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：

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

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}

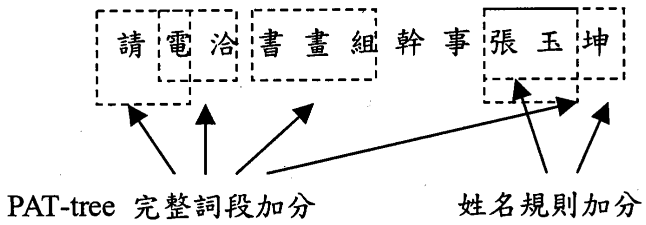
仔細觀察上述資料之後，可發現左右接字集均很大者比較有可能是一個詞彙。因此、我們稱一個左右接字集均大於 2 者為一個完整詞段。

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

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

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 $(S',S) = \dfrac{|S \cap S'|}{|S|}$　　詞彙衝突率 $(S',S) = \dfrac{|\text{conflict}\ (S,S')|}{|S|}$

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

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

The Design of Sem-Syn Initial Grammar
In Chinese Grammatical Inference

Hsue-Hueh Shih

Department of Foreign Languages and Literature
National Sun Yat-sen University, Taiwan, R.O.C.
E-mail address: hsuehueh@mail.nsysu.edu.tw

Abstract

This paper describes our on-going project on grammatical inference for Chinese. We here emphasize on the design of our sem-syn initial grammar that is a set of stochastic context-free rules and whose probabilistic parameters will be iteratively re-estimated in a corpus-based inference technique. Manually developing and maintaining a grammar for a NLP system has long been regarded as a painful and endless job. Besides, this conventional approach usually results in a grammar with limited coverage. With large bodies of text corpora available on computers, corpus-based grammatical inference (GI) techniques seem to provide a promising solution to the problems. An initial grammar is one of the important components in GI techniques and its function is to facilitate the inference process to proceed. In this paper, we describe the design of our sem-syn initial grammar and how it corresponds to the information given in Sinica Corpus on which our inference system is based. We also give a brief introduction to our Chinese grammatical inference system, showing how the system will use the sem-syn initial grammar to generalize structure from the Corpus.

1. Introduction

Grammar plays an indispensable role in most NLP systems. Conventional handcrafted approach to grammar construction and maintenance has been regarded as painful and endless work. Besides, this conventional approach usually results in a grammar with limited coverage. With the increasing availability of large text corpora in machine-readable form, Grammatical Inference (GI) [Pereira and Schabes, 1992] has surged to provide a promising solution to the problems. In GI, there are two components which are essential to the inference process, namely inference algorithms and initial grammars. An inference algorithm takes the responsibility of learning grammatical knowledge from a set of language samples; whereas an initial grammar is regarded as seed knowledge needed for the inference process to proceed. A well-designed initial grammar can assist the inference algorithm to produce a generalized grammar.

Initial grammars utilized in GI may be classified into four types. First, a null grammar, or an empty grammar, learns all its rules gradually from a set of training sentences in the course of the inference process. Using this approach, inference systems usually start to learn simple grammatical structures from short sentences. This is done by sorting the training data by length and presenting them to the inference systems in an ascending length order [Carroll and Charniak, 1992]. The main disadvantage of this approach is that it tends to acquire a large number of specific rules rather than a set of general ones. This lack of generality leads to an uncontrollable grammar size.

Second, a seed grammar, or core grammar, consists of a set of manually produced rules. Acquisition of new rules proceeds in the course of parsing the training sentences. The pattern of required new rules is often linguistically restricted in order to narrow down the number of plausible candidates [Osborne and Bridge, 1994] [Hindle, 1989]. One of the limitations in this method is that only one parse analysis is allowed to trigger the acquisition process. As a result, new rules are acquired based solely on the first parse available. This conflicts the fact that natural language is ambiguous, and thus all possible analyses should be taken into consideration.

Third, an initial grammar may consist of all possible rules, which are generated using a predefined set of non-terminals and terminals. The form of the rules is usually limited in Chomsky Normal Form and each rule is given a random initial probability. The inference process iteratively modifies the probabilities according to the frequency of use of the rules in the parses of a training set [Pereira and Schabes, 1992]. One disadvantage of this approach is the arbitrary non-terminal labeling of the inferred grammar, which may be linguistically implausible and therefore may weaken the ability for subsequent interpretation of analyzed sentences.

An alternative to these three types of initial grammar is the hybrid grammar used in the Explicit-Implicit technique [Briscoe and Waegner, 1992] [Shih, Young and Waegner, 1995]. The hybrid grammar consists of two sets of production rules, namely explicit and implicit. The explicit rules, like the core grammar mentioned above, are manually produced; the implicit part is similar to the third type of initial grammar but with headedness constraint [Jackendoff, 1977] imposed on the rules. The former aims to analyze general syntactic structure of the target language, whereas the later is responsible for analyzing the sentences the former fails to generate, including the ill-formed. This hybrid initial grammar aims to obtain the merits of the previous two approaches.

In Section 2, we describe the design of our sem-syn initial grammar that is a modified hybrid grammar tailored for Chinese inference. This is followed by in a brief introduction to our Chinese inference system in Section 3, and our conclusion and future work outlined in Section 4.

2. The Design of Sem-Syn Hybrid Grammar

In our system, parts-of-speech rather than words in Sinica Corpus are used in the inference process; therefore, before designing the initial grammar, it is important to examine the POS set utilized in the corpus. There are two issues to be taken into consideration here.

The Chinese Knowledge Information Processing (CKIP) group in Academia Sinica used to classify Chinese words into 178 parts-of-speech for their dictionary of 80,000 entries [CKIP, 1993]. This large set of POSs aims to describe the phenomenon in detail that semantic interpretation of Chinese words has strong influence on the structures of sentences. For instance, “房子” in the sentence “房子蓋好了” is originally the object of the verb “蓋好”, but is moved to the beginning of the sentence because it is a definite noun (we know which house it is). This large set of POS was later reduced to 46 for Sinica Corpus [CKIP, 1995]. This reduction was done by merging some semantically similar words into one. For instance, the semantically intransitive verb “重” in the sentences “這箱子很重” and “這箱子重十公斤” was originally given two different POSs (vh11 and vh12 respectively) because of their different syntactic behaviors. Nevertheless, these two together with other five POSs carrying different syntactic forms of the verb were assigned the same POS, vh, in the corpus.

The second issue is about words carrying different syntactic forms as arguments. In Sinica Corpus, a verb, which takes various syntactic forms as its arguments, is not given different

POSs if the semantics interpretation of the verb in those forms does not change. Here, we call it one-word-one-tag policy. According to this policy, if a verb can take both a clause (the verb is called 句賓述詞) and a noun phrase (the verb is called 單賓述詞) as its arguments, it will only be given one POS which carries a larger element(here it is marked as 句賓述詞). For instance, the verb “討論” can take both a noun phrase and a clause as the arguments in the sentences “我們討論明年的計畫” and “我們討論明年大家是否能申請計畫”, although its POS in the corpus is ve(句賓述詞).

