資訊檢索查詢之自然語言處理

陳光華

國立臺灣大學圖書館學系

khchen@nlg.csie.ntu.edu.tw

【摘要 Abstract】

資訊檢索系統的發展已行之有年，然而使用者查詢的方式仍是以關鍵詞爲基礎，並且結合布林邏輯運算。對於使用者而言，最自然的查詢方式是使用自然語言查詢所需要的資訊。尤其當語音輸入技術達到實用階段時，自然語言查詢的需求，將益形重要。少數具有自然語言查詢方式的系統，使用的僅是樣式比對（Pattern Matching）的技術，無法知道各詞彙之間的關係，當然也就無法知道各詞彙扮演的角色。本文提出一種自然語言剖析的技術，可以快速並且穩定地剖析自然語言，分析各詞彙的關係與彼此的角色，而建構這種剖析技術所需的資源僅僅是具有詞類標記的語料庫。在分析大規模的實驗結果之後，根據Crossing準則，平均準確率爲81%；根據PARSEVAL準則，平均的召回率與精確率皆爲33%。

Many researchers have devoted to developing information retrieval systems for a long time. However, the form which users use to retrieve information is still keyword-based queries with Boolean operators. From user's point of view, the most natural way to issue queries is using they mother tongue. That is, users are more likely to apply natural language queries to retrieve useful information. Some systems accept natural language queries, but the technique they use is simple pattern-matching. Such kind of technique does not discover the relationship among the keywords, and then does not understand the roles which these keywords play in natural language sentences. This paper proposes a new parsing technique, which can quickly and stably parse natural language sentences. Moreover, this parsing technique can find out the relationship among keywords and the roles of each grammatical constituents. The resource needed to construct a parser based on the proposed parsing technique is tagged text corpora. After a large-scale experiment, the recall and the precision are 33% based on the PARSEVAL criterion. The accuracy is 81% based on the Crossing criterion.

〔關鍵字 Keywords〕：

資訊檢索；自然語言處理；剖析技術

Information Retrieval；Natural Language Processing；Parsing Technology

一、緒論

自古以來，人類探詢資訊的行爲就已經非常普遍。例如，由天空的色彩與雲狀，得知天氣的好壞；由太陽的位置，得知時間的訊息；由莊稼的收成，得知地利的良莠；由他人的表情與姿勢，得知言下之意。尚書堯

典也曾記載堯帝命羲氏與和氏推算日月星辰運行的規律：「乃命羲和親若昊天曆象日月星辰敬授人時」(註1)。

時序進入科學昌明的時代，人類對於資訊的需求益形殷切，培根說：「知識就是力量」(註2)，這代表人類不斷追尋資訊的潛在動力。而知識的形成卻是需要資訊不斷的累積、分析與合成。如今，世界最大的圖書館-美國國會圖書館，館藏量已達九千萬件（註3），而隨著網際網路的蓬勃發展，人類面臨了資訊爆炸的問題，獲得有用資訊的代價越來越高。如何協助讀者或是尋求資訊的人們取得有用的資訊，成爲圖書館研究領域中非常重要的課題。資訊檢索系統的研究已經進行很長的一段時間，其目的在於協助使用者從浩瀚的資訊汪洋，取得有用、適用的資訊。目前的檢索系統提供的檢索方式不外乎關鍵字或是索引詞彙加上布林運算子的組合。對於使用者而言，最自然的查詢方式是使用自然語言查詢需要的資訊。少數具有自然語言查詢方式的系統，使用的僅是樣式比對的技術，在資訊檢索的環境裡，樣式指的是關鍵詞或是索引詞彙。關鍵字比對的優點是：

- 處理速度快
- 簡單

但是，此法也有重大的缺點，例如無法知道各詞彙之間的關係，當然也就無法知道各詞彙扮演的角色。查詢句中各詞彙到底是主語、謂語、還是賓語，無法透過簡單的關鍵字比對得到解答。想要解決這些問題，必須對查詢句做進一步的處理，也就是剖析（Parsing）的工作，而且在資訊檢索的環境中，這樣的處理時間不能過長。

有相當多的專家學者提出各種不同剖析自然語言的演算法，早期規則式的剖析技術需要語言學專家整理分析語法規則，耗費極大的人力、物力。當規則式系統的規模不大的時候，通常相當有效率；然而當系統逐漸成長（規則越來越多），規則間的不一致現象造成系統發展者很大的困擾，也使得系統效率越來越差。最後往往只能處理許多語言的核心現象，通常無法處理真實生活的用語。這也是有些人戲稱這一類的剖析系統是玩具系統（Toy System）的原因。

近來使用語料庫（註4）進行自然語言的研究逐漸活絡起來。主要的原因是，語料庫提供真實的語料，透過觀察大量的語料庫，可知道人們生活起居語言使用的情形。其次，電腦科技的突飛猛進，有快速的微處理器足以進行複雜的語言模型計算；有高容量的儲存媒體得以儲存訓練語料，使得早在60年代就已經存在的研究方法，再度呈現新的生命。

筆者根據語料庫研究方法的脈絡，提出一個快速的、初步的自然語言剖析技術（Parsing Technology），基本上具有下列的特性：

- 使用詞類資訊，進行自然語言剖析
- 剖析技術容易移轉至其他語言
- 建立二元剖析樹
- 剖析速度快

本文第二節介紹筆者提出的剖析技術的基本觀念；第三節則說明根據資訊理論，如何度量詞彙與詞彙之間的關係；第四節討論使用Crossing與PARSEVAL兩種準則，評估實驗的結果；第五節提出剖析演算法，並且說明訓練語料庫以及測試語料庫，最後進行一連串的實驗；第六節說明這個剖析技術可能的後續研究與應用；第七節做一個簡短的結論。

