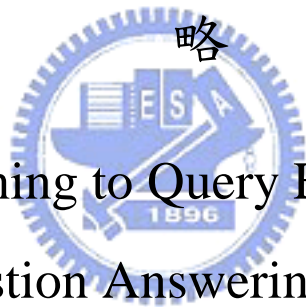


國立交通大學

電機資訊學院在職專班數位圖書資訊組

碩士論文

中文問答系統—以網路為基礎之查詢詞擴充策



Web-Based Learning to Query Expansion for Chinese
Question Answering System

研究生：許長進

指導教授：梁 婷

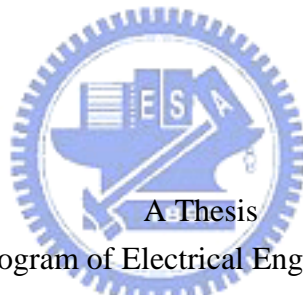
中華民國九十五年六月

中文問答系統—以網路為基礎之查詢詞擴充策略
Web-Based Learning to Query Expansion for Chinese Question
Answering System

研究生：許長進
指導教授：梁 婷

Student : Chang-Chin Hsu
Advisor : Tyne Liang

國立交通大學
電機學院與資訊學院專班 數位圖書資訊學程
碩士論文



Submitted to Degree Program of Electrical Engineering and Computer Science
College of Computer Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science
in
Digital Library
June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

中文問答系統—以網路為基礎之查詢詞擴充策略

研究生：許長進

指導教授：梁 婷

國立交通大學 電機學院與資訊學院 數位圖書資訊學程（研究所）碩士班

摘 要

網路有如資訊的海洋。然而遺憾的是，人們在網路中尋找自己感興趣的答案常如大海撈針。傳統的關鍵詞檢索方式多無法解決使用者的查詢意圖。本論文提出一個應用網路資料和查詢詞擴充技術的中文問答系統。我們提出法則式的問句樣式機制以分析問句的意圖。另一方面，有別於一般中文問答系統擴充詞多來自事先所設定的相關詞資料庫，本論文所提的查詢詞擴充技術乃是應用現成的網路語料，進行相關詞探勘。我們利用對應演算法將訓練問句和搜尋結果進行非名詞關鍵詞與查詢詞擴充。

爲了檢驗所提的方法，我們以 383 個問句做為訓練資料，進行查詢詞擴充探勘，並另以 80 個問句作測試，所得到的搜尋結果比一般關鍵詞搜尋在使用者所需要閱讀的篇數明顯減少，實驗結果顯示系統效能為每個問題所花的 human effort 2.03 和 MMR 0.765。

關鍵詞：中文問答系統、問句類型分析、查詢詞擴充、關鍵詞擴充

Web-Based Learning of Query Expansion for Chinese Question Answering System

Student : Chang-Chin Hsu

Advisor : Dr. Tyne Liang

Degree Program of Electrical Engineering and Computer Science

National Chiao Tung University

ABSTRACT

Searching in the Web is just like searching in a sea. Traditional query resolution is based on inefficient keyword search. This thesis proposes a Chinese question answering system by using web corpus and query expansion. We propose a rule-based query processing method to detect the query type. On the other hand, we propose new query expansion which is unlike traditional one based on predefined thesaurus. The presented query expansion is based on web corpus by aligning the training questions and the search-results returned from a search engine.

In order to verify the proposed method we use 383 questions for training and 80 questions for testing. The results show that the proposed expansion technique yields better performance than the keyword-based search in terms of less human efforts per question 2.03 and MMR 0.765.

Keyword : Chinese question answering system 、 question type extraction 、 query expansion 、 keyword expansion

誌 謝

首先要感謝我的指導教授—梁婷老師，身為一個在職專班的學生，工作與學業要兼顧是一件很困難的事，並且我的工作與經驗也都非相關領域，因此對自然語言與資料擷取的技術並不熟悉，但是老師確不厭其煩的指導我，並帶我進入這一個領域，學習到更有效的方法去執行論文研究，因此我真的很感謝指導教授—梁婷老師

其次要感謝資料擷取研究室中每一位學長姐與同學的幫助，幫我找出論文中不夠完善之處，並且給予建議。更要感謝黃立泓同學的幫助，適時的伸出援手，幫我解決許多相關的問題。

還有感謝大學同學杜宗原，幫助我修飾我的論文。還有每一位數圖組老師的教導與同學的幫助，讓我完成我的論文。最後感謝中央研究院詞庫小組所提供的中文斷詞系統。



目錄

中文摘要	i
英文摘要	ii
誌謝	iii
目錄	iv
表目錄	v
圖目錄	vi
第一章 緒論	1
1-1 背景	1
1-2 中文問答系統組成要素.....	1
1-3 動機.....	2
第二章 相關研究	4
第三章 系統學習問句中查詢詞的轉換.....	7
3-1 系統架構	7
3-2 問句斷詞與擷取答案段落	9
3-3 問句分析	9
3-3-1 疑問詞分類	9
3-3-2 萃取問句樣式.....	10
3-4 分析答案段落與查詢詞和關鍵詞轉換擴充.....	12
3-4-1 使用文字組合的技術去學習非名詞關鍵詞與查詢詞的轉換.....	12
3-4-2 鄰近等級.....	14
3-4-3 結合鄰近等級與配對等級.....	15
3-5 產生查詢句.....	16
第四章 實驗與分析	18
4-1 實驗步驟	18
4-2 擷取問句樣式的成效	18
4-3 實驗結果與評論非名詞關鍵詞與查詢詞擴充.....	20
第五章 未來工作	23
參考文獻	24
附錄一 測試的問句與答案.....	26
附錄二 問句法則例句.....	29

表目錄

表 1：問句與答案，輸入與輸出範例.....	9
表 2：疑問詞分類.....	10
表 3：問句法則.....	11
表 4：候選擴充詞與關鍵詞的配對等級範例.....	14
表 5：問句答案類型是”什麼地方 生長”的鄰近詞加總.....	15
表 6：將表 4 與 5 結合取得加權平均值的最後等級.....	15
表 7：非名詞關鍵詞與查詢詞轉換擴充.....	17
表 8：訓練資料與測試資料問句類型統計.....	18
表 9：訓練資料問句數量、問句樣式及從搜尋引擎擷取的答案段落統計.....	18
表 10：問句樣式評論結果（一）.....	19
表 11：問句樣式評論結果（二）.....	19
表 12：前五個問句和問句樣式與評論.....	19
表 13：各類查詢句的 MRR 值與改進程度.....	20
表 14：各類查詢句的 MMR 值依問句的類別分類.....	20
表 15：查詢詞擴充在 Recall at R 與 Human Effort 的結果.....	21
表 16：評論非名詞關鍵詞與查詢詞擴充 HE per question 的結果，依問句的類別分類.....	22



圖目錄

圖 1：一般的問答系統.....	1
圖 2：線上中文問答系統架構圖.....	7
圖 3：資料訓練系統流程圖.....	8



第一章 緒論

1-1. 背景

網路是資訊的海洋，令人遺憾的是，人們在網路中尋找到自己真正感興趣問題的答案確如大海撈針。通常使用者在搜尋引擎的文字框中輸入關鍵詞，搜尋引擎按照可信度的大小順序取回一系列含有該關鍵詞的文字資料與 URL(Uniform Resource Locator)。使用者再依次瀏覽每一個 URL，以找尋真正需要的資料。(關毅 et al., 2004)。

資料搜尋傳統的方式是依賴關鍵詞做搜尋，通過關鍵詞檢索來找尋所要的資料及文件。但是，關鍵詞檢索並不能真正立即獲取所要的知識或是立即回答使用者的問題，所以我們需要新的技術來加強檢索的功能。近年來問答系統在資料檢索、自動學習、自然語言處理等研究領域都受到了注意(Huang et al., 2004)。

問答系統可以根據用戶的問題找出確切的答案。我們希望這個答案不只是與問題所搜尋出來的整個文獻有關，而是更精確地滿足用戶需求文獻的一部份或一句話(李季, 2004)。