A sem-syn initial grammar in our system is designed to handle the complexities mentioned above. Similar to the Explicit-Implicit technique, the grammar is divided into two components: semantically-oriented and syntactically-oriented rules. The semantically-oriented part of the initial grammar, which is manually developed, is the core part of the grammar responsible for capturing general semantically-consistent structures. For instance, if a verb is classified as active intransitive verb (va: 動作不及物述詞) in the corpus, there will be a semantically-oriented rule: VP -->va, regardless of its syntactic behavior or its other possible POSs. The following is an example showing how the rules look like:

```
VP1 [active +] --> va       /* 大軍出動了    */
VP1 [active +] --> vb PP    /* 授粉給雌蕊    */
VP1 [active -] --> vh       /* 這箱子很重    */
VP1 [active -] --> vi PP    /* 醉心於政治    */
```

The feature *active* is utilized to indicate whether the verb is an active (動作) or a stative (狀態) verb. Note that the stative intranstive verb vh also has other syntactic form that carries an NP (十公斤) as mentioned earlier in this section, but we leave it to the syntactically-oriented rules to handle since its form is inconsistent with the definition of vh.

The syntactically-oriented part, designed to generate from rule templates, is to handle the syntactic behaviors of a verb, accommodate the structures that are excluded due to the one-word-one-tag policy, and even deal with ill-formed sentences in the corpus. Like implicit rules with headedness constraint in the Explicit-Implicit technique, the syntactically-oriented rules will be generated from the following templates:

```
NT  -->  NT NT
NT  -->  NT  T
NT  -->  T  NT
NT  -->  T  T
```

Where NT is a non-terminal and T is a terminal (part-of-speech) symbols used in the grammar. However, implicit rules are a set of non-recursive rules with limited generating power. It is believed that the syntactically-oriented rules need to handle the majority of the training data due to the complexity of POSs in the corpus mentioned at the beginning of this section. Therefore, it is desirable to loosen the headedness constraint so that the bar level of a head daughter can be equal to or less than that of its mother. This results in a set of recursively syntactically-oriented rules.

The two set of rules will be put together to form our sem-syn initial grammar with a set of corresponding random probabilities. This grammar will then be ready for our inference process to proceed.

3. The Chinese Grammatical Inference System

We are currently developing a corpus-based grammatical inference system for Chinese.

The system consists of three components:

- Initial Grammar

The sem-syn hybrid grammar mentioned in the previous section is currently under development. A grammar development environment tool called GDE [Carroll, Briscoe and Grover 1991] is employed to develop our semantically-oriented rules in GPSG formalism, whereas syntactically-oriented rules will be generated using four templates shown in Section 2. Both parts of rules will then be converted into context-free rules and given random initial probabilities to meet the requirement of the stochastic inference algorithm mentioned below.

- Corpus

The pre-tagged primary school textbooks from Sinica Corpus are used for both training and testing. These data will be manually phrase-bracketed as a tree bank to provide the phrasal information during inference process. The phrase-bracketed test set will be used to examine the bracketing accuracy of the parses generated by the inferred grammar.

- Inference Algorithm

The system utilizes a chart-based Inside-Outside algorithm [Waegner, 93]. It is a stochastic inference algorithm that can take a stochastic context-free grammar as a source and iteratively re-estimates the set of probabilistic parameters of the grammar. Analogous to the forward and backward probabilities in conventional Hidden Markov Model(HMM), this algorithm define the inside(e) and outside(f) probabilities as:

$$e(s,t,I) = P(S=^{*}>O(s),\cdots..,O(t)/G),$$

$$f(s,t,I) = P(S=^{*}>O(1),\cdots,O(s-1),I,O(t+1),\cdots,O(T)/G)$$

where e(s,t,I) is the probability of the non-terminal symbol I generating the observation O(s),‥‥,O(t), and f(s,t,I) the probability of I being generated but not involved in generating the observation O(1),‥‥,O(s-1), and O(t+1),‥‥,O(T). G is the grammar, T is the total number of elements in the observation O(1),‥‥,O(T), and 1≦s≦t≦T. The inside probabilities are computed bottom-up, and outside probabilities are computed top-down. Like training the transition and emission probabilities in HMM, the values of e and f of non-terminal symbols can be used to re-estimate rule probabilities of the grammar in the similar fashion.

In our system, the set of probabilistic parameters of the sem-syn grammar will be iteratively re-estimated from all legitimate parses that conform with the bracketing constraints in our training tree bank (it is called supervised training). The inference process finishes when a change in total log probability of training sentences is less than a set threshold.

4. Conclusion and Future Work

We have outlined our on-going research on the design of the sem-syn initial grammar for the Chinese inference system. Unlike its European counterparts, Chinese language has its structure complexity, which have lead to a different design on the initial grammar for grammatical inference.

The sem-syn initial grammar has been under development in the project. We will soon start to build up the tree bank for the supervised training and develop the inference system. We hope that the inferred grammar will not only reflect the sentence structure in the training set, but also can predict the unseen sentences (test data) which in some sense are of the same nature as the training data. This expectation will be verified by experiments in which the test sentences are analyzed using the inferred grammar and their results (parses) are examined using the test tree bank.

References

Briscoe, E. and Waegner, N. (1992) Robust stochastic parsing using Inside-Outside algorithm. In AAAI Symposium on Statistic Applications to Natural Language.

Carroll, G. and Charniak; E. (1992) Learning probabilistic dependency grammars from labelled text. In AAAI fall Symposium Series: Probabilistic Approach to Natural Language, pp. 25-32.

Carroll, J., Briscoe, E., and Grover, C. (1991) A development environment for large natural language grammars. TR. 233, Computer Laboratory, Cambridge University, England.

CKIP (Chinese Knowledge Information Processing Group), (1995) Introduction to Sinica Corpus: A tagged balance corpus for Mandarin Chinese.TR95-02. Taipei: Academia Sinica.

CKIP (Chinese Knowledge Information Processing Group), (1993) Analysis of Chinese parts-of-speech.TR93-05. Taipei: Academia Sinica.

Hindle, D. (1989) Acquiring disambiguation rules from text. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pp. 118-125.

Osborne, M. and Bridge, D. (1994) Learning unification-based grammars using the spoken English corpus. In R.C. Carrosco and J. Oncina, editors, Proceedings of the Second International Colloquium on Grammatical Inference, pp. 260-270, Alicante. Springer-Verlag.

Pereira F. and Schabes Y, (1992). Inside-Outside re-estimation for partially bracketed corpora. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, pp. 128-135.

Shih, H-H., Young, S., and Waegner, N. (1995) An inference approach to grammar construction. Computer Speech and Language, 9:235-256

Waegner, N. (1993). Stochastic Model for Language Acquisition. PhD thesis, Cambridge University, England.

應用於電子商場營業類別查詢之語音辨識系統

陳建宏、黃英峰、陳榮貴

中華電信研究所 應用科技研究室
桃園縣楊梅鎮 326 民族路 5 段 551 巷 12 號（應科 5857）
Tel: (03)4245857 Fax:(03)4244167
Email: {cch, engfong, jkchen}@ms.chttl.com.tw

摘 要

本文針對網際網路(World Wide Web)上的資料庫查詢，提供除了鍵盤輸入外，另一個更便捷迅速的語音查詢方式。在現階段, 本系統主要是應用在電子商場營業類別查詢，但也適合其他類似的網際網路上之資料庫查詢。本系統有兩大優點：首先，我們提出一個架構在網際網路上之主從式(Client-Server)語音辨認系統架構，將大量的資料及語音辨認運算放在伺服器端，解決在瀏覽器端進行語音辨認時需要大量資料下載及計算能力不足的問題。 再者，我們也提出一個快速部分匹配(Fast Partial Matching, FPM) 語音辨認演算法，此演算法適合對已經過整理的資料庫，以標題辨認的方式查詢，且容許不完整和含有多餘內容的語音輸入，增加使用者的方便性。本系統對上網查詢生活資訊，提供了一個更方便迅速的查詢方式。

