第5章 語言模型應用於語音文件摘要

具時序性的語音文件(Spoken Documents)，例如廣播新聞、演講錄音等等，不易直接瀏覽，如果透過文字的呈現比較容易進行查詢。這個需求，可以透過語音辨識技術解決。整段語音的對應辨識文字可能很冗長，如果想要快速地了解語音的主題，若有簡短的描述將會一目了然。而這樣的需求，我們可以透過文件摘要技術解決。在語音文件摘要過程中，我們擁有許多語言或語音的資訊，如何充分利用這些資訊是一個研究的重點。本章主要是透過對摘要過程建立機率生成架構，並將語言模型技術應用於其架構[Chen *et al.* 2007]。

5.1 語音文件摘要介紹

文件摘要(Document Summarization)方式大致可分為兩類，摘錄式(Extractive)與非摘錄式(Non-Extractive or Abstractive)摘要。摘錄式摘要透過設定欲呈現的摘要比例(Summarization Ratio)，直接從文件中抽取重要的詞、片語、語句或段落來組成摘要結果，可能產生不通順的語句；非摘錄式摘要是依據文件主題直接重寫摘要，需要考慮許多自然語言資訊，例如語意表示和文法限制等。目前也有學者是採用語句抽取(Sentence Extraction)加上語句壓縮(Sentence Compaction)的方法達到重寫的目的[Kikuchi *et al.* 2003]。由於非摘錄式摘要仍有一定的難度，所以現階段的自動文件摘要的相關研究多以摘錄式摘要為主，本章亦是著重於以語句為單位的摘錄式摘要方法之探討。

語音文件摘要(Spoken Document Summarization)與一般文件摘要差異在於語音文件需要先經過語音辨識，得到自動轉寫文件(Automatic Transcription)後，再使用文件摘要技術進行摘要。圖 5-1 為語音文件摘要流程圖。一般而言，我們會使用文字語料(Text Corpus)統計資訊來輔助決定自動轉寫文件的詞彙與語意資訊

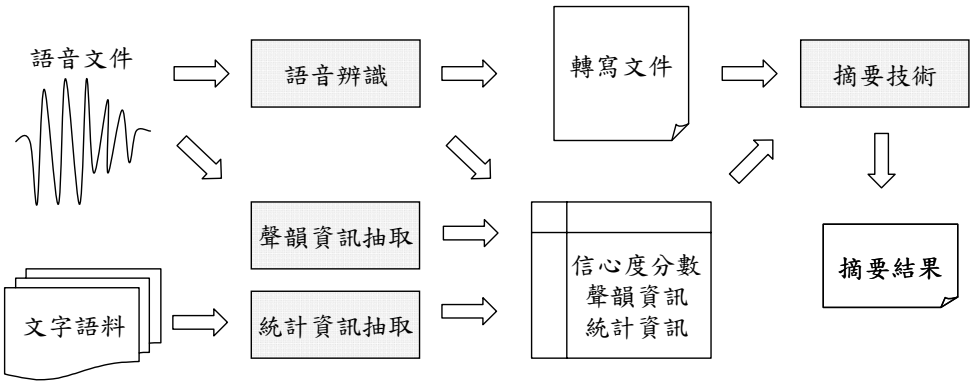

圖 5-1 語音文件摘要流程圖

是否合理。除此之外，因為辨識結果可能有錯誤，直接使用轉寫文件的文字結果會有些問題，我們可以使用辨識過程中計算的信心度分數(Confidence Score)判斷語句的正確性。語音文件的另一個特點是除了轉寫文件的文字資訊外，我們亦可使用語音本身的聲韻資訊(Prosodic Information)，例如音高(Pitch)、停頓(Break)、持續時間(Duration)等來輔助文件摘要技術，例如詞的能量(Energy)可能暗示著某種重要性。

摘錄式摘要方法有許多種，例如以文件結構為基礎的摘錄方法(Document Structure-based Approach)，依據詞或語句所在的位置決定其重要性。比如新聞語料中，第一句或是最後一句可能是最重要的，所以選擇其為摘要的話可能會有不錯的結果。這類方法簡單且直覺，但前提是文件需要具有一定結構性才適用。或是以統計值為基礎的摘錄方法(Statistics-based Approach)，其使用統計資訊來決定語句的重要性，例如詞頻數(Term Frequency, TF)、反文件頻數(Inverse Document Frequency, IDF)、語音辨識信心度分數、聲韻資訊等。然後再透過如向量空間模型(Vector Space Model, VSM)、潛藏語意分析(Latent Semantic Analysis, LSA)模型或各種分類器(Classifier-based Approach)的使用來進行摘要，關於模型的詳細說明可參考[陳怡婷 2006]。除此之外，我們亦可以採用機率生成架構(Probabilisitc Generative Framework)進行摘要[Chen *et al.* 2007]。機率生成架構主要包含了語句生成模型(Sentence Generative Model)與語句事前機率模型(Sentence Prior Model)兩部分，將於下一節介紹。

