DSP Final

2012 Spring

1. 三個 key elements 分別是 Speech recognition and understanding、Discourse analysis 跟 Dialogue management，以下分別說明

(a) Speech recognition and understanding

這個 element 負責辨識使用者發出的聲音是在說什麼（Recognition），辨識完成之後要理解這段話是在說什麼。（Understanding）

辨識的方法之前已學過很多，其中一種方法就是去 train triphone 的 HMM model，然後用這些 model 去算 observation（即使用者發出的聲音）的機率，找到機率最高的一段路徑，就得到這句話。

理解的方法也有數種不同方法，以純粹應用的角度出發，可以使用 semantic frame，定義一些 semantic class，系統在跟使用者對話時只要把 frame 填滿就好，例如飛机訂票系統就可定義 Flight 這個 semantic class

[Flight]:
- [Airline]
- [Origin]
- [Destination]
- [Date]
- [Flight No]

這樣的結構只要配合 keyword spotting 的技巧就容易實作

不過若要做更廣泛的應用，這個方法不能完善，使用 CFG 來 parse 辨識出來的結果是比較接近人類處理自然語言的方法。

(b) Discourse analysis

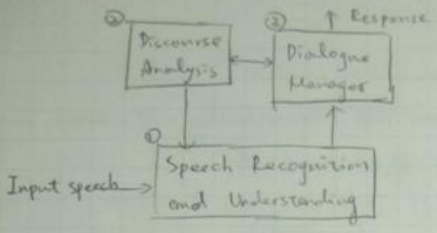
discourse 是指系統對使用者語者，或者對這段對話的理解。Discourse analysis 這個 element 是負責將 speech recognition & understanding 理解的結果整合到目前對整段對話的理解。如果沒有這塊，我們會說這個系統有瞬間失憶的問題。一種做 discourse analysis 的方法就是用 dialogue states，跟 (a) 的 semantic frames 很像，一樣有一個表格表示對話的資訊，而使用者說的話若有新的有用資訊就填入這個表格。

(c) Dialogue management

這個 element 像是整個系統的腦袋，如果缺少這塊，就會是個腦殘的系統。它主要負責產生對話跟使用者互動。

Dialogue manager 要負決定要如何跟使用者互動，也就是產生 action。

接下來說明三者如何連接，可參考此圖



使用者說的話 X_n 經由辨識之後，產生某些語意的理解 F_n，F_n 跟由再之前的理解 S_{n-1} 一起考慮得到新的理解 S_n，而由考慮 S_n 產生最恰當的回應 A_n。

以數學式來看

$$A_n^* \approx \underset{A_n, S_n}{\operatorname{argmax}}\, P(A_n|S_n) \sum_{F_n} P(S_n|F_n, S_{n-1})\, P(F_n|X_n, S_{n-1})$$

2. MLLR 是把所有的 model 裡的 Gaussian 分群，也就是把所有 μ_{jk} 分群（μ_{jk} 表示某個 model 裡第 j 個 state 的第 k 個 Gaussian 的 mean），我們可根據這些 model 的相似度（data-driven or knowledge-driven），並輔以 training data 來分群，只要有足夠 training data 量的 μ_{jk} 就可自成一群。

而 MLLR 調整的方式是 $\hat{\mu}_{jk} = A\mu_{jk} + b$，A 跟 B 是用 EM algorithm 算出來的，式子如 $[A_i, b_i] = \underset{A,b}{\operatorname{argmax}}\, \text{prob}[\bar{O}|A, A_i, b_i]$ for a class c_i，也就是說，只要一個 class 裡的某個 Gaussian 有 training data，我們就可以用這個 training data 算出這個 class 的 A_i、b_i，然後這個 class 裡的每個 μ_{jk} 都算 $A_i\mu_{jk} + b_i$ 變成新的值。

因為一群 phoneme 的 model 會一起調，所以就算沒有 adaptation data，只要它那群裡有某個 model 有 data 就可以調。

3. 在做 Voice-based information retrieval 時，通常是先做語音辨識把 query 從語音轉成文字，再去跟 document 比對（document 多為文字，若是語音形式也要先轉回文字），這個比對的過程就可選擇用 word、subword 或 keyword 當做 indexing feature。接下來分別說明三者的優缺點

① word：比對 query 跟 document 出現的 word

優點：word 的語意比較明確，比較不會造成 ambiguity。

缺點：(a) 中文的結構複雜難以處理，例如台灣大學→台大；支那→核那。

(b) document 跟 query 轉成文字時可能有錯誤，有 error propagation。

(c) 有 OOV 問題。

(d) 字跟字的 segmentation 不易，例如 腦科→電腦科學

② Subword unit：用 phone 或 syllable 當作搜尋單位，例如英語先把某可用音節字等搜尋。

優點：(a) 所需的 database 小，word 往往有效篇個，但 phone 或 syllable 只要幾百幾千個而已。

(b) 解決 OOV 問題，因為不管什麼新詞，都可由 subword unit 組成。

缺點：有 ambiguity 問題，例如

③ keyword：每篇 document 都有一些 keyword，根據這些 keyword 來 index document 只要從 query 中取出 keyword，跟所有 document 取出的 keyword 比對取出有相同 keyword 的 document 即可。

從 document 中取得 keyword 的方法有兩種，一種是人工取，另一種是自動取，但人工取成本太高時，自動取又很困難，故如何從 document 中取得關鍵字是這方法的瓶頸。

