以共現資訊爲基礎增進中學英漢翻譯試題與解答之詞彙對列

Using Co-Occurrence Information to Improve Chinese-English Word Alignment in Translation Test Items for High School Students

黃昭憲[†] 張裕淇[↕] 劉昭麟[‡] 曾元顯[↑]

Chao-Shainn Huang Yu-Chi Chang Chao-Lin Liu Yuan-Hsien Tseng

[†↕‡]國立政治大學資訊科學系 [↑]國立台灣師範大學資訊中心

National Chengchi University National Taiwan Normal University

{[†]g9707, [↕]g9824, [‡]chaolin}@cs.nccu.edu.tw, [↑]samtseng@ntnu.edu.tw

摘要

本文探討我國中學程度翻譯試題的中文試題與英文解答的詞彙對列問題。我們先利用漢英字典作爲基礎，找尋互爲翻譯的漢英詞對；然後利用未被對列的剩餘詞彙之間的共現關係，以五種過去在文獻中探索過的計分方式，來尋找與界定更多的互譯詞彙組合。在超過 17,000 道試題爲基礎的測試中，我們以人工檢視計分機制給予高分的部分對列詞彙的正確性。實驗結果顯示，進一步利用未對列詞彙的方法，可以把對列成果的 F measure 從 76.9%提高到 83.7%。

Abstract

We study the word alignment between the Chinese problems and the English answers for the English-Chinese translation tests at the high school level in Taiwan. After applying a dictionary-based approach, we attempted to take advantage of the information about co-occurrence of unaligned words to find more translation pairs. We explored five scoring methods that were previously proposed in the literature. We ran our experiments with more than 17,000 test items, and checked the aligned word pairs that were assigned high scores. Experimental results showed that we could improve the F measure for the alignment task from 76.9% to 83.7% with the best performing scoring method.

關鍵詞：詞彙對列、電腦輔助翻譯、遺留字詞、新詞對擷取

1. 緒論

外語試題的翻譯工作，往往會利用大量的人力資源和時間來進行人工翻譯，以得到翻譯品質優良的中文試題。然而人工翻譯會因爲不同的翻譯者，對相同的英文句型有不同地見解，而產生不一致的翻譯結果。因此，倘若能利用機器翻譯(machine translation)的技術來輔助人工英譯的工作，便可能提升翻譯速度和翻譯結果的一致性。

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 128-142, Puli, Nantou,Taiwan, September 2010.

爲了提升機器翻譯結果的品質，我們希望透過詞彙對列(word alignment)的技術，了解翻譯不同語文時詞彙順序相異的現象，以便讓翻譯系統先進行詞序互換的動作，再進行字詞的翻譯，使整個翻譯句更加通順且一致。常見的詞彙對列技術是以辭典爲基礎來進行對列工作，主要的步驟是將中文句子經過斷詞系統之後，再逐一的把中文字詞透過漢英辭典查詢，並且記錄所得到的英文翻譯集合，最後將英文翻譯集合中的英文單字與英文句子進行比對，以了解詞序互換的現象。然而，這樣的做法會受限於辭典內部的詞彙量，連帶影響詞彙對列的效果。Mihalcea 和 Pedersen[18]提出了幾點改進詞彙對列整體效能的建議，其中一點就是盡可能的整合所有可用資源，他們更提出了遺留詞彙如果可以透過有效的評估與擷取機制，則可得到大量有用的資源。國內學者多年以來也有很多詞彙對列的相關研究，例如柯與張以辭典爲基礎的對列[13]；白等學者[12]利用平行語料庫來擷取多字詞語(multiword expressions)，當找出中英文平行句對間可能互爲翻譯的字串之後，作者利用正規化頻率(normalized frequency)來對中英文詞對組合評分，並把英文字串切割成多個共同子序列(common subsequences)，計算其出現頻率後，利用 Dice 係數(Dice coefficient)來產生量化分數[16]，接著將這些可能互爲翻譯的詞彙組合進行排序(ranking)，最後探討並設立其門檻值，挑選出最佳的英文目標候選字(target candidate words)。

我們對原始詞彙對列技術進行改良，包含常見的一個中文字詞對應一個英文字詞此類之對列方式，將原先辭典內部擁有的複合字資訊(如片語)加以利用，促使詞彙對列模組可進行一個中文字詞對應多個連續英文字詞之對列情況。我們發展的對列模組，以中文字詞當作索引字，透過漢英辭典來得到相對應的英文字詞，若漢英辭典中沒有收錄其中文字詞，我們便無法進行對列的動作，所以辭典本身的資訊量與整個系統效能有著密不可分的關係。但也因爲如此，我們不能爲了眾多的中文字詞，而一味地追求龐大的辭典，所以我們試圖利用史丹佛詞類標記器(Stanford part-of-speech tagging)[20]來進行原詞還原(lemmatization)，對原始的英文句子中的各個單詞，進行詞性處理(如時態問題、複數問題等)；並且也將中文句子透過中研院中文斷詞系統，產生數個中文字詞。由於語言的特性，一個概念，常可有多種辭彙可以使用，因此在對應的同時，我們將透過同義詞詞林[4]來進行中文字詞的擴充，以便擴大對應的機會。