5.2 機率生成架構

在摘錄式摘要的機率生成架構(Probabilistic Generative Framework)中，代表文件 D 主題的重要語句 S_i 可以透過其事後機率 $P(S_i \mid D)$ 排名選出：

$$P(S_i \mid D) = \frac{P(D \mid S_i)P(S_i)}{P(D)} \tag{5-1}$$

$P(S_i \mid D)$ 是給定文件 D，語句 S_i 的事後機率。在這邊我們不直接對文件 D 建立模型，而透過貝氏定理轉換成三個機率 $P(D \mid S_i)$、$P(S_i)$ 與 $P(D)$。$P(D \mid S_i)$ 是語句 S_i 產生文件 D 的機率，$P(S_i)$ 是語句 S_i 的事前機率，$P(D)$ 是文件 D 的事前機率。由於 $P(D)$ 不會影響排名結果，故在此可以忽略。所以我們要探討的是估測文件機率 $P(D \mid S_i)$ 的語句生成模型(Sentence Generative Model)與估測語句機率 $P(S_i)$ 的語句事前機率模型(Sentence Prior Model)。

5.2.1 語句生成模型

語句生成模型方面，我們首先假設文件 D 中的詞 w 是獨立的，所以文件機率可以表示成文件中的詞機率的連乘積：

$$P(D \mid S_i) = \prod_{w \in D} P(w \mid S_i)^{n(w,D)} \tag{5-2}$$

$P(w \mid S_i)$ 是給定語句 S_i 產生詞 w 的機率，$n(w, D)$ 是詞 w 在文件 D 裡出現的次數。過去已經有許多語句生成模型被提出，例如使用逐字比對(Literal Term Matching)的隱藏式馬可夫模型(Hidden Markov Model, HMM)[Chen et al. 2006]。逐字比對指的是使用文件中明確的詞比對[Lee and Chen 2005]：

$$P_{HMM}(D \mid S_i) = \prod_{w \in D} \left[\lambda P(w \mid S_i) + (1-\lambda)P(w \mid C)\right]^{n(w,D)} \tag{5-3}$$

$P(w \mid S_i)$ 是給定語句 S_i，詞 w 的機率，$P(w \mid C)$ 是從大量文字語料 C 估測詞 w 的機率，用以平滑化 $P(w \mid S_i)$：

$$P(w \mid S_i) = \frac{n(w, S_i)}{\sum_{w'} n(w', S_i)} \tag{5-4}$$

$$P(w \mid C) = \frac{n(w, C)}{\sum_{w'} n(w', C)} \tag{5-5}$$

$n(w, S_i)$是詞w在語句S_i裡出現的次數，$n(w, C)$是詞w在語料C裡出現的次數。λ用來調整詞w在語句S_i或背景語料C機率的比例，可用期望值最大化法(Exceptation Maximumm, EM)估測[Dempster *et al.* 1977]。

由於語句S_i的長度通常很短，估測出來的詞機率$P(w \mid S_i)$可能不太準確，我們可以使用關聯性模型(Relevance Model, RM)來輔助估測詞機率[Croft and Lafferty 2003; Smucker et al. 2005; Chen et al. 2006b]。在摘錄式摘要中，每一語句S_i都有其所屬的相關文件集(Relevant Document Set)R_{S_i}，且語句S_i的關聯性模型RM_{S_i}可定義成從相關文件集R_{S_i}隨機地選擇文件D_l，並且從文件D_l隨機地選擇出詞w的機率$P(w \mid RM_{S_i})$。由於語句S_i的相關文件集R_{S_i}不易取得，所以我們採用局部性回饋(Local Feedback)的概念[Baeza-Yates and Ribeiro-Neto 1999]，以語句S_i當作查詢(Query)，透過資訊檢索系統找出其相關文件集合R_{S_i}，所以關聯性模型$P(w \mid RM_{S_i})$可表示成：

$$P(w \mid RM_{S_i}) \approx \sum_{l=1}^{L_i} P(D_l \mid R_{S_i}) P(w \mid D_l) \tag{5-6}$$

L_i是文件S_i的相關文件數量，可限制只使用前L_i篇相關文件，實驗中設定$L_i = 5$；$P(w \mid D_l)$是給定文件D_l，詞w的機率，且

$$P(w \mid D_l) = \frac{n(w, D_l)}{\sum_{w} n(w, D_l)} \tag{5-7}$$

$n(w, D_l)$是詞w在文件D_l裡出現的次數。$P(D_l \mid R_{S_i})$是相關文件集R_{S_i}產生文件D_l

的機率，且

$$P(D_l \mid R_{S_i}) \approx \frac{P(D_l)P(S_i \mid D_l)}{\sum_{r=1}^{L_i} P(D_r)P(S_i \mid D_r)} \tag{5-8}$$