1-2. 中文問答系統組成要素

一般的問答系統包含三個構成要素(鄭實福 et al., 2002; 李季, 2004)見圖 1：問題分析、資訊檢索、答案抽取。

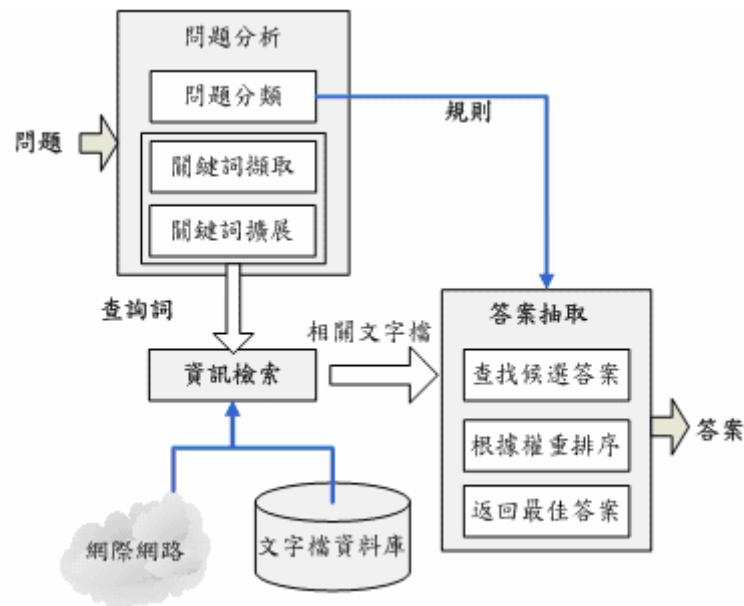


圖 1 一般的問答系統

中文問答系統先將問句斷詞。問句分析的目的在擷取可以明確表達問句的問句樣式 (Question Pattern)，並分析出問句類型 (Question Type) 類型包含人、事物、時間、地點、數量等。之後在擷取問句的關鍵詞，並將關鍵詞擴展，通過問題分析而取得到的關鍵詞集需提交資訊檢索系統中檢索。

資訊檢索的任務就是將取得的關鍵詞集至文件資料庫中搜尋出相關的文件資料。或也可以使用網路上的搜尋引擎如：Google，將關鍵詞集送至搜尋引擎做搜尋擷取出結果。在 TREC 會議中就不要求每個問答系統都要有自己的資訊檢索模組 (鄭實福 et al., 2002)。將這些文件資料送至答案抽取中處理。

在答案抽取這一個階段中必須處理這些網頁，抽取出相關的答案，這些答案可能是一句話或是一段文字，依據問句類型與這些預選答案中的相關度計算出權重，並根據權重排序找出最佳的答案。

1-3. 動機

一般找尋答案的方式是使用關鍵詞搜尋，但是關鍵詞檢索並不能真正立即獲取所要的知識或者立即回答使用者的問題 (Huang et al., 2004)。所以我們產生了動機，設計一個有效的方法去分析自然語言問句，以問句樣式中的主要詞為查詢詞，並擷取出問句的關鍵詞，這些關鍵詞又分為名詞關鍵詞與非名詞關鍵詞，因為名詞在問句中有其代表性，將其轉換會影響答案搜尋結果，所以我們將選出非名詞關鍵詞與查詢詞做擴充並搜尋答案。

我們以問句”岳飛在哪裡出生？”做範例，使用關鍵詞搜尋擷取出”岳飛”與”出生”作為關鍵詞，並送至 Google 做搜尋，我們會發現所回傳的結果並不是很正確。如果我們找出問句樣式為”哪裡出生”，查詢詞為”出生”並且轉換成”出生於”，再送至 Google 做搜尋將會得到更好的結果。

論文中我們將挑戰一般人們最常使用的關鍵詞搜尋，設計一個中文問答系統，系統分為兩個部分：線上中文問答系統與訓練系統，在訓練階段中，訓練系統時會先自動將中文問句斷詞並至網路上擷取答案片段，分析問句擷取出每一個問句的問句樣式，將答案片段中在答案附近的鄰近詞萃取出來，最後再根據語言學與統計的技術計算出查詢詞、非名詞關鍵詞與鄰近詞的鄰近等級 (Proximity Rank) 與配對等級 (Alignment Rank)，根據權重計算出每一個查詢詞、非名詞關鍵詞與鄰近詞的相關等級。在執行階段中，線上中文問答系統會將自然語言的問句分析，擷取出問句樣式，分別找出與查詢詞、非名詞關鍵詞相關等級最高的詞作為擴充詞。

在第二章中主要是討論國內外相關論文與技術，第三章介紹中文問答系統的架構與

原理、流程並且分析問句做非名詞關鍵詞與查詢詞的轉換，第四章敘述實驗流程與分析結果，蒐集問句輸入系統中，分析結果並評論之。第五章說明未來工作。



第二章 相關研究

目前問答系統的研究可分為三大類：基於常問問題集的問答系統，基於百科知識的問答系統，開放領域的問答系統(秦兵 et al., 2003)。依據答案來源的研究方向可以分為以網際網路上的資訊為潛在的答案來源以及以 TREC 的 Q&A 測試文集為答案來源等兩大類(Huang et al., 2004)。問答系統處理的三個步驟：問題分析、資料檢索、答案抽取(鄭實福 et al., 2002; 李季, 2004)。本論文中，我們主要是針對中文問句的分析與處理，利用網路上的資源進行訓練與擷取答案。

問題分析包含問句類型分類與查詢詞轉換，這些對擷取答案段落與文件是很重要的。超過 70% 的錯誤歸因於無效的問句類型分類、關鍵字選擇、查詢詞擴充(Moldovan et al., 2002)。所以在我們的中文問答系統提出一個有效的問句類型分類規則，依據疑問詞的類別將問句分類，在查詢詞擴充上使用鄰近等級與配對等級加權計算出相關等級，並選出前兩名相關度最高的詞為擴充詞。

在問答系統中使用到的有效技術有(Lee et al., 2005)：

1. 斷詞與標記
2. 具名實體辨識
3. 完全與淺的語法分析
4. 語義角色/關聯標明



論文中的問答系統使用到 1 與 4 項技術，斷詞是使用中央研究院的中文斷詞系統同時也會自動給詞類標記，並使用語意角色的技術將疑問詞做分類，擷取問句答案的類型。

目前中文資料檢索仍然面對以下需要解決的難題(Huang et al., 2004)：

1. 使用文件中出現的字詞或簡單的數學式無法輕易地表達概念，如何決定或表達隱藏於文件中的知識是相當困難的。
2. 一般文件幾乎是中英文共存，因此中文檢索系統需提供處理雙語的能力。
3. 中文檢索在處理斷詞方面不同於西方文字，中文句子的斷詞要困難得多。

我們所開發的中文問答系統，針對第 3 項上我們開發了一個學習的功能，在中文斷詞發生斷詞錯誤時，可以使用人工的方式建立新詞，再下一次斷詞時系統就會自動修正，彌補中文斷詞的錯誤。

我們將答案段落中在答案鄰近的字切開使用「N-gram」索引法(曾元顯, 2004)，N-gram 為文件中任意 N 個連續字元，「中國社會」N=2 時產生「中國」、「國社」、「社會」三個索引詞，可排除或降低「字元法」中類似「中國」與「國中」的字串順序問題，可省去「詞彙法」中維護詞庫的煩惱，在研究中是將 N=2 and 3，因為 2grams 時重複性太高，很容易不相關的詞，切詞後產生重複，當 3grams 會減少這種問題的出現機率。

中文的問答系統中，問句使用自然語言處理去取得文件，擷取查詢詞與查詢詞擴充是影響搜尋結果的主要關鍵。在資訊檢索中，查詢詞擴充通常是基於查詢詞的語意關係，查詢詞擴充根據問題的特性可以使精確度提升。中文問答系統中查詢詞擴充是根據問句的類型(Yu et al., 2005)。

例如：

Q2: 泰山的**高度**是多少？(How high is the Taishan?)

