應用平行語料建構中文斷詞組件

Applications of Parallel Corpora for Chinese Segmentation

王瑞平 劉昭麟
Jui-Ping Wang Chao-Lin Liu

國立政治大學資訊科學系
Department of Computer Science, National Chengchi University
{g9916, chaolin}@cs.nccu.edu.tw

摘要

不同於直接提供中文斷詞服務，網路上的開放軟體讓人們可以利用自有的訓練語料來訓練中文的斷詞模型，藉以實踐自己的斷詞功能。如若可以全人工方式建構斷詞訓練語料，則以目前的機器學習技術所訓練出來的模型，常常可以達到相當好的斷詞效果。然而，實務上全人工的標記工作常常是難以提供足夠多的訓練語料。本文利用中英平行語料與各類辭典，搭配中文未知詞和近義詞的偵測，先建構一個粗略的斷詞器，藉以產生訓練語料，最後再利用網路上的開放軟體來建構中文斷詞服務。在目前的實驗中，雖然依照我們的程序所得的斷詞服務未能立即獲得優於知名的中文斷詞服務的成效，但是表現卻相去不遠；我們所提出的訓練語料產生程序提供了一個一般人可以考慮的選擇。

Abstract

Instead of directly providing the service of Chinese segmentation, some open-source software allows us to train segmentation models with segmented text. The resulting models can perform quite well, if training data of high quality are available. In reality, it is not easy to obtain sufficient and excellent training data, unfortunately. We report an exploration of using parallel corpora and various lexicons with techniques of identifying unknown words and near synonyms to automatically generate training data for such open-source software. We achieved promising results of segmentation in current experiments. Although the results fell short of outperforming the well-known Chinese segmenters, we believe that the proposed approach offers a viable alternative for users of the open-source software to generate their own training data.

關鍵詞：機器學習，語料標記，機器翻譯

1. 緒論

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。中文斷詞技術大致可分為法則式斷詞法[10]及統計式斷詞法[5][16][26]。統計式斷詞法在訓練斷詞模型時若以大量高品質的訓練語料進行訓練，則通常可有好的斷詞效能，但因為通常透過人工斷詞所得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。因此我們建立一個透過以下程序來提供中文斷詞服務的系統：首先透過查詢各類辭典的方式來產生中英平行語料之所有中文句的各種斷詞組合，並將錯誤斷詞組合去除，藉以產生訓練語料；然後再將所產生的訓練語料提供給網路上的開放軟體去訓練斷詞模型，以建構中文斷詞服務。

在本論文後續內容，我們將所建立的提供中文斷詞服務的系統[1]簡稱為我們的系統。而當使用者提供未斷詞語料給我們的系統時，系統會以訓練好的斷詞模型對未斷詞語料斷詞。

中文斷詞存在兩個重要問題：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，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。Li[17]等利用非監督式（unsupervised）訓練的方法處理交集型歧異，本研究則透過英漢翻譯的資訊去處理交集型歧異。

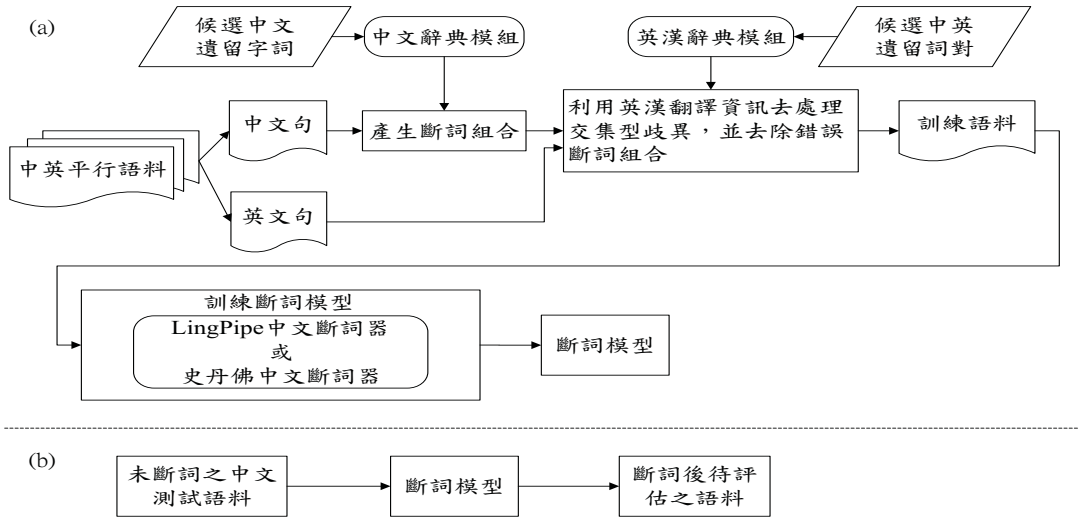
未知詞指的是未收錄於辭典中的詞彙，例如人名、地名、組織名等。在日常生活中人們會不斷創造出新的詞彙，故未知詞經常會出現在文章中。斷詞系統在對未知詞斷詞時通常會出現錯誤斷詞的情形，所以如果想要提升斷詞效能，則處理未知詞問題會是必要的工作。在處理未知詞問題相關研究方面，Chen 等[11]利用以語料庫為基的學習法(corpus-based learning approach)去產生規則以進行未知詞偵測。擷取未知詞時，Chen 等[12]針對屬於未知詞一部份的單字詞，會判斷該單字詞是否可以和相鄰的詞彙進行合併。

