應用不定長度特徵之條件隨機域於口語不流暢語流修正

Disfluency Correction of Spontaneous Speech using Conditional Random Fields with Variable Length Features

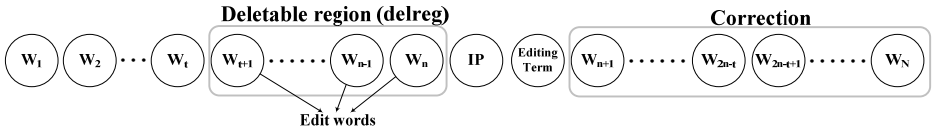
葉瑞峰 [2]、吳宗憲 [1]、吳維彥 [1]
[1]Dept. of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan
[2]Dept. of Computer Science and Information Engineering, Far East University, Tainan County, Taiwan

摘要

針對口語化語音中之不流暢語流(disfluency)現象，本文提出以不定長度特徵之條件隨機域。利用狀態轉移特徵函數、觀測特徵函數以及相對應之參數，針對不流暢語流進行修正。其中觀測特徵函數可整合多種知識來源，包括前後文相關特徵、不流暢相關特徵以及圖樣符合相關特徵。在狀態方面我們使用可變動長度單位，包括詞、字元串集(chunk)以及句子三種不同狀態。在評估上，則使用現代漢語口語對話語料庫(MCDC)做為訓練以及測試語料。其中被修正詞(editing word)錯誤率為 17.3%，相對於 DF-gram、HMM、最大熵以及 N-gram 加校正之混合模型的方法分別降低了 11.7%、8.7%、8%以及 3.9%。在給定中斷點的情況下，被修正詞錯誤率為 6.1%。實驗證明所提之模型優於其他方法，並可有效偵測並修正口述語言中之不流暢語流。

1. 緒論

要應用語音技術於人機介面上，語音辨識則為最重要且核心之技術之一。近十年來，語音辨認技術已臻於成熟且蓬勃發展。目前的語音辨認系統對於朗讀的語音輸入辨認效果極佳，然而要實際應用，必須考慮口語化語音[1]。而口語化語音常會伴隨著非正規化(ill-formed)以及不流暢語流(disfluency)，這些現象會造成目前辨識系統的錯誤率大幅度提高，以至於無法應用於日常生活[2]。而參雜著不流暢語流之辨識後文字，也會使得使用者極不容易閱讀，對使用者造成困擾[3]。編輯不流暢語流結構共可區分為以下四個部份如圖一所示。



圖一 編輯不流暢語流之結構

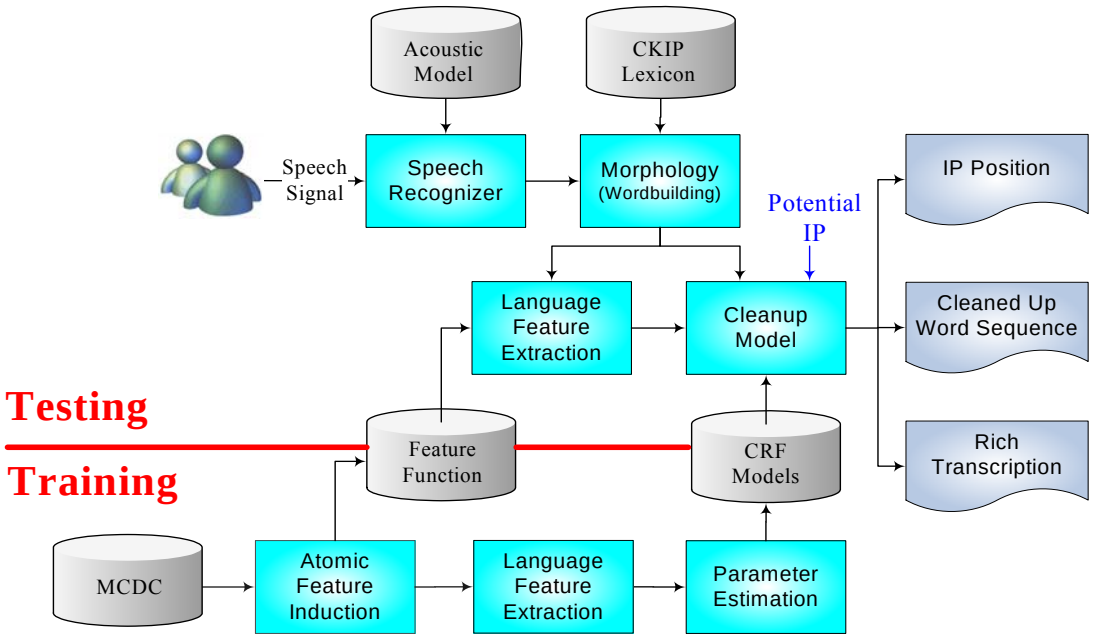
編輯不流暢語流包括三種型別：重複(Repetition)、修正(Repair) 和重開始(Restart)，其定義如下。重複即語者重複語句的某個部份，也就是可刪除區域與修正區域的語句重複。修正即語者將語句的某個部份做修正。也就是可刪除區域將取代修正區域並改變它的意思。重開始:語者將未完成的語句中斷並重新開始另一句。也就是中斷點前面的部分全都是可刪除區域。

相關的研究在國外方面， ISCI 以及 SRI 等國際研究中心利用語言模型以及韻律模型偵測不流暢語流[4]、結合基於詞和詞性的語言模型解決重複[5]和使用隱藏事件語言模型直接對不流暢語流進行統計式分析以及利用不流暢語流語言模型(DF-gram)來預測是否出現不流暢現象[6]以及使

用最大熵模型以及隱藏馬可夫模型修正不流暢語流[7]。John Bear 應用不同知識來源來針對不流暢語流進行偵測及修正[8]，Anand Venkataraman 使用人工訂定之規則來判斷不流暢語流[9]、Matthias Honal 利用噪音頻道(Noisy Channel)的觀念，運用不同特徵訓練出統計模型並以線性組合來修正之[10]。Charniak and Johnson 建立一基於詞性特徵之分類器來預測可被刪除區域[11]。Nakatani and Hirschberg 利用聲學、音韻學以及語言特徵建立一決策樹模型來偵測重複[12]。Snover 等人以及 Joungbum 等人利用轉換學習(Transformation-Based Learning)偵測不流暢語流[13][14]。日本的 Furui 則致力於口語化語音辨識之研究[15]。國內方面，中研院針對不流暢語流的語音特性做分析[16]，台灣大學研究關於不流暢語流中斷點偵測之特徵[17]、交通大學電信工程系則針對自發性中文語音建立辨識系統[18]以及自發性對話語音辨識做研究[19]。近年來，成功大學也投入大量研究能量於口語對話系統中不流暢語音之語音動作型態模型化與驗證[20]以及運用語言模型與校正模型來對編輯不流暢語流做修正[21]。本文則針對編輯不流暢語流提出一利用條件隨機域之修正方法。

