國立交通大學

資訊學院資訊科技（IT）產業研發碩士專班

碩士論文

中文文本中限定性抽象名詞指代消解

Definite abstract anaphora resolution in Chinese Texts

研 究 生：程俊樺

指導教授：梁　婷　博士

中 華 民 國 一 百 年 一 月

中文文本中限定性抽象名詞消解

Definite abstract anaphora resolution in Chinese Texts

研 究 生 ：程俊樺　　Student：Jyun-Hua Cheng

指導教授 ：梁　婷　博士　　Advisor：Dr. Tyne Liang

國 立 交 通 大 學

資訊學院資訊科技（IT）產業研發碩士專班

碩 士 論 文

A Thesis

Submitted to College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Industrial Technology R & D Master Program on

Computer Science and Engineering

Jan 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年一月

中文文本中限定性抽象名詞指代消解

研究生：程俊樺　　指導教授：梁　婷　博士

國立交通大學資訊學院產業研發碩士專班

摘　　要

在文本中，指代是一種常見的詞彙替換，用以指示先前所提到的事物。在中文文件裡，指代現象包括有代名詞指代、零指代以及名詞指代，其參照對象可為抽象描述或實體名稱。在本論文中，我們針對限定性的抽象名詞指代，提出一個以小句為單位的指代消解程序。利用同義詞詞林、中研院八萬目詞辭典及網路搜尋相關詞等資源，進行指代詞辨識、辨識特徵萃取。我們建立有限狀態機，以進行指代詞辨識，在 1538 個實例中達到 90%辨識正確率。我們萃取四種類型共十個特徵，包括位置特徵、距離特徵、詞彙特徵和語義特徵，做為回指對象的挑選依據。我們分別以支援向量機分類器和權重計算法來進行指代消解，並以基因演算法求出最佳特徵組合。實驗結果顯示在 241 個抽象名詞指代消解，支援向量機分類器在小句符合的正確率是 40.66%，長句符合的正確率是 68.46%，權重計算方法在小句符合的正確率是 42.32%，長句符合的正確率是 70.54%。

Definite abstract anaphora resolution in Chinese Texts

Student：Jyun-hua Cheng　　Advisor：Dr. Tyne Liang

Industrial Technology R & D Master Program of
Computer Science College
National Chiao Tung University

ABSTRACT

Anaphora is a common phenomenon in written texts, denoting the use of terms referring the mentioned entities previously. There are pronominal anaphora, zero-anaphora, and nominal anaphora in Chinese texts. The referents can be abstract or entities. In this thesis, we focus on studying definite abstract noun anaphora, and we propose a clause based anaphora resolution procedure. Furthermore, anaphora identification and feature selection are done by using CLINE, CKIP lexical and Google search results etc. The anaphora recognition achieves 90% precision using finite state machine in 1538 instances. Furthermore, we extract four types of features to classify candidate antecedents including position features, distance features, lexicon features and semantic features. These features are used for building SVM classifiers and weighted model on resolving anaphora. The best features set are found by a genetic algorithm. In 241 definite anaphora instances, the SVM classify achieves 40.66% on correct clause position and 68.46% on correct sentence position. The weighted method achieves 42.32% on correct clause position and 70.54% on correct sentence position.

致謝

碩士生涯即將結束，也代表一個新的旅程即將開始。這個蛻變需要不斷的成長與學習，如同我之所以能完成這篇碩士論文一樣。我的成長來自於老師的教導，讓我知道自己的成長空間有多麼廣大和自己不足的地方。我所學到的知識都是來自於老師指導，讓我知道錯誤的地方和下一步該往那個方向前進。在學期間，我曾因為個人因素迷失了自己，我不知道自己的方向，是老師給了我「腳前的燈，路上的光」。因此我第一個想感謝的人是把我們當她的小孩一樣照顧和給予我們指導的梁老師。

典松學長、哲青學長、冠熙學長和淳齡學姊感謝你們給予我課業上的援助，你們在開會時所報告的內容和報告技巧是讓我獲益良多的來源。家祺、紹宜和博榮，你們和我一起修課、面對考試和假日留在實驗室為大學部的學弟妹們 demo 的日子，我永生難忘。奕賢、荃權和宏達，實驗室有了你們總是會熱鬧許多，是你們帶給實驗室歡樂的氣氛。

感謝在學期間我所遇到的每個人，不論是指導過我的老師們、實驗室的成員、其它實驗室的戰友、公司實習的主管和支持我的家人，有你們的提攜與陪伴是我最珍惜的時光。

目錄

表目錄

圖目錄

第一章　緒論

1.1　研究動機

指代(anaphora)是一種在文本或對談中常見的語言使用方法，可以用代名詞或指示詞回指先前所提到的事或物。指代的使用可以使行文簡潔流暢，避免詞彙重複出現。在中文文本中，常見的指代現象有代名詞指代、名詞指代和零指代。例如下面範例：

代名詞指代範例：*適之先生*與中研院的關係相當深遠,他是第一屆院士...

名詞指代範例：*適之先生*辭世已有三十二年...這位譽滿天下、謗亦隨之的哲人...

零指代範例：*這位譽滿天下、謗亦隨之的哲人*，φ靜靜的躺在這偏僻小公園的一隅，φ再也聽不見人世間的褒貶...

從這三種範例中，我們可以看見代名詞指代和名詞指代具有明顯的指代詞出現。相反的，零指代現象沒有指代詞出現，其指代的現象多依據句法結構。例如在零指代現象範例中，其未出現的指代詞是指先前已提過的主詞或受詞。

在文本中，將這些出現與否的指代詞進行回復到先前所提的事物稱之為指代消解，是文本理解中一項重要的工作。這項工作包括指代詞辨識和回指對象的辨識。基本上指代詞有三種類型，分別是代名詞、指示代詞(Demonstrative)與限定性描述(又稱有定描述“Definite description”)[1]。常見的代名詞有「他」、「她」、「它」、「我們」、「他們」..等；指示代詞是「這」、「那」..等；限定性描述是「這+名詞片語」、「那+名詞片語」..等。根據我們以 20 篇報導文章的統計指代詞的類型分佈，其中代名詞佔 54%、指示代詞佔 6%、限定性描述佔 40%。

至於回指對象的辨識，目前的研究多著重在實體的辨識，如前述的名詞指代範例中的「哲人」。對於回指對象類型屬於抽象事物的研究卻少見有深入的探討。因此，在本篇論文中我們針對指代詞是抽象名詞的限定性指代進行研究，提出有效的指代消解程序。我們利用外部資源包括同義詞詞林[23]、中研院八萬目詞辭典(CKIP 辭典)、中央研究院

斷詞系統[1]、和 Google 搜尋輸出[2]進行語義分析，並以基因演算模組工具 Pygene[3]選取最佳特徵組合，再以支援向量機分類器 LIBSVM[2]進行回指對象的挑選。

本論文的其餘內容包括在第一章中我們詳細定義所探討的問題以及相關研究。第二章是描述語料的建立過程、標記原則和統計資訊。第三章則詳細的敘述所提的指代消解過程，包括指代詞辨識、特徵萃取、回指對象辨識、以及實驗分析。最後在第四章中我們總結本論文的工作和後續的研究方向。

1.2 限定性名詞指代定義

如前所述，在中文文本中限定性指代，其指代詞需為包含指代定詞及名詞片語。指代定詞是指在中研院平衡語料庫詞類標記為「Nep」，包括「這、此、其、那、什麼、其中、個中、甚、啥、哪、斯、甚麼」等詞彙。此種限定性指代可對應到英文文本中的限定描述指代(definite description anaphora)，是指有出現「the」後接名詞片語的指代現象[3]。在中文文本中，熊學亮與劉東虹[4]這二位語言學家指出漢語並沒有定冠詞，限定性名詞難以鑑別，他們將限定性名詞「the+名詞片語」與指示詞「this/that+名詞片語」視為相同。在本論文中我們將只針對中文文本中有出現「這+名詞片語」為我們限定性指代的探討對象，其中的「名詞片語」限定在抽象名詞，抽象名詞的界定是根據 CKIP 辭典中屬於「非實體」類別的名詞，類別細目請參閱表 1。在我們的語料 885 篇文章中約有 3500 個「這+名詞片語」個限定性名詞指代出現，其中約有 1500 左右的名詞片語是 CKIP 辭典標為抽象名詞。例如在「這個方案...」，其中「方案」在 CKIP 辭典中是歸類為抽象類別「法則(Principles)」，是我們要處理的限定性抽象指代。