因為我們只使用前L篇相關文件，所以是一個近似的結果，其中$P(D_l)$可設定為平均分布(Uniformly Distributed)或是根據文件於檢索系統的相關排名估測：

$$P(D_l) = \frac{(1/l)}{\sum_{r=1}^{L_i}(1/r)} \tag{5-9}$$

l是文件D_l在相關文件集合R_{S_i}的排名，排名越高，機率越大。此外，$P(S_i \mid D_l)$可使用隱藏式馬可夫模型估測：

$$P_{HMM}(S_i \mid D_l) = \prod_{w \in S_i} [\lambda P(w \mid D_l) + (1-\lambda)P(w \mid C)]^{n(w,S_i)} \tag{5-10}$$

與式 (5-3)不同，式 (5-10)是針對檢索系統的文件D_l建立隱藏式馬可夫模型。有了$P(w \mid RM_{S_i})$之後，我們可以結合隱藏式馬可夫模型與關聯性模型：

$$P_{HMM+RM}(D \mid S_i) = \prod_{w \in D} \begin{bmatrix} \lambda(\gamma P(w \mid S_i) + (1-\gamma)P(w \mid RM_{S_i})) \\ +(1-\lambda)P(w \mid C) \end{bmatrix}^{n(w,D)} \tag{5-11}$$

γ是隱藏式馬可夫模型中的$P(w \mid S_i)$與關聯性模型之間的比重。

除了逐字比對，我們亦可使用概念比對(Concept Matching)的詞主題混合模型(WTMM)[Chen and Chen 2007]作為語句生成模型。概念比對指文件與語句在概念上相似，而所使用的詞不一定相同[Lee and Chen 2005]：

$$P_{WTMM}(D \mid S_i) = \prod_{w \in D} \left[\sum_{w_j \in S_i} \alpha_{j,i} P(w \mid M_{w_j}) \right]^{n(w,D)} \tag{5-12}$$

$\alpha_{j,i}$是詞w_j在句子S_i中所佔比例：

$$\alpha_{j,i} = \frac{n(w_j, S_i)}{\sum_w n(w, S_i)} \tag{5-13}$$

$n(w_j, S_i)$是詞w_j在句子S_i出現的次數。$P(w \mid M_{w_j})$是詞主題混合模型M_{w_j}產生詞w的機率：

$$P(w \mid M_{w_j}) = \sum_{k=1}^{K} P(w \mid T_k) P(T_k \mid M_{w_j}) \tag{5-14}$$

K是主題個數，$P(w \mid T_k)$是詞w發生於主題T_k的機率，$P(T_k \mid M_{w_j})$是給定詞w_j，主題T_k發生的機率，模型可使用期望值最大法訓練，詳細訓練方式可參考 第 4 章 詞主題混合模型與位置相關語言模型。

5.2.2 語句事前機率模型

對於語句而言，其事前機率的估測仍舊是一個尚未解決的問題。過去使用機率生成架構時多是假設語句事前機率為平均分布(Uniformly Distributed)。然而，在語音文件中，會被摘要出的語句必有其重要性，所以語句不應該是相同的機率分布，而可能跟許多資訊有關，例如語句在文件中的位置、語句在語言中的合理性、語音辨識率或是語句裡的聲韻資訊等。我們擁有這些資訊，但是它們之間的關聯我們並不清楚，為了能夠整合這些特徵而不需要額外的資訊，我們採用最大熵值(Maximum Entropy)的方式結合。傳統的條件最大熵值模型(Conditional Maximum Entropy Model)所採用的特徵定義的是歷史詞序列與目前詞的關係，而由於我們想要估測的是語句事前機率$P(S_i)$，所以採用完整語句特徵的整句最大熵值(Whole Sentence Maximum Entropy, WSME)模型較能符合我們的需求[Rosenfeld *et al.* 2001]。使用整句最大熵值模型估測的語句機率$P(S_i)$可以用指數型(Exponential Form)表示：

$$P(S_i)=\frac{1}{Z}P_0(S_i)\exp\left(\sum_j \lambda_j f_j(S_i)\right) \tag{5-15}$$

$P_0(S_i)$是語句S_i的任意初始機率，$f_j(S_i)$是事先定義好的第j種語句特徵，λ_j是對應語句特徵$f_j(S_i)$的權重值，Z是正規化常數(Normalization Constant)：

$$Z=\sum_{S_i} P_0(S_i)\exp\left(\sum_j \lambda_j f_j(S_i)\right) \tag{5-16}$$

整句最大熵值模型訓練時期望使得模型機率分布$P(S_i)$與初始機率分布$P_0(S_i)$的KL距離(Kullback–Leibler Divergence)最小，

$$D(P(S_i)\|P_0(S_i))=\sum_{S_i} P(S_i)\log\frac{P(S_i)}{P_0(S_i)} \tag{5-17}$$

且所有特徵滿足限制(Constraint)：

$$E_p[f_j]=K_j \tag{5-18}$$

$E_P[f_j]$為語句特徵f_j在整句最大熵值模型$P(S_i)$的期望值，K_j是語句特徵f_j在經驗分布(Empirical Distribution)的期望值。關於整句最大熵值模型與KL距離的關係，可參考 附錄B 整句最大熵值模型。我們可使用迭代方式求解，如改善迭代調整法(Improved Iterative Scaling, IIS)[Berger 1997]或是廣義迭代調整法(Generalized Iterative Scaling, GIS)[Darroch and Ratcliff 1972]。關於迭代調整法詳細說明可參考[蔡文鴻 2005]。