2. 系統架構

本文所提之系統架構如下所示:



圖二，不流暢語流修正系統之系統架構

各模組之功能如下:

語音辨識器模組 (Speech Recognizer):利用 HTK 將語音信號進行語音辨識並得到音節網(Syllable Lattice)。其中我們定義 157 個次音節模型(sub-syllable models)以及 11 個填空詞模型(filler models)。構辭模組 (Morphology):將經過語音辨識器所得之音節網對照詞典得到相對應的詞網(Word Lattice)。

語言特徵擷取模組 (Language Feature Extraction):根據特徵，對詞網中語言上的資訊特徵擷取。

子特徵推導 (Atomic Feature Induction):使用語料訓練出最小單元的特徵，稱之為子特徵。

參數估測 (Parameter Estimation): 對於不同的特徵及其相對應的參數值，估計出這些參數。

修正模型 (Cleanup Model): 根據詞絡、可能的中斷點以及擷取之語言特徵配合對應的參數對詞絡作修正。

最後，詞絡經過修正模型修正之後，我們將得到中斷點資訊、修正後結果、以及辨識後結果。而系統流程分為訓練和測試兩部份，分別如下:

訓練部份--首先，從 MCDC 語料中經由子特徵堆導得到子特徵。這些子特徵則成為我們修正模型中的特徵函數。最後則是對語料進行語言特徵擷取並對所有的特徵函數進行參數估測，這些參數即為測試時修正模型之參數。

測試部分--語音訊號經由辨識器進行語音辨識後，得到音節絡，將此音節絡配合詞典經由構辭後會得到詞絡，然後對此詞絡做語言特徵擷取。最後，修正模型則根據詞絡、可能的中斷點以及所擷取之語言特徵，配合訓練所得到的參數模型，找出中斷點資訊、修正後結果、以及辨識後結果並將其輸出。

而修正不流暢語流之流程以式子(1)表示，一語音訊號 X 輸入後，我們要得到其相對應之最佳狀態序列 S，於是引入詞序列 W 此參數，之後在條件獨立的假設下，得到最後的式子，也就是從語音訊號我們可經過辨識器得到詞序列，而後我們從詞序列找到相對應的狀態序列。在本論文中，我們使用不定長度特徵之條件隨機域得到 $P(\mathrm{S}|\mathrm{W})$。

$$\begin{aligned}\hat{\mathrm{S}} &= \arg\max_{\mathrm{S}} P(\mathrm{S}\,|\,\mathrm{X}) \\ &= \arg\max_{S}\left(\sum_{W} P(\mathrm{S}|\mathrm{W},\mathrm{X})\mathrm{P}(\mathrm{W}|\mathrm{X})\right) \qquad (1)\\ &\cong \arg\max_{S}\left(\sum_{W} P(\mathrm{S}|\mathrm{W})\mathrm{P}(\mathrm{W}|\mathrm{X})\right)\end{aligned}$$

3. 不定長度特徵之條件隨機域

條件隨機域

條件隨機域為一種無向圖(Undirected Graphical)的模型，可被用來估算給予一觀測序列，得到相對應的狀態序列其交集的機率分佈。其概念是以隨機域為基礎，加上全域被限制於 X 這個條件，稱之一條件隨機域，X 為觀測序列。正式的來說，我們定義一圖 G=(S,E)為一無向圖，S 為所有點之集合，每個點皆為隨機變數，我們可將某個點 Sv 看成狀態序列 Y 上的某個狀態 Yv。若每個隨機變數 Yv 都遵守馬可夫原則，也就是說給予 X 和所有其他隨機變數 Y{u|u≠v}的條件之下，得到隨機變數 Yv 的機率

$$p(\mathrm{Y_v}\,|\,\mathrm{X},\mathrm{Y_u},u\neq v,\{u,v\}\in \mathrm{V}) \qquad (2)$$

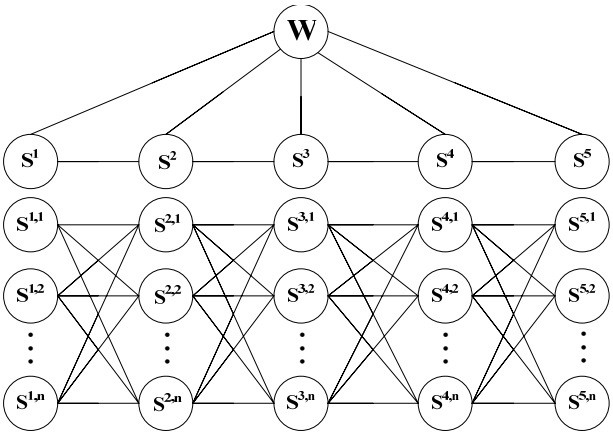
相等於給予 X 和 Yv 的鄰居點的條件之下，得到隨機變數 Yv 的機率，

$$p(\mathrm{Y_v}\,|\,\mathrm{X},\mathrm{Y_u},u=neighbor(v),\{u,v\}\in \mathrm{V}) \qquad (3)$$

則(X,Y)為一條件隨機域。

理論上來說，圖 G 的結構可以是任意的，然而，當用來對序列建模時，最簡單且最普通的圖的結構則為形成一個簡單的一階鏈(First-Order chain)，如下圖所示，其中 W 為觀測值，S 為

狀態序列。



圖三，條件隨機域示意圖

於是在給予觀測序列 X，得到對應狀態序列 S 的機率為:

$$P(S|W)=\frac{1}{Z}exp\left(\sum_t\sum_k \lambda_k f_k\left(s^{(t-1)},s^{(t)},\mathrm{W}\right)+\sum_t\sum_k \mu_k g_k\left(s^{(t)},\mathrm{W}\right)\right) \tag{4}$$