在處理中文文本中的限定性指代，我們觀察到回指的對象與指代詞之間的關係有二種常見的現象[3]。第一種關係稱為「直接回指/顯性回指(Direct anaphora)」，如範例 1a

[1] 中央研究院斷詞系統 http://ckipsvr.iis.sinica.edu.tw

[2] Google http://www.google.com.tw

[3] PyGene http://www.freenet.org.nz/python/pygene

中的指代詞「這項收費停車方案」與參照對象「學生收費停車方案」有相同的中心語(Head Noun)。

範例 1a：淡大自本學期開始，實施*學生收費停車方案*。這項收費停車方案，規定淡大學生停放汽車一學期一千六百元。

第二種回指現象稱為「間接回指/隱性回指(Indirect anaphora)」或稱「搭橋參照(bridging reference)」，如範例 1b 的指代詞「這個方案」與參照對象「推薦甄選」二者並沒有相同的中心語，需要藉由語義推導或聯想詞彙才能將二者關係聯繫起來[3][5]。因此處理這種回指對象的辨識較直接回指處理具有挑戰性。然而更困難的回指處理是當回指對象是一段敘述，如範例 1c 中需要界定回指對象敘述的邊界。

範例 1b：大考中心制定*推薦甄選*是為使整個考試制度多元化，不會局限在聯考一途。以前就有些高中向大學推薦優秀學生，去年有兩位錄取的學生是來自於聯考錄取率很低的高中。我們現在把這個方案推廣出去，應該較易為大家接受。

範例 1c：*中華奧會之所以打算走向結合企業的發展，主要希望能以民間團體的性質，改變奧會的體質，大力推動體育休閒的正當社會風氣，提供較多的選手和教練就業和福利*，目前這個方案正在策畫階段...

範例 1c 中是一段敘述用來描述或解譯指代詞的意思，這種也是一種「bridging」。在漢語中，王道英[6]將這種回指歸類為「總括型隱性回指」；在英語中，Byron[7]將這種以非名詞構成的回指對象，通稱為「抽象參照(Abstract referents)」。

表 1　CKIP 辭典非實體名詞類別

非實體類別名稱	普通名詞	專有名詞	地方詞
特徵 (Characteristics)	5434	102	119
文明 (Enlightenment)	1402	13	4
法則 (Principles)	938	11	1
社會活動 (Social_actiyities)	384	14	0
法人 (Corporation)	323	75	2476
名稱 (Nomenclature)	175	392	1
狀況 (Situations)	552	1	8
社會關係 (Social_relation)	57	0	0
財務關係 (Monetary_relation)	614	0	0
權力 (Authority)	125	0	0
疾病 (Illness)	389	0	0
時間 (Temporal_relation)	316	0	1
事件 (Events)	971	35	0
總合	11680	643	2612

1.3　指代消解相關研究

抽象指代的回指對象可以是名詞片語或者是語篇上所指的命題、概念、事實或事件[8]。要消解這些抽象指代現象，首先需要知道有幾種指代詞可以用於回指抽象實體。我們可以藉由一些線索，如英語中「it」、「this/that」、「this/that+名詞片語」或「the+名詞片語」四種形式來判斷。至於中文文本裡常見的線索是「這/那」、「這/那+名詞片語」。

研究漢語抽象指代的語言學者們都專注於探討指代詞與參照對象之間的關聯性與現象。其中王道英等人[6]將關聯性分為「包括型隱性回指」、「聯想型隱性回指」與「總括型隱性回指」三種類型。包括型隱性回指，如「*天然氣*這項乾淨的能源」，例子中的指代詞「能源」與參照對象「天然氣」之間存在著整體與部件關係；聯想型隱性回指，如「*天下沒有白吃的午餐*，這一句右派經濟學家所說出的格言」，例子中的指代詞「格言」與參照對象「天下沒有白吃的午餐」是根據常識產生關聯；總括型隱性回指，如「*如果以色列從約旦河西岸及加薩走廊撤離以及敘利亞從黎巴嫩撤軍，他將願意從科威特撤*

舉。這項計劃已被西方領袖拒絕」，例子中指代詞「計劃」是對先前提及事情做總結或者評論。

熊學亮等人[4]統計漢語抽象回指的指代詞與參照對象類型分佈情形，如表 2。其中語料是來自《魯迅文集》和報導評論共計 31059 個字，包含 108 個抽象回指。這些抽象回指的參照對象裡面，有 51%的參照對象出現在指代詞的前一個長句，37%的參照對象與指代詞在同一長句，12%的參照對象與指代詞是在同一段落或者跨段。

表 2　漢語指示語與參照對象分佈表[4]

指代詞類型	參照對象類型			
	句群	長句	小句	動詞片語
這	17	14	32	1
這+名詞片語	1	12	19	3
那	1	4	3	1
總計	19	30	54	5

至於英文的隱性回指消解分為二個方向，一種是名詞指代消解[9][10][11][12][13]。另一種是在消解代名詞指代時，將指示詞回指抽象參照納入處理範圍[7][14][15]。

Poesio et al.[9]利用 corpus-based 方法消解參照對象與指代詞是屬於同義關係和整體-部件關係。同義字是來自 WordNet，整體-部件關係是使用句法結構來擷取成對的名詞組，再計算每個名詞組的 mutual information 數值做為參考依據。句法結構包括「the NP of NP」、「NP of NP」、「NP's NP」和「NP N」四種，只對每一個名詞片語的中心語進行處理。測試語料是 38 個同義或整體-部件關係的實例，recall 是 66.7%，precision 是 72.7%。

Strube et al.[10]是利用決策樹分類器來消解限定性名詞指代、專有名詞指代與代名詞指代回指名詞片語的情形。分類的特徵有 14 個，包含語義角色特徵(主詞或受詞)、詞彙類型特徵(代名詞或專有名詞..等)、性別數量一致特徵、語義類別特徵(實體詞或抽象詞)、距離特徵、字串完全一致特徵和字串部分符合特徵。語料是 242 篇有關於歷史事件與人物相關的德文短篇，共 36924 個字。使用 10-fold 交叉驗証方法得到的 recall 是

56.65%，precision 是 84.96%。其中只單純消解限定性名詞指代所得到的 recall 是 22.47%，precision 是 69.26%。

Bunescu[11]利用 Web-based 的方法消解隱性回指和聯想型隱性回指，如森林和樹木。以指代詞、參照對象和二個詞彙做為搜尋關鍵字，擷取出回傳的文件數量再計算 mutual information。Recall 是 22.7%，precision 是 53%。

Gasperin and Vieira[12]消解葡萄牙文的隱性回指名詞指代，方法是先從語料庫中產生指代詞的相似詞詞群，以這個詞群做為參照對象的候選詞。相似詞的判斷方法是依據三種句法樣式，分別是「subject/verb」表示兩個詞彙是相同動詞的主詞、「verb/object」表示兩個詞彙都是相同動詞的受詞和「modifier/noun」表示兩個詞彙的修飾詞相同，只要符合其中一種即視為相似詞。測試語料是 95 個隱性回指實例，recall 是 33.3%，precision 是 52.7%。

Poesio et al.[13]利用多層感知器來消解整體-部件關係的限定性名詞指代。分類器使用距離和權重二種類型的特徵。距離特徵指的是詞彙相似度，計算方式有二種，一種是將指代詞和參照對象組合成關鍵字，如「the wheel of the car」，送到 Google 搜尋引擎，擷取出回傳的網頁數量。另一種是尋找指代詞和參照對象在 WordNet 的共同上位詞，再計算這兩個詞彙到共同上位詞所需要的連結次數。權重特徵包括二個部分，第一部分是候選參照對象與指代詞的長句距離，第二部分是考慮候選參照對象是否在指代詞前五個長句內第一次出現或者在整篇文章第一次出現。語料庫是來自 GNOME 裡面的 153 個整體-部件關係的隱性指代實例，使用 10-fold 交叉驗証方法進行實驗。實驗分別使用 Google 和 WordNet 進行比較，前者得到的 F-measure 是 76.3%，後者是 75.8%。

