以語法分析為輔建立新聞名詞知識庫

楊昌樺　陳信希

國立台灣大學資訊工程學系

chyang@nlg.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

摘要. 本文針對新聞文件中出現之具名實體──人名為出發，以語法分析選定同位語結構為線索，輔助建立新聞知識庫。我們採用中文新聞文件為測試題材，經由分詞詞性標記和具名實體辨識處理。在標記人名及其前後文詞性資訊後，統計常出現在人名標記之前的同位語結構。由演算法統計後的結果篩選出常用於敘述新聞人物之稱謂 (title)，藉以建立新聞領域中人物實體之知識本體雛型，以協助文件解析及機器翻譯等應用。

1　緒論

新聞媒體每天針對不同的「人、事、時、地、物」，產生許多報導和敘述，尤其在網路化的環境更是突破了時效與場合的限制，新聞文件得以電子化的型式讓大眾選閱。人們可以隨時隨地獲得最新的訊息，並追蹤感興趣事件之後續發展。由於網際電子化新聞具有上述的時空便利性，成為發展自然語言處理的基礎，使得我們能以最快的速度收集新聞文件，進而建立豐富的語料庫。近來新聞語料提供了許多系統及研究所需要的知識來源，如文件摘要系統(Chen *et al*., 2003; Lee and Ker, 2001)、跨語言資訊檢索(NTCIR-4 Test Collections, 2004)[1]、中文語料庫的建構(Ma *et al*., 2002)等等。本文擬針對新聞文件所出現之具名實體，以語法分析選定同位語結構為線索，輔助建立新聞知識庫。

從知識本體 (Ontology)[2]的角度來看，新聞文件敘述人們這個主要實體 (entity)在各個領域(domain)(如：國際、政治、經濟、娛樂)所參與的事件(events)。在這項前提之下，如何從新聞文件中找出實體，成為處理新聞文件最基本的工作。新聞文件中傳遞資訊的對象是一般的大眾，因此具名實體(Named Entity)的描述通常採用完整的名字，而少用指涉的方式。例如，記者使用「台東東河鄉的果農王金發先生」來取代「隔壁的老王」。既然"具名"是新聞文件中敘述之實體所具備的重要特性，因此我們可套用具名實體辨識(Named Entity Recognition; 以下簡稱NER)模型，來找出新聞文件中的人名、地名、組織名──這些名稱通常代表了新聞文件中最重要的實體。

從語法分析的角度來看，由新聞文件可歸納出許多具名實體所形成的名詞片語(NP; noun phrase)。以新聞文件中常出現的一個樣版(template)，「某人在什麼時候表示…」為例，我們節錄2004/6/30的三則新聞：

(東森新聞報)　日本首相**小泉純一郎**6月29日表示…
(中國日報)　新黨主席**郁慕明**30日表示…
(聯合新聞網)　聯發科發言人**喻銘鋰**昨天表示…

這三則新聞所描述的對象，除了人物的姓名之外，姓名前面還附上其隸屬的組織名與稱謂。組織名加上稱謂形成了另一個名詞片語，與接在後面的人名形成語法上所稱的同位語 (apposition)架構，在中研院詞庫小組所建構中文句結構樹資料庫[3]敘述的基本原則[4]中，同位語與名詞中心語等義，因此與名詞中心語構成雙岔結構。若以人名當作主要的名詞中心語，我們可以從中央研究院中文句結構樹 (以下簡稱

[1] http://research.nii.ac.jp/ntcir-ws4/data-en.html

[2] 「知識本體」參照自 http://bow.sinica.edu.tw/

[3] http://rocling.iis.sinica.edu.tw/ROCLING/Treebank/Treebank_cf.htm

[4] http://godel.iis.sinica.edu.tw/CKIP/treebank/index.html

TreeBank)中觀察到很多 形成雙岔結構的同位語 現象，如取出字串 (s1)、(s2)兩個具有相同同位語結構 NP→NP + Nba[5] (正式專有名詞)的部分剖析樹如下：

(s1) 經濟部次長江丙坤
NP(apposition:NP (property:Nca:經濟部|Head:Nab:次長)
|Head:Nba:江丙坤
)

(s2) 三重市長蔡火石
NP(apposition:NP (property:Nca:三重|Head:Nab:市長)
|Head:Nba:蔡火石
)