我們從中英平行語料中擷取未知詞、新的中英詞對，藉此處理未知詞問題與提升利用英漢翻譯資訊去處理交集型歧異的效果。擷取中英詞對與未知詞之大略流程如下：首先從中英平行語料中擷取候選中英遺留詞對、候選中文遺留字詞。之後利用可能性比例(likelihood ratios)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。

2. 系統架構

2.1 系統流程與架構

我們的系統之流程與整體架構如圖一所示，而流程中各步驟的詳細方法會在後續章節詳加說明；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



圖一、系統的流程與架構

[1] 因工作時程等因素，我們所建立的提供中文斷詞服務的系統目前並沒有線上版本。

的候選中文遺留字詞擴充至中文辭典模組。在提供我們的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，並將錯誤斷詞組合去除。得到訓練語料後，我們利用 LingPipe 中文斷詞器[18]及史丹佛中文斷詞器[23]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. 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，而在這兩個模組中都包含一般辭典與專業辭典這兩種類別。中文與英漢辭典模組的各辭典之列表、辭典詞彙數統計如表一、表二所示。關於「英漢技術名詞辭典」與「中文技術名詞辭典」及「加入近義詞之英漢合併辭典」的建置會在後續內容中說明。

本研究從國家教育研究院學術名詞資訊網[8]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。

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[6]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[13]合併成「英漢合併辭典」，以增加英文詞彙的中文翻譯數目。我們針對「英漢合併辭典」中的各個英文詞彙，從中央研究院現代漢語一詞泛讀系統[7]（以下簡稱一詞泛讀）及 E-HowNet[14]取得該詞彙的中文翻譯近義詞集後，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」中的英文詞彙之中文翻譯進行整合，就完成「加入近義詞之英漢合併辭典」的建置。

4. 產生訓練語料

4.1 產生各種斷詞組合

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$C_{1:n}$（$C_{1:n} = C_1\ C_{2...}C_n$），並依照下頁圖二的步驟來產生句子的各種斷詞組合。以下為圖二中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

表一、中文辭典模組之辭典列表、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

表二、英漢辭典模組之辭典列表、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

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。

圖二、產生句子的各種斷詞組合的步驟

選集合，在$Cand_i$內會存放字串$C_{1:i}$的各種斷詞組合。

在圖二步驟 3(b)中之「包含單字詞的詞彙組合」的定義為：當某詞彙組合包含單字詞，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」。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。若語料中的許多中文句都產生大量的斷詞組合，會使訓練語料過於龐大，造成訓練斷詞模型時消耗大量時間、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進行說明。在圖二步驟 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如圖三所示。

在圖二步驟 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。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)，因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、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

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:，

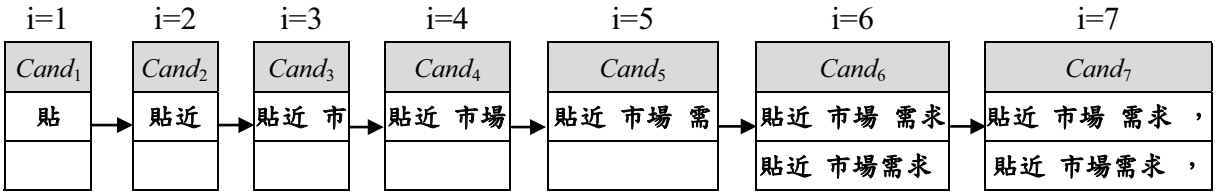
↓

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				

圖三、產生「貼近市場需求，」之 V_i



圖四、各階段的 $Cand_i$ 的內容

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」後等同於$C_{1:6}$，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」後等同於$C_{1:6}$，所以也會被加入至$Cand_6$。執行步驟 3、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。圖四則是各階段的$Cand_i$的內容。

