中文文本人名辨識問題之研究

李振昌　李御璽　陳信希

國立台灣大學資訊工程學研究所
E-mail: hh_chen@csie.ntu.edu.tw

摘要

現今的中文斷詞系統已有不錯的成果，但是仍然存在著一些問題。專有名詞的辨識即是其中之一。本論文整合法則式與統計式的方法來解決其中人名辨識的問題。我們將人名分成中式人名與音譯人名兩大類並詳細討論其所採行的策略。在大規模不同新聞版面的實驗中，我們得到平均81.46%的精確率與91.23%的召回率。

1. 緒論

中文斷詞是一個存在已久的問題。迄今已有許多文章被發表在國內外各個重要的會議中並且也都有不錯的正確率。綜觀而言，斷詞系統可略分為兩大陣營：法則式與統計式。法則式的作法優點在於所需的空間較小。而其缺點則是執行的速度較慢，同時有些規則可能會互相衝突。陳克健教授(1992)所提出的作法，其斷詞的結果得到很好的正確率。相對的，統計式的作法優點是不需大量的人力介入且其執行的速度也較快。而其缺點則是大量的語料取之不易，其統計資料也相當佔空間。此種作法也同樣得到很好的正確率(Chang, *et al.*, 1991; Chiang, *et al.*, 1992; Lin, *et al.*, 1993)。雖然斷詞的正確率很高，然而專有名詞的辨識問題仍是其瓶頸的所在。專有名詞與其他名詞不同的地方在於專有名詞容易隨著時間的改變而新增與刪減。因為專有名詞的數量極多，因此我們無法建構一個辭典來容納所有的專有名詞。

不同的專有名詞有著不同的特性。若要解決此一問題，針對不同型態的專有名詞設計不同的方法應是較為有效的作法。我們優先考慮人名，因為在專有名詞中人名佔了最大宗。就人名而言，可以分成兩大類：中國人名與外國人名。因為除了一些亞洲國家人名與中國人名類似之外大部份的外國人名皆是音譯人名，所以我們再將它重新分成兩大類：中式人名與音譯人名。雖然都是人名，但是音譯人名有著與中式人名截然不同的特性，因此需要分開來討論。本文不僅是要將這兩類的專有名詞找出來，同時還要賦予它們適當的類別。亦即中式人名者若被視為音譯人名則算是錯誤。

同樣的，若是組織名被視為人名，我們也視其為錯誤。在未來的中文處理上，這樣的處理將是非常有用的。

2. 語料庫

由於我們的系統是整合法則式與統計式的作法，因此除了一些經驗法則外也需要一些統計的資料。這一節將介紹我們所使用的語料庫。第一個語料庫是一個平衡分佈的中文語料庫。這是依據LOB英文語料庫的格式來建立的。它包含了113,647個詞，合191,173個字，是一個平衡而且以人工校定過的的中文語料庫。第二個語料庫則是從報紙所抽取出來的，共含約二百六十萬個字。由於只經過初步斷詞再加上報紙中的專有名詞數量非常多，因此這個語料庫斷詞上的錯誤會比預期的高。第三個訓練語料庫是一個人名的語料庫，總共包含了219,738個中式人名，合661,512個字。第四個訓練語料庫則是一個音譯人名的語料庫。它是由一本書"洋名洋名任你選"(黃玉眞, 1992)中取出來的，共有2,692個音譯人名(其中有1,414個男性的名字與1,278個女性的名字)。

要知道一個系統的實驗結果好壞與否，合適的測試資料是很重要的。我們從自由時報的語料庫中隨機選取了六個檔案。每一個檔案代表一個不同的版面。這個自由時報的語料庫與我們第二個訓練的語料庫是不同的。詳細的統計資料如下所示：

檔案一： 政治版(有許多與立法院有關的新聞)
共有23,695個詞，合36,059個字
檔案二： 社會版(有許多與警員，罪犯有關的新聞)
共有61,846個詞，合90,011個字
檔案三： 娛樂版(有許多與電視、電影、明星有關的新聞)
共有38,234個詞，合55,459個字
檔案四： 國際版(有許多與國際事件有關的新聞)
共有19,049個詞，合29,331個字
檔案五： 經濟版(有許多與股市、銀行有關的新聞)
共有39,008個詞，合54,124個字
檔案六： 運動版(有許多與運動的新聞)
共有36,971個詞，合54,124個字

從這六個版面來看，由於含有大量的專有名詞且類別不盡相同，因此會有許多不同的狀況發生。如此複雜的測試資料應能顯示出一個相當客觀的實驗結果。

3. 中式人名的辨識

3.1 定義與分析

中式人名的辨識簡單來說就是將文章中所有的中式人名都找出來。中式人名是由兩個部份所組成：姓與名。大部份的姓是一個字，名是兩個字。少部份的姓是兩個字至四個字，名也有少部份是一個字。中式的姓可分為下列三種型態：

型態一：單姓，如"趙"、"錢"、"孫"、"李"。
型態二：複姓，如"歐陽"、"上官"。
型態三：冠夫姓，如"蔣宋美玲"中的"蔣宋"。

姓的特性是並非所有的中文字都可以成為姓，只有少部份的字可成為姓。在我們的資料系統中，只有399個字被視為姓。型態三的姓氏是一種較特別的姓。一般而言，姓氏是一個字或兩個字。但因有型態三的姓氏，使得姓氏的字數可能高達四個字。型態三的姓氏通常是屬於女性所有。這是一個相當重要的線索，可以避免許多不必要的錯誤。

名字的結構相較於姓就顯得較為簡單。它只有一個字或者是兩個字。但是它與姓不同之處在於所有的中文字都有可能成為名字，因此其複雜度並不亞於姓。在中式人名中唯一較特別的線索便是姓氏。然而並非所有的字皆可為姓氏，所以我們便假設當有一字是單字詞且可為姓氏時，則此字便有可能是一個中式人名的開端，然後再依其他線索判斷此一候選者是否成立。

3.2 作法回顧

在此之前，已有一些相關研究發表在國內外會議中(Chang, *et al.*, 1991; Wang, *et al.*, 1992)。張等人的作法是將中式人名的辨識問題加入斷詞系統中一併處理。而另一作法則是以頭銜為最重要的線索。本節將簡單介紹一下張等人的作法。

