國立交通大學

資訊科學系

碩士論文

利用 Ontological Chain 解決跨語言資訊檢索系統中的翻譯歧義性問題

Resolving Translation Ambiguity By Ontological Chain for Cross Language Information Retrieval

研 究 生：梁哲瑋

指導教授：柯皓仁　博士

楊維邦　博士

中 華 民 國 九 十 三 年 六 月

利用 Ontological Chain 解決跨語言資訊檢索系統中的翻譯歧義性問題

Resolving Translation Ambiguity By Ontological Chain for Cross Language Information Retrieval

研 究 生：梁哲瑋　　Student：Je-Wei Liang

指導教授：楊維邦　　Advisor：Wei-Pang Yang

柯皓仁　　Hao-Ren Ke

國 立 交 通 大 學

資 訊 科 學 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Computer and Information Science

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer and Information Science

June 2004

Hsinchu, Taiwan, Republic of China

中華民國九十三年六月

Resolving Translation Ambiguity Using Ontological Chain for Cross Language Information Retrieval

Student: Je-Wei Liang Advisor: Dr. Hao-Ren Ke, Dr. Wei-Pang Yang

Institute of Computer and Information Science

National Chiao Tung University

ABSTRACT

Bilingual dictionaries have been commonly used for query translation in cross-language information retrieval(CLIR). However, the problem of translation ambiguity happens in query translation. Recent studies suggest traversing WordNet for selecting appropriate translations. This paper proposes an ontological chain approach to resolve translation ambiguity. First, we find the most smilar ontology nodes for each query. Second, we construct a semantic graph according to the semantic distances between these nodes. And finally we select the connected component with the highest score as our ontological chain. We show that our approach reaches 81% effect of monolingual information retrieval systems. When there are many candidate translations, our system performs better than monolingual information retrieval system.

Keywords: Cross Language Information Retrieval; Query Translation; Word Sense Disambiguation; Ontology; Ontological Chain

利用 Ontological Chain 解決跨語言資訊檢索中的翻譯歧義性問題

研究生: 梁哲瑋　　指導教授: 柯皓仁博士，楊維邦博士

國立交通大學資訊科學研究所

摘要

翻譯檢索問句為本的跨語言資訊檢索系統會遭遇到翻譯歧義性的問題，目前解析歧義性的方法主要有同義詞典為本和語料庫為本的方法，前者的涵蓋範圍不夠，詞鍵關係過少；後者構需要耗費龐大成本來建構語料庫。本論文提出一套知識本體鏈(Ontological Chain)的方法，解決跨語言資訊檢索系統中翻譯歧義性(Transilation Ambiguity)問題。運用知識本體表示專家建構的領域知識(Domain Knowledge)，從知識本體相關的節點延伸出知識本體鏈，替每個中文詞鍵找到最適當的英文翻譯。本論文以英國聖安德魯大學照片資料集(The Eurovision ST Andrews Photographic Collection,簡稱 ESTA)兩萬八千篇影像和照片說明為例，實作一個跨語言資訊檢索系統。本系統的平均準確率可達 49%，並且達到單語言資訊檢索系統的 81%效能。

關鍵字： 跨語言資訊檢索、翻譯檢索問句、解析詞鍵歧義、知識本體、知識本體鏈

致謝

感謝兩年來指導教授柯皓仁老師以及楊維邦老師在各方面的悉心指導和照顧，以及提供了許多寶貴的建議，讓我體驗了完成一項研究的整個過程，在生活上也給了我許多啟示。

感謝實驗室的學長們深夜還時常在實驗室給予指導和熱心的討論，以及對論文耐心的修訂和各方面的協助，讓我能順利完成論文的每個細節。特別感謝葉鎮源學長以及鄭培成學長提供許多關鍵的建議和協助。

感謝實驗室的同學和學弟妹們，你們適時的幫助與實驗室和樂的氣氛是使我完成論文的一大動力，也讓這段獨一無二的過程充滿回憶。

最後要感謝親愛的家人無條件的支持，使我能專心致力於研究，讓我能心無旁騖，順利完成學業。僅將此篇論文獻給他們。

目錄

圖目錄

表目錄

公式目錄

第一章　簡介

第一節 跨語言資訊檢索系統

近年來，網際網路的普及，使得數位資訊的傳播跨越國度的限制；持續累積的多樣化資訊，儼然已成為一個巨大、分散且資訊豐富的多語言資料庫。各種語言寫成的文件都可能包含使用者需要的資訊。因此，除了母語之外，使用者也有檢索外語文件的需求。然而，傳統的搜尋引擎(Search Engine)與資訊檢索系統(Information Retrieval System)僅就單一語言的文件作考量；亦即，檢索問句(Query)與文件皆須使用相同的語言來表達，並沒有考慮到檢索問句與文件分屬不同語言的可能性。因此如何跨越語言的障礙，以達到跨語言資訊檢索的目的，顯然是個迫切需要解決的重要課題。

語言的差異，使得資訊的取得多了一道障礙。使用者往往不知道檢索問句在另一個語言中的正確翻譯，或是在某領域的適當翻譯。例如，「男人」在雙語字典中有許多翻譯，如 man、male、gentleman 等等，要檢索穿著十九世紀愛德華風格服裝(Edwardian Dress)的男人，比較適合的英文翻譯是 gentleman，但是想檢索正在處理魚的男人，適合的翻譯則是 fisherman 或是 man。使用者由於語言、文化以及背景知識的差異，往往無法下達最適合的檢索問句。

網際網路上的資訊使用各種不同語言寫成，依據 ETHNOLOGUE[1] 目錄上的統計(圖 1)，語言使用人口數的前幾名，依次為中文、英文、印度文及西班牙文等等。然而，根據 2001 年的估算(圖 2)，網頁使用語言的前幾名，依次為英文、日文、德文等等。由此可知，網際網路上約有 80%的網站為英文網站，卻有將近 40%的網際網路使用者並非以英文為母語。

[1] http://www.ethnologue.com

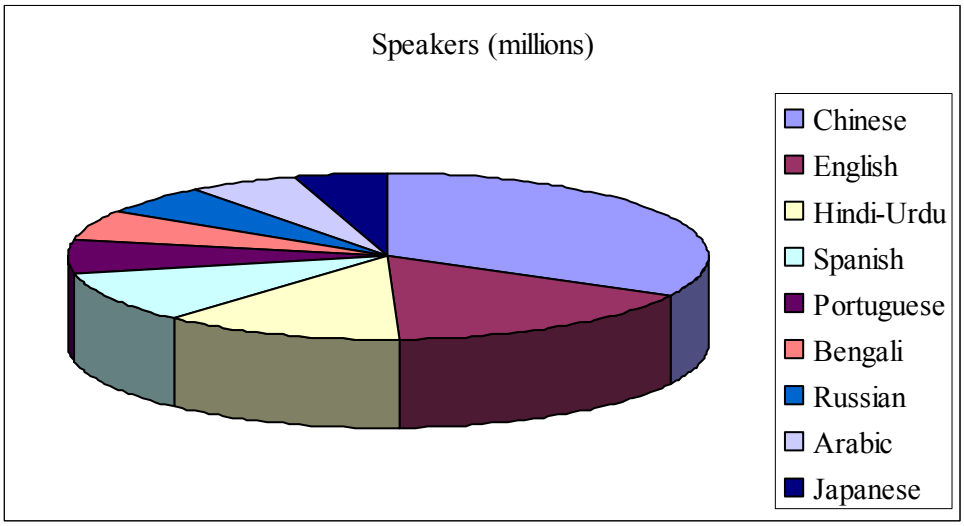

圖 1: 2001 年網際網路上主要語言的使用人口數統計

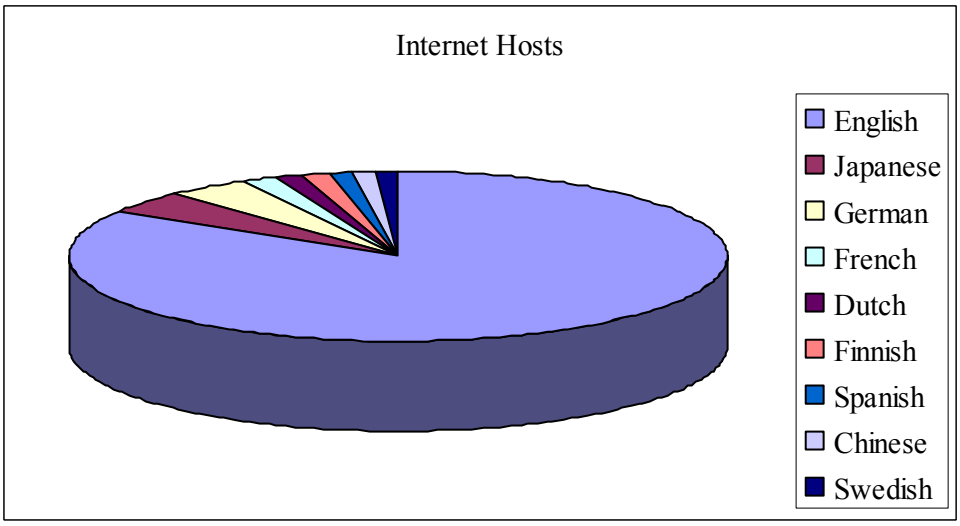

圖 2: 2001 年網際網路上網頁使用的語言統計

跨語言資訊檢索(Cross-Language Information Retrieval, 簡稱 CLIR)的目的即是消除語言的差異，使得使用者可以利用本身熟悉的語言，檢索其他語言的文件。CLIR 的應用相當廣泛。舉例來說，跨國犯罪常需要各國協同作業，然而語言的差異使得各國間犯罪資訊取得困難。歐洲各國因此而發展一套 AVENTINUS[2] (Advanced Information System for Multilingual Drug Enforcement)

[2] http://www.aventinus.de/

系統，以協助警方取得相關的緝毒與執法資訊。此系統中收集歐洲各國有關毒品、犯罪和嫌疑犯的多語言資料，並可以使用歐盟任何一種語言進行檢索。CLIR 系統亦可應用於數位典藏，例如，數位圖書館或數位博物館皆收藏大量的外語數位館藏，應用 CLIR 系統可以提供使用者使用熟悉的語言來查詢外語文件。除此之外，若將跨語言資訊檢索技術應用於搜索引擎(Search Engine)，便可容許使用者以其最熟悉的語言文字表達本身的資訊需求，並提供由各種語言所描述的相關資訊。

本論文以 ImageCLEF2004 [ImageCLEF04]資料集為例，實作一個可以實際應用於數位圖書館館藏檢索的跨語言資訊檢索系統，提供使用者以中文查詢條件檢索英文館藏資料。舉例來說，若使用者想查詢早期英國農耕的方式，然而受限於自身的外語能力無法精確地利用英文描述檢索問句時，跨語言資訊檢索系統便可幫助使用者達到檢索的目的。圖 3 為使用「犁」與「耕種」作為檢索問句所得到的結果。使用者不需具備相關的外語知識，即可查詢到蘇格蘭地區以馬耕種的相關圖片及敘述。由此可知，如何讓使用中文的使用者方便且快速檢索英文文件乃是本論文要探討的問題。

說明:
Man standing in just-opened furrow; single horse harnessed to plough; wire fence bounding field extends into centre distance.

圖 3: 使用「犁」及「耕種」跨語言檢索所得到的英國耕作相關文件

既然牽涉兩種以上的語言，因此檢索問句或者文件集兩者之一必須進行翻譯，如此檢索問句與文件集就屬於同一種語言，之後的處理方式和單語言資訊檢索相同。目前用於處理跨語言資訊檢索的相關技術主要可分為翻譯檢索問句(Query Translation)與翻譯文件集(Document Translation)兩類。翻譯文件的作法所

需的處理時間隨文件的不同而有極大的差異，而且計算量過於龐大，極少有系統採用這種作法，比較實際而且主流的作法是遵循翻譯檢索問句的研究。

第二節 研究動機

一般來說，中文跨語言資訊檢索系統除了要達到傳統資訊檢索系統的目的，更要能處理跨語言的問題。其中，會產生歧義性的模組主要有三部分(圖4)：斷詞歧義性，翻譯歧義性以及詞鍵歧義性。本論文探討前兩個歧義性，包括中文斷詞的問題、檢索問句過短導致語意不明確的問題以及翻譯時所產生的歧義性問題：

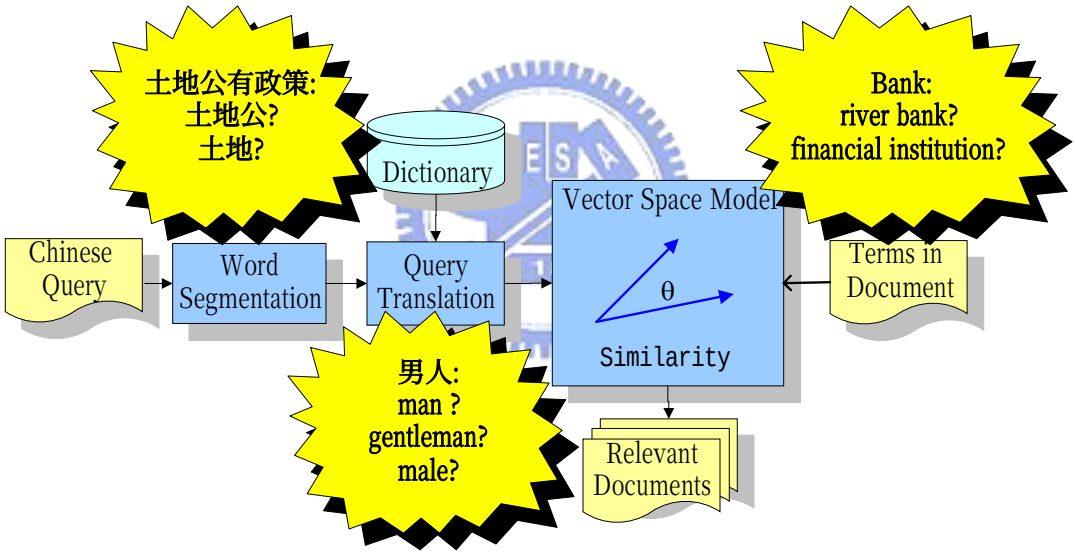

圖 4: 跨語言資訊檢索系統中，歧義性可能發生的模組

(1) 中文斷詞的問題

由 2003 年 ImageCLEF [ImageCLEF04]結果可知中文對英文的跨語言檢索效能比歐洲語系還要低，此乃因為拼音語系的詞使用空白分隔，因此判斷空白字元可得到詞，亦即最小有意義的單元。然而，亞洲語言如中文、日文、韓文等等，詞和詞之間並沒有決定性的分隔符號；以中文來說，斷詞的結果會影響

到語意的判斷，因此中文斷詞(Chinese Word Segmentation)也是必須考慮的問題。

(2) 檢索問句過短導致語意不明確的問題

使用者下達的檢索問句通常很短，導致無法判斷明確的語意。例如，使用者只有輸入「聖安得魯大學」作為檢索問句，系統無法得知使用者想檢索的是關於聖安得魯大學的學生職員的文件，或是有關校園景色的文件。因此，當檢索問句過短時，會造成語意資訊不足以決定正確的語意。

(3) 翻譯時所產生的歧義性問題

通常，一個中文字可能有很多相對應的英文翻譯，而適合的翻譯乃是取決於檢索問句的語境或是文件集的語境。例如，男人可以翻譯成 man，可以翻譯成 gentleman，何者是最適當的翻譯必須由前後文決定。

第三節 研究目的

本論文之研究目的在於探討跨語言資訊檢索系統的相關技術，並且針對第二節所述的三個問題提出解決方案：1) 使用雙語字典為本的斷詞方法以解決中文斷詞問題；2) 使用查詢自動擴展(Query Expansion)解決檢索問句過短的問題；3) 使用知識本體鏈(Ontological Chain)來解決翻譯歧義性的問題。同時，本論文亦以 ImageCLEF2004 資料集為例，應用上述解決方案，實作建構一個功能完整的跨語言資訊檢索系統。