在完成第一階段的詞彙對列之後，我們進一步探討中英文平行句對中沒有成功對應的字詞(Null Alignment, 以下稱之爲遺留字詞)。詞彙對列主要可產生四種結果，「完全對列(無任何遺留字詞)」、「只遺留中文字詞」、「只遺留英文字詞」和「中英都有遺留字詞」，我們將針對「只遺留中文字詞」及「只遺留英文字詞」兩種結果，來進行停用詞列表的選取和遺留字詞利用的討論；並對「中英都有遺留字詞」進行新詞對的擷取，我們將應用曾元顯等人[10]的方法，來評估並擷取遺留詞彙之對列。在此過程中，我們將設立一信心分數做爲門檻值，以擷取出新的詞對組合，並整合至原始的辭典，重新進行詞彙對列工作，以改善詞彙對列模組。我們將在第 2 節中介紹實驗語料的來源，第 3 節則詳細說明詞彙對列的流程，並利用共現(co-occurrence)觀念來對遺留字詞進行新詞對的擷取，在第 4 節中則是系統的效能評估，最後在第 5 節進行總結。

表 1 高中英文試題範例

試題類型	試題範例
引導式翻譯	提示句子：據聞這明星結過七次婚。 _____ the _____ _____, the movie star has married seven times. 答案：As; story; goes
整句式翻譯	提示句子：原子鐘需要數百萬元來製造。 _____ 答案：Atomic clocks cost several million dollars to make.
連貫式翻譯	提示句子：(1)不射殺幼熊是仁慈的舉止。 _____ 提示句子：(2)一位新聞記者甚至畫了一幅有關此事件的卡通。 _____ 答案：(1)Not shooting the bear cub is a kindly/kind act. (2)A newspaper reporter even drew a cartoon of this incident.

2. 研究語料介紹

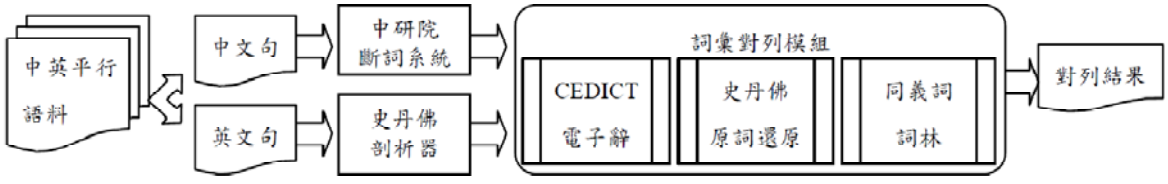
我們利用高中程度英文句子來當作訓練語料。想獲得高中程度的中英文平行語料，最直接的方法可以從「課文」和「試題」兩個方向來獲取，但是「課文」的資訊會因爲前後文的關係，以篇章段落作爲翻譯單位，其相對應之中文翻譯句經過潤飾後，時常並非爲一句英文對應一句中文，而會有一對多、多對一，甚至是多對多的情況產生。假若我們無法準確地將這些句子進行斷句，則我們就無法獲取較爲完美的平行句對，這會影響詞彙對列的效能，因此我們並不以「課文」作爲訓練語料。

另一個方向，我們知道英文試卷常是由許多試題類型所組合而成，其中我們感興趣的有「引導式翻譯」、「整句式翻譯」和「連貫式翻譯」，以表1爲例。上述三種題型結構都帶有中文提示句子，接著要求學生依提示句子的敘述，撰寫出英文答案。我們將會透過自動化程序擷取其中文句子，並且透過答案欄中的資訊，還原其英文句子，以獲得大量的平行句對。我們的實驗主要從三民出版社[1]所發行的高中英文試題光碟，進行訓練語料的擷取，從中獲得9954句中英文平行句對。

3. 研究方法

3.1 詞彙對列技術

詞彙對列模組的流程圖如圖一所示，首先我們進行斷詞，將一句中文句子經由中研院斷詞系統[2]，斷成數個中文詞，形成以字詞爲單位的集合，英文句子則是透過空白隔開進行斷字，並將英文句子以句爲單位，利用史丹佛剖析器[20]得到各英文單字之詞性標



圖一 前處理與詞彙對列模組流程

表 2 中英平行句對斷詞結果

原始中英平行句對	
中文	你可以用遙控器改變許多功能。
英文	You can change a lot of functions with the remote control .
中英平行句對斷詞結果	
中文	你 可以 用 遙控器 改變 許多 功能 。
英文	PRP/You VP/can VB/change DT/a NN/lot IN/of NNS/functions IN/with DT/the remote NN/control ./.

輸入：中英文平行句對

輸出：中英平行句對之相對應詞序[英文詞序/中文詞序]

處理程序：
前處理：將來源中文句子透過中研院斷詞系統進行斷詞，得到來源中文字詞
將目標英文句子透過史丹佛剖析器得到每個目標英文字的詞性標記

步驟 1：以來源中文字詞作爲索引字，透過 CEDICE 電子辭典互爲翻譯的英文字詞，並對目標英文句子進行比對
步驟 2：把尚未完成對列之目標英文字，利用史丹佛詞性標記器對目標英文字進行詞性還原，並重複一次 Step 1
步驟 3：把尚未完成對列之來源中文字詞，透過同義詞詞林獲得來源中文字詞之同義詞，再將同義詞當作索引字重複一次 Step 1
步驟 4：以 0 表示對列過程中無法產生連接

