板中，每一個語意元素欄位中所填入的擷取字詞是正確的，沒有因誤判而錯置或切割錯誤的情形。我們的實驗結果共有約10萬筆資料的輸出，然而在有限的人力與時間下，我們無法完全核對所有輸出結果。所以，我們以四種採樣方式，選取部份區間的輸出資料進行人工檢驗核對，希望以此推測所有輸出資料的可能評估結果。

我們規劃的四種採樣方式為：(1)連續區間(continuous interval): 涵蓋從1998年1月到2003年6月的所有文本，共有27,541個擷取目標；(2)隨機(不重複)選取(random item): 從1981年1月到1997年12月的文本中，隨機選取1200個命令句；(3)規則間隔區塊(regular block): 從1981年1月到1997年12月的文本中，每隔約4300個句子就選取連續的120個句子為一個區塊，共選取10個區塊，1200個句子；(4) 隨機(不重複)區塊(random block): 從1981年1月到1997年12月的文本中，以連續的120個句子為一個區塊，隨機選取10個區塊。

此外，我們以擷取目標為單位，評估系統在擷取結果上的成效。當系統所擷取出來的各項欄位資訊均符合擷取模板中應該對應的語意元素時，才視為正確的擷取結果。因此，只要有一個欄位發生錯誤，包括不正確的空白、錯置、不正確的字詞切割等，即視為一錯誤的擷取。

在評估的量度方面，我們採用精確度(precision)、回收率(recall)、及F-measure。其中，精確度(p)的算式為正確擷取數與系統擷取數之比率；回收率(r)的算式為正確擷取數與應擷取數之比率；而F-measure的算式為$2pr / (p+r)$，我們採取p與r相同的權重($\beta = 1$)，表示對精確度與回收率採取同樣的重視。

4.2 實驗結果

本研究的實驗結果列於表二。從各項實驗資料可以看出，系統在『總統府人事任免公報』領域約20幾年的資料上，有著不錯的精確度及回收率，印證了以型態辨識的方法應用在主題導向資訊擷取上的可行性。本系統的精確度可達約98%，顯示系統有著高準確度的擷取執行能力；而回收率可達約97%，顯示我們掌握了絕大部份的擷取型態與關鍵字。因此，相信使用者透過擷取模板的給定、擷取型態與關鍵字的建立、系統的運作等執行方式，就能針對特定主題領域的文本，產生具有實用價值的擷取結果。

表二：實驗結果數據

	Continuous Interval (1998-2003)	**Random Item (1981~1997)**	**Regular Block (1981~1997)**	**Random Block (1981~1997)**
應擷取數	27541	2159	2424	1706
系統擷取數	27340	2137	2401	1699
正確擷取數	26916	2102	2370	1671
精確度	98.45%	98.36%	98.71%	98.35%
回收率	97.73%	97.36%	97.77%	97.95%
F-1	98.09%	97.86%	98.24%	98.15%

4.3 結果討論

根據我們的觀察與分析，系統在辨識語句及擷取結果時產生失誤的原因有三種：

(1) 部分擷取型態沒有掌握: 導致某些語句能被其他較鬆的擷取型態對應，而產生錯誤的辨識與擷取，影響到系統的精確度與回收率。另外，也可能使某些語句無法被系統中的任何擷取型態對應，而沒有擷取任何資訊，影響到系統的回收率。

(2) 部分職稱關鍵字沒有掌握: 在擷取模板中，職稱的語意元素是屬於 RTE，是系統依賴以進行比對與辨識的關鍵字之一種。在二十幾年的公報中，我們沒有掌握到所有的職稱，致使部分語句辨識錯誤，影響到系統的精確度與回收率。

(3) 擷取內容含關鍵字: 在擷取模板中，人員姓名及組織單位的語意元素是屬於 EOE，必須依賴緊接其後的 ROE 或 RTE，比對辨識到特定關鍵字後，進行狀態轉移到下一個狀態。假如 EOE 所對應的字詞中含有其後的 ROE 或 RTE 的關鍵字，就會造成辨識上的錯誤，影響系統的精確度與回收率。例如，在人員姓名(EOE)中含有其後的 ROE 關鍵字"為"，或在組織單位(EOE)中含有其後的職稱(RTE)關鍵字，如"審計部審計"或"XXX 委員會委員"等。