語音語句可使用的特徵有很多種，例如在文字上有*N*連詞特徵、*N*連類別特徵等；語句上有觸發對特徵、語句長度特徵等；其對應語音上則有辨識信心度特徵、聲韻資訊特徵等；在文件中則有語句位置特徵等。在[Chan and Togneri 2006]中，其使用的語料包含了正確的詞性標籤(Part-of-Speech, POS)及聲韻標記

表 5-1 語音文件摘要採用的語句特徵

特徵	描述	關係	類型
F1	語句位置倒數	文件-語句	語言
F2	二連詞機率平均	語句-詞	語言
F3	詞辨識信心度平均	語句-詞	語音
F4	詞音高平均	語句-詞	語音
F5	詞能量平均	語句-詞	語音

(Prosodic Label)，如聲調(Accent)、停頓(Break)等，Chan的作法是採用以語句中的詞為單位的單連及二連特徵，例如詞與詞性配對(Word-POS Pair)之單連及二連特徵、詞性與聲調配對(POS-Accent Pair)之單連及二連特徵等。

而我們的作法是選擇了五種特徵：第一種特徵(F1)是語句在文件中的位置(Location)，一般而言，文件的前幾句話的重要性可能較高，也較能代表整個文件的語意。第二種特徵(F2)是二連詞語言模型分數(Bigram Language Model Score)，透過語言模型的分數，我們能夠決定語句在語言中的合理性。第三種特徵(F3)是辨識信心度(Confidence Score)，我們使用辨識系統計算出的事後機率當作信心度。第四(F4)與第五種(F5)特徵是語音的聲韻資訊，我們分別抽取語音的音高(Pitch)與能量(Energy)值，加以使用，如 表 5-1 所示。位置特徵是考慮語句的位置，其分數估測方式為語句位置除以文件包含的語句位置和的倒數，所以能表示語句與文件的關係：

$$F(S_i)=\frac{\sum_{S_i\in D}L_D(S_i)}{L_D(S_i)} \tag{5-19}$$

$L_D(S_i)$表示語句S_i在文件D裡的位置。其餘特徵則是語句中的每一個詞都會有對應的分數，如語言模型的詞機率、辨識系統的詞信心度分數、語音中的詞音高值與詞能量值等，表示語句本身的特性。我們初步地使用語句S裡每個詞分數的算術平均當作語句的特徵：

$$F(S_i) = \frac{\sum_{w \in S_i} Score(w, S_i)}{|S_i|} \tag{5-20}$$

$|S_i|$是語句的長度，$Score(w, S_i)$是語句S_i中的詞w的特徵分數。此外，位置與詞機率都是語言特性上的特徵，而信心度、音高與能量都是屬語音特性上的特徵。有別於Chan的作法是每一句訓練語句的特徵數量不同，我們的作法則是每一語句使用固定五種特徵[Chen et al. 2007]。定義了語句特徵之後，我們可以訓練最大熵值模型，即特徵對應的參數λ。

5.3 摘要實驗設定與結果

5.3.1 摘要實驗語料

摘要測試語料蒐集自News 98 新聞網[News 98]，包含 2001 年 8 月 1 日至 8 月 24 日中午 12:00~13:00 的FM廣播新聞，共 200 則，並有對應的正確人工轉寫文件，詳細內容如 表 5-2 所示。新聞語音經過大詞彙連續語音系統辨識後，辨識字錯誤率為 14.17%，再經過斷句處理產生自動轉寫文件。測試語料的自動摘要評估標準答案部分，由三位國立台灣大學文學院大三以上的學生，分別對此 200 則廣播新聞的人工轉寫文件產生人工標註摘要，摘要的結果可分為依語句重要性排名的句排名形式與依特定比例重寫的摘要兩種，我們使用句排名結果當作參考文件，圖 5-2 為一則廣播新聞人工標註摘要範例[Ho 2003]。我們進一步將 200 則新聞等分成發展集(Development Set)與評估集(Evaluation Set)各 100 則。發展集用於調整模型所需的參數，於發展集將參數調整到最佳後，再將此參數設定用於評估集。機率生成模型或是語句事前機率模型所使用的語言模型的訓練文字語料皆來自中央社(CNA)新聞[LDC]。訓練隱藏式馬可夫模型(HMM)與關聯性模型(RM)的語料為 2001 年八月；其中隱藏式馬可夫模型所需的背景語言模型(BLM)機率 $P(w|C)$ 的訓練語料為 2000 年至 2001 年；詞主題混合模型(WTMM)的訓練語料則來自 2002 年八月至十月。訓練語料詳細資訊如 表 5-3 所示。

表 5-2 摘要測試語料之統計資訊

新聞時間	2001 年 8 月 1 日~2001 年 8 月 24 日
新聞則數	200 則
總語音長度	1.61 小時
平均每則新聞語音長度	28.96 秒
總字數/總詞數(人工轉寫)	28,235 字/16,424 詞
平均每則字數/詞數(人工轉寫)	約 141 字/約 82 詞