關鍵詞：語音辨認(Speech Recognition) 部分匹配(Partial Matching)
主從式(Client-Server)

1. 前言

目前網際網路(internet)蓬勃發展，特別是 WWW(World Wide Web)不但應用在網際網路，連公司內部的企業內網路(intranet)也都使用瀏覽器(browser)的方式來操作。我們相信未來個人電腦會成為每個家庭的必備家電用品，而音效卡和麥克風會成為個人電腦的必備週邊設備；同時，上網查詢生活資訊將是日常生活中常會去做的事。網路上很多應用都會使用到資料庫的查詢, 當可查詢的資料項目數目多到網頁無法列表顯示出來時，一般使用者便須敲入所要查詢的字串，由文字檢索系統去查出最可能的幾個答案，再由使用者挑選正確的答案。如果網站的資料庫能提供語音查詢功能，將使得上網查詢變得更輕鬆，甚至不用學電腦（中文輸入）也可以查。

一般的檢索方式主要有兩類，一類是所謂的全文檢索，例如 GAIS 的 WWW 檢索系統[1]，這樣的系統是在文章內容中搜尋，使用者輸入的字串基本上沒有什麼限制，若以語音來輸入，必須用聽寫機的方式來輸入任意字串，此種方式雖然應用的範圍較廣，但語音辨認的效果較難得到令人滿意的結果；另一類則是針對標題進行搜尋，例如 HiNet 上的智慧型電子商場(IEM), 也就是 HiGo 所提供的商品查詢[2]，即以營業項目或廠商名稱為搜尋對象，以查得相關廠商的超連結。此類應用通常使用在已經過整理的資料庫上的查詢，因為詞彙量有限，甚至可以用詞彙辨認的方式來作語音辨認，以目前的技術，此類的語音辨認的效果已可達實用的階段。

本文的重點是介紹我們如何在 WWW 的環境上建立一套實用的語音查詢系統，現階段將查詢的範圍設定為經過整理的資料庫的標題，例如人事資料庫的人名；問題題庫的題目等。主要內容包括：描述如何以主從式的方式在 WWW 上架構這樣的語音查詢系統；同時考慮到使用者語音輸入時的方便性，這裡我們也探討容許不完整語音輸入的辨認技術。本系統可應用的範圍包括在企業內網路(intranet)上的人事資料查詢，分機號碼查詢；或網路上的股票查詢，天氣查詢，颱風天是否上班的查詢，以及時刻表查詢等等，舉凡此類資料庫的查詢皆是。本文以 HiGo 的商品查詢為我們的實際應用系統，瀏覽器則使用 PC 及 Netscape。

2. 系統描述

一般而言此類系統所要查詢的內容，並不在網頁的內容中，且通常是動態的，隨時在改變，並且其數量也不小。若要下傳到瀏覽器這端來進行辨認，要花不少時間，不是很實際的做法。再考慮到瀏覽器端的 PC，可能不是每部機器的效能都可滿足語音辨認的需求。因此我們決定將辨認器分為兩部分，前端處理放在瀏覽器端，辨認器的主要部分則放在伺服器端，也就是在網路上以主從式的方式來建立我們的辨認系統。使資料庫及主要運算都在伺服器上，故可同時解決大量資料下傳及運算能力不足的問題。

下面幾節是系統方塊圖以及語音辨認技術的介紹：

2.1. 系統方塊圖

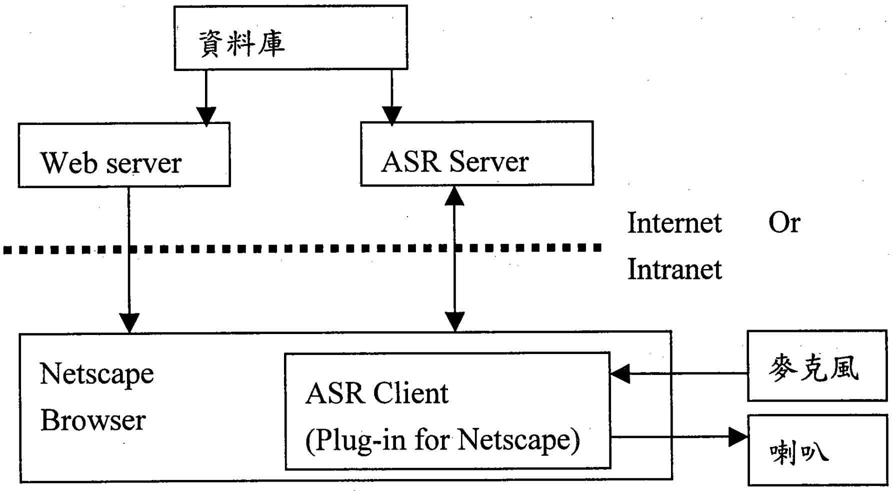


圖 一.不含防火牆的系統網路架構圖

圖一是此系統的方塊圖，運作方式描述如下：首先 Netscape 瀏覽器連上 Web Server，下載我們的首頁，此首頁包含 ASR(Automatic Speech Recognition) Server 的網路位置，及各項設定值。瀏覽器根據此首頁啟動先前已經安裝好的 Plug-in(ASR Client)，ASR Client 提示使用者輸入語音，經錄音並初步求取特徵值，ASR Client 再經由網路將特徵值傳送至語音辨認伺服器(ASR Server)，語音辨認伺服器根據傳來的資料選擇適當的辨認範圍，進行語音辨認，辨認結果連同超連結資料再送回 ASR Client，使用者可在 ASR Client 所顯示的畫面上得到所要的查尋的結果，或經使用者確認後點選連結到相對應的網頁去，進而從此網頁得到結果，或繼續查詢直到所要的結果得到為止。

至於與瀏覽器結合的部分，本階段我們選擇目前相當普遍的組合—Windows 95(或 NT) 搭配 Netscape。因為必須使用到低階的錄放音功能，我們選擇以 Plug-in 的方式來撰寫 ASR Client 端。此 Plug-in 基本上是一個 C 語言所寫成的動態連結程式庫(DLL)，只能在 X86 的 PC 上執行，所以目前我們的系統也只能在 X86 的 PC 上執行。但低階的錄放音功能是用 Windows 提供的低階函式介面撰寫，故任何在 Windows 上裝妥的音效卡皆可正常使用。

2.2. 主從式的網路架構

視 ASR Client 和 ASR Server 之間是否有防火牆相隔，系統的架構也相對的有所不同。茲分成企業內網路(intranet) 和網際網路(internet)來討論：

2.2.1.企業內網路

ASR Client 和 ASR Server 在網路上，以 TCP/IP 的協定，用 TCP mode 的 SOCKET 互相連接。在企業內網路的應用下，通常有快速的網路傳輸速度，而且 ASR Client 可以直接連接到 Server，中間不需透過防火牆，如圖一所示。為了獲得最佳的使用效果，也

就是最快的反應速度，我們訂定了 ASR Client 和 ASR Server 間的協定，使他們可用交談式的方式來傳遞資料。當 ASR Client 端一邊在錄音、求取語音特徵參數時，即可一邊將語音特徵參數往 ASR Sever 送。等錄音完成，ASR Client 於很短的延遲時間後即可收到 ASR Server 送回的辨認結果。