二、自然語言剖析

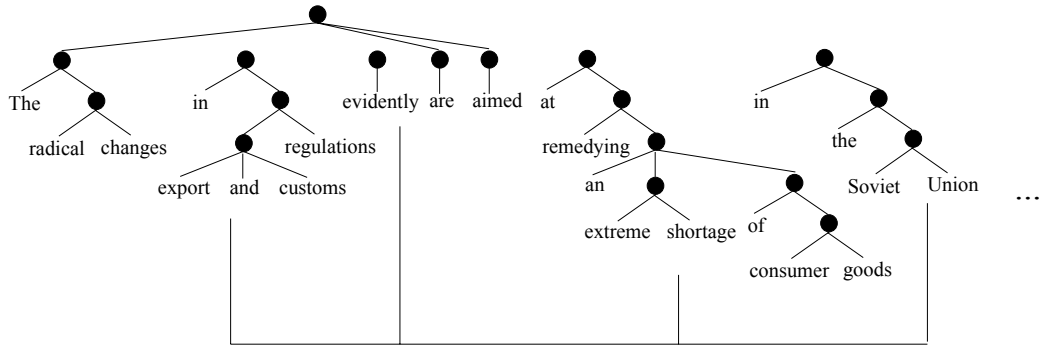
剖析自然語言一直是瞭解語言最初步的工作，然而卻不是簡單的工作。一個非常簡單的句子爲例：

Jordan watched a girl with a telescope.

介詞組with a telescope可以修飾名詞組a girl或是修飾動詞組watched a girl。雖然在多數的情況下，人類能夠瞭解上面的句子指的是Jordan用望遠鏡看一位女子，而非看著一位攜帶望遠鏡的女子。但是要電腦正確地判斷，卻需要許多細微而複雜的知識。例如，電腦得知道望遠鏡是一種可用來觀望的工具，因此與動詞watched的關連性比較高，而且必須知道一般女子攜帶望遠鏡的情形比較少見。

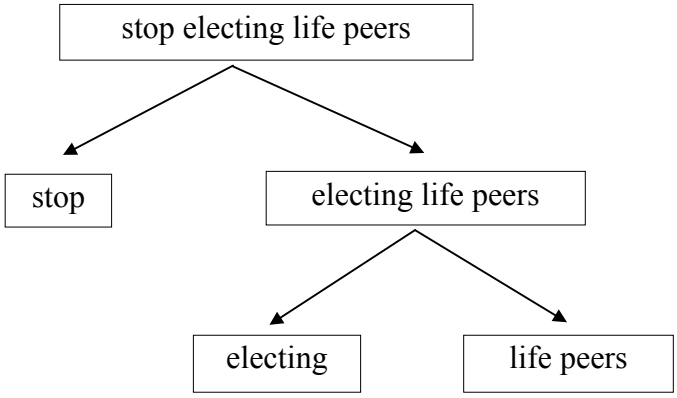
事實上，有不少專家學者提出不同的辦法（註5），盡可能地解決這類詞組併接（Phrases Attachment）的問題，然而自然語言畢竟過於複雜，剖析的結果常常是一座剖析森林（Parsing Forest）而非一棵剖析樹（Parsing Tree）。以Hindle提出的Fidditch剖析程式爲例，剖析下面這個句子時，產生的就是一座剖析森林：

> The radical changes in export and customs regulations evidently are aimed at remedying an extreme shortage of consumer goods in Soviet Union …



無法併接的語法成分

圖一、Fidditch剖析器產生的剖析森林



圖二、自然語言的二元樹狀結構

圖一表示相對的剖析森林，圖中清楚的表示筆者稱之爲森林的原因，因爲許多語法成分無法互相併接，構成一棵完整的樹，因此成爲一棵棵的樹。企圖使用語意的知識解決併接問題的努力，仍然持續地進行中。但是，如何適切地表達知識卻又是另一個重大的研究課題。（註6）而語料庫卻提供了另一扇窗戶，許多語料庫語言學專家（註7）發現若能夠使用大量的語料庫，建構適當的統計式語言模型，許多語法與語意等等的語言知識都能夠用數值化的統計參數表現。對於筆者於本文提出的剖析技術，也將由這一個角度出發。（註8）

當看到一系列的詞彙時，人類如何建構這些詞彙彼此之間的關係。以“stop electing life peers”爲例（註9），可以明確地知道“life”與“peers”的關係最密切，接著是“electing”與“life peers”，然後才是“stop”與“electing life peers”。換另一個角度，也可以視爲將這個句子重複地分成兩段，直到最後的一段剩下兩個詞彙爲止。按照這樣的關係可繪出一棵樹狀結構，或許也可稱之爲剖析樹（爲了方便起見，筆者爾後將直接以剖析樹稱之），請見圖二。前述的例子直接以詞彙本身作爲考慮的因素，然而，以LOB Corpus有49,426個詞彙爲例（註10），考慮詞彙兩兩之間的關係，總共有2,442,880,050個統計參數，以LOB Corpus的大小而言，這樣的參數量過大，參數值的信賴度比較低，下面將以詞類取代詞彙，這個動作相當於將詞彙分類。LOB Corpus總共有134個詞類，參數量可以有效地降低，信賴度也比較高。筆者以前述的句子加以說明：

stop_VB electing_VBG life_NN

peers_NNS（註11）

對應的樹狀結構爲[[VP：stop_VB [VP：electing_VBG [NP：life_NN peers_NNS]]]。這種逐步決定詞彙（或是詞類）間關係的作法有許多好處：

- 產生的二元剖析樹，可以在不同的情況下，轉換爲三元或是其他型態的剖析樹。以此觀點，甚至可以稱二元剖析樹爲正規剖析樹結構（Canonical Tree Structure）。
- 只要擁有一份詞類標記的語料庫，就可以建構這樣的剖析程式。以此觀點，可以將其視爲與語言無關的剖析技術。
- 不需要內建的語法規則（無論是定率式或是機率式），因此剖析的速度非常快。