其中 $f_k\left(s^{(t-1)},s^{(t)},\mathrm{W}\right)$ 為整個觀測序列和在狀態序列中位置 t-1 的狀態轉到位置 t 的狀態轉移特徵函數定義為；

$$f_k\left(s^{(t-1)},s^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s^{(t-1)}=s\wedge s^{(t)}=s'\\ 0 & otherwise\end{cases} \tag{5}$$

而 $g_k\left(s^{(t)},\mathrm{W}\right)$ 為狀態序列中位置 t 和觀測序列的狀態特徵函數:

$$g_k\left(s^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s^{(t)}=s\wedge \mathrm{W}^{(t)}=w\\ 0 & otherwise\end{cases} \tag{6}$$

λ_k 以及 μ_k 為從訓練語料中估出來的參數，Z 是正規化係數，為所有可能的狀態序列的機率總和，其定義如下式所示:

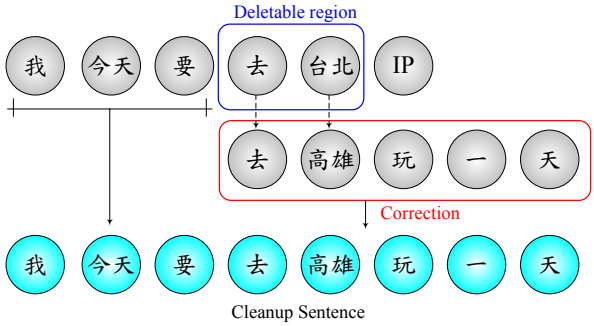
$$Z=\sum_{W,s} exp\left(\sum_t\sum_k \lambda_k f_k\left(s^{(t-1)},s^{(t)},\mathrm{W}\right)+\sum_t\sum_k \mu_k g_k\left(s^{(t)},\mathrm{W}\right)\right) \tag{7}$$

就修正不流暢語流而言，我們可以把觀測序列 X 當作是以詞構成的序列，狀態序列 Y 為以 1 和 0 組成之序列，若在觀測序列中位置 t 的詞所對應的狀態為 0，則代表此詞為一被修正詞，需將其刪除；若為 1，則表示位置 t 的詞為流暢部分，並將其保留。

.

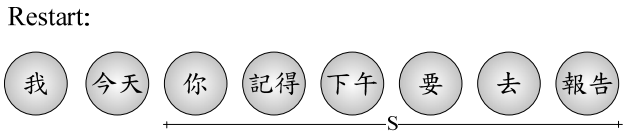
不定長度特徵

在不流暢語流中的"修正"型別中，其發生的情形為語者語句部分有誤，故會修正之。我們從語料觀察到修正時常會存在著圖樣一致性(Pattern Matching)的現象，也就是修正與被修正的兩個部份其句法結構相似，這些圖樣的組成可能是一個片語(phrase)、或是一個字元串集(chunk)，如圖四。於是我們先將某些詞的單元進一步合併為字元串集單元。



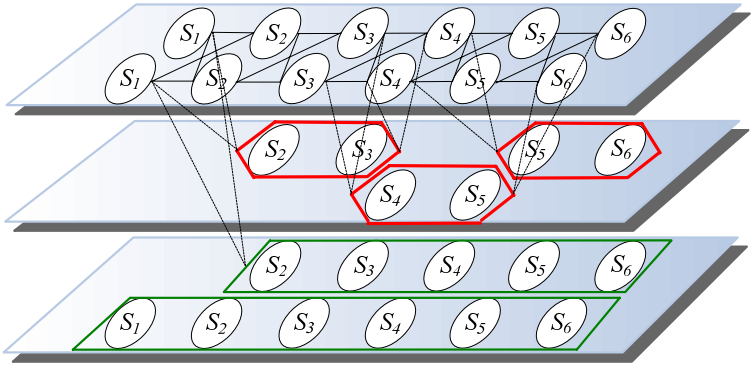
圖四，編輯不流暢語流之修正

而在"重開始"此種型別的不流暢語句中，因為使用者是將原句放棄並重新開始另一語句，故在其結構中幾乎都會有包含另一語句的現象發生，也就是說若一輸入語句包含另一語句，則是"重開始" 此種型別的不流暢語句的機率大增，如圖五。若我們能先將某些詞的單元進一步合併為句子單元，則對於修正不流暢語流之"重開始"型別，將會大大減少被修正辭誤判(False Alarm)的錯誤率。



圖五，編輯不流暢語流之 restart

原本條件隨機域的狀態序列皆以詞為單位，而本論文提出一種以條件隨機域為基礎之方法—不定長度特徵之條件隨機域，對於一輸入之詞序列，於找出最佳路徑狀態序列之前，先將可合併之字元串集及句子合併，觀念似於語音辨識時所做構詞的部份。而狀態部份則從原來的詞，增加為共三層狀態，分別是詞、字元串集以及句子，再根據合併後的狀態絡找出最佳狀態序列，然後依此狀態序列得到最後的修正結果。如下圖六所示:



圖六，不定長度特徵之條件隨機域

在給予觀測序列 W，得到對應狀態序列 S 的機率為:

$$P(S|W)=\frac{1}{Z}exp\left(\sum_{t}\sum_{k}\sum_{p,q}\lambda_k f_k\left(s_p^{(t-1)},s_p^{(t)},\mathrm{W}\right)+\sum_{t}\sum_{k}\sum_{c}\mu_k g_k\left(s_p^{(t)},\mathrm{W}\right)\right) \quad (8)$$

其中 p,q 為層次個數，在此共有詞、字元串集以及句子這三種層次，而每一層次包含兩種狀態，一種為 0，也就是此層次為可刪除區域，修正時需將其刪除；另一種則是 1，即為流暢部分，修正時將其保留。其中 $f_k\left(s_p^{(t-1)},s_p^{(t)},\mathrm{W}\right)$ 為整個觀測序列和在 p 層次狀態序列中位置 t-1 的狀態轉到位置 t 的狀態轉移特徵函數，而 $g_k\left(s_p^{(t)},\mathrm{W}\right)$ 為 p 層次狀態序列中位置 t 和觀測序列的狀態特徵函數。