第四節 論文架構

本論文首先於第二章介紹各項相關研究，包括檢索問句翻譯系統、解析翻譯歧義性、單語言資訊檢索系統；接著，第三章闡述本論文所提出的跨語言資訊檢索系統、各項模組的功能、採用的技術及解決方案。此章節將分別介紹系統架構、翻譯檢索問句流程及解析詞鍵歧義性。第四章針對本論文提出的知識本體鍵為本的跨語言檢索系統進行效能評估及實驗；最後，第五章總結本論文，並探討及說明未來的研究發展方向。

第二章　相關研究工作

本章依照本論文採用之技術，說明相關的研究工作。其中，第一節說明翻譯檢索問句(Query Translation)相關的研究工作，主要分為字典(Dictionary-based)為本以及語料庫為本(Corpus-based)的方法。第二節說明解析詞鍵歧義的相關研究工作(Word Sense Disambiguation)。

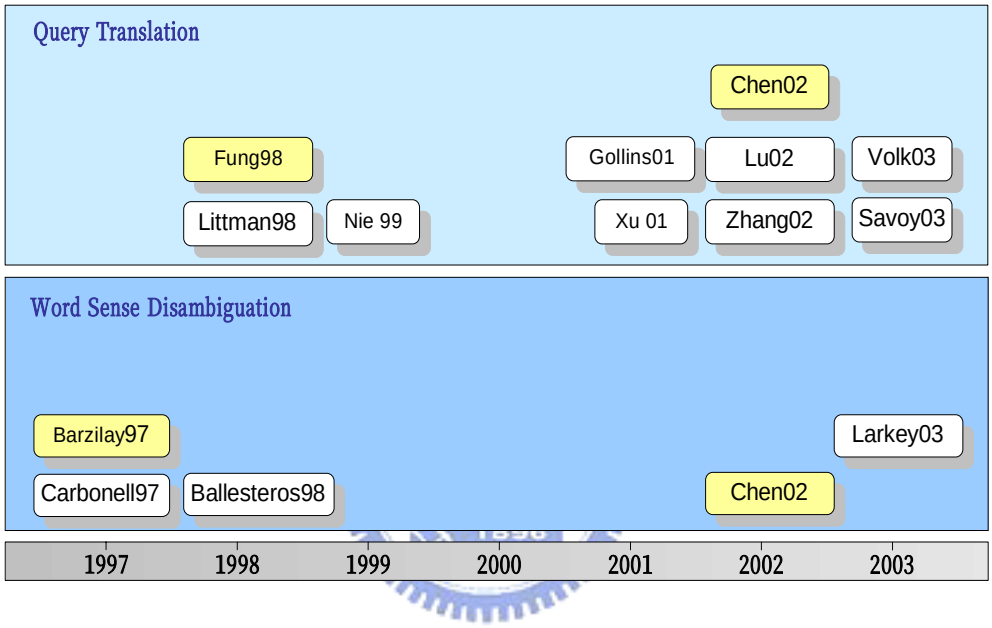

圖 5: 本論文相關的研究工作

如

圖 5 所示，跨語言檢索系統所要解決的問題主要包含以下兩個部分：

(1) 翻譯檢索問句

翻譯檢索問句將檢索問句翻譯成文件集所使用的語言，以使用單語言查詢技術。主要有以字典為本的方法，相關研究包含 [Chen02]、[Gollins01]、[Volk03]，及以語料庫為本的方法，相關研究包含 [Fung98] 、[Lu02]。

(2) 解析詞鍵歧義性

每一個詞鍵有許多意思，如何選擇最適切的意思，乃是重要的研究議題。解決方式主要有三種方法，以字典為本的方法，相關研究有[Barzilay97][Carbonell97]；監督式訓練方法及非監督式的方法。

第一節 翻譯檢索問句

翻譯檢索問句的目的，乃是將原始語言(Source Language)所寫成的檢索問句翻譯成文件所屬的目的語言(Target Language)。主要的方法有兩種，一個是以字典為本，一個以語料庫為本。以字典為本的方法是從字典所有可能的翻譯裡面選擇適當的翻譯。以語料庫為本的方法則是從平行語料庫(Parallel Corpus)，或是比較語料庫(Comparable Corpus)中學習出正確的翻譯。其中，平行語料庫為一中英文對照之語料庫，且每一句中文都有其相對應的英文語句，因此可以由字在語句中的相對位置，判斷該字在另一個語言的正確翻譯。比較語料庫中的對應則是一篇中文文件對應到一篇英文文件，缺乏句與句間的對應關係，因而無法從位置關係來判斷翻譯。

本節分為兩部分，第一部分說明以字典為本的檢索問句翻譯方法，第二部分則介紹以語料庫為本的檢索問句翻譯方法。

2.1.1 以字典為本的翻譯檢索問句方法

以字典為本的方法利用查詢雙語字典的方式，將原始語言翻譯成目的語言。然而，一個詞鍵可能有多個翻譯結果，因此，便需要有選擇翻譯的策略。相關研究中所採用的策略主要有以下幾種：1) 選擇排列第一的翻譯：字典羅列的翻譯中，第一個翻譯通常為一般狀況下最常使用的意思；2) 選擇所有的翻譯：將所有的翻譯都視為正確，但存在有翻譯歧義性問題；3) 選擇最佳的 N 個翻譯：藉

由判斷檢索問句語境，作為選擇該字最適合問句語境的翻譯。

[Chen02] 採用選擇最佳 N 個翻譯的策略。其所使用的漢英雙語詞典匯集多部現有的電子版詞典，包括致遠漢英詞典 2.2 版、LDC 雙語詞典及英漢雙語詞典等，共有 20 萬個詞彙。[Chen02] 將一個中文詞集合翻譯成相對應的英文詞集合；主要有兩個步驟，先產生一個英文詞集合的初始集合，再依據這個初始集合產生完整的英文詞集合，茲將[Chen02]的做法簡述如下。

首先，從中文詞集合中挑選出存在於雙語詞典中且只具有單一英文翻譯之中文詞；這些中文詞在翻譯成英文時，並不需要解決「翻譯歧義性」的問題，可以將這些英文翻譯當成英文語境的初始集合。如果中文詞集合中，並不存在具有此特性的中文詞集合，或是具有單一英文翻譯之中文詞個數太少時，則找出前 k 個(k<=*10*)具有兩個英文翻譯的中文詞。此 k 個中文詞的英文翻譯會構成 2^k 種組合，公式 1 定義兩個英文詞之間的交互資訊(Mutual Information, MI)值，用以計算兩個英文詞的關聯程度。接著，依照公式 2，可以計算出兩兩詞彙間 *MI* 值的總和，當成是這一組英文翻譯的 *MI* 值。同時，找出具有最大的 *MI* 值總和的英文翻譯組合，即 $ew_1, ew_2, \ldots ew_k$，並將這組翻譯加入英文詞初始集合中。

$$MI(ew_i, ew_j) = \log_2 \frac{P(ew_i, ew_j)}{P(ew_i)P(ew_j)} \approx \log_2 \frac{f(ew_i, ew_j)}{f(ew_i)f(ew_j)} \times N$$

公式 1: 計算兩個英文詞鍵間的交互資訊值[Chen02]

$$\arg\max_{ew_1 ew_2, \ldots, ew_k} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} MI\ (ew_i, ew_j)$$

公式 2: 2^k 組合中選出最適當的翻譯的方法[Chen02]

建立起英文詞的初始集合後，對於剩下尚未經過翻譯之中文詞，依照它們在雙語字典中找到的英文翻譯個數，由少至多依序排列，再利用公式 3 從英文翻譯個數最少的中文詞處理起。假設 i-1 個中文詞已被翻譯成英文，並放入英

文詞初始集合中，集合 S 中應該已存在了 i-1 個英文詞，這個初始集合作為英文語境集合的初始值。第 i 個中文詞 cw_i 在雙語詞典中可查到 n 個英文翻譯，分別是 $ew_{i1}, ew_{i2}, ..., ew_{in}$，透過公式 2，對每一個 $ew_{ij}\,(j=1...n)$ 分別和目前的英文語境集合計算 *MI* 值，此 *MI* 值是從 ew_{ij} 與集合中的每個英文詞間 *MI* 值的總和所得到。最後，利用公式 3，比較 $ew_{i1}, ew_{i2}, ..., ew_{in}$ 個別算出的 *MI* 值，選 *MI* 值最高的 ew_{ij} 作為中文詞 cw_i 最適當的英文翻譯。將這個被選出之 ew_{ij} 加入英文語境集合 S 中，更新英文語境集合，再處理下一個中文詞，直到所有中文詞處理完為止。表 1 為一個將一個中文詞集合翻譯成英文詞集合的例子。

$$\arg\max_{j} \sum_{k=1}^{i-1} MI(ew_{ij}, ew_k)$$

公式 3: 當第 i 中文詞被翻譯成英文的選擇公式[Chen02]

Synonym Set in Cilin	打、拍、撫摸、搔、摸
Sense Vector	踢、叫好、罵、樂團、安打、自信心
English Version	Play, applaud, abuse, band, bingle, confidence

表 1: 中文詞集合翻譯為英文詞集合的例子 [Chen02]

2.1.2 語料庫為本的翻譯檢索問句方法

[Fung98] 使用比較語料庫(Comparable Corpus)的語境，學習出中文字詞的英文翻譯。[Fung98] 假設一個字的語境可由其前後文(Context)的文字所決定，亦即意思相同的中文字和英文字，他們在文件中會擁有類似的語境。如表 2 所示，「流感」與「flu」具有相似的語境。舉例來說，「流感」前後文中出現 147 次「病毒」，相對地，「flu」前後文中亦出現 26 次「virus」。

English Word	Frequency	Chinese Word	Frequency
bird	170	病毒	147
virus	26	市民	90
spread	17	香港	84
people	17	感染	69
government	13	證實	62
avian	11	表示	62
scare	10	發現	56

表 2: 「流感」與「flu」前後文的字頻統計[Fung98]

由上可知，中文字語境與其英文翻譯的語境共同出現次數，以及語境詞的順序，皆可用來當作翻譯的依據。例如，「病毒」在中文語境的出現次數很高，其英文翻譯「virus」在英文語境中出現次數也很高，因此，「病毒」與「virus」即是橋接「流感」和「flu」的強烈線索。

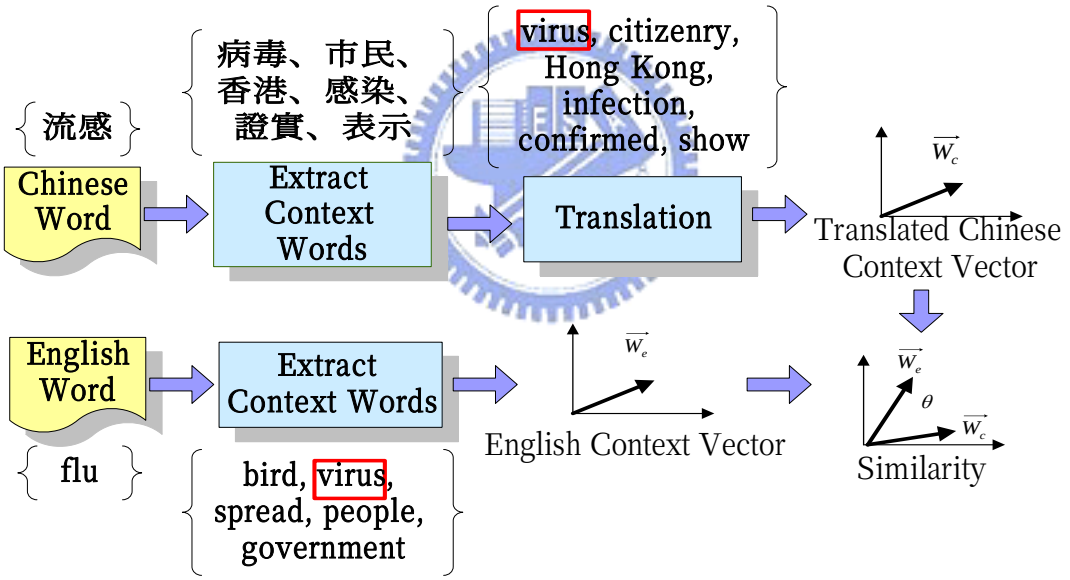

圖 6: 中文詞鍵翻譯成英文的流程圖[Fung98]

為計算中文詞與英文詞相似度的流程，對於一個中文字，例如「流感」，取出比較語料庫中該字的中文語境，亦即每篇中文文件中「流感」前後的文字，將這個中文語境利用雙語字典翻譯成英文，採用向量空間模型(Vector Space Model)將翻譯過後的語境以向量表示；而英文方面也是從語料庫中取出「flu」的英文語境，採用向量空間模型將一個語境表示成語境向量。英文語境向量和

翻譯過後的中文語境向量位於同一個向量空間，可以計算兩個向量間 Cosine 值衡量相似程度。

[Fung98] 亦提出了幾種不同的模型(S0-S7)以計算向量中詞鍵權重與相似度。S0 考慮考慮詞鍵頻率(Term Frequency, TF)[Salton83]，即詞鍵在語境中的出現次數，如公式 4：

$$S0(W_c, W_e) = \frac{\sum_{i=1}^{t}(w_{ic} \times w_{ie})}{\sqrt{\sum_{i=1}^{t} w_{ic}^{\ 2} \times \sum_{i=1}^{t} w_{ie}^{\ 2}}}$$

$$w_{ic} = TF_{jc}$$

$$w_{ie} = TF_{je}$$

公式 4: 考慮詞頻的語境相似度公式[Fung98]

然而，語境詞鍵頻率只計算該詞鍵在語境的頻率，並無考慮詞鍵於文件集中所出現的頻率。舉例來說，[Fung98] 之研究所使用的語料庫為香港明報(HK Standard/Mingpao Corpus)，其中出現頻率最高的字為「Hong Kong」，但是這不表示Hong Kong就是某個中文字的翻譯。消除這類問題最常用的方法是逆向文件頻率(Inverse Document Frequency, IDF)[Salton83] 。以此例來說，「virus」與「Hong Kong」的IDF值分別是1.81及1.23，「病毒」與「香港」的逆向文件頻率則為1.92和0.81。S1針對每個字都可以給一個權重 $W_{ij} = TF_{ij} \times IDF_i$ ，修正的相似度函數如公式 5：

$$S1(W_c, W_e) = \frac{\sum_{i=1}^{t}(w_{ic} \times w_{ie})}{\sqrt{\sum_{i=1}^{t} w_{ic}^{\ 2} \times \sum_{i=1}^{t} w_{ie}^{\ 2}}}$$

$$w_{ic} = TF_{jc} \times IDF_i$$

$$w_{ie} = TF_{je} \times IDF_i$$

公式 5: 考慮 TF*IDF 的語境相似度公式[Fung98]

除此之外，Dice係數[Frakes92]也被用來比較相似程度。

$$S2(W_c, W_e) = \frac{2\sum_{i=1}^{t}(w_{ic} \times w_{ie})}{\sum_{i=1}^{t} w_{ic}^{\ 2} + \sum_{i=1}^{t} w_{ie}^{\ 2}}$$
$$w_{ic} = TF_{jc} \times IDF_i$$
$$w_{ie} = TF_{je} \times IDF_i$$

公式 6: 考慮 Dice 係數的語境相似度公式[Fung98]

兩個權重相似度公式可以互相組合，定義 $S3 = S1 \times S2$ 。

S1用來比較簡短檢索問句(short query)和一篇文件的相似程度，S2則用來比較兩篇文件內容的相似程度。此外，橋接中英文字的種子品質相當重要，首先中文斷詞就會引入一些種子詞鍵的模糊性，而中英翻譯又會引入更多模糊性。[Fung98] 針對此現象提出每個翻譯配對皆引進信心權重(Confidence Weighting)的計算方式；假設一個英文字i_e是中文字i_c第k個候選翻譯，則將權重除以k。S4, S5為考量此情形時的相似度計算方式，如公式7以及公式8；而 $S6 = S4 \times S5$ 為考量S4及S5組合時的狀況。

$$S4(W_c, W_e) = \frac{\sum_{i=1}^{t}(w_{ic} \times w_{ie})/k_i}{\sqrt{\sum_{i=1}^{t} w_{ic}^{\ 2} \times \sum_{i=1}^{t} w_{ie}^{\ 2}}}$$
$$w_{ic} = TF_{jc} \times IDF_i$$
$$w_{ie} = TF_{je} \times IDF_i$$

