

新聞文件摘要之研究

李祥賓 柯淑津

東吳大學資訊科學系

{ms8805, ksj@cis.scu.edu.tw}

摘要

本文主要以兩種摘要技巧對路透社新聞文件進行自動文件摘要處理，分別為由語句出現的位置來判斷其重要性，以及擴充標題詞彙兩種摘要技巧。我們對文件進行分析探討，找出文件主題通常是佔據了哪些位置，以擷取這些重要位置的句子為摘要。另外，我們認為標題對於文件是相當重要的，因此我們透過 WordNet 找尋標題的相關詞彙，對標題詞彙進行擴充，來找出更多與標題相關的字，增加標題的重要性，進而協助在文件中找尋與標題較相關的摘要語句。在實驗評估方面，我們提出一個以分類方式進行自動摘要評估的想法，並建立了一個分類系統來評估摘要結果。最後，本文提出了一種綜合擴充標題詞彙與重要位置的摘要方法，此方法得到 71.9% 分類精確率的實驗結果，相較於 65.6% 的基準分類精確率改善了 9.6% ($(71.9\% - 65.6\%) / 65.6\% = 9.6\%$)。

1. 簡介

在資訊科技發達的今日，文件已由傳統的書面呈現方式，轉化成數位方式包裝。這些資訊再藉由無遠弗屆的網際網路傳播到世界上各個角落，因此使用者可以輕易地透過網路獲得所需求的資訊。

資訊流通的便利性，雖然帶來了豐富的資源，但同時，也引進了另一個問題：「資訊氾濫」。網路使用者可能都有過這樣的經驗：當瀏覽線上文件時，發現過多的文件讓使用者無法一一詳盡閱讀全文，而只觀看文件的標題又無法掌握到文件的內容，進而判斷出此文件是否符合自己需求。目前新聞網站的線上新聞就是一個相當好的例

子。如果這些網路新聞在總覽時，能適切地提供精簡摘要來協助使用者選擇自己所需的文件。將有助於降低網路傳輸量，進而提升網路服務品質。

這類的文件摘要工作最先是由專業人員以人工方式來進行，雖然人工摘要的品質相當好，但遇到大量且更新快速的網路新聞，用這樣的方法就顯得緩不濟急。然而，自動文件摘要的技術正是解決這個難題的利器。自動文件摘要技術是擷取文章內重要的訊息出來，經過組合產生較短的摘要，讓使用者可快速地明白這篇文章的主旨，藉此節省使用者的閱讀時間，而能較快地判斷該篇文章是否為自己想要的文件。

過去文獻中，已有許多相關於自動文件摘要的研究，而本研究將針對下列兩種摘要策略進行研究與探討：由位置來判斷文件主題與擴充標題詞彙。並設計實驗來驗證這些摘要策略是否可擷取出品質良好的摘要內容。另外，對於文件摘要的成效評估，本研究提出了一個新的想法，以應用系統來評估摘要結果。我們將摘要結果取代原本文件，進行分類處理。再以分類結果來驗證摘要成效。假若，我們的摘要確實能由文件中擷取重要資訊，相較於用全文來進行分類，在分類效果上應該會不相上下或是有更好的精確率。

本篇文章共分為六節，第二節針對過去有關於文件摘要研究的文獻進行回顧。第三節介紹本文所使用的研究資源，包括路透社新聞語料與含標記詞義訊息的布朗語料庫。第四節主要探討本研究所使用的兩種摘要策略：由位置來判斷文件主題以及擴充標題詞彙。另外，在此節中，我們提出一個以分類系統來評估摘要成效的想法。第五節針對本文提出的摘要方法設計一系列實驗，以路透社新聞語料進行摘要處理，並將摘要結果送交分類系統，再對分類結果進行討論。最後，我們在第六節提出結論以及未來研究方向。

2. 相關研究

過去對於文件摘要的研究，多以單一文本為對象，也就是只針對一篇文章內容來進行摘要處理，應用不同的技巧，來表現出文件中的重要資訊。其中，有些研究透過計算

各詞彙在文件中所擁有的權重($tf \times idf$)，藉此權重值可找出文件中較具重要性的詞彙，進而擷取出含有重要詞彙的語句來形成摘要(Forsyth and Rada, 1986)。有的研究則是針對文章進行資訊擷取(information extraction)處理，找出文章內的人名、地名、組織名等專有名詞，再對這些專有名詞與新聞內文設定不同的權重，進而擷取出文件中的重要語句(邱中人, 2000)。有些研究則考慮語句在文件中的位置，認為出現在某些特定位置的句子，常常較具重要性，可以直接擷取出來當作摘要(Hovy and Lin, 1997)。

而相對於單一文本的研究，有些學者致力於多文本的摘要研究，他們對報導同一事件的多篇新聞文稿歸納它們的相似處，以及辨識出彼此相異的地方，做成該事件的摘要(McKeown and Radev, 1995; Barzilay, McKeown and Elhadad, 1999; Chen and Huang, 1999)。多文本摘要研究常用的技術，除了上面所談到的慣用於單一文本摘要處理的方法外，McKeown 等人則是為相對於恐怖份子的新聞設定了樣版(template)，樣版用於擷取多篇文章內的資訊，如報導來源、報導日期、事件發生日期、事件發生的情形等等詳細的資訊；這些擷取出的資訊，經判斷與比對處理後，會送至摘要製作模組以產生出一個多文本的摘要(McKeown and Radev, 1995)。

多數的摘要研究在針對文件的重要內容做相關統計時，大部分是以字詞作為處理單位。這些方法大多都是找尋文章內重要的字詞，再依這些字詞來進行進一步的處理，找出重要語句來形成摘要。有的則利用自然語言處理技術，如：片語、暗示字 (cue word)、上下文 (discourse) 處理等來協助辨認文件中的重要語句 (Marcu, 1999)。這些研究認為某些特定片語後面接的句子有某程度的重要性，如：“結論...”，因此包含這些特定片語的句子會依照其上下文關係，經過處理後，被擷取出來形成摘要。

但是，因為字詞的參數空間大，且存在一詞多義與同義字等問題；因此，有些學者認為應該跳脫字詞層次，以字詞所蘊藏的概念或語意來取代字詞本身，彙整成文件的概念主題，進而找出表達文件的重要概念。在 1999 年，Hovy 等人使用 WordNet 同義辭典來做概念之間的相關性聯結(Hovy and Lin, 1999)。Woods 使用片語的分析來組織字詞成為一個概念性的架構(Woods, 1997)。但這類的方法往往需要有強大的語言

知識做為輔助，而這種資源通常是相當不容易獲得的。

3. 研究資源

本文在研究過程中，使用「WordNet」、「路透社新聞語料」與「含標記詞義訊息的布朗語料庫」等資源來進行摘要實驗。對於 WordNet 的介紹，請參見 Miller 等人的文章(Miller et al., 1993; Fellbaum, 1998)。在下面，我們將對後兩項資源逐一介紹。

3-1 路透社新聞語料

路透社新聞語料庫(Reuter Corpus)，由 Lewis 在 1992 年所收集，目前此語料庫是文件分類研究中最常使用的語料庫之一，其內容取自 90 年間的路透社新聞文章，總共含超過兩萬篇標註分類類別的文件(Hayes and Weinstein, 1990; Lewis and Ringuette, 1994)，而從最初的版本演變到現在共有五個版本，各版本的差異在 Yang 的論文中有很詳盡的比較(Yang, 1999)。

本文選擇版本 3 的語料做為我們進行摘要處理時的實驗語料。此版本的路透社新聞語料包含 7789 篇訓練文件與 3309 篇測試文件，共分為 93 種類別。新聞語料中多數的文件被歸類到一種類別，但被歸類在二種類別以上的文件也不在少數。經過統計後，我們得知在訓練語料與測試語料裡，每篇文件的平均類別數分別為 1.23 和 1.24。

在路透社新聞語料中，每篇文件的內容長短不一，有些文件可能只有幾個句子，有些卻可達到十多句以上。經過統計後，在訓練語料與測試語料中，每篇文件的平均句長分別為 6.8 句與 7.3 句；而從整個語料來看，每篇文件的標題與內文平均長度分別為 7.4 和 126.9 個詞彙。

3-2 含標記詞義訊息的布朗語料庫

布朗語料庫(Brown Corpus)是在資訊檢索相關研究上，常被使用的語料庫之一。本文所使用的布朗語料庫，主要詞彙已標記出其所屬的 WordNet 詞義。這個布朗語料庫是由普林斯頓大學認知科學實驗室(Cognitive Science Laboratory)，利用 WordNet 內的詞義架構，以人工方式為此語料內的詞彙進行詞義標記，此語料庫可由

<http://www.cogsci.princeton.edu/~wn/> 網址獲得，主要分為 brown1, brown2, brownv 三個部分，詳細的資料列於表 1。

表 1：具標記詞義的布朗語料庫---詳細資訊。

名稱	檔案個數	詞義標記範圍	已標記詞義的詞彙
brown1	103	名詞、動詞、形容詞與副詞	106,725
brown2	83	名詞、動詞、形容詞與副詞	86,414
brownv	166	動詞	41,525
Total	352		234,664

從表 1 內容，可看出這個布朗語料庫共有 352 個檔案。因 WordNet 只包含名詞、動詞、形容詞與副詞這四種詞性，因此，只能針對這四種詞性的詞彙進行詞義標記工作。在整個語料庫中，總共有 234,664 個詞彙被標記上 WordNet 詞義。我們將利用這個布朗語料庫，所提供的詞彙與詞義資訊，來協助本研究進行摘要處理。

4. 研究方法探討

在本節，我們將研究方法區分為摘要與評估兩部分進行討論。在 4-1 至 4-2 小節中，我們說明本文在進行摘要研究時所使用的兩種方法。另外，本研究提出一個以分類方式來評估摘要成效的想法，並針對我們所使用的分類系統，在 4-3 小節中進行詳細的說明。

4-1 由位置來判斷文件主題

在一般文件的內容架構上，文件的主題句通常會佔據著某些特定的位置。因此本研究希望能透過考量位置的重要性來找出文件的主題所在。Edmundson 認為包含標題字的語句，會與文件主題較為相關(Edmundson, 1969)。另外，最重要的語句通常會出現在文件較前面或是較後面的位置。此外，Hovy 等人也認為文件主題常佔據特定位置，他們曾針對電腦相關的新聞文件進行研究，提議將標題字與簡介內的字視為文件的重要詞彙，並應用這些重要詞彙來統計分析出文件的重要位置(Hovy and Lin, 1997)。

本研究將針對訓練語料來進行文件內重要語句位置判定。而重要位置判定的方法運用了 Edmundson 與 Hovy 等人的看法，利用文件的關鍵字來統計分析在文件內的哪一個位置，其語句的內容通常具有較高重要性。

在關鍵字的選取方面，我們運用傳統資訊檢索所常用的 $tf \times idf$ 技巧，來找出文件裡的重要詞彙，再從重要詞彙中抽取出一定比例的詞彙來做為關鍵字，而從文件中找出重要詞彙的方法，如公式 1、2 所示。假設一個文件 d ，以及出現在 d 中的詞彙 t ，我們定義 t 在 d 中所具有的權重值， $W(t, d)$ ，為 t 在 d 中的出現次數 $tf_{t,d}$ 乘以詞彙 t 本身的重要性 idf_t ，再除以 d 中詞彙出現的最高詞頻。接著，我們將權重值較高的詞彙視為文件 d 的重要詞彙。

$$W(t, d) = \frac{tf_{t,d} \times idf_t}{\text{Max}_t (tf_{t,d})}, \quad (1)$$

$$idf_t = \log \left(\frac{T}{df_t} + 1 \right), \quad (2)$$

$tf_{t,d}$: 文章 d 中此詞彙 t 的字頻，

df_t : 詞彙 t 出現在訓練語料中的文章數，

T : 訓練語料中所有文章總數。

接著，使用這些關鍵字來進行重要位置的判定。我們假設這些關鍵字的重要性比文件內其餘的詞彙高，而且一個重要的語句應該會包含較多的關鍵字。

我們針對給定的一個文件 d ，如果詞彙 t 屬於文件 d 的關鍵字，則將擁有權重值 $W(t, d)$ 。之後，再以句子為單位，計算出文件內句子 s 所擁有的權重值 $Score(s, d)$ ，計算句子權重值的方法如同公式 3 所示。

$$Score(s, d) = \sum_{t \in s \cap \text{關鍵字}_d} W(t, d), \quad (3)$$

t : 出現在文件 d 中的詞彙。

當我們求出文件中每一個句子的權重值後，依句子的權重值高低就可求得文件中重要位置的排行。我們為訓練語料內的所有文件進行上述的處理，就可以知道每一篇

文件中重要句子的位置分佈。最後，我們統計出每個位置在文件內所擁有的重要性排行。進而利用此結果擷取出重要位置所在的句子以形成摘要。綜合上述，我們將此摘要方法的處理過程加以整理，列出於圖 1。

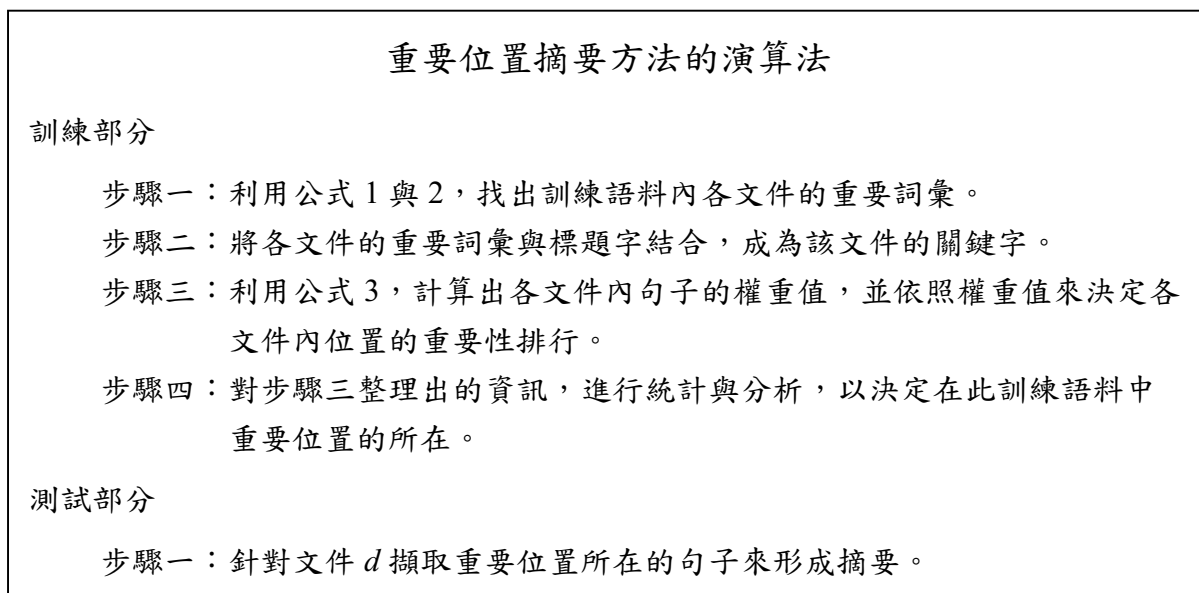


圖 1：重要位置摘要方法的演算法。

4-2 擴充標題詞彙

標題通常與文件主題具有較高的相關程度，我們認為文件中的句子若包含較多標題詞彙，則此句子與文件主題會具有較高的相關程度，因此可擷取作為摘要的內容。但是，標題內容往往因過於精簡，使得我們無法從中獲得足夠的資訊。而且在一篇文章中，作者可能會使用不同的詞彙來表達同一事物或動作，這樣的寫作風格雖增加了詞彙的多樣性，卻使得我們在進行重要詞彙比對過程中，比對到相同詞彙的機率大為降低。對於這樣的現象，我們認為文章作者雖然可能使用多樣性的詞彙來表達同一主題，不過這類相互可替代的詞彙在語意上應該具有高度相關意義。因此，我們希望找出文章作者有可能使用的相關詞彙。

為了找出與標題內容具相關語意的詞彙，我們藉由 WordNet 的豐富語意網絡，利用關聯性指標找出一個詞彙的同義詞與上義詞。另外，我們認為在解釋詞彙詞義的定義中所使用到的詞彙，應該會與此詞彙具有一定的相關程度。因此，我們利用

WordNet 來擴充標題詞彙，透過收集與標題詞彙較為相關的同義詞、上義詞與定義內的詞彙，並將這些詞彙視為標題詞彙，來解決文件內容使用多樣性詞彙所帶來的影響。

在 WordNet 內，有些一詞多義的詞彙會擁有多個詞義。因此在擴充詞彙之前，我們必須知道詞彙的正確詞義，也就是說我們需要先解決詞義歧異的問題。本研究使用了兩個方法來針對標題詞彙進行詞義辨識，分別是以 WordNet 的定義與文件詞彙的重疊性來進行詞義歧異辨識，與利用語料庫詞義出現機率進行詞義歧異辨識。

4-2-1 以 WordNet 進行詞義歧異辨識

這個方法是假設在標題中一個詞彙若具有某項詞義，此詞義所衍生出來的相關詞彙，在文件中出現的機率應該會高過於其它詞義的衍生詞彙。於是我們針對標題內的名詞、動詞與形容詞，使用 WordNet 來找出每個詞彙在不同詞義下所擁有的同義詞與上義詞，以及各詞義在定義內所使用的詞彙，再將三者聯集成一個詞彙集合。接著，將集合內的詞彙與文件內容進行比對，我們把出現重覆性最高的集合所代表的詞義，設定為此詞彙的詞義。圖 2 表示此方法對於詞彙所形成的集合，圖中的詞彙有 n 種詞義，而這 n 種詞義分別有各自的同義詞、上義詞與定義內的詞彙，因此會有 n 個詞彙集合。

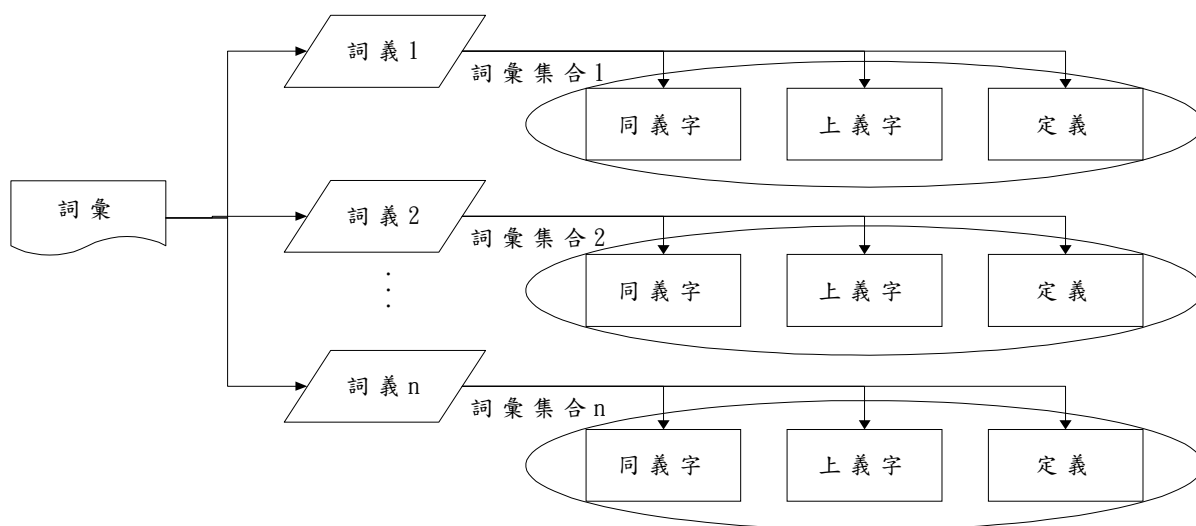


圖 2：詞彙在 WordNet 中所擁有的各種詞義與其相關詞彙集合。

當我們利用 WordNet 對標題詞彙作詞義辨識時，一個詞彙下的所有詞彙集合有可能在文件內出現次數都一樣，使得無法藉著出現次數多寡來判斷正確詞義。此時，

我們便借重第二種詞義辨識方法，利用語料庫詞義出現機率進行詞義歧異辨識。

4-2-2 利用語料庫詞義出現機率進行詞義歧異辨識

第二個詞義辨識方法則是給予一個詞彙在語料庫中最常出現的詞義。本文所使用的語料是已標記詞義的布朗語料庫(Brown Corpus)，本文在 3-2 小節中對此語料庫有詳細的說明。此語料庫內的詞彙使用了 WordNet 來進行詞義標記，我們對這些詞彙進行統計，列出詞彙在不同詞義下出現的機率。因此在辨識詞義的過程中，我們直接挑選出現頻率最高的詞義，來作為這個詞彙出現在標題時所具有的詞義。藉由這樣的處理方式，我們便可輔助以 WordNet 辨識詞義的不足，而提高成功辨識率。

利用上述兩種方法進行詞義辨識後，就可以針對標題詞彙的詞義來擴充其相關的詞彙。詞彙擴充的結果，就是將標題詞彙的上義詞、同義詞與定義內的相同詞性詞彙，擷取出來形成此標題詞彙的相關字。我們將標題詞彙與擴充後的詞彙視為重要詞彙，接著，計算每個句子所擁有的重要詞彙數，作為句子的分數，然後再挑出較高分數的句子來作為摘要。擴充標題詞彙摘要方法的演算法，我們整理於圖 3。

擴充標題詞彙摘要方法的演算法

- 步驟一：針對文件 d 的標題詞彙 t ，使用 WordNet 找出其詞義的詞彙集合。
- 步驟二：將詞彙 t 的所有詞彙集合與文件 d 內的詞彙進行比對，計算各集合的出現次數。
- 步驟三：計算各詞彙的最高頻詞義與次高頻詞義之比例值 r ，若 r 大於某預先設定的門檻值 h ，則設定詞彙 t 的詞義為出現次數最高的集合所代表的詞義；否則，直接給予語料庫中出現頻率最高的詞義。
- 步驟四：利用辨識出的詞義，來擴充標題詞彙，收錄該詞義的同義字、上義字與定義內同詞性的詞彙。
- 步驟五：結合標題詞彙與擴充後的詞彙，作為文件的重要詞彙。
- 步驟六：計算文件 d 內句子 s 所擁有的重要詞彙數，挑選出擁有較多重要詞彙的句子作為摘要內容。

圖 3：擴充標題詞彙摘要方法的演算法。

4-3 評估方法

自動摘要的評估是一件相當困難且主觀的工作。一般而言，評估文件摘要成效的方法可分為兩種作法。第一種作法是由公正的第三者對摘要內容進行判斷，決定摘要內容是否恰當，或是涵蓋的資訊是否充足。第二種作法則是將摘要內容應用於其他不同的工作上，觀察此工作的表現來決定摘要的成效(Mani and Bloedorn, 1999)。

本研究提出一個評估的想法，此想法較偏向於 Mani 等人所提出的第二種摘要評估作法。我們針對英文文件所進行的摘要實驗，實驗資料為路透社的新聞語料，這些新聞語料內的文件均已標註其所屬的分類類別。我們認為如果本研究的摘要成果是具有正向意義的，那麼經過摘要處理的文件，因為其摘要內容是由文件中較重要的句子所組成，這些句子所表達的資訊應該足夠代表該文件。因此若將此摘要結果送由分類系統來進行分類處理，相較於用全文來進行分類，在分類效果上應該會不相上下或是有更好的精確率。為因應這個想法，我們需要一個分類系統來協助我們進行摘要的評估工作。

本研究根據 Ker 和 Chen (2000)所提出的分類方法中，挑選了其中一個方法來針對文件全文建立分類系統。首先，我們將語料中的詞彙全部轉成小寫字母，並刪除停用字，再把這些留下來的詞彙予以原形化(stemming)處理，最後所得的詞彙就是特徵字(feature)。接著，我們針對訓練語料中每個類別 c 所擁有的文件，彙總其特徵字 f 來進行統計處理，給予它們應有的權重值 $W(f, c)$ ，而計算權重值的方式如公式 4, 5 所示。

$$W(f, c) = tf_{f,c} \times idf_f, \quad (4)$$

$$idf_f = \log_2\left(\frac{T}{df_f}\right), \quad (5)$$

$tf_{f,c}$: 特徵字 f 出現在類別 c 的頻率，

T : 類別的總數，

df_f : 特徵字 f 出現過的類別總數。

給予一篇測試語料的文件 d ，我們在決定所屬的類別時，必須針對文件內所含的各類別特徵字進行權重值的加總，以得到文件在不同類別下的總權重值 $R(c, d)$ 。接著，統計各類別的總權重值，將文件歸屬於總權重值最高的類別 $Class(d)$ ，計算方式如公式 6, 7 所示。

$$R(c, d) = \sum_{f \in F_c} tf_{f,d} \times W(f, c), \quad (6)$$

$$Class(d) = \arg \max_c R(c, d), \quad (7)$$

$tf_{f,d}$ ：特徵字 f 在文件 d 出現的頻率，

F_c ：類別 c 的特徵字集合。

由上述的方法，我們建立了一個分類系統。我們將未經過摘要處理的路透社新聞測試語料由分類系統處理，得到精確率為 65.6% 的分類效果，我們以此精確率來當作基準精確率，來做為摘要成效比較的基準。

5. 文件摘要實驗

為驗證本研究所提出的文件摘要方法之效果，我們以路透社的新聞語料(版本 3)為實驗資料設計了一系列的實驗，來觀察各摘要方法的成效。首先，我們先介紹資料的前置處理。接著，將分別介紹各種摘要實驗方法，包含實驗設計與實驗結果等，最後對實驗結果進行討論。

5-1 前置處理

在進行摘要處理之前，我們對於實驗資料的前置處理主要有下列幾個步驟：首先必須去除停用字(stopword)，以減少不具重要意義的詞彙。接著是原形化的處理，這步驟主要考量英文詞彙具有不同詞性的轉換。如 relate、relation、relative，這三個詞彙雖具有相似的意義，但在統計處理上會視為不同的詞彙；而原形化能將這類的詞彙變成較短的字根，如 relate，因此可彙整此類詞彙在文件中所佔有的重要性。

5-2 由位置來判斷文件主題

我們運用在 4-1 小節中提出的方法來分析文件的重要位置所在。針對所有訓練語料中

的文件皆挑出五個重要句子，再依照各位置的重要性與出現頻率來進行整理，進而分析重要位置所在。另外，由於路透社新聞語料的文件平均長度約為七個句子，因此我們認為摘要的長度約在三個句子左右較為合適。所以，我們只找出文件內的三個重要位置。我們從統計數據中，找出了三個出現次數較高的位置。按出現次數由高而低地排列下來，分別為第二段第一句、第一段第一句與第三段第一句（簡稱為 P_2S_1 ， P_1S_1 ， P_3S_1 ），如表 2 所示。

表 2： P_1S_1 、 P_2S_1 與 P_3S_1 出現在文件內的重要性與次數。

句子位置\重要性	一	二	三	四	五	總計
P_2S_1	1584	2227	1303	626	355	6095
P_1S_1	3142	1338	617	412	364	5873
P_3S_1	385	975	1692	1012	631	4695

接著，從表 2 內容，我們發現在總數 7789 篇的訓練語料中，第一段第一句出現在文件內最重要句子的情形就多達 3142 次，依此可明顯判斷出此位置是此新聞語料中的最重要位置。至於第二段第一句與第三段第一句，這兩個位置的句子出現在文件內次要句子的次數分別為 2227 次與 975 次，明顯地區別了兩者的重要程度，因此次要位置應屬於第二段第一句。藉由對這三個位置的分析，我們得知在路透社新聞語料的訓練文件中，重要句子出現的位置依排行分別為第一段第一句、第二段第一句以及第三段第一句。

另外，我們由每篇文件提出前五重要句子，總共得到 30800 多個句子；但由於某些文章本身並不足五句，導致我們提出的句子數目比預計的數目少了許多。然而，除了我們上述討論的三個位置總共佔了 16663 個句子外，其餘位置總共出現了約 14200 次，但是這些位置的出現次數都不高，因此我們不對這些位置做進一步的處理。

5-2-1 實驗設計

本研究設計了四組實驗來驗證我們經由統計分析後，所得出的重要位置對於摘要是否

有其正向意義。這些實驗的主要考量為依照不同位置的重要性，分別擷取出不同句數的摘要。

此外，我們考慮在路透社新聞語料中每個類別所擁有文件的數量多寡不一，文章長度也不太一樣（如表 3 所示，類別 acq 文件的平均句長為 5.3 句，而類別 money-fx 文件的平均句長則為 8.9 句）。在這樣的情況下，重要位置分析結果是否適用於不同的分類，我們無法確切得知。於是，本研究針對文件數量較多的三個類別 earn、acq 以及 money-fix，進行相同的實驗，希望能得知在不同類別的文件內，重要句子出現的位置是否一致。綜合上述考量，我們所設定的四組實驗如表 3 所示，其中，實驗一以整個測試語料進行實驗，而其他三組實驗則各自針對不同類別進行相同處理。

表 3：重要位置摘要方法實驗設定。

組別	受測文件	平均文件長度(句)	文件數量(篇)
實驗一	整個測試語料	7.3	3309
實驗二	類別 earn	7.4	1176
實驗三	類別 acq	5.3	776
實驗四	類別 money-fx	8.9	207

註：四組實驗的位置選取設定皆是相同的，分別以 $\{P_1S_1\}$ 、 $\{P_1S_1, P_2S_1\}$ 、 $\{P_1S_1, P_2S_1, P_3S_1\}$ 為摘要內容。

5-2-2 實驗結果與討論

本研究針對重要位置摘要方法所設計的四組實驗，經過分類系統處理後，所得到的精確率如表 4 所示。由於製作出的摘要包含的句子數量不同，使得摘要佔全文的長度比例也不同。我們對此列出不同實驗設定所製作出的摘要，其內容所佔全文的長度比例，來評估資訊量減少的情形。

從表 4 中，我們可以觀察到實驗二與實驗三所得到的數據，皆高於其它兩組。對此，我們認為這樣的情形是由於製作分類系統時，我們所使用的方法對於文件數量較多的類別會產生較好的分類辨識效果，因此對於數量較多的類別 earn 與 acq 而言，

其分類成效會較高。

表 4：重要位置摘要方法的實驗結果。

重要位置選取	實驗一		實驗二		實驗三		實驗四	
	精確率	長度	精確率	長度	精確率	長度	精確率	長度
P ₁ S ₁	55.4%	13.8%	70.2%	13.5%	76.8%	19.0%	46.4%	11.3%
P ₁ S ₁ 與 P ₂ S ₁	66.9%	26.1%	94.4%	26.4%	77.7%	35.4%	65.7%	20.2%
P ₁ S ₁ 、P ₂ S ₁ 與 P ₃ S ₁	65.9%	37.2%	95.2%	39.2%	73.2%	48.1%	67.1%	28.4%
全文	65.6%	100.0%	95.8%	100.0%	76.7%	100.0%	68.1%	100.0%

從實驗一、二與四的結果來看，只選取最重要位置的句子為摘要，此摘要佔全文的長度比例未達三分之一，在這樣的摘要長度下，摘要所能包含的資訊並不充足；因此，得到的精確率與全文的分類成效相比，均有相當大的差距。當摘要內容增加第二與第三重要位置後，摘要長度可達到全文的四分之一至三分之一；這樣的摘要長度可擷取出較充足的資訊，再加上摘要內容是由重要位置的句子所組成；因此，此摘要所得到的分類成效已經接近各組實驗的全文分類精確率。

另外，我們觀察類別 acq 所進行的實驗，發現其摘要長度為全文的三分之一至五分之一時，這樣的摘要所帶來的資訊量是比較充足的，因此實驗成效已經突破了基準精確率。但是增加擷取第三重要位置後，摘要的長度比例增為 48.1%，已相當接近全文的一半；但是由分類精確率反之下降了 4%來看，顯示了過長的摘要並不一定會帶來更多重要的資訊。

除此之外，本研究進行了另一項實驗，我們從測試語料各文件中隨機選取了三分之一的句子來形成摘要，並進行分類評估，得到了 55.7% 的分類精確率。此精確率與實驗一的成效相比較，我們可以看出文件內的确有特定重要位置的存在，才會使得由重要位置所形成的摘要在分類成效上比起隨機選取的方式改善了 20.2% 左右。

從上述這些實驗中，我們認為對於不同來源的文件，只要經過統計分析，就可以得知其重要位置所在，進而藉此資訊來找出重要句子。而從實驗結果中，我們認為一

個摘要的產生，它的內容若可以包含前兩重要的位置——第一、二段的第一句，在摘要長度與重要性上是較好的考量。

5-3 擴充標題詞彙

擴充標題詞彙主要是以標題內的名詞、動詞與形容詞為對象進行處理，因此我們必須得知詞彙的詞性，才可以擷取這三種詞性的詞彙來進行詞彙擴充。本文使用由麻省理工學院的 SLS(Spoken Language Systems Group)在 1993 年所發表的詞性標記工具 (<http://www.sls.lcs.mit.edu/sls>)，是一種以規則為本(rule-based)的標記詞性方法。

接著，針對標題內的名詞、動詞與形容詞，我們使用了在 4-2-1 與 4-2-2 小節所提出的兩個方法來進行詞義辨識。並藉由辨識出的詞義，來為標題詞彙進行擴充，納入與標題詞彙相關的 WordNet 同義詞、上義詞與定義內容。

5-3-1 實驗設計

首先我們利用 4-2-1 小節所提出的第一個方法來為標題詞彙辨識詞義。我們將這些標題詞彙直接與 WordNet 所擁有的詞彙進行比對，找出標題詞彙的各種詞義。

在這個步驟中，我們先從 23,200 多個標題詞彙中，挑出了 19,600 多個屬於名詞、動詞與形容詞的詞彙。我們利用這些詞彙與 WordNet 進行比對後，總共有 14,600 多個詞彙出現於 WordNet 中。至於那些不包含在 WordNet 內的詞彙，我們經過觀察，發現其中有許多是屬於專有名詞，例如 IBM、COMPAQ 等等；因為這些詞彙未出現於 WordNet 中，我們無法對它們進行詞彙擴充處理。接著，我們從 14,600 多個詞彙中，找出了 69,300 多種詞義，平均一個詞彙有 4.7 種詞義存在。我們藉由 WordNet 內的詞彙網絡，找出了各詞義的同義詞、上義詞以及其定義內與此標題詞彙詞性相同的詞彙，整理成一個一個的詞彙集合。

當各詞義的詞彙集合整理出來後，我們利用這些詞彙集合與文件內容做重覆性比對，計算每個集合中的詞彙被比對到的次數。針對一個標題詞彙，我們選出次數最高的集合所代表的詞義作為此詞彙的詞義。然而，當比對次數最高的集合超過一個的情形，我們便無法判斷此詞彙該歸屬於哪一個詞義。這時我們便利用第二種詞義辨識法

賦予此詞彙在語料庫中最常出現的詞義，以完成對標題詞彙進行詞義辨識的工作。

當成功地判斷出標題詞彙的詞義之後，我們便可依據這些判斷出的詞義，來為各詞彙進行詞彙擴充的工作。詞彙擴充的方式主要是從 WordNet 中，將各已知詞義的同義詞、上義詞與定義內相同詞性的詞彙擷取出來，藉由這些擴充詞彙來補強原有的標題詞彙，增加標題詞彙的影響力。

下面，我們舉個例子做個簡單的說明。表 5 是一篇節錄後的文章，以方形框起來的字是內文中出現的標題字，而以黑底呈現的字則是擴充後的字。我們觀察第七段第一句，此句與標題詞彙比對後，只有一個詞彙 baker 出現在此句；而在擴充詞彙後，我們在此句中找出了 treasury 的上義詞 {funds, finances}，以及 shift 的上義詞 {modifications}，藉著這些相關詞彙使得第七段第一句變得與標題較為相關。

表 5：詞彙擴充前後的文件範例。

文件位置	內容
標題	U.S. TREASURY'S BAKER SAYS RATE SHIFTS ORDERLY..
P ₁ S ₁	WASHINGTON, April 9 - Treasury Secretary James Baker said that changes in exchange rates have generally been orderly and have improved the prospects for a reduction in external imbalances to more sustainable levels.
P ₂ S ₁	In remarks before the IMF's policy-making Interim Committee, Baker reiterated a Group of Seven statement last night that the substantial exchange rate changes since the Plaza agreement 18 months ago have "now brought currencies within ranges broadly consistent with economic fundamentals."
P ₇ S ₁	Baker also urged the International Monetary Fund's executive board to review possible modifications to the Fund's compensatory financing facility before the annual meeting this fall.
註：{consistent}是 orderly 的同義詞，{change, modification}是 shift 的上義詞，{funds, finances}是 treasury 的上義詞。	

接著，摘要的製作是將經過擴充後的標題詞彙視為文件的重要詞彙，並給予一權重值，藉此對文件內的句子進行重要句子的選取。在這裡，我們將標題詞彙與擴充後的詞彙所擁有的權重值皆設定為 1，將兩者視為具有相同的重要性。在設定好權重值

後，我們利用 4-1 小節的公式 4 去計算文件內每個句子所擁有的權重總值，擷取權重總值較高的句子做為文件摘要。在摘要長度考量上，我們擷取前三分之一的句子（最多三句）作為摘要。

5-3-2 實驗結果與討論

從實驗結果中，我們發現在總數 3309 篇的測試語料文件中，只有 2916 篇文件能成功地以擴充標題詞彙的方法，找出重要句子以形成摘要。其他的 393 篇文件因為標題詞彙與擴充後的詞彙並沒有出現在文件內文裡，使得我們在計算句子權重值時，文件內所有的句子權重值皆為 0，因而無法擷取出任何句子來形成摘要。這說明了擴充標題詞彙摘要方法在某些文件上無法發揮其功效。表 6 列出擴充標題詞彙摘要方法的實驗結果，若以所有 3309 篇文件來計算，摘要結果可得到 61.9%的分類精確率。

表 6：擴充標題詞彙摘要的實驗結果

	文件數量	分類精確率
成功給出摘要的文件	2916	70.3%
無法給出摘要的文件	393	0.0%
所有文件	3309	61.9%

不過，如果我們在評估摘要成效時，只考慮那些有摘要產生的 2916 篇文件，我們可以得到 70.3%的分類精確率。因此，從上述的實驗結果，我們驗證了擴充標題詞彙摘要方法所產生的摘要，其摘要內容的確包含了文件內的重要句子，才能得到較佳的分類成效。更值得一提的是在詞義辨識部分，我們的方法不需繁複的計算與疊代，就可以對標題詞彙找出其最有可能詞義，來協助擴充標題詞彙摘要方法的進行。

5-4 結合擴充標題詞彙與重要位置摘要方法

在 5-3 小節中，我們利用擴充標題詞彙方法來擷取文件內的句子以形成摘要。但在詞彙比對過程中，我們發現並不是所有的文件內容都會出現標題詞彙，這樣的情形使得有些文件因為不能在內容中比對到標題詞彙，而無法擷取出任何句子出來，因此這篇

文件的摘要內容將會是空白的。為了彌補擴充標題詞彙摘要方法的不足，我們參照了 5-2 小節重要位置摘要方法的實驗結果，決定利用此摘要方法來加以輔助。

5-4-1 實驗設計

本研究設計了一組實驗，我們先利用擴充標題詞彙摘要方法來對所有文件進行處理，過程中所使用的各項參數與 5-3 小節中有相同的設定。接著，我們找出摘要內容為空白的文件，以重要位置摘要方法來進行處理，我們直接擷取出文件中的第一段第一句、第二段第一句以及標題來作為該文件的摘要內容。綜合這兩種摘要方法，我們便可以對所有文件擷取出重要句子。

5-4-2 實驗結果與討論

我們製作出的摘要，經過分類系統處理後，得到了 71.9% 的分類精確率，相較於 65.6% 的基準精確率提升了 9.6% ($(71.9\% - 65.6\%) / 65.6\% = 9.6\%$)。這顯示出我們結合擴充標題詞彙的摘要方法，並以重要位置為輔助，所製成的摘要，確實包含文件的重要內容，使得評估成效能有大幅度的改善。

6. 結論與未來方向

由第五節各種摘要的實驗結果，我們可以發現重要位置摘要實驗，得到接近基準精確率的實驗結果，說明了文件內的确存在著與主題相關的特定重要位置。另外，我們由實驗結果中，驗證了摘要長度若能達到文件的三分之一，將可提供足夠的重要資訊給予使用者。

在擴充標題詞彙摘要實驗中，我們利用了 WordNet 的特性，藉著它的豐富語意網絡對標題進行詞彙擴充，增加了標題的影響力，找出與標題較相關的語句作為摘要內容。不過在實驗過程中，我們發現這個摘要方法有其不足之處，主要是由於文件的內容不一定有著標題詞彙或其相關詞彙的存在。因此，我們提出了一種綜合擴充標題詞彙與重要位置的摘要方法，藉由結合兩個方法來協助所有的文件都能產生出重要的摘要內容，這樣的作法得到了令人滿意的摘要成效，將分類精確率提升了 9.6%。

未來，我們將試著把本文所提出的摘要方法應用於中文文件上，以測試其強健性

(robustness)；我們認為挑選重要詞彙與重要位置兩個摘要方法，應用於中文上應該會有不錯的摘要成效。不過，由於目前在中文方面，沒有一部類似 WordNet 架構的辭典。因此，可能無法使用擴充標題詞彙方法來對中文文件進行摘要處理。

本文的摘要研究，主要是以英文文件為對象，進行單文本的摘要處理。然而，隨著網際網路的盛行，使得我們可以輕易取得多樣化、多語言的資訊。因此，我們希望在未來，能夠把研究範圍擴展至多文本的摘要處理，甚至多語言摘要處理。

參考文獻

- Barzilay R., K. R. McKeown and M. Elhadad, "Information Fusion in the Context of Multi-Document Summarization," In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999, pp. 550-557.
- Chen H. H. and S. J. Huang, "A Summarization System for Chinese News from Multiple Sources," In Proceeding of the 4th Information Retrieval for Asia Language 1999, pp. 1-7.
- Edmundson H. P., "New Methods in Automatic Extracting," Journal of the ACM, Vol. 16, No. 2, 1969, pp. 264-289.
- Fellbaum C., WordNet : An Electronic Lexical Database, The MIT Press, 1998.
- Forsyth and Rada, "Adding an Edge in Machine Learning: Applications in Expert Systems and Information Retrieval," Ellis Horwood Ltd, 1986, pp. 198-212.
- Hayes P. J. and S. P. Weinstein, "Construction A System for Content-based indexing of a database of new stories," In Proceedings of 2nd Annual Conference on Innovation Applications of AI, 1990.
- Hovy E. and C. Y. Lin, "Automated Text Summarization in Summarist," Advances in Automatic Text Summarization, The MIT Press, 1999.
- Hovy E. and C. Y. Lin, "Identifying Topic by Position," In Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP), Washington, DC, 1997.

- Ker S. J. and J. N. Chen, "A Text Categorization Based on Summarization Technique," In Proceeding of NLPIR Workshop of ACL2000, 2000, pp. 79-83.
- Lewis D. and M. Ringuette, "Comparison of two Learning Algorithms for Text Categorization," In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.
- Mani I. and E. Bloedorn, "Summarizing Similarities and Difference Among Related Documents," Information Retrieval, Vol. 1, No. 1, 1999, pp.35-67.
- Marcu D., "Discourse Trees are Good Indicators of Importance in Text," Advances in Automatic Text Summarization, The MIT Press, 1999.
- McKeown K. and D. R. Radev, "Generating summaries of multiple news articles," In Proceedings on the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995, pp. 74-82.
- Miller G. A., R. Beckwith, C. Fellbaum, D. Gross and K. Miller, "Introduction to WordNet: An On-line Lexical Database," In Proceedings of the fifteenth International Joint Conference on Artificial Intelligence, 1993.
- Woods W. A., "Conceptual Indexing: A Better Way to Organize Knowledge," Sun Labs Technical Report: TR-97-61, editor, Technical Reports, 1997.
- Yang Y., "An Evaluation of Statistical Approaches to Text Categorization," Information Retrieval. Vol. 1, 1999, pp. 69-90.
- 邱中人, "中文新聞摘要", 碩士論文, 清華大學資訊工程系, 2000.

使用關聯法則為主之語言模型於擷取 長距離中文文字關聯性

Association Rule Based Language Models for Discovering Long Distance Dependency in Chinese

簡仁宗 陳鴻儀

國立成功大學資訊工程學系

Email : jtchien@mail.ncku.edu.tw

摘要

本論文提出一種能擷取長距離資訊的語言模型，它可以擷取多詞彙之間的關聯性，擷取的方式是使用資料探勘中十分流行的 Apriori 演算法，傳統上 n-gram 語言模型只能在 n-gram 視窗內擷取到有限距離的資訊，較長距離的資訊也就因此而流失，然而這些失去的長距離資訊對於語言模型是十分重要的，所以如何克服 n-gram 模型缺乏長距離資訊一直是非常熱門的研究課題，觸發序對就是其中一種有效的方法，其主要功能是在擷取長距離之詞序對資訊，也就是建立起詞與詞之間的關聯性，然而我們所提出的關聯法則技術能擷取多元詞組間的關聯性，可以說是進一步改良詞組數並建立更長距離資訊，而實驗結果也顯示本論文方法比起傳統觸發序對獲得較低的 perplexity，此關聯法則技術也可以有效的與其他模型調整及模型平滑化的技術結合，在語言模型的效率改善方面能有更良好的效果，最後本論文也將提出的語言模型成功的應用在語音辨識與文件分類上，並建立一套個人化之新聞瀏覽器之展示系統。

1. 簡介

拜硬體技術不斷進步的貢獻之下，一般人會很理所當然的使用自動櫃員機提款或是利用自動空調設備來控制室內的溫度，而這都是由於電腦的自動化管理讓生活變的如此便利，正所謂“科技始終來自人性”，推動科技進步的那隻幕後

的黑手就是建立在“使人便利”的基礎之上，但是電腦自從在發明之初就存在一個與人性背道而馳的缺點，與它們的溝通需要透過一個特定的按鍵裝置，比方說要與個人電腦溝通就必須透過鍵盤或滑鼠等裝置，事實上這是使許多人對電腦望之卻步的原因，要學習如何使用鍵盤與電腦做溝通就等於是強迫人去學習一種“電腦語言”，這與“使人便利”的原則當然是互相違背的，但是反過來說如能讓電腦學習人類的語言，使電腦能更接近人類，也就能使其與人類生活的結合更加緊密，進一步如果電腦能透過語言的學習而具備了閱讀的能力，我們就可以讓電腦為我們過濾亦或分類每天所需閱讀的文件，比方說可以應用在於 e-mail 廣告過濾或是新聞文件分類等等，就可以讓電腦為我們省下更多的時間。

要克服電腦與人在語言上的鴻溝，在語言技術的領域有了聲學模型(acoustic model)與自然語言模型(natural language model)的產生，而這兩項技術的發展在國外已經行之有年，台灣自西元一九八二年起便開始有了中文聲學模型方面的研究，許多研究單位包括台清交成等大專院校，以及工研院、交通部、中研院、中華電信等都積極的投入研究的工作並且擁有了十分豐碩的研究成果，而在聲學模型已日益成熟的基礎下，自然語言模型的發展也備受矚目，誠如前文所述，語音技術發展的最終目的就是要將電腦與人類的溝通便利化，而要達到這個目的，將語音模型與自然語言模型做結合是必須的，我們的論文主要就是著墨於自然語言模型的探討，我們將會對自然語言模型中的一項十分成功且廣泛運用的技術 n-gram 語言模型做介紹，並且分析其在傳統上的缺點與改進技術，而本論文也將會針對 n-gram 模型其中一項缺點-長距離資訊的缺乏，提出一套新的改進方法，並且結合其他改進方法，進而發展出一套較有效率的 n-gram 模型，我們將會將其應用在結合聲學模型做語音辨識和文件分類的領域之上，期望對其正確率有一定幅度的改善。

而自然語言模型方面在現今有許多不同的發展，依其內容主要分為三個方向，一、根據語言學所發展出的文法(grammar)分析，二、以知識為基礎而發展的語言資料庫，三、著重於統計而發展出的 n-gram 模型。而我們主要是著墨於統計式的 n-gram 模型，在第二章中，我們將對 n-gram 模型做詳細的介紹，並對其缺點加以探討，第三章中將會介紹傳統上針對 n-gram 模型的缺點所衍生出的改進方法，並且提出一種能擷取長距離資訊的語言模型，將它應用在語音辨識或

新聞文件分類的系統上有一定幅度的幫助。

2. n-gram 語言模型簡介

目前 n-gram[11]模型的探討於各相關學術會議及期刊論文上已有相當多的文獻發表，顯示各種研究機構對此一領域的發展有相當大的期許，故投身於其中，而在各方都致力於改進 n-gram 模型之下，n-gram 模型在效能上已獲得相當不錯之成果，在本章中我們將會對 n-gram 模型的基本概念做一簡單之介紹。

2.1 n-gram 模型之應用

一般而言 n-gram 語言模型通常應用於貝式分類器(Bayes classifier)，扮演著事前機率(priori probability)或是可能性(likelihood)的角色，以語音辨識為例子而言，假設有一段聲學訊號(acoustic signal) X ，我們的目標是去找尋出此訊號最有可能的對應文句(sentence) S^* ，使用貝式分類架構是找出最佳事後機率的文句

$$S^* = \underset{S}{\operatorname{argmax}} P(S|X) = \underset{S}{\operatorname{argmax}} P(X|S) \cdot P(S) \quad (1)$$

其中 n-gram 模型 $P(S)$ 扮演著事前機率的角，透過聲學模型計算可獲得一段文句的聲學模型分數 $P(X|S)$ ，再透過語言模型計算可獲得此文句的語言模型分數 $P(S)$ ，將兩機率相乘求得最佳化之文句，即為此聲學訊號最有可能之對應文句。

就文件分類的領域而言，給定一篇文件 d ，目標是去找尋此篇文件所屬的類別 c (category)，假設我們總共定義了 k 個類別，並且使用這些類別所屬的文件訓練好不同類別的 n-gram 模型 L_1, L_2, \dots, L_k ，使用貝式分類器求得此篇文件所屬的類別 c^* 可寫成

$$c^* = \underset{c}{\operatorname{argmax}} P(c|d) = \underset{c}{\operatorname{argmax}} P(d|c) \cdot P(c) \quad (2)$$

在這邊 n-gram 模型 $P(d|c)$ 扮演的是可能性量測的角色，透過語言模型機率計算可以獲得 $P(d|c)$ 的值，而假設所有類別出現的機率是均等，只要能使 $P(d|c)$ 最佳化的類別語言模型，即為此文件 d 為最有可能對應之類別。

2.2 n-gram 模型之建立

語言模型主要的功能是在評估一段文句出現的機率，假設有一文句 S 其長度為 T 並且是由一段詞序列 $W_1 W_2 W_3 \dots W_T$ 所組成，則 S 出現的機率可以寫成

$$\begin{aligned}
P(S) &= P(W_1, W_2, \dots, W_T) \\
&= P(W_1)P(W_2 | W_1) \dots P(W_T | W_1, W_2, \dots, W_{T-1}) \\
&= \prod_{i=1}^T P(W_i | W_1, W_2, \dots, W_{i-1})
\end{aligned} \tag{3}$$

但是此種方法在計每一個詞的條件機率時都要牽涉到前面所有的詞序列，使得計算量太大而無法實現，為解決這個問題所以有 n-gram 模型的產生，在 n-gram 模型中，它是假設一個詞出現的機率只跟前面 n-1 個詞有關，因此(3)式可以近似為

$$P(S) = P(W_1, W_2, W_3, \dots, W_T) \cong \prod_{i=1}^T P(W_i | W_{i-n+1}^{i-1}) \tag{4}$$

其中 W_{i-n+1}^{i-1} 代表 $W_{i-n+1}W_{i-n+2} \dots W_{i-1}$ 詞序列，如此一來使用 n-gram 可以大量節省計算時間與記憶體，讓實用性大為提高。而一般在建立 n-gram 機率模型 $P(W_i | W_{i-n+1}^{i-1})$ 最直覺的方法就是統計在詞序列 $W_{i-n+1}W_{i-n+2} \dots W_{i-1}$ 後出現 W_i 的次數再除以詞序列 $W_{i-n+1}W_{i-n+2} \dots W_{i-1}$ 在訓練文集中出現的次數，也就是

$$P(W_i | W_{i-n+1}^{i-1}) = \frac{C(W_{i-n+1}^i)}{C(W_{i-n+1}^{i-1})} = \frac{C(W_{i-n+1}^i)}{\sum_{W_i} C(W_{i-n+1}^i)} \tag{5}$$

其中 $C(W_i^j)$ 代表 W_i^j 在訓練文集中出現的次數。

2.3 n-gram 模型之評估

基本上在評估一個 n-gram 模型的效果時常使用 perplexity[12]這個評估標準，而事實上它是在計算機率模型的 entropy，entropy 在訊息理論上指的是將機率 P 乘以資訊 $-\log P$ ，應用在 n-gram 模型的評估則表示為：

$$\begin{aligned}
H_p &= -P \log P \\
&= -\lim_{Q \rightarrow \infty} \frac{1}{Q} \sum_{W_1 W_2 \dots W_Q} P(W_1 W_2 \dots W_Q) \log P(W_1 W_2 \dots W_Q)
\end{aligned} \tag{6}$$

其物理意義表示在計算一個 n-gram 模型的 entropy 時，必須先將詞典中的詞做組合，形成為無限長的詞序列 $W_1 W_2 \dots W_Q$ ，並且將所有的可能詞序列計算其機率與資訊的乘積後加總，即可得到此 n-gram 模型的 entropy。但事實上不容易實現如此複雜的計算，必須假設可以提供一段足夠長的詞序列來代表所有的詞序列組

合，這種假設在統計學上稱為此詞序列為 ergodic，故(8)式可改寫為

$$H_p = -\left(\frac{1}{Q}\right) \log P(W_1 W_2 \dots W_Q) \quad (7)$$

而 perplexity 的定義為

$$perplexity = 2^{H_p} \quad (8)$$

perplexity 代表了 n-gram 模型中的平均分支因數(average branching factor)，perplexity 越低代表 n-gram 模型在做機率評估時，所遇到的分支越少，也就是此模型的效率越好。

2.4 n-gram 模型的缺點

n-gram 模型長久已來就存在著三個重要的問題，也是研究 n-gram 模型的人一直努力的目標，我們分述如下：

1. 訓練文集與測試文集領域上之差距(domain mismatch)：

n-gram 模型在建立時，必須要有一訓練文集來統計出此模型的機率，因此 n-gram 模型受制於它的訓練文集，當訓練文集不平均時可能會使 n-gram 模型較偏向某種領域(domain)，假設我們的訓練文集是財經類的新聞，但是此 n-gram 模型的目的是用來測試政治新聞，那麼就會造成較大的誤差，在這方面通常會使用較一般化的平衡文集作為訓練文集來解決這個問題。但矛盾的是如果我們使用較為平衡的文集訓練出我們的 n-gram 模型，此 n-gram 模型用來測試某些特定領域的新聞是否恰當？事實上我們希望在測試政治新聞時我們的 n-gram 模型是偏向政治類的，測試財經新聞時 n-gram 模型是偏向財經類的，為了要完成這項需求，就必須對 n-gram 模型再做改進，使其具有調整之效果[3][4][5][7]。

2. 訓練文集不足(data sparseness)：

n-gram 模型在訓練時，並不能保證訓練文集能夠包含所有詞的組合，以至於所訓練出來的機率模型某些詞組相連的機率為零，或是因為訓練文集的不平衡，造成統計出來的機率模型並不夠一般化，而為了解決這個問題，就有平滑化技術的產生，在參考文獻[2]中對傳統上受歡迎的平滑化技術有詳盡的說明。

3. 長距離資訊(long distance)缺乏之問題：

n-gram 模型在計算上的優勢是在於它使用了 n-gram 視窗(n-gram window)做為基礎，節省了大量的記憶體與運算時間，但也因為使用了這個概念使得 n-gram 模型只能擷取到視窗之內的資訊，長距離的資訊就因此而流失，而這些流失的資訊很可能會造成 n-gram 模型測試時相當程度的誤差，故如何擷取長距離的資訊一直都是 n-gram 模型中相當受到矚目的研究的課題。一般而言目前 n-gram 模型的研究均以解決此三項問題為主，本論文針對上述第三項長距離資訊的擷取提出改進方法，期望能提昇 n-gram 模型的效果。

3. n-gram 模型改進方向

針對 n-gram 模型的問題已經有許多論文提出改善的方法，在本章中，我們將針對幾組熱門的解決方式做簡介。3.1 節是快取 n-gram 模型(cache n-gram model)與混合式 n-gram 模型(mixture n-gram model)[3]的介紹，此項技術是為了要使 n-gram 模型更符合測試文集領域所發展出來。3.2 節我們將介紹一個在平滑化技術上十分受到歡迎且有效的 Witten-Bell 平滑化技術[13]。3.3 節是對於觸發序對(Trigger pair)[8][9]的簡介，觸發序對是在擷取長距離資訊的一種有效的方法，可以用來補償 n-gram 模型長距離資訊的不足，而本論文也將提出一種改進觸發序對的方法為對照組，並在實驗中做比較研究。

3.1 快取(cache)n-gram 模型與混合式(mixture)n-gram 模型

為了要使 n-gram 模型能夠更符合測試時的領域，所以產生了模型調整的概念，它的概念是基於一篇文章或是一段文句會有一個近似的主題，比方說棒球類的新聞就比較偏向運動類的領域，與其他類別的新聞(如財經新聞)就有一段相當大的差距，而希望能在做測試時，利用文章前面出現的資訊，動態的調整我們的 n-gram 模型，使得我們的模型更能符合我們測試文集的領域，基於此種概念，就有快取模型與混合式模型的技術產生。

快取 n-gram 模型顧名思義就是相同的詞序列會在鄰近的時間點上不斷出現，比方說我們的測試文集是一篇有關金融股的新聞，也就是說此篇文件“金融股”這段詞序列會不斷出現，透過我們的詞典，會將此詞序列斷詞為“金融”與“股”

兩個詞，此時若我們在第一次測試到此詞序列時，將“金融”後面接“股”的機率提高，自然可以增強我們模型的準確性，在快取模型中會保留一塊快取記憶體，而做文件測試時，會將最近測試過的文句拿來訓練出快取 n-gram 模型 P^c 將其與原始的統計模型 P^s 做結合，我們將模型機率用(9)式表示

$$P(W_1 W_2 \dots W_T) = \prod_{i=1}^{T+1} [(1-\mu)P^s(W_i | W_{i-n+1}^{i-1}) + \mu P^c(W_i | W_{i-n+1}^{i-1})] \quad (9)$$

其中 μ 代表結合比重。

而在本論文中，我們使用的是文句階層混合式 n-gram 模型(sentence-level mixture n-gram model)，在每經過一文句後，就利用此文句所提供的資訊調整混合模型的比重參數。我們是利用奇摩網站已分類好的新聞，做為我們的分類群組。而我們會依據分群過後之文集訓練出對應於各群組之 n-gram 模型，在這邊以 P_k 表示第 k 個群組的 n-gram 模型，而在做測試時，使用權重 λ_k 將各群之模型做組合成為測試用的 n-gram 模型，也就是說假設有一文句 S 為 $W_1 W_2 W_3 \dots W_T$ ，則此文句出現的機率為

$$P(S) = \prod_{i=1}^{T+1} P(W_i | W_{i-n+1}^{i-1}) = \prod_{i=1}^{T+1} \sum_{k=1}^m \lambda_k P_k(W_i | W_{i-n+1}^{i-1}) \quad (10)$$

其中 m 代表混合數個數，但為此模型還須做兩點改進，第一、為了避免每個群組中的訓練文集太少，造成資料稀疏(data sparseness)，每個單一群組模型需要再結合一個一般化的模型(general model)，用以增加模型的可靠度，第二、在測試時可能會有無領域(nontopic)的文集存在，所以我們又必須將一般化模型加入，視為一個無領域的群組，在此我們將一般化模型以 P_G 表示，故上式可改寫為

$$P(S) = \sum_{k=1}^{m,G} \lambda_k \prod_{i=1}^{T+1} [\alpha_k P_k(W_i | W_{i-n+1}^{i-1}) + (1-\alpha_k) P_G(W_i | W_{i-n+1}^{i-1})] \quad (11)$$

其中 α_k 為第 k 個群組模型與一般化模型的組合權重。在混合式 n-gram 模型中，有兩個權重 α_k 及 λ_k 存在，基本上混合式 n-gram 模型是依據前文來動態的調整此二權重，在初始時會使用少數保留文集估測出其初始值，測試時會在每一文句結束時再去做一次權重的調整，而調整的動作可以分別寫成(12)(13)式

$$\alpha_k^{new} = \frac{1}{\sum_{l=1}^{N_k} T_l} \sum_{l=1}^{N_k} \sum_{i=1}^{T_l} \frac{\alpha_k^{old} P_k(W_i | W_{i-n+1}^{i-1})}{\alpha_k^{old} P_k(W_i | W_{i-n+1}^{i-1}) + (1-\alpha_k^{old}) P_G(W_i | W_{i-n+1}^{i-1})} \quad (12)$$

其中 T_l 代表在文句 l 的詞數， N_k 表示在群組 k 的總文句數。

$$\lambda_k^{new} = \frac{1}{N} \sum_{i=1}^N \frac{\lambda_k^{old} P_k(W_1, \dots, W_{T_i})}{\sum_{j=1}^{m,G} \lambda_j^{old} P_j(W_1, \dots, W_{T_i})} \quad (13)$$

其中 N 代表調整的總文句數。權重的調整的主要根據測試時文件所出現的資訊，混合式 n-gram 模型會依前文在每個群組模型出現的機率為權重，動態的調整測試模型的組合權重，比方說在測試文件中不斷提到金融消息，混合式 n-gram 模型就會將模型逐步的調整到財經領域，再利用這調整過後之 n-gram 模型繼續測試後面的文句，然後再將測試而得的新資訊繼續做調整，這種遞迴式的做法是一種稱為資訊結構(Information structure)的概念。

3.2 Witten-Bell 平滑化技術

在平滑化問題上，我們引進了十分廣泛應用且受到歡迎的 Witten-Bell 平滑化技術[13]做為加強我們 n-gram 模型的基礎，平滑化技術主要建立於將 n-gram 模型中沒有訓練到的詞序列機率模型使用(n-1)-gram 模型做補償，也就是

$$P_{\text{interp}}(W_i | W_{i-N+1}^{i-1}) = \lambda_{W_{i-N+1}^{i-1}} P(W_i | W_{i-N+1}^{i-1}) + (1 - \lambda_{W_{i-N+1}^{i-1}}) p_{\text{interp}}(W_i | W_{i-N+2}^{i-1}) \quad (14)$$

這是一個遞迴式的定義，所有的 n-gram 模型都必須利用(n-1)-gram 模型做補償，其中 $\lambda_{W_{i-N+1}^{i-1}}$ 代表的是合併 n-gram 與 (n-1)-gram 之權重，而 Witten-Bell 平滑化技術對此一權重有一個特殊的估測方式，在這邊先對符號做以下的定義

$$N_{1+}(W_{i-n+1}^{i-1}, \cdot) = |\{W_i : C(W_{i-n+1}^{i-1} W_i) > 0\}| \quad (15)$$

$N_{1+}(W_{i-n+1}^{i-1}, \cdot)$ 代表在 W_{i-n+1}^{i-1} 後可接的詞數，其中下標「1+」代表是連接一個詞以上之意。權重因數定義為

$$1 - \lambda_{W_{i-n+1}^{i-1}} = \frac{N_{1+}(W_{i-n+1}^{i-1}, \cdot)}{N_{1+}(W_{i-n+1}^{i-1}, \cdot) + \sum_{W_i} C(W_{i-n+1}^{i-1} W_i)} \quad (16)$$

即為 Witten-Bell 的 n-gram 模型建立方式，其物理意義表示在統計 W_{i-n+1}^{i-1} 出現次數時，如果 W_{i-n+1}^{i-1} 後面可接的詞數越少，我們給 $P(W_i | W_{i-n+1}^{i-1})$ 較大的權重，反之則使用較多的(n-1)-gram 做補償，假設在做 bigram 統計時，詞典中有一詞為「類神經」，我們發現在訓練文集中「類神經」後都接「網路」一詞，此時就不需要太多的 unigram 做補償，這是因為此名詞有獨特性，後面幾乎都接少量特定的詞，而若欲統計一詞「幾乎」後可接詞的 bigram 機率，可能會發現訓練文集中其後可接

的詞非常多，此時 unigram 的權重可以適度加大，以彌補可能較多的資訊損失，使語言模型的準確性提高。

3.3 觸發序對 (Trigger Pair) 演算法

在自然語言中，存在著許多高度關聯性的詞組，比方說“醫生、“護士”或是“陽光”、“熱”等就經常出現於同一句子之中，但由於它們通常在句子中並不相連，所以 n-gram 模型並沒有辦法擷取到這些詞之間的關聯資訊，因此就有了觸發序對的產生，觸發序對的設計主要在於解決長距離資訊彌補 n-gram 模型的不足的問題，觸發序對由於其沒有演算法與資料結構可以快速的對資料庫做求取，故觸發序對會限制本身為“序對”、即若有一辭典 V ，觸發序對會對其中所有可能的詞序對做考慮，如此一來可將促發序對的總個數控制於 $|V|^2$ 內。