他們是將中式人名的辨識問題加入斷詞系統中，並以斷詞的概念來賦予每個中式人名的候選者一個機率值。再利用中式人名的候選者與其他可能的詞組來互相競爭，選出機率值最高者為其結果。簡單來說，中式人名的候選者被視為一詞，而此一詞的機率值是由另一公式所提供與一般詞的機率值來源不同。此種作法的優點是一氣呵成。將斷詞與中式人名的辨識同時完成，但前提是需要兩個能互相匹配的訓練語料庫。然而在他們作法中忽略了型態三的姓氏問題。亦即冠夫姓並不在討論的範圍之內。根據他們在後續的文章(彭, 張, 1993)中所發表的數據(測試資料共有3,500個詞)：

姓氏的精確率與召回率分別是86.7%與96.3%，而名的精確率與召回率則分別是78.3%與85.5%。

3.3 我們的策略

3.3.1 前處理(Preprocessing)

我們的基本假設是：中式人名除非在辭典中有出現，否則大都會被斷成一連串的單字詞。一般而言，文章中不應有太長的單字詞串。所以每當文章中有一連串單字詞出現時，很可能就是斷詞錯誤發生的地方，此處就有可能有一個中式人名。因此我們的基本作法是：每當遇到一個單字詞，若同時此一單字詞又可當作姓氏，則使用隨後的策略來判斷是否為中式人名。因此斷詞在我們的系統中是必要的前處理。此與張等人的作法有所不同。

3.3.2 字的變異性(Variance)

本系統最重要的部份就是去估計一候選者的成績。由(彭, 張, 1993)可以得到下列公式：

$$P(W,GN) = P(GN) * P(W|GN)$$

P(GN)表示姓氏出現的機率，而P(W,GN)則表示W這個字當姓氏的機率。由於他們是以斷詞的角度來看這個公式與我們的作法不同，因此我們不能全盤採用。但也因此得到一靈感。

我們的作法是每當有單字詞可成為姓氏時，便取後面連接的兩個字將其視之為名。然後用一公式來估計其成為中式人名的成績。若高於某一標準，則視之為中式人名。因此，我們需要一個能確實反映出一個字能成為一個姓氏或名字的成績的公式。所以我們就將上一公式加入另一項參數，考慮一個字的使用性。舉例來說，"傳"與"暉"這兩個字在我們的訓練語料庫中成為名字的次數一樣多。若前述公式不加修改的話，則此二字的成績將會是一樣的。但是很明顯的"傳"這個字的使用性比"暉"大得多(如常見的"傳染"、"傳播"、"傳話"等詞)。而"暉"這個字的使用性就相對小得多(只有一個較罕見的詞"暉映")。換句話說，這兩個字的變異性不相同("傳"比"暉"大得多)。雖然此二字當名字的次數一樣多，但是它們當名字的強度應不相同。若以成績而言，"傳"應較低，而"暉"應較高。所以我們就對每一個字的成績除上述的公式外，再除以這個字的一般用法的次數，來求得一個較為合理的成績。

3.3.3 基本模型(Baseline Model)

這一節將列出我們所採用的基本模型。經由前述的討論，我們修改其公式得到下列模型：

$$P(C_1)*P(C_2)*P(C_3) \tag{1}$$

$$\Rightarrow P(C_1)*\frac{1}{\&C_1}*P(C_2)\times\frac{1}{\&C_2}*P(C_3)*\frac{1}{\&C_3} \tag{2}$$

$$\approx\frac{\#C_1}{\&C_1}*\frac{\#C_2}{\&C_2}*\frac{\#C_3}{\&C_3} \tag{3}$$

公式(1)是根據前述的公式將三個字分別成為姓氏與名的機率值相乘。然後於每個字的機率之後考慮每個字的變異性，便演變成為公式(2)，再加以簡化成為公式(3)。其中，$\#C_i$表示C_i這個字當姓氏或名的次數，$\&C_i$表示C_i這個字當其他用途的次數。根據公式(3)以及每種不同型態姓氏的特性，可以得到下列不同的模型：

模型一：針對型態一的姓氏

$$\frac{\#C_1}{\&C_1}*\frac{\#C_2}{\&C_2}*\frac{\#C_3}{\&C_3}>Threshold_1 \tag{A}$$

$$\frac{\#C_1}{\&C_1}>Threshold_2\ \&\ \frac{\#C_2}{\&C_2}*\frac{\#C_3}{\&C_3}>Threshold_3 \tag{B}$$

模型二：針對型態二的姓氏

$$\frac{\#C_2}{\&C_2}*\frac{\#C_3}{\&C_3}>Threshold_4 \tag{A}$$

模型三：針對型態三的姓氏

$$\frac{\#C_{11}}{\&C_{11}}*\frac{\#C_{12}}{\&C_{12}}*\frac{\#C_2}{\&C_2}*\frac{\#C_3}{\&C_3}>Threshold_5 \tag{A}$$

$$\frac{\#C_{11}}{\&C_{11}}*\frac{\#C_{12}}{\&C_{12}}>Threshold_6\ \&\ \frac{\#C_2}{\&C_2}*\frac{\#C_3}{\&C_3}>Threshold_7 \tag{B}$$

每個不同的模型是針對不同的姓氏型態。其中，模型二是針對複姓的部份。由於複姓部份並不如單姓一般容易出現，且像"歐陽"、"上官"這種詞除了當姓氏之外並不太可能會成為別的詞，所以我們在此省略姓氏的成績，而直接以後面的字當名字與否來決定。

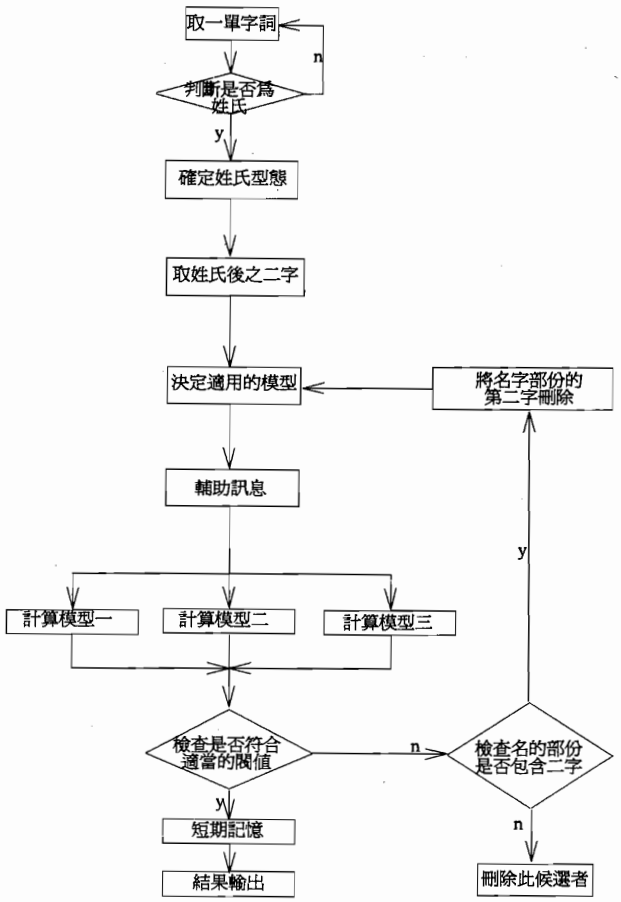
以上各個模型所討論的都是名字部份是兩個字的情形。對於單名，我們的處理方式是：每當有中式人名的候選者其成績低於標準時，我們將其名字部份的第二個字刪除掉，再以類似的模型來估計其成績。也就是說，不考慮名字第二個字的影響。而此時的閥值與原先的閥值不同。之所以先

