結合統計與規則的多層次中文斷詞系統

陳鍾誠、許聞廉

中央研究院資訊科學研究所

E-mail : johnson@iis.sinica.edu.tw

FAX: 886-2-27824814

摘要

本論文設計了一套結合 PAT-tree 的統計資訊與規則比對以進行多層次斷詞的方法，用以解決一般斷詞系統中未知詞不容易被斷出的問題，並提出一組以召回率(recall)和衝突率(conflict)為基準的多層次斷詞評估方法，用來評估本系統的斷詞正確率。召回率定義為標準斷詞集合中被系統斷出的百分比，衝突率則是系統斷詞與標準斷詞交叉重疊的比率。本系統之實驗以中央研究院平衡語料庫為標準斷詞語料，該語料庫共有 455 萬詞，我們取其中 265 萬詞為訓練語料，剩下的 190 萬詞為測試語料。實驗結果在訓練語料上的詞彙召回率為 96.9%、衝突率為 0.50%在測試語料上的詞彙召回率為 96.7%、衝突率為 0.50%。本實驗說明了經由 PAT-tree 與規則比對兩者混合使用，可使召回率有相當程度的提升，這證明了在未知詞的處理上，PAT-tree 與規則比對有互補的效果。

1. 簡介

斷詞是中文語言處理的最基礎工作，由於中文的詞彙之間並沒有明顯的斷詞符號，因而有許多研究提出不同的方法處理斷詞問題，並取得了不錯的實驗成果。然而對於未知詞的處理，一直仍是斷詞系統的一大困難。人名、地名、術語、專有名詞、簡稱、文言等詞彙經常無法盡收在辭典當中，這是目前斷詞系統的主要錯誤原因。

目前未知詞的處理方法大致上可分為兩類，一類是使用構詞律以辨認未知詞的方

法[Lin1993]，另一類則是使用詞雙連(bigram)的統計以辨認未知詞的方法[Chen1997]。構詞方法在具有明確詞首或詞尾的詞彙上表現較好，而詞雙連統計方法在強健性(robustness)上的表現的較好，但是對於不具明顯詞首或詞尾的較長詞彙而言，則這兩種方法都難以正確辨認。

針對上述缺點，本論文引進了一種稱為 PAT-tree 的資料結構，用以辨識較長的未知詞，並進行斷詞。PAT-tree 是一種統計字串之出現次數的資料結構 [Gonnet 1992]，不論一字串有多長，PAT-tree 都能記下該字串的出現次數，其方法是將文件視為一個無限長的字串，並對每一字串的出現位置建立索引。實驗證明使用 PAT-tree 來抽取一文件中的關鍵詞，對於出現次數高的詞彙而言，可以得到相當不錯的結果[Chien 1997]，因此、我們合併 PAT-tree 與規則比對的方式來解決未知詞的斷詞問題。

本論文第二節描述一個將 PAT-tree 與規則比對結合的多層次斷詞法，第三節提出多層次斷詞的正確率評估方法，第四節描述實驗的結果，最後對本系統斷詞的某些正確與錯誤案例進行分析與檢討。

2. 結合統計、PAT-tree 與規則的多層次斷詞方法

本系統使用一個二維矩陣來記錄一個句子中任何子字串的斷詞機率，矩陣座標(i,j)的格子點上所記錄的是從第 i 個到第 j 個字的字串之斷詞機率。程式首先利用斷詞語料庫的統計次數來設定初始機率，並用動態規劃決定第一層的最佳斷詞，接著幾層都使用規則比對與 PAT-tree 的統計次數來調整機率值，並計算每一層的最佳斷詞機率與斷詞點，這樣的一種多層次的斷詞方法，對文法結構比較鬆散的中文來說，是一種不完全的剖析(parsing)。

機率式斷詞系統在計算最佳斷詞時，必須使用某些假設，即是所謂的機率模型。機率式斷詞有許多不同的機率模型，本系統採用的是一種完全獨立式的機率模型[張俊盛 91]，其模型如下：

$$\max \quad P(W_1, W_2, .., W_k \mid C_1, C_2, .., C_n)$$
$$\approx \max \quad P(W_1) * P(W_2) * .. * P(W_k)$$
$$= \max \quad \prod_{i=1}^{k} P(W_i)$$