在統計觸發序對之前，我們必須訂定一個觸發序對的視窗大小，而觸發序對的選取主要是依據平均相互資訊(average mutual information)，簡稱 AMI ，它是用來評估兩個詞 W_i 和 W_j 之間的關聯性大小， AMI 以數學式表示如下

$$\begin{aligned}
 AMI(W_i; W_j) = & P(W_i, W_j) \log \frac{P(W_i, W_j)}{P(W_i)P(W_j)} + P(W_i, \overline{W}_j) \log \frac{P(W_i, \overline{W}_j)}{P(W_i)P(\overline{W}_j)} \\
 & + P(\overline{W}_i, W_j) \frac{P(\overline{W}_i, W_j)}{P(\overline{W}_i)P(W_j)} + P(\overline{W}_i, \overline{W}_j) \frac{P(\overline{W}_i, \overline{W}_j)}{P(\overline{W}_i)P(\overline{W}_j)}
 \end{aligned} \tag{17}$$

其中 $P(W_i, W_j)$ 代表 W_i 、 W_j 出現在同一視窗的機率， $P(W_i, \overline{W}_j)$ 代表在同一個視窗中只出現 W_i 而沒出現 W_j 的機率。透過 AMI 評估標準，我們將其選為觸發序對，以符號 $(W_i \rightarrow W_j)$ 表示。當序對選取完畢後，必須要對每個觸發序對計算其相互資訊 MI (mutual information)，用對數表示之如下

$$MI(W_i; W_j) = \log \frac{P(W_i, W_j)}{P(W_i)P(W_j)} \tag{18}$$

如果 W_i 和 W_j 是相互獨立的話，則 $MI(W_i, W_j) = 0$ ，相互資訊反映了觸發序對中兩個詞相互間的資訊變化。而觸發序對並無法單獨使用[14]，因為它只能反映出詞與詞的資訊變化，所以我們必須將其與 unigram 做結合，如此一來所獲得的資訊

比起 n-gram 模型多了長距離的資訊，為了方便起見使用對數表示為

$$\log P(S) = \sum_{i=1}^T \log P(W_i) + \sum_{i=T}^2 \sum_{j=i-1}^{\max(1, i-ws)} MI-Trigger(W_j \rightarrow W_i) \quad (19)$$

其中 $\log P(W_i)$ 即為 unigram 模型機率， ws 代表 window size，現在在我們的論文中將 window size 定為文句長度，也就是說在我們論文中的觸發序對是文句階層的觸發序對(sentence-level trigger pair)，代表我們只能擷取同一文句中的觸發序對資訊。現在我們必須將觸發序對加入 n-gram 模型之中做為長距離資訊擷取之輔助，透過線性插補(linear interpolation)的方式，我們可以一權重 a_i 將其做合併，也就是

$$\log P_{MERGED}(S) = \sum_{i=1}^k a_i \cdot \log P_i(S) \quad (20)$$

其中 $0 \leq a_i \leq 1$ 且 $\sum_{i=1}^k a_i = 1$ ，在這邊我們有兩個模型機率存在

1. $P_1(S) = P_{n-gram}(S)$ 為 n-gram 模型對文句 S 所估測出之機率。

2. $P_2(S) = P_{MI-Trigger-pair}(S)$ 為觸發序對模型對文句 S 所估測出之機率。

透過(20)式的計算，我們可以使用觸發序對計算出一段文句的機率，且此機率有長距離資訊存在，比起傳統的 n-gram 模型在資訊擷取上為佳。

4. 關聯法則與其應用

在這邊我們引入了一個在資料探勘(Data Mining)領域受到十分廣泛運用的 Apriori 演算法[1]，此演算法可以用來建立關鍵詞的關聯法則，舉例而言，假設有一組交易紀錄資料庫，此資料庫記錄著每筆交易所包含的商品，關聯法則所要擷取的就是每個商品間的相互關係，也就是說我們想知道一筆交易出現了某種商品後，還有哪些商品是可能出現在同一筆交易紀錄之中，如果說商家從關聯法則中知道顧客買了商品甲後，還有很大的機率會去買商品乙，則可將商品甲與商品乙放在附近增加顧客的方便性與商家的業績。

4.1 Apriori 演算法

假設我們有一組新聞文件資料庫 D ，裡面包含了 $|D|$ 篇文章，每篇文章均是辭典 $L = \{w_1, w_2, \dots, w_n\}$ 的子集合，用上面的例子解釋就是各種商品的集合之意，而關聯法則以 $X \Rightarrow Y$ 的型式表示，其中 X 、 Y 均是 L 的子集合(subset)且互

相獨立，如果在所有包含 X 的文章中有 $c\%$ 同時也包含了 Y ，則我們可以稱關聯法則 $X \Rightarrow Y$ 存在於資料庫 D 中的信賴度(confidence)為 c ，此外若有 $s\%$ 的文章同時包含 X 與 Y ，則我們可稱關聯法則 $X \Rightarrow Y$ 以支持度(support) s 存在於資料庫 D 中，換句話說，信賴度是一種量測關聯法則強弱的標準，而支持度則是表示統計上出現的頻率，事實上我們實作時會訂定信賴度與支持度的門檻，我們擷取出來之關聯法則的信賴度與支持度均必須大於此門檻。

以下即為擷取關聯法則的演算法流程，是以資料探勘中的 Apriori 演算法做為基礎所改寫而成若我們以簡單的例子說明之，假設我們共有三詞，分別以 a 、 b 、 c 代表，Apriori 演算法就是在找尋此三詞的關聯性，它的概念就是先將這些詞兩兩為一組建立序對集合 (a,b) ， (a,c) ， (b,c) ，並且對資料庫搜尋每一序對，是否同時出現在於同一文句中，假設只有 (a,b) ， (b,c) 序對符合這項條件，則將 (a,c) 刪除，此時我們建立 (a,b) ， (b,c) 的關聯法則，此關聯法則的層級(step) 為二，不過我們必須計算其信賴度與支援度，例如我們可以計算同一篇文章出現詞 a 且出現詞 b 的機率，即為其信賴度。而我們會再將剩餘下之序對 (a,b) ， (b,c) 做結合成為 (a,b,c) 並搜尋訓練文集中 (a,b,c) 會不會同時出現於一文句中，如果沒有則刪除之，有則可以計算給定任二詞，出現第三詞的機率，此時稱關聯法則的層級為三，找尋出的關聯法則之層級是依其詞數而定，層級越大，代表其關聯法則中的字數越多。上述只是概念性的做法，Apriori 演算法事實上會做一些節省時間的動作，而其中最重要的部分就是建立雜湊樹(hash tree)的資料結構以節省更多的運算時間。

Apriori 演算法

- 1) $L_1 = \{ \text{words} \mid \text{counts is large than support threshold} \};$
- 2) for ($k = 2$; $L_{k-1} \neq \emptyset$; $k++$) do begin
- 3) $C_k = \text{unification}(L_{k-1}); // \text{produce new candidate sets}$
- 4) for all sentence f do begin
- 5) $C_t = \text{subset}(C_k, f); // \text{candidates contained in sentence } f$
- 6) for all candidate $c \in C_t$ do
- 7) Increment count of candidate c ;
- 8) end
- 9) $L_k = \{ c \in C_k \mid \text{count of candidate is large than support threshold} \}$
- 10) end
- 11) Answer = $\bigcup_k L_k$;

表一、關聯法則演算法

使用我們改變後的 Apriori 演算法，我們可以獲得詞與詞之間的關聯法則，而關聯法則的形式如下

$WordSeq \Rightarrow Y$ confidence = $c\%$ support = $s\%$ 代表出現詞序列

$WordSeq$ 的文章中有 $c\%$ 的機率會出現 Y ，而有 $s\%$ 的文章同時包含了 $WordSeq$ 與 Y 。以下為利用西元二千零一年十二月二十八號到西元二千零一年十二月三十一號期間的政治新聞所擷取出的兩條關聯法則範例，第一條關聯法則的層級為二，第二條的層級則為三

(1) 小三通 \Rightarrow 大陸

confidence = 90% support = 6.25%

(2) 大陸 小三通 \Rightarrow 兩岸

confidence = 100% support = 2.25%

以上例而言，出現“小三通”後，新聞語料有 90% 的機率也會出現“大陸”這個詞，這個關聯法則佔了總文句 6.25%，如果“大陸”與“小三通”同時出現後，新聞語料會有 100% 的機率也會出現“兩岸”，而此關聯法則佔了總文句 2.25%。

4.2 關聯法則為主之 n-gram 模型機率計算

為了使關聯法則更能反映語言模型的特性，我們將關聯法則做一稍微的變動，我們將其使用相互資訊的形式表示，就有如觸發序對一般對任一關聯法則 $WordSeq \Rightarrow B$ ，我們計算其相互資訊並用對數表示之

$$MI - Association(WordSeq; B) = \log \frac{P(WordSeq, B)}{P(WordSeq)P(B)} \quad (21)$$

如使用關聯法則為長距離資訊的輔助的 n-gram 模型機率，為了方便以對數表示為

$$\log P(S) = \sum_{i=1}^T \log P(W_i) + \sum_{i=1}^T MI - Association(W_1 W_2 \dots W_{i-1} \Rightarrow W_i) \quad (22)$$

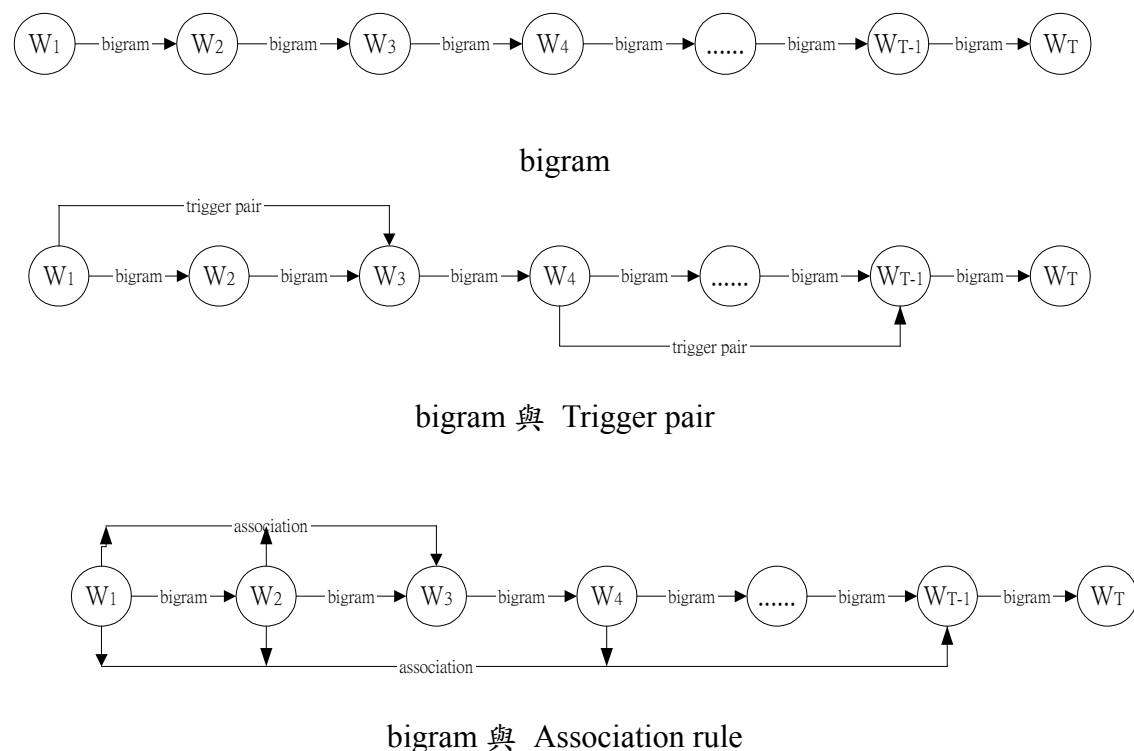
其中 $P(W_i)$ 代表 unigram 模型之機率。 $MI - Association(W_1 W_2 \dots W_{i-1} \Rightarrow W_i)$ 代表使用詞序列 $W_1 W_2 \dots W_{i-1}$ 所找出之最大層級之關聯法則 $WordSeq \Rightarrow W_i$ 的相互資訊，如同觸發序對一般 ws 代表 window size，在這邊我們也將 window size 定為文句長度，也就是說我們的關聯法則是文句階層的關聯法則(sentence-level

association rule)。如同觸發序對一般，我們要將關聯法則與傳統 n-gram 模型做結合，如同(20)式，此時

1. $P_1(S) = P_{n\text{-gram}}(S)$ 為 n-gram 模型對文句 S 所估測出之機率。
2. $P_2(S) = P_{MI\text{-Association}}(S)$ 為關聯法則模型對文句 S 所估測出之機率。

4.3 關聯法則與觸發序對之比較

使用關聯法則做資訊擷取與觸發序對最大的不同在於我們透過關聯法則可以獲得多元詞組(multi-word)之間的關聯性，而觸發序對只能擷取詞與詞之間(word pair)的關聯性，互相比較之下我們的方法是較強健的，圖一為關聯法則與觸發序對之示意圖，圖中箭頭表示關聯性。由圖中我們可以清楚的看出傳統的 bigram 模型只能由前面所出現的詞來對目前所出現的詞做機率評估，即 n-gram 模型文句間的關聯性是循序的，且受制於 n-gram 視窗之大小，而觸發序對則可跳脫此關聯性，只要是同一段文句中所出現的詞都可以有相互間的關聯性存在，不過觸發序對模型的限制在於只能擷取詞與詞之間的關係，而我們所提出之關聯



圖一、bigram 模型、Trigger pair 與 Association rule 之比較

法則序對則可以將關聯性擴大，變成多元詞組間的相互關係，可以說是觸發序對的延伸研究。

5. 實驗

5.1 實驗資料庫

為了將本論文方法時實現在中文系統中，首先必須製作了一套詞典，詞典中主要部分是 CKIP 中文詞庫[15]，它主要是利用國語日報辭典中約四萬目詞的原始資料加以分類，並且附加部分的語法及語意訊息在其中，本論文只使用到詞出現的頻率，並無使用到語法與語意資訊，取出其中一、二、三、四詞的部分作為我們的基本辭典，並不定期由人工更新我們這部詞典的新詞。

另外我們準備了兩組基本的實驗資料庫，第一是 CKIP 平衡語料庫，這是一個十分一般化的語料庫，有近二千萬字之現代漢語語料及超過五百萬字之古代漢語語料。另外我們在西元二千零一年四月十日及四月十六日擷取民視即時新聞、中央社、中時電子報、電子新聞網、聯合新聞網、ETtoday 與鉅亨網等網站的新聞文件共 3118 篇，包含有科技、社會、休閒、國際、體育、影視、政治及財經等八大類別，依日期將其區分為訓練文集(四月十日到四月十四日共 2234 篇)和測試文集(四月十五、十六日共 884 篇)。並且我們以 bigram 模型來驗證本研究方法。

5.2 不同語言模型之實驗結果

我們會對平滑化技術做評估，將傳統 n-gram 模型與加入 Witten-Bell 平滑化技術的 n-gram 模型做比較，在這邊我們使用 CKIP 平衡語料庫加上新聞訓練文集做訓練，並對測試文集做評估，表二的第一列為傳統 n-gram 模型所得之結果，第二列則為加入 Witten-Bell 改進後之結果。另外我們會在本小節中對混合式模型的效能做評估，在混合式模型方面，我們使用 CKIP 平衡語料做為一般化模型之訓練文集，並經由新聞訓練文集分類好的八個群組訓練群組模型，對測試文集做測試，所測得之 perplexity 如表二第三列所示。接下來是對觸發序對與原始模型的效能做比較，在這邊我們使用 CKIP 平衡語料庫加上新聞訓練文集做訓練，

並對測試文集做評估，所得之結果列於表二的第四列。

Bigram(Baseline)	258.8
Bigram + Witten-Bell	193.5
Bigram + Mixture n-gram	201.5
Bigram + MI-Trigger pair	237.5

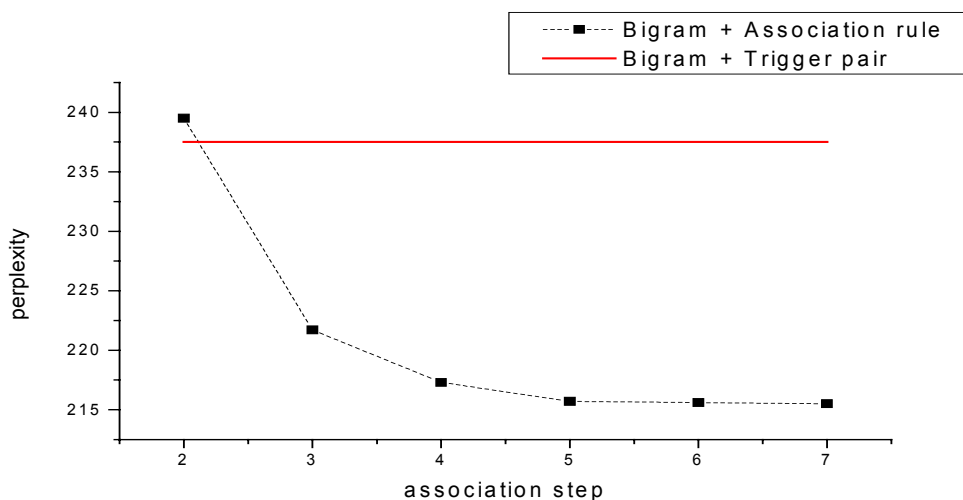
表二、不同語言模型改進技術之 perplexity 比較

5.3 關聯法則之模型效能

在本實驗中，一開始我們將會測試關聯法則依最大層級數對於 perplexity 之影響，在這邊我們使用 CKIP 平衡語料庫與新聞訓練文集做訓練，並且對測試文集做測試，表三即為所得之結果，圖二為其與觸發序對的圖形化表示，可以看出隨著最大層級的增加，perplexity 有一定程度的下降，由此項觀察可以證明我們所提出的相互資訊為基礎的關聯法則對於 n-gram 模型的改進有相當程度的改善，並且層級數較高的情形下，模型的效能優於觸發序對，而我們也發現最大層

Association Step	2	3	4	5	6	7
Bigram + Association	239.5	221.7	217.3	215.7	215.6	215.5

表三、依關聯法則最大層級不同所測得之 perplexity 比較



圖二、觸發序對與關聯法則不同層級所得之 perplexity 比較

級到達五就發生了飽和狀態，在此之後隨著層級數增加，perplexity 並不會有顯著的改善，故在本論文之後的實驗我們都將關聯法則的最大層級定為五，用以節省記憶體與運算時間。我們會將所有提到的技術相互結合做比較，在這邊使用為 CKIP 平衡語料庫加上新聞訓練文集做訓練，而後對測試文集做 perplexity 之評估比較，結果如表四所示，在這邊可以發現我們的方法可以有效的結合平滑化技術與混合式模型，對於 n-gram 模型的改進可以更進一步。

Bigram (Baseline) + Witten-Bell	193.5
Bigram + Mixture n-gram + Witten-Bell	178.3
Bigram + Witten-Bell+MI-Trigger Pair	168.8
Bigram + Mixture n-gram + Witten-Bell+ MI-Trigger pair	160.4
Bigram + Witten-Bell+ MI-Association	148.2
Bigram +Mixture n-gram + Witten-Bell+ MI-Association	135.8

表四、不同結合技術之 perplexity 比較

5.3 語音辨識之實驗

我們將 n-gram 語言模型與語音辨識的工作做結合，語音辨識是以隱藏式馬可夫模型(HMM)為基礎，特徵參數為二十六階語音特徵參數，由 12 階的 MFCC、12 階的 delta MFCC、delta log energy 與 delta delta log energy 所組成，語音訊號的取樣頻率為 8kHz，解析度為 16 bits，音框大小為 256 點(23.22ms)，音框位移大小為 85 點(7.74ms)。所使用的語音資料庫為 Mandrain Across Taiwan(MAT) 所提供的 MAT-160。測試語料由不確定人數之男性及女性所錄音之國語獨立詞與文句透過電話錄音共 500 句，供做便是測試用。表五為使用上述語料所做出之辨識結果，第一列(Baseline)為單純使用音節模型辨識技術所得之結果，第二列(Bigram)為音節模型辨識分數再加上語言模型辨識分數所得之結果，第三列(Bigram + MI-Trigger pair)則是先對語言模型使用觸發序對改進後所得之分數，再與音節模型分數合併所得之結果，第四列(Bigram + MI-Association) 則是先對語言模型使用關聯法則補償技術後所得之分數，再與音節模型分數合併所

得之辨識率，上述語言模型與聲學模型分數之合併比重均為 1:1，並且語言模型事先都經過 Witten-Bell 平滑化技術解決其平滑化問題。

Baseline	51.33
Bigram	51.86
Bigram + MI-Trigger pair	52.31
Bigram + MI-Association	52.92

表五、不同語言模型所測得之音節正確率(%)

5.4 文件分類之實驗

在本實驗中，我們將透過所提出的改進方法對文件分類的工作做模擬，在這邊我們使用新聞訓練文集對八種不同的領域分別訓練出 n-gram 模型，成為原始模型，另外加入觸發序對與關聯法則成為改進後之模型，由此二模型為基礎進行文件分類模擬工作之正確率比較，表六為所得之結果，由表中可以觀察到我們改進過後的模型在文件分類的工作上比起傳統之 n-gram 模型在分類的正確率上有小幅度的改進，而我們認為改進幅度並不如預期的主要原因是由於我們所選取的新聞文件有未確定性(ambiguous)的問題，有些新聞文件在網頁上是分類於政治領域，但實際上若將其分類於財金領域也未嘗不可，這些文件造成我們在模擬分類的實驗時錯誤率的增加。

	科技	社會	休閒	國際	體育	影視	政治	財經	平均
Bigram(Baseline)	64.8	77.1	69.8	72.6	84.6	75.4	86.9	72.1	75.4
MI-Trigger pair	66.6	78.3	69.6	70.9	85.2	76.0	86.9	74.8	76.0
MI-Association	66.8	78.3	70.8	71.9	85.2	75.8	88.6	75.1	76.6

表六、不同語言模型所做之文件分類正確率(%)

5.5. 個人化新聞文件瀏覽器

在論文最後，我們將以自然語言模型為發展基礎，透過模型機率的評估，對

於網路新聞文件做線上分類，並依據個人閱讀的習慣，建立出一套符合個人需求之新聞文件瀏覽器，期望能藉由這套系統增加一般人在閱讀新聞文件上的效率，圖三即為此瀏覽器在視窗上的執行時的畫面，藉由畫面中的“文件更新”按鈕，此瀏覽器會透過網路獲得最新之新聞文件，而“啟動學習”的按鈕則會將閱讀過的文件資訊加入語言模型之中，最後由語言模型的預測分數將新聞文件做排序，使用者可能較有興趣的新聞會優先列在標題欄內，假設使用者是一位籃球迷對於 NBA 的相關報導十分的關心，在六月十四日 NBA 總冠軍賽正打的火熱時在運動類新聞選取了幾篇相關的文件，如圖三所示，當此使用者在六月十五日瀏覽運動類新聞時，系統會優先將 NBA 相關的文件列於標題欄，如圖四所示。

6. 結論及未來研方向

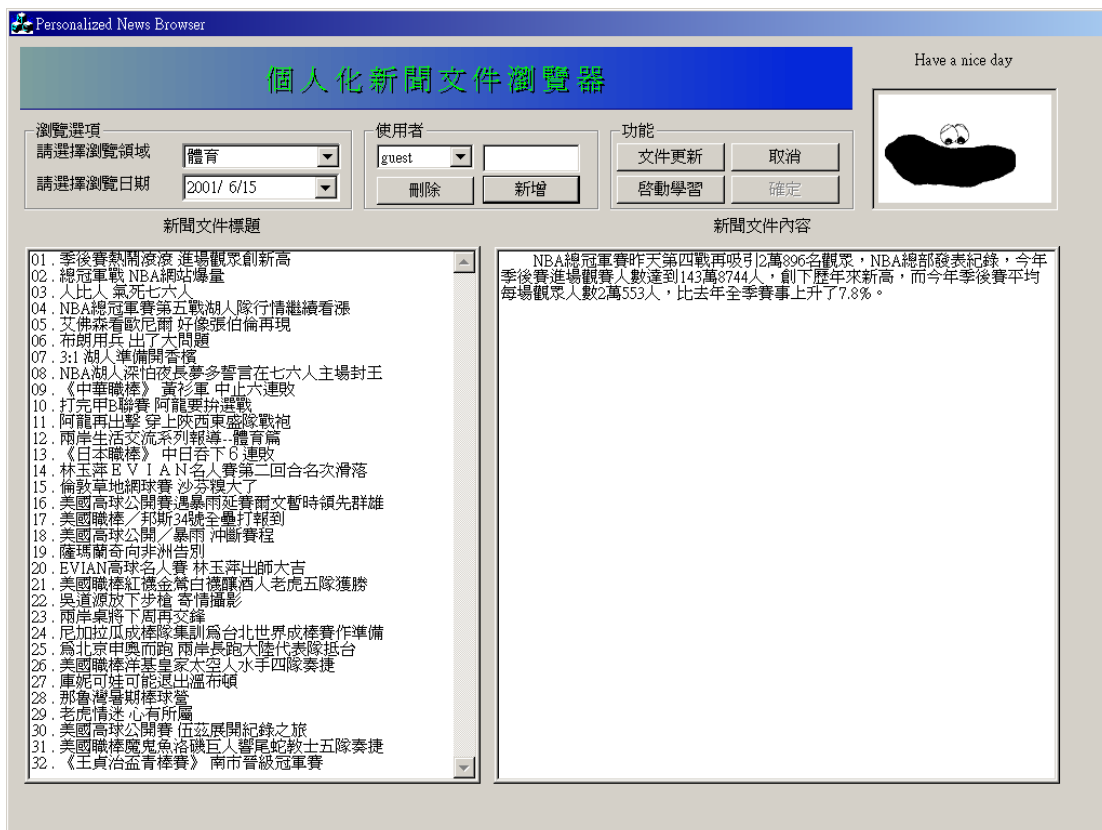
在本論文中我們對於傳統的 n-gram 模型做了完整的介紹，從模型的建立與評估到模型缺點的探討，都有一套完整的說明，而我們也針對每項 n-gram 模型的缺點介紹了近幾年來一些較為受歡迎的解決方式，包括了混合式 n-gram 模型、Witten-Bell 平滑化技術與觸發序對等，而在我們論文中的實驗，也證明了這些方法對於改進 n-gram 模型是十分有效的。

另外我們也在論文中提出了一個關聯法則的技術，此方法是透過一個在資料探勘上十分受歡迎的 Apriori 演算法，利用文句結構的特性，使用文句前面所提供的資訊來建立文句中前後文的關聯法則，將其用於 n-gram 模型的改善上可以得到不錯的效果，在實驗方面本方法可以有效降低 perplexity，證明我們的改進過後之 n-gram 模型比起傳統的模型效能為佳，最後我們將其應用在語音辨識與文件分類的工作上在正確率上也有一定幅度的改善。

未來在 n-gram 模型上的研究應不僅止於解決傳統 n-gram 模型上的缺陷，傳統上的 n-gram 模型是從統計的概念發展而出，嚴格來講並不是一個完整的自然語言，只能說是其中的一個重要的部分，未來必須有效的結合語言文法與知識背景的語言學才能算是真正的語言模型，而如何將三者融合[10]是一個十分困難



圖三、個人化新聞文件瀏覽器展示介面(2001/6/14)



圖四、個人化新聞文件瀏覽器展示介面(2001/6/15)

的問題，因為 n-gram 模型是機率模型，語言文法則是一個人工定的條例，並不是一組機率模型，有人透過剖析(Parsing)將其機率分析而出，這是一門重要且艱深的學問，而知識背景的語言學又更複雜了，只有人工定的分數沒有機率模型，要將其與 n-gram 模型結合則又必須花費更多的功夫，但是唯有克服這個困難，才能夠大幅度的提昇語言模型的效率。

參考文獻

- [1] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, Proceedings of the 20th VLDB Conference, Santiago-Chile, pp.487-499, 1994。
- [2] S. F. Chen and J. Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling”, Computer Speech and Language, vol.13, 359-394, 1999。
- [3] P. R. Clarkson and A. J. Robinson, “Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache”, Proc. of ICASSP, pp.799-802, 1997。
- [4] R. Iyer and M. Ostendorf, “Relevance weighting for combining multi-domain data for n-gram language modeling”, Computer Speech and Language, vol.13, pp.267-282, 1999。
- [5] R. M. Iyer and M. Ostendorf, “Modeling long distance dependence in language : Topic Mixtures Versus dynamic cache models”, IEEE Transaction on speech and audio processing, vol.7, January 1999。
- [6] F. Jelinek and R. L. Mercer, “Interpolation estimation of Markov source parameters from sparse data”, Proceedings of the workshop on pattern recognition in Practice, North-Holland, Amsterdam, The Netherlands, pp.381-397, May 1980。
- [7] D. Klakow, “Selecting Articles from the Language Model Training Corpus”, Proc of ICASSP, pp.1695 –1698, 2000。
- [8] R. Rosenfeld, “A maximum entropy approach to adaptive statistical language model”, Computer Speech and Language, vol 10, pp.187-228, 1996.
- [9] R. Lau, R. Rosenfeld, and S. Roukos, “Trigger-based language models: A

- maximum entropy approach” , in Proc. Int. Conf. Acoustics, Speech, Signal Processing, vol. II, pp. 45–48. , 1993
- [10] M. Meteer and J. R. Rohlicek, “Statistical language modeling combining N-gram and context free grammars” , in Proc. Int. Conf. Acoustics, Speech, Signal Processing, vol. II, pp. 37–40 , 1993.
- [11] C. D. Manning, H. Schutze, “Foundations of statistical natural language processing”, Massachusetts Institute of Technology pp.315-407, 1999 ◦
- [12] L. Rabiner and B.H. Juang, “Fundamental of Speech Recognition”, Prentice Hall, pp.321-387, 1993 ◦
- [13] I. H. Witten and T. C. Bell, “The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression.”, IEEE Transactions on Information Theory , vol.37, pp.1085-1094, 1991 ◦
- [14] G. D. Zhou and K. T. Lua, “Interpolation of n-gram and mutual-information based trigger pair language models for Mandarin speech recognition”, Computer Speech and Language, vol. 13, pp.125-141, 1999 ◦
- [15] CKIP, <http://godel.iis.sinica.edu.tw>, 中央研究院資訊科學研究所詞庫小組 ◦

多篇文件自動摘要系統

沈健誠，張俊盛

清華大學資訊工程研究所

mr884354@cs.nthu.edu.tw, jschang@cs.nthu.edu.tw

摘要

目前大部分的摘要系統為單篇文章摘要系統，雖然能提示個別文章的要點，卻無法把性質相近的文章集成摘要。能否夠發展一個多篇文章摘要系統，將敘述相同事件的文章統合成一篇摘要？如此一來，兩三個句子就能把文章的文意清楚而簡潔的表達出來，讓使用者能在一分鐘之內，明瞭這幾篇文章是否符合資訊需求，以縮短其蒐集的時間，更有效率的吸收網路上的大量資訊。

我們的目標在於發展一個多篇文章摘要系統，系統所產生的摘要能滿足以下兩個條件：指示性簡單摘要，和查詢主題相關，能因應使用者的查詢而有所改變。

為了達成此目標，我們將探討句子的指示性和查詢主題相關性，並選出重要性高而且相互獨立的句子，然後將不重要的小句刪除，以得到最終摘要。

我們針對 NTCIR 的 248 篇文章和 50 個查詢標題作實驗，所得到的摘要縮減比率為 95% 以上。整體而言，產生的摘要都能指示出幾篇相關新聞以及查詢主題的要旨。

1. 簡介

1.1 研究動機與目的

隨著網際網路的蓬勃發展，網路資訊（如電子報，政府公告，企業概況報導）的數量與日遽增，常常會有不同文章描述相同事件的情況，造成讀者無謂的時間浪費。摘要系統能夠將文章濃縮成精華片段，讀者只要閱讀摘要，就能瞭解此篇文章所敘述的事件和目的。因此，好的摘要系統能夠縮短使用者的閱讀時間，讓使用者在短時間內閱讀更多文章，吸收更多有用的資訊，精確地得到所需的情報。

目前大部分的摘要系統，針對單篇文章提供摘要，無法把性質相近的文章集成摘要，讀者仍然需要一篇一篇的篩選。如果能夠發展一個多篇文章摘要系統，將敘述相同事件的文章統合成一篇摘要，在兩、三個句子之內，把文章的文意清楚而簡潔的表達出來（提示性摘要），讓使用者能在一分鐘之內，明瞭這幾篇文章是否符合資訊需求，以縮短其蒐集的時間，更有效率的吸收網路上大量資訊。

因此，我們的目標在於發展一個多篇文章摘要系統，它所產生的摘要能滿足以下兩個條件：

1. 數篇文章的綜合摘要，而且是最簡潔的提示性摘要，最好能夠用兩三句就將這幾篇文章的主題事件描述出來。在目前的工商業社會中，資訊成長的速度遠超過想像，所以記者都會把新聞都寫得儘量簡短，只描述單一事件，並把文章的重點集中在某一段；往往只要簡短的一兩句，就能將文章的意義表達清楚。本系統的目的，便是從文章中找出足以代表整篇文章的句子，經過修飾後即成為真正的摘要。
2. 和查詢主題（topic）相關，能因應使用者的查詢而有所改變，以符合使用者的真正需求。

1.2 摘要如何產生

目前的摘要產生方式有兩種：摘述法 (Extraction) 與重述法 (Abstraction)。摘述法就是將文章中的關鍵句抽取出來，將其組合成為摘要；重述法則是做摘要的人將文章的要義以另外的文句寫下。目前大部分的摘要系統多採用摘述的方法，因為它比較簡單，只要句子取的好，就能達到指示文章的目標。

摘要的產生方向也有兩種：通用型 (Generic) 與查詢主題相關型 (Topic-Related)。對通用型摘要而言，不管使用者所在意的問題為何，相同文章一律產生相同的摘要；它比較固定，而且作業上比較簡單。而查詢主題相關摘要，是由使用者的查詢 (query 或 topic) 顯示出的使用者關心的部分所構成。

由以上敘述可得知，查詢主題相關摘要較符合使用者的需求，而利用摘錄方法所產生的摘要能夠兼顧效率與準確性。

1.3 多篇文件摘要系統的相關研究

摘要系統的研究，已行之有年；隨著網際網路的蓬勃發展，其重要性日益增高，但大多數的摘要系統是為單篇文件而設。多篇文件摘要系統的研究在最近一兩年來，逐漸引起大家的重視。McKeown 和 Radev(1998) 在它們的系統中除了地名，人名，組織名的辨識，還用新聞摘要語料庫，學習摘要的產生規則，最後利用文章合成技術 (Text Generation) 來產生摘要。

Mani 和 Bloedorn (1999) 以分析，重組，合成三步驟來產生摘要，並利用 WordNet 來計算句子間的關係，進一步尋找出各篇文章間的同和相異處；此外，他們還利用 Spreading Activation 演算法將與查詢主題相關的句子放入摘要中，進而產生查詢主題導向 (Topic-Oriented) 的摘要。

Chen 和 Huang (1999) 將中國時報，中央日報，中時晚報及工商時報網站上的新聞，以 Complete-Link clustering (Salton, 1989) 的方法，依事件分類，然後計算句子間的相似度，產生重點式與瀏覽式摘要。

有鑑於中文的 WordNet 尚未發展完全，目前也沒有中文新聞摘要的語料庫，所以我們希望能在沒有任何訓練及參考資料的情形下，以統計方式找出最合適的句子，並進一步將句子中不重要的部分剔除，完成最精簡的摘要。

2. 摘要的生成

2.1 摘要的條件

好的摘要應該簡潔、清楚，對文章有強烈的代表性及提示性。讀者光憑摘要，就能瞭解本文所描述的事件，輕易分辨這些文章是否為其所需。此外，摘要系統必須能從文章中挑出讀者想知道的部分；換句話說，摘要應該隨使用者的查詢需求而有所調整。

大部分的摘要系統，都是從文章中挑出關鍵句，將其合成摘要。選到適當的句子，摘要就成功了一半。評量句子的方法，大多為 tfidf 的變形，或是句子間的詞彙鏈結 (Lexical Chain)。多篇文章摘要的來源文章，彼此之間都有一定程度的相似，因此本系統以句子間的相似度為主。

2.2 摘要產生步驟

大多數的摘要，都是以選句為基礎，本系統亦不例外；在選句之前，要先做好斷句的工作；除此之外，句子間的字彙鏈結是以句中的動詞和名詞為主，所以必須有良好的斷詞和詞性標注 (Part-of-Speech Tagging)。為句子計算分數時，必須考量其提示性和查詢主題相關度。在選取句子的部分，我們希望摘要中的句子不但有高度的代表性，而且彼此獨立。最後，

我們將探討如何進一步把摘要中不重要的小句刪除，讓摘要更簡潔。

2.3 斷句斷詞

斷句方面，以句號（。），分號（；），驚嘆號（！）和問號（？）做為斷句的準則；換言之，一個完整的句子是以上述四個標點符號作結尾。而句子中的逗號（，）則為小句的分隔符號，在句長縮短的步驟中，我們考慮將不重要的小句剔除，進一步減少摘要的長度。

斷詞方面，則以 Chen (2000) 所發展的斷詞工具，對句子作斷詞和詞性標注，以名詞和動詞，作為句子評分的主要依據。

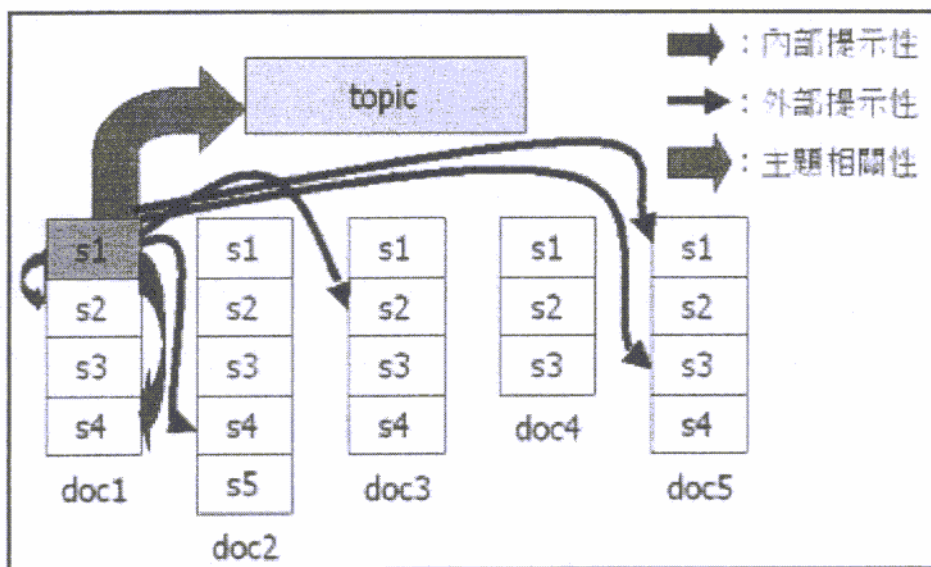
2.4 句子評分 (Sentence Scoring)

哪些句子才是關鍵句？很明顯的，為了回答這個問題，我們必須對句子作評分，決定何者為關鍵句；而關鍵句的考量有兩個方向：

1. 在此篇文章中的代表性：此句子是否足以代表本篇文章的意義。
2. 和查詢主題 (topic) 的相關聯程度：此句子是否和查詢主題有關係。

假設有一句子 S，其分數以兩方面來考慮：

- A. 提示性 (Indicativeness)：S 對其他句子的提示程度；對 S 所在文章內的其他句子的提示程度，稱為內部提示性，對其他相關文章內的句子提示程度稱為外部提示性。
- B. 主題相關性 (Topic Relevance)：句子與查詢主題的關聯程度。



圖一. 句子的提示性與查詢主題相關性

2.4.1 句子的提示性

要評斷一個句子提示其他句子的程度，最簡單的辦法便是計算兩句子的詞重複的數目。在大部分的情況下，句中的動詞和名詞（尤其是專有名詞）比較能夠代表整個句子的主要意義，因此在本系統中，我們以兩個字（含）以上的動詞與名詞作為計算相似度的基本單位；也就是說，如果兩個句子有一個以上的相同動詞或名詞，我們便說這兩個句子間有某種程度的相互提示性。

在本系統中，我們還將提示性分為內部提示性與外部提示性。內部提示性指的是此句對於同一文章內其他句子的提示程度，外部提示性則是此句對於其他文章內句子的提示程度。至於提示性的分數計算，我們採用三種方法：

2.4.1.1 方法 A1：

- 句子間的提示性 I_{ij} ：兩句子 S_i 與 S_j 之間的相關係數，若 S_i 和 S_j 在同一篇文章內，稱為內部提示性；若 S_i 和 S_j 在同一查詢主題（topic）

的不同文章內，則稱之為外部提示性

$$I_{ij} = \frac{V_{ij} + N_{ij}}{V_i + N_i}$$

V_{ij} ： S_j 和 S_i 內相同的動詞個數（重複出現不計）

N_{ij} ： S_j 和 S_i 內相同的名詞個數（重複出現不計）

V_i ： S_i 的動詞總數

N_i ： S_i 的名詞總數

在此公式之下，如果兩句子內同時出現的動詞與名詞數越高，兩者之間便有越高的提示性。為了正規化（normalize），因此必須以 S_i 的動詞術語名詞數總和當分母； S_i 中和 S_j 相同的重要字詞比例，即為 S_i 對 S_j 的提示性。

- 句子 S_i 的提示性分數：假設 S_i 對 m 個句子（ $S_j, j=1\sim m$ ）有提示性，則 S_i 的分數為 I_{ij} 的總和。提示性又分為外部提示性（ OI_i ）和內部提示性（ II_i ），其計算方式如下：

$$OI_i = \sum_{j=1}^m I_{ij}, S_i \text{和} S_j \text{在不同文章內}$$

$$II_i = \sum_{j=1}^m I_{ij}, S_i \text{和} S_j \text{在同一篇文章內}$$

例如：

S_1 = 『為配合司法院大法官會議四四五號解釋，並進一步保障人民集會遊行自由，內政部在會商相關部會後，已研議完成「集會遊行法」部分條文修正草案，除刪除現行集遊法第四條「集會、遊行不得主張共產主義或分裂國土」的條文，同時也以更為具體、明確的文字，完備集會、遊行之申請採取許可制的相關規定。』

S_2 = 『為配合司法院大法官會議解釋，並進一步保障人民集會遊行自由，內政部在會商相關部會後，已經研議完成「集會遊行法」部分條文修正草案。』

兩句共有 20 個相同動詞與名詞（其中”集會”在 S_1 出現四次，在實驗一中算成一個），而 S_1 的動詞和名詞總個數為 36，所以

$$I_{12} = 20 / 36 = 0.5556$$

2.4.1.2 方法 B1

- 句子間的提示性 I_{ij} ： S_i 和 S_j 為相異兩句

$$I_{ij} = \frac{N_{ij}}{\sqrt{N_i N_j}} + \frac{V_{ij}}{\sqrt{V_i V_j}}$$

N_{ij}, V_{ij} ：同時出現在 S_i 和 S_j 中的名詞（動詞）總數

N_i, N_j ：出現在 S_i (S_j) 中的名詞數目

V_i, V_j ：出現在 S_i (S_j) 中的動詞數目

- 句子 S_i 的提示性：同實驗一，假設 S_i 對 m 個句子 ($S_j, j=1\sim m$) 有提示性，則 S_i 的分數為 I_{ij} 的總和

以 S_1 和 S_2 為例， S_1 和 S_2 之間的提示性為

$$I_{12} = \frac{8}{\sqrt{16*8}} + \frac{12}{\sqrt{20*12}} = 0.7071 + 1.0607 = 1.7678$$

2.4.2 查詢主題相關性 (Topic Relevance)

本系統的目的是『產生與查詢主題相關的摘要』，因此除了考量句子本身對於文章的代表性，還要考慮句子跟查詢主題之間的相關性。我們以 NTCIR-2 查詢主題 Concepts 欄位中的詞為關鍵詞，並採用資訊檢索 (Information Retrieval) 的方式，來對句子的查詢主題相關性做評分。查詢主題的範例如下：

ID	SECTION	CONTENT
1	Number	CIRB010TopicZH001
1	Title	集會遊行法與言論自由
1	Question	查詢集會遊行法中有關主張共產主義或分裂國土規定之修正與討論。
1	Narrative	相關文件內容應敘述集會遊行法原本對主張共產主義或分裂國土之限制，其是否符合憲法中對言論自由等基本人權的保障，大法官對此議題的相關解釋，學者專家的討論與看法，以及集會遊行法條文的修改現況。
1	Concepts	集會遊行法、集會遊行、集遊法、憲法、言論自由、保障、共產主義、分裂國土、大法官會議、立法、修正條文。

表一. 查詢主題 (topic) 範例

在此步驟，我們參考資訊檢索的方式，採用了四個方法。

2.4.2.1 方法 A2

帶入如下公式

$$R_i = \frac{\# \text{ of (Terms in } S_i \text{ I Concept terms in topic)}}{\# \text{ of (Concept terms in topic)}} \times C$$

一個句子中出現了越多查詢主題內的關鍵詞，它和查詢主題的相關程度就越高。為了和句子的提示性分數平衡 (I_i 的值大多在 0~10 之間， R_i 原始值在 0~1 之間)，所以必須乘以常數 C 。(本系統中， C 的預設值為 10)，例如：

$S_2 =$ 『為配合司法院大法官會議解釋，並進一步保障人民集會遊行自由，內政部在會商相關部會後，已經研議完成「集會遊行法」部分條文修正草案。』

上述句子中包含了『集會遊行』，『集會遊行法』，『保障』，『大法官會議』四個 <topic 1> 的關鍵詞，而 <topic 1> 共有 11 個關鍵詞，因此其主

題相關性的評分為 $R_2 = \frac{4}{11} * 10 = 3.6364$

2.4.2.2 方法 B2

柏克萊大學在 TREC-2 (Text Retrieval Evaluation Conference) 中提出一種機率統計式的文件評分方法，在 TREC-5 (中文查詢) 與 NTCIR-2 (中文與英文查詢) 都有不錯的效果。在本方法中，引用此公式來表達句子 S_i 與查詢主題間的關連程度，其公式如下：

$$\begin{aligned} R_i &= \log O(R | S_i, Q) + K \\ &\approx \log \frac{P(R | S_i, Q)}{P(\bar{R} | S_i, Q)} + K \\ &\approx -3.51 + 37.4 * X_1 + 0.330 * X_2 \\ &\quad + (-0.1937) * X_3 + 0.929 * X_4 + K \end{aligned}$$

$P(R | S_i, Q)$: S_i 和查詢 (Query) 相關的機率

$P(\bar{R} | S_i, Q)$: S_i 和查詢 (Query) 不相關的機率

K : 爲了使分數大於零，所加上的常數，在本篇預設值爲5

X_1, X_2, X_3, X_4 : 此公式的四個參數，計算方法如下

$$\begin{aligned} X_1 &= \frac{1}{\sqrt{N+1}} \sum_{i=1}^N \frac{qtf_i}{ql+35} \\ X_2 &= \frac{1}{\sqrt{N+1}} \sum_{i=1}^N \log \frac{dtf_i}{dl+80} \\ X_3 &= \frac{1}{\sqrt{N+1}} \sum_{i=1}^N \frac{ctf_i}{cl} \\ X_4 &= N \end{aligned}$$

N : <文件> 和 <查詢> 中相符的字詞數

qtf : 字詞在 <查詢> 中出現的頻率

ql : <查詢> 所包含的字詞總數

dtf : 字詞在 <單一文件> 中出現的頻率

dl : <單一文件> 所包含的字詞總數

ctf : 字詞在 <所有文件> 中出現的頻率

cl : <所有文件> 包含的字詞總數

由於此公式適用於較長的整篇文件，因此用於長度較短的句子時，會有分數小於零的情況發生。為了和句子的提示性分數平衡，必須加

上一個常數 K (在本系統中, K 的預設值為 5)。

2.4.3 句子的分數：一個句子的重要性包含了內部提示性，外部提示性，與查詢主題相關性，計算公式如下：

$$score_i = r_{out} * OI_i + r_{in} * II_i + (1 - r_{out} - r_{in}) * R_i$$

OI_i ：外部提示性

II_i ：內部提示性

R_i ：查詢主題相關性

r_{out} ：外部提示性分數所佔的比例，介於 0~1 之間

r_{in} ：內部提示性分數所佔的比例，介於 0~1 之間

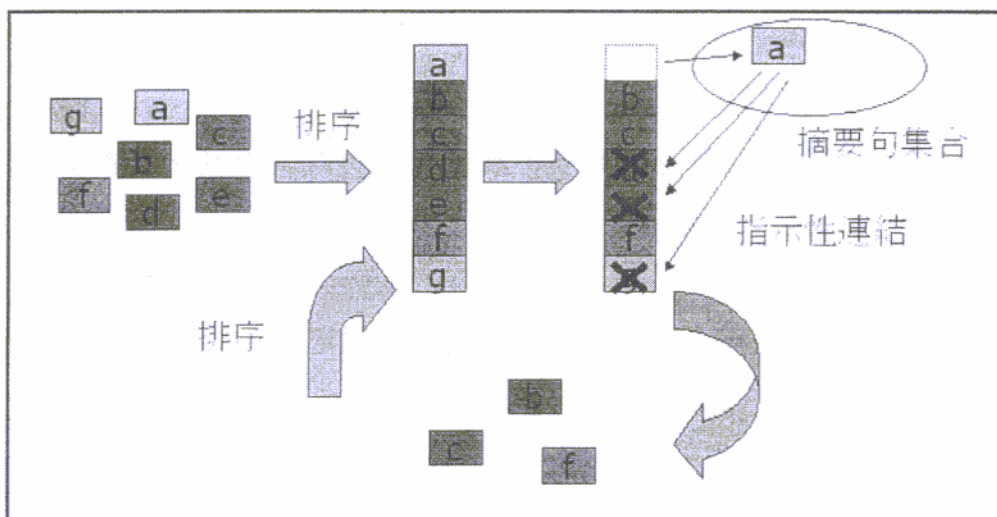
在本系統中，我們給予內部提示性，外部提示性與查詢主題相關度相同的重要性，因此 $r_{out} = r_{in} = 1/3$ 。

2.5 關鍵句選取 (Key Sentence Selection)

良好的摘要句應該兼顧『提示性』和『相互獨立』的原則；被選入的句子本身能夠對其他句子有強烈的提示性，足以代表其他句子和文章，而句子相互之間的提示性越少越好，避免重複敘述的情況發生。為達此目的，使用如下的方法：

1. 將同一群組內不同文章的所有句子，依照句子的分數由大排到小，形成一個句子名單 (list)
2. 把分數最大的句子 a 從名單中取出，放入摘要句集合中
3. 與 a 有一定程度相互提示性 (大於某個下限值) 的句子，從名單中移除
4. 重複步驟 1，將剩下的句子重新排列選取

如此一來，便可以選出具有高提示性又相互獨立的句子。



圖二. 選句方法示意圖

2.6 句子縮短 (Sentence Reduction)

為了進一步縮短摘要長度，減少讀者閱讀摘要的時間，我們必須將句子中無關緊要的部分刪除。通常一個句子內包含了數個小句，以逗號(,) 分隔之。如果某小句符合下列兩條件，則將其從摘要中剔除：

1. 毫無標題性：標題通常能代表文章的主要文意；如果以某字詞為開頭的小句常常出現在標題中，即具有標題性。若某小句的開頭為字詞 W ，則其標題性計算公式如下

$$T_w = \frac{\text{標題內的小句以}W\text{開頭的次數}}{\text{所有文章(包含文章標題)內的小句以}W\text{開頭的次數}}$$

在此我們以 NTCIR-2 的所有文章(共十三萬篇)為訓練資料，將所有小句的標題性算出，當小句的 $T_w \approx 0$ 時，表示毫無標題性，可視為不重要的小句。

2. 與查詢主題不相關：如果小句內不包含任何一個查詢主題的關鍵詞(Concepts 欄位內的字詞)，則認為與查詢主題不相關。

換言之，如果一個小句的標題性趨近於 0，且不包含任何一個查詢

主題關鍵詞，則將其從摘要中刪除之。除此之外，某些格式的小句可以直接刪去：

1. 括弧內的字：如【XXXXX】、(XXXXX)等文字。大部分的新聞文件，會在開頭標示『【媒體/地名/記者名/報導】』(例如『【記者賴廷恆台北報導】』)。此類句子跟文章的意義完全無關，可以直接刪除。而小括號內的文字往往做為前面名詞或動詞的註釋(例如『來自全世界各地對「金學」(即指對金庸小說的研究)學有專精的學者專家』)或是將相對時間修正為絕對時間(例如『馬友友昨(八)日展開忙碌的行程』)，一般而言，容易造成重複敘述的情況，違反了摘要的精簡原則，必須剔除。
2. 三個字以下的小句：如『畢竟，但是，因此，其次，基本上，事實上，不過，此外，另外，其實……』等小句，往往是連接詞或語氣詞，作為前後句的接續之用，沒有實質的意義。雖然刪除之後偶而會造成文句的不通順，但是對於摘要的『提示性』毫無影響，因此我們可以放心的刪除之。四個字以上的小句可能會包含『某人說』(如『卡特說』)，『某單位表示』，或內含動詞名詞的小句，因此予以保留。

例如：第六篇～第十篇(和<查詢主題 2>相關)原本的摘要如下：

既然台灣已經是一個主權獨立的國家，何以民進黨還要追求獨立呢？林義雄並為維持現狀下一定義就是不發生戰爭，不刺激中共打台灣，如果改國號會引發中共的攻擊，美國的反對，就是破壞現狀，當然會慎重處理。在年底立委選舉時，林義雄認為，各政黨都不可能迴避統獨議題，否則是不負責任，至於此議題對民進黨的利弊，林義雄表示難以判斷，但他相信，維護台灣安全的政黨，是能得到人民的支持，這方面民進黨比其他政黨更易得到人民的肯定。

上述摘要文章中，以黑體字標注的這幾個小句，不包含查詢主題的

關鍵詞，而且以『就是』、『當然』、『但』、『否則』為開頭的小句完全沒有在標題出現過，所以這幾個小句被視為不重要，可刪除之。經過句子刪除後，產生的摘要如下：

既然台灣已經是一個主權獨立的國家，何以民進黨還要追求獨立呢？林義雄並為維持現狀下一定義就是不發生戰爭，不刺激中共打台灣，如果改國號會引發中共的攻擊，美國的反對。在年底立委選舉時，林義雄認為，各政黨都不可能迴避統獨議題，至於此議題對民進黨的利弊，林義雄表示難以判斷，維護台灣安全的政黨，是能得到人民的支持，這方面民進黨比其他政黨更易得到人民的肯定。

經過小句刪除作業後的摘要比較簡短，而且大部分的情況下，文章尚屬流科，也不影響指示原文的效果。

3. 實驗資料與結果

3.1 實驗資料

本系統的目的為產生查詢主題相關連 (topic-related) 的多篇文件摘要，所以實驗資料必須包含文章與查詢主題 (topic) 的詳細描述。本系統專門針對中文查詢與文章，因此我們使用 NTCIR-2 的中文資料，針對每個查詢主題 (topic)，從大會公佈的標準答案中取出五篇文章；第 30 和 45 個查詢主題的標準答案只有四篇，所以實驗用的文章總篇數為 248 篇。

3.2 比較式評估

本文以人工評估，比較幾種計分方法所產生的摘要優劣及長度。比較項目包含了提示性的計分方式，查詢主題相關性的計分方式，關鍵詞是否分割成覆疊性雙連字，以及小句刪除之後的滿意度。

3.2.1 提示性計算方式的差異

我們在計算提示性時，採用了兩種不同計分方式。方法 A1 和方法

B1 的比較結果如下：

	較好的查詢主題 (topic) 編號	50 篇摘要總長度 (字數)
方法 A1 較佳	1,3,4,6,9,10,11,12,14,17,20,24,26,30,31,32,33,34, 37,39,42,48,50 (共 23 篇)	(方法 A1) 8073
方法 B1 較佳	2,5,7,8,13,16,18,19,23,25,27,28,35,38,40,41,43, 44,49 (共 19 篇)	(方法 B1) 8582
相同	15,21,22,29,36,45,46,47 (共 8 篇)	

表二 提示性的計算方式比較

在大部分的情況下，方法 A1 所選出來的句子較短，而且提示性和查詢主題相關性較高。

以 **topic 34**—威而剛的副作用為例，方法 A1 和方法 B1 所產生的摘要各為：

<方法 A1>：輝瑞並警告說，因胸痛而服用含有磷酸鹽藥物的病人，若再服用威而鋼甚至會致命。報導指出，目前已有三十名男子因為服用陽痿治療藥「威而鋼」不幸死亡，另外還有七十人發生嚴重副作用。輝瑞藥廠同時強調，全世界已有至少一百萬人服用過「威而鋼」，其中大多數是中年男性，該廠將與衛生部密切配合，就極少數死亡案例進行調查。

<方法 B1>：食品藥物管理局及輝瑞藥廠都表示，他們正在調查六名使用者死亡的原因，食品藥物管理局的聲明中說：「我們仍然相信這種藥物對於它的病症及病人安全有效」。報導指出，目前已有三十名男子因為服用陽痿治療藥「威而鋼」不幸死亡，另外還有七十人發生嚴重副作用。

在上述例子中，兩個摘要都有提到查詢主題的主要意義，但是方法 A1 所產生的摘要的字數較少，是比較好的摘要。

雖然在某些情況下，方法 B1 的摘要較好，但是所用的字數卻比較多，例如 **topic 7**—卡特訪台：

<方法 A1>：二十年來，美國和臺灣、美國和北京及臺灣和北京的關係，都有很大的進步。呂秀蓮認為，卡特應為台灣民主運動的受挫負責，並要求卡特就此對台灣人民道歉。卡特並認為每個國家都有自己的問題，美國不應該負這個責任。

<方法 B1>：卡特在昨日的離華記者會上仍然重申他當初決定與台灣斷交並沒有錯誤，他表示他在來台之前心理有些害怕，但是他認為要讓台灣的民眾了解為什麼他當初要做這個決定，現在雖然有許多人對於斷交的決定持不同的看法，但是卡特仍認為這個決定增進了美、中、台三方的關係。卡特說，中共當初承諾，要以和平方式解決兩岸問題，他不認為中共未來會對台灣採取軍事行動，他並重申，兩岸問題應由兩岸以和平方式解決，外人不適合，也不應該介入。

上述兩摘要所描述的都是卡特訪台的相關事件，雖然著重之處不同。方法 B1 所產生的摘要篇幅較大，因此較方法 A1 的摘要詳盡。

由表二及上述例子可得知，方法 A1 和方法 B1 所形成的摘要滿意度相差不大 (23:19)，但是方法 A1 產生出的摘要篇幅縮得較短 (8073:8582 個字)；因此，方法 A1 的摘要較合乎我們的需要。

3.2.2 查詢主題的相關性比較

接下來，我們針對兩種查詢主題相關性的評分方法做比較 (關鍵詞直接使用，未拆散成覆疊性雙連字)。

	較好的查詢主題 (topic) 編號	50 篇摘要總長度 (字數)
方法 A2 較佳	1,10,12,13,14,18,20,23,24,25,32,34,38,42,43,44,47,48,49,50 (共 20 篇)	(方法 A2) 6530
方法 B2 較佳	4,16,17,28,29,30,35,37,41 (共 9 篇)	(方法 B2) 6902
相同	2,3,5,6,7,8,11,15,19,21,22,26,27,31,33,36,39,40,45,46 (共 21 篇)	

表三 查詢主題相關性計算方式的比較

從評估結果中發現，方法 A2 的滿意度遠比方法 B2 (利用 Berkeley 的 IR 公式) 來得好 (20:9)，可能因為 Berkeley 的 IR 公式，是專門處理字數較多的整篇文章，對於長度較短的句子而言，容易造成分數低落 (小於 0) 的情況發生；雖然最後的分數加上一個常數 K ，但這種齊頭式平等的加分法，會造成分數的不平衡，真正和查詢主題相關的句子，因而無法被突顯出來，較長的句子可能會比較佔便宜。由滿意度的比較和篇幅長短的比較，我們認為方法 A2 形成的摘要較符合我們的需求。

3.2.3 句子縮短的影響

接下來我們針對句子縮短前後的摘要，比較其字數與句子縮短後的接受程度。

	較好的查詢主題 (topic) 編號	50 篇摘要總長度 (字數)
縮短前較佳	10,21,30,31,47 (共 5 篇)	(縮短前) 8073
縮短後較佳	3,4,6,8,9,11,12,13,14,15,16,17,18,19,20,24,25,26,30,33,34,35,36,38,39,40,44,45,48,50 (共 30 篇)	(縮短後) 6902
相同	1,2,5,7,22,23,27,28,29,37,41,42,43,46,49 (共 15 篇)	

表四 句子縮短前後的比較

在五十個查詢主題的摘要中，有五個被認為是縮短之後結果變差的。雖然被刪除的小句不多，但通常刪去之後會造成整個句子的不連貫，使句子本身變得不清楚；這種情況通常出現在句子前後有邏輯論證的關係時，例如 topic 10—庫藏股制度，原本的摘要為：

庫藏股制度究竟是利多，還是利空？未來，大股東可以決定動用公司資本公積及保留盈餘進場買回自家股票，最高可達一〇%，形成變相減資；不過，由於子公司買賣母公司股票完全不必規範，使得這套防弊措施能否奏效尙待觀

察。

第三句的重點在於後半句，因此在後半句被刪除的情況下，第三句的前半段就變得較不清楚，而且和第二句有重複現象，喪失了本身的獨立性和提示性。第二句則是刪除了前半句形成的後果，使得一般人無法瞭解此事件會形成什麼樣的影響，和事件本身的重要性。

句子縮短後的摘要，達到了 14.5% 的縮減率，和九成的滿意度，效果相當良好。未來我們希望將刪除的範圍從小句擴大到字詞，並做一些小句間的關連分析，希望在追求縮減率的同時，也能將句子的原意完整的保留下來。

3.2.4 斷詞 vs. 雙連字

另外，我們來比較另外一種查詢主題相關性評分方式的差異；對於查詢主題關鍵詞，未經處理而直接使用，與拆成覆疊性雙連字再套入查詢主題相關性的計算方法，有什麼樣的不同？哪一種方式比較好？

我們針對查詢主題相關性的計算方法：使用關鍵詞，和將關鍵詞拆成覆疊性雙連字（Overlapping Bigram）做比較，其結果如下：

	較好的查詢主題（topic）編號
斷詞（方法 A2）	1,3,7,11,22,26,34,35,39,45,46,49
拆成雙連字（方法 A2）	4,9,12,14,17,18,19,20,24,33,38,44,47
相同	2,5,6,8,10,13,15,16,21,23,25,27,28,29,30,31,32,36,37,40,42,43,48,50

表五 關鍵詞是否拆成覆疊性雙連字的比較

在此我們發現，有一半以上的摘要是相同的，其他互有優劣的摘要在選取的三句中往往有一兩句相同，即使完全不同者，兩邊的滿意度

都算可以接受。以 **topic 17**—流浪狗問題為例：

<關鍵詞>：動物保護法從去年十月底立法實施以來，已撥出五千三百萬元經費給各縣市政府改善收容所設備，而且也提供巴比妥酸鹽給收容所，進行流浪犬安樂死的人道處理。對於臺南縣政府發生溺斃、電殛流浪犬收容所作法，農委會依動物保護法的規定，中央主管機關並不能開出罰單，只能要求地方政府對行為人開出五萬元的罰單。由農委會及省農林廳補助六百萬元興建的「花蓮縣吉安鄉流浪犬中途之家」，第一期硬體工程已經完工。

<雙連字>：動物保護法從去年十月底立法實施以來，已撥出五千三百萬元經費給各縣市政府改善收容所設備，而且也提供巴比妥酸鹽給收容所，進行流浪犬安樂死的人道處理。對於臺南縣政府發生溺斃、電殛流浪犬收容所作法，農委會依動物保護法的規定，中央主管機關並不能開出罰單，只能要求地方政府對行為人開出五萬元的罰單。行政院農業委員會決定全面整頓流浪犬處理問題並將興建現代化的流浪犬中途之家，同時統一採購流浪犬安樂死用麻醉劑，提供給地方政府人道處理流浪犬之用。

因此，我們可以得到一個結論：不論關鍵詞有無被拆成覆疊性雙連字，結果都差不多。在 NTCIR-2 的探討會議中有提到：因為關鍵詞整理得太好，所以直接用關鍵詞來做資訊檢索（Information Retrieval）的查詢反而比利用其他欄位（如 Title，Question，Narrative 等，請參閱表一）的其他方法好。在此也是相同的情形，所以在查詢主題有很多關鍵詞的情況下，似乎不需要多此一舉，拆成覆疊性雙連字。不過除了 NTCIR-2 以外，網路搜尋引擎收到的查詢主題往往只有兩三個關鍵詞；此時，將系統找到的關鍵詞拆成雙連字，可能會有一定的改進。也可避免斷詞的錯誤所造成的負面影響。

4. 結論

4.1 主題相關的多篇摘要

本篇論文提出了一個多篇文件自動摘要系統，在此系統中我們利用簡單的斷詞斷句工具，以 NTCIR-2 的文章和查詢主題為實驗對象，針對句子的提示性和查詢主題相關性作分析，選出富有提示性又相互獨立的句子，將句子內不重要的小句刪除後，產生提示性摘要。

本篇摘要在提示性的計算上，採用了方法 A1；在查詢主題相關性的計算方面，採用方法 A2；不重要的小句也已經刪除。原文共有 197423 個字，五十篇摘要的總字數為 6902 個字，平均每篇 138 字；壓縮比為 3.5 %。

小句的刪除大致上不影響『提示性摘要』的需求，只有少數情況下，如 **topic 32**—腸病毒的摘要，效果不太理想。刪除後，摘要縮得太短。為此我們設定了字數下限，以保持足夠的指示性。當摘要的字數小於某一個值時（例如：60 字），則停止刪除小句的動作，以免刪除過多小句，喪失摘要的原意。如此摘要便能維持較適宜的狀況。

4.2 未來研究方向

本系統在詞彙的比對上，只使用了最簡單的斷詞和句子間的詞彙比對。然而文章內的句子寫法千變萬化，不同的語詞，不同的語法，可能形成相同的意念。在黃聖傑（1999）的多篇摘要研究中，試圖利用同義詞來處理這些現象，卻發現效果不大。這是因為很多新的專有名詞不會出現在辭典中，偏偏這些新詞往往是文章的重心所在。如果要解決這問題，可能需要作更深入的語意分析。

另外，目前的摘要作法還存在一個很大的問題：各摘要的篇幅差距頗大（52~349 字之間，標準差為 63.8），可能的原因為文章中句子的長

短不一，因此我們必須做長句分割，在語句停頓的地方將逗號改成句號（You Yu-Ling）。此外，也希望將句子縮短的步驟作得更徹底，除了小句之外，不重要的詞也可以一併刪除；更進一步。將長詞換成同等意義的短詞（例如：行政院長→閣揆，清華大學→清大）。

本文的研究，主要是針對新聞為主，因為一篇新聞往往只描述一個事件，有利於摘要的形成。我們希望除了新聞外，也能將這些方法推廣到其他文體（例如技術報告），產生具有指示性的多篇的摘要。

參考文獻

1. Chinatsu Aone, Mary Ellen Okurowski, James Gortlinsky. 1998. Trainable, Scalable Summarization Using Robust NLP and Machine Learning. In 36th Annual Meeting of the COLING-ACL, pp. 62-66.
2. Mark Wasson. 1998. Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications. In 36th Annual Meeting of the COLING-ACL, pp. 1364-1368.
3. Regina Barzilay, Kathleen R. McKeown and Elhadad. 1999. Information Fusion in the Context of Multi-Document Summarization. In 37th Annual Meeting of the ACL, pp. 550-557.
4. Adam Berger, Vibhu O. Mittal. 2000. Query-Relevant Summarizations using FAQs. In 38th Annual Meeting of the ACL, pp. 294-301.
5. Hongyan Jing. 2000. Sentence Reduction for Automatic Text Summarization. In Proceedings of the 6th ANLP / 1st NAACL, Section 1, pp. 310-315.
6. Hongyan Jing and Kathleen R. McKeown. 2000. Cut and Paste Based Text Summarization. In Proceedings of the 6th ANLP / 1st NAACL, Section 2, pp. 178-185.
7. Inderjeet Mani, Eric Bloedorn. 1999. Summarizing Similarities and Differences Among Related Documents. In Information Retrieval, Vol. 1, pp. 35-67.
8. Weiquan Liu and Joe Zhou. 2000. Building a Chinese text summarizer with phrasal chunks and domain knowledge. In ROCLING XIII, pp. 87-96.
9. D. R. Radev and K. R. McKeown. 1998. Generating Natural Language Summaries from Multiple On-line Sources. In Computational Linguistics,

Vol. 24, No. 3, pp. 469-500.

