國立政治大學資訊科學系

Department of Computer Science

National Chengchi University

碩士論文

Master's Thesis

應用平行語料建構中文斷詞組件

Applications of Parallel Corpora for Chinese Segmentation

研 究 生：王瑞平

指導教授：劉昭麟

中華民國一百零一年八月

Aug 2012

應用平行語料建構中文斷詞組件

Applications of Parallel Corpora for Chinese Segmentation

研 究 生：王瑞平　　Student：Jui-Ping Wang

指導教授：劉昭麟　　Advisor：Chao-Lin Liu

國立政治大學

資訊科學系

碩士論文

A Thesis

submitted to Department of Computer Science

National Chengchi University

in partial fulfillment of the requirements

for the degree of

Master

in

Computer Science

中華民國一百零一年八月

Aug 2012

致謝

在作為研究生的這兩年中，體驗到了許多大學時未曾接觸過的新事物，也學習到了不少新知識。首先我要感謝我的指導教授 劉昭麟老師；對於不夠主動積極的我，是老師讓我瞭解主動積極的態度的重要性。在作研究方面，謝謝老師不厭其煩地糾正我所犯下的許多錯誤，並引導我到正確的方向。真的很感謝老師這兩年來的教誨與照顧。

在 MIG 實驗室度過的這段時間，感謝各位 MIG 成員對我的關懷與照顧。謝謝建良學長、裕淇學長與怡軒學姐，在我遇到挫折時，因為有學長姐的鼓勵，所以使我能繼續向前邁進。謝謝同學柏廷、家琦，當我在課業或研究等事情上遇到困難時，總是能給予我最大的幫助。謝謝學弟瑋杰、孫暐的關心與在實驗室的大小事情上的協助。謝謝大學部成員的恰恰、刀片、翅膀、鴻源的幫忙。

然後我要感謝媽媽對我無微不至的照顧；聽到妳對我說「早點休息」時，心裡總是覺得很溫暖。也謝謝姊姊給我的鼓勵，讓我能更堅定地朝完成研究所學業的目標前進。

最後也感謝口試委員　高照明老師 與 張嘉惠老師 的指導。

2012 年 7 月　王瑞平

機器智能實驗室

應用平行語料建構中文斷詞組件

摘要

在本論文，我們建構一個基於中英平行語料的中文斷詞系統，並透過該系統對不同領域的語料斷詞。提供我們的系統不同領域的中英平行語料後，系統可以自動化地產生品質不錯的訓練語料，以節省透過人工斷詞方式取得訓練語料所耗費的時間、人力。

在產生訓練語料時，首先對中英平行語料中的所有中文句，透過查詢中文辭典的方式產生句子的各種斷詞組合，再利用英漢翻譯的資訊處理交集型歧異，將錯誤的斷詞組合去除。此外本研究從中英平行語料中擷取新的中英詞對與未知詞，並分別將其擴充至英漢辭典模組與中文辭典模組，以提升我們的系統之斷詞效能。

我們透過兩部分的實驗進行斷詞效能評估，而在實驗中會使用三種不同領域的實驗語料。在第一部分，我們以人工斷詞的測試語料進行斷詞效能評估。在第二部分，我們藉由漢英翻譯的翻

譯品質間接地評估我們的系統之斷詞效能。由實驗結果顯示，我們的系統可以有一定的斷詞效能。

Applications of Parallel Corpora for Chinese Segmentation

Abstract

In this paper, we construct a Chinese word segmentation system which based on Chinese-English Parallel Corpus to save time and manpower, and the corpora in different domains can be segmented by our system.

By providing Chinese-English Parallel Corpus to our system, training corpus can be automatically produced by our system. Then segmentation model can be trained with the produced training corpus. We use Chinese translation of words in English parallel sentences to solve overlapping ambiguity. We extract translation pairs and unknown words from Chinese-English Parallel Corpus.

In evaluation, two different experiments are conducted, and experimental data in three domains are used to evaluate segmentation performance in two experiments. In the first experiment, manually annotated Chinese sentences are used as testing data. In the second experiment, segmentation performance is indirectly indicated by translation quality. Experimental results show that our system achieves acceptable segmentation performance.

目錄

圖目錄

表目錄

第一章　緒論

1.1 研究背景與動機

詞為最小有意義且能夠自由運用的語言單位[2]，而英文與中文在取得句子中的詞的方法上有所不同：在英文可以用空白去斷出英文句中的各個詞，中文則需要透過斷詞這個步驟來取得中文句中的各個詞。因為在機器翻譯、資訊檢索等相關領域上，都需要先對語料進行中文斷詞處理才能進行後續的工作，所以對於中文自然語言處理，中文斷詞是一項非常重要且基礎的工作。

中文斷詞技術大致上可以分為法則式斷詞法以及統計式斷詞法。近來許多採用統計式斷詞法的研究，都能獲得不錯的斷詞效能。不過統計式的斷詞法在訓練斷詞模型時會需要大量的訓練語料，而因為通常透過人工斷詞所得到的訓練語料才能有較高的品質，所以高品質的訓練語料往往不易取得。此外在不同的需求下，使用者可能會提供不同領域的語料給斷詞系統，所以一個中文斷詞系統會需要對不同領域的語料皆有不錯的斷詞效能。但若是使用某一個領域的語料所訓練出的斷詞模型，去對其他不同領域的測試語料進行斷詞的話，可能會因為斷詞模型與其他不同領域的測試語料之間的性質差異大，導致斷詞效能不佳。因此本研究建構一個基於中英平行語料的斷詞系統；提供我們的系統各個不同領域之中英平行語料，就可自動化地得到品質不錯之訓練語料，以節省透過人工斷詞得到訓練語料所需的大量人力與時間；之後該系統會利用所得之訓練語料去訓練斷詞模型，並以斷詞模型對該領域的語料進行斷詞。

中文斷詞存在以下兩個重要問題：斷詞歧異性問題、未知詞問題。斷詞歧異性問題是指當一個中文字串可以被斷成數種的斷詞組合時，則包含該字串的句子在斷詞後可能會被斷成不符合句意的錯誤斷詞結果，進而影響斷詞效能。斷詞歧異性問題包含組合型

歧異(combination ambiguity)和交集型歧異(overlapping ambiguity)，在本研究中我們只著重處理交集型歧異。交集型歧異是當一個中文字串「ABC」可以被斷成「AB/C」及「A/BC」時（A、B、C 皆為單一中文字，AB 與 C 之間的斜線代表詞彙間的斷詞點），則「AB」、「BC」會有共同的交集「B」，如此就會形成交集型歧異，而我們稱「ABC」為交集型歧異字串。為了提升斷詞效能，本研究透過英漢翻譯的資訊去處理交集型歧異。未知詞指的是未收錄於辭典中的詞彙，例如人名、地名、組織名等。在日常生活中人們會不斷創造出新的詞彙，因此不太可能存在一部辭典能包含所有新的詞彙，故未知詞經常會出現在文章中。斷詞系統在對未知詞斷詞時通常會出現錯誤斷詞的情形，所以如果想要提升斷詞效能，則處理未知詞問題會是必要的工作。本研究則透過詞性序列規則去篩選出未知詞。

1.2 研究方法

我們的系統之大略架構為：首先藉由中英平行語料來自動化地得到品質不錯的訓練語料，並利用該訓練語料訓練斷詞模型。之後透過斷詞模型對測試語料斷詞。

在處理交集型歧異時我們會利用英文詞彙的中文翻譯進行對應；而因為英漢辭典中的英文詞彙之中文翻譯有限，所以為了提升利用英漢翻譯的資訊去處理交集型歧異的效果，本研究透過 E-HowNet[24]與一詞泛讀[11]去取得英文詞彙的中文翻譯近義詞，以擴充英文詞彙之中文翻譯數量。在產生訓練語料時，我們對中文語料中的句子，透過查詢中文辭典的方式，得到該句的各種斷詞組合。之後利用英漢翻譯的資訊去處理交集型歧異，將英文詞彙的中文翻譯對應到的斷詞組合視為正確斷詞組合，並去除錯誤的斷詞組合，藉此提升訓練語料的品質。利用英漢翻譯的資訊去處理交集型歧異的原因是：透過英文詞彙的中文翻譯，可以挑選出符合英文陳述的正確中文斷詞組合。得到訓練語料後，我們利用 LingPipe 中文斷詞器[31]及史丹佛中文斷詞器（Stanford Chinese Segmenter）[38]訓練斷詞模型。

我們從中英平行語料中擷取未知詞，藉此處理未知詞問題，以提升我們的系統之斷詞效能；我們從中英平行語料中擷取新的中英詞對來提升利用英漢翻譯的資訊去處理交集型歧異的效果，並藉此提升我們的系統之斷詞效能。以下則為擷取中英詞對與未知詞之大略流程：首先對所有中英平行句對，利用英文詞彙的中文翻譯對中文句斷詞後，在英文句會有「英文遺留字詞」，中文句會有「中文遺留字詞」。透過 PAT-tree 抽詞程式對「中文遺留字詞」進行初步詞彙擷取並以停用詞列表過濾後，就得到「候選中文遺留字詞」；而我們把由「英文遺留字詞」與「候選中文遺留字詞」所構成的詞對稱為「候選中英遺留詞對」。之後利用可能性比例與共現頻率對「候選中英遺留詞對」進行篩選，將通過篩選的「候選中英遺留詞對」視為正確詞對，加入至英漢辭典模組；利用詞性序列規則對「候選中文遺留字詞」進行篩選，將通過篩選的「候選中文遺留字詞」視為未知詞，加入至中文辭典模組。

為了評估我們的系統之斷詞效能，本研究共使用科學文章類型的科學人、C300、C220 及新聞文章類型的新聞語料與會話文章類型的廣播會話語料這三種不同領域之各個語料進行實驗，而關於各種語料的來源會在後續章節詳述。本研究之實驗則分為兩大部分。在第一部分，因為我們沒有測試語料之斷詞標準答案，所以我們對測試語料進行人工斷詞以作為斷詞標準答案，並透過召回率、精確率、F_1-measure 三個評估指標進行斷詞效能評估。在第二部分，我們利用統計式機器翻譯系統「Moses」[33]進行漢英翻譯實驗，並藉由翻譯品質的好壞，來間接地評估斷詞效能之好壞。

1.3 論文架構

在第一章我們介紹研究背景與動機、研究方法，第二章回顧中文斷詞之相關研究與基於英漢雙語平行語料進行斷詞的相關研究，第三章針對辭典模組與加入近義詞之英漢合併辭典建置進行介紹，第四章說明我們的系統的架構與介紹訓練斷詞模型的工具，第五章詳述產生訓練語料的方法，第六章說明實驗語料的來源及介紹以人工斷詞測試語料評估

斷詞效能、以漢英翻譯的翻譯品質評估斷詞效能這兩部分實驗的實驗流程與實驗結果分析，第七章為結論與未來展望。

第二章　文獻探討

本章共分為兩小節。在 2.1 節回顧中文斷詞之相關研究，2.2 節介紹基於英漢雙語平行語料進行中文斷詞的相關研究。

2.1 中文斷詞之相關研究

本節共分為四個小節。在 2.1.1 節、2.1.2 節分別回顧法則式斷詞法、統計式斷詞法之相關研究，2.1.3 節回顧處理斷詞歧異性、未知詞問題之相關研究，2.1.4 節為斷詞標準不一問題之相關研究。

2.1.1 法則式斷詞法之相關研究

法則式的斷詞法會利用辭典，並搭配規則進行斷詞。Chen[19]提出了利用經驗法則（heuristic rules）處理斷詞歧異性問題的方法。該方法當斷詞歧異性問題發生時，會根據辭典及 determinative-measure compounds rule 產生以該字詞開頭的連續三個詞的所有詞組。產生所有詞組後，會利用六條經驗法則去挑選符合規則的詞組。

2.1.2 統計式斷詞法之相關研究

在統計式斷詞法中，有許多研究將中文斷詞視為字元標記的工作，而在其中較廣泛被使用的技術有條件隨機域模型（Conditional Random Fields）、隱藏式馬可夫模型（Hidden Markov Models）、感知器（Perceptron）等。以下為統計式斷詞法之相關研究介紹。

Wang 等[40]結合 character-based discriminative model 和 character-based generative model 這兩種模型以進行斷詞；該研究中提到了比起使用單一模型，如果將兩種模型結合，能夠有更好的斷詞效能。

Jiang[28]使用 cascaded linear model 進行斷詞與詞性標記(POS tagging)；cascaded linear model 為兩層的結構，在內層利用以字元為基礎的(character-based) 感知器作為核心，並在外層的線性模型（linear model），將感知器的輸出結果作為特徵(feature)，搭配語言模型（language model）等其他特徵一起去訓練模型。結果顯示比起單獨使用感知器的斷詞模型，cascaded linear model 不管在斷詞還是結合斷詞與詞性標記的工作上都能有更好的正確率。

以下介紹國內一些採用統計式的斷詞法的研究。詹嘉丞[14]提出一個針對非繁體中文字進行處理的方法，使得斷詞系統遇到非繁體中文字也能斷詞。該研究利用判斷模組處理中文人名與日文人名，並將繁簡日韓漢字對應到繁體字，以便之後能使用繁體的已知詞進行候選詞判斷。在斷詞效能上，採用 bigram 機率模型時，F-Measure 可以達到 94.16 %。

朱怡霖[6]採用交疊式(interleaving)方式將中文斷詞與專有名詞辨識兩項工作整合。與管流式(pipeline) 方式不同，採用交疊式方式，會把所有候選詞保留住，並利用人名、地名、組織名等辨識模組辨識出可能的候選詞，到最後再選出由候選詞組成的斷詞組合中的最佳斷詞組合。

林筱晴[8]認為與傳統語料庫相比，web 擁有更大的資料量，且具有即時性；所以該研究將 web 當成一個大型語料庫，將搜尋引擎所提供的 page count 作為詞頻套用至 likelihood ratio test，以辨識人名、地名、組織名這三種型態的未知詞；在斷詞歧異性問題方面，則是利用 word-based bigram model 進行處理。

利用隱藏式馬可夫模型進行中文斷詞時，許多研究會使用外部資源或是結合其他的機器學習演算法來提高斷詞效能。但林千翔[9]不使用任何外部資源，而是應用特製化（specialization）的概念，將長詞優先斷詞法與隱藏式馬可夫模型結合，使得隱藏式馬可夫模型能夠帶有斷詞歧義性及未知詞的資訊，進而提升斷詞效能。

羅永聖[18]透過兩階段的方式進行斷詞。在第一階段，透過查詢辭典得到斷詞候選句後，利用條件隨機域模型從斷詞候選句中去除機率較低的句子；在第二階段會利用語言規則處理人名、二字詞拆解等問題，最後再利用條件隨機域模型選出最好的候選句做為最後的斷詞結果。

上述提到的研究中，有部分研究透過查詢辭典並搭配規則的方式來產生所有的斷詞組合[8][14][18]，並在產生所有斷詞組合後，利用馬可夫 bigram 機率模型去處理斷詞歧異性問題[8][14]，或利用條件隨機域模型去處理斷詞歧異性問題[18]；他們的作法是透過機率模型算出所有斷詞組合的機率，再選擇所有斷詞組合中機率值最高的斷詞組合作為正確的斷詞組合。而與他們的作法不同，我們產生各種斷詞組合後，不透過機率模型，而是利用英漢翻譯的資訊去找出正確的斷詞組合。

綜觀上述所提到的各種統計式的斷詞法，幾乎都會需要使用大量的訓練語料；而因為現在公開提供使用的人工斷詞的訓練語料不多，所以他們所使用的訓練語料大多是中研院平衡語料庫[3]或 SIGHAN Bakeoff 2[37]所公開的 4 種訓練語料(由中央研究院（Academia Sinica）、香港城市大學（City University of Hong Kong）、北京大學(Peking University)及微軟亞洲研究院（Microsoft Research）所提供)。我們希望能透過系統化的流程來自動地產生訓練語料，這樣在訓練斷詞模型時就可以不用侷限於少數幾種公開提供使用的語料。

2.1.3 斷詞歧異性問題與未知詞問題之相關研究

在處理斷詞歧異性問題的研究中， Li[30]等人利用非監督式（unsupervised）訓練的方法處理交集型歧異。該研究利用非監督式的方式去訓練 Naive Bayesian 分類器，將判斷交集型歧異的問題轉換成二元分類的問題後，搭配 ensemble learning，藉由多數的分類器的投票結果去決定最後的斷詞結果。之後利用 5759 筆人工標記的交集型歧異字串的測試集進行實驗，結果顯示該方法能夠有 94.3%的正確率。

以下介紹一些處理未知詞問題的研究。利用以辭典為基的斷詞法對未知詞進行斷詞的話，未知詞會被切成幾個較小的單位，而 Chen 等人[20]觀察到大多數的未知詞的詞構中都會包含單字詞。不過單字詞除了可能是未知詞的一部份之外，也有可能是單獨使用的已知詞。所以他們利用以語料庫為基的學習法(corpus-based learning approach)去產生偵測單獨使用的已知詞的規則，符合規則的單字詞為單獨使用的已知詞，而不符合規則的單字詞即為未知詞的一部份。

Chen 等人[21]在 2002 年的研究中提出了一種擷取未知詞的方法；該研究在擷取未知詞時，會針對屬於未知詞一部份的單字詞，判斷該單字詞是否可以和相鄰的詞彙進行合併。該研究使用「形態規則」(morphological rules)與「統計規則」(statistical rules)去擷取未知詞，並利用結構正確性(structure validity)、句法正確性(syntactic validity)、區域一致性 (local consistency)三條準則去驗證所擷取出的結果是否為未知詞。

2.1.4 斷詞標準不一問題之相關研究

由於「中文詞」的定義每個人並不相同，且不同的工作可能適合不同的斷詞標準，所以各個斷詞系統之間或各個人工標記語料庫(manually annotated corpora)之間可能擁有不

同的斷詞標準。以下介紹處理各個人工標記語料庫之間斷詞標準不一的問題之相關研究。

為了處理人工標記語料庫之間斷詞標準不一的問題，Jiang[29]提出了自動化地將斷詞標準轉換成另一種斷詞標準的方法。Jiang 使用以來源語料庫(source corpus)所訓練的來源分類器(source classifier)對目標語料庫(target corpus)斷詞，將該斷詞結果作為引導資訊(guide information)，使用引導資訊與目標語料庫訓練出目標分類器(target classifier)。之後以目標分類器將句子從來源語料庫的斷詞標準轉換成目標語料庫的斷詞標準。

2.2 基於英漢雙語平行語料進行斷詞的相關研究

在進行漢英或英漢機器翻譯時，需要先對英漢雙語平行語料中之中文語料進行斷詞，才能進行後續的動作，故對漢英或英漢機器翻譯工作而言，中文斷詞是重要的工作；而在進行漢英或英漢機器翻譯時，有些研究在對中文語料斷詞時會運用英漢雙語平行語料中的英漢雙語資訊，藉此提升翻譯的品質。

Huang[26]使用基於chinese language model與bilingual STM的模型去進行斷詞。在訓練的過程，利用IBM model 2去估算翻譯機率(translation probability) 以及對列機率(alignment probability)，以建立bilingual STM。實驗結果顯示使用bilingual segmentation演算法的召回率（recall）為0.794，比起單語（monolingual）的斷詞工具Ketitool的召回率來得更好。在中翻英的工作上，使用bilingual segmentation演算法所得到的BLEU分數為0.235，也優於使用Ketitool的BLEU分數(0.223)。

Ma[32]提出了一種適用於不同領域的雙語平行語料的斷詞法。該方法為：首先建立 1 對 n 關係的雙語辭典（bilingual 1-to-n dictionary），該辭典的格式為一個英文詞彙對列至多個中文字。該研究將中文句斷成由中文字組成的字串後，再查詢 1 對 n 關係的雙語

辭典，若辭典的某詞對的英文詞彙有包含在英文句，且該英文詞彙所對列到的多個中文字有包含在中文句，則將該多個中文字結合成一個中文詞彙。

第三章　系統架構

在本章我們於 3.1 節介紹我們的系統之流程與架構，於 3.2 節介紹訓練斷詞模型時所使用的軟體工具。

3.1 系統流程與架構

本系統的整體架構與流程如圖 3.1 所示，而流程中各步驟的詳細說明會在後續章節加以介紹；首先，將從中英平行語料中所篩選出的候選中英遺留詞對擴充至英漢辭典模組，將從中英平行語料中所篩選出的候選中文遺留字詞擴充至中文辭典模組。提供本系統中英平行語料後，我們透過查詢中文辭典模組中的辭典之方式，對語料中的每一句中文句產生該句的各種斷詞組合。而為了得到較少錯誤的訓練語料，我們藉由查詢英漢辭典模