上列公式中的 $P(W_i)$ 表示第 i 個詞彙的斷詞機率。

由於辭典所收錄的詞彙不一定符合語料庫的斷詞原則，因此、本系統首先統計辭典中各個詞彙在訓練語料庫中各種斷法的次數以作為斷詞機率的初始值。我們以 $T(W_i)$ 表示在訓練語料庫中 W_i 一詞的出現次數總和， $T(W_{i,c})$ 表示在訓練語料庫中 W_i 一詞被斷成 C 斷法的次數。以這些統計次數為基礎，配合上 PAT-tree 之建立以及規則的比對，本系統即可進行多層次的斷詞。目前本系統的斷詞層次共可分為語料統計層、詞彙層與短語層等三層，以下我們使用 "請電洽書畫組幹事張玉坤" 為例句，以便對各層次作進一步的說明。

第 1 層 ：語料統計層

首先設定 $P(|W_i|)$ 機率初始值如下：

$P(|W_i|) = T(|W_i|) / T(W_i)$　　(說明：$T(|W_i|)$ 表示 W_i 一詞左右都被斷開的次數)

接著即可利用動態規劃計算最佳斷詞機率，例句的最佳斷詞結果如下：

max(P(請電洽書畫組幹事張玉坤))

= P(|請|)*P(|電|)*P(|洽|)*P(|書畫|)*P(|組|)*P(|幹事|)*P(|張|)*P(|玉|)*P(|坤|)

因此、最佳的斷詞是 "請|電|洽|書畫|組|幹事|張|玉|坤"

第 2 層 ：詞彙層

步驟 1 ：利用詞彙層規則比對，以提升符合詞段的得分。

在例句 "請電洽書畫組幹事張玉坤"中，經規則比對後會發現 "張玉" 符合二字姓名規則 "fname[1..2]:u" 的條件，"張玉坤" 符合三字姓名規則 "fname[1..2]:u:u"的條件[1]。因此，在本步驟裏，例句中的 "張玉" 和 "張玉坤" 都會被加分 ，加分情形請參考圖一，加分的幅度由規則寫作人員指定，在此範例中、"張玉" 得 12 分，"張玉坤" 得 16 分。

[1]規則中的 fname 表示姓氏，[1..2]表示姓氏可為一字到二字，u 表示任意單字

步驟 2 ：利用 PAT-tree 的統計提升最小完整詞段的得分。

一個字串是否會是一個詞彙，可由其出現頻率與左右接字集的大小來判斷[Chien1997]。假設在文章中出現有兩個句子 --"程式設計改採物件導向方法" 與 "新一代物件導向資料庫"，則程式可利用 PAT-tree 計算所有詞段的左右接字集，例如：

"物件導向"　字串共出現兩次，其左接字集為{採、代}，右接字集為{方、資}

"物件導"　字串共出現兩次，其左接字集為{採、代}，右接字集為 {向}

"件導向"　字串共出現兩次，其左接字集為{物}　，右接字集為 {方、資}

"件導"　字串共出現兩次，其左接字集為{物}　，右接字集為 {向}

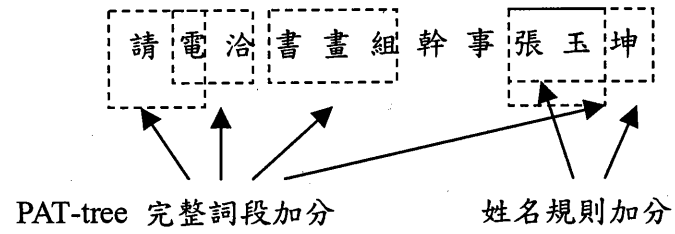
仔細觀察上述資料之後，可發現左右接字集均很大者比較有可能是一個詞彙。因此、我們稱一個左右接字集均大於 2 者為一個完整詞段。

令 $L(W_i)$ 代表 W_i 的左接字集，$R(W_i)$ 代表 W_i 的右接字集，$Score(W_i)$ 為 W_i 的得分，本系統所採用的完整詞段加分公式如下

$$Score(W_i) \leftarrow Score(W_i) + 10 + \log_2(|L(W_i)|) + \log_2(|R(W_i)|)$$

公式中的數字 10 為左右接字集均大於 2 的完整詞段之基本分數，$\log(|L(W_i)|)$ 為左接亂度，$\log(|R(W_i)|)$為右接亂度。

經由 PAT-tree 的查詢，可計算出例句中的 "請電","請電洽","電洽","書畫組","張玉坤" 等都是完整詞段，但是為了做漸進的多層次斷詞，本步驟只對相對較短的詞彙作分數的提升，所以、只有 "請電","電洽","書畫組","張玉坤" 的分數會被提升，圖一顯示了本階段的加分的情形。