10. Yu-Jin Chen. 2000. Scalable Summarization for Chinese Text. National Tsing-Hua University, master thesis.
11. Yu-Ling You. 2000. Toward Defining Discourse Unit in Chinese Discourse. In Language researching and teaching. (in press)
12. 黃聖傑, 1999. 多文件自動方法摘要研究. 台灣大學資訊工程研究所碩士論文, 台北.
13. 楊允言, 謝清俊, 陳淑美, 陳克健. 1992. 中文文件自動分類之研究. 中華民國八十二年第六屆計算語言學研討會論文集, pp.217-233.
14. 楊允言, 張俊盛, 陳克健. 1993. 文件自動分類及其相似性排序. 清華大學資訊科學研究所碩士論文, 新竹.

基於階層式類神經網路之自動新聞文件分類方法

陳彥呈 蔣榮先

國立成功大學資訊工程系

chenyc@ismp.csie.ncku.edu.tw

jchiang@mail.ncku.edu.tw

摘要

文件分類是一項決定一篇文件是否屬於一個或多個已事先定義好的類別之工作，而自動化分類則可以有效地幫助分類的處理。在本篇論文中，我們提出了一個以階層混合式的專家模組(hierarchical mixture of experts model)為基礎的文件分類方法。這個模組使用了分割－克服原理(divide-and-conquer principle)，在一個事先定義好的階層架構下定義較小的分類問題，而最後的分類器則是使用類神經網路中的倒傳遞網路來完成分類機制。另外，在特徵選取(feature selection)上，我們也做了一些有別於傳統方法的改變。最後，我們以部份路透社(Reuters-21578)的新聞性文件做為測試資料，實驗結果顯示我們所提出的方法能有效地改善文件分類的正確率。

1. 緒論

近幾年來，隨著網路技術不斷地進步，有用的資訊也相對地大量成長中。雖然網路上舉手可得的資訊方便人們對資訊的取得與傳遞，但是當網路資訊量愈來愈大時，如何有效、且快速地取得有用的資訊，便成為非常重要的事情。此時，文件分類(text categorization)技術，即透過演算法分析一電子文件後，將其分配(assign)給一或多個類別(categories)，便扮演著其中重要的角色。

傳統的文件分類工作都是由某個領域的人類專家(human experts in domain)所完成。但是，隨著文件數量快速地成長，對於專家而言，這樣的工作就變得更困難了。在這種情況下，文件的自動分類就顯得更加重要了。

很多在做文件分類的方法中，例如使用規則庫(rule-based)、知識庫(knowledge-based)、或樣本庫(instance-based)．．．等，都是依賴大量的樣本來決定和文件有關的規則或知識。一般而言，這些樣本集合必須由那些對應用領域有深入認識的專家來訂定與建立，也因此，這些方法常常因為相關樣本建立得不足或不完全，使得規則或知識也就相對地不齊全，因此，就無法對文件做全盤性的樣本比對，以致於造成了分類上的困難。

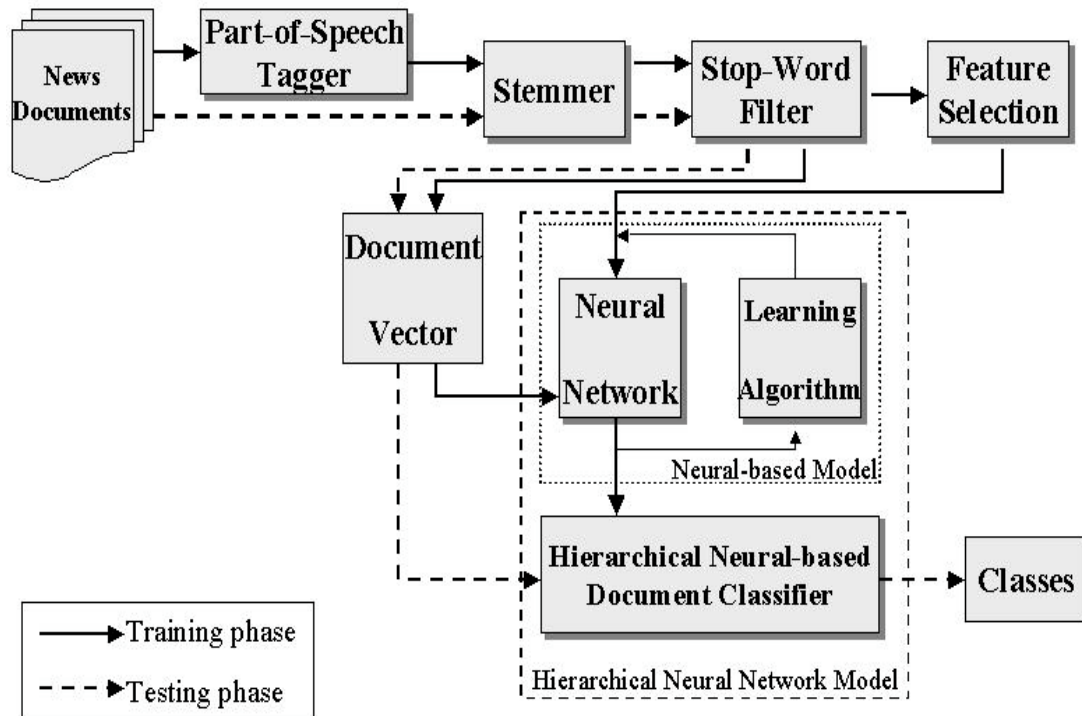
在本篇論文中，主要的動機在於改善目前文件分類的方法，我們不以關鍵字的存在否來決定一篇文件應屬於那一個或多個類別。進一步的，我們採用以類神經網路為基礎的階層式架構的機器學習的方法來決定文件的歸屬。而且，經由這樣學習的方法，可以使文件分類系統更容易地應用到其他的領域。

本篇論文除了緒論外，第二節將介紹我們所提的階層式模組，第三節將介紹特徵及訓練樣本集的選取，第四節則針對我們所使用的路透社新聞性資料集所做的一些自動化文件分類實驗的結果與分析。最後，我們為本篇論文提出總結。

2. 階層式模組

圖一所示，是我們所提出的自動化文件分類的完整模組。一個文件分類系統(text categorization system)的主要工作流程，是先用一組訓練樣本集來訓練系統中的文件分類器；然後再藉由已訓練好的分類器對測試樣本中的新文件做自動化分類的動作。在圖一的實線箭頭部份是整個文件分類的詳細訓練過程，首先決定一組已由專家分類好的樣本集，從此樣本集中，經過一連串的前處理程序後，選擇一組最能代表及識別(identification)此類別的特徵集(feature set)。並以向量方式表示之，如此就可得到一個以特徵向量表示的樣本組，而在階層式類神經網路模組中，主要是希望能透過每一個樣本組來訓練其所屬的分類器，使其能很正確地將每一個樣本分到正確的類別去。經過一連串的反覆學習後，我們得到一組已訓練好、具有相當辨識程度的分類器，以供測試階段時使用。

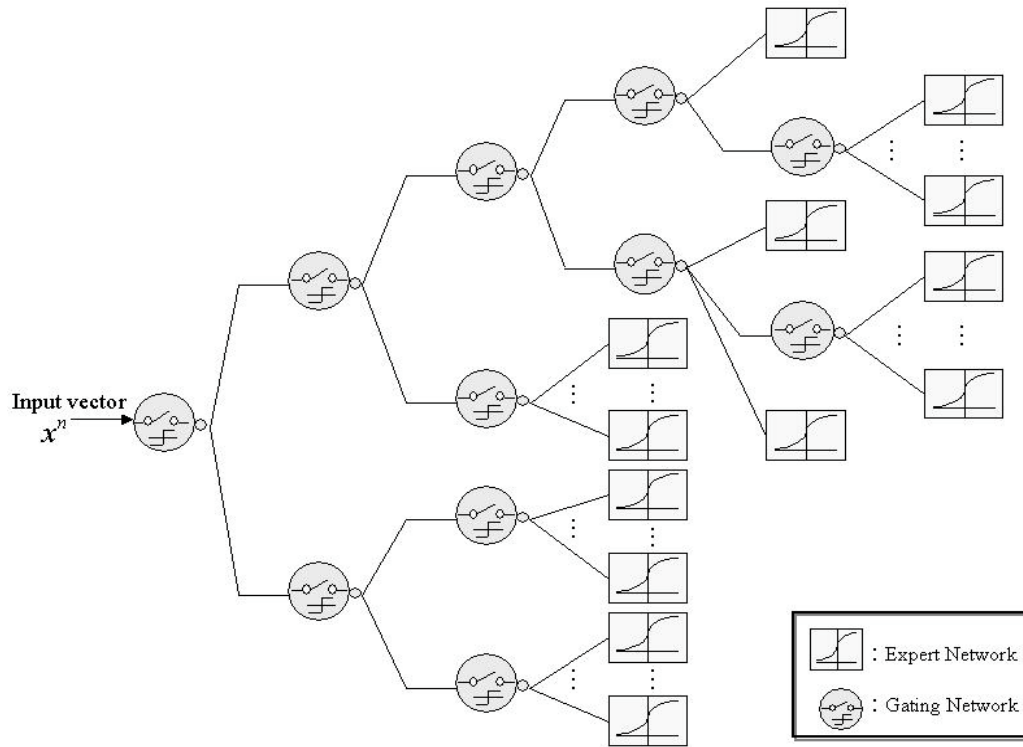
圖一中的虛線箭頭部份則是整個測試流程，起初也是將一新文件經過一連串的前序處理後，再依特徵集轉換成向量形式，最後透過階層式類神經網路模組，以決定新文件所屬的類別。



圖一 本論文所提出之自動化文件分類模組

在圖一用虛線方塊所圍成的，就是我們所提出的階層式類神經網路模組，其詳細的架構如圖二，此模組的主要的靈感是來自於 Jordan 和 Jacobs[1993]所提出的階層式混合的專家模型(hierarchical mixture of experts, HME model)。HME 模式是以分割-克服原理(divide-and-conquer principle)為基礎，其主要的想法是將一個大問題分割成若干個容易解決的小問題，然後再結合這些小問題的解答，以得到一般化的解答。而在分類一個減少範圍上(reduced domain)，HME 模型是經由將輸入空間(input space)劃分成一巢狀、順序的區域，然後訓練特定的較小分類器，以此求得一個分類問題的答案。HME 模型包含兩個基本的元件：閘門網路(gating networks)和專家網路(expert networks)。這些元件的結構類似於樹狀結構

(tree structure)，其內部節點是閘門、樹葉節點是專家。圖二就是我們提出的一個五層的階層式模組架構圖。



圖二 本論文所提出之階層架構圖

在我們的模型中，每個閘門所表示的是一份文件的一般概念，假如文件中包含著所表示的概念，則網路的輸出是 1，否則為 0。而專家所表示的是特定的類別[Ruiz, 1999]。所有的文件都以向量表示之。整個分類工作是由根節點(root node)開始，假如閘門的輸出值為真，則第二層的節點都會被啟動，如此的程序持續至它到達一個樹葉節點。

對於閘門和專家網路，由於類神經網路中的倒傳遞網路(back-propagation, BP Network)具有學習正確率高、理論簡明[Zurada, 1992]。因此，我們決定使用三層的倒傳遞類神經網路，其輸入層包含了 N 個特徵，隱藏層包含了 $(2N/3)$ 個節點，而輸出層為單一個節點。而在神經元的架構中，我們使用 S 形函數(sigmoid function)作為轉換函數。此函數具有微分容易的優點，可配合降梯度法則來調整

神經元間的權重，此函數當自變數趨向正負無限大時，函數值趨近於常數，其函數值域在 $[0,1]$ 之間。

3. 特徵選取和訓練資料集選取

一般而言，文件大部份都是人們以自然語言所書寫而成的，這些文件中的文字所要表達的，則是人們的想法與意見。我們相信在這些想法與意見中，主要是由一些重要的觀念所組成的，而我們認為文字中的名詞字詞最能表達一個觀念的形成。因此，在特徵選取過程中，我們首先使用了由 Eric Brill [1993]所提出的詞性分析器(part-of-speech tagger)為每個英文字標示其詞性資訊，然後選擇名詞集合的關鍵字詞。接下來則必須使用 stop word 過濾器模組，將上述所選取標示名詞的關鍵字詞中，過濾一些不足以代表文件本身特性字詞，以避免在接下來的處理過程中，引入太多不必要的雜訊(noise)。在做完 stop word 的處理後，其他剩下的名詞字集還不能算是最後想要的特徵集。因為根據人們的寫作習慣，對於那些出現頻率太過於頻繁或過於貧乏的字，通常都沒有太大的義意及重要性，對於符合這兩種情形的字集，我們可以經由字詞頻率—反文件頻率(term frequency and inverse document frequency , TFIDF)的分析而將其過濾掉，如此處理後所剩下的部份，我們稱之為特徵字詞(feature words)，這些字詞才是最重要的精華。

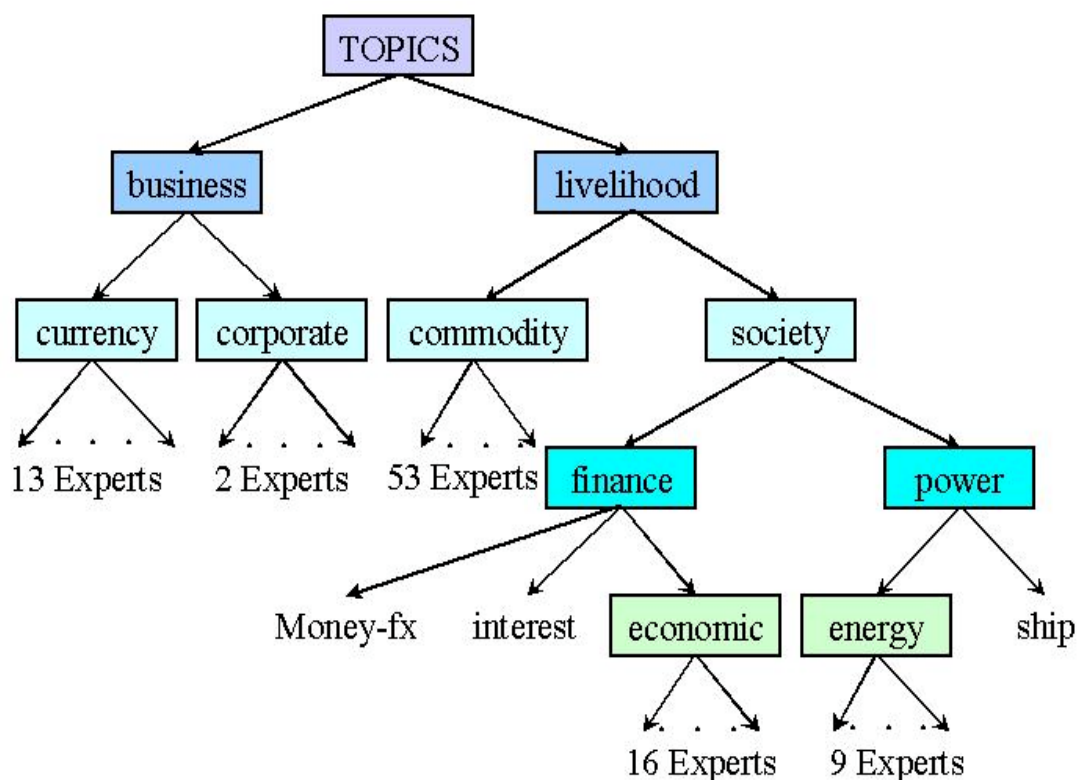
此外，在訓練分類器方面，對於同一類別的正負訓練樣本選取上，若兩者的選取差距過大，造成過度地不平均，很有可能會造成分類器在學習上的誤差，以致於造成最後分類上的錯誤。因此，對於訓練樣本的選取也是不可忽視的工作之一。在這一方面，我們採用了由 Ruiz [1999]所提出的“類別區(category zone)”的概念來選取訓練樣本集，其基本做法為選取屬於此類別的文件為正樣本，而選取最靠近此類別、卻不屬於此類別的文件做為負樣本。這樣的觀念，最早是來自於 Singhal 等人[1997]為文件繞送(text routing)所提出來的想法。

4. 實驗結果與分析

4.1 資料集

本實驗所使用的測試資料集，是由 David D. Lewis [1996]和路透社人員所共同整理而成的路透社新聞性文件—Reuters-21578。在這個資料集中，總共包含了 21578 篇文章，分為五大類別 (EXCHANGES, ORGS, PEOPLE, PLACES, TOPICS)，我們只拿五大類別中的 TOPICS 類別做為實驗之用。在這個類別中，包含了 135 個子類別，為了階層式模組的訓練及測試的需要，我們只選擇包含三篇文章以上的子類別做為測試類別。最後，我們使用了 96 個子類別、10555 篇文章作為實驗用的資料集。

對於 96 個子類別的階層架構，我們使用了 [陳彥呈, 2000] 所提出的架構圖，其架構如圖三。它基本的建構概念是依據文件在各類別之間的分佈來分析類別間的關連性所建立起來的。



圖三 在 TOPICS 中，96 個子類別的階層式架構圖

4.2 結果

在評估我們的模組效能之前，我們要先針對我們的模組提出兩個問題：1) 在同樣使用類神經網路方法的情況下，有使用階層式架構和沒有使用階層式架構的效能差異。2) 我們所提出的階層式架構和目前幾個有名的分類方法比較，其優劣為何？

在本實驗中所使用的評估方法，為在資訊擷取中最常被大家使用的正確率 (precision)、召回率(recall)和 F_1 評估方法。

表格一所示，是我們所提出的階層式方法和沒有使用階層架構的方法的比較 [Manevitz, 2000]，由表格中，我們可以很清楚地看出來，我們所提出的階層式方法，大大地提昇了分類的正確性。

表格一 使用階層式架構 V.S. 沒有使用階層式架構的平均效能比較

Class	NN (Hadamard)			NN (Frequency)			Proposed approach		
	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1
Earn	0.800	0.763	0.781	0.805	0.282	0.418	0.837	0.851	0.844
Acq	0.598	0.483	0.534	0.363	0.332	0.347	0.850	0.844	0.847
Money-fx	0.641	0.470	0.542	0.420	0.546	0.475	0.826	0.841	0.833
Grain	0.394	0.439	0.415	0.355	0.408	0.379	0.831	0.862	0.846
Crude	0.505	0.573	0.537	0.410	0.566	0.476	0.853	0.867	0.860
Trade	0.600	0.547	0.573	0.513	0.561	0.536	0.842	0.847	0.845
Interest	0.416	0.616	0.496	0.405	0.583	0.478	0.836	0.899	0.866
Ship	0.328	0.492	0.393	0.400	0.376	0.388	0.827	0.865	0.846
Wheat	0.446	0.588	0.507	0.430	0.400	0.414	0.839	0.886	0.862
Corn	0.451	0.236	0.310	0.434	0.247	0.315	0.830	0.892	0.860
Average (top 10)	0.517	0.520	0.508	0.453	0.430	0.422	0.837	0.866	0.851
Average (all)							0.867	0.892	0.879

表格二所示，則是我們所提出的方法和兩個著名的分類方法的比較—決策樹 (decision tree) [Weiss, 1999]和 k-NN 方法[Aas, 1999]。由表格中，我們可以知道，我們所提出的模組在某些類別上，其效能比其他兩種方法好。而在正確率及召回率上的成長，也比其他兩種方法要來得穩定。

表格二 我們所提出的階層式分類模組和決策樹及 k-NN 之比較

Class	k-NN (k=30)			Decision Tree			Proposed approach		
	Recall	Precision	F ₁	Recall	Precision	F ₁	Recall	Precision	F ₁
Earn	0.950	0.920	0.935	0.953	0.966	0.978	0.837	0.851	0.844
Acq	1.000	0.910	0.953	0.961	0.953	0.957	0.850	0.844	0.847
Money-fx	0.920	0.650	0.762	0.771	0.758	0.764	0.826	0.841	0.833
Grain	0.960	0.700	0.810	0.953	0.916	0.934	0.831	0.862	0.846
Crude	0.820	0.750	0.783	0.926	0.850	0.886	0.853	0.867	0.860
Trade	0.890	0.660	0.758	0.812	0.704	0.754	0.842	0.847	0.845
Interest	0.800	0.710	0.752	0.649	0.933	0.766	0.836	0.899	0.866
Ship	0.850	0.770	0.808	0.769	0.861	0.812	0.827	0.865	0.846
Wheat	0.690	0.730	0.709	0.972	0.831	0.894	0.839	0.886	0.862
Corn	0.350	0.760	0.479	0.982	0.821	0.894	0.830	0.892	0.860
Average (top 10)	0.823	0.756	0.788	0.879	0.879	0.879	0.837	0.866	0.851
Average (all)	0.792	0.818	0.805	0.878	0.878	0.878	0.867	0.892	0.879

5. 結論

本論文主要是在文件分類上，提出一個結合機器學習方法的階層式模組，並且使用了詞性分析器，以擷取出真正有意義的特徵字詞。最後，我們將我們的方法和其他方法做比較。從實驗的結果我們得知，我們所提出的階層式模組確實能提高正確率和召回率。

本論文的未來研究方向主要有特徵的選取，在使用類神經網路做為分類模組時，特徵選取的好壞會直接影響到分類的正確性。此外，我們也希望在類別區上尋求其他的方法，以期能求得更合適的訓練樣本集。

參考文獻

- Aas, K., and Eikvil, L., "Text categorization: A Survey", *Report No. 941, Norwegian Computing Center*, June, 1999. ISBN 82-539-0425-8
- Eric Brill, "A Corpus-based Approach to Language Learning", *phD Dissertation, University of Pennsylvania*, 1993.
- Jordan, M. I., and Jacobs, R. A., "Hierarchical Mixtures of Experts and the EM algorithm", *Technical Reports A. I. Memo No. 1440, Massachusetts Institute of Technology*, 1993
- Koller, D., and Sahami, M., "Hierarchical Classifying Documents Using very few Words", in *ICML-1997: Proceedings of the 14th International Conference on Machine Learning*, 1997, pages 170-178.

Lewis, D. D., "Reuters-21578 Text Categorization Test Collection Distribution", in *AT&T Labs – Research*, 1996.

Manevitz, L. M., and Yousef, M., "Document classification on neural networks using only positive examples", *ACM SIGIR*, 2000, pages 304-306.

Ng, H. T., Goh, W. B., and Low, K. L., "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization", in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, pages 67-73.

Ruiz, M. E., and Srinivasan, P., "Combining Machine Learning and Hierarchical Indexing Structure for Text Categorization", in *Proceedings of the 10th ASIS/SIGCR Workshop on Classification Research*, 1999.

Ruiz, M. E. and Srinivasan, P., "Hierarchical Neural Networks for Text Categorization", in *Proceedings of the 22nd ACM SIGIR International Conference on Information Retrieval*, 1999, pages 281-282.

Singhal, A., Mitra, M., and Buckley, C., "Learning Routing Queries in a Query Zone", in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, pages 25-32.

Weiss, S.M., Apte, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T., Hampp, T., "Maximizing text-mining performance", *IEEE Intelligent Systems*, Volume: 14 Issue: 4, July-Aug, 1999, pages 63-69.

Zurada, Jacek M., "Introduction to Artificial Neural Systems", *West Publishing Company, USA*, 1992.

陳彥呈, 蔣榮先, "基於階層式類神經網路之自動文件分類模式", 第八屆模糊理論及其應用會議, 2000.

Using Chi-square Testing in Modeling Confusion Characteristics for Robust Phonetic Set Generation

Yeou-Jiunn Chen⁽¹⁾ and Chung-Hsien Wu⁽²⁾

(1) Advanced Technology Center, Computer & Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan, R.O.C.

(2) Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.

chenyj@itri.org.tw, chwu@csie.ncku.edu.tw

Abstract

A phonetic representation of a language is used to describe the corresponding pronunciation and synthesize the acoustic model of any vocabulary. In order to obtain better phonetic representation, context-dependent units are used to model co-articulation effects between phones and have been broadly in speech recognition. However, this representation generally increases the number of recognition units. A phonetic representation with smaller phonetic units such as SAMPA-C for Mandarin Chinese can be applied to reduce the number of recognition units. Nevertheless, smaller phonetic units such as SAMPA-C will contain confusion characters and generally degrade the recognition performance. In this paper, a statistical method based on chi-square testing is used to investigate the confusion characteristics among phonetic units and develop a more reliable phonetic set, named modified SAMPA-C. Finally, experiments on continuous Mandarin telephone speech recognition were conducted. Experimental results show an encouraging improvement on recognition performance can be obtained. In addition, the proposed approaches represent a good compromise between the demands of accurate acoustic modeling.

1. Introduction

From the viewpoint of speech recognition, a phonetic representation is functionally defined by the mapping of the fundamental phonetic units of a language to describe the corresponding pronunciation and synthesize the acoustic model of any vocabulary. In the past years, context-dependent units have been broadly used to model the co-articulation effects such as triphone models, which consider both left and right phonemes at the same time. However, this representation generally increases the number of recognition units. Approaches for designing a smaller number of phonetic units are needed in the context-dependent based recognition.

In recent years, many phoneme-based phonetic representations have been used such as International Phonetic Alphabet (IPA) [1], Speech Assessment Methods Phonetic Alphabet (SAMPA) [2], and SAMPA for Chinese (SAMPA-C) [3]. Among these representations, SAMPA-C is more flexible and consistent than other phoneme-based phonetic representations for Mandarin Chinese. However, in SAMPA-C, several phonetic units with short duration are not easy to be distinguished and therefore degrade the recognition performance.

For Mandarin speech, the confusion characteristics can be found and analyzed in syllable-dependent, subsyllable-dependent, or phoneme-dependent situation. In a training database, syllable-dependent confusion characteristics are difficult to extract due to the sparse data problem. In contrast, the inconsistent phoneme segment in the training data is also not suitable to detect the phone-dependent confusion characteristics. The misdetected phones will result in misrecognition of syllables/subsyllables. Consequently, the phone-dependent confusion characteristic is not helpful for the analysis and representation of confusion characteristics of SAMPA-C based Mandarin speech recognizer. Therefore, the subsyllable is chosen as a compromising unit for the analysis of subsyllable-dependent confusion characteristics.

In this paper, based on the statistical hypothesis, the χ^2 (chi-square) testing [4] is an alternative test for evaluating dependence, which does not assume normally distributed probabilities. The underlying principle is to compare the observed frequencies with the expected frequencies. For investigating the effects of the confusion characteristics, the χ^2 statistic is used to examine the consistencies of two probabilistic distributions and the statistical decision criteria are applied to evaluate the statistical evidence for the confusion degree of two subsyllables. According to the analysis result, a less confusable phonetic set, namely modified SAMPA-C, is applied to develop a new Mandarin speech recognizer and compared to the original SAMPA-C.

The architecture for constructing the recognition model is shown in Fig. 1 and can be divided into two processes: development process and evaluation process. In the development process, an acoustic training database is collected and classified statistically for establishing SAMPA-C based recognition models. By analyzing the output distributions of confusion models, the confusion characteristics are extracted and used to generate the modified SAMPA-C. Moreover, using decision tree, the context-dependent models are generated for evaluating the performance. In the evaluation process, two continuous Mandarin speech recognition systems are developed and used to evaluate the syllable recognition rates using SAMPA-C and modified SAMPA-C HMM-based recognition models, respectively.

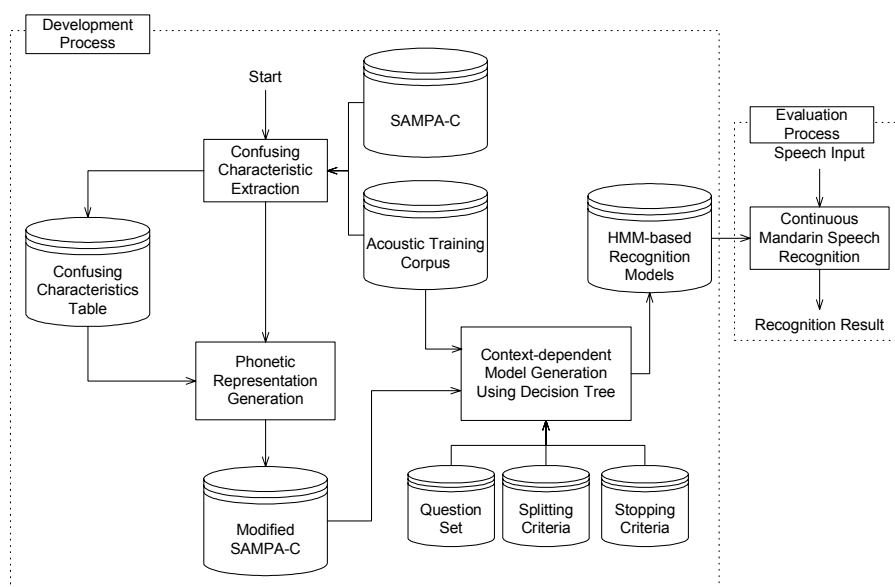


Fig. 1. Overall scheme for developing the HMM-based recognition models using modified SAMPA-C

2. Analysis of Confusion Characteristics

To accurately represent the confusion characteristics of Mandarin speech, the subsyllables are used as the basic units in the analysis process and extracted from the recognition outputs generated by the SAMPA-C based syllable recognizer. In this analytic procedure, 50 context-independent left-to-right HMMs with 4 states and 12 mixtures are built as the basic recognition models. 1551 utterances generated by 80 speakers in Mandarin Speech Database Across Taiwan (MAT) are used for advance analysis. In the following tests, the effect of the confusion characteristics between every two subsyllables is considered.

2-1 Testing for subsyllable-dependent confusion characteristics

To clarify the subsyllable-dependent confusion characteristics, the training and misrecognized data are used and divided into several categories, which are defined as subsyllable attributes (SA). For each SA, the numbers of occurrences and misrecognitions generated by the recognizer are accumulated. Then, these two corresponding frequency distributions of the training and misrecognized data, treated as SA distributions, are utilized to quantitatively analyze the confusion degree by using the χ^2 testing. The χ^2 value, which is greater than a threshold of the predefined significance level, implies that the SA distributions can be regarded as different. Accordingly, several subsyllables are treated as confusable and need further discrimination. The formula to calculate the χ^2 value is defined as follows.

$$\chi^2 = \sum_{i=1}^N \frac{(M_i - E_i)(M_i - E_i)}{E_i} \quad (1)$$

where N is the number of SAs, M_i is the number of misrecognitions of the i -th SA and E_i is the expected value of M_i and can be defined as

$$E_i = W_i \frac{\sum_{j=1}^N M_j}{\sum_{j=1}^N W_j} \quad (2)$$

where W_i is the number of appearances of the i -th SA.

The effects of confusion characteristics are analyzed and extracted from the recognition outputs generated by the SAMPA-C based syllable recognizer. Table I and Table II show two SA distributions of INITIALs and FINALs represented by SAMPA-C, respectively. It is clear that “d” and “V:” has the largest number of appearances in INITIALs and FINALs. However, the tendency of “dC” and “IM” was misrecognized frequently more than that of “d” and “V:”, respectively. “dC”, “IM”, “d”, and “V:” are the Mandarin syllables represented by SAMPA-C. As a result for a Mandarin speech recognizer, the confusion characteristics seems to strongly depend on the subsyllables. Next, since insufficient training data happen for some SAs, the χ^2 testing conditions might not be satisfied. Thus, the following two conditions in each SA have to be considered [5].

- (1) The percentage of the expected value over five is above 80%.
- (2) All expected values are more than one.

In Table I and Table II, the χ^2 values are 164 and 97 for INITIALs and FINALs, respectively. It is clear that the χ^2 value is greater than 5% of the significance level. Therefore, the analyzed results

show significant evidences that the confusion characteristics of INITIALs and FINALs can be regarded as subsyllable-dependent.

Table I. SA distributions of INITIALs represented by SAMPA-C, χ^2 value = 164, $p \leq 0.05$

INITIAL	NULL	b	p	m	f	d	t	n	l	g	k
Number of appearances	141	65	46	47	17	227	71	87	97	59	56
Number of misrecognition	49	27	11	10	6	57	38	23	26	21	14
INITIAL	h	dC	tC	C	dZ	tS	S	R	dz	ts	s
Number of appearances	80	54	51	49	62	56	72	52	57	49	56
Number of misrecognition	18	51	31	19	32	51	38	11	36	44	20

Table II. SA distributions of FINALs represented by SAMPA-C, χ^2 value = 97, $p \leq 0.05$

FINAL	NULL	a:	O:	V:	ai	ei	aU	ou	aM	@M	aN	VN	r
Number of appearances	38	72	8	194	60	40	61	53	69	52	59	56	8
Number of misrecognition	11	32	3	33	24	14	36	29	13	20	18	14	2
FINAL	i:	ja:	jE	jai	jaU	jou	jEM	IM	jaN	IN	u:	wa:	wO:
Number of appearances	41	15	37	4	30	29	47	34	25	43	58	21	47
Number of misrecognition	21	11	11	2	6	11	25	25	6	22	13	8	22
FINAL	wai	wei	waM	w@M	waN	wVN	y:	yE	yEM	yM	yN		
Number of appearances	23	57	58	38	27	53	17	23	24	14	16		
Number of misrecognition	8	9	23	20	12	4	11	11	12	4	8		

2-2 Examination of confusable phonetic set

According to the previous analysis, the misrecognition happens in some specific SAs. In general, the misrecognition is caused by the incorrect pronunciation or the confusable phonetic set. The incorrect pronunciation is due to inarticulacy such as the retroflexion in Mandarin speech. For examples, the “tS” and “IN” is usually pronounced as “ts” and “IM” in INITIALS and FINALS, respectively. Thus, in this paper, the confusion characteristic of each recognition units in the SAMPA-C based recognizer has to be examined and the phonetic set should be redefined. Table III shows some examples of SA distributions of confusions for recognition units in SAMPA-C. The upper two measures show the χ^2 values are greater than 5% of the significance level and the phoneme will cause the subsyllable-dependent confusion according to the χ^2 testing. On the other hand, the lower two measures show the χ^2 values are smaller than 5% of the significance level and these subsyllables possess less confusion characteristic.

Table III. Comparison of SA distributions of syllables represented by concatenating (+) phonetic units in SAMPA-C

Subsyllable	d	d+C	d+Z	d+z
Num. of appearances	227	54	62	57
Num. of misrecognition	57	51	32	36
χ^2 value = 55, $p \leq 0.05$				

(a)

Subsyllable	y+E	y+E+M	y+M	y+N
Num. of appearances	23	14	14	16
Num. of misrecognition	11	4	4	8
χ^2 value = 1.65, $p \geq 0.05$				

(b)

2-3 Determination of confusable phones

Given a subsyllable A , the subsyllable-dependent confusion characteristic between subsyllables A and B can be analyzed in Table IV, which show the four possible outcomes for a given trial. The confusion relationship between subsyllables A and B can be shown in Fig. 2. According to this representation, the χ^2 testing serves as a way to quantify the confusion between

these two distributions. Hence, based on the four outcomes in Table IV, the χ^2 testing can be applied to determine the degree of confusion between subsyllables A and B and is given by

$$\chi^2 = \sum_{\text{cells}} \frac{(f_{ij} - E_{ij})(f_{ij} - E_{ij})}{E_{ij}} \quad (3)$$

where f_{ij} is the observed frequency. E_{ij} is the expected frequency and defined as

$$E_{ij} = f_{i0} \frac{f_{0j}}{\sum_{k=1}^2 f_{k0}} \quad (4)$$

where f_{i0} is the totals of the i -th row and f_{0j} is the totals of the j -th column. If the value in Table IV is small, Yate's correction method is used to estimate a robust χ^2 value [6]. Therefore, the confusable phone, which causes the subsyllable-dependent confusion, can be found. Table V shows some examples of confusion measure. In this table (a) and (b) have high confusion contrast to (c) and (d). Accordingly, subsyllable "d+C" and subsyllable "U+N" are likely confused with subsyllable "C" and subsyllable "i+U+N", respectively.

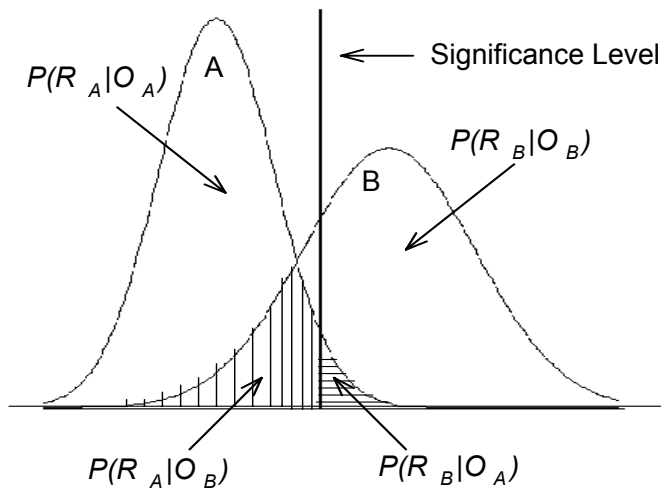


Fig. 2. Confusion relationship of subsyllables A and B

Table IV. Four possible outcomes for a given trial

		Recognition Result	
		R_A	R_B
Observations	O_A	$P(R_A O_A)$	$P(R_B O_A)$
	O_B	$P(R_A O_B)$	$P(R_B O_B)$

Table V. Examples of confusion measure (number of appearances)

		Recognition Result	
		d+C	C
Observations	d+C	3	23
	C	3	30
χ^2 value = 0.00265, $p \geq 0.05$			

(a)

		Recognition Result	
		d+C	d
Observations	d+C	3	12
	d	0	170
χ^2 value = 23.16, $p \leq 0.05$			

(c)

		Recognition Result	
		U+N	i+U+N
Observations	U+N	49	0
	i+U+N	7	8
χ^2 value = 0.00146, $p \geq 0.05$			

(b)

		Recognition Result	
		U+N	V+N
Observations	U+N	49	3
	V+N	2	42
χ^2 value = 76.98, $p \leq 0.05$			

(d)

3. Design of the Modified SAMPA-C

Based on the analysis of confusion characteristics, several confusion subsyllables caused by the confusable phonetic representation can be extracted. The confusable phonetic representation can be automatically detected using the above process. In our experimental results, the automatic speech recognition based on SAMPA-C cannot model the rapid variation between subsyllables. This is because that the confusion always occurs in the short duration between two subsyllables and the phonetic units representing the short phones cannot model this short duration well. Accordingly, a longer phonetic representation similar to subsyllable units is adopted to eliminate the confusion between two confusable subsyllables. These unsuitable phonetic units are manually analyzed. Each unit is concatenated with other phonetic unit to form a new, longer phonetic unit. The testing process is performed on the new representation iteratively. Finally, a modified SAMPA-C phonetic set, which suitably represent Chinese pronunciation is obtained and listed in Table VI. The original SAMPA-C phonetic set is also listed in Table VI for comparison. The phonetic units with boldface are the newly defined units. For example, the new phonetic unit “G” is defined by concatenating the phonetic units “d” and “C.” The total number of phonetic units in the modified SAMPA-C becomes 52 compared to 45 in the original SAMPA-C.

Table VI. Modified SAMPA-C and the examples with the corresponding Chinese characters and PinYin

Modified SAMPA-C	Examples by PINYIN	Modified SAMPA-C	Examples by PINYIN
G(d+C)	GIN (晶 jing1)	z(d+z)	zI: (子 zi3)
Q(t+C)	Qi: (七 qi1)	c(t+s)	cu@M (村 cun1)
X(C)	XiaU (小 xiao3)	aN(a+N)	laN (狼 lang2)
Z(d+Z)	ZUN (中 zhong1)	aM(a+M)	maM (慢 man4)
C(t+S)	Ca: (茶 cha2)	iU(I+U)	XiUN (兄 xiong)

4. Experimental Results

In the experiment setup, a Mandarin Speech Across Taiwan (MAT) telephone speech database, pronounced by 160 speakers (81 males, 79 females), with 8,237 files (sampling rate of 8kHz) was employed. Another speech database with 500 utterances was also collected and used as the testing data. In the following experiments, 12 Mel-Frequency Cepstrum Coefficient (MFCC), 12 delta MFCC, one delta log energy, and one delta delta log energy are extracted as a 26-dimension feature vector.

In the first experiment, the SAMPA-C based recognizer and the modified SAMPA-C based recognizer were built for the comparison of recognition performance. In these systems, the context-independent models were adopted and the subsyllable recognition rates of INITIALS and FINALS for the two systems are listed in Table VII.

Table VII. Recognition rates using SAMPA-C and modified SAMPA-C, respectively

	SAMPA-C	Modified SAMPA-C
INITIAL	55.86%	75.08%
FINAL	66.53%	67.26%

For Mandarin speech, the confusion effects of INITIALS are more obvious than that of FINALS. Due to the channel distortion of telephone network, the unvoiced INITIAL part with short duration is easy to be misrecognized. Therefore, the confusion between INITIALS can be discriminated using the modified SAMPA-C and the recognition performance can be improved

significantly.

Moreover, another phonetic representation set is also developed for evaluating the confusion characteristics analysis. This phonetic representation with 58 fundamental subsyllables [7-9] was adopted in this experiment. With the same training database, the distributions of misrecognition for subsyllable “dC”, “C”, “dZ”, and “d” are shown in Fig. 3. The subsyllable “dC” is usually misrecognized to “C”. However, the subsyllable “C” is not usually misrecognized to “dC”. It is difficult to detect the confusion characteristic of subsyllable “dC”. In our approach, the χ^2 value of “dC” compared with other subsyllables is shown in Fig. 4. The confusion characteristic of subsyllable “dC” can be detected. For the significance level, the subsyllable “C” usually confused with subsyllable “dC”.

In the next experiment, the context-dependent models were applied for evaluation and the experimental results are shown in Fig. 5. It is clear that the modified SAMPA-C can achieve an encouraging recognition performance, which is better than that obtained using the SAMPA-C. Especially, for the context-dependent models, the confusion between syllables can be efficiently discriminated and the recognition performance can also be improved.

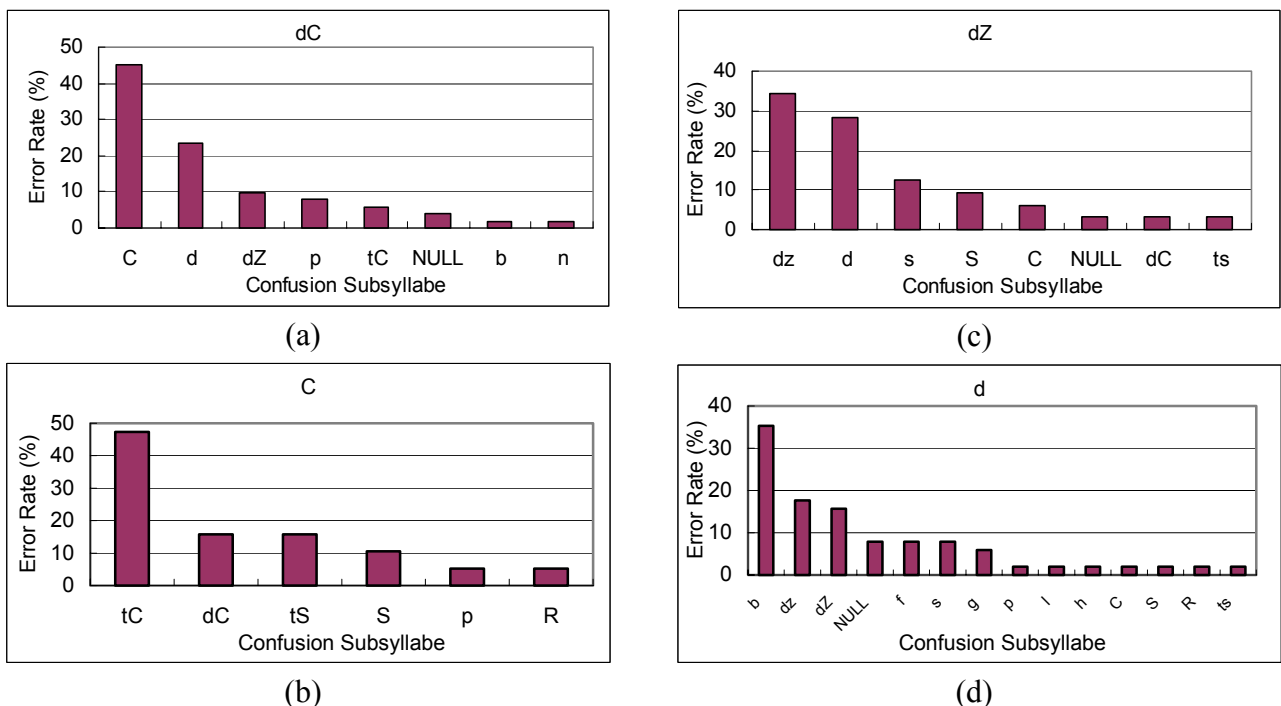


Fig. 3. Distributions of error rate for subsyllables (a) dC, (b) C, (c) dZ, and (d) d.

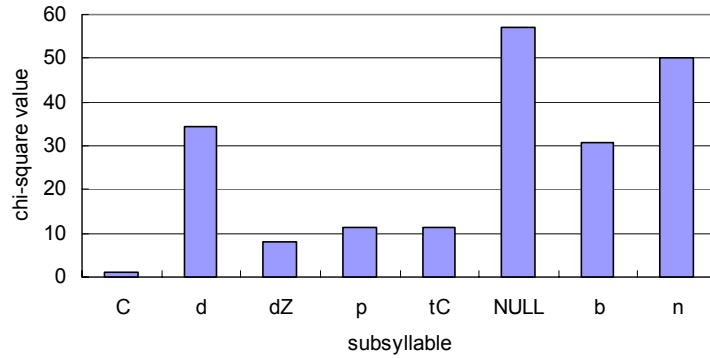


Fig. 4. χ^2 value of “dC” compared with other subsyllables

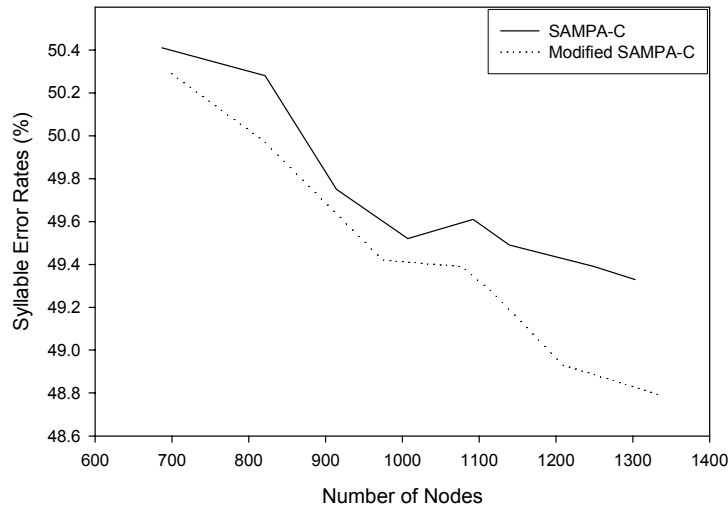


Fig. 5. Syllable error rates with respect to SAMPA-C and modified SAMPA-C based recognition system.

In order to evaluate the performance of different phonetic representations, we conducted experiments on three continuous syllable recognition model types. Three forms of subsyllabic units – right-context dependent INITIAL/FINAL (RCD-IF), SAMPA-C based tri-phones, and modified SAMPA-C based tri-phones were conducted to evaluate the syllable recognition rates (SRR). Table VIII shows the experimental results and the modified SAMPA-C based approach outperformed the other two types.

Table VIII. Syllable recognition rates using RCD IF, SAMPA-C based tri-phones, and modified SAMPA-C based tri-phones

	RCD IF	SAMPA-C Tri-phones	Modified SAMPA-C Tri-phones
No. of Nodes	675	754	812
SRR	46.12%	43.23%	50.23%

5. Conclusions

In this paper, the confusion characteristics for Mandarin speech using SAMPA-C were analyzed. The confusion characteristics generated with respect to confusable phonetic set can be discriminated by incorporating a statistical categorical data analysis method without any model assumption. Redefining the phonetic set, the effect of the confusion characteristics can be reduced and the recognition performance can be improved significantly. Hence, a modified SAMPA-C is proposed to provide a corresponding phonetic representation for building more reliable recognition models. Experimental results show that the proposed approaches give an encouraging improvement. For the portability to other languages, the proposed procedure can be easily applied to detect the confusion phonetic units of that language. Accordingly, a more reliable phonetic set for that language can be obtained.

6. Acknowledgment

The authors would like to thank the National Science Council, R.O.C., for its financial support of this work, under Contract No. NSC89-2614-H-006-004-F20. The paper is also a partial result of Project 3XS1B11 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, R.O.C.

7. References

- [1] R. H. Mathews, Mathews' Chinese-English Dictionary, Caves, 13th printing, 1975.
- [2] J. Wells, *EAGGLES Handbook on Spoken Language Systems(DRAFT) – SAMPA computer readable phonetic alphabet*, <[http:// www.phon.ucl.ac.uk/home/sampa/home.htm](http://www.phon.ucl.ac.uk/home/sampa/home.htm)>, 1997.
- [3] F. Seide, and N. J. C. Wang, "Phonetic modeling in the Philips Chinese continuous-speech recognition system", *Proc. of ISCSLP'98*, 1998, pp. 54-59.
- [4] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, 1990.

- [5] W. G. Cochran, "Some methods for strengthening of common tests", *J. of the International Biometric Society*, 1954, pp. 417-451.
- [6] W. J. Krzanowski, *Principles of Multivariate Analysis*. Oxford University Press, New York, 1988.
- [7] S. H. Chen, Y. F. Liao, S. M. Chiang, and S. Chang, "An RNN-based pre-classification method for fast continuous Mandarin speech recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 1, 1998, pp. 86-90.
- [8] R. Y. Lyc, I. C. Hong, J. L. Shen, M. Y. Lee, and L. S. Lee, "Isolated Mandarin based-syllable recognition based upon the segmental probability model", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 3, 1998, pp. 293-299.
- [9] C. H. Wu, Y. J. Chen, and G. L. Yan, "Integration of phonetic and prosodic information for robust utterance verification", *IEE Proceedings-Vision, Image and Signal Processing*, Vol. 147, 2000, pp. 55-61.

Pitch Marking Based on an Adaptable Filter and a Peak-Valley Estimation Method

Jau-Hung Chen and Yung-An Kao

Advanced Technology Center, Computer and Communication Research Laboratories,
Industrial Technology Research Institute, Chutung 310, Taiwan
Email: chenjh@itri.org.tw, kya@itri.org.tw

Abstract

In a text-to-speech (TTS) conversion system based on the time-domain pitch-synchronous overlap-add (TD-PSOLA) method, accurate estimation of pitch periods and pitch marks is necessary for pitch modification to assure an optimal quality of the synthetic speech. In general, there are two major issues on pitch marking: pitch detection and location determination. In this paper, an adaptable filter, which serves as a bandpass filter, is proposed for pitch detection to transform the voiced speech into a sine-like wave. Based on the sine-like wave, a peak-valley decision method is investigated to determine the appropriate part (positive part and negative part) of the voiced speech for pitch mark estimation. At each pitch period, two possible peaks/valleys are searched and the dynamic programming is performed to obtain the pitch marks. Experimental results indicate that our proposed method performed very well if correct pitch information is estimated.

1. Introduction

In past years, the approach of concatenative synthesis has been adopted by many text-to-speech (TTS) systems [1]–[6]. The concatenative synthesis uses real recorded speech segments as the synthesis units and concatenates them together during synthesis. Also, the time-domain pitch-synchronous overlap-add (TD-PSOLA) [6] method has been employed to perform prosody modification. This method modifies the prosodic features of the synthesis unit according to the target prosodic information. Generally, the prosodic information of the speech includes pitch (the fundamental frequency), intensity, and duration, etc. For a synthesis scheme based on TD-PSOLA method, it is necessary to obtain a pitch mark for each pitch period in order to assure an optimal quality of the synthetic speech. The pitch mark is a reference point for the overlap of the speech signals.

It is useful to have a speech synthesizer with various voices for speech synthesis. Sometimes it is also important for a service-providing company to have a synthesizer with the voice of its own employee or the speaker of its favorite. For conventional TTS systems, however, it is a professional but tedious job to create a new voice. Recently, corpus-based TTS systems have been appreciated which use a large amount of speech segments. Some approaches selected the speech segments as the candidates of synthesis units. Establishing the synthesis units includes speech segmentation, pitch estimation, pitch marking, and so on. However, pitch marking is very labor-intensive among them if there involved no automatic mechanism.

In general, there are two major issues on pitch marking: pitch detection and location determination. Compared to pitch detection [7]-[14], few papers have been presented for pitch marking [15][16], which is also a difficult problem because of the great variability of the speech signals. Moulines *et al.* [15] proposed a pitch-marking algorithm based on the detection of abrupt changes at glottal closure instants. At each period, they assumed that the speech waveform could be represented by the concatenation of the response of two all-pole systems. On the other hand, Kobayashi *et al.* [16] used dyadic wavelet for pitch marking. The glottal closure instant was detected by searching for a local peak in the wavelet transform of the speech waveform.

In this paper, we propose a pitch-marking method based on an adaptable filter and a peak-valley estimation method. The block diagram is shown in Fig. 1. The input signals are constrained to the voiced speech because only the periodic parts are interested. We introduce an adaptable filter, which serves as a bandpass filter, to transform the voiced speech into a sine-like wave. The autocorrelation method is then used to estimate the pitch periods on the sine-like wave. Also, a peak-valley decision method is presented to determine which part of the voiced speech is suitable for pitch mark estimation. The positive part (the speech with positive amplitude) and the negative part (the speech with negative amplitude) are investigated in this method. This is motivated from Fig. 2(a), which displays an example of waveform having the negative part reveals explicit periodicity. In general, it could synthesize better speech quality if the pitch marks are labeled at the positions of extreme points (peaks and valleys) of the speech. At each pitch period, two possible peaks/valleys are searched. Finally, the pitch marks are obtained by the dynamic programming by calculating the pitch distortion.

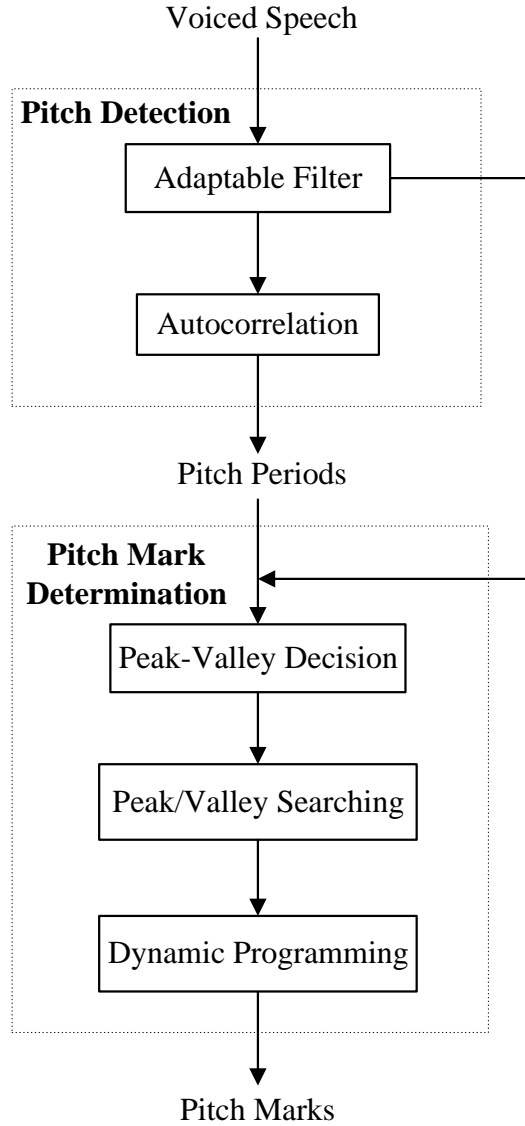


Figure 1: Block diagram of the proposed pitch-marking method.

2. Pitch Detection Using an Adaptable Filter Followed by Autocorrelation Method

The proposed adaptable filter serves as a bandpass filter in which its pass band is from 50 Hz to the detected fundamental frequency, up to 500 Hz, of the voiced speech. The adaptable filter is achieved by the following three steps.

Step 1. It computes the FFT (Fast Fourier Transform) to transform the voiced speech into the frequency domain.

Step 2. The fundamental frequency, f_0 , is detected by searching the first peak of the spectral contour.

Step 3. The IFFT (Inverse FFT) is invoked over the passband between 50 Hz and f_0 to obtain the filtered speech.

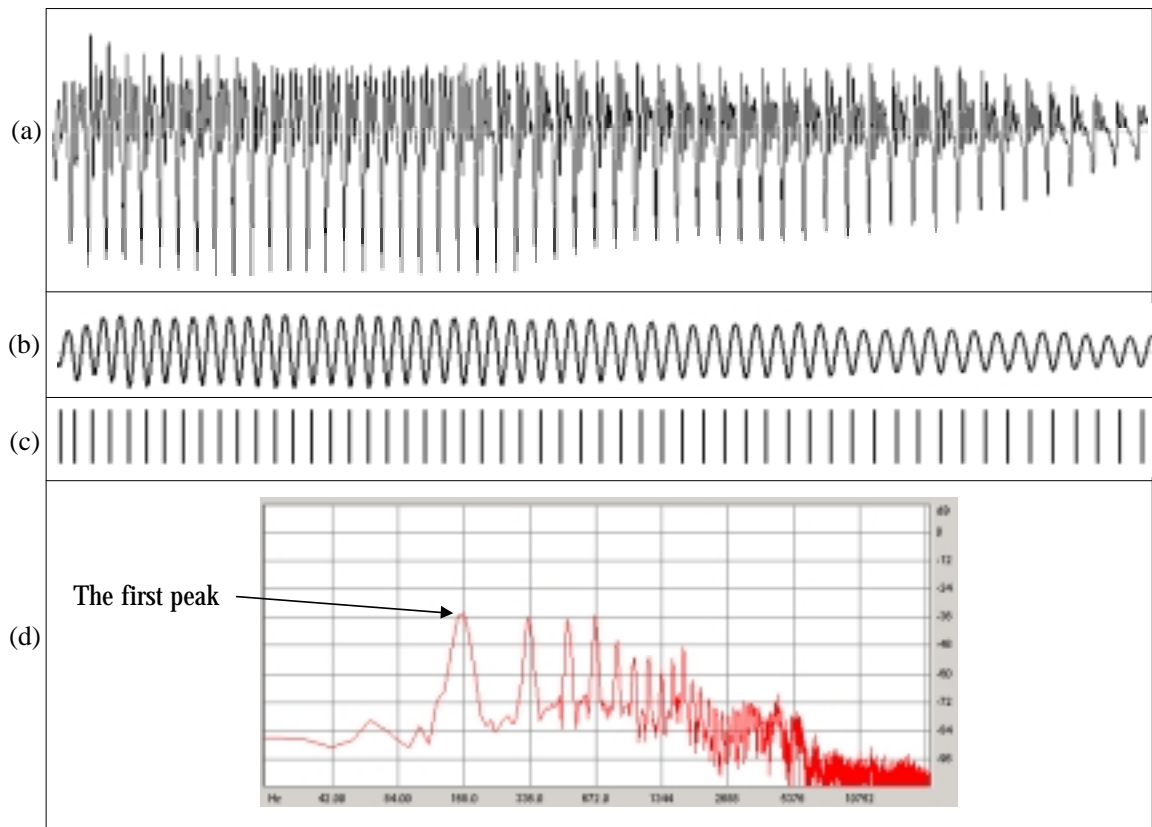


Figure 2: Results of the adaptable filter and pitch mark determination. (a) Waveform of the voiced speech with explicit periodicity on the negative part. (b) Waveform of the filtered speech. (c) Detected pitch marks. (d) Spectral contour (note that the frequency axis is not linearly plotted).

An example of the adaptable filter is displayed in Fig. 2. Panel (a) and (b) shows the waveforms of the original speech and the filtered speech, respectively. It can be seen that the filtered speech is generally a sine-like wave that reveals clear periodicity than that on the original speech waveform. For a frame in the middle of the voiced speech, the spectral contour is depicted in panel (d). Note that the frequency axis is not linearly plotted for the reason of inspecting the first spectral peak. The first peak was found at 168 Hz, which is the fundamental frequency. Finally, the pitch periods are obtained by analyzing the filtered speech using the conventional autocorrelation method.

3. Pitch Mark Determination Using a Peak-Valley Decision Method and Dynamic Programming

3-1 Peak-Valley Decision

From observations, we found that the voiced speech, $s[\cdot]$, is synchronous with the filtered speech, $o[\cdot]$, either at peaks or at valleys. For the case illustrated in Fig. 2 (a) and 2 (b), they are synchronous at valleys having explicit periodicity instead of those at peaks. As a result, the pitch marks could be easily determined at the negative part than those at the positive part. In the following, peak-valley decision method calculates two costs by summing the amplitudes of $s[m]$, where m represents the position of the local extreme point of $o[\cdot]$ over each pitch period:

$$C_{peak} = \frac{1}{N_{peak}} \cdot \sum_{n=1}^{N_{peak}} s[Pos_{peak}[n]] \quad (1)$$

$$C_{valley} = \frac{-1}{N_{valley}} \cdot \sum_{n=1}^{N_{valley}} s[Pos_{valley}[n]] \quad (2)$$

where the symbols are defined as follows:

C_{peak} : Cost estimated at the peaks of $o[\cdot]$.

C_{valley} : Cost estimated at the valleys of $o[\cdot]$.

N_{peak} : Total number of the peaks of $o[\cdot]$.

N_{valley} : Total number of the valleys of $o[\cdot]$.