現在的問題是如何度量詞類之間的「關係」。筆者將於下一節詳細討論這一個關鍵的問題。

三、詞類的關聯性

對於給定的詞類序列，$t_1, t_2, \ldots, t_{n-1}, t_n$，如果想要將這串詞類序列分成兩部份，如何決定這一刀要切在什麼地方？請考慮 t_i與 t_{i+1}之間的第 i 個位置，假如這個位置是可能性最大的切分點，根據 Shannon 資訊理論的觀點（註 12），能夠在 t_i後面出現的詞類數目以及後面能夠接 t_{i+1}的詞類數目，相對地會比其他的詞類配對多。除此之外，t_i與 t_{i+1}共同出現的頻率也同樣相對地比較低。依照上面這些基本觀念，可以比較正式地使用數學式定義正向熵（Forward Entropy, FE）、逆向熵（Backward Entropy, BE）、以及共容資訊（Mutual Information, MI）。

定義一：正向熵（FE）。詞類t_i的正向熵為

$$FE(t_i) = -\sum_{t_j \in Tagset} P(t_j|t_i)\log P(t_j|t_i)$$

定義二：逆向熵（BE）。詞類t_i的逆向熵為

$$BE(t_i) = -\sum_{t_j \in Tagset} P(t_j|t_j)\log P(t_i|t_j)$$

這裡*tagset*表示使用的詞類標記集合，由於筆者使用LOB Corpus，因此，也就是134個詞類標記。根據定義一與定義二，當可以在t_i後面出現的詞類數目以及後面可以接t_{i+1}的詞類數目越多時，也就是選擇性越大時，亂度越大，能量越高，正向熵與逆向熵也就越高。

定義三：共容資訊（MI）。詞類t_i與t_j的共容資訊為

$$MI(t_i,t_j) = \log\frac{P(t_j|t_i)}{P(t_j)} = \log\frac{P(t_i,t_j)}{P(t_i)P(t_j)}$$

共容資訊的意義是，當t_i與t_j經常一起在語料庫出現，聯合機率$P(t_i,t_j)$會甚大於$P(t_i)\times P(t_j)$，因此$MI(t_i,t_j)$會甚大於0；當t_i與t_j出現的方式是背道而馳時，$MI(t_i,t_j)$會甚小於0；當彼此沒有什麼關係時（以機率論的術語而言，也就是互相獨立），因此$P(t_i,t_j)\approx P(t_i)\times P(t_j)$，所以$MI(t_i,t_j)$接近於0。

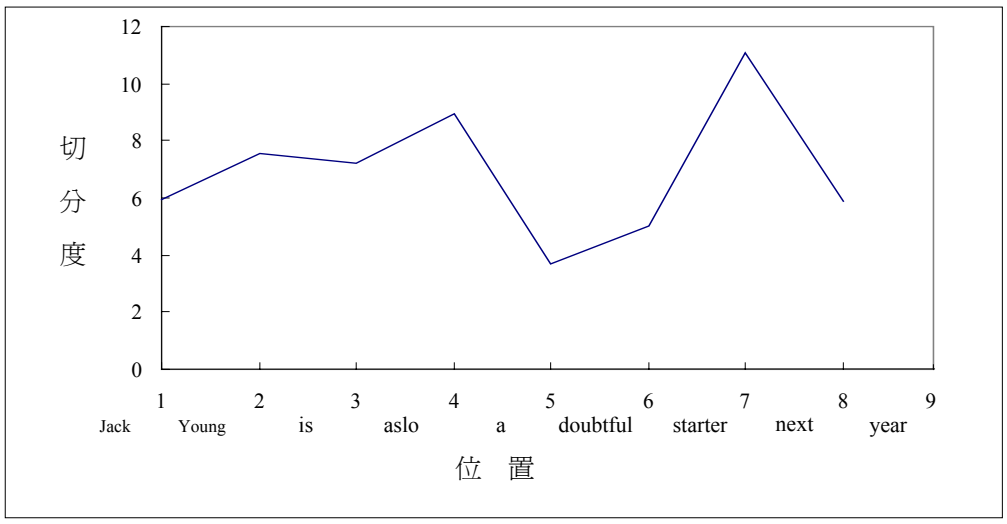
定義四：切分度（Cuttability Measure, CM）。詞類t_i與t_{i+1}的切分度為

$$CM(t_i,t_{i+1}) = FE(t_i) + BE(t_{i+1}) + MI(t_i,t_{i+1})$$

根據定義四，詞類之間的切分度越高，表示這個位置越有可能被切開。也就是當接收一串詞類序列後，經過上述的計算程序，可以知道這些詞類相對的切分度高低。接著從切分度最高的位置開始分割，然後再由分成兩段的詞類序列中，選擇切分度最高的位置，進行切分的工作。依此程序，直到剩下兩個詞類標記爲止。筆者再以下面的例子仔細說明：

> Jack_NP Young_NP is_BEZ also_RB a_AT doubtful_JJ starter_NN next_AP year_NN ._.

圖三表示的是這個句子的切分度。其中切分度最高的地方是starter next之間的位置，因此就由這個位置下第一刀，依序是also與a之間、Young與is之間、is與also之間、Jack與Young之間、next與year之間、doubtful與starter之間、a與doubtful之間的位置。因此相對的剖析樹爲[[[[Jack_NP Young_NP] [is_BEZ also_RB]] [[a_AT doubtful_JJ] starter_NN]] [next_AP year_NN]]。



圖三、範例的切分度圖示。

在前面的討論中，可以發現，這種剖析技術所需要的訓練語料只是大量的詞類標記語料庫，而這一類型的語料庫是目前各種語言語料庫中，比較容易取得的資源。例如，LOB Corpus擁有一百萬詞，經過標記的英國英語語料；Brown Corpus（註13）擁有一百萬詞，經過標記的美國英語語料；我國中央研究院已經完成三百萬詞平衡式漢語標記語料庫。（註14）下列的演算法表示整套剖析技術。