公式 7: 考慮信心權重的相似度公式[Fung98]

$$S5(W_c, W_e) = \frac{2\sum_{i=1}^{t}(w_{ic} \times w_{ie})/k_i}{\sum_{i=1}^{t} w_{ic}^{\ 2} + \sum_{i=1}^{t} w_{ie}^{\ 2}}$$

$$w_{ic} = TF_{jc} \times IDF_i$$

$$w_{ie} = TF_{je} \times IDF_i$$

公式 8: 考慮信心權重和 Dice 係數的相似度公式[Fung98]

Model	English	Chinese	Score
S0	Lei	流感	0.18111
	flu	流感	0.08888
	Tang	流感	0.08589
	AP	流感	0.08141
S4	flu	流感	0.12088
	Lei	流感	0.09758
	Beijing	流感	0.06866
	poultry	流感	0.06583
S5	flu	流感	0.08629
	China	流感	0.04009
	poultry	流感	0.02816
	Beijing	流感	0.0245
S6	flu	流感	0.01043
	poultry	流感	0.00185
	China	流感	0.00184
	Beijing	流感	0.00168
S7	flu	流感	0.00767
	poultry	流感	0.00196
	Beijing	流感	0.00167
	China	流感	0.00139

表 3: 不同的相似度公式對「流感」的計算結果

[Fung98]適用於雙語字典查不到，但是經常在語料庫中出現的中文字和英文字，可以使用新詞粹取工具來找這些字。爲了從香港明報語料庫中學習出「流感」的翻譯，首先從新聞語料庫中選擇118個字典查不到的英文字作為可能的翻譯，使用相似度公式S3~S6，從表 3可以看出最好的相似度公式是S6和S7。測試其他不在字典，但是經常在語料庫中出現的中文字，找出未知的中文字和英文字作比對。其中排除斷詞歧義性和翻譯歧義性，例如林是個姓，也可以是森林這個雙字詞的一部分，所以這個字具有斷詞歧義性。從表 4可以得知沒有成功找到翻譯的只有葉利欽和農曆，其餘的中文字都可以找到正確的翻譯。另外，禽流感在英文中稱為「bird flu」，但是中文使用「禽」這個字而不是「鳥」，所以「禽」是「流感」的語境向量，被翻譯成英文語境向量成為「poultry」，因此翻譯會出現

「poultry」。

Score	English	Chinese	Score	English	Chinese
0.008421	Teng-hui	登輝	0.004275	Kalkanov	珠海
0.007895	SAR	特區	0.00355	poultry	鴨
0.007669	Flu	流感	0.003519	SAR	葉利欽
0.007588	Lei	鴨	0.003481	Zhuhai	建華
0.007283	Poultry	家禽	0.003407	Prime Minister	林
0.006812	SAR	建華	0.003407	President	林
0.00643	Hijack	登輝	0.003338	Flu	家禽
0.006218	Poultry	特區	0.003324	apologies	登輝
0.005921	Tung	建華	0.00325	DPP	登輝
0.005527	Diaoyu	登輝	0.003206	Tang	唐
0.005335	Prime Minister	登輝	0.003202	Tung	梁
0.005335	President	登輝	0.00304	Leung	梁
0.005221	China	林	0.003033	China	特區
0.004731	Lien	登輝	0.002888	Zhuhai	農曆
0.00447	Poultry	建華	0.002886	Tung	董

表 4: [Fung98]從新聞語料庫尋找中英翻譯的結果

第二節　解析詞鍵歧義性

本節介紹詞鍵語意的問題以及解決方法，2.2.1 介紹三種解析詞鍵歧義性的方法，2.2.2 介紹同義詞典方法用到的兩個資源，英文文件使用 WordNet，而中文文件則使用同義詞詞林。2.2.3 說明一個混合英文 WordNet 以及中文同義詞詞林，建立中英對照的 WordNet，並且用來解析詞鍵歧義。2.2.4 則說明語彙鏈結解析詞鍵歧義的方法。

2.2.1 建構中英對照的WordNet解析詞鍵歧義性

判斷詞鍵語意的問題稱為詞鍵歧義性解析(Word Sense Disambiguation, 簡寫為WSD)，主要是針對一個具有歧義性的詞，從這個詞(Word Form)可能擁有的所有詞義(Word Meanings)類別中，分辨出它目前在文章中所表現的詞義。解決詞義歧義性的方法分為三種，一種是直接利用字典或同義詞典所提供的詞義資訊。第二種方法是監督式(Supervised)的訓練方法，利用已標定好詞義標記的語料庫，訓練出每個詞義的語境，比較語境間相似度後辨別出正確的詞義。這種方法需要有規模夠大的語料庫，且在語料庫的標定工作上，通常需要大量的人力介

入，因此，語料庫的取得是這個方法的一大瓶頸。第三種方法是，由未經任何標定處理的語料庫raw corpus）中，訓練出可用來區辨詞義的資訊，此種方法是非監督式(unsupervised)的訓練方法。本節介紹的兩個相關研究都是屬於第一種，也就是利用同義詞典提供的詞義資訊。

解決詞義歧義性的問題，同義詞典即提供一個方便而完整的詞義分類資訊來源，它將所有同義的字或詞集合在一起，成為一個詞義類別，這個詞義類別的定義以及所涵蓋的詞義範圍，就可由集合中這些同義字或詞的共同性得知。當然，不同的同義詞典間，其所定義的詞義類別個數與範圍可能會有所出入。先前做過的許多研究，通常都是藉由同義詞典來提供詞義的分類項目及詞義資訊。例如在中文方面的同義詞典有「同義詞詞林」。在英文解決詞義歧義性問題的研究上是利用 Roget's International Thesaurus [Kipfer01]或是 WordNet [Miller95]。

2.2.2 WordNet 以及同義詞詞林簡介

WordNet [Miller95]是在 1990 年由 George A. Miller 等人所提出，是普林斯頓大學的一個計畫，該計畫被稱為「英語詞彙資料庫」(WordNet)，屬於同義詞典(Thesaurus)的一種。它使用同義詞集合(Synonym Sets，或稱 Synsets)來描述和分類詞鍵及概念。它和一般同義詞典的不同處在於，它比同義詞典增加了更多的訊息和知識，在 WordNet 每個同義詞集合間，都有一些關聯性指標(Relational Pointers)以同義詞集為節點，透過語意關係建立節點間的連結，就形成了詞彙語意的關係網絡。「關聯性指標」是指如「上下位」關係(Hyperonymy-Hyponymy)，例如圖 7 中「非洲國家」(African Country)是一種「國家」(Country)，所以國家是非洲國家的上位詞；同理，英格蘭是歐洲國家的下位詞。其他還有「反義」(Antonym)關係、「導致」(Cause)關係等多達數十種語意關係。在 WordNet 架構中，依照詞性分成名詞(noun)、動詞(verb)、形容詞(adj)、副詞(adv)等四類，

每一類各有其關聯性指標，但這些指標都只指向同一詞性的同義詞集合，而不指到屬於不同詞性的同義詞集合。WordNet 針對這四個詞性，共分成四十四個大類，將近十萬個同義詞集合。

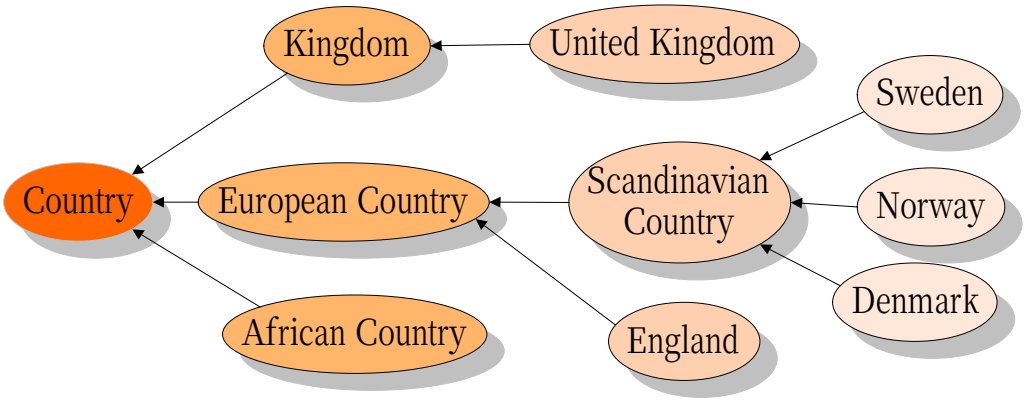

圖 7: WordNet 結構的範例[Miller95]

以WordNet 為例，可利用每個詞義集合(Synset)中所包含的詞，及這些詞在這個詞義集合中的定義(Definitions)和例句(Glosses)等，來區別詞義集合間的差異。此外，WordNet 中由上位詞、下位詞等關聯性指標所建立起之階層式架構，可用來計算兩個詞彙之間的概念距離(Conceptual Distance)或者概念密度（conceptual density），利用這些計算方式，可以比較詞義間的相似度，進而對這些詞彙進行解決詞義歧義性的工作。

「同義詞詞林」是由大陸學者編輯，收錄了近七萬的詞彙，全部按詞義編排。本書除了以詞義為分類原則，也兼顧詞類。它把詞語分為大、中、小類三級，共分12 個大類，94 個中類，1428 個小類，小類下再依同義原則劃分詞群，分成3925 個詞群。圖 8的例子中「人」是大類，「男女老少」是中類，「老人」是小類別，而小類之下還會有詞群。[Chen02]的研究對中研院平衡語料庫標定詞義標記是以1428 個小類做為詞義標定時的詞義標記。同義詞詞林中對詞類的分類大致是：屬於為A 和D 大類的詞大部份是名詞，屬於E大類的大部份是形容詞，屬於F和J 大類的大部份為動詞，屬於K 大類的為助語，L 大類則為敬語及問候語。

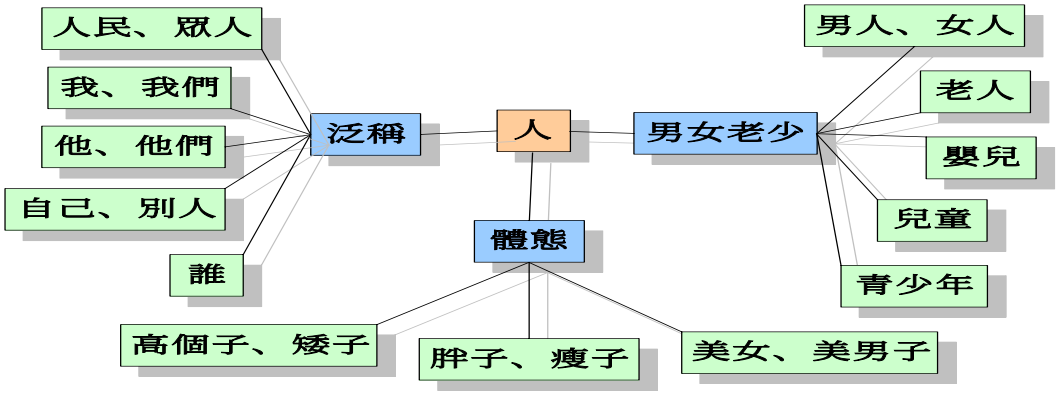

圖 8: 同義詞詞林的例子[Chen02]

2.2.3 建構中英對照的WordNet解析詞鍵歧義性

[Chen02] 使用了五個資源，包括「WordNet」、「同義詞詞林」、「中研院平衡語料庫」、「SemCor」語料庫以及中英字典。整合這五個語言學資源建立一個中英文對照的 WordNet，可以用來解析詞鍵歧義性。除了 2.2.2 節介紹的 WordNet 和同義詞詞林外，以下先說明另外兩個語言學資源「中研院平衡語料庫」以及「SemCor」語料庫。

「中央研究院平衡語料庫」（簡稱「研究院語料庫」，Sinica Corpus）它是世界上第一個具有完整詞類標記的中文語料庫，由中央研究院資訊所、語言所詞庫小組完成的。1997 年中研院所開放的版本已具有五百萬詞的規模。此語料庫專門針對語言分析而設計的，每個文句都依詞斷開，並標示詞類。語料的蒐集也盡量做到現代漢語分配在不同的主題和語式上，分為六大類，「哲學」（10%）、「科學」（10%）、「社會」（35%）、「藝術」（5%）、「生活」（20%）「文學」（20%）。[Chen02] 以自動的方式為其加標詞義標記。

SemCor(Semantic Concordance) 是一部具有詞類和詞義標記的小規模語料庫，其來源是從知名的布朗語料庫（Brown Corpus）中擷取出一小部份，以 WordNet 的同義詞集合為標記，為每個字加標上詞義標記。由於布朗語料庫本

身已標有詞類標記，再加上人工為其所標定的詞義標記，因此，所建構出的 SemCor 是一個同時具有詞類及詞義標記的英文語料庫。

圖 9是建立中英對照WordNet的流程，步驟主要如下：

1. 對中研院平衡語料庫標定詞義標記，建立一部可提供詞義關係資訊的中文語料庫。
2. 訓練出每個中、英文詞義的語境，並將中文語境轉成以英文表現。
3. 建立同義詞詞林之詞義標記與WordNet之synsets間的對應關係表。
4. 建立中文部份的詞彙知識庫，並進而與英文的WordNet連結，建構成一部可雙向查詢的英中詞彙知識庫。

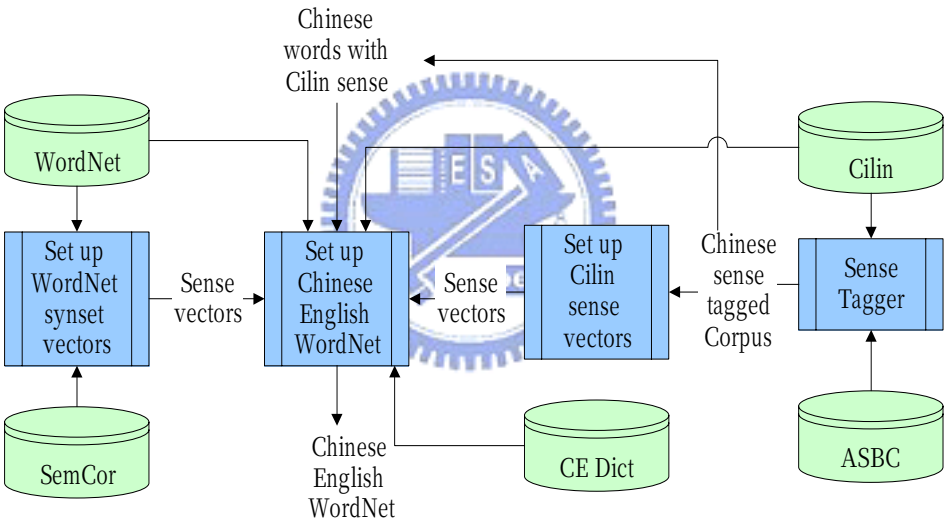

圖 9: 建立中英對照 WordNet 的流程圖[Chen02]

[Chen02]也是使用語境向量代表語意，但是和[Fung98]不同的是語境向量由 WordNet 的同義詞集合定義和例句計算得來，而不是來自語料庫。WordNet 中的定義和例句，是用來解釋每一個同義詞集合所代表的含義，以及某英文詞在此同義詞集合時的用法與例句，再加上每個同義詞集合中均列出了屬於此集合詞義的所有英文詞。因此，剛好可藉由這些資訊建立起每個同義詞集合之語境向量。除了把定義和例句中的 Stopwords 去除之外，剩下的字全部當作此同義詞集合的上下文資訊(Contexts)。把從 SemCor 和 WordNet 中訓練所得的上

下文資訊集合起來，所產生的每個 synset 之語境向量，就同時包含了這兩個資源所提供的訊息。最後，計算出上下文詞群中的每個詞與此同義詞集合出現在語料庫中的交互資訊值，當作語境向量中每個元素的值。在比較中文語境向量與英文語境向量和[Fung98]的做法類似，將中文語境翻譯成為英文語境，然後計算英文語境向量和翻譯過後的語境向量之間的相似程度。

用上述方法建立中英對照的 WordNet 之後，可以應用這個資源來解析詞鍵歧義性。針對每一個中文詞，使用中英對照 WordNet 對應到許多的英文同義詞集合，從中挑出最適合的一個同義詞集合作為這個中文字真正的意思，作到解析詞鍵歧義性。首先定義兩個同義詞集合之間的交互資訊，例如：同義詞集合 A 中包含有 a1,a2,...,am m 個詞，同義詞集合 B 中有 b1,b2,...,bn n 個詞，則 A 與 B 間的 MI 值即是：

$$MI(synset_A, synset_B) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} MI(a_i, b_j)}{m \times n}$$