$Pos_{peak}[n]$: Position of the n -th peak of $o[\cdot]$.

$Pos_{valley}[n]$: Position of the n -th valley of $o[\cdot]$.

The peak-valley decision is made as follows: If $C_{peak} > C_{valley}$ then the positive part (peak) of $s[\cdot]$ is adopted for the evaluation of pitch mark. Otherwise, the negative part (valley) of $s[\cdot]$ is adopted.

3-2 Pitch mark determination Based on Dynamic Programming

Once the adoption of the peak or valley has been decided, say peak, the positions of pitch marks are determined by picking the peaks of $s[\cdot]$. For the i -th pitch period, P_i , two highest peaks in the corresponding voiced speech are searched. Suppose the highest and the second highest peaks are located at L_{i1} and L_{i2} , respectively. It might occur that the second one is absent. For this case, we let $L_{i2} = L_{i1}$. For all the detected peaks, the determination of pitch mark is then performed based on dynamic programming. The distortion of pitch period, $d_i(j,k)$, and its accumulation, $A_i(j)$, are defined as follows:

$$d_i(j,k) = \left| L_{ij} - L_{(i-1)k} \right| - P_i + g(j,k), \text{ for } i=2, \dots, PN \quad (3)$$

$$A_i(j) = \min \left\{ \begin{array}{l} d_i(j,1) + A_{i-1}(1), \\ d_i(j,2) + A_{i-1}(2) \end{array} \right\}, \text{ for } i=2,3,\dots,PN \quad (4)$$

where PN is the total number of pitch period and $j, k=1,2$. In Equation (3), $g(j,k)$ is a penalty function represented as

$$g(j,k) = \begin{cases} 0, & \text{if } j = 1 \text{ or } k = 1 \\ \frac{1}{PN}, & \text{otherwise} \end{cases} \quad (5)$$

The penalty function is introduced here due to the preference of the highest peak.

The search path of the dynamic programming is illustrated in Fig. 3. The peak locations (pitch marks) can be obtained by back tracing the peak sequence corresponding to the smallest value of $A_i(1)$ and $A_i(2)$. An example of the results of pitch marking is shown in Fig. 2(c). Similar procedures described above can be applied to the case of “valley”.

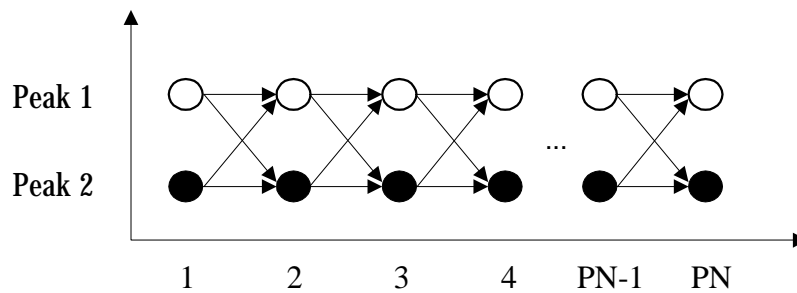


Figure 3: Illustration of the peak-picking search path of the dynamic programming.

4. Experiments and Results

4-1 Experimental environment

A continuous speech database was established which provides the basic synthesis units of our Mandarin Chinese TTS system. This database is composed of 70 phrases and their lengths are between 4 to 6 Chinese characters. It includes an amount of 436 tonal syllables comprising the required 413 basic synthesis units. A native female speaker read them in normal speaking style. The speech signals were then digitized by a 16-bit A/D converter at a 44.1k Hz sampling rate. The syllable segmentation was manually done in order to obtain the precise boundaries of voiced speech and unvoiced speech. The total duration of the 436 voiced speech is about 2.1 minutes. For each syllable, the voiced speech was used to test the proposed methods. The frame size used in the adaptable filter was set to 4096 speech samples (92.8 ms).

For the voiced speech, the waveforms along with the pitch marks obtained from our

pitch-marking program were visually displayed. The pitch marks were then checked and corrected by an experienced person through a friendly interface. For the evaluation of the experiments, we obtained 436 sets of human-labeled pitch marks, denoted as H , which comprises 23868 pitch marks.

4-2 Performance of the pitch marking method

The results of the peak-valley decision were verified by human judgment on visual displays. A success rate of 99.1% is obtained (4 of the 436 results were disagreed). For the female speaker, we found that 97.2% of the voiced segments reveal clear periodicity on the negative parts.

The proposed method generated 23860 pitch marks, denoted as I , without any duplication. The success rate of the pitch marking method is defined as follows:

$$\text{Correct rate} = \frac{|\{x \mid x \in I \text{ and } x \in H\}|}{|H|} \times 100\% \quad (6)$$

As shown in Table 1, a success rate of 97.2% is obtained (baseline), in contrast with the 95% and 97% success rates of the methods of [15] and [16], respectively. However, we found that most of the errors are resulted from the incorrect results of pitch detection. Most of the pitch errors are due to large changes of pitch locating at the boundaries of the voiced speech. Providing correct pitch information, our method leads to a success rate of 99.5%.

Table 1: Success rate of the pitch-marking method.

Condition	Baseline	Using correct pitch
Success rate	97.2%	99.5%

5 Conclusions

In this paper, a preliminary work on pitch marking has been proposed. We present the adaptable filter combined with the autocorrelation method for pitch detection. On the other hand, a peak-valley decision method is introduced to select either the positive or the negative parts for evaluation of pitch mark. Also, a dynamic-programming-based pitch mark determination method is demonstrated where two peaks/valleys are searched at each period. In the experiments, our pitch-marking method achieves 97.2% success rate.

Furthermore, a high success rate of 99.5% is obtained providing correct pitch information.

Acknowledgement

This paper is a partial result of Project 3XS1B11 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

References

1. Hamon, C., E. Moulines, and F. Charpentier, "A diphone synthesis based on time-domain prosodic modifications of speech," in Proc ICASSP, 1989, pp.238-241.
2. Iwahashi, N. and Y. Sagisaka, "Speech segment network approach for optimization of synthesis unit set," Computer Speech and Language, 1995, pp.335-352.
3. Shih, C. L. and R. Sproat, "Issues in text-to-speech conversion for Mandarin," in Computational Linguistics and Chinese Language Processing, vol.1, 1996, pp.37-86.
4. Chen, S. H., S. H. Hwang and Y. R. Wang, "An RNN-based prosodic information Synthesizer for Mandarin text-to-speech," IEEE Trans. on Speech and Audio Processing, Vol. 6, No. 3, 1998, pp. 226-239.
5. Chou, F. C. and C. Y. Tseng, "Corpus-based Mandarin speech synthesis with contextual syllabic units based on phonetic properties," in Proc. ICASSP, 1998, pp.893-896.
6. Charpentier, F. J. and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in Proc. ICASSP, 1986, pp. 2015-2020.
7. Rabiner, L. R., M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A Comparative performance study of several pitch detection algorithms," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, 1976, pp. 399-417.
8. Rabiner, L. R., "On the use of autocorrelation analysis for pitch detection," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, 1977, pp. 24-33.
9. Noll, A. M., "Cepstrum pitch determination," J. Acoust. Soc. Amer., Vol. 47, 1967, pp. 293-309.
10. Markel, J. D., "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust., Vol. Au-20, 1972, pp. 367-377.

11. Barnard, E., R. A. Cole, M. P. Veal, and F. A. Alleva, "Pitch detection with a neural-net classifier," *IEEE Trans. On Signal Processing*, vol. 39, No. 2, 1991, pp. 298-307.
12. Kadambe, S., G. F. Boudreaux-Bartels, "A comparison of wavelet functions for pitch detection of speech signals," in *Proc. ICASSP*, 1991, pp.449-452.
13. Barner, K. E., "Colored L-1 filters and their application in speech pitch detection," *IEEE Trans. On Signal Processing*, Vol. 48, No. 9, 2000, pp. 2601-2606.
14. Huang, H. and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," in *Proc. ICASSP*, 2000, pp.1523-1526.
15. Moulines, E., F. Emerard, D. Larreur, J. L. Le Saint Milon, L. Le Faucheur, F. Marty, F. Charpentier, and C. Sorin, "A real-time French text-to-speech system generating high-quality synthetic speech," in *Proc. ICASSP*, 1990, pp.309-312.
16. Kobayashi, M., M. Sakamoto, T. Saito, Y. Hashimoto, M. Nishimura, and K. Suzuki, "Wavelet analysis used in text-to-speech synthesis," *IEEE Trans. on Circuits and Systems-II, Analog and Digital Signal Processing*, Vol. 45, No. 8, 1998, pp. 1125-1129.

Optimization of HMM by The Tabu Search Algorithm

Xiao-dan Mei^{*}, Sheng-he Sun^{*}, Jeng-shyang Pan^{**} and Tsong-Yi Chen^{**}

^{*} Dept. of Automatic Test, Measurement and Control, Harbin Institute of Technology, China

^{**} Dept. of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan

^{**} jspan@cc.kuas.edu.tw, chentso@cc.kuas.edu.tw

Abstract

In this paper, the tabu search algorithm is employed to train Hidden Markov Model (HMM) to search out the optimal parameter structure of HMM for automatic speech recognition. The proposed TS-HMM training provided a mechanism that allows the searching process to escape from the local optimum and obtain a near global optimum. Experimental results show that the TS-HMM training has a higher probability to find the optimal model parameters than the traditional algorithms.

Keywords: tabu search, Hidden Markov Model, speech recognition, global optimum

1. INTRODUCTION

HMM is a highly robust statistical method and widely used for automatic speech recognition. It is a powerful algorithm used to estimate the model parameters and can achieve high performance [1,2,3]. Once a structure of the model is given, the model parameters are obtained automatically by feeding training data. The HMM model parameters take an important role in a HMM based speech recognizer because they can characterize the behavior of the speech segments and affect the system recognition accuracy directly.

Many heuristic algorithms are developed to optimize the model parameters in order to describe the trained observation sequences better, such as the forward-backward method [4] and the gradient method [5]. However, all these methods start from an initial guess and at last converge to a local optimum in practice. Few methods can escape from the local optimum to obtain the global optimum.

The tabu search algorithm [6] is the generalized heuristic global search technique with short-time memory, and suitable for solving many nonlinear optimization problems. The basic idea of the tabu search approach is to explore the search space of all feasible solutions by a sequence of moves. The spirit of this method is embedded in its short-term memory process. The elements of the move from the current solution to its selected neighbor are partially or completely recorded in the tabu list for forbidding the reversal of the replacement in future iterations. The search would cycle between the first encountered local optimum and its neighbor without this assurance.

In this paper, the tabu search algorithm is utilized to HMM training to search out the optimal structure of HMM for automatic speech recognition. The proposed TS-HMM training provided a mechanism that allows the searching process to escape from the local optimum and to obtain a near global optimum.

In Section 2 of this paper, the definition of HMM is given, then the tabu search algorithm is described in Section 3. The TS-HMM training algorithm is presented in Section 4. Simulation results are shown in Section 5 and conclusions are given in Section 6.

2. HIDDEN MARKOV MODEL

HMM is a probability model used to represent the statistic property of the stochastic process and is characterized by the model parameters. The stochastic process in speech recognition is the finite-length stochastic sequences called observation symbol, denoted by $\mathbf{O} = \mathbf{o}_1 \mathbf{o}_2 \cdots \mathbf{o}_M$, where M is the dimension of the observation symbol. One HMM with N states (S_1, S_2, \cdots, S_N) can be characterized by the parameter set $\lambda = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}\}$, where

(1) $\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_N]$ is the initial distribution. It is used to describe the probability distribution of the observation symbol in the initial moment when $t = 1$, namely

$$\pi = P(q_1 = S_i) \quad i = 1, 2, \cdots, N \quad (1)$$

it must satisfy $\sum_{i=1}^N \pi_i = 1$. (2)

(2) $\mathbf{A} = \{a_{ij} \mid i, j = 1, 2, \cdots, N\}$ is the transition probability distribution matrix. Its element at row i , column j is the probability a_{ij} of transition from current state i to next state j , namely

$$a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i) \quad (3)$$

it must satisfy the following condition:

$$\sum_{j=1}^N a_{ij} = 1 \quad (4)$$

(3) $\mathbf{B}=\{b_{ik} \mid i=1, 2, \dots, N, k=1, 2, \dots, M \}$ is the observation symbol probability distribution matrix in the discrete HMM. Its element at row i , column k is the probability b_{ik} of observation symbol with index k emitted by current state i and must satisfy the following condition:

$$\sum_{k=1}^M b_{ik} = 1. \quad (5)$$

As above-mentioned, HMM is used to approximate the probability of each observation symbol existing in the current state. When π , \mathbf{A} , \mathbf{B} are given, the probability $P(\mathbf{O}|\lambda)$ of the HMM system generating one random observation symbol can be calculated. Three essential problems of HMM must be solved, they are:

- (i) how to effectively calculate the probability $P(\mathbf{O}|\lambda)$;
- (ii) how to select the optimal state sequence when the model λ is given;
- (iii) how to adjust the model parameter λ to make the probability $P(\mathbf{O}|\lambda)$ higher.

People often employ the Forward algorithm to solve the first problem and the Viterbi algorithm to solve the second one. For the third problem, people use Gradient algorithm. This paper aims at solving the last problem, we use the tabu search algorithm to search the optimal model parameters λ .

3. THE TABU SEARCH ALGORITHM

The tabu search algorithm, which was proposed by Glover [6], is a generalized heuristic global search technique with short-time memory. Its basic idea is to explore the search space of all feasible solutions by a sequence of moves and to forbid some search directions at the present iteration in order to avoid cycling and jump off local optima. The elements of a move from the current solution to its selected neighbor are partially or completely recorded in the tabu list for the purpose of forbidding the reversal of the replacement in a number of future iterations.

The tabu search approach begins with test solutions generated randomly and their corresponding objective function values are computed. If the best of these solutions is not tabu or if it is tabu but satisfies the aspiration criterion, then select this solution to be the new current solution to generate test solutions for next iteration. It is called aspiration criterion if the test solution is a tabu solution but the objective value is better than the best value of all iterations. The tabu search algorithm is given as follows:

Tabu Search Algorithm ()

```
{
    generate the initial solutions;
    calculate the current solution and the best solution;
    while termination criterion not reached
    {
        generate the test solutions in the neighborhood of the current solutions;
        calculate the corresponding objective values;
        update the current solution and the best solution;
        update the tabu list;
    }
}
```

4. THE TS-HMM ALGORITHM

In this paper, the configuration of HMM is a five states left-right model and the speech feature vectors are vector quantized into the codebook with the size of 256. So A is a 5-by-5 matrix and B is a 5-by-256 matrix. As shown in Fig. 1, this model can represent speech signal whose properties change over time in a successive manner.

Due to the configuration of the model, some transitions between states do not exist so that the corresponding elements in matrix A are constantly zero and these elements will not be encoded when performing search.

The optimal model parameters searching problem must be mapped to the tabu search algorithm before it can be used. The mapping procedure is described as follow:

In TS-HMM training, the model is encoded into a string of real numbers between 0 and 1, and of course they satisfy the equation (4) and (5). As shown in Fig. 2, this string is composed of two parts: SA and SB . These two parts are composed of the rows of matrices A and B respectively.

A solution of this algorithm is defined as s_l consisting of a set of real numbers like the one shown in Fig. 2. The probability $p_n(O|\lambda_n)$ of the HMM solution λ_n which generates the training observation sequences $O = o_1 o_2 \cdots o_M$ must be calculated as the objective function value.

The initial test solutions are generated randomly. After the first iteration, the test solutions are generated from the best solution of current iteration by swapping two indices randomly. The tabu list

memory stores the swapped indices only. It is a tabu condition if the swapped indices to generate the new test solution from the best solution of current iteration are the same as any records in the tabu list memory.

Let $\theta_l = \{\lambda_{l1}, \lambda_{l2}, \dots, \lambda_{lN_s}\}$ to be the set of the test solutions, let $\lambda_c = \{\lambda_c(1), \lambda_c(2), \dots, \lambda_c(N)\}$ and $\lambda_b = \{\lambda_b(1), \lambda_b(2), \dots, \lambda_b(N)\}$ be the best solution of current iteration and the best solution of all iterations respectively, let $V_l = \{v_1, v_2, \dots, v_{N_s}\}$, v_c and v_b denote the set of objective function values for test solutions, the objective function value for the best solution of current iteration and the objective function value for the best solution of all iterations, respectively, where v_l is the objective function value for solution λ_l , $1 \leq l \leq N_s$. The algorithm is given as follows:

Step 0. Set the tabu list size T_s , the number of test solutions N_s and the optimum number of iterations

l_m . Set the iteration counter $i = 1$ and insertion point of the tabu list $t_l = 1$. Generate N_s solutions $\theta_l = \{\lambda_{l1}, \lambda_{l2}, \dots, \lambda_{lN_s}\}$ randomly, calculate the corresponding objective values $V_l = \{v_1, v_2, \dots, v_{N_s}\}$ and find out the current best solution $\lambda_c = \lambda_j$, $j = \arg \max_l (v_l)$, $1 \leq l \leq N_s$.

Set $\lambda_b = \lambda_c$ and $v_b = v_c$.

Step 1. Copy the current best path λ_c to each test solution λ_l , $1 \leq l \leq N_s$. For each test solution λ_l , $1 \leq l \leq N_s$, generate two random integers r_1 and r_2 , $1 \leq r_1 \leq N$, $1 \leq r_2 \leq N$, $r_1 \neq r_2$. Generate the new test solutions by swapping $\lambda_l(r_1)$ and $\lambda_l(r_2)$. Calculate the corresponding objective values v_1, v_2, \dots, v_{N_s} for the new test solutions.

Step 2. Sort v_1, v_2, \dots, v_{N_s} in increasing order. From the best test solution to the worst test solution, if the test solution is a non-tabu solution or it is a tabu solution but its objective value is larger than the best value of all iterations v_b (aspiration level), then choose this solution as the current best solution λ_c and choose its objective value as the current best objective value v_c , go to step 3; otherwise, try the next test solution. If all test solutions are tabu solutions, then go to step 1.

Step 3. If $v_c < v_b$, set $\lambda_c = \lambda_b$ and $v_c = v_b$. Insert the swapped indices of the current best solution λ_c into the tabu list. Set the inserting point of the tabu list $t_l = t_l + 1$. If $t_l > T_s$, set $t_l = 1$. If $i < l_m$, set $i = i + 1$ and go to step 1; otherwise, record the best path index and terminate the algorithm.

5. SIMULATIONS

10 experiments are conducted to validate the algorithm proposed in this paper. We recorded each word's pronunciation 10 times. Then we have 100 training observation sequences. For each word, we used the tabu search algorithm and the forward-backward algorithm to train the HMM respectively, and then we can obtain two sets of HMM model parameters and compare them. In this paper, the length of the tabu list $T_s = 20$, the threshold of the probability $P_{th} = 0.17\%$, the number of the iteration $I_m = 800$, the number of the solutions in each iteration $N_s = 20$. The initial model parameters are created randomly and are normalized to satisfy the equation (4) and (5).

In each experiment, the HMM training using the forward-backward algorithm will be terminated when the increase of the average log probability less than 0.00001 and TS-HMM training will be terminated after 800 iterations.

In this paper, we compared the HMMs trained by the tabu search algorithm and the forward-backward algorithm respectively. Simulation results are shown in Table 1. They are made up of two parts: P_s and P_d . P_s denotes the average log probability of the HMM generated by the 10 training observation sequences of this HMM and P_d denotes the average log probability of the HMM generated by the other 90 training observation sequences of the other HMMs.

As shown in Table 1, the HMMs trained by the tabu search algorithm that have higher average log probabilities than the HMMs trained by the forward-backward algorithm except experiment #6. It means the HMMs trained by the tabu search algorithm can better describe and recognize the training observation sequences. The experiment #6 is not satisfying because the better optimum is not encountered during searching, thus the whole search procedure is not globally optimal.

6. CONCLUSIONS

This paper proposes the TS-HMM training method. The tabu search algorithm is employed to repair the HMM model parameters λ and make $P_n(O|\lambda)$ highest. The simulation results indicated that TS-HMM training has a higher probability in finding the global optimal parameters with better performance than the forward-backward algorithm. Besides, parallel implementation of TS algorithm can be employed to reduce searching time such that its searching time can compare with other heuristic algorithms.

REFERENCE

1. L. R. Rabiner, "A tutorial on Hidden Markov Models and Selected applications in speech recognition," *Proceeding of IEEE*, , Vol. 77, No. 2, 1989, pp. 257~285.
2. Joseph Picone, "Continuous speech recognition using Hidden Markov Models," *IEEE ASSP Mag.*, Vol. 7, No. 7, 1990, pp. 26~41.
3. D. Burshtein, "Robust parametric modeling of duration in Hidden Markov Models," *IEEE Trans. on Speech & Audio Processing*, , Vol. 4, No. 3, 1996, pp. 240~242.
4. F. Jelinek, "Continuous speech recognition by statistical methods," *Proceeding of IEEE*, Vol. 64, No. 4, 1976, pp. 532~556.
5. S. E. Levinson, L. R. Rabiner and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *The Bell System Technical Journal*, April 1983, pp. 1035~1074.
6. F. Glover and M. Laguna, "Tabu search," *Kluwer Academic Publishers*, 1997.

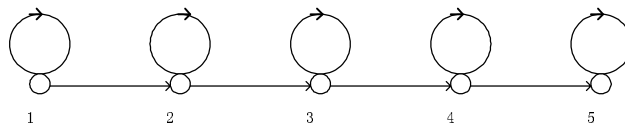


Fig. 1. A five states left-right model

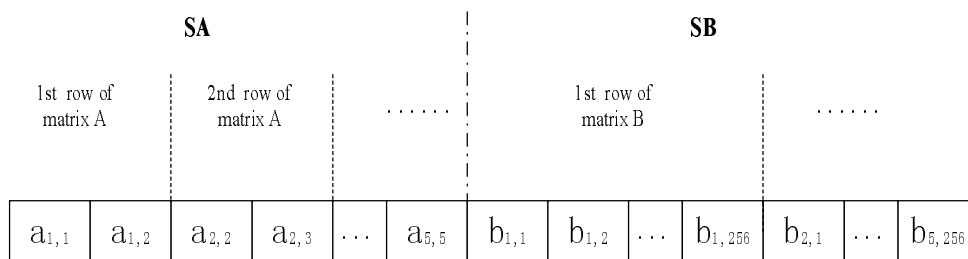


Fig. 2. The string representation of HMM

Table 1 The comparison of average log probability obtained with two algorithms

Experiment	TS		Forward-Backward	
	P_s	P_d	P_s	P_d
#1	-3.3463	-9.0765	-4.2946	-8.9249
#2	-4.8036	-9.4363	-4.8116	-8.4035
#3	-4.6056	-8.4663	-5.6599	-8.3756
#4	-3.5379	-8.3139	-4.3562	-7.9967
#5	-4.6579	-9.9391	-5.1033	-7.6877
#6	-4.5324	-9.3661	-4.3394	-8.6031
#7	-3.2752	-9.3218	-4.7167	-8.4162
#8	-3.6225	-9.3123	-4.3607	-8.2275
#9	-3.8032	-9.6469	-4.5107	-9.2521
#10	-4.3190	-8.2152	-4.4864	-7.8755

簡易影片字幕文字辨識法及其詢答應用

林川傑 劉哲嘉 陳信希

國立臺灣大學資訊工程學研究所

{cjlin, jjliu}@nlg2.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

摘要

影片字幕通常反應影片部份內容，可以輔助影片內容檢索。雖然在新的影片格式如 MPEG2 以上，字幕內容可以輕易取得，但是仍有大量早期的影片，需要進行影片文字辨識，才能擷取字幕內容。本文提出一套簡易的中文字幕辨識法，包括影像擷取、字幕尋找、背景去除、字元切割、光學文字辨識、及後處理。我們以 Discovery Channel 影片作為訓練和測試的資料，以兩部影片作集外測試，其辨識率分別為 82.3% 和 85.9%，而集內測試可以達 94.2% 的正確率。在 Pentium-4 1.7G，256M RAM，40G 7200 轉速的 IBM 硬碟等配備下，處理平均 495MB 大小的影片，需要 29 分 11 秒。這套影片文字辨識法，對於影片數位圖書館的建立，以及後續的影片內容檢索有很大的助益。本文以影片檢索和詢答系統為例，說明影片文字辨識的應用。

1. 緒論

在多媒體的資訊時代，影片數量相當龐大，也隱藏豐富的知識，如何有效的檢索與擷取影片內容，就成為重要的考量要素之一。在著名的數位圖書館計畫 Infomedia (Wactlar, 2000)，影片資料庫的管理使用就是個顯明的例子。由於每一幕影片片段透過聲音和影像傳遞重要信息，因此以語音檢索或影像相似性比對，找出相關的影片片段，是最直接的方式。但是在相關的影像相似性比對技術，還未完全成熟之前，影片字幕仍是重要的檢索依據。在新的影片格式如 MPEG2 以

上，字幕內容可以輕易取得，但是仍有大量早期的影片，需要進行影片文字辨識，才能擷取字幕內容。本論文擬提出一套簡易的中文字幕辨識法，擷取影片中的文字，供後續的檢索使用。

光學文字辨識的研究歷史很早，且已經有很好的成果。紙本的文字資料透過掃描器輸入成影像檔，透過文字辨識系統的處理，將影像檔辨識成文字檔。另外，手寫文字辨識也有突破性的發展。相對的，影片文字辨識比傳統的文字辨識挑戰性高，主要的原因是後者所辨識的格式，大多是白底黑字，而影片上的文字大多出現在複雜的背景上，並且字通常不大，前者會遇到解析度較差及複雜背景的問題。

過去已有些論文與影片文字辨識相關，Wu 等人(1997, 1998)嘗試以 connected component 的方式尋找圖片中的文字，其方法在圖片中的結果不錯。但應用在影片中的文字尋找時，會因為影片中的文字大多有著複雜的背景，造成字會與其它圖形物件相連在一起，因而產生不好的結果。Lienhart 等人(1998, 2000)則利用 color segmentation、contrast segmentation、geometry analysis、texture analysis 等方法尋找影片中的文字。Li、Doermann 和 Kia (2000)採用類神經網路的方式，來找尋影片中的字串。Li 和 Doermann (1999) 也利用多張影像的整合，來加強文字影像的解析度。在影片的整合部份，Smith 和 Kande (1997)利用字幕、影像的移動及臉部辨識等方法，來簡化影片的大小。Sato 等人(1998)利用文字的修補及多張影像的文字擷取，來提昇影片文字辨識的正確率。

本文以影片字幕的文字辨識為主，研究的對象是中文字。論文共分十節，第二節將介紹影片文字辨識可能的問題，以及系統架構。第三節至第八節分別描述系統每個模組採用的策略和方法，並以 Discovery Channel 影片為訓練和測試材料，實驗各個模組的效能。第九節討論影片字幕文字辨識結果在詢答系統上的應用。第十節做總結，並探討未來的方向。

2. 影片文字辨識系統架構

影片中會出現的文字有兩種：一種是字幕文字(subtitles or captions)，一種是畫面文字(texts in image)。字幕文字常出現在畫面上特定的地方，例如字幕和標題常出現在影片的下方橫書，而中文的人名頭銜等文字通常出現右方或左方直書。畫面文字指的則是出現在字幕文字外畫面中的文字，例如商店的招牌，車子的車牌號碼等。它本身本來就是畫面的一部份，因此不僅會因鏡頭的移動而改變位置，而且在與字幕等重合時，會被字幕所遮蓋。由於字幕正是影片中旁白或是對話的文字呈現，字幕往往提供了影片內容的重要資訊，因此本文的重點是“字幕文字”的處理。

影片字幕文字辨識首先面對的問題就是如何擷取畫面，以及如何去除背景。一般影片中的文字，大多出現在很複雜的畫面上，因此系統第一步工作就是要去除複雜的背景，並將影像轉成白底黑字，再交給文字辨識軟體辨識。最後文字辨識後的結果，再運用自然語言處理技術來提昇正確率。圖一顯示整體系統架構圖。

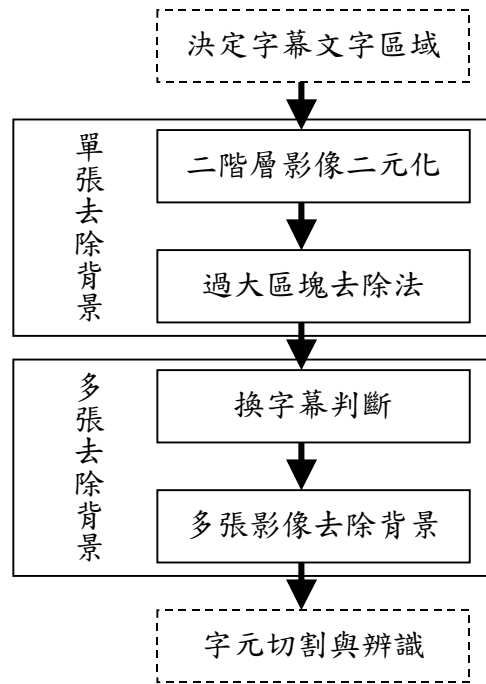
為了訓練及測試這個系統的效能，實驗的素材取自於 Discovery Channel 影片，內容包羅萬象，有自然生態、歷史文化、科技新知、軍事飛行、旅遊冒險、生活集錦等知識性題材。在第九節將結合詢答系統，對此影片集提出問題，透過字幕辨識結果，快速尋找相關的影片片段並做評估。

3. 字幕文字區域之決定

“字幕文字”的特徵如下：(1) 呈垂直或水平排列；(2) 字的本身和影片會有強烈的對比色，一般會有明顯的邊框；(3) 一定是在影片前方，不會被影片畫面所遮蓋；(4) 會連續出現兩字以上；(5) 不會太大，一般不會高於影片高度的1/3；通常也不會太小，因為太小，人類也無法識別；(6) 固定的高度(或寬度)與字體大小；(7) 固定的顏色。我們根據這些特徵來尋找字幕的位置。

3.1 影像二元化

在進一步處理影像文字之前，我們會先將影片轉成二元化影像(Binary



圖一、影片文字辨識系統架構圖

Image)。這個步驟是做影片文字處理過程中常用的方法，它可以幫助將複雜的背景單純化，並讓字幕文字更易於顯現出來。

我們在撥放影片的過程中擷取出影像畫面來，一秒取 2 幕，並將之存為 BMP 檔。在 BMP 檔中，圖片上每一點的色彩均是以其 RGB 值來記錄。所謂 RGB 值，就是該色彩由多少亮度的紅(Red)、綠(Green)、藍(Blue)色光所合成，記成 RGB(色值,色值,色值)，其中色值的範圍由 0~255，0 表最暗(無此色彩)，255 表最亮。

利用影像中各點的 RGB 值，我們就能將原影像轉換成二元化的影像。演算法如下：

設定二元化門檻值 SegColorScore

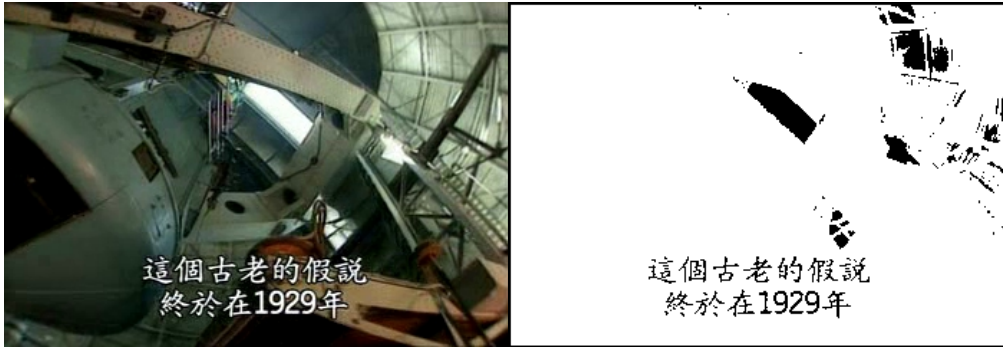
對影像中的每一點而言：

若 該點色彩 R 值、G 值、B 值均 $>$ SegColorScore

則 更改該點色彩為黑色(RGB(0,0,0))；

否則 更改該點色彩為白色(RGB(255,255,255))。

其中 SegColorScore 值在實驗中設定為 190，這是多次實驗所得結果最好的經驗值。圖二為影像二元化結果的範例。我們可以很清楚地看到，字幕文字在二元化影像中更容易與背景分離開來。此次實驗中字幕文字在原影片中為白色字體、黑色邊框，轉換為二元化影像後成為黑色字體。



圖二、影像二元化範例

3.2 決定字幕文字區域

將影像二元化之後，接著要決定畫面中字幕文字的區域在哪裡。這裡我們利用了字幕文字的另一項特性：在橫書的字幕中，若在字幕上劃上一條水平線，則此線會通過不少的直豎筆劃。由於在文字書寫的習慣上，直豎筆劃的寬度大致一定。而且影片畫面的其他地方，也不容易出現此種多個連續相同寬度的黑色區塊，我們便可利用這個資訊，計算得知字幕所在位置。

考慮在同一水平高度位置 $height_i$ 的各點，將連續相鄰的黑色點視為同一區段 (segment)，則可得在 $height_i$ 的水平位置上的黑點區段集合為 $SEGMENT_i = (segment_{i1}, segment_{i2}, \dots)$ 。考慮各 $segment_{ij}$ 中所含的黑點數，若相鄰區段所含黑點數相差不超過一定值 δ 時(本實驗中 δ 值設為 3)，則將這些區段視為同一組 (group)，如此可得在 $height_i$ 的水平位置上的黑點組集合為 $GROUP_i = (group_{i1}, group_{i2}, \dots)$ 。令 $Seg(group_{ij})$ 為構成此 $group_{ij}$ 的區段個數，我們定義 $height_i$ 為字幕區域的可能分數 $SASA_i$ (Score as Subtitle Area) 為：

$$SASA_i = \sum_{j=1}^{|GROUP_i|} Seg(group_{ij}) \times \log_2 Seg(group_{ij}) \quad (1)$$

考慮如下的範例：point 列中以 0 表示白點，1 表示黑點。

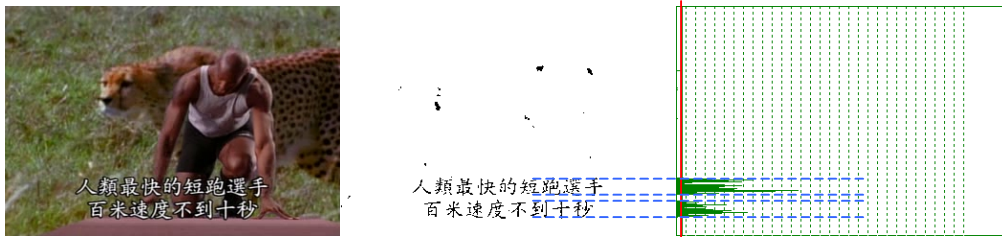
```

point: 0011101111100110001110111111101111111100111101110110111111
segment: --111-22222--33---444-5555555-6666666666--7777-888-99-AAAAAA
group:  --1                -2                --3                -4
Seg(group):  =4                =2                =3                =1

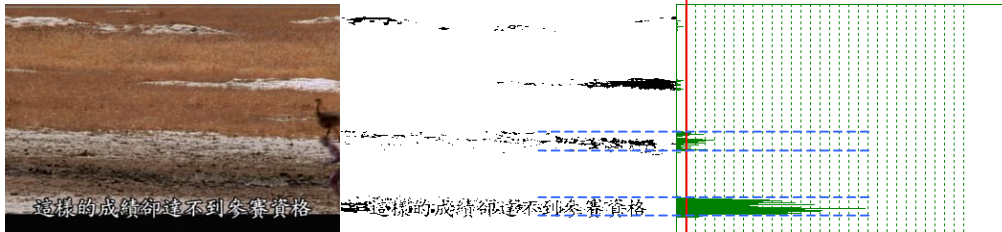
```

這個例子的 $SASA$ 分數為 $4 \log_2 4 + 2 \log_2 2 + 3 \log_2 3 + 1 \log_2 1 = 14.75$ 。

若影像畫面總高度有 m 列，算出各列的 $SASA$ 值，求其平均值 \overline{SASA} 。我們



圖三、決定字幕文字區域結果範例一



圖四、決定字幕文字區域結果範例二

將各列中 *SASA* 值高於平均值者視為字幕區域，如此便可決定出字幕出現在畫面中的位置了。圖三和圖四為決定字幕文字區域的範例，其中左圖為原影像圖片，中圖為其二元化圖片，右圖則為各列之 *SASA* 值。右圖中縱軸表畫面高度，橫軸即各列 *SASA* 值之大小，垂直紅線為平均值 \overline{SASA} 的所在，而水平的藍色虛線即為判斷所得的字幕文字區域。

3.3 實驗評估

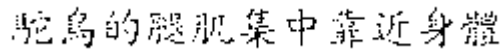
字幕文字區域的實驗資料來自三部 Discovery Channel 的影片：「閃電」、「動物之最」和「鯨魚探奇」，其中各出現 69、66 和 71 行字幕。實驗結果如表一所示，正確率各為 76.7%、39.8% 和 82.0%，召回率幾乎可以達到 100.0%。如同圖四所示，在畫面中央的碎石石子路，因為出現多個連續相同寬度的色塊，因此被誤判為字幕文字區域。對於不正確的字幕文字區域，尚可在 OCR 處理的過程中，因為與標準字庫過低的相似度而被過濾。因此這裡高召回率就比正確率要來的重

表一、字幕尋找結果評估

字幕個數	實際	系統判斷	正確	正確率	召回率
閃電	69	90	69	76.7%	100.0%
動物之最	66	161	64	39.8%	97.0%
鯨魚探奇	41	50	41	82.0%	100.0%



圖五、SegColorScore 設為 140 的影像二元化結果



圖六、SegColorScore 設為 180 的影像二元化結果

4. 單張影像去除背景法

調整 SegColorScore 的值將影像二元化時，我們發現一個有趣的現象。若是將 SegColorScore 值調得太低，會有字幕文字和殘留背景交雜重疊的情形出現；但若是 SegColorScore 值調得太高，得到的字幕文字又會過於模糊破碎，不利於 OCR 的進行。在第 3 節所用的值為 190，就只能留下較模糊的字幕文字影像。

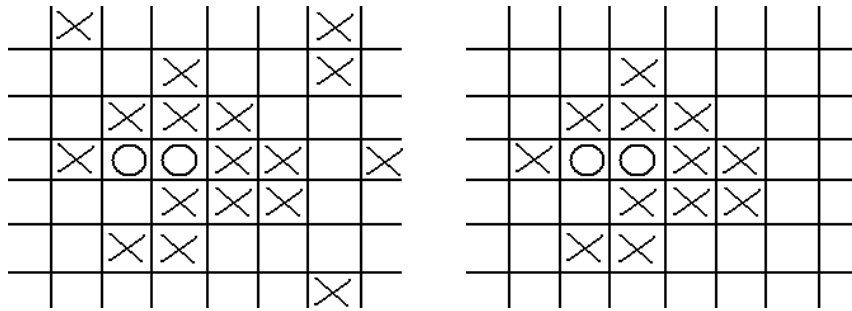
為了讓 OCR 結果能更準確，就需試著留下清晰的文字，並做去除背景的動作。本節先介紹在單張圖片中，利用一些文字與背景畫面不同的特性，將背景去除的方法。下一小節再討論利用多張同字幕文字的畫面去除背景的方法。

4.1 二階層影像二元化

在第 3 節中，我們設定了 SegColorScore 值。色彩 RGB 三值均高於 SegColorScore 的點會被轉為黑色，而其中一值低於 SegColorScore 者則被轉為白色。然而 SegColorScore 的值大小，影響著二元化後的字幕清晰度。例如圖五和圖六分別為 SegColorScore 設為 140 和 180 的結果，可見 SegColorScore 設為 140 時，字幕文字清晰，但是有過多背景殘存下來。而 SegColorScore 設為 180 時，背景雖多已去除，但所留下來的字幕文字較為模糊。

這裡我們提出了一種新的方法，稱為二階層的影像二元化方法。利用二個不同的 SegColorScore 值所得的二元化影像，我們可以把字幕文字清晰的留下，且背景去除。方法如下：

將同一張圖片利用二個高低不同的 SegColorScore 值(分別為 HiSegColorScore 和 LowSegColorScore)，所得的二元化影像重疊在一起，參考圖七左圖。其中 ○ 為 HiSegColorScore 值所得到的黑點處，而 × 則表示



圖七、二階層影像二元化示意圖

駝鳥的腿肌集中靠近身體

圖八、由圖五及圖六二階層影像二元化的結果

LowSegColorScore 值所得的黑點處，值得注意的是○也會是 LowSegColorScore 值所得的黑點處。接著我們只保留與○相連接的×黑點區塊，其餘未與任何○相連的×區塊均改為白色點，結果如圖七右圖。圖八即為圖五和圖六利用二階層影像二元化所得的結果，為更清晰的字幕文字圖。

4.2 過大區塊去除法

當字幕文字區域出現高亮度的背景時，第 4.1 小節的做法便不足以去除之。如果此背景為一大片的高亮度區塊(二元化後則大片黑色區塊)，則我們提出一個方法在單張圖片中清除此一過大區塊。對於零碎的小區塊，下一節會試著利用多張同字幕的圖片資訊來清除背景。

圖九上圖為字幕文字區域中有過大黑色區塊的範例，下面為提出的演算法：

Range = (字幕文字區域高度) ÷ 4;

Total = Range × Range × 0.9;

對字幕文字區域中每一黑點**均做如下檢查：**

檢視以該黑點為左上角、邊長為 Range 的正方形區域，

若 正方形區域中黑點數 \geq Total (即此區域中有九成以上均為黑點)

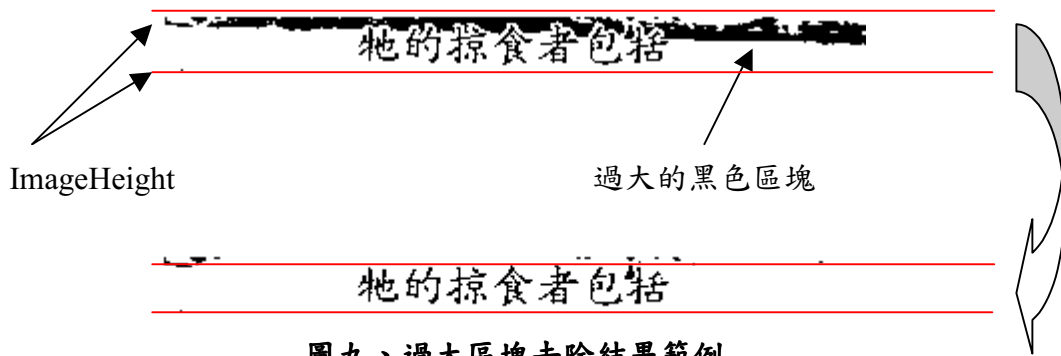
則 把該黑點及與其相連接的黑色區塊都改成白色

結束

圖九下圖為其去除過大區塊後的結果影像。

5. 多張影像去除背景法

雖然利用單張去除背景的方法已可很有效地去除大部份的背景，但小區塊高



圖九、過大區塊去除結果範例

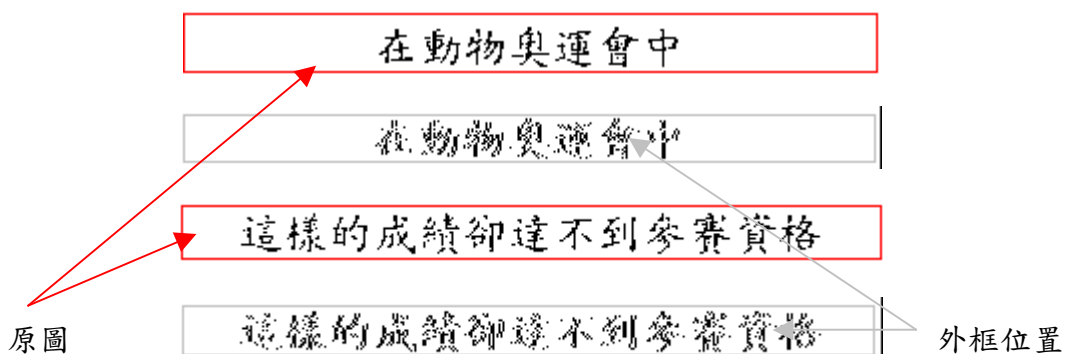
亮度的背景卻不易被清除。如圖九下圖在字幕文字周圍仍有一些背景圖案存在。

字幕文字的另一項特性是：它不會隨著畫面鏡頭的移動而改變位置，然而背景圖案卻會隨之更動。利用這項特性，把同一字幕文字的二元化影像重疊在一起，留下出現頻率高的黑點，即為字幕文字部份。Sato 等人(1998)就是利用這樣的想法，去除移動中的背景。

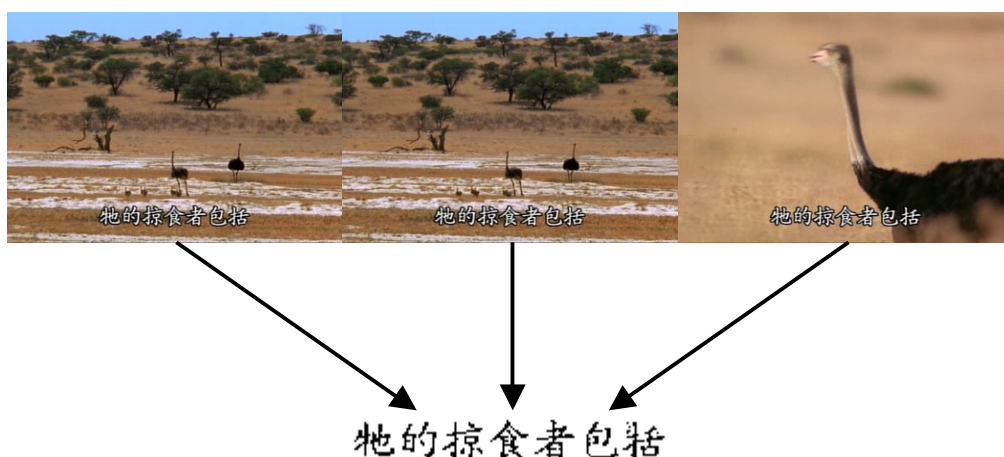
要用多張影像去除背景，就得先判斷那幾張影像為同一字幕。此後兩小節分別介紹判斷更換字幕，以及多張影像去除背景的方法。

5.1 換字幕判斷

在換字幕的判斷上，我們保留了字形的外框，藉以偵測字幕文字是否已經更換。以圖十為例，我們先將每一張圖中所有黑色區塊的外框位置記錄下來。當讀入下一張圖後，將其外框位置與前一張的外框位置做比較。若位置不同的比率超過某一門檻值 SceneChangeScore 時，就將其判斷為字幕轉換點。實驗得 SceneChangeScore=0.6 時，有最佳結果。



圖十、字幕外框範例



圖十一、多張去背結果範例

表二、字幕轉換點判斷結果

	換字幕次數	判斷錯誤次數	正確率
閃電	69	0	100.0%
動物之最	66	3	95.5.0%
鯨魚探奇	41	0	100.0%

我們也做了一個小小的評估。同樣以第3節實驗所用的三部影片，來看看各行字幕的轉換點是否判斷正確。由表二結果可知判斷正確的成功率相當高。

5.2 多張背景去除法

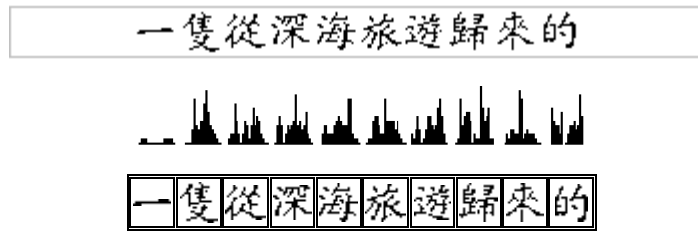
對於同一字幕文字的連續圖片，我們利用多張影像去除背景的方法來得到一個字幕文字區域大小、內容僅含字幕文字的圖片。令 NumFrames 為此連續圖片的張數，考慮字幕文字區域中的任一點位置。若在連續圖片中該點位置有九成 ($\text{NumFrames} \times 0.9$) 以上的圖片出現黑點，則在結果字幕文字圖片上該點位置亦設為黑色，否則設定為白色。

圖十一為多張影像去除背景的結果範例，可以看到在背景部份比圖十清除地更乾淨。

6. 字元切割

經由前幾節對影像處理的步驟，現在每一個字幕文字都已有對應的白底黑字結果字幕文字圖片，接下來就可以用傳統 OCR 的方法辨識出字幕中的文字。

OCR 的第一步是決定每個字元的邊界。由於我們先前決定字幕文字區域



圖十二、字元切割結果範例

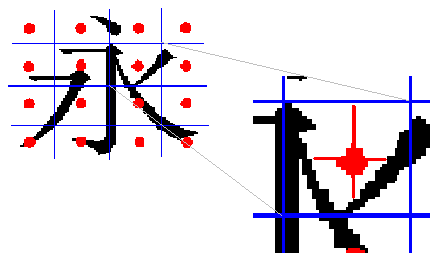
時，等於已經決定了各字元的上下邊界，因此現在所要判斷的，是各字元的左右邊界。

字元切割的方法，大多都用垂直投射的方法(Lu, 1995)：如圖十二所示，對於結果字幕文字圖片的每一個水平位置，將不同高度上的所有黑色點投影到水平線上。在中文字與字之間因為有空間，因此會出現投影量為 0 的間隔 (gap)。由於中文字元大多在正方形區域內，真正是字元間隔的兩間隔之間寬度(即字的寬度)也會大約等於結果字幕文字圖片的高度 ImageHeight (即字的高度)。我們的方法如下：若兩間隔間的距離為 $\text{ImageHeight} \times 0.7 \sim \text{ImageHeight} \times 1.4$ 之間，則此兩間隔切出一個字元。圖十二的最下圖即是字元切割的結果範例。

7. 文字辨識

OCR 的研究主要分 on-line 及 off-line 兩種。而早在 1970 年代，就有許多的研究針對 on-line 手寫或是簡單的印刷字體辨識，到了 1980 年代 off-line 的研究才慢慢變多，而 off-line 的辨識系統又主要分統計式模型及結構分析兩種。論文中所採用的是統計式模型，為 Oka 在 1982 年所提出的方法。

以圖十三為例，首先將讀入的影像檔等分為 16 區塊，由每一區塊中點開始，觀察其上下左右四個方向。如果在這區塊中該方向上有黑點存在，則記錄特徵值為 1，否則為 0。如此一來共可以得到 64 個(16 區塊 \times 4 個方向)特徵值。




圖十三、記錄影像特徵值(Oka, 1988)


探索遺傳學的奇異世界

0000003-1-01.bmp: (56)探 (52)權抓微 (51)撇攏很多育擺
0000003-1-02.bmp: (58)孝素 (57)幸索 (56)案業 (55)希考常 (54)途
0000003-1-03.bmp: (56)遠速 (53)達遺逝遺道 (52)道情運
0000003-1-04.bmp: (60)傳 (52)博 (51)偉 (50)佈搏佛格 (49)踏彈像
0000003-1-05.bmp: (59)學卡層 (51)銀峰單軍旁革帶
0000003-1-06.bmp: (59)的勾鄉 (52)稀將哺特豹 (51)均擠
0000003-1-07.bmp: (60)奇 (53)槍青賽考 (52)希老奔逢旁
0000003-1-08.bmp: (60)異 (52)具其提隻 (51)週各姿域農
0000003-1-09.bmp: (59)世 (53)親摺甘種 (52)奇發音實奮
0000003-1-10.bmp: (63)界 (55)善 (53)輩華 (52)谷才在像毒舞

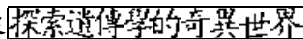
圖十四、OCR 辨識後所得前十名的候選字集

我們蒐集了一些文字圖片做為標準字庫以做比對之用，將這些圖片也都以特徵值的方式記錄下來。當一個新的文字影像要做辨識時，首先找出其特徵值，再與標準字庫中的特徵值做比對。我們以計算相同特徵值個數來做為計分依據，當對應的特徵值相同時，相似度加一分，因此相似度的分數會界於 0 到 64 之間，分數越高代表兩個影像越相似。

下面舉一個簡單的例子：設有一文字影像為，分別和標準字庫中的「傳」與「博」字做比對，其特徵值分別如下：

 10101000100011011101001000001011111010100100010111111111111001111
傳: 1010100000001101010101110000010011111010100100010111111111111001111
博: 1110100000011111010001000100100111101100010001011111111111101101

比對結果與標準字庫的「傳」字相似度為 60 分，與「博」字相似度為 50 分，因此「傳」成為此文字影像的第一名候選字。

圖十四是影像經過 OCR 辨識後所得前十名的候選字集。可以看到第一名的候選字即為正確答案的比例非常高，而且正確答案也都出現在前十名的候選字集中。下一節我們會利用 OCR 後處理的方法，選出不在第一名的正確答案。

僅選取第一名做為辨識結果的實驗評估數據記錄在表三中，其中影片「萬象雜誌：基因的秘密」為集內測試，而「金字塔之王」、「埃及豔后」影片則為集外測試。實驗數據顯示，以 OCR 辨識第一名做為答案的正確率，集內測試可達 91.5%，集外測試也可達到 78.5%和 81.5%，已有不錯的成績。

表三、OCR 辨識實驗結果

影片	TOTAL	CORRECT	ERROR	MISS
基因的秘密	809	739(91.5%)	69 (8.5%)	0
金字塔之王	684	537(78.5%)	110(16.1%)	37(5.4%)
埃及豔后	750	611(81.5%)	86(11.5%)	53(7.1%)

8. OCR 後處理

在前面的小節中我們知道 OCR 第一名的集內測試正確率約為 90%，而前十名的正確率約為 95%。所以我們的目標就是希望我們能將 OCR 出來的字正確率能逼近 95%，要讓第二名以後的正確答案能被選中，而原本就辨識第一名即為正確答案的則保持不變。所以如何提高辨識率，是本節主要要介紹的課題。

8.1 基本後處理方法

每個文字影像辨識出來的結果都取其前十名做候選字，並有相對應的相似度分數，圖十四中標示為 (分數)候選字。首先我們將分數與第一名差在 4 分以上(包含 4 分)的候選字剔除，以減少後處理比對時所帶來的雜訊(在圖十四中以灰色表示被剔除的候選字)。接下來從第一個字開始，每次連續看三個文字影像(令其為 ABC)，查詢其候選字組 A_iB_j 或 B_jC_k 是否在字典中為二字詞或是多字詞的一部份。若是，則分數為兩候選字相似度相乘，否則分數為零。比較所有 A_iB_j 和 B_jC_k 所得分數的高低，若最高分的候選字組為 A_iB_j ，就選定 A_iB_j 為 AB 的辨識結果，然後由第三個文字影像 C 開始，重覆前面的步驟(看 CDE)。若是最高分的候選字組為 B_jC_k ，則選擇第一個文字影像 A 的第一名候選字 A_1 為其辨識結果，然後由第二個文字影像 B 開始，重覆前面的步驟(看 BCD)。

8.2 後處理實驗策略

為了了解在做後處理時，是否有必要考慮所有相似字的任意組合，或是第一名的候選字有其重要性，甚至候選字組是否出現在字典中的資訊有何幫助，我們提出了三種不同的後處理策略，並且和僅取第一名候選字、以及僅以長詞優先法則所得的結果做比較。

[策略一]在選取文字影像的候選字時，考慮所有的候選字組合。

[策略二]在兩兩一組查詢字典的時候，其中一個候選字一定是第一名的候選字。

例如在辨識文字影像 AB 時，只查詢 A_1B_1 、 A_1B_2 、...、 A_2B_1 、 A_3B_1 、...

這樣的做法是為了增加對第一名候選字的信任。

[策略三]連續看四個文字影像，若其所有的候選字組合中出現一組在字典中為四字詞，則以這四字詞為其辨識結果。否無則再連續看三個文字影像，是否有三字詞的候選字組合。如果有則選為辨識結果，無則接著以策略二的方式選擇結果。

8.3 實驗評估

標準字庫的蒐集來自六部 Discovery Channel 的影片(「動物之最」、「蛇類奇觀」、「萬象雜誌：基因的秘密」、「天然景觀—落磯山脈」、「神戶大地震」、「達爾文之島」)，總共有 7,818 個文字影像檔，得到 2,256 不同字的特徵值。

表四到表六分別是針對三部影片做字幕文字辨識所得的實驗結果，其中「萬象雜誌：基因的秘密」影片為集內測試，而「金字塔之王」、「埃及豔后」影片則為集外測試。表格中各欄位資訊解釋為：

TOTAL: 影片中字幕總字數

CORRECT: 辨識正確的字數

ERROR: 文字出現在標準字庫中但辨識錯誤的字數

MISS: 文字未收錄在標準字庫中的字數

Improve: OCR 後處理所得的改進值

最佳優先: 選取第一名候選字為辨識結果

長詞優先: 選取候選字組中出現在字典最長詞者為辨識結果

表四、OCR 後處理結果(影片「萬象雜誌：基因的秘密」)

	TOTAL	CORRECT	ERROR	MISS	Improve
最佳優先	809	739(91.5%)	69(8.5%)	0	-----
長詞優先	809	753(93.1%)	56(6.9%)	0	1.6%
策略一	809	751(92.8%)	58(7.2%)	0	1.3%
策略二	809	759(93.8%)	50(6.2%)	0	2.3%
策略三	809	762(94.2%)	47(5.8%)	0	2.7%

表五、OCR 後處理結果(影片「金字塔之王」)

	TOTAL	CORRECT	ERROR	MISS	Improve
最佳優先	684	537(78.5%)	110(16.1%)	37(5.4%)	-----
長詞優先	684	544(79.5%)	103(15.1%)	37(5.4%)	1.0%
策略一	684	546(79.8%)	101(14.8%)	37(5.4%)	1.3%
策略二	684	559(81.7%)	88(12.9%)	37(5.4%)	3.2%
策略三	684	563(82.3%)	84(12.3%)	37(5.4%)	3.8%

表六、OCR 後處理結果(影片「埃及豔后」)

	TOTAL	CORRECT	ERROR	MISS	Improve
最佳優先	750	611(81.5%)	86(11.5%)	53(7.1%)	-----
長詞優先	750	614(81.9%)	83(11.1%)	53(7.1%)	0.4%
策略一	750	635(84.5%)	62(8.3%)	53(7.1%)	3.0%
策略二	750	640(85.3%)	57(7.6%)	53(7.1%)	3.8%
策略三	750	644(85.9%)	53(7.1%)	53(7.1%)	4.4%

由表四到表六可以發現，策略三是所有方法中效果最好的。它在集外測試可有 82.3%和 85.9%的正確率，在集內測試更可達到 94.2%。

另外因為標準字庫僅蒐集 2,265 個字，在集外測試中就分別有 5.4%和 7.1%的字幕文字無法做比對。擴充標準字庫絕對是未來的重要工作之一。

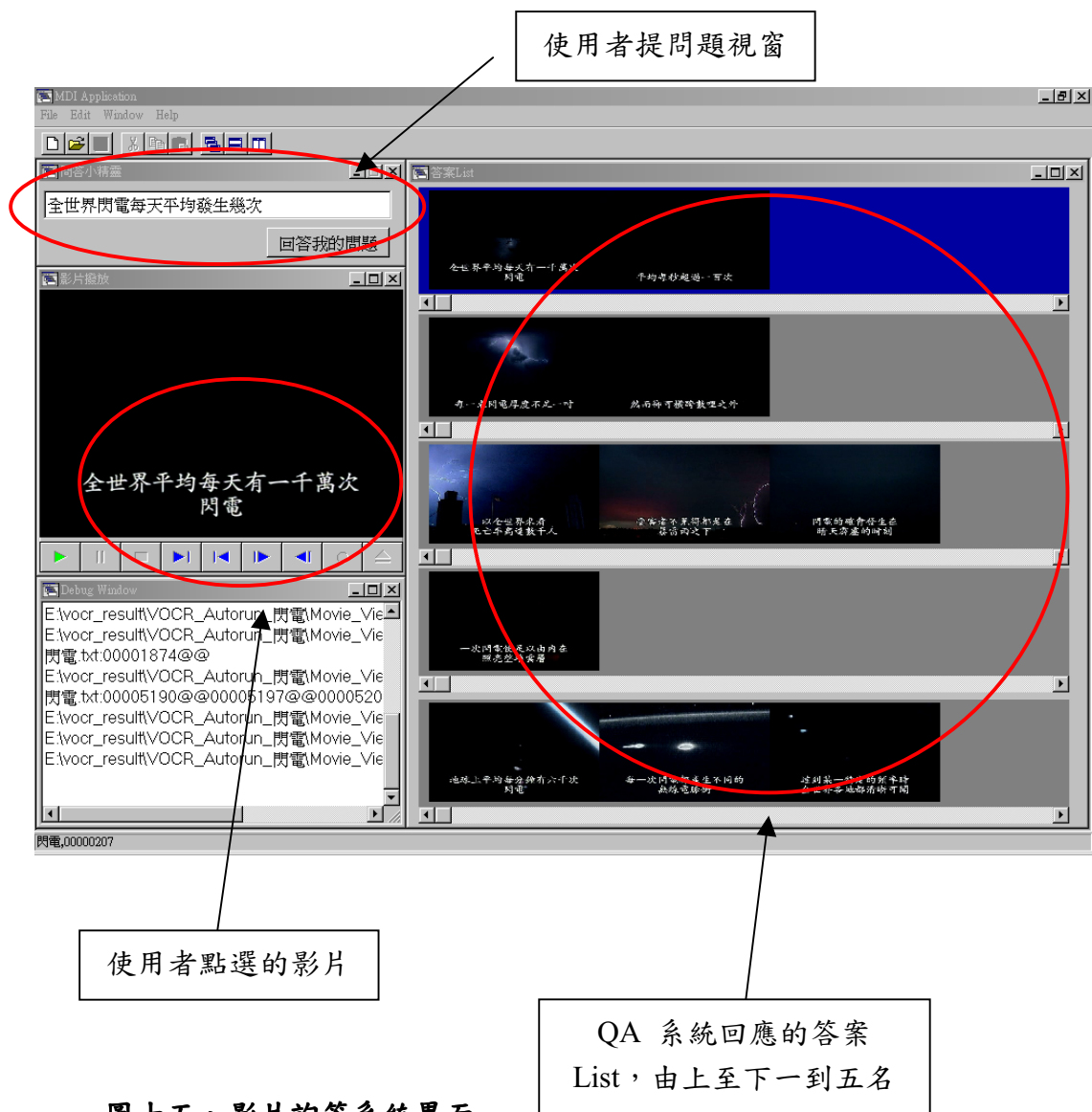
將 OCR 後處理策略三的實驗結果與未做 OCR 後處理(最佳優先)的實驗結果做分析，比較結果記錄在表七之中。其中“True to False”表示原本依最佳優先判斷正確、策略三卻判斷錯誤的情形，而“False to True”表示原本依最佳優先判斷錯誤、策略三可判斷正確的情形。由表七可見 OCR 後處理僅會造成 0.7%左右多餘的錯誤，卻能多判斷正確 3.0%到 5.2%的文字，由此可知後處理的幫助。

表七、策略三與最佳優先之結果比較分析

	Total	Result	True to True	True to False	False to True	False to False
基因的秘密	809	94.2%	738(91.2%)	6(0.7%)	24(3.0%)	41 (5.1%)
金字塔之王	647	87.0%	532(82.2%)	5(0.8%)	31(4.8%)	79(11.2%)
埃及豔后	697	92.4%	608(87.2%)	3(0.4%)	36(5.2%)	50 (7.2%)

9. 影片檢索與詢答系統

藉由先前各節所描述的影片 OCR 方法，我們已經能將影片中的字幕擷取出來。接著就可結合資訊檢索技術，或是問題詢答技術，發展查詢影片的檢索與詢答系統。



圖十五、影片詢答系統界面

9.1 影片詢答系統

圖十五是本實驗室所發展的影片詢答系統的界面。使用者可以輸入感興趣的問題，系統所找到的答案會陳列在右方視窗中，每一名答案均列出它的代表字幕畫面讓使用者參考。若使用者想要觀看其中一個答案的原始影片，可點選其畫面，系統就會自動將影片撥放出來。

詢答系統部份的技術，採用的是 Lin 等人(2001)對各類異質的資料進行詢答的技術中，針對 Video OCR 所提的方法。其中特別需要注意的是，因為搜尋的文本是文字辨識以後的結果，而由表四和表五得知，文字辨識的效能是 82.3%以上，尚未能完全正確。因此傳統詢答技術所用到的字串比對(不論是關鍵詞、同義詞，甚至是語意關係樹等)，都必須考慮到 OCR 產生的錯誤。因此含答文句的

計分方式，就要引入 OCR 相似度的分數：

$$\begin{aligned} score(qw_i, pw_j) &= 0 \quad \text{if } |qw_i| \neq |pw_j| \\ &\text{else} = \left(\frac{\sum_{k=1}^{|qw_i|} Ocr(qc_k, pc_k)}{|qw_i|} \right) \times weight(qw_i) \end{aligned} \quad (2)$$

其中 qw_i 和 pw_j 是做字串比對的兩個詞， $|qw_i|$ 表示詞 qw_i 中的字元個數， qc_k 是詞 qw_i 的第 k 個字元 (pw_j 的表示法同 qw_i)。 $Ocr(qc_k, pc_k)$ 是 qc_k 和 pc_k 的 OCR 相似度分數，為第 7 節特徵值比對所得分數除以 64，以使值的範圍落在 0~1 之間。 $weight(qw_i)$ 則為原先 qw_i 和 pw_j 相同時所得的分數。

9.2 實驗評估

9.2.1 問題的來源

問題的來源是 Discovery 繁體中文網站(<http://chinese.discovery.com>)，其“教育工程”內所擺放的影片與相關問題。這個網站提供一個免費而龐大的影像記錄片庫給教師使用，每個節目鎖定一個主題，讓老師在課堂上透過影像與特別編製的教師手冊、特別設計的活動、以及相關的網路資源，輔助學生在課程內或課程外的學習。

經由該網站上的資料，挑選了幾個影片的相關問題，來對我們的詢答系統做評估。至於為何要用該網站上的問題，主要原因是因為其問題較具公平性、一般性。其中，影片的片名有「大象」、「木星」、「哈伯望遠鏡：太空的奧秘」、「蛇之眼」、「鯨」及「地球科學面面觀：閃電」等。

9.2.2 詢答系統準確率

我們在此使用 MRR(Mean Reciprocal Rank)評估詢答系統的準確率，這是在詢答系統評比(TREC QA-Track)中所用的評量方法(Voorhees, 2000)。

在六部影片中共有 43 個問題，其結果如表八所示。MRR 分數為 0.1848 ($0.1848 = (4 + 5/2 + 3/3 + 1/4 + 1/5) / 43$)，答題率為 32.6% (14/43)。

表八、影片詢答系統評估結果

第一名	第二名	第三名	第四名	第五名	沒答出來
4	5	3	1	1	29

為何此處 MRR 只有 0.1848，觀察問題後，主要為下列幾點原因：

- (1) 與問題有關的關鍵字文字未收錄在標準字庫中。

例如問題「冰雹如何形成？」中的「雹」字。

- (2) 問題的用詞與影片中的用詞不一樣。

例如問題「木星繞行太陽一週需時多久？」，在影片中出現的是「木星環繞太陽一周，須地球時間十二年。」

- (3) 需要更精準的問句規則處理。

以問題「閃電可以到達多熱的程度？」為例，目前系統在處理以“多”字來詢問程度的問題時，僅試著找尋屬於數量的答案，而不是更精確地找尋溫度描述詞“華氏五萬度”來做為答案。

- (4) 需引入常識或理解文字。

以問題「歷史上第一位做閃電實驗的人是誰？」為例。影片中有提到 1752 年，富蘭克林做了閃電實驗，但是並沒有提到“第一位”這樣的字眼，系統並無法得知他就是第一人。

扣除掉第一點影片文字辨識系統的錯誤，目前的詢答系統大多只做到關鍵字比對與依問題類型尋找專有名詞答案的程度。而本文中所用的問題大多需要很多其它的相關知識或是語意上的分析，才能夠回答的出來。而問題的類型也偏重於 Why 及 How，這類型的問題本來就比較難解。因此，未來這部分的研究是個很大的挑戰。

10. 結論

本文介紹 Video OCR 的所有步驟，從影片的畫面擷取、尋找畫面中字幕文字區域的位置、去除背景、字元的切割、OCR 及透過自然語言的技術提高 OCR 的辨識率，並介紹了 Video OCR 的一些應用。Video OCR 的辨識率在集內測試，約有九成以上的正確率，而集外測試也有八成以上的正確率。

從得到的結果回去看錯誤的地方大概分下列幾種：一、收集的字元不夠多；二、背景去的不夠乾淨；三、字元切割錯誤；四、OCR 後處理錯誤。以下我們一點一點來討論。

第一個問題是收集的字元不夠多。目前我們所收集的字元共有 7,818 個文字影像檔，其中共有 2,256 個不同的字，而一般的常用字共有 5,401 個，所以很多字辨識不出來。

第二個問題是背景去除地不夠乾淨。在大部分的情況下，背景都可以經由單張影像去除背景的方法，加上多張連續同字幕畫面去除背景的方法來清除。但是遇到不會移動的背景，加上字幕後的背景又是破碎的黑白色素混雜時，往往就無法順利清除乾淨。而未去除的背景會影響後面的字元切割及 OCR 辨識。

第三個問題是字元切割錯誤，這常常是因為背景沒去除乾淨所造成。

第四個問題是 OCR 後處理錯誤。這個問題主要也是來自第一和第二個問題，因為原本所收集的字元中就沒有出現，也就不會出現在候選字中，所以常會將原本辨識為第一名正確的字，因為前後字的關係反而被辨識錯了。

此外，本次實驗資料均取材自 Discovery Channel 的節目，字幕的字型、顏色或是大小都比較一致。未來在處理其他來源的影片文字時，就必須再更進一步地探討不同字型、不同格式所帶來的影響。

在未來的工作中，也將試著以現有字型(例如標楷體等)建立標準字庫，以更完整的字元集來作實驗。找出更好的去除背景方法以及套用更好的 OCR Model，並且和以現有的語言模型做後處理結果來比較，以期能更進一步利用這樣的工具去挖掘出影片所帶有的資訊。

參考文獻

Discovery Channel, <http://chinese.discovery.com/>.