考慮名字為兩個字的情形，是因為單名的比例較雙名來得低，自然應該優先考慮雙名的情形。

3.3.4 閥值(Threshold)

閥值的決定方式是將我們的人名語料庫中的每個人名，以我們的模型來計算其成績。選取一值能使我們的人名語料庫中，99%的人名成績皆大於此值。凡是小於此數字者，皆視為不合格。

對於模型一與模型三而言皆有兩個式子。其原因是為了避免一些姓氏的候選者的影響力太大，而造成一些不必要的錯誤。例如"陳踢了王一腳"這個例子。"陳"這個字當姓的次數很多，因此當姓氏的成績也是很高。雖然"踢了"這兩個字當名字的次數少且其成績也很低，但是因為"陳"的成績太高，因此會將"陳踢了"視為人名。為了要防範這種錯誤，所以我們就將姓與名的成績分開來考慮，以避免因為某一方的成績太高而互相影響。



圖一、 辨識流程

除了基本的統計模型外，文章中還有許多輔助訊息可以使用。我們將這些訊息分成三個層面來探討。第一級是字的層面，用到了一些性別

(gender)與相互訊息(mutual information)的訊息。第二級是句子層面，用到了一些標點符號(punctuation marks)與頭銜(title)的訊息。第三級是段落層面，用到了短期記憶(cache memory)的觀念。整個中式人名的辨識流程如圖一所示。其所使用的輔助訊息在下面幾節中將一一說明。

3.3.5 性別(Gender)

中式人名有一種特殊的習慣，有些名字的姓氏是由兩個姓氏所組成的。例如"許林鹽梅"這個名字中，"許"與"林"都是常見的單姓。這種姓氏與我們一般所稱的複姓是不同的。複姓是一個單獨存在的姓氏，而前述的情形則是由兩個姓氏所組成的。此種情形一般是出現在結過婚的女人身上。因為當女人結過婚後通常會將其丈夫的姓氏加在自己的姓氏之前，所以若要解決此一問題，性別則是一個很大的關鍵。

前述此種姓式是由兩個姓氏所組成的。然而是否兩個姓氏在一起就是此種情形呢？有許多的反例。例如"毛高文"、"羅張"等，其姓名的前二字皆可成為姓氏卻不是上述的情形。所幸此種姓名的特性是專屬於女性所有。我們只要先檢查此一候選者的性別，便知道其是否應屬冠夫姓的姓名。例如"毛高文"便屬於男性，他就不會被當成型態三的姓名處理。

中國人取姓名有一習慣，男性與女性的名字形式不同，用字更是不同。因此造成有些字絕少在男性的名字中出現。同樣的，女性也有類似的情形。例如下列的一些字：

屬於男性的字：豪、霸、宏、志、強、正、昌、光、輝、雄
屬於女性的字：佩、月、玉、如、秀、佳、怡、芬、芳、女

上述屬於男性的字是一些陽剛性較強的字，較常在男性名字中出現。而屬於女性的字是一些較陰柔的字，較常在女性人名中出現。由此點我們得到一個構想：統計每個字出現在男性人名與女性人名中的次數，再利用這些數據去計算人名的名字部份當男性與女性的次數，以次數多者為依歸來決定性別。

3.3.6 由雙字詞組成的候選者

中式人名的候選者並非永遠是由一串單字詞所組成，有時候也可能是由非單字詞所組成的。例如"陳建中的功課永遠是那麼好"的例子中，"建中"就是一個詞，而此處"陳建中"正是一個中式人名。然而根據觀察，當名字的部份是由詞所組成的時候會有一些例外的情形。例如"陳家世清白，絕不會犯法..."的例子。若以基本模型的作法，則"陳家世"將會辨識成一個中式人名，而且也是非常合理的人名。但問題在於"陳"是一個人名的簡寫，

名字的部份已被省略了，"家世清白"是另有用途。所以若是考慮上下文，則"家世"不應是一個人名。

就原先的假設而言認為一般的中式人名將會被斷詞成為一串單字詞。而這一串單字詞通常是不應以單字詞的方式存在，所以才會使用統計式的模型去猜那些單字詞可成為一個中式人名。而今有些人名不是被斷成單字詞，而是以雙字詞型式存在。若依原先的假設，則應先檢查此一雙字詞是否應存在。若不應存在於此，則可視為與原先的被斷成單字詞時的狀況相同，我們就不做任何特殊處理。反之若此雙字詞恰巧應在此出現，則原先的假設不符，須做另外特殊的處理。

如何能知道一雙字詞是否應存在，是此一問題的關鍵。所幸前人已有許多在此方面的研究(Church and Hanks, 1990)可供參考。我們採用了相互訊息(mutual information)這個公式來判斷兩個詞在訓練語料庫中同時出現的機會有多大。其公式如下：

$$MI(W_1, W_2) = MI(W_2, W_1) = \log_2 \frac{P(W_1, W_2)}{P(W_1) \times P(W_2)}$$

3.3.7 標點符號(Punctuation Marks)

根據我們的觀察，人名出現在句首或句尾的機會比較大，因為人名常為一個句子的主詞或受詞，而且在句首或句尾時也有助於人名的邊界辨識。例如"...焦仁和。"這個句子。以一般的情形而言，"焦仁"會被辨識為人名。但是此時"焦仁"後面還有一個字"和"。一個人名與句尾間有一個字存在是一件很奇怪的事情，所以"和"這個字理應被括入人名中。但若是採用規則的方式來強制規定此字必屬於此人名所有也不太合理，因此我們採折衷的方法。我們給予此字額外的加分，由統計模型決定其是否能因此被辨識為人名。