4.2 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。

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，我們利用英文句中各詞彙之中文翻譯集合中的中文翻譯去對應斷詞組合中的中文詞彙；如果某英文詞彙的中文翻譯集合中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。

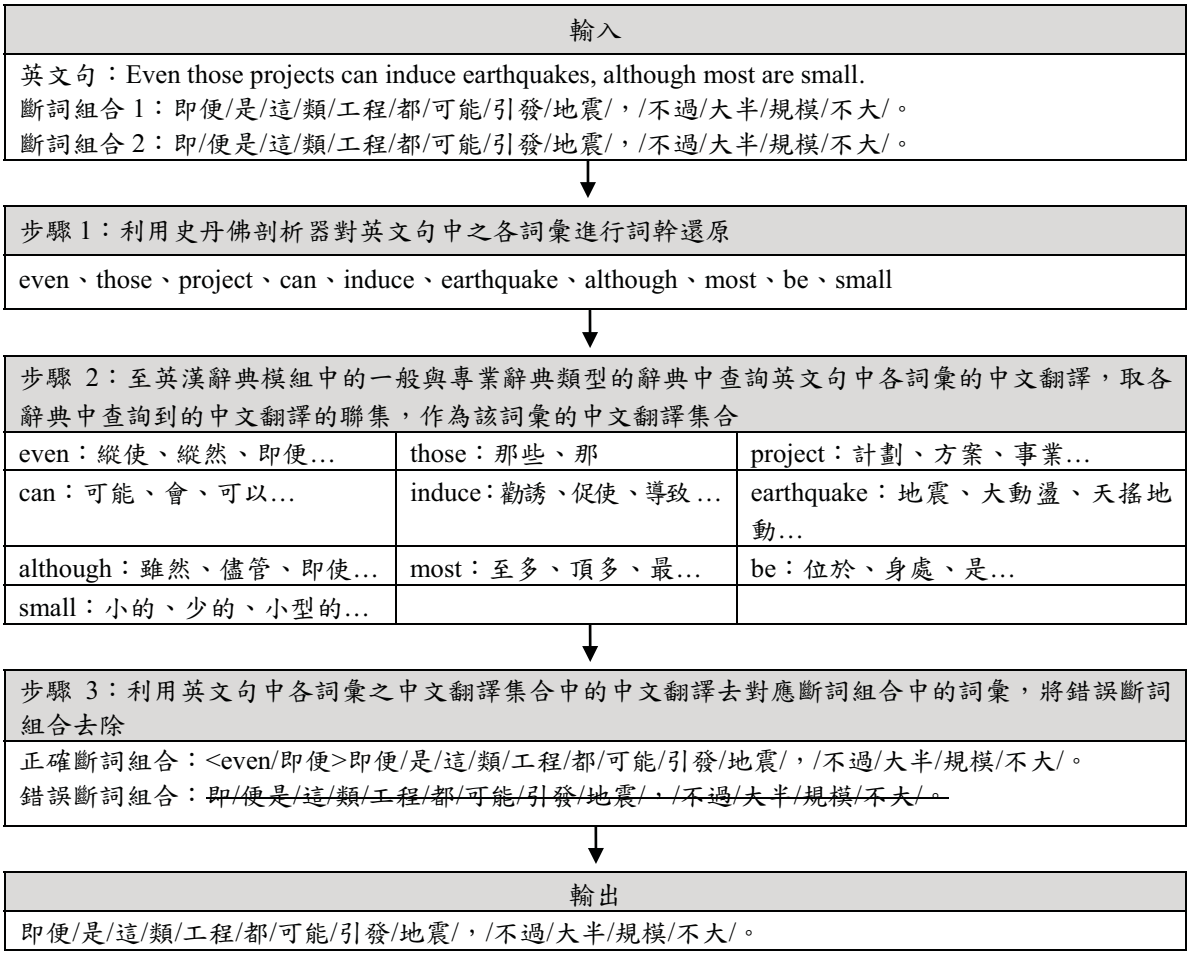
以下藉下頁圖五說明處理交集型歧異的整體流程。中文句「即便是這類工程都可能引發地震，不過大半規模不大。」包含了交集型歧異字串「即便是」(「即便是」可被斷成「即便/是」及「即/便是」)，而圖五中的斷詞組合 1、斷詞組合 2 為該中文句的各個斷詞組合。經過步驟 1、 2 後，我們取得了英文句中各詞彙的中文翻譯集合，如“even”的中文翻譯集合包含「縱使」、「縱然」、「即便」等詞彙。在步驟 3 正確斷詞組合的部分，我們標記<詞幹還原後的英文詞彙/中文詞彙>的意思是利用左側的詞幹還原後的英文詞彙之中文翻譯集合中的中文翻譯，可以對應到右側的中文詞彙，例如< even/即便>的意思是 “even”是經過詞幹還原後的詞彙，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。