(s1)和(s2)這兩個句子片段的剖析皆滿足以下兩個條件(c1)、(c2)：

(c1) LHS(left-hand side)的NP和Nba為同位語、(c2) LHS的NP繼續剖析為Nca(專有地方名詞)+ Nab(個體名詞)；

並且具有性質(*1)：

(*1) 兩個連續標記成Head的名詞具有同位語法架構。

同樣的情況也出現在TreeBank中「瑞典好手貝爾格斯壯」、「琉森藝術家羅芙布姆」等字串的剖析中。另外，在下面兩個同樣從TreeBank取出語法結構為NP→NP + Nba的例子(s3)和(s4)字串中：

(s3) 湖人主力史卡特和魔術
NP(apposition:NP (property:Nba:湖人|Head:Nad:主力)
|Head:Nba(DUMMY1:Nba:史卡特|Head:Caa:和|DUMMY2:Nba:魔術)
)
(s4) 清太祖努爾哈齊
NP(apposition:NP (Head:Nba:清太祖)
|Head:Nba:努爾哈齊
)

(s3)第一個NP同位語分別剖析成「湖人」和「主力」，其詞性分別為Nba(專有名詞)和Nad(抽象名詞)。由「主力」開始連續標記成Head，因而具有性質(*1)連續同位語架構。不過與(s1)和(s2)相較之下只符合條件(c1)，並沒有符合(c2)。(s4)的「清太祖」、「努爾哈齊」連續標記成Head可視為滿足性質(*1)，但是「清」和「太祖」並沒有分開[6]，因此與(s1)和(s2)相較之下條件(c1)和(c2)無法成立。

然而(s4)的例子仍具有語法剖析上的歧義性，根據CNS14366[7]中文分詞標準，分詞的單位的性質(*2)是：

(*2) 具有獨立意義且扮演固定詞性的字串。

因此，我們也需藉由對Head是否扮演固定詞性進行分析。換句話說，我們需觀察滿足性質(*1)的第一個Head，是否也滿足性質(*2)。為了簡化討論範圍，本文不考慮(s4)這種情形的例子。

---

[5] 關於詞性次分類的定義參照自(中文詞庫小組, 1993)

[6] 使用 CKIP 自動斷詞系統(http://godel.iis.sinica.edu.tw/CKIP/r_content.htm)處理類似的字串「清太宗皇太極」，會將「清太宗」分詞成「清」和「太宗」。

[7] http://www.sinica.edu.tw/~ndaplib/channels/dlm_paper/0910-05.pdf，CNS14366 國家標準 (Huang, 1998)。

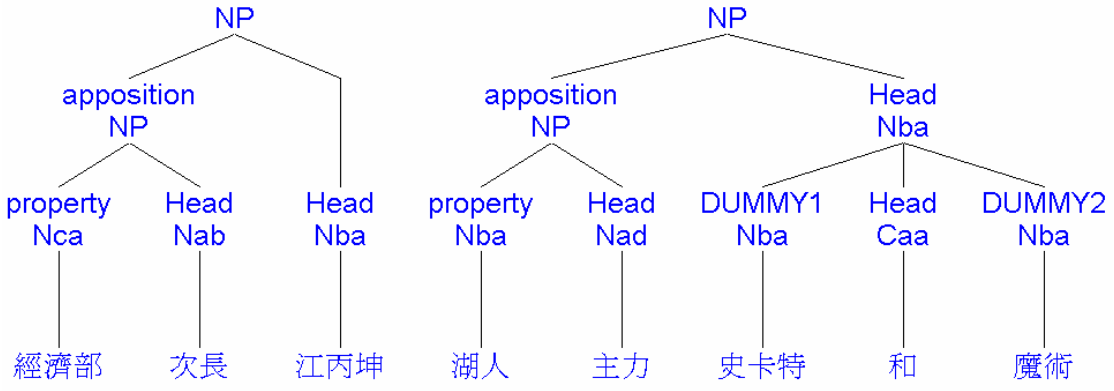

圖1 字串「經濟部次長江丙坤」和「湖人主力史卡特和魔術」的剖析樹

對照(s1)與(s3)的剖析樹，如圖1所示。可發現在人名之前的名詞片語所構成的兩個單位中，後者具有稱謂的語意特性。根據這樣的觀察，我們提出兩個假設(r1)及(r2)。在包含稱謂的名詞片語，其前後兩個單位：