另外，頓號是一個非常有用的符號。頓號兩邊的詞，其性質是一樣的。例如"林亦宏、王為墜、莊朝焰..."中的三個人名就是用頓號隔開的。由於頓號的這種用法，使我們可用頓號來挽救一些成績極低的候選者。例如上例中，"王為墜"這個人名，其名字部份的成績極低。所以若非有頓號存在，這個人名是不會被找出來的。所以當頓號出現時，若在頓號的另一邊是人名時，則此一候選者的成績應加一些分數。

3.3.8 頭銜(Title)

頭銜是一個非常有用的訊息。有些作法就是以頭銜為主要的訊息來辨識中式人名(Wang, *et al.*,1992)。通常頭銜的前面或後面會有人名伴隨出現。例如"李登輝總統"就是人名與頭銜一起出現的例子。但這並不表示有頭銜的存在就一定會有人名伴隨在一起。例如"向總統報告"則是一個典型的例子。同時頭銜也有可能出現在人名之前。例如"總統李登輝"。雖然如此，頭銜仍有一個很大的好處：它能幫助我們確定人名的右邊界。例如"李煥常到..."。"李煥常"常被視為一個中式人名，這是因為右邊界切錯。但若是換成"李煥院長常到..."這個句子，則其結果會變成"李煥"為一個中式人名。這是因為有頭銜的存在使右邊界切對了。所以當有這種情形出現時，我們就給在頭銜前後的中式人名候選者加分，表示此類的候選者應有更高的機會成為中式人名。

3.3.9 短期記憶(Cache Memory)

根據我們的觀察，由於所在的上下文不同使得有些人名在某些狀況下會被辨識正確，而在某些情形下卻又辨識錯誤。通常這種人名是成績較低者在正常情形下無法超過閥值而不被辨識，但有時候可能是因為有其他訊息的幫忙而使得其成績超過閥值。我們覺得若能藉著其中辨識正確的結果將錯誤的糾正過來，將能減少許多不必要的錯誤，因此才有了此種想法。

由於需要依據正確的人名來更正錯誤的人名，因此勢必要有一個地方將所有的人名都存放起來以茲比對。Kuhn(1988)所提出的一個短期記憶的概念，給了我們一個靈感去利用短期記憶來存放我們所辨識出來的人名。但是否所有的人名都有必要存放起來比對呢？由於文章中通常每一段有自個的主題，當主題換了內容就不一樣了。拿不同主題的內容來互相比較似乎不太適合。所以每當一段落結束時，短期記憶的內容也要全部清除，不要讓別的段落內容來影響新的辨識結果。

什麼時候才會利用短期記憶來檢查辨識結果呢？我們的作法是每當有兩個已辨識的相似人名時，便需要檢查是否需要更正。何謂相似？就是每當有兩個人名的前兩個字相同時，此二人名便稱為相似的人名。因為我們主要是尋找兩個可能是同一個人名，但因為某些原因而多一個字或少一個字，所以只要前兩個字相同者都有嫌疑。當兩個相似的人名出現後，總共有以下四種不同的處理方法：

(1) $C_1C_2C_3$與$C_1C_2C_4$同時存在，但C_1C_2才正確
(2) $C_1C_2C_3$與$C_1C_2C_4$同時存在，但兩者皆正確
(3) $C_1C_2C_3$與C_1C_2同時存在，但$C_1C_2C_3$才正確
(4) $C_1C_2C_3$與C_1C_2同時存在，但C_1C_2才正確

其中C_i是表示一個字。從上面的四種情形看來，很明顯的兩兩矛盾：情形(1)與(2)互相矛盾，而情形(3)與(4)也互相矛盾。短期記憶的作法重點正在這裡：找出適合的情形，才能將錯誤更正而不會將正確的弄錯。

什麼時候會發生情形(1)？例如"李煥常去打高爾夫球"的句子中，"李煥常"容易被辨識成為一個人名。而同樣的在"李煥及林洋港..."的句子中，"李煥及"也容易被辨識成為一個人名。此時短期記憶中就有兩個類似的人名出現："李煥常"與"李煥及"。此一例子是屬於情形(1)。而其它的例子如"邱永漢、邱永強兩兄弟..."的句子中，就有"邱永漢"與"邱永強"兩個相似的人名被辨識出來。但此時是屬於情形(2)。(3)與(4)也是類似的情況。

要更正錯誤必須先知道哪一個人名是正確的。知道那個人名是正確的之後，才能做出正確的判斷。我們的作法是對每個在短期記憶中的人名給予一個權重(weight)。總共有兩種權重：高權重與低權重。當有一人名被辨識出來而且我們能確定此一人名的右邊界時，此一人名便有高權重。而當此一人名在句尾或者緊接著一個頭銜時，此人名的右邊界便可確定。例如"焦仁和秘書長..."或"...焦仁和。"的句子中，其右邊界便可確定。除此之外的人名皆只是低權重而已。

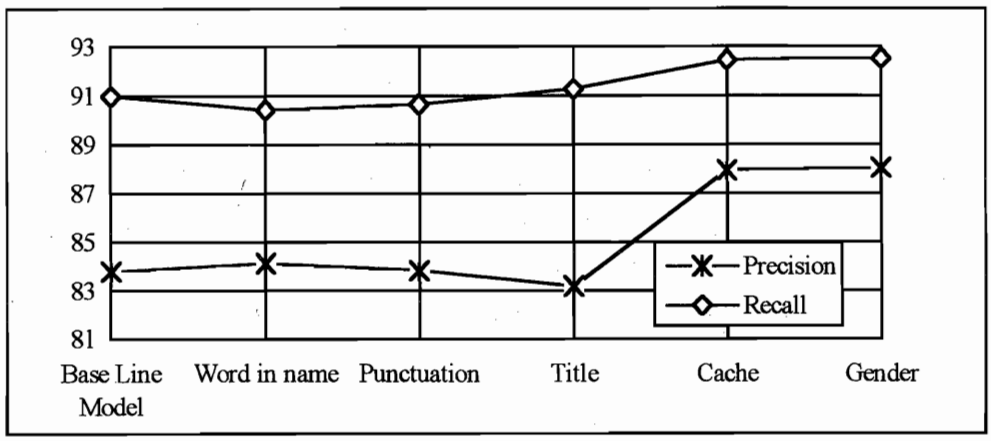
我們的作法就是由兩個類似的人名的權重來決定誰是正確的人名。基本上是以高權重者為正確的人名。當發生兩個人名都是高權重時，我們就視此二人名皆正確。若兩個人名皆是低權重時，則看第三個字的成績如何。若第三個字當人名的成績夠高，則視其為正確人名。若不夠高則將第三個字去掉，也就是情形(1)。例如"李煥常"與"李煥及"就是此類情形。