演算法： BuildTree(*W*, *T*, *i*, *j*, *tree*)
Input: a given word sequence, $W = w1, w2, ..., wn-1, wn$, and its corresponding tag sequence, $T = t1, t2, ..., tn-1, tn$. $i = 1$; $j = n$.
Output: a binary parsing tree, *tree*.
if (i == j) tree= "wi_ti"
else if (i == j-1) tree= "[wi_ti wj_tj]";
else
    $k = \arg\max_{i \le l < j} CM(t_l, t_{l+1})$;
    BuildTree(*W*, *T*, *i*, *k*, *ltree*);
    BuildTree(*W*, *T*, *k*+1, *j*, *rtree*);
    tree= "[ltree rtree]";
endif

而演算法所需要的參數值*FE*、*BE*、*MI*、或是*CM*，可以由訓練語料庫統計取得。

四、如何評估剖析結果

筆者已經說明如何使用標記語料庫建構剖析程式，然而如何評估這樣的剖析方式到底是好還是壞？Black在1991年提出了兩種評估準則（註15），一爲PARSEVAL；另一爲Crossing。前者是比較嚴格的準則；後者卻是比較鬆的準則。下面是這兩種準則的定義。

定義五：PARSEVAL。
PARSEVAL考慮兩個廣泛使用於資訊檢索的兩種評估方式：精確率（Precision）與召回率（Recall）。

$$\text{Precision} = \frac{\#\text{of correct constituents}}{\#\text{of produced constituents}}$$

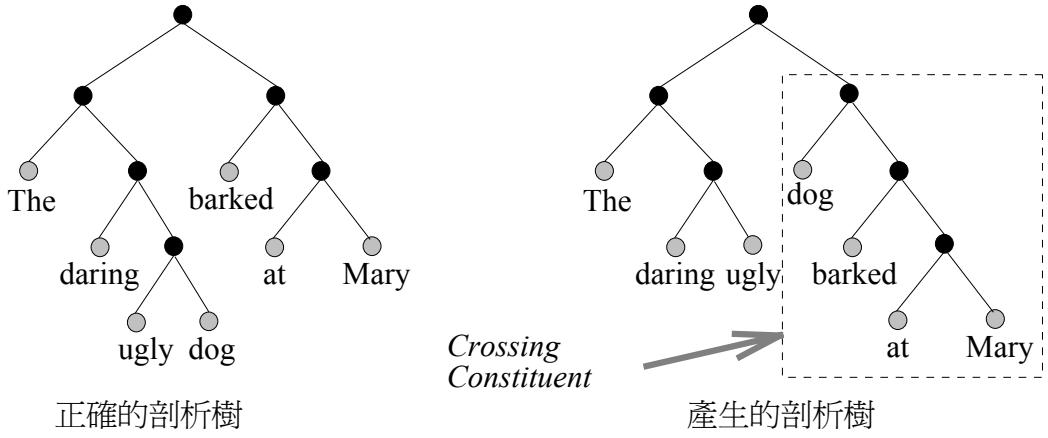
$$\text{Recall} = \frac{\#\text{of correct constituents}}{\#\text{of constituents in treebank}}$$

例如，"the daring ugly dog barked at Mary"的正確剖析樹爲[[the_AT [daring_JJ [ugly_JJ dog_NN]]] [barked_VBD [at_IN Mary_NP]]]，若剖析程式產生的剖析樹爲[[the_AT [daring_JJ ugly_JJ]] [dog_NN [barked_VBD [at_IN Mary_NP]]]] and [[the_AT [daring_JJ [ugly_JJ dog_NN]]] [barked_VBD [at_IN Mary_NP]]]。正確剖析樹與產生剖析樹的語法成分都是六個，但是產生剖析樹與正確剖析樹相同的語法成分有兩個，也就是[at_IN Mary_NP]與[barked_VBD at_IN Mary_NP]，則精確率與召回率皆爲2/6 = 0.33 (33%)。

定義六：Crossing。
Crossing是正確剖析樹與產生剖析樹之間不一致的語法成分的個數。

以前面的句子爲例，產生剖析樹的語法成分有 [at Mary], [barked at Mary], [dog barked at Mary], [daring ugly], [the daring ugly], 以及 [the daring ugly dog barked at Mary] 等六

個；而正確剖析樹的語法成分有 [at Mary], [barked at Mary], [ugly dog], [daring ugly dog], [the daring ugly dog], 以及 [the daring ugly dog barked at Mary]。其中不一致的語法成分只有[dog barked at Mary]，而[daring ugly]包含於[daring ugly dog]，並不算是不一致的語法成分。因此，這個產生剖析樹的Crossing是1。圖四比較清楚地描述Crossing的觀念。



圖四、語法成分的Crossing

表一、語料庫的統計資料

	段落數	句子數	字數
LOB Corpus	18,678	54,297	1,157,481
SUSANNE Corpus	1,967	6,920	150,053

一般而言，單獨計算Crossing的個數並不公平，因爲每一個句子的語法成分個數不一，產生的語法成分越多，則Crossing數目就越可能相對增多。比較好的方式是將語法成分的個數減去Crossing的個數後，再除以語法成分的個數，筆者稱之爲準確率（Accuracy）。依照這種計算方式，前述例子的準確率爲(6-1)/6=0.83=83%。事實上，這幾種評估準則都有各自的缺點，以前面的句子爲例，召回率與精確率爲33%，但是Crossing卻爲1，依此計算的準確率爲83%，同樣的剖析樹應用不同的準則，然而結果高低相差太多，這表示這兩種方式並不是非常適合作爲這一類研究的評估準則。筆者正在