圖一 ： 詞彙層的加分情況示意圖

步驟 3 ：利用各詞段的得分對機率矩陣進行機率提升。

在本步驟、我們利用詞段的得分對機率進行機率調整，其調整公式如下

$$P(W_i) \leftarrow Max\ (P(W_i), \frac{Score\ (W_i)}{100})$$

接著再次以進行最佳化、可得到詞彙層的最佳斷詞為"請|電洽|書畫組|幹事|張玉坤"

第 3 層 ：短語層

短語層的作法同詞彙層，不同的是使用的規則為短語規則，另外、PAT-tree 會對較長的詞段進行加分，例句在此層會被斷為 "請電洽|書畫組|幹事張玉坤"。

最後、將各層的斷詞合併，則形成多層次的斷詞 "(請|電洽)(書畫|組){幹事(張|玉|坤)}"

3. 正確率的評估方法 -- 「召回率」與「衝突率」

對於斷詞正確率的比較，目前並沒有一個客觀的比較基礎，雖然近來有些標記語料庫可以作為斷詞的比較基準(benchmark) [黃居仁 1995]，但是由於每個人所認為的正確斷詞都是不同的，因此在標準斷詞的認定上會產生許多爭議，一般來說、斷詞的爭議點主要在於詞彙斷點的層次上，而非斷詞點的位置，例如在 "中山南北路上的牛肉麵店共有三十五家" 這個句子中，"中山南北路" 可斷成 "中山|南|北|路"、"中山|南北|路" 或 "中山南北路"，但是、若使用多層次的斷詞方法將該句子斷成 "[中山(南|北)路]上|的[(牛肉|麵)店]共|有(三十五|家)" 則較不容易引起爭議。

因此、為了避免 "何謂一個完整詞彙？" 這個問題的主觀性影響評估的結果，我們定義了「召回率」(recall) 和「衝突率」(conflict) 這兩個多層次斷詞的衡量標準，為了定義召回率與衝突率，我們首先必須先作一些名詞與符號的定義如下：

詞段相互衝突 ：若兩個詞段 (i_1,j_1) 與 (i_2,j_2) 互相跨越 (亦即、$i_1 < i_2 < j_1 < j_2$)，則稱這兩個詞段相互衝突。

無歧義詞段集合 ：若詞段集合 F 中，任兩個詞段都不會互相衝突，則稱 F 為一個無歧義詞段集合。

詞段與詞段集合相互衝突 ：令 (i, j) 代表一個詞段，F 為一詞段集合，若 (i, j) 與 F

中的任一詞段相互衝突，則稱 (i, j) 與 F 相衝突。

標準詞段集合 S：S 為標準斷詞語料庫所提供的多層次且無歧義之詞段集合。

系統詞段集合 S'：S' 為自動斷詞系統所建構的多層次且無歧義之詞段集合。

共同詞段集合 S∩S'： S 與 S' 之交集。

衝突詞段集合 conflict(S,S')： S 中所有與 S' 相衝突的詞段所成的集合。

詞段集合大小 |S|：|S| 為詞段集合 S 的總詞段數。

詞段集合字元總數 length(S)：length(S) 為詞段集合 S 的總字元數

接著定義多層次斷詞的「召回率」與「衝突率」如下：

詞彙召回率 $(S',S)=\frac{|S\cap S'|}{|S|}$ 詞彙衝突率 $(S',S)=\frac{|\text{conflict}\ (S,S')|}{|S|}$

字元召回率 $(S',S)=\frac{length\ (S\cap S')}{length\ (S)}$ 字元衝突率 $(S',S)=\frac{length\ (\text{conflict}\ (S,S'))}{length\ (S)}$

召回率與衝突率兩者之和必然小於或等於 1，但不一定等於 1，下列範例說明了召回率與衝突率的計算方式。

範例：「召回率」與「衝突率」的計算方法

以 "土地公有政策" 為例句，假設：

標準斷詞為： {[土地(公有)]政策} 系統斷詞為： (土地|公)有|政策

則標準詞段集合、系統詞段集合、共同詞段集合與衝突詞段集合如下：

S = {土地, 土地公有, 土地公有政策, 公有, 政策}, |S| = 5, length(S) = 16

S' = {土地, 土地公, 公, 有, 政策}, |S'| = 5, length(S') = 9

S∩S' = {土地, 政策}, |S∩S'| = 2, length(S∩S') = 4

conflict(S,S') = {公有, 土地公有}, |conflict(S,S')| = 2, length(conflict(S,S')) = 6

如此可計算出詞彙召回率與衝突率如下：

詞彙召回率 = 2/5 = 40% 詞彙衝突率 = 2/5 = 40%

字元召回率 = 4/16 = 25% 字元衝突率 = 6/16 = 37.5%

由上述範例可看出，一個斷點錯誤可能會造成數個詞彙未被召回，並造成數個衝突，例如 "土地公" 這個錯誤的斷詞造成了 "公有" 與 "土地公有" 兩個詞彙未被斷出，也