Li, Huiping and Doermann, David (1999). "Text Enhancement in Digital Video Using Multiple Frame Integration." *Proceedings of SPIE, Document Recognition IV*, pp. 1-8.

Li, Huiping; Doermann, David and Kia, Omid (2000). "Automatic Text Detection and

- Tracking in Digital Video.” *IEEE Transactions on Image Processing*, Vol. 9, No. 1, pp. 147-156.
- Lienhart, Rainer and Wernicke, Axel (2000). “On the Segmentation of Text in Videos.” *IEEE Int. Conference on Multimedia and Expo (ICME2000)*, Vol. 3, pp. 1511-1514, also as *Technical Report MRL-VIG00005*.
- Lienhart, Rainer and Wolfgang, Effelsberg (1998). “Automatic Text Segmentation and Text Recognition for Video Indexing.” *Technical Report TR-98-009, Praktische Informatik IV*, University of Mannheim.
- Lin, Chuan-Jie; Chen, Hsin-His; Liu, Che-Chia; Tsai, Jin-He and Wong, Hong-Jia (2001). “Open-Domain Question Answering on Heterogeneous Data.” *Proceedings of Workshop on Human Language Technology and Knowledge Management*, ACL.
- Lu, Y. (1995). “Machine Printed Character Segmentation – An Overview.” *Pattern Recognition*, Vol. 28, pp. 67-80.
- Oka, R. I. (1982). “Handwritten Chinese-Japanese Characters Recognition by Using Cellular Feature.” *Proc. 6th Int. Joint Conf. on Pattern Recognition*, pp. 783-785.
- Sato, Toshio; Kanage, Takeo; Ellen K.Hughes; Smith, Michael A. and Satoh, Shin’ichi (1998). “Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption.” *ACM Multimedia Systems Special Issue on Video Libraries*.
- Smith, Michael A. and Kande, Takeo (1997). “Video Skimming and Characterization Through the Combination of Image and Language Understanding Technique.” *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 775-781.
- Voorhees, (2000) “QA Track Overview (TREC) 9.” [on-line]
Available: <http://trec.nist.gov/presentations/TREC9/qa/index.htm>
- Wactlar, H., (2000) “Informedia - Search and Summarization in the Video Medium.” *Proceedings of Imagina 2000 Conference*.
- Wu, Victor and Riseman, Edward M. (1998). “TextFinder: An Automatic System to Detect and Recognize Text in Images.” *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 21, No. 11.
- Wu, Victor; Manmatha, R. and Riseman, Edward. M. (1997). “Finding Text in Images.” *Proceedings of the 2nd intl. conf. on Digital Libraries*. pp. 1-10.

中文語料庫構建及管理系統設計

馬偉雲 謝佑明 楊昌樺 陳克健

中央研究院資訊科學研究所

{ma, morris, kchen}@iis.sinica.edu.tw

ms8903@[cis.scu.edu.tw](mailto:ms8903@cis.scu.edu.tw)

摘 要

一個中文帶詞類標記的平衡語料庫，在中文自然語言的研究與應用上是不可或缺的角色，然而要構建一個數量且高品質的語料庫往往需要投入大量的人力及時間，為了提升構建的效率以及提高語料庫管理的機能，在管理方面，我們建立了以文本為單位的資料庫系統作為語料庫的架構，並開發一管理介面。構建方面，我們設計了一套構建流程以及開發了四個子系統來幫助我們完成構建語料庫的工作。構建語料庫的第一步是語料蒐集，為此我們設計了一個語料蒐集介面，能夠蒐集網路上豐沛的電子文件資源，並在某些特定網址來源當中自動分析其文本格式。第二步是語料的斷詞及標記，我們透過未知詞擷取模組作為斷詞標記的前處理，大幅提高了斷詞標記程序的正確性，減少其後人力校正的負擔。最後一步是人工檢驗，我們設計了操作簡便的人工檢驗介面，並結合詞典與舊版本的語料庫提供使用者參考來做出正確的判斷，完成斷詞、詞類與句子的編修工作。

1. 簡介

語料庫為本 (corpus-based) 的研究是近年來語言學及計算語言研究的一個重要發展 [Church, Mercer93]、[Chen94]、[Huang95]。

建立帶詞類標記的平衡語料庫是一個浩大的工程，但也是自然語言研究的基礎工程 (infrastructure)。其效應可由現存語料庫，如布朗，LOB，London-Lund 等所衍生的大量研究成果得到證明。

「中央研究院平衡語料庫」簡稱「中研院平衡語料庫」(Sinica Corpus)，是世界上第一個有完整詞類標記的漢語平衡語料庫。於 1994 年公開提供給國內外學術研究使用；以期在使用過程中得到回饋，在完成目標規模前可以做必要的修正。1997 年開放的研究院語料庫 3.0 版已經達到五百萬目詞的預計規模。我們期望在 2003 年能夠達到一千萬目詞的規模。

建構一個中文的平衡帶詞類標記的語料庫，包括語料的收集、語料的整理(包含語料清潔、為語料分類、加詞類標記等等)、人工的校定。從早期的建構經驗中，由於缺乏合適的工具，我們遭遇了以下的困難：(1) 早期我們以檔案的形式作為語料的最小單位，一份檔案通常包含數十篇不同的文本，文本的格式屬性以符號配合文字在文本之前表示之。這樣以檔案為單位的架構對整體語料的管理及統計相當不便，同時對人工校對的工作分配而言，也比較沒有彈性。(2) 大量語料的蒐集、維護、分類、校定交由各人以檔案的方式加以處理，並無統一的處理介面，形成管理上的紊亂。(3) 過去我們使用詞庫小組自行開發的系統 [Chen, Liu92] 將語料加以斷詞標記，卻發現由於文本當中未知詞的存在，使得系統的斷詞表現大幅下降，而必須倚靠事後大量的人力來加以合分詞。(4) 人工校正時，由於斷詞以及詞類標記時常有歧異現象發生，校正者沒有工具立即檢驗相關的用法或範例，造成判斷上的困難，使得有時候斷詞標記的校對品質因人而異。這些問題除了造成在管理上的困難之外，同時在人工校正的過程中花費了大量的人力及時間，在斷詞標記的一致性上也不易維持。

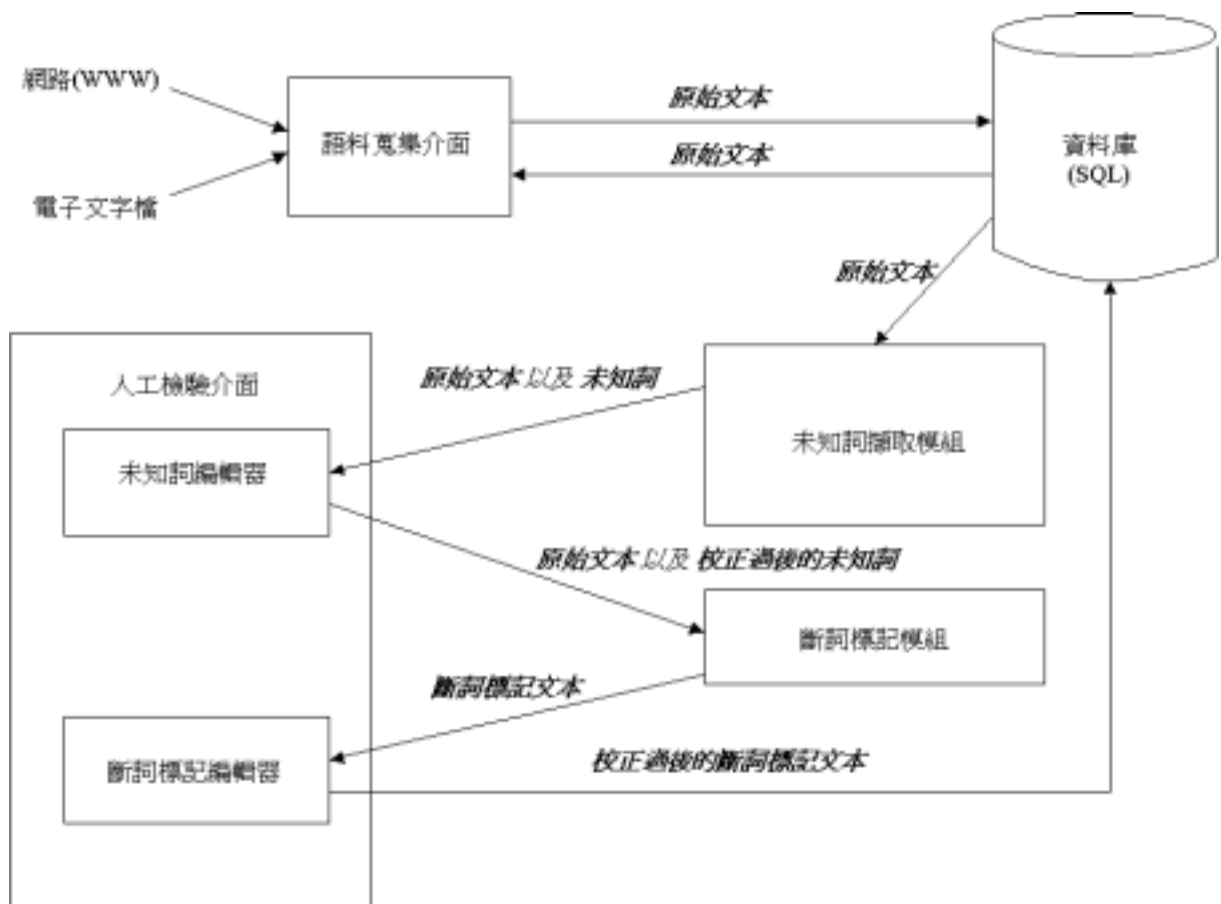
因此，我們認為要有效率的建構一個高品質的平衡語料庫，必須要訂定出一套嚴

謹的管理方式，發展出一套理想的構建流程以及開發一些合適的工具，才有可能達成。在語料的管理上，我們捨棄以往以檔案為單位的管理方式，而改用文本為單位，並採用資料庫 (database) 的架構儲存，一篇文本即為資料庫的一筆記錄 (record)，其格式屬性以此記錄的欄位 (field) 加以儲存。在語料的蒐集上，除了直接自相關單位取得語料之外，網路上的各類文本也是一重要來源，為此我們設計了一套語料蒐集介面，能夠將網址上的文本擷取下來，在某些特定網址上能自動分析其文體作者等格式，減輕人力的負擔。在語料的斷詞標記方面，未知詞的存在使得斷詞標記模組的錯誤率大增，因此，我們提出一改善的作法，在斷詞標記之前，先利用自行開發的未知詞擷取模組來擷取文本的未知詞，之後再加以斷詞標記，如此使得斷詞的品質大幅度的提高。最後是人工檢驗的部分，為了確保語料庫的品質，有必要將語料庫做一地毯式的人工檢驗，我們在 windows 環境下開發了一套由滑鼠操作的人工檢驗介面，除了操作簡便快速之外，最大的特色在於系統連結了詞典資料庫以及之前版本的語料庫，讓檢驗者能快速的查詢詞典中相關的資訊，也可以在先前版本的語料庫當中查詢相關的範例，來幫助檢驗者做出正確的判斷。

本文第二節將詳述建立平衡語料庫的系統設計架構及流程。第三節對系統在實務上的表現上作一討論。最後是結論。

2. 建立平衡語料庫的系統設計

本系統主要包含四個子系統，分別是語料蒐集介面、未知詞擷取模組、斷詞標記模組、以及人工檢驗介面。其中語料蒐集介面以及人工檢驗介面是在 windows 的環境，而未知詞擷取模組及斷詞標記模組是在 linux 的環境下運作。子系統間相互的溝通均是採用 client-server 的模式。系統流程圖如圖一。



圖一：系統流程圖

首先語料蒐集介面將使用者所指定的網址抓回原始文本(text)，並針對特定網址自動分析其文體作者等格式，經過使用者檢驗無誤後，將原始文本存入資料庫，準備後續斷詞及詞類標記工作，由於文本中包含各式各樣的未知詞，造成斷詞及詞類標記的困擾，因此後續處理的第一步是先經過未知詞擷取模組擷取未知詞，再將原始文本及抽取出的未知詞送交斷詞標記模組將原始文本加以斷詞標記，最後以人工逐句檢查，經過此一連串的程序以及最後的人工檢驗，得到斷詞標記後的文本，可以存入資料庫當中作為語料。

這樣的設計流程具有如下的優點：(1) 未知詞擷取模組在斷詞標記模組之前即已將未知詞標示出來，提高斷詞標記的正確率，降低人力修正的負擔。(2) 針對同一篇文本，蒐集語料和人工檢驗可以由不同的人擔任，這是因為這兩項工作所需的專業知

識不盡相同，因此在設計上以資料庫為中繼站將兩者分離，在人力的分配上這樣的設計提供了更大的彈性。(3)自動化的處理提高語料處理的效率及一致性。(4)人工檢驗介面提供完整的詞典以及舊有的標記訊息和範例，幫助使用者做判斷，提高資料的正確性。

2.1 語料蒐集介面

隨著網際網路的蓬勃發展，大量文本也以電子化的形式呈現，每個人可以輕易地閱讀這些文本，不受空間及時間的限制。從蒐集語料的觀點出發，也可以善用環境中的這項特點，設計一個使用介面，在網路環境裡獲得語料，標記屬性特徵資訊，將成果以資料庫形式儲存。

早期蒐集語料的方式是以電子文字檔為主，必須事先定義其內容格式，之後使用者將大量的語料以文本檔案的形式加以分類管理，再由程式員撰寫程式進行各項統計查詢的動作。面對數量眾多的語料，使用者必須自行紀錄工作的流程與進度，修正錯誤時也必須用傳統的方法找到該錯誤檔案再加以修正，容易造成版本不統一的問題。

為了改善這些問題，使蒐集語料的工作更有效率，我們設計的介面，必須符合下列二個要求：(1)統一版本及減少人工時程(2)利用網路改善工作流程。為了達成以上兩個要求，語料蒐集介面提供以下功能：(1)文本擷取(2)修改屬性特徵(3)自動分類資料擷取。

2.1.1 統一版本及減少人工時程

在蒐集語料的過程中，使用者除了找到文本之外，還必須判斷其屬性特徵加以人工標註，包括文類、文體、語式、主題、媒體、作者姓名、性別、國籍、母語、出版單位、出版地、出版日期、版次〔Chen94,Huang95〕。系統擷取語料時會事先自動判別文本的相關屬性特徵，當無法擷取到所有的屬性特徵時，則以視窗形式呈現的介面（如圖二所示），提供了線上修改的功能，使用者可直接在圖形介面上進行屬性特徵的

修改，避免文書編輯上的困擾，也能統一修改的格式與製作出來的版本，減少對單一文本處理的時程。



圖二：語料蒐集視窗程式介面

2.1.2 利用網路改善工作流程

目前網際網路上的電子文件普遍以 HTML 格式存在，使用者可透過本介面直接輸入目的網頁之網址，由本系統過濾掉不必要的 HTML TAG，呈現本文部分供使用者審核編輯；並由事先定義好、並撰寫於程式中的規則，判斷該文本的屬性特徵，再由使用者加以確認及修改，如此可減少繁瑣的文字檔案編輯工作；這些事先定義好的規則有賴於程式員對目的網頁的原始結構加以分析。如圖二所示之介面，使用者至中時電子報 (<http://news.chinatimes.com/>) 擷取新聞文本；程式員根據中時電子報 HTML 內文格式

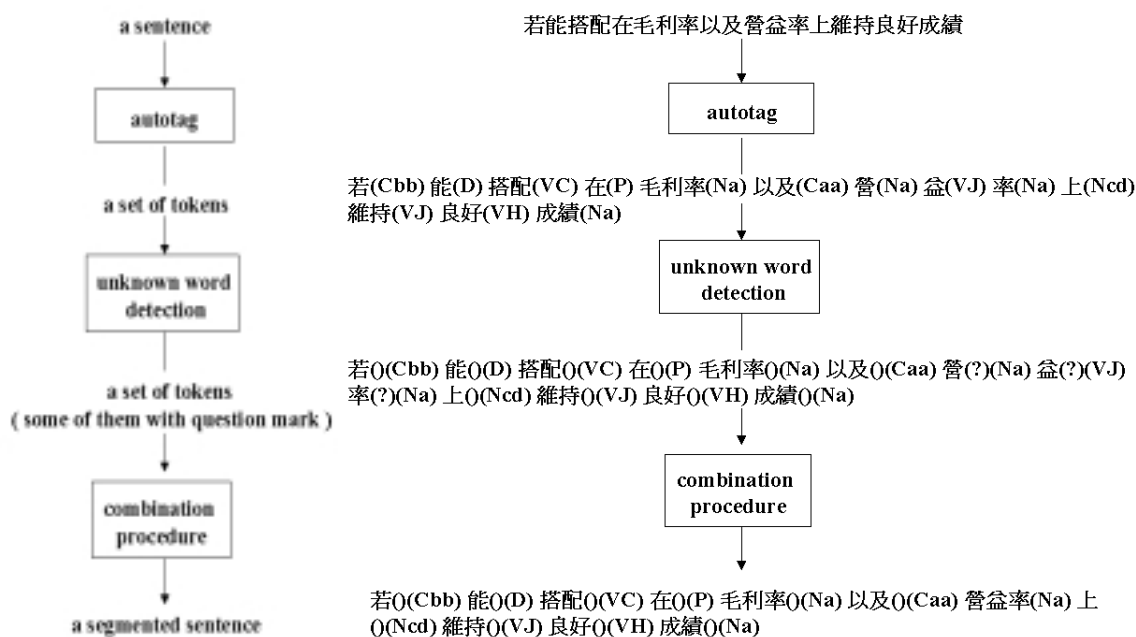
撰寫擷取的程式碼。使用者處理好語料文本後，該文本將直接透過本介面存進資料庫，讓後續的語料庫標記程式能夠直接銜接，不必再進行文字檔的轉換工作。

2.2 未知詞擷取模組

從以往到現在處理斷詞及詞類標記的過程中，大多數的錯誤來自於未知詞的出現，例如：人名、縮寫、複合詞等在文本中出現的機率相當高，自動斷詞及標記是以詞典資料為本，因此只要有一個未知詞出現就會造成斷詞及標記的數個錯誤。

目前大多數的未知詞擷取作法與斷詞程序相連結，先將文本做斷詞，未知詞由於詞典未收錄，故會被斷成切割過短的單字或字組，之後再由統計式〔Sproat90〕、〔Smadja93, 96〕、〔Wu93〕、〔Chang97〕或法則式〔Yeh91〕、〔Lin93〕的技術加以合併這些短字組成為未知詞。

本系統所採用的未知詞擷取模組的演算法大體上亦採取這樣的概念，除了與斷詞程序相連結之外，在斷詞完畢後，會再經過一未知詞偵測程序〔Chen, Bai97〕，決定那些單字是未知詞的一部份，那些是本來就能夠獨用的單字。屬於未知詞一部份的單字之後能夠啟動合併程序，有機會結合前後字組合併成未知詞。如圖三。



圖三：未知詞擷取模組流程圖及範例

當一個句子經過斷詞以及未知詞偵測程序後，未知詞會被切分為較短的成分，正常的單字詞會被上下文規律區分出來，因此我們可以區分出何者屬於正常的單字詞，何者屬於未知詞的成分，之後再利用統計及構詞律等合併程序得到未知詞，並根據未知詞的內部結構猜測其詞類〔Chen, Bai, Chen97〕。

如圖三的範例所示，當輸入的句子「若能搭配在毛利率以及營益率上維持良好成績」經過斷詞標記後得到「若(Cbb) 能(D) 搭配(VC) 在(P) 毛利率(Na) 以及(Caa) 營(Na) 益(VJ) 率(Na) 上(Ncd) 維持(VJ) 良好(VH) 成績(Na)」，接下來經過未知詞偵測程序，判斷出「營(Na)」、「益(VJ)」、「率(Na)」可能是未知詞的成分(以問號標示)，經過合併程序將之合併成為未知詞「營益率」，再分析「營益率」的內部結構猜測其詞類為「Na」。

在合併程序的部分，我們同時採用統計式以及法則式的技術，統計方面，不同於前人從整體語料獲得詞彙組成的統計資訊，而是從單篇文本中獲得統計資訊，這是因為我們認為大多數的未知詞在單篇文本的統計比在整體語料的統計上更有意義。

法則式的部分，我們利用構詞學與構句學的理论以及觀察到的現象，加以訂定若干合併的規則及限制，例如某些詞性標籤或是如百家姓的訊息等等都是法則所規範的對象。

2.3 斷詞標記模組

經過未知詞擷取模組之後，雖然已經可以以一個斷詞完成並標記好的句子形式呈現，然而所擷取出來的未知詞當中仍有錯誤發生的可能，因此有必要以人工校定的方式直接增刪或修改這些未知詞，再利用斷詞標記模組參考這些人工確認後的未知詞，將原始文本重新斷詞標記，得到較為正確的結果。

因此未知詞擷取模組將未知詞傳送給人工檢驗介面，如「營益率(Na)」。由人工判斷是否要修改或增刪這些未知詞。之後再將這些正確的未知詞以及原始文本傳送給斷詞標記模組。

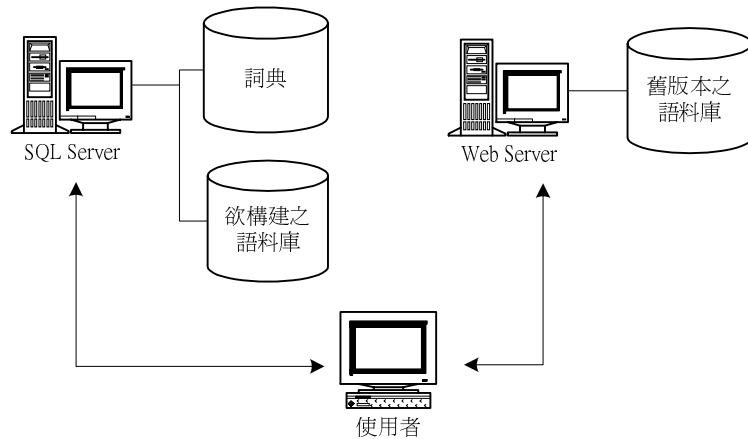
我們的斷詞標記模組是以詞典資料為本，因此當有了某一文本的未知詞資訊後，立即動態產生一部未知詞詞典的資料結構來幫助我們斷詞標記，我們可視作新增了一部未知詞的詞典和原有詞典搭配，在斷詞標記的程序中同時參考這兩部詞典，來完成這一文本的斷詞標記工作。

2.4 人工檢驗介面

經由前面章節的介紹，我們瞭解到如何透過網際網路收集 WWW 上的文本做為語料，再經由未知詞擷取模組及斷詞標記模組將文本加以斷詞標記。以上動作均採用自動化處理的方式，不僅有效率而且保持了斷詞標記的一致性，達到百分之九十五以上的正確結果。然而，在斷詞及詞類標記的過程中，因為有切分歧義、標記歧義與未知詞判斷的困難，所以，標記出來的詞類結果難免會有些許錯誤的地方。為了達到高品質的語料庫，人工檢驗的過程還是不能避免。因此，需要一個系統來輔助處理斷詞標記後的檢驗動作。該系統需提供簡單的介面與簡易的操作方式，來節省人力及時間。並能夠取得已有的詞典、語料標記資訊做為參考，來幫助使用者完成斷詞、詞類與句子的編修。

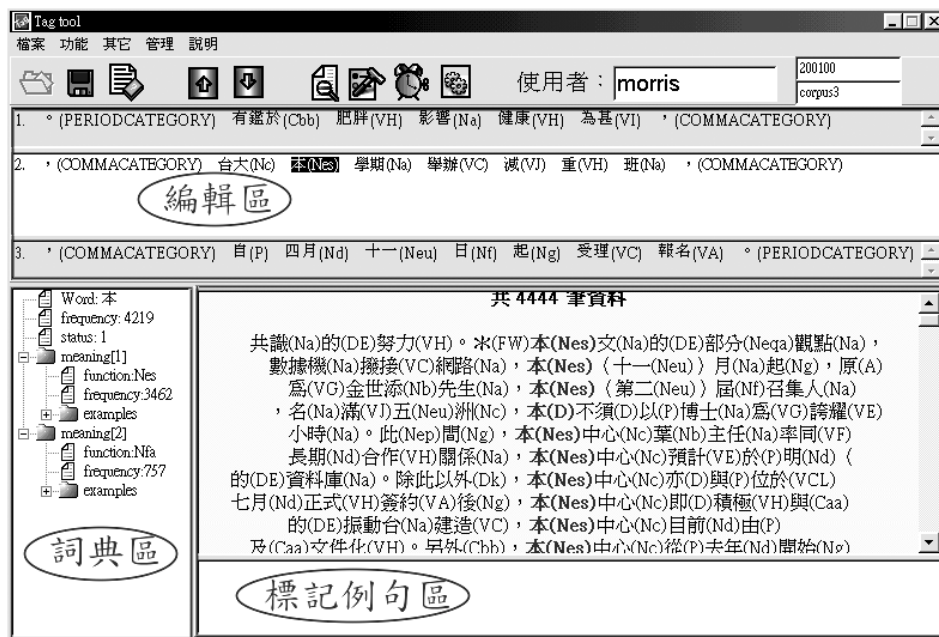
2.4.1 系統功能

本系統的開發平台是在微軟作業系統上。使用者可以在任何一台電腦上，透過區域網路的連線，連上後端語料庫及詞典的資料庫伺服器，來選取欲編輯的文本與查詢某詞的詞典資料等。該後端資料庫伺服器是以微軟 SQL Server 所建構，系統並連結以 Web Server 建構的舊版本語料庫，提供相關例句供使用者參考。如圖四所示。



圖四：系統與資料庫分佈架構

語料編輯介面規劃三個區域來表示這些訊息，分別是編輯區、字典區與標記例句區，所呈現的畫面如圖五所示。



圖五：人工檢驗畫面

現在就分別對這三個主要區域加以描述：

(1) 編輯區

該區所編輯的對象是斷詞標記後的文本，經由使用者選取文本後載入所呈現出來的。在這裡可以看到有三個小區域，分別代表上一句、編輯句與下一句，分別以不同

的顏色與字體來區分，這樣劃分的目的在提供使用者瞭解編輯過與將要編輯的句子內容，好讓使用者盡量達到詞類標記的一致性。

當使用者要進行修改時，所有的功能，都可以在主功能表中看到。我們將常用的部份以快速鍵或快速按鈕表示，以增加使用者編輯的效率。使用者可以在編輯列中利用滑鼠去選取欲處理的詞或是配合鍵盤的左右鍵去選取，然後配合快速鍵或是按下滑鼠右鍵依不同目的去選擇欲執行的項目。這些功能可以分類如下：

(a)詞的修改：包含了合分詞、改詞類、改特徵、去特徵等功能。其中，在合分詞的部份，系統會列出合分詞後的詞有哪些詞類，供使用者選擇。倘若只有一種選擇，系統會自行加入，盡量不讓使用者自行輸入，避免往後詞類標記不一致的狀況發生。在改詞類的部份，作法亦同上述的合分詞。

(b)句子修改與詞字數統計：在斷詞標記的過程中，有可能會有不正確的斷句結果產生，這時就需要合分句的功能。另外，使用者可以統計到目前為止文本當中已檢驗了多少個詞與字，系統本身也會在文本完稿之後重新編排句子序號並加以統計該文本的總詞字數，以提供使用者參考。

(c)記憶功能：在編輯的過程中，使用者會遇到某種錯誤可能一再的發生，並不希望每次都去改相同的錯誤。因此，系統提供了規則檔記錄的功能，紀錄使用者合分詞與改詞類的過程，當使用者對某詞的詞類討論出最後的定論，即可根據規則檔的記錄將目前的編輯句自動修正。系統也提供批次更新的功能作全文修正。批次更新的處理對象是全文內容，在全文當中找出符合規則的詞並加以列出，供使用者選擇是否更新取代之。

(d)提醒的功能：目的在告知使用者目前編輯句中，有哪些詞是歧義切分詞或詞綴，以不同的方式加以顯示，讓使用者注意到在處理該詞時要特別小心。

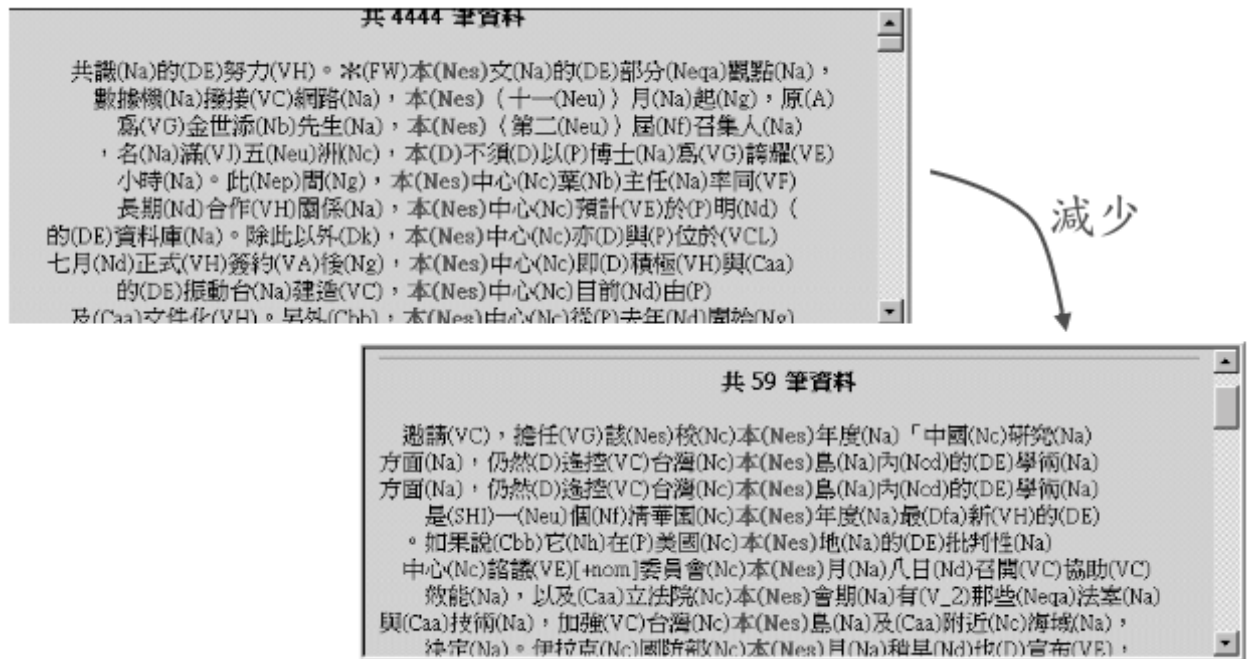
(2) 字典區

在資料庫伺服器中，我們建立了一個以詞為基礎的詞典資料庫，使用者可以輸入一個欲查詢的詞，系統會回應使用者該詞的詞類有哪些？頻率是多少？相關例句有哪

些？我們將這些結果顯示在字典區，以樹狀結構的方式呈現出來，讓使用者清楚地了解到該詞有多少詞類，以更確定詞類標記的對錯。

(3) 標記例句區

在這裡，我們整合了中央研究院開放出來供各界使用的網頁版本語料庫查詢介面 (<http://www.sinica.edu.tw/ftms-bin/kiwi.sh>)於系統當中。該語料庫的語料平衡分配在不同的主題和語式上，總共約有五百萬目詞之多，為現代漢語無窮多的語句中一個具代表性的樣本。該網頁版的語料查詢系統，提供使用者多樣式的查詢，在顯示的結果中，包含每個文句與其斷詞標記後的結果。本系統將查詢的步驟簡化，讓使用者只要在本系統中的編輯句裡，先點選欲查詢的詞，再按下滑鼠右鍵選擇查例句，就會列出與該詞相關的句子，且每個例句都是標記過的。如此一來，使用者可以參考這樣的結果來進行詞類標記的判斷。此外，有時為了避免過多的例句產生，系統提供了前後詞類的相關限制功能，以減少輸出，讓查詢到的結果更符合我們的需要，同時節省查詢時間。如圖六所示，以「本」作為查詢的關鍵字可得到 4444 筆例句，若限制前後的詞類為「Nc」及「Na」，查詢的例句數目則大幅精減為 59 筆。



圖六：精減相關例句查詢結果

2.5 整合後的語料庫構建管理系統介面

本系統以視窗介面整合了斷詞標記模組、未知詞擷取模組與人工檢驗介面，提供使用者文本選取與查詢介面、未知詞編輯介面、人工檢驗介面與語料類別修改介面，針對這四個介面內容描述如下：

(1) 文本選取與查詢介面

透過 2.1 節所描述的語料蒐集介面所抓來的文本，都會放在後端的語料庫伺服器中。當使用者欲檢驗文本時，可透過本介面以條件式的查詢方式從語料庫伺服器中取出。另外，系統也提供使用者以檔案的型式處理。參考畫面如下：

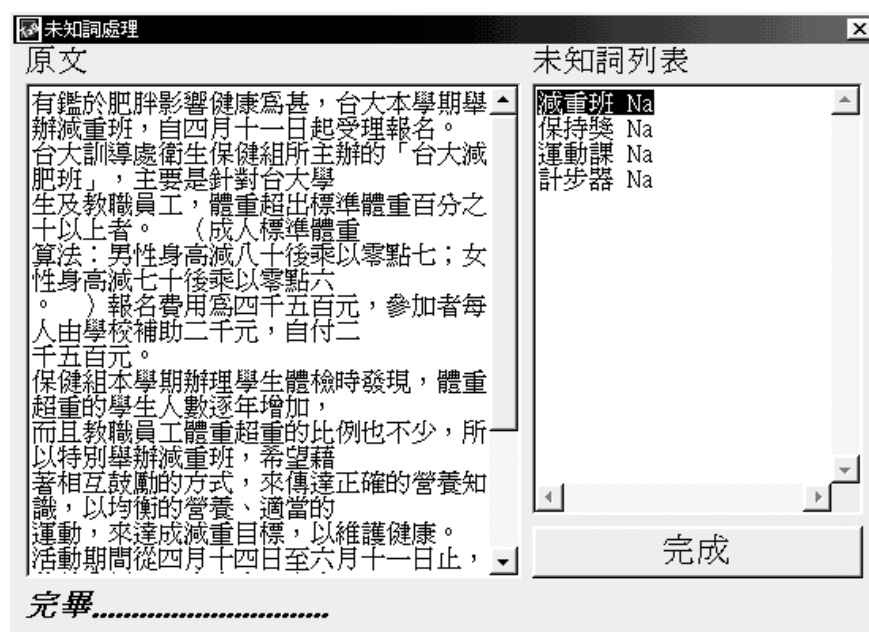


圖七：文本選取與查詢介面

(2) 未知詞編輯介面

在文本確定之後，倘若該文本已有斷詞標記過的資料，會直接取出進行上次未完的編輯動作。反之，系統會先將此未標記過的原始文本進行未知詞擷取，使用者可針

對擷取出來的未知詞進行新增、刪除及修改等編輯動作，如圖八所示。再將原始文本及編輯過後的未知詞送到斷詞標記模組加以斷詞標記。



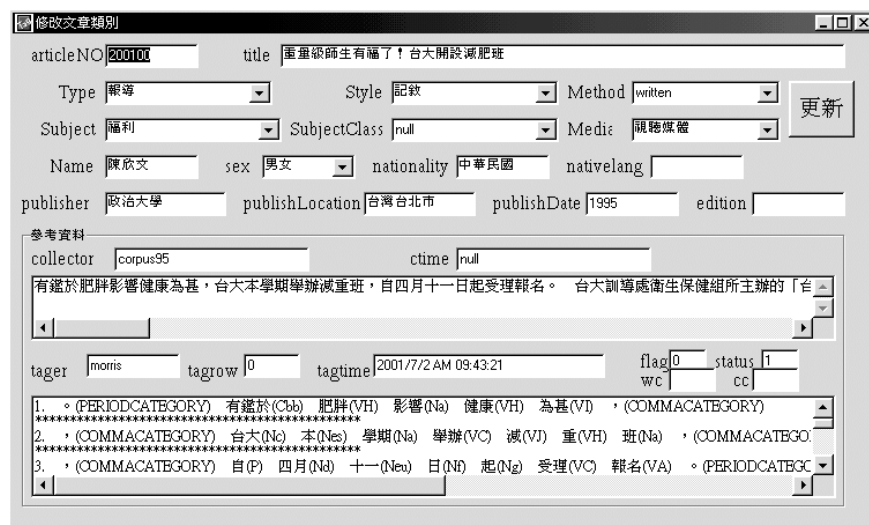
圖八：未知詞編輯介面

(3) 人工檢驗介面

請參照 2.4 節中的說明。

(4) 語料類別修改介面

當使用者發現文本的類別說明不太正確或有需要補充的時候，可以透過該介面進行修改。如圖九所示。



圖九：語料類別修改介面

2.5.1 系統輔助功能

本系統中尚有一些其它輔助功能的設計，分別說明如下：

- (a) Help 功能：提供一些參考資料，如分詞標準，新修訂的詞類，與以前做過的疑難雜症的判例等，供使用者參考，目的在讓使用者於詞類標記時有個統一的標準。
- (b) 提醒儲存功能：當使用者換句編輯或離開系統時，若所編修的文本有易動，系統會提醒使用者要做儲存的動作。
- (c) 復原修改的功能：目前提供單次復原修改。

3. 系統成效

前述各個系統均已開發完成，並實際運作近半年的時間。在語料庫的維護與管理上，由於採用文本為單位的資料庫，因此管理維護上更簡易也更有系統；語料蒐集的部分，過去完全由人工來標示文本的格式、特徵，平均一篇文本約花五分鐘左右的時間，現在由於許多網址的 HTML 格式是固定的格式，可由語料蒐集介面自動標示文本的格式、特徵，人工只做必要的修改，故大幅縮短所需時間，平均在二十秒之內可校對完一篇原始文本的格式特徵。語料的斷詞標記部分，過去由於未知詞的存在，斷詞標記的效果大打折扣，使得人工校正時通常要花大量的時間將未知詞的部分做合分詞的修正，平均一篇六百字的文本需時四十分鐘。現在經過未知詞擷取模組的前處理，使得斷詞標記的表現有著顯著的改善，特別是有許多文本重複著大量的未知詞，例如新聞文本。斷詞標記的改善使得之後的人工校正更加的容易且快速，平均一篇六百字的文本縮短其校正時間為十五分鐘。

4. 結論

本論文提出一套構建及管理語料庫的系統設計，可以大幅縮短構建時程以及減少所花費的大量人力資源，經過實際運作近半年的時間，這樣的構建方式在語料蒐集、語料整理、人工校對等方面在成效上跟過去相比均有相當的改善。

本系統的後續研究方向主要有三：(1) 語料蒐集的格式判斷：目前我們只針對特定的網址來源，對固定的 HTML 格式抽取出文本的特徵，並無法自動的分析判斷所有網址的 HTML 格式。(2) 未知詞擷取模組的持續改善：未知詞擷取模組針對新聞類目前已達九成左右的正確率，然而對於其他文類或是長度過長的文本，由於未知詞在這些文本的統計訊息較不明顯，擷取出來的效果仍有持續改善的空間。(3) 將系統應用到古文或其他少數方言的語料庫構建。

參考文獻

- Church, K. W. and R. L. Mercer, "Introduction to the Special Issue on Computational Linguistics Using Large Corpora," *Computational Linguistics*, Vol.19, No.1, 1993, pp.1-24.
- Chen, Keh-jiann, Shing-huan Liu, Li-ping Chang and Yeh-Hao Chin, "A Practical Tagger for Chinese Corpora," *Proceedings of ROCLING VII*, 1994, pp.111-126.
- Hsu, Hui-li and Chu-Ren Huang, "Design Criteria for a Balanced Modern Chinese Corpus," *Proceedings of ICCPOL'95*, Hawaii.
- Smadja, Frank, "Retrieving Collocations from Text: Xtract," *Computational Linguistics*, Vol.19, No. 1, 1993, pp. 143-177.
- Smadja, Frank, McKeown, K.r. and Hatzivasiloglou, V. "Translating Collocations for Bilingual Lexicons: A Statistical Approach," *Computational Linguistics*, Vol. 22, No.1,1996

- Chang, Jing Shin, "Automatic Lexicon Acquisition and Precision-Recall Maximization for Untagged Text Corpora," National Tsing-Hua University, P.h.d. thesis, 1997
- Wu, M. W. and Su, K. Y. "Corpus-based Automatic Compound Extraction with Mutual Information and Relative Frequency Count," Proceedings of ROCLING VI, Nantou, Taiwan, ROC, Sep. 1993 pp.207-216
- Sproat, Richard and Shin, Chilin "A Statistical Method For Finding Word Boundaries In Chinese Text," Computer Processing of Chinese & Oriental Language, Vol. 4, No. 4, March 1990.
- Yeh, Ching-Long and Lee, His-Jian, "Rule-Based Word Identification for Mandarin Chinese Sentences – A Unification Approach," Computer Processing of Chinese & Oriental Languages, Vol. 5, No.2, March 1991.
- Lin, M.Y., Chang, T. H. and Su, K. Y., "A preliminary study on unknown word problem in Chinese word segmentation," Proceedings of 1993 R.O.C. Computational Linguistics Conference, Taiwan, 1993, pp.119-137.
- Chen, Keh Jiann, Bai, Ming Hong, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," Proceedings of ROCLING X, Taipei, Taiwan, ROC, 1997, pp.159-174.
- Chen, Keh Jiann and Liu Shing Huan, "Word Identification for Mandarin Chinese Sentences," Proceedings of COLING-92, vol. I, 1992, pp. 101-107.
- Chen, C.J., M.H. Bai, & K.J. Chen, 1997, "Category Guessing for Chinese Unknown Words," Proceedings of 4th Natural Language Processing Pacific Rim Symposium(NLPRS'97) pp.35-40.
- 詞庫小組, "中文詞類分析 (三版)," CKIP Technical Report no.93-05.
- 詞庫小組, "中央研究院平衡語料庫的內容與說明 (修訂版)," CKIP Technical Report no.95-02/98-04.

Design, Compilation and Processing of CUCall: A Set of Cantonese Spoken Language Corpora Collected Over Telephone Networks

W.K. LO, P.C. CHING, Tan LEE and Helen MENG

The Chinese University of Hong Kong

wklo@ee.cuhk.edu.hk, pcching@ee.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk, hmmeng@se.cuhk.edu.hk

Abstract

The design and compilation of the CUCall telephone speech corpora is described in this paper. Speech database is an indispensable resource for research and development of state-of-the-art spoken language technology. These speech recognition systems rely greatly on a huge amount of well-designed and appropriately processed speech data for parameters training. On the other hand, as telephony applications are becoming more demanding and complicated, natural language interface is gaining more popularity than the traditional touch tone operation. Therefore, large telephone speech databases are required for such system building. Separate speech corpora are needed for telephone systems since there exist significant differences due to the channel difference. In this paper, we will describe the design and processing of a set of spoken language corpora for Cantonese that are collected over fixed line as well as mobile telephone networks. The corpora are intended as a versatile set of training data for general purpose application systems that adopt a statistical approach to spoken language processing. The designed set of corpora will be made up of over 1000 speaker calls.

1 Introduction

Speech data collected over telephone network is an essential resource for telephone based spoken language systems. The increasing penetration of remote system or service access over telephone networks has created a great driving force for collecting a huge amount of telephone speech data from a large speaker population and for different languages. Since the current state-of-the-art speech recognition techniques are statistically based, the availability of annotated data is particularly important. In general, the greater the amount and coverage of the data, the better the speech applications developed. In order to build a spoken language system over telephone network, the speech data has to be collected over telephone network and properly transcribed. The goal of this work¹ is to collect and compile a set of general purpose Cantonese telephone speech data from a large group of people of both genders. With the availability of this set of corpora, the rapid growth in the spoken language applications over telephone networks for the Cantonese speaking community is made possible.

Over the past decades, many telephone based spoken language systems have been developed with great success. They all take advantage of the existence of several spoken language corpora compiled in recent years. Examples include the Jupiter from MIT [25], HMIHY² from AT&T [17] and the European Union projects such as ACCeSS [27] and ARISE [28] etc. Nowadays, there are quite a large number of companies that make use of simple automatic telephone service systems to reduce the cost of employing human operators. Many of them have upgraded or wish to upgrade their touch-tone based system to speech enabled versions. It is obvious that continuous efforts are needed to enhance these services via speech technologies as much as possible.

For building telephone speech recognition systems, there has long been a great demand on Cantonese telephone speech data. This work is an initial effort to collect a set of Cantonese spoken language corpora over telephone network. It is targeted to provide some versatile data for public use. It aims to enrich the infrastructure for spoken language technology by providing the speech community with well-designed corpora in Cantonese. The compiled database will enable the integration of Cantonese speech technology to many of the existing telephone based interactive systems.

¹<http://dsp.ee.cuhk.edu.hk/speech/cucall.html>

²How may I help you? is the service offered by AT&T.

1-1 Background

There has been much effort in spoken language corpora development over the past decades. These include the TIMIT [8], Resource Management [15], Wall Street Journal [14], Air Travel Information Service [16] etc. from the United States. In Europe, there are the EUROM1 [21] and SpeechDat [3] etc. They contain microphone data and telephone data as well. From the early adaptation of microphone corpora to network versions like NTIMIT [5] and the collection of real-world telephone data such as MACROPHONE [1], CALLHOME [30], SpeechDat [3], POLYPHONE [29] etc., there is an abundant amount of data available for the western languages. The availability of these telephone corpora has successfully helped drive the research and development of telephone-based speech technologies of these languages.

For Asian languages, there has been limited investment spent on corpora development. Much effort came from Japan, for example those reported in [6, 7, 13]. For Chinese language, speech database collection has only started relatively recently. More widely used databases include microphone speech corpora such as the USTC95 [19], HKU96 [2, 24], HKU99 [4], CMSC [22] and others [23]; and telephone speech corpora such as MAT-160 and MAT-2000 [18, 20, 31]. These telephone data become valuable resources for many voice-activated telephony applications development.

Among the many Chinese dialects, Cantonese is one of the most popular Chinese dialects used in the southern China. Development of spoken language corpora has just started within the past decade [9, 10, 11]. It began with some small-scale corpus collection for specific projects. There is great shortage in Cantonese speech corpora to drive the growth and advancement of Cantonese speech technologies.

In 1997, the development of CUCorpora³ [9, 11, 12] was initiated at the Chinese University of Hong Kong. CUCorpora is the first large-scale Cantonese spoken language corpora that are made available for public access. It is designed to cover both phonetically based content and common task oriented and application-specific content. The present work on telephone speech data compilation is a momentous extension of this effort. The vast variation of operator network protocols in Hong Kong⁴ yet enrich the content of the

³<http://dsp.ee.cuhk.edu.hk/speech>

⁴Hong Kong has a large number of mobile network operators offering different kinds of network services using different protocols. This includes the GSM900, GSM1800, TDMA, CDMA.

corpora. Since the speakers will have to call our server to activate the data collection process, the resultant corpora are thus code named CUCall. The availability of the invaluable CUCall will undoubtedly nourish the booming technologies to a greater extent.

1-2 Paper organization

The paper is organized as follows. The design of the corpora materials will be described in detail first in Section 2. The design selection of the major parts of data will be elaborated. After that, actual collection process is presented. From the recording system setup down to the collection process, every detail of the process will be given. In Section 4, the post-processing of the captured data will be explained. The validation, transcription as well as the organization procedures are described. We will then provide some initial analysis on the designed corpora materials. Finally, conclusions are made in Section 6.

2 Corpora Design and Organization

The design of the CUCall has been based on our previous experience with CUCorpora. The concepts behind stay the same. Like CUCorpora, CUCall comprises of linguistically oriented and application-specific data. In CUCall, we take a step forward to include spontaneous conversations and short paragraphs data. These will altogether make up two major parts in the corpora:

1. Phonetically oriented continuous speech data that focus on:
 - (a) coverage through carefully designed corpora materials; and
 - (b) different speaking styles from short paragraphs and free form spontaneous conversation style.
2. Application-oriented short phrases and digit strings.

Figure 2 shows an overview of the organization of the CUCall telephone spoken language corpora.

2-1 Phonetically-oriented data

2-1-1 Phonetic coverage oriented

The phonetically oriented data in the CUCall is based on the design of the CUCorpora with some variations. This part of the data set is made up from sentences and short paragraphs. The materials for the sentences are based on the test and training materials of the CUSENT corpus in CUCorpora and the short paragraphs are excerpted from local newspapers.

Sentences The sentences are chosen to be phonetically rich in the sense that they constitute complete coverage of bi-phone class context. The selection of sentences was detailed in [9, 11]. It was implemented as a semi-automatic process where human intervention is included to decide on the readability of the automatically selected sentences.

Short paragraphs The short paragraphs attempt to emphasize more on the variations of the speaking behaviour and characteristics. For short paragraphs, the selection is solely based on the readability of the paragraphs without taking into consideration of the phonetic content. It aims to enrich the sentence data as well as provide data that bears very different speaking style. Table 1 shows the amount of data for each of these types.

Table 1: The number of reading materials for each type of the phonetically oriented data.

material	number
sentences	5719
short paragraphs	90

2-1-2 Speaking style oriented

For collecting speech data of different speaking styles, the design of CUCall included specifically short paragraphs and conversation parts.

Short paragraphs While the short paragraphs can enrich the phonetic coverage as mentioned in 2-1-1, the data collected in this part is believed to be very different from that of the stand alone sentences. There are many different speaking phenomena being

exaggerated when people reading a section of long text materials. These include correction, hesitation, breathing, long pause etc. Therefore, these recorded materials can also serve the purpose of representing another kind of speaking style in addition to enriching the phonetic content of the sentence corpus.

Spontaneous conversation In the CUCall corpora, a new type of speech data to be collected is the spontaneous conversation type of utterances. These data are collected with the aim to obtain the characteristics of various speakers when prompted to speak in an unprepared manner. There are expected delay, hesitation, correction and skipped words etc. In addition, there are also many colloquials, pronunciations and agrammatical sentences that will not be found in normal read speech. These will provide us with invaluable data for the study of the variation of speaking characteristics under different situations.

The design of “prompts” for this part of data collection has been carefully planned. It is implemented as a single round dialogue between the speaker and the system. Since the speakers are free to answer anything to the prompts, the phonetic content is uncontrollable. The major consideration here is to ensure that there is a high proportion of speakers capable of responding to the prompts. Due to the lengthy nature of the recording process, some speakers are expected to skip these prompts intentionally while some may be too enthusiastic to give very long answers. Several points are considered during the design of prompts:

1. The prompts must be simple enough that “spontaneous” response is possible. Calculation, memory recall or questions requiring accuracy are not suitable.
2. The prompts must have different answers from different speakers so as to increase the variations of the collected data. It would be even better if the same speaker will give different answers at different time.
3. The responses to the prompts may be either long or short.
4. Both for legal purpose and encouraging speakers to answer, the content of the answer must be irrelevant to privacy of the speakers.

Based on the considerations mentioned above, we have carefully designed six prompts. These prompts are carried out at the end of each of the collection sessions. It is done

this way because by that time, the speaker will be more familiar with the recording process. This will then reduce the probability of making mistake since unprepared types of responses are usually “error” prone. In Figure 1, the six prompts are listed for reference (in English translation, because of the colloquial nature of the Cantonese prompt, not all words are writable in characters).

Figure 1: The Cantonese prompts (with English translation) for spontaneous spoken response collection.

-
-
1. 請簡單介紹一吓你而家身處嘅環境，例如，你而家係乜嘢地方，身邊有乜嘢人，有乜嘢事物等等。
 2. 請講一吓你就讀過嘅學校，例如中學、小學，或者各樣嘅進修課程。
 3. 請問，你而家係咪使用緊手提電話？
 4. 請講一吓你居住嘅地區，同埋屋村名稱或者街道名稱。
 5. 除咗廣東話之外，請問你仲會講乜嘢語言？
 6. 請講一吓你最經常乘搭嘅交通工具，同埋所前往嘅目的地。
-
-

1. Would you please describe the environment of your recording, such as where are you, anybody nearby and anything happening?
 2. Which schools have you been studying at? Such as primary and secondary school. Did you study other short courses of any kinds?
 3. Are you using a mobile phone? (*this is intentional for a short yes-no answer*)
 4. In which district of the city do you live? And what is the name of the estate or street?
 5. Besides Cantonese, what other languages do you speak?
 6. What kinds of transportation do you take the most frequently and where do you go?
-
-

2-2 Application-specific data

The CUCall corpora also contain digit strings as well as application-specific short phrases in some specific domains. The design of the digit corpus is similar to that of the CUD-IGIT [9] corpus. In CUCall, the reading materials include all of the single digits together with some random generated long digit strings. This makes up a small-scale digit string

corpus collected over telephone network from a large number of speakers.

The short phrase materials are designed with reference to CUCorpora. Phrases are chosen from various reading materials including names of listed companies and their abbreviations, name of foreign currencies, district names and major housing estates in Hong Kong together with the navigation commands adopted from the CUCMD [9, 11] corpus. These phrases cover the financial domain, navigation commands, as well as major local places. They could be used when building command based speech applications for the related domains. Table 2 lists the amount of corresponding type of phrases.

Table 2: The amount of different types of phrases for the application-specific data.

material	amount
name of places (districts & housing estates)	228
listed companies	1085
foreign currencies	37
navigation commands	90
Total	1440

3 Data Collection Process

The data collection is facilitated by using an automatic call centre type telephone server system. The overall set-up is shown in Figure 3. This server system allows the speakers to call in and then read the provided materials. It is also equipped with the usual navigating features with a touch-tone telephone system.

3-1 Telephone Server

The telephone server is a cluster of computers with one file server and two computer telephony servers (see Figure Figure 4). The file server has a large 64 GB harddisk and is directly connected to the two telephony servers over a 100 Mbps isolated ethernet.⁵ The

⁵This is intentionally set up to improve the security and robustness of the systems. The cluster of computer connected in their own network could eliminate the interference of possible network traffic from other irrelevant processes.

computer telephony servers are equipped with a Dialogic D/41-ESC four port telephony cards for telephone network connection.

There are eight ports available on the Dialogic D/41-ESC card, but only two ports are used. This is sufficient for our current scale of speech collection. Also, additional ports may be used as backup during system maintenance or occasional system breakdown. Furthermore, we can also even out the potential analog channel discrepancies among the different ports by intermittently changing the answering ports over the course of data collection.

3-2 Collection Process

The actual collection process was implemented in several steps:

1. Preparation of the reading materials;
2. Distribution of the reading materials;
3. Accepted speakers call to the telephone server.
4. Return of filled questionnaires from speakers.

Preparation of reading materials The reading materials are mixtures of phrases and sentences described in Section 2. Each part is randomly shuffled and printed out on paper. Every 10 to 30 successful calls will give a complete set for that part of the corpora. In order to differentiate against different gender and different kinds of telephone networks, the reading materials are prepared and distributed in four parallel streams: male mobile, male fixed-line, female mobile and female fixed-line. At the end of each of the prompt sheets, there is a short questionnaire to enable the collection of information about the speaker's age group, telephone network operator (for mobile phone) or type of telephone (whether they are using extension line or direct line).

Distribution of prompt sheet The prepared prompt sheets are distributed through recruited agents. They pass the reading materials to candidate speakers. After recording, the speakers then return the prompt sheet with questionnaire duly completed to the agent and then the agent pass them back to us for processing. The adoption of an agent based distribution network allows an efficient collection process while we could indirectly control the speaker community by choosing appropriate agents.

Speakers call The speakers will make call to our telephone server at any time they so wish. The server would answer the calls whenever it is idle. The speakers are then requested to jot down a generated serial number for bookkeeping purpose. After that, our server program will prompt the speaker by the item numbers on the prompt sheet and then wait for the speakers' speech data with an automatic silence detector. After the speakers have read the prompted item (or time-out if the speakers do not say anything), the data is immediately stored on to the server's hard disk. This prompting process repeats until the last item is finished. The server then reminds the speakers to fill out the questionnaire and hang up subsequently.

Questionnaire return After the agents have collected the prompt sheet, the serial number and questionnaire results are entered into our database for bookkeeping and analysis purposes. Up to this point the collection process is completed and the data are kept for later post-processing.

4 Post-Processing of Data

The most important part of a spoken language corpora development process is the post-processing of the collected speech data. The collected data need to be accurately annotated with necessary labels and organized properly for easy distribution and usage. Based on our previous experience from developing the CUCorpora [9], we have carefully designed the post-processing procedure for the telephone speech data. Figure 5 illustrates the general flow of the post-processing procedures.

Validation of the calls Among the large number of calls received, there is a small percentage of useless data. It may be due to the reason that the speakers give up reading after a short while, the recording environment is too noisy that the silence detector failed totally, or even the system broke down. Based on the serial numbers, we validate all of the calls by checking if there is reasonable amount of data being recorded. If the call is finished properly, the information of the speaker provided on the questionnaire is entered into our speaker database anonymously.

Phonemic transcription of the validated data A major effort in spoken language corpora development is annotation. This is the most important and labour intensive process. In our case, all of the validated data will be transferred using cassette tapes to our contracted professional transcribers. They will listen to the recording tapes and provide Cantonese phonemic transcriptions to all data or mark them as noise wherever applicable. Those successfully transcribed data will then be accompanied by the corresponding phonemic transcription when distributed.

Partitioning and distribution of the collected data The transcribed data will then be partitioned according to the different parts (e.g. digit strings, short phrases, sentences, spontaneous conversation etc.). The partitioned data will be organized into different directories according to different speakers. The phonemic transcription will also be provided in the form of LSHK⁶ transcription symbols. These organized directories of speech data and transcription will be printed on to compact disk for distribution.

5 Data Analysis

In this section, some statistical information of the designed corpora reading materials will be presented. Although there are many expected discrepancies from the actual data that are collected, these statistics can still give an overview of the characteristics of the designed corpora. The discrepancies between the designed materials and the recorded data are mainly due to the reason that there are many speakers who read colloquial and 'lazy' pronunciations, mis-read of materials (e.g. insertion, deletion and substitution of words), and mis-use of the recording systems (e.g. start reading before the recording actually started, stop reading before all of the materials are read, etc.). These could only be analyzed after all data have been transcribed. Detailed statistical analysis of the actual collected data will be released after the information has been prepared.

Table 3 shows the basic information for different parts of the corpora. From this table, it can be observed that out of the 1600 common tonal syllables in Cantonese, the sentence materials have covered over 85% of the syllables. In the short paragraphs corpus, even though the tonal syllable coverage is not as high as that of the sentence recording, we are

⁶Linguistic Society of Hong Kong.

Table 3: Statistical information of the reading materials for the phonetically oriented and application-specific parts of the corpora.

Part	# per speaker	# tonal syl.	# base syl.	syllable count
Phonetically oriented corpora				
sentences	50 (out of 5719)	2251	1030	4 to 31
short paragraphs	3 (out of 90)	768	418	23 to 120
Application-specific corpora				
1-digit string	10	N.A.	N.A.	N.A.
7-digit string	5	N.A.	N.A.	N.A.
8-digit string	5	N.A.	N.A.	N.A.
16-digit string	5	N.A.	N.A.	N.A.
phrases	48 (out of 1440)	562	344	2 to 8

expected to obtain speech data in the form of sentences of length ranging from 23 to 120 characters. These could give us a number of important and unique characteristics in long utterances.

Figure 6 gives another way to look at the properties of the designed reading materials in the sentences and paragraphs parts of the corpora. These are the frequency-of-frequency (FOF) scattered plots for the base and tonal syllables in these parts of the corpus. The FOF plots show the distributions of the occurrences of the syllables. From these figures, it is observed that the content of the corpora is reasonably distributed. While there are some frequently occurred syllables and also some rarely occurred syllables, the majority of the syllable occurrences lie in the middle range. This could then enable us to obtain a normal distribution for the syllables in these parts of the corpora.

For the application-specific corpora, information shown in Table 3 can give us an idea of what is being collected for the database. We have some randomly generated digit strings of various lengths. They should cover most of the common applications where digit strings are needed to be recognized. These may include getting identity card number, telephone number, credit card numbers etc. The 7-digit, 8-digit, 16-digit strings together with the single digits are targeted for these applications. However, since digit strings are so general that continuous digit string data can definitely be applied to other areas of applications.