Byron[16]針對 TRAIN93 對話語料庫的「It」和「That」指代詞進行特徵標記與統計分析。此語料一共出現 376 個指代詞，其中 50%參照對象是名詞片語和代名詞，21%參照對象是抽象參照，29%的指代詞沒有參照對象。這些特徵包含指代詞、指代詞是否為重心、指代詞是否為主詞、指代詞的位置(主要子句或者附屬子句)、參照對象、參照對象是否為主詞、參照對象的位置和參照對象與指代詞之間的距離。記錄每個指代詞的特徵後，再以特徵的組合計算出現機率，例如對指代詞「it」而言，它出現在主要子句，

它的語法角色不是主詞，它的參照對象是名詞片語，它的參照對象在主要子句，這四種特徵同時出現的機率是 97.7%。

Byron[7]更進一步將語義規則納入消解的方法。語義規則是根據指代詞所連接的動詞來選擇適當參照對象，例如指代詞出現在動詞「load」的受詞位置時，指代詞的屬性須是「可移動的目標」，因此候選參照對象也必須是「可移動屬性的物體」。語義規則所參考到的屬性是建立於特定的語料之下，語料內容都是「貨運工作」相關的對話記錄。消解的方法是先利用語義規則過濾不適合的候選詞，當有多個候選詞時，則選擇距離指代詞最近的候選詞為參照對象。實驗語料共 10 篇包含 180 個指代現象，加入語義規則進行消解的正確率為 72%，未使用語義規則的正確率為 51%。

Navarretta[15]利用 Byron[7]所提及的語義規則與 Eckert and Strube[17]所提出的文章結構分析來消解指示詞指代。文章結構分析是經由人工進行標記，標記過的文章結構如同一個樹狀圖。其中父節點是文章的主體，子節點是文章的次主題。結構標記目的是用於減少候選參照對象的個數。語料庫中有 277 個實例，precision 是 71.84%，在不使用文章結構分析的情況下，precision 是 39.3%。

Strube and Muller[14]利用決策樹分類器消解代名詞指代和指示詞指代。分類是依據 25 個特徵，其中 16 個特徵是 Strube et al.[10]和 Muller et al.[19]所建立的名詞指代消解與代名詞指代消解。另外 9 個特徵是為了處理抽象參照而產生的，分別是參照對象類型特徵(名詞片語、動詞片語或者長句)、指代詞的參照對象是否傾向於名詞片語、指代詞的參照對象是否傾向於動詞片語、指代詞的參照對象是否傾向於長句、指代詞是否較有可能是名詞片語距離特徵和詞頻特徵。語料是使用 20 篇對話記錄共 3275 個長句有 1250 個實例，recall 是 40.72%，precision 是 56.74%，F-measure 是 47.72%。

第二章　語料建立與分析

我們使用的語料是來自於中研院平衡語料庫，以報導記敘類型的文章為主要處理對象。接下來的 2.1 小節描述語料蒐集細節與標記原則；2.2 小節是標記的結果與分析。

2.1 語料蒐集與標記原則

標記步驟與原則如下：

步驟一：檢查文章中每個小句是否出現「這(Nep)」，過濾掉含有「這是」和「這(Nep)」與「是(SHI)」中間未出現任何普通名詞(Na)、專有名詞(Nb)、地方詞(Nc)與動詞(V)的情形。

步驟二：尋找指代定詞所描述的對象，並將其標註為指代詞。

步驟三：從含有指代詞的小句往前尋找回指對象，並且記錄回指對象的位置。尋找過程中，優先考慮以名詞片語為主，若無法找到合適的名詞片語，則進行步驟四。

步驟四：以小句為一個單位，尋找出參考的小句編號並且記錄範圍。

標記過程中，我們排除以下三種情形：

(a) 指示代詞：「這」之後並沒有接續名詞片語的情形，例「這是」、「這可能是」..等。

(b) Discourse-New：出現指代詞，但並沒有回指對象，例「這個社會是不是病了」。

(c) Cataphora：出現指代詞，但參照對象並沒有在指代詞之前，例「這十項關鍵技術包括…」。

表 3 是一個標記範例，其中包括長句編號、小句編號、指代詞、參照對象、參照對象的小句參照範圍。範例中，第六小句出現限定性名詞指代「這項計劃」後，接著往前尋找參照對象，找到第四和第五小句，並將其標為參照對象。

表 3　限定性名詞指代標記範例

長句編號	小句編號	小句	指代詞	參照範圍
1	1	海珊在八月二日入侵科威特的十天之後，		
1	2	將他佔領科威特的行動與巴勒斯坦人的大業相提並論，		
1	3	他表示，		
1	4	如果以色列從約旦河西岸及加薩走廊撤離以及敘利亞從黎巴嫩撤軍，		
1	5	他將願意從科威特撤軍。		
2	6	這項計劃已被西方領袖拒絕，	計劃	4-5
2	7	但阿拉伯世界許多人已將海珊視為巴勒斯坦大業的英雄以及以色列的死敵。		

2.2 標記結果與分析

標記的文章篇數共 885 篇，其中有 3350 個限定性名詞指代。我們根據 CKIP 辭典中的非實體類別詞彙，找出指代詞屬於非實體的限定性名詞指代，再去除 Discourse-New 和 cataphora 的實例共 35 個，符合的有 1538 個實例。

在本研究中，我們定義長句單位是利用標點符號「。？！；」切割，小句單位是利用標記符號「。？！；，」切割。表 4 為語料統計資訊。

表 4　語料統計

語料類型	篇數	詞彙數	小句數	長句數
報導記敘文章	885	59985	82783	24062

指代詞可能包含多個參照對象，我們以文章中第一次提及的參照對象為準，統計它們的參照對象範圍，列於表 5。其中單一小句表示參照對象在一個小句內，二個以句(含)以上表示參照對象是一段敘述。

表 5　參照對象類型統計

參照對象類型		
二個小句(含)以上	單一小句 (含名詞/動詞片語)	加總
896 (59%)	607(41%)	1503

表 6 和表 7 是以參照對象的第一個小句為起點，分別計算參照對象與指代詞之間的小句和長句距離。間距為零表示二者在相同小句或長句，例如「*中東*這個地區」，指代詞是「地區」，參照對象是「中東」。指代詞與參照對象的距離超過三個長句以上的原因是在報導性質的部分文章中，通常在文章第一句話會說明文章所要描述的事件，而接下來數個長句都在講述事件的內容過程。例如報導運動比賽類型文章的開頭講述比賽的名稱與比賽發生的時間，接下來在數個長句都在描述比賽過程，最後作者再以指代詞回指到文章開頭的比賽名稱。

表 6　指代詞與參照對象起點的小句距離

小句距離	次數	百分比	累計百分比
0	83	6%	6%
1	304	20%	26%
2	267	18%	44%
3	206	14%	58%
4	149	10%	68%
5	100	7%	75%
6	67	4%	79%
7	56	4%	83%
8	31	2%	85%
(包含)9 以上	240	15%	100%
總計	1503	100%	

表 7　指代詞與參照對象起點的長句距離

長句間距	次數	百分比	累計百分比
0	532	35%	35%
1	575	38%	73%
2	177	12%	85%
3	81	5%	90%
4	38	3%	93%
5	27	2%	95%
(包含)6 以上	73	5%	100%
總計	1503	100%	

根據我們的語料統計，抽象名詞指代與參照對象的距離在三個長句以內的佔 90%左右，參照對象的位置分佈順序是「前長句(38%)＞同長句(35%)」，這個結果與熊學亮等人[4]所做的英漢語抽象回指的位置統計相同。另外，語料中有 25%實例的指代詞有多個參照對象，如範例 2a。