(r1) 後者為比較廣泛的概念，不是專有名詞；
（次長、市長、主力）

(r2) 前者為特定的實體，通常是專有名詞。
（經濟部、三重、湖人）

(r2)應用的對象的是同位語名詞片語組中的第一個詞，在語法框架下具property特性。(r1)應用的對象是同位語名詞片語組的第二個詞，這一類詞彙也是NER模型中常用來偵測人名的線索(Chen *et al.*, 1998)。這類詞彙除了會出現在人名之前，也會出現出現人名之後。但是在中文句結構的分析中，人名之後所出現的稱謂會與該人名形成名詞片語，而不是上述分析中所出現之雙岔結構同位語現象，如下列兩個從TreeBank取出的名詞片語組：

NP(property:Nba:胡適 |Head:Nab:先生)；
NP(property:Nbc:周 |Head:Nab:老師)。

同樣的現象在新聞人名的表達上則較少出現，而且格式較為固定，通常是敘述是不需要具名或是不適合具名的新聞人物，如節錄2004/8/10的兩則新聞：

(東森新聞報) 王姓**民眾**在花蓮縣…
(中時電子報) 台北市某國小的徐姓男**老師**涉嫌…
(聯合新聞網) 台南市廢五金業馬姓**少東**…

這兩則新聞針對描述對象所使用的表達方式，除了沒有列出完整的姓名外，也不適用(r1)及(r2)所提出的假設。

本文分析出現在人名標記之前的同位語，篩選出常用於敘述新聞人物之稱謂結構，內容共分成四節，第一節描述目標與語言現象，第二節提出方法、實驗題材，並作評估。第三節顯示所自動擷取的知識庫和瀏覽介面。最後第四節作一總結。

2 研究方法與實驗

根據(r1)和(r2)這兩個假設，在不使用中文剖析器(Parser)的情況下，我們設計以下字串檢驗的方法。由於沒有高準確性之中文剖析器輔助，我們將觀察的對象調整成三連字串(trigram)ABC，在C為人名的前提下，測試問題(q1)及(q2)是否滿足：

(q1) A是常出現的名詞且符合(r2)；

(q2) 當A、C成立時，B滿足(r1)的可能性。

我們使用大量未經人工標記之新聞文件，經由分詞與詞性標記系統處理後，用finite-state的方式取得三連字串(trigram)──依序輸入字串Ａ、Ｂ及Ｃ，當state滿足C為人名且A為名詞時，即取出此trigram作為候選字串。接著設計演算法(詳見2.3節)，篩選候選字串，留下高頻率之字串集合。

2.1 實驗新聞語料

實驗所使用的語料來自三個新聞來源，分別是中央社從2002/1/1到2002/12/31共24,342則新聞、中時新聞從2002/1/1到2002/12/31共82,606則新聞、以及聯合新聞從2002/04/09到2003/06/26共100,617則新聞[8]。新聞的統計資訊如表1所示，詞數為經分詞與詞性標記系統處理(詳見2.2節)後得之。統計顯示中央社每則新聞平均詞數最多、聯合新聞則數最多、中時新聞平均單則新聞詞數最少。合計的語料詞數規模為1 : 2.33 : 3.86。

表1 新聞文件數量與詞數分季統計表

	中央社		中時新聞		聯合新聞	
	新聞則數	詞數	新聞則數	詞數	新聞則數	詞數
2002第一季	7,663	3,340,600	20,114	7,879,821		
2002第二季	6,500	3,048,492	21,514	6,946,475	21,694	8,429,921
2002第三季	5,290	2,901,751	20,436	6,528,886	30,961	13,054,258
2002第四季	4,889	2,676,842	20,542	6,577,950	28,912	13,522,185
2003第一季					10,215	6,048,806
2003第二季					8,835	5,159,618
合計	24,342	11,967,685	82,606	27,933,132	100,617	46,214,788
數量比例	1	1	3.39	2.33	4.13	3.86
平均詞數(比)	491.65 (1)	338.15(0.69)	459.3139			

2.2 分詞與詞性標記系統

本實驗使用台大自然語言處理實驗室分詞與詞性標記系統(以下簡稱Tagger)，處理上述所有新聞文件。該系統核心採取Tigram Markov模型，整合入NER[9]模組(Chen *et al*., 1998)，其中本實驗著重之人名辨識核心是在系統中是(Chen and Lee, 1996) 所提出的架構。該系統針對以句子為單位之字串如「財政部次長林宗勇表示，」產生如下之分詞與詞性標記結果：

財政部(Nc) 次長(Na) 林宗勇(Nb_PERSON) 表示(VE) ，(COMMACATEGORY)