The other application-specific data collected in this corpora are phrases of various kinds (see Section 2). The phrases from the various different domains are mixed and shuffled for each of the speakers so as to increase the variation in the collection data. From Table 3, it may be found that the acoustic coverage of the phrase part is not as good as that of sentences and paragraphs. Since these data are designed for use in the designated domains, phonetic coverage is not the major concern during corpus design. Nevertheless, the base syllable coverage for these phrase is not far deviated from the complete Cantonese syllable inventory.

Regarding the amount of data in the corpora, a rough estimation has been made. Up to the time of writing, there have been over 1,000 successful calls received. These calls give a total of around 200 hours of data covering all sorts of acoustic events (speech, silence, noise, background etc). Among this volume of recording, the sentence, speaking style, short phrases and digit parts roughly contains 84, 40, 28 and 59 hours of recording. We are currently post-processing these data and they will be made available for public release in the near future when the data are processed.

6 Conclusions

In this paper, the design and data collection process for a telephone spoken language corpora is presented. Details about the post-processing and preliminary analysis of the data are given. Based on the previous experience in microphone speech data collection, this work is extended to collect telephone speech data so as to provide sufficient materials for the building of statistical spoken language systems. The corpora are again divided into two parts: phonetically oriented data and application-specific data. In this work, we have further extend our previous design to include also short paragraph for encompassing speaking characteristics when people reading long materials. Furthermore, we have also included some free-form open questions or prompts for obtaining speaking characteristics in spontaneous speech. Spontaneous speech presents new challenges to speech recognition and the collected data is a valuable resource for investigating possible solutions.

7 Acknowledgements

This project is developed with the support from the Innovation and Technology Fund (AF/96/99). We are grateful to industrial sponsors: Group Sense Limited and SmarTone Mobile Communication Limited. We would also like to thank the Hong Kong Blind Union for helping us transcribes the telephone speech data.

References

- [1] J. Bernstein, K. Taussig, and J. Godfrey, "MACROPHONE, an American English telephone speech corpus for the polyphone project," *Proceedings of 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 81-84, 1994.
- [2] C. Chan, "Design considerations of a Putonghua database for speech recognition," *Proceedings of the Conference on Phonetics of the Languages in China*, pp. 13-16, Hong Kong, 1998.
- [3] H. Hoge, H.S. Tropic, R. Winski, H. van den Heuvel, R. Haeb-Umbach, and K. Choukri, "European speech databases for telephone applications," *Proceedings of 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1771-1774, 1997.
- [4] Q. Huo, and B. Ma, "Training material considerations for task-independent sub-word modeling: design and other possibilities," *Proceedings of 1999 Oriental CO-COSDA Workshop*, pp. 85-88, 1999.
- [5] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," *Proceedings of 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 109-112, 1990.
- [6] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, pp. 357-363, Elsevier Science, 1990.
- [7] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, and T. Watanabe, "Construction of a large-scale Japanese speech database and its management sys-

- tem,” *Proceedings of 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 560-563, 1989.
- [8] L. Lamel, R. Kassel, and S. Seneff, “Speech database development: design and analysis of the acoustic-phonetic corpus,” *Proceedings of DARPA Speech Recognition Workshop*, pp. 100-109, 1986.
- [9] Tan Lee, W.K. Lo, P.C. Ching, and Helen Meng, “Spoken language resources for Cantonese speech processing,” *to appear in Speech Communication*, Elsevier Science, 2001.
- [10] W.K. Lo, K.F. Chow, Tan Lee, and P.C. Ching, “Cantonese databases developed at CUHK for speech processing,” *Proceedings of the Conference on Phonetics of the Languages in China*, pp. 77-80, Hong Kong, 1998.
- [11] W.K. Lo, Tan Lee, and P.C. Ching, “Development of Cantonese spoken language corpora for speech applications,” *Proceedings of the First International Symposium on Chinese Spoken Language Processing*, pp. 102-107, Singapore, 1998.
- [12] W.K. Lo, Helen Meng, and P.C. Ching, “Sub-syllabic acoustic modeling across Chinese dialects,” *Proceedings of the Second International Symposium on Chinese Spoken Language Processing*, pp. 97-100, Beijing, 2000.
- [13] K. Ohtsuki, T. Matsuoka, T. Mori, K. Yoshida, Y. Taguchi, S. Furui, and K. Shirai, “Japanese large-vocabulary continuous speech recognition using a newspaper corpus and broadcast news,” *Speech Communication*, vol. 28, pp. 155-166, Elsevier Science, 1999.
- [14] D. Paul, and J. Baker, “The design of the Wall Street Journal based CSR corpus,” *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*, Morgan Kaufmann, 1992.
- [15] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett, “The DARPA 1000-word resource management database for continuous speech recognition,” *Proceedings of 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 651-654, 1988.
- [16] P. Price, “Evaluation of spoken language systems: The ATIS domain,” *Proceedings of the Third DARPA Speech and Natural Language Workshop*, Morgan Kaufmann, 1990.

- [17] G. Riccardi, A.L. Gorin, A. Ljolje, and M. Riley, "A spoken language system for automated call routing," *Proceedings of 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1143-1146, 1997.
- [18] C.Y. Tseng, "A phonetically oriented speech database for Mandarin Chinese," *Proceedings of 1995 International Congress of Phonetics Sciences*, vol. 3, pp. 326-329, 1995.
- [19] R. Wang, D. Xia, J. Ni, and B. Liu, "USTC95-A Putonghua corpus," *Proceedings of the Fourth International Conference on Spoken Language Processing*, vol. 3, pp. 1894-1897, 1996.
- [20] H.C. Wang, "Speech research infra-structure in Taiwan," *Proceedings of 1999 Oriental COCODA Workshop*, pp. 53-56, 1999.
- [21] R. Winski, and A. Fourcin, "A common European approach to assessment, corpora and standards," in *Advanced Speech Applications: European Research on Speech Technology*, K. Varghese, S. Pflieger, and J.P. Lefvre Eds., pp. 25-79, Springer-Verlag, 1994.
- [22] Y. Wu, "Chili Mandarin speech corpus," *Newsletter of ISCSLP98 Special Interest Group: Linguistic Database and Tools*, pp. 1-3, 1998.
- [23] J. Zhang, "Notes on speech corpora of standard Chinese in China," *Newsletter of ISCSLP98 Special Interest Group: Linguistic Database and Tools*, pp. 4-5, 1998.
- [24] Y.Q. Zu, W.X. Li, M.C. Ho, and C. Chan, "HKU96-A Putonghua corpus (CDROM version)," *HKU96 corpus*, Department of Computer Science, University of Hong Kong, Hong Kong, 1996.
- [25] V. Zue, S. Seneff, J.R. Glass, J. Polifroni, C. Pao, T.J. Hazen, and L. Hetherington, "JUPITER: a telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, vol. 8, is. 1, pp. 86-96, 2000.
- [26] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, pp. 351-356, Elsevier Science, 1990.
- [27] <http://www.wcl.ee.upatras.gr/access/access.htm> Automatic Call Center Through Speech Understanding System.

- [28] <http://www.compuleer.nl/arise.htm> Automatic Railway Information Systems for Europe.
- [29] <http://www.icp.grenet.fr/ELRA/home.html>, European Language Resources Association.
- [30] <http://www ldc.upenn.edu>, Linguistic Data Consortium.
- [31] http://rocling.iis.sinica.edu.tw/ROCLING/MAT/index_cf.htm, Mandarin Across Taiwan corpus.

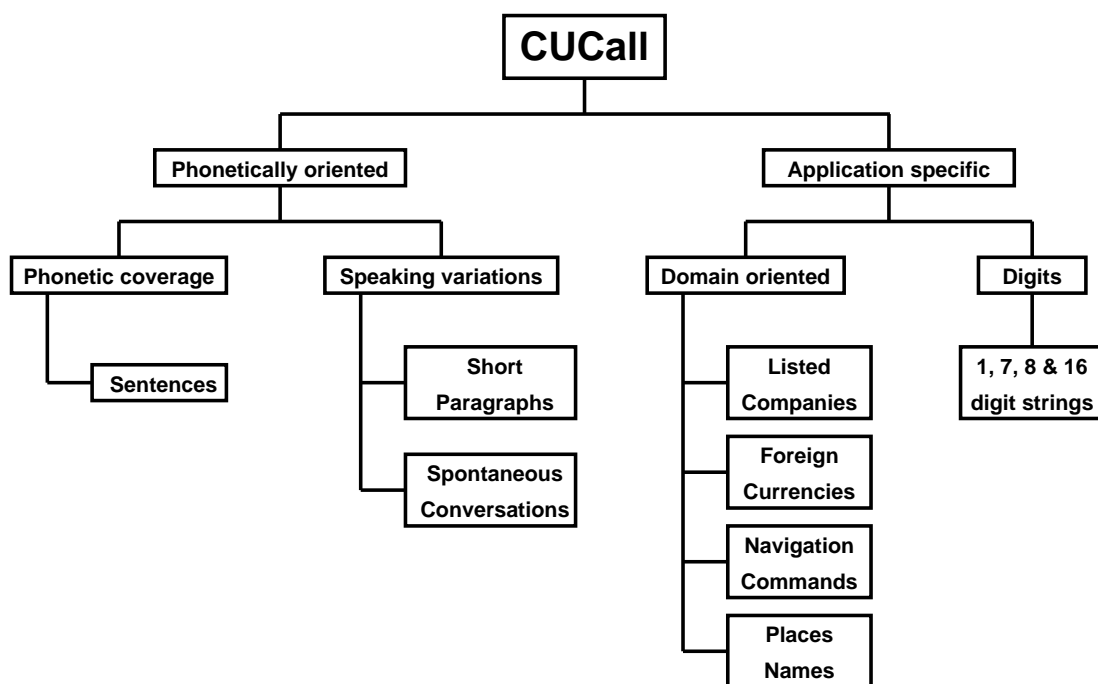


Figure 2: This is an overview of the organization of the CUCall telephone spoken language corpora for Cantonese.

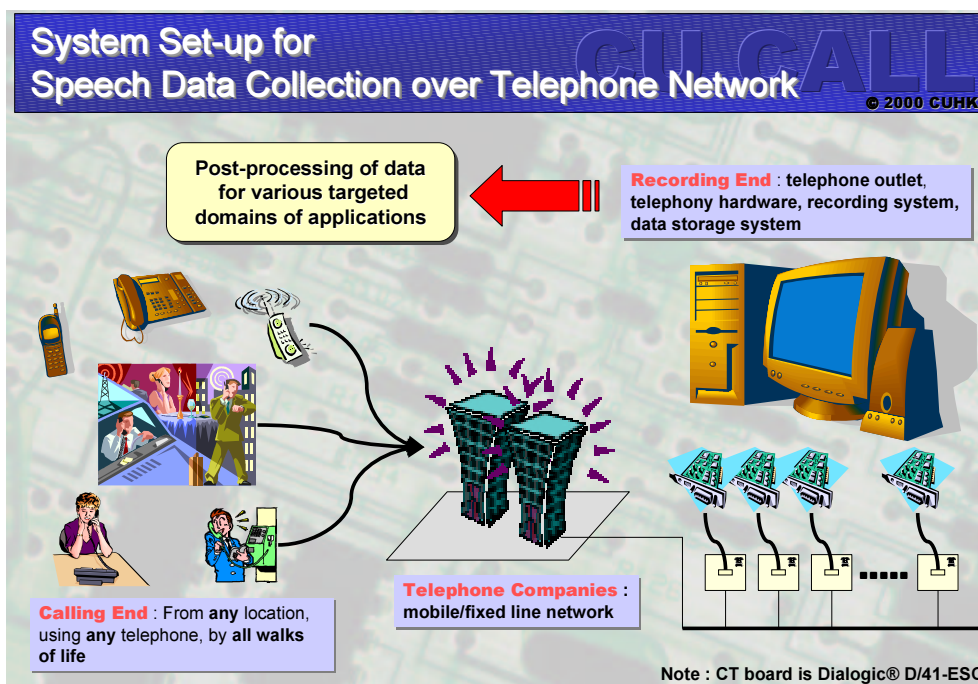


Figure 3: The data collection process for the CUCall corpora over the telephone networks.

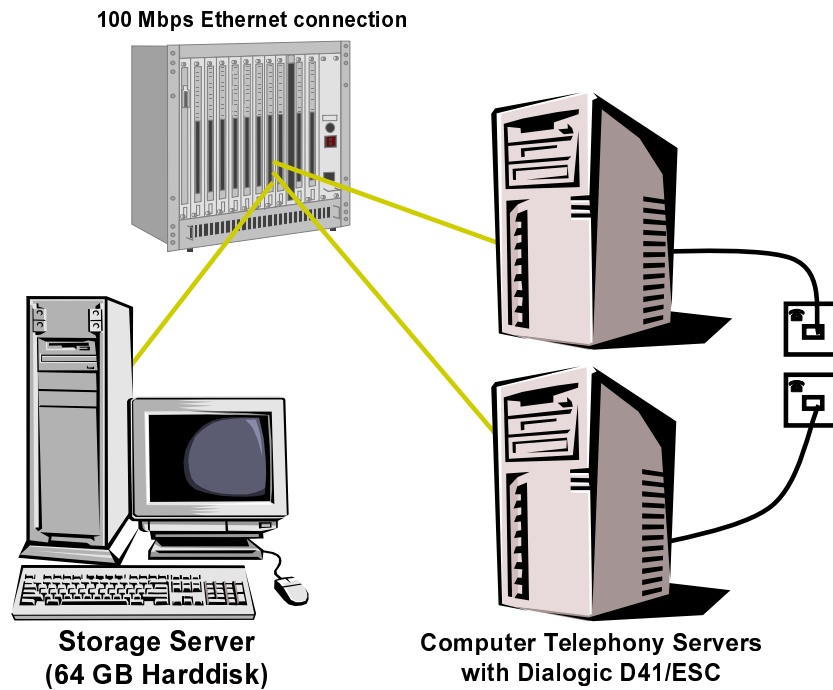


Figure 4: The telephone server setup for corpora data collection.

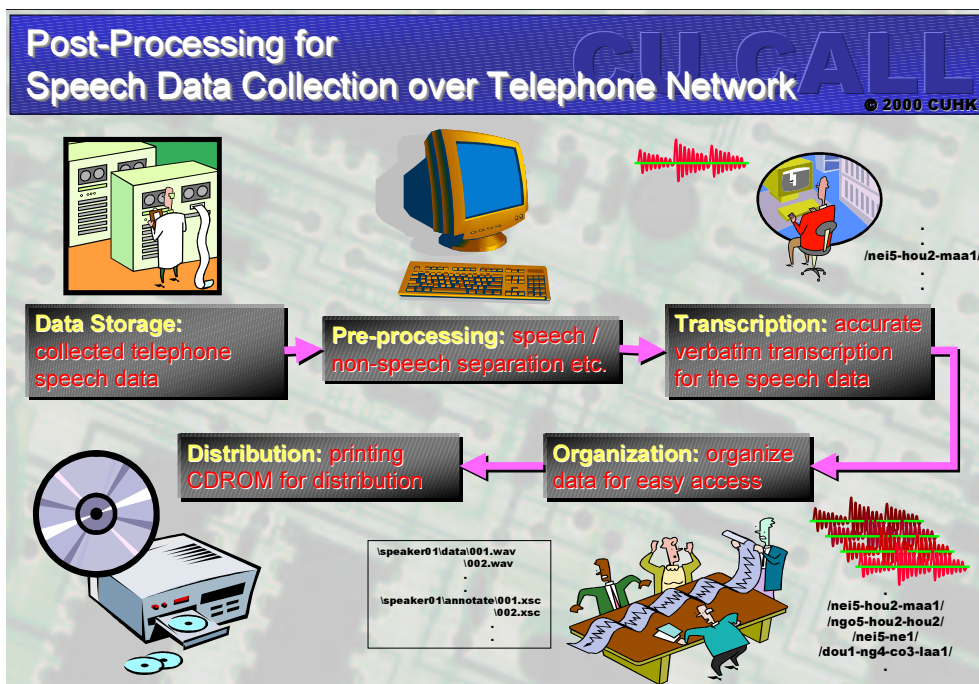
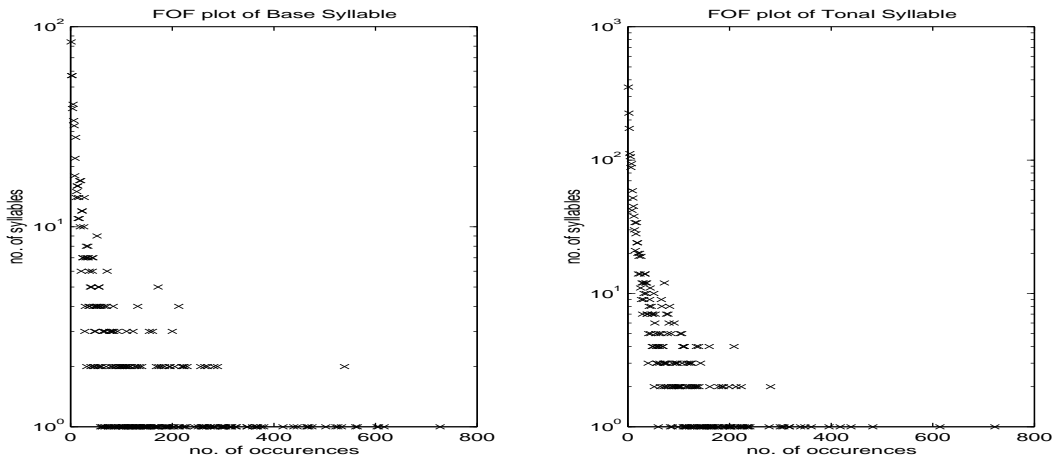
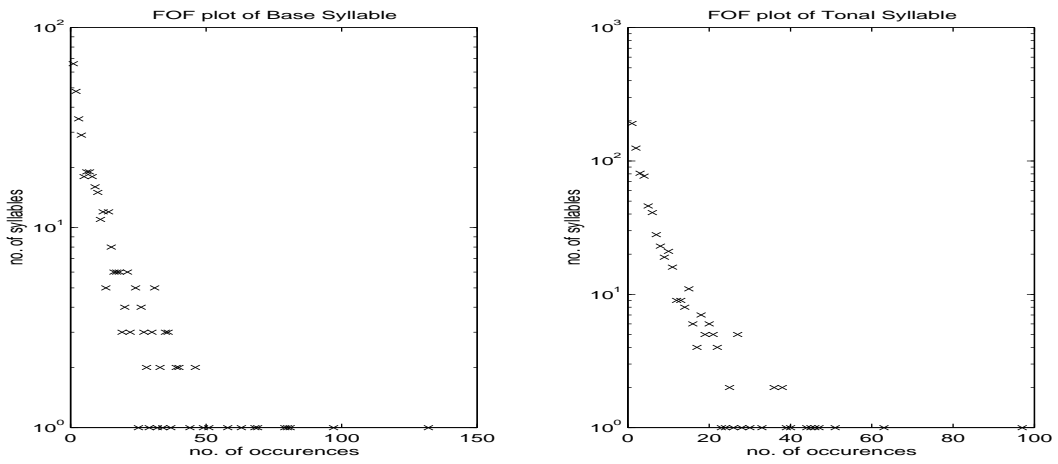


Figure 5: Data post-processing for the CUCall corpora.



(a)



(b)

Figure 6: Scatter plots showing the frequency-of-frequency statistics for syllables in (a) the sentence and (b) the paragraph reading materials.

An Empirical Study of Zero Anaphora Resolution in Chinese Based on Centering Model

Ching-Long Yeh and Yi-Jun Chen
Department of Computer Science and Engineering
Tatung University
40 Chungshan N. Rd. 3rd. Section
Taipei 104
R.O.C.
chingyeh@cse.ttu.edu.tw d8806005@cseserv.cse.ttu.edu.tw

Abstract

In this paper, we describe the creation of Chinese zero anaphora resolution rules by performing experiments. The rules were constructed based on the centering model. In the experiments, we selected several texts as testing examples. We compared the referents of zero anaphors in the testing texts identified by hand with the ones resolved by using an algorithm employing a resolution rule. Three rules were used to carry out the experiment. The results show that the rule considering grammatical role criteria and domain knowledge obtained the best result: 85% of zero anaphors in the test texts were correctly resolved. We investigate problems of miss-resolution of zero anaphors in the test text and propose solution to deal with them.

1. Introduction

In Chinese text, anaphors are frequently eliminated, termed zero anaphor (ZA) hereafter, due to their prominence in discourse [LT81]. For example in (1), the topic of the utterance (1a) is 電子股 ‘Electronics stocks,’ which is eliminated in the second utterance and the topic of utterance (1c), 證券股 ‘Securities stocks,’ is eliminated in the utterance (1d).

- (1) a. 電子股ⁱ 受 美國高科技股 重挫 影響 ,
Electronics stocks were affected by high-tech stocks fallen heavily in America.
- b. ⁱ 今日 持續 下跌 ;
(Electronics stocks) continued falling down today.
- c. 證券股^j 也 相對回應 ,
Securities stocks also had respondence.
- d. ^j 盤 中 陸續 下殺 至 跌停。
(Securities stocks) fell by close one after another on the market.

A simple rule, Rule 1, can be formulated by observing the phenomenon of topic chain in Chinese text. This rule can be used to correctly resolve the referent of the ZA in

(1a), for example.

Rule 1: If a ZA occurs in the topic position of utterance i , then its antecedent is the topic of utterance $i-1$.

In general, zero anaphors in Chinese can occur in any grammatical slot with an antecedent that may occur in any grammatical slot, regardless of their distance [LT79]. Thus Rule 1 is obviously insufficient to account for the resolution of ZAs.

Within the theories of discourse, Centering is a computational model, which has been developed as a methodology for the explanation of the local coherence and its relationship to attentional state at the local level and focuses on pronominal and nominal anaphora [GJW83, GJW95]. It is formalized as a system of constraints and rules, which can, as part of a computational discourse model, act to control inference [JW81]. In the centering model, each utterance in a discourse segment has two structures associated with it, called Forward-Looking and Backward-Looking centers, which correspond approximately to Sidner's potential foci and discourse focus [Sid79]. Forward-Looking Centers, C_f , is a set of discourse entities in an utterance, and Backward-Looking Center, C_b , is a special member of this set, which is the discourse entity that the utterance most centrally concerns. Our analysis is based on this computational model to resolve the intersentential ZAs.

In this paper, we aim at formulating rules for the resolution of zero anaphors in Chinese. We start with a rule, Rule 2, formulated by employing the centering model.

Rule 2: For each utterance U_i in a discourse segment U_1, \dots, U_m : If $C_b(U_i)$ is realized by a zero anaphor in U_{i+1} then the $C_b(U_{i+1})$ must be realized by $C_b(U_i)$.

We performed an experiment by using an algorithm employing this rule to see how the ZAs in news text are resolved. The initial result showed that about half of the ZAs could not be correctly resolved. Consequently we considered adding other constraints, such as grammatical role criteria and semantic knowledge, to enhance the rule and get better results. We repeated the experiment and the result showed that about 85% can be correctly resolved by using the new rule. The remaining 15% errors of the ZAs resolution occur because of the lack of sufficient semantic knowledge and the character of locality of centering model. We further investigate these situations and propose an approach to solve the problem.

In the next section we describe the nature of zero anaphora in Chinese. In Section 3, we describe the centering model, and we illustrate the result of the

empirical study we made by observing the industry news we collected in Section 4. The discussion and implementation are in Section 5 and 6, respectively, and finally conclusions and future works are made.

2. Zero Anaphora in Chinese

In Chinese, anaphors can be classified as zero, pronominal and nominal forms, as exemplified in (2) by i , 他 and 那個人, respectively [Chen87]¹. Zero anaphors are generally noun phrases that are understood from the context and do not need to be specified.

- (2) a. 張三^{*i*} 驚慌的往外跑，
Zhangsan frightened and ran outside.
- b. ^{*i*} 撞到 一個人^{*j*}，
(He) bumped into a person.
- c. 他^{*i*} 看清了 那人^{*j*} 的長相，
He saw clearly that person's appearance.
- d. ^{*i*} 認出 那人^{*j*} 是誰。
(He) recognized who that man is.

According to [LT81], zero anaphors can be classified as intrasentential or intersentential. Intrasentential zero anaphora occur mainly in topic-prominent constructions, namely, sentences having a topic but not a subject such as the in (3). In this sentence, the noun phrase, 房子 (house), is the topic while the subject is not present. In sentences of this sort, subjects, in general, refer to general classes or unspecified noun phrases. In English, *you*, *they* (or more formally *one* is used in this function. This kind of zero anaphor occurs specifically in topic-prominent constructions; they have nothing to do with entities in previous sentences in discourse.

- (3) 房子 蓋好了
The house, (someone) has finished building it.

In the intersentential case, antecedent and anaphors are located in different sentences. Depending upon the distance between the sentences containing antecedent

¹ We use a f_a^b to denote a zero anaphor, where the subscript a is the index of the zero anaphor itself and the superscript b is the index of the referent. A single without any script represents an intrasentential zero anaphor. Also note that a superscript attached to an NP is used to represent the index of the referent.

and anaphor, it can further be divided into two types: immediate and long distance. The former is where the sentence containing the antecedent is immediately followed by the one containing the anaphor, such as f_1^j in (4b) and f_1^k in (4d). For the long distance type, the sentence containing the antecedent and anaphors, on the other hand, are not in immediately succeeding order, such as f_1^i in (4e).

- (4) a. 螃蟹ⁱ 有 四對 步足^j
A crab has four pairs of feet.
- b. f_1^j 俗稱 「腿兒」
(They) are commonly called "tuier."
- c. 由於 每條 「腿兒」 的 關節^k 只能 向下 彎曲
Since every "tuier"'s joint can only bend downwards,
- d. f_1^k 不能 向 前後 彎曲
(it) can't bend backward or forwards.
- e. f_1^i 爬行 時
(When) (it) crawls,
- f. f_2^i 必須 先用 一邊 步足 的 指尖 抓地
(it) must use the tips of feet on one side to grasp the ground.
- g. f_3^i 再用 另一邊 的 步足 直伸 起來
(It) then uses the feet on the other side to move upwards.
- h. f_4^i 把 身體 推 過去
(It) pushes (its) body towards one side.

Since Chinese has no inflection, conjugation, or case markers, the pronominal system is relatively simple, as shown in Table 1 [LT81]. A third-person pronoun can be used to replace an intersentential zero anaphor, except for first- and second-person pronouns, without changing the meaning of the sentence. Though the resulting meaning of each sentence is unchanged, the whole discourse becomes less coherent.

Table 1: Pronominal system in Chinese

Number	Person	Pronoun
singular	first	我
singular	second	你, 妳
singular	third	他, 她, 它
plural	first	我們
plural	second	你們, 妳們
plural	third	他們, 她們, 它們

3. Centering Model

Centering has its computational foundations established by Grosz and Sidner [Gro77, Sid79] and were further developed by Groze, Joshi and Weinstein [GJW83, GJW95]. Within the framework of the centering model, each utterance U in a discourse segment has two structures associated with it, called forward-looking centers, $C_f(U)$, and backward-looking center, $C_b(U)$. The forward-looking centers of U_n , $C_f(U_n)$, depend only on the expressions that constitute that utterance. They are not constrained by features of any previous utterance in the discourse segment (DS), and the elements of $C_f(U_n)$ are partially ordered to reflect relative prominence in U_n . The more highly ranked an element of $C_f(U_n)$, the more likely it is to be $C_b(U_{n+1})$. The highest ranked element of $C_f(U_n)$ that is realized² in U_{n+1} is the $C_b(U_{n+1})$.

The set of forward-looking centers, C_f , is ranked according to discourse salience. The highest ranked member of the set of forward-looking centers is referred to as the preferred center, C_p .³ The preferred center of the utterance U_n represents a prediction about the C_b of the following utterance U_{n+1} and is the most preferred antecedent of an anaphoric or elliptical expression in U_{n+1} . Hence, the most important single construct of the centering model is the ordering of the list of forward-looking centers [WIC94, SH96].

3.1 Constraints and rules

In addition to the structures for centers, C_b , and C_f , the theory of centering specifies a set of constraints and rules [WIC94, GJW95].

Constraints

For each utterance U_i in a discourse segment U_1, \dots, U_m :

1. U_i has exactly one C_b .
2. Every element of $C_f(U_i)$ must be realized in U_i .
3. Ranking of elements in $C_f(U_i)$ guides determination of $C_b(U_{i+1})$.
4. The choice of $C_b(U_i)$ is from $C_f(U_{i-1})$, and can not be from $C_f(U_{i-2})$ or other prior sets of C_f .

² An utterance U , realizes c if c is an element of the situation described by U , or c is the semantics interpretation of some subpart of U .

³ The notion of preferred center corresponds to Sider's notion of expected focus [Sid83]

Backward-looking centers, C_b s, are often omitted or pronominalized and discourses that continue centering the same entity are more coherent than those that shift from one center to another. This means that some transitions are preferred over others. These observations are encapsulated in two rules [WIC90, WIC94, GJW95]:

Rules

For each utterance U_i in a discourse segment U_1, \dots, U_m :

- I. I. If any element of $C_f(U_i)$ is realized by a pronoun in U_{i+1} then the $C_b(U_{i+1})$ must be realized by a pronoun also.
- II. Sequences of continuation are preferred over sequence of retaining; and sequences of retaining are to be preferred over sequences of shifting.

Rule I represents one function of pronominal reference: the use of a pronoun to realize the C_b signals the hearer that the speaker is continuing to talk about the same thing. Psychological research and cross-linguistic research have validated that the C_b is preferentially realized by a pronoun in English and by equivalent forms (i.e. zero anaphora) in other languages [GJW95].

Rule II reflect the intuition that continuation of the center and the use of retentions when possible to produce smooth transitions to a new center provide a basis for local coherence. The transition states are further described in the next section.

3.2 Transition states

The typology of transitions from U_{i-1} to U_i is based on two factors: whether the $C_b(U_i)$ is the same as $C_b(U_{i-1})$, and whether this discourse entity, $C_b(U_i)$, is the same as the $C_p(U_i)$:

1. $C_b(U_i) = C_b(U_{i-1})$, or $C_b(U_{i-1})$ is undefined.
2. $C_b(U_i) = C_p(U_i)$

If both (1) and (2) hold then a pair continuations across U_n and across U_{n+1} . If (1) holds but (2) does not then the utterances are in a retaining transition, which corresponds to a situation where the speaker is intending to shift onto a new entity in the next utterance. If (1) does not hold then the utterances are in one of the shifting transition states depending on whether or not (2) holds. The definition of transition states is summarized in Table 2 [WIC94].

Table 2: Transition states

	$C_b(U_i) = C_b(U_{i-1})$ or $C_b(U_{i-1})$ is undefined	$C_b(U_i) \neq C_b(U_{i-1})$
$C_b(U_i) = C_p(U_i)$	CONTINUE	SMOOTH-SHIFT
$C_b(U_i) \neq C_p(U_i)$	RETAIN	ROUGH-SHIFT

For illustration purpose, consider the example (1) in Section 1; in the Table 3, the centering structures contain C_b , C_f and C_p where the set of C_f are partially ordered to reflect relative prominence in each utterance. The first two transition states of (1a) and (1b) are CONTINUE corresponding to the two factors, “ $C_b(U_i) = C_b(U_{i-1})$, or $C_b(U_{i-1})$ is undefined” and “ $C_b(U_i) = C_p(U_i)$, or $C_b(U_i)$ is undefined.” In (1c), the transition state is RETAIN because of “ $C_b(U_{1c}) \neq C_p(U_{1c})$.”. SMOOTH-SHIFT is the last transition state of example (1) while “ $C_b(U_{1d}) = C_p(U_{1d})$ ” and “ $C_b(U_{1d}) \neq C_b(U_{1c})$ ” hold.

Table 3: Centering structures and transition states for example (1)

(1a)	C_b : undefined	CONTINUE
	C_f : [電子股, 美國高科技股]	
	C_p : 電子股	
(1b)	C_b : 電子股	CONTINUE
	C_f : [ZA (電子股)]	
	C_p : ZA (電子股)	
(1c)	C_b : 電子股	RETAIN
	C_f : [證券股]	
	C_p : 證券股	
(1d)	C_b : 證券股	SMOOTH-SHIFT
	C_f : [ZA (證券股), 盤]	
	C_p : ZA (證券股)	

4. Experiment and Result

This paper is concerned with resolving the problem of zero anaphora in Chinese using the centering model. In this section, we first describe the methodology of zero anaphora resolution we adopted based on centering. Second, we explain how to apply our rules and represent the results of applying the different rules to the test texts.

4.1 Experiment for zero anaphora resolution

The task of zero and nominal anaphora resolution is performed after the semantic interpretation phase that converts the syntactic structure of a sentence into a semantic representation form such as the logic form [JA94]. After semantic interpretation, an anaphor becomes a parameter in a logic form. For example, the logic form of the (5b) is 新鮮(). The task of anaphora resolution is to find out the referent of the omission in the logic forms.

- (5) a. 張三 買了 一顆 蘋果^{*i*}
 Zhangsan bought an apple.
- b. ^{*i*} 很 新鮮
 (It) is very fresh.

Recall that the centering model, an utterance, U_i , is associated with a set of forward-looking centers, C_f , with each element an entity in U_i . The highest ranked element in the set, C_p , becomes the prediction of backward-looking centers, C_b , of the following utterance, which is zeroed if it does not violate syntactic constraints, such as the object of a prepositional phrase [LT81]. Therefore to apply the centering model for zero anaphora resolution, the essential task is to rank the elements in the set. The task of ranking elements is determined according to certain rules, for example Rule 2 described previously in Section 1. In this paper, our goal is to develop effective rules to obtain better result.

We performed an experiment to examine the effectiveness of using a rule for the resolution of zero anaphors. In the experiment, we selected a number of industry news as the test texts. Table 4 summarizes the total news, paragraphs, utterances, zero anaphors and words in the test texts.

Table 4: Summary of test texts

	Paragraphs	Utterances	Words	Zero Anaphors
1	4	36	199	25
2	3	26	229	9
3	4	31	213	13
4	4	29	213	15
5	3	27	208	11
6	4	35	282	15
7	3	28	234	14
8	3	27	289	12
Total	28	239	1867	115

In the experiment, we first of all identify by hand the referent of each zero

anaphor occurring in the texts. Then we compute the referents of zero anaphors identified by using an algorithm employing a resolution rule. The computed result is then compared with the one by hand to see the correction rate of the resolution rule. The correction rate of a resolution rule is defined as below.

Correction rate: Assume that m ZAs occur in n utterances. The correction rate of a resolution rule is the number of referents of ZAs resolved by an algorithm employing the resolution rule that are identical to the ones identified by hand.

The experiment is performed repeatedly by replacing new rules and it is stopped until promising result is obtained. The initial result of using Rule 2 shows that only 55% of the ZAs are correctly resolved, which is obviously not effective enough. The errors occurs in the initial result may be that Rule 2 does contain enough semantic knowledge. In the following, we propose other rules to replace Rule 2 and compare the results.

4.2 Results of using other rules

Grosz *et al.*, in their paper [GJW95], assume that grammatical roles are the major determinant for ranking the forward-looking centers, with the order “*Subject > Object(s) > Others*”. In Chinese, the concept of subject seems to be less significant while the topic in a sentence appears to be crucial in explaining the structure of ordinary sentences in the language [LT81]. By adopting the concept of grammatical roles and topic-prominence in Chinese, we order the grammatical roles in Chinese with topic having the highest priority as shown in Figure 1. The subject and objects occurring in an embedded clause, that is, *Secondary Subject* and *Secondary Objects*, are give lower priority.

*Topic > Main Subject > Direct Object >
Secondary Subject > Secondary Objects*

Figure 1: Grammatical role criteria

By adding the grammatical role criteria to Rule 2, we obtain a new rule, Rule 3:

Rule 3: For each utterance U_i in a discourse segment U_1, \dots, U_m : If $C_b(U_i)$ is realized by a ZA in U_{i+1} and no other noun phrase having higher priority of grammatical role criteria than the ZA then the $C_b(U_{i+1})$ must be realized by $C_b(U_i)$.

Rule 3 is used to verify if the order of the elements in grammatical role criteria we assumed is helpful to raise the correction rate of zero anaphora resolution. We further developed another rule, Rule 4, by considering the domain knowledge corresponding to the test texts.

Rule 4: For each utterance U_i in a discourse segment U_1, \dots, U_m : If $C_b(U_i)$ is realized by both specific nouns in the lexicon and a ZA having the highest priority of grammatical role criteria in U_{i+1} then the $C_b(U_{i+1})$ must be realized by $C_b(U_i)$.

In Rule 4, in addition to grammatical role criteria, we further add the lexical semantic knowledge to the nouns specified in the lexicon. The experiment results of using these rules are investigated as follows.

4.3 Experiment results using three rules

The experiment is performed three times by using Rule 2, 3 and 4, respectively. The first experiment employs the simplest rule, Rule 2, as described in Section 1. Since Rule 2 does not have constraint to order elements in C_f , here we take the surface order of entities from left to right in the utterance. After performing the experiment, the correction rate is 55%, which is obviously not satisfied. In the second experiment, we employed an enhanced rule, Rule 3, and the correction rate is 62%. The result is better but it is still not significant. In the third experiment, we used a further enhanced rule, Rule 4, the correction rate becomes 85%, which is more promising. The results are summarized in Table 5.

Table 5: Summary of experiment results using three rules

	Rule 2	Rule 3	Rule 4
ZAs correctly resolved	63	71	98
Correction Rate	55%	62%	85%

5. Discussions

We have performed experiments on ZA resolution by using three rules with different complexities. The result is promising to some extent; however, there are still 15% of ZAs in the test texts can not be correctly resolved. In the following, we investigate the problems and propose methods to deal with them. One problem is because of insufficient semantic knowledge, namely domain ontology. In the lexical database,

one word may have several word senses and there is a set of synonyms for each sense [MBF+90]. Besides, one word may have hypernyms, hyponyms, coordinate sisters, and other relationship to another word, e.g., 大同 ‘Tatung,’ is a hyponym of 電子股 ‘Electronics stocks,’ and 上市公司類股 ‘listed securities,’ is a hypernym of 電子股 ‘Electronics stocks.’ If the domain ontology contains sufficient lexical and semantic knowledge, it would be helpful to analyze a discourse by understanding the context.

Another problem is with the locality of $C_b(U_i)$ as mentioned in Constraint 4 in Section 3.1. The centering model only accounts for local coherence, that is, the computation of C_b and C_f is confined within successive utterances. Thus the rules we proposed in Section 4 can only deal with immediate zero anaphors. For zero anaphors having their antecedents outside this scope, the rules would be ineffective. Worse yet, the miss-resolution of a long distance zero anaphora would fail to resolve the following zero anaphors, or *error chaining* [SH96]. To solve this problem, we extend the referent set of $C_b(U_i)$ to be the collection of entities occurring in utterances previously in the discourse, that is, $U_1 \dots U_{i-1}$. The referent of a long distance zero anaphor is then determined by examining the elements in the extended referent set. The algorithm for resolving long distance zero anaphors is described as below.

Description: A long distance zero anaphor z is found in the current utterance U_i and then it enters the following procedure. Assume that the extended referent set is E . A temporary set, *temp_set*, is used to record the elements in E that satisfy the semantic constraints of z .

Procedure:

For each element e in E do

If e satisfies the semantic constraints of z , then add e to *temp_set*.

end for;

If there is one element in *temp_set* then return the element as the result;

else return the element in *temp_set* having longest distance from z as the result.

The semantic constraints we used in the above procedure come from the selectional restrictions of the main verb in utterance U_i [JA94]. This kind of restrictions can be used to select the referents of zero anaphors in the topic position. On the one hand, in the sentence which the topic and subject are identical, the zero anaphor in the topic position is restricted by the semantics of the main verb. On the other hand, for sentences with both topic and subject, the topic is frequently moved from the object position of the sentence. Thus zero anaphors of this sort are restricted by the main verb as well. We ignore the selectional restrictions of other syntactic

constructs such as coverb and adjective phrases because the objects or heads of these kinds of phrases can not be zeroed according to syntactic constraints in Chinese [LT81]. Consider, for example, the long distance zero anaphor f_2^i in (6d). Before entering the above procedure, assume that the extended referent set, {市場人士ⁱ, 央行^j, 匯率^k, 台幣^l}, was obtained, where the first two elements satisfy the selectional restrictions of the main verb of (6d), 預期. Here the first one is selected because it is in a more prominent position.

- (6) a. 市場人士ⁱ 擔心 央行^j 會 再度 干預 匯率^k ,
 People on the market worry that Central Bank will intervene the exchange rate again.
- b. f_1^i 不敢 輕易 搶匯 ,
 (They) are afraid to enter the exchange market.
- c. 台幣^l 匯率^k 緩步 走低 ,
 The NTD's exchange rate stops to slowly fall down.
- d. f_2^i 預期 央行^j 不會 輕易 讓 新台幣^j 貶值。
 (They) expect that Central Bank of China will not let NTD be depreciated.

6. Implementation

The goal of this paper is to resolve zero anaphors occurring in discourses based on the centering model. A discourse is a sequence of utterances exhibiting coherence [GJW95]. The resolution of zero anaphors in a discourse is therefore divided into two parts. First, we process each utterance in turn and identify zero anaphors occurring in the utterance. Then we apply a zero anaphor resolution algorithm to resolve the referents of the zero anaphors.

The first part consists of tasks of word segmentation, parsing and semantic interpretation. An input utterance is fragmented into word sequence, and after parsing and semantic interpretation, the semantic form is obtained. Therefore, in this part, the input is a sequence of utterances and the output is the corresponding sequence of semantic forms. Zero anaphors with the information of either immediate or long distance are represented as arguments in the semantic forms. Basically, a zero anaphor is considered an immediate one. But if there are linguistic cues accompanied with the utterance, such as the utterance is the beginning of a new full sentence, and it has initial adverbial connectives, *etc.*, then the zero anaphor is considered a long distance

case. In the second part, the resolution procedure examines each zero anaphor in turn. If an immediate zero anaphor is found, then apply the resolution rules described in Section 4. Otherwise, if it is a long distance zero anaphor, then apply the procedure as described in Section 5. The system architecture is shown in Figure 2.

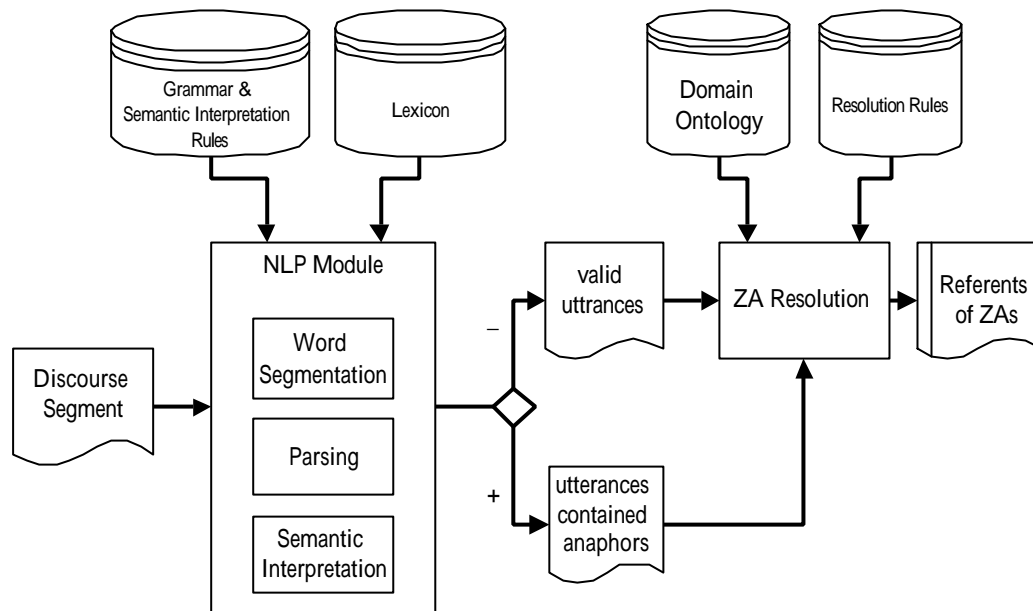


Figure 2: System architecture

In the system the NLP Module carries out in order the work of word segmentation, parsing and semantic interpretation by consulting the lexicon and the syntactic and semantic rules. This module corresponds to the first part described previously in this section. According to the segmentation standard proposed by Academia Sinica [HCC96, HC+97], we built a small lexicon for the test texts, and employ a simple algorithm of word segmentation. The algorithm is according to a strategy that prioritizes the longest word first. The syntactic grammar rules we construct are from the utterances of the test texts and refer to Sinica Corpus and Auto tag program [SC01, CKIP99] and then the parser corresponding to the grammar rules is build as a sentence-level parser in DCG [GM89]. Each utterance within an input discourse segment is converted into a syntactical structure by the parser and the output structure is interpreted to produce the semantic form, which includes the entities in the utterance and is also used to judge whether the utterance contains zero anaphors or not.

ZA resolution by consulting the domain ontology and resolution rules is the

second part of our system. If an input utterance contains a zero anaphor, then apply the resolution rules described in Section 4 to obtain the referent of the zero anaphor. Currently, the ZA Resolution only deals with the immediate zero anaphors. We will extend the algorithm to include the resolution of long distance zero anaphors described in Section 5.

7. Conclusions

In this paper, we performed the experiments on zero anaphora resolution in Chinese based on centering model. In the experiments, 85% of zero anaphors in the test texts were correctly resolved. The remaining zero anaphors were miss-resolved because of lack of sufficient domain knowledge and occurrence of long distance zero anaphors. Since the centering model only focuses on local coherence in discourse, we therefore propose to extend the referent set of a zero anaphor to include all entities occurring previously in the discourse. Though the experiment results are promising to some extent, we found that there are problems that are worth further study. First we need to build domain ontology to get better resolution. Second, the phenomenon of error chaining is inherent in zero anaphors resolution. Thus an effective method is needed to account for this problem. The method we proposed in Section 5 is a step towards solving this problem. Third, the test texts used in this paper were selected from industry news. We will further extend our experiment to include texts from other domains.

References

- [Chen87] P. Chen. 1987. *Hanyu lingxin huizhi de huayu fenxi* (a discourse approach to zero anaphora in chinese) (in chinese). *Zhongguo Yuwen* (Chinese Linguistics), pages 363-378.
- [CKIP99] CKIP. 1999. 中文自動斷詞系統 (Auto tag), Academic Sinica.
- [GJW83] B. J. Grosz A. K. Joshi and S. Weinstein, 1983. Providing a unified account of definite noun phrases in discourse. *Proc. of 21st Annual Meeting of the ACL*
- [GS86] B. J. Grosz and C. L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, No 3 Vol 12, pp. 175-204.
- [GJW95] B. J. Grosz, A. K. Joshi and S. Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics* 21(2), pp. 203-225.
- [GM89] G. Gazdar and C. Mellish. 1989. *Natural Language Processing in PROLOG – An Introduction to Computational Linguistics*, Addison-

- Wesley.
- [Gro77] B. J. Grosz 1977. The representation and use of focus in dialogue understanding *Technical Report 151*, SRI International.
- [HCC96] Chu-Ren Huang, Keh-Jiann Chen and Li-li Chang. 1996. Segmentation Standard for Chinese Natural Language Processing. *Proceedings of the 1996 International Conference on Computational Linguistics (COLING 96)*, pp.1045-1048. Copenhagen, Denmark.
- [HH76] Halliday, M. A. K. and Hasan, R. 1976. *Cohesion in English*. (English Language Series, 9). London, Longman.
- [JA94] James Allen. 1994. *Natural Language Understanding 2nd ed.*, The Benjamin/Cummings Publishing Company, Inc.
- [JW81] Aravind K. Joshi and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure – centering. In *Proc. International Joint Conference on Artificial Intelligence*.
- [Kat97] Boris Katz. 1997. From Sentence Processing to Information Access on the World Wide Web. *1997 AAAI Spring Symposium*.
- [LT79] Charles N. Li and Sandra A. Thompson. 1979. Third-person pronouns and zero-anaphora in Chinese discourse. In T. Givon, editor, *Syntax and Semantics: Discourse and Syntax*, volume 12, pages 311-335. Academic Press.
- [LT81] Charles N. Li and Sandra A. Thompson. 1981. *Chinese Chinese – A Functional Reference Grammar*, University of California Press.
- [MBF+90] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, vol. 3(4), pp. 235--244.
- [SC01] 中央研究院現代漢語平衡語料庫 (Sinica Corpus). 2001. Academic Sinica. <http://www.sinica.edu.tw/>
- [SH96] Strube, M. and U. Hahn. 1996. *Functional Centering. Proc. Of ACL '96*, Santa Cruz, Ca., pp.270-277.
- [Sid79] C. L. Sider. 1979. *Toward a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Ph.D. thesis, MIT.
- [Sid83] C. L. Sider. 1983. Focusing in the comprehension of definite anaphora. *Computational Models of Discourse*, MIT Press.
- [WIC90] Walker, M. A., M. Iida and S. Cote. 1990. Centering in Japanese discourse. *Proc. Of COLING-90*, Appendix, 6pp.
- [WIC94] Walker, M. A., M. Iida and S. Cote, 1994. Japan Discourse and the Process of Centering. *Computational Linguistics*, 20(2): 193-233.
- [Yeh95] Ching-Long Yeh 1995. *Generation of Anaphors in Chinese*, Ph.D. dissertation, University of Edinburgh

中文動詞自動分類研究¹

Automatic Classification of Chinese Unknown Verbs

曾慧馨、劉昭麟、高照明、陳克健

政治大學語言所、政治大學資訊系、台灣大學外文系、中央研究院資訊所

huihsin@iis.sinica.edu.tw; chaolin@cs.nccu.edu.tw; zmgao@ccms.ntu.edu.tw; kchen@iis.sinica.edu.tw

Abstract

We present a new method for automatic classification of Chinese unknown verbs. The method employs the instance-based categorization using the k-nearest neighbor method for the classification. The accuracy of the classifier is about 70.92%.

Keyword: unknown word, lexical similarity

1. 緒論

自然語言處理中重要的步驟是將中文文件斷詞並附加詞類標記；在斷詞標記的過程中會遇到的一個問題為未知詞的存在。現行的斷詞標記系統以辭典為基礎輔以構詞的規則訊息進行斷詞標記，但因為語言的特性之一「無窮盡的創造力」，無法窮舉出所有的詞彙；一本好的辭典也不應該無止盡的擴大所收錄的詞彙，因此如何辨識處理辭典中不存在的詞彙就成了一個重要的課題。

1-1 研究動機與目標

前人對於未知詞的探討重點集中在名詞細目的辨認上，如組織名、人名、地名辨識等(李振昌(1993)，李振昌、李御璽與陳信希(1994)人名辨識等等)。僅有Chen、Bai與Chen(1997)利用前綴(prefix)、後綴(suffix)的訊息處理全部的未知詞，正確率約為76%，而白明宏、陳超然與陳克健(1998)使用Chen、Bai與Chen(1997)所提出的方法，再利用前後文的訊息來補強Chen、Bai與Chen(1997)方法不足之處，將正確率提高至83.83%。在動詞辨識正確結果不高的情況下，本論文將處理重心放在未知動詞的辨識處理上，並且希望將這種處理未知動詞的方法在未來可以轉移處理名詞與形容詞。

動詞不管在任何文法理論中，在剖析句子時都是位於最中心的部分，若動詞為未知詞，勢必將影響句子剖析的正確性。現代漢語的動詞結構繁複，內部規則複雜，若無足夠的語言訊息完全無法判斷其分類，我們認為動詞自動分類研究至

¹ 本論文中程式設計感謝馬偉雲、楊昌樺學長提供意見；兩位評審老師惠賜意見，特此致謝。

今無法提高正確率的主因為動詞繁複的內部結構。

我們的目標為將動詞自動分類到中研院詞庫小組(1993)的詞類架構上，動詞的詞類分類共有 15 類，但並非每一類都具有孳生性。有些類別如功能詞一般，屬於封閉性詞類，封閉性詞類為該分類中的詞彙不會增加，而在中研院詞庫小組的分類中 15 類中有 9 類是具有孳生性的分類；這 9 類分類中的動詞詞彙，會隨著語料庫的增長而增多，我們希望將未知動詞自動分類到這 9 類動詞分類中，這九類為動作不及物動詞(VA)、動作及物動詞(VC)、動作及物動詞+地方賓語(VCL)、動作雙賓動詞(VD)、動作句賓動詞(VE)、分類動詞(VG)、狀態不及物動詞(VH)、狀態使動動詞(VHC)、狀態及物動詞(VJ)。

1-2 研究方法

本論文中未知詞的定義為不存在辭典中的詞彙。陳克健、陳超然(1997)分析未知詞的種類為兩種，第一種為封閉性，這一類型雖然在數量上可能為無數個，但是可用規則語法(Regular Expression)來產生與辨識，如：西元一九九九年(時間)、一千兩百七十二(數字)、二七八八三七九九(電話)等。第二類則為開放性，這一類的未知詞很難用規則語法來表達，複合詞即屬這一類。白明宏、陳超然與陳克健(1998)在分析中研院平衡語料庫後歸納出未知詞主要的分類為略語、專有名詞、衍生詞、複合詞與數字型複合詞。

未知動詞通常為複合詞，由兩個以上的組成成分組合而成，這種組成成分我們稱為詞基(base)²。趙元任(1968)、Li 與 Thompson (1981)與湯廷池(1988)提及漢語的複合詞具有特定的內部句法結構；如：「欺敵」，由「欺」與「敵」這兩個詞基組成，兩個詞基之間的關係為動賓結構。雖然詞基是有限的，但是詞基與詞基的組合數量龐大，且組成成分間的語意關係複雜，因此造成了我們無法將所有的未知動詞收錄進字典中。

在本論文中我們利用相似法來判斷動詞的分類，尋找未知動詞的相似詞，計算未知動詞與相似詞之間的相似度，再將這些相似詞依照詞類分組。從每個詞類當中取出 K 個相似詞出來，將這些相似詞的分數予以平均，得到未知動詞到每個詞類的平均距離，未知動詞的詞類即與其距離最相近的詞類。

²Sproat 與 Shih (1996) 稱內部的處理單位為詞根(root)，Chen、Bai 與 Chen (1997)稱處理的單位為前綴(prefix)與後綴(suffix)。我們則稱處理單位為詞基(base)，並採用 Katamba (1993:45) 對詞基(base)所下定義：“...a base is any unit whatsoever to which affixes of any kind can be added....In other words, all roots are bases. Bases are called stems only in the context of inflectional morphology.” 我們在此處決定使用詞基為我們切割的單位的原因在於詞基的定義較詞根 (root)、詞幹 (stem) 寬鬆。未知動詞被我們斷詞系統切分出來很多單位，我們並不確定這些單位真正的意義，因此我們希望選用一個最寬鬆的定義可以涵蓋所有被斷詞系統所切分的單位。

1-3 語料分析與處理

我們在此介紹未知動詞的特性與可猜測未知動詞詞類的可能因素。首先，討論未知動詞的特性。未知動詞為複合詞，通常由數個具有孳生性的詞基所組成，本身語言具有高透明性。例如，未知動詞「求新」與「講錯」相對於列入辭典中的「忐忑」、「侷促」這一類的詞彙多具有語意透明性，並且可以從其組成成分預測出該詞的語意。

其次，我們認為有兩個因素可預測未知動詞的分類。一、語意。語意相近的詞彙，所屬的詞類應類似。我們將同義詞詞林中的語意類與中研院詞庫小組(1993)詞類作對應，中研院詞庫小組詞類有 45 類。平均來說，同義詞詞林一個語意類僅對應到詞庫小組 1.97 種詞類，即一個語意類中的詞彙共有的詞類數量。因此我們認為語意因素可左右詞彙的詞類。二、結構。結構通常會限定組成的詞類，若結構為“VC+Na”的未知動詞，通常會組成 VA 詞類，因為在這個未知動詞的內部結構中已經出現了一個普通名詞(Na)來滿足前面的動作及物動詞(VC)所要求的論元，在這種情形下通常會形成不及物動詞，因此我們認為結構會影響到動詞的詞類。

在本篇論文中我們利用這些線索尋找與未知動詞相似的詞彙，來預測未知動詞所屬的詞類。

2. 實驗方法

我們利用相似法來判斷動詞的分類，尋找未知動詞的相似詞，計算未知動詞與相似詞之間的相似度，再將這些相似詞依照詞類分組。從每個詞類當中取出 K 個相似詞出來，將這些相似詞的分數予以平均，得到未知動詞到每個詞類的平均距離，未知動詞的詞類即與其距離最相近的詞類。

2-1 相似法 (Instance-Based Categorization)

我們在這節說明如何使用相似法來預測動詞的分類。未知動詞的特性之一為組成成分屬於常用詞且語意明確，例如：試印、講完。這兩個詞彙都無法在辭典中查詢到，但我們卻很清楚的可以從字面上得知這兩個動詞的語意，而且這樣的組合方式是非常具有孳生性的，可以繼續孳生「唱完」、「說完」等等各樣的詞彙。

根據我們對未知動詞語料的觀察，未知動詞的組成雖然有一定的模式，但因為語言的複雜度，無法將所有的規則條列出來。因此我們在這邊使用相似法，將每個訓練語料中的未知動詞都當作是一條規則，當有新的未知動詞出現時，將其與所有的動詞做比較，測量新的未知動詞與訓練語料中的動詞的相似度，新的未知動詞與訓練語料中的動詞越相似時，新的未知動詞越有可能屬於與其相似動詞的詞類。例如：講完與唱完。若「講完」我們訓練語料中的動詞，「唱完」為我

們的未知動詞。未知動詞的第二個組成成分與訓練語料中的例子相同都為「完」，因此我們僅需要得知「講」與「唱」的相似度，若「講」與「唱」分屬的詞類相似度高，則表示「講」與「唱」的結構類似；若「講」與「唱」的語意相似程度高的話，則「唱完」的動詞分類則很可能與「講完」相同。

使用相似法的好處在於相似法所尋找的相似詞，若相似度高的話，不僅可以預測詞類分類，同時也可以預測語意與結構分類。當兩個詞彙相似度高時，表示這兩個詞彙的詞類、語意類與結構必定相似。

我們在本節中首先介紹語意與詞類相似度的測量方法，接下來說明相似詞的選取與未知動詞詞類的預測。

2-2 相似度測量

在本論文中我們使用知網作為語意測量的工具，中央研究院中文句結構樹測量詞類相似度，介紹如下。

一、知網為一雙語(中文、英文)的知識性辭典，由董振東與董強編撰完成收錄約十一萬條詞條，知網系統中包含有中英雙語知識辭典、中文簡體知識辭典、中文繁體知識辭典、概念特徵、動態角色與屬性、詞類表、反義關係表、對義關係表、標示符號與說明、知網管理程序等。我們在本節當中將介紹如何使用知網計算語意相似度與評量方法。

二、中央研究院中文句結構樹資料庫 1.0 中包含了十個檔案，三萬八千七百二十五棵中文結構樹，含有二十三萬九千五百三十二個詞詞彙，每一句結構樹，標示漢語句法與語意訊息，詞類標記與斷詞標記系統四十五個標記不同，結構樹中的標記是由四十五個標記細分而成。在本節中我們利用中研院中文句結構樹測量詞類的相似度。

2-2-1 語意相似度測量

知網約選用了一千七百多個義原來定義中英雙語知識辭典中的每個詞，並且建有描述各個義原之間的關係的分類樹。例如：「讀書」一詞由「從事」、「學」與「教育」三個義原定義而成，知網中並有分類樹表示「從事」、「學」與「教育」三個義原之間的關係。

一般來說，一個詞在知網中可能擁有多個詞條，原因在於詞彙的多義性，因此我們在這邊定義兩個詞 $Word_1, Word_2$ 間的相似度相等於兩個詞各屬的詞條間最大相似度。

$$\text{HowNetSimScore}(Word_1, Word_2) = \max \text{HowNetSimScore}(Word_1 - \text{Entry}_x, Word_2 - \text{Entry}_y)$$

其次，每一個詞條可能由一到八個義原定義而成，如「讀書」一詞由「從

事」、「學」與「教育」三個義原定義而成，在知網標記義原的規則中，在詞條的所有定義義原中，第一個義原一定是主要意義分類，形成概念間的上下位關係(is-a relation)，第二個以後的義原為次要區分與詞彙之間的關係就不確定，依照知網標記決定。計算兩個詞條間相似度時主要義原與整個詞彙之間的關係十分重要，必須與其他的次要義原分開計算。因此

$$\begin{aligned} & \text{HowNetSimScore}(\text{Word}_1 - \text{Entry}_x, \text{Word}_2 - \text{Entry}_y) \\ &= w_1 * \text{PrimaryScore}(\text{Sem}_{x,1} \cap \text{Sem}_{y,1}) \\ &+ w_2 * \text{SecondaryScore}((\text{Sem}_{x,2} \dots \text{Sem}_{x,n}), (\text{Sem}_{y,2} \dots \text{Sem}_{y,m})) \end{aligned}$$

知網中有描述義原與義原之間的階層關係的分類樹，我們在這邊利用描述義原之間關係的分類樹來幫助我們計算義原間的相似度。陳克健、陳超然(1997:270)認為兩個語意類的相似度在於兩個語意類在分類樹交集節點的語意訊息量(Information Content)，將整個詞分類架構看成一個訊息系統，一個語意類 Sem (相當於知網中的義原)的訊息量定義為 Entropy(System)-Entropy(Sem)。我們在這邊使用陳克健、陳超然(1997)計算語意訊息量的方法來計算知網中各義原的訊息量。

知網中兩個義原的相似度為這兩個義原所交集節點的語意訊息量，所得到語意訊息量越高表示這兩個義原越相似，因此第一部份的相似度定義如下：

$$\begin{aligned} & \text{PrimaryScore}(\text{Sem}_{x,1} \cap \text{Sem}_{y,1}) \\ &= \text{InformationContent}(\text{Sem}_{x,1} \cap \text{Sem}_{y,1}) / \text{Entropy}(\text{System}) \\ &= (\text{Entropy}(\text{System}) - \text{Entropy}(\text{Sem}_{x,1} \cap \text{Sem}_{y,1})) / \text{Entropy}(\text{System}) \end{aligned}$$

而第二部份的相似度的定義為：

$$\begin{aligned} & \text{SecondaryScore}((\text{Sem}_{x,2} \dots \text{Sem}_{x,n}), (\text{Sem}_{y,2} \dots \text{Sem}_{y,m})) \\ &= \left(\left(\sum_{i=2}^n \text{Max}_{j=\{1 \dots m\}} ((\text{InformationContent}(\text{Sem}_{x,i} \cap \text{Sem}_{y,j}) / \text{Entropy}(\text{System})) / (n-1)) \right) \right) \\ &= \left(\left(\sum_{i=2}^n \text{Max}_{j=\{1 \dots m\}} ((\text{Entropy}(\text{System}) - \text{Entropy}(\text{Sem}_{x,i} \cap \text{Sem}_{y,j})) / \text{Entropy}(\text{System})) \right) \right) / (n-1) \end{aligned}$$

我們令(n>=m)，也就是第一個詞條的定義的義原多於或等於第二個詞條的義原，從第一個詞條中第二個義原開始，每個義原與第二個詞條中的每個義原計算相似度，第一個詞條中每個義原留下與第二個詞條義原相似分數最高的組合，將第一個詞條中每個義原得到的分數平均，就是我們所定義的第二部份的相似度。

以上兩式中各項皆除以 Entropy(System) 是為維持相似值介於 0,1 之間。

2-2-2 詞類相似度測量

我們將 1.0 版中的句結構樹中歸納出規則，並統計每條規則出現的頻率，如圖 1 可歸納出右邊的三條規則，規則之前的數量表示規則出現的次數。下圖為中研院中文句結構樹的範例：

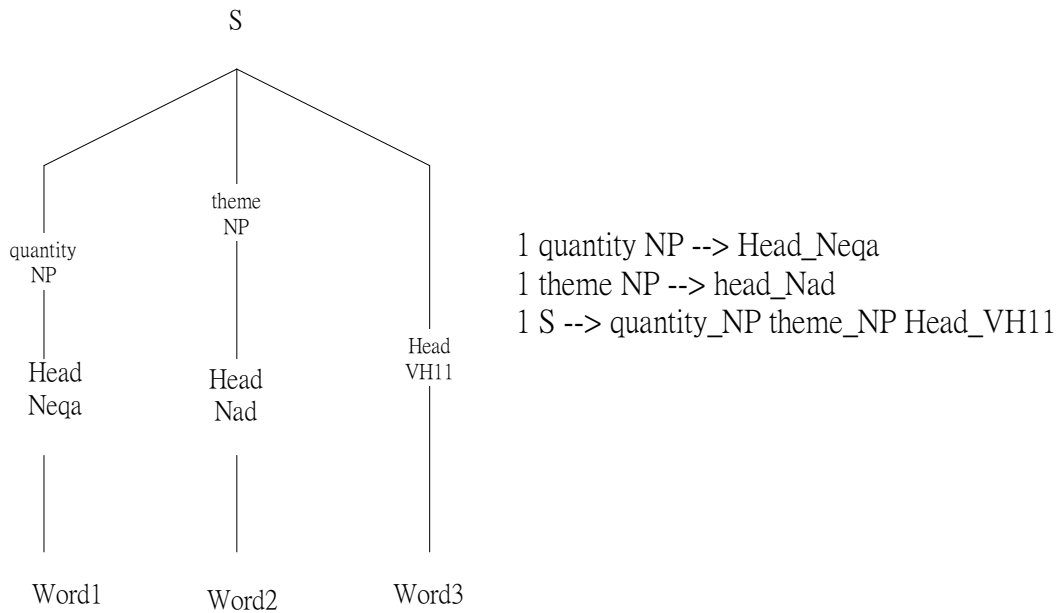


圖 1 中文句結構樹狀圖與歸納規則

每一個詞類的向量由各父節點與兄節點出現的頻率組成，先為放入各父節點的頻率，再依次放入兄節點的頻率，若該個節點沒出現在該詞類中，則放入為 0。定義如下：

$$i = \{VA, VAC, VB, VC, VCL, \dots, Na, Nb, \dots, A, \dots, P, \dots\}$$

$$\overrightarrow{\text{Category}_i} = \langle \text{freq}(\text{parent node}_1), \text{freq}(\text{parent node}_2), \dots, \text{freq}(\text{parent node}_n), \text{freq}(\text{sibling node}_1), \text{freq}(\text{sibling node}_2), \dots, \text{freq}(\text{sibling node}_m) \rangle$$