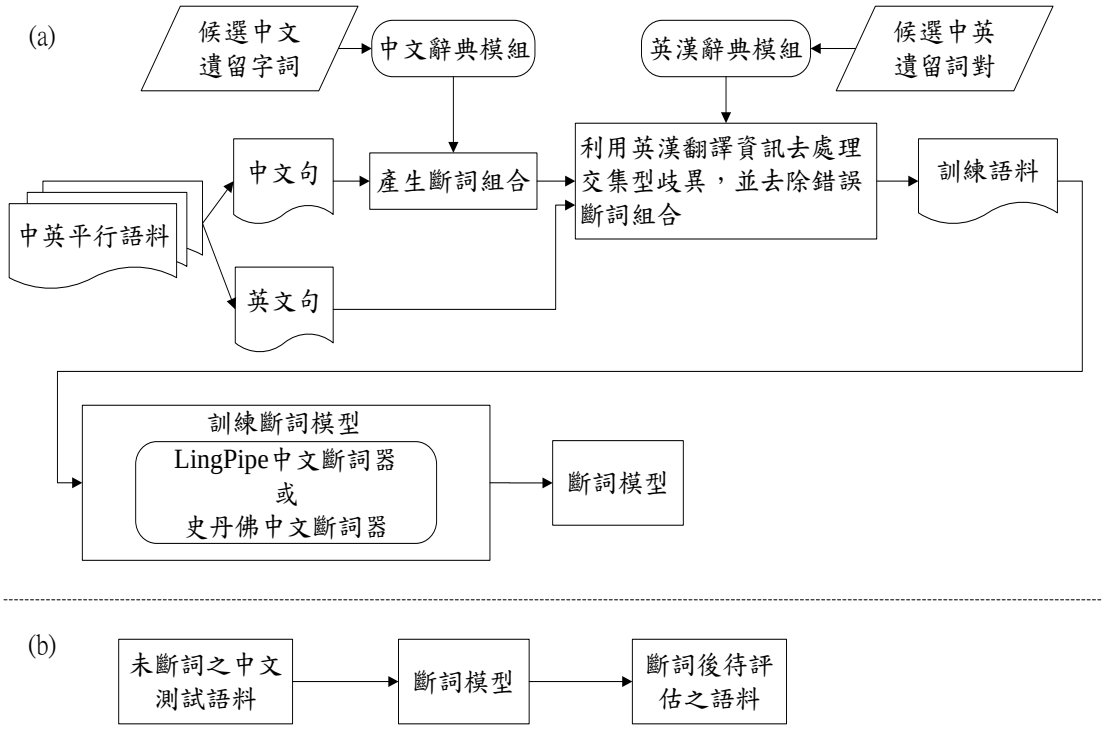

圖 3.1 系統的流程與架構

組中的辭典之方式來利用英漢翻譯的資訊去處理交集型歧異，將所產生的斷詞組合中的錯誤斷詞組合去除。得到訓練語料後，我們利用 LingPipe 中文斷詞器及史丹佛中文斷詞器訓練斷詞模型；透過上述兩種工具去訓練斷詞模型時，除了提供這兩種工具訓練語料之外，也可以加入外部辭典一起訓練。最後利用所得到的斷詞模型將未斷詞測試語料進行斷詞，得到已斷詞之語料。

3.2 斷詞模型訓練工具

本研究利用史丹佛中文斷詞器以及 LingPipe 中文斷詞器這兩種軟體工具進行斷詞模型的訓練。

史丹佛中文斷詞器是基於條件隨機域實做而成的斷詞器。我們將訓練語料提供給史丹佛中文斷詞器進行訓練，在訓練完成後會得到斷詞模型。而在 2008 年 5 月 21 號之後的版本，史丹佛中文斷詞器可以在訓練斷詞模型時藉由加入外部辭典的方式來增加 lexicon–based features 至條件隨機域模型中。藉由增加 lexicon–based features，可以增加斷詞器斷詞時的一致性(consistency)[22]。

LingPipe 中文斷詞器在進行斷詞時會將詞彙間未正確插入空白的情形視為錯誤，並透過斷詞模型，將空白插入至兩詞彙間以修正該錯誤。我們將訓練語料提供給 LingPipe 中文斷詞器進行訓練，就可以得到斷詞模型。而 LingPipe 中文斷詞器也可於訓練斷詞模型時加入外部辭典，以提供更多資訊給斷詞模型。

第四章　辭典模組介紹與加入近義詞

本章共分為兩個節次。在 4.1 節介紹我們的系統之辭典模組，4.2 節介紹加入近義詞的方法與「加入近義詞之英漢合併辭典」的建置。

4.1 辭典模組介紹

我們的系統之辭典模組包含英漢辭典模組與中文辭典模組，而這兩個模組中都包含一般辭典與專業辭典兩種類別。英漢辭典模組中的專業辭典類別包含「英漢技術名詞辭典」，中文辭典模組中的專業辭典類別包含「中文技術名詞辭典」、世界人名翻譯大辭典，關於「英漢技術名詞辭典」與「中文技術名詞辭典」的建置會在下一段落說明。而英漢辭典模組中的一般辭典類別包含「加入近義詞之英漢合併辭典」與懶蟲簡明英漢詞典[17]，關於「加入近義詞之英漢合併辭典」的建置會在 4.2 節詳細說明。中文辭典模組的一般辭典類別則包含教育部國語辭典[13]、成語詞典[7]及高級漢語大詞典[13]。中文辭典模

表 4.1 中文辭典模組之辭典詞彙數統計

中文辭典模組		
辭典類別	辭典名稱	中文詞彙數
一般辭典	教育部國語辭典	157704
一般辭典	成語詞典	13947
一般辭典	高級漢語大詞典	54467
專業辭典	中文技術名詞辭典	804053
專業辭典	世界人名翻譯大辭典	648612

表 4.2 英漢辭典模組之辭典詞彙數統計

英漢辭典模組			
辭典類別	辭典名稱	英文詞彙數	中文詞彙數
一般辭典	加入近義詞之英漢合併辭典	99805	3729292
一般辭典	懶蟲簡明英漢詞典	121525	323766
專業辭典	英漢技術名詞辭典	586075	804053

組之辭典詞彙數統計、英漢辭典模組之辭典詞彙數統計分別如上頁表 4.1、表 4.2 所示。

本研究從國家教育研究院學術名詞資訊網[12]下載了 138 個技術名詞檔案，並將其整合成「英漢技術名詞辭典」。「英漢技術名詞辭典」的內容格式為一個英文技術名詞對應一個中文技術名詞的形式，而「中文技術名詞辭典」是只取「英漢技術名詞辭典」中的中文技術名詞整合而成。

4.2 加入近義詞之英漢合併辭典建置

當中文句出現交集型歧異時，我們會利用英漢辭典中的英文詞彙之中文翻譯去進行比對，所以為了提高利用英漢翻譯的資訊去處理交集型歧異的效果，會須要增加英文詞彙的中文翻譯詞彙數目；我們參考[10]的作法將牛津現代英漢雙解詞典[1]和 Dr.eye 譯典通線上字典[23]合併成「英漢合併辭典」，以增加英文詞彙的中文翻譯詞彙數目。除了利用英漢辭典的中文翻譯詞彙，本研究也加入中文翻譯詞彙的近義詞，來擴充英漢辭典的中文翻譯詞彙之數目。本研究使用中央研究院現代漢語一詞泛讀系統（以下簡稱一詞泛讀）及 E-HowNet 去尋找中文翻譯詞彙的近義詞。

我們於 4.2.1 節中介紹利用一詞泛讀尋找近義詞的方法，4.2.2 節中介紹利用 E-HowNet 尋找近義詞的方法，4.2.3 節介紹建置「加入近義詞之英漢合併辭典」的流程。

4.2.1 利用一詞泛讀尋找近義詞

本研究參考[10]的作法去利用一詞泛讀取得近義詞。將中文詞彙輸入至一詞泛讀，一詞泛讀會傳回該詞彙的近義詞群。我們把英文詞彙的各個中文翻譯輸入至一詞泛讀，將一詞泛讀傳回的各個中文翻譯的近義詞群合併後，即為該英文詞彙的中文翻譯近義詞集。以下以單字“quietness”為例進行說明，而“quietness”的中文翻譯為「安靜」、「肅靜」、「平靜」與「樸素」。我們將「安靜」、「肅靜」、「平靜」與「樸素」逐一輸入至一詞泛讀，

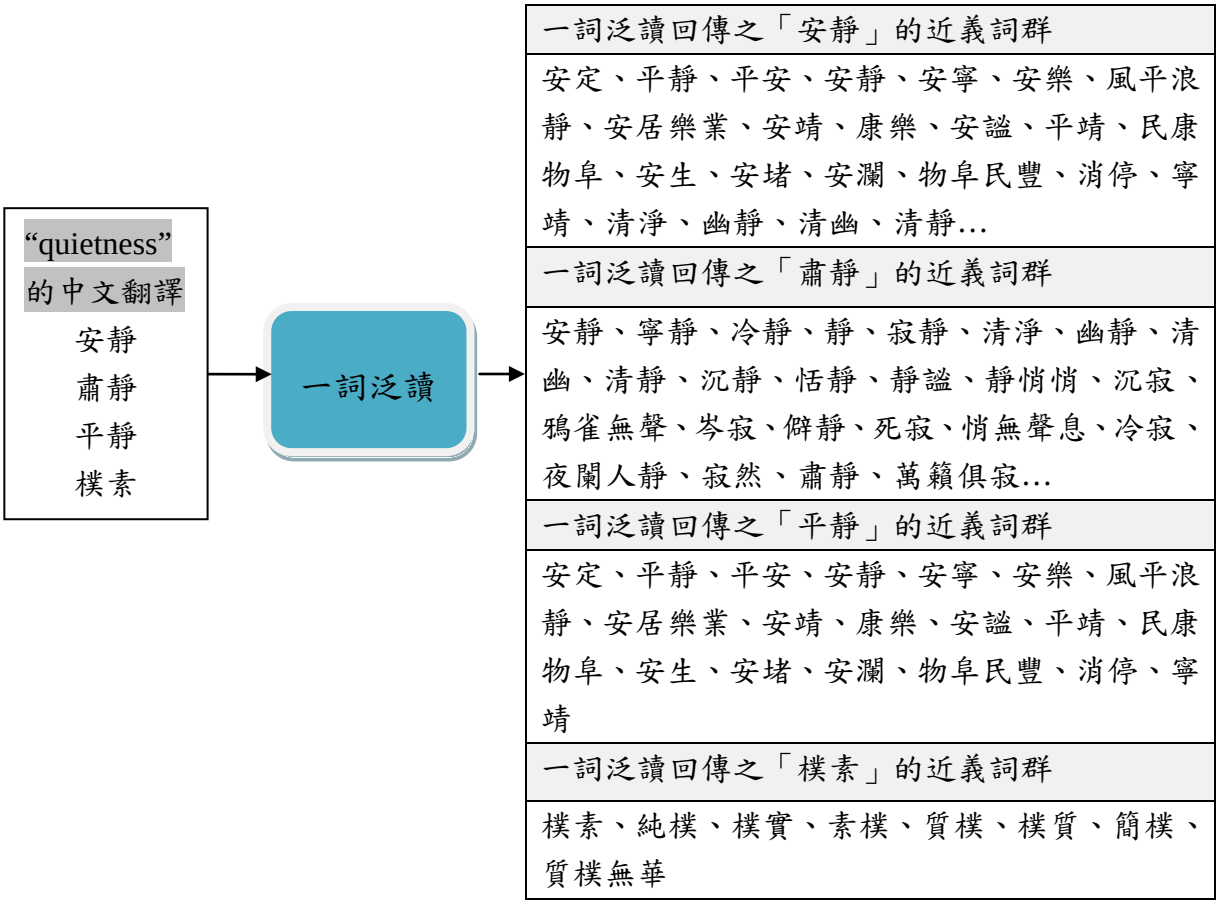

圖 4.1 輸入“quietness”的中文翻譯後一詞泛讀回傳之近義詞群

而一詞泛讀回傳的結果如圖 4.1 所示。最後把系統傳回的這四個中文翻譯的近義詞群合併，就會得到“quietness”的中文翻譯近義詞集。

4.2.2 利用 E-HowNet 尋找近義詞

在 E-HowNet 中，用來定義及描述詞彙之語義（概念）的單位為義原或簡單概念(simpler concept)，而以下將介紹 E-HowNet 詞彙之定義結構。下頁圖 4.2、圖 4.3 中的 TopLevelDefinition 和 BottomLevelExpansion 標記中含有描述詞彙語義的表示式，而 BottomLevelExpansion 表示式是 TopLevelDefinition 表示式之更細一步的意義擴充[10]；BottomLevelExpansion 表示式可以有以下幾種構成方式：由一個上位詞概念與許多特徵所構成、由一個概念所構成、由一個義原所構成[25]。由一個上位詞概念與許多特徵所

```
<Word item = "懼高症">
        <WordFreq>7</WordFreq>
        <WordSense id="1">
            <English>acrophobia</English>
            <Phone>ㄐㄩˋ　ㄍㄠ　ㄓㄥˋ</Phone>
            <PinYin>ju4 gao1 heng4</PinYin>
            <SyntacticFunction>
                <POS>Nad</POS>
                <Freq>7</Freq>
            </SyntacticFunction>
            <TopLevelDefinition>
                {disease|疾病:CoEvent={fear|害怕:cause={high|高}}}
            </TopLevelDefinition>
            <BottomLevelExpansion>
                {disease|疾病:CoEvent={fear|害怕:cause={high|高}}}
            </BottomLevelExpansion>
        </WordSense>
 </Word>
```

圖 4.2 E-HowNet 詞彙的定義結構之範例一

```
<Word item = "樹幹">
        <WordFreq>56</WordFreq>
        <WordSense id="1">
            <English>tree trunk</English>
            <Phone>ㄕㄨˋ　ㄍㄢˋ</Phone>
            <PinYin>shu4 gan4</PinYin>
            <SyntacticFunction>
                <POS>Nab</POS>
                <Freq>56</Freq>
            </SyntacticFunction>
            <TopLevelDefinition>{BodyPart({tree|樹})}</TopLevelDefinition>
            <BottomLevelExpansion>{BodyPart({tree|樹})}</BottomLevelExpansion>
        </WordSense>
</Word>
```

圖 4.3 E-HowNet 詞彙的定義結構之範例二

構成之構成方式如上頁圖 4.2 所示，BottomLevelExpansion 表示式中的義原「disease|疾病」為「懼高症」的上位詞概念，而「CoEvent={fear|害怕:cause={high|高}}」、「cause={high|高}」為其特徵。此外在 E-HowNet 中含有兩種關係：語意角色(Semantic Role)以及函數(function)。函數可以將一個概念轉換成另一個新的概念，如在上頁圖 4.3「樹幹」之 BottomLevelExpansion 表示式部分，函數 BodyPart( )可以將義原 tree|樹所構成的概念，轉換成另一個新的概念。語意角色則是用來建構兩個參數間的主題關係(thematic relation)、性質屬性(property attribute)[25]。

在詞彙的相似度計算上，我們參考了 Liu 與 Li[16]於 2002 年提出的方法。Liu 與 Li 透過計算兩詞彙的概念語義運算式之相似度的方式來得到兩個詞彙之間的相似度，本研究則透過計算兩詞彙的 BottomLevelExpansion 表示式之相似度的方式來得到兩個詞彙之間的相似度。我們的想法為：因為 BottomLevelExpansion 表示式是用來描述詞彙之語義，所以兩個互為近義詞的詞彙應該會有相似的 BottomLevelExpansion 表示式。

在本研究中我們只會利用 BottomLevelExpansion 表示式去計算詞彙相似度，而不會利用到 TopLevelDefinition 表示式，所以在本章以下內容中我們將 BottomLevelExpansion 表示式簡稱為「表示式」。在計算表示式之相似度時我們首先會對表示式進行擷取。為了簡化計算的複雜度，我們不從表示式中擷取結構較複雜的特徵，而僅擷取義原、由義原與修飾義原之函數所構成之組合（以下簡稱義原函數組合）來代表該表示式。例如對於上頁圖 4.2 中的詞彙「懼高症」，我們會擷取「懼高症」之上位詞概念「disease|疾病」與特徵中的義原「fear|害怕」、「high|高」，對於上頁圖 4.3 中的詞彙「樹幹」會擷取「BodyPart({tree|樹})」這個義原函數組合。

首先在義原的相似度計算上，我們沿用 Liu 與 Li 在 2002 年所使用的公式，即以下公式(1)：

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha} \tag{1}$$

公式(1)中的p_1、p_2為兩個義原，α則是可調節的參數。在廣義知網知識分類體系（E-HowNet Taxonomy）中包含了兩棵子樹(subtree)：{entity|事物}、{relation|關係}，義原為{entity|事物}中的節點，函數則為{relation|關係}中的節點。我們將 d 定義為 p_1、p_2在{entity|事物}中的路徑長度(d 為一整數)，若以圖 4.4 中的義原「sky |空域」、「the Pacific Ocean |太平洋」為例，則這兩個義原的 d 為 3。

Liu 與 Li 將義原的相似度計算作為概念語義運算式的相似度計算之基礎，但我們發現若只以義原的相似度計算作為表示式之相似度計算之基礎，可能會將相似度不高的詞彙當做近義詞。例如以下頁圖 4.5(b)中的兩個詞彙為例，因為「出聲」、「口吻」的義原皆為「speak|說」，若只以義原作為表示式相似度計算之基礎，則這兩個詞彙的表示式之相似度為 1，所以相似度不高的「出聲」會被視為「口吻」之近義詞。因此本研究在計算表示式相似度時，不只以義原作為表示式相似度計算之基礎，而另外考慮了函數對於概念的影響，以義原或義原函數組合的相似度計算作為表示式相似度計算之基礎。義原或義原函數組合的相似度之計算則如公式(2)所示。

$$Sim(c_1, c_2) = W_f \times \frac{\alpha}{d + \alpha} \tag{2}$$

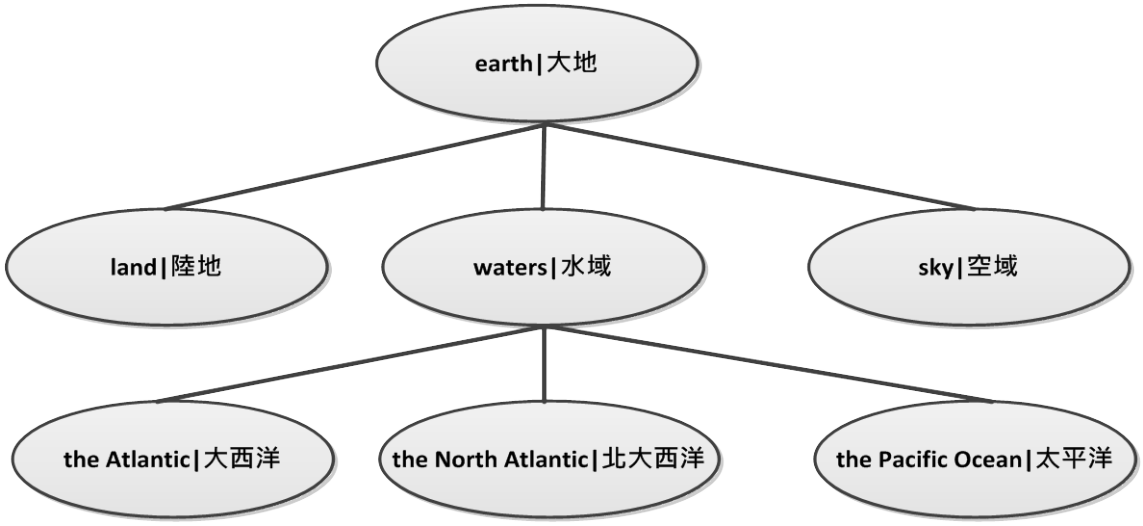

圖 4.4 義原於{entity|事物}中的距離之範例

(a)