問句類型: **高度**(Num_height)

查詢詞: {泰山 **高度**}

查詢詞的擴充詞: {千米, 米, 英尺, 厘米}

系統特色是依據查詢詞的類型擴充，根據系統預先所設定的擴充詞擴充，系統的精確度取決於所設定的擴充詞。在我們的系統中也是使用查詢詞擴充的方式，但是擴充的相關詞來源不同，相關詞是由搜尋引擎所擷取的答案片段分析統計所得到的。

網頁搜尋引擎問答(王雅婷, 2003)，研究中提供兩種問答方式，一是提供使用者使用自然語言查詢。二是提供關於人與物的問題型態選擇題。先將使用者的自然語言問題分類，型態大略可分為人、事、時、地、物，後再將分類後的問句做適當的轉型，當網路在進行資料擷取時可傳回更適合與更相關的資料。例如：“陳水扁的老婆是誰？”由於“老婆”一詞較口語化，網路上的尊稱多為“妻子”、“女士”。在問句分析中將問句的停詞與常用詞挑出，找出關鍵詞，將查詢字串做轉型，例如：問題分類為“人”，判斷查詢詞中是否有“寫”的關鍵詞，如果有將“寫”轉成“作者”。處理好的查詢關鍵詞使用 GAIS 搜尋引擎進行網頁資料的擷取。取回的資料計算其特有詞的數量，計算出比重與分數並找出答案。研究中使用的轉型與查詢詞的擴充其意義相同，都是使用相關詞。分類後依據問句型態找出相關詞對關鍵詞作轉型擴充是這一個研究的特點。

所謂的停詞(Stop word)往往是一些文法用詞，或者是些常見但不重要的名詞等等，這些停詞出現的頻率很高，容易影響分類或斷詞的準確性，例如：的、與、是...等。常用詞(Common word)若有一串詞網頁中出現次數很高，代表此字串可能代表某種語文上的意義以致常被使用，有些常出現的字也是，如果可以去除掉這些詞，就可以留下有用的詞(王雅婷, 2003)。我在系統中是把停詞與常用詞結合起來統稱為停詞，分成兩部份利用，在擷取問句樣式時系統會將這些這些詞移除，但是在鄰近等級與配對等級時是使用加權平均法來做計算，把含有這些詞的詞等級降低，使真正有用的詞等級上升。

問答系統之查詢擴充—以網路資源為本的非監督式學習策略(王怡嘉, 2004)。在訓練過程中，必須先自網路上收集所需的訓練資料，依據語言學的知識及統計學上的技術從問句訓練資料中抽取出可清楚表達答案類別的問句樣式，最後再透過問句與答案段落之間的關聯性統計以計算出最適合每個問句樣式的擴充查詢詞組。在執行階段，輸入的問句之問句樣式會被轉換成其對應的擴充查詢詞，用以增加搜尋引擎擷取出答案的機會。

其特色是問句所使用的擴充詞並非系統建置者所預設的，擴充詞是由搜尋引擎所得到的結果分析而來，因此所得的擴充詞有助於對搜尋答案效能的提升。在我們的系統中學習此研究中查詢詞擴充的方法，並且把非名詞的關鍵詞也擴充，增加系統效能。

基於常問問題的中文問答系統(秦兵 et al., 2003)，系統必須先建立一個候選的問題集，然後計算使用者所提問句的語意相似度，在候選的問題集中找尋相似的問句，並將答案返回給使用者。在實驗中發現基於關鍵詞的句子相似度計算，基於語意的句子相似度計算提高了問題匹配的準確率。基於常問問題的中文問答系統可作為其他的問答系統的一部份，當使用者提出問句時，可以先去資料庫中找尋是否有相關的題目，如果有就可以直接把答案返回給使用者。這樣可以減少系統的處理問句的數量，提高系統效率。系統的精確度取決於候選問題集的範圍與類別、數量。

以網際網路內容為基礎之問答系統”Why”問句研究(沈天佐 et al., 2003)，問題的答案是”原因”，”原因”有不同的型態，可能是一句片語、一個子句、一個句子，甚至跨越句子的範圍。答案文件的取得，是將問句的停詞與標點符號全部移除，使用”AND”的方式至 Google 搜尋，所以文件必須包含所有的關鍵詞，並在文件中找尋表達因果關係的資訊，使用因果 patterns 比對與 Penn Treebank 之 PRP 標記來擷取答案。這類研究方式的特色是使用詞意分析與標記，並針對問題的因果關係分析文件擷取答案。

開放領域的中文自動問答系統(林川傑, 2004)，自動回答非限定領域的自然語言問句，中文問句類型定義有十一類，問句類型分類規則共有 136 條，針對不同的問句類型，論文中提出不同的候選答案找尋方式。短答問題的候選詞多半來自於具名實體辨識系統的輸出，或是在語意辭典中的下位詞等。長答問題的候選答案則是藉由特定句型所抽取出的詞組，並可針對特別問句類型限制組中所帶的語意等。不同的問句類型亦有著不同的候選答案排名策略。各種分數計算、權重設定及排名策略等都將由實驗結果來決定最好的組合。中文問句類型定義與分類，候選答案分數的多層次排名策略、以及短答問題及長答問題的可能候選答案為何為此系統的特色。我們學習這一個論文中的將問題分類與建立問句類型分類規則，分為 4 類 24 條問句類型分類規則。

在我們的研究方法上，分析自然語言問句取得問句樣式與關鍵詞，資訊檢索以 Google 為資源並擷取出答案段落，最後再透過問句與答案段落之間的關聯性統計以計算出最適合每個問句樣式與關鍵詞的擴充查詢詞組，增加查詢詞組至搜尋引擎擷取答案的機會。

第三章 系統學習問句中查詢詞的轉換

3-1 系統架構

在本論文中我們將提出以非名詞關鍵詞與查詢詞轉換擴充的線上中文問答系統，此系統共分為兩個部份，第一個部份是線上中文問答系統，架構圖如圖 2，第二部分是學習非名詞關鍵詞與查詢詞轉換擴充的訓練系統，架構圖如圖 3。

系統以網路為基礎，使用語言學與統計學的理论去分析處理問句，並且根據問句和答案與問句類別處理後自動擷取出問句樣式再擷取出含有答案與關鍵詞的答案段落，利用這些答案段落使系統能夠自動去學習，分析這些答案段落與關鍵詞、查詢詞之間的關係，有效的將非名詞關鍵詞與查詢詞做擴充，增加系統至搜尋引擎搜尋答案的精確度。