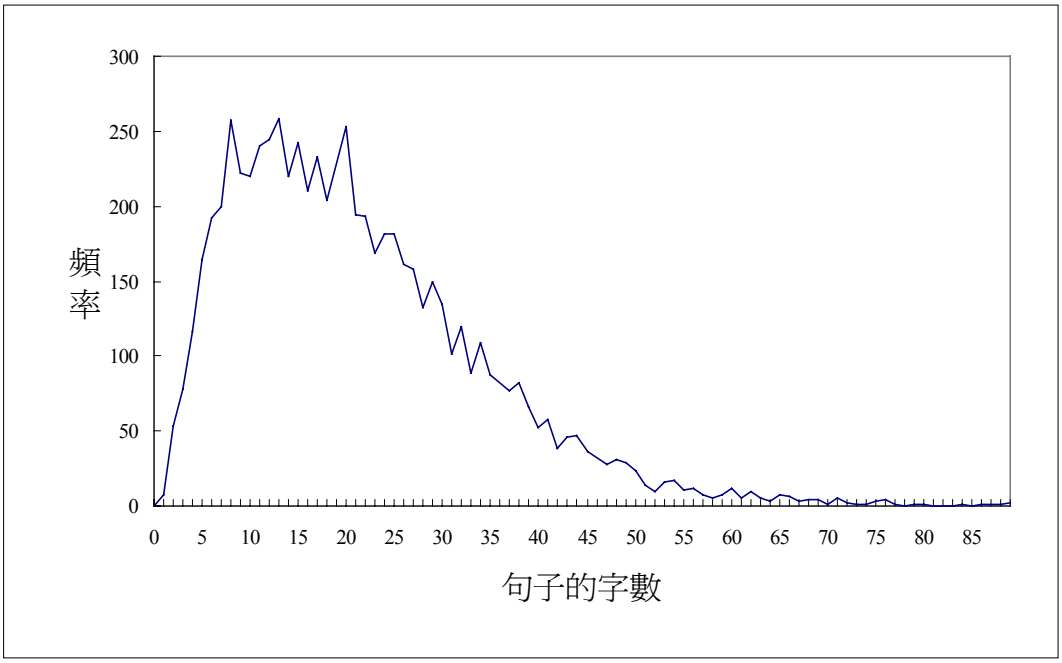
構思其他的評估準則，將另文討論有關的想法。

五、實驗的素材與結果分析

筆者使用LOB Corpus語料庫作爲訓練語料庫，而測試語料庫爲SUSANNE Corpus（註16），這兩個語料庫詳細的統計資料如表一所示。在前面已經介紹過LOB Corpus，這裡不再贅述。SUSANNE Corpus比較特別，SUSANNE代表的是Surface and Underlying Structural Analyses of Naturalistic English。其具有樹狀結構的訊息，也就是通稱的TreeBank。SUSANNE Corpus的建構者Sampson博士取材Brown Corpus的A類（新聞報導）、G類（書信、傳記、回憶錄）、J類（學術論文）、與N類（歷險小說、西部小說）等不同類型的文章，加注剖析樹的資訊。正因爲SUSANNE Corpus是樹狀語料庫，可以作爲正確的答案和實驗的結果比對，所以筆者選擇該語料庫作爲測試語料庫。由於訓練語料庫與測試語料庫並不相同，這類型的實驗稱爲開放式測試（Open Test）。圖五是SUSANNE Corpus的部份語料，共分六欄分別代表索引標號、狀態、詞類標記、詞彙、詞彙原形、與剖析樹。（註17）

索引標號	狀態	詞類標記	詞彙	詞彙原形	剖析樹
A01:0010a	-	YB	<minbrk>	-	[Oh.Oh]
A01:0010b	-	AT	The	the	[O[S[Nns:s.
A01:0010c	-	NP1s	Fulton	Fulton	[Nns.
A01:0010d	-	NNL1cb	County	county	.Nns]
A01:0010e	-	JJ	Grand	grand	.
A01:0010f	-	NN1c	Jury	jury	.Nns:s]
A01:0010g	-	VVDv	said	say	[Vd.Vd]
A01:0010h	-	NPD1	Friday	Friday	[Nns:t.Nns:t]
A01:0010i	-	AT1	an	an	[Fn:o[Ns:s.
A01:0010j	-	NN1n	investigation	investigation	.
A01:0020a	-	IO	of	of	[Po.
A01:0020b	-	NP1t	Atlanta	Atlanta	[Ns[G[Nns.Nns]
A01:0020c	-	GG	+<apos>s	-	.G]
A01:0020d	-	JJ	recent	recent	.
A01:0020e	-	JJ	primary	primary	.
A01:0020f	-	NN1n	election	election	.Ns]Po]Ns:s]
A01:0020g	-	VVDv	produced	produce	[Vd.Vd]
A01:0020h	-	YIL	<ldquo>	-	.
A01:0020i	-	ATn	+no	no	[Ns:o.
A01:0020j	-	NN1u	evidence	evidence	.
A01:0020k	-	YIR	+<rdquo>	-	.
A01:0020m	-	CST	that	that	[Fn.
A01:0030a	-	DDy	any	any	[Np:s.
A01:0030b	-	NN2	irregularities	irregularity	.Np:s]
A01:0030c	-	VVDv	took	take	[Vd.Vd]
A01:0030d	-	NNL1c	place	place	[Ns:o.Ns:o]Fn]Ns:o]Fn:o]S]
A01:0030e	-	YF	+.	-	.O]