下節敘述將Nc、Na、 Nb_PERSON等詞性，或具名實體標記統稱為「詞性碼」。

[8] 聯合新聞語料日期之收集較前兩家晚一至二季，可保留偵測新人名之擴充性。

[9] NER 辨識系統之使用可參照 http://nlg.csie.ntu.edu.tw/

2.3 演算法

實驗新聞語料經由Tagger處理後，採取計算詞頻的方式，來篩選新聞人物常使用的同位語稱謂 。其演算法詳述如下：

(1) 從Tagger結果中找出任何9-gram[10]，令其為XXXABCXXX，當其中的mid-trigram(亦即ABC)滿足第一字A的詞性碼第一碼為N，且第三字 C 的詞性碼為Nb_PERSON。

(2) 蒐集每一種tri-gram與其詞性配對<A, tagA, B, tagB, C, tagC>形成集合S，並統計其出現頻率 $freq_{<A, tagA, B, tagB, C, tagC>}$ 。

(3) 統計第二詞及詞性配對<B, tagB>出現的頻率$freq_{<B, tagB>}$。

(4) 篩選B詞長度大於2，tagB第一碼為N，且$freq_{<B, tagB>}$大於β [11]。

(5) 人工檢驗選出之B詞是否為稱謂，通過檢驗之B詞集合為S_B。

(6) 取子集合S'={s ls屬於S, 且s中對應的<B, tagB>屬於S_B、$freq_{<A, tagA, B, tagB, C, tagC>}$＞1}。以通過檢驗之稱謂篩選S並保留其詞頻>1之資訊。子集合說明篩選之同位語片語除人名 C前出現常見之稱謂B外，同樣的同位語片語ABC必須出現兩次以上。)

2.4 評估分析

表2列出三家媒體選出的稱謂字數分別為65、103及125，比例是1 : 1.58 : 1.92。中央社新聞篩選出的稱謂正確率達100%、中時新聞為97.01%、聯合新聞為93.60%。我們發現篩選的稱謂詞數，跟新聞語料分詞的數量，呈現正相關，但不是以等比增長。這顯示新聞人物的稱謂不會毫無限制地增加，應該會限定在一個範圍內。

第2.3節演算法(1)-(5)步驟，找出所有trigram ABC中限定A、C的詞性之後，B的詞性碼第一碼為N的詞。(因此B在這樣的限定下tag可能是Na、Nb、等合法的名詞。)接著檢驗問題(q2)，我們觀察B是否符合(r1)為廣泛的稱謂概念(通常是Na)。觀察由中央社篩選出的65個名詞列出如表3所示，列出的65個詞在Tagger皆標記成Na，因此假設(r1)大致成立。若將演算法步驟(4)的β設成50，則會篩選出102個詞，其中B不是Na或意思有誤的有8組，分別是：

第67名的「批評」	(Na;	頻率99)、
第71名的「要求」	(Na;	頻率89)、
第75名的「支持」	(Na;	頻率81)、
第78名的「政務委員」	(Nb;	頻率70)、
第83名的「土耳其」	(Nc_LOCATION;	頻率65)、
第87名的「國小」	(Nc;	頻率61)、
第90名的「颱風」	(Na;	頻率59)、
第97名的「行政長官」	(Nb;	頻率53)，

其中除「政務委員」、「行政長官」應是Tagger的錯誤，其他三種錯誤有動詞名物化如「批評」、地名如「土耳其」、事物擬人化如「颱風」等，系統正確率因此降至94.12%。

[10] 取 9-gram 可保留觀察前後文之擴充性。

[11] β 為本實驗自訂之參數，β=100 次，要求稱謂字串出現頻率要超過 100 次。

表2 自動篩選稱謂之結果統計表

	自動選出 B字(比)	人工檢驗 正確字數	正確率	正確稱謂 詞頻第一名	正確稱謂 詞頻第二名	錯誤稱謂 詞頻第一名
中央社	65(1)	65	100.00%	立委(3,656)	總統(3,119)	
中時新聞	103(1.58)	100	97.01%	局長(2,742)	立委(2,414)	選區(166)
聯合新聞	125(1.92)	117	93.60%	總經理(5,671)	董事長(5,599)	分析(824)