造成了與 "公有" 一詞的衝突，因此是一個嚴重的錯誤，如此、本方法雖然沒有所謂的『權值(weight)』概念，但實際上卻能衡量出錯誤的嚴重程度。

根據召回率，我們可以計算出所有原先未被斷出的詞彙中，究竟有多少比例可被輔助方法 A 斷出來，這個比例稱之為輔助方法 A 的未知詞召回率，其計算方法如下：

$$\text{輔助方法}A\text{的未知詞召回率} = \frac{\text{使用方法}A\text{之後的召回率} - \text{使用方法}A\text{之前的召回率}}{1.0 - \text{使用方法}A\text{之前的召回率}}$$

下節中所述的實驗即是以召回率和衝突率作為斷詞評估的標準。

4. 實驗說明與結果

本實驗使用中央研究院資訊科學所詞庫小組的平衡語料庫作為斷詞語料庫[黃居仁95]，該語料庫共有 455 萬詞，我們將此語料庫依據檔名的開頭字母，均分成大小約略相等的兩個等分，前半部(共 265 萬詞) 作為訓練語料，後半部(共 190 萬詞) 作為測試語料，同時、本實驗也採用該語料庫所提供的辭典，共有 78410 個詞彙。由於該辭典的詞彙標記無法符合本系統進行規則比對時的需求，因此、我們採用中央研究院資訊科學所語言系統實驗室的詞類標示，以進行規則比對。

為了比較我們方法的好壞，我們另外使用了長詞優先的方法作為對照，此方法是以查辭典的方式，總是斷出最長的詞彙，這是一個很簡單且純粹使用辭典的斷詞方法，主要目的是用來對照出本系統改進的程度。

為了瞭解規則比對與 PAT-tree 的個別功效，本實驗首先測試只使用第一層的『單純語料統計』方法之功效，接著測試以 PAT-tree 加分後的『多層次 PAT-tree』方法之功效，然後移除 PAT-tree ，改用『多層次規則比對』進行測試，最後才測試完整的規則與 PAT-tree 『多層次混合使用』的功效。

本實驗所有的方法都會事先將連續的數字及英文字母斷成單一詞，以避免因為這個問題所造成的立足點不平等現象，影響實驗的客觀性。另外、本實驗所使用的規則共有 37 條，主要用於處理人名、地名、組織名等未知詞，數量相當少，因此在規則比對上仍有許多的改進空間。

表一列出了訓練語料庫與測試語料庫的大小，表二列出了本系統在訓練語料上的召回率與衝突率，表三列出了本系統在測試語料上的召回率與衝突率。

	總詞彙數	總字元數
訓練語料庫	2642601	3912242
測試語料庫	1871566	2798064

表一 ： 訓練與測試語料庫的大小

	詞彙召回率	字元召回率	詞彙衝突率	字元衝突率	未知詞召回率
長詞優先（對照組）	93.61%	88.57%	0.56%	0.84%	---
單純語料統計	94.28%	89.52%	0.30%	0.46%	基準
多層次 **PAT-tree**	96.28%	92.80%	0.54%	0.91%	34.5%
多層次規則比對	95.58%	92.01%	0.43%	0.72%	22.7%
多層次混合使用	96.89%	94.14%	0.50%	0.84%	45.5%

表二 ： 訓練語料的斷詞召回率與衝突率

	詞彙召回率	字元召回率	詞彙衝突率	字元衝突率	未知詞召回率
長詞優先（對照組）	93.54%	88.42%	0.57%	0.84%	---
單純語料統計	94.19%	89.32%	0.32%	0.49%	基準
多層次 **PAT-tree**	96.04%	92.37%	0.51%	0.85%	31.8%
多層次規則比對	95.41%	91.64%	0.45%	0.73%	20.9%
多層次混合使用	96.65%	93.70%	0.50%	0.83%	42.4%

表三 ： 測試語料的斷詞召回率與衝突率

5. 斷詞結果分析

為了瞭解對本系統的優缺點，我們從測試結果中摘選出一些正確與錯誤的案例，以便進行分析，以下所有案例中、前面的句子為語料庫所提供的斷詞，後面的句子為本系統所斷出來的多層次斷詞，案例中以框線標示出斷詞錯誤的部分。