圖二 詞彙對列演算法

記(Part-Of-Speech tags)，如表 2 所示。接著我們給定中英文句斷詞後的各個字詞，由左至右依序標記上詞序，以便詞彙對列模組透過各個步驟來完成平行語料的對應，整體的演算法如圖二所示。圖二只列出粗略的步驟，我們在本節次中的其他小節中會詳述相關細節。

3.1.1 以字典爲基礎

接下來將斷詞過後的中英文平行句對進行詞彙對列。詞彙對列已有許多技術，在此我們選用以辭典爲基礎的對列技術作爲我們的出發點，此方法主要是將索引字詞，透過雙語辭典進行查詢，將會得到大量的翻譯候選字，利用這些翻譯候選字和目標字詞進行字詞相似度的比對。而此種技術相當依賴雙語辭典的翻譯品質，並且我們所需要的辭典必須明確的指出，該索引字詞有可能被翻譯成哪些目標字詞，而非一般的翻譯辭典(利用其他字詞來「說明」索引字詞的詞意)。因此我們選用CEDICT電子辭典[14]當作索引字詞資料庫，此辭典含有97184個中文索引詞(本模組所使用的版本時間爲2010-02-22 06:12:50 GMT)，辭典內部的格式以行爲單位，每行所儲存的資訊由中文字詞、漢語拼音和英文翻譯候選字

```
…
郵政 [you2 zheng4] /postal/
郵政局 [you2 zheng4 ju2] /postal bureau/
油脂 [you2 zhi1] /grease/oil/fat/
卣 [you3] /wine container/
友 [you3] /friend/
優 [you3] /grievous/relaxed/
有 [you3] /have/there is/there are/exist/be/
…
```

圖三 CEDICT 漢英電子辭典格式範例

表 3 一對一之詞彙對列

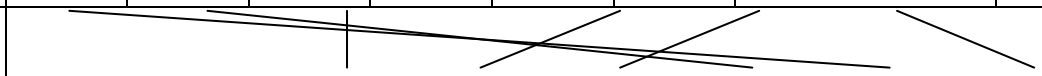
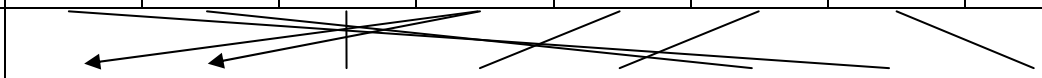
中文	醫院	裡	很少	有	男	護士	。	
								
英文	There	are	few	male	Nurses	in	hospitals	.

表 4 整合一對多之詞彙對列

中文	醫院	裡	很少	**有**	男	護士	。	
								
英文	**There**	**are**	few	male	Nurses	in	hospitals	.

群所組成，且每個翻譯候選字以「/」符號隔開，如圖三所示。

由圖三我們可以得知，此辭典除了一對一的中英翻譯以外，還包含了片語和複合字的資訊，例如：「有」對應的英文翻譯除了「have、exist、be」這類的單字詞外，還包含了「there is、there are」這類的複合字或片語。針對此類型的資訊，對列模組在進行對列的時候，將會把複合詞和片語這類多個單字組合而成的字串，當作對列過程的首要任務，如此一來便可以從字對字的對列，擴展成一個中文字詞可以對應多個英文字詞。假若詞彙對列模組，其只能一個中文字詞對應一個英文字詞，我們以表3作爲例子，可以發現中文字詞「有」，並無法成功的和「there are」產生連結，這是因爲「there are」必須兩個字詞連結在一起，才能與「有」這個中文字詞互爲翻譯。因此我們可以善加利用辭典內部的資訊，來增加一個中文字詞可以對應到多個英文字詞的能力，進而得到表4這樣的對列結果(目前模組只能處理「there are」這類連續的英文字詞，「not only…but also」這類並非連續字串的字詞結構，我們暫時無法處理)。因此我們的系統便可依此類型辭典作爲出發點，將中文字詞分別透過CEDICT電子辭典進行相對應翻譯字詞的查詢，如表5的例子所示。

在上述對列的過程中，我們還將所有的英文字詞轉化成小寫(如英文句中的開頭第一個字「You」會被轉化成「you」)，由左至右依序把中文字詞透過CEDICT電子辭典，進行一個中文字詞對應一個或多個英文字詞的對列。若產生成功的對列結果，則在英文詞序的後方標記上其對應的中文詞序並以「/」隔開，表示此英文字詞已完成對列，不需要再與其他中文字詞進行比對。以表6爲例，例句中共有十二個英文字詞，其中只有九個字詞可以成功對列。

3.1.2 以原詞還原爲基礎

在以辭典爲基礎的詞彙對列中，對列模組以原始的英文句子當作對列目標，主要的原因是我們所使用的辭典內部會有許多複合詞的資訊，如「撲克牌」所對應的翻譯詞爲「playing card」，如果直接先對原始英文句子進行還原處理(lemma)，則會把「playing」還原成「play」，這樣便無法正確的把兩組詞彙產生連結，因此我們會先把原始英文句進行以辭典爲基礎的對列之後，才針對還沒有完成的英文字詞進行原詞還原(lemmatization)以

表 5 詞彙對列(透過辭典查詢)

詞序	1	2	3	4	5	6	7	8	9	10	11	12
英文	You	can	change	a	lot	of	functions	with	the	remote	control	.

詞序	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
中文	你	可以	用	遙控器	改變	許多	功能	。
翻譯字詞	you	can may possible	use	remote control	change alter transform	many much a lot of	function capability	.