在特徵函數方面我們共分為三類，分別是上下文相關、不流暢相關以及圖樣符合相關觀測特徵函數，我們一一介紹如下。

上下文相關觀測特徵函數:上下文相關觀測特徵函數乃是根據所觀測點位置之上下文關係取得一觀測範圍並以 N-gram 的概念所建立之特徵函數並以樣板(Template)的形式來表示。

不流暢相關特徵函數:我們從 Apriori 演算法擷取出關聯法則，以此關聯法則為我們的不流暢相關特徵函數。譬如我們找出"去台北=>去高雄"此一關聯法則，我們即可定義觀測特徵函數如下:

$$g_1\left(s_p^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s_p^{(t)}=s\wedge W^{t-k}\text{出現}"\text{去台北}"\wedge W^{t+k}\text{出現}"\text{去高雄}"\\ 0 & otherwise\end{cases} \quad (9)$$

故若有符合此特徵函數，則"去台北"被刪除機率則大增。所有擷取到的的關聯法則皆可以此方式定義出。

圖樣符合相關觀測特徵函數:因修正(repair)此種不流暢語流會有圖樣符合(Pattern Matching)的情形，故我們訂定了圖樣符合相關觀測特徵函數，例子如下

$$g_2\left(s_p^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s_p^{(t)}=s\wedge P^{t-1}=P^{t+1}\wedge P^{t}=P^{t+2}\\ 0 & otherwise\end{cases} \quad (10)$$

也就是當出現連續兩個圖樣其詞性序列一致時，極有可能是不流暢現象。因修正之編輯詞個數集中在 1-2 個，故我們定義時皆以 2 個詞的詞性的組合當做圖樣符合相關觀測特徵函數。我們總共有 37 種詞性，故會組合出 1369 種特徵函數。因填空詞出現在語料中，所以我們加上位移 k 的考慮如下:

$$g_3\left(s_p^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s_p^{(t)}=s\wedge P^{t-1}=P^{t+1+k}\wedge P^{t}=P^{t+2+k}\\ 0 & otherwise\end{cases} \quad (11)$$

於是我們利用這些特徵函數配合參數估測得到各權重植,之後利用此修正模型產生與輸入句相對應的狀態序列，並對原句進行修正。

子特徵推導演算法

在建立不定長度特徵之條件隨機域時，我們必須先將不同的層次確認出來。在字元串集(chunk)

部分，我們使用的是 Apriori 演算法[22]，如找出的關聯法則包含互相鄰近的詞，則我們將這些詞組合為一字元串集；在構成句子部份，我們是以動詞配合其必要論元為一組成句子之主要成分來對輸入語句判斷是否包含另一語句。

動詞為一個句子的中心，句子的構成常被視為動詞本身語態的擴展延伸。動詞的歸納整理關係到詞彙知識及句法的表達。於是，我們利用中研院詞庫小組(CKIP)所研究之中文詞類分析[23]，其對於漢語的詞類分析及相對應的詞彙結構提出完整的看法。對於其中每一類動詞皆整理出其必要論元(argument)。我們假設句子的主成分由這些動詞與其必要論元所構成，若一輸入語句之詞性序列 x = [x1, x2, …, xm]和某一動詞以及其必要論元詞性序列 y=[y1, y2, …, yk]存在共同子序列(common subsequence)z，且 z=y，則我們判斷輸入語句包含一句子，範圍從 y1 到 xm。

假設現在有 x, y 兩個序列，若存在一個序列 z 同時為 x 與 y 的子序列，那麼 z 便稱為 x, y 的相同子序列（common subsequence）。舉個例子，假設我們輸入語句"你 今 我 明天 要 去 台北"其詞性序列 x 為[Nhaa, Ndabd, Nhaa, Nddb, Dbab, VA11,Nca]，一動詞 VC1 配合其必要論元之詞性序列 y 為[Nh, VA11,Nca]，我們則可找到其共同子序列 z=y，於是我們判斷[Nhaa, Nddb, Dbab, VA11,Nca]此序列為一句子。

參數估算

對於不定長度特徵之條件隨機域之最大對數似然法參數估測(ML)，我們定義機率為

$$p(s\,|\,W,\Theta)=\frac{1}{Z}exp\left(\sum_{t}\sum_{k}\sum_{p,q}\lambda_k f_k\left(s_p^{(t-1)},s_q^{(t)},\mathrm{W}\right)+\sum_{t}\sum_{k}\sum_{p}\mu_k g_k\left(s_p^{(t)},\mathrm{W}\right)\right) \tag{14}$$

從訓練資料的集合中來估算出使得訓練資料的 log-似然度最大的一組參數 $\Theta=(\lambda_1,\lambda_2,\cdots;\mu_1,\mu_2,\cdots)$。於是我們將條件隨機域的 $p(s\,|\,W,\Theta)$ 代入 log-似然度函數的定義式:

$$\begin{aligned}\mathrm{L}(\Theta)&=\log\prod_{\mathrm{W},s}\mathrm{p}(\mathrm{s}|\mathrm{W},\Theta)^{\tilde{p}(W,s)}\\&=\sum_{W,s}\tilde{p}(W,s)\,log\;p(s\,|\,W,\Theta)\\&=\sum_{W,s}\tilde{p}(W,s)\,log\left(\frac{1}{Z}exp\left(\sum_{t}\sum_{k}\sum_{p,q}\lambda_k f_k\left(s_p^{(t-1)},s_q^{(t)},\mathrm{W}\;\right)+\sum_{t}\sum_{k}\sum_{p}\mu_k g_k\left(s_p^{(t)},\mathrm{W}\;\right)\right)\right)\end{aligned} \tag{15}$$

經過整理之後得到下式

$$\sum_{W,s}\tilde{p}(W,s)\left[\sum_{t}\sum_{k}\sum_{p,q}\lambda_k f_k\left(s_p^{(t-1)},s_q^{(t)},\mathrm{W}\right)+\sum_{t}\sum_{k}\sum_{p}\mu_k g_k\left(s_p^{(t)},\mathrm{W}\right)\right]-\sum_{W}\tilde{p}(W)\,log\,Z \tag{16}$$