範例 2a：美國總統布希今天與蘇聯外長謝瓦納茲會談後宣布，他計畫於二月十一日至十三日在莫斯科與蘇聯總統戈巴契夫舉行高峰會，希望屆時雙方能簽訂旨在大幅削減長程核子武器的*裁減戰略武器條約*；他並決定暫時擱置限制對蘇貿易的修正案，擬提供蘇聯多達十億美元的農產品出口貸款，以供購買美國農產品，解決蘇聯糧荒。布希是在謝瓦納茲及國務卿貝克的陪伴下，在白宮玫瑰花園舉行記者會中作上述宣佈。謝瓦納茲讚譽布希的宣布是項非常重要的聲明，並表示他期盼見到布希與戈巴契夫簽訂*裁減戰略武器條約*。這項條約將規定兩國各裁減三十的戰略武器。貝克表示，這項條約還有一些高度技術性問題待解決，但他認為這些障礙可望及時在二月的高峰會前解決。

表 8 是依據 CKIP 辭典類別來統計指代詞出現在類別的次數，以及它們的參照對象類型次數。表中一共有 14 個類別，包括 CKIP 辭定義的 13 種類別和動名詞類別，當詞彙未出現在辭典中且被標記為動名詞時我們歸類到動名詞類別。詞彙範例是取自於我們的語料庫中出現次數較多的詞彙。

表 8　參照對象長度分佈次數

指代詞的類別名稱	參照對象長度		詞彙範例
	單一小句	二個小句(含)以上	
特徵	115	235	想法、行為
文明	71	134	問題、決議
法則	109	122	方式、制度
社會活動	84	22	比賽、會議
法人	43	5	社會、學校
名稱	10	2	職位、名字
狀況	10	120	情況、現象
社會關係	1	8	關係、情誼
財務關係	11	17	經費、收入
權力	4	1	政權、主權
疾病	9	2	病、病變
時間	30	35	期間、階段
事件	53	102	行動、過程
動名詞	57	91	調查、會談
總數	607	896	

由表 8 的統計結果，我們可以得知以「狀況」類別的詞彙做為指代詞時，參照對象有 92%是非名詞片語的類型。而以「社會活動」和「法人名稱」類別的詞彙做為指代詞時，參照對象的類型偏向於名詞片語。

表 9 是每一個類別的詞彙做為指代詞時，它們的參照對象的平均小句數。

表 9　非實體名詞指代詞的參照對象次數與平均長度

指代詞的類別名稱	指代詞出現次數	參照對象平均長度(單位：小句)
特徵	350	2.23
文明	205	2.79
法則	231	2.44
社會活動	106	1.21
法人	48	1.46
名稱	12	1.09
狀況	130	3.34
社會關係	9	2.7
財務關係	28	2.71
權力	5	1.4
疾病	11	1.73
時間	65	2.25
事件	155	2.79
動名詞([+nom])	148	2.07

第三章　指代消解

指代消解包括指代詞辨識和指代詞的參照對象辨識。指代詞辨識是判斷句子中是否存在名詞指代。參照對象辨識部分，我們以多種特徵和支援向量機進行辨識。

圖 1 為所提的指代消解架構圖，其中 Google 搜尋結果已經事先儲存至資料庫。

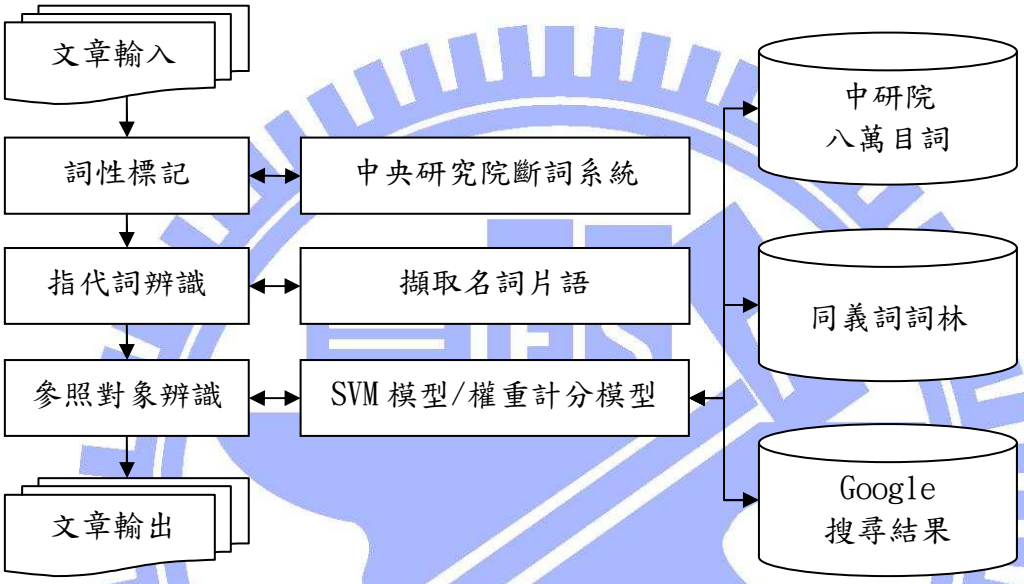


圖 1　系統架構圖

3.1 指代詞辨識

在本論文中所處理的指代詞是指句中含有『這(Nep)＋ 名詞片語/動詞』句型。首先，我們利用斷詞系統[4]將輸入的文章進行詞性標記，使用簡單規則辨識文章中名物化動詞與動詞轉修飾語的情形，接著建立有限狀態機來擷取名詞片語/動詞。以下為指代詞辨識與判斷中心語的步驟：

[4] 中央研究院-中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw

步驟 1：若“這(Nep)”出現在句子中且不含「這是」

步驟 2：名物化動詞處理、動詞轉修飾語處理

步驟 3：尋找名詞片語，若有名詞片語出現，則到步驟 5

步驟 4：尋找“這(Nep)”之後的第一個動詞，若存在則到步驟 5，若沒有擷取到任何動詞則辨識失敗，跳到步驟 6

步驟 5：將片語中最後一個詞彙視為中心語

步驟 6：結束

在 885 篇已標記的報導語料中我們發現，指代詞不僅僅是名詞片語，也包括了名物化動詞和動詞的情形。語料中 1538 個含有指代詞的小句裡有 148 個指代詞是動詞，因此我們參考馬偉雲[20]所使用的規則，將句子中的名物化動詞詞性轉成名詞。名物化動詞的處理規則如下：

(1).「是」前面的動詞詞性轉成名詞。

例 3a: *這(Nep) 種(Nf) 微量(VH) 的(DE) 調整(VC) 是(SHI) 被(P) 容許(VE) 的*

(2).「的」後面的動詞詞性轉成名詞。

例 3b：*這(Nep) 種(Nf) 傳統(Na) 的(DE) 繼承(VC)*

(3). 兩兩連續動詞中，若第一個動詞詞性為「VC」(動作及物動詞)，則第二個動詞詞性轉成名詞。

例 3c：*這(Nep) 項(Nf) 仲介(VC) 服務(VC)*

根據詞庫小組技術報告[21]的描述，我們知道中文裡大部分動詞可以轉化為名詞的修飾語，條件如下：

(4).「的」前面的動詞詞性轉成形容詞。

例 3d：*這(Nep) 個(Nf) 打拼(VA) 的(DE) 過程(Na)*

完成文章前處理的步驟後，我們建立有限狀態機找出指代定詞搭配名詞片語的組合，有限狀能機是改良自游基鑫[22]。

FSM 可以辨識的名詞片語如下：

這(Nep) 個(Nf) 國際(Nc) 金融(Na) 中心(Nc)

這(Nep) 種(Nf) 方式(Na) 進行(A) 的(DE) 消費(Na) 商品(Na) 交易(Na)

這(Nep) 項(Nf) 國小(Nc) 教師(Na) 心得(Na) 公開(A) 發表會(Na)

當句子中沒有出現任何一般名詞、專有名詞與地方詞時，我們會找出「這(Nep)+[量詞(Nf)/數詞(Neu)]+動詞」的句型，因此在例 3e 的小句中找出的片語是「這個妥協」。