圖五、SUSANNE Corpus的部份語料



圖六、測試語料庫句長分佈

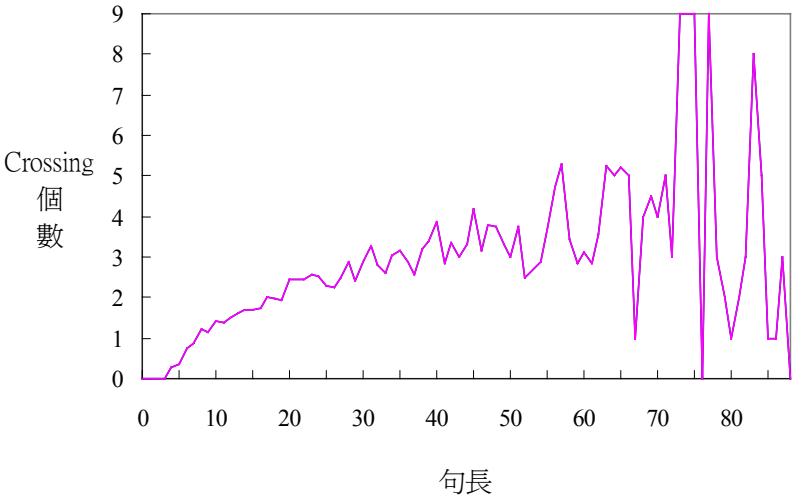
整個實驗程序如下所示：

- 抽取SUSANNE Corpus的每一個句子作爲測試句；相對應的剖析樹作爲標準答案。
- 根據右結合律將標準多元剖析樹轉換爲二元剖析樹。
- 使用本文提出的剖析演算法剖析每一個測試句。
- 使用第四節提出的評估準則度量產生的剖析樹。

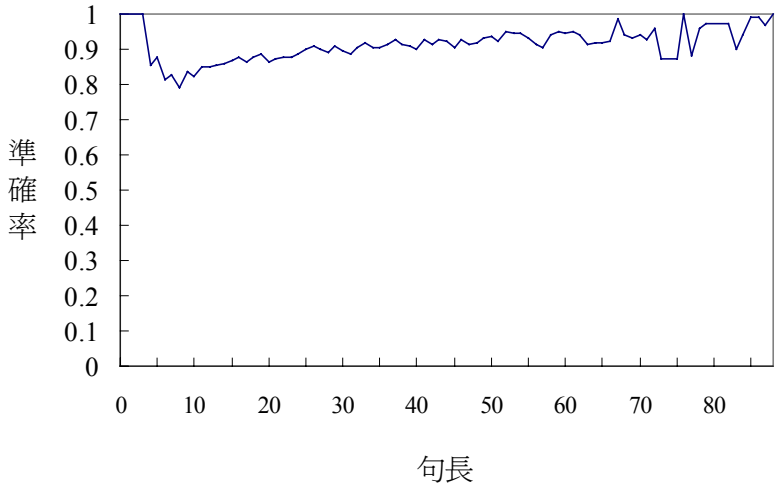
這裡另外要注意的是，SUSANNE Corpus的詞類標記與LOB Corpus的詞類標記並不相同，還必須將SUSANNE Corpus使用的詞類標記轉換爲LOB Corpus的詞類標記。由於SUSANNE Corpus的詞類標記原本就是根據LOB Corpus的詞類標記修改的，所以這樣的轉換過程並沒有很大的困難。

首先檢視測試語料的句長分佈情形，圖六顯示大部份的句子長度爲4個字到20個字，但是最短爲2個字，最長有109個字。句子的型態變化相當懸殊，句長變化也很大。因此使用傳統的剖析程式，無法快速地剖析句子，不適用資訊檢索系統查詢的應用。

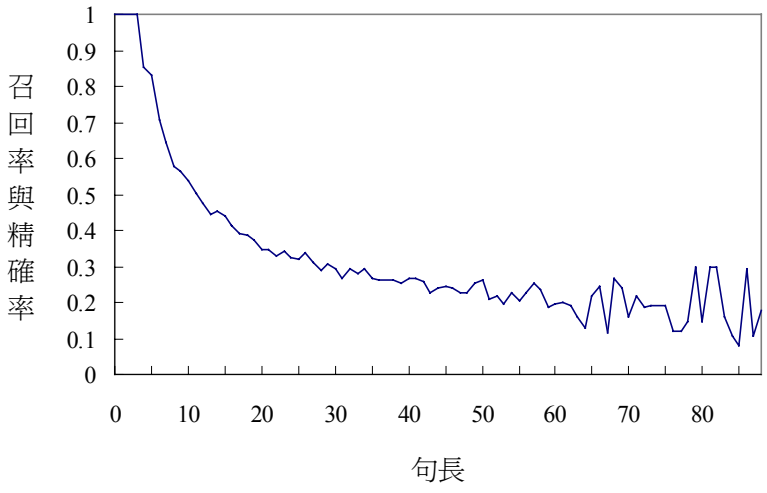
實驗的結果展示於圖七、圖八、與圖九。圖七表示的是剖析程式平均的Crossing個數；圖八是平均的準確率；圖九則是召回率與精確率。要注意的是，因爲本文提出的剖析程式產生的剖析樹是二元剖析樹，因此在定義五中的精確率與召回率的分母是相同的，所以召回率與精確率是完全一樣。比較圖七與圖九，可再一次看到PARSEVAL



圖七、測試語料平均的Crossing個數



圖八、測試語料平均的準確率



圖九、測試語料平均的召回率與精確率

與Crossing不同的表現行爲，這也回應筆者認爲有必要發展新的評估準則的看法。

由於4個字到20個字的句子佔了絕大多數，筆者特別將這個句長範圍內的句子，進一步地分析剖析程式的效能。下面將以三種不同的句長範圍，分別是4-40、4-25、以及10-20個字，觀察其平均表現行爲，結果列於表二。剖析程式在4-40句長的範圍內有相當一致的表現，準確率平均爲81%，召回率與精確率平均爲33%。這說明本文提出的剖析技術剖值得信賴。另外還有一個重要的因素，也就是剖析時間，必須特別說明。實驗使用的電腦設備是CPU爲 SPARC-5、記憶體32MB 的 Sun 相容工作站，剖析整個SUSANNE Corpus總共用了3477.18秒，平均每個句子使用0.50秒，平均每個字使用0.02秒。這樣的處理時間基本上是符合資訊檢索系統的需求。