表 6 詞彙對列(以辭典爲基礎的對列結果)

詞序	**1/1**	**2/2**	**3/5**	**4/6**	**5/6**	**6/6**	**7**	**8**	**9**	**10/4**	**11/4**	**12/8**
英文	**you**	**can**	**change**	**a**	**lot**	**of**	**functions**	**with**	**the**	**remote**	**control**	**.**

詞序	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
中文	你	可以	用	遙控器	改變	許多	功能	。
翻譯字詞	**you**	**can** may possible	use	**remote control**	**change** alter transform	many much **a lot of**	function capability	.

執行對列動作。

而執行原詞還原的目的是，英文字詞常在不同的詞性或是不同的時態下，原始的單字會有一些變化。因此在第二步驟的詞彙對列中，我們把尚未完成對列的英文字詞，透過史丹佛詞性標記器(Stanford Part-Of-Speech tagger)來進行原詞處理，期盼加強詞彙對列的對列能力。

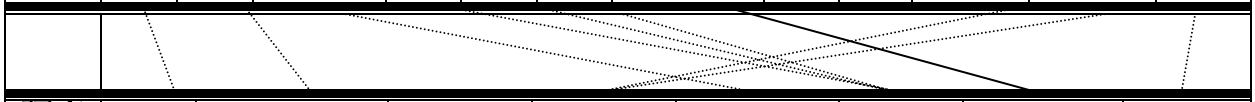
主要的概念在於，一句英文句子常會因爲時態(如過去式、現在進行式和未來式)，或者是前後文的因素來改變其單字的型態(如複數型態等)，而我們所用的辭典內部則多爲原始形態。因此，我們對輸入的英文句子進行原詞處理，以提升詞彙對列的機會。爲了在原詞處理可以得到較佳的效果，們先利用史丹佛剖析器將英文字詞進

functions/NNS	→	function
with/IN	→	with

圖四 史丹佛詞類標記器原詞還原之結果

表 7 詞彙對列(以原詞還原爲基礎的對列結果)

詞序	**1/1**	**2/2**	**3/5**	**4/6**	**5/6**	**6/6**	**7/7**	**8**	**9**	**10/4**	**11/4**	**12/8**
英文	**you**	**can**	**change**	**a**	**lot**	**of**	**functions**	**with**	**the**	**remote**	**control**	**.**
原詞還原							**function**	**with**				



詞序	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
中文	你	可以	用	遙控器	改變	許多	功能	。
翻譯字詞	you	can may possible	use	remote control	change alter transform	many much a lot of	function capability	.

表 8 詞彙對列(透過同義詞詞林查詢)

詞序	1	2	3	4	5	6	7	8
中文	你	可以	用	遙控器	改變	許多	功能	。
同義詞群編號			同義詞詞林 第一步			Dj083 用,用度,用費,花費,花項,用項,… Fc061 用,啖,服,偏,飽食,… Hc252 收錄,錄取,齒錄,收齒,用,任人唯親,… Hj281 用,以,使,動,下,祭,自用,… Jc051 欲,用,得,待,提供,… Kb052 給,受,以,用,拿,將,… Kc092 故此,故而,從而,故,用,… 第二步		

行詞類標記，我再將其英文字詞和詞類標記一起透過史丹佛詞類標記器作型態上的(morphological)原詞還原。

我們針對表 6 中尚未完成對列的字詞進行原詞還原，第七個英文字「functions」複數型態(其詞性標記爲 NNS)，透過史丹佛詞類標記器還原成原型「function」，而「佛詞類標記器還原成原型「function」，而「with」仍然是「with」進行標記，如圖四所示。左邊爲輸入字串「英文字詞/詞類標記」，右邊則爲輸入字串「還原後的英文字詞」。接著進行第二輪的詞彙對列，整體的效果如表 7 所示。

3.1.3 以同義詞詞林爲基礎

```
…
Aa042,他們
Aa042,她們
Aa042,彼等
Aa051,自己
Aa051,自身
Aa051,本身
Aa051,我
…
```

圖五 同義詞詞林內部

承上一個小節的結果，由於中文字詞的多樣性，往往一個中文字詞的意思，可以用其他中文字詞來表示，舉例來說，「用」在某些狀況下，也可以用「以」、「使用」和「花費」等來代替。爲了增加對應率，我們採用哈爾濱工業大學訊息檢索實驗室同義詞詞林擴充版(以下稱爲同義詞詞林)來進行中文詞彙擴展。此版本內含 66,697 個中文字詞，且將這些字詞分爲 5,353 個類別，其內容如圖五所示，類別編號和其類別的中文字詞以逗號隔開。基於效率的問題，我們一樣只針對尚未對列之中文字詞進行擴展。第一步、將尚未進行連結的中文字詞查詢同

表 9 詞彙對列(e)

詞序	1/1	2/2	3/5	4/6	5/6	6/6	7/7	**8/3**	9	10/4	11/4	12/8
英文	you	can	change	a	lot	of	functions	**with**	the	remote	control	.
原詞還原							function	**with**				

詞序	1	2	3	4	5	6	7	8
中文	你	可以	用	遙控器	改變	許多	功能	。
翻譯候選字			以 給 受 花費 用項 …	查詢 CEDICT 辭典	/use/according to/so as to/in order to/by/**with**/because/… /to/for/for the benefit of/give/allow/… /receive/accept/suffer/subjected to/… /expense/cost/spend/expenditure/ /items of expenditure/expenditures/ …			