公式 9: 兩個同義詞集合的交互資訊定義[Chen02]

例如要解析「國際組織犯罪」的語意，在中英對照WordNet中，「國際」可查到兩個同義詞集合，「組織」可查到兩個同義詞集合，「犯罪」有三個同義詞集合。在表 5中分別列出兩兩同義詞集合間的*MI* 值。沒有列出*MI* 值的集合表示在語料庫中找不到這兩集合的交互資訊。對每一個中文詞，從它所有的同義詞集合與其它詞的同義詞集合配對中，選出一個最高*MI* 值。例如就「國際」而言，它在表中所列出的所有*MI* 值中，最高的*MI* 值4.394 是出現在syn11 中，因此對「國際」一詞就選擇syn11 當作語意。同樣地，「組織」的最高*MI* 值4.394 是出現在syn22 中，「犯罪」的最高*MI* 值3.899 出現在syn31 中，因此對這兩個詞就分別選擇syn22 與syn31 作為語意。

		國際		組織		犯罪		
		syn11	syn12	syn21	syn22	syn31	syn32	syn33
國際	syn11			1.517	4.394	1.233	0.444	1.583
	syn12			0	0	0	0	0
組織	syn21	1.517	0			-0.061	0.028	-0.536
	syn22	4.394	0			3.899		0.417
犯罪	syn31	1.233	0	-0.061	3.899			
	syn32	0.444	0	0.028	0			
	syn33	1.583	0	-0.536	0.417			

表 5: 解析「國際組織犯罪」語意的例子[Chen02]

2.2.4 利用語彙鏈結解析語意歧義性

語彙鏈結(Lexical Chain) [Barzilay97]是文章中具有相同意義或是直接、間接關係的字詞所組成的集合，每個語彙鏈結代表文章中所描述的一個概念(Concept)。建立語彙鏈結的主要步驟如下：

1. 挑選候選的名詞。
2. 對於每個候選的詞鍵，針對每個語彙鏈結，衡量該詞鍵所代表的語意與語彙鏈結中每個詞鍵的語意關聯度，藉此找出相關聯的語彙鏈結。
3. 如果找到適當的語彙鏈結，便將該詞鍵加入該語彙鏈結中；如果沒有找到的話，便建構新的語彙鏈結。

[Barzilay97]根據WordNet中詞鍵之間的關聯結構來衡量關聯強度，若某鏈結為同義詞關係，則給予10 分；完全關係(Holonym)給予7 分；上位詞關係則給予4 分。

以下說明如何建構語彙鏈結。

Mr. *Kenny is the* ***person*** *that invented an anesthetic* ***machine*** *which uses* ***micro-computers*** *to control the rate at which an anesthetic is pumped into blood. Such* ***machines*** *are nothing new. But his* ***device*** *uses two* ***micro-computers*** *to achieve much closer monitoring of the* ***pump*** *feed the anesthetic into*

表 6: 建立語彙鏈結的原始文章[Barzilay97]

語彙鏈結只考慮名詞，以表 6 文章為例，對於第一個名詞「Mr.」建構一個獨立的語彙鏈結，接著考慮第二個字詞「person」。由 WordNet 中可知該字具有三種不同的涵義：(1)人類(human being)；(2)人的身體(a person's body)以及(3)人稱，文法上的分類(grammatical category of pronouns and verb forms)。考慮所有可能的鏈結組合，如

圖 10 會產生三種可能。

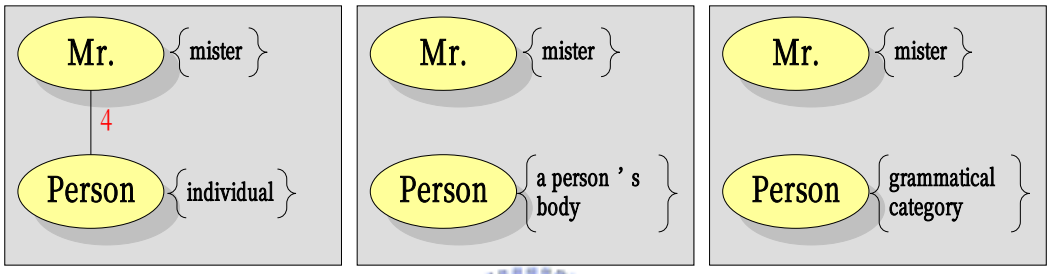

圖 10: Mr.和 Person 的三種可能語意組合[Barzilay97]

其中，「person」語意為人類(human being)時是「Mr.」的上位詞，因此它們之間的關聯強度為4 分，person的另外兩種語意和Mr.沒有任何關係，因此沒有分數。再接著考慮第三個字「machine」，它在WordNet中有五種語意：(1)有效率的人(an efficient person)，例如這個拳擊手是一種專長打鬥的人，英文是：The boxer was a magnificent fighting machine。這句的「machine」指的是人而不是機器；(2)機械或電子裝置；(3)很有效率的機構或是組織；(4)控制政黨的幾個人，黨機器之意；(5)裝置。由WordNet可知「machine」作為有效率的人時和「person」之間有包含關係，也就是說「machine」和「Mr.」也有間接關係。所以可能的關係如

圖 11。

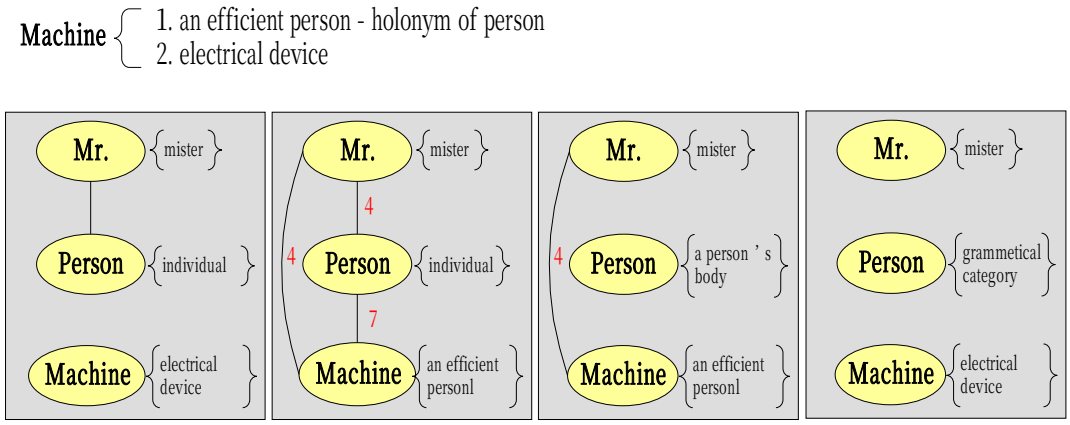

圖 11: 「machine」、「person」和「Mr.」 之間的可能關係[Barzilay97]

上例中最後建構出的語彙鏈結有兩種可能，如圖 12所示，衡量「machine」 在上圖的鏈結強度為11 分，下圖的強度是30 分。因此以下圖作為語彙鏈結結果。圖 12中清楚地看到「Mr.」與「person」在同一個鏈結，「machine」、「micro-computer」、「device」以及 「pump」則在另一個鏈結，可知語彙鏈結可以反映出某字詞在文件中的語意。

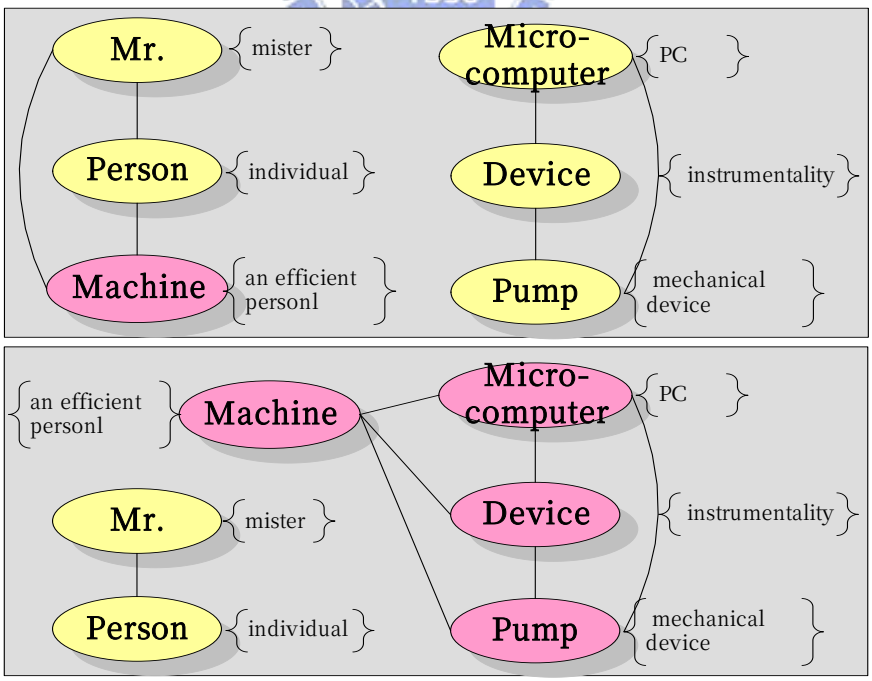

圖 12: 語彙鏈結的結果[Barzilay97]

第三章　跨語言資訊檢索系統之設計

本章將探討本論文設計之跨語言資訊檢索系統(Cross Language Information Retrieval,簡稱 CLIR)。第一節介紹跨語言資訊檢索系統的所有模組以及模組間的運作方式；第二節介紹翻譯檢索問句模組；第三節介紹知識本體的建立以及利用知識本體鏈來解析翻譯歧義性問題；第四節介紹英文的單語言資訊檢索系統。

第三節　系統架構

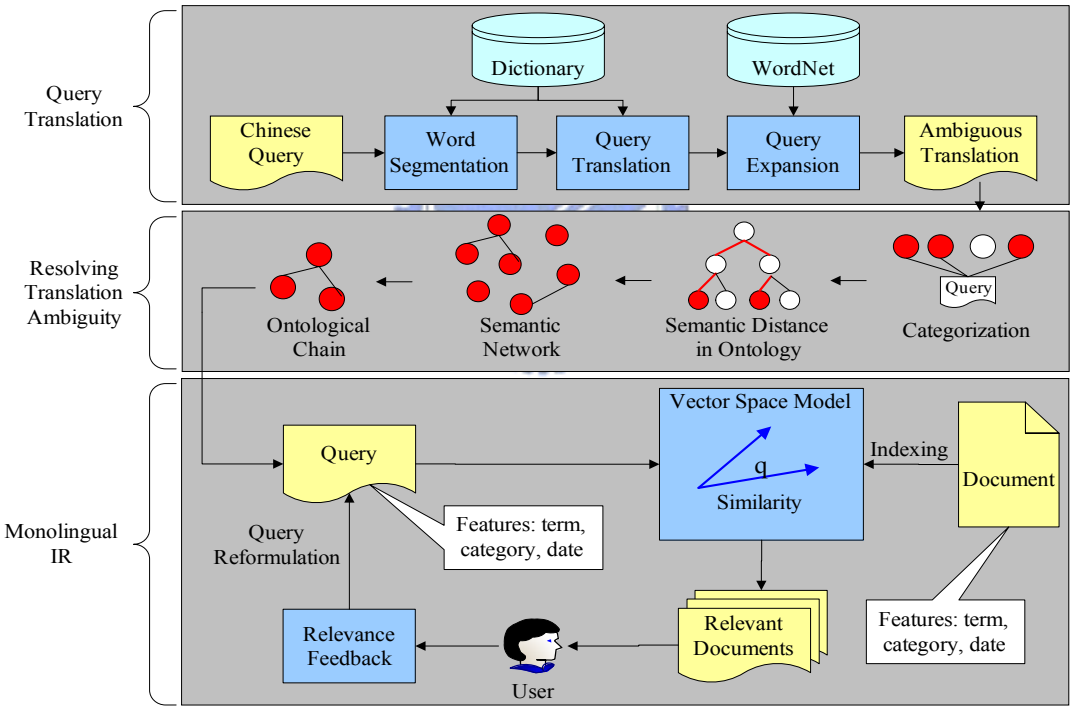

圖 13: 本論文系統架構圖

本論文提出的跨語言資訊檢索系統如

圖 13 所示，包括三個模組：

1. 翻譯檢索問句(Query Translation)：將使用者的中文檢索問句翻譯為英

文檢索問句，取所有可能的英文翻譯，以及其同義詞，上位詞，下位詞。

2. 解析翻譯歧義性(Resolving Translation Ambiguity)：查詢知識本體，找出和檢索問句最相關的節點來建立知識本體鏈，利用此知識本體鏈對於每個中文查詢詞，取出最適當的英文翻譯。
3. 單語言資訊檢索系統(Monolingual Information Retrieval System)：將解析過後的英文查詢輸入英文單語言資訊檢索系統中，找出最相似的文件。

第二節 翻譯檢索問句

翻譯檢索問句主要有三個步驟，首先將中文檢索問句作中文斷詞(Chinese Word Segmentation)，找出最小有意義單元；其次使用雙語字典翻譯檢索問句中的所有詞鍵(Query Translation)；最後利用 WordNet 擴展查詢詞(Query Expansion)。經過這三個步驟，中文的檢索問句將會被翻譯成英文檢索問句，但是這個英文檢索問句包含了所有的可能翻譯和所有相關英文詞，所以是語意模糊的。

3.2.1 中文斷詞

字(Word)在英文語言裡面是最小有意義單元，而字的邊界可以用空白字元或標點符號來判斷，也就是每個英文字都是用空白或標點符號隔開。但是在中文語言裡，詞(Phrase)才是最小有意義單元，字並無法包含正確的語意，例如「羅馬」這個詞，如果分成單獨的字「羅」或是「馬」，並無法代表「羅馬」的語意，所以字並不是中文最小有意義單元，必須從檢索問句中準確找出中文詞的邊界才可得知使用者的意圖。

本論文混合使用雙語字典以及語料庫作為中文斷詞的依據，首先針對一個

句子產生二字詞以及三字詞的所有可能組合，從雙語字典中查詢每個詞，如果該詞可以翻譯成為英文，則取該詞為斷詞結果。

3.2.2 翻譯查詢詞

使用兩個中英翻譯軟體的字典檔案，包括 Linux 的開放原始碼字典軟體 pyDict 以及遠東 21 世紀字典。一個中文字可能會有一個或以上的英文翻譯，要判斷檢索問句的正確翻譯方式必須考慮檢索問句的語境(Context)，以及文件集的語境。在這個步驟無法判斷兩者的語境，所以無法找出適當的翻譯，而是選擇所有可能的翻譯。

3.2.3 查詢自動擴展

使用者的中文翻譯問句可以使用雙語字典翻譯成英文問句，再使用 WordNet 將英文查詢詞的同義詞，上位詞以及下位詞作查詢擴展。一個英文詞有許多意思，如果只用關鍵字比對，會無法找到相關的字詞。如

圖 14，使用者想找「fish」相關的文件，如果只用「fish」作為關鍵字，只會找到「feeding fish」的文件。但是透過 WordNet 可以得知「fish」有五個意思：

1. 「魚」：上位詞是動物，下位詞有青魚（Herring）和鮭魚（Salmon）等等。
2. 「魚肉」：上位詞是食物，下位詞是可以在盤中煎的魚（Panfish）
3. 「雙魚座」：上位詞是人（Person）。
4. 「找尋」：上位詞是搜尋（Search）。
5. 「釣魚」：上位詞是補捉（Catch）。

利用 WordNet 做查詢擴展可以得到「salmon」、「herring」、「catching fish」等等和「fish」相關的文件，若純粹關鍵字比對則無法達成這種效果。