(1) 在|尋找|頂夸克|的|實驗|中|所|獲致|的|數據

在|尋找[頂(夸|克)]的|實驗|中|所|獲致|的|數據

分析：「夸克」，「頂夸克」 都是辭典未收錄的詞，但因在語料庫中出現許多次，

因此被系統以 PAT-tree 加分將之斷開了。

(2) 自|台北市|體育場|經|敦化北路|再|轉進|民權東路

自|台北市|體育場|經(敦|化|北|路)再|轉進(民權|東|路)

分析：「敦化北路」，「民權東路」 都是辭典未收錄的詞，卻因規則比對成功而正確的斷開了。

(3) 經|瑞典|皇家|科學院|評定|為|一九九四年|克列佛|獎|得主

經|瑞典|皇家|科學院|評定|為(一九九四|年)克|列|佛|獎|得|主

(4) 其|前身|為|臺灣|總督府|殖產局|草湳坡|製|茶|試驗場

其|前身|為|臺灣|總督|府|殖產|局|草|湳|坡|製|茶|試驗場

(5) 因此|稱為|「|楊密爾思|規範場論|」

因此|稱為|「(楊|密|爾)思|規範場論|」

分析：本例為人名辨識錯誤的情形，系統誤認「楊密爾」是一個三字中文人名，因而將「楊密爾」斷開。

(6) 教育部|首次|設置|母語|研究|著作|獎補助|辦法

(教育|部)首次|設置(母|語)研究|著作|獎(補助|辦法)

分析：本例的錯誤是由 PAT-tree 所引起的，因為在該檔案中「獎補助」 只出現一次，然而「補助辦法」的出現次數多且左右字集大，使得 PAT-tree 統計認為「補助辦法」應該形成一個詞。

案例 (1)~(2) 是正確斷詞的案例，(3)~(4) 是詞彙未被斷出的案例，(5)~(6) 是搶詞所造成的錯誤，由以上分析可歸納出下列四種主要的錯誤類型- 1.出現次數少的複合動詞與複合名詞常無法正確辨認 2.規則誤認某個段落所造成的錯誤 3. PAT-tree 誤認某個段落所造成的錯誤 4.辭典收錄詞、PAT-tree 建構詞與規則建構詞之間的搶詞所造成的錯誤。

6. 結語

我們建議以多層次斷詞的評量方法取代單層次斷詞的評量方法，並以 「衝突率」與「召

回率」作為多層次斷詞的評估依據，如此、可以降低 "何謂正確的斷詞？" 這個問題的爭議性，以使各個斷詞系統之間能有比較的基礎，並使中文自然語言在高層處理上能夠比較不受單層次斷詞系統的限制。

為了克服未知詞不容易斷詞的問題，我們引進了 PAT-tree 的結構，並與規則導向的方法配合，使許多未知詞能被比較有效的斷開，本實驗證明在多層次斷詞的問題上，藉由付出少許衝突率為代價，可以使得召回率有相當程度的提升。

對於只在文章中出現一次的未知詞而言，PAT-tree 是無能為力的，因此、這些未知詞常無法被本系統所斷出，然而、利用詞首字與詞尾字，常可解決掉這方面的問題，因此、進一步加入詞首與詞尾，可望能補強本系統這方面的弱點。另外、如何解決搶詞與誤認的問題，則是尚待進一步研究的課題。

參考文獻

張俊盛、陳志達、陳順德, "限制式滿足及機率最佳化的中文斷詞系統", 中華民國八十年第四屆計算語言學研討會論文集, 1991, 147-165 頁.

黃居仁、陳克健、張莉萍、許蕙麗, "中央研究院平衡語料庫簡介", 中華民國八十四年第八屆計算語言學研討會論文集, 1995, 81-99 頁.

Chien, Lee-Feng "PAT Tree-Based Keyword Extraction for Chinese Information Retrieval," The ACM SIGIR Conference, Philadelphia, USA, 1997 , pp. 50-58.

Gonnet, Gaston H., Richardo A. Baeza-yates and Tim Snider, "New Indices for Text : PAT-trees and Pat Arrays," Informational Retrieval Data Structure & Algorithm, Prentice Hall, 1992, pp. 66-82.

Lin, Ming-Yu, Tung-Hui Chiang and Keh-Yih Su "A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," Proceeding of ROCLING VI, 1993, pp.119-137.

Chen, Keh-Jiann and Ming-Hong Bai "Unknown Word Detection for Chinese by a Corpus-based Learning Method," Proceeding of ROCLING X, 1997, pp. 159-171