例 3e：*這(Nep) 個(Nf) 妥協(VA) 從此(D) 一發不可收拾(VH)*

擷取出指代詞並將最後的詞彙視為指代詞的中心語，中心語是用來判斷是否是抽象名詞，若中心語是動名詞時也視為抽象名詞。

3.2 指代詞辨識結果分析

實驗語料是中研院平衡語料庫中以報導記敘類型的文章共 885 篇，經過人工標記後共有 1538 個句子符合抽象名詞指代。我們對 1538 個句子進行指代詞辨識，有 89.99% 的準確率可以正確擷取出中心語。辨識失敗的情形歸納成以下三種：

(1)動詞名物化問題：利用定理所設計的名物化處理無法涵蓋所有名物化動詞的類型，例「參加(VC) 這(Nep) 次(Nf) 義工(Na) 培訓(VC)[+nom] 」。在例句中正確的指代詞是「義工培訓」，而我們擷取到「義工」。當我們把所有語料庫中標記為名物化動詞的詞性轉成名詞後，有限狀態機的正確率可達到 93.7%。

(2)句子結構問題：利用有限狀態機並沒有辦法分辨出子句的結構，例「這(Nep) 項(Nf)由(P)國際(Nc)自動機(Na)工程(Na)學會(Nc)中華民國(Nc)分會(Nc)舉辦(VC)的(DE)超級(A)省(VJ)油(Na)車(Na)比賽(Na)」。在例句中，正確的指代詞是「超級省油車比賽」，而我們擷取到的是「國際自動機工程」。

(3)倒裝句型：在報導記敘性質的文章中，少部分的句子省略一些字或者標點符號造成辨識上的錯誤，例「這(Nep) 事(Na) 研究院(Nc) 方面(Na) 也(D) 漫無頭緒(VH)」。在例句中，正確的指代詞是「事」，而我們擷取到的是「事研究院方面」。

3.3 特徵萃取

如第一章的範例顯示，參照對象的語法結構可以是片語、小句或者數個小句。基於這個原因我們以小句做為參照對象的單位。根據表 7 語料統計的結果顯示 90%的參照對象在三個長句以內，因此我們只考慮這個範圍內的所有小句。當指代定詞之前有出現名詞或動詞時，例「*中東*這個地區」,「中東」也視為一個候選小句。

表 10 是我們提出的特徵，分成位置特徵、距離特徵、詞彙特徵和語義特徵四種類型共十個特徵。

表 10　候選小句特徵表

特徵類別	特徵描述
位置	候選小句與指代小句在相同長句
	候選小句與指代小句在相同小句
距離	候選小句與指代小句的長句距離
	候選小句與指代小句的小句距離
詞彙	候選小句的詞彙包含所有指代詞彙
	候選小句的詞彙包含部分指代詞彙
	候選小句的動詞包含指代小句的動詞
語義	候選小句與指代小句的相似度
	候選小句與指代小句的長句主題詞一致
	候選小句的詞彙是文章的高頻詞

每一個特徵的特徵值是零或一，用來表示候選參照對象是否符合此特徵。其中長句距離特徵、小句距離特徵和相似度特徵需要計算出門檻值。本研究將使用卡方檢定(chi-square)來決定特徵的門檻值。在進行詞彙比對或相似度計算時，使用到的詞彙是詞性標記為普通名詞、專有名詞、地方詞、時間詞與動詞五種，在比較時只考慮字是否相同並未考慮詞性是否相同。

參照對象和指代小句的相似度是利用公式(1)計算求得。

$$\text{Dice\_Coefficient} = \frac{2|C \cap A|}{|C|+|A|} \qquad (1)$$

C：候選小句的詞彙集合

A：指代小句的詞彙集合

計算詞彙集的交集方法是先將指代小句的詞彙進行相關詞集擴充，再計算候選小句的詞彙出現在相關詞集的次數。相關詞集是來自於三個外部資源，分別是同義詞詞林[23]、CKIP 辭典和 Google 搜尋結果。

同義詞詞林共有 77270 個詞彙，每個詞彙至少都有一個代碼，依照代碼可以將詞彙區分成大類、中類、小類、詞群和原子詞群五種類別。我們同樣以卡方檢定求出那一種類別的詞彙適合當相關詞集，下表 11 是同義詞詞林的詞集範例。

表 11　同義詞詞林的原子詞群範例

原子詞群代碼	詞彙
Di20C01	行為、行止、作為、行事、表現、行、所作所為、一言一行

CKIP 辭典共有 14935 個詞彙分屬於 13 個非實體類別，在擴充相關詞集時會選擇指代詞在類別中的詞彙，CKIP 辭典範例請參考表 8。

此外對於未出現在同義詞詞林與 CKIP 辭典的詞彙，我們以 Google 搜尋引擎的結果做為擴充詞彙的依據。搜尋的關鍵字是指代詞和指代詞的長句主題詞，取出搜尋結果前 100 個不重複的 snippets 進行斷詞和計算高頻詞，並同樣以卡方檢定求出適合的高頻詞數量來做為相關詞集擴充。

計算交集的公式如下：

$$|C \cap A| = \sum \text{Related}(c_i, a_j) \tag{2}$$

$$\text{Related}(c_i, a_j) = \begin{cases} 1, \text{if } c_i = a_j \text{ or} \\ \quad \text{CKIP}(c_i) = \text{CKIP}(a_j) \text{ or} \\ \quad \text{TONYI}(c_i) = \text{TONYI}(a_j) \text{ or} \\ \quad a_i \text{ in web}(C) \\ 0, \text{otherwise} \end{cases} \tag{3}$$

CKIP(x)　：詞彙 x 在 CKIP 辭典中的類別名稱

TONYI (x)　：詞彙 x 在同義詞詞林的類別

Web (C)　：指代詞 C 的搜尋結果

我們假設指代詞和參照對象的位置會是在文章的同一個段落，因此利用長句主題詞[23]為每一個長句擷取出一個主題，當參照對象的長句主題詞與指代小句的長句主題詞

相同時視為同一個段落。主題詞辨識在 188 個長句測試結果得到的正確率是 86.84%，召回率是 68.51，F-指標是 76.59%。

文章中重要的部分往往會被作者重複論述，相對的指代詞出現機會也會增加。因此我們計算整篇文章的詞頻並選取出高頻名詞與動詞集合視為最有可能被作者重複論述的目標。當候選參照對象的詞彙出現在高頻詞則表示有高頻詞特徵，高頻詞的詞彙個數決定方法是利用卡方檢定。

3.4 分類模組建立

我們從 879 篇文章隨機選擇出 80%的文章做為計算特徵的門檻值和建立分類器的訓練語料，在剩餘的 20%語料中選擇出參照對象在三個長句以內的實例做為測試語料，語料的分佈下表 12。

表 12　訓練與測試語料相關資訊

	文章篇數	實例總數
訓練語料	708	1226
測試語料	171	241

語料庫中指代詞的參照對象有小句和長句二種類型，我們的目的是找出參照對象在文章中的位置。辨識參照對象的方法分成二種，分別是建立分類器和計算特徵權重。本研究將用這二種不同的方式在相同語料上，找出最適合用於抽象指代消解的方法。語料中有 25%的指代詞實例有二個以上的參照對象，因此我們分別以第一次提及的參照對象和最接近指代詞的參照對象建立出二種分類器。接下來我們以「由前往後」來表示，以第一次提及的參照對象為正例所建立的分類器；「由後往前」來表示，以最接近指代詞的參照對象為正例所建立的分類器，負例是從第一次提及的參照對象之前的實例。如果實例的第一次提及的參照對象在文章的第一個小句，則不會產生負例。

使用 LIBSVM 分類工具分別針對以單一小句為單位、連續二小句為單位、小句/長句為單位和長句為單位訓練出四種分類器模組。每一個模組的參照對象正負例個數都不同，因此特徵門檻值需要分別計算。