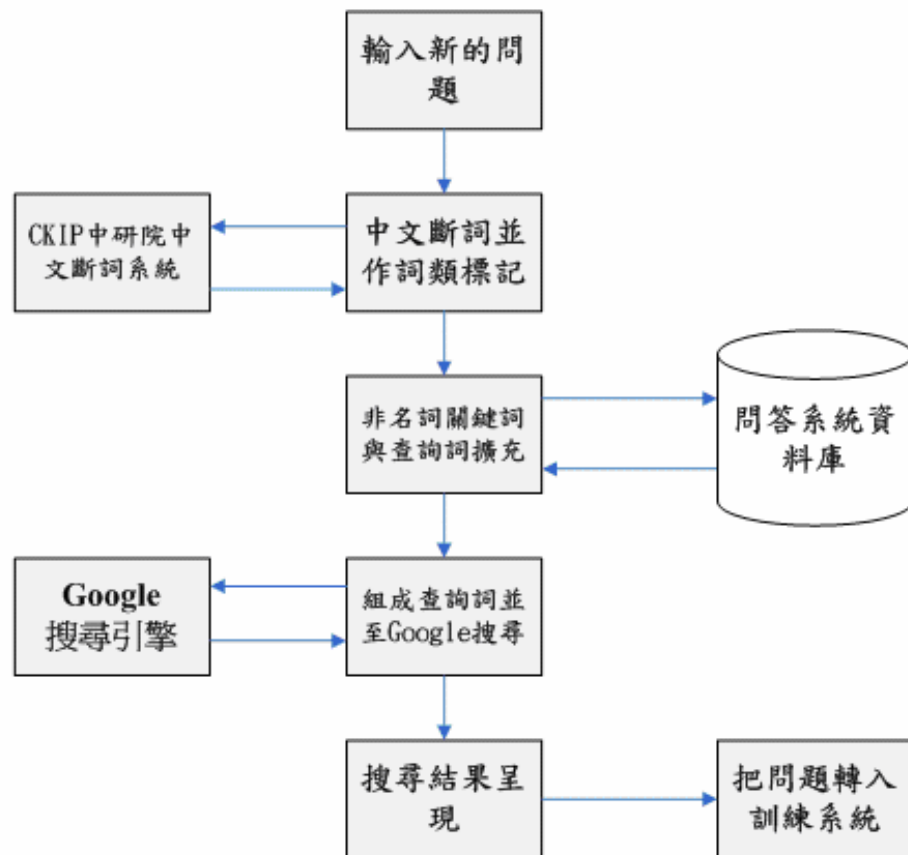


圖 2：線上中文問答系統架構圖

當線上中文系統取得問句的搜尋結果後可以把問句轉換到學習資料中，訓練系統會把這一個新的問句進行處理。

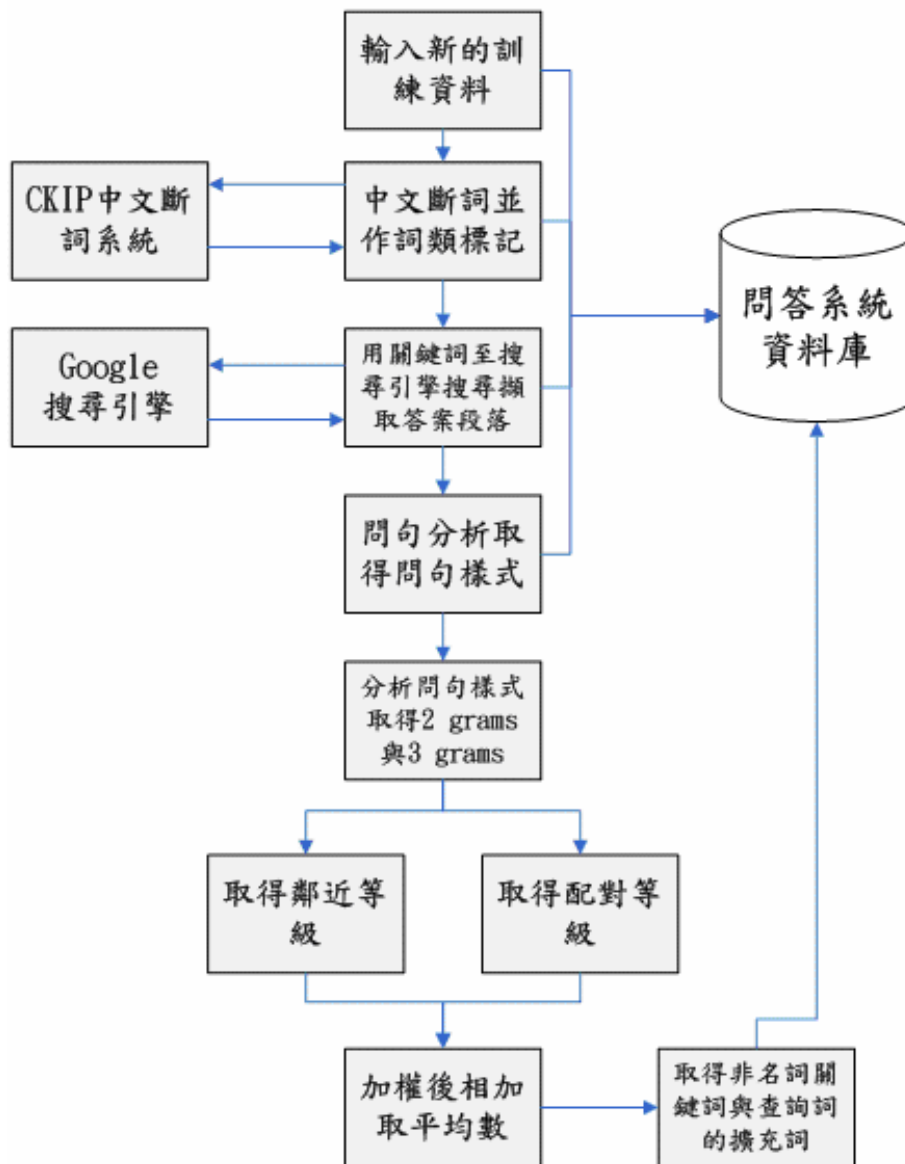


圖 3：資料訓練系統流程圖

系統主要處理的問句類型分為五類：人、時間、地點、數量、事與物。針對答案單一與絕對的問題做處理。

我們將系統分為四個階段，在 3-2 中我們將描述如何將問句做斷詞並將關鍵詞送至搜尋引擎擷取出答案段落。3-3 中我們將描述如何分析問句依據我們所制定的規則萃取出問句樣式。3-4 中我們將分析問句的答案段落，並且有效的擷取出查詢詞與關鍵詞的擴充詞。最後在 3-5 中我們將擷取的結果處理後組合成搜尋引擎所能處理的方程式。

3-2 問句斷詞與擷取答案段落

搜集大量的問句與答案作為系統的學習資料。利用中文斷詞系統做標記，所給的詞類標記是根據中研院平衡語料庫詞類標記集之簡化詞類與線上斷詞服務採用之精簡詞類。當有時中文斷詞系統所斷的詞會有錯誤，所以在這一個階段中，系統會出現一個建立新詞的服務，可以修正錯誤的斷詞，只要做過一次修正後下次遇到相同的詞句就會自動做修正。當系統回收的資料量越多時同時系統在問句斷詞的精確度也會升高。當問句中含有特殊的名詞時使用者可以使用”[“符號，將特殊名詞前後括起，系統就不會將這一個特殊的名詞做斷詞。

接下來系統會先找出問句的關鍵詞，至 Google 搜尋取得答案段落，步驟如下：

1. 從問句與答案對中找出關鍵詞(k_1, k_2, \dots, k_n, A)，並且將停詞移除。
2. 將(k_1, k_2, \dots, k_n, A)組合至搜尋引擎中處理。
3. 擷取搜尋引擎前面 30 個結果，保留文字部分，把標點符號與英文字母移除。
4. 將處理過後的結果取出一段含有答案與任一關鍵詞的答案段落。

範例如表 1：

表 1：問句與答案，輸入與輸出範例

(Q, A)	AP
豐田汽車的創始人是誰? Answer:(<u>喜一郎</u>)	這就是TOYOTA創始人 <u>豐田喜一郎</u> 在全力以赴開發日本第一輛國產汽車時的思想 ...
(k_1, k_2, \dots, k_n, A)	利三郎同意 <u>豐田喜一郎</u> 於1937年8月27日另立門戶成立“豐田汽車工業株式會社” ...
豐田汽車，創始人，喜一郎	...豐田創始人: <u>豐田喜一郎</u> 豐田汽車公司是世界 ...

3-3 問句分析

在這一個章節中，我們主要是針對問句的問句類型與問句樣式做分析。

3-3-1 疑問詞分類

當建立一個新的問句時系統會分析問句，自動找出最合適的疑問詞並且依據這一個疑問詞的類別來決定問句真正的問句類型，疑問詞分類如表 2。

表 2：疑問詞分類

疑問詞分類	疑 問 詞
人	誰、那位、什麼人、何人
事物	哪些、那些、那樣、何許、何種、哪種、那種、何謂、什麼是、甚麼是
時間	多久、何時、幾時、哪天、幾時、哪一天
地	哪兒、哪裡、何在、那裡、何處、何地
數量值	若干、多少、幾道、幾個、幾人、幾年、幾邊形、幾歲、幾種、幾座、幾條、第幾、幾位
模糊疑問詞	幾、什麼、甚麼、何、多、哪、那

如果疑問詞是模糊疑問詞，系統會依據疑問詞後面所連接的詞來決定問句的類型。範例如下：

例一：

如果模糊疑問詞後面連接的是一個數量如：一、二...等，系統會把再根據下一個名詞來決定問句類型。

疑問詞 = 模糊疑問詞 + 數量 + 名詞

母親節是哪一天？->母親節(N) 是(Vt) 哪(DET) 一(DET) 天(N)

疑問詞 = 哪一天

問句類型 = 時間

系統會自動的將“哪一天”組合起來，作為這一個問句的問句類型。

例二：

如果模糊疑問詞後面連接的是一個名詞，系統就會依據名詞來決定問句類型。

疑問詞 = 模糊疑問詞 + 名詞

什麼人將中國造紙術傳入歐洲？->什麼(DET) 人(N) 將(P) 中國(N) 造紙術(N) 傳入(Vt) 歐洲(N)

疑問詞 = 什麼人

問句類型 = 人

系統會自動的將“什麼人”組合起來，作為這一個問句的問句類型。

3-3-2 萃取問句樣式

我們建立一些問句的法則分析問句萃取問句樣式。問句樣式格式設定如下：

問句樣式 = 疑問詞 + 主要詞

查詢詞定義：

1. 問句樣式 = 疑問詞 + 主要詞，查詢詞 = 主要詞。
2. 問句樣式 = 模糊疑問詞 + 跟隨的名詞，查詢詞 = 跟隨的名詞。
3. 問句樣式 = 疑問詞，查詢詞 = 疑問詞。