2.2.2.網際網路

在網際網路的應用下，必須考慮到一般公司均會加裝防火牆。對於 WWW 而言通常就是裝一個 Proxy，使所有的瀏覽器要下載公司外的首頁資料都必須透過 Proxy。如此我們在公司內的 ASR Client 便無法直接連接到在公司外的 ASR Server，那麼如何讓 ASR Client 可以將特徵值送到 ASR Server，並取回辨認結果呢？一個解決的方法是依據 HTTP(Hypertext Transfer Protocol)協定[3]，使用其 POST 的方式。依此法 ASR Client 連接上 Proxy 並將要傳送給 ASR Server 的資料一次全部傳送給 Proxy，Proxy 便會將資料轉送給 ASR Server；且將 ASR Server 送回的結果轉送給 ASR Client。這個方法有一個限制是：資料必須一次全部送出去，才等送回的結果。因此 ASR Client 的做法改為先完成錄音並求得全部特徵值，才將特徵值 POST 給 ASR Server，然後等待辨認結果回來。同時，ASR Server 端也必須配合。為了使原來的 ASR Server 能夠同時應付企業內網路和網際網路的需要，ASR Server 保持不變，另外多執行了一個模擬的 CGI(Common Gateway Interface)，簡稱 SCGI。SCGI 模擬 HTTP Server 上的 CGI 接收 ASR Client POST 過來的資料，再以交談式的方式和 ASR Server 連接，並傳遞資料給 ASR Server 及取得辨認結果。然後 SCGI 再以 POST 的方式把辨認結果傳回給 Client。系統方塊圖如圖二所示：

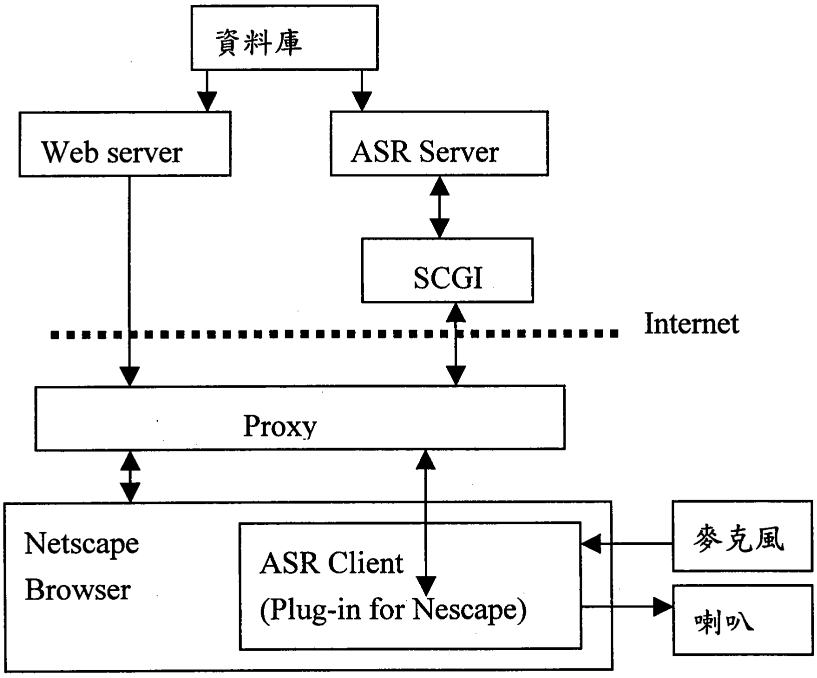


圖 二. 含有防火牆的系統網路架構圖

2.3. 語音辨認

在語音辨認的部分，雖然是標題式的辨認，但使用者往往不知道確切的文字內容，在此情況下，要正確無誤念出整個標題，實屬不可能，最常碰到的問題少掉一些或有多餘的內容，或兩者兼有，例如：

標題：　　電視電腦遊樂器
語音輸入：電視遊樂器

或

標題：　　電腦組件
語音輸入：電腦週邊設備

為了讓系統更好用，必須配合使用者的習慣，給使用者最少的限制同時又能得到良好的正確率。為此，我們採用了「部分匹配(Partial Matching ,PM)」的觀念，以解決上述問題。

本系統查詢的標題大約有 3000 個, 字數從 1 到 11 個字都有。下面幾節我們首先介紹本系統所使用的語音特徵向量及語音模型；然後在語音辨認方法方面, 分別介紹部分匹配(Partial Matching, PM)演算法以及快速部分匹配(Fast Partial Matching, FPM)演算法。

2.3.1.語音特徵參數及語音模型之介紹

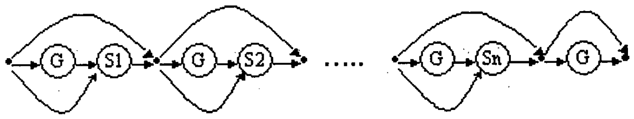
本系統的語音是麥克風輸入，取樣頻率為 16KHz。語音特徵參數向量由 25 個係數組成，分為 12 個 Mel-scale 反頻譜係數(cepstrum)、12 個一階微分的反頻譜係數、一個一階微分能量值。

語音模型採用目前自動語音辨認系統(Automatic Speech Recognition, ASR)最流行之隱藏式馬可夫模型(Hidden Markov Model, HMM)[L.R. Rabiner, 1989]，每個狀態中之機率函數採用 Gaussian mixture density， 混合(mixture) 的數目為 4。每一個國語音節包含子音部分及母音部分，子音部分的狀態數(state)為 3，母音部分的狀態數(state)為 5。

2.3.2.部分匹配(Partial Matching, PM)演算法

本法與一般的詞彙辨認類似，但為了克服多字或少字的困難，須對每一標題構建新的 HMM 網路，然後將輸入語音與每一個標題的 HMM 網路做最佳化匹配，求出屬於該標題的分數來，以分數最高的幾個標題當做候選標題，再由使用者去挑選。接下來是我們第一種方法的介紹：

1. 對每一個標題找出相對應的音節模型($S_1S_2…Sn$)，配合 Garbage Model 組合出如下的 HMM 網路:



圖三. 有順序限制的標題HMM網路圖

這個網路使標題中的任一音節都可被跳過，而且在任何音節之間都可加入Garbage model，但被跳過的音節都必須按其在標題中的順序出現。

2. 對上述 HMM 網路用輸入語音資料去比對出維特比路徑(Viterbi Path)，得到對數相似度(log likelihood)來求該標題的分數(score)。對於一個標題的分數求法如下：

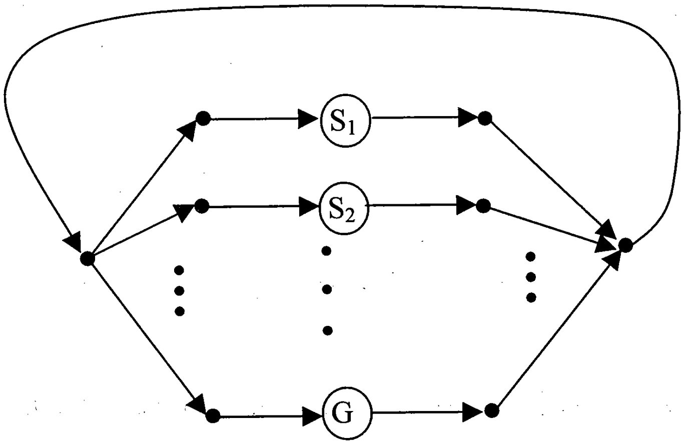
$$\text{Score} = \sum_{t}^{T} \log(f(O_t \mid \lambda_{St}))$$