總括而言，我們希望能利用此一概念來將一些重複的錯誤。利用可能被正確辨識的部份來加以更正，以減少一些不必要的錯誤。

3.4 實驗結果

圖二是每次加入一輔助訊息後的效益走勢圖。從圖二可以看出每一輔助訊息對整個系統的影響程度。根據觀察，前三條輔助訊息對系統的幫助有限。這是由於在我們的測試資料中符合這三類的情形並不多，以至於效果並不顯著。最後的兩條訊息則效果較為明顯。

表一則是最後中式人名辨識的實驗結果。從表一可以知道對大部份的檔案而言，精確率或召回率都相當不錯。但是檔案四與檔案五的精確率並不理想，因此需要詳細的討論為什麼這兩個檔案會如此特殊。



圖二、 每次加入一輔助訊息後的效益走勢圖

表一、 中式人名辨識實驗結果

	正確的姓名總數	系統辨識的總數	正確	錯誤	遺失	精確率	召回率
檔案一	641	657	600	57	41	91.3242	93.6037
檔案二	1628	1631	1533	98	95	93.9914	94.1646
檔案三	666	655	565	90	101	86.2595	84.8348
檔案四	148	220	141	79	7	64.0909	95.2703
檔案五	176	213	158	55	18	74.1784	89.7727
檔案六	694	780	662	118	32	84.8718	95.3890
合計	3953	4156	3659	497	294	88.0414	92.5626

檔案四是報導國際新聞的國際版，其中有許多專有名詞(特別是國家名很多)。而此類專有名詞常常斷詞錯誤，因此會造成干擾。例如"立陶宛"這個國名。在這個例子中"立陶宛"被斷成三個單字詞，而"陶宛"是一個非常合理的人名，所以造成了大量的錯誤。

另一錯誤是一些像中式人名的音譯人名。例如"魏斯特"、"艾琳達"等等。此類的音譯人名擁有與中式人名相同的特性，所以被視為中式人名。但由於我們的目的不僅是要將人名找出來，同時還要確定這些人名的屬性。所以雖然這些是正確的人名，但因為並不是中式人名，所以我們仍視其為錯誤。以上二類的錯誤，佔了檔案四所有的錯誤中的50%。

同樣的，在檔案五也有類似的問題。檔案五是經濟版，其中有許多的股票名稱與術語與中式人名很像。例如"華隆"。同時在這個檔案裡有一些不常見的姓氏。例如"應"這個姓氏。由於這個姓氏不在我們的姓氏集合

中，所以就無法辨識。還有這個檔案較為特殊的地方是"萬元"這兩個字。由於其出現的頻率也很高，因此也被當成人名看待，所以也造成許多錯誤。

總而言之，在所有錯誤中因為音譯人名被誤認為中式人名而造成的錯誤總共佔了總錯誤的20%。而因為一些罕見的姓氏沒有包括在我們的姓氏中而造成的錯誤總共佔了總錯誤的14%。其餘的錯誤則是一些非人名的候選者被辨識為人名或者是邊界辨識錯誤。

4. 音譯人名的辨識

4.1 定義與分析

首先我們要先知道什麼是音譯人名？例如"Picasso"一般會被翻譯成"畢卡索"這就是一個音譯人名。一般而言，外國人名都是以音譯的方式來翻譯。然而並非所有的外國人名都是以相同的方式來翻譯。一般而言西方語言的人名都是以音譯的方式來翻譯，而東方語言的人名則未必。像韓國人名就類似中式人名，可以用中式人名來看待。而日本人名則是完全不同的型態。例如"梁井新一"、"犬養毅"等。此類人名因為不屬於音譯人名的範圍，所以並不在我們的考慮範圍內。

音譯人名的結構方面比起中式人名來得複雜得多。在中式人名中，有姓氏，而且姓氏的用字有固定的字集。另外，名字的長度也有限制。所以結構方面較為清楚。相較起來，音譯人名就沒有類似的結構。所以音譯人名勢必需要另闢蹊徑，找別的線索。

4.2 基本想法

我們主要目標在於音譯人名，音譯人名的翻譯是以發音為其根本。所以要解決此一問題，發音是一個很重要的線索。發音是由音調與注音符號所組成的。而音譯人名的音調通常與原文的發音不太相同，所以音調的部份我們不予考慮，而只考慮其注音符號。中文的發音習慣與英文的發音習慣不太一樣，而這一點正是我們所憑藉的線索。

根據我們的觀察，由於中西方之間的發音習慣不同，造成音節的排列順序也不同。也就是說，注音符號的排列順序在中西方的語言中會有極大的差異。例如"史蒂芬席格"的注音符號排列順序是"ㄕ ㄉㄧ ㄈㄣ ㄒㄧ ㄍㄜ"，而此種注音符號排列順序在中文是非常罕見的。所以由注音符號的排列順序，便可以看出是否為西方語言所翻譯的字串。

現在我們原先的想法碰到了一個問題，要知道哪些注音符號的順序屬於中文或是西方語言的翻譯。因為參數量大自然需要大量的訓練語料庫。然而此種大量的音譯資料的訓練語料庫取得困難，所以我們必須要對原先的構想作一些修正，以至於有了以下的一些條件。

4.2.1 音譯人名的字集

音譯人名有一種很有趣的現象：所有的同音字只有固定的幾個字會被選來當音譯人名的字。如下例：

Richard Macs	理查馬克斯	娌茶螞剋鷥
George Bush	喬治布希	瞧製怖稀

在上例中，每個英文人名後接的第一個字串是可能的譯名，而第二個字串則是我們所找的反例。一個英文名字可能有許多種翻譯結果，但決不會如上例一般有"娌茶螞剋鷥"或"瞧製怖稀"等的翻譯結果。在上例中，每個英文名字後接的兩個中文字串是同音字串，如"理"與"娌"就是同音字，其餘對應的字，也都是同音字。從這裡可以看出，中文的同音字雖多，但是會被選為音譯人名者並不多。所以我們挑選出一個字集，其中包含了所有曾被音譯人名使用過的字，並以此為基本線索。