表 10 詞彙對列完畢之情況

情況	對列結果	遺留字詞	數量
完全對列 (無任何遺留字詞)	你 爲什麼 站 在 這裡 ？ Why are you standing here ?		0
只遺留中文字詞	不斷 地 在 生活 中 學習 。 Learning constantly from everyday life .	地 中	2
只遺留英文字詞	價錢 一定 很 高 。 The price must be very high .	The	0 1
中英都有遺留字詞	因爲 它 有 很多 咖啡因 。 Because it has a lot of caffeine .	很多 a lot of	1 3

表 11 對列結果統計

情況	數量	整體比例(%)
完全對列(無任何遺留字詞)	21	0.2
只遺留中文字詞	194	2.0
只遺留英文字詞	107	1.0
中英都有遺留字詞	9631	96.8

義詞詞林，以得到同義詞群編號。第二步、以此同義詞編號，一樣透過同義詞詞林來獲得中文同義詞。如表 8 所示，中文字詞「用」可以在第一步中得到七個同義詞群編號，在第二步時，則利用這七個編號，去查出屬於該類別的中文字詞，最後可以得到相似於「用」的同義詞。將這些中文同義詞透過漢英辭典，進行相對應翻譯字詞的查詢，之後逐字與英文字詞和原詞還原進行比對，若有相符的字詞，一樣標記上其對應的中文詞序，整體的最後結果以表 9 爲例。最後我們將沒有成功對列的英文字詞「the」，以「0」表示它無法找到其對應的中文字詞。

3.2 遺留字詞利用

中英文平行句對經過詞彙對列之後的結果，主要分成「完全對列」、「只遺留中文字詞」、「只遺留英文字詞」和「中英都有遺留字詞」共四種情況，如表 10 所示。然而，如表 11 所示，在高中英文共 9954 組平行句對中，實際上「完全對列」的句對數量相當稀少。但從另一方面來想，若其中英文句子確實互爲翻譯句，且平行句對並無過多的翻譯潤飾，那我們便大膽假設這些遺留字詞其實是互爲對列的，因此我們想從其他三個情況探討是否有可再度利用的資訊，有助於我們得到更多「完全對列」的句對數目。

3.2.1 停用詞列表與遺漏詞修補

針對「只遺留中文字詞」的情況，其代表的是詞彙對列技術有誤？還是屬於中文語法中的語助詞？(代表爲虛詞的一種，常置於句子的尾端或是在句子中間，可以表示特定的語氣，或是當作暫時停頓的功能。如：了、呢、嗎、乎、也等。[1])，我們將此類的字詞頻率統計後，其結果如表 12 所示。

表 12 中文遺留字詞次數統計

中文字詞	次數
的	55
個	11
在	11
了	11
嗎	7
要	6
都	6
…	…

[1] 引用自教育部重編國語辭典修訂本
http://dict.revised.moe.edu.tw/cgi-bin/newDict/dict.sh?idx=dict.idx&cond=%BBy%A7U%B5%FC&pieceLen=50&fld=1&cat=&imgFont=1

以次數最高的中文字詞「的」爲例，其隸屬於兩大類的詞性，分別爲結構助詞和句尾助詞。結構助詞通常出現在各種詞性的後面，如形容詞(美麗的女孩)、名詞或代名詞(如：我的實驗報告)、修飾片語或是子句(她看過的那個人)和副詞(慢慢的說)；句尾助詞則用來表示肯定或是加強的語氣，例如「這樣子是不對的！」這類的句型。

在此我們提出兩種看法，一種是斷詞系統的錯誤，另一種代表「的」真的屬於中文句尾助詞。經過人工觀察的結果，我們發現中研院斷詞系統時常把「的」，斷成一個獨立的字詞；也就是說，中研院斷詞系統時常把帶有「的」之詞彙斷成兩個字詞，如「開心的」會被斷成「開心」和「的」、「我們的」會被斷成「我們」和「的」，也因爲這種狀況，將讓詞彙對列模組在進行辭典查詢時，會有一些失誤的情況。所以我們重新對中研院斷詞系統所產生的字詞集合，進行一些修正。主要的步驟是重新檢視斷詞後的結果，當「的」被斷爲單一字詞後，系統會試探性地將「的」與前一個字詞合併，接著與辭典內部的資訊進行比對，假若合併後的字詞有出現在辭典內部，則把兩個字詞串連起來。

以表 13 爲例，在進行修補之前，我們的系統會依序把中文字詞透過辭典查詢，並以同義詞詞林來協助我們獲得更多對列的線索。在這一個例子中，「你的」是正確的斷詞，如果被斷成「你」和「的」的話，就會因爲我們的字典裡面「的」可以是「的確」的意思，又「的確」的一個英文翻譯是「really」，所以會把獨立的「的」連結到「really」，當中英詞彙間產生連結之後，便表示以後的字詞便不能進行干涉，而失去「really」其實應該和「真是」產生正確的對應。而透過修補的動作，我們便可以避開這種錯誤的連結，把「你」和「的」這兩個中文字詞綁在一起，並且與「your」成功的連結在一起，並且完成「really」和「真是」之間的連線。