4.3 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、未知詞，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處理訓練語料中的未知詞問題。在擷取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」。之後我們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，利用詞性序列規則對「候選中文遺留字詞」進行篩選。

4.3.1 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，之後再利用英文句各詞彙之中文翻譯集合中的中文翻譯對中文句進行斷詞。在斷詞後，中文句會有未被斷詞的「中文遺留字



圖五、處理交集型歧異的整體流程

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」。對於所有「中文遺留字詞」，我們使用 PAT-tree 抽詞程式**[21]**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，許多錯誤的結果都含有停用詞，如「會不」、「確的」；因此對於以 PAT-tree 抽詞程式所擷取出的各結果，我們藉由停用詞列表將其中包含停用詞的結果去除後，其餘的結果即為「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」。

4.3.2 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[20]，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。可能性比例的公式如下：

$$\text{Likelihood ratio}\,(\mathrm{c},\mathrm{e}) = \log\lambda = \log\frac{\mathrm{b}(Fce, Fc, p)\mathrm{b}(Fe - Fce, N - Fc, p)}{\mathrm{b}(Fce, Fc, p_1)\mathrm{b}(Fe - Fce, N - Fc, p_2)} \quad (1)$$

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（共現頻率代表候選中英遺留詞對中的中文詞、英文詞共同出現之句對數），N為中英平行語料的總句數。我們將信心水準(confidence level)訂為 99.5%，則臨界值(critical value)為 7.88。當$-2\log\lambda$超過 7.88 時，代表「候選中文遺留字詞」與「英文遺留字詞」是有關連的。

以下透過表三說明如何利用可能性比例與共現頻率進行篩選，而表三中的候選中英遺留詞對已依照共現頻率大小由大到小進行排序（當共現頻率相等時再依照$-2\log\lambda$大小由大到小進行排序）。假設共現頻率的門檻值為 3，則雖然詞對「越高 increase」之共現頻率大於或等於 3，但因進行可能性比例檢測後其$-2\log\lambda$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$-2\log\lambda$大於 7.88，所以這 4 個詞對會被視為新的中英詞對並加入至英漢辭典模組中。

4.3.3 利用詞性序列規則進行篩選

表三、候選中英遺留詞對之共現頻率與$-2\log\lambda$對應表

排名	候選中英遺留詞對	共現頻率	$-2\log\lambda$
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

我們發現「候選中文遺留字詞」可分成三大類：第一類為「已知詞」，第二類為「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[9]），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。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」。

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。

斷詞系統遇到未知詞時會將未知詞斷成幾個較小的單位。我們藉由去除辭典的部分詞彙的方式，將這些詞彙當作未知詞，所以這些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所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」所構成的詞彙組合「房地　產」即為「未知詞候選詞彙組合」。