表 5-3 摘要訓練語料之統計資訊

模型	BLM	HMM, RM	WTMM
時間	2000 年~2001 年	2001 年八月	2002 年八月~十月
新聞則數	322,295	14,178	39,027
總字數	155,272,922 字	4,447,396 字	17,686,770 字
總詞數	90,924,247 詞	2,756,723 詞	10,670,436 詞
平均每則長度(字)	約 481 字	約 313 字	約 453 字
平均每則長度(詞)	約 282 詞	約 194 詞	約 273 詞

編號：[1]N200108011200-01
請將下列新聞中的每一句依重要性排名，1-最重要、2-次重要、依此類推（排名請用阿拉伯數字）。
（1）桃芝颱風重創花蓮
（2）光復鄉大興村死傷慘重
（4）感觸最多的是在當地送信長達十七年的郵差鄭順發
（7）村子裡頭平常天天見面打招呼的老朋友
（3）一夕之間天人永隔
（5）災後頭一天送信到大興村
（6）鄭順發的心情已經不是複雜兩個字能夠形容

請為本則新聞重寫一個約 34 字左右的摘要：
桃芝風災造成光復鄉大興村的嚴重死傷 老郵差災後送信時感觸良多

圖 5-2 人工轉寫文件摘要範例

5.3.2 實驗評估

摘要的評估方式可分為主觀與客觀評估。主觀評估是由評估者根據自己的想法，如摘要是否表達出文件重點、摘要是否流暢等來決定摘要好壞。客觀評估，多是以數學方式計算結果，例如餘弦測量(Cosine Measure)等。本論文是採用ROUGE(Recall-Oriented Understudy for Gisting Evaluation)評估摘要結果[Lin 2004]。ROUGE是一種召回導向(Recall-Oriented)的主旨(Gist)評估，主要是透過計算自動摘要結果與人工摘要結果之間的重疊單元(Overlapping Unit)個數，如*N*

連詞或是詞序列，來評估摘要的品質。例如ROUGE-*N*是一種*N*連詞召回測量(*N*-gram Recall Measure)。一篇自動摘要文件 S 的ROUGE-*N*評估可以定義成：

$$ROUGE-N=\frac{\sum_{M\in\mathbf{M}_R}\sum_{gram_n\in M}Count_{Match}(gram_n)}{\sum_{M\in\mathbf{M}_R}\sum_{gram_n\in M}Count(gram_n)} \tag{5-21}$$

N 表示使用*N*連詞為重疊單元，$\mathbf{M}_R$ 是文件 S 對應的參考摘要文件集，例如不同專業人員所標註的參考摘要文件，M 是某一人員針對文件 S 標註的參考摘要文件，$Count_{Match}(gram_n)$ 是文件 S 與參考摘要文件 M 最多共同出現的*N*連詞個數，$Count(gram_n)$ 是參考摘要文件 M 出現的*N*連詞個數。所以，對於某個*N*連詞而言，$Count_{Match}(gram_n)\le Count(gram_n)$。本論文採用 $N=2$ 的二連詞為重疊計算單元的ROUGE-2 評估，圖 5-3 是使用ROUGE-2 的文件摘要範例。得到每一篇自動摘要文件的ROUGE評估值之後，再取平均值得到最後的ROUGE結果。除了ROUGE-*N*之外，ROUGE評估亦包含了使用最長共同子序列(Longest Common Subsequence)為重疊單元的ROUGE-L、或是使用略二連詞(Skip Bigram)為重疊單元的ROUGE-S等。