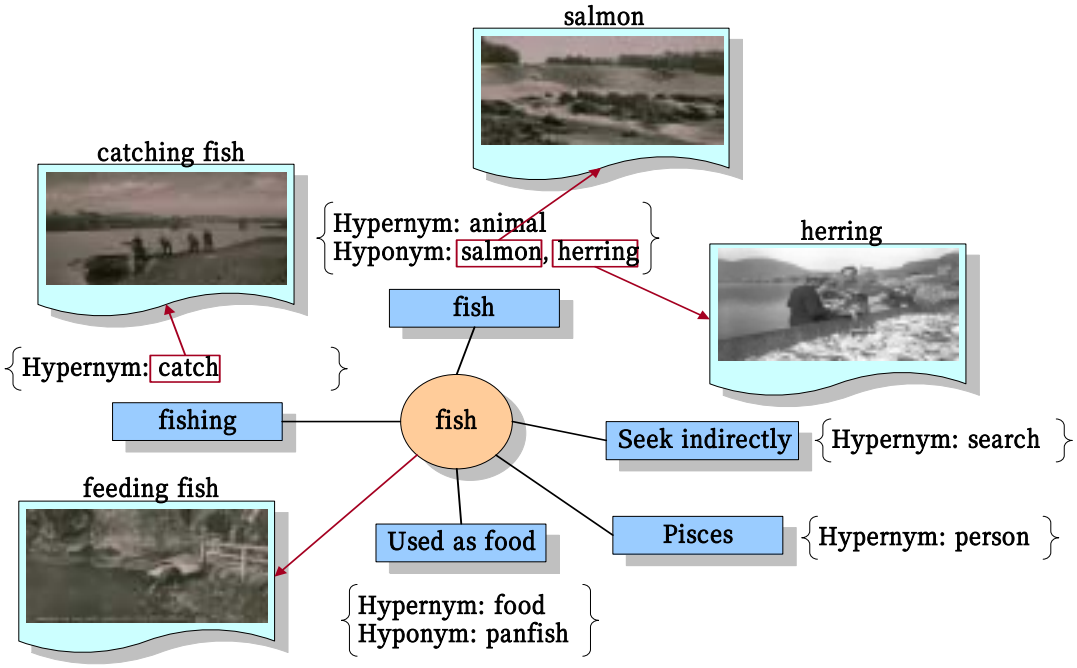

圖 14: fish 的同義詞、上位詞、下位詞

利用 WordNet 做查詢擴展會將英文查詢詞擴展成許多相關的詞，但是其中只有某些意思是適用的。例如「fish」可以當魚肉或是魚，如果魚肉和魚的意思全部取用，則會出現雜訊過多的問題。爲了避免雜訊，將原始的查詢詞視為最重要的詞，而擴展後的同義詞，上位詞，下位詞等等給定比較低的權重，權重的定義如公式 10：

$$\overrightarrow{q_m} = \alpha\vec{q} + \frac{\beta}{|D_s|}\sum_{\forall \overrightarrow{d_j} \in D_s}\overrightarrow{d_j}$$

公式 10: 查詢擴展公式

$\overrightarrow{q_m}$ 為擴展後的查詢向量，$\vec{q}$ 為原始查詢向量，D_s 是同義詞集合，d_j 是每個同義詞的向量。也就是當某個詞有 N 個同義詞，則原始詞的比重為 1，每個同義詞的比重皆為 1/N，上位詞和下位詞也是同樣的方法。如圖 15，「fish」有 4 個同義詞，每個權重都是 1/4，23 個上位詞，每個上位詞的權重為 1/23，56 個下位詞，每個下位詞的權重為 1/56。圖 15 也可知檢索問句會被翻譯成許多模糊的英文查詢詞，所以會有翻譯歧義性的問題。

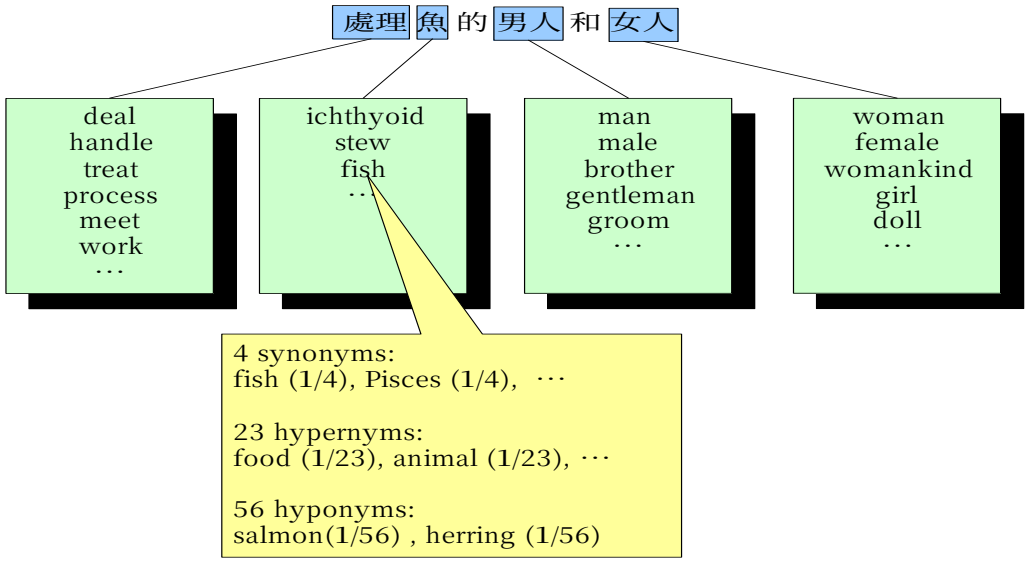

圖 15: 翻譯歧義性問題例子

第三節 解析翻譯歧義性

本節介紹本論文最主要處理的問題，解析翻譯歧義性。3.3.1 介紹知識本體；3.3.2 介紹本論文的知識本體建構流程；3.3.3 說明本論文提出的知識本體鏈方法解析翻譯歧義性。

3.3.1 知識本體(Ontology)簡介

知識本體，亦稱本體論、知識分類等，定義為概念化的明確規範說明(specification of a conceptualization) [Gruber93]。可視為一種分類法，是一個正式的(Formal)且明確的(Explicit)規格，旨在說明可以共享的概念(Shared Conceptualisation)；知識本體是「特定領域」中的概念描述，包含此特定領域的重要基本概念與彼此間的關係。在此定義中，特別強調特定領域，意謂著知識本體所要處理的資訊並不是涵蓋所有領域知識，而是集中在某一個特定的知識領域上來做分析。不僅定義出此特定領域中的重要概念，亦可呈現出概念間的關係，包括垂直的階層關係、水平的對等關係及群組的相依關係。藉由知識本體的建立，領域知識(Domain Knowledge)中可能的重要觀念和概念間的關係得以被清楚的定義。

知識本體在資訊檢索系統中也有許多應用，資訊檢索系統常使用語意網路（Semantic Network）來表示概念之間的關係。一個文件包含了許多概念，而這些概念之間又有關係。例如一篇提到「漁工」的文件可能包含了「漁船」、「漁夫」等等不同的概念。而「漁工」、「漁船」、「漁夫」這些概念都有關聯，所以可以建構類似WordNet的關係，「漁工」的廣義(General)概念是「漁業」，而「漁業」的狹義(Specific)概念是「漁工」、「漁船」、「漁夫」等等。因此可以定義廣義和狹義兩種關係。當檢索系統找到「漁工」相關文件時，可以透過廣義關係來找出「漁船」、「漁夫」是相關的概念。

圖 16就是一個知識本體的例子。

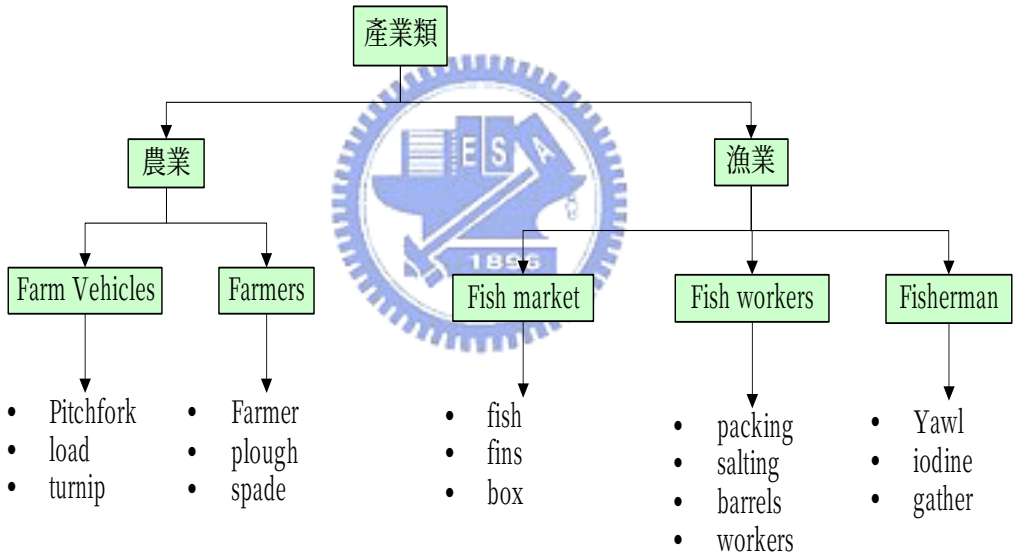

圖 16: 知識本體的例子

3.3.2 知識本體的建構

ImageCLEF2004[ImageCLEF04]的資料集如圖 17，每個文件都會屬於一個以上的分類。總共可以整理出 946 個分類，而且每個類別之間都有關係，例如漁工(Fish Worker)和漁業有關連，漁夫(Fisherman)和漁業也有關連。當使用者檢索漁工的文件，使用者可能對漁業相關的文件有興趣，所以可以透過知識本

體的搜尋來檢索出漁夫相關文件。而這些分類可以被整理成有關係的階層結構，也就是知識本體。如圖 18 所示，將有關係的分類結合，由下而上(Bottom-Up)建立知識本體。

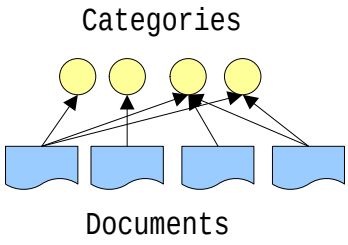

圖 17: ImageCLEF2004 資料集的結構 [ImageCLEF 04]

知識本體的樹葉節點(Leaf Node)是 CLEF2004 資料集給定的文件分類，而內部節點(Internal Node)則為專家建立的知識本體。每個樹葉節點可以用統計的方式粹取出最重要的關鍵詞。農夫(Farmers)和農用車(Farm Vehicles)可以形成農業的類別，而魚市場(Fish Market)、漁工 (Fish Workers)以及漁人(Fishermen)可以形成漁業的類別，而農業和漁業又可以形成產業的類別。農夫類別下面最重要的關鍵字是農夫和耕種(Plough)，漁工類別最重要的關鍵字是包裝(Packing)，醃(Salting)等等。

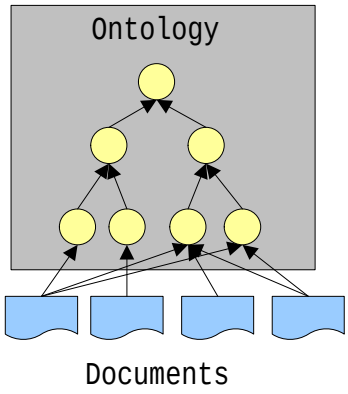

圖 18: 本論文的知識本體結構

每個類別的關鍵字可以用統計的方法(公式 11)自動粹取出來。先定義關鍵

字對類別的權重 W_{ij} 為詞鍵頻率(Term Frequency)和逆向分類頻率(Inverse Category Frequency)的乘積。

$$W_{ij} = Tf_{ij} \times Icf_i$$
$$Tf_{ij} = \text{Number of Occurrence of term i in concept j}$$
$$Icf_i = \frac{n}{cf_i}$$
$$cf_i = \text{Number of concepts that contain term i}$$

公式 11: 統計知識本體的關鍵字公式

下表是關鍵詞粹取的例子，使用上述方法，對知識節點大學(Universities)可以粹取出關鍵字大學(Universities)、成績(Scores) 、劍橋大學(Cambridge)、牛津大學(Oxford)等等關鍵字，對於大學圖書館(University Libraries)可以粹取出圖書館(Libraries)等字。

Ontology Node	Keyword	Relevance
Universities	universities	83.03
	scores	48.41
	cambridge	46.88
	oxford	26.83
University Libraries	libraries	15.76
	fergusson	8.42
	arts	8.12
	comfortable	7.26
	reading	6.67
	sofas	6.59
	galleries	5.80

表 7: 知識本體的關鍵字粹取結果。

3.3.3 建立知識本體鏈

當一個檢索問句包含許多語意時，檢索的結果會包含各種語意的結果，本論文提出了知識本體鏈(Ontological Chain)的方法來找出最適當的語意。由於每

個分類之間會有關聯性，只用單純的文件分類方式將檢索問句分類到知識本體節點下，相關的資訊可能會流失。本論文使用知識本體鏈來表示知識本體節點之間的關係，找出檢索問句中最主要的概念，以及解析翻譯歧義性。建立知識本體鏈主要有幾個步驟：

1. 找出和檢索問句最相關的 12 個知識本體節點。
2. 利用兩個節點在知識本體的距離算出 12 節點中兩兩之間的關聯性。可以獲得一個語意網路。
3. 從語意網路中找出所有連通成份(Connected Component)，取出權重總合最大的一個連通成份作為知識本體鏈。
4. 對於知識本體鏈的節點，擴展他的兄弟節點(Sibling Node)。
5. 從知識本體鏈的節點中計算英文查詢詞的交互資訊，交互資訊大於某個門檻值則取作為翻譯。

以下詳述這五個步驟：

1.首先要找出和檢索問句最相關的 12 個知識本體節點，本論文提出了一個相似度公式，如公式 12，對於檢索問句 Q 和一個知識本體樹葉節點 L_i 相關程度可以定義為

$$Sim(L_i, Q) = \sqrt{\frac{\sum_{j=1}^{N} t_{ij}^{\ 2}}{N}}$$

公式 12: 檢索問句和知識本體節點的相似度定義

其中 t_{ij} 為中文查詢詞的英文翻譯出現次數，N 為不同的中文查詢詞個數。檢索問句 Q 對於知識本體的所有樹葉節點皆計算相似度，取出前 12 個最相關的節點。

2. 建立語意網路

任意兩個知識本體節點的相關程度可以用兩個點在知識本體內的距離來決定。兩個知識節點在語意空間的距離定義為 K/D，其中 K 是常數，D 是兩個點在知識本體中的距離。如圖 19，「herring」和「fish processing」在知識本體內的路徑長度為 3，所以在語意空間的距離為 K/3。 上個步驟產生的 12 個知識本體節點兩兩之間計算相關程度，將相關的節點作為語意網路的點，節點間的距離作為語意網路的邊，可以得到一個語意網路如圖 20。

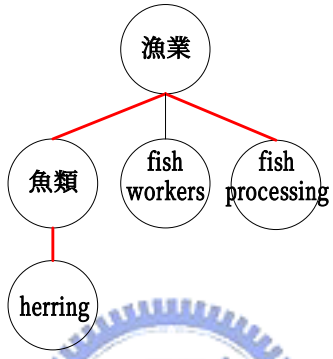

圖 19: 語意空間的距離例子

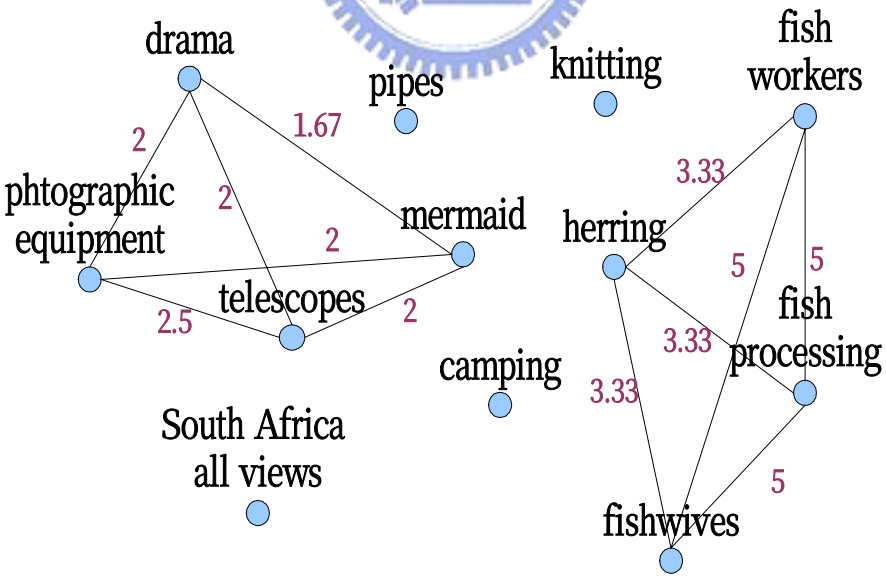

圖 20: 語意網路的例子

3. 找出連通成分：

語意網路的結構是個圖(Graph)，可以從中找出連通成分(Connected Component)作為檢索問句的最主要概念。利用 Union-Find [Trajan75]演算法如表 8，先將圖中的每個頂點(Vertex)都視為一個集合，一個一個邊加入，對於每個邊，取出兩端頂點所在的兩個集合，兩個頂點有邊相連代表兩個頂點有關聯，所以將兩個集合取聯集變成新的集合，舊的兩個集合被刪除。

```
for (each vertex v in V)
    Makeset(v): put v in its own set
  for (each edge (u,v) in E)
    if (find(u) != find(v))
        union(u,v)
```