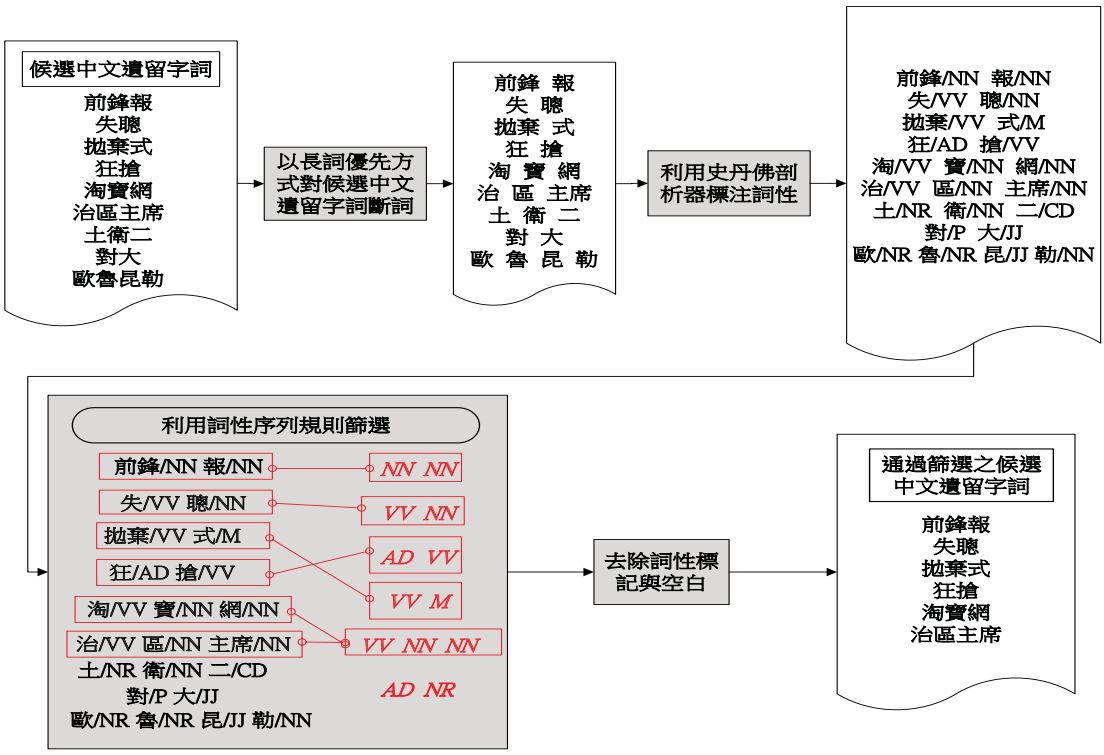
我們透過下頁圖六之各個步驟來建立詞性序列規則表。在圖六中步驟 1，我們將 N 取 10，把辭典切割成十等份。以下我們對步驟 3 到 6 進行說明：在第 k 回合，我們將原始中文辭典的第 k 份去除，所以在辭典之第 k 份中的詞彙會被當成未知詞；對語料斷詞後，出現在語料中之第 k 份中的詞彙會被斷成「未知詞候選詞彙組合」。在步驟 5，本研究利用史丹佛剖析器[4]對語料標注詞性，而標注時所使用的字典模型為 xinhuaFactored.ser.gz。在標注詞性後，語料中的「未知詞候選詞彙組合」之詞性序列規則即為該詞彙之詞性序列規則。例如「房地　產」經過詞性標注後變為「房地/NN　產/NN」，則「房地產」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」）

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. 從標注詞性後的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$R_1, R_2,\ldots, R_N$的結果

圖六、建立詞性序列規則表的步驟

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。

以下我們藉圖七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；若以「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大於或等於門檻值（詞性序列規則的出現次數）的各個詞性序列規則（圖中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、空白去除就會得到「淘寶網」、「治區主席」這兩個候選中文遺留字詞。最後我們將通過篩選之候選中文遺留字詞視為未知詞，將其加入至中文辭典模組。



圖七、利用詞性序列規則篩選候選中文遺留字詞之範例

5. 實驗結果與分析

5.1 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 C300、C220 與廣播會話(BroadCast Conversation)語料，實驗語料句數統計如表四所示，而以下將對上述提到的語料的來源及我們對語料所做的處理進行說明。

表四、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

田侃文[3]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，而我們沿用這 63256 句句對進行實驗。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[3]轉換成中英平行句對後進行合併，就得到新聞語料。

Tseng等人[25]於Patent Machine Translation Task at the NTCIR-9[22]（以下簡稱NTCIR-9 PatentMT）時對 100 萬句專利平行語料進行前處理後得到了兩種英漢雙語訓練語料C300[2]、C220，而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、Part 2 語料的檔案中擷取中英平行句對，並將短句(長度小於 6 之句子)、重複的句對及句中特殊符號去除。最後將GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 1 、Part 2 語料之中英平行句對（已去除重複句對、短句）進行合併後就得到廣播會話語料。