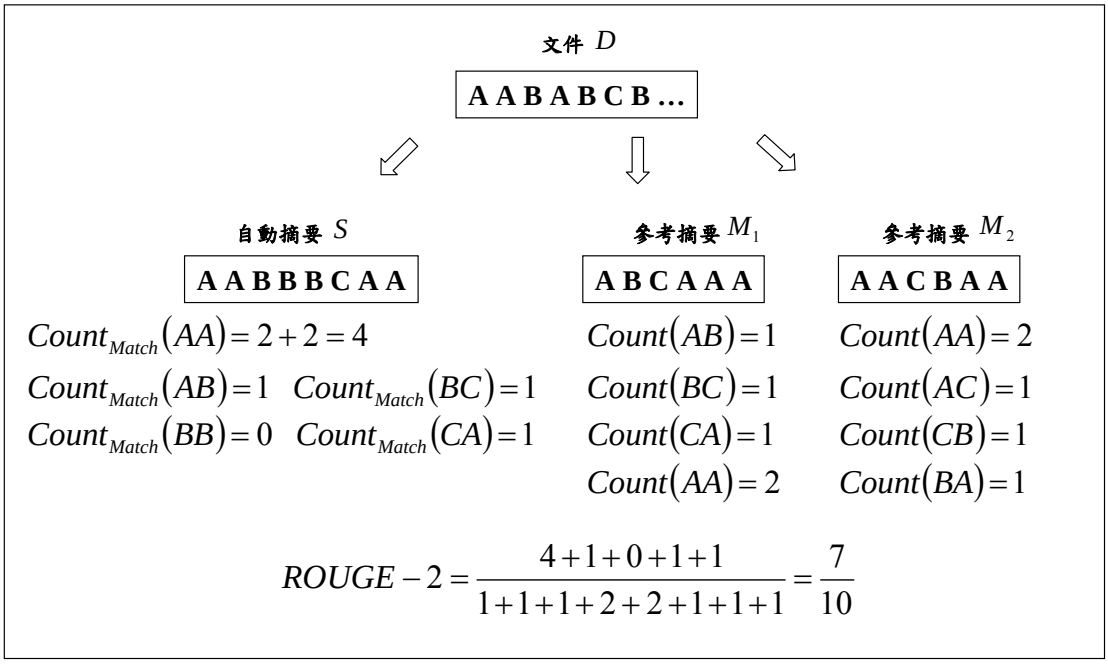

圖 5-3 使用 ROUGE-2 的文件摘要評估範例

5.3.3 摘要實驗結果

我們於本節中呈現機率生成架構中，使用不同語句生成模型及配合不同事前機率模型的實驗結果。

5.3.3.1 語句生成模型實驗結果

於機率生成架構中，我們首先假設事前機率的機率分布為平均分布，也就是僅使用語句生成模型。表 5-4 與 表 5-5 是使用不同模型及不同摘要比例下的實驗結果，如向量空間模型(VSM)、隱藏式馬可夫模型(HMM)、隱藏式馬可夫模型結合關聯性模型(HMM-RM)與詞主題混合模型(WTMM)等。向量空間模型(VSM)是使用詞頻數(TF)與反文件頻數(IDF)建立語句向量與文件向量，再使用餘弦測量(Cosine Measure)估測兩種向量的相似度，再根據相似度進行排名。於結果中我們觀察到，生成模型如隱藏式馬可夫模型(HMM)，在大部分的摘要比例下，的確能夠比傳統的向量空間模型(VSM)來的好。此外，結合了關聯性模型的隱藏式馬可夫模型能夠表現更好，因為關聯性模型能夠輔助解決原先語句生成模型因為

表 5-4 基本模型與語句生成模型於發展集結果

摘要比例	VSM	HMM	HMM-RM	WTMM
10%	0.2653	0.3084	0.3369	0.3245
20%	0.3103	0.3467	0.3757	0.3522
30%	0.3331	0.3734	0.3725	0.3509
50%	0.4436	0.4768	0.4779	0.4577

表 5-5 基本模型與語句生成模型於評估集結果

摘要比例	VSM	HMM	HMM-RM	WTMM
10%	0.3073	0.2932	0.3182	0.3248
20%	0.3188	0.3191	0.3264	0.3324
30%	0.3593	0.3705	0.3671	0.3816
50%	0.4485	0.4732	0.4774	0.4581

語句長度太短而語句模型估測不可靠的問題，實驗中使用檢索系統輸出的前 5 篇文件當作相關文件集。使用詞主題混合模型(WTMM)亦有改善，因為其使用了隱藏的主題來表示詞與詞的關係，透過詞模型的結合，進一步表示由詞組成的語句及文件的關係，實驗中採用 2 個隱藏主題數建立詞關聯。

5.3.3.2 語句事前機率模型實驗結果

首先我們針對不同的語句特徵估測事前機率。由於語句特徵分數並非機率，所以我們先以文件為單位，將語句特徵值進行正規化(Normalization)的動作，使得每一摘要文件裡所有語句的某特徵值和為 1，並當作語句 S_i 的事前機率 $P(S_i)$：

$$P(S_i)=\frac{Score(S_i)}{\sum_{S_j\in D} Score(S_j)} \tag{5-22}$$

$Score(S_i)$ 是語句 S_i 的某特徵值分數。此外，語句生成模型所產生的文件機率 $P(D\mid S_i)$ 為詞機率的連乘積，與語句事前機率 $P(S_i)$ 的值域範圍有差距，所以我們進一步透過權重值 α 調整：

$$P(S_i\mid D)=\frac{P(D\mid S_i)^{\alpha}P(S_i)}{P(D)} \tag{5-23}$$

我們先於發展集中調整權重值 α，達到最佳之後，再將 α 用於評估集。

我們首先嘗試僅使用語句特徵分數的機率分布進行摘要實驗，即不使用語句生成模型。實驗結果如 表 5-6 與 表 5-7 所示。我們很明顯地發現，使用位置特徵的結果相當地好，這是因為語句位置特徵代表語句於文件的位置，能夠直接表示語句與文件的關係，例如第一句可能是重要的語句，而這種關係與新聞的呈現方式一致，所以相當有效。其他特徵是由語句中的個別詞分數平均所估測，只能表現語句於語言或語音的關係，而沒有考慮到語句與文件的關係，所以效果不甚理想。