我們的作法是：以此字集為根本，掃瞄輸入的句子，並找出其中所有由字集中的字所組成的字串，視這些字串為候選者。字集在此就扮演著與姓氏在中式人名辨識中一樣的角色：負責引發系統開始動作。再由其餘的策略來確定。

4.2.2 音節的條件

所謂音節條件如前述。由於音譯人名長度的不確定再加上注音符號共有37個，若想藉統計的方式與大型語料庫合作找出音譯人名的注音符號排列形式，則總共需要的參數量實在太大了。因此在沒有足夠的訓練語料庫之下，此一條件必須作一些必要的修正。

我們手中可茲訓練音節條件的訓練語料庫只有2,692個音譯人名。數量太小，所以我們便縮小條件的適用範圍。我們不檢查整個字串的注音符號排列方式，而只檢查每個字串的頭尾兩個字。方式是檢查此二字的注音符號是否曾為音譯人名中的頭尾二字的注音符號。亦即若C_h與C_t為某字串的第一字與最後一字，而C_h的注音符號不曾為音譯人名中的第一字的注音符號，則此字便不符合音節條件。同樣的，C_t也是類似的方式來檢查。

使用此一條件的目的在於輔助我們確定候選者的邊界，同時剔除一些由字集中的字所組成的候選者。例如"各國"、"令人"、"二發"等。其用字皆屬於字集中的字，但是卻明顯不是音譯人名，由其發音便可以明顯的得知此點。有了音節條件，此種問題便可以有效的解決。

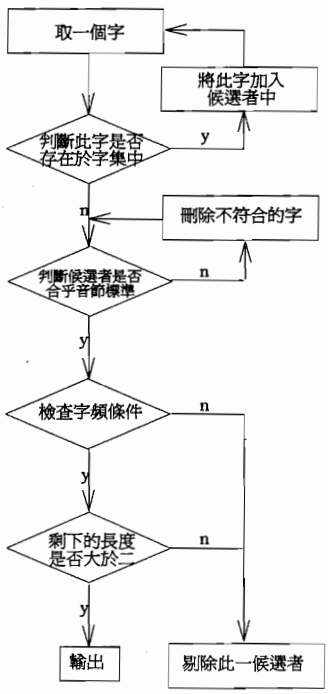
4.2.3 字頻

字頻這個條件，在中式人名中是一個很基本且重要的條件。但是在音譯人名中，有著與音節條件相同的困難：訓練語料庫太小及參數太多的問題。因此，我們也勢必要縮小整個適用範圍，以免因為訓練量的不足而造成反效果。

由於兩個字的候選者特別多而且在兩個字時的參數量較少，所以我們便將此一條件應用在兩個字的候選者中。為什麼兩個字的候選者最多？因為兩個字最容易符合字集的條件而成為候選者。每當有兩個字的候選者出現時，我們便檢查此一候選者的字頻的總和。亦即此二字出現在音譯人名中的次數，若大於一閥值，則算是通過此一標準，反之則不符合此一標準。閥值的取得方法與中式人名的閥值取得方法類似，可由訓練語料庫中取得。

4.3 辨識流程

圖三是音譯人名的辨識流程。



圖三、 音譯人名的辨識流程

辨識流程首先由左至右掃瞄一輸入的句子，將其中所有由字集中的字所組成的字串找出來，而這些字串便是候選者。再利用音節條件與字頻條件來篩檢不符合的候選者。

4.4 實驗結果

表二是音譯人名辨識在不同檔案的實驗結果。附錄列出六個版面的樣本語料，及組織名辨識結果可供參考。有些檔案的精確率與召回率都還不錯，有些檔案則有待改進。就平均而言，召回率還不錯，精確率則尚待改進。在此我們將仔細討論一些問題。

表二、 音譯人名辨識之實驗結果

	正確的音譯人名總數	系統辨識總數	正確辨識總數	錯誤	遺失	精確率	召回率
檔案一	52	64	34	30	18	53.1250	65.3846
檔案二	9	88	6	82	3	6.81818	66.6667
檔案三	238	300	180	120	58	60.0000	75.6303
檔案四	301	301	230	71	71	76.4120	76.4120
檔案五	34	152	26	126	8	17.1053	76.4706
檔案六	214	300	134	166	80	44.6667	62.6168
合計	848	1205	610	595	238	50.6224	71.9340

從平均召回率來看，我們用來選取候選者的線索，其效果還不錯，大部份的音譯人名都能被選取出。造成召回率不夠高的原因，是因為有許多的音譯人名與中式人名的形式類似。例如"魏斯特"、"艾琳達"等音譯人名，有如同中式人名的姓氏，名字的部份也符合中式人名的要求：兩個字。因為我們的系統是採取中式人名辨識與音譯人名辨識一起執行。所以這些音譯人名便被視為中式人名。雖然我們的音譯人名也有辨識出來，但是競爭結果中式人名贏了，所以被視為中式人名。這個因素是造成音譯人名的召回率無法達到應有的水準的重要原因之一。

造成精確率如此不理想的原因有很多，我們大致將所有的錯誤分成三大類：一般專有名詞的錯誤、邊界錯誤、其他。所謂的一般專有名詞是指一些品牌名、綽號等的專有名詞。因為看來像音譯人名，因而被辨識成音譯人名，因此造成錯誤。例如"黛安芬"、"雪佛蘭"等。因為此二者皆為音譯文字，因此符合我們所有條件。但因不是人名，所以屬於錯誤的辨識結果。另外，如"東尼"、"哈姆"、"阿蓮"等是一些綽號，也有與音譯人名類

似的特性。由於我們的策略其實對所有音譯文字都適用，而上述的例子都有與音譯人名相同的特性，所以都被辨識出來了。

另一型的錯誤是邊界錯誤。其實此種音譯人名幾乎已經被辨識出來，但是可能多了幾個字或少了幾個字而造成錯誤。由於是因為名字的邊界錯誤而造成的，所以我們稱之為邊界錯誤。例如"狄恩史密斯[都]"、"(拉)瑞強森"。[都]是多出來的字，而(拉)是少掉的字。最後一型是其他。可能是一些根本與音譯文字無關的字串。例如"利多"、"連拉"等。