得到各個詞類的向量後，我們利用下列公式計算詞類與詞類之間的相似程度，所得的分數介於 0~1 之間，1 表示完全相同，0 表示完全不相同。

$$CategoryScore(\overrightarrow{Category_i}, \overrightarrow{Category_j}) = \frac{\overrightarrow{Category_i} \cdot \overrightarrow{Category_j}}{|\overrightarrow{Category_i}| * |\overrightarrow{Category_j}|}$$

我們列出部分 VH 類的動詞與各類動詞的相似度於表格 1。除了 VH 類下的分類 VHC 類外，VH 類動詞與 VI 類相似程度最高，VH 類與 VI 類兩者皆為狀態動詞，他們的差別僅在於可接的論元數量。VI 類為類單賓動詞，基本上也是不及物動詞，但是 VI 類的動詞在語意上可接受一個論元，但該論元的位置不出現在動詞之後，通常使用一個介詞將論元引介出來。而 VH 類與 VA 類的相似程度為次高，VH 類與 VA 類同屬不及物動詞，他們的差別僅在於動作與狀態的區分。

表格 1 詞類相似度(部分)

詞類 1	詞類 2	相似度
VH	VA	0.674
VH	VC	0.611
VH	VD	0.643
VH	VE	0.540
VH	VG	0.591
VH	VH	1.000
VH	VI	0.736
VH	VJ	0.655
VH	VHC	0.852

2-3 相似詞的選取

在使用相似法來預測動詞分類的過程中，三個主要的步驟。一為未知動詞的相似詞的選取，二為測量未知動詞與相似詞的相似度，三為決定未知動詞的詞類。

首先，當一個新的未知動詞出現時，我們並不知道哪些訓練語料的動詞與新的未知動詞較相似，因此理論上我們必須計算每個訓練語料中的動詞與新的未知動詞的相似度，尋找出相似度較高的相似詞作為新的未知動詞預測詞類的依據，計算新的未知動詞與訓練語料中動詞的定義如下：

$$\begin{aligned} \text{If Word} &= \text{wordbase}_1 + \text{wordbase}_2 + \text{wordbase}_3 + \dots + \text{wordbase}_n \\ \text{Sim}(\text{Word}_{\text{unknown}}, \text{Word}_{\text{known}}) & \\ &= \text{weight}_1 * \text{Sim}(\text{wordbase}_{1\text{-unknown}}, \text{wordbase}_{1\text{-known}}) \\ &+ \text{weight}_2 * \text{Sim}(\text{wordbase}_{2\text{-unknown}}, \text{wordbase}_{2\text{-known}}) \\ &+ \dots \\ &+ \text{weight}_n * \text{Sim}(\text{wordbase}_{n\text{-unknown}}, \text{wordbase}_{n\text{-known}}) \end{aligned}$$

若採用這種方法必須計算訓練語料中的每一個詞彙與我們未知動詞的相似度，將會浪費許多不必要的計算時間，因此僅就訓練語料中與新的未知動詞前詞基相同與後詞基相同的相似詞為計算標的。尋找到前詞基相同與後詞基相同的相似詞後，第二步需計算這些選取出來的相似詞中與新的未知動詞詞基相異的部分的相似度。計算兩個詞彙相似度的方法，如下；

$$\begin{aligned} & \text{Sim}(\text{Word}_{\text{unknown}}, \text{Word}_{\text{known}}) \\ &= w_1 * \text{Score}_1 + w_2 * \text{Score}_2 \\ &= w_1 * \text{HowNetSimScore}(\text{Base}_i, \text{Base}_j) \\ &+ w_2 * \text{CategoryScore}(\text{category}(\text{Base}_i), \text{category}(\text{Base}_j)) \end{aligned}$$

$\text{Word}_{\text{known}}$ 為相似詞

Base_i 為未知動詞與相似詞相異的詞基

Base_j 為相似詞與未知動詞相異的詞基

最後一個步驟是決定未知動詞的詞類。我們已有了一群相似詞，同時每個相似詞也有與未知動詞的相似分數。先將這些相似詞依照詞類分組，從每個詞類當中取出 K 個相似詞出來，將這些相似詞的分數予以平均，得到未知動詞到每個詞類的平均距離，未知動詞的詞類即與其距離最相近的詞類。我們將在下一節測試語意相似度中的比重、語意與詞類的比重以及 K 值的大小對正確率的影響。

3. 實驗結果

中研院平衡語料庫第三版五百萬詞內，未知詞經由人工標記為動詞者有 9170 個，佔 18.2%，名詞類有 40455 個，佔 80.3%，非謂形容詞有 761 個，佔 1.5%。本論文的處理範圍為這 9170 個未知動詞，即不存在辭典中的動詞。從中抽取出 1000 個動詞當作測試語料(Final Test Set)，其餘的未知動詞當作訓練語料(Training Set)，測試語料的正確答案為人工標記的詞類。

在相似法中需要討論下列三點。一、調整語意相似度中的主要義原與次要義原間的比重，二、調整語意與詞類兩種相似度的比重，三、調整 K 值的大小，使整個系統的正確率達到最佳狀態。

正確率的定義為：

$$\text{正確率} = \frac{\text{猜測正確的未知動詞}}{(1000 - \text{無法猜測的未知動詞})}$$

3-1 語意比重相似度評量

我們首先要固定兩個變數，語意與詞類的比重與 K 值，才能觀察出相似度比重的變化對正確率的影響。因此先給予 K=1，語意與詞類比重為 1 與 0，依照相似度比重的變化對正確率的影響製成下表。

表格 2 相似度比重與正確率變化表

語意與詞類比重(語意,詞類)	語意相似度比重(w_1, w_2)	正確率
(1,0)	(1,0)	54.04%
(1,0)	(0.9,0.1)	57.58%
(1,0)	(0.8,0.2)	57.70%
(1,0)	(0.7,0.3)	57.58%
(1,0)	(0.6,0.4)	56.97%
(1,0)	(0.5,0.5)	56.85%
(1,0)	(0.4,0.6)	56.23%
(1,0)	(0.3,0.7)	55.87%
(1,0)	(0.2,0.8)	56.11%
(1,0)	(0.1,0.9)	56.11%
(1,0)	(0,1)	56.09%

由上表中可看出主要義原的比重為 0.8 與次要義原的比重為 0.2 時可以得到最高的正確率，因此在本實驗中我們使用 0.8 與 0.2 作為主要義原與次要義原的比重。

3-2 語意與詞類比重評量

我們將相似度比重設定 w_1 為 0.8 與 w_2 為 0.2，K=1，觀察語意與詞類比重的變化對正確率的影響。

表格 3 語意與詞類比重與正確率變化表

語意與詞類比重(語意,詞類)	語意相似度比重(w_1, w_2)	正確率
(1,0)	(0.8,0.2)	57.70%
(0.9,0.1)	(0.8,0.2)	58.41%
(0.8,0.2)	(0.8,0.2)	58.30%
(0.7,0.3)	(0.8,0.2)	58.85%
(0.6,0.4)	(0.8,0.2)	58.85%
(0.5,0.5)	(0.8,0.2)	59.40%
(0.4,0.6)	(0.8,0.2)	59.62%
(0.3,0.7)	(0.8,0.2)	60.40%

(0.2,0.8)	(0.8,0.2)	60.51%
(0.1,0.9)	(0.8,0.2)	60.62%
(0,1)	(0.8,0.2)	48.97%

經由上表的觀察，我們不能放棄詞類分數或語意分數，因為當詞類分數或語意分數比重為 0 時，所得到的正確率皆低。另外，我們發現約有一成的未知動詞無法猜測，原因在於沒有相似的例子或知網中沒有收錄該詞彙，造成沒有猜測結果的情形。

3-3 K 值變化

經由上述的實驗，我們現在將相似度比重設定為 0.8 與 0.2，而語意與詞類的比重設定為 0.9 與 0.1，來觀察 K 值的變化對正確率的影響。

表格 4 K-value 與正確率變化表

語意與詞類比重	語意相似度比重	K-value	正確率
(0.1,0.9)	(0.8,0.2)	1	60.62%
(0.1,0.9)	(0.8,0.2)	2	65.82%
(0.1,0.9)	(0.8,0.2)	3	66.70%
(0.1,0.9)	(0.8,0.2)	4	67.48%
(0.1,0.9)	(0.8,0.2)	5	67.15%
(0.1,0.9)	(0.8,0.2)	6	67.26%
(0.1,0.9)	(0.8,0.2)	7	67.59%
(0.1,0.9)	(0.8,0.2)	8	67.81%
(0.1,0.9)	(0.8,0.2)	9	68.14%
(0.1,0.9)	(0.8,0.2)	10	68.25%
(0.1,0.9)	(0.8,0.2)	20	68.14%
(0.1,0.9)	(0.8,0.2)	30	67.59%
(0.1,0.9)	(0.8,0.2)	40	67.48%
(0.1,0.9)	(0.8,0.2)	50	68.25%

在這邊我們發現，相似例子的多少對整體正確率的變化有影響，因此我們將字典中的動詞放入我們的訓練語料中，使得訓練語料增多，觀察正確率的變化。從表格 5 中我們發現，當訓練語料為未知動詞加上字典中的動詞時，正確率

隨之增長。但在表格 5 與 4 的比較中發現，當 k=1 和 2 時，表格 5 的正確率比表格 4 的正確率低。我們認為可能的解釋原因在於字典中有一小部分的詞彙不具語意透明性，當我們僅尋找少數相似的詞彙時，容易造成誤判，使正確率降低造成的結果。

表格 5 K-value 與正確率變化表

語意與詞類比重	語意相似度比重	K-value	正確率
(0.1,0.9)	(0.8,0.2)	1	58.59%
(0.1,0.9)	(0.8,0.2)	2	65.12%
(0.1,0.9)	(0.8,0.2)	3	67.65%
(0.1,0.9)	(0.8,0.2)	4	67.97%
(0.1,0.9)	(0.8,0.2)	5	67.86%
(0.1,0.9)	(0.8,0.2)	6	68.49%
(0.1,0.9)	(0.8,0.2)	7	68.49%
(0.1,0.9)	(0.8,0.2)	8	68.49%
(0.1,0.9)	(0.8,0.2)	9	68.81%
(0.1,0.9)	(0.8,0.2)	10	68.91%
(0.1,0.9)	(0.8,0.2)	20	70.60%
(0.1,0.9)	(0.8,0.2)	30	70.50%
(0.1,0.9)	(0.8,0.2)	40	70.71%
(0.1,0.9)	(0.8,0.2)	50	70.92%
(0.1,0.9)	(0.8,0.2)	60	70.71%
(0.1,0.9)	(0.8,0.2)	70	70.60%
(0.1,0.9)	(0.8,0.2)	80	70.39%
(0.1,0.9)	(0.8,0.2)	90	70.18%
(0.1,0.9)	(0.8,0.2)	100	70.07%
(0.1,0.9)	(0.8,0.2)	200	68.81%

4. 結果分析

我們在本節分析猜測錯誤的未知動詞、相似法無法處理之未知動詞與我們蒐集語料的問題。

4-1 猜測錯誤之未知動詞分析

在我們實際觀察相似法猜測錯誤的例子當中，參見下表(詳列於在附錄中)，我們條列出來猜測錯誤的例子。從這些例子當中我們觀察幾個現象：一、大部分的猜測錯誤的例子為較罕見的詞語，如：言趣、黏結之類。使用者在使用這些詞語之時，若無語境實在也很難判斷出正確的分類。二、有部分的詞彙已經詞彙化了，如：高挑。該詞彙無法從詞彙的組成成分觀察出該詞彙的意義來。這些動詞的處理方式，不適用於以上所提出的方法。

表格 6 猜測結果

未知動詞	系統猜測詞類	正確詞類
言趣	VH	VA
捐掉	VC	VD
晨運	VC	VA
夷平	VH	VC
黏結	VC	VH
自屬	VC	VG
練唱	VC	VA

4-2 相似法無法處理之未知詞

我們以相似法在處理未知動詞之時，觀察到當訓練語料僅為未知動詞時，約有一成的未知動詞無法辨識。而當訓練語料的數量增大時，不能處理的動詞數量便降低了。

表格 7 無法處理之未知動詞數量變化

訓練語料	不能處理動詞的數量
未知動詞	96
未知動詞+字典	34

我們可以從上表中發現，當我們的訓練語料增加時，的確可以縮減不能處理

的未知動詞。我們在下面分析這三十四個不能處理的未知動詞，並試著尋找解決無法預測詞類的問題。

表格 8 無法預測分類的未知動詞分類

無相似分數	潰腫起來、大挪移、一決勝負、直垂到、激盪、鬧雙胞、大買單、下油鍋、昇進到、起酒疹、大收紅、上山下海、泫然淚下、游手好閑、匯寄到、歸併到、暗藏玄機、唉聲嘆氣、安天樂命、黯然無語、大飽耳福、轉增資
無相似詞彙	蕞爾小邦、遊客如織、蝶躞、潸然淚下、叱吒、洞房花燭、吃飽喝足、上山下海、商調至、松蘿垂掛、喁喁情話、萬民歸心、克己復禮

我們可以從上表中大概歸納出未知動詞不能處理的情形。一、找到相似的詞彙，但無法計算出相似度。

如：「泫然淚下」尋找到兩個相似詞。

泫然欲滴,0,VH

泫然欲泣,0,VH

但是因為「淚下」與「欲滴」無法計算相似度，造成了無法判斷「泫然淚下」的詞類。二、完全無法找到相似詞彙，如：「蕞爾小邦」。不能處理的四字未知動詞大多為 VH 類的成語，因此我們猜測這些不能處理的四字動詞為 VH 類，三字動詞若為 V+N 結構則為 VA 類，字尾為趨向補語的詞彙詞類與補語之前的動詞相同，剩餘的都猜測為 VH 類。

4-3 語料分析

我們討論三個語料問題。一、未知詞的定義與抽取未知詞的方法。二、中研院平衡語料庫中標記的一致性。三、中研院平衡語料庫標記的模糊地帶。

4-3-1 未知詞定義與抽取未知詞的方法

我們在這邊討論未知詞的定義與抽取出來的未知動詞所衍生的一些問題。首先，本文未知詞的定義為不存在於字典中的詞彙，並且假設未知詞應具有語意透明性，即我們可以從字面上得到該詞彙的語意，但是在我們所收集的未知動詞中，有一小部分並不屬於這種類型，例如：中的(一箭中的)、夯築、向邇、離去、過飛、熏繞、絜靜、馱彼等。

不具語意透明性詞彙的最佳的處理方式是將其增入辭典中，根據我們對語料

的觀察與分析，發現這類型詞彙出現的原因大多是作者引用到了非現代漢語的詞彙，或非常用詞與字，我們認為解決這部分詞彙最好的方法就是將這一類型的詞彙全部收錄字典中。

4-3-2 中研院平衡語料庫標記的一致性

在我們觀察訓練語料中，發現有標記不統一的現象，這讓我們很難將這一部份的語料歸納出任何的結論，例如：「verb_i+不了」這種結構，在 verb_i 屬動作動詞的情況下，我們發現有部分的標記人員將「verb_i+不了」這種結構的動詞的分類標記成 verb_i 的分類，即仍屬動作動詞，另外有部分的人則將「verb_i+不了」標記成為一個狀態動詞，論元結構分類不改變。如：「抵擋不了」標記為 VJ 類(狀態單賓動詞)，「阻擋不了」標記為 VC 類(動作單賓動詞)，但「抵擋」與「阻擋」在中研院詞庫小組詞知識辭典中的詞類皆屬 VC 類(動作單賓動詞)。

我們推測這樣的標記方法是部分標記人員認為「不了」會使整個動詞狀態化，但是不會改變整個動詞的論元結構，因此標記人員將這樣的組合給予狀態動詞，而另外一部分的認為加上「不了」後，並不會影響整個動詞的動作與狀態的分類，則給予該 verb_i 原先的分類。

由於標記規則的不統一，我們無法從中歸納出任何規則，這些標記不一致的詞彙影響到我們使用相似詞判斷動詞分類的正確率。我們也認為這種類型的詞彙的確很難去決定分類，但希望有個統一的規則，可以將這種類型的詞彙給予一致性的標記。我們也希望藉由我們從這個角度的觀察與提出討論，爾後進一步修改中研院平衡語料庫中的詞類標記，使得語料庫標記更為一致。

4-3-3 中研院平衡語料庫標記模糊地帶

在我們觀察訓練語料時，發現有許多未知動詞具有兩種以上的分類，我們認為這一類的動詞出現的原因在於本身語意上的模糊，讓標記人員不易判斷該類動詞的所屬標記。下表 9 列出未知動詞具有兩種以上標記的詞彙，出現最多的為兼有動作與狀態的未知動詞，其次為及物性的模糊，如：及物與不及物的模糊與單賓與雙賓的模糊。

表格 9 標記模糊詞彙

動作與狀態模糊	出現次數	詞彙
兼有 VA 與 VH	43	一爭長短、三拖四拉、大放光明、大開眼界、小反彈、不忍卒說、天搖地動、斗轉星移、左等右等、交互為用、共霑法益、同床共枕、合作來合作去、在握、安享天年、行俠仗義、呼朋喝友、居無定所、忽前忽後、背井離鄉、家

		傳戶曉、站得住、笑吟吟、鬥鬧熱、剪徑伏擊、密鑼緊鼓、接應不暇、清火、連戰皆墨、速審速結、喧天動地、惡語相向、發人省思、亂停、滑漏過去、蓄勢待撲、蝕甚復圓、語冠全場、誤蹈法網、趕流行、撫今思昔、講道完、斷去、藏垢納汙
兼有 VC 與 VJ	17	出不了、交織出、伴隨有、抵免、抵抗不了、附設有、看不順眼、要不了、浪費掉、配置有、停放有、牽連到、脫離不了、著有、裝設有、過不了、應證
及物性模糊		
兼有 VA 與 VC	17	大吵、大降、生產出來、存下來、折回來、狂襲過來、延伸出來、拖出去、放飛、流傳下來、洗來洗去、捲來捲去、教改、淡出、聊開、散布開來、著墨、解解饞、蓋下來、聯合起來、隱去
兼有 VH 與 VJ	3	互敬、住不起、嚇慘、憨拙
兼有 VC 與 VD	13	付得起、保留給、致上、展現給、租到、退回給、配起來、推介紹、移轉給、設計給、散播給、轉述給、轉移到

動詞分類的模糊性影響了正確率的高低，有些動詞可為動作也可為狀態，但是這裡所謂的正确答案並沒有把這類動詞所有可能的分類標記出來。在這種的情況下，若他們僅標記該動詞為動作動詞的一類，而我們僅猜測為狀態動詞的一類，使得正確答案與我們預測的分類不同。儘管我們預測了其中一項可能的分類，但仍須計算為猜測錯誤，使得我們的正確率降低。

例如：「站得住」一詞可以當 VA 類與 VH 類的動詞，但若標記者僅將「站得住」標記為 VH 類，而我們的系統卻猜測「站得住」為 VA 類，雖然我們的系統猜測正確，但是因為正確答案僅給予「站得住」一個可能的詞類，沒有標記出來第二個可能的詞類。這類標記模糊的動詞是我們系統無法提高正確率的原因之一。

這一類的詞彙大概的歸納為下列幾點。一、成語性的詞彙多具有 VA 類別與 VH 類別。二、特殊的後綴詞，如：「不了」與「有」。若該詞彙擁有這樣的後綴詞，就可能具有狀態與動作的身份。三、部分加了「趨向補語」的動詞，也可能具有不及物與及物兩種特性。四、加上了「給」為後綴的動詞，也具有單賓與雙賓的特性。

我們將來希望將這一部分的詞彙歸納出規則，若屬於這類型的動詞，我們就可以自動給予一種以上的分類。

5. 結論

我們本文中利用相似法來判斷動詞的分類，尋找未知動詞的相似詞，計算未知動詞與相似詞之間的相似度，再將這些相似詞依照詞類分組。從每個詞類當中取出 K 個相似詞出來，將這些相似詞的分數予以平均，得到未知動詞到每個詞類的平均距離，未知動詞的詞類即與其距離最相近的詞類，正確率為 70.92%。

分析猜測錯誤的未知動詞中，大部分的詞彙為比較罕見的用詞或是已經詞彙化的詞語，我們建議將這一部分的未知詞收錄於字典中。其次，將部分無法預測分類的未知詞收錄辭典中，如成語『蕞爾小邦』。另外，我們也期待當訓練語料增多與知網收錄詞彙增多時，可以處理另外一部份目前無法得到相似分數或無法尋找到相似詞的未知詞。

相似法容易受到語料中錯誤訊息的干擾，因此中研院平衡語料庫中標記的不一致性與部分詞彙本身的模糊性都影響到我們未知動詞自動分類的正確率。我們希望中研院平衡語料庫中標記不一致的語料與標記模糊語料的處理方式能夠得到改善，也期待改善後的結果能夠影響我們動詞分類系統的效能。

6. 參考文獻

中文

中央研究院詞知識庫小組。「技術報告 9305：中文詞類分析」。南港：中央研究院詞知識庫小組，1993。

---。「技術報告 9601：『搜』文解字---中文詞界研究與資訊用分詞標準」。南港：中央研究院詞知識庫小組，1996。

---。「技術報告 9502/9804：中央研究院平衡語料庫的內容與說明」。修訂版。南港：中央研究院詞知識庫小組，1998。

白明弘、陳超然、陳克健。「以語境判定中文未知詞詞類的方法」，《第十一屆計算機語言學會論文集》，1998，頁 47-60。

李振昌。「中文文本專有名詞辨識問題之研究」。台北：台灣大學資訊工程研究所碩士論文，1993。

李振昌、李御璽、陳信希。「中文文本人名辨識問題之研究」，〈第七屆計算機語言會會議論文集〉，1994，頁 203-222。

李坤霖。「網際網路 FAQ 檢索中意圖萃取及語意比對之研究」。台南：成功大學資訊工程研究所碩士論文，2000。

陳克健、洪偉美。「中文裡「動名」述賓結構與「動名」偏正結構的分析」，《第八屆計算機語言學會論文集》，1996，頁 1-29。

陳克健、陳超然。「語料庫為本的中文複合詞構詞律模型研究」，《漢語計量與計算研究》，編輯：鄒嘉彥、黎邦洋、陳偉光、王士元，1997，頁 283-305。

梅家駒、竺一鳴、高蘊琦、殷鴻翔。「同義詞詞林」。香港：商務印書館，1984。

- 湯廷池。《漢語詞法句法論文集》。台北：學生書局，1988。
- 董振東、董強。知網---中文信息結構庫。<<http://www.keenage.com>>，2000。
- 。事件關係與角色轉換庫。<<http://www.keenage.com>>，2000。
- 趙元任。《中國話文法》。丁邦新譯。香港：中文大學，1980。
- 賴育昇、李坤霖、吳宗憲。《網際網路 FAQ 檢索中意圖萃取及語意比對之研究》。
<第十三屆計算機語言學研討會>，2000，頁 135-156。

西文

- Chen, Chao-Jan, Ming-Hung Bai and Keh-Jiann Chen. "Category Guessing for Chinese Unknown Words," Proceedings of the Natural Language Processing Pacific Rim Symposium, 1997, pp. 35-40.
- Chen, Keh-Jiann and Ming-Hong Bai. "Unknown Word Detection for Chinese by a Corpus-based Learning Method," Computational Linguistics and Chinese Language Processing vol3 no. 1, 1998, pp. 27-44.
- . "Knowledge Extraction for Identification of Chinese Organization Names," Proceedings of the second Chinese Language Processing Workshop, 2000, pp. 15-21.
- Li, Charles and Sandra Thompson. "Mandarin Chinese: A Functional Reference Grammar". Berkeley: University of California Press, 1981.
- Resnik, Philip. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, pp. 448-453.
- . "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language," Journal of Artificial Intelligence Research XI, 1998, pp. 95-130.
- Resnik, Philip and Mona Diab. Measuring Verbal Similarity. Technical Report: LAMP-TR-047//UMIACS-TR-2000-40/CS-TR-4149/MDA-9049-6C-1250. University of Maryland, College Park, 2000.
- Sproat Richard and Shilin Shih. "A Corpus-Based Analysis of Mandarin Nominal Root Compound," Journal of East Asian Linguistics 5, 1996, pp. 49-71.

Weischedel, Ralph, Marie Meteer, Richard Schwartz, Lance Ramshaw and Jeff Palmucci. "Coping with Ambiguity and Unknown Words through Probabilistic Model," *Computational Linguistics* 19, 1993, pp. 359-382.

附錄

未知動詞	系統猜測詞類	正確詞類
迎來	VA	VC
言趣	VH	VA
捐掉	VC	VD
晨運	VC	VA
夷平	VH	VC
黏結	VC	VH
自屬	VC	VG
練唱	VC	VA
傾盡	VC	VH
夾處	VC	VJ
車拼	VA	VC
迷向	VC	VA
齊薰	VH	VC
走穩	VH	VA
攏好	VH	VC
耐燃	VC	VH
嗤鼻	VH	VA
申領	VH	VC
正價	VA	VH
堵回	VC	VCL
承標	VA	VC
高挑	VC	VH
漂忽	VJ	VA
大登	VH	VC
抽背	VA	VC
交詮	VE	VC
報來	VH	VC
吹響	VH	VC
中第	VH	VA
悠到	VCL	VC
轉劇	VA	VH
燒昏	VC	VHC
人治	VC	VH
自定	VH	VC

重登	VC	VCL
轉走	VC	VA
折拗	VC	VH
實徵	VC	VH
提味	VH	VA
湊熱	VH	VA
偵錯	VC	VA
曝身	VA	VH
避靜	VH	VA
擁至	VC	VCL
步踵	VCL	VC
自食	VA	VC
藏諸	VC	VCL
孝親	VA	VH
對招	VC	VA
營築	VH	VC
試走	VC	VA
給就	VC	VD
易記	VC	VH
無誨	VC	VH
寶肝	VH	VA
起腳	VH	VA
哭窮	VH	VA
時起	VC	VH
無念	VJ	VH
回傳	VD	VC
啞口	VA	VH
湧退	VC	VA
鼓足	VH	VC
互映	VA	VH
自野	VA	VH
正對	VH	VC
不輟	VH	VA
承續	VH	VC
板結	VC	VH
對中	VC	VJ

獨統	VA	VH
攘外	VH	VA
租住	VC	VCL
教懂	VH	VC
沾黏	VH	VC
抗震	VH	VA
並稱	VE	VG
削竿	VC	VH
運賣	VC	VD
增殖	VC	VA
惡用	VC	VJ
作美	VH	VA
明化	VHC	VH
撤到	VC	VCL
畫方	VH	VA
採到	VCL	VC
怒張	VH	VA
虛矯	VH	VC
為亂	VH	VA
酌進	VH	VC
轉遊	VA	VCL
移葬	VCL	VC
撲上	VC	VA
倒盡	VH	VJ
近身	VH	VA
自云	VA	VE
叮穩	VC	VA
廣作	VG	VC
唸作	VG	VC
雙殺	VC	VA
解壓	VC	VA
加烈	VC	VH
從眾	VA	VH
靜棲	VA	VCL
哭倒	VC	VA
收擔	VC	VA

悅動	VC	VH
跌斷	VC	VHC
互連	VH	VJ
觀照	VA	VC
拒容	VJ	VA
超產	VA	VC
大誇	VH	VC
盲動	VA	VH
整錯	VH	VC
共學	VC	VA
澆到	VC	VCL
搞懂	VH	VE
裝得了	VJ	VC
破記錄	VH	VA
斷不了	VH	VC
碰不得	VH	VC
管起來	VC	VE
定出來	VA	VC
問清楚	VC	VH
遞上來	VC	VA
妨害到	VJ	VC
運下來	VA	VC
大震撼	VC	VJ
侵占去	VA	VC
檢回來	VA	VC
沒法度	VC	VH
往回收	VC	VA
伸進來	VA	VC
哄上去	VA	VC
傾銷至	VC	VCL
抬回來	VA	VC
切進來	VA	VC
拗脾氣	VA	VH
合八字	VH	VA
行的通	VC	VH
巡禮完	VC	VCL

播下去	VA	VC
暗下來	VA	VH
當不了	VC	VG
惹不得	VH	VJ
漲退潮	VH	VA
跑進來	VA	VCL
走著瞧	VC	VA
自宣布	VA	VE
談戀愛	VH	VA
喜愛上	VC	VJ
大辯論	VC	VE
窩藏進	VC	VCL
穿進去	VA	VC
減輕到	VC	VH
溜進去	VA	VCL
搬下去	VA	VC
大販賣	VC	VD
實習用	VC	VH
激增到	VC	VJ
起不了	VH	VJ
減少到	VC	VH
大肚子	VA	VH
縮小到	VC	VH
經得住	VC	VJ
試開工	VC	VH
容不進	VC	VJ
拖出來	VA	VC
傳遞至	VCL	VC
量出來	VA	VC
久旱未雨	VH	VA
學有專攻	VC	VH
驚醒過來	VA	VHC
得意滿面	VA	VH
金盆洗手	VH	VA
易朝換主	VA	VH
照射下來	VH	VA
春光外泄	VC	VH

跌足搥嘆	VE	VH
蹈光養晦	VH	VA
展現出來	VA	VC
天人交戰	VH	VA
提報上來	VA	VC
言者諄諄	VA	VH
睽隔已久	VH	VA
糾結起來	VA	VH
隱善揚惡	VH	VA
生聚教養	VH	VA
增添進來	VA	VC
聽不進去	VC	VA
學無止境	VC	VH
塵灰撲撲	VC	VH
消受得起	VC	VJ
流於言表	VA	VH
大傷腦筋	VA	VH
表露出來	VA	VC
先馳得點	VH	VA
昏迷過去	VA	VH
應付得了	VJ	VC
大顯神威	VH	VA
搜竭枯腸	VH	VA
共體國艱	VH	VA
發監執行	VC	VA
拋頭露臉	VC	VH
乘興而去	VA	VCL
曬曬太陽	VC	VA
回復過來	VA	VH
傲霜鬥雪	VA	VH
以策萬全	VA	VH
雕梁畫棟	VA	VH
大行其道	VE	VH
持之有故	VH	VA
視而未見	VC	VA
竄高伏低	VH	VA
弄虛作假	VC	VA

堆積下來	VA	VC
決定不了	VC	VE
放縱出來	VC	VA
鴨子聽雷	VA	VH
經受不起	VC	VJ
藉題發揮	VC	VA
傳染開來	VC	VH
遙不可及	VJ	VH
滲濕開來	VC	VH
妙筆生花	VA	VH
期待出來	VC	VH
隨而俱之	VA	VH
碰觸不得	VH	VC
知恩圖報	VC	VH
生存下去	VA	VH
未述其詳	VH	VA
據實以告	VE	VA
另當別論	VA	VH
勞而不獲	VC	VH
再蹈覆轍	VH	VA

統計式片語翻譯模型

張俊盛

游大緯

國立清華大學資訊工程研究所

jschang@cs.nthu.edu.tw

摘要

機器翻譯是自然語言處理研究上最重要的課題之一，在過去運用機器翻譯比較成功的例子，多是特定的領域文件的翻譯。近來因為網際網路與搜尋引擎的盛行，大家開始重視機器翻譯在跨語言檢索（Cross Language Information Retrieval）中的角色。在跨語言檢索的問題上，通常是對查詢字詞或片語，進行翻譯（Query Translation）。然而翻譯的結果必須和欲搜尋的文件庫的有高度的相關性，才能達到檢索的效果。目前的查詢關鍵詞翻譯的做法，或者採用現成的翻譯軟體，或者使用一般性的雙語詞典，都無法產生和文件相關的翻譯。因此我們希望能夠透過統計式機器翻譯的做法來進行查詢關鍵詞的翻譯，以提高跨語言檢索的效率。在這篇論文中，我們提出新的統計式片語翻譯模型，並進行實驗，證實能改進原有的統計式機器翻譯模型的缺點，提升片語對應與翻譯的效率。

1. 簡介

機器翻譯是自然語言處理研究上最重要的課題之一，有助於幫助使用者跨越語言與文化的障礙。在過去運用機器翻譯比較成功的例子，多是特定的領域文件的翻譯，如技術性的使用手冊、氣象報告、國際機構的官方文件。近來因為網際網路與搜尋引擎的盛行，大家開始重視機器翻譯

在跨語言檢索 (Cross-Language Information Retrieval)，可能扮演的角色。

在特定領域的文件翻譯上，機器翻譯系統主要是以句子為單位，進行處理。在跨語言檢索的問題上，可以採取「文件翻譯」(document translation)，或者「查詢資訊翻譯」(query translation)的做法 (McCarley 1999)。目前大部分的研究者都採取查詢關鍵詞翻譯的做法。例如，在 NTCIR-2 的英到中的資訊檢索評估活動 (Kando et al. 2001) 中的一個查詢主題中，就提供以下的英文關鍵詞，試驗參與的系統，找到相關中文新聞文件的能力：

- Assembly Parade Law
- Parade and Demonstration
- Constitution
- Freedom of speech
- Indemnification
- Communism
- Country separation
- Council of Grand Justices
- Legislation
- Amendments

查詢關鍵詞的翻譯涉及詞彙語義解析 (Word Sense Disambiguation) 的問題 (Ide and Veronis 1998, Chen and Chang 1998) 與片語的翻譯 (Phrase Translation) 的問題，和一般性翻譯很重要的不同點，在於翻譯的結果，是要拿來在一個文件庫 (Text Collection) 中搜尋文件。所以翻譯的詞義解析與翻譯的詞彙選擇 (Lexical Choice) 必須和文件庫的語料有高度的相關性。以上述關鍵詞中的 demonstration 為例，我們就必須翻譯成新聞中常見的「示威」而不能翻譯成「示範」。

目前學者研究跨語言檢索的做法，大致上分為兩種：

1. 利用市場上販售的翻譯軟體（Gey and Chen 1997, Kwok 2001）
2. 使用一般性的雙語詞典（Oard 1999, Kwok 2001）

這兩種做法，很明顯的都不容易產生和文件庫相關的翻譯。這一點對於音譯的專有名詞，特別明顯。Kwok 就指出使用現成翻譯軟體和一般性雙語詞典，不能得到 Michael Jordan 在文件庫的正確音譯「麥可喬丹」，顯然是跨語言檢索研究的一大問題。

為了提高翻譯和文件庫的相關性，Chen 等（1999）將詞彙共現機率（occurrence statistics）導入翻譯詞彙選擇的考慮中。有鑒於音譯專有名詞在跨語言檢索的重要性，也有研究者提出了一些統計或規則式的做法，將音譯轉換成原始專有名詞（Knight and Graehl 1997, Chang et al. 2001）。這些做法，雖然對於跨語言檢索有一定的效果，但缺乏比較全面性而嚴謹的理論架構，也因此影響到改進發展的空間。

我們認為要做好跨語言檢索中的查詢關鍵詞的翻譯，必須有一套全面而嚴密的方法，發展適用的機器翻譯模型。在機器翻譯的做法中，範例為本做法（Example-based Approach）和統計式機器翻譯，都是比較資料導向（data-driven）的做法，比較能夠產生和資訊檢索文件庫相關的翻譯。其中又以 IBM Watson 研究中心的 Brown 等（1988, 1990, 1993）提出的統計式機器翻譯做法，在理論上較為嚴謹，在架構與做法上較為明確可行。

因此我們希望能夠透過一種新的統計式對應與機器翻譯做法（Statistical Alignment and Machine Translation）來進行查詢關鍵詞的翻譯，為跨語言檢索的查詢詞翻譯提供一個比較有效而解嚴謹的做法。在這篇論文中，我們提出一種新的翻譯對應機率（Alignment Probability）

的做法，並進行實驗。實驗的結果證實新的模型的確能改進片語對應與翻譯的效率。

2. 統計式機器翻譯模型

機器翻譯早期是以逐字翻譯加上局部的位置調整的直接做法 (Direct Approach)，後來逐漸轉成主要是以句法分析為基礎的轉換式的做法 (Transfer Approach)。在 1980 年代末，研究的趨勢比較傾向實證式的做法 (Empirical Approach)，以翻譯的範例或平行語料庫為本，發展機器翻譯系統。Brown 提出的語料庫為本之統計式做法，在理論的架構最為完備。在 Brown 的統計式機器翻譯模型下，原文 S 和譯文 T 的翻譯機率 (Translation Probability) $Pr(T|S)$ ，可以分解成以下的三個機率函數：

- (a) 詞彙翻譯機率 (Lexical Translation Probability)

$$Pr(S_i | T_j)$$

- (b) 孳生機率 (Fertility Probability)

$$Pr(a | b)$$

- (c) 位置扭曲機率 (Distortion Probability)

$$Pr(i | j, k, m)$$

其中

S_i 為 S 的第 i 個字

T_j 為 T 的第 j 個字

a 為 S_i 的長度

b 為 T_j 的長度

k 為 S 的長度

m 為 T 的長度

Brown 等使用加拿大國會議事錄的英法平行語料庫，證實透過反覆交替的「期望值估計」與「最佳化」演算法 (Expectation and Maximization

Algorithm)，可以得到這三個簡單的機率函數的統計估計值。其「最佳化」的步驟，就是在目前的機率函數估計值下，求取最可能的翻譯對應。而「期望值估計」的步驟，就是以所有的雙語語料樣本的最佳的翻譯對應為根據，估計三個機率函數值。

透過 EM 演算法，統計式機器翻譯模型中的翻譯機率函數的估計值可趨於收斂。在雜訊通道模型 (Noisy Channel Model) 下，結合翻譯機率函數，與目標語的 N-gram 語言模型 (Language Model)，可以用搜尋演算法，如束限搜尋法 (Beam Search) 求最佳機率值的方式，產生翻譯。

3. 適用於片語對應與翻譯的統計式模型

Brown 原始模型中的位置扭曲機率，是基於每一字的翻譯目標位置和其他字無關的假設。在獨立事件的假設下，某一個翻譯對應 (alignment) 方式的機率，在位置方面而言，是所有字的和對應字的位置形成的位置扭曲機率值的乘積。實際上，每一字的翻譯目標位置和其他字的翻譯位置有高度的相關性。如果 $S_i, i' \neq i$ 都不對應到 T_j ，則 S_i 對應到目標位置 j 的機率幾乎為 1

$$Pr(j | i, k, m) \cong 1 \text{ 若 } Pr(j | i', k, m) = 0, i' \neq i$$

因此獨立假設下的機率，幾乎大部分的情況下都是過低的估計。即便是很可能的翻譯對應方式，其機率值還是一連串位置扭曲機率的乘積，因此常趨於非常的小的數值。例如，檢視三字英文五中文字的片語樣本，最可能翻譯對應 A^* 下的三個字 $S_1 S_2 S_3$ 翻譯目標位置，分別是

$$S_1 \rightarrow \{T_1, T_2\}$$

$$S_2 \rightarrow \{T_3, T_4\}$$

$$S_3 \rightarrow \{T_5\}$$

也就是 $A^* = (0, 12, 34, 5)$ (第一個 0 代表所有的中文字都有對應，沒有中文字無法對應到英文字的情況)。在 $k = 3$ 及 $m = 5$ 的片語樣本中，翻譯對應為 A^* 的情況約佔 35%。直接估計 A^* 的最大可能估計值 (Maximum Likelihood Estimation)，得到

$$Pr_{MLE}(A^*) = 0.35$$

然而在機率獨立的假設下

$$Pr(A^*) = P(1|1,3,5) P(2|1,3,5) P(3|2,3,5) P(4|2,3,5) P(5|3,3,5)$$

即使以較高的位置扭曲機率值 (0.6) 估計 $P(j|i,3,5)$ ，其乘積仍然過低，遠低於合理的估計值：

$$Pr(A^*) < (0.6)^5 = 0.046656 \ll 0.35$$

$$Pr(A^*) \ll Pr_{MLE}(A^*)$$

為了更精確合理的估計翻譯目標位置的機率，我們提出了直接估計整體翻譯配對位置與字數的做法。在此做法下，孳生機率和位置扭曲機率合併成為翻譯對應機率 (Alignment Probability)。因此不再獨立考慮個別的字的位置、翻譯目標位置、孳生的字數，而是以翻譯對應來一併考慮。在這樣的想法下，我們將原文 S 和譯文 T 的翻譯機率 $Pr(T|S)$ ，分解成以下的兩個機率函數：

(a) 詞彙翻譯機率 (Lexical Translation Probability)

$$Pr(T(A_i) | S_i)$$

(b) 翻譯對應機率 (Alignment Probability)

$$Pr(A | k, m) = Pr(A_0, A_1, A_2, \dots, A_k | k, m)$$

其中

S_i 為 S 的第 i 個字

$T(A_i)$ 為 T 中對應到 S_i 的部分

A_0 為 T 中沒有對應到 S 的部分的標號

A_i 為 T 中對應到 S_i 的部分的標號, $i > 0$

k 為 S 的長度

m 為 T 的長度

如果個別字 S_i 的對應字在 T 中為連續，則我們可以用對應目標位置

的起點 $B(A, i)$ 與終點 $E(A, i)$ ，來簡化對應關係的表達，也就是

$$S_i \rightarrow T(A_i) = \{ T_{B(A,i)}, T_{B(A,i)+1}, T_{E(A,i)} \}$$

$$A_i = \{ B(A,i), B(A,i)+1, \dots, E(A,i) \}$$

4. 實驗

我們進行了一系列的實驗，以驗證我們提出的新的片語翻譯模型的效果與可行性。透過實驗，我們想了解新模型有關的下列幾個問題：

1. 以翻譯對應機率替代孳生機率和位置扭曲機率，是否可以得到較正確的對應分析？
2. 翻譯對應是否集中在幾種樣式，而不是許多個別對應目標位置的排列組合？翻譯對應機率的參數量，會不會過多，導致估計的速度會不會過慢？
3. 翻譯對應機率的參數量和樣本數量，相較之下，其機率值的統計可靠度會不會過低？
4. 訓練後的機器翻譯模型，應用到跨語言檢索的可行性高或低？

4.1 實驗的設計與起始機率值的設定

由於不易取得大量雙語片語的語料，我們採用 BDC 漢英字典(BDC 1992) 的片語條目作為實驗的原始材料。為了配合實驗的目標，並簡化問題，我們首先去掉英文多於 3 個詞的條目，但中文長度不限。另外我們也去掉中文的四字成語條目。這些條目的翻譯，常常不是字面翻譯，去掉之後，可以降低資料的雜訊。原始資料經過整理之後，我們得到 96,156 筆可用的英中片語翻譯的記錄。我們以 (P_n, Q_n) ， $n = 1, N$ 來代表這組語料。

在試驗中，我們以 EM 演算法，來得到第三節所提出的辭彙翻譯機率、翻譯對應機率。我們採取了和一般不同，但類似 Och 等人 (2000)

對於 IBM 機率模型的改進實驗的做法。其目的都是希望加速機率的估計。

1. 開始的時候，我們採取 Brown 模型原有的位置扭曲機率。在 EM 演算法的第二輪之後才開始使用新模型的翻譯對應機率。
2. 我們假設英中片語翻譯時，英文和中文字的順序一致的機率較高。所以第一輪運算機率模型的位置扭曲機率不用一般常用的平均分布 $Pr(j | i, k, m) = 1/m$ ，而採用無母數的統計法，令位置扭曲機率的值如下：

$$Pr(i | j, k, m) = 1 - \left| \frac{j-0.5}{m} - \frac{i-0.5}{k} \right| \quad [1]$$

其中 i = 英文字位置， k = 英文字總數， j = 中文字位置， m = 中文字總數。

S	T	i	k	j	M	$Pr(j i, k, m)$
flight	8	1	2	1	4	0.875
flight	字	1	2	2	4	0.875
flight	飛	1	2	3	4	0.625
flight	行	1	2	4	4	0.375
eight	8	2	2	1	4	0.375
eight	字	2	2	2	4	0.625
eight	飛	2	2	3	4	0.875
eight	行	2	2	4	4	0.875

表 1 位置扭曲機率的無母數統計

對於每一筆雙語片語，我們假設每個英文字可以翻譯成其中任何一個中文字，但是其機率會因位置不同而異。例如某一筆記錄是 2 個英文字翻譯成 4 個中文字，我們可以得到 8 個英中文字的任意配對。每一個配對的位置扭曲機率和公式 1 的 $Pr(j | i, k, m)$ 值成正比。例如，對語料中雙語片語 (flight eight, 8 字飛行)，我們用公式 1 可以計算得到如表 1 的任意詞彙配對的位置扭曲機率。

有了任意配對的位置扭曲機率後，我們就可據此估計語料庫片語中的任何英文字 E 和中文字 C 間的翻譯機率 $\Pr(C|E)$ ，公式如下：

$$\Pr(C|E) = \frac{\sum_{n=1}^N \sum_{i=1}^k \sum_{j=1}^m \delta(E, P_n(i)) \delta(C, Q_n(j)) \Pr(j|i, k, m)}{\sum_{n=1}^N \sum_{i=1}^k \sum_{j=1}^m \delta(E, P_n(i)) \Pr(j|i, k, m)} \quad [2]$$

其中 $P_n(i)$ 為 P_n 之第 i 字， $Q_n(j)$ 為 Q_n 之第 j 字， $k = |P_n|$ ， $m = |Q_n|$

$\delta(x, y) = 1$ 若 $x = y$ ， $\delta(x, y) = 0$ 若 $x \neq y$

S_i	T_j	i	k	j	m	$\Pr(j i, k, m)$	$\Pr(T_j S_i)$	$\Pr(T_j S_i) \Pr(j i, k, m)$
flight	8	1	2	1	4	0.875	0.00797	0.00697
flight	字	1	2	2	4	0.875	0.00797	0.00697
flight	飛	1	2	3	4	0.625	0.25770	0.16106
flight	行	1	2	4	4	0.375	0.16901	0.06338
eight	8	2	2	1	4	0.375	0.02903	0.01089
eight	字	2	2	2	4	0.625	0.04839	0.03024
eight	飛	2	2	3	4	0.875	0.06774	0.05927
eight	行	2	2	4	4	0.875	0.06774	0.05927

表 2 位置扭曲機率與詞彙翻譯機率的估計值

公式 2 的用意在於加總 E 和 C 的在所有片語中的機率值，並除以 E 和所有中文的機率值的總和，使得 $\Pr(C|E)$ 的機率值介於 0 和 1 之間。依據公式 2 所得到的機率值，我們可以估計任何片語內任意字的配對的機率值。表 2 列出表 1 的任意配對的詞彙翻譯機率。

4.2 EM 演算法的第一輪計算

第一次的對應最佳化

有了起始的機率函數估計值，我們就可以進行 EM 演算法中的最佳化步驟。我們採取簡單的貪婪法 (Greedy Method) 來求取每一組雙語片語 (P_n, Q_n) 的最佳對應。我們假設簡單的孳生模型：一個英文可以對應到 0 到多個中文字，而每個中文字只能對應到最多一個英文字。有了片語內的詞彙翻譯與位置扭曲機率的起始估計值與其乘積 (如表 2)，我們就可以逐次選取最高機率值者，產生英文和中文字的配對，並根據假設的孳生模型，排除其他的英文字和此中文字的配對。反覆的執行上述步驟，直到沒有剩餘的中文字，或機率值低於某一個門檻值 (threshold) 為止。若有剩餘的中文字，就視為沒有對應到英文字。最低對應的機率門檻值，可以避免信賴度太低的錯誤對應，也有助於導入 0 對 1, 0 對多的孳生模式。經過實際抽樣觀察之後，以 0.008 為門檻值，可去掉大部分低信賴度的錯誤配對。再回到 “flight eight” 的例子，由表 2 的機率值，我們可得到如表 3 的對應方式 (0, 34, 12)。

S_i	T_i	i	j	k	m	$Pr(j i,k,m)$	$Pr(T_j S_i)$	$Pr(T_j S_i) Pr(j i,k,m)$
flight	飛	1	3	2	4	0.625	0.25770	0.16106
flight	行	1	4	2	4	0.375	0.16901	0.06338
eight	8	2	1	2	4	0.375	0.02903	0.01089
eight	字	2	2	2	4	0.625	0.04839	0.03024

表 3 (flight eight, 8 字飛行) 之最佳對應 (0, 34, 12)

期望值的估算 - 翻譯對應機率函數

經過機率最佳化求取最可能的對應方式後，我們就可以拋棄個別字的位

置扭曲機率，導入新的翻譯對應機率模型，直接估計整個對應方式的機率值。我們依照片語的英中文字數，統計出英中文字數 k 與 m 固定下，各種對應方式 A 的機率：

$$\Pr(A|k,m) = \frac{\text{count}(A \text{ 為 } (\mathbf{S}, \mathbf{T}) \text{ 的對應})}{\text{count}(k = |\mathbf{S}|, m = |\mathbf{T}|)} \quad [3]$$

k	m	A			$Pr(A k,m)$
		A_0	A_1	A_2	
2	4	0	12	34	0.572025052
2	4	0	123	4	0.121317560
2	4	0	1	234	0.085479007
2	4	0	1234	0	0.078056136
2	4	0	0	1234	0.065066110
2	4	0	124	3	0.020992809
2	4	0	2	134	0.016585479
2	4	0	3	124	0.007886801
2	4	0	34	12	0.005915101
2	4	0	13	24	0.004059383
2	4	0	134	2	0.003363489
2	4	0	23	14	0.002551612
2	4	0	4	123	0.002319647
2	4	1	0	234	0.002087683
2	4	0	234	1	0.001855718

表 4 兩字對四字片語的翻譯對應機率值最高的前 15 名

在實驗中，EM 演算法的第一輪自動的發掘出 601 種對應方式。以兩字對四字片語而言，有 38 種方式。表 4 列出依照機率由高到低排列的前 15 名對應方式。由表 4 可以觀察到幾點：

1. 機率估計的結果，和我們的認知沒有出入：
最可能的片語翻譯的順序是保留原文的順序。
同一英文字翻譯的目標位置是連續的。
一個英文字最可能翻譯到 2 個中文字。
2. 對應安排的機率值集中在少數的幾個樣式上。最可能的對應，佔了 90% 以上的機率。
3. 對應機率函數收斂的速度很快。

表 5 列出 2 對 4 字片語對應機率值前 5 名的實際例子。

S	T	T(A ₀)	S ₁	T(A ₁)	S ₂	T(A ₂)
T-shaped antenna	T 形天線		T-shaped	T 形	antenna	天線
X-ray examination	X 光檢查		X-ray	X 光	examination	檢查
irresistible force	不可抗		irresistible	不可抗	force	力
Unwritten law	不成文法		unwritten	不成文	law	法
Central Asia	中亞細亞		Central	中	Asia	亞細亞
mutual non-interference	互不干涉		mutual	互	non-interference	不干涉
undesirable element	不良少年		undesirable	不良少年	element	
unalterable truth	不易之論		unalterable	不易之論	truth	
come soon	不日放映		come		soon	不日放映
a desperado	不逞之徒		a		desperado	不逞之徒

表 5 二字到四字片語，最可能的 5 種對應方式的實例

期望值的估算 – 詞彙翻譯對應機率

在統計對應方式的機率的同時，我們同樣的也拿 4.2 節最佳化的結果，估計英文字翻譯成不同中文字的機率。我們採取和第一輪不一樣的做法，不再考慮英文字對應到中文單字的機率，而是考慮每一個英文字在片語中，所對應到的中文字串。這些中文字串大部分的情況是連續的，而且是詞典裡常見的詞項。當然也有少數的例子，英文的對應目標是空字串、

不連續字串、不能獨用的詞素 (bound morpheme) 等等情況。我們以“\$empty\$”來代表英文字對應到空字串的情況。考慮資料不足 (data sparseness) 的可能，我們導入“\$any\$”來代表英文字對應到訓練外的任意中文字串的情況，並採用 Good-Turing 的平滑化方法 (smoothing method) 來估計\$any\$的翻譯機率。

E	C	Pr (C E)
flight	飛行	0.6480231012
flight	飛	0.1411528654
flight	航空	0.0602616768
flight	\$empty\$	0.0296114718
flight	航	0.0296114718
flight	分	0.0041786956
flight	分隊	0.0041786956
flight	飛班機	0.0041786956
flight	飛航	0.0041786956
flight	飛機	0.0041786956
flight	航飛	0.0041786956
flight	黑	0.0041786956
flight	群	0.0041786956
flight	\$any\$	0.0000009248

表 6 “flight”翻譯成不同中文字串的機率

表 6 列出 flight 翻譯成不同中文字串的機率，包括一般的詞、詞素、\$empty\$、\$any\$。在這一輪的期望值估計中，flight 對應到\$empty\$的機率估計值 0.0296114718 仍然過高。只要翻譯對應機率如表 4 的(0,0,1234)

和 (1,0,234) 的機率，以及 \$any\$ 機率的估計值估計得合理，我們期望在 EM 演算法的以後的幾個輪迴中，兩者互相競爭的情況下，\$empty\$ 機率的估計值會逐漸的降低，而趨近合理的區段。

4.3 EM 演算法的第二輪計算

在第一輪的期望值估計之後，我們可以再次的求取片語的最可能對應方式。在第二輪的運算當中，我們不再使用公式 1 的位置扭曲機率，而是採用已經估計出來的整體性的翻譯對應機率。

S	T	A ₀	A ₁	A ₂	T(A ₀)	T(A ₁)	T(A ₂)	Pr(T S,A)
flight eight	8 字飛行	0	34	12		飛行	8 字	0.0000788100
flight eight	8 字飛行	0	3	124		飛	8 字行	0.0000000051
flight eight	8 字飛行	12	34	0	8 字	飛行	\$empty\$	0.0000000007
flight eight	8 字飛行	12	3	4	8 字	飛	行	0.0000000003
flight eight	8 字飛行	2	3	14	字	飛	8 行	0.0000000001

表 7 Pr(8 字飛行| flight eight) 機率值最高之前 5 名

第二輪運算中，我們對每一筆雙語片語 (S, T)，依據其英文和中文字數，考慮相符的所有的對應方式 A，計算其翻譯機率 Pr(T | S, A)。對於某一對應方式 A，Pr(T | S, A) 為 A 的機率和由 A 所決定的詞彙配對 (S_i, T(A_i)) 的機率乘積：

$$\Pr(T | S) = \max_A \Pr(T | S, A) = \max_A \Pr(A | k, m) \prod_{i=1}^k \Pr(T(A_i) | S_i)$$

因此最可能的對應 A* 可由下列公式決定

$$\begin{aligned}
A^* &= \arg \max_A \Pr(T | S, A) \\
&= \arg \max_A \Pr(A | k, m) \prod_{i=1}^k \Pr(T(A_i) | S_i)
\end{aligned}
\tag{4}$$

其中 $k = |\mathbf{S}|$, $m = |\mathbf{T}|$

以 $(\mathbf{S}, \mathbf{T}) = (\text{flight eight}, \text{8 字飛行})$ 為例，對於不同的對應 \mathbf{A} ，其翻譯機率的計算如下：

$\mathbf{A} = (0, 12, 34)$:

$$\Pr(\mathbf{T} | \mathbf{S}, \mathbf{A}) = \Pr(0, 12, 34 | 2, 4) \Pr(\text{8 字} | \text{flight}) \Pr(\text{飛行} | \text{eight})$$

$\mathbf{A} = (0, 34, 12)$:

$$\Pr(\mathbf{T} | \mathbf{S}, \mathbf{A}) = \Pr(0, 34, 12 | 2, 4) \Pr(\text{飛行} | \text{flight}) \Pr(\text{8 字} | \text{eight})$$

$\mathbf{A} = (0, 3, 124)$:

$$\Pr(\mathbf{T} | \mathbf{S}, \mathbf{A}) = \Pr(0, 3, 124 | 2, 4) \Pr(\text{飛} | \text{flight}) \Pr(\text{8 字行} | \text{eight})$$

$\mathbf{A} = (2, 34, 1)$:

$$\Pr(\mathbf{T} | \mathbf{S}, \mathbf{A}) = \Pr(2, 34, 1 | 2, 4) \Pr(\text{飛行} | \text{flight}) \Pr(\text{8} | \text{eight})$$

$\mathbf{A} = (12, 34, 0)$:

$$\Pr(\mathbf{T} | \mathbf{S}, \mathbf{A}) = \Pr(12, 34, 0 | 2, 4) \Pr(\text{飛行} | \text{flight}) \Pr(\text{\$empty\$} | \text{eight})$$

表 7 列出 (flight eight, 8 字飛行) 的幾個最高翻譯機率值的對應方式。表 7 的數值顯示第二輪的統計估計值已經相當的收斂，可以導出正確的對應分析 $\mathbf{A}^* = (0, 34, 12)$ 。

5. 實驗結果與評估

我們進行的實驗，證明了新的統計式片語翻譯模型確實可行，能產生相當正確的對應分析。新模型中導入的翻譯對應機率的參數不會過度的膨脹，因此 10 萬筆的資料就可以估計出相當可靠的各項機率值。由於新的模型，避免了許多機率值的乘積，EM 演算法的花費的時間較少，機率函數的收斂速度也比較快。

S	T	第一輪結果			第二輪結果		
		T(A ₀)	T(A ₁)	T(A ₂)	T(A ₀)	T(A ₁)	T(A ₂)
association football	A 式足球			A 式足球		A 式	足球
delay flip-flop	D 型正反器			D 型正反器		D 型	正反器
I demodulator	I 信號解調器			I 信號解調器		I 信號	解調器
Disgraceful act	不友好行動			不友好行動		不友好	行動
disregard to	不拘於		不拘於			不拘	於
secret ballot	不記名投票			不記名投票		不記名	投票
bearer stock	不記名股票		不記名票	股		不記名	股票
false retrieval	不實檢索			不實檢索		不實	檢索
used car	中古車		中古車			中古	車
infix operation	中序運算		中序運算			中序	運算

表 8 第二輪運算之後翻譯結果變好的例子

由實驗的結果觀察，以翻譯對應機率替代孳生機率和位置扭曲機率，確實可以得到比較正確的對應分析。在對應的問題比較困難的幾個情況仍然能夠做出正確的分析：

1. 比較偏離常態的罕用翻譯，如 association 通常翻譯成「協會」、「學會」。而 association football 中卻翻譯成類音譯的「A 式」。
2. 翻譯非常分散，沒有定譯的虛詞或輕動詞 (light verb)，如 make、take、to 等。
3. 和原文不一致的翻譯順序，如 (flight eight, 8 字飛行)。

我們檢視對應分析的結果，特別觀察這幾種困難的情況，比較其第一輪和第二輪分析的結果。我們發現 Brown 原始模型不盡理想，使得第一輪的許多分析不正確。這些情況到了第二輪時，使用了新模型的分析

後，大部分很明顯的已經扭轉到正確的分析。部分的例子，請參見表 8。

為了評估實驗的效能，我們使用 Och 等人(2000) 的評估方法。我們從實驗第二輪結果中，隨機抽取 100 個樣本 (包含 2 個英文字及 3 個英文字的樣本各 50 個)，由人工對這些樣本做對應的標示，以作為參考答案。標示分為 2 種：S (sure) 和 P (possible)，S 表示確定的對應，P 表示可能的對應，且 $S \subseteq P$ 。將實驗的結果與人工標示的參考答案比較，我們可以得到以下的召回率(recall)、準確率(precision)與錯誤率(error rate)：

$$recall = \frac{|AI \ S|}{|S|} = \frac{185}{212} = 0.873$$

$$precision = \frac{|AI \ P|}{|A|} = \frac{226}{273} = 0.828$$

$$AER(S, P; A) = 1 - \frac{|AI \ S| + |AI \ P|}{|A| + |S|} = 1 - \frac{185 + 226}{273 + 212} = 0.153$$

6 討論

我們就實驗的結果，進一步討論平滑化的改進做法。我們也分析統計式片語翻譯模型的可能的應用。

6.1 其他平滑化方法

由訓練語料得到各項翻譯的機率後，我們可以用這些機率，繼續對應訓練以外的其他片語，或是進行片語的翻譯。此時，我們可能會因為資料不足，而遇到訓練資料以外的情況，例如

(flight attendant 空服員)

我們的訓練語料，並沒有 (flight, 空) 的詞彙配對，來正確的分析 (flight attendant 空服員) 的對應。當然此時我們可以應用 flight 對\$any\$的機率。但是\$any\$的機率是平均的分配，無法考慮到有少數訓練外的狀況比較可

能，而大多數訓練外的狀況相當不可能。這些少數比較可能的狀況和中文縮寫的特性有關。另外有些中文字容易孳生很多的同義或近義，也會造成相同的效應。

中文的使用有縮寫的現象，所以訓練內的詞彙配對如 (flight, 航空) 的部分翻譯 (flight, 航) 與 (flight, 空) 在訓練外出現的可能性不低，而非部份翻譯的 (flight, 員) 則趨近於 0。同樣的，在 (attendant, 服務員) 配對例子中，部分翻譯 (attendant, 服員)、(attendant, 服)、(attendant, 服務) 的可能性也顯然高於 (attendant, \$any\$) 的平均值。另外翻譯有部份重疊的情況，也應給予較高的平滑機率。例如訓練內的詞彙配對有 (preservation, 保留)，而 (preservation, 保持) 與 (preservation, 保護) 雖然沒有出現在訓練語料中，其可能性仍然高於其他完全沒有重疊的翻譯配對。

如果沒有這樣的考慮，對於 (flight attendant 空服員) 的對應，詞彙翻譯機率就會全然都使用 $Pr($any$|flight)$ 和 $Pr($any$|eight)$ 的機率值，無法區隔可能與不可能的配對。如此將流於由翻譯對應機率 $Pr(A|2,3)$ 來決定一切。在這種狀況下，由於兩字到三字片語的最高可能對應為 (0, 12, 3)，我們很可能得到以下的不完全正確的對應分析：

(flight, 空服)

(attendant, 員)

若能考慮翻譯部分符合的條件，給予 (flight, 空) 與 (attendant, 服員) 較高的平滑機率估計值，則比較容易得到正確的對應分析，如

(flight, 空)

(attendant, 服員)

目前我們正實驗以英文到中文單字以及英文到中文雙字的兩組 LTP，來合成機率估計值。實驗的目標在於，讓部分字相符的對應，可以透過單字 LTP 模型得比較合理的估計值。

6.2 統計式片語翻譯模型的應用

我們提出的新的統計式片語對應語翻譯的模型，可以應用於翻譯跨語言檢索中的關鍵詞。新模型也可以在平行語料庫擷取雙語的片語，提供建立語料庫相關的翻譯詞典，作為翻譯與術語管理的基本工具。

6.2.1 在跨語言檢索上的應用

在跨語言檢索的研究顯示通常不到一半的查詢的關鍵詞組，可以在雙語詞典中查到適當的翻譯。當詞典沒有收錄詞組時，我們必須逐字翻譯，通常每字都有許多的翻譯，而只有部分和查詢主題相關。統計式的片語翻譯模型，可以發揮幾種作用：

1. 模型中的詞彙翻譯機率，提供比雙語詞典單字詞條更豐富的諸多可能性。若不能全部採用，也可取機率值最高、同時文件頻率（document frequency）適中的翻譯。這個做法簡單可行，也可稍稍減少翻譯和文件庫相關性不大的缺點。
2. 透過雜訊通道模型，結合翻譯機率函數，與文件庫所訓練出來的 N-gram 語言模型，可以用搜尋演算法，產生機率最佳化的翻譯。這個做法可以大大的提升翻譯結果的文件庫相關性。

明確的來說，以訓練出來的辭彙翻譯機率和翻譯對應機率，我們更容易求取查詢關鍵詞組 S 的最佳翻譯 T^* 。我們可以使用下列公式：

$$\Pr(T|S) = \max_A \Pr(T|S, A)$$

$$\begin{aligned}
T^* &= \arg \max_T \Pr(T | S) \Pr(T) \\
&= \arg \max_T \max_A \Pr(T | S, A) \Pr(T) \\
&= \arg \max_T \max_A \Pr(A | k, m) \prod_{i=1}^k \Pr(T(A_i) | S_i) \Pr(T)
\end{aligned}$$

其中 $\Pr(T)$ = 翻譯 T 的語言模型機率

k = S 的字數

m = T 的字數

$\Pr(A | k, m)$ 的模型簡化了 $\Pr(T | S)$ 的計算，使得我們更容易以分岔與限制演算法 (Branch and Bound Algorithm) 搜尋機率最大化的 T^* 值。我們可以由最高機率值的 $\Pr(A | k, *)$ 、 $\Pr(* | S_i)$ 組合的解 (solution) 開始搜尋，建立搜尋的高限 (upper bound)，再利用 $\Pr(* | S_i)$ 、 $\Pr(T)$ 的 N-gram 模型的最高機率值，達到限制搜尋範圍的效果。在不遺漏最佳解的情況下，縮短搜尋的時間。

6.2.2 在雙語片語對應的應用

學者大多認為統計式機器翻譯最有應用潛力的地方，在於雙語詞典的編輯與機率翻譯詞典的發展。當然在這樣的考慮下，詞彙對應的發現，不限於單字詞，而應及於多字的片語 (Kupiec 1993)。

利用新發展出來的模型，我們提出一套逐句進行的片語對應做法。以新的統計式片語翻譯模型為中心，我們可以透過下列的步驟，在英中例句中擷取對應的英中片語。對例句中的英文的基本片語 P 我們可以計算其最佳的翻譯對應 A 與翻譯目標 $T_{j+1, j+m}$ ：

$$\begin{aligned}
&\arg \max_{A, j, m} \Pr(A | n(i), m) \Pr(T_{j+1, j+m} | P, A) \\
&= \arg \max_{A, j, m} \Pr(A | n(i), m) \Pr(T_{j+1, j+m} | S_{b, e}, A) \\
&= \arg \max_{A, j, m} \Pr(A | n(i), m) \prod_{k=1}^{n(i)} \Pr(T_{j+B(A, K), j+E(A, k)} | S_{b+k-1})
\end{aligned}$$

其中 A 為 n 個英文對應到 m 個中文的任何可能的對應

$Pr(A | n, m)$ 為 A 的機率函數值

S_{b+k-1} 為基本片語 P 的第 k 個字

$T_{j+B(A,k), j+E(A,k)}$ 為 P 的第 k 個字在翻譯對應 A 下的翻譯目標

逐句式的雙語片語對應演算法

輸入：英文句子 S 中文句子 T

輸出：一組 k 個雙語片語 $(P_1, Q_1), (P_2, Q_2), \dots, (P_k, Q_k)$

- (1). 以詞性分析程式 (Part of speech tagger) 分析英文句子 S 的單字的詞性，並取得原形化後的字根 (lemma)。
- (2). 以淺型剖析 (shallow parsing) 的做法，擷取句子的基本片語 (basic phrase) : P_1, P_2, \dots, P_k 。令 $P_I = S_{b(i), e(i)}$, $|P_b| = n(i)$
- (3). 對每一個基本片語 P_I 計算其最佳的翻譯對應 A 與翻譯目標 $T_{j^*(i)+1, j+m^*(i)}$ 。做法是以分岔與限制 (Branch and Bound Algorithm) 搜尋最大化下列公式的 $A^*(i), j^*(i), m^*(i)$

$$\arg \max_{A, j, m} Pr(A | n(i), m) \prod_{k=1}^{n(i)} Pr(T_{j+B(A,k), j+E(A,k)} | S_{b+k-1})$$

- (4). 對每一個基本片語 P_I 輸出 $(P_I, T_{j^*(i)+1, j+m^*(i)})$

7. 結論與未來的研究方向

雖然統計式機器翻譯的研究，已經有十多年的歷史，在本研究中我們發現仍然有很大的改進空間，特別是在片語的對應與翻譯方面。我們提出新的統計式片語模型來進行片語的對應，並可應用於查詢關鍵詞的翻譯，以提高跨語言檢索的效率。我們在實驗中，初步驗證新的模型比起

Brown 的原始模型，確實可以在計算效率與對應效果上，有很大的改進。

統計式的做法確實對翻譯辭彙的整理、編輯有很大的效用。在這篇論文中，我們也討論統計式片語翻譯模型，應用於關鍵詞翻譯與雙語例句的片語對應的做法。

我們認為未來統計式雙語對應與機器翻譯，在跨語言檢索應該還有很大的空間可以發揮。幾個可能的研究方向包括：

1. 導入句法的訊息，對於不同的名詞、動詞、形容詞片語，訓練不同的統計式模型，可能可以導致更好的對應與翻譯的效果。
2. 導入詞彙語義解析的做法，先行分析查詢主題的語意範疇，縮小翻譯選擇的範圍。
3. 將語義的限制，如 Wordnet 的上位詞導入詞彙翻譯機率中，以配合詞彙語義解析的做法。

致謝

本文之研究受到國科會編號 892420H007001 計畫之補助。作者也非常感謝匿名審查者所提供之寶貴建議。

參考文獻

1. BDC 1992 *The BDC Chinese-English electronic dictionary* (version 2.0), Behavior Design Corporation, Taiwan.
2. Brown, P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Mercer R. L., and Roosin P. S. 1988 *A Statistical Approach to Language Translation*, In Proceedings of the 12th International Conference on Computational Linguistics, Budapest, Hungary, pp. 71-76.