<Word item = "兵士">

<BottomLevelExpansion>{member({army|軍隊})}</BottomLevelExpansion>

</Word>

<Word item = "兵種">

<BottomLevelExpansion>{kind({army|軍隊})}</BottomLevelExpansion>

</Word>

(b)

<Word item = "出聲">

<BottomLevelExpansion>{speak|說}</BottomLevelExpansion>

</Word>

<Word item = "口吻">

<BottomLevelExpansion>{style({speak|說})}</BottomLevelExpansion>

</Word>

圖 4.5 計算W_f之範例詞彙

$$W_f = \begin{cases} 1, & \text{如果} c_1 \text{及} c_2 \text{皆為義原} \\ Sim(f_1, f_2), & \text{如果} c_1 \text{及} c_2 \text{皆為義原函數組合} \\ \delta, & \text{如果} c_1 \text{及} c_2 \text{其中之一為義原函數組合} \end{cases} \quad (3)$$

$$Sim(f_1, f_2) = \frac{\alpha}{f_d + \alpha} \quad (4)$$

上頁公式(2)中的c_1、c_2為義原或義原函數組合，W_f的計算方式則如公式(3)所示。當c_1及c_2皆為義原時，我們將W_f設為 1，此時公式(2)等同於計算義原相似度的公式(1)。當c_1及c_2皆為義原函數組合時（如圖 4.5 (a)之 member({army|軍隊})、kind({army|軍隊})），則W_f的值為c_1之函數(f_1)與c_2之函數(f_2)間的相似度$Sim(f_1, f_2)$，而函數間的相似度可由公式(4)得到；公式(4)中的f_1、f_2為兩個函數，f_d定義為f_1、f_2在{relation|關係}中的路徑長度(f_d為一整數)。當c_1或c_2的其中之一為義原函數組合時（如圖 4.5 (b)之 speak|說、style({speak|說})），則因為該義原函數組合對應到義原，所以該義原函數組合中的函數所對應的函數為空，此時我們參考 Liu 與 Li 的作法，將W_f的值設為常數δ（任一非空值與空值的相似度）。

以下介紹詞彙的表示式相似度之計算方法。當兩個詞彙的表示式皆為義原或義原函數組合時，則詞彙的表示式相似度可由公式(2)得到。當兩個詞彙中至少一個詞彙之表示式是由一個上位詞概念與許多特徵構成時，我們利用公式(5) 計算詞彙的表示式的相似度。當詞彙之表示式是由一個上位詞概念與許多特徵構成時，我們覺得表示式中的上位詞概念是描述此詞彙之概念的主要部分，所以將上位詞概念稱為「主要概念描述」，而由各個特徵中的義原或義原函數組合所構成的集合則稱為「次要概念描述」。在公式(5)中的$Sim_1(SS_1, SS_2)$為兩個詞彙表示式的「主要概念描述」的相似度，$Sim_2(SS_1, SS_2)$為兩個詞彙表示式的「次要概念描述」的相似度。在公式(5)中的權重β_i的設定上，因為我們覺得「主要概念描述」的重要性大於「次要概念描述」，所以設定$\beta_1 > \beta_2,\ \beta_1 + \beta_2 = 1$ 。

$$\begin{aligned} Sim(SS_1, SS_2) &= \sum_{i=1}^{2} \beta_i Sim_i(SS_1, SS_2) \\ &= \beta_1 Sim_1(SS_1, SS_2) + \beta_2 Sim_2(SS_1, SS_2) \end{aligned} \tag{5}$$

我們在利用公式(5)計算詞彙相似度時會分成以下兩種情形進行計算，下頁圖 4.6 則為這兩種情形的範例，在圖中沒有用灰底標示的部分是主要概念描述，有用灰底標示的部分則是次要概念描述。情形 1：當兩個詞彙的表示式的構成方式都是由一個上位詞概念與許多特徵構成時（如圖 4.6 中的「僕人」、「女傭」兩詞彙），則兩詞彙表示式的$Sim_1(SS_1, SS_2)$可由公式(2)得到，在$Sim_2(SS_1, SS_2)$計算上，我們沿用[16]中的集合的相似度計算之演算法進行計算，而集合中的元素為義原或義原函數組合。情形 2：當一個詞彙的表示式是由一個上位詞概念與許多特徵構成，另一個詞彙的表示式是由一個義原或義原函數組合構成時（如圖 4.6 中的「僕人」、「假冒」兩詞彙），我們將由一個義原或義原函數組合構成的表示式中的義原或義原函數組合視為「主要概念描述」，將該表示式的「次要概念描述」視為空值；兩詞彙表示式的$Sim_1(SS_1, SS_2)$也是由公式(2)得到，而因為我們定義集合與空值的相似度為常數δ，所以$Sim_2(SS_1, SS_2)$為δ。

僕人之表示式：{human|人:predication={engage|從事:content={affairs|事務:CoEvent={engage|從事}},location={family|家庭},agent={~}}}		
女傭之表示式：{human|人:predication={engage|從事:content={affairs|事務},location={family|家庭},agent={~}},gender={female|女}}		
假冒之表示式：{fake|偽}		
（情形 1）	主要概念描述	次要概念描述
僕人之擷取後的表示式：	human|人	engage|從事、affairs|事務、family|家庭
女傭之擷取後的表示式：	human|人	engage|從事、affairs|事務、family|家庭、female|女

（情形 2）	主要概念描述	次要概念描述
僕人之擷取後的表示式：	human|人	engage|從事、affairs|事務、family|家庭
假冒之擷取後的表示式：	fake|偽	

圖 4.6 計算詞彙相似度之兩種情形的範例

介紹了計算兩詞彙的表示式相似度之計算方法後，以下我們透過下頁圖 4.7 中的英文詞彙“servitor”為例，說明如何取得英文詞彙的中文翻譯近義詞集。首先對“servitor”的所有中文翻譯之表示式與所有 E-HowNet 中文詞彙之表示式進行擷取，而擷取後的表示式如圖 4.7 中灰底標記所示。之後將“servitor”的所有中文翻譯之表示式，一一與所有 E-HowNet 中文詞彙之表示式計算相似度後，將相似度高於門檻值的 E-HowNet 中文詞彙視為近義詞。

以下為我們設定各個公式中所存在之可調節的參數與相似度之門檻值的過程。在各個公式中所存在之可調節的參數分別為：α、δ、β_1、β_2。我們將α的值限制在 1.6 或 2.0，δ的值限制在 0.05 或 0.1 或 0.2，因為我們設定$\beta_1 > \beta_2$, $\beta_1 + \beta_2 = 1$，所以將(β_1, β_2)的值限制在(0.9,0.1)、(0.8,0.2)、(0.7,0.3)、(0.6,0.4)這四種。相似度之門檻值則限制在 0.9、0.8、0.7。之後針對在不同的α、δ、(β_1, β_2)、相似度之門檻值設定下所得的結果，我們用人工方式比較結果中的部分英文詞彙之中文翻譯近義詞，以選出較佳的參數組合。最後我們設定α為 1.6，δ為 0.05，β_1為 0.6，β_2為 0.4，相似度之門檻值為 0.9。

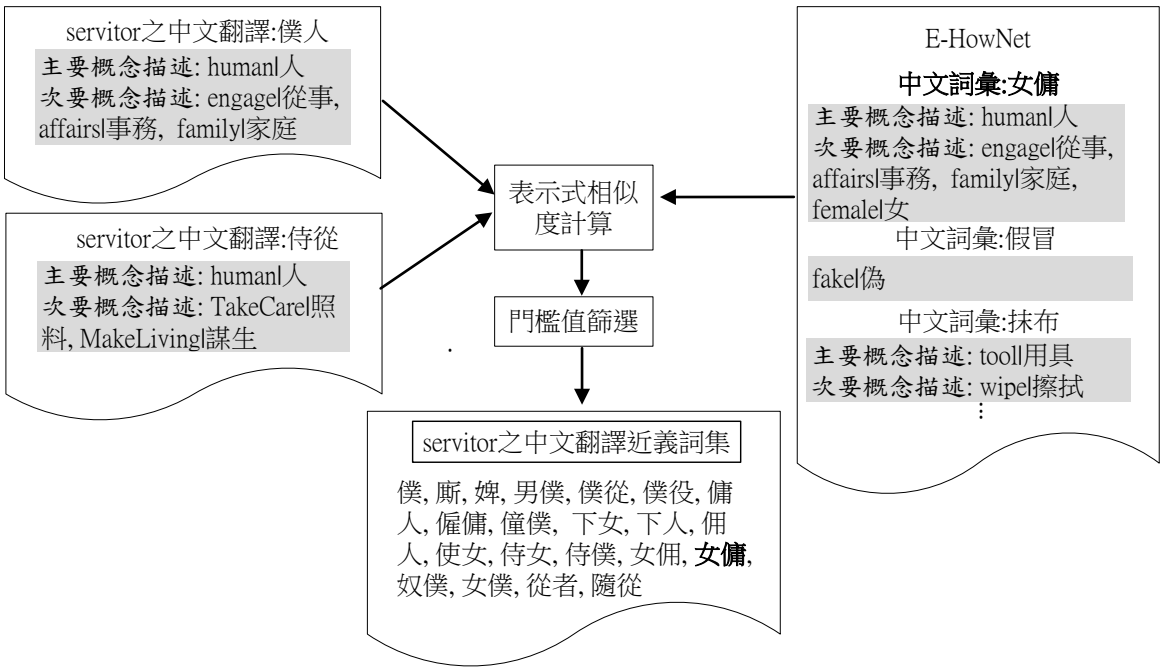

圖 4.7 藉由表示式相似度計算取得中文翻譯近義詞集之流程

4.2.3 辭典建置流程

以下為建置「加入近義詞之英漢合併辭典」之流程：我們對「英漢合併辭典」中的各個英文詞彙，依照在 4.2.1、4.2.2 節所述的方法從一詞泛讀及 E-HowNet 取得該詞彙的中文翻譯近義詞集後，我們把從一詞泛讀及 E-HowNet 得到的中文翻譯近義詞集與「英漢合併辭典」的英文詞彙之中文翻譯詞彙進行整合，就完成「加入近義詞之英漢合併辭典」的建置。

第五章　產生訓練語料

第五章主要介紹產生訓練語料與擷取中英平行語料中的中英詞對與未知詞的方法。5.1 節介紹產生句子的各種斷詞組合的方法，5.2 節介紹如何利用英漢翻譯的資訊處理交集型歧異，並去除錯誤的斷詞組合，5.3 節介紹擷取中英詞對與未知詞的方法。

5.1 產生各種斷詞組合

中文句可以看成是由字所組成的字串，而隨著組合成句子的詞彙的不同，會形成不同的斷詞組合。因此我們針對未斷詞語料中的每句中文句，透過查詢中文辭典的方式，產生由不同的詞彙所組成的句子之各種斷詞組合，藉此得到訓練語料。我們產生中文句的各種斷詞組合的目的為希望在訓練斷詞模型的過程中，透過大量語料的統計現象，來得到較佳的斷詞模型。我們將句子表示成字串 $C_{1:n}$（$C_{1:n} = C_1\ C_{2...}C_n$），並依照圖 5.1 的步驟來產生句子的各種斷詞組合。以下為圖 5.1 中 V_i 與 $Cand_i$（i=1 to n）的定義。V_i 為詞彙集合，在 V_i 內會存放句子中所有以 C_i 開頭的詞彙。$Cand_i$ 為候選集合，在 $Cand_i$ 內會

1. 針對句子中的每一個字 C_i（i=1 to n）查詢中文辭典模組的辭典中是否包含句子中以該字開頭的不同長度之字串（字串的長度為 1 to n-i+1），若包含則將該字串加入 V_i 。
2. 將 i 的初始值設為 1。
3. (a).如果 V_1 中的某一詞彙等同於 $C_{1:i}$，則把該詞彙加入至 $Cand_i$。
   (b). for j =1 to i-1, $i > 1$
   如果 $Cand_j$ 中的某一斷詞組合加上 V_{j+1} 中的另一詞彙後，不含有「包含單字詞的詞彙組合」，並且等同於 $C_{1:i}$，則把該斷詞組合加入至 $Cand_i$。
4. 如果 i 不等於 n，則把 i 遞增 1，並重回到步驟 3。如果 i 等於 n，則 $Cand_i$ 內的所有斷詞組合即為該句子的各種斷詞組合。

圖 5.1 產生句子的各種斷詞組合的步驟

存放字串 $C_{1:i}$ 的各種斷詞組合。

在上頁圖 5.1 步驟 3(b)中提到的「包含單字詞的詞彙組合」的定義為：當某詞彙組合中包含單字詞，且該詞彙組合可以結合成一個詞彙時，則該詞彙組合為「包含單字詞的詞彙組合」。例如「科學/家」這一個詞彙組合包含了單字詞「家」，且「科學/家」可以結合成詞彙「科學家」，則「科學/家」為「包含單字詞的詞彙組合」。我們發現若句子內含有許多「包含單字詞的詞彙組合」時，會產生大量的斷詞組合。如「一家民間公司提議用鐵粉在部分海洋施肥」這句中文句，包含了「一/家」、「民/間」、「公/司」、「提/議」、「鐵/粉」、「部/分」、「海/洋」、「施/肥」這些「包含單字詞的詞彙組合」，而在不去除含有「包含單字詞的詞彙組合」之斷詞組合的情況下，最後該中文句會產生 256 組的斷詞組合。若語料中的許多中文句都會產生大量的斷詞組合，就會使得訓練語料變得過於龐大，造成在訓練斷詞模型時會消耗大量時間、資源。因此在步驟 3(b)我們不將含有「包含單字詞的詞彙組合」的斷詞組合加入 $Cand_i$，藉此去除含有「包含單字詞的詞彙組合」之斷詞組合。

以下我們以「貼近市場需求，」這一句子為例，對產生句子的各種斷詞組合的步驟

字串 $C_{1:i}$						
C_1:貼	C_2:近	C_3:市	C_4:場	C_5:需	C_6:求	C_7:，

查詢中文辭典中是否包含句子中以 C_i 開頭的不同長度之字串，若包含則將該字串加入 V_i

詞彙集合 V_i						
V_1	V_2	V_3	V_4	V_5	V_6	V_7
貼	近	市	場	需	求	，
貼近		市場		需求		
		市場需求				

圖 5.2 產生「貼近市場需求，」之 V_i

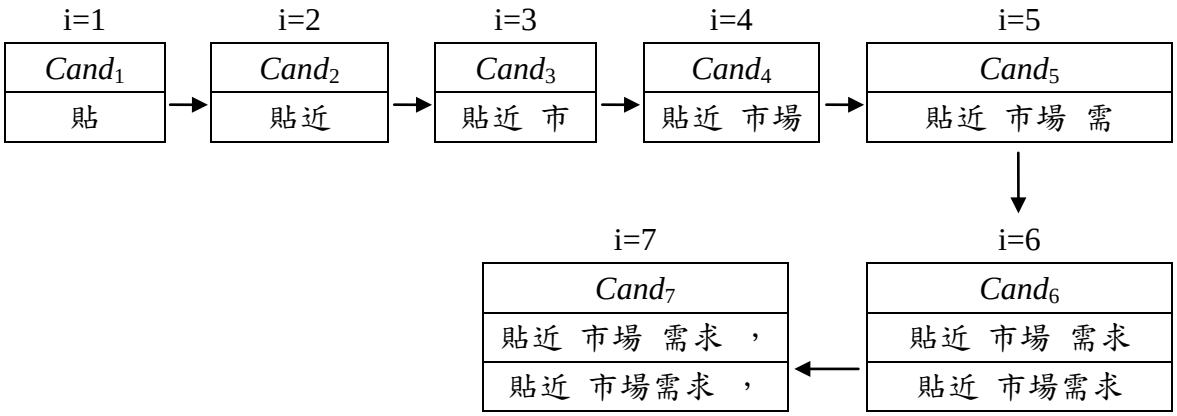

圖 5.3 各階段的 $Cand_i$ 的內容

進行說明。在上頁圖 5.1 中步驟 1，會針對「貼」、「近」、「市」…「，」一一去查詢中文辭典模組的辭典中是否包含句子中以該字開頭的不同長度之字串。若以「貼」為例，會查詢辭典中是否包含「貼」、「貼近」、「貼近市」等字串，若辭典中有包含，則表示該字串為一詞彙，所以該字串會被加入至 V_1；此外若 C_i 為標點符號，我們則把它視為存在於辭典中的單字詞，將其加入至 V_i。最終的 V_i 則如上頁圖 5.2 所示。

在圖 5.1 步驟 3 中的 i 代表不同的階段，而在各個階段會產生字串 $C_{1:i}$ 之各種斷詞組合。在 i 等於 1 時，在步驟 3(a)會檢查 V_1 中是否有詞彙等同於 $C_{1:1}$，而因為 V_1 中的「貼」等同於 $C_{1:1}$，所以會被加入至 $Cand_1$。i 等於 2 時，在步驟 3(a)會查詢 V_1 中是否有詞彙等同於 $C_{1:2}$，而 V_1 中的「貼近」等同於 $C_{1:2}$，所以會被加入至 $Cand_2$；在步驟 3(b)，「貼」加上「近」後會形成「貼 近」，為含有「包含單字詞的詞彙組合」的斷詞組合，所以「貼 近」不會被加入至 $Cand_2$。重複執行步驟 3、步驟 4 到 i 等於 6 時，在步驟 3(b)，$Cand_5$ 中的「貼近 市場 需」加上「求」後會含有「需 求」這個「包含單字詞的詞彙組合」，所以不會被加入至 $Cand_6$；而 $Cand_4$ 中的「貼近 市場」加上 V_5 中的「需求」會等同於 $C_{1:6}$，所以會被加入至 $Cand_6$； $Cand_2$ 中的「貼近」加上 V_3 中的「市場需求」會等同於 $C_{1:6}$，所以也會被加入至 $Cand_6$。重複執行步驟 3、步驟 4 到 i 等於 7，則 $Cand_7$ 內的所有斷詞組合就是句子之各種斷詞組合。圖 5.3 則是各階段的 $Cand_i$ 的內容。

5.2 利用英漢翻譯的資訊處理交集型歧異

在產生句子的各種斷詞組合後，本研究利用英漢翻譯的資訊去處理交集型歧異。我們利用英漢翻譯的資訊去處理交集型歧異的原因為：當一個句子有交集型歧異時，透過英文詞彙的中文翻譯，可以挑選出符合英文陳述的正確斷詞組合。例如有交集型歧異的句子「一旦有機會」可以被斷成「一旦/有機/會」、「一旦/有/機會」，而透過英文詞彙“chance”的中文翻譯「機會」可以挑選出正確的斷詞組合「一旦/有/機會」。挑選出正確的斷詞組合之後，我們會去除錯誤的斷詞組合，以得到較少錯誤的訓練語料。

在對交集型歧異進行處理前，需要先取得英文句中的各個詞彙的中文翻譯集合，而以下說明取得英文句中的各個詞彙的中文翻譯集合的流程。首先我們透過英文句中的空白對英文句進行斷詞，以取得句子中的各個詞彙。英漢辭典模組的各辭典中之英文詞彙大多以原形表示，但英文句中的各個詞彙並不一定為原形，而可能為不同的時態（如過去式、未來式等），或者是複數形態或大寫形態。所以為了讓英文句中的非原形之詞彙也可以對應到辭典中的詞彙，我們利用史丹佛剖析器[5]對英文句中的各英文詞彙進行詞幹還原(lemmatization)後，再到英漢辭典模組中的一般與專業辭典類型的各辭典中查詢該詞彙的中文翻譯。之後取一般辭典類型的各辭典中查詢到的中文翻譯與專業辭典類型的各辭典中查詢到的中文翻譯的聯集，作為該英文詞彙的中文翻譯集合。

以下介紹處理交集型歧異的方法。給定含有交集型歧異字串「ABC」（A、B、C 皆為單一中文字，而「ABC」可以被斷成「A/BC」或「AB/C」）的中文句之各個斷詞組合與該中文句所對應的英文句，我們利用英文句中的各個詞彙之中文翻譯集合的中文翻譯去對應斷詞組合中的中文詞彙；如果某個英文詞彙的中文翻譯集合之中文翻譯對應到斷詞組合中的詞彙 AB，則將包含「AB/C」的斷詞組合視為正確斷詞組合，而包含「A/BC」的斷詞組合則是錯誤斷詞組合，所以我們會去除包含「A/BC」的錯誤斷詞組合。