之後我們對 log-似然度函數偏微參數 λ_k：

$$\begin{aligned}\frac{\partial L(\Theta)}{\partial\lambda_k}&=\sum_{W,s}\tilde{p}(\mathrm{W},s)\sum_{t=1}^{n}\sum_{p,q=1}^{l}f_k\left(s_p^{t-1},s_q^{t},\mathrm{W}\right)\\&\quad-\sum_{W,s}\tilde{p}(W)p(s\,|\,W,\Theta)\sum_{t=1}^{n+1}\sum_{p,q=1}^{l}f_k\left(s_p^{t-1},s_q^{t},W\right)\\&=E_{\tilde{p}(W,s)}\left[f_k\right]-E_{p(s|W,\Theta)}\left[f_k\right]\end{aligned} \tag{17}$$

若要求得全域最大值則須令式子(17)為零，求得 Θ 解。不過一般來說，這是不可行的，因為將 log-

似然度函數經偏微分後設為零來解出參數Θ，未必為封閉解。故本文採用一些反覆(iterative)形式的技巧取得使 log-似然度最大之參數。因 IIS 具有較快收斂之優點，故本論文是利用 IIS 演算法來進行參數估測。IIS 演算法是以 GIS 演算法為基礎改變而成，其優點為收斂速度較 GIS 演算法快。我們以此為基礎配合 Lafferty 等人提出的動態規劃法來估算參數。

Lafferty 等人[24]觀察到對於一個鏈狀結構的條件隨機域(CRFs)，給予觀測序列 W 所得到狀態序列 s 的條件機率 p(s|W)可簡單的用矩陣的形式來表示。對於觀測序列 W 中的每一個位置 t，我們分別定義了一個 $|\kappa|\times|\kappa|$ 的矩陣隨機變數 $\mathrm{M}_t(W)=\left[M_t(s',s\,|\,W)\right]$，$\kappa$ 為狀態的種類個數。

$$\mathrm{M}_t(s',s\,|\,W)=exp\left(\sum_k\sum_{p,q}\lambda_k f_k\left(s_p^{t-1}=s',s_q^t=s,W\right)+\sum_k\sum_p\mu_k g_k\left(s_p^t=s,W\right)\right) \tag{18}$$

每個 $\mathrm{M}_t(W)$ 可視為表示在時間 t 時，模型中每個轉移的權重。於是我們可以將未正規化的條件機率 $P^*(s\,|\,W)$ 表示為矩陣的連乘積:

$$P^*(s\,|\,W)=\prod_{t=1}^{n+1}M_t\left(s_p^{t-1},s_q^t\,|\,W\right) \tag{19}$$

而正規化係數 Z(W)，只和觀測序列 W 有關，為長度 n+1 時所有可能之狀態序列組合:

$$Z(W)=\left(M_1(W)M_2(W)\cdots M_{n+1}(W)\right)_{\text{start,stop}}=\left[\prod_{t=1}^{n+1}M_t(W)\right]_{\text{start,stop}} \tag{20}$$

故正規化後的條件機率 p(s|W)可表示為:

$$P(s\,|\,W)=\frac{\prod_{t=1}^{n+1}M_t\left(s_p^{t-1},s_q^t\,|\,W\right)}{\mathrm{Z(W)}} \tag{21}$$

我們利用反覆(iterative)的形式，每一回合更新一次參數:

$$\lambda_k=\lambda_k+\Delta\lambda_k \tag{22}$$

$$\mu_k=\mu_k+\Delta\mu_k \tag{23}$$

其中每一回合更新的值

$$\Delta\lambda_k=\frac{1}{S}log\frac{E_{\tilde{p}(s|W,\Theta)}\left[f_k\right]}{E_{\tilde{p}(W,s)}\left[f_k\right]} \tag{24}$$

$$\Delta\mu_k=\frac{1}{S}log\frac{E_{\tilde{p}(s|W,\Theta)}\left[g_k\right]}{E_{\tilde{p}(W,s)}\left[g_k\right]} \tag{25}$$

在(24)式以及(25)式中，S 為某一訓練資料其包含之特徵函數的總數在全部訓練資料中最大的。

$$\mathrm{S} = max_s \left(\sum_t \sum_k \sum_{p,q} f_k \left(s_p^{(t-1)}, s_q^{(t)}, \mathrm{W} \right) + \sum_t \sum_k \sum_p g_k \left(s_p^{(t)}, \mathrm{W} \right) \right) \tag{26}$$

$E_{\tilde{p}(W,s)}[f_k]$ 為特徵函數 f_k 其訓練資料分佈的期望值:

$$E_{\tilde{p}(W,s)}[f_k] = \sum_{W,s} \tilde{\mathrm{P}}(\mathrm{W},s) \sum_{t=1}^{n+1} \sum_{p,q=1}^{l} f_k \left(s_p^{t-1}, s_q^t, W \right) \tag{27}$$

$E_{\tilde{p}(s|W,\Theta)}[f_k]$ 為預估測分佈的期望值:

$$E_{\tilde{p}(s|W,\Theta)}[f_k] = \sum_{W,s} \tilde{\mathrm{P}}(\mathrm{W}) P(s|W) \sum_{t=1}^{n+1} \sum_{p,q=1}^{l} f_k \left(s_p^{t-1}, s_q^t, W \right) \tag{28}$$

Lafferty 等人提出的動態規劃法(dynamic programming)即是利用 $P(s|W)$ 可表示為矩陣 $M_t(W)$ 的形式，故式子(28)可表示為:

$$\begin{aligned} E_{\tilde{p}(s|W,\Theta)}[f_k] &= \sum_{W,s} \tilde{\mathrm{P}}(\mathrm{W}) P(s|W) \sum_{t=1}^{n+1} \sum_{p,q=1}^{l} f_k \left(s_p^{t-1}, s_q^t, W \right) \\ &= \sum_{W} \tilde{\mathrm{P}}(\mathrm{W}) \sum_{t=1}^{n+1} \sum_{p,q=1}^{l} \sum_{s',s} f_k \left(s_p^{t-1} = s', s_q^t = s, W \right) \\ &\quad \times \frac{\alpha_t(s'|W) M_t(s',s|W) \beta_{t+1}(s|W)}{Z(W)} \end{aligned} \tag{29}$$

其中 $\alpha_t(W)$ 以及 $\beta_t(W)$ 為向前(forward)和向後(backward)向量，定義如下:

$$\alpha_0(s|W) = \begin{cases} 1 & \text{if s = start} \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

以及

$$\beta_{n+1}(s|W) = \begin{cases} 1 & \text{if s = stop} \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