問句樣式分為 4 類 23 條法則，問句法則是依據疑問詞鄰近的詞分析產生，問句範例見附註二，法則列表如表 3：

表 3：問句法則

法則	疑問詞位置	問句樣式	問句結構	
法則 1	句首	第一類 疑問詞 + 名詞	疑問詞 + 名詞 + 動詞，當有數個連續名詞出現時以最後的名詞為主要詞。	
法則 2	句尾		語助詞 + 名詞 + "是"或"在"或"是在" + 疑問詞，當有數個連續名詞出現時以最後的名詞為主要詞。	
法則 3	句首		疑問詞 + "是" + 名詞 + ... + 問句最後的名詞，疑問詞後沒有任何的動詞	
法則 4	句首		疑問詞 + 語助詞 + 名詞	
法則 5			名詞 + 副詞 + 停詞 + 疑問詞	
法則 6			名詞 + 疑問詞 + 名詞	
法則 7			名詞 + 停詞	名詞 + "的"
			+ 疑問詞 +	停詞 + 名詞 + "的"
				"的" + 名詞
法則 8	句尾		名詞 + 停詞 + 疑問詞	
法則 9			名詞 + 停詞	+ 疑問詞 + 名詞
			介詞	
			動詞	
法則 10			停詞 + 名詞 + 疑問詞	
法則 11		名詞	+ 疑問詞是數量定詞 +	
			量詞 + 名詞	
法則 12	句首	第二類 疑問詞 + 動詞	疑問詞 + 動詞	
法則 13	句首		名詞	停詞 + 名詞
			疑問詞 +	副詞
				介詞
法則 14			動詞 + 語助詞 + 名詞	+ 停詞 + 疑問詞
		動詞 + 語助詞		

			動詞 + 名詞
法則 15			動詞 + 停詞 + 疑問詞
法則 16	句尾		動詞 + 介詞 + ... + 疑問詞
法則 17	句尾		動詞 + 疑問詞，動詞前不可以為停詞
法則 18			動詞 + 名詞 + 疑問詞
法則 19			名詞 + 停詞 + 動詞 + 疑問詞
			名詞
法則 20			名詞 + 介詞 + 疑問詞 + 動詞
			停詞 + 動詞
法則 21			名詞 + "是"或"在"或"是在" + 疑問詞 + 動詞
法則 22			名詞 + 疑問詞 + 動詞 + 名詞
			停詞
法則 23		第三類 疑問詞 + 量 詞	疑問詞是數量定詞 + 量詞，相連前後無名詞
法則 24		第四類 疑問詞	特殊問句如： 誰「挾天子以令諸侯」？

3-4 分析答案段落與非名詞關鍵詞和查詢詞轉換擴充

訓練系統找出最好的非名詞關鍵詞與查詢詞轉換擴充詞。主要的處理步驟如下：

1. 應用配對演算法去處理問句與答案段落，取得根據問句樣式與關鍵詞所得配對組合 (Alignment) 的統計。
2. 尋找在答案附近出現頻率最高的詞，並計算出統計與排序。
3. 將 1 與 2 的結果加總平均並做排序。找出最佳的轉換擴充詞。

3-4-1 使用文字組合的技術去學習非名詞關鍵詞與查詢詞的轉換

在這一個章節中，我們使用將使用 3-2 所取得的答案段落根據(問句，答案段落)根據下面的步驟過濾成有用的資料，並且使用 Competitive Linking Algorithm (Melamed, 1997) 做(問句，答案段落)選對。

1. 將答案段落中的答案以<ANS>標記答案，並且取出其左右鄰近的六個中文字，移除非中文字。
2. 以在 3-3 取得的問句樣式並且以 k_1 代替，並且分析問句取得關鍵詞，並且移除掉名詞，剩下來的關鍵詞以 $k_2, k_3, k_4, \dots, k_n$ 來表示。
3. 將中文字串重新組合成 2 個中文字(2grams)與 3 個中文字(3grams)的組合詞，用

$a_1, a_2, a_3, \dots, a_m$ 來表示。

4. 使用 Competitive Linking Algorithm(CLA)處理所得到的關鍵詞與組合詞。

1.將所有(問句, 答案)的組合中,所有的非名詞關鍵詞、查詢詞與組合詞,計算出LLR(k_i, a_j),使用統計學中Logarithmic Likelihood Ratio(LLR)計算出 k_i 與 a_j 反應出來的相互關係。

Log-likelihood ratio : LLR(x,y)

$$LLR(x ; y) = -2\log_2 \frac{p_1^{k_1} (1 - p_1)^{n_1 - k_1} (1 - p_2)^{n_2 - k_2}}{p^{k_1} (1 - p)^{n_1 - k_1} (1 - p)^{n_2 - k_2}}$$

k_1 = number of pairs that contain x and y simultaneously ◦

k_2 = number of pairs that contain x but do not contain y ◦

n_1 = number of pairs that contain y

n_2 = number of pairs that does not contain y

$p_1 = k_1/n_1$ $p_2 = k_2/n_2$ $p = (k_1+k_2)/(n_1+n_2)$

2.當(k, a)的 LLR 值小於 7.88 時系統就會丟棄

3.其餘的(k, a)繼續做下面的流程 4 至流程 7。

4.完成LLR的計算後將(問句 ; 答案)的(k_i, a_j)對做排序列表。

5.如果與先前選擇不衝突時可以選擇未計算的(k, a)及將計算好的(k, a)列表是由上到下遞減的。

6.停止執行當(k, a)對計算完畢。

7.產出所有的問句與答案段落的 aligned pairs 列表。

8.紀錄每一個 aligning 組合(k, a)的總數。

9.選取每一個 k 擇前 n 名的組合詞, $t_1, t_2, t_3 \dots, t_n$, k 為問句樣式或非名詞關鍵詞。

LLR 統計方式是一個廣泛有效被利用來確認兩個詞的相關性,如果兩個詞在統一個問句發生的頻率高,它們的 Alignment counterparts 的頻率也會高。範例結果如表 4。

表 4：候選擴充詞與關鍵詞的配對等級範例

	疑問詞與關鍵詞	擴充	統計
K	生長	生長	27
K	生長	長在	13
K	生長	生長在	12
K	生長	的生	8
K	生長	生長於	8

Q	發明	發明	103
Q	發明	明的	23
Q	發明	電話	20
Q	發明	的發	20

Q	出生	斯坦	22
Q	出生	出生於	10
Q	出生	生在	10
Q	出生	出生在	10
Q	出生	生於	10

K=關鍵詞, Q=問句樣式

3-4-2 鄰近等級

鄰近等級是計算每一對(問句, 答案段落), 答案段落中在答案附近出現的組合詞數目(Proximity Count), 並且把這些做排序如表 5, 由這一個方式可以得知那些詞在答案的附近是最容易出現的, 通常這些詞的出現頻率跟答案的相關性也成正比。

表 5：問句答案類型是”什麼地方 生長”的鄰近詞加總

組合詞	出現次數	鄰近等級
生長	16	1
長在	10	2
生長在	9	3
長於	7	4
生長於	7	4
間帶	5	5
海岸	5	5
沼澤	5	5
潮間	5	5
潮間帶	5	5

3-4-3 結合鄰近等級與配對等級

把所求出的結果鄰近等級與配對等級結合並求出最後等級，並把結果依大小排列如表 6。

$$\text{擴充詞等級} = (X * \text{配對等級} + Y * \text{鄰近等級}) / Z$$

系統中 X 與 Y 我們是設定 X=4 Y=6 Z=X+Y

表 6：將表 4 與 5 結合取得加權平均值的最後等級

組合詞	配對等級	鄰近等級	平均	最後等級
生長	1	1	1	1
長在	2	2	2	2
生長在	3	3	3	3
長於	4	6	4.8	5
生長於	4	4	4	4
間帶	5	7	5.8	7
海岸	5	7	5.8	7
沼澤	5	7	5.8	7

系統會自動將所擷取的詞做整合，整合的方式有下列幾項：

1. 把最後等級中含有關鍵詞與問句樣式的組合詞移除。
2. 相同最後等級的擴充詞如果有相似保留主要的詞，其餘移除。

間帶	7	6	6	7
潮間	7	6	6	7
沼澤	7	6	6	7
潮間帶	7	6	6	7

相同最後等級的擴充詞(間帶、潮間、潮間帶)只會取[潮間帶]