每個分類器的正負例個數與特徵門檻值計算結果見下表 13、14。

表 13　由前往後分類器特徵門檻值

由前往後分類器					
特徵擷取單位	單一小句	連續二小句	單一小句(註一)	單一長句(註二)	單一長句
長句距離	≦1	≦1	≦1	≦1	≦1
小句距離	≦3	≦3	≦4		
相似度	≧0.21	≧0.26	≧0.52	≧0.54	≧0.76
文章高頻詞	前四個高頻詞	前四個高頻詞	前三個高頻詞	前五個高頻詞	前三個高頻詞
同義詞詞林類別	詞群	詞群	詞群	小類	小類
Web 詞群個數	50	50	50	100	50
正例個數	2848	2848	497	1006	1511
負例個數	2283	2283	361	879	1281

註一：人工標記的參照對象為小句所建立的正例。

註二：人工標記的參照對象為二個小句以上的實例視為長句，所建立的正例。

表 14　由後往前分類器特徵門檻值

由後往前分類器					
特徵擷取單位	單一小句	連續二小句	單一小句(註一)	單一長句(註二)	單一長句
長句距離	≦1	≦1	≦1	≦1	≦1
小句距離	≦3	≦3	≦3		
相似度	≧0.52	≧0.73	≧0.55	≧0.54	≧0.74
文章高頻詞	前四個高頻詞	前三個高頻詞	前三個高頻詞	前二個高頻詞	前三個高頻詞
同義詞詞林類別	大類	小類	詞群	中類	小類
Web 詞群個數	100	100	50	100	100
正例個數	2357	2357	790	806	1394
負例個數	1937	1937	633	645	1034

註一：人工標記的參照對象為小句所建立的正例。

註二：人工標記的參照對象為二個小句以上的實例視為長句，所建立的正例。

除了上述使用 LIBSVM 建立分類器之外，還使用權重計分法建立出另一種指代消解程序。權重計分法是先給予每一個特徵 0.1 分的權重，每個候選參照對象都會考慮是否符合十種特徵，特徵見下表 15。此方法會撰擇分數最高的候選參照對象做為起始邊界，因此從分數最高的小句到指代小句之間，我們視為參照對象。接著，我們再利用基因演算法模組來決定每一個特徵權重值，其中我們設定特徵值總合為 1，並觀察那些特徵會得到較高權重。

表 15　權重計分法特徵表

特徵	分數
候選小句與指代小句在相同長句	f_1
候選小句與指代小句在相同小句	f_2
候選小句與指代小句的長句距離小於 1	f_3
候選小句與指代小句的小句距離小於 3	f_4
候選小句的詞彙包含所有指代詞彙	f_5
候選小句的詞彙包含部分指代詞彙	f_6
候選小句的動詞包含指代小句的動詞	f_7
候選小句與指代小句的相似度大於 0.52	f_8
候選小句與指代小句的長句主題詞一致	f_9
候選小句的詞彙是文章的高頻詞之一	f_{10}

$$1 = f_1 + f_2 + f_3 + f_4 + f_5 + f_6 + f_7 + f_8 + f_9 + f_{10} \quad (4)$$

我們計算小句符合與長句符合的方式如下圖 2，其中「Sn」表示第 n 個長句，「Cn」表示第 n 個小句，A 表示指代詞出現的小句位置，參照對象是 C4 和 C5 小句，「○」表示分類器辨識為參照對象，「×」表示分類器辨識為非參照對象。

圖 2　參照對象辨識計算方式

每一個分類器和權重計分法都利用基因演算法進行特徵挑選，最後的結果如下表 16、17。

表 16　從前往後的回指辨識結果

辨識法	SVM 分類法				權重計分法
特徵擷取單位	單一小句	連續二小句	先小句後長句*	單一長句	單一小句
相同長句位置	1	1	(1,1)	1	0.1
相同小句位置	0	0	(0,0)	0	0
長句距離	0	0	(1,1)	0	0.1
小句距離	0	0	(1,0)	0	0
詞彙完全相同	1	0	(0,0)	0	0.1
詞彙部分相同	1	1	(1,1)	1	0.2
動詞相同	0	1	(1,1)	0	0.1
小句相似度	1	1	(1,1)	1	0.1
主題詞一致	0	0	(1,1)	0	0
高頻詞	0	0	(0,0)	1	0.2
小句符合	21.46/27.53	19.43/24.29	35		34.42/41.76
長句符合	43.72/47.37	44.94/46.56	38.19	42.9/43.27	57.49/58.3

註 1：*表示此方法僅列最佳特徵組合的辨識結果。

註 2：(A,B)表示 A 為小句分類器的特徵值，B 為長句分類器的特徵值。

註 3：X/Y 表示 X 為用所有特徵的辨識結果，Y 為最佳特徵的辨識結果。

表 17　從後往前的回指辨識結果

辨識法	SVM_分類法				權重計分法	
特徵擷取單位	單一小句	連續二小句	先小句後長句*	單一長句	單一小句	單一長句*
相同長句位置	1	1	(1,1)	1	0.1	0.1
相同小句位置	0	0	(0,0)	0	0	0
長句距離	0	0	(1,1)	0	0.1	0.1
小句距離	0	0	(0,0)	0	0	0
詞彙完全相同	1	0	(0,1)	0	0.1	0.1
詞彙部分相同	1	1	(1,1)	1	0.2	0.2
動詞相同	0	0	(0,1)	0	0.1	0.1
小句相似度	1	1	(1,1)	1	0.1	0.1
主題詞一致	0	0	(1,1)	0	0	0
高頻詞	1	0	(0,0)	1	0.2	0.2
小句符合	39/40.66	39/39	34.71		38.59/42.32	
長句符合	65.14/68.46	65.14/65.14	38.03	42.7/53.65	69.29/70.54	60.34

註 1：*表示此方法僅列最佳特徵組合的辨識結果。

註 2：(A,B)表示 A 為小句分類器的特徵值，B 為長句分類器的特徵值。

註 3：X/Y 表示 X 為用所有特徵的辨識結果，Y 為最佳特徵的辨識結果。

3.5 實驗結果與分析

「從前往後」的正確率偏低的原因是，當文章內容出現多個與指代詞相關的描述時，最接近指代詞的描述句才是正確的答案，如下例。其中 C9 小句出現指代詞「問題」，答案是 C7 小句。若「從前往後」進行指代消解時，會先找到 C5 小句，但該小句並非是指代詞所指示的對象。

例：C1：至於資金方面，

C2：謝森中則表示，

C3：由於全球性資金緊俏，

C4：利率上揚是無法避免的趨勢，

C5：但為了顧及國內產業融資問題，

C6：央行將維持現行適度寬鬆的貨幣政策。

C7：*至於外傳央行與立委們將在會中討論中央銀行改隸總統府一事*，

C8：謝森中則表示，

C9：餐會中並沒有討論這個問題，

使用「單一小句」所建立的分類器其正確率比其它分類器好，它的長句符合正確率是 68.46，而它的小句正確率偏低的原因是，候選小句的詞彙與指代小句的詞彙之間沒有關聯，如下例，其中 C1 到 C5 是 C6 的參照對象。

例：C1：*中華奧會之所以打算走向結合企業的發展*，

C2：*主要希望能以民間團體的性質*，

C3：*改變奧會的體質*，

C4：*大力推動體育休閒的正當社會風氣*，

C5：*提供較多的選手和教練就業和福利*，

C6：目前這個方案正在策畫階段…。

候選小句 C5「提供較多的選手和教練就業和福利」與指代小句 C6「目前這個方案

正在策畫階段」二者的詞彙沒有關聯。而候選小句 C4「大力推動體育休閒的正當社會風氣」的詞彙「社會」與「風氣」與指代小句 C6 的「階段」在同義詞詞林的相同類別、候選小句 C4 的「社會」詞彙是文章的高頻詞之一、候選小句 C4 的「推動」詞彙是搜尋結果的高頻之一，因此 C4 被選為答案。但 C5 也是答案之一，所以在此例中只有長句符合。