表3 中央社新聞稱謂篩選統計表

稱謂	頻率	稱謂	頻率	稱謂	頻率	稱謂	頻率	稱謂	頻率
立委	3656	立法委員	618	執行長	352	副院長	198	大使	143
總統	3119	副主席	560	國務卿	351	資政	194	里長	141
市長	1991	縣長	549	院長	344	司長	187	首相	138
主席	1836	候選人	511	會長	335	負責人	185	參議員	134
秘書長	1155	處長	491	召集人	316	副總統	184	秘書	132
主委	1102	教授	481	外長	295	經理	184	總幹事	125
黨主席	1094	領袖	454	祕書長	291	行政院長	179	委員長	125
局長	1087	代表	452	署長	277	校長	178	鄉長	120
部長	1038	總經理	443	國防部長	270	領導人	175	議員	112
董事長	1014	次長	420	夫人	263	總裁	173	研究員	107
主任	1009	議長	410	市議員	260	專家	151	常委	105
發言人	966	理事長	400	顧問	235	副總理	147	官員	105
總理	734	委員	363	檢察官	214	組長	146	記者	102

表4 新聞人物同位語結構字串篩選統計表

		中央社		中時新聞		聯合新聞	
篩選數(比)		2,399(1)		3,927(1.64)		5,463(2.28)	
前50名	1.專有地名	16	32%	14	28%	14	28%
	2.專有組織名	4	8%	6	12%	5	10%
	3.普通組織名	10	20%	9	18%	4	8%
	4.普通名詞	16	32%	17	34%	24	48%
	5.斷詞錯誤	4	8%	4	8%	1	2%
	6.無法判斷					2	4%
正確字串頻率第一名		台北市長馬英九		台北市長馬英九		美國總統布希	
正確字串頻率第二名		美國總統布希		美國總統布希		台北市長馬英九	
正確字串頻率第三名		中國國民黨主席連戰		法務部長陳定南		國家主席江澤明	

表4列出三家媒體篩選出的同位語結構字串數分別為2,399、3,927、5,463，比例是1：1.64：2.28，數量上依然跟新聞語料分詞的數量呈現正相關。雖然倍數不是以等比增長，但是較稱謂筆數的比例來的放大，顯示同樣一種稱謂可能可以成功套用在多筆結構上，如「民進黨立委蔡同榮」、「國民黨立委陳學聖」等。

我們分別針對三家新聞媒體語料，抽出出現頻率前50名同位語結構字串，針對trigram的A作型態的判斷，共分成六種類型：

(1) 專有地名：　　如台北、美國等。

(2) 專有組織名：　　如中國國民黨、民進黨等。

(3) 普通組織或地名(Nc)：　如總統府、行政院等。

(4) 普通名詞(Na)：　　如國家(主席)、法務(部長)等。

(5) 斷詞錯誤：　　如國民(黨主席)、(親)民黨(副主席)。

(6) 無法判斷：　　如橘子(董事長)──橘子標記成名詞，其實是專有組織名、鴻海(董事長)──鴻海標記成人名，其實是專有組織名。

結果顯示各家媒體使用專有名詞的部分(1)+(2)的比例約為40%，使用普通名詞的比例約為50%，錯誤及無法判斷的比例在10%以下。

回到問題(q2)，我們觀察A是否符合(r2) 為特定的實體，由實驗結果顯示僅有 40%的機會為專有名詞，許多Na及Nc都可以用來限定稱謂，例如用「集團」(Na)限定「董事長」、用「總統府」(Nc)限定祕書長。

另外，使用三連字串的設計來偵測字串本身即有限制，如果是以名詞片語「政治大學(Nb)新聞系(Nc)教授(Na)」為例，其同位語架構中的A，即包含了開頭的Nb和Nc兩部份，因此需要擴充2.3節演算法的步驟(2)，蒐集四連字串才能取得。以中央社的新聞為例，以擴充後的演算法，透過四連字串可選出861筆資料，茲列出前十名如表5所示。使用五連字串可選出308筆資料，前十名如表6所示。使用六連字串[12]可選出112筆資料，前十名如表7所示。

針對使用不同長度的n-gram所獲得的結果顯示，使用的n-gram越大，獲得的詞數就越少，而且所需要的計算時間隨之增長；若欲透過增加語料庫規模來增加獲得的詞量，也會增加計算的時間。表5、表6及表7列出結果中標示錯誤的地方，是由於Tagger將「副總」分成一個詞，以及沒有收錄「(親)民黨」詞彙所造成。