表二、部份實驗結果

句長	4 - 40	4 - 25	10 - 20
句子數目	5905	4373	2400
平均句長	19.09	14.71	15.02
平均Crossing數目	2.655	2.327	2.482
準確率	0.821	0.801	0.804
召回率（精確率）	0.319	0.355	0.316

六、可能的後續應用

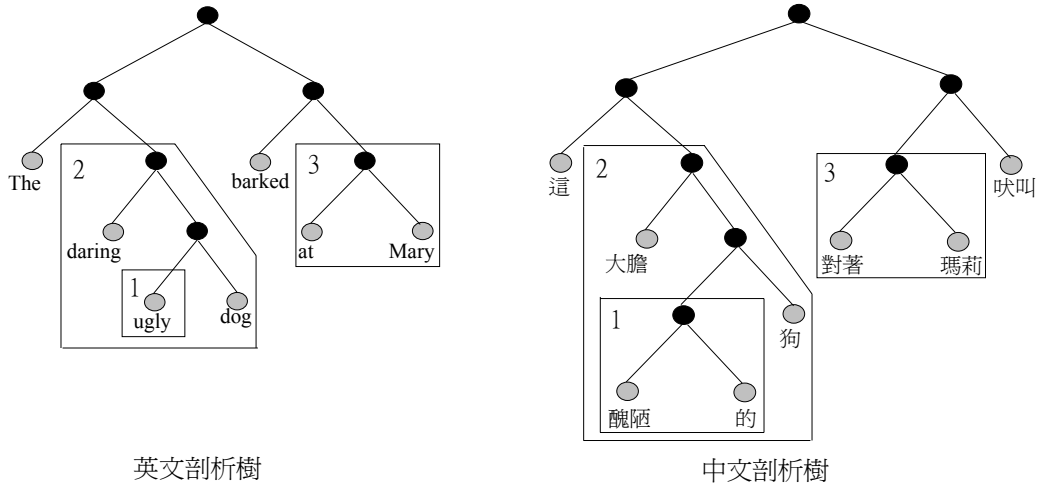
本文提出的剖析技術，除了可以用於資訊檢索系統之外，一個重要的應用是作爲建構樹狀語料庫的輔助工具。筆者曾經提及Sampson博士建構SUSANNE Corpus用了相當長的時間以及許多的人力，才完成一份樹狀語料庫，這樣的問題同樣也發生於其他語言樹狀語料庫的建構過程。若是有一種剖析程式能夠將標記過的語料做初步的剖析，然後再交由語言專家做進一步的修正，可以有效地降低所需的人力與時間。

資訊檢索中有關索引項目的建立，有所謂的單詞索引（Single Term Index）以及片語索引（Phrase Term Index）。自動建構片語索引有賴對文件進一步的分析，本文的剖析技術可提供這方面的協助。甚至於雙語索引詞彙的自動建構，也可以使用這樣的剖析技術，筆者以下面的例子做個說明：

The daring ugly dog barked at Mary.

這 大膽 醜陋 的 狗 對著 瑪莉 吠叫。

圖十顯示片語之間的對應關係。若使用剖析程式分別處理中文與英文句子，然後使用片語層次的對列技術（Alignment Technology）（註18）適當地處理雙語剖析樹，最後，抽取對應的片語結構，建立一部雙語術語對譯辭典。



英文剖析樹　　中文剖析樹

圖十、中英文對應的剖析樹結構

七、結語與討論

資訊檢索系統的研究發展已經有很長的一段時間，各種的檢索模型與回饋機制也不斷推陳出新。但是，對於使用者而言，查詢的方式仍然是以關鍵字爲基礎的查詢（無論是受控制的抑或不受控制的）。目前風起雲湧的WWW風潮，將資訊檢索的需求推到前所未有的高峰，如果仔細探究目前WWW上提供的檢索系統（如Alta Vista、Lycos、Excite等等系統），在檢索技術上並沒有超越以往獨立的資訊檢索系統，只不過是換了一副誘人的面貌，增添了更多種媒體資訊，變成了開放的系統，並且加入了許多商業氣息罷了。

展望未來，自然語言仍然是人類最便於使用的溝通工具。因此利用自然語言作爲資訊檢索系統的查詢方式，是無法避免的，尤其在語音輸入的技術達到實用的階段，這樣的需求將更形殷切。如何提供比目前樣式比對更好的技術，進一步地分析自然語言，無疑是一項重要的研究課題。本文企圖在這個研究方向作一個初步的嘗試。筆者提出的自然語言剖析技術，在自然語言理解的研究領域裡，只能算是最初步的分析；然而，對於資訊檢索的應用，這卻又算是比較高階的處理方式。因爲，能夠比較準確地分析重要的詞彙，提供詞彙之間的關係（主語、謂語或是賓語）。重要的是，建構本文提出的剖析

程式，需要的資源並不多，只要擁有詞類標記的語料庫。而詞類標記語料庫是目前除了原始語料庫之外，最容易取得的語料庫。

本文使用LOB Corpus訓練建構剖析程式所需的統計參數，同時進行大規模的實驗，總共6,920句長短不一的語料（SUSANNE Corpus）用於實驗。結果顯示，這樣的剖析技術具有高度的穩定性（參見表二），同時所需的處理時間不多，相當適用於資訊檢索系統的應用。筆者另外也討論到這樣的剖析技術其他有關的應用，其中最值得注意的是，可作爲標記語言庫轉換爲樹狀語料庫的前端處理系統（Frontend System），協助建構語料的樹狀結構，縮短發展樹狀語料庫所需的時間，並且減少所需的人力。這是一種良性循環的過程：當擁有更豐富的語料資源，可以發展更好的自然語言處理技術；新發展的技術又可以協助建構更好的語料庫。在這樣的循環過程中，逐漸逼近學術研究發展的目標。