表 8: 找出連通成分的演算法

舉個例子來說，如圖 21，總共有五個頂點 A、B、C、D、E，四個邊(A,C) 、(A,D)、(C,D) 、(B,E)。表 9 是詳細的過程，首先 A、B、C、D、E 是五個獨立的集合，加入(A,C)之後，A、C 所在的集合作聯集，也就是剩下：{A,C}、{B}、{D}、{E}。加入(A,D)以後 A、D 所在的集合作聯集，剩下{A,C,D}、{B}、{E}。接著加入{B,E}，B、E 所在的集合作聯集，得到結果{A,C,D}、{B,E}。所以上圖得到了兩個連通成分，其中{A,C,D}的權重總合為 15.0，{B,E}的總合為 2.5，所以取{A,C,D}作為知識本體鏈。

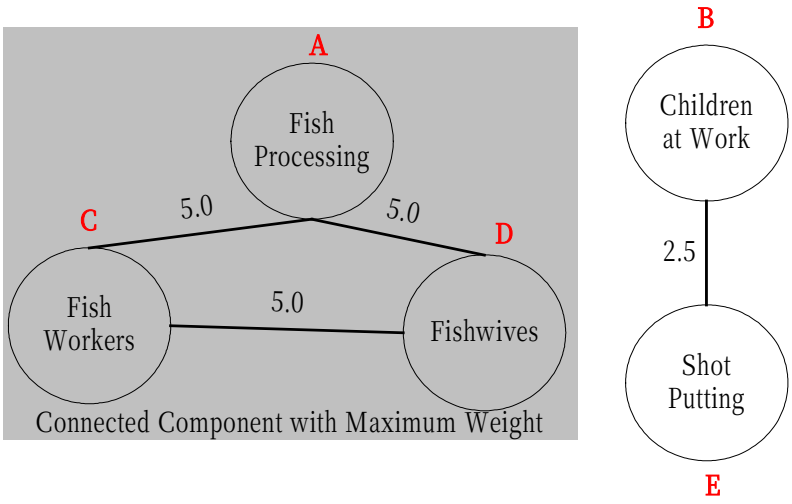

圖 21: 求語意網路中連通成分的例子

Make-set	{A}, {B}, {C}, {D}, {E}
Edge(A,C)	find(A)={A}, find(C)={C} union({A},{C})
Result	{A,C},{B},{D},{E}
Edge(A,D)	find(A)={A,C}, find(D)={D} union({A,C},{D})
Result	{A,C,D},{B},{E}
Edge(B,E)	find(B)={B}, find(E)={E} union({B},{E})
Result	{A,C,D}, {B,E}

表 9: 求連通成分的過程

6. 對於知識本體鏈的節點，擴展他的兄弟節點(Sibling Node)。也就是對於知識本體鏈中的每個節點，尋找他在知識本體中的兄弟節點，然後加入原本的知識本體鏈。
7. 從知識本體鏈的節點中計算英文查詢詞的交互資訊。交互資訊大於某個門檻值則取作為翻譯。

第四節 單語言資訊檢索系統

本節介紹本論文的單語言資訊檢索系統，總共有四個步驟，3.4.1 說明取動詞和名詞原形；3.4.2 建立向量空間模型；3.4.3 定義向量的相似程度；3.4.4 定義使用者相關度回饋。

3.4.1 取名詞及動詞原形

英文的名詞有複數形，動詞有現在進行式、過去式、過去完成式等等時式變化，字的外觀不同但意思相同。例如 Man 和 Men 不能視為兩個不同的意思，Get 和 Got 也不該視為不同的意思。將名詞和動詞轉換為原形(Stemming)再做索引可以解決這種問題。

本論文混合 Porter's Stemming Algorithm [Porter80]，以及字典查詢兩個方法。Porter 的方法使用語言學家的經驗法則取原形，速度快但是會有錯誤，例

如去掉 ity 的規則會讓 City 變成 C，University 變成 Univers，造成錯誤。為了避免錯誤，本論文先用 Porter 的方法取原形，再查詢字典確定原形是合法存在的英文字，如果是才取這個結果，若不存在就取這個字的原始形式。

3.4.2 建立向量空間模型

本論文的單語言資訊檢索系統採用向量空間模型(Vector Space Model) [Salton 83]來表示檢索問句與文件集。該模型將文件和檢索問句表示為向量空間的向量，兩個向量在空間之中會有夾角，夾角越大表示兩個向量相似程度越小，反之則越相似。檢索問句可以用向量空間的一個向量表示，計算檢索問句向量和所有文件向量的夾角就可算出檢索問句和所以文件的相似程度，再依照此相似度排序作為檢索結果。

1）文件向量表示法

本論文採用三種特徵值來表示 ImageCLEF2004 文件集中的文件向量，包括關鍵字(Term)、文件所屬分類(Category)以及時間特徵(Temporal Feature)。如公式 16 所示，$\overrightarrow{d_j}$ 是文件集中，第 j 個文件的向量表示法，向量前 n 個元素代表該文件包含的關鍵字；第 $n+1$ 到第 $n+m$ 個元素代表文件所在的分類；第 $n+m+1$ 到第 $n+m+k$ 個元素代表文件所在的年代。每個關鍵字在文件向量 $\overrightarrow{d_j}$ 的權重使用 *TF*IDF*[Salton83]權重定義如公式 13 所示，$tf_{i,j}$ 是文件 j 中，詞鍵 i 出現的次數；N 是文件集中的文件總數；n_i 是包含詞鍵 i 的文件總數。

$$W_{i,j} = \frac{tf_{i,j}}{\max tf_{i,j}} \times \log \frac{N}{n_i}$$

公式 13：詞鍵對文件向量的權重公式[Salton83]

文件所屬類別在文件向量$\overrightarrow{d_j}$的權重使用布林權重定義如公式 14，當文件 j 屬於類別 i 時，$W_{i,j}$ 為 1，反之則為 0。文件出版年代在文件向量$\overrightarrow{d_j}$的權重定義亦使用布林權重，如公式 15 所示，若文件 j 的出版年代為 i，則$W_{i,j}$為 1，反之則為 0。

$$\begin{cases} W_{i,j} = 1, \text{if document j belongs to category i} \\ W_{i,j} = 0, \text{if document j doesn't belong to category i} \end{cases}$$

公式 14: 類別對文件向量的權重公式

$$\begin{cases} W_{i,j} = 1, \text{if document j was published in year i} \\ W_{i,j} = 0, \text{if document j wasn't published in year i} \end{cases}$$

公式 15: 出版年代對文件向量的權重公式

$$\overrightarrow{d_j} = < \underbrace{w_{1,j}, w_{2,j}, \ldots, w_{n,j}}_{\text{Terms}}, \underbrace{w_{c1,j}, w_{c2,j}, \ldots\ldots, w_{cm,j}}_{\text{Categories}}, \underbrace{w_{t1,j}, w_{t2,j}, \ldots\ldots, w_{tk,j}}_{\text{Temporal Feature}} >$$

公式 16: 單語言檢索系統中文件的向量表示法

2）檢索問句向量表示法

本論文採用三種特徵值來表示檢索問句向量，包括關鍵字、文件所屬分類以及時間特徵。如公式 17 所示，$\overrightarrow{q}$ 是檢索問句的向量表示法，向量前 n 個元素代表檢索問句包含的詞鍵；第 $n+1$ 到第 $n+m$ 個元素代表檢索問句所在的分類；第 $n+m+1$ 到第 $n+m+k$ 個元素代表文件所在的年代。每個詞鍵對檢索問句的權重可以定義為詞鍵頻率和逆向文件頻率的乘積(公式 18)；分類對檢索問句

的權重使用布林權重；檢索問句的年代權重亦使用布林權重，並且定義了三種運算(Operation)：某年之前(Before)、某年之中(In)以及某年之後(After)。例如表 10 中 $\overrightarrow{d_1}$ 的出版年代是 1901 年，$\overrightarrow{d_2}$ 則是 1898 年。當使用者想找 1900 年之前的文件，則定義"1900 年之前"的運算，也就是查詢向量 $\overrightarrow{Q}$ 中，1900 之前的年代權重值為 1，1900 以後的年代權重值為 0。$\overrightarrow{Q}$ 和 $\overrightarrow{d_1}$ 的向量內積為 0，$\overrightarrow{Q}$ 和 $\overrightarrow{d_2}$ 的向量內積為 1，根據公式 19 的相似度定義可知檢索問句和文件 $\overrightarrow{d_1}$ 的相似度為 0，和文件 $\overrightarrow{d_2}$ 的相似度則大於 0。

$$\overrightarrow{q} = < \underbrace{w_{1,j}, w_{2,j}, \ldots, w_{n,j}}_{\text{Terms}}, \underbrace{w_{c1,j}, w_{c2,j}, \ldots, w_{cm,j}}_{\text{Categories}}, \underbrace{w_{t1,j}, w_{t2,j}, \ldots, w_{tm,j}}_{\text{Temporal Feature}} >$$

公式 17: 單語言檢索系統中，檢索問句的向量表示法

$$W_{i,q} = \frac{tf_{i,q}}{\max tf_{i,q}} \times \log \frac{N}{n_i}$$

公式 18: 詞鍵對檢索問句的權重公式[Salton83]

	1897	1898	1899	1900	1901	1902	1903	…
$\vec{D}_1$	0	0	0	0	1	0	0	…
$\vec{D}_2$	0	1	0	0	0	0	0	…
$\vec{Q}$	1	1	1	0	0	0	0	…
$\vec{D}_1 \bullet \vec{Q} = 0 \; ; \; \vec{D}_2 \bullet \vec{Q} = 1$								

表 10: 時間特徵的向量內積實例

3.4.3 相似度計算

計算兩個向量的夾角可以得到兩個文件的相似程度，$\cos\theta$ 定義為向量內積除以兩個向量的長度，而 $\cos\theta$ 越大代表夾角越大。檢索問句與文件之間的相

似程度定義如公式 19 所示，相似程度越高代表檢索問句和文件越相關，排序檢索問句和所有文件的相似度即為最後的檢索結果。

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{\left|\vec{d_j}\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^{\ 2}} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^{\ 2}}}$$

公式 19: 單語言檢索系統中的相似度計算公式[Salton83]

3.4.4 使用者相關度回饋

使用者相關度回饋(User Relevance Feedback) 是使用者對於檢索結果的回應，可以引導系統的檢索方向，進而提高檢索效能。例如，使用者可以根據檢索結果，指出哪些檢索出的文件跟他的檢索問句相關，而哪些又是完全不相關，將此訊息回饋給系統，使用公式 20 重新調整檢索問句，將相關文件的詞鍵增加權重；將不相關文件的詞鍵減少權重，再使用新的檢索問句作進一步的檢索。

$$\vec{q_m} = \alpha \vec{q} + \frac{\beta}{\left|D_r\right|} \sum_{\forall \vec{d_j} \in D_r} \vec{d_j} - \frac{\gamma}{\left|D_n\right|} \sum_{\forall \vec{d_j} \in D_n} \vec{d_j}$$

公式 20: 使用者相關度回饋公式[Rocchio 71]

修正過後的檢索問句 $\vec{q_m}$ 是原始查詢向量 $\vec{q}$ 加上正例向量再扣除負例向量。其中，D_R 表示相關的文件集合，D_N 表示不相關的文件集合。α，β，γ 分別為可調整之參數。

第四章　實驗結果分析與評估

本章針對本論文提出的跨語言資訊檢索系統作效能評估及討論。第一節說明本實驗採用的資料集；第二節介紹本實驗使用的檢索主題；第三節介紹本實驗效能評估的條件；第四節表列實驗結果；第五節討論本論文方法在各種檢索主題下的效能表現和優缺點。

第一節　實驗資料集

跨語言資訊檢索的評估語料包括資料集、檢索主題、和參考答案三部份。評估跨語言資訊檢索的組織有 TREC[1]（美國貿易部 NIST 主辦），CLEF［CLEF04］（歐盟所支援的數位圖書館計畫），和 NTCIR[2]（日本文部科學省下的情報資訊研究所 NII 主辦）…等。這些組織首先委由專家收集語料庫，例如新聞、生物、醫學資料…等，從文件集中挑出相關的文件或是關鍵字作為檢索主題，針對每個檢索主題從文件集中挑出所有相關的答案。參與測試的各系統必須在規定的測試集中進行實驗，並將每個檢索問句檢索出的前 100 篇相關文件送回，供大會計算系統效能。ImageCLEF[ImageCLEF04]主要以召回率(Recall)及準確率(Precision)作為主要的測量準則，本論文預定參加 CLEF 舉辦的 ImageCLEF2004 Bilingual Ad hoc Tasks 評估。

ImageCLEF2004 資料集為英國聖安德魯大學歷史照片集（The Eurovision ST Andrews Photographic Collection, ESTA），包含了該大學圖書館 28,133 張照片館藏，大部分為蘇格蘭地區的照片。該館藏預估有超過三十萬張影像，取了大約百分之十的影像作為評估效能之用。

[1] Text Retrieval Conference, avalible at http://trec.nist.gov/
[2] NII-NACSIS Test Collection for IR Systems, avalible at http://research.nii.ac.jp/ntcir/index-en.html

ImageCLEF對於每篇文件中均以標準通用標誌語言(Standard Generalized Mark-Up Language，簡稱SGML)及文件型態定義檔（Document Type Definition，簡稱DTD)加上標籤(Tag)，以便系統進行剖析(Parsing)工作。每張照片都有說明文字(caption)，包含了八個欄位:(1)標題、(2)簡短標題、(3)唯一的記錄編號、(4)影像內容的文字說明、(5)拍攝日期、(6)照片來源(大部分是人名或是公司名稱)、(7)拍攝地點(可能是城市或國家)、(8)照片的額外資訊。這28133張照片的說明包含了44,085個詞，所有詞的出現總數為1,348,474次；最長的照片說明是316字，平均照片說明的長度是48個字。所有的照片說明都用英式英文寫成，而且包含了口語詞彙(Colloquial Expressions)和歷史術語。81%照片說明包含了所有八個欄位，其他照片的說明則不完整。在大部分的狀況下，圖片說明都是符合文法的句子，一個句子大約15個字。82%的照片是黑白照片。如圖 22，資料集的年代分佈集中於1930年代。如圖 23，資料集中每篇文件的分類平均約三個，最多九個分類，最少一個分類。