「只遺留英文字詞」，這類的情況與「只遺留中文字詞」類似，我們一樣討論次數較高的英文字詞，統計所遺留的英文字詞次數，如表 14 所示。其中最高次數的字詞爲「the」，在 107 句只遺留英文字詞的情況中，將近有八成的句對無法將「the」產生正確的連結。以「the」爲例探討其所屬的詞性，分別常見於冠詞類別(article)或是副詞類別(adverb)[3]。其中「the」當作冠詞時，常可翻譯成「這(個)、那(個)、這些、那些」，這些翻譯大多數都出現在名詞之前。而實際上發現，平行句對的中文句子有一定比例不對「the」進行翻譯，如「You can change a lot of functions with the remote control .」其中文翻譯可爲「你可以用遙控器改變許多功能。」，或是「你可以用『這個』遙控器改變許多功能。」。實際上，在這種情況下「the」應該和「remote control」合併，並且與中文字詞「遙控器」進行對列，如表 15 所示。

表 13 利用中文遺漏詞進行修補

若無進行修補								
東尼	，	你	的	房間	真是	亂	。	
Tony	,	your	room	is	really	messy	.	
修補之後								
東尼	，	你的	房間	真是	亂	。		
Tony	,	your	room	is	really	messy	.	

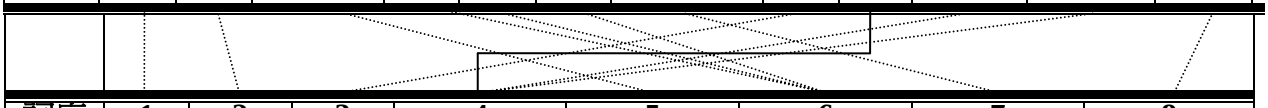
表 14 英文遺留字詞次數統計

英文字詞	次數
the	81
to	55
a	27
is	17
that	16
…	…

[3] 引用自譯典通線上詞典 http://www.dreye.com:8080/axis/ddict.jsp?ver=big5&dod=0102&w=the

表 15 利用英文遺漏詞進行修補

詞序	1/1	2/2	3/5	4/6	5/6	6/6	7/7	8	9	10/4	11/4	12/8
英文	you	can	change	a	lot	of	functions	with	the	remote	control	.
原詞還原							function	with				



詞序	1	2	3	4	5	6	7	8
中文	你	可以	用	遙控器	改變	許多	功能	。

依照表 11 所表示，絕大多數平行句對的詞彙對列結果都屬於「中英都有遺留字詞」的情況，主要是因爲我們的詞彙對列模組以漢英辭典爲基礎，所以假若其英文字詞並未收錄在辭典內部，就一定無法產生正確的對列。如同
表中的例句所表示的，中文句子的遺留字詞爲「很多」，英文句子所遺留的字詞爲「a lot of」，其實應該是互爲對列的組合，但是因爲目前的辭典，對於「很多」這個中文字詞，只有如表 16 這些翻譯。儘管我們透過同義詞詞林，依舊無法得到我們所期望的翻譯-「a lot of」。

如何利用剩餘字詞，使詞彙對列更佳完善？實際上詞彙對列結束後，「中英都有遺留字詞」的組合情況共有四種，如表 17 所示。由於我們的模組主要是以中文字詞當作索引詞，且可以完成一個中文字詞對應多個英文字詞的對應方式，所以在這邊還將英文句子中連續的遺留字詞視爲一個單位，與中文字詞進行配對。

3.2.2 對列計算

在此我們參照曾元顯等[10]所使用的計分公式，來篩選「中英都有遺留字詞」之詞對組合，以便擴張辭典，重新進行詞彙對列模組，達到提升詞彙對列整體的效果。在[10]中，作者利用大量專利文書的平行語料，擷取出互爲翻譯的詞彙，使得既有的上百萬條雙語詞庫，再增加約 20%的新詞彙。在此我們選用其中五個詞對擷取的評估公式，分別爲點互

表 16 辭典內部不慎完美之資訊

中文字詞	翻譯字詞
很多	/very many/very much/great/

表 17 遺留字詞的情況

遺留字詞	對列結果	遺留字詞
一對一	我 在 想 她 是 誰 。	想
	I am wondering who she is .	wondering
一對多	眼見爲憑 。	眼見爲憑
	Seeing is believing .	Seeing is believing
多對一	我們 用 有毒 的 化學 藥品 來 殺死 老鼠 。	的 化學 藥品
	We use poisonous chemicals to kill rats .	chemicals
多對多	自助 旅行 使 你 可以 遇見 當地 的 人 。	自助 使 可以 的
	Self-arranged traveling enables you to meet local people .	Self-arranged enables

表 18 實驗語料來源統計

語料	語言	句對數目	詞彙數目	總詞彙數目(tokens)	平均句長
國中平行語料	中文	7360	4876	57346	7.8
	英文		6420	59767	8.1
高中平行語料	中文	9954	10679	120509	12.1
	英文		12867	130908	13.2
科學人雜誌	中文	2685	9279	70411	26.2
	英文		10504	68434	25.5