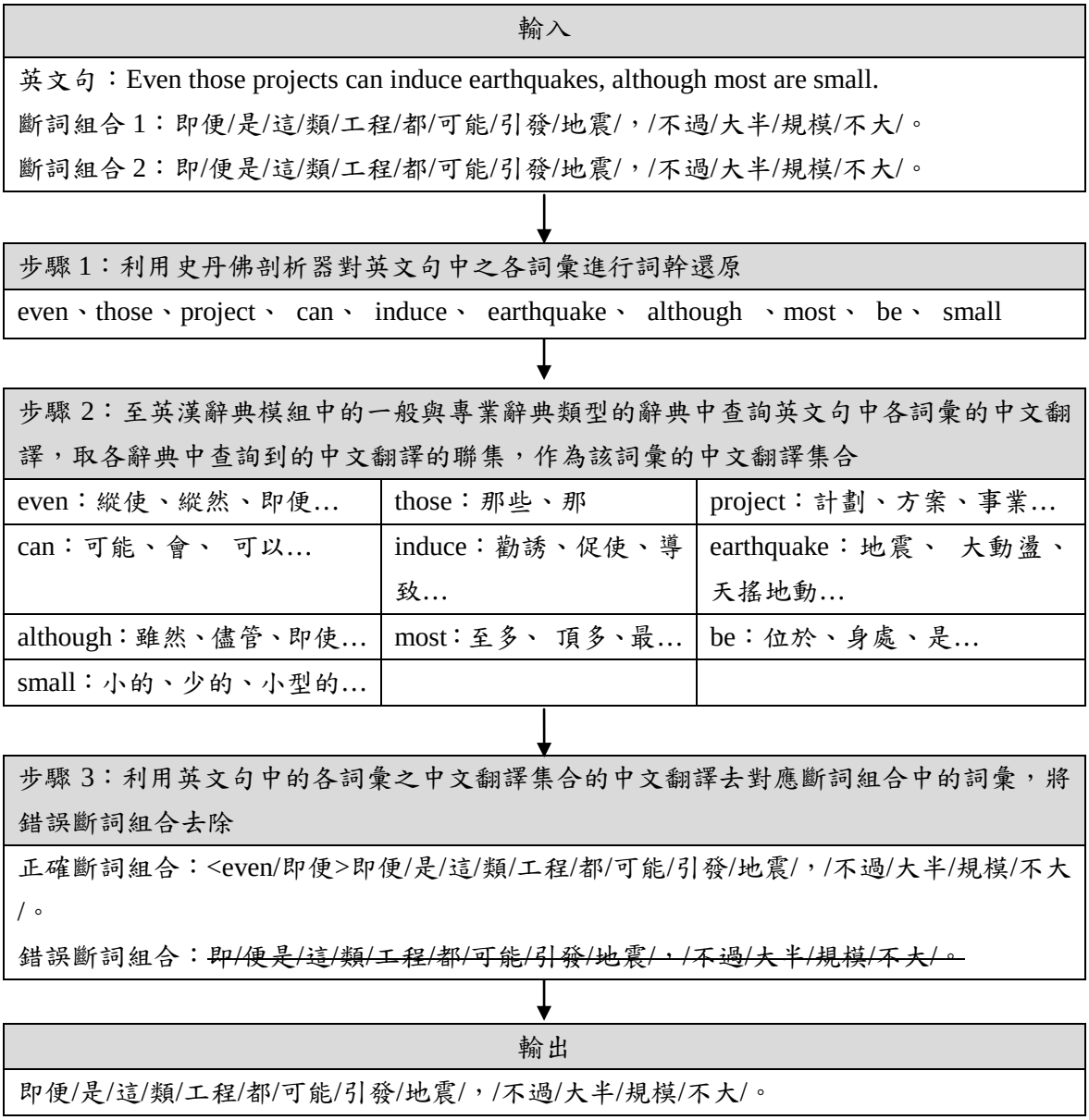

圖 5.4 處理交集型歧異的整體流程

以下藉圖 5.4 說明處理交集型歧異的整體流程。中文句「即便是這類工程都可能引發地震，不過大半規模不大。」包含了交集型歧異字串「即便是」(「即便是」可以被斷成「即便/是」及「即/便是」)，而圖 5.4 中的斷詞組合 1、斷詞組合 2 為該中文句的各個斷詞組合。經過步驟 1 及步驟 2 後，我們取得了英文句中各詞彙的中文翻譯集合，如“even”的中文翻譯集合包含「縱使」、「縱然」、「即便」等詞彙。在步驟 3 正確斷詞組合的部分，我們標記<詞幹還原後的英文詞彙/中文詞彙>的意思是利用左側的詞幹還原後的英文詞彙之中文翻譯集合中的中文翻譯，可以對應到右側的中文詞彙，例如< even/即便>的意思

是 “even”是經過詞幹還原後的詞彙，而“even”的中文翻譯集合中的中文翻譯會對應到「即便」;因為“even”的中文翻譯集合中的中文翻譯可以對應到斷詞組合 1 中的「即便」，所以在步驟 3 我們將包含「即便/是」的斷詞組合視為正確斷詞組合，並去除包含「即/便是」的錯誤斷詞組合。

5.3 擷取中英詞對與未知詞

本研究從中英平行語料中擷取新的中英詞對，以擴充英漢辭典模組的詞對數量，藉此提高利用英漢翻譯資訊處理訓練語料中的交集型歧異之效果，與提升我們的系統之斷詞效能。此外本研究也從中英平行語料中擷取未知詞，藉此處理訓練語料中的未知詞問題，以提升我們的系統之斷詞效能。在擷取中英平行語料中的中英詞對與未知詞時，首先我們會從語料中擷取「候選中英遺留詞對」、「候選中文遺留字詞」。之後對於「候選中英遺留詞對」，我們利用可能性比例(likelihood ratios)與詞對之共現頻率進行篩選；若通過篩選，則將該詞對視為新的中英詞對，擴充至英漢辭典模組中。另外我們也利用詞性序列規則對「候選中文遺留字詞」進行篩選，通過篩選的「候選中文遺留字詞」則視為未知詞，擴充至中文辭典模組中。

本節共分為三個節次，在 5.3.1 節介紹如何從語料中擷取「候選中英遺留詞對」、「候選中文遺留字詞」，在 5.3.2 節詳細說明如何利用可能性比例與詞對之共現頻率對「候選中英遺留詞對」進行篩選，在 5.3.3 節詳細說明如何利用詞性序列規則對「候選中文遺留字詞」進行篩選。

5.3.1 擷取「候選中英遺留詞對」與「候選中文遺留字詞」

我們首先藉由查詢英漢辭典模組的方式來取得英文句的各個詞彙之中文翻譯集合，之後再利用英文句的各個詞彙之中文翻譯集合的中文翻譯對中文句進行斷詞。在斷詞後，中

表 5.1 PAT-tree 抽詞程式所擷取出之結果

擷取出之結果	詞頻
劍橋	10
會不	10
歐斯	10
確的	10
飛利浦	9
火劫學說	8

文句會有未被斷詞的「中文遺留字詞」，英文句會有無法在中文句中找到對應詞彙的「英文遺留字詞」。對於中文句中的所有「中文遺留字詞」，我們使用 PAT-tree 抽詞程式**[35]**進行初步的詞彙擷取。我們發現利用 PAT-tree 抽詞程式所擷取出的結果中，許多錯誤的結果都會含有停用詞，如表 5.1 中的「會不」、「確的」；因此對於以 PAT-tree 抽詞程式所擷取出的結果，我們藉由停用詞列表將其中包含停用詞的結果去除後，我們稱其餘的結果為「候選中文遺留字詞」。由同一平行句對的「候選中文遺留字詞」及「英文遺留字詞」所產生的詞對則稱為「候選中英遺留詞對」。然後因為我們希望得到的是新的中英詞對與未知詞，所以我們去除包含於英漢辭典模組中的辭典之「候選中英遺留詞對」及包含於中文辭典模組中的辭典之「候選中文遺留字詞」。

5.3.2 利用可能性比例與共現頻率進行篩選

因為可能性比例可用於分析兩個詞的關連度[34]，而由較有關連的「候選中文遺留字詞」與「英文遺留字詞」所形成的「候選中英遺留詞對」有較大的機會為正確的中英詞對，所以本研究利用可能性比例對「候選中英遺留詞對」進行篩選。

我們首先對「候選中文遺留字詞」(c) 與「英文遺留字詞」(e) 進行 H1、H2 兩個假設:

$$\text{H1: } P(e|c) = p = P(e|\bar{c}) \tag{6}$$

$$\text{H2: } P(e|c) = p_1 \neq p_2 = P(e|\bar{c}) \tag{7}$$

$$p = \frac{Fe}{N} \tag{8}$$

$$p_1 = \frac{Fce}{Fc} \tag{9}$$

$$p_2 = \frac{Fe - Fce}{N - Fc} \tag{10}$$

H1 表示兩個詞之間是獨立的，H2 表示兩個詞之間是相依的。Fe為在所有英文句中「英文遺留字詞」出現的句數，Fc為在所有中文句中「候選中文遺留字詞」出現的句數，Fce為「候選中英遺留詞對」的共現頻率（共現頻率為候選中英遺留詞對中的中文詞與英文詞共同出現的句對數，而中文詞與英文詞共同出現的意思是：中文詞出現在某平行句對的中文句，且英文詞也出現在該句對的英文句），N為中英平行語料的總句數。

我們利用可能性比例檢驗 H1、 H2；假設機率分佈為 binomial distribution，則 $\mathrm{b}(k, n, x) = \binom{n}{k} x^k (1-x)^{n-k}$。

而可能性比例的公式如下：

$$\text{Likelihood ratio } (c, e) = \log\lambda = \log \frac{\mathrm{b}(Fce, Fc, p)\mathrm{b}(Fe - Fce, N - Fc, p)}{\mathrm{b}(Fce, Fc, p_1)\mathrm{b}(Fe - Fce, N - Fc, p_2)} \tag{11}$$

$$= \log\mathrm{L}(Fce, Fc, p) + \log\mathrm{L}(Fe - Fce, N - Fc, p)$$

$$-\log\mathrm{L}(Fce, Fc, p_1) - \log\mathrm{L}(Fe - Fce, N - Fc, p_2)$$

在公式(11)中，$\mathrm{L}(k, n, x) = x^k (1-x)^{n-k}$。

表 5.2 候選中英遺留詞對之共現頻率與−2logλ對應表

排名	候選中英遺留詞對		共現頻率	−2logλ
1	石墨薄膜	graphene	11	65.154
2	奈米碳管	nanotube	10	55.323
3	線寬	feature	7	27.043
4	波束	beams	7	24.219
5	越高	increase	3	6.230
6	損失	major	1	1.152

我們將信心水準(confidence level)訂為 99.5%，則臨界值(critical value)為 7.88。當−2logλ超過 7.88 時，代表接受 H2，此時「候選中文遺留字詞」與「英文遺留字詞」是有關連的。

除了利用可能性比例做為篩選「候選中英遺留詞對」的條件外，我們也將「候選中英遺留詞對」的共現頻率作為門檻值來對「候選中英遺留詞對」進行篩選。我們將共現頻率作為第一篩選條件，可能性比例檢驗為第二篩選條件，而以下為篩選的大略流程：首先我們會將候選中英遺留詞對依照共現頻率之大小由大到小進行排序，當共現頻率相等時再依照−2logλ之大小由大到小進行排序。而在篩選時，首先判斷該「候選中英遺留詞對」的共現頻率是否大於或等於我們設定的門檻值，若通過會再對該「候選中英遺留詞對」進行可能性比例檢驗，若該詞對之−2logλ超過 7.88，則將該詞對視為正確的詞對，將其篩選出。不過若由某候選中文遺留字詞或某英文遺留字詞所形成的許多詞對都被篩選出的話，則我們只取包含該候選中文遺留字詞或英文遺留字詞的排名最高之詞對。

以下透過表 5.2 說明如何利用可能性比例與共現頻率進行篩選，而表 5.2 中的候選中英遺留詞對已依照上一段落所述方法依序依照共現頻率、−2logλ大小由大到小進行排序。假設將共現頻率的門檻值設為 3，則表 5.2 中的詞對“越高 increase”雖然共現頻率大於或等於 3，但因進行可能性比例檢測後其−2logλ小於 7.88，所以該詞對會被視為錯誤的詞對。而「石墨薄膜 graphene」、「奈米碳管 nanotube」、「線寬 feature」、「波束 beams」

之共現頻率皆大於或等於 3 且進行可能性比例檢測後其−2logλ大於 7.88，所以這 4 個詞彙會被視為新的中英詞對並加入至英漢辭典模組中。

5.3.3 利用詞性序列規則進行篩選

我們觀察了所擷取出的「候選中文遺留字詞」後，發現「候選中文遺留字詞」可分成以下 3 大類：第一類為存在於辭典中的「已知詞」，第二類為不存在於辭典中的「未知詞」，第三類為「不是詞彙的中文字串」，例如「我搶」。中文詞彙通常會擁有特定之構詞結構（如並列式、偏正式等結構[15]），而不是任意地由幾個中文字進行組合就可構成；我們稱由不同詞性之詞素所組成的規則為詞性序列規則，而詞彙之構詞結構可由不同詞性序列規則所構成，例如「名詞　動詞」這個詞性序列規則是由名詞與動詞之詞素組成，而偏正式結構可由「名詞　動詞」所構成。對於辭典中的各個詞彙，本研究設計了一套流程去取得構成辭典詞彙之構詞結構的各個詞性序列規則，之後利用所取得的詞性序列規則去對「候選中文遺留字詞」進行篩選。利用詞性序列規則篩選「候選中文遺留字詞」的原因是：當構成「候選中文遺留字詞」的構詞結構之詞性序列規則符合構成辭典詞彙之構詞結構的詞性序列規則時，表示「候選中文遺留字詞」所擁有的構詞結構符合辭典中的詞彙之構詞結構，因此我們認為該「候選中文遺留字詞」較可能為未知詞，而非「不是詞彙的中文字串」。我們將通過篩選的「候選中文遺留字詞」視為未知詞，將其加入至中文辭典模組，以擴充詞彙數量。

為了利用詞性序列規則去篩選「候選中文遺留字詞」，首先需建立詞性序列規則表。建立詞性序列規則表後，我們利用詞性序列規則的出現次數作為門檻值，並利用通過門檻值的詞性序列規則對「候選中文遺留字詞」進行篩選。

為了取得詞彙的詞性序列規則，需要先將詞彙切割成幾個小單位，再對其標注詞性。而因為斷詞系統遇到未知詞時會將未知詞斷成幾個較小的單位，所以我們藉由去除辭典

1. 將原始中文辭典切割成 N 等份
2. for k =1 to N
3. 將原始中文辭典中的第 k 份去除
4. 利用去除掉第 k 份的中文辭典對語料進行斷詞
5. 利用史丹佛剖析器對已斷詞的語料標注詞性
6. 從語料中取得各詞彙之詞性序列規則，統計各個詞性序列規則的出現次數並記錄於R_k中
7. 合併上述$R_1, R_2,..., R_N$的結果

圖 5.5 建立詞性序列規則表的步驟

的部分詞彙的方式，將這些詞彙當作未知詞；若這些詞彙出現在語料中，則該詞彙經過斷詞處理後會被斷成幾個較小的單位。本研究把由這幾個較小的單位構成的詞彙組合稱為「未知詞候選詞彙組合」。比方說我們將「房地產」由辭典中去除，使其成為未知詞。而「房地產」經過斷詞後被斷成「房地」、「產」兩個小單位，由「房地」、「產」構成的詞彙組合「房地　產」即為「未知詞候選詞彙組合」。

我們透過圖 5.5 之各個步驟來建立詞性序列規則表。在圖 5.5 中步驟 1，我們將 N 取 10，把辭典切割成十等份。以下我們對步驟 3 到 6 進行說明：在第 k 回合，我們將原始中文辭典的第 k 份去除，所以在辭典之第 k 份中的詞彙會被當成未知詞；對語料斷詞後，出現在語料中之第 k 份中的詞彙會被斷成「未知詞候選詞彙組合」。在步驟 5，本研究利用史丹佛剖析器對語料標注詞性，而標注時所使用的字典模型為 xinhuaFactored.ser.gz。對語料標注詞性後，語料中的「未知詞候選詞彙組合」之詞性序列規則即為該詞彙之詞性序列規則。例如「房地　產」經過詞性標注後變為「房地/NN　產/NN」，則「房地產」之詞性序列規則為“ NN　NN ”。不過史丹佛剖析器在不同的語境下，對相同的「未知詞候選詞彙組合」可能會標注不同的詞性，如「房地　產」也可能被標注為「房地/NN　產/VV」，所以一個詞彙的詞性序列規則可能不只一種。在步驟 6 我們對各個經過詞性標注後的未知詞候選詞彙組合(如「房地/NN　產/NN」)進行擷取，就取得各個詞彙之詞性序列規則；而在統計詞性序列規則時，我們將詞彙之可能的各種

表 5.3 詞性序列規則表的內容格式

詞性序列規則	出現次數
NN　NN	6238
VV　NN	3596
AD　VV	3579
VV　M	213
VV　NN　NN	156
AD　NR	55
NN NN VV NN	19
NR NN CD	6

詞性序列規則都納入統計。最後我們將R_1到R_{10}的結果進行合併，就完成詞性序列規則表的建置。

表 5.3 為詞性序列規則表的內容格式；在篩選「候選中文遺留字詞」時我們將詞性序列規則的出現次數做為門檻值，以出現次數大於或等於門檻值的各個詞性序列規則對「候選中文遺留字詞」進行篩選。假設我們將門檻值設為 30，則表 5.3 中用紅色粗體標示的詞性序列規則為出現次數大於或等於門檻值的規則，我們會利用這些詞性序列規則對「候選中文遺留字詞」進行篩選。

下頁圖 5.6 為利用詞性序列規則篩選候選中文遺留字詞之範例。以下我們藉圖 5.6 說明利用詞性序列規則篩選候選中文遺留字詞的整體流程。首先透過中文辭典以長詞優先方式對候選中文遺留字詞進行斷詞，再利用史丹佛剖析器標注詞性，就可取得各個候選中文遺留字詞之詞性序列規則；如果以圖 5.6 中之「前鋒報」為例，因為「前鋒報」經過斷詞、標注詞性後變成「前鋒/NN　報/NN」，所以「前鋒報」之詞性序列規則為「NN NN」。之後我們透過詞性序列規則表中各個詞性序列規則（圖 5.6 中以紅色斜體標示的規則）進行篩選，將詞性標記、空白去除就得到通過篩選之候選中文遺留字詞；例如在圖 5.6 中，透過詞性序列規則表中的詞性序列規則「VV　NN　NN」篩選出「淘/VV　寶/NN　網/NN」、「治/VV　區/NN　主席/NN」之後，將詞性標記、空白去除就得到「淘寶網」、「治區主席」這兩個候選中文遺留字詞，而「淘寶網」為未知詞，治區主席則為