4.5 討論

在我們所提出的策略中，其實是以中西方語言發音習慣的不同再加上音譯文字的用字習慣，來分別文章中的的音譯文字與一般的中文字串。從平均召回率來看，此一策略似乎成功的將音譯人名都涵蓋住了。現在的問題就是名字的邊界問題。我們採用的策略不夠理想，以至於精確率還有很大的改進空間。另外還有人名與其他音譯名詞的歧義問題也是尚待改進之處。

由我們的錯誤中觀察得知，我們所使用的策略有相當的潛力來擴張到對所有的音譯專有名詞。因為有許多音譯專有名詞被成功的辨識出來。但是有一點須注意的是：一般的外文專有名詞與音譯人名不同。一般的外文專有名詞再翻譯時常是音譯與意譯併行。例如"George Town"若是翻譯成人名，應是全部音譯成"喬治唐"。而若是表示一城市，則可能變成"喬治城"。意即最後一字是意譯的結果。這一點是需要在未來克服的問題。

表三、 人名辨識之實驗結果

	正確人名總數	系統辨識總數	正確辨識總數	錯誤	遺失	精確率	召回率
檔案一	688	721	643	78	45	89.1817	93.4593
檔案二	1634	1719	1539	180	95	89.5288	94.1860
檔案三	902	955	762	193	140	79.7906	84.4789
檔案四	449	521	402	119	47	77.1593	89.5323
檔案五	202	365	183	182	19	50.1370	90.5941
檔案六	912	1080	838	242	74	77.5926	91.8860
合計	4787	5361	4367	994	420	81.4587	91.2262

由於我們是將中式人名辨識與音譯人名辨識兩系統同時運作，並且賦予每一辨識結果一適當的屬性(即此一人名是中式人名或音譯人名)，難免

會有些人名被誤認。因此我們另外統計，若不論其是否為中式人名或音譯人名，只要是正確的人名便視為正確時，其結果詳列於表三。

5. 結論

本論文提出一些策略用來解決一些在現今斷詞系統中存在已久的問題：人名辨識。我們不僅將這些專有名詞從文章中找出來，同時還賦予這些專有名詞一個適當的屬性。有了這些屬性，不僅斷詞系統可以使用，後續的系統也可以利用這些屬性。

為了瞭解這些策略加入斷詞系統後對斷詞系統的影響，我們也做了一個較小型的實驗來驗證一下。我們從平衡語料庫中，每個不同的文體中隨機取一個檔案，總共有15個檔案，21,181個詞，合39,544個字，作為測試資料。NTU斷詞系統對此測試資料的原始精確率為85.79%，召回率為91.79%。加入人名辨識系統後，精確率提高1.09%變成86.88%，而召回率提高0.33%變成92.12%。提高程度並不如預期中的高，其原因是因為原本此一測試資料中的人名就不多，所以影響並不大。

在我們歸納所有的錯誤後發現，由於訓練語料庫的不適當所造成錯誤非常的多。因此一個量大且正確的訓練語料庫是非常重要的，這將是未來一個很重要的工作。我們在第4節中所提出來的策略，雖然是針對音譯人名。但是如第4節所述，其實是相當有潛力擴張成對所有的音譯文字。不過要注意的是有些專有名詞的翻譯方式是音譯與意譯並行，這也是未來一個重要的研究方向。

誌謝

本研究部份由國科會計畫NSC83-0408-E002-019支持，在此誌謝。

參考文獻

Chang, J.S., *et al.* (1991). "Word Segmentation through Constraint Statisfication and Statistical Optimization." *Proceedings of ROCLING*, 1991, pp. 147-165.

Chang, J.S., *et al.* (1991). "A Multiple-Corpus Approach to Identification of Chinese Surname-Names." *Proceedings of Natural Language Procssing Pacific Rim Symposium*, 1991, pp. 87-91.

Chen, K.J. and Liu, S.H. (1992). "Word Identification for Mandarin Chinese Sentences." *Proceedings of COLING*, 1992, pp. 101-107.

Chiang, T.H., *et al.* (1992). "Statistical Models for Word Segmentation and Unknown Word Resolution." *Proceedings of ROCLING*, 1992, pp. 121-146.

Church, K.W. and Hanks, P. (1990). "Word Association Norms, Mutual Information and Lexicography." *Computational Linguistics*, 1990, Vol. 16, No. 1, pp. 22-29.

Kuhn, R. (1988). "Speech Recognition and the Frequency of Recently Used Words: a Modified Markov Model for Natural Language." *Proceedings of COLING*, 1988, pp. 348-350.

Lin, M.U., Chiang, T.H. and Su, K.Y. (1993). "A Preliminart Study on Unknown Word Problem in Chinese Word Segmentation." *Proceedings of ROCLING*, 1993, pp. 119-141.

Wang, L.J., *et al.* (1992). "Recognizing Unregistered Names for Mandarin Word Identification." *Proceedings of COLING*, 1992, pp. 1239-1243.

彭載衍,張俊盛 (1993). "中文辭彙岐義之研究--斷詞與詞性標示." *Proceedings of ROCLING*, 1993, pp. 173-193.

黃玉眞 (1992). *洋名洋名任你選*. 學習出版有限公司, 台灣, 1992.

附錄・樣本測試語料

我們由六個不同的報紙版面中各抽出一小段測試結果。附錄中列出這些樣本測試結果。

(1) 政治版

海外異議份子羅益世被控叛亂罪，台灣高等法院昨日宣判，審判長劉士元認為，羅益世主張台灣獨立的言行並不構成預備叛亂罪，但有觸犯刑法煽惑他人犯罪（犯叛亂罪）嫌疑。該罪因屬一審法院管轄，故高院諭知本案管轄錯誤，將卷證移送桃園地院審理。這是台灣四十餘年來海外異議人士涉嫌叛亂罪中，司法機關首度作成不構成叛亂罪的認定。羅益世叛亂案是由高院的審判長劉士元，法官洪清江、薛爾毅組成合議庭審理，審判長劉士元指出，檢察官將羅益世鼓吹台獨言論依預備叛亂罪起訴，合議庭審理後認為，從起訴事實中並未見羅益世有台獨的具體行為，因此並不構成叛亂罪。檢察官陳耀能在起訴書中指出，七十八年八月二日，羅益世入境參加世台會第十六屆年會，趁機推動「台灣獨立」，八月十一日在高雄市立圖書館中興堂參加該會活動時，公開宣稱：「現在主張一台一中，把台灣放在前面，經過公民投票，這個意思也同樣是說台灣要獨立」。八月十七日及廿一日，羅益世又先後在彰化縣平和國小及台中市健行國小操場，參加「歡迎李憲榮返鄉說明會」時，公然鼓吹「台灣獨立」，向群眾聲稱：「台灣的主權是屬於台灣人民，絕對不是在中國」「其實國際上已經將台灣列為一個獨立的國家，他們要等我們台灣人自己來爭取台灣獨立」。