5.2 擷取中英詞對與未知詞之實驗

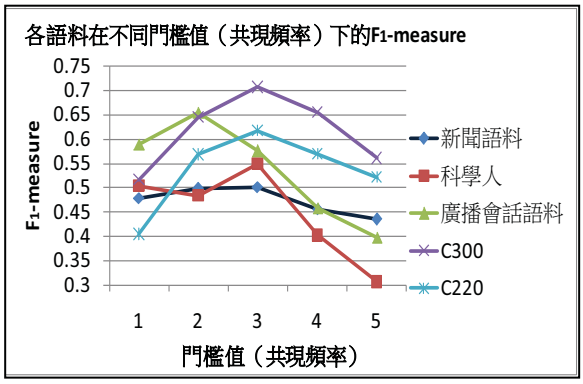
在本實驗中我們從科學人、新聞語料、廣播會話語料、C300、C220 這五種語料中擷取中英詞對、未知詞，並評估其效果。首先我們從各語料中擷取候選中英遺留詞對、候選中文遺留字詞，而所擷取出的候選中英遺留詞對、候選中文遺留字詞之數量如表五所示。

我們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。我們以詞性序列規則的出現次數作為門檻值來取得不同的詞性序列規則，再透過所取得之各個詞性序列規則對候選中文遺留字詞進行篩選。之後對於所有被篩選出的候選中文遺留字詞，我們以人工的方式檢測其是否為未知詞。在結果評估上，則使用標準定義的

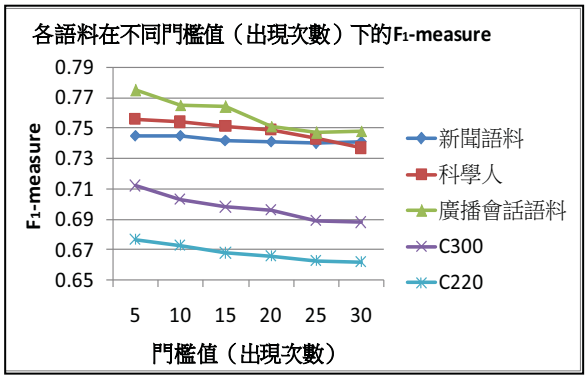
表五、候選中文遺留字詞、候選中英遺留詞對數量統計

語料名稱	候選中英遺留詞對數量	候選中文遺留字詞數量
科學人	5410	2484
新聞語料	3502	2475
廣播會話語料	831	356
C300	9326	4619
C220	7798	3469

[2] 雖然在[25]中並沒有記錄對 C1140 的進一步處理，但 Tseng 等人得到 C1140 後還有對 C1140 進行篩選，再利用篩選完的語料進行實驗。而 C1140 經篩選後約剩 30 萬句，故本研究以 C300 代替 C1140。



圖八、不同門檻值(共現頻率)下所得的 F1-measure



圖九、不同門檻值(出現次數)下所得的 F1-measure

精確率（Precision）、召回率（Recall）、F1-measure 這三個指標進行評估。

圖八是以不同門檻值去對各實驗語料之候選中英遺留詞對進行篩選所得的 F1-measure。如圖八所示，在新聞語料、科學人、C300、C220 部分，F1-measure 最高的都是門檻值為 3 之結果，故我們分別把這四種語料之以門檻值為 3 所篩選出的結果加入至英漢辭典模組。在廣播會話語料部分，F1-measure 最高的是門檻值為 2 之結果，所以我們把以門檻值為 2 所篩選出的結果加入至英漢辭典模組。

圖九是以不同門檻值去對各實驗語料之候選中文遺留字詞進行篩選所得的 F1-measure。如圖九所示，在新聞語料的部分，門檻值為 5 或 10 時有相同的F_1-measure。而門檻值為 5 之結果的召回率為 0.915，門檻值為 10 之結果的召回率為 0.907。因為我們希望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，並把以門檻值為 5 所篩選出的結果加入至中文辭典模組。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 所篩選出的結果加入至中文辭典模組。

5.3 以人工斷詞測試語料評估斷詞效能之實驗

5.3.1 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中英平行語料中切割而來，所以我們並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斷詞標準答案，以進行斷詞效能的評估。