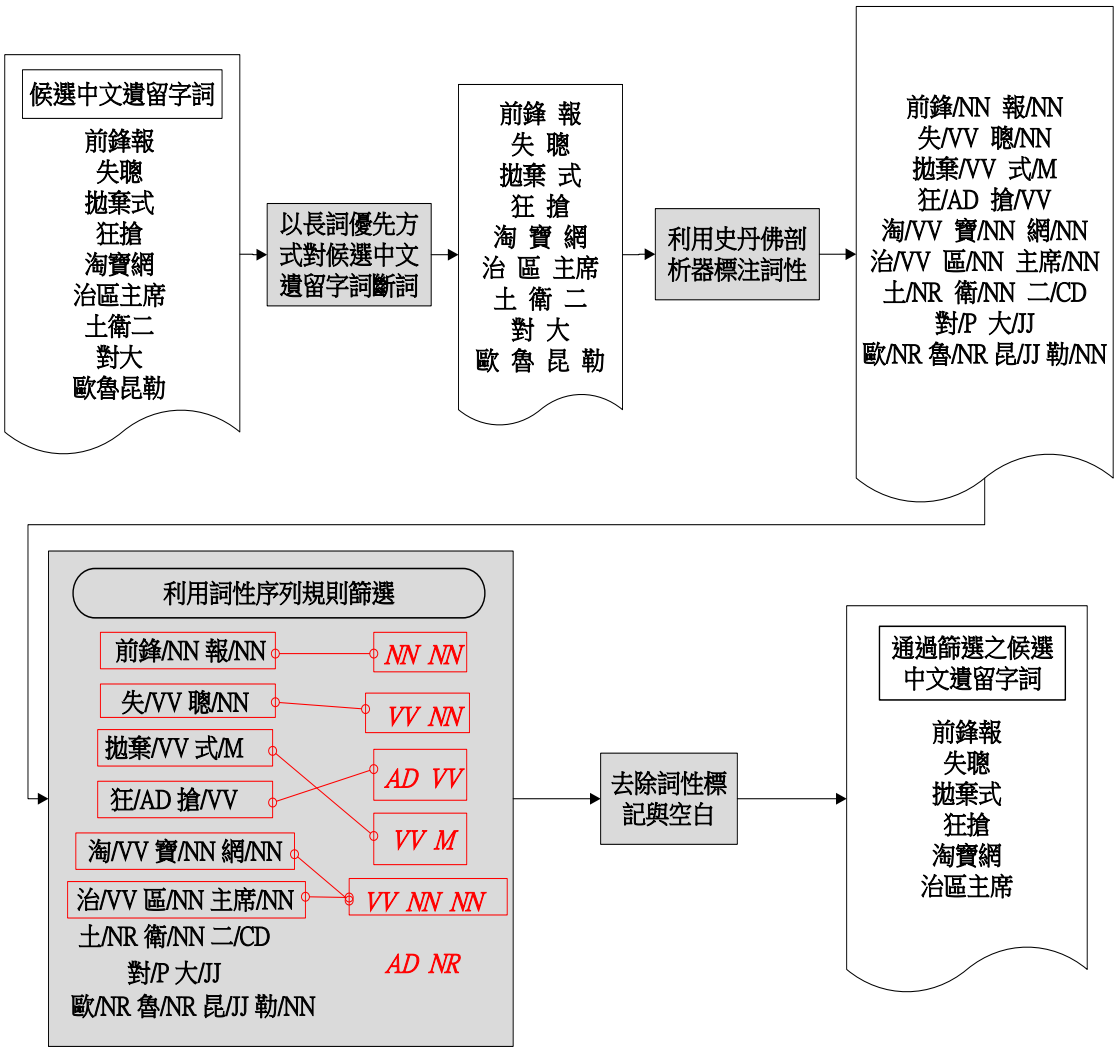

圖 5.6 利用詞性序列規則篩選候選中文遺留字詞之範例

「不是詞彙的中文字串」。最後我們將通過篩選之候選中文遺留字詞視為未知詞加入至中文辭典模組。

第六章　實驗結果與分析

本章共分為 4 個節次；在 6.1 節介紹實驗語料來源，6.2 節介紹擷取中英詞對與未知詞之實驗。然後在 6.3 節、6.4 節的實驗中分別以不同的方式利用不同領域語料去評估本系統之斷詞效能：在 6.3 節中，利用人工斷詞測試語料去評估斷詞效能，6.4 節中藉由漢英翻譯之品質好壞去間接地評估斷詞效能。

6.1 實驗語料來源

本研究使用的實驗語料皆為中英平行語料，而我們根據中英平行語料之中文語料是繁體中文或簡體中文將語料分為兩大類；繁體中文的部分有科學人雜誌中英對照電子書（以下簡稱科學人）以及新聞語料，簡體中文的部分則是有專利平行語料之 C300、C220 與廣播會話(BroadCast Conversation)語料，實驗語料句數統計如表 6.1 所示，而以下將對上

表 6.1 實驗語料句數統計

語料	句數
科學人	63256
新聞語料	54002
C300	296748
C220	222250
廣播會話語料	24351

表 6.2 繁體中文類型的實驗語料之統計

語料	文章數	範圍
科學人	1745	2003 年到 2009 年
自由時報中英對照讀新聞	1420	2005 年 2 月到 2011 年 11 月
雙語網站知識管理平台新聞	737	2005 年 8 月到 2007 年 12 月
聯合新聞網中英對照新聞	519	2008 年 8 月到 2011 年 10 月
美國之音雙語新聞	1299	2005 年 8 月到 2010 年 12 月

表 6.3 Chinese Broadcast Conversation Parallel Text - Part 1、Part 2 語料之統計資料

目錄編號	語料名稱	檔案數量	原始中英句對數	中英句對數(去除短句、重複句對後)
LDC2009T02	GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 1	39	14806	11590
LDC2009T06	GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 2	33	16017	12761

述提到的語料的來源及我們對語料所做的處理進行說明。

田侃文[4]於 2009 年使用英漢文句對列技術，將科學人之 1745 篇文章轉換成 63256 句中英平行句對，而我們將直接沿用這 63256 句平行句對進行實驗。我們將自由時報中英對照讀新聞、雙語網站知識管理平台新聞、美國之音雙語新聞及聯合新聞網中英對照新聞這四種英漢雙語語料，利用英漢文句對列技術[4]轉換成中英平行句對後進行合併，就得到新聞語料。而繁體中文類型的實驗語料之統計如上頁表 6.2 所示。

Tseng 等人[39]於Patent Machine Translation Task at the NTCIR-9[36]（以下簡稱 NTCIR-9 PatentMT）時對原始 100 萬句專利平行語料進行了前處理後得到兩種不同的英漢雙語訓練語料C300[1]、C220，我們直接沿用這兩種語料進行實驗。我們從所購買的 Linguistic Data Consortium之GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 1 語料、GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 2 語料的原始檔案中擷取中英平行句對，並將短句(長度小於 6 之句子)、重複出現的句對及句中一些特殊符號去除，而表 6.3 為GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 1 、Part 2 語料之統計資料。最後我們將GALE Phase 1 Chinese Broadcast Conversation

[1] 雖然在[39]中並沒有記錄對 C1140 的進一步處理，但在 NTCIR-9 Patent MT 時 Tseng 等人得到 C1140 後還有對 C1140 作進一步地篩選，再利用篩選完的語料進行實驗。而 C1140 經篩選後約剩 30 萬句，所以在本研究中以 C300 代替 C1140。

Parallel Text - Part 1 、Part 2 語料之去除重複句對、短句後的中英平行句對進行合併就得到廣播會話語料。

6.2 擷取中英詞對與未知詞之實驗

本研究為了評估透過 5.3 節所述的方法去擷取中英詞對與擷取未知詞的效果，將分別於 6.2.1 節、6.2.2 節中介紹擷取中英詞對之實驗及擷取未知詞之實驗。

6.2.1 擷取中英詞對之實驗

在本實驗中我們從科學人、新聞語料、廣播會話語料、C300、C220 中擷取中英詞對，並評估其效果。我們首先從各語料中擷取候選中英遺留詞對，而所擷取出的候選中英遺留詞對之數量如表 6.4 所示。之後我們依照 5.3.2 節所述之方法透過可能性比例與共現頻率對候選中英遺留詞對進行篩選，並利用人工的方式去檢測以不同的共現頻率作為門檻值所篩選出的結果：在科學人、新聞語料、廣播會話語料的部分，我們對篩選出的不同共現頻率之所有候選中英遺留詞對都進行人工檢測，但在 C300、C220 的部分，因為篩選出的共現頻率為 2、共現頻率為 1 的候選中英遺留詞對數量皆在數千以上，所以對於共現頻率為 2、共現頻率為 1 的候選中英遺留詞對，我們從每 100 名中取前 50 名進行檢測。我們使用精確率（Precision）、召回率（Recall）、F_1-measure 三個評估指標進行評估，各評估指標的定義如下頁公式(12)-(14)所示。

表 6.4 候選中英遺留詞對數量統計

語料名稱	候選中英遺留詞對數量
科學人	5410
新聞語料	3502
廣播會話語料	831
C300	9326
C220	7798

$$精確率 = \frac{篩選出的正確候選中英遺留詞對之數量}{篩選出的候選中英遺留詞對之數量} \tag{12}$$

$$召回率 = \frac{篩選出的正確候選中英遺留詞對之數量}{所有正確候選中英遺留詞對之數量} \tag{13}$$

$$\mathrm{F_1-measure} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{14}$$

表 6.5 以不同的共現頻率作為門檻值之篩選結果（新聞語料）

門檻值(共現頻率)	P	R	F_1
1	0.314	1.000	0.478
2	0.347	0.874	0.497
3	0.403	0.659	0.500
4	0.393	0.544	0.456
5	0.684	0.317	0.434

表 6.6 以不同的共現頻率作為門檻值之篩選結果（科學人）

門檻值(共現頻率)	P	R	F_1
1	0.337	1.000	0.504
2	0.368	0.703	0.483
3	0.782	0.422	0.548
4	0.826	0.265	0.401
5	0.856	0.187	0.306

表 6.7 以不同的共現頻率作為門檻值之篩選結果（廣播會話語料）

門檻值(共現頻率)	P	R	F_1
1	0.416	1.000	0.588
2	0.579	0.751	0.654
3	0.747	0.468	0.575
4	0.821	0.318	0.458
5	0.841	0.260	0.397

表 6.8 以不同的共現頻率作為門檻值之篩選結果（C300）

門檻值(共現頻率)	P	R	F_1
1	0.347	1.000	0.516
2	0.541	0.797	0.644
3	0.697	0.714	0.706
4	0.757	0.576	0.654
5	0.749	0.447	0.560

表 6.9 以不同的共現頻率作為門檻值之篩選結果（C220）

門檻值(共現頻率)	P	R	F_1
1	0.253	1.000	0.404
2	0.415	0.899	0.567
3	0.512	0.772	0.616
4	0.566	0.573	0.569
5	0.602	0.460	0.521

上頁之表 6.5、表 6.6、表 6.7 與表 6.8、表 6.9 分別是以不同的共現頻率作為門檻值去對新聞語料、科學人、廣播會話語料、C300、C220 之候選中英遺留詞對進行篩選的結果，各表中之 P 代表精確率，R 代表召回率，F_1 代表 F_1-measure。如表 6.5、表 6.6、表 6.8、表 6.9 數據所示，在新聞語料、科學人、C300、C220 部分，F_1-measure 最高的都是門檻值為 3 之結果，我們分別把這四種語料之以門檻值為 3 所篩選出的候選中英遺留詞對加入至英漢辭典模組。如表 6.7 數據所示，在廣播會話語料部分，F_1-measure 最高的是門檻值為 2 之結果，所以我們把廣播會話語料之以門檻值為 2 所篩選出的候選中英遺留詞對加入至英漢辭典模組。下頁表 6.10 為上述提到的各語料之被加入至英漢辭典模組的候選中英遺留詞對。

由表 6.5、表 6.6、表 6.7、表 6.8、表 6.9 中數據可看出，在各實驗語料的結果中，雖然以共現頻率 1、2 作為門檻值所得到的結果可以得到不錯的召回率，但因為所篩選出的共現頻率為 1、共現頻率為 2 的候選中英遺留詞對大部分都不是正確的中英詞對，所以造成無法有好的精確率；而在部分實驗語料的結果中（科學人、廣播會話語料、

表 6.10 被加入至英漢辭典模組的各語料之候選中英遺留詞對

科學人之篩選出的候選中英遺留詞對
哈伯 Hubble、通常 typically、樹突 dendritic、目前 current、蛋白 protein、普恩蛋白 prion、首次 first、暴脹 inflation、史丹佛 Stanford、訊息 signal、造影 imaging、關鍵 crucial、探測車 rover、永續 sustainable、衰變 decay …
新聞語料之篩選出的候選中英遺留詞對
胡錦濤 Hu、新華 Xinhua、歐盟 EU、微軟 Microsoft、臉書 Facebook、北約 NATO 恐怖 terrorist、奧運 Olympic、德州 Texas、總理 Prime、史丹福 Stanford、塔利班 Taliban、援引 quote、歐巴馬 Obama、馬英九 Ma …
廣播會話語料之篩選出的候選中英遺留詞對
宋楚瑜 James、登輝 Lee、大家 everyone、衛視 Satellite、馨田 xintian、民進黨 dpp 哈馬斯 Hamas、擦鞋 shoeshine、國民黨 kmt、雲林 yunlin、雙劍 shuangjian、角度 perspective、寨子 Zhaizi、剛才 earlier、京廣 jingguang …
C300 之篩選出的候選中英遺留詞對
情況 case、蜂窩 cellular、相應 respective、編程 program、微粒 microparticle、治療 therapeutically、介導 mediate、側壁 sidewall、標識 identify、直鏈 linear 寡核甘酸 oligonucleotide、電機 motor、映射 map、具體 particular…
C220 之篩選出的候選中英遺留詞對
本文 herein、標識 identification、相應 respective、引物 primer、尋呼 paging、碼元 symbol、外周 peripheral、轉染 transfect、制備 produce、藥物 agent、市售 commercially、物理 physically、反義 antisense、升高 increase…

C300)，以共現頻率 4、5 作為門檻值所得到的結果則可以得到不錯的精確率，但因為篩選出的正確候選中英遺留詞對過少，使得召回率低下。

6.2.2 擷取未知詞之實驗

在本實驗中我們從科學人、新聞語料、廣播會話語料、C300、C220 中擷取未知詞，並評估其效果。我們首先從各語料中擷取候選中文遺留字詞，而所擷取到的候選中文遺留字詞數量如下頁表 6.11 所示。之後對於各種實驗語料，依照 5.3.3 節所述方法建立詞性

表 6.11 候選中文遺留字詞數量統計

語料名稱	候選中文遺留字詞數量
科學人	2484
新聞語料	2475
廣播會話語料	356
C300	4619
C220	3469

序列規則表。之後利用詞性序列規則的出現次數做為門檻值來取得不同的詞性序列規則，再透過所取得各個詞性序列規則對候選中文遺留字詞進行篩選。對於所有候選中文遺留字詞，我們透過人工的方式檢測其是否為未知詞，並藉由精確率（Precision）、召回率（Recall）、F_1-measure 三個評估指標來評估透過詞性序列規則對候選中文遺留字詞進行篩選的效果，而公式(15)-(17)為各評估指標的定義。

$$精確率 = \frac{篩選出的正確候選中文遺留字詞之數量}{篩選出的候選中文遺留字詞之數量} \tag{15}$$

$$召回率 = \frac{篩選出的正確候選中文遺留字詞之數量}{所有正確候選中文遺留字詞之數量} \tag{16}$$

$$F_1\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

表 6.12、表 6.13、表 6.14 與下頁表 6.15、表 6.16 分別為以詞性序列規則出現次數作為門檻值，並利用通過不同門檻值之各個詞性序列規則去對新聞語料、科學人、廣播會話語料、C300、C220 之候選中文遺留字詞進行篩選所得到的結果，各個表中的 P 代表精確率，R 代表召回率，F_1 代表 F_1-measure。由表 6.12 的數據可看出，在新聞語料的部分，門檻值為 5 或 10 時有相同的F_1-measure，如此會遇到召回率與精確率的取捨

表 6.12 以通過不同門檻值之詞性序列規則進行篩選的結果（新聞語料）

門檻值 (出現次數)	P	R	F_1
5	0.621	0.915	0.745
10	0.625	0.907	0.745
15	0.625	0.896	0.742
20	0.624	0.890	0.741
25	0.625	0.885	0.740
30	0.630	0.883	0.741

表 6.13 以通過不同門檻值之詞性序列規則進行篩選的結果（科學人）

門檻值 (出現次數)	P	R	F_1
5	0.632	0.943	0.756
10	0.631	0.937	0.754
15	0.631	0.928	0.751
20	0.631	0.921	0.749
25	0.629	0.907	0.743
30	0.627	0.894	0.737

表 6.14 以通過不同門檻值之詞性序列規則進行篩選的結果（廣播會話語料）

門檻值 (出現次數)	P	R	F_1
5	0.722	0.837	0.775
10	0.724	0.812	0.765
15	0.725	0.808	0.764
20	0.722	0.784	0.751
25	0.720	0.776	0.747
30	0.722	0.776	0.748

表 6.15 以通過不同門檻值之詞性序列規則進行篩選的結果（C300）

門檻值 (出現次數)	P	R	F_1
5	0.627	0.822	0.712
10	0.625	0.804	0.703
15	0.624	0.793	0.698
20	0.625	0.787	0.696
25	0.621	0.773	0.689
30	0.622	0.770	0.688

表 6.16 以通過不同門檻值之詞性序列規則進行篩選的結果（C220）

門檻值 (出現次數)	P	R	F_1
5	0.645	0.713	0.677
10	0.644	0.704	0.673
15	0.643	0.695	0.668
20	0.643	0.690	0.666
25	0.643	0.685	0.663
30	0.643	0.682	0.662

(trade-off)問題，而因為我們希望能取得較多正確候選中文遺留字詞，所以我們取召回率較高的門檻值為 5 之結果，並把新聞語料之以門檻值為 5 所篩選出的候選中文遺留字詞加入至中文辭典模組。如上頁表 6.13、表 6.14 與表 6.15、表 6.16 結果所示，在科學人、廣播會話語料、C300、C220 部分，F_1–measure 最高的是門檻值為 5 之結果，故我們分別把這四種語料之以門檻值為 5 所篩選出的候選中文遺留字詞加入至中文辭典模組。下頁表 6.17 為上述提到的被加入至中文辭典模組的各語料之候選中文遺留字詞。

由表 6.12、表 6.13、表 6.14、表 6.15、表 6.16 可以發現，除了廣播會話語料的結果外，在其他實驗語料的結果中，召回率會隨著門檻值的提升而逐漸下降，這是因為隨著門檻值的提升，可以用來篩選的詞性序列規則就越少，使得篩選出的正確候選中文遺

表 6.17 被加入至中文辭典模組的各語料之候選中文遺留字詞

科學人之篩選出的候選中文遺留字詞
三維流形、世記、丙泊酚、中性伴子、丹納基、乘員、乙基汞、九維、乳突病毒亞琛、亞裔、亞馬遜河流、傷齒龍、優缺點、優先主義、內布拉、冰釘、匿名認證、卡羅萊納、卡西佛隕石坑、反演化論、單株、地理定位、地理藏寶 …
新聞語料之篩選出的候選中文遺留字詞
七宗罪、世博會、世衛、中寫、丹尼葉、主業會、亞特蘭蒂斯號、伍大偉、休佛佛瑞斯特、佛瑞西尼、克欽族、兩伊、冬奧會、前首、前鋒報、台糖、台鐵、台長艾德、地理紀錄器、塔塔汽車、夏姆席克、姚福信、密德塞克斯、寶萊塢 …
廣播會話語料之篩選出的候選中文遺留字詞
京廣、什葉派、低度管理、冉廣岐、勞動密集型、勢氣、北影、千手觀音、半厘反分裂、外患罪、天蟬、天雲山、女配角、宗記社、寫字樓、岳忠、崇奐、崇明徐祖遠、炒房團、瑞芳、皇宮博物館、直通車、祝希娟、葉大鷹、趙少威 …
C300 之篩選出的候選中文遺留字詞
丁氧環酮、丁羥甲苯、丁基橡膠、三環類、三線態、三芳基琉、主題詞、乙基纖維、乳果糖、代謝拮抗劑、倒立槽、假馬齒莧、偶聯、偽像、傅利葉、傍軸、像差儀、免疫球蛋白、免疫療法、內窺鏡、內核空間、分枝桿菌、分流翼 …
C220 之篩選出的候選中文遺留字詞
乳酸桿菌、乾酪乳桿菌、交叉反應、交感神經、伴侶蛋白、低密度脂蛋白、低熔點、假單胞菌、傳輸功率、光刻工藝、光熱敏、助粘劑、勃母石、卡波西、卡維地洛、反轉錄病毒、吉非貝、吉姆薩、啤酒酵母、噬菌粒、增容 …

留字詞數量下降。而在表 6.12、表 6.13、表 6.14、表 6.15、表 6.16 的結果中，在不同門檻值下所得到的各個精確率之間的最大差距只有 0.009(新聞語料的結果中的門檻值 5、30)，由此可看出不管是哪一種實驗語料，精確率在不同的門檻值下都有相近的水準。