5. (a) 每種 index feature type 用 j 表示，在做 information retrieval 前先為每個 d 建一個 vector，假設 index feature $j=1$ 是 word，vector 如下

$$\vec{d}_j = \begin{matrix} w_1 \\ w_2 \\ \vdots \\ w_t \\ \vdots \\ w_L \end{matrix}\begin{bmatrix} \\ \\ \\ z_{jt} \\ \\ \\ \end{bmatrix}_{j=1}, \quad \text{其中 } z_{jt} = \underbrace{(1+\ln[c_t])}_{w_t \text{ 在 } d \text{ 中出現次數}} \cdot \ln\left(\frac{N}{N_t}\right) \leftarrow N \text{ 是全部的文章數}, \ \leftarrow \text{含 term } t \ (w_t) \text{ 的文章數}$$

z_{jt} 表示 w_t 的重要性

這個 vector 的意義在於可為每個 term 定出一個分數，如此一來就可以知道 d 跟哪些 term 較相關

(c) 當使用者產生 query 時，亦把 query 用相同的方法轉成 vector ($\vec{g}_j$)，然後計算

$$\arg\max_{d_i} R(g, d_i) = \arg\max_{d_i} \sum_j w_j \cdot R_j(\vec{g}_j, \vec{d}_{ij}), \text{ where } R_j(\vec{g}_j, \vec{d}_{ij}) = \frac{\vec{g}_j \cdot \vec{d}_{ij}}{|\vec{g}_j||\vec{d}_{ij}|}$$

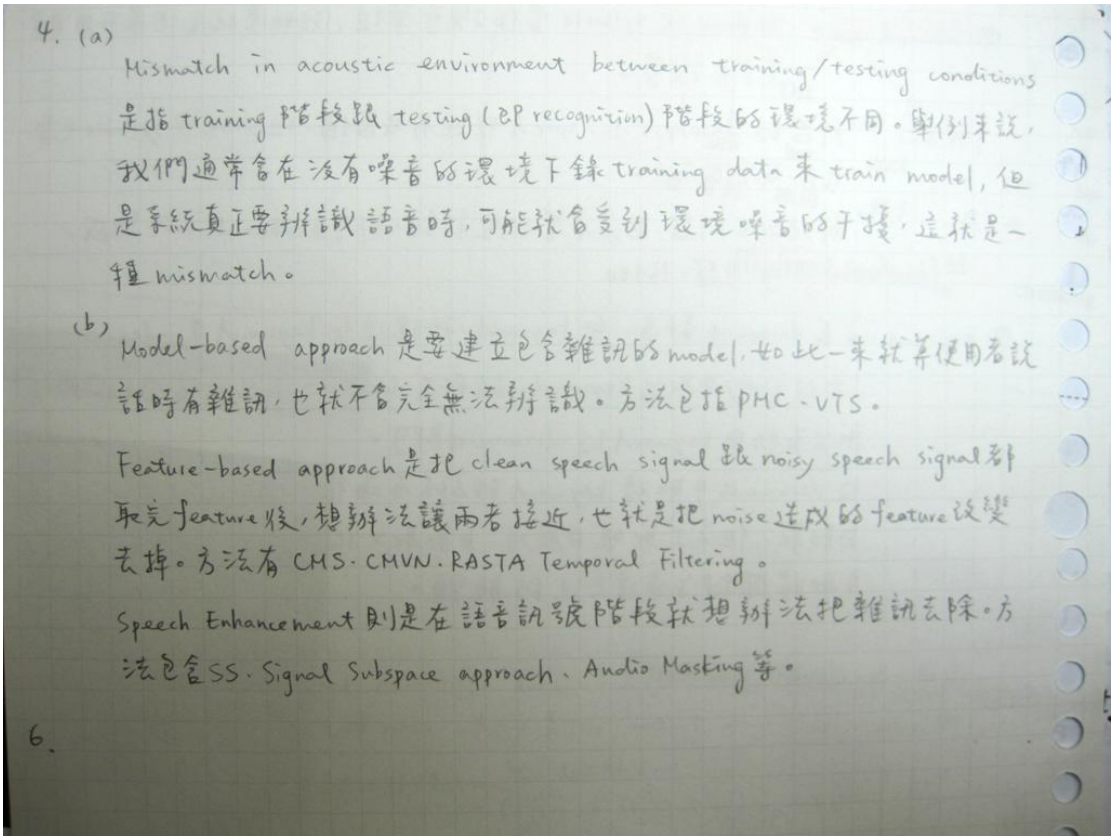
d_i 指第 i 個 document

(b) Term frequency $= 1+\ln[c_t]$，c_t 是某個 term 在 document 中出現次數，因為 c_t 可以很大，故取 ln 把值壓下來，而加 1 是為了讓 $c_t=1$ 時此項不為 0。

($1+\ln[c_t]$ 只是一種做法，可以自己調整)

Inverse document frequency $= \ln\left(\frac{N}{N_t}\right)$，N 是全部文章數目，$N_t$ 是包含 term t 的文章數目，IDF 可反映一個 term 的重要性，越少出現的 term 的 IDF 越高。

B98901160 電機四 賴柏任



4. (a)

Mismatch in acoustic environment between training/testing conditions 是指 training 階段跟 testing (ASR recognition) 階段的環境不同。舉例來說，我們通常會在沒有噪音的環境下錄 training data 來 train model，但是系統真正要辨識語音時，可能就會受到環境噪音的干擾，這就是一種 mismatch。

(b)

Model-based approach 是要建立包含雜訊的 model，如此一來就算使用者說話時有雜訊，也就不會完全無法辨識。方法包括 PMC、VTS。

Feature-based approach 是把 clean speech signal 跟 noisy speech signal 都取完 feature 後，想辦法讓兩者接近，也就是把 noise 造成的 feature 改變去掉。方法有 CMS、CMVN、RASTA Temporal Filtering。

Speech Enhancement 則是在語音訊號階段就想辦法把雜訊去除。方法包含 SS、Signal Subspace approach、Audio Masking 等。

6.