其遞迴關係為:

$$\alpha_i(W) = \alpha_{t-1}(W) M_t(W) \tag{32}$$

和

$$\beta_i(W) = M_{t+1}(W) \beta_{t+1}(W) \tag{33}$$

與特徵函數 f_k 相似，特徵函數 g_k 其預估測分佈的期望值為:

$$E_{\tilde{p}(s|W,\Theta)}[g_k] = \sum_{W} \tilde{\mathrm{P}}(\mathrm{W}) \sum_{t=1}^{n+1} \sum_{p=1}^{l} \sum_{s} g_k \left(s_p^t = s, W \right) \times \frac{\alpha_t(s|W) \beta_t(s|W)}{Z(W)} \tag{34}$$

利用這些表示法來算出特徵函數的期望值，我們只需要使用動態規劃法計算每個可能的 $p(s_{t-1}, s_t \mid W)$ 的值，而不用計算整個模型的分布 $p(s_1, \cdots s_n \mid W)$。

4. 實驗與討論

語音辨識器乃採用 HMM Tool Kit (HTK)(http://htk.eng.cam.ac.uk/)，其中定義了 157 個次音節模型以及 11 個填空詞 (filler) 模型。對 TCC-300 (http://rocling.iis.sinica.edu.tw/ROCLING/MAT/Tcc_300brief.htm) 以及 MCDC 語料庫 (http://www.aclclp.org.tw/use_mat_c.php#mcdc)辨識率如表一，

表一，TCC-300 及 MCDC 辨識率

	Acc.	Del.	Sub.	Ins.
TCC-300 (Top 1)	89.51	0.15	9.55	0.79
TCC-300 (Top 3)	89.76	0.15	9.32	0.77
TCC-300 (Top 5)	90.38	0.15	8.82	0.65
MCDC (Top 1)	52.83	7.79	32.35	16.36
MCDC (Top 3)	53.27	7.75	32.32	16.32
MCDC (Top 5)	53.92	7.75	31.42	16.26

依據 MCDC 語料庫中，文字轉寫的標籤(tag) 以 mcdc-01、mcdc-02、mcdc-03 以及 mcdc-05 此四組為訓練語料，mcdc-09、mcdc-10、mcdc-25 以及 mcdc-26 此四組為測試語料。在全部 8 組語料中，包含不流暢語流之語句佔全部之 52.14%，也就是約 2 句對話就有一句具有不流暢語流之現象，顯示出對話時發生不流暢語流之情形在對話中十分常見。訓練語料與測試語料中發生不流暢語流情形如下表所示：

表二，語料中包含之不流暢語流現象之計數

	Repair	Repetition	Restart
訓練語料	107	318	94
測試語料	142	180	246

實驗中所用的比較對象共使用了四個方法分別是以最大熵模型(Maximum Entropy)、結合語言模型與校正模型、以詞為基礎之條件隨機域模型以及不流暢語流語言模型(DF-gram)。系統評估方面，則以 Rich04[25]中的被修正詞錯誤率(edit word error rate)以及中斷點錯誤率(edit IP error rate)作為比較的標準，如以下式子(35)與(36)所示。

$$Error_{EWD} = \frac{n_{M-EWD} + n_{FA-EWD}}{n_{EWD}} \tag{35}$$

$$Error_{IP} = \frac{n_{M-IP} + n_{FA-IP}}{n_{IP}} \tag{36}$$

表三為結合語言模型與校正模型的被修正詞錯誤率，其中在詞性 tri-gram 配合校正模型在 α=0.25 時效果為最佳，故之後和不定長度特徵之條件隨機域比較時我們皆以此為基準。

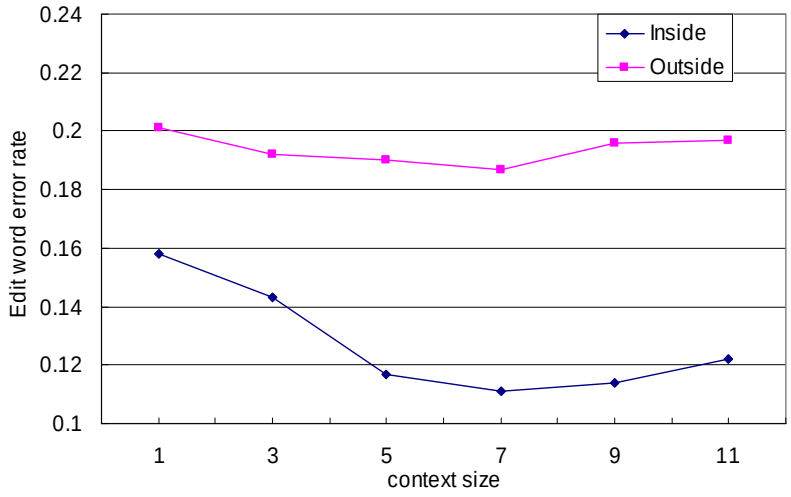
表三，結合語言模型與校正模型不同層次之被修正詞錯誤率

	Human generated transcription (REF)	Speech-to-text recognition output (STT)

	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EWD}$	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EWD}$
2-gram+ alignment	0.17	0.12	0.29	0.40	0.32	0.72
2-gram+ alignment[1]	0.09	0.15	0.24	0.36	0.32	0.68
2-gram+ alignment[2]	0.10	0.21	0.31	0.34	0.54	0.88
3-gram+ alignment	0.16	0.12	0.28	0.31	0.35	0.66
3-gram+ alignment[1]	0.09	0.12	0.21	0.32	0.32	0.64
3-gram+ alignment[2]	0.07	0.16	0.23	0.28	0.30	0.58

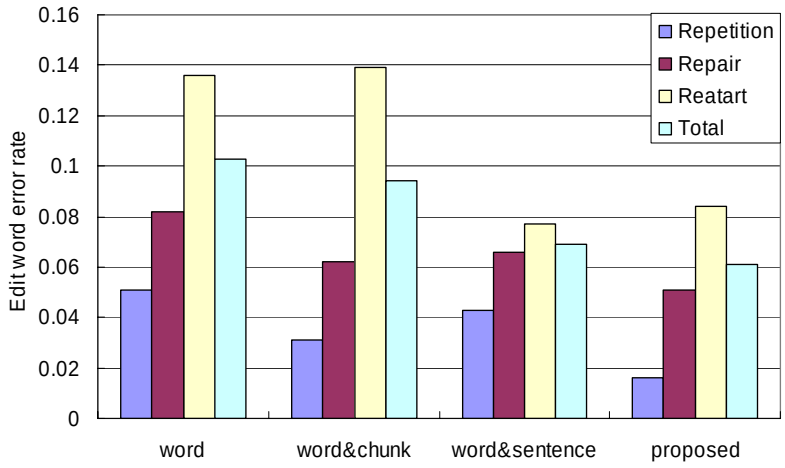
[1]: word class based on part of speech (POS) [2]: word class based on the semantic class