其中o_t是一個音框的特徵向量，s_t是指最佳路徑中 t 音框所對應的狀態(state)，λ_{St}則是該狀態的模型參數。以相同做法為每一標題求得分數。

3. 找出分數最高的前幾名來。

在我們的系統中 Garbage Model 的選擇，使輸入語音匹配到正確音節的部分對該正確音節模型所得到的對數相似度(log likelihood)，會高於對 Garbage Model 所得到的對數相似度(log likelihood)高。故匹配到越多音節的標題，整體分數也越高。

選用圖三的 HMM 網路結構的原因是基於我們認為:一個有意義的短詞裡的音節組合，一般在念的時候也會按其固定順序。這個網路有這個順序限制，對有意義的短詞有較好的辨認效果。但對於標題中有兩個以上的短詞，使用者念的時候又沒有一定順序時，將造成只有其中一個短詞被匹配到。為了改善這個缺點，可將 HMM 網路改成如圖四的結構：



圖四. 無順序限制的標題 HMM 網路圖

圖四的結構使不按順序的輸入也都可被匹配到。另外為了減少因為網路限制太鬆，使不正確的標題也容易匹配到高的分數的問題，可在最佳路徑的音節連接到另一音節的地方，根據此兩音節在該標題中是否前後相鄰來給於不同的加分。假設最佳路徑中有 M 個音節($S_1S_2...S_M$)，此標題的整體分數計算方式如下：

$$\text{Score} = \sum_{t}^{T}\log(f(O_t \mid \lambda_{St})) + \sum_{l=1}^{M-1} Q(S_l, S_{l+1})$$

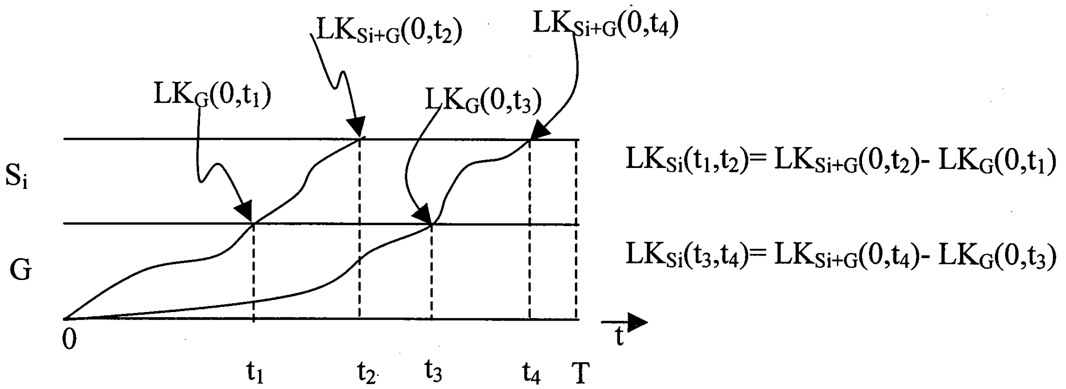
$$Q(S_l, S_{l+1}) = \begin{cases} q_0, \text{當}S_l\text{和}S_{l+1}\text{在標題中是前後相連} \\ q_1, \text{當}S_l\text{和}S_{l+1}\text{在標題中不是前後相連} \end{cases} \quad \text{且 } q_0>q_1$$

2.3.3.快速部分匹配(Fast Partial Matching, FPM)演算法

在 2.3.2.節中所提的 PM 方法因為要對每一標題去計算分數，當所要查詢的標題數目增大時，計算時間會成比例增加。當數目多到一個程度時，系統的反應速度便不符合實際使用需求。在這節中我們提出一個快速部分匹配演算法，以預處理的方式，避免一個一個比對。因此，大幅提昇比對速度的目的。此演算法分四個步驟：步驟 1. 從輸入語音找出 N 個最可能的音節區段(Syllable islands); 步驟 2. 根據這些音節以快速查詞的方法，找出有這些音節出現的 L 個標題來；步驟 3. 用步驟 1 所得到的 likelihood 來對在步驟 2 中找出來的標題進行快速評比(fast ranking)，將 L 個標題再縮小範圍到 K 個標題；(4). 若有需要可再對這 K 個標題進行仔細評比(detail ranking)。 4 步驟分述如下：

步驟 1. 尋找音節區段**(Syllable islands)**

在這裡要把輸入語音對國語 414 個音節作比對(warping)，每一個音節都要找出 n 個音節區段(Syllable islands)。這些音節區段彼此重疊的音框數不可超過設定值，比如說須少於音節區段音框數的四分之一。通常一個 標題中出現的相同音節數目不會多於 3 個，因此 n 等於 3 到 5 之間即可。如圖五. 所示，以 level-building 的方式將 Garbage model(G)和音節模型(S_i)疊起來，進行維特比(viterbi search)比對，計算出每一個時間點的累積對數相似度(log likelihood)值，並記錄每條路徑從 Garbage model(G)跳到音節模型(S_i)的時間點：



LK_{Si+G}(0,t2)
LK_{Si+G}(0,t4)
LK_G(0,t1)
LK_G(0,t3)
Si
G
0
t
t1
t2
t3
t4
T
LK_{Si}(t1,t2)= LK_{Si+G}(0,t2)- LK_G(0,t1)
LK_{Si}(t3,t4)= LK_{Si+G}(0,t4)- LK_G(0,t3)

圖五. Garbage model(G)和音節模型(S_i)的 level-building 之比對路徑圖

圖中 $LK_{Si+G}(0,t_2)$表示包含 Garbage model 及 syllable model，從 0 比對到 t_2所得到的累積對數相似度(log likelihood)；$LK_{Si}(t_1,t_2)$ 表示單單 syllable model，從 t_1 比對到 t_2所得到的累積對數相似度(log likelihood)；$LK_G(t_0,t_2)$則表示單單比對 Garbage model 得到的累積對數相似度(log likelihood)。音節區段的資訊包括起始時間(t_1)、終止時間(t_2)以及該區段對音框數做過正規化的 log likelihood rate($LR_{Si}(t_1, t_2)$)。參考[Seiichi Nakagawa,1997]，$LR_{Si}(t_1, t_2)$的求法如下：

$$LR_{Si}(t1,t2)=\frac{LK_{Si}(t_1,t_2)-LK_G(t_1,t_2)}{t_2-t_1}$$

因為 $LK_{Si}(t_1, t_2)= LK_{Si+G}(0, t_2)- LK_G(0, t_1)$，且 $LK_G(t_1, t_2)= LK_G(0, t_2)- LK_G(0, t_1)$，所以

$$LR_{Si}(t1,t2)=\frac{LK_{Si}(0,t_2)-LK_G(0,t_2)}{t_2-t_1}$$

接著我們要求得 t_2=0 到 T 的所有 LR_{Si}，從中挑出 n 個的步驟如下：

1. 首先挑出具有最大 $LR_{Si}(t_1, t_2)$值的那個音節區段。
2. 從其餘音節區段挑出次大的音節區段，並測試其與其他已挑中的音節區段是否重疊的音框數不超過設定值，若是則選用；否則捨棄。
3. 重複步驟 2 直到選到 n 個音節區段或沒音節區段符合為止。

然後從所有音節，共有 414*n 的音節區段, 中挑出最大的 N 個來，做為下一步驟使用。

步驟 2. 快速查標題

這裡是利用本所研發的快速查詞演算法[Eng-Fong Huang,1994]，來查出包含上一步驟所求出的音節之所有標題來。此演算法僅使用音節編號，並沒有用到任何其他的 likelihood 值，其運算速度相當快，即使輸入的音節區段數目多到 50 個，而待搜尋的標題數目多達數萬的情況下，此部份的搜尋時間也只佔整個語音辨認的極小部分的時間。因此標題數目多寡雖會影響此步驟的時間，但幾乎不影響整體的辨認時間。假設輸入的音節編號有：