使用「單一長句」所建立的分類器是將參照對象以一個長句為單位，所以當實際的參照對象只是一個名詞片語時，會包含太多不相關的詞彙，使得辨識率大為降低。

使用「先小句後長句」所建立的二個分類器，其辨識正確率只有 36%左右的原因是，沒有辦法判斷指代詞回指的參照對象類型是小句或長句。若參照對象類型是長句，當小句分類器辨識出小句正例，即使小句正例在長句答案範圍內，仍算是錯誤。

權重計分法比單一小句分類器佳的原因是權重計分法會撰擇出與指代詞最相關的句子，而分類器只能判斷相關或不相關，如下例。

例：C1：ＰＣ基礎課程深獲同仁迴響將再度開課・中心日前舉辦的ＰＣ基礎課程如何熟悉及操作您的ＰＣ？

C2：深獲同仁熱烈迴響，

C3：謹在此向各位使用者致謝。

C4：然在感謝之餘，

C5：*中心亦警覺到ＰＣ基礎課程的重要性*，

C6：*故預計於下年度（８３上半年再次開設相同的課程）*，

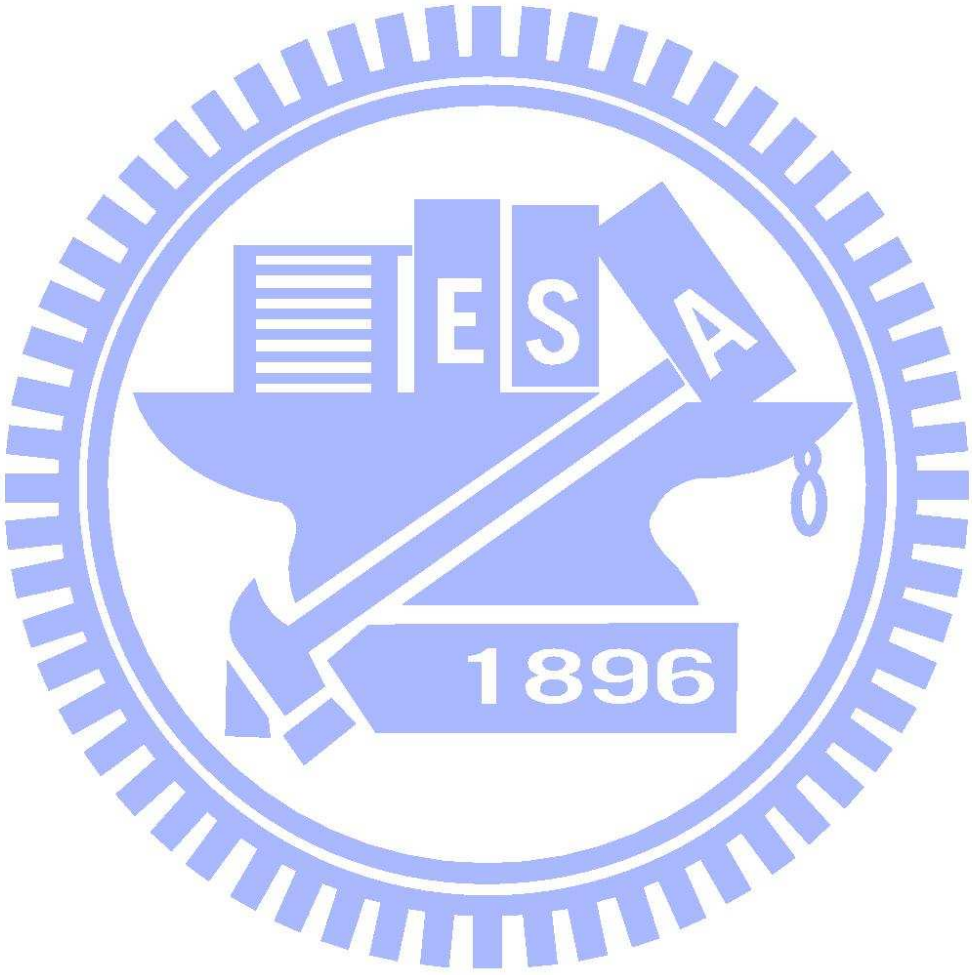
C7：以滿足同仁的需求。

C8：這次ＰＣ基礎課程一系列的講座，

權重計分方法找到得到最高的 C6 小句，C6 小句的「課程」符合部分詞彙特徵、C6 小句的「課程」是文章高頻詞之一。而分類器會先找到 C7 小句並視為正例，原因是 C7 小句的「同仁」是高頻詞之一。C6 和 C5 小句也是分類器的正例。

分類器方法與特徵權重方法都判斷錯誤的情況是，報導性質文章會論述到人名、疾病名稱、諺語等問題，如「『天才是九十九分的努力加一分的才氣』這種話銘記在心…」，

其中「這種話」指的是前面的「天才是九十九分的努力加一分的才氣」。或者是如「在大清帝國這個時代中..」，其中「這個時代」指的是「大清帝國」。



第四章　結論與未來工作

本論文研究二種不同方法來消解抽象名詞指代，利用不同的參照對象建立分類器來處理抽象名詞指代會遇到的問題。經實驗數據分析顯示，以權重計分法在處理抽象名詞指代時，可以得到較好的消解結果。

分析成果與貢獻如下：

1. 建立抽象名詞指代語料
2. 分析抽象名詞的參照對象類型
3. 利用自動化特徵擷取方式消解抽象名詞指代

後續研究有以下幾種方向：

1. 分析文章結構：將文章的內容依照主題與次主題建立出樹狀結構，此方法可以改善候選參照對象的邊界問題。
2. 加入句法結構特徵：使用中文剖析樹找出詞彙的句法位置，如主詞、受詞或附屬小句位置。將這些特徵應用於分類器上，增加指代詞回指名詞片語的辨識正確率。
3. 提升名物化動詞辨識率：使用規則方式並無法辨識出所有名物化動詞情形，導致無法辨識出真正的指代詞。
4. 提升指代詞辨識率：我們在擷取指代詞時發現當指代詞前出現子句時，有限狀態機並無法正確擷取出指代詞。因此使用剖析樹判斷出子句可以幫助我們提升辨識率。

參考文獻

[1] 孔芳，周國棟，朱巧明，錢培德，”指代消解綜述”，計算機工程，第 36 卷第 8 期，2010。

[2] Chih-Chung Chang, Chih-Jen Lin, LIBSVM : a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[3] Renata Vieira, Susanne Salmon-Alt, Emmanuel Schang, “Multilingual Corpora Annotation for Processing Definite Descriptions”, In Proceedings of the Third International Conference on Advances in Natural Language Processing(PORTAL), Faro, Portugal, LNAI 2389, pp. 249-258, June 2002.

[4] 熊學亮，劉東虹，”論証文中抽象實體的回指研究”，四川外海學院學報，第 23 卷第一期，2007。

[5] 熊學亮，劉東虹，”論証文中抽象實體的回指研究”，四川外海學院學報，第 23 卷第一期，2007。

[6] 王道英，韓蕾，“「這」、「那」類的隱性回指”，西南民族大學學報（核心期刊），第 4 期，2006。

[7] Donna K. Byron, “Resolving Pronominal Reference to Abstract Entities”, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), pp. 80-87, 2002.

[8] Nicholas Asher, Reference to abstract objects in discourse, Kluwer Academic Publisher, 1993.

[9] Massimo Poesio, Tomonori Ishikawa, Sabine Schulte im Walde, Renata Vieira, “Acquiring Lexical Knowledge for Anaphora Resolution”, In Proceedings of the 3RD Conference on Language Resource and Evaluation(LREC), Las Palmas, May 2002.

[10] Michael Strube, Stefan Rapp, Christoph Muller, “The Influence of Minimum Edit Distance on Reference Resolution”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, Penn., 6-7, pp. 312-319, July 2002.

[11] Razvan Bunescu, “Associative Anaphora Resolution: A Web-Based Approach”, In Proceedings of EACL 2003 workshop on The Computational Treatment of Aanphora, Budapest, pp. 47-52, 2003.

[12] Gasperin Caroline and Vieira Renata, “Using Word Similarity Lists for Resolving Indirect Anaphora”, In Proceedings of Association for Computational Linguistics Workshop on Reference Resolution and its Applications, pp. 40-46, 2004.