圖七，為使用最大熵模型，再給予中斷點情形下的被修正詞錯誤率，關於特徵函數則採用與條件隨機域相同的特徵，可以看到當 n=3 時為最佳，同樣是因為當上下文觀測範圍長度太短時，所需要的資訊不足；若是太長時，太多的特徵反而會造成混淆。故我們以此 n=3 為基準和我們提出之方法做比較。



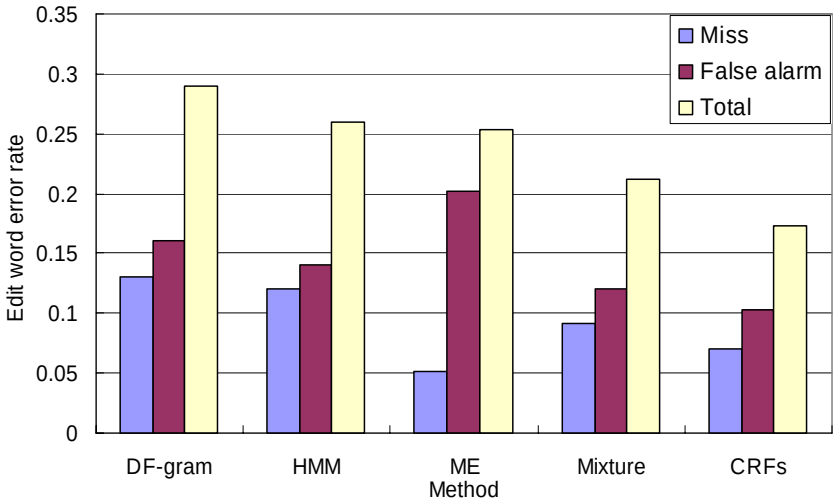
圖七，最大熵模型被修正詞錯誤率

圖八為條件隨機域分別使用不同層次所得之結果，word 代表只使用到詞的單元。word+chunk 即代表多使用了字元串集(chunk)這個單元，最後則是本論文提出的不定長度特徵之條件隨機域。我們可以看到當使用多單元時，不論二個或是三個，其錯誤率皆有降低，且對於修正(Repair)此種不流暢語流型別而言，我們可以從實驗發現使用字元串集這個層次是有幫助的；對於重開始(Restart)而言，使用句子此層次也有相當的助益。由實驗可看出我們所提出之方法對於降低被修正詞錯誤率有顯著的效果。



圖八，使用不同層次之 CRFs 在不同型別之不流暢現象被修正詞錯誤率

圖九為使用我們提出之方法與四種比較系統在文字輸入的被修正詞錯誤率，表四為使用我們提出之方法與四種比較系統在文字以及語音輸入的被修正詞錯誤率實驗證明本論文所提之模型優於其他方法。

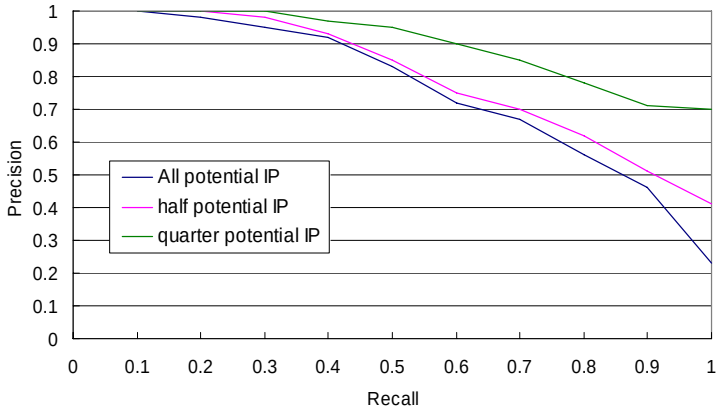


圖九，文字輸入之被修正詞錯誤率比較圖

表四，各種方法於文字與語音之被修正詞錯誤率

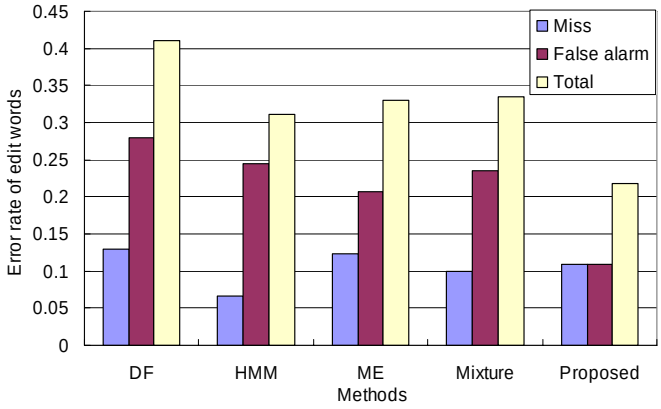
	Human generated transcription (REF)			Speech-to-text recognition output (STT)		
	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EW}$	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EWI}$
DF-gram	0.13	0.16	0.29	0.37	0.346	0.71
ME	0.05	0.20	0.25	0.14	0.52	0.66
HMM	0.12	0.14	0.26	0.34	0.35	0.68
3-gram+ alignment	0.09	0.12	0.21	0.32	0.32	0.64
Proposed	0.07	0.10	0.17	0.25	0.35	0.60

圖十為在給予不同個數的潛在中斷點下，對中斷點之求全率(Recall)與求準率(Precision)曲線。可以發現如果潛在中斷點越準確的話，對於找到中斷點的效能會越高。



圖十，求全率與求準率曲線

最後，對於複雜的口語不流暢語流也就是在語句中存在有兩個或兩個以上之不流暢現象現象做處理，其結果如圖十一所示。可以發現在複雜的口語不流暢語流情形，所提之方法仍可有效的修正不流暢語流。