表 5-6 使用語句特徵分布於發展集結果

摘要比例	F1	F2	F3	F4	F5
10%	0.4028	0.0654	0.0521	0.0940	0.1581
20%	0.4085	0.0942	0.0625	0.1389	0.1871
30%	0.3653	0.0995	0.0896	0.1802	0.2342
50%	0.4354	0.2621	0.2965	0.3493	0.3725

表 5-7 使用語句特徵分布於評估集結果

摘要比例	F1	F2	F3	F4	F5
10%	0.4502	0.0957	0.0489	0.1478	0.1954
20%	0.4381	0.1072	0.0545	0.1575	0.2178
30%	0.4472	0.1337	0.0734	0.2304	0.2515
50%	0.4478	0.2618	0.2793	0.3741	0.3778

5.3.3.3 機率生成架構實驗結果

接著我們整合語句生成模型與語句事前機率模型，表 5-10、表 5-11、表 5-10 與表 5-11 分別是以結合關聯性模型之隱藏式馬可夫模型與詞主題混合模型為語句生成模型，並搭配不同的語句特徵分數正規化的語句事前機率分布的實驗結果。根據此結果，我們可以觀察到，使用位置特徵(F1)機率分布與語句生成模型的結合，能夠進一步增加摘要的效果，如 表 5-10 使用HMM-RM+F1 可達到48.64%的正確率，相較於單獨使用HMM-RM的 31.82%及僅使用F1 特徵的45.02%，相對改善為 52.85%及 8.04%。此外，使用結合關聯性模型之隱藏式馬可夫模型(HMM-RM)為語句生成模型時，以能量特徵(F5)為事前機率分布在摘要比例 10%時能有些許改善；使用詞主題混合模型(WTMM)為語句模型時，以辨識信心度(F3)與能量特徵(F5)為事前機率在摘要比例 10%與 20%時能有些許改善。我們認為，能量特徵(F5)表示語句的強調，能夠表現語句的重要性。然而，使用二連詞語言模型分數特徵(F2)則無明顯改善，我們認為原因可能是生成模型產生的文件機率，亦是語言模型分數，而兩者在此似乎無加成性的效果。

表 5-8 整合關聯性模型之隱藏式馬可夫模型結合語句特徵機率分布於發展集結果

摘要比例	HMM-RM	F1	F2	F3	F4	F5
10%	0.3369	0.4370	0.3375	0.3355	0.3465	0.3468
20%	0.3757	0.4579	0.3670	0.3737	0.3851	0.3833
30%	0.3725	0.4237	0.3730	0.3734	0.3854	0.3762
50%	0.4779	0.4754	0.4756	0.4791	0.4762	0.4783

表 5-9 詞主題混合模型結合語句特徵事前機率分布於發展集結果

摘要比例	WTMM	F1	F2	F3	F4	F5
10%	0.3245	0.4140	0.3221	0.3384	0.3345	0.3342
20%	0.3522	0.4285	0.3481	0.3679	0.3614	0.3619
30%	0.3509	0.4024	0.3424	0.3506	0.3532	0.3532
50%	0.4577	0.4616	0.4491	0.4573	0.4556	0.4559

表 5-10 整合關聯性模型之隱藏式馬可夫模型結合語句特徵機率分布於評估集結果

摘要比例	HMM-RM	F1	F2	F3	F4	F5
10%	0.3182	0.4864	0.3029	0.3146	0.3172	0.3211
20%	0.3264	0.4724	0.3151	0.3228	0.3216	0.3261
30%	0.3671	0.4687	0.3517	0.3639	0.3590	0.3627
50%	0.4774	0.4761	0.4625	0.4761	0.4763	0.4773

表 5-11 詞主題混合模型結合語句特徵事前機率分布於評估集結果

摘要比例	WTMM	F1	F2	F3	F4	F5
10%	0.3248	0.4692	0.3128	0.3276	0.3179	0.3260
20%	0.3324	0.4507	0.3179	0.3352	0.3250	0.3331
30%	0.3816	0.4203	0.3753	0.3821	0.3725	0.3801
50%	0.4581	0.4709	0.4583	0.4619	0.4602	0.4598

因為我們不清楚這些特徵彼此之間的關聯，所以我們嘗試使用整句最大熵值(Whole-Sentence Maximum Entropy)模型整合這些特徵。由於我們要強調的是文件中重要且將被摘要出的語句，所以我們初步地從發展集的參考摘要文件集中選出摘要比例 20%的語句當作新的訓練語料(亦曾嘗試使用摘要比例 10%與 30%的語句，但是效果較差)。實驗中我們設定 K_j 為特徵 f_j 在所有訓練語句的分數平均值，並且以不同語句特徵為初始機率分布 $P_0(S_i)$，而使用其餘特徵限制訓練整句最大熵值模型。表 5-12、表 5-13、表 5-14 與 表 5-15 分別是結合不同語句生成模型及使用整句最大熵值法訓練的語句事前機率模型實驗結果，括弧內的數值是比較以特徵分數正規化為語句事前機率分布的改善幅度。我們可以發現，透過整句最大熵值模型整合不同的特徵，在大部分的情況下，能夠有些改善，例如，於整合關聯性模型之隱藏式馬可夫模型(HMM-RM)中，以二連語言模型特徵(F2)、信心度特徵(F3)、音高特徵(F4)及能量特徵(F5)為初始機率時，在低摘要比例 10%、20%時皆能有所進步，這可能是因為位置特徵(F1)的影響使其變高。然而以位置特徵(F1)為初始機率分布時，其他特徵亦能帶來一些改善。在詞主題混合模型中亦有類似的情況，惟在以語言模型特徵(F2)為初始機率分布時，沒有改善。雖然整體而言，改善幅度並不顯著，但我們認為整合這些特徵是有必要的，且整句最大熵值模型能夠滿足這個需求。