綜合以上三個表格所提供之觀察，發現同位語元素Head B仍是緊鄰在人名之前，因此驗證了本文所提出演算法之可行性，但是property A的長度限制會是問題所在。同樣的問題也說明了設計一個良好的中文剖析器會遇到的困難──因為無法保證同位語結構中A能允許的詞數能有多長。

表5. 中央社新聞人物同位語結構四連字串前十名列表

正確	<A_1, tagA_1, A_2, tagA_2, B, tagB, C, tagC>								次數
√	中共	Nb	國家	Na	主席	Na	江澤民	Nb_PERSON	319
√	中央	Nc	委員會	Nc	祕書長	Na	林豐正	Nb_PERSON	149
√	中共	Nb	國家	Na	副主席	Na	胡錦濤	Nb_PERSON	147
√	中央	Nc	委員會	Nc	秘書長	Na	林豐正	Nb_PERSON	139
√	發展	Na	委員會	Nc	主委	Na	陳健治	Nb_PERSON	103
√	政策	Na	委員會	Nc	執行長	Na	曾永權	Nb_PERSON	93
√	高雄	Nc_LOCATION	市議會	Nc	議長	Na	黃啟川	Nb_PERSON	79
√	臺北	Nc_LOCATION	市長	Na	候選人	Na	李應元	Nb_PERSON	68
X	黨團	Na	副總	Na	召集人	Na	秦慧珠	Nb_PERSON	67
√	高雄	Nc_LOCATION	市長	Na	候選人	Na	黃俊英	Nb_PERSON	64

[12]由於 2.3 節演算法步驟(1)從語料庫中保留了 9-gram，因此本實驗可偵測字串之最長長度，是使用六連字串的所偵測出的 AAAABCXXX。

表6. 中央社新聞人物同位語結構五連字串前十名列表

正確	<A_1, tagA_1, A_2, tagA_2, A_3, tagA_3, B, tagB, C, tagC>										次數
√	組織	Na	發展	Na	委員會	Nc	主委	Na	陳健治	Nb_P	103
√	中國國民黨	Nb_O	中央	Nc	委員會	Nc	祕書長	Na	林豐正	Nb_P	62
√	國民黨	Nb_O	中央	Nc	委員會	Nc	秘書長	Na	林豐正	Nb_P	56
√	中央	Nc	政策	Na	委員會	Nc	執行長	Na	曾永權	Nb_P	54
√	民進黨	Nb_O	臺北	Nc_L	市長	Na	候選人	Na	李應元	Nb_P	52
√	中國國民黨	Nb_O	中央	Nc	委員會	Nc	秘書長	Na	林豐正	Nb_P	38
√	中國	Nc_L	大陸	Nc	國家	Na	主席	Na	江澤民	Nb_P	33
X	立法院	Nb_O	黨團	Na	副總	Na	召集人	Na	秦慧珠	Nb_P	32
√	白宮	Nb_O	國家	Na	安全	Na	顧問	Na	萊斯	Nb_P	31
√	經濟	Na	研究	Na	中心	Nc	主任	Na	林毅夫	Nb_P	30

表7. 中央社新聞人物同位語結構六連字串前十名列表

正確	<A_1, tagA_1, A_2, tagA_2, A_3, tagA_3, A_4, tagA_4, B, tagB, C, tagC>												次數
√	中國國民黨	Nb_O	組織	Na	發展	Na	委員會	Nc	主委	Na	陳健治	Nb_P	55
√	國民黨	Nb_O	組織	Na	發展	Na	委員會	Nc	主委	Na	陳健治	Nb_P	34
X	民黨	Nb	立法院	Nb_O	黨團	Na	副總	Na	召集人	Na	秦慧珠	Nb_P	30
√	中國	Nc_L	經濟	Na	研究	Na	中心	Nc	主任	Na	林毅夫	Nb_P	30
√	國民黨	Nb_O	中央	Nc	政策	Na	委員會	Nc	執行長	Na	曾永權	Nb_P	28
√	中國國民黨	Nb_O	中央	Nc	政策	Na	委員會	Nc	執行長	Na	曾永權	Nb_P	23
√	行政院	Nb_O	農業	Na	委員會	Nc	主任	Na	委員	Na	范振宗	Nb_P	19
√	美國	Nc_L	聯邦	Na	準備	Na	理事會	Na	主席	Na	葛林斯潘	Nb_P	17
X	民黨	Nb	立院	Nb_O	黨團	Na	副總	Na	召集人	Na	秦慧珠	Nb_P	14
X	民黨	Nb	立法院	Nb_O	黨團	Na	副總	Na	召集人	Na	李鴻鈞	Nb_P	13