我們不以一些網路上開放使用的有斷詞標準答案之測試語料（如由中央研究院、香港城市大學等所提供的測試語料[24]）去評估我們的系統的斷詞效能之原因是：若將科學人等實驗語料提供給我們的系統，再以所得的各個斷詞模型對網路上開放使用的測試語料進行斷詞並評估斷詞效能，則可能因為實驗語料與測試語料並非是相同領域的語料，導致得到不精確的評估結果；此外一些網路上開放使用的訓練語料（與網路上開放使用的測試語料同領域）並非是中英平行語料，故無法提供給我們的系統去得到斷詞模型。因此我們不使用網路上開放使用的有斷詞標準答案之測試語料進行斷詞效能評估。

本實驗之產生訓練語料的方式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)。

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]與斷章取義斷詞系統[27]、SCS 之 Pku 及 Ctb 斷詞模型、ICTCLAS 漢語分詞系統[15]（以下簡稱 ICTCLAS）作為我們的系統之比較的對象。而除了評估我們的系統之斷詞效能外，我們在 5.3.2 節分析透過本研究提出的加入未知詞及中英詞對或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未加入辭典這兩種類型去進行實驗，而在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。

$$\text{精確率} = \frac{\text{系統斷出的正確詞數}}{\text{系統斷出的詞數}} \quad (2)$$

$$\text{召回率} = \frac{\text{系統斷出的正確詞數}}{\text{參考答案中的所有詞數}} \quad (3)$$

$$\mathrm{F_1-measure} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (4)$$

我們使用精確率(Precision)、召回率(Recall)、F_1-measure 這三個評估指標去評估斷詞效能，公式(2)-(4)為各指標的個別定義；而在下頁表六中的 P 代表精確率，R 代表召回率，F_1 代表 F_1-measure。

5.3.2 實驗結果與分析

下頁表六為我們的系統對不同領域語料之斷詞效能。在表六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，而在表六廣播會話語料之中研院斷詞系統、斷章取義斷詞系統的結果部分我們則將其標示為「-」。

以下為我們的系統與其他斷詞系統或斷詞模型之斷詞效能比較。在表六科學人部分，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，我們的系統的最高 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的其他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。

在表六的結果中，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、廣播會話語料的部分，比起沒有利用英漢翻譯資訊處理交集型歧異的結果之 F_1-measure，有利用英漢翻譯資訊處理交集型歧異的結果之 F_1-measure 皆能提升，而其中 F_1-measure 提升最多的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（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。

由表六數據可看出，在訓練斷詞模型時未加入辭典的情況下，在所有實驗語料的部分，比起沒有加入未知詞與中英詞對的結果之 F_1-measure，有加入未知詞與中英詞對的結果之 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

表六、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	

異的結果（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提升斷詞效能。但在訓練斷詞模型時加入辭典的情況下，並不是所有的有加入未知詞與中英詞對的結果之 F_1-measure 皆高於沒有加入未知詞與中英詞對的結果之 F_1-measure。

以下藉由表六數據來比較訓練斷詞模型時加入辭典與未加入辭典的情況下，我們的系統對不同領域語料進行斷詞之斷詞效能。在新聞語料、廣播會話語料的部分，並不是所有訓練斷詞模型時加入辭典的結果之 F_1-measure 都可優於未加入辭典的結果之 F_1-measure。而在科學人的部分，則是所有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。

5.4 以漢英翻譯的翻譯品質評估斷詞效能之實驗

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，所以中文斷詞效能的好壞可能會影響到最後的翻譯品質。因此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我們的系統的斷詞效能。

5.4.1 實驗流程設計

我們於本實驗中使用不同領域之中英平行語料進行實驗，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。對科學人、新聞語料、廣播會話語料這三種語料，我們從語料中切割出 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。

本研究透過統計式機器翻譯系統「Moses」[19]去進行實驗。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中文訓練語料作區別。而實驗的流程大略為：首先我們將英漢訓練語料提供給我們的系統來得到各個斷詞模型；之後，我們使用所得到的各個斷詞模型對測試語料、英漢訓練語料之中文句進行斷詞。最後將英漢訓練語料之英文句、英漢訓練語料之已斷詞中文句提供給 Moses 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的翻譯模型進行翻譯。

在 5.4.2 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。在 C300、 C220 的部分，我們另外將 Tseng 等在 NTCIR-9 PatentMT 利用優化資料進行評估所得之 BLEU 分數最高的結果（在 C300 的部分 BLEU 分數最高的為 Z16，在 C220 的部分 BLEU 分數最高的為 Z18*）作為我們的系統之比較對象。在翻譯結果的評估上，則使用 BLEU 和 NIST 這兩個指標進行評估。