致謝

筆者感謝英國Sampson博士熱心提供SUSANNE Corpus作爲學術研究之用。筆者還要感謝台灣大學資訊工程學研究所陳信希教授提供電腦設備以及語料供筆者進行實驗。感謝台灣大學圖書館學系陳雪華教授與林珊如教授的寶貴意見。另外筆者還要謝謝台灣大學資訊工程學研究所兩位博士候選人，李御璽與邊國維。

【附註】

註1： 孔安國，尚書孔傳（台北市：新興書局，民國66年），頁4。

註2： 大眾書局編輯部，名家語錄（高雄市：大眾書局，民國61年），頁158。

註3： 國立編譯館主編，圖書館學與資訊科學大辭典（台北市：漢美，民國84年），頁1515。

註4： 基本上，語料庫是日常生活使用的語言，包括報紙、小說、文學作品等等的集合。至於要蒐集那類型的語言資料，端賴語料庫建構者的籌畫與爾後的使用方式。由收集的語言種類，語料庫可分爲單語語料庫、雙語語料庫、多語語料庫；以收集的方式，可分爲平衡式語料庫與非平衡式語料庫，也就是考量語料型態分佈的情形；由整理的方式，可分爲原始語料庫、標記語料庫、與樹狀語料庫。

註5： Hindle, D., "User Manual for Fidditch, A Deterministic Parser," Naval Research Laboratory Technical Memorandum 7590-142, Naval Research Laboratory, Washington, D.C., 1983.

Chen, K.H. and Chen, H.H., "A Rule-Based and MT-Oriented Approach to Prepositional Phrases Attachment," Proceedings of the 16th International Conference for Computational Linguistics (COLING-96), 1996, pp. 166-171.

註6： Jordan, P.W., “Using Terminological Knowledge Representation Languages to Manage Linguistic Resources,” cmp-lg/9605024 (URL：http://xxx.lanl.gov/), 1996.

Agirre, E; Arregi, X; Artola, X; Diaz de Ilarraza, A. and Sarasola, K., “Lexical Knowledge Representation in an Intelligent Dictionary Help System,” Proceedings of the 15th International Conference for Computational Linguistics (COLING-94), 1994, pp. 544-550.

Light, M. and Schubert, L., “Knowledge Representation for Lexical Semantics: Is Standard First Order Logic Enough?” cmp-lg/9412004 (URL：http://xxx.lanl.gov/), 1994.

註7： Chen, K.H. and Chen, H.H., “Extracting Noun Phrases for Large-Scale Text Corpora: A Hybrid Approach and Its Automatic Evaluation,” Proceedings of the 34th Annual Meeting of Association for Computational Linguistics (ACL-94), 1994, pp. 234-241.

Brill, E., "A Simple Rule-Based Part of Speech Tagger," Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, 1992, pp. 152-155.

Cutting, D.; Kupiec, J.; Pedersen, J and Sibun, P., "A Practical Part-of-Speech Tagger," Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, 1992, pp. 133-140.

Church, K., "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas, 1988, pp. 136-143.

註8： Garside, R. and Leech, F., “A Probabilistic Parser,” Proceedings of Second Conference of the European Chapter of the ACL, 1985, pp. 166-170.

Brill, E. and Marcus, M., "Automatically Acquiring Phrase Structure Using Distributional Analysis," Proceedings of the DARPA Conference on Speech and Natural Language, 1992, pp. 155-159.

Chen, H.H. and Lee, Y.S., “Development of a Partially Bracketed Corpus with Part-of-Speech Information Only,” Proceedings of the Third Workshop on Very Large Corpora, 1995, pp. 162-172.

註9： 這個例子取材於SUSANNE Corpus的第一個句子（標題）。本文使用的測試語料庫是以美國英語爲體裁的SUSANNE Corpus，所以筆者提及的例子也都是由SUSANNE Corpus摘錄的。我國中央研究院也已經推出漢語平衡語料庫，筆者也計畫於不久的將來使用這份語料庫從事相關的研究。

註10： LOB Corpus全名是Lancaster-Oslo/Bergen Corpus，是由建構該語料庫的三所大學校名定名的。有關該語料庫詳細的說明可以參考LOB Corpus使用手冊：

Johansson, S., The Tagged LOB Corpus: Users' Manual, Bergen: Norwegian Computing Centre for the Humanities, 1986.

註11： stop_VB中的VB是stop的詞類標記，也就是原形動詞；而VBG表示動名詞；NN表示單數普通名詞；NNS表示複數普通名詞。

註12： Shannon, C.E. and Weaver, W., The Mathematical Theory of Communication, University of Illinois Press, Urbana, IL, 1949.

註13： Francis, N. and Kucera, H., Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for Use with Digital Computers, Department of Linguistics, Brown University, Providence, R. I., U.S.A., original ed. 1964, revised 1971, revised and augmented 1979.

註14： 詞庫小組，中央研究院漢語平衡語料庫的內容與說明，技術報告95-02，（台北市：中央研究院資訊科學研究所，民國84年）。

註15： Black, E. et al., "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars," Proceedings of the Workshop on Speech and Natural Language, 1991, pp. 306-311.

註16： Sampson, G., English for the Computer, Oxford University Press, 1995.

註17： 圖五表示一個句子的有關資訊，而第六欄表示該詞彙在剖析樹中相對的位置（以「.」符號表示詞彙的位置）。因此，必須將第六欄整個串起來，才能夠瞭解剖析樹的結構。

註18： 在進行句子剖析之前，必須先處理對應中英文句子的排列問題，而片語層次的對列問題，也是重要的研究課題。這方面有關的討論，請參見：

Chen, K.H. and Chen, H.H., "Aligning Bilingual Corpus: Especially for Language Pairs from Different Families," Information Science: An International Journal, 4(1995): 57-81.

Matsumoto, Y.; Ishimoto, H. and Utsuro, T., “Structural Matching of Parallel Texts," Proceedings of the 31st Annual Meeting of ACL, Ohio, USA, June 22-June 26, 1993, pp. 23-30.