訊息(pointwise mutual information)[17]、相關分析(correlation coefficient)[17]、可能性比例 (likelihood ratios)[17] 、 Dice 係數 (Dice coefficient)[16] 和分數累積 (fractional count,*FC*)[10]，其中累積詞對組合在平行句對中出現的翻譯機率(並非次數)，則是這邊所指的分數累積。其公式表達如下：

$$FC(c,e) = \sum_{for\ all\ i, s.t.(c,e)\in sp(i)} P_{sp(i)}(c,e) \quad (1)$$

公式(1)中的 sp(i)代表第 i 組中英句對中，同時出現中文字詞 c 和英文字詞 e，而 $P_{sp(i)}(c,e)$則表示第 i 組句對中(c,e)互爲詞對組合的機率。我們以表 17 中，多對多情況的句對作爲例子，我們從中發現中文遺留字詞共有四個(自助、使、可以、的)，英文遺留字詞有兩個(Self-arranged、enables)，我們將其交錯配對，並且以中文遺留字詞的數量作爲分母，如＜自助－Self-arranged＞、＜自助－enables＞這兩個組合，各別可獲得 1/4 的分數，其他的遺留字詞透過一樣的方式進行計算，最終將統計訓練語料中所有的遺留字詞進行累加的動作，透過分數累積的計算，當分數越高時，代表兩詞彙的相關性也越高。其他計分方法的實驗結果在第 4.3.2 節。

4. 系統效果評估

4.1 實驗語料來源

本實驗所使用的平行語料有三大部分，分別爲國中英文語料[6]-[9]、高中英文語料和科學人雜誌中英對照電子書[5](以下簡稱爲科學人雜誌)，將分別進行詞彙對列處理，並且利用對應率當作門檻值加以組合進行測試，實驗語料之統計如表 18 所示。國中英文語料的部分主要是從多個線上資源整理而來，我們從該網頁以人工方式擷取出中英平行句子，其內容多是課文、試題和例句等語料。高中英文語料則是第 2 節中所介紹之平行語料。科學人雜誌則是彙整了從 2002 年 3 月創刊號至 2006 年 12 月共 110 篇，並利用呂明欣[3]簡易中英語句對列，從中取出 2685 句。

4.2 實驗設計

我們試圖計算詞彙對列的召回率和精確率，但是由於三份訓練語料(國中英文、高中英文和科學人雜誌)皆無正確之對列答案，因此我們並無法利用自動化的程序來計算出召回率和精確率。參考[15]文獻中所提出的方法，隨機從詞彙對列的結果中取出一定的比

例，再由人工來進行檢測，以句爲單位計算其召回率和精確率，如公式(2)、公式(3)所示。我們將會把各句所計算出的數值進行加總，並且除上挑選出來的句對數，以代表該語料的平均召回率和精確率。

$$召回率=\left(\frac{正確的對列數量}{英文字詞總數目}\right) \quad (2) \qquad 精確率=\left(\frac{正確的對列數量}{模組產生的對列數量}\right) \quad (3)$$

4.3 實驗結果與分析

4.3.1 未利用遺留字詞之對列結果比較

依實驗設計流程，隨機抽選數句的對列結果，觀察第一階段(未利用遺留字詞)詞彙對列模組所產生的連線正確性，計算其召回率與精確率。一開始我們先針對同義詞擴充來進行分析，由於同義詞擴充可以使我們獲得更多的對列機會，但相對的也可能獲得錯誤的對列結果，結果如表 19 所示。從數據可以知道，我們僅犧牲了 4%的精確率，換得召回率 16%的提升。而在不同語料方面，整體結果如表 20 所示，對列結果以國中語料效果最佳，接著依序爲高中語料和科學人雜誌。由於在對列的過程中，我們是以辭典內部的資訊做爲基礎，再對尚未完成對列的字詞，以原詞還原和同義詞詞林進行擴充，所以平均精確率可以達到九成。

4.3.2 遺留字詞之結果分析

以高中語料爲例，將已完成詞彙對列的 9954 組中英平行句對，透過 **3.2.2** 節中所述方法進行對列計算。我們將實驗分爲兩個方向，首先是一個中文字詞對一個英文字詞，這樣子的組合總共得到了 115000 餘個詞彙組合，並且依照五組評估公式所產生的數值進行排序，由於詞彙組合過度的龐大，不可能依序進行人工檢驗，因此我們選出各組評估公式的前 100 名進行人工檢驗。人工檢驗的評分方式採，「完全正確」、「部分正確」和「錯

表 19 有無同義詞詞林之比較

語料	句對總數	抽選句數	平均召回率	平均精確率
高中語料	9954	500	62%	91%
高中語料(未使用同義詞)	9954	500	46%	95%

表 20 隨機抽選句數與其對列結果

語料	句對總數	抽選句數	平均召回率	平均精確率
國中語料	7360	400	71%	95%
高中語料	9954	500	62%	91%
科學人雜誌	2685	50	54%	84%

表 21 人工檢測排序後前 100 名組合(一對一)

	相關分析	點互訊息	分數累積	Dice 係數	可能性比例
完全正確	34	35	3	38	37
部分正確	15	14	0	15	10
錯誤	51	51	97	47	53
正確率(完全正確/100)	34%	35%	3%	38%	37%

表 22 人工檢測排序後前 100 名組合(利用停用詞列表過濾)

	相關分析	點互訊息	分數累積	Dice 係數	可能性比例
完全正確	34	35	82	38	37
部分正確	15	14	15	15	10
錯誤	51	51	3	47	53
正確率(完全正確/100)	34%	35%	82%	38%	37%

表 23 人工檢測排序後前 100 名組合(一對多)

	相關分析	點互訊息	分數累積	Dice 係數	可能性比例
完全正確	19	14	77	21	21
部分正確	16	23	10	18	20
錯誤	65	63	13	61	59
正確率(完全正確/100)	19%	14%	77%	21%	21%