6.3 以人工斷詞測試語料評估斷詞效能之實驗

在本節中主要介紹利用不同領域之人工斷詞測試語料去評估我們的系統的斷詞效能之實驗。在 6.3.1 節說明整體實驗流程，6.3.2 節分析與討論實驗結果。

6.3.1 實驗流程設計

我們將於本實驗中使用科學文章類型的科學人、新聞文章類型的新聞語料、會話文章類型的廣播會話語料這三種不同領域的實驗語料。在本實驗，我們從實驗語料中抽取出兩百句當作測試語料，實驗語料的其餘部分提供給我們的系統去產生訓練語料。由於科學人、新聞語料、廣播會話語料的測試語料都是直接由原始中英平行語料中切割而來，所以我們手邊沒有測試語料之斷詞標準答案。因此我們對兩百句測試語料進行人工斷詞，並以人工斷詞的結果當作正確的斷詞標準答案，以進行斷詞效能的評估。

本實驗之產生訓練語料的方式如圖 6.1 所示，由有或沒有利用英漢翻譯的資訊去處理交集型歧異之兩種情況去與有或沒有加入未知詞及中英詞對之兩種情況進行組合，故最後有 4 種產生訓練語料的方式。訓練斷詞模型的工具則是有 LingPipe 中文斷詞器(以下簡稱為 LPS)以及史丹佛中文斷詞器(以下簡稱為 SCS) 。

除了評估我們的系統之整體斷詞效能之外，我們會在 6.3.2 節分析透過本研究提出的加入未知詞及中英詞對與利用英漢翻譯的資訊去處理交集型歧異的方法能否提升斷詞效能。此外為了評估在訓練斷詞模型時加入外部辭典對斷詞效能的影響，我們將分別就訓練斷詞模型時加入辭典與訓練斷詞模型時未加入辭典這兩種類型去進行實驗，而在

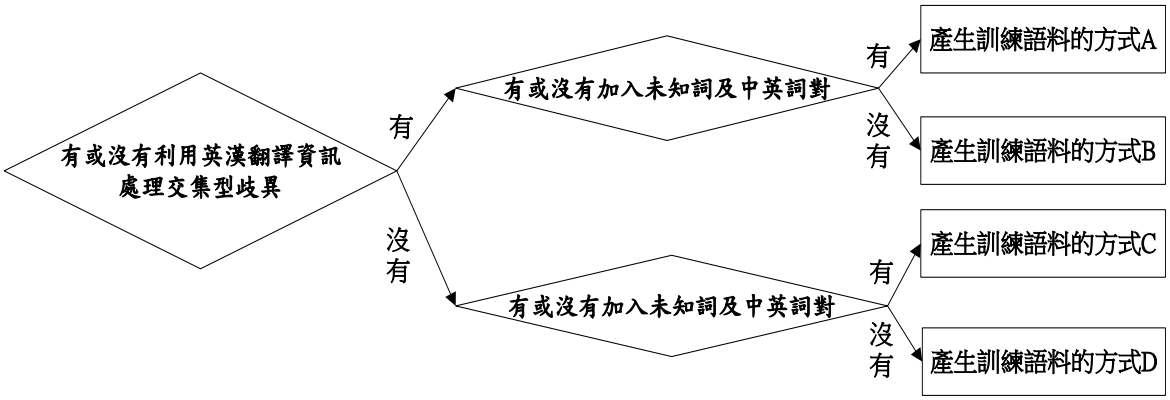

圖 6.1 產生訓練語料之方式

訓練斷詞模型時所加入的辭典包含了中文辭典模組中的所有辭典。

為了比較我們的系統與其他斷詞系統或斷詞模型間的斷詞效能差異，我們將中研院斷詞系統[2]與 Yahoo 開發的斷章取義斷詞系統[41]、SCS 之 Pku 及 Ctb 斷詞模型、ICTCLAS 漢語分詞系統[27]（以下簡稱 ICTCLAS）作為我們的系統之比較的對象。

在斷詞效能評估方面，我們共使用了精確率（Precision）、召回率（Recall）、F_1-measure 三個評估指標去評估斷詞效能，而在下頁表 6.18 中的 P 代表精確率，R 代表召回率，F_1 代表 F_1-measure；以下為評估指標的個別定義。

$$\text{精確率} = \frac{\text{系統斷出的正確詞數}}{\text{系統斷出的詞數}} \tag{18}$$

$$\text{召回率} = \frac{\text{系統斷出的正確詞數}}{\text{參考答案中的所有詞數}} \tag{19}$$

$$F_1\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{20}$$

表 6.18 不同領域語料之斷詞效能

訓練斷詞模型時未加入辭典												
訓練工具	加入未知詞與中英詞對	利用英漢翻譯資訊處理交集型歧異	廣播會話語料			科學人			新聞語料			
			P	R	F_1	P	R	F_1	P	R	F_1	
LPS	沒有	沒有	0.776	0.809	0.792	0.793	0.834	0.813	0.724	0.801	0.761	
		有	0.788	0.818	0.803	0.806	0.843	0.824	0.727	0.803	0.763	
	有	沒有	0.778	0.810	0.794	0.797	0.834	0.815	0.732	0.801	0.765	
		有	0.792	0.820	0.806	0.815	0.847	0.831	0.737	0.803	0.769	
SCS	沒有	沒有	0.792	0.827	0.809	0.762	0.897	0.824	0.679	0.863	0.760	
		有	0.808	0.842	0.825	0.781	0.909	0.840	0.689	0.871	0.769	
	有	沒有	0.801	0.832	0.816	0.778	0.906	0.837	0.681	0.864	0.762	
		有	0.812	0.843	0.827	0.799	0.919	***0.855***	0.710	0.883	***0.787***	
訓練斷詞模型時加入辭典												
訓練工具	加入未知詞與中英詞對	利用英漢翻譯資訊處理交集型歧異	廣播會話語料			科學人			新聞語料			
			P	R	F_1	P	R	F_1	P	R	F_1	
LPS	沒有	沒有	0.819	0.791	0.805	0.792	0.805	0.798	0.742	0.786	0.764	
		有	0.834	0.805	0.820	0.820	0.828	0.824	0.749	0.793	0.770	
	有	沒有	0.818	0.790	0.804	0.797	0.805	0.801	0.753	0.784	0.768	
		有	0.836	0.806	0.821	0.819	0.822	0.820	0.762	0.794	0.778	
SCS	沒有	沒有	0.802	0.832	0.817	0.772	0.818	0.794	0.681	0.863	0.762	
		有	0.823	0.851	***0.837***	0.792	0.834	0.812	0.688	0.870	0.768	
	有	沒有	0.796	0.826	0.811	0.784	0.822	0.802	0.682	0.864	0.763	
		有	0.819	0.845	0.832	0.790	0.830	0.810	0.705	0.880	0.783	
其他斷詞系統或斷詞模型												
			廣播會話語料			科學人			新聞語料			
			P	R	F_1	P	R	F_1	P	R	F_1	
中研院斷詞系統			–	–	–	0.878	0.932	***0.904***	0.854	0.923	***0.887***	
斷章取義斷詞系統			–	–	–	0.754	0.739	0.746	0.743	0.753	0.748	
SCS 之 Pku 斷詞模型			0.870	0.884	***0.877***	0.839	0.867	0.852	0.815	0.853	0.834	
SCS 之 Ctb 斷詞模型			0.846	0.869	0.857	0.827	0.868	0.847	0.832	0.878	0.855	
ICTCLAS			0.849	0.887	0.868	0.785	0.841	0.812	0.758	0.848	0.801	

6.3.2 實驗結果與分析

上頁表 6.18 為本系統對不同領域語料之斷詞效能。此外我們也將表 6.18 之各結果的精確率、召回率的數據，改成用系統斷出的詞數、系統斷出的正確詞數、參考答案的所有詞數來表示，並將結果收錄於附錄中。表 6.18 中之加入未知詞與中英詞對欄位表示產生訓練語料時是否加入未知詞及中英詞對，利用英漢翻譯資訊處理交集型歧異欄位表示產生訓練語料時是否利用英漢翻譯的資訊去處理交集型歧異。在表 6.18 各個實驗語料的實驗數據中，我們將我們的系統之最高 F_1-measure 與其他的斷詞系統或斷詞模型中的最高 F_1-measure 用紅色粗體加斜體標示。因為從 LDC 購買的廣播會話語料有版權問題，所以我們並沒有利用中研院斷詞系統、斷章取義斷詞系統對其進行斷詞，而在表 6.18 廣播會話語料之中研院斷詞系統、斷章取義斷詞系統結果部分我們則將其標示為「-」。

以下為我們的系統與其他的斷詞系統或斷詞模型的斷詞效能比較。在表 6.18 科學人部分，我們的系統的最高 F_1-measure 為 0.855，高於 SCS 之 Pku、Ctb 斷詞模型、ICTCLAS、斷章取義斷詞系統之 F_1-measure，比起斷詞效能最佳的中研院斷詞系統之 F_1-measure 低了 0.049。在新聞語料部分，我們系統的最高 F_1-measure 為 0.787，比起斷詞效能最佳的中研院斷詞系統之 F_1-measure 低了 0.1，但高於斷章取義斷詞系統之 F_1-measure。在廣播會話語料的部分，我們系統的最高 F_1-measure 為 0.837，低於 SCS 之 Pku、Ctb 斷詞模型、ICTCLAS 之 F_1-measure，但我們的系統與 Pku、Ctb 斷詞模型、ICTCLAS 的 F_1-measure 之差距皆在 0.04 以內。

由以上分析可看出，在三種實驗語料的結果中，我們的系統之最佳斷詞效能都無法優於所有其他的斷詞系統或斷詞模型之斷詞效能。但在科學人、廣播會話語料部分，我們的系統之最高 F_1-measure 與斷詞效能最佳的其他斷詞系統或斷詞模型之 F_1-measure 的差距都在 0.05 以內，且我們的系統之最高 F_1-measure 都在 0.835 以上，因此我們覺得這顯示了我們的系統能夠有一定的斷詞效能。

表 6.19 未利用與利用英漢翻譯資訊處理交集型歧異所得到之斷詞結果

正確斷詞結果：但/其他/人/還有/疑慮/。 正確斷詞結果：笑/的/神經/迴路/存在/於/大腦/的/非常/古老/區域/，	
未利用英漢翻譯資訊處理交集型歧異所得到之斷詞結果	利用英漢翻譯資訊處理交集型歧異所得到之斷詞結果
但/其/他人/還有/疑慮/。	但/其他/人/還有/疑慮/。
笑/的/神經/迴路/存/在於/大腦/的/非常/古老/區域/，	笑/的/神經/迴路/存在/於/大腦/的/非常/古老/區域/，

在表 6.18 的結果中，不論訓練斷詞模型時加入辭典或未加入辭典，在科學人、新聞語料、廣播會話語料部分，比起沒有利用英漢翻譯資訊處理交集型歧異的實驗結果之 F_1-measure，有利用英漢翻譯資訊處理交集型歧異的實驗結果之 F_1-measure 皆能提升，而其中 F_1-measure 提升最多的實驗結果為訓練斷詞模型時未加入辭典的情況下，新聞語料部分之利用 SCS 訓練斷詞模型，且有加入未知詞及中英詞對的結果（F_1-measure 由 0.762 提升至 0.787）。因此我們覺得這顯示了與沒有利用英漢翻譯資訊處理交集型歧異相比，有利用英漢翻譯資訊處理交集型歧異應能夠使斷詞效能提升。表 6.19 則為測試語料中的一些未利用與利用英漢翻譯資訊處理交集型歧異所得到之斷詞結果，而表 6.19 中的兩句句子在利用英漢翻譯資訊處理交集型歧異後都可變成正確斷詞結果。

由表 6.18 可看出，在訓練斷詞模型時未加入辭典的情況下，不論實驗語料是科學人或新聞語料還是廣播會話語料，比起沒有加入未知詞與中英詞對的實驗結果之 F_1-measure，有加入未知詞與中英詞對的實驗結果之 F_1-measure 皆能提升，其中 F_1-measure 提升最多的為新聞語料部分之利用 SCS 訓練斷詞模型，且有利用英漢翻譯資訊處理交集型歧異的結果（F_1-measure 由 0.769 提升至 0.787）。因此我們覺得這顯示了在訓練斷詞模型時未加入辭典的情況下，有加入未知詞與中英詞對應可以對斷詞效能的提升有一定的幫助。下頁表 6.20 則為測試語料中的一些未加入與加入未知詞與中英詞對所得到之斷詞結果，而表 6.20 中的兩句句子在加入未知詞與中英詞對後都可變成正確斷

表 6.20 未加入與加入未知詞與中英詞對所得到之斷詞結果

正確斷詞結果：他們/想要/或/不會/反對/有/一個/像/史達林/那樣/的/領袖/在/今日/領導/國家/。 正確斷詞結果：大嶼山/由此/開始/成/為/新港/旅遊/新/熱點/。	
未加入未知詞與中英詞對所得到之斷詞結果	加入未知詞與中英詞對所得到之斷詞結果
他們/想要/或/不會/反對/有/一個/像/史/達林/那樣/的/領袖/在/今日/領導/國家/。	他們/想要/或/不會/反對/有/一個/像/史達林/那樣/的/領袖/在/今日/領導/國家/。
大嶼/山/由此/開始/成/為/新/港/旅遊/新/熱點/。	大嶼山/由此/開始/成/為/新港/旅遊/新/熱點/。

詞結果。但我們發現在訓練斷詞模型時加入辭典的情況下卻不是所有的有加入未知詞與中英詞對的實驗結果之 F_1-measure 皆高於沒有加入未知詞與中英詞對的實驗結果之 F_1-measure，而以下我們將討論造成此結果的可能原因。

因為我們所加入的候選中文遺留字詞可能包含「不是詞彙的中文字串」，所以加入未知詞與中英詞對後會有以下兩種情況發生，情況 1：在未加入未知詞與中英詞對時所得到的錯誤斷詞結果，在加入未知詞與中英詞對後變成正確的斷詞結果；情況 2：在未加入未知詞與中英詞對時所得到的正確斷詞結果，在加入未知詞與中英詞對後變成錯誤的斷詞結果。分析斷詞結果後，我們覺得造成在訓練斷詞模型時加入辭典的情況下，加入未知詞與中英詞對的實驗結果之 F_1-measure 低於沒有加入未知詞與中英詞對的實驗結果之 F_1-measure 的可能原因如下：原本在訓練斷詞模型時未加入辭典的情況下，情況 1 的出現次數大於情況 2 的出現次數，則斷詞效能有所提升；但在訓練斷詞模型時加入辭典的情況下，因為某些句子會出現下頁圖 6.2 之加入的未知詞與辭典詞彙衝突的情況（即未加入未知詞與中英詞對時會得到錯誤斷詞結果的中文句，在訓練斷詞模型時未加入辭典的情況下，加入未知詞（「旅遊年」包含在所加入的未知詞中）與中英詞對後會得到正確斷詞結果，但在訓練斷詞模型時加入辭典的情況下，因為受到辭典中的詞彙「旅遊」、「年」影響，使得加入未知詞與中英詞對後卻不會得到正確斷詞結果），造成情況 1 的出現次數反而小於情況 2 的出現次數，因此使得斷詞效能下降。

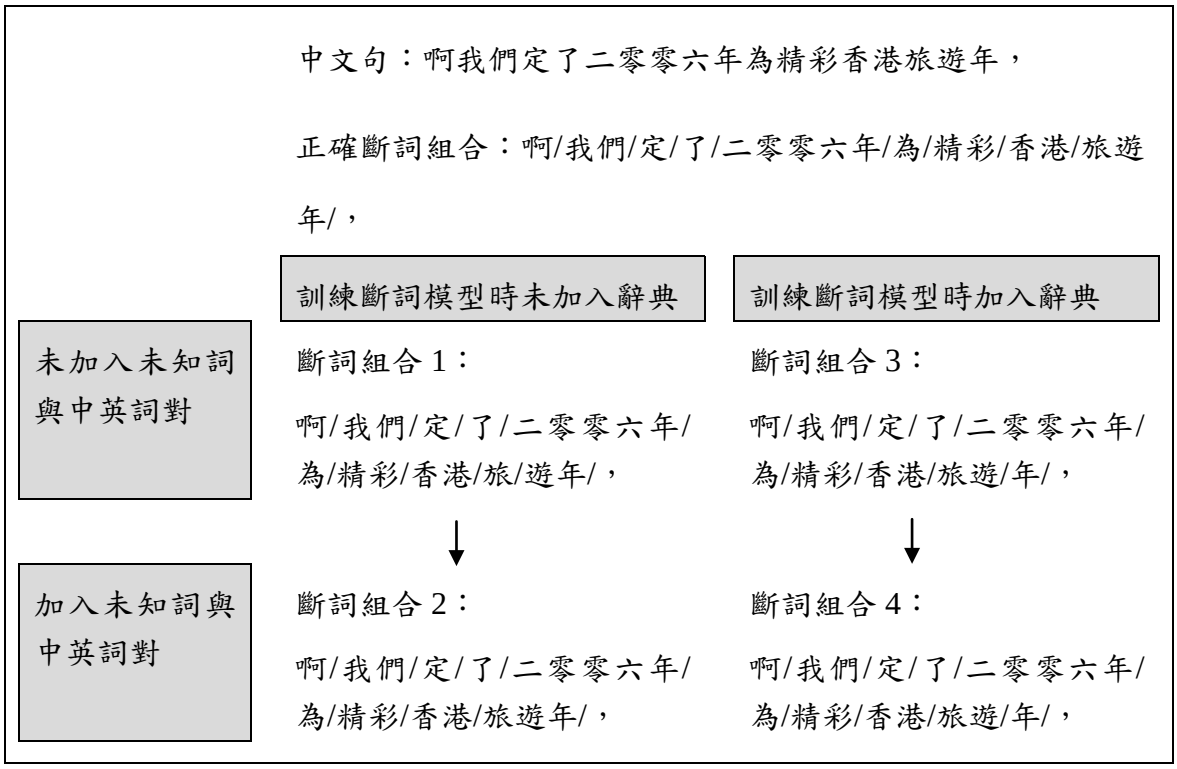

圖 6.2 加入的未知詞與辭典詞彙衝突的情況下對斷詞結果之影響

在以下我們比較訓練斷詞模型時加入辭典與未加入辭典的情況下，本系統對不同領域語料進行斷詞之斷詞效能。在表 6.18 新聞語料的部分，除了利用 SCS 訓練斷詞模型，且有利用英漢翻譯資訊處理交集型歧異的兩組結果外，其他的結果皆是比起訓練斷詞模型時未加入辭典的情況下，在訓練斷詞模型時加入辭典的情況下能有較高的 F_1-measure。在廣播會話語料的部分，則是除了利用 SCS 訓練斷詞模型，有加入未知詞與中英詞對且沒有利用英漢翻譯資訊處理交集型歧異的結果之外，其他的結果都是比起在訓練斷詞模型時未加入辭典的情況下，在訓練斷詞模型時加入辭典的情況下有較高的 F_1-measure。但在科學人的部分，訓練斷詞模型時加入辭典的結果之 F_1-measure 皆無法優於未加入辭典的結果之 F_1-measure。綜合以上可看出，訓練斷詞模型時加入辭典的結果不一定能夠比未加入辭典的結果有更好的斷詞效能。以下我們將對斷詞結果進行分析，以進一步討論訓練斷詞模型時加入辭典不一定能提升斷詞效能的原因。

範例(a)

中文句		斷詞標準答案
當她在週日大約正午時分上了「維珍藍」航空公司的飛機時，		當/她/在/週日/大約/正午/時分/上/了/「/維珍藍/」/航空公司/的/飛機/時/，
訓練斷詞模型時未加入辭典	斷詞結果	是否為正確斷詞結果
	當/她/在/週日/大約/正午/時分/上/了/「/維/珍/藍/」/航空公司/的/飛機/時/，	否
訓練斷詞模型時加入辭典	斷詞結果	是否為正確斷詞結果
	當/她/在/週日/大約/正午/時分/上/了/「/維珍藍/」/航空公司/的/飛機/時/，	是

範例(b)

中文句		斷詞標準答案
有人提議復原孟德爾的蜂巢，		有人/提議/復原/孟德爾/的/蜂巢/，
訓練斷詞模型時未加入辭典	斷詞結果	是否為正確斷詞結果
	有人/提議/復原/孟德爾/的/蜂巢/，	是
訓練斷詞模型時加入辭典	斷詞結果	是否為正確斷詞結果
	有人/提議/復原/孟/德爾/的/蜂巢/，	否