ㄉㄧㄢ，ㄋㄠ，ㄑㄧ，ㄘㄞ…

會被搜尋到的標題可能是：

電腦遊戲	11··
電腦配件及耗材	11····1

其中「11··」和「11····1」代表該標題有音節出現的位置。

步驟 3. 快速評比**(fast ranking)**

根據步驟 2 所找出的標題，我們可在較少量的範圍，對這些標題評分(scoring)，再選出最後可能的候選標題。因為對一個標題而言，出現在其中的音節區段可能有相互重疊的情形發生，必須先進行音節區段的時間匹配，將互有重疊而分數較低的音節區段剔除。最後才根據匹配到的音節區段來計算標題分數。

標題分數的求法，我們實驗過以下三種方法，假設匹配到此標題的音節區段有 P 個：

(1). 音節區段的 LR(likelihood rate)除以其音框數作正規化，然後所有匹配到的音節區段的正規化 LR 再平均：

$$\text{score} = \frac{\sum_{i}^{P} \frac{LK_{Si}(t_1^i, t_2^i) - LK_G(t_1^i, t_2^i)}{t_2^i - t_1^i}}{P} \qquad (1)$$

(2). 音節區段的 LR 除以其音框數作正規化，然後所有匹配到的音節區段的正規化 LR 全部加起來：

$$\text{score} = \sum_{i}^{P} \frac{LK_{Si}(t_1^i, t_2^i) - LK_G(t_1^i, t_2^i)}{t_2^i - t_1^i} \qquad (2)$$

(3). 音節區段的 LR 不作正規化，直接把所有匹配到的音節區段的 LR 全部加起來：

$$\text{score} = \sum_{i}^{P} LK_{Si}(t_1^i, t_2^i) - LK_G(t_1^i, t_2^i) \qquad (3)$$

在我們初步的實驗中，發現第三個方式有最好的辨認結果。我們的推論是：在我們這樣的查詢工作中，有越多音節及音框匹配到的標題，越有可能是所要的標題。從(3)式可看出匹配到的音節數以及每一音節所含的音框數越多，累積的 score 就越高。最後按 score 高低找出前 K 個標題。

步驟 4. 仔細評比(detail ranking)

步驟 3 找出的前 K 個標題，已是最終所要的答案了。但若有需要，可再用前述的 PM 演算法再對這 K 個標題進行仔細評比。

2.3.4.語音辨認方法的結論

快速部分匹配(Fast Partial Matching, FPM)演算法利用多段式的做法將搜尋的範圍逐漸縮小。原本用 PM 需要數分鐘才能算完的計算，用 FPM 卻可即時(real time)完成。FPM 主要計算時間是花在音節區段的產生，這部分的運算時間和標題數目大小無關；而其他步驟所花的時間雖和標題數目大小有關，但因在 FPM 中所佔的時間比例相當小，故整體而言 FPM 的計算速度幾乎不受標題數目多寡所影響。

以實際使用的觀點來看，也就是假設使用者不會故意輸入很多多餘的語音，從初步實驗結果，可看出本系統有不錯的表現。因為本系統尚未完全完成，故本文尚未能提供完整的實驗數據。

3. 結語和展望

本文針對網際網路上的資料庫語音查詢系統，提出一個架構在網際網路(World Wide Web)上之主從式(Client-Server)語音辨認系統架構，解決在瀏覽器端進行語音辨認時需要大量資料下載及計算能力不足的問題。 同時我們也提出一個快速部分匹配(Fast Partial Matching, FPM) 語音辨認演算法，此演算法適合對已經過整理的資料庫，以標題辨認的方式查詢，且容許不完整及含有多餘內容的語音輸入。本系統對上網查詢生活資訊，提供一個更方便迅速的查詢方式。

本系統尚有許多工作未完成，除了將這些工作做得更完整之外，未來我們將朝兩個方向繼續努力：一是提高產生音節區段的準確度；另一個則是在標題的評比(ranking)方面加入詞庫及新詞的知識，以提高整體語音辨認的正確率。

致謝：

在此感謝中華電信研究所王所長、鄭副所長、張光耀主任及劉繼謚總計畫主持人對本項研究的支持與幫助，並感謝同仁在相關問題討論上給予的協助與建議。

參考文獻：

1. http://gais.cs.ccu.edu.tw/cwww2.html 。
2. http://higo.hinet.net/chinese2.htm 網頁中的商品查詢。
3. <draft-ietf-http-v11-spec-07> 可從 http://www.ics.uci.edu/pub/ietf/http/ 取得。
4. Eng-Fong Huang, Chian-Hung Chen and Hsiao-Chuan Wang, "New Search Algorithms for Fast Syllable Hypothesization and Lexical Access in a Large-Vocabulary Mandarin Polysyllable Word Recognizer", Computer Processing of Chinese and Oriental Languages, Vol. 8, No. 2, Dec 1994, pp. 211-225.
5. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recogniton",Proc.IEEE, 77 (2):257-286,February 1989.
6. Seiichi Nakagawa, Konstantin P. Markov, "Speaker Verification Using Frame and Utterance Level Likelihood Normalization", Proc. ICASSP, Vol.II, pp.1087-1090, 1997.

A Way to Extract Unknown Words Without Dictionary from Chinese Corpus and Its Applications

Yih-Jeng Lin, Ming-Shing Yu, Shyh-Yang Hwang and Ming-Jer Wu
(林義証)　(余明興)　(黃世陽)　(吳明哲)
Department of Applied Mathematics
National Chung-Hsing University, Taichung, Taiwan

Abstract

We propose a way to detect the unknown words from the corpus. We call such unknown words Chinese frequent strings(CFS). The strings could be the combinations of some common Chinese words that are defined in a traditional dictionary. Such Chinese frequent strings appear more than once in some Chinese texts. The method we proposed can automatically detect such strings without using any lexicon, and no word segmentation is needed.

We retrieve 55,518 Chinese frequent strings (reached for 13-gram in character) from a corpus consisting of 536,171 characters. To show that the strings we got are useful, we use these strings in Chinese phoneme-to-character and character-to-phoneme tasks. The test corpus contains manually-tagged phonetic symbols for each character. The correctness of the phoneme-to-character test is 96.5% and the correctness of the character-to-phoneme test is 99.7%. We make an MOS test about the determination of prosodic segments. The MOS score is 4.66 relative to the prosodic segments in spontaneous speech. This shows that the strings we retrieved are helpful in this aspect.

Keywords: *unknown words, phoneme-to-character, character-to-phoneme, prosodic segment*

1. Introduction

An intact Chinese electric dictionary is required in the processing of Chinese natural language (Chang 1997). Such dictionary plays an important role in machine translation, text-to-speech system, speech recognition system, and intelligent Chinese input methods. Yet there are many unknown words or new words coming into being in the world. The unknown words are various and diversified. Such as nominal compounds, verbal compounds, personal names, organization names and their abbreviations (Chang 1997, Chen 1997). There are many works focus on this problem recently (Chang 1994, Chang 1997, Chang etc. 1994, Chen and Bai 1997, Chen and Bian 1997, Chen 1994, Chien 1995, Sun 1994, Wu 1994).