因為擷取型態、關鍵字的不足所產生的失誤，可透過補足擷取型態、關鍵字的方式來解決，而這也反應以知識工程法建構的資訊擷取系統，處理的範圍與系統的正確性受到文本領域知識涵蓋程度的直接影響。例如，由系統發展者觀察部分文本而建置的30幾種擷取型態與130個關鍵字在20年的人事任免公報中可能只有少數遺漏，所以，系統可以達成相當高的精確度與回收率。這些系統辨識所需之文本領域資訊的建置，並不需要特定的背景與知識，只要具備一般的中文能力既可勝任。這是因為政府人事任免公報為單一主題的制式文件，語言表達的變化空間相對較小，所以，達成極高精確率與回收率的可能性較大。另外，由"擷取內容含關鍵字"因素所造成的錯誤，則是屬於系統辨識機制如何因應字詞屬性模糊性的問題。我們認為這種辨識問題大致可以用兩種方式改善。第一種是考慮更多的檢查條件，訂定更嚴謹的狀態轉移條件。第二種是以文本領域知識對於擷取內容(EOE)的語意元素加上條件定義，如字數限制。

5 結論及未來研究方向

我們以中文非結構性文本為擷取對象，發展高正確率的主題導向資訊擷取系統，透過擷取模板的建立、多層次擷取型態的訂定、及有限狀態自動機的轉換，系統展現出高度實用的價值。我們採用知識工程的方式來建立資訊擷取系統，雖然在政府人事任免公報領域有足夠的描述與處理能力，但是如果過度依賴使用者來建立擷取型態與關鍵字，就容易造成使用者的負擔，也造成系統可攜性(portability)的不足。未來我們希望透過與使用者互動或是自動學習的方式，建立擷取型態與關鍵字，以提高擷取的正確率與系統的可攜性。另外，我們也將考慮進一步了解與評估中文斷詞與named entity辨識技術對本系統的可能助益。最後，本系統將繼續在公報涵蓋時間的完整性上努力，建置從民國37年第1期到最新出版之公報的全面擷取資料庫。以系統可擴充性(scalability)的角度而言，本系統的辨識機制沒有任何預期的困難，但在辨識知識上，可能必須建置一些新舊文本中可能出現的(系統尚未具備的)辨識型態與職稱關鍵字。

參考文獻

[1] Information extraction: a multidisciplinary approach to an emerging information technology: international summer school, SCIE-97, Frascati, Italy, July 14-18, 1997.

[2] Jim Cowie , Wendy Lehnert. Information Extraction, *Communications of the ACM,* 39 (1), pp. 80-91, 1996.

[3] Applet, D. E. and Israel, D. J. Introduction to Information Extraction Technology. *In Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1999.

[4] N. Kushmerick, D. Weld, and R. Doorenbos. Wrapper Induction for information extraction. *In Proceedings of the 15th International Joint Conference on AI (IJCAI-97),* pp. 729-737, 1997.

[5] Chun-Nan Hsu and Ming-Tzung Dung. Generating Finite-State Transducers for Semi- Structured Data Extraction from the Web, *Journal of Infromation Systems, Special Issue on Semi-structured Data*, Vol.23, No.8, pp. 521-538, 1998.

[6] I. Muslea, S. Minton, and C. Knoblock. STALKER: Learning Extraction Rules for Semi-structured, Web-based Information Sources. *In Proceedings of AAAI-98 Workshop on AI and Information Integration*, AAAI Press, Menlo Park, California, 1998.

[7] Chia-Hui Chang and Chun-Nan Hsu. Automatic Extraction of Information Blocks Using PAT Trees. *In Proceedings of 1999 National Computer Symposium (NCS-1999)*, Tamking University, Tamsui, Taiwan, 1999.
[8] Ellen Riloff. Automatically Constructing a Dictionary for Information Extraction Tasks. *Proceeding of the Eleventh National Conference on Artificial Intelligence,* pp.811-816, 1993.
[9] D. Applet, J. Hobbs, D. Israel, M. Kameyama, M. Tyson. The SRI MUC-5 JV FASTUS Information Extraction System. *Proceedings of the Fifth Message Understanding Conference*, 1993.
[10] Ralph Grishman, and Beth M. Sundheim. Message Understanding Conference-6 : A Brief History. *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96),* Copenhagen, Denmark, 1996.
[11] 易文韜, 樹狀HTML文件之資訊擷取, 碩士論文, 台大資工, 指導教授: 許永真, 民國86年。
[12] 呂紹誠, 網際網路半結構性資料擷取系統之設計與實作, 碩士論文, 中央資工, 指導教授: 張嘉惠, 民國89年。
[13] 游基鑫, 中文資訊擷取環境建構與同指涉問題之研究, 碩士論文, 台大資工, 指導教授: 陳信希, 民國89年。
[14] 張嘉洋, 古文獻中資訊擷取之研究, 碩士論文, 台大資工, 指導教授: 歐陽彥正, 民國87年。
[15] C.-H. Chang, Information Extraction: A Pattern Mining Approach for Free-Form Text, *Proceedings of 2003 The Joint Conference on AI, Fuzzy System, and Gray System*, Taipei, Taiwan, 2003.
[16] 總統府人事任免公報, *URL :* *www.president.gov.tw/2_report/layer2.html*