3. Brown, P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., and Roosin P. S. 1990 *A Statistical Approach to Machine Translation*, Computational Linguistics, 16/2, pp. 79-85.
4. Brown, P. F., Della Pietra S. A., Della Pietra V. J., and Mercer R. L. 1993 *The Mathematics of Statistical Machine Translation: Parameter Estimation*, Computational Linguistics, 19/2, pp. 263-311.
5. Chang, J. S. et al. 2001. *Nathu IR System at NTCIR-2*. In Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, pp. (5) 49-52, National Institute of Informatics, Japan.
6. Chang, J. S., Ker S. J., and Chen M. H. 1998 *Taxonomy and Lexical Semantics – From the Perspective of Machine Readable Dictionary*, In Proceedings of the third Conference of the Association for Machine Translation in the Americas (AMTA), pp. 199-212.
7. Chen, H.H., G.W. Bian and W.C. Lin. 1999. *Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval*. In Proceedings of the 37th Annual Meeting of the Association for Computation Linguistics, pp 215-222.
8. Dagan, I., Church K. W. and Gale W. A. 1993 *Robust Bilingual Word Alignment or Machine Aided Translation*, In Proceedings of the Workshop on Very Large Corpora Academic and Industrial Perspectives, pp. 1-8.
9. Fung, P. and McKeown K. 1994 *Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping*, In Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA), pp. 81-88, Columbia, Maryland, USA.
10. Gale, W. A. and Church K. W. 1991 *Identifying Word Correspondences in Parallel Texts*, In Proceedings of the Fourth DARPA Speech and Natural Language Workshop, pp. 152-157.
11. Gey, F C and A. Chen. 1997. *Phrase Discovery for English and Cross-Language Retrieval at TREC-6*. In Proceedings of the 6th Text Retrieval Evaluation Conference, pp 637-648.

12. Ide, N. and J Veronis. 1998. *Special Issue on Word Sense Disambiguation*, editors, Computational Linguistics, 24/1.
13. Isabelle, P. 1987 *Machine Translation at the TAUM Group*, In M. King, editor, Machine Translation Today: The State of the Art, Proceedings of the Third Lugano Tutorial, pp. 247-277.
14. Kando, Noriko, Kenro Aihara, Koji Eguchi and Hiroyuki Kato. 2001. Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, National Institute of Informatics, Japan.
15. Kay, M. and Röscheisen M. 1988 *Text-Translation Alignment*, Technical Report P90-00143, Xerox Palo Alto Research Center.
16. Ker, S. J. and Chang J. S. 1997 *A Class-base Approach to Word Alignment*, Computational Linguistics, 23/2, pp. 313-343.
17. Knight, K. and J Graehl. 1997. *Machine Transliteration*, In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of ACL European Chapter, pp. 128-135.
18. Kupiec, Julian. 1993 *An Algorithm for finding noun phrase correspondence in bilingual corpus*, In ACL 31, 23/2, pp. 17-22.
19. Kwok, K L. 2001. *NTCIR-2 Chinese, Cross-Language Retrieval Experiments Using PIRCS*. In Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, pp. (5) 14-20, National Institute of Informatics, Japan.
20. Longman Group 1992 *Longman English-Chinese Dictionary of Contemporary English*, Published by Longman Group (Far East) Ltd., Hong Kong.
21. McCarley, J. Scott. 1999. *Should we Translate the Documents or the Queries in Cross-Language Information Retrieval?* In Proceedings of the 37th Annual Meeting of the Association for Computation Linguistics, pp 208-214.
22. Melamed, I. D. 1996 *Automatic Construction of Clean Broad-Coverage Translation Lexicons*, In Proceedings of the second Conference of the Association for Machine Translation in the Americas (AMTA), pp. 125-134.
23. Nagao, M. 1986 *Machine Translation: How Far Can it Go?* Oxford University Press, Oxford.

24. Oard, D W and J. Wang. 1999. *Effect of Term Segmentation on Chinese/English Cross-Language Information Retrieval*. In Proceedings of the Symposium on String and Processing and Information Retrieval. <http://www.glue.umd.edu/~oard/research.html>.
25. Och, Franz Josef and Hermann Ney. 2000. *Improved Statistical Alignment Models*. In Proceedings of the 38th Annual Meeting of the Association for Computation Linguistics.
26. Pirkola, A. 1998. *The Effect of Query Structure and Dictionary Setups in Dictionary-based Cross-Language Retrieval*. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 55-63.
27. Smadja, F., McKeown K., and Hatzivassiloglou V. 1996 *Translating Collocations for Bilingual Lexicons: A Statistical Approach*, Computational Linguistics, 22/1, pp. 1-38.
28. Utsuro, T., Ikeda H., Yamane M., Matsumoto M., and Nagao M. 1994 *Bilingual Text Matching Using Bilingual Dictionary and Statistics*, In Proceedings of the 15th International Conference on Computational Linguistics, pp. 1076-1082.
29. Wu, D. and Xia X. 1994 *Learning an English-Chinese Lexicon from a Parallel Corpus*, In Proceedings of the first Conference of the Association for Machine Translation in the Americas (AMTA), pp. 206-213.

Metaphor, Inference, and Conceptualisation* :
On the Development of V-diao Construction in Mandarin
Louis Wei-lun Lu (呂維倫)
Graduate Institute of Linguistics, National Taiwan University
weilunlu@taiwan.com / r89142004@ms89.ntu.edu.tw

Abstract

Keywords: Construction Grammar, metaphorical transfer, conceptualization

V-diao constructions, according to their semantics, fall into three categories:

- A) Physical disappearance from its original position, with the V slot filled by physical verbs, such as *tao-diao* “escape”, *diu-diao* “throw away”, and so on.
- B) Disappearance from a certain conceptual domain, rather than from the physical space, with the V slot filled by less physically perceivable verbs, such as *jie-diao* “quit”, *wang-diao* “forget” and the like.
- C) The third category of V-diao involves speaker’s subjective, always negative, attitude to the result. Examples include: *lan-diao* “rot”, *ruan-diao* “soften”, *huang-diao* “turn yellow”, and so forth.

The meaning in Type C constructions cannot be gained by simply putting their component parts together, so in this study, I shall term V-diao as a construction (Goldberg 1995) rather than merely a resultative compound (Li and Thompson 1981).

Metaphor, as a mechanism of semantic change (Sweetser 1990, Bybee, Perkins and Pagliuca 1994, Heine, Claudi and Hunnemeyer 1991), is a plausible account of the polysemy between Type A and B. Type A denotes disappearance from physical space, while Type B disappearance from the conceptual space. I thus speculate on the mapping relation between the physical and the abstract, conceptual domain.

Other than metaphor, pragmatic inference is claimed to be a major mechanism of semantic change (Hopper and Traugott 1993, Bybee, Perkins and Pagliuca 1994). In such changes, context plays a crucial role. Frequent use of a grammatical or lexical unit in a particular context may lead to the inference that the context is an integral part of its meaning. The development of Type C V-diao may relate to frequent co-occurrence of negative verbs and -diao. (The reason why only negative verbs are allowed in the construction will be further addressed in the next section.) Consequently, negative connotation may spread to the entire construction and give rise to the constructional meaning.

* I’m grateful to Dr. Lily I-wen Su for her insightful comment on this paper.

There also exists a cognitive constraint on its applicability. The construction does not allow verbs with positive connotation in the V slot. This is because, the semantics of the construction cannot contradict the metaphor it is based on (Huang and Chang 1996). Also, it cannot override, either, the orientational metaphor based on human experiential basis (Lakoff and Johnson 1980): GOOD IS UP; DOWN IS BAD.

Hopefully, this study can serve as a valid argument for the interaction of our language use and grammar, and the conceptual basis of human language.

1 *V-diao* as a Construction

V-diao is traditionally termed as a resultative compound (Li and Thompson 1981). However, in this study, I shall avert this conventional terminology and treat it as a construction instead, because it actually denotes something more than what its components literally give. In this paper, I shall adopt the definition of Goldberg (1995), and Fillmore, Kay, and O'Connor (1988), to define a "construction" as follows: A construction refers to a form-meaning pair, the meaning of which cannot be strictly predictable from its component parts or from other previously established constructions. It may specify not only syntactic, but also lexical, semantic, and pragmatic information.

But as a construction, what is its constructional meaning? Also, what are the driving forces of the emergence of constructional meaning? Furthermore, a selectional restriction seems to stand in what properly fits into the V-slot. In this paper, I shall attempt to look into the constructional meaning, its driving forces, and finally, the selectional constraint of the verb.

V-diao construction comprises a verb (be it action or stative), and a verbal suffix *-diao*. It gives the final state of the agent, if used intransitively, and of the receiver of the action, in transitive cases. It may represent: A) Physical disappearance of an entity from its original position, B) Disappearance from a certain conceptual domain,

and C) Speaker's subjective evaluation on the result of an event, as in (1)-(3)

respectively.

(1) ta qiaoqiao pao-diao le
 he quietly run away CRS
 "He ran away quietly."

(2) ta jie-diao le nage huai
 xiguan
 he get rid of Perf that bad habit
 "He got rid of that bad habit."

(3) diennau zuotien huai-diao le
 computer yesterday break down CRS
 "The computer broke down yesterday."

I shall begin this paper with a close look at the semantics of the foregoing types of *V-diao*, especially the last one, since Type C constructions involve an intriguing phenomenon: the interpretation of negative results cannot be gained directly from the compositional parts of a construction, as indicated in Goldberg (1995). Later I shall further look into how the constructional meaning emerges.

1.1 Physical Disappearance

It is reported that a suffix in a resultative verb compound indicates the result of an action (Li and Thompson 1981). The first kind of *-diao* gives the final state, i.e., physical absence, of the agent or the patient. Mostly this kind of *-diao* is affixed to easily perceivable physical action verbs such as *pao* "run", as in (1), *diu* "throw", *shao* "burn", and so on.

1.2 Disappearance from a Conceptual Domain

The second sort of *V-diao* denotes also the result of an action. However, this differs from type A in the sense that it represents a less "concrete" disappearance. It is often attached to low transitive verbs, without obvious physical motion, and accompanies an abstract noun phrase. Consider example (2) again:

(2) ta jie-diao le nage huai
 xiguan
 he get rid of Perf that bad habit
 "He got rid of that bad habit."

A bad habit is an abstract entity. The giving up of it is almost undetectable. But how can we perceive its existence and disappearance? Also, from where does the habit disappear?

This has everything to do with our conceptual system. We experience many things, through sight and touch, as having distinct physical shapes and boundaries. We thus tend to project physical shapes and boundaries on them, conceptualising them as entities and imposing on them physical characteristics such as existence and disappearance, even though we can never really feel them with our hands or sense them with our eyes or nose (Lakoff and Johnson 1980).

Therefore, in this case, a habit is conceptualised as a physical entity. It can be done away with, fade out, and finally disappear from our conceptual domain as

physical things do from the physical space. Thus, Type B seems to represent the final state of usually a non-physical action, i.e., an abstract entity being done away with and finally disappearing from one's conceptual domain.

1.3 Negative Evaluation on the Result

Type C *V-diao* denotes a somewhat negative evaluation on the result in question. It often co-occurs with verbs with negative connotation, such as *lan-diao* “rotten”, *si-diao* “die”, *shu-diao* “lose”, and et cetera. However, its negative meaning does not seem to come from the preceding verb in every case. Consider the following instances (4) and (5):

- | | | | | | | |
|-----|--|------------|----------|-----|---------|-----|
| (4) | binggan | ruan-diao | jiu | bu | hauchi | le |
| | cookie | soften | PARTICLE | not | tasty | CRS |
| | “Cookies won’t taste good if becoming soft.” | | | | | |
| (5) | cai | huang-diao | jiu | bu | xinxien | le |
| | vegetable | yellow | PARTICLE | not | fresh | CRS |
| | “Vegetables won’t be fresh if they turn yellow.” | | | | | |

In (4) and (5), the words *huang* “yellow” and *ruan* “soft” do not themselves carry negative meanings, but the constructions clearly involve one’s unfavourable attitude to the final state of vegetables and cookies. The constructional meaning, which carries negative attitude, cannot be gained from the compositional parts (Goldberg 1995), in this case, *-diao* and the verb preceding it. In the following sections, I shall examine the semantic change of *-diao*, and try to account for the

emergence of the constructional meaning of *V-diao*.

1.4 Data and Methodology

Two main sources provide the examples of the expressions discussed in this research. The written source mostly comes from the Academia Sinica Corpus. The spoken source comprises the Taida Spoken Corpus, and another eight hours of data from Professor Lily I-wen Su. An approximate total of sixteen hours of conversational Mandarin is adopted to serve the purpose of this study.

2 Metaphorical Relation

It is argued that, when a grammatical meaning is derived from its source, there often exists a metaphorical relation between the two meanings (Sweetser 1990, Bybee, Perkins and Pagliuca 1994). Such semantic change takes place to serve certain functional end in grammar and discourse, as indicated by Heine, Claudi and Hunnemeyer (1991:48):

We try to demonstrate that metaphorical transfer forms one of the main driving forces in the development of grammatical categories; that is, in order to express more “abstract” functions, concrete entities are recruited.

Similarly, Goldberg considers metaphor a mechanism to develop polysemous construction. Her study on the *way* construction indicates that metaphor be a plausible cause of semantic change, since it involves a “metaphorical self-created path” (1995: 203). This corresponds to my observation on *V-diao*: a metaphorical

transfer takes place when the construction proceeds from the physical domain to a conceptual domain, denoting metaphorical disappearance.

2.1 From Type A to B: Metaphor at Work

The above claim seems to be the case in the development of *-diao*. The meaning of Type A is the most concrete and physical, since it indicates a salient result after some physical action is carried out. Type B, on the other hand, denotes disappearance from our mental space instead of from the physical space. Now consider (6), a typical instance of such metaphorical transfer:

- (6) a. ta xiang pao keshi pao-bu-diao
He think run but run-not-away
“He tried to escape but failed.”
- b. zhuan ge shiwan pao-bu-diao
earn PARTICLE a hundred thousand run-not-away
“(Someone) should earn more than a hundred thousand dollars.”

Pao-bu-diao in (6a) denotes the unsuccessful outcome of the agent’s escape.

The agent fails to escape and will not disappear. In (6b), it means that, the landmark “a hundred thousand” is certain to be met. However, not every single case of Type B has a counterpart in A. Actually, most Type B constructions do not. *Pao-bu-diao* is simply a case employed to illustrate the metaphorical relation between the two polysemous constructions. In most cases of type B constructions, the V slot is filled by less physical verbs, such as *jie* “get rid of” in (2), *hulue* “ignore”, *wang* “forget”,

and so on.

2.2 Summary

In this section, I have shown that the physical “resultative compound” *V-diao* has undergone a metaphorical transfer, and develops the sense of disappearance from a conceptual domain. Thus, it makes perfect sense to speculate that the polysemy of the construction is at least partly contributed by metaphor, since disappearance is a common feature of Type A and B. This corresponds to the observation of Goldberg (1995). Figure 1 is representative of the mapping relation between Type A and B *V-diao*:

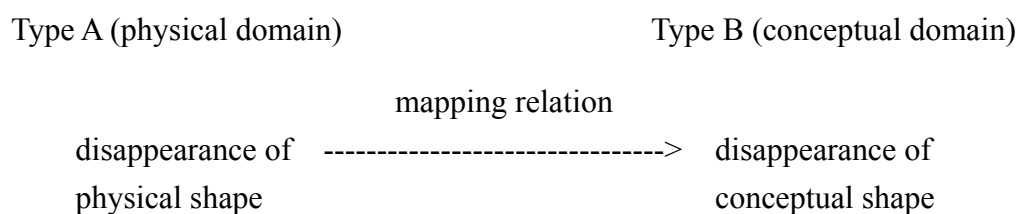


Figure 1 Metaphorical Transfer Between Type A and B *V-diao*

3 Pragmatic Inference

Other than metaphor, pragmatic inference is claimed to be a major mechanism of semantic change (Hopper and Traugott 1993, Bybee, Perkins and Pagliuca 1994). In such changes, context plays a crucial role. Frequent use of a grammatical or lexical

unit in a particular context may lead to the inference that the context is an incorporated part of its meaning. Goossens' research on Old English modals (1982) reports that, there were rarely "real" epistemic markers in OE, and that possibility markers frequently combined with adverbs to express epistemic functions. That is, speakers could have generalised and have extracted the epistemic meanings from the context and have imposed them on modals. This suggests that frequent co-occurrence with a particular context may "colour" the semantics of a gram.

It is highly likely that the final stage of development of *V-diao* is based on such mechanism. I have argued for the existence of the constructional meaning in 1.3. Now let us see how language use and context collaborate to lead to the constructional meaning in this case.

3.1 From Type B to C: Emergence of Constructional Meaning

In Type C construction, the sense of disappearance retains, but there seems to be something more than the combination of the verbal sense and disappearance. In general, these constructions involve undesirable assessment from the speaker. That is, the speaker obviously does not favour the consequence of the change of state.

It is noteworthy that Type C constructions can be further divided into two subtypes by the verbs in the V slot: 1) Verbs with negative connotation, such as *lan*

“rot”, *si* “die”, *po* “break”, *shu* “lose”, and so on. 2) Neutral verbs, such as *huang* “yellow”, *ya* “croak”, *ruan* “soft”, and so on. This classification highly pertains to the emergence of constructional meaning. Let us see how.

Initially, only the former constructions are formed. They simply denote a metaphorical disappearance, labeled Type B. As frequency of use increases, speakers tend to associate the construction with the adverse image. Such frequent collocation of verbs and the suffix may invite the inference that the constructions are used to express one’s unfavourable appraisal of the situation at issue. The context is thus “semanticized” (Hopper and Traugott 1993:75), and becomes an integral part of the construction. Consequently, the construction may accommodate neutral stative verbs in the V slot and still gain a negative interpretation. See (4) and (5) again for illustration:

- | | | | | | | |
|-----|--|------------|----------|-----|---------|-----|
| (4) | binggan | ruan-diao | jiu | bu | hauchi | le |
| | cookie | soften | PARTICLE | not | tasty | CRS |
| | “Cookies won’t taste good if becoming soft.” | | | | | |
| (5) | cai | huang-diao | jiu | bu | xinxien | le |
| | vegetable | yellow | PARTICLE | not | fresh | CRS |
| | “Vegetables won’t be fresh if they turn yellow.” | | | | | |

“Yellow” and “soft” themselves do not signal any adversity. The adverse meaning is subtly signalled and triggered by the frequent occurrence of negative verbs in the construction. In other words, the constructional meaning, i.e. speaker’s negative attitude, derives neither from the suffix denoting disappearance nor the verb

preceding it, but could have been generalised from the constant collocation of negative words and *-diao*. Now even neutral verbs may fit into the V slot and indicate negative assessments.

3.2 Summary

Pragmatic inference is one of the driving forces of semantic change, and I have proven that it is at crucial play in the development of *V-diao* construction as well. First only verbs that result in physical and conceptual disappearance may occur in the construction. Then a group of verbs with negative connotation prompts a deduction of constructional meaning. Consequently, the negative sense of the verbs has transferred onto the entire construction, and the constructional meaning is drawn: the speaker's undesirable appraisal of the result. The following figure illustrates the development path from Type B to C:

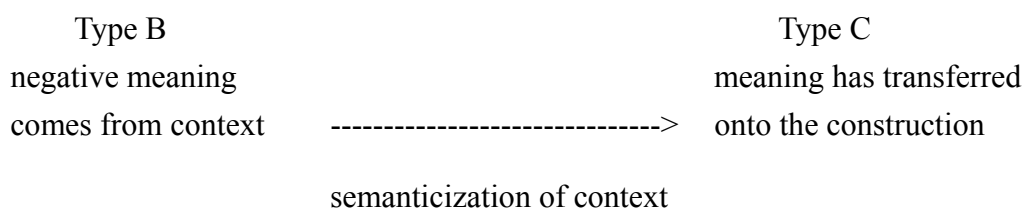


Figure 2 Semanticization of Context and Emergence of Constructional Meaning

4 Conceptual Structure and Selectional Restriction

As the construction develops its polysemy, it comes to be used in increasingly wider contexts. At the beginning it only accommodates physical verbs and denotes physical disappearance. It further proceeds to tolerate less physical verbs and metaphor allows a sense of conceptual disappearance. Finally, it may apply to a variety of stative verbs to express speaker attitude. Nevertheless, in spite of its seemingly free occurrence, some restriction still exists. Consider the following pairs for illustration:

- | | | | | | | |
|-----|----|---------------------------|----------------|--------|-------------|-----|
| (7) | a. | wo | zhengge | ren | sha-diao | le |
| | | I | entire | person | stun-Suffix | CRS |
| | | “I was entirely stunned.” | | | | |
| | b. | *wo | congming-diao | | le | |
| | | I | smart-Suffix | | CRS | |
| (8) | a. | dongxi | langfie-diao | | le | |
| | | thing | waste-Suffix | | CRS | |
| | | “The thing is wasted.” | | | | |
| | b. | *dongxi | jenxi-diao | | le | |
| | | thing | cherish-Suffix | | CRS | |

From the above pairs, it is evident that the V slot does not allow verbs with positive connotation. It seems that the semantics of positive verbs clashes with that of the entire construction. Why is this the case? What is basis of such selectional restriction?

4.1 Metaphorical Basis of Selectional Restriction

Previous studies on Mandarin *-qilai* constructions claim that the development of grammatical units cannot contradict the metaphor that they are based on, and that the collocation of *-qilai* and verbs are conceptually restricted on a semantic basis (Chang 1994, Huang and Chang 1996). My following observation on *V-diao* corresponds to this claim.

I have argued for metaphor as the driving force of semantic change from Type A to Type B *V-diao*. Further, this metaphorical transfer obeys the orientational metaphor GOOD IS UP; BAD IS DOWN (Lakoff and Johnson 1980:16):

Physical basis for personal well-being: Happiness, health, life, and control— the things that principally characterize what is good for a person— are all UP.

The physical and experiential basis for DOWN IS BAD is also evident in our language use and conceptual system. Synchronically, the most basic meaning of *diao* is physical dropping / falling and involves downward movement. It follows that *diao* can relate to something bad in our conceptual system. Be it grammaticalised or not, *diao* should never override the conceptual restriction to modify something good. In other words, if the metaphor DOWN IS BAD is truly at work in the emergence of the construction, it seems rather natural for the construction not to accommodate a verb with positive connotation. Thus, conceptual / cognitive restriction can fully

account for the intrinsic incompatibility of positive verbs in *V-diao* construction.

The above semantic restriction is critical in the development from Type B to C *V-diao*, and without it, the rise of constructional meaning would be impossible. The constructional meaning is language users' generalisation from a previous existing pattern. The constraint must have existed prior to the formation of constructional meaning. Otherwise, without such a selectional restriction, the construction would fail to emerge, since positive verbs would intervene. Therefore, it is justified to say that this constraint metaphorically shapes, or at least helps to shape, the constructional meaning.

4.2 Summary

In this section, the incompatibility of positive verbs and *-diao* is closely examined from a semantic viewpoint. The meaning of *diao* metaphorically constrains the verb types it co-occurs with, which proves the metaphorical nature of our conceptual system. Also, such selectional restriction results in the emergence of constructional meaning. The metaphorical condition on constructional meaning thus reflects the interaction between grammar and human conceptual system.

5 Conclusion

In this study, I have classified *V-diao* constructions according to their semantics, and have explained the constructional meaning. In the second section, metaphorical transfer is argued to be an important mechanism in the development of the construction. Furthermore, I have discussed how pragmatic inference enables language users to arrive at the constructional meaning. Figure 2 shows the different stages of *V-diao* construction and their change of mechanism.

Finally, a selectional restriction on the V slot exists. The exclusion of positive verbs is conceptually conditioned by the semantics of *diao*. This suggests, the semantic change and grammaticalisation process of a grammatical unit, or a construction, is conditioned by human physical and experiential basis. Hopefully, this study may serve as a valid argument for the interaction of our language use and grammar, and the conceptual basis of human language.

TYPE A		TYPE B		TYPE C
physical	<u>(metaphor)</u>	conceptual	<u>(inference)</u>	negative evaluation

Figure 3 Development of *V-diao* and Change of Mechanism

Reference

- Bybee, Joan L., Revere Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: The University of Chicago Press.
- Chang, Shen-ming. 1994. V-qi-lai Constructions in Mandarin Chinese: A Study of Their Semantics and Syntax. M. A. Thesis. National Tsing Hua University.
- Fillmore, Charles J., Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone*. *Language* 64:501-38
- Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
- Goossens, Louis. 1982. On the Development of the Modals and of the Epistemic Functions in English. *Papers from the Fifth International Conference on Historical Linguistics*, ed. by Anders Ahlqvist, 74-84. Amsterdam: Benjamins.
- Heine, Bernd, Ulrike Claudi, and Friederike Hunnemeyer. 1991. From Cognition to Grammar -- Evidence from African Languages. In Traugott and Heine eds., Vol. 1, 149-87.
- Hopper, Paul, and Elizabeth C. Traugott. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.
- Huang, Chu-ren and Shen-ming Chang. 1996. Metaphor, Metaphorical Extension, and Grammaticalization: A Study of Mandarin Chinese -qilai. *Conceptual Structure, Discourse, and Language*. ed., by Adele Goldberg. CSLI.
- Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live by*. Chicago: University of Chicago Press.

Li, Charles, and Sandra Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Los Angeles: University of California Press.

Sweetser, Eve Eliot. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.

心理動詞「想」、「認為」、「以為」與「覺得」的語義區分及訊息表達

--以語料為本的分析方法

巫宜靜

劉美君

國立清華大學

國立交通大學

d898702@oz.nthu.edu.tw mliu@cc.nctu.edu.tw

摘要

動詞語義的分析與表達是處理語言知識的核心課題。藉由對近義詞的探索，有助於了解動詞語義的重要特徵。本研究即以表示心理活動的動詞--「想」、「認為」、「以為」與「覺得」為例，試圖釐清動詞所表達的語法與語義關係。以語料為本的方式，透過詞義區分，我們認為以「模組 屬性動詞語意表徵模式」(MARVS)來表達這組詞彙，可以說明其在句構搭配上的異同，符合一般對動詞語義決定其句法行為的看法，提供比一般詞典所載更豐富的語義訊息。

1. 前言

「想」、「認為」、「以為」與「覺得」這組詞，皆可用以表達說話者對事物的看法，詞典中往往以這組詞來相互定義與解釋，例如《重編國語辭典》(教育部重編國語辭典編輯委員會 1982)、《漢語常用詞用法詞典》(李曉琪等 1997)、《現代漢語詞典》(中國科學院研究所詞典編輯室 1988)等等，這組詞也多半被收錄在同義詞或近義詞詞典中，如《近義詞應用詞典》(陳炳昭 1991)、《同義詞詞林》(梅家駒等 1986)、《現代漢語同義反義詞典》(吳海 1996)等。由詞典中，這組詞所表示的意思可以整理如下：

【認為】：對某一事物經分析思考後所作的判斷。如：我不認為他會出席會議。

【以為】：(1)認為，看成，理解是。如：不以為苦，反以為樂。

(《漢語常用詞用法詞典》：「常和實際情況相反」)

(《近義詞應用詞典》：「『以為』著重於自己認為。多指自己的想法、看法，主觀性強；『認為』指對人或事物確定看法、做出判斷，客觀性較強。」)

(2) 用為、用作。左傳·昭公十五年：司晉之典籍，以為大政。

【想】：(1) 思索、思考。如：想辦法、想不出所以然。

(2) 欲、要、打算、希望。如：想結婚、想出國。

(3) 推測、猜度。如：料想、推想、猜想。

(4) 認為、覺得。如：你想這樣對不對？我想你應該回家一趟。

(5) 思念、懷念。如：想念。

(6) 似、像。

唐·李白·清平調三首之一：雲想衣裳花想容，春風拂檻露華濃。

【覺得】：(1) 感到。如：我覺得有點冷。

(2) 認為。如：我覺得這樣做最好。

(《現代漢語詞典》：「語氣較不肯定」)

(《漢語常用詞用法詞典》：「語氣較輕」)

然而，這組詞，其間的差異除了在抽象的「主觀性」、「客觀性」與語氣上的差異之外，在句法以及語用表現上究竟有何異同，其內部的語意結構、屬性為何，為本文探討的重點。在中研院詞庫小組含有五百萬語料的平衡語料庫中，這組詞出現的筆數分別為「想」6,005筆，「認為」4,132筆，「以為」719筆，「覺得」4,441筆。以下將先區分這組詞個別的語意，觀察具有近似語意的「想」、「覺得」、「認為」和「以為」，在功能分布、參與角色特性、與其他辭彙的搭配情形，最後嘗試以這組詞在 Huang *et al* (2000)所提出之「模組 屬性動詞語意表徵模式」(Module-Attribute Representation of Verbal Semantics, MARVS)的模組來說明這組近義詞的異同。

2. 語意區分

當一個詞可以解讀為一個以上的意思時，這些不同的意思，究竟是這個詞的不同的意義(sense)，或者只是同一個或某幾個意義的不同義面(facet)，向來是語言學家所關切的問題。Ahrens, Chang, Chen & Huang (1998), Lin & Ahrens (2000)為名詞訂出了區分準則，以下我們將參考該標準，嘗試界定動詞「想」、「覺得」、「認為」和「以為」這組詞個別詞項的意義與義面，擷取其語義相近的意義或義面來做比較。

2-1 詞義區分標準 (Ahrens, Chang, Chen & Huang 1998, Lin & Ahrens 2000)

區分個別語義(sense)和義面(meaning facet)的準則：

區分語義:

- (1) 在同一個語境下，不能出現兩個（或多個）個別的語義；
- (2) 沒有別的核心語義可以衍伸出該語義來，或者很難清楚地界定哪一個是核心語義。

區分義面:

- (1) 在同一個語境下，可以出現另一個義面；
- (2) 是核心語義或其他義面的衍伸；
- (3) 語意類別相同的詞，會有類似的語義衍伸出相關的義面來。

2-2 「想」、「覺得」和「以為」的語義區分

「想」、「覺得」、「認為」和「以為」這組詞，除了「認為」之外，「想」、「覺得」和「以為」，都具有多重意思，以下將逐一敘述。

2-2-1 「想」

從平衡語料庫中我們觀察到在現代漢語使用中，「想」的意思可能至少有以下八種，採用以近義詞互釋的方式，可如下表示：

想 1：表「思索」

- 1a. 狼來了！怎麼辦呢？我們一起來看，一起來聽，一起來<想>。
- b. 我就說：「我要怎麼做你才會高興呢？」<想>了一<想>，她說：「要我高興就陪我打麻將。」
- c. 對極了！我<想>不通的問題就在這裡。當你要用錢去賺錢時，就不能用錢去助人。

這類「想」常搭配思索的問題如(1)中各句，也常見思考後所得出的答案由「說」引介出來，如(1b)中之「要我高興就陪我打麻將。」。

想 2表「希望」、「想要」

- 2a. 也許你比較<想>知道的是，今年景氣好不好？
- b. 台灣不少堅強女性，不但要主宰自己的生活也<想>改進社會。
- c. 我不說，你也知道我心裡想什麼嗎？還不是<想>我出人頭地，封妻萌子。
- d. 我<想>你做個溫柔、可愛、聽話的好姑娘，不多嘴多舌。

這類「想」，表示意欲；多半可由近義詞「希望」、「想要」代換。其內容可以由其後緊接著的動詞組顯示，如(2a)和(2b)，亦可由子句表達，如(2c)和(2d)。

想 3表「回憶」、「回想」

- 3a. 夢的最後我似乎有上前去跟妳打招呼聊天的樣子。但是我醒來後卻始終<想>不起來到底有沒有，只記得車窗外，妳我的過去一直在經過。
- b. 有些語言，初聽時微痛，然而，事後一<想>，卻痛不可當，因為說話的人在語言的刀刃上抹上了毒藥。

這類「想」，常搭配「起來」表示回憶一個已知而被遺忘了的事件、訊息，如例句(3a)。事件的已然性質往往可以從語境中得知，如(3a)中的「夢...醒來後」與例句(3b)中的「初聽...事後」。

想 4 表「懷念」

- 4a. 一跟她說話，她就有一點兒想哭。原來這是她第一次離開家，
她 <想> 母親，她 <想> 乾酪肉腸，她 <想> 她的狗。
- b. 謝謝老師。我上一年級的時候，心裡有點兒害怕，雖然在學校上課，可是
一直 <想> 家。老師知道我不習慣，常常走來跟我說話。

這類表示懷念、思念的「想」，其後常緊接著名詞組表示懷念或思念的對象。這個對象帶給懷念者的往往是美好的回憶，而目前卻與懷念者分離。

想 5 表「推測」、「認為」

- 5a. 我 <想> ，不管這世界再怎麼變，做人的原則始終是不會改變的。
- b. 他的妻子嫦娥，看到后羿越來越驕傲，心裡 <想> ：如果讓他吃了不死的藥，那麼人們就要痛苦不堪了。

這類「想」，表達的是經過推測、思索後判斷所得的信念，其後常接子句作為賓語，並且和其句賓之間常有停頓。

想 6 表意外

- 6a. 天保陡然憤恨不已，狠力朝昏暗中想像的可惡男人踢去，不 <想> 踢在一架破風車的支腳上，疼得滋出一身熱汗。
- b. 大學生的兼差工作琳瑯滿目，但你可能沒 <想> 過，人體模特兒現在也成了大學生打工的新出路。
- c. 雖有說不盡的惆悵，我卻放下了懸掛的心。實在 <想> 不到，貝珍不僅深愛著東尼，而且也懂得如何去愛。

這類「想」搭配否定詞「不」與「沒」，形成「不想」、「沒想過」、「想不到」，與「沒想到」等等，表示「出乎意料之外」的語意。

想 7 表「設想」

- 7a. 看電視是很好的休閒活動，可是我們也要為靈魂之窗<想>— <想>，別為了看電視而損壞了它。
- b. 你得替人家<想>— <想>，那生身父母，雖然沒有撫養她，可是他們也有骨肉之情啊！

這類「想」常搭配「為」與「替」引介須顧及而為之「設想」、「著想」的對象，如例句(7a)中之「為靈魂之窗」與例句(7b)中之「替人家」。須顧及的因素往往也伴隨著出現，如例句(7a)中的「別為了看電視而損壞了它。」，與例句(7b)中的「那生身父母，雖然沒有撫養她，可是他們也有骨肉之情啊！」

想 8 表「聯想」

8. 每當我從全自動電腦洗衣機中取出甩乾的衣服時，常使我 <想> 及人生。
乾衣筒祇有在高速旋轉時，才會把水份甩得乾乾淨淨。人生也是。

這類「想」常搭配「及」與「到」，形成「想及」、「想到」等。其所表達的語意是由某一個事件引發另一個事件的成形。

以上「想」的八種意思當中，後三者「想」的語意並非單由「想」可得知，而是在語境中和其他詞彙搭配下產生的結果，如「想 6」搭配否定詞，「想 7」搭配「為」與「替」，「想 8」搭配「及」和「到」等。因此，「想」主要的語義可能是前五種意思，其在語料庫中出現的頻率如表（一），圖示如下頁圖（一）：

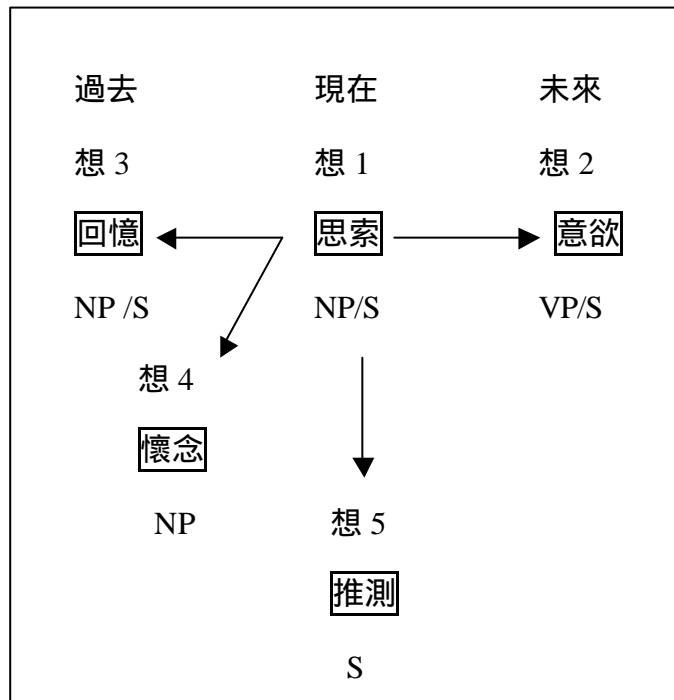
表（一）：「想」出現的頻率

詞項	想 1	想 2	想 3	想 4	想 5
筆數	907	3177	140	73	1504

下圖（一）中假設以表示思索的「想 1」為「想」的中心語義，考慮因素之一

乃是基於「想 1」與其他語義的「想」比較起來，在時間參照上，比較自由。「想 2」，表示意欲、希望，指的常是尚未發生的事情(irrealis)。「想 3」、「想 4」分別表示回憶與懷念，指的常是已經發生過的事情(realis)。而表示推測的「想 5」，和「覺得」、「認為」與「以為」具有近似語義，雖然所需推測的事情可能已經發生，可能尚未發生，但在其後賓語搭配的種類上，不如「想 1」來得多。因此，在頻率上，「想 1」出現的次數並非最多，但基於時間參照時間與後接賓語種類的考量，以「想 1」為「想」的核心語意。

圖(一):「想」主要的語義



2-2-2 「覺得」

「覺得」的意思，至少有以下兩種：

覺得 1 表示「感覺到」

- 9a. 不知不覺船已經出了三峽。李白 <覺得> 眼前一亮，立刻看到天空地闊，江面寬大；果然當天晚上，就到了江陵。
- b. 欣賞他、讚美他、支持他，孩子自然會成長起來，而做母親的人也會 <覺得> 開心。

覺得 2 表示「認為」

10a. 孟子看見人家讀書，就學著讀起書來。孟母 <覺得> 這地方對於孩子很好，才住下來，不再搬家。

b. 我 <覺得> ，一個成功的現代人就是要學習。

表示認為的「覺得 2」，常接子句當賓語，而表示感覺的「覺得 1」較常接動詞組當賓語。上述例句中，「覺得 1」和「覺得 2」的意思似乎清楚可辨，然而在以下例句中，他們的界線卻顯得模糊許多，用「感覺到」或者「認為」來代換，似乎皆可：

10c. 他看到陶師們把用剩的泥土丟棄， <覺得> 十分浪費，於是把這些泥土收集起來，再加淘洗。

d. 四分之一，卻配備了比交響樂團音量大幾倍的打擊樂器，這種比例，一看看就 <覺得> 不合理。

因此，「覺得」可能只有一個語義(sense)，為表示「感覺」、「感到」的「覺得 1」，而與「認為」等具有近義的「覺得 2」為「覺得 1」的語義延伸，為其義面之一。

2-2-3 「以為」

「以為」有三個語義：表示「以之為」、「認為」與「錯誤地認為」，前二者搭配的句式分別為 NP 以及具有詢問語義的疑問句。

以為 1 表「以之為」

語料庫中這類意思的「以為」，均為古文，所接之賓語皆為名詞組：

11a. 這是宋朝歐陽修文忠集裡的詩詞。詞中所提的鷓鴣凌晨先雞而鳴，農家 <以為> 下田之候，俗稱催明鳥。

b. 租界馬路四通，城內道途狹隘；租界內異常清潔，車不揚塵，居之者幾 <以為> 樂土，城內雖有清道局，然城河之水，穢氣觸鼻，僻靜之區，坑廁接踵，較之租界，幾有天壤之別。

以為 2 表「認為」

這類意思的「以為」，常和「如何」等疑問詞搭配，且語義為詢問，如(12a)。出現於直述句時，以表示第一人稱的「我個人」或「個人」為主語如(12b)與(12c)。

12a. 但為了掌握工作狀況，保持人和的政治關係面是應該的工作行為，您 <以為> 如何？

12b. 我個人<以為>，他是屬於比較理性的指揮。

12c. 經過以上七點質疑之後，個人<以為>要重塑城鎮風貌可能性的過程應有如下如下幾項方案，可作為進一步努力的參考...

語料中，含疑問詞「什麼」的問句，亦有使用「認為」替換回答的情形，如(12d)中畫線部分所示。

12d. 古人說的鬼多半指死去的祖先，所以祭拜祖先又叫祭鬼。那麼古人 <以為> 鬼神到底是什麼樣子呢？他們認為鬼神就像法官一樣，處罰做壞事的人...

以為 3 表「錯誤地認為」

在此語義表達，說話者(speaker)認為主事者 (Agent)錯誤地相信某信念。

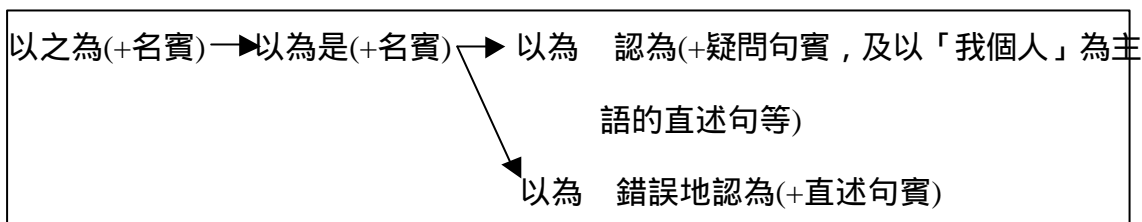
13a. 不了解這個循環，男人和女人都很容易開始懷疑他們的愛。女人會 <以為> 他不再愛她。

13b. 小敏最近常到我這兒來。開始我 <以為> 她是怕我一個人寂寞，來和我做伴兒的。後來發覺事情有點不對頭。她總抱怨這個學期很累。我開始和她玩笑說那是說那是她伺候丈夫太盡心盡意的結果。

例句(13a)中，「以為」的域外論元(external argument)，主事者(Agent)「女人」認為「他不再愛她」為真，而說話者認為非真。例句(13b)中的主語「我」認為「她是怕我一個人寂寞」為真，而說話者認為非真。因為主語和說話者為同一人，故有自我發現實情的「後來發覺事情有點不對頭」語句出現。

「以為」以上三種語義衍伸，可能先有以名詞組為賓語的「以之為」、「以為是」，再分成和「認為」近義，以及和「錯誤地認為」近義的兩個不同的意思的「以為」：

圖(二)：「以為」的語義衍伸



由於大致成互補分布狀態，故「以為」應該有三個語義(sense)：

表(二)：「以為」的語義

<p>「以為」 sense1：「以之為」+名賓(NP)</p> <p>sense2：「認為」+疑問句賓(S[+Q])</p> <p>sense3：「錯誤地認為」+直述句賓(S[-Q])</p>
--

以上我們從語料庫中，觀察「想」、「覺得」、「認為」和「以為」的所表達的意思，逐一列出「想」、「覺得」和「以為」所表達的多種意思，再根據詞義區分標準 (Ahrens, Chang, Chen & Huang 1998, Lin & Ahrens 2000)，將「想」分成五個語義：「想 1」表「思索」；「想 2」表「意欲」；「想 3」表「回憶」，「想 4」表「懷念」，「想 5」表「推測」、「認為」。並且將「覺得」所表達的兩個意思「感覺到」和「認為」，視為同一個語義下的兩個義面。最後將「以為」區分成三個語義：「以為 1」表「以之為」，「以為 2」表「認為」，而「以為 3」表「錯誤地認為」。如此，我們可以清楚地界定和「認為」具有近似語意者為「想」的第五個語義、「覺得」的第二的義面，以及「以為」的第二個語義。

由於「覺得」的第二個義面，如上所示，有時不易和另一個義面明確區分開來，且「以為」的第三個語義「錯誤的認為」亦和「認為」相關，因此在以下的觀察與探討中，也同時涵蓋了「覺得」整個語義的兩個義面，以及「以為 2」和「以

為 3」兩個語義。

3. 觀察

在此部份我們將觀察這組近義詞--「想」,「覺得」,「以為」與「認為」,在功能分布上和參與的角色特性上,以及與其他句式或辭彙搭配上的異同。

3-1 功能分布

在功能分布上,如下表所示,這組詞在句中皆多作為述語(predicate)功能。除了「認為」有極少數的名物化情形之外,其餘三個詞皆無名物化用法。和關係句搭配作為名前修飾語的情形亦不多見。「想」和「以為」能帶少數的名詞組當賓語,而「認為」和「覺得」不能帶名詞組作為賓語。此外,「想」和「認為」後接句賓與述賓筆數的百分比,差量較大。

表(三):「想」、「認為」、「以為」與「覺得」語法功能分布比較

		想	認為	以為	覺得	
做述語	總數	(99.1%)	(99.98 %)	(98 %)	(100%)	
	後接賓語	總數	(89.5 %)	(98%)	(99 %)	(100%)
		後接句賓	(78.9%)	(84.5 %)	(73.2 %)	(64.3 %)
		後接述賓	(10.2%)	(13.5 %)	(26.4 %)	(35.7 %)
		後接名賓	(0.4%)	0	(0.4 %)	0
	後不接賓語	(10.5 %)	(1.98 %)	(1.0 %)	0	
做修飾語		(0.9%)	0	(2.0 %)	0	
名物化		0	0.02	0	0	

3-2 參與角色特性

依照中文詞知識庫(1993)的分類,「覺得」屬於狀態句賓述詞(VK1),所帶論元角色為經驗者(Experiencer)終點(Goal)之外,「想」、「認為」和「以為」皆為屬於動作句賓述詞(VE2),帶有主事者(Agent)與終點(Goal)兩論元。在搭配的主語

方面，「想」的主語為[有生] ([+animate])的主事者(Agent)，而「認為」的主語可為[無生]([-animate])、[表見解]([+opinion])的主事者。「以為」的主語亦多為[有生]的主事者，但可有[無生]表示立場等作為主事者(Agent)。

表(四)：論元結構比較

詞項 類別	想、認為、以為	覺得
動詞類別	VE2	VK1
論元結構	Agent<*<Goal	Experiencer<*<Goal

表(五)：「想」、「認為」與「以為」的主事者(Agent role) 屬性比較

詞項 論元屬性	想	認為	以為
主事者	[有生]	[有生]； [無生,表見解]； [有權威]	[有生]； [無生, 表立場]

「認為」的主語可帶[有權威] ([+authority])，而「想」的主語比較不含此特徵。其比較，如下表(六)。「想」有近九成的主語，皆為人稱代名詞，而「認為」以人稱代名詞為主語者，不到一成。語料中「想」只有一筆出現主語為「音樂家」，屬於有特殊專長者，然而「音樂家」和「檢察官」、「醫師」等比較起來，似乎「檢察官」、「醫師」等在權威性方面稍高一籌。此外語料中亦顯示「想」以機構、場所為主語者，多為「學校」與「家裡」，其權威性更是遠遠不及「認為」可帶主語如「總統府」、「法院」、「教育局」等等。「認為」除了能如以下(14)中畫線部分以權威人士為主語之外，猶能以「觀念」、「觀點」、「看法」、「輿論」等為主語，分別如以下(15)至(17)句中所示。顯示「想」的主語須為[有生] ([+animate])，而「認為」的主語可以為[無生]([-animate])、[表見解]([+opinion])。

表(六)：「想 5」與「認為」主語比較

主語類別		範例	認為	想 5
人	人稱	我、我們、你、你們、他、她、他們...	9%	89.36%
	一般人	小朋友、孩子、學生、兒子、母親、家長、女性、讀者、觀眾、工作者、工作者、經營者、百姓、人名.....	35.5%	7.98%
	有特殊專長、地位之人	學者、專家、學家、教授、醫師、檢察官、縣官、黨工、官員、職稱，職稱+人名	29%	0.07%
業界		學界、電腦業、..類、警方、校方	7%	0
機構、組織、場所		政府、總統府、黨、黨中央、機關、機構、陸委會、署、司、鎮公所、教育局、縣市、法院、學校、學生會、醫院、雜誌、社會、家裡	13.5%	0.13%
思想、意見		觀點、看法、輿論	5.5%	0
其他			0.5%	2.46%

- (14) 以色列參謀總長蕭姆隆認為，伊拉克動用化學武器的時刻已越來越接近。
- (15) 血壓的控制什麼程度才算理想？而什麼程度才需長期服藥呢？目前觀念認為早晨睡醒時與下午二時到四時所測的血壓，若兩天所測的血壓值，均發現收縮壓超過一五毫米水銀柱或是舒張壓超過一毫米水銀柱，就需要開始治療。
- (16) 語言派：第一個派別將社會建構等同於語言活動，這個觀點認為所謂的男女特質是由男女在其主體位置的確認過程中所決定的，而語言活動在這確認過程中起著主要作用，
- (17) 自從黃加進發現澎湖古沉船後，政府相關部門受到很大壓力，民間輿論認為，

沉船為重要文化資材，教育部明知「國寶」就在自家門口，卻還不積極「搶救」，實在有失職守。

故，在語用上，「認為」多搭配較具有權威性的人士、機關與職稱。並且，可由下表(七)中「認為」常和法相副詞搭配，看出「認為」對事件的非實然(irrealis)要求較高。

表(七)：「認為」後接賓語類型

「認為」後接賓語類型			百分比%
句型 (23.50%)		疑問句	3.0
		比較句型	9.0
		假設句	3.0
		因果句	0.5
		轉折句	2.0
		評價句	4.5
		事實陳述句	1.5
含副詞 (71.5%)	法相副詞 (52.50%)	[+epistemic]	32.5
		[+deontic]	20.0
	其他副詞 (19%)	評價副詞	3.5
		程度副詞	11.0
		時間副詞	3.5
	數量副詞	1.0	
其他(5%)			5.0
總計			100.0%

3-3 搭配限制

在體(aspect)和狀語的搭配上，亦各有不同。分別如下表(八)至(十)所示（星號「*」表示該搭配方式不被接受，問號「？」表示接受度存疑）。

表(八)：與體之搭配關係比較

動詞 體	想 5	認為	覺得	以為
了	*想了	*認為了	覺得了	*以為了
過	想過	*認為過	*覺得過	*以為過
著	想著	*認為著	*覺得著	*以為著
一直	一直想	一直認為	一直覺得	一直以為
在	在想	*在認為	*在覺得	*在以為

表(九)：與表示反覆、多次等句式之搭配關係比較

動詞 句式	想 5	認為	覺得	以為
V 越 V	想越想	*認為越認為	*覺得越覺得	*以為越以為
V - V	想一想	*認為一認為	*覺得一覺得	*以為一以為
再 V	再想	*再認為	*再覺得	*再以為
不只一次地 V	不只一次地想	*不只一次地認為	*不只一次地覺得	*不只一次地以為

表(十)：與狀語搭配關係比較

動詞 狀語	想 5	認為	覺得	以為
一路	一路想	*一路認為	*一路覺得	*一路以為
一致	*一致想	一致認為	一致覺得	一致以為
一貫	*一貫想	一貫認為	*一貫覺得	*一貫以為
主觀地	?主觀地想	主觀地認為	*主觀地覺得	主觀地以為
直覺地	?直覺地想	直覺地認為	*直覺地覺得	直覺地以為

此外，只有「想 5」能帶「得」字補語。而在與否定詞「別」的搭配方面，「以為」常與「別」搭配，能和「別」搭配的為表示「錯誤地認為」的「以為 3」。表示「認為」的「以為 2」亦不能和「別」共現。「覺得」不可和「別」搭配，「想 5」

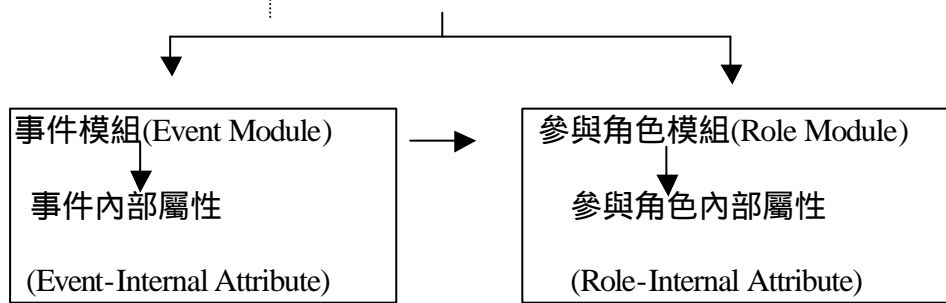
與「認為」亦通常不和「別」搭配使用。

4. 動詞語意表達式--MARVS

關於動詞語義訊息的表達，Huang *et al* (2000) 提出了「模組 屬性動詞語意表徵模式」(Module-Attribute Representation of Verbal Semantics, MARVS)，主張每一詞義即代表一個特殊的事件結構，具有其獨特的表徵內涵。此表徵模式將動詞語意分為兩個模組，事件模組(event module)與角色模組(role module)。各模組之下則含括細部的屬性界定。其基本架構如下：

圖三：模組--屬性表達式

動詞 – 語義(Sense)_i – 事件訊息(Eventive Information)



在 MARVS 中，動詞所能表達的事件類型主要由五種基本事件模組組成，用以描述事件的過程(process)、狀態(state)、階段(stage)，事件的長短，以及事件的端點(boundary)，其表達方式如下表(十一)所示：

表(十一)：五個基本事件模組

表達方式		/////	^^^^^	—	/
事件模組	端點 (boundary)	過程 (process)	階段 (stage)	狀態 (state)	瞬時 (punctuality)

只能呈現一種單純事件形態的動詞，為表示核心事件(nuclear event)的動詞，可以同時指涉事件端點的，為表示簡單事件(simples event)動詞，而可指涉過程或過程完成後的狀態的動詞，為表示複合事件(composite event)的動詞。

5. 解釋

於 MARVS 架構中的觀察，顯示「認為」、「以為」和「覺得」在事件類型中，較可能為表示狀態的動詞，其中「覺得」帶有起始點，而「想 5」比較可能為無端點的表示過程的動詞。

表(十二)：「想」、「覺得」、「認為」和「以為」在為 MARVS 中的表達式

詞項	「想 5」	「覺得」	「認為」	「以為 2」	「以為 3」
事件模組	////	_____	_____	_____	_____
事件內部屬性	[-控制] 說者報導 主事者 的信念	[-控制] 說者報導 經驗者的 信念	[-控制] 說者報導 主事者的信 念	[-控制] 說者報導 主事者的 信念	[+控制] 說者指出 主事者的信念 有誤
參與角色 模組 與 參與角色 內部屬性	主事者 [有生]	經驗者 [有生]	主事者 [有生] [無生,表意見] 終點： [表確定]	主事者： 第一人稱 終點：[非表疑問] 主事者： 第二人稱 終點：[表疑問]	主事者 [有生] 終點 [非表疑問]

若如上表(十二)所示，「想 5」為不帶端點，單純表示過程的動詞，則可以說明上表(八)中，「想 5」可以不和「了」搭配，而和表示事件正在進行的「在」與表示持續的「一直」和「著」搭配。同時也可以說明上表(九)中，只有「想 5」可以和表示反覆、多次的狀語搭配。「覺得」為帶有起始端點，表示起始狀態的動詞，可以說明上表(八)中，這組詞中，只有「覺得」可以和「了」搭配使用。再者，若只有「以為 3」帶有[+控制]([+control]) 的屬性，則可以說明這組詞中，只有「以為 3」能和「別」搭配出現。「認為」的角色「終點」(Goal) 為若帶有[表確定]([+affirmation]) 的徵性，則吾人便不難預測在語用上為凸顯其確定性，搭配的主事者多半是具有某方面權威的人士、機關與職稱。

此外，透過事件屬性中「說者報導主事者(Agent)/經驗者(Experiencer)的信念」與「說者指出主事者(Agent)的信念有誤」，可以區分出「想」、「覺得」、「認為」和「以為」這組詞的語意相似處，同時也可以區分出「以為」的細部語意差異來。否則我們可能不容易理解詞典中所載「以為」具有「認為」的語意，但卻又「常和實際情況相反」。再者，如果能以參與角色模組搭配角色內部屬性來表徵辭彙語意，而將「以為₂」中不同屬性的主事者搭配不同屬性的對象，則辭典中所述「『以為』主觀性較強」可更清楚地表達出來，同時也呈現了「以為」不止能夠搭配第一人稱，也可以搭配第二人稱的實際使用現象。

6. 結論

「想」、「覺得」、「認為」以及「以為」這組在詞典中常用以互相釋義的近義詞，其所表達的意思與實際使用上的異同處，並不容易從既有詞典所載中得知。由以語料為本的方式，從含有大量語料的語料庫中，觀察「想」、「覺得」、「認為」與「以為」這組近義詞的所表達的意思以及出現的語境，我們區分出「想」有五個語義，「覺得」有兩個義面，「以為」有三個語義，而「認為」只有一個語義。具有相近意義者為「想₅」、「覺得」、「認為」和「以為₂」。由他們搭配詞語以及句法表現，以「模組 屬性動詞語意表徵模式」(MARVS)表達，可以讓我們對這組詞彙語義的訊息與樣貌，得到更多的認識與瞭解。在 MARVS 中，「想₅」為表示過程(process)的動詞，「覺得」為表示起始狀態(inchoative state)的動詞，「認為」和「以為」皆為表示均質狀態(homogenous state)的動詞。表示「錯誤地認為」意思的「以為₃」和表示「認為」的這組近義詞，不同之處在於後者為「說話者旨在報導主事者或經驗者的信念」，而前者為「說話者旨在指出主事者或經驗者的信念為誤」。說話者對主事者或經驗者信念的看法，在此扮演極為重要的語義區隔。透過 MARVS 對事件模組、事件屬性，角色模組以及角色屬性等分層的表達方式，可以對這組詞的語義、句法行為以及語用上，得到較為合理的解釋與預測。符合詞彙

語義學家(Pustejovsky 1995, Levin 1993, Atkins 等 1988) 對「動詞句法行為，特別是論元表達，決定於動詞語義」的共同看法。

中文參考書目：

中文詞知識庫 1993 《中文詞類分析》(技術報告 93-05) 台北：中央研究院。

中國科學院研究所詞典編輯室 1988 《現代漢語詞典》3 刷 香港：商務印書館香港分館。

王克仲，孫秉德，房聚棉 1993 《古今詞義辨析詞典》 哈爾濱：黑龍江人民出版社。

吳海主編 1996 《現代漢語同義反義詞典》 北京：學苑出版社。

李曉琪等編 1997 《漢語常用詞用法詞典》 北京：北京大學出版社。

張麗麗，陳克健，黃居仁 2000 <漢語動詞詞彙語義分析：表達模式與研究方法>《中文計算語言學期刊》 Vol. 5.1.pp. 1-18.

教育部重編國語辭典編輯委員會 1982《重編國語辭典》 台北市：臺灣商務印書館。

梅家駒，竺一鳴，高蘊琦，殷鴻翔 1986 《同義詞詞林》2 刷 香港：商務印書館香港分館。

陳炳昭 1991 《近義詞應用詞典》2 刷 北京：語文出版社。

英文參考書目：

Ahrens, Kathleen, Li-Li Chang, Keh-jiann Chen, Chu-Ren Huang. 1998. "Meaning Representation and Meaning Instantation for Chinese Nominals," *Computational Linguistics and Chinese Language Processing*, 3, pp. 45-60.

Lin, Charles Chien-Jer, and Kathleen Ahrens. 2000. "Calculating the Number of Senses: Implications for Ambiguity Advantage Effect during Lexical Access," *Seventh International Symposium on Chinese Languages and Linguistics Proceedings*, pp.

141-156.

Atkins, B.T., J. Kegl, and Beth Levin. 1988. "Anatomy of a Verb Entry: from Linguistic Theory to Lexicographic Practice," *International Journal of Lexicography I*, pp. 84-126.

Huang, Chu-Ren, Kathleen Ahrens, Li-Li Chang, Keh-Jiann Chen, Mei-Chun Liu, and Mei-Chih Tsai. 2000. "The Module-Attribute Representation of Verbal Semantics: From Semantics to Argument Structure," *Computational Linguistics & Chinese Language Processing*. Vol. 5.1, pp. 19-46.

Levin, Beth. 1993. *Verb Classes and Alternation*. Chicago: University of Chicago Press.

Liu, Mei-Chun, Chu-Ren Huang, Charles Lee, Ching-Yi Lee. 2000. "When Endpoint Meets Endpoint: A Corpus-based Lexical Semantic Study of Mandarin Verbs of Throwing," *Computational Linguistics & Chinese Language Processing*. Vol. 5.1, pp 81-96.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge: The MIT Press.

網路資源：

中央研究院現代漢語平衡語料庫 <http://www.sinica.edu.tw/ftms-bin/kiwi.sh/>

教育部國語推行委員會重編國語辭典修訂本 <http://www.edu.tw:81/mandr/>