3 知識庫整理與視覺化瀏覽介面

目前由演算法自動擷取出來的各報社人名知識庫 (如中央社的2,399筆資料)，透過知識本體介面對這些資料進行瀏覽，可以協助系統或使用者了解與整理所擷取的知識。圖2及圖3分別使用中央社新聞及中時新聞，以樹狀結構的方式，對照兩個新聞資料庫所收錄的人名實體及事件關係。由左側視窗Root「新聞人物」，首先列出通過檢測之新聞人物稱謂成為children(trigram的B部分)(如中央社的65筆資料)，接著列出該稱謂的所有property(trigram的 A部分)，圖示中同一個node的children以顯示五筆為上限。右側上半部列出具有稱謂人物後面常出現之事件，例如點選「民進黨」後出現之動詞片語如所列。下半部分則列出點選稱謂下特定property所代表人物後面常出現之事件。

以圖2為例，「立委」人物的後面常出現的動詞依序有有「表示」、「指出」、「質詢」、等等，與立委分類中的「民進黨立委」，後面常出現動詞的順序大致相符，顯示中央社報導的立委行為有一定的模式。圖3同樣是以「立委」為例，但中國時報的動詞順序便有些許不同，不過大致上仍是相符。另外觀察的window size若放大到2，也可計算出常出現的動詞片語有「昨天召開」、「昨日表示」、等等。