3. 如果配對等級是空值，當鄰近等級是前兩名一樣並入計算，計算方式如下：

$$Top_n = \text{名次} \rightarrow Top_1 = 1$$

$$\text{擴充詞等級} = (Y * \text{鄰近等級} + 30 + (Top_n * 10)) / Z$$

如此當配對等級是空值時，系統會自動加上 $30 + (Top_n * 10)$ 去作加權平均

4. 兩個停詞的組合可能會轉變成有效詞，所以系統在停詞或組合詞是使用加權平均的方式來計算，只會使等級下降，增加積分是由系統管理者根據實驗結果判斷並給積分。

例如：“的”是一個停詞，他出現在兩個字的組合詞與三個字的組合詞可以有不一樣的設定。

如：你的、我的...等”的”出現在兩個字組合詞的後面，百分之九十是無效詞，我們就可以設定積分等於 45 分。但是如果出現在三個字組合詞時又有不一樣結果，只有當”的”出現在三個字組合詞的中間時百分之三十是無效詞，所以就可以設定積分為 15 分。計算方式如下：



$$\text{無效詞積分} \\ (\text{無效百分比} / 10) * 5 = SW$$

當出現無效詞時產生積分後的平均等級計算方式：

$$\text{擴充詞等級} = (X * \text{配對等級} + Y * \text{鄰近等級}) + (SW) / Z$$

$$\text{無效詞積分是可以累加的 } SW_{all} = SW_1 + SW_2 + \dots + SW_n$$

$$\text{擴充詞等級} = (X * \text{配對等級} + Y * \text{鄰近等級}) + (SW_{all}) / Z$$

我們使用這一個方法可以把一些無效或無用詞的 $Rank_{avg}$ 積分增加使最後等級下降，讓系統能夠真正擷取到有用的詞，增加系統的精確度。

3-5 產生查詢句

我們先將問句分析擷取出問句樣式與關鍵詞，並且將問句樣式與非名詞的關鍵詞做擴充，擴充出 2 個擴充詞，使用布林邏輯與 Google 的可以判讀的格式轉換成查詢句。範例如表 7：

表 7：非名詞關鍵詞與查詢詞轉換擴充

問句	人類在什麼出現後進入歷史時代？		
類型	詞	擴充詞(一)	擴充詞(二)
問句樣式	什麼 出現	發明	紀錄
名詞關鍵詞	人類		
名詞關鍵詞	歷史時代		
非名詞關鍵詞	進入	記載	紀錄
布林邏輯	(出現 OR 發明 OR 紀錄)AND"人類"AND"歷史時代"AND(進入 OR 發明 OR 紀錄)		
Google	(出現 發明 紀錄)"人類""歷史時代"(進入 發明 紀錄)		

有些問句並沒有適當可擴充的關鍵詞，所以在非名詞關鍵詞與查詢詞全擴充的查詢句中，只有查詢詞的擴充。在做完上面的步驟後最後系統就會將這些結果送至搜尋引擎。

為了使搜尋結果減少誤差，系統還有一個相關答案的功能，因為有些答案是翻譯字，而因為翻譯的不同會有不同的答案，如：

“凱薩大帝”與“凱撒大帝”，“阿爾卑斯山”與“丫爾卑斯山”

但是這些答案也並非是錯誤或不正確，為了使系統不會遺漏掉這些答案，系統管理者可以將相關的答案新增至資料庫中，搜尋答案時系統就會自動進行比對。



第四章 實驗與分析

4-1 研究步驟

在系統剛建立完成時，必須建立訓練資料，所以必須去搜集相關的問句與答案(Q, A)，這些問句可以涵蓋地理、歷史、地球科學、音樂、國文、數學...等領域。

我們在網路上如：奇摩知識家¹、蕃薯藤的小蕃薯問號小博士²等網站與一些書本如：高中歷史與地理參考書、一般考試用書等書籍上找尋到 383 個符合的中文問句與答案，分為五種類別，如表 8：

表 8：訓練資料與測試資料問句類型統計

問句類型	訓練資料	測試資料
人	97	18
事物	96	17
時間	61	15
地	64	16
數量	65	14
總數	383	80

將這些問句輸入到系統中，根據問句分析取得問句的問句樣式，並且從 Google 取回 30 個搜尋結果，總共取回 10,805 個答案段落。如表 9：

表 9：訓練資料問句數量、問句樣式及從搜尋引擎擷取的答案段落統計

訓練資料	問句與答案	問句樣式	答案段落
統計數量	383	289	10,805

我們也找了 80 個問句與答案來做系統測試，類型統計如表 8：

4-2 擷取問句樣式的成效

我們將討論系統擷取問句樣式，並且分析錯誤。一位主要研究語言學的研究所同學，一位是我數位圖書資訊組的同學，一位是沒有語言學基礎的公司同事，完全是用他對語言的感覺去評斷。這裡我們所要評論是問句的問句樣式，我把測試資料所擷取出的問句樣式列表給她們評分，如果覺得是正確的就給”Good”，如果不正確就給”Bad”，實驗結果如表 10。實驗結果中如果三個人都給”Good”就給”Very Good”，如果三個人都

¹ 奇摩知識家<http://tw.knowledge.yahoo.com/>。

² 小蕃薯的問號小博士<http://kids.yam.com/why/>。

給”Bad”就給”Very Bad”，只有一個人給”Bad”就給”Good”，只有一個人給”Good”就給”Bad”，實驗結果如表 11。

表 10：問句樣式評論結果 (一)

測試資料	測試者(一)	測試者(二)	測試者(三)
“Good”總計	73	76	77
“Good”百分比	91.25%	95%	96.25%

表 11：問句樣式評論結果 (二)

測試資料	Very Bad	Bad	Good	Very Good
統計數量	3	3	1	73
百分比	3.75 %	3.75 %	1.25%	91.25%

擷取問句樣式錯誤分析，表 12 是系統對測試資料中前五題所擷取出的問句樣式。

表 12：前五個問句和問句樣式與評論

問句	問句樣式	評論
紅樹林生長在什麼地方？	什麼地方 生長	Very Good
美國獨立戰爭是發生在何時？	何時 發生	Very Good
法拉利的標誌上是哪一種動物？	哪一種 動物	Very Good
戰國末期那一國滅六國統一中國？	那一 國滅六國	Very Bad
[居禮夫人]發現了什麼元素？	什麼 元素	Very Good

問句樣式經過評分後，我們分析系統錯誤的原因，並且把原因歸類如下：

1. CKIP 中文斷詞發生錯誤。

例如：

問句：戰國末期那一國滅六國統一中國？

中文斷詞：戰國(N) 末(N) 期(N) 那(DET) 一(DET) 國滅六國(N) 統一(Vt) 中國(N)

問句樣式：那一 國滅六國

“國滅六國”斷詞錯誤，應該為”國(N) 滅(Vt) 六(DET) 國(N)”，問句樣式=那一國。

2. 特例問句，中文斷詞的標記所給的詞類出現兩個相同動詞，系統取第一個出現的動詞，但是主要的動詞應該是第二個動詞。

例如：

問句：蘇聯在何時正式瓦解？

中文斷詞：蘇聯(N) 在(Vt) 何時(N) 正式(Vi) 瓦解(Vi)

問句樣式：何時 正式

應該取”瓦解”為主要詞，但是系統卻取”正式”，問句樣式=何時瓦解。

3.系統在分析問句錯誤，沒有把”出”設定為停詞，或沒有將詞做正確組合產生新詞。

例如：

問句：英國黛安娜王妃是在哪裡出車禍死亡的？

中文斷詞：英國(N) 黛安娜(N) 王妃(N) 是(Vt) 在(Vt) 哪裡(N) 出(Vt) 車禍(N) 死亡(Vi) 的(T)

問句樣式：哪裡 出

應該取”出車禍”為主要詞，但是系統卻只取”出”，未建立新詞，問句樣式=哪裡出車禍 或 哪裡 車禍。

4-3 實驗結果與評論非名詞關鍵詞與查詢詞擴充

在這一個章節中我們將評論非名詞關鍵詞與查詢詞擴充的效果。將測試資料中的 80 個問句實際至 Google 搜尋檢索前 10 筆回傳結果作分析。

我們以 Mean Reciprocal Rank(MRR)做評量依據。MMR 計算結果表 13，顯示使用查詢詞擴充搜尋可以得到很明顯的進步。

表 13：各類查詢句的 MRR 值與改進程度

查詢	關鍵詞搜尋	查詢詞擴充	非名詞關鍵詞與查詢詞擴充
MRR	0.716	0.757	0.765
改進		4.1%	4.9%

依類別來分析如表 14，問句類型 = “地”與”事物”時有最好的效果，問句類型 = “時間”時效果最差。當非名詞關鍵詞與查詢詞一起擴充時間句類型 = “地”與”事物”還是有很好的效果，但是在問句類型 = “人”時效果是最佳 100%，分析其原因是因為”人”類別中需要做非名詞關鍵詞擴充的問句不多，並且正好她們都是在第一個搜尋結果就找到答案，所以其 MMR 值才會為”1”

表 14：各類查詢句的 MMR 值依問句的類別分類

	人	事物	地	時間	數量
關鍵詞搜尋	0.668	0.803	0.802	0.529	0.773
查詢詞擴充	0.752	0.823	0.770	0.647	0.785
非名詞關鍵詞與查詢詞擴充	1	0.875	0.9	0.504	0.444