表 5-12 整合關聯性模型之隱藏式馬可夫模型結合整句最大熵值模型於發展集結果

摘要比例	F1	F2	F3	F4	F5
10%	0.4394 (+0.0024)	0.2846 (-0.0529)	0.3663 (+0.0308)	0.3699 (+0.0234)	0.3663 (+0.0195)
20%	0.4605 (+0.0026)	0.3076 (-0.0594)	0.4109 (+0.0372)	0.4075 (+0.0224)	0.4084 (+0.0251)
30%	0.4290 (+0.0053)	0.3510 (-0.0220)	0.3919 (+0.0185)	0.3819 (-0.0035)	0.3934 (+0.0172)
50%	0.4738 (-0.0016)	0.4554 (-0.0202)	0.4732 (-0.0059)	0.4758 (-0.0004)	0.4754 (-0.0029)

表 5-13 詞主題混合模型結合整句最大熵值模型於發展集結果

摘要比例	F1	F2	F3	F4	F5
10%	0.3989 (-0.0151)	0.3061 (-0.016)	0.3566 (+0.0182)	0.3530 (+0.0185)	0.3566 (+0.0224)
20%	0.4248 (-0.0037)	0.3227 (-0.0254)	0.3831 (+0.0152)	0.3790 (+0.0176)	0.3814 (+0.0195)
30%	0.4053 (+0.0029)	0.3613 (+0.0189)	0.3662 (+0.0156)	0.3628 (+0.0096)	0.3662 (+0.0130)
50%	0.4583 (-0.0033)	0.4607 (+0.0116)	0.4517 (-0.0056)	0.4509 (-0.0047)	0.4510 (-0.0049)

表 5-14 整合關聯性模型之隱藏式馬可夫模型結合整句最大熵值模型於評估集結果

摘要比例	F1	F2	F3	F4	F5
10%	0.4907 (+0.0043)	0.3196 (+0.0167)	0.3760 (+0.0614)	0.3736 (+0.0564)	0.3763 (+0.0552)
20%	0.4749 (+0.0025)	0.3194 (+0.0043)	0.3757 (+0.0529)	0.3717 (+0.0501)	0.3759 (+0.0498)
30%	0.4641 (-0.0046)	0.3441 (-0.0076)	0.3814 (+0.0175)	0.3761 (+0.0171)	0.3769 (+0.0142)
50%	0.4822 (+0.0061)	0.4581 (-0.0044)	0.4822 (+0.0061)	0.4741 (-0.0022)	0.4757 (-0.0016)

表 5-15 詞主題混合模型結合整句最大熵值模型於評估集結果

摘要比例	F1	F2	F3	F4	F5
10%	0.4615 (-0.0077)	0.3464 (+0.0336)	0.3574 (+0.0298)	0.3526 (+0.0347)	0.3451 (+0.0191)
20%	0.4445 (-0.0062)	0.3344 (+0.0165)	0.3580 (+0.0228)	0.3519 (+0.0269)	0.3457 (+0.0126)
30%	0.4252 (+0.0049)	0.3581 (-0.0172)	0.3838 (+0.0017)	0.3816 (+0.0091)	0.3816 (+0.0015)
50%	0.4723 (+0.0014)	0.4608 (+0.0025)	0.4615 (-0.0004)	0.4631 (+0.0029)	0.4618 (+0.0020)

5.4 本章結論

於本章中，我們簡介了語音文件摘要的過程，並且使用機率生成架構(Probabilistic Generative Framework)進行摘要。機率生成架構可分成兩個部分，語句生成模型(Sentence Generative Model)及語句事前機率模型(Sentence Prior Model)。過去已有一些語句生成模型被提出，如隱藏式馬可夫模型及詞主題混合模型，實驗亦展示了其效果比向量空間模型好。本論文提出了使用整句最大熵值模型(Whole Sentence Maximum Entropy, WSME)作為語句事前機率模型，並嘗試整合語言及語音上的特徵，如二連詞分數(Bigram Score)，位置資訊(Position Information)、辨識信心度(Confidence Score)、音高(Pitch)及能量(Energy)等。由初步實驗結果發現，使用整句最大熵值模型整合額外的特徵資訊能夠提升語音摘要正確率，且語句於文件位置的特徵對於新聞語料相當有效。