We will propose a method to extract unknown words from corpus in Sect. 2. In Sect. 2, we also try to find some unknown words in a Chinese corpus with phonetic symbol for each character by our proposed method. And we applied the words we extracted to three applications in Sect. 3 to show that the words we get are reasonable and useful. We will discuss about our method in Sect. 4. We make some conclusions in Sect. 5.

2. The Proposed Method

Let's give definition to a Chinese unknown word at the beginning of this section. A Chinese unknown word is a Chinese string that is used frequently by the people. There are about 13000 Chinese characters. The number of combinations of several characters is very large. But there could be very few combinations meaningful. Such meaningful combinations are Chinese words in the lexicon or unknown words if they do not appear in the lexicon. It is more appropriate to use Chinese frequent strings (CFS) instead of Chinese unknown words. For example, "不得不去讀書" (can not help but study) is a Chinese frequent string since we find such Chinese pattern in many Chinese texts. It is not a word but a combination of some words that presents some meaningful idea.

2.1 Extracting Chinese Frequent Strings from Corpus

If a combination of some characters is a meaningful pattern, such combination will very likely appear more than once in a large corpus. We make a Chinese string pattern be a Chinese frequent string if it appears twice or more in the corpus. Since not all the strings are meaningful. It is very important to let a computer make a decision that which pattern is meaningful.

Our method is divided into two steps. The first step is searching for all characters in the corpus to determine which patterns appear more than one time. Such patterns will be gathered into a database that is so called "may be a word" database. The entries in the "may be a word" database are the strings and their number of occurrences. The second step is to find the raw frequency for each entry in the database mentioned just now. The raw frequency of a entry is the frequency of self appearance of the entry. And we can decide which patterns are meaningful patterns according to the raw frequency.

2.2 Constructing the Database of "may be a word"

The first step is to construct the database "may be a word" from corpus. In this step we collect each pattern that the number of characters of the pattern is less than or equal to 15. The number of characters of every entry in the database is not greater than 15. This is because that a breath group contains no more than 13 characters according to the experiments in our group (Pan 1998, Jen 1997). We need to find the strings of length two or more than 13. The reason will be explained later.

The frequency of each entry in the database is greater than or equal to two. Yet not all the entries are the patterns we want. Some of them are nonsense. For instance, consider the following fragment:

> "從二次大戰以後，自然科學的重要，幾乎沒有人不知道了。戰爭固然需要自然科學，和平更需要自然科學。「原子能和平用途」的目標，差不多是全世界文明人類的希望所寄託的。要達到這個目標，自然科學的研究是唯一的途徑。但我覺得現在世俗有一種誤解，就是許多人都以爲自然科學的用處只在增進物質上的文明，而和人類的精神文明沒有關係。要討論青年的求學問題，似乎應該先在這點上做匡謬正俗的功夫。所謂精神文明，當指人類道德的觀念和社會的組織各端而言。譬如仁義的敷教，法律的制定，都是精神

文明範圍以內的事。自然科學愈進步，則根據這些科學而制定的法律必更適合於文明人群的需要，至於「博愛之謂仁」，「行而宜之之謂義」，也不是每個人照著自己的意思可以定得好的。”

There are 36 patterns that appear more then once. They are listed in the first column of Table 1. There are some patterns that we may not wish to treat them as the unknown word. Such as “然科”, “然科學”, and “然科學的”, etc. However, The computer has no idea to get rid of such items. They can work well by computing. The rest task of our method for extracting unknown words is to identify the patterns automatically using a computer. This the second step in our proposal.

2.3 Extracting Unknown Words form Database “may be a word”

We have constructed a database that each entry may be a new word in the previous subsection. The main idea of this subsection is to make sure which pattern we indeed need. We will exclude a pattern whose appearance is due to the appearance of a longer pattern. For instance, all the occurrences of “然科” are caused by the occurrences of “自然科”. So “然科” will not be an unknown word. Again since “自然科學” are brought in by “自然科學的” and “需要自然科學”. The frequency of “自然科學” appear by itself is only 1. The frequency that a possible word appear by itself is shown in the third column of Table 1.

We could extract the entry which frequency of self appearance is more than once to be as our unknown words according to the third column of Table 1.

2.4 Implementation

We use a fraction of the Academia Sinica Balance Corpus as our training data (Chen 1996). The corpus we used is raw text. There is no information of segmentation. The total number of characters is 536,171. And the number of entries in “may be a word” database is 115,140. We applied the method mentioned in subsection 2.3 to decide which entry in the database is a real word that appears twice or more truly in the training corpus. We get 55,518 words. Appendix shows some high frequency words we extracted which can not be found in an ordinary dictionary.

Table 1. The listing of string patterns.

String Pattern	Occurrences	Frequency of self appearance
人類	3	1
人類的	2	2
之謂	2	2
文明	6	1
文明人	2	2
目標	2	2
自然	6	0
自然科	6	0
自然科學	6	1
自然科學的	3	3
沒有	2	2
制定	2	2
和平	2	2
明人	2	0
法律	2	2
科學	7	1
科學的	3	0
要自	2	0
要自然	2	0
要自然科	2	0
要自然科學	2	0
神文	3	0
神文明	3	0
然科	6	0
然科學	6	0
然科學的	3	0
精神	3	0
精神文	3	0
精神文明	3	3
需要	3	1
需要自	2	0
需要自然	2	0
需要自然科	2	0
需要自然科學	2	2
學的	3	0
類的	2	0

3. System Applications

In order to make sure that the words we retrieved are useful, we applied the words we retrieved to three tasks, namely Chinese character-to-phoneme(CTP) conversion, Chinese phoneme-to-character(PTC) conversion, and the determination of breath groups in an input

Chinese sentence.

3.1 The Chinese Character-to-Phoneme Task

The Mandarin text-to-speech system needs a character-to-phoneme system to get the correct syllables (Lin 1998, Ouh-Young 1985, Pan 1998). To have high performance of Chinese character-to-phoneme, we applied the words we retrieved in this aspect. The lexicon we used is the combination of Academia Sinica dictionary and the words we retrieved from the training corpus.

There are four principles in our CTP task. They are (1) the number of words should be minimized. (2) The number of characters in a word should be as many as possible. (3) The number of mono-character word should be minimized. (4) The probability of the combination of words should be high.

We performed an outside test for the Chinese character-to-phoneme task. The test corpus consists 82,610 characters with corresponding phonetic symbols for each character. To ensure the correctness, we check the phonetic symbols manually. The correctness of outside test is 99.7%.

3.2 The Chinese Phoneme-to-Character Task

The second work is the reverse of the work mention in subsection 3.1. That is Chinese phoneme-to-character task. This task is more difficult. There are some studies in this problem (Hsu 1995, Ho 1997). As we know, the best performance is 96% that is done by Hsu in Academic Sinica (Hsu 1995). Two possible applications of Chinese phoneme-to-character are Chinese speech-recognition system and Chinese input system. There are many methods to solve the problem. Three approaches have been proposed for this problem, (1) is the statistical approach, (2) is the grammatical approach, and (3) is the semantic template.

We applied the lexicon mentioned in subsection 3.1 to finish the task. We also use the same four principles in CTP to the task of PTC.

The test of Chinese phoneme-to-character is inside test. The test corpus contains 536,171 characters with corresponding phonetic symbols. The correctness of this test is 96.5%. Some errors like “它”, “他”, “她”, “牠” are not identified by our system. Such an identification is generally considered an difficult problem. However, the proper names we have trained could be identified by our strategies.