範例(c)

中文句		斷詞標準答案
真正重要的是分析，		真正/重要/的/是/分析/，
訓練斷詞模型時未加入辭典	斷詞結果	是否為正確斷詞結果
	真正/重要/的/是/分析/，	是
訓練斷詞模型時加入辭典	斷詞結果	是否為正確斷詞結果
	真正/重要/的是/分析/，	否

圖 6.3 在訓練斷詞模型時加入辭典與未加入辭典的情況下所得之斷詞結果

分析了斷詞結果後，我們發現若在訓練斷詞模型時加入辭典，則利用斷詞模型進行斷詞時會受到辭典中的詞彙的影響，使得所得到的斷詞結果可能不同於訓練斷詞模型時未加入辭典的情況下所得之斷詞結果，如此會造成以下兩種情況發生。第一種情況為：對某一中文句進行斷詞時，在訓練斷詞模型時未加入辭典的情況下會得到錯誤斷詞結

果，但在訓練斷詞模型時加入辭典的情況下則會得到正確斷詞結果；上頁圖 6.3 之範例(a)為第一種情況的例子，因為辭典中包含詞彙「維珍藍」，所以在訓練斷詞模型時加入辭典的情況下會得到正確斷詞結果，而上頁圖 6.3 中之斷詞標準答案是以人工斷詞方式取得。第二種情況為：對某一中文句進行斷詞時，在訓練斷詞模型時未加入辭典的情況下會得到正確斷詞結果，在訓練斷詞模型時加入辭典的情況下則會得到錯誤斷詞結果。上頁圖 6.3 之範例（b）、範例（c）為第二種情況的例子；在範例（b）、範例（c）中訓練斷詞模型時未加入辭典的情況下皆會得到正確斷詞結果，但在範例（b）中訓練斷詞模型時加入辭典的情況下，因為辭典中包含詞彙「孟德爾」及「德爾」，所以會將中文句斷成錯誤斷詞結果；在範例（c）中訓練斷詞模型時加入辭典的情況下，因為辭典中包含較不常見之詞彙「的是」，所以會將中文句斷成錯誤斷詞結果。我們覺得因為受到上述第二種情況的影響，所以訓練斷詞模型時加入辭典的結果之斷詞效能可能差於訓練斷詞模型時未加入辭典的結果之斷詞效能，例如若對測試語料進行斷詞後，上述之第二種情況的出現次數多於第一種情況的出現次數的話，則訓練斷詞模型時加入辭典的結果之斷詞效能會差於訓練斷詞模型時未加入辭典的結果之斷詞效能。

6.4 以漢英翻譯的翻譯品質評估斷詞效能之實驗

在 6.3 節我們透過人工斷詞之測試語料去評估斷詞效能，但如此可能因斷詞者之斷詞正確率，以及斷詞者之斷詞標準與系統的斷詞標準不同等因素而影響到評估之準確度，例如「他當上副主席了。」這句中文句，在不同的斷詞標準下會有「他/當上/副主席/了/。」、「他/當上/副/主席/了/。」兩種可能斷詞結果；當斷詞者以自己的斷詞標準將其斷成「他/當上/副主席/了/。」，而我們的系統或其他斷詞系統依照不同斷詞標準將其斷成「他/當上/副/主席/了/。」時，就會形成斷詞錯誤。因此除了透過 6.3 節的方式來評估本系統的斷詞效能外，我們也透過漢英翻譯的翻譯品質來評估斷詞效能。

在進行漢英機器翻譯時，需要先對中文語料進行斷詞才能進行後續處理；所以對於漢英機器翻譯，中文斷詞會是一件重要的基礎工作，而中文斷詞效能的好壞可能會影響到最後的翻譯品質。雖然在[22]中有提到斷詞效能越佳不一定保證翻譯品質越好，但在[22]的一些實驗結果中也可以看到，斷詞效能較好之斷詞器能夠有較好之翻譯品質，例如採用長詞優先斷詞的 MaxMatch 斷詞器之斷詞效能優於將句子斷成一個個字的 CharBased 斷詞器之斷詞效能，且 MaxMatch 斷詞器之翻譯品質也優於 CharBased 斷詞器之翻譯品質；所以我們假設在大多數的情形下利用斷詞效能較佳的系統所斷出的中文訓練語料進行翻譯模型訓練，能夠有較好的漢英翻譯之翻譯品質，以利用漢英翻譯之翻譯品質的好壞去間接地評估本系統的斷詞效能。

6.4.1 實驗流程設計

我們將於本實驗中分別使用不同領域之中英平行語料進行漢英翻譯實驗，而所使用的實驗語料有：科學文章類型的 C220、 C300、科學人與新聞文章類型的新聞語料以及會話文章類型的廣播會話語料。由於 NTCIR-9 PatentMT 並未提供測試語料的正確答案，所以我們以 NTCIR-9 PatentMT 提供的有正確答案之 2000 句優化資料（tuning data）作為 C300、C220 之測試語料去進行翻譯效能評估。對科學人、新聞語料、廣播會話語料這三種語料，我們則從語料中切割出 2000 句作為測試語料，其餘的部分則作為訓練翻譯模型之訓練語料。

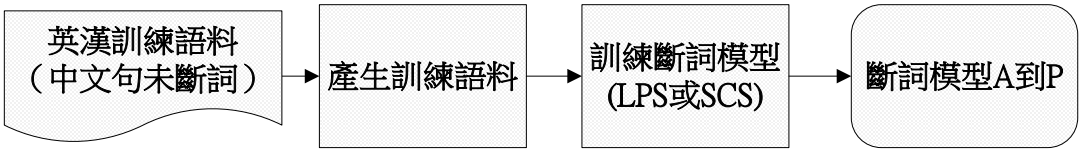

圖 6.4 得到斷詞模型的流程

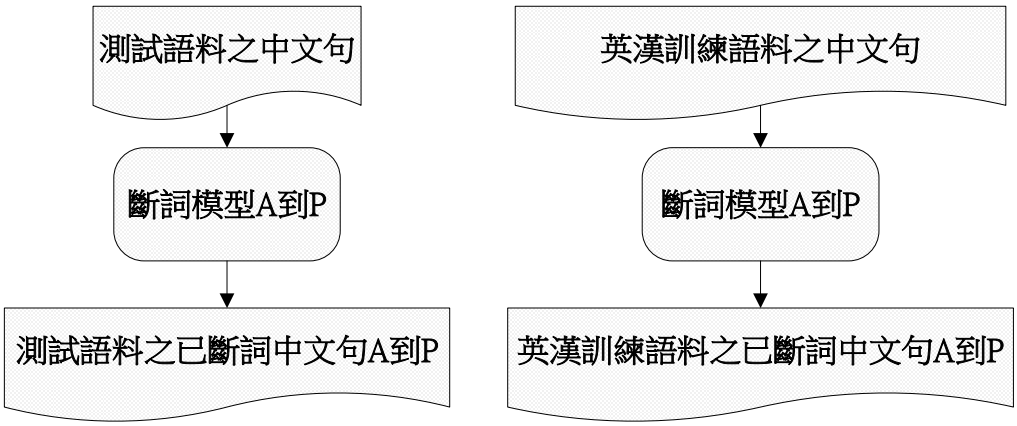

圖 6.5 測試語料與英漢訓練語料之中文句的斷詞流程

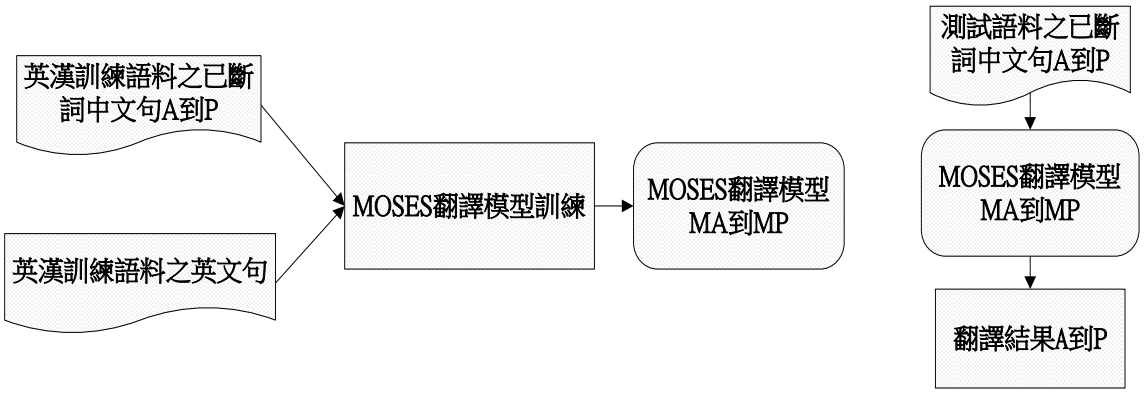

圖 6.6 得到翻譯結果的流程

本研究透過統計式機器翻譯系統「Moses」去進行實驗。在上頁圖 6.4、圖 6.5 中，我們將用來訓練翻譯模型之中英平行語料稱為英漢訓練語料，以跟我們的系統所產生的已斷詞中文訓練語料作區別。而各個實驗的流程大略為：首先我們依照圖 6.4 的流程來透過我們的系統得到各個斷詞模型；與 6.3 節相同，為了評估加入外部辭典進行訓練對於斷詞效能的影響，在利用 LPS 或 SCS 訓練時會分成訓練斷詞模型時加入辭典與訓練斷詞模型時未加入辭典兩種類型。而產生訓練語料的方式也與 6.3 節相同，由有或沒有利用英漢翻譯的資訊去處理交集型歧異之兩種情況去與有或沒有加入未知詞及中英詞對之兩種情況進行組合，故有 4 種產生訓練語料的方式。得到斷詞模型後，我們依照圖

6.5 的流程對測試資料、英漢訓練語料之中文句進行斷詞。最後依照上頁圖 6.6 的流程進行翻譯模型訓練，將測試語料之已斷詞中文句提供給所得到之翻譯模型進行翻譯。

在 6.4.2 節我們將 SCS 之 Pku 斷詞模型、Ctb 斷詞模型及 ICTCLAS 作為我們的系統之斷詞效能比較對象，並藉由翻譯品質去間接評估我們的系統與 SCS 之 Pku 斷詞模型、Ctb 斷詞模型及 ICTCLAS 的斷詞效能。在 C300、C220 的部分，我們另外將 Tseng 等人在 NTCIR-9 PatentMT 利用優化資料進行評估所得的結果之翻譯品質（以下簡稱 Tseng.PatentMT 的結果之翻譯品質）作為透過我們的系統所得之翻譯品質的比較對象。在翻譯結果的評估上，我們則使用 BLEU 和 NIST 兩個指標進行評估。

6.4.2 實驗結果與分析

表 6.21 為 Tseng.PatentMT 的結果之翻譯品質，表中用紅色斜體標示的組別為利用 C220 作為訓練語料的結果，其餘的組別為利用 C300 作為訓練語料的結果；在表 6.21 利用 C300 作為訓練語料的結果中，BLEU 分數最高的是 Z16；而利用 C220 作為訓練語料的結果中，BLEU 分數最高的是 Z18*。下頁表 6.22 為 C300、C220 之漢英翻譯實驗結果，表 6.23 則為科學人、新聞語料、廣播會話語料之漢英翻譯實驗結果；在表 6.22、表 6.23 中，我們將我們的系統之最高 BLEU 分數與其他斷詞系統或斷詞模型中的最高 BLEU 分

表 6.21 Tseng.PatentMT 的結果之翻譯品質

排序	組別	NIST	BLEU	排序	組別	NIST	BLEU	排序	組別	NIST	BLEU
1	*Z18**	7.6120	0.2604	8	*Z2*	6.7831	0.2203	15	Z11	5.1946	0.1487
2	*Z17**	7.3990	0.2514	9	Z15	6.4609	0.2050	16	Z12	4.9405	0.1467
3	*Z16**	7.4346	0.2500	10	Z14	6.3295	0.1995	17	*Z6*	3.3105	0.0674
4	*Z1*	7.3911	0.2486	11	Z13	6.0342	0.1918	18	*Z9*	3.0507	0.0643
5	Z16	7.3778	0.2407	12	*Z5*	6.2146	0.1911	19	*Z7*	2.8805	0.0616
6	Z18	7.3822	0.2403	13	*Z4*	6.1948	0.1811	20	*Z8*	2.7636	0.0602
7	Z17	7.2038	0.2325	14	*Z3*	5.8907	0.1730	21	*Z10*	2.9282	0.0551

數用紅色粗體加斜體標示。

以下我們透過漢英翻譯的品質去間接地評估我們的系統之斷詞效能。在表 6.22 中 C300 的實驗結果部分，我們的系統之最高 BLEU 分數(0.2398)，高於 ICTCLAS 之 BLEU 分數（0.2350），但比上頁表 6.21 中同樣是利用 C300 作為訓練語料的 Z16 之 BLEU 分數（0.2407）低了 0.0009。在表 6.22 中 C220 的實驗結果部分，我們的系統之最高 BLEU 分數(0.2541)，高於 ICTCLAS 之 BLEU 分數(0.2527)，但比表 6.21 中同樣是利用 C220 作為訓練語料的 Z18*之 BLEU 分數（0.2604）低了 0.0063。

表 6.22 C300、C220 之漢英翻譯實驗結果

訓練斷詞模型時未加入辭典						
訓練工具	加入未知詞與中英詞對	利用英漢翻譯資訊處理交集型歧異	C300		C220	
			NIST	BLEU	NIST	BLEU
LPS	沒有	沒有	7.3614	0.2371	7.5545	0.2521
		有	7.4188	***0.2398***	7.5927	0.2541
	有	沒有	7.3496	0.2375	7.5195	0.2498
		有	7.3985	0.2393	7.5962	***0.2541***
SCS	沒有	沒有	7.1789	0.2310	7.4979	0.2496
		有	7.2094	0.2304	7.4834	0.2486
	有	沒有	7.3080	0.2357	7.4267	0.2455
		有	7.1315	0.2289	7.4922	0.2498
訓練斷詞模型時加入辭典						
訓練工具	加入未知詞與中英詞對	利用英漢翻譯資訊處理交集型歧異	C300		C220	
			NIST	BLEU	NIST	BLEU
LPS	沒有	沒有	7.2314	0.2327	7.4434	0.2473
		有	7.2476	0.2328	7.3050	0.2415
	有	沒有	7.3083	0.2355	7.4408	0.2513
		有	7.2315	0.2317	7.3192	0.2418
SCS	沒有	沒有	7.2877	0.2351	7.4877	0.2489
		有	7.3032	0.2376	7.4769	0.2532
	有	沒有	7.3442	0.2378	7.4428	0.2512
		有	7.3295	0.2359	7.4540	0.2503
其他斷詞系統或斷詞模型						
			C300		C220	
			NIST	BLEU	NIST	BLEU
ICTCLAS			7.3104	***0.2350***	7.5012	***0.2527***

在表 6.23 科學人的部分，我們的系統之最高 BLEU 分數（0.0793），比起 ICTCLAS 的 BLEU 分數(0.0813)低了 0.0020，但比起 SCS 之 Ctb 斷詞模型的 BLEU 分數(0.0651)

表 6.23 科學人、新聞語料、廣播會話語料之漢英翻譯實驗結果

訓練斷詞模型時未加入辭典								
訓練工具	加入未知詞與中英詞對	利用英漢翻譯資訊處理交集型歧異	科學人		新聞語料		廣播會話語料	
			NIST	BLEU	NIST	BLEU	NIST	BLEU
LPS	沒有	沒有	4.1036	0.0746	3.9775	0.0717	3.7987	0.0994
		有	4.1178	0.0770	3.9836	***0.0719***	3.8622	0.1024
	有	沒有	4.0959	0.0764	3.9385	0.0697	3.7938	0.1002
		有	4.1494	0.0778	3.9588	0.0695	3.8495	0.1014
SCS	沒有	沒有	3.8661	0.0692	3.8752	0.0685	3.8250	0.1020
		有	4.1493	***0.0793***	3.9331	0.0704	3.8611	0.1035
	有	沒有	4.1230	0.0775	3.8653	0.0672	3.8012	0.1017
		有	4.1582	0.0772	3.9689	0.0695	3.8306	0.1025
訓練斷詞模型時加入辭典								
訓練工具	加入未知詞與中英詞對	利用英漢翻譯資訊處理交集型歧異	科學人		新聞語料		廣播會話語料	
			NIST	BLEU	NIST	BLEU	NIST	BLEU
LPS	沒有	沒有	3.9924	0.0736	3.8998	0.0667	3.6675	0.0957
		有	4.0716	0.0757	3.9606	0.0675	3.7303	0.0988
	有	沒有	4.0010	0.0739	3.8884	0.0661	3.6602	0.0961
		有	4.0592	0.0755	3.9510	0.0680	3.7066	0.0965
SCS	沒有	沒有	3.9924	0.0737	3.8678	0.0673	3.7815	0.0993
		有	4.1498	0.0780	3.9228	0.0698	3.8762	***0.1044***
	有	沒有	4.1044	0.0770	3.8653	0.0672	3.7593	0.0996
		有	4.1563	0.0788	3.9381	0.0696	3.8531	0.1017
其他斷詞系統或斷詞模型								
			科學人		新聞語料		廣播會話語料	
			NIST	BLEU	NIST	BLEU	NIST	BLEU
SCS 之 Pku 斷詞模型			4.2462	0.0806	4.1131	0.0720	3.9019	0.1001
SCS 之 Ctb 斷詞模型			3.8329	0.0651	4.1411	***0.0738***	3.9263	0.1005
ICTCLAS			4.1883	***0.0813***	4.0367	0.0733	3.9316	***0.1067***

高了 0.0142。在上頁表 6.23 新聞語料的部分，我們的系統之最高 BLEU 分數（0.0719），比起 SCS 之 Ctb 斷詞模型的 BLEU 分數（0.0738）低了 0.0019。在表 6.23 廣播會話語料的部分，我們的系統之最高 BLEU 分數（0.1044），比起 ICTCLAS 斷詞器之 BLEU 分數（0.1067）低了 0.0023，但比起 SCS 之各個斷詞模型之 BLEU 分數皆高出 0.004 左右。

由以上分析可看出，在科學文章類型之科學人、C300 或新聞文章類型之新聞語料或會話文章類型之廣播會話語料的部分，我們的系統之最佳翻譯品質都略差於其他斷詞系統或斷詞模型中的最佳翻譯品質，而在 C300 的部分，我們的系統之最高 BLEU 分數跟其他斷詞系統或斷詞模型中的最高 BLEU 分數之差距只有 0.0009。所以我們覺得這間接顯示了我們的系統可以有一定的斷詞效能。

以下我們藉由表 6.22、上頁表 6.23 的數據來分析透過本研究提出的加入未知詞及中英詞對與利用英漢翻譯的資訊去處理交集型歧異的方法是否能提升斷詞效能。在訓練斷詞模型時未加入辭典的情況下，所有實驗語料當中，只有廣播會話語料的部分，有利用英漢翻譯資訊處理交集型歧異的實驗結果之 BLEU 分數皆高於沒有利用英漢翻譯資訊處理交集型歧異的實驗結果之 BLEU 分數。在訓練斷詞模型時未加入辭典的情況下，在任何一種實驗語料的結果中，都並非所有的有加入未知詞與中英詞對的實驗結果之 BLEU 分數皆高於沒有加入未知詞與中英詞對的實驗結果之 BLEU 分數。所以由以上分析可看出，利用英漢翻譯資訊處理交集型歧異或加入未知詞與中英詞對不一定能提升斷詞效能。

以下我們針對表 6.22、上頁表 6.23 中訓練斷詞模型時加入辭典與訓練斷詞模型時未加入辭典的實驗結果進行分析。在表 6.22 中的 C300、C220 之利用 LPS 訓練斷詞模型的部分，除了 C220 之有加入未知詞與中英詞對且沒有利用英漢翻譯資訊處理交集型歧異的結果之外，其他的結果都是訓練斷詞模型時加入辭典之情況下的 BLEU 分數比起訓練斷詞模型時未加入辭典之情況下的 BLEU 分數來得低；但在利用 SCS 訓練斷詞模型的部分，除了 C220 之沒有加入未知詞與中英詞對且沒有利用英漢翻譯資訊處理交集型