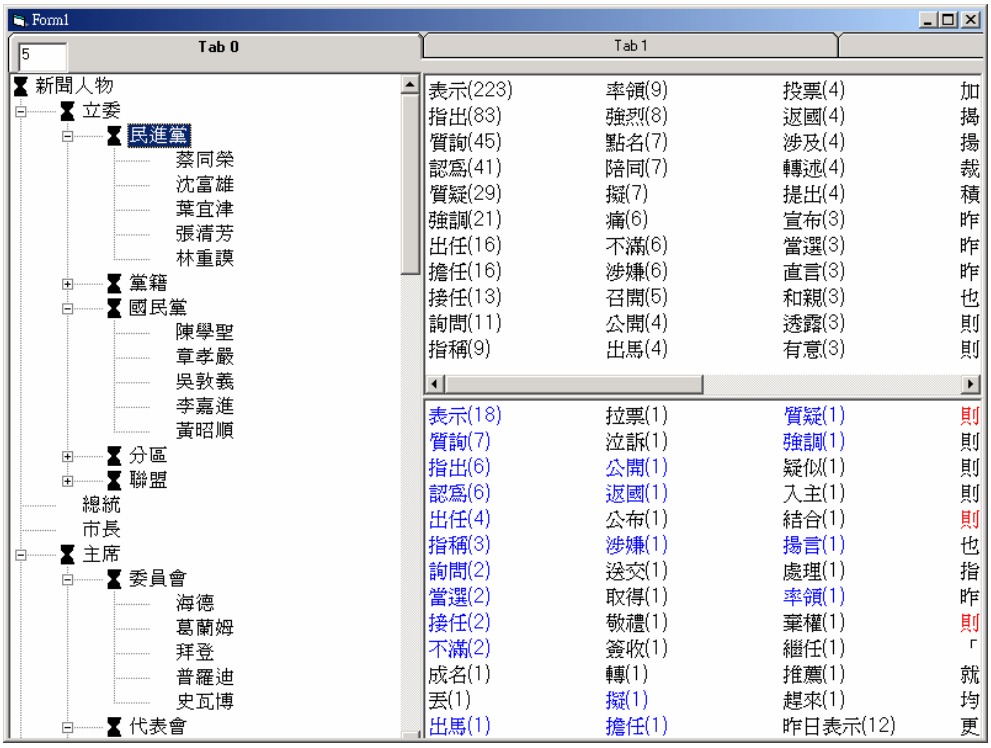

圖2 瀏覽中央社新聞人名之知識庫介面

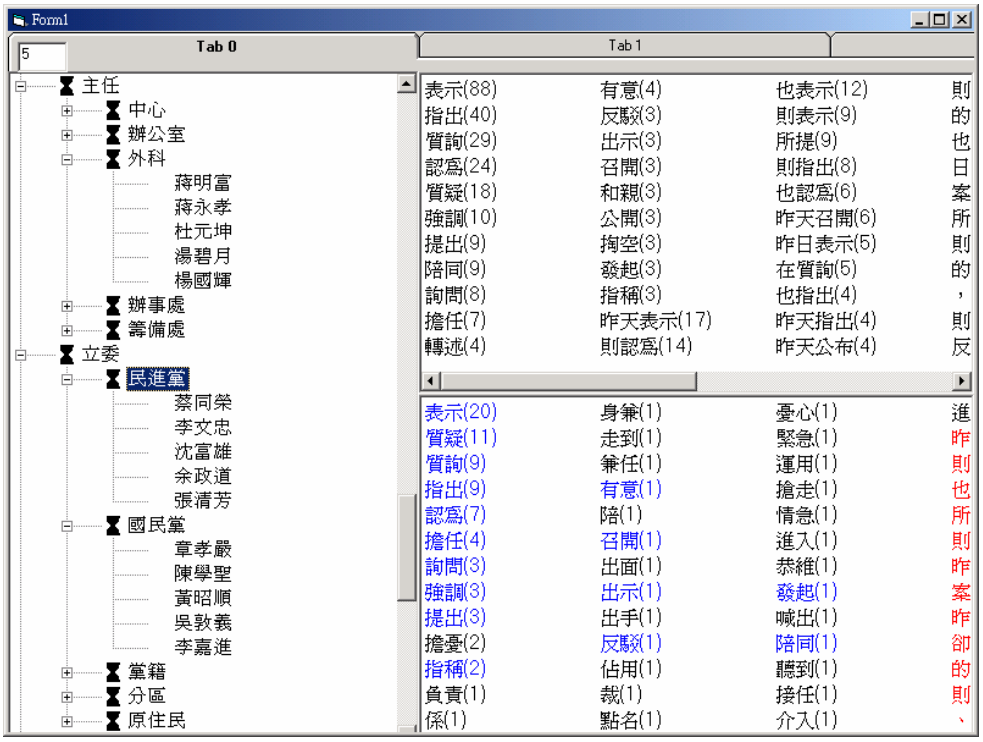

圖3 瀏覽中時新聞人名之知識庫介面

4 結論

新聞媒體透過網路環境突破時空的限制，隨著時間不斷的推移提供更多更新的文件。一個與時並進的自然語言系統，也可以利用網路新聞的這項特質，獲取最新的字詞統計資訊，以延伸出各種形式的應用，如對n-gram機率的估計、衡量(單語或跨語言)字詞間的相關度、建立同義詞集或知識本體(Ontology)等等。

透過本文敘述的方法所建立的人名實體與事件關係的雛型，將可以協助新聞文件的解析和機器翻譯的應用。例如在一個敘述多個實體的事件報導中，運用本知識庫所建立的概念關係，可以挑出適合的指涉人物，以解決指涉問題。例如提及「立委」和「主任」之後，誰是接著敘述「涉嫌...」事件所指涉的對象。另外，如英文的NER及剖析系統技術都較中文純熟，如果能用同樣的機制針對英文新聞建立知識庫，將可對不同語言，但屬於新聞領域知識庫所包含的實體及事件，透過連結的方式，建立跨語言的對應關係。

註謝

本文部分成果為國科會計畫NSC 93-2752-E-001-001-PAE補助。

參考文獻

Hsin-Hsi Chen, June-Jei Kuo, Sheng-Jie Huang, Chuan-Jie Lin and Hung-Chia Wung (2003). "A Summarization System for Chinese News from Multiple Sources." Journal of the American Society for Information Science and Technology, 54(13), pp. 1224-1236.

Hsin-Hsi Chen, Yung-Wei Ding and Shih-Chung Tsai (1998). "Named Entity Extraction for Information Retrieval." Computer Processing of Oriental Languages, Special Issue on Information Retrieval on Oriental Languages, 12(1), pp. 75-85.

Hsin-Hsi Chen and Jen-Chang Lee (1996). "Identification and Classification of Proper Nouns in Chinese Texts." Proceedings of 16th International Conference on Computational Linguistics, Copenhagen, Denmark, August 5-9, pp. 222-229..

李祥賓、柯淑津 (2001)，「新聞文件摘要之研究」，第十四屆計算語言學研討會論文集，pp. 23-42，台南成功大學。

馬偉雲、謝佑明、楊昌樺、陳克健 (2001)，「中文語料庫構建及管理系統設計」，第十四屆計算語言學研討會論文集，pp. 175-191，台南成功大學。

中文詞知識庫小組 (1993) ，「中文詞類分析(三版)」。