(2) 社會版

二名警員遭槍殺後，台北市警局所屬的刑警大隊、城中分局及大安分局分別出動大批警力協助緝凶，由於現場歹徒只遺留了一部山葉黑色追風重型機車，車號〇〇七四七二八乃據此查出牌照係失竊，而再據引擎號碼查知車主李立雄及其犯有殺人未遂前科的胞弟李立中可能涉嫌。但由於歹徒逃逸時，曾攔下一部〇八〇三六六六白色裕隆計程車逃離，雖查出車主是陳來福所有，但迄至昨晚，透過各交通電台及無線電計程車台，呼籲其出面說明，但仍未見其到案。警方昨天上午首先根據目擊證人楊姓學生提供，在離紹興南街廿公尺處，找到歹徒丟棄的山葉黑色追風重機車，經查牌照為陳學斌所有，乃由大安分局警備隊員，前往青田街二巷，陳某住處詢問。就讀某私立大學，當時正在睡覺的陳某，對於警方詢問後，始知車牌失竊，他指出，由於他的機車引擎故障，機車停在樓下已有幾個月未騎用，他前晚七時許返家，發現車牌還在，昨天上午十時許警察到家中找他，他才知道車牌失竊。而且由於上周一發生車禍，膝蓋處、腳踝部還裹上石膏，他的身高一七七公分，體重六十公斤，與目擊證人描述不符，警方已予排除。接著警方又拆下引擎，經以雷射化驗未被變造過，而查出車主是住在新莊市的李立雄所有，前往李某住處時，其父稱李某從事木工，機車在昨天被其弟弟李立中（廿四歲）騎用，尚未返家，至於李立中從事何種行業，李父稱只知道他在台北工作，並不知從事何種行業，而且警方經查其前科資料，發現李立中犯有殺人未遂前科，而將他列入涉嫌重大，正全力緝捕中。

(3) 娛樂版

由年代公司負責人邱復生與學者電影公司蔡松林共同組成的大學電影公司，昨（二十三）日宣佈成立，大學公司將結合香港的技術及智力，引進香港技術。希望在三年內扳回台片的劣勢，大學將以發行西片及在港投資拍片為主。第一部推出的國片為「火燒島」，西片為「時空攔劫」。學者公司仍將維持代理港片發行及百分之七十的拍片量；年代公司原有的年代國際股份有限公司，則走國際公司路線，在香港成立和盛公司及香港年代公司，前者已殺青由大陸演員鞏俐主演的「大紅燈籠高高照」；後者則代理港劇，「年代」原有的歡樂電影部拍片業務，則移轉至大學公司。「大學」在港已邀約張婉婷等多位傑出導演，為公司籌備新戲，西片則以千萬美元取得代理美國新興起的卡洛可公司，兩年內的影片在台發行權。影片包括席維斯史特龍的影片、「魔鬼終結者續集」等。三月一日首開推出「時空攔劫」；第一部國片「火燒島」則於青年節推出。

(4) 國際版

伊拉克新聞社週一報導，海珊總統在傳達給伊朗總統拉夫桑加尼的訊息當中，強調改善兩伊關係的必要性。海珊在一封紀念伊朗回教革命十二週年的信中說：「我們希望眞主看在伊朗回教徒致力獻身宗教的份上，將成功賜給他們。」他也告訴拉夫桑加尼：「我們希望兩國的關係能確保人民的福祉，並且繼續和睦地發展下去，尤其當兩國人民是團結在相同的宗教及信仰之下。」海珊並向眞主求助：「使那些行走在正義和信仰上的回教徒團結起來吧，迎頭痛擊一切暴政、腐敗、異教徒的陰謀和傲慢。」海珊在週末過後對伊朗的和平方案做了回應，拉夫桑加尼對此回應表示失望。

(5) 經濟版

朱耕稼一直被市場視為指標之一的華隆集團股票昨日紛紛展開強勢，但是該集團的幾個主要「靈魂人物」卻表示並未大力做多，甚至還在外界打聽是誰在「動」他們的股票，而由進出表觀察，華隆旗下的幾個號子在旗下股票上升呈現賣超，如此不尋常的現象，值得進一步觀察。由於華隆旗下股票在上週的反彈過程中，並未呈現相對強勢，而昨日的華隆、華隆特、嘉富、農林等股同步轉強，顯得格外搶眼，讓投資人有華隆集團做多的聯想。但是昨日該集團的主要操盤人卻表示對旗下股票轉強「毫不知情」，甚至還向外界打聽誰在「進貨」，頗耐人尋味。但是據了解，該集團的股票由於向來是自家人強勢主控，握有生殺大權，按理市場人士應不會輕易試圖「對做」，但是不排除有心人士看準該集團尚未吃貨完畢，而先行上車想抬個轎子。亦有業內推測，由於華隆集團的動向已成為市場關注的焦點，製造自家券商賣超而在其他號子進貨為一種避開他人耳目的方法，另外由於華隆在外界形象較為偏多，而國華證券週年慶的「一日反轉」，使一些懷有預期心理的投資人慘遭套牢，亦不排除日後會以明出暗進或明進暗出的方式操作，以避免擔負「重任」的壓力，所以進出較不願聲張，使外界無法捉摸其動向，或許可為另一種解釋。

(6) 運動版

來自田徑協會的消息，第三屆世界室內田徑錦標賽，昨天我國短跑女將王惠珍在六十公尺複賽跑出七秒四二未擠進決賽，男子跳遠賴正全因腳痛未參加。王惠珍前天在六十公尺預賽，驚險地以七秒四零搭上複賽列車，成為亞洲唯一進入複賽的選手，但她昨天在複賽跑出七秒四二，排名第十四名。賴正全前天在男子六十公尺預賽跑出六秒九五遭淘汰後，腳傷疼痛難耐，昨天決定退出跳遠賽。北京亞運跳遠冠軍大陸選手陳尊榮及季軍黃庚的成績分別為七公尺五七與七公尺五四，與決賽七公尺九十標準相差甚遠，都遭淘汰。今天王惠珍將參加二百公尺預賽，中華代表隊預計十三日返國。