5.4.2 實驗結果與分析

表七、下頁表八分別為 C300、C220 與科學人、新聞語料、廣播會話語料之漢英翻譯實驗結果；在表七，我們將我們的系統之最高 BLEU 分數與 Z16、Z18*的 BLEU 分數用紅色粗體加斜體標示；在表八，則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。在表七中 C300 的實驗結果部分，我們的系統之最高 BLEU 分數，高於 ICTCLAS 之 BLEU 分數，但比同樣是利用 C300 作為訓練語料

表七、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
其他斷詞系統、Z18*、Z16			C300		C220	
			NIST	BLEU	NIST	BLEU
ICTCLAS			7.3104	0.2350	7.5012	0.2527
Z18*			—	—	7.6120	***0.2604***
Z16			7.3778	***0.2407***	—	—

表八、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
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***

的 Z16 之 BLEU 分數低了 0.0009；在 C220 的實驗結果部分，我們的系統之最高 BLEU 分數，高於 ICTCLAS 之 BLEU 分數，但比同樣是利用 C220 作為訓練語料的 Z18*之 BLEU 分數低了 0.0063。由表八的數據可看出，在科學人的部分，我們的系統之最高 BLEU 分數，比 ICTCLAS 的 BLEU 分數低了 0.002，但比 SCS 之 Ctb 斷詞模型的 BLEU 分數高了 0.0142。在新聞語料的部分，我們的系統之最高 BLEU 分數，比起 SCS 之 Ctb 斷詞模型的 BLEU 分數低了 0.0019。在廣播會話語料的部分，我們的系統之最高 BLEU 分數，比 ICTCLAS 斷詞器之 BLEU 分數低了 0.0032，但比 SCS 之各個斷詞模型之 BLEU 分數皆高出 0.003 左右。

由以上分析可看出，在科學文章類型之科學人、C300 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。

6. 結論

在本篇論文中，我們建立一個透過以下程序來提供中文斷詞服務的系統：首先透過查詢中文辭典的方式來產生中英平行語料之所有中文句的各種斷詞組合，並利用英漢翻譯的資訊將錯誤斷詞組合去除，藉以產生訓練語料；最後再將所產生的訓練語料提供給開放軟體去訓練斷詞模型，以建構中文斷詞服務。

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，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，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。此外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。

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。由實驗結果可看出，在四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有一定的斷詞效能。

致謝

本研究承蒙國科會研究計畫 NSC-100-2221-E-004-014 的部份補助，謹此致謝。我們感謝評審對於本文的各項指正與指導，限於篇幅因此不能在本文中全面交代相關細節。

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] 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。
[4] 史丹佛剖析器，http://nlp.stanford.edu/software/lex-parser.shtml [2012/2/26]。
[5] 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。
[6] 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。
[7] 現代漢語一詞泛讀，http://elearning.ling.sinica.edu.tw/introduction.html [2011/8/26]。
[8] 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]。
[9] 構詞篇（下），http://chcs-opencourse.org/chcs/full_content/A21/pdf/03.pdf [2012/2/27]。
[10] 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.
[11] 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.
[12] 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.
[13] Dr.eye譯典通字典, http://www.dreye.com/ [2011/8/26].
[14] E-HowNet, http://ckip.iis.sinica.edu.tw/taxonomy/taxonomy-doc.htm [2011/8/26].
[15] ICTCLAS漢語分詞系統, http://ictclas.org/ [2012/7/1].
[16] 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.
[17] 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.
[18] LingPipe, http://alias-i.com/lingpipe/ [2011/8/26].
[19] Moses, http://www.statmt.org/moses/ [2011/12/22].
[20] C. D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, 1999, MIT Press.
[21] PatTree 中文抽詞程式, http://www.openfoundry.org/of/projects/367/ [2012/3/16].
[22] Patent Machine Translation Task at the NTCIR-9, http://ntcir.nii.ac.jp/PatentMT/ [2012/3/11].
[23] Stanford Chinese Segmenter, http://nlp.stanford.edu/software/segmenter.shtml [2011/8/26].
[24] SIGHAN Bakeoff 2, www.sighan.org/bakeoff2005/ [2011/12/22].
[25] 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.
[26] 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.
[27] Yahoo!斷章取義API, http://tw.developer.yahoo.com/cas/ [2011/11/2].