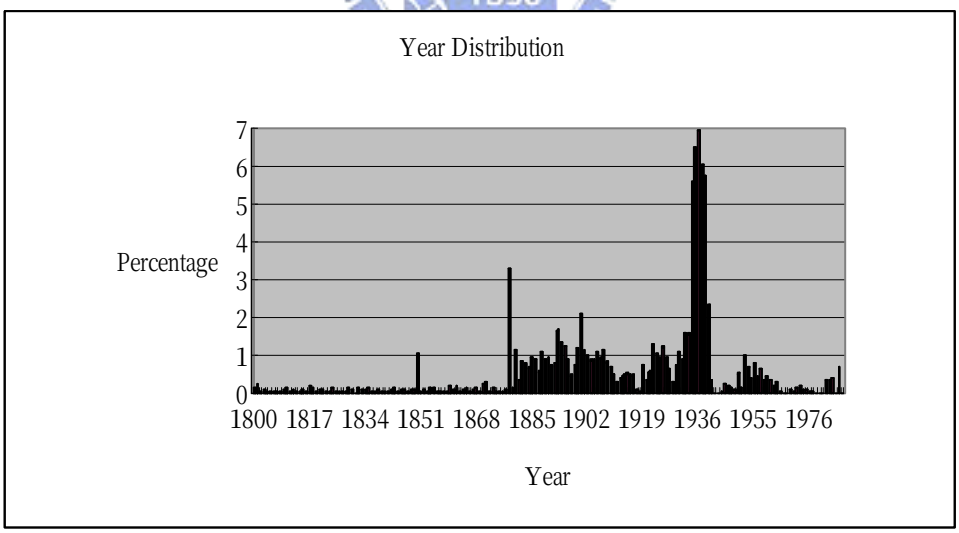

圖 22: ImageCLEF2004 文件集中每篇文件的年代分布

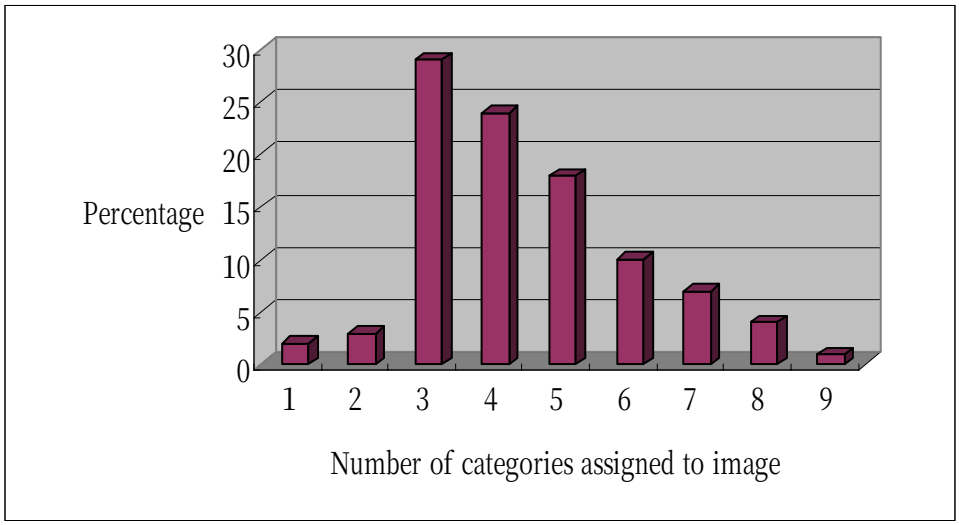

圖 23: ImageCLEF2004 文件集中每篇文件的分類數目

第二節　檢索主題

ImageCLEF 挑選了 50 個英文檢索問句，或稱檢索主題，並且會翻譯成六種不同的語言包括荷蘭文、西班牙文、德文、法文、義大利文和中文。檢索問句由專家建議產生，從圖書館的存取紀錄檔案(Access Log)分析經常被搜尋的主題，依據這些常被搜尋的主題產生檢索問句來測試跨語言檢索系統，包括比較廣義的概念(Broad Concepts) 以及狹義(Narrow Concept)的概念、專有名詞(Proper Names)、複合字(Compound Words)、簡寫(Abbreviations)、文法上的變形(Morphological Variants)以及成語(Idioms)等等不同的形式。ImageCLEF2004 提供的檢索問句如表 11：

編號	中文檢索問句	英文檢索問句
1	Thomas Rodger 拍攝的牧師肖像	Portrait pictures of church ministers by Thomas Rodger
2	1908 年四月拍攝的羅馬照片	Photos of Rome taken in April 1908
3	John Fairweather 拍攝的安德魯教堂	Views of St. Andrews cathedral by John Fairweather
4	George Middlemass Cowie 拍攝的穿軍服的男人	Men in military uniform, George Middlemass Cowie
5	北愛爾蘭的漁船	Fishing vessels in Northern Ireland
6	加拿大英屬哥倫比亞區的風景	Views of scenery in British Columbia,

		Canada
7	埃及廟宇的外部景觀	Exterior views of temples in Egypt
8	劍橋學院或大學的建築	College or university buildings, Cambridge
9	英格蘭燈塔照片	Pictures of English lighthouses
10	倫敦的繁忙街景	Busy street scenes in London
11	蘇格蘭 Bute 地區風景的綜合明信片	Composite postcard views of Bute, Scotland
12	1879 年 Tay Bridge 的火車災難	Tay Bridge rail disaster, 1879
13	1939 年聖安德魯斯高爾夫球公開賽	The Open Championship golf tournament, St. Andrews 1939
14	1954 年伊麗莎白女王的母親訪問 Crail Camp	Elizabeth the Queen Mother visiting Crail Camp, 1954
15	二戰轟炸造成的損害	Bomb damage due to World War II
16	約克大教堂的照片	Pictures of York Minster
17	聖安德魯斯北街的所有景觀	All views of North Street, St. Andrews
18	1900 年以前拍攝的愛丁堡城堡的照片	Pictures of Edinburgh Castle taken before 1900
19	列隊行進或遊行中的人們	People marching or parading
20	背景有高架橋的河流	River with a viaduct in background
21	十字型戰爭紀念碑	War memorials in the shape of a cross
22	跳傳統蘇格蘭舞的照片	Pictures showing traditional Scottish dancers
23	天鵝在湖上的照片	Photos of swans on a lake
24	正在揮動球桿的高爾夫球員	Golfers swinging their clubs
25	運河上的船隻	Boats on a canal
P1	處理魚的男人和女人	Men and women processing fish
P2	教堂裡的座位	Seating inside a church
P3	大雅穆斯海灘	Great Yarmouth beach
P4	成為廢墟的英國城堡	Ruined castles in England
P5	博物館的展覽品	Museum exhibits

表 11: ImageCLEF2004 檢索主題 [ImageCLEF04]

第三節　相關程度評估

要決定一篇文件和檢索問句是否有關聯是很主觀的，每個人可能有不同的專業背景，搜尋的經驗也不同，會有不同的詮釋。最理想的狀況是每個專家都看過每篇文件，再一篇一篇決定是否相關，但是在文件數目過多的情形下，這種方式並不切實際。ImageCLEF 採用 TREC 和 CLEF、NTCIR 等等的做法，先

產生候選文件，稱為 pool，候選文件的產生是由參加者的前 n 個相關文件取交集，這個方法是假設每位參加者找到的前幾名文件都是相關的。ImageCEF 也使用了 NTCIR 的方法來補充 Pooling 法，亦即讓專家作互動式搜尋(Interactive Search and Judge, ISJ) 來保持 Pool 的品質。相關程度主要是從影像得知，用說明文字做輔助。相關程度的評估使用了四個集合(Qrels)，即嚴格/寬鬆(Strict/Relaxed)、交集/聯集(Union / Intersection)，比較嚴格的集合可以用來評估高準確率的任務，比較寬鬆的集合用來評估高召回率的任務。

之前的結果(ImageCLEF2003)顯示，翻譯檢索問句為本的多語言影像檢索對單語言來說可以達到高準確率。 對中文來說，專有名詞的翻譯是有幫助的，對其他索引典為本的查詢自動擴展也可以提高準確率。這個 Ad Hoc Task 可以不用使用影像特徵。主要評估幾種不同的方法對於跨語言檢索系統效能的影響：1. 不同的檢索問句翻譯方法(例如雙語字典查詢之於機器翻譯)；2. 自動查詢擴展(全域之於區域)；3. 不同的檢索模型(Retrieval Model)；4. 不同的索引方式(Indexing Methods)；5. 手動和自動的使用者相關度回饋。

ImageCLEF2004 使用幾個標準來評估效能：(1)尚未內插的平均準確率；(2)前 100 個影像沒有任何相關(Failed Topics)；(3)100 張影像包含相關的比率(Precision at 100)；(4)相關的圖片在前 100 被找到的比率(a Normalized Precision at 100，不會受到答案集的長度影響)。

準確率 (Precision)、召回率 (Recall)的定義如表 12，檢索的結果和相關程度有 A、B、C、D 四種關係，A 代表檢索出來而且相關，B 為檢索出來但是不相關，C 為沒有檢索出來但是相關，D 為沒有檢索出來而且不相關。準確率定義為檢索出來的文件中相關文件的比例，也就是 $\frac{|A|}{|A| \cup |B|}$ ，召回率定義為相關的文件被檢索出來的比例，也就是 $\frac{|A|}{|A| \cup |C|}$ 。

	相關	不相關
被檢索	A	B
未被檢索	C	D

表 12: 檢索出的文件和相關程度的四種可能關係

ImageCLEF 採用的評估方式為平均準確率，也就是每當一個相關的文件被檢索出來時，計算當時的準確率，再取平均。其中相關的文件但是沒有被檢索出來的準確率是 0。以表 13 為例，正確答案總共有九篇文件，第一欄是檢索出來的並且是相關文件的排名，檢索出來的文件中排名第 3、6、7、8、16、20、22、24，的八篇是檢索出來而且相關的，也就是還有一篇相關但是沒有被檢索出來。計算每一篇相關文件被檢索出來時的準確率，例如第三篇文件是相關而且被檢索，前兩名是被檢索出來但是不相關，所以該點的準確率是 1/3 也就是 33%；而總共九篇相關文件到第三篇文件時被檢索出一篇，所以該點的召回率是 1/9=11%。依此類推，最後算十一點準確率的平均值，但是有一篇相關文件沒被檢索出來，所以取平均時必須除以 9 而非 8。

平均準確率只是針對單一檢索問句計算，對於整個系統的效能評估必須考慮多個檢索問句，因此 ImageCLEF 採用 Mean Average Precision(MAP)的算法，對於每一個查詢主題都計算平均準確率(Average Precision)，最後再將所有查詢主題的平均準確率加以平均，也就是 MAP。

排名	準確率	召回率
3	0.33	0.11
6	0.33	0.22
7	0.43	0.33
8	0.50	0.44
16	0.31	0.56
20	0.30	0.67
22	0.32	0.78
24	0.33	0.89
平均準確率	0.36	

表 13: 平均準確率的計算例子。

第四節　實驗結果

本論文使用了表 11 的 30 個檢索主題作為檢索問句，實驗了三個模型：英文單語言檢索(Mono-Lingual IR)、字典為本的跨語言檢索(Dictionary-based CLIR)以及知識本體鏈為本的跨語言檢索(Ontological Chain-based CLIR)。其中單語言檢索由專家將中文檢索問句翻譯為英文，因此這個模型不會有翻譯歧義性的問題，以此作為基底評估標準(Baseline)；字典為本的模型只用雙語字典翻譯，並且取所有的可能翻譯，因此有翻譯歧義性的問題；知識本體鏈模型使用本論文提出的知識本體鏈方法來解決翻譯歧義性問題。本節評估 1) 檢索結果中前一百篇的準確率和召回率; 2) 前一百篇的 11-Point Precision/Recall; 3) 使用者回饋對平均準確率的影響。

針對檢索結果中前一百篇的效能評估方面，由表 14 可以看出純粹使用雙語字典翻譯檢索問句的跨語言資訊檢索系統準確率只有 5%，而加入知識本體鏈之後可以提升到 10%; 召回率也由 63%提升到 83%。由於準確率的計算方式是檢索出的相關文件除以所有檢索出的文件總數，所有檢索出的文件總數固定為 100，但是相關的文件可能不到 100，如表 13 的例子中，相關文件只有九篇，因此上例準確率最高只能到 9%，無法到 100%。

	單語言檢索	字典為本	知識本體鏈
準確率	12.49%	5.3%	10.62
召回率	89.23%	63.98%	83.37%

表 14: 前一百篇檢索結果的準確率和召回率

在前一百篇的 11 點準確率/召回率方面，由於相關文件的個數沒有正規化，所以前一百篇的平均準確率很低，意義不太大，因此使用本章第三節介紹的平均準確率(Average Precision)以及 MAP (Mean Average Precision) 來避免相關文件個數的影響。表 15 是三種模型的 11 點平均準確率；表 16 是三種模型的 MAP 值實驗結果，從表 15 可以看出使用知識本體鏈可以從 49%提升到 55%，並且達到單語言資訊檢索效能的 92%。

檢索模型 / 召回率	字典為本跨語言檢索	單語言檢索	知識本體鏈跨語言檢索
0	0.3489	0.5877	0.3305
0.1	0.3134	0.5564	0.3292
0.2	0.3078	0.4863	0.3340
0.3	0.2457	0.4243	0.3563
0.4	0.2570	0.3993	0.3531
0.5	0.2394	0.3571	0.3277
0.6	0.2405	0.3659	0.3132
0.7	0.2321	0.3526	0.3084
0.8	0.2330	0.3122	0.3030
0.9	0.2254	0.2926	0.3038
1.0	0.2163	0.2906	0.3044
平均準確率	0.2600	0.4023	0.3240

表 15: 三種模型的平均準確率比較

	單語言檢索	字典為本	知識本體鏈
MAP	60.63%	49.18%	55.81%

表 16: 前一百篇檢索結果的 MAP 值

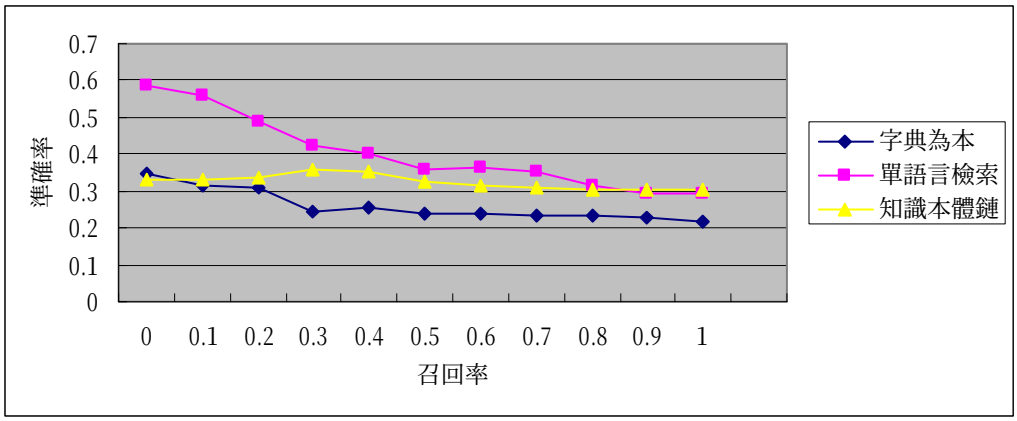

圖 24 是系統的準確率/召回率相對圖，可以看出效能最佳的是單語言資訊檢索，而知識本體鏈的方法效能較雙語字典比對為佳。

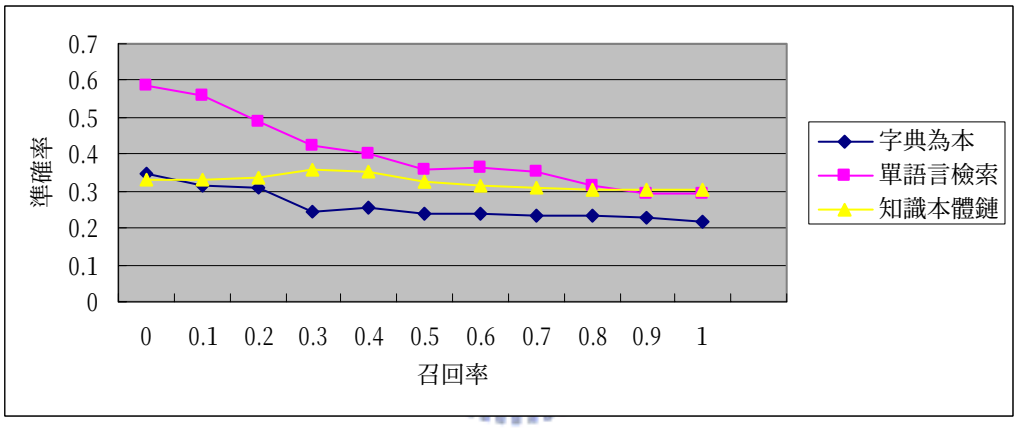

圖 24: 11 點準確率/召回率相對圖

在探討使用者回饋對平均準確率的影響方面，此功能需要兩種相關度回饋，包括正例(也就是被檢索出的相關文件)和反例(也就是被檢索出的不相關文件)。本論文採用自動的方式評估使用者相關度回饋功能：對於每個檢索問句的每篇檢索結果交由評估的程式標示相關以及不相關，從相關的文件中隨機取出 4 篇文件作為正例；不相關的文件中隨機取出 2 篇文件作為反例，以模擬使用者挑選正例和反例的流程。圖 25 是自動評估的結果，可知使用者回饋次數越多，可以進一步提升準確率。