[13] Massimo Poesio, Rahul Mehta, Axel Maroudas, Janet Hitzeman, “Learning to Resolve Bridging References”, In Proceedings of Annual Conference for Association of Computational Linguistics, pp. 143-150, 2004.

[14] Michael Strube, Christoph Müller, “A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue”, In Proceedings of the 41st Annual Meeting of Association for Computational Linguistics(ACL), pp. 168-175, 2003.

[15] Costanza Navarretta, “Resolving Individual and Abstract Anaphora in Texts and Dialogues”, In Proceedings of the 20th International Conference of Computational Linguistics (COLING), Geneva, Switzerland, pp. 233-239, 2004.

[16] Donna K. Byron, James F. Allen, “Resolving Demonstrative Anaphora in the TRAINS93 Corpus”. In New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2), pp. 68-81, 1998.

[17] Miriam Eckert, Michael Strube, “Dialogue Acts, Synchronizing Units and Anaphora Resolution”, Journal of Semantics 2000, 17, pp. 51-89, 2000.

[18] Christoph Muller, Stefan Rapp, Michael Strube, “Applying Co-Training to Reference Resolution”, Porceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp352-359, July 2002.

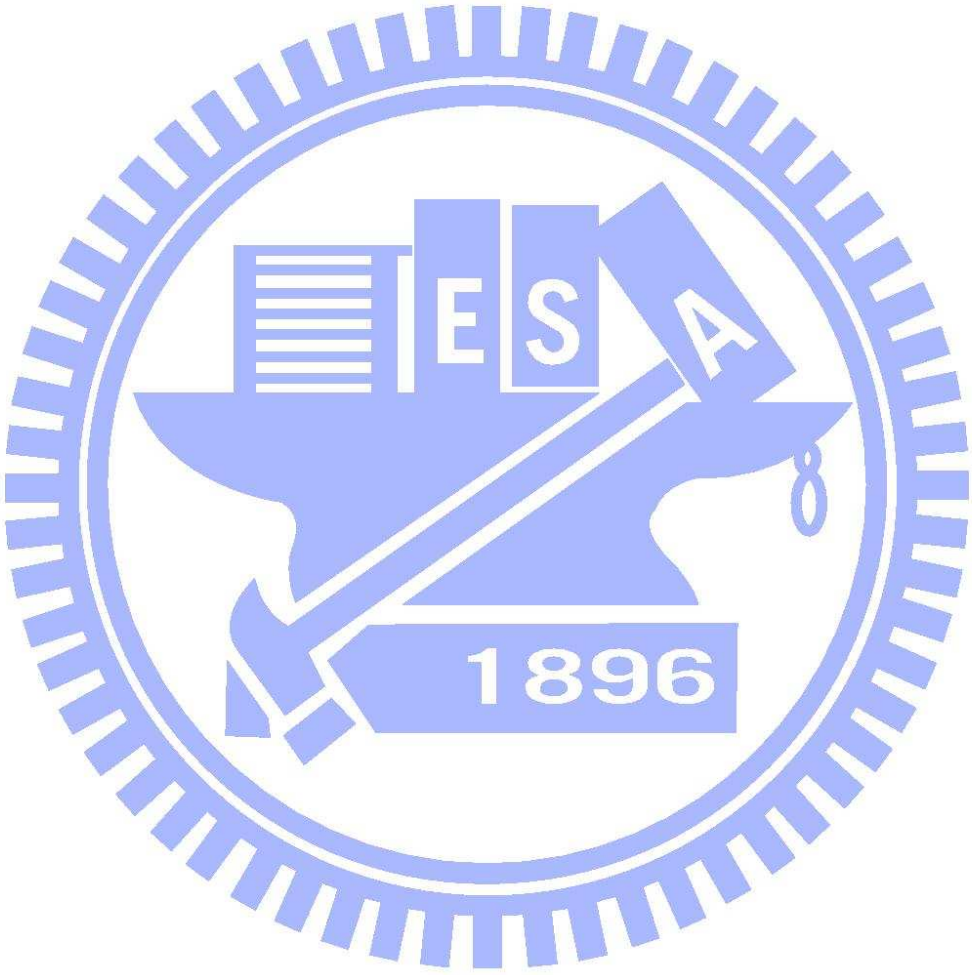
[19] 馬偉雲，”中文動詞名物化判斷的統計式模型設計”，ROCLING XVIII: Conference on Computational Linguistics and Speech Processing (ROCLING), Hsinchu, Taiwan, 2006。

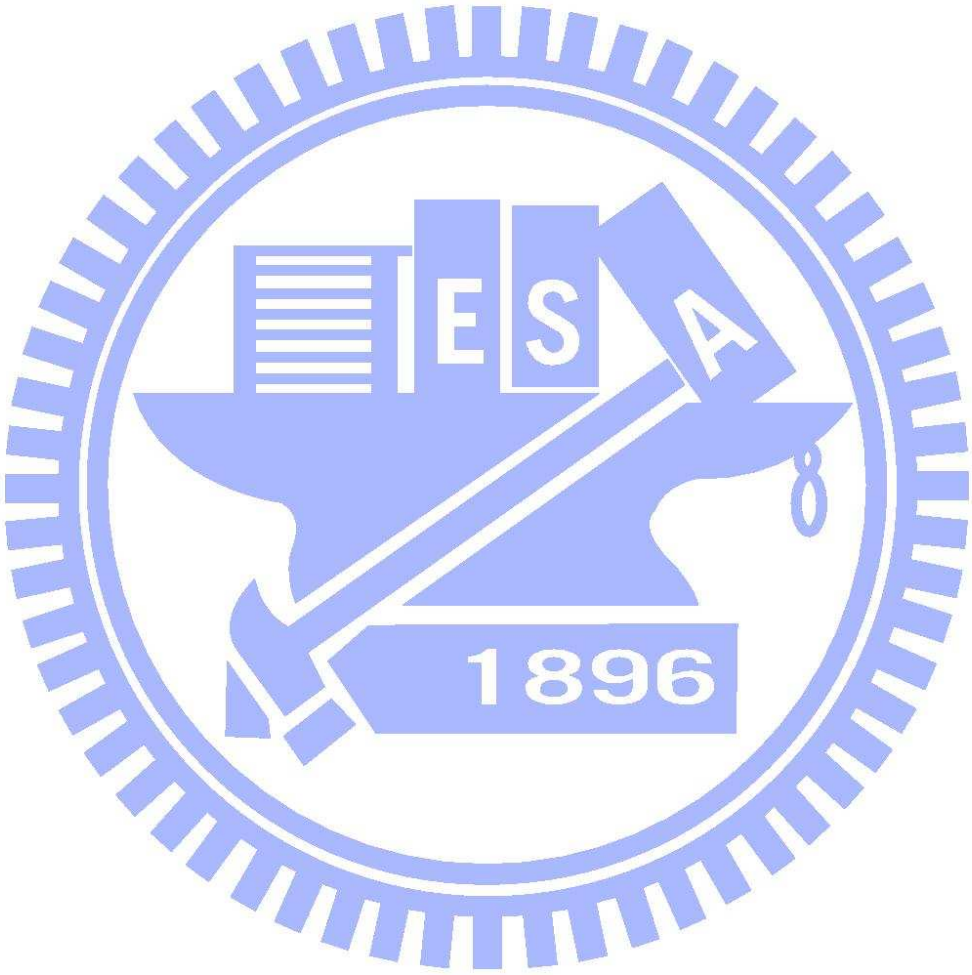
[20] 中央研究院詞知識庫小組，技術報告 9305：中文詞類分析-第三版，1993。

[21] 游基鑫，“中文資訊擷取環境建構與同指涉問題之研究”，台灣大學，碩士論文，2000。

[22] Tyne Liang, Shan-Chun Pan, Kwan-His Chen, “Sentence-based Topic Identification and Its Applications in Chinese Texts”, National Computer Symposium, Taipei, Taiwan, 2009.

[23] 梅家駒, 竺一鳴, 高蘊琦, 殷源翔，同義詞詞林，臺灣東華書局股份有限公司，1997年3月。




E
S
A
1896