圖十一，複雜不流暢語流編輯詞錯誤率比較圖

5. 結論與未來工作

本文提出一不定長度特徵之條件隨機域統計式模型修正不流暢語流，配合觀測特徵函數以及狀態轉移特徵函數找出最佳狀態序列，最後根據最佳狀態序列判斷句中某詞是否須刪除。共有三種不同之狀態單元，分別為詞、字元串集以及句子。而在觀測特徵函數的產生選取上，因人工定義過於耗時且不夠強健，故利用 Apriori 演算法產生特徵函數。由實驗中我們得到所提之方法被刪除詞錯誤率為 17.3%，相對於不流暢語流語言模型、隱藏式馬可夫模型、最大熵模型以及 N-gram 加校正之混合模型的方法分別降低了 11.7%、8.7%、8%以及 3.9%。可以發現所我們所提之方法能夠有效降低被刪除詞錯誤率。同樣的在中斷點錯誤率的實驗中也可得到相同的論證。

以下介紹未來可深入探討與研發之方向:

結合語音參數： 對於修正不流暢語流部份，我們主要抽取之參數皆為語言參數，若未來能夠將語

音參數一起考慮當作特徵函數，應可有效降低中斷點錯誤率。

解決部分辭和音節合併（Contraction）的問題：部分詞及音節合併的情形常出現於自發性口語中且對於語音辨認有一定程度之影響。

6. References

1. Byrne, W., D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J.Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Z. Wei-Jin, "AutomaticRecognition of Spontaneous Speech for Access to Multilingual Oral HistoryArchives," IEEE Trans. on Speech and Audio Processing, Vol. 12, No. 4, pp.420-435, 2004.
2. Kahn, J.G., M. Ostendorf and C. Chelba" Parsing Conversational Speech Using Enhanced Segmentation." Proc. HLT-NAACL, 2004. pp. 125-128.
3. Soltau, H., , B. Kingsbury, , L. Mangu, , D. Povey, , G. Saon, and D. Zweig, " The IBM 2004 Conversational Telephony System for Rich Transcription." In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05). (2005), 205-208.
4. Stolcke, A., E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, and Y. Lu." Automatic detection of sentence boundaries and disfluencies based on recognized words," In Proc. International Conference on Spoken Language Processing, pages 2247--2250, 1998.
5. Liu, Y., E. Shriberg, and A. Stolcke. "Automatic disfluency identification in conversational speech using multiple knowledge sources," In Proc. Eurospeech, volume 1, pages 957—960, 2003.
6. Stolcke, A. and E. Shriberg. "Statistical language modeling for speech disfluencies". In Proceedings of the International Conference of Acoustics, Speech, and Signal Processing, 1996.
7. Liu, Y., E. Shriberg, A. Stolcke, M. Harper"Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection." Eurospeech 2005.
8. Bear, J., J. Dowding, and E. Shriberg, "Integrating multiple knowledge sources for detecting and correction of repairs in human computer dialog," in Proc. of ACL, 1992, pp. 56–63.
9. Stolcke, A., W. Wang, D. Vergyri, V. R. R. Gadde, and J. Zheng, "An efficient repair procedure for quick transcriptions," in Proc. Intl. Conf. Spoken Language Processing, (Jeju, Korea), October 2004.
10. Honal, M., and T. Schultz, , "Correction of disfluencies in spontaneous speech using a noisy-channel approach," In EUROSPEECH-2003, 2781-2784.
11. Charniak, E. and M. Johnson. "Edit detection and parsing for transcribed speech," In Proceedings of the North American Chapter of the Association for Computational Linguistics annual meeting, pages 118--126, 2001.
12. Nakatani, C. and J. Hirschberg. "A corpus-based study of repair cues in spontaneous speech." Journal of the Acoustical Society of America, pages 1603--1616, 1994.
13. Snover, M., B. Dorr, and R. Schwartz. "A lexically-driven algorithm for disfluency detection". In Proceedings of Human Language Technology Conference / North American Chapter of the

Association for Computational Linguistics annual meeting, 2004.

14. Kim, J., S. E. Schwarm, and M. Ostendorf," Detecting structural metadata with decision trees and transformation-based learning." Proceedings of HLT/NAACL 2004, (2004), 137–144.
15. Furui, S., K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira"Toward the realization of spontaneous speech recognition – Introduction of a Japanese priority program and preliminary results –", Proc. ICSLP2000, Beijing.
16. Tseng, S.-C, " Repairs and Repetitions in Spontaneous Mandarin," In Proceedings of Workshop on Disfluency in Spontaneous Speech (DISS 03). Ed. Robert Eklund. Gothenburg Papers in Theoretical Linguistics 90. Pp. 71-74. University of Gothenburg.
17. Lin, Che-Kuang , Tseng, Shu-Chuan , Lee, Lin-Shan, "Important and new features with analysis for disfluency interruption point (IP) detection in spontaneous Mandarin speech", In DiSS-2005, 117-121.
18. 羅應順，自發性中文語音基本辨認系統之建立，國立交通大學電信工程所碩士論文，民國 94 年。
19. 徐文翰，自發性對話語音辨識之初步研究，國立交通大學電信工程所碩士論文，民國 93 年。
20. Wu, Chung-Hsien; Gwo-Lang Yan. "Speech act modeling and verification of spontaneous speech with disfluency in a spoken dialogue system". In Speech and Audio Processing, IEEE Transactions on Volume 13, Issue 3, May 2005 Page(s):330 – 344.
21. Yeh, Jui-Feng and Chung-Hsien Wu."Edit Disfluency Detection and Correction Using a Cleanup Language Model and an Alignment Model,"accepted by IEEE Trans. Audio, Speech, and Language Processing, 2006.
22. Chien, Jen-Tzung,"Association Pattern Language Modeling." IEEE Transactions on Audio, Speech, and Language Processing : Accepted for future publication Volume PP, Issue 99, 2005 Page(s):1-10
23. 中研院詞庫小組技術報告 93-05 中文詞類分析。
24. Lafferty, J., A. McCallum, and F. Pereira. "Conditional random fields: probabilistic models for segmenting and labeling sequence data." In ICML, 2001.
25. The EARS Fall 2004 Rich Transcription Evaluation Plan August 30, 2004。