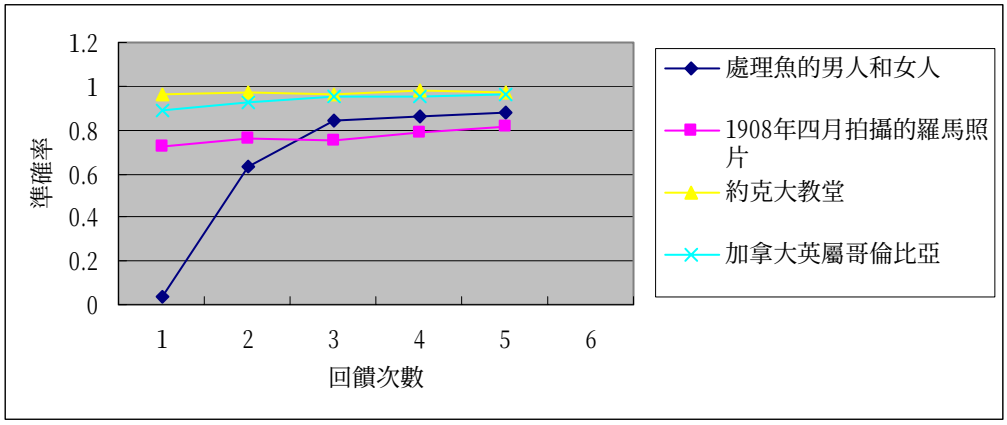

圖 25: 使用者相關度回饋次數與準確率關係圖

第五節　討論

第三節的評估是對 30 個檢索問句的檢索結果取平均，但是對於每一個單獨的檢索問句，效能的排名並非固定的。本節討論幾種狀況，在某些狀況下知識本體鏈的方法會表現的特別好，甚至比單語言資訊檢索高；而某些狀況下知識本體鏈的方法會比雙語字典比對差，但不會差太多。

4.4.1 檢索問句包含廣泛概念

當檢索問句是比較廣泛的概念時，例如「處理魚的男人和女人」，英文正確翻譯是「man and woman processing fish」。由於向量空間模型中每個詞鍵都是獨立的，也就是文件中的「男人」一詞權重很高，就可能被檢索出來，而且權重過高時，它的相似度可能會比包含「男人」、「女人」、「處理」、「魚」四個詞的文件還要高。而且當檢索問句比較廣泛時，通常有許多個翻譯，例如「處理」在本論文的字典中有 27 個翻譯；「男人」有 12 個翻譯；「女人」有 17 個翻譯；「魚」而只有 3 個翻譯。針對字典的翻譯個數來看，上述句子有

$27\times12\times17\times3=16524$ 種不同的語意，可以看出來這個檢索問句中，有三個詞是廣泛的概念，並且有嚴重的翻譯歧義性問題。

表 17 中可以看到以字典為本的 CLIR 效能準確率是 0%，這個結果並不令人意外，因為該模型只有查詢雙語字典，取所有可能的翻譯，有 16524 種語意組合，會產生嚴重的翻譯歧義性問題。因此這個查詢主題對於這個方法來說是一個失敗主題(Failed Topic)，也就是前 100 篇檢索出來的文件中沒有任何一篇相關。

	單語言檢索	字典為本	知識本體鏈
平均準確率	5.14%	0%	35.74%

表 17: 處理魚的男人和女人檢索結果的平均準確率

由表 17 中的單語言檢索系統在本例中表現也是不佳，這個方法使用「man and woman processing fish」作為檢索問句，並不需要字典比對，不會有翻譯歧義性問題。經過統計，語料庫中 28133 篇文章中包含「男人」的文件有 1720 篇；「女人」800 篇；「處理」有 7 篇；「魚」有 130 篇。可以看出「男人」、「女人」都是很一般性的概念，因此某篇文章要是有「男人」和「女人」並且權重高，就可能會被檢索出來，而真正和「處理魚」相關的文件則可能排名到後面。圖 26 是這個檢索問句的準確率/召回率相對圖。

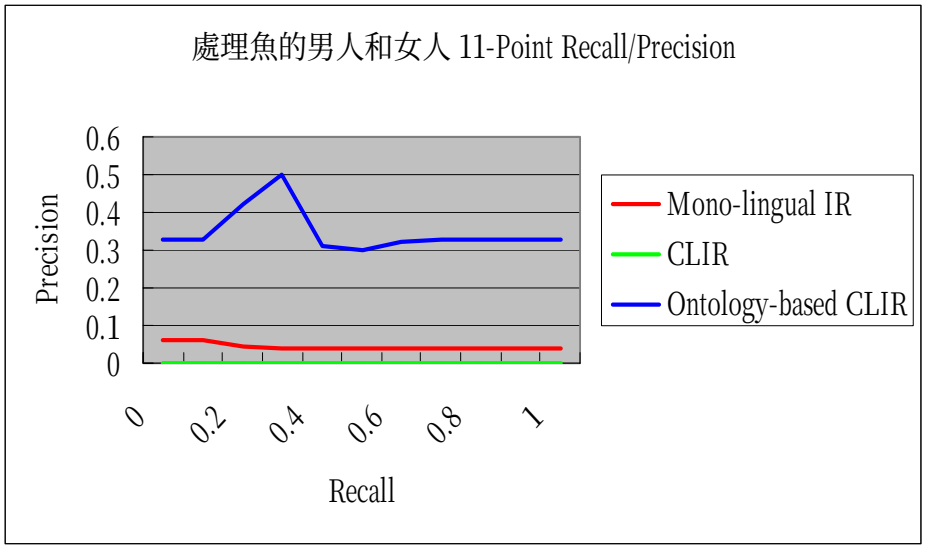

圖 26: 處理魚的男人和女人準確率/召回率相對圖

4.4.2 檢索問句包含特定概念

另一類的檢索問句則是語意非常明確，例如「1908 年四月羅馬拍攝的照片」，英文檢索問句則為「Photos of Rome taken in April 1908」。和上個例子不同的是，同時出現「Rome」、「April」、「1908」的文章並不多，因此只要三個詞鍵都出現的文章容易被排名到前面，事實上三個詞鍵都出現的文章就是相關的文章。

	單語言檢索	字典為本	知識本體鏈
平均準確率	67.25%	59.34%	54.34%

表 18 是這個檢索問句的平均準確率，圖 27 是這個檢索問句的準確率/召回率相對圖。可以看出由於語意相當明確，單語言資訊檢索的準確率很高。而字典中「1908」不需要翻譯，「四月」只有一種翻譯，「羅馬」只有三種翻譯，因此這個檢索問句的翻譯歧義性並不高，只有三種不同的語意組合。因此只要使用關鍵字比對就可以得到不錯的效果，單語言資訊檢索和字典翻譯的效能都不錯。而知識本體鏈會擴展其他相關的節點，在此例中，檢索問句和知識本體中的義大利景色(Italy All View)最相關，會擴展其他和義大利相關的景色，準確率就沒其他兩個方法高。但是知識本體鏈在這個例子也有單語言檢索的 80%效能，平均準確率也有 54%，雖然沒有其他兩者高，但是也只比字典比對少 5%。

	單語言檢索	字典為本	知識本體鏈
平均準確率	67.25%	59.34%	54.34%

表 18: 1908 年四月羅馬拍攝的照片檢索平均準確率

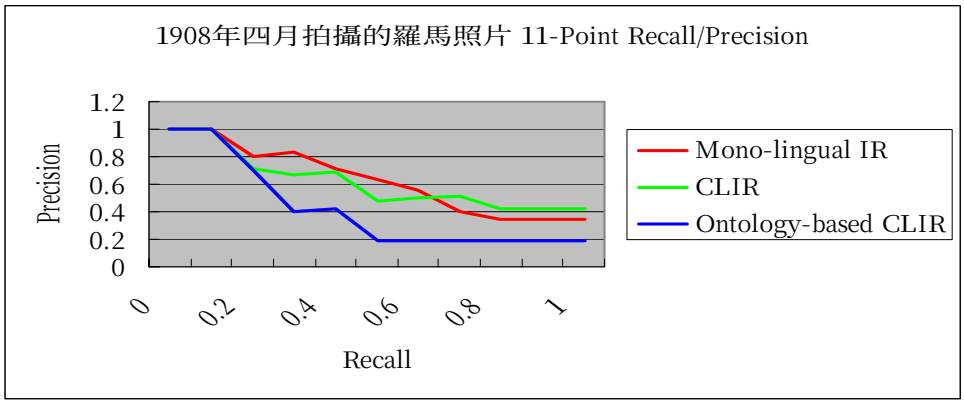

圖 27 1908 年四月羅馬拍攝的照片準確率/召回率相對圖

總結上面兩個比較極端的例子，可以發現知識本體鏈在處理高度翻譯歧義性時，可以有很好的表現，以 16524 種翻譯組合為例來說，效能比單語言檢索高 30%，比字典比對高 35%。而在歧義性較低的檢索問句，例如只有 3 種翻譯組合，則效能會稍微低一點，準確率比字典比對低 5%。因此本論文提出的知識本體鏈在翻譯歧義性高的狀況效能很高，在歧義性低的狀況下效能稍差。

第五章　結論與未來研究方向

本章總結本論文並說明未來的研究方向。第一節討論本論文提出的知識本體鏈應用在跨語言資訊檢索系統的效益、優點和限制；第二節說明未來可能的研究方向。

第一節　結論

本論文提出了一個跨語言資訊檢索系統的架構，並且針對 ImageCLEF2004 的資料集實作跨語言資訊檢索系統，使用者可以使用中文檢索問句檢索資料集中的照片以及英文照片說明。對於跨語言資訊檢索系統中會出現歧義性的問題，使用中英字典為本的斷詞方法解決斷詞歧義性，並且提出一個利用知識本體鏈解析翻譯歧義性的方法。實驗得知，本論文的方法在高度翻譯歧義性的狀況下可以達到很好的效能，並且可以比單語言檢索系統高出 30%，比字典比對高出 35%；在低度翻譯歧義的環境下表現稍差，比字典比對低 5%左右；所有狀況的平均可以達到單語言檢索系統的 81%效能。將跨語言檢索系統運用於英國聖安德魯大學圖書館的照片館藏，顯示跨語言資訊檢索系統可以應用到數位典藏領域，幫助使用者跨越語言的藩籬。

第二節　未來研究方向

跨語言資訊檢索系統中主要有三個模組可能隱含歧義性問題：中文斷詞模組，檢索問句翻譯模組以及文件索引模組。中文斷詞模組的歧義性問題使用中英字典為本的斷詞方法解決；翻譯模組的歧義性問題使用知識本體鏈的方法解決，但是對於文件索引模組的歧義性問題由於在 ImageCLEF2004 資料集和查詢主題集影響不大，本論文沒有針對這方面做改良。但是當查詢過短時，解析詞

鍵的歧義性就相當重要。例如 bank 有銀行和河岸的意思，當檢索問句是「銀行領錢」時，翻譯成英文後可以判斷「bank」和「money」的關係來解析語意，這在本論文的跨語言檢索系統中可以達成。但是當檢索問句只有「銀行」兩字，翻譯過後「bank」沒有任何前後文當語境，這時就需要靠語意索引，這是未來值得研究的主題。

目前本論文使用的單語言資訊檢索系統中，以使用詞鍵對文件的關係作為向量空間的元素，也就是詞鍵是用來作為索引的單位。但是要加入語意索引必須把每個字的意思視為不同，當「machine」有五個意思時，必須將五個意思視為不同的元素作索引，換句話說，「machine」的第一個意思和「machine」第二個意思雖然是相同的字但是由於意思不同，視為不同的字，因此會有兩個不同的元素，可以解決索引時的詞鍵歧義性。對每個字的所有語意作索引是未來值得研究的方向。

參考文獻

[Ballesteros98] L. Ballesteros and W.B. Croft, “Resolving ambiguity for cross language retrieval,” *Proc. 21st annual international ACM SIGIR conference on Research and development in information retrieval,* pp.64-71, 1998.

[Barzilay97] R. Barzilay and M. Elhadad, “Using Lexical Chains for Text Summarization,” *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, 1997.

[Carbonell97] J. Carbonell, Y. Yang, R. Frederking, R.D. Brown, Y. Geng, and D. Lee, "Translingual Information Retrieval: A Comparative Evaluation," *Proc. Fifteenth International Joint Conference on Artificial Intelligence* Vol 1, pp. 708-715, 1997.

[Chen02] H.H. Chen, C.C. Lin and W.C. Lin, “Building a Chinese-English wordnet for translingual applications,” *ACM Transactions on Asian Language Information Processing* vol. 1, Issue 2, pp.103-122, 2002.

[CLEF04] Cross Language Evaluation Forum, avalible at http://clef.iei.pi.cnr.it:2002/2004.html

[Frakes92] W.B. Frakes, R. Baeza-Yates, *Information Retrieval, Data Structures & Algorithms*. Prentice Hall, 1992.

[Fung98] P. Fung, L.Y. Yee, ”An IR Approach for Translating New Words from Nonparallel, Comparable Texts,” *Proc. of the 36th Annual Conference of the Association for Computational Linguistics*, pp. 414-420, 1998.

[Gruber93] T. R. Gruber, “A translation approach to portable ontologies,” *Knowledge Acquisition*, pp. 199-220, 1993

[ImageCLEF04] Cross Language Evaluation Forum, avalible at http://ir.shef.ac.uk/imageclef2004/

[Kipfer01] B.A. Kipfer and R. L. Chapman, *Roget's International Thesaurus*. , HarperResource, 2001.

[Larkey03] L.S. Larkey and M.E. Connell, “Structured Queries, Language Modeling, and Relevance Modeling in Cross-Language Information Retrieval,” *Information Processing and Management* Special Issue on Cross Language Information Retrieval, 2003.

[Littman98] M.L. Littman, S.T. Dumais, and T.K. Landauer, “Automatic cross-language information retrieval using latent semantic indexing,” *Cross-Language Information Retrieval*, pp. 51 – 62, 1998.

[Lu02] W.H. Lu, L.F. Chien and H.L. Lee, “Translation of web queries using anchor text mining,” *ACM Transactions on Asian Language Information Processing* ,Vol 1, Issue 2, pp.159-172, 2002

[Miller95] G. Miller, "Wordnet: A Lexical Database for English,” *Proc. of Communications of CACM*, 1995.

[Miller99] D.R.H. Miller, T. Leek, R.M. Schwartz, ”A hidden Markov model information retrieval system,” *Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information*, pp. 214-221, 1999.

[Nie99] J.Y. Nie, M. Simard, P. Isabelle and R. Durand , ”Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web,” *Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information*, pp. 74-81, 1999.

[Porter80] M. F. Porter, “An algorithm for suffix stripping,” *Program*, Vol. 14, No. 3, pp. 130-137, 1980

[Rocchio71] J. Rocchio, “Relevance Feedback in Information Retrieval,” Prentice-Hall, Inc., 1971.

[Salton83] G. Salton and M. J. McGill, “*Introduction to Modern Information Retrieval* , “ McGraw-Hill, 1983

[Savoy03] J. Savoy ,”Cross-language information retrieval: experiments based on CLEF 2000 corpora,” *Information Processing & Management* ,Vol. 39, Issue 1, pp. 75-115, 2003.

[Trajan75] R.E. Tarjan, “Efficiency of a Good But Not Linear Set Union Algorithm,” Journal of the ACM, Vol 22, Issue 2, pp. 215-225, 1975.

[Xu01] J. Xu, R. Weischedel, and C. Nguyen, ”Evaluating a probabilistic model for cross-lingual information retrieval,” *Proc. 24th annual international ACM SIGIR conference on Research and development in information retrieval* , pp. 105-110, 2001

[Zhang02] Y. Zhang and P. Vines, “Improved use of Contextual Information in Cross-language Information Retrieval,” *ACDS,* 2002.