我們也檢驗查詢詞有效回收的 R 值(Recall at R)與 HE(Human Effort)。其中 HE 定義為搜尋結果中答案的位置，計算結果如表 13。表 13 顯示查詢詞擴充當答案在等級 1 至

等級 3 時有最好的效果，回收率為 90%，所以我們 R 取 3。查詢詞與非名詞關鍵詞擴充當答案在等級 1 至等級 3 時有最好的效果，有回收率有 88.46% 所以我們 R 取 3。

表 15 顯示，在查問詞擴充時 $HE\ per = 1.9$ ，查詢詞與非名詞關鍵詞擴充時 $HE\ per = 2.03$ ，都小於關鍵詞搜尋的 $HE\ per = 2.187$ ，證明使用查問詞擴充或查詢詞與非名詞關鍵詞擴充能改進搜尋的效能。

在查詢詞與非名詞關鍵詞擴充中，因為有些問句中並沒有非名詞關鍵詞，所以測試的資料只有 26 筆資料。

表 15：查詢詞擴充在 Recall at R 與 Human Effort 的結果

等級	Rank count			Recall at R (%)		
	關鍵詞搜尋	查詢詞擴充	非名詞關鍵詞 與查詢詞擴充	關鍵詞搜尋	查詢詞擴充	非名詞關鍵詞 與查詢詞擴充
1	46	50	17	57.5	62.5	65.38
2	12	11	3	72.5	76.25	76.92
3	11	11	3	86.25	90	88.46
4	3	3	1	90	93.75	92.30
5	3	1	0	93.75	95	92.30
6	1	2	0	95	97.5	92.30
7	0	0	1	95	97.5	96.15
8	0	1	0	95	98.75	96.15
9	1	0	0	96.25	98.75	96.15
10	0	0	0	96.25	98.75	96.15
No answers	3	1	1			
Human Effort	175	152	53			
# of questions	80	80	26			
HE per question	2.187	1.9	2.03			

在表 16 我們依據問句類別來劃分，找出各類別的變化，查詢詞與非名詞關鍵詞擴充在類別”地”與”人”時對搜尋的幫助就比較大。查詢詞擴充在類別”地”與”數量”時對搜尋的幫助就比較大。在大部分的實驗結果中”時間”類別的效果都是最差的，在訓練資料中類別”時間”的問句也是最少的，所以可得知當訓練資料越少時系統的準確度就會越低。

表 16：評論非名詞關鍵詞與查詢詞擴充 **HE per question** 的結果，依問句的類別分類

	人	事物	時間	地	數量
關鍵詞	2.27	2.41	2.8	1.43	2
查詢詞擴充	1.94	2	2.26	1.62	1.64
非名詞關鍵詞與查詢詞擴充	1	2.12	3.5	1.2	2.33



第五章未來工作

未來的工作針對非名詞關鍵詞與查詢詞的擴充上，希望能夠建立一個相關詞的資料庫，結合問答系統資料庫計算出與查詢詞或非名詞關鍵詞相關度最高的擴充詞。增加中文問答系統的效能上，希望能在中文問答系統前設置一個基於常問問題集的問答系統，過濾已處理過的問題，提升系統效能。在回覆給使用者的答案上，希望能使用關鍵詞與問句樣式分析搜尋引擎所回傳資料，抽取出問句中最可能的答案，提供給使用者。



參考文獻

Huang , Gai-Tai and Yao, Hsiu-Hsen. “Chinese Question-Answering System”, 計算機科學技術學報, 2004 年第 4 期, pp. 479-488, 2004.

Lee, Cheng-Wei et al., ASQA: A Hybrid Architecture for Answering Chinese Factoid Questions, Hsinchu.Taiwan, 2005.

Melamed, I. Dan. “A Word-to-Word Model of Translational Equivalence”, In *Proceedings of the 35st Annual Meeting of the Association for Computational Linguistics*, pp. 490-497, 1997.

Moldovan, D., Pasca, M., Harabagiu, S., and Surdeanu M. “Performance Issues and error Analysis in an Open-Domain Question Answering System”, In *Proceedings of the 40st Annual Meeting of the Association for Computational Linguistic*, pp. 33-40, Philadelphia, PA, USA, July 2002.

Yu, Zheng-Tao et al., “Query Expansion For Answer Document Retrieval in Chinese Question Answering System”, *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics,Guangzhou*, pp. 18-21, August 2005.

鄭實福、劉廷、秦冰、李生, ”自動問答總述”, 中文信息學報, 16 卷 6 期, 46-52 頁, 2002。

王亭雅, 網頁搜尋引擎問答, 中正大學資工所, 碩士, 民國九十二年七月。

王怡嘉, 問答系統之查詢擴充—以網路資源為本的非監督式學習策略, 交通大學資訊科學所, 碩士, 民國九十三年十二月。

曾元顯, 資訊檢索與知識探勘, 台北, 民國九十三年七月十五日。

沈天佐、林川傑、陳信希, ”以網際網路內容為基礎之問答系統“Why”問句之研究”, 第十五屆計算語言學研討會論文集, 211-229 頁, 2003。

李季, ”淺談中文問答系統”, 計算機科學技術學報, 2004 年第 1 期, 71-72 頁, 2004。

林川傑, 中文開放領域自動問答系統之研究, 國立臺灣大學資訊工程學研究所, 博士, 2004。

秦冰、劉廷、王洋、鄭實福、李生, 基於常問問題的中文問答系統研究, 哈爾濱·大陸,

2003。

關毅、王曉龍、趙岩、趙健，面向專業網站的中文問答系統研究，2003。



附錄一 測試的問句與答案

No.	問句	Rule	問句類型	問句樣式
1	紅樹林生長在什麼地方？	法則 15	地	什麼地方 生長
2	美國獨立戰爭是發生在何時？	法則 15	時間	何時 發生
3	法拉利的標誌上是哪一種動物	法則 7	事物	哪一種 動物
4	茶神是指誰	法則 19	人	誰 指
5	鄭和總共下西洋幾次	法則 11	數量	幾次 下西洋
6	[西遊記]的作者是誰	法則 2	人	誰 作者
7	那一位美國總統解放黑奴	法則 1	人	那一位 總統
8	[都江堰]是誰修建的	法則 22	人	誰 修建
9	中國國民黨的黨徽是何人設計的	法則 7	人	何人 設計
10	誰建立了羅馬帝國	法則 12	人	誰 建立
11	岳飛被誰陷害	法則 20	人	誰 陷害
12	誰第一個踏上月球	法則 1	人	誰 第一個
13	汽車是誰發明的	法則 22	人	誰 發明
14	[古巴危機]是發生在哪一位總統時	法則 9	人	哪一位 總統
15	[雅虎]的創辦人是哪位	法則 8	人	哪位 創辦人
16	南非的立法首都是哪裡	法則 2	地	哪裡 首都
17	[居禮夫人]出生於哪裡	法則 16	地	哪裡 出生
18	[都江堰]位於哪一省	法則 17	地	哪一省 位於
19	美國的大峽谷位於那一州	法則 17	地	那一州 位於
20	[出埃及記]中的應許之地是指何處	法則 19	地	何處 指
21	哪裡被喻為[北方威尼斯]	法則 13	地	哪裡 喻為
22	[世界法庭]位在何處	法則 19	地	何處 位在
23	[阿茲特克人]原居住在哪裡	法則 15	地	哪裡 居住
24	阿根廷首都是哪裡	法則 8	地	哪裡 首都
25	[九一八事變]發生在民國幾年	法則 10	時間	幾年 民國
26	韓戰發生在什麼時候	法則 15	時間	什麼時候 發生