3.3 The Determination of Prosodic Segments

The third task we do is the determination of prosodic segments or breath groups (Lopez-Gonzalo 1997, Pan 1998, Jen 1997) in a sentence. The prosodic segments are important for a Mandarin text-to-speech system. To show the words we extracted are useful in this aspect, we try to decide the prosodic segments in input Chinese sentence. We have analyzed the real speech to get the prosodic segments. We found that a prosodic segment is about 7 to 10 characters. And a prosodic segment contains one or more words. No word will be divided to belong to different prosodic segments.

The evaluation we take is objective MOS evaluation (Wei 1997). The evaluation data are two paragraphs, say paragraph A and paragraph B. Two versions of the prosodic segments of these two paragraphs are provided. One version is obtained by analyzing the real speech and the other is got by our system. There are forty-two undergraduate students in the Department of Chinese Lecture in our university making the evaluation. The students are divided into two groups. One group takes the evaluation in paragraph A with prosodic segments determined according to real speech and paragraph B with prosodic segments determined by our system. The other group takes the evaluation in paragraph A with prosodic segments determined by our system and paragraph B with prosodic segments determined according to real speech. There are totally 50 sentences for every student.

The results are 3.834 for the prosodic segments of real speech and 3.572 for the prosodic segments of our system. The relative score is 3.572/3.834*5=4.66. This means that the words we retrieved make good help in determining the prosodic segments. Since some meaningful patterns or proper names can be identified by our system. The characters in such a pattern should belong to the same prosodic segment.

4. Discussion

The features of our extraction of unknown words can be summarized in the followings

(1) Our method can extract any unknowns for corpus. Such unknowns could be people names, origination names, verbal compounds, foreign translated nouns, and so no.

(2) We need no dictionary while extracting unknown words. And no Chinese word segmentation is needed in preprocessing.

(3) We could find long word patterns in one pass.

(4) There is not any rule applied in our system. The generation of rules very often requires many human works.

(5) The computation of frequency for each word is very accurate in our system. We have proposed a method to compute the frequency of occurrence for each word.

(6) There is no part-of-speech information needed. The use of part-of-speech information generally requires a large amount of human involvement.

5. Conclusions

We have proposed a robust way to extract unknown words form corpus. We also show that the words we retrieved are very useful. Such words make good help in Chinese phoneme-to-character, Chinese character-to-phoneme, and the generation of prosodic segments in a Mandarin TTS system.

Some future works we want to do are in the following three aspects.

(1) To extract more unknown words in a large corpus.

(2) Developing a robust system for Chinese phoneme-to-character task.

(3) Developing a high performance Mandarin text-to-speech system.

References

[1] C. H. Chang, "A Pilot Study on Automatic Chinese Spelling Error Correction," Communication of COLIPS, Vol.4, No.2, pp.143-149, 1994.

[2] J. S. Chang, "Automatic Lexicon Acquisition and Precision-Recall Maximization for Untagged Text Corpora," Ph.D. Thesis, National Ching-Hua University, 1997.

[3] J. S. Chang, S. D. Chen, S. J. Ker, Y. Chen, and J. Liu, "A Multiple-Corpus Approach to Recognition of Proper Names in Chinese Texts," Computer Processing of Chinese and Oriental Languages, Vol. 8, No.1, pp. 75-85,1994.

[4] K. J. Chen and M. H. Bai, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," Proceeding of ROCLING X, pp.159-174, 1997.

[5] H. H. Chen and G. W. Bian, "Proper Name Extraction from Web Pages for Finding People in Internet," Proceeding of ROCLING X, pp.143-158, 1997.

[6] K. J. Chen, C. R. Hunag, L. P. Chang, and H. L. Hsu, "SINICA CORPUS: Design Methodology for Balanced Corpora," Proceeding of PACLIC 11th Conference, pp.167~176, 1996.

[7] H. H. Chen and J. C. Lee, "The Identification of Organization Names in Chinese Texts," Communication of COLIPS, Vol.4, No.2, pp. 131-142, 1994.

[8] L. F. Chien, "尋易(Csmart)—A High-Performance Chinese Document Retrieval System," Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages, ICCPOL'95, pp. 176-183, 1995.

[9] Eduardo Lopez-Gonzalo, Jose M. Rodriguez-garcia, Luis Hernandez-Gomez, and Juan M. Villar, "Automatic Prosodic Modeling for Speaker and Task Adaptation in Text-to-Speech," Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pp. 927-930, 1997.

[10] W. L. Hsu, "Chinese Parsing in a Phoneme-to-Character Conversion System Based on Semantic Pattern Matching," International Journal on Computer Processing of Chinese and Oriental Languages 40, pp.227-236, 1995.

[11] T. H. Ho, K. C. Yang, J. S. Lin, and L. S. Lee, "Integrating Long-Distance Language Modeling to Phoneme-to-Character Conversion," Proceeding of ROCLING X, pp.287-229, 1997.

[12] Y. J. Lin and M. S. Yu, "An Efficient Mandarin Text-to-Speech System on Time Domain," to appear in IEICE Transactions on Information and Systems in June 1998.

[13] M. Ouh-Young, "A Chinese Text-to-Speech System," Master thesis, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, June 1985.

[14] N. H. Pan, "Prosody Model for Syllable Energy and Intonation in a Mandarin Text-to-Speech System," Master thesis, Department of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan, June 1998.

[15] W. T Jen, "Prediction Models for Syllable Duration in a Mandarin Text-to-Speech System," Master thesis, Department of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan, June 1997.

[16] M. S. Sun, C. N. Huang, H. Y. Gao, and Jie Fang, "Identifying Chinese Names in Unrestricted Texts," Communication of COLIPS, Vol.4, No.2, pp. 113-122, 1994

[17] H. N. Wei, "An Approach to the Measurement of Fondness and Similarity on Speech," Master thesis, Department of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan, June 1997.

[18] Zimin Wu and Gwyneth Tseng, "Chinese Text Segmentation for Text Retrieval: Achievements and Problems," JASIS, 44(9), pp. 532-542, 1994.

Appendix: Some unknown words we extracted from corpus

Unknown Words	Frequency	Unknown Words	Frequency
的人	114	在日常生活中	17
他說	74	中央研究院的	8
也是	71	五十而知天命	8
他的	70	生活與工作的	7
這是	65	人活在世界上	6
這個	62	民族學研究所	6
這種	59	每個人都可以	6
也有	53	近代史研究所	6
爲了	49	己欲立而立人	4
的時候	79	人力資源的開發	8
的問題	68	自我的重新探索	6
事實上	53	金車文教基金會	5
這就是	51	做爲一個現代人	5
的工作	48	本院數理組院士	4
的生活	44	在生活與工作中	4
這樣的	42	研討會議程如下	4
有一位	39	學者社區的形成	4
的方法	39	七十而從心所欲	3
的關係	39	人與人之間的關係	5
一九九二	24	山豬窟垃圾掩埋場	5
我們可以	20	數學研究所研究員	5
山西商人	19	人多的地方不要去	4
我們應該	18	天下無不是的父母	4
的年輕人	18	不是一件容易的事	3
我們必須	17	本院歷史語言研究所	4
一個人的	15	不可或缺的靈魂人物	3
千萬不要	15	民族學研究所研究員	3
有很大的	15	一個完全不同的人格	2
這是一個	15	一個發出菩提心的人	2
學者的社區	13	歷史語言研究所研究員	4
人與人之間	12	一場沒有終止線的競爭	2
的觀點來看	11	人生最大的本錢是尊嚴	2
我們可以說	9	不要擔心別人不了解我	2