歧異之結果外，其他的結果都是訓練斷詞模型時加入辭典之情況下的 BLEU 分數高於訓練斷詞模型時未加入辭典之情況下的 BLEU 分數。

在表 6.23 中的科學人部分，除了利用 SCS 訓練斷詞模型的其中兩組結果之外，其他的結果皆是訓練斷詞模型時加入辭典的情況下之 BLEU 分數低於訓練斷詞模型時未加入辭典的情況下之 BLEU 分數。在廣播會話語料部分，除了利用 SCS 訓練斷詞模型，沒有加入未知詞與中英詞對且有利用英漢翻譯資訊處理交集型歧異的結果外，其餘的結果皆是訓練斷詞模型時加入辭典的情況下之 BLEU 分數低於訓練斷詞模型時未加入辭典的情況下之 BLEU 分數。在新聞語料部分，則除了利用 SCS 訓練斷詞模型，有加入未知詞與中英詞對的兩組結果外，其餘的結果皆是訓練斷詞模型時加入辭典的情況下之 BLEU 分數低於訓練斷詞模型時未加入辭典的情況下之 BLEU 分數。

由以上分析可看出，不管是在哪種實驗語料的結果部分，訓練斷詞模型時加入辭典的結果之翻譯效能都不一定優於訓練斷詞模型時未加入辭典的結果之翻譯效能，我們覺得這間接地顯示訓練斷詞模型時加入辭典的結果不一定能夠比未加入辭典的結果有更好的斷詞效能，這也符合了在 6.3.2 節我們所觀察到的結果：訓練斷詞模型時加入辭典不一定能夠提升斷詞效能。

第七章　結論與未來展望

本章為本論文之結論與未來展望。在 7.1 節介紹本研究之結論，7.2 節說明未來之可能改進方向。

7.1 結論

在本篇論文中，我們建構一個基於中英平行語料的斷詞系統。提供我們的系統不同領域之中英平行語料後，系統可以自動化地產生品質不錯的訓練語料，並使用所產生的訓練語料訓練斷詞模型，再利用斷詞模型對該領域之語料斷詞。

我們的系統所用的辭典模組包含英漢辭典模組、中文辭典模組。而因為在利用英漢翻譯資訊處理交集型歧異時會使用英文詞彙之中文翻譯去對應斷詞組合，所以為了提升利用英漢翻譯資訊處理交集型歧異的效果，本研究透過 E-HowNet、一詞泛讀去尋找英文詞彙的中文翻譯之近義詞，並建置加入近義詞之英漢合併辭典。

在產生訓練語料時，我們的系統會對中英平行語料的每句中文句產生句子的各種斷詞組合，並利用英漢翻譯的資訊處理交集型歧異，將錯誤的斷詞組合去除。我們從中英平行語料擷取出「候選中文遺留字詞」、「候選中英遺留詞對」，把利用可能性比例、共現頻率篩選出的「候選中英遺留詞對」視為新的中英詞對，加入至英漢辭典模組。把利用詞性序列規則篩選出的「候選中文遺留字詞」視為未知詞，加入至中文辭典模組。

在以人工斷詞測試語料評估斷詞效能之實驗中，本研究針對科學文章類型之科學人、新聞文章類型之新聞語料、會話文章類型之廣播會話語料這三種不同領域之語料進行實驗。在科學人、廣播會話語料部分，我們的系統之最高 F_1-measure 與斷詞效能最佳的其他斷詞系統或斷詞模型之 F_1-measure 的差距都在 0.05 以內，且我們的系統之最高的 F_1-measure 都在 0.835 以上。因此我們覺得這顯示了我們的系統能夠有一定的斷詞效

能。另外由實驗結果可發現，在訓練斷詞模型時未加入辭典的情況下，有利用英漢翻譯資訊處理交集型歧異或有加入未知詞與中英詞對的結果之斷詞效能都能提升。而在訓練斷詞模型時加入辭典的情況下，可能因為受到加入的未知詞與辭典詞彙衝突的情況影響，導致加入未知詞與中英詞對的結果之斷詞效能並沒有都優於未加入未知詞與中英詞對的結果之斷詞效能。我們也評估訓練斷詞模型時加入辭典對斷詞效能的影響，而實驗結果顯示訓練斷詞模型時加入辭典不一定能夠提升斷詞效能。

因為使用人工斷詞測試語料進行評估可能會因為斷詞者之斷詞正確率與斷詞者之斷詞標準不同於斷詞系統的標準而影響到評估之準確度，所以本研究另外進行了以漢英翻譯的翻譯品質評估斷詞效能之實驗，藉由翻譯品質去間接地評估我們的系統的斷詞效能。本實驗之實驗語料是科學文章類型之科學人、C300、C220 與新聞文章類型之新聞語料與會話文章類型之廣播會話語料。由實驗結果可以發現，在 4 種實驗語料的結果中，我們的系統之最佳翻譯品質都略差於其他斷詞系統或斷詞模型中的最佳翻譯品質，我們覺得這間接顯示了我們的系統可以有一定的斷詞效能。另外分析了各實驗結果之翻譯品質後，我們發現利用英漢翻譯資訊處理交集型歧異或加入未知詞與中英詞對不一定能提升斷詞效能，以及訓練斷詞模型時加入辭典不一定能夠提升斷詞效能。

7.2 未來展望

在辭典模組的部分，我們希望能蒐集到更多專業辭典（如醫學辭典、地名辭典等），以提升對某特定領域語料的斷詞效能及對專有名詞的斷詞正確率。而除了利用英漢翻譯資訊處理交集型歧異外，若也能利用英漢翻譯資訊處理組合型歧異，相信能對斷詞效能的提升有所幫助。

在擷取中英詞對的方面，因為我們利用可能性比例、共現頻率對候選中英遺留詞對進行篩選的效果並不是很好，所以或許能探討利用一個以上的分析公式(例如可能性比

例、點互訊息(pointwise mutual information)、相關係數（correlation coefficient）等公式)加上共現頻率來進行篩選的效果，是否能比僅利用可能性比例、共現頻率進行篩選的效果來得好。此外或許也可以利用一些專有名詞辨識（named entity recognition）的軟體工具來辨識出是專有名詞的英文遺留字詞或候選中文遺留字詞後，再以可能性比例、共現頻率等作為輔助的資訊對候選中英遺留詞對進行篩選，如此或許可提升篩選的效果。

在利用詞性序列規則篩選候選中文遺留字詞時，目前較難篩選出長度較長的一些中文專有名詞或構詞結構較多變化的地名、組織名等詞彙。如果在利用詞性序列規則進行篩選時，同時以「山」、「社」等關鍵字來判斷詞彙是否為地名或組織名，或許能篩選出較多的地名、組織名。另外因為取得詞性序列規則前需要先對中文語料、候選中文遺留字詞標注詞性，故詞性標注的準確率可能會影響篩選候選中文遺留字詞的效果；所以除了使用史丹佛剖析器進行詞性標注外，或許也可以使用其他剖析器（如伯克利剖析器(Berkeley Parser)）去進行詞性標注，以分析利用不同的剖析器進行詞性標注是否會對篩選候選中文遺留字詞的效果造成影響。

在實驗中我們共使用科學文章、新聞文章、會話文章類型這三種不同領域的中英平行語料進行實驗，如果可以取得更多不同領域的大量中英平行語料，我們就能夠更全面性地評估我們的系統的斷詞效能；而實驗語料的數量除了 C300、C220 較為充足外，其他的語料之數量都略顯不足。如果可以取得更大量的中英平行語料，就能分析以不同數量之訓練語料進行訓練是否會對斷詞效能造成影響。在利用人工斷詞測試語料進行評估之實驗部分，因為我們目前只有用自己自行人工斷詞之 200 句測試語料進行實驗，所以為了更客觀的進行斷詞效能評估，我們需要取得有專家標記的斷詞標準答案之測試語料來進行評估。

參考文獻

[1] 牛津現代英漢雙解詞典，http://startdict.sourceforge.net/Dictionaries_zh_TW.php [連結已失效]。

[2] 中央研究院中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw/ [2011/11/2]。

[3] 中央研究院現代漢語標記語料庫 4.0 版簡介，http://db1x.sinica.edu.tw/cgi-bin/kiwi/mkiwi/mkiwi.sh [2011/12/22]。

[4] 田侃文，*英漢專利文書文句對列與應用*，國立政治大學資訊科學所，碩士論文，2009。

[5] 史丹佛剖析器， http://nlp.stanford.edu/software/lex-parser.shtml [2012/2/26]。

[6] 朱怡霖，*中文斷詞與專有名詞辨識之研究*，國立臺灣大學資訊工程學研究所，碩士論文，2002。

[7] 成語詞典，http://yeelou.com/huzheng/stardict-dic/zh_TW/ [2011/3/30]。

[8] 林筱晴，*語料庫統計值與網際網路統計值在自然語言處理上之應用：以中文斷詞為例*，國立臺灣大學資訊工程學研究所，碩士論文，2004。

[9] 林千翔，*基於特製隱藏式馬可夫模型之中文斷詞研究*，國立中央大學資訊工程研究所，碩士論文，2006。

[10] 莊怡軒，*英文技術文獻中動詞與其受詞之中文翻譯的語境效用*，國立政治大學資訊科學所，碩士論文，2011。

[11] 現代漢語一詞泛讀，http://elearning.ling.sinica.edu.tw/introduction.html [2011/8/26]。

[12] 國家教育研究院學術名詞資訊網，http://terms.nict.gov.tw/download_main.php [2011/8/26]。

[13] 掌印辭典整理，http://www.palmstamp.com/forum/viewthread.php?tid=832&page=1#pid6847 [2011/8/26]。

[14] 詹嘉丞，*中文斷詞系統中非繁體中文詞彙之處理*，國立台灣海洋大學資訊工程所，碩士論文，2009。

[15] 構詞篇(下)，http://chcs-opencourse.org/chcs/full_content/A21/pdf/03.pdf [2012/2/27]。

[16] 劉群、李素建，基於《知網》的辭彙語義相似度計算，*中文計算語言學期刊*，第七卷第二期，59-76，2002。

[17] 懶蟲簡明英漢詞典，http://yeelou.com/huzheng/stardict-dic/zh_TW/ [2011/3/30]。

[18] 羅永聖，*結合多類型字典與條件隨機域之中文斷詞與詞性標記系統研究*，國立臺灣大學資訊工程學研究所，碩士論文，2008。

[19] Keh-Jiann Chen and Shing-Huan Liu, Word Identification for Mandarin Chinese Sentences, *Proceedings of the 15th International Conference on Computational Linguistics*, 101-107, 1992.

[20] Keh-Jiann Chen and Ming-Hong Bai, Unknown Word Detection for Chinese by a Corpus-based Learning Method, *International Journal of Computational linguistics and Chinese Language Processing*, Vol. 3, Num. 1, 27-44, 1998.

[21] Keh-Jiann Chen and Wei-Yun Ma, Unknown Word Extraction for Chinese Documents, *Proceedings of the 19th International Conference on Computational Linguistics*, 169-175, 2002.

[22] Pi-Chuan Chang, Michel Galley, and Christopher D. Manning, Optimizing Chinese Word Segmentation for Machine Translation Performance, *Proceedings of the 3rd Workshop on Statistical Machine Translation*, 224-232, 2008.

[23] Dr.eye譯典通字典, http://www.dreye.com/ [2011/8/26].

[24] E-HowNet, http://ckip.iis.sinica.edu.tw/taxonomy/taxonomy-doc.htm [2011/8/26].

[25] E-HowNet Technical Report, http://rocling.iis.sinica.edu.tw/CKIP/paper/Technical_Reprt_E-HowNet.pdf [2012/6/21].

[26] Chung-Chi Huang, Wei-Teh Chen, and Jason S. Chang, Bilingual Segmentation for Alignment and Translation, *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, 445-453, 2008.

[27] ICTCLAS漢語分詞系統, http://ictclas.org/ [2012/7/1].

[28] Wenbin Jiang, Liang Huang, Qun Liu, and Yajuan Lü, A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging, *Proceedings of 46th Annual Meeting on Association for Computational Linguistics: HLT*, 897-904, 2008.

[29] Wenbin Jiang, Liang Huang, and Qun Liu, Automatic Adaptation of Annotation Standards:ChineseWord Segmentation and POS Tagging – A Case Study, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 522-530, 2009.

[30] Mu Li, Jianfeng Gao, Changning Huang, and Jianfeng Li, Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation, *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, 1-7, 2003.

[31] LingPipe, http://alias-i.com/lingpipe/ [2011/8/26] .

[32] Yanjun Ma and Andy Way, Bilingually Motivated Domain-Adapted Word Segmentation for Statistical Machine Translation, *Proceedings of the 12th Conference of the European Chapter of the ACL*, 549-557, 2009.

[33] Moses, http://www.statmt.org/moses/ [2011/12/22].

[34] C. D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, 1999, MIT Press.

[35] Pat-Tree 中文抽詞程式, http://www.openfoundry.org/of/projects/367/ [2012/3/16].

[36] Patent Machine Translation Task at the NTCIR-9, http://ntcir.nii.ac.jp/PatentMT/ [2012/3/11].

[37] SIGHAN Bakeoff 2, www.sighan.org/bakeoff2005/ [2011/12/22].

[38] Stanford Chinese Segmenter, http://nlp.stanford.edu/software/segmenter.shtml [2011/8/26].

[39] Yuen-Hsien Tseng, Chao-Lin Liu, Chia-Chi Tsai, Jui-Ping Wang, Yi-Hsuan Chuang, and James Jeng, Statistical approaches to patent translation - Experiments with various settings of training data, *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access - PatentMT*, 661-665, 2011.

[40] Kun Wang, Chengqing Zong, and Keh-Yih Su, A Character-Based Joint Model for Chinese Word Segmentation, *Proceedings of the 23th International Conference on Computational Linguistics*, 1173-1181, 2010.

[41] Yahoo!斷章取義API, http://tw.developer.yahoo.com/cas/ [2011/11/2].

附錄Ⅰ 不同領域語料之斷詞效能（以詞數表示）

在本附錄中的各表格中，P 代表精確率，R 代表召回率。表中各項結果的 P、R 我們不以數據形態表示，而是以系統斷出的詞數、系統斷出的正確詞數、參考答案的所有詞數來表示。所以表中 P 的表示方式為 $\frac{\text{系統斷出的正確詞數}}{\text{系統斷出的詞數}}$，R 的表示方式為 $\frac{\text{系統斷出的正確詞數}}{\text{參考答案中的所有詞數}}$。

訓練斷詞模型時未加入辭典								
訓練工具	加入未知詞與中英詞對	利用英漢翻譯資訊處理交集型歧異	廣播會話語料		科學人		新聞語料	
			P	R	P	R	P	R
LPS	沒有	沒有	2284/2944	2284/2822	1471/1856	1471/1763	1488/2066	1488/1857
		有	2309/2929	2309/2822	1487/1846	1487/1763	1491/2051	1491/1857
	有	沒有	2287/2940	2287/2822	1470/1844	1470/1763	1487/2032	1487/1857
		有	2314/2921	2314/2822	1494/1833	1494/1763	1491/2023	1491/1857
SCS	沒有	沒有	2334/2947	2334/2822	1581/2076	1581/1763	1602/2360	1602/1857
		有	2377/2941	2377/2822	1602/2052	1602/1763	1617/2347	1617/1857
	有	沒有	2348/2933	2348/2822	1598/2055	1598/1763	1604/2355	1604/1857
		有	2378/2930	2378/2822	1620/2028	1620/1763	1640/2309	1640/1857

訓練斷詞模型時加入辭典								
訓練工具	加入未知詞與中英詞對	利用英漢翻譯資訊處理交集型歧異	廣播會話語料		科學人		新聞語料	
			P	R	P	R	P	R
LPS	沒有	沒有	2233/2727	2233/2822	1420/1794	1420/1763	1460/1967	1460/1857
		有	2273/2724	2273/2822	1459/1779	1459/1763	1472/1964	1472/1857
	有	沒有	2230/2727	2230/2822	1420/1782	1420/1763	1456/1934	1456/1857
		有	2275/2722	2275/2822	1449/1770	1449/1763	1474/1934	1474/1857
SCS	沒有	沒有	2347/2926	2347/2822	1442/1869	1442/1763	1603/2353	1603/1857
		有	2402/2917	2402/2822	1470/1856	1470/1763	1615/2346	1615/1857
	有	沒有	2330/2926	2330/2822	1449/1849	1449/1763	1605/2352	1605/1857
		有	2385/2911	2385/2822	1463/1851	1463/1763	1634/2319	1634/1857

	廣播會話語料		科學人		新聞語料	
其他斷詞系統或斷詞模型	P	R	P	R	P	R
中研院斷詞系統	–	–	1643/1872	1643/1763	1714/2007	1714/1857
斷章取義斷詞系統	–	–	1303/1729	1303/1763	1398/1881	1398/1857
SCS 之 Pku 斷詞模型	2496/2868	2496/2822	1528/1822	1528/1763	1584/1943	1584/1857
SCS 之 Ctb 斷詞模型	2453/2901	2453/2822	1530/1849	1530/1763	1630/1958	1630/1857
ICTCLAS	2502/2946	2502/2822	1483/1889	1483/1763	1575/2077	1575/1857

附錄Ⅱ 口試問題與建議之記錄

在本附錄中記錄了口試時三位口試委員所提出的問題與建議，以及對於各問題與建議的回答內容。

問題或建議 1	可直接比對 E-HowNet 中文詞彙之表示式與英文詞彙之中文翻譯的表示式是否完全相同，若完全相同則將該 E-HowNet 中文詞彙視為近義詞。透過這樣的方法得到的近義詞會較精準。
回答	本研究並不是以取得精準的近義詞為主要目標。本研究尋找英文詞彙的中文翻譯近義詞的目的是提高利用英漢翻譯資訊處理交集型歧異的效果，而在利用不精準之英文詞彙的中文翻譯近義詞去中文句進行比對時，我們認為中文句出現該近義詞的機會並不大，如此就不會造成錯誤情形發生。

問題或建議 2	在用 PAT-tree 抽詞程式擷取詞彙時有設定 frequency 或詞的長度等參數嗎?
回答	使用 PAT-tree 抽詞程式擷取詞彙時我們把 minFreq（詞的最小詞頻）設為 2。詞的最小長度設為 2，詞的最大長度則設為 12。

問題或建議 3	在以人工斷詞測試語料評估斷詞效能之實驗中，是否都會從各實驗語料中切割 200 句出來作為測試語料?
回答	如論文 6.3.2 節所述，對於每種實驗語料，我們都會從語料中切割 200 句出來作為測試語料。

問題或建議 4	sememe 的中文寫法為「義原」還是「義元」？
回答	在知網官方網頁中的知網簡介等內容中所使用的都為「義原」而非「義元」。故我們認為 sememe 的中文寫法為「義原」。

問題或建議 5	在擷取未知詞時是否有考慮到直接音譯的中文詞彙？
回答	對於直接音譯的中文詞彙（如「布魯克」等），我們也是透過詞性序列規則進行篩選。但因為直接音譯的中文詞彙有很多種不同的構詞結構，所以在利用詞性序列規則篩選時，能夠篩選出的直接音譯的中文詞彙之數量可能偏低。

問題或建議 6	在中英平行語料的使用上，只有利用到句對的資訊，還是也有用到整篇文章的資訊？
回答	我們在篩選候選中英遺留詞對時有使用到由整篇文章所得到的資訊（共現頻率）。

問題或建議 7	可以把目前所使用的方法之不完善的地方列出，當作未來改進的方向。
回答	已於論文 7.2 節將其列出。

問題或建議 8	應提早說明系統的整體架構。
回答	已將系統的整體架構提前於論文第三章中進行說明。

問題或建議 9	可用較 benchmark 的方法來評估斷詞效能(如使用 SIGHAN Bakeoff 2 所公開的 4 種語料之測試語料）。
回答	SIGHAN Bakeoff 2 所公開的 4 種語料都非中英平行語料,所以無法提供給我們的系統產生訓練語料。而若使用科學人、新聞語料等實驗語料所訓練出的斷詞模型去對 SIGHAN Bakeoff 2 所公開的 4 種語料之測試語料斷詞，則因為不知道斷詞模型與測試語料是否為相同領域，故難以進行精確的斷詞效能評估。故我們認為 SIGHAN Bakeoff 2 所公開的 4 種語料之測試語料不適合用來評估我們的系統的斷詞效能。