27	[美國獨立宣言]是在何時通過	法則 21	時間	何時 通過
28	[西羅馬帝國]何時滅亡	法則 22	時間	何時 滅亡
29	蘇聯在何時正式瓦解	法則 21	時間	何時 正式
30	[六四天安門]發生在哪一年	法則 15	時間	哪一年 發生
31	[北大西洋公約組織]何時成立	法則 22	時間	何時 成立
32	兒童節是每年的哪一天	法則 8	時間	哪一天 每年
33	[居禮夫人]發現了什麼元素	法則 15	事物	什麼元素 發現
34	[高速公路電子收費系統]感應器是使用什麼系統	法則 17	事物	什麼系統 使用
35	硬度最高的礦物是什麼	法則 2	事物	什麼 礦物
36	戰國末期那一國滅六國統一中國	法則 6	事物	那一 國滅六國
37	世界上最長海底隧道是哪一個	法則 8	事物	哪一個 隧道
38	世界最長的吊橋是哪一座橋	法則 7	事物	哪一座 橋
39	人類在什麼出現後進入[歷史時代]	法則 21	事物	什麼 出現
40	[歐洲屋脊]是指哪一座山	法則 9	事物	哪一座 山
41	[以色列人]信奉什麼宗教	法則 17	事物	什麼宗教 信奉
42	歐亞的界山是哪一座山	法則 7	事物	哪一座 山
43	阿里山海拔高度幾公尺	法則 11	數量	幾公尺 高度
44	宋美齡逝世時享年幾歲	法則 17	數量	幾歲 享年
45	一副象棋有多少棋子	法則 7	數量	多少 棋子
46	[中山高速公路]全長幾公里	法則 11	數量	幾公里 全長
47	[雪山隧道]全長多少公里	法則 11	數量	多少 公里 全長
48	[十字軍]共東征幾次	法則 17	數量	幾次 東征
49	[豐田汽車]的創始人是誰	法則 2	人	誰 創始人
50	誰有[台灣第一名模]之稱	法則 13	人	誰 之稱
51	電影[黃金眼]的男主角是誰	法則 2	人	誰 男主角
52	霹靂布袋戲中[刀狂劍癡]是指誰	法則 19	人	誰 指
53	江戶幕府是由誰建立的	法則 20	人	誰 建立

54	誰是世界上第一個太空旅遊者	法則 3	人	誰 旅遊者
55	誰代表清政府與日本簽訂馬關條約	法則 12	人	誰 代表
56	中央山脈北起哪裡	法則 19	地	哪裡 北起
57	英國黛安娜王妃是在哪裡出車禍死亡的	法則 21	地	哪裡 出
58	萊茵河起源於哪裡	法則 19	地	哪裡 起源於
59	熊貓產於中國的何處	法則 16	地	何處 產
60	元宵節蜂炮是在哪裡施放	法則 21	地	哪裡 施放
61	西楚霸王項羽在哪裡兵敗自刎	法則 9	地	哪裡 兵敗
62	[拜占庭帝國]亡於哪一個帝國	法則 6	事物	哪一個 帝國
63	那一個山脈有[台灣屋脊]之稱	法則 1	事物	那一個 山脈
64	那一個島是世界上面積最大的	法則 1	事物	那一個 島
65	[台灣意象]票選活動第一名是什麼	法則 8	事物	什麼 第一名
66	三國鼎立局面形成的關鍵是什麼	法則 2	事物	什麼 關鍵
67	青蛙的幼蟲是什麼	法則 2	事物	什麼 幼蟲
68	[拜占庭帝國]在何時滅亡	法則 21	時間	何時 滅亡
69	美國的越戰自何時開始	法則 20	時間	何時 開始
70	美國總統甘迺迪在何時遇刺身亡	法則 21	時間	何時 遇刺
71	希特勒在什麼時候自殺身亡	法則 21	時間	什麼時候 自殺
72	[友訊科技]成立於何時	法則 16	時間	何時 成立
73	夏威夷在何時成為美國第五十洲	法則 21	時間	何時 成為
74	[台北市立動物園]的門票全票多少錢	法則 6	數量	多少 錢
75	南北韓的軍事分界線是北緯幾度	法則 10	數量	幾度 北緯
76	2004 年台灣男性的平均壽命是幾歲	法則 8	數量	幾歲 平均壽命
77	萊茵河全長多少公里	法則 11	數量	多少 公里 全長
78	奧運會多久舉行一次	法則 22	數量	多久 舉行
79	古代中國人將一天分為幾個時辰	法則 9	數量	幾個 時辰
80	[排雲山莊]有多少床位	法則 7	數量	多少 床位

附錄二 問句法則例句

No.	問句	中文斷詞	法則	問句類型	問句樣式
1	哪一個宗教的教徒不吃豬肉？	哪(DET) 一(DET) 個(M) 宗教(N) 的(T) 教徒(N) 不(ADV) 吃(Vt) 豬肉(N)	法則 1	事物	哪一個宗教
2	請問埃及的首都是哪裡？	埃及(N) 的(T) 首都(N) 是(Vt) 哪裡(N)	法則 2	地	哪裡首都
3	誰是中國歷史上唯一的女皇帝	誰(N) 是(Vt) 中國(N) 歷史(N) 上(N) 唯一(A) 的(T) 女(N) 皇帝(N)	法則 3	人	誰 女皇帝
4	什麼類型的動物最常被用於癌症研究	什麼(DET) 類型(N) 的(T) 動物(N) 最(ADV) 常(ADV) 被(P) 用於(Vt) 癌症(N) 研究(N)	法則 4	事物	什麼類型動物
5	員工配股不得轉讓之期限最長是幾年	員工(N) 配股(Vi) 不得(ADV) 轉讓(Vt) 之(T) 期限(N) 最(ADV) 長(Vi) 是(Vt) 幾年(N)	法則 5	數量	幾年期限
6	宋代那一位皇帝寫的書法有[瘦金體]	宋代(N) 那(DET) 一(DET) 位(M) 皇帝(N) 寫(Vt) 的(T) 書法(N) 有(Vt) 瘦金體(Nb)	法則 6	人	那一位皇帝
7	矮靈祭是哪個民族的傳統祭典？	矮靈(N) 祭(Vt) 是(Vt) 哪(DET) 個(M) 民族(N) 的(T) 傳統(N) 祭典(N)	法則 7	事物	哪個民族
8	[台灣意象]票選活動第一名是什麼	台灣意象(Nb) 票選(Vt) 活動(Vt) 第一(DET) 名(M) 是(Vt) 什麼(DET)	法則 8	事物	什麼第一名
9	李安以哪一部電影獲得奧斯卡最佳導演	李安(N) 以(P) 哪(DET) 一部(N) 電影(N) 獲得(Vt) 奧斯卡(N) 最佳(A) 導演(N)	法則 9	事物	哪一部電影
10	美國南北戰爭發生在西元幾年	美國(N) 南北戰爭(N) 發生(Vt) 在(Vt) 西元(N) 幾年(N)	法則 10	時間	幾年西元
11	萊茵河全長多少公里	萊茵河(N) 全(DET) 長(Vi) 多少(DET) 公里	法則 11	數量	多少公里全長
12	誰發明了電燈泡？	誰(N) 發明(Vt) 了(ASP) 電燈泡(N)	法則 12	人	誰發明

13	哪裡被喻為[北方威尼斯]	哪裡(ADV) 被(P) 喻為(Vt) 北方威尼斯(Nb)	法則 13	地	哪裡 喻為
14	全球紅樹林分佈的北界是哪裡?	全球(N) 紅樹林(N) 分佈(Vi) 的(T) 北界(N) 是(Vt) 哪裡(N)	法則 14	地	哪裡 分佈
15	美國獨立戰爭是發生在何時?	美國獨立戰爭(N) 是(Vt) 發生(Vt) 在(Vt) 何時(N)	法則 15	時間	何時 發生
16	台灣的沙岸主要分布於哪裡	台灣(N) 的(T) 沙岸(N) 主要(A) 分布(Vi) 於(P) 哪裡(N)	法則 16	地	哪裡 分布
17	[高速公路電子收費系統]感應器是使用什麼系統	高速公路電子收費系統(Nb) 感應器(N) 是(Vt) 使用(Vt) 什麼(DET) 系統(N)	法則 17	事物	什麼系 統使 用
18	[金屋藏嬌]成語的典故出自中國哪一朝	金屋藏嬌(N) 成語(N) 的(T) 典故(N) 出自(Vt) 中國(N) 哪(DET) 一朝(ADV)	法則 18	時間	哪一朝 出自
19	中國歷史中的「成吉思汗」是指誰?	中國(N) 歷史(N) 中(POST) 的(T) 成吉思汗(N) 是(Vt) 指(Vt) 誰(N)	法則 19	人	誰指
20	唐朝至何時才完成統一大業?	唐朝(N) 至(P) 何時(N) 才(ADV) 完成(Vt) 統一(Vt) 大業(N)	法則 20	時間	何時 完成
21	鄧麗君在何時逝世?	鄧(N) 麗(Vi) 君(N) 在(Vt) 何時(N) 逝世(Vi)	法則 21	時間	何時 逝世
22	自由女神是哪一國送美國?	自由(Vi) 女神(N) 是(Vt) 哪(DET) 一(DET) 國(N) 送(Vt) 美國(N)	法則 22	人	哪一國 送
23	[1英吋]是多少公分?	1英吋(Nb) 是(Vt) 多少(DET) 公分(M)	法則 23	數量	多少公 分
24	誰「挾天子以令諸侯」?	誰(N) 挾天子以令諸侯(N)	法則 24	人	誰