誤」三種等級，「完全正確」透過谷歌線上辭典進行人工比對，如＜羅密歐－romeo＞、＜漁業－fishing＞這類我們則判定爲完全正確，「部分正確」則表示其英文遺留字詞與中文遺留字詞有部分符合，如＜車禍－car＞、＜蛋塔－tarts＞此種情況，其餘我們則歸類爲「錯誤」。統計情況如表 21 所示。從表 21 中發現，使用分數累積做爲評估公式時，相較於其他四種評估公式，前 100 個詞彙組合只有三組是正確的。其最主要的原因爲大量停用詞，如中文字詞「了」、「個」和「的」，英文字詞「the」、「to」和「a」占據了前面的排名。

因此，我們將 3.2.1 節中所得到的停用詞列表當作過濾條件，針對排序中的詞對，只要是停用詞列表內部的詞彙就進行刪除。結果得到了 56000 餘個詞彙組合，並重新檢測前 100 個詞彙組合的正確率，結果如表 22 所示。

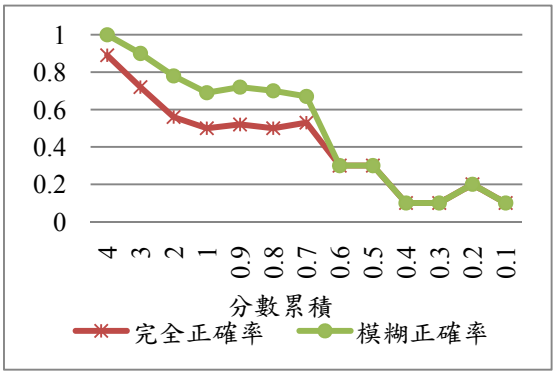
我們發現分數累積獲得相當大的改善，前 100 名的詞彙組合近乎八成二是完全正確的，並且遙遙領先其餘的評估公式。而另一個方向爲一個中文字詞對多個英文字詞組合，如同一對一的做法，在 9954 組中英文平行句對中可獲得 24000 餘個詞彙組合，我們一樣對前 100 名的組合進行人工檢測，結果如表 23 所示。

從表 22 和表 23 發現，目前效果最佳的對列計算公式爲「分數累積」，我們進一步的去對分數累積進行人工檢測，依照分數累積的數值當作區間，我們對超過 1 的詞彙組合全部進行人工檢測(一對一共有 468 組詞對，一對多共有 1333 組詞對)，其餘詞彙組合以 0.1 作爲分組組距進行抽樣檢測，並且計算完全正確率(完全正確／檢測數)和模糊正確率((完全正確＋部分正確)／檢測數)，如圖六和圖七所示。

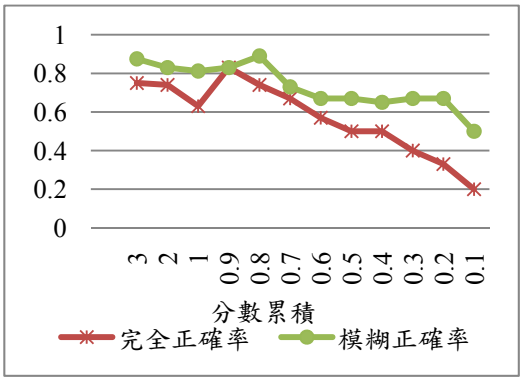
從兩個圖的趨勢看來，當分數越來越小時，整體的正確率也逐漸變低。在此，我們將對不同程度的語料，依「分數累積」超過 1 做爲門檻值，將符合門檻值的詞對組合當作新的翻譯詞組，主要是因爲當「分數累積」數值超過 1 的同時，便表示是透過兩組句對以上所累積出來數值，我們更大膽的假設這類的情況並非偶然，這些新詞組將會整合至原始的辭典，以改善詞彙對列模組的效能。

4.3.3 以遺留字詞修正詞彙對列結果比較

接著我們以上一節中所說明之步驟，進行詞彙對列模組的補強。評估方式與之前的步驟相同，隨機挑選出一定數量的對列結果，進行人工檢測的動作，並且將其召回率和精確率與補強前的結果進行比較，整體的實驗結果如圖八所示。我們可以看到，在國中語料

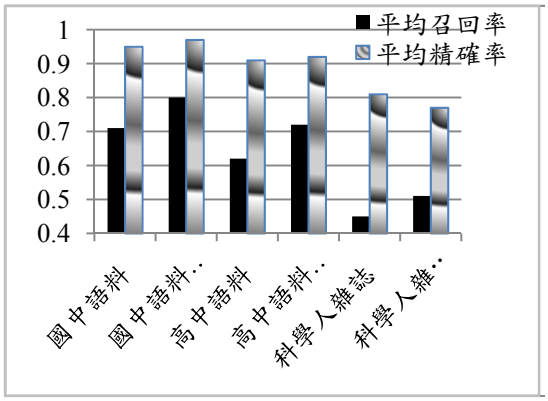


圖六 一對一正確率趨勢圖



圖七 一對多正確率趨勢圖

和高中語料的部分，召回率都進步了約一成左右，主要是因爲我們從遺留字詞中所獲得的新詞組，是由語料本身中所擷取出來的，如＜節食－going on a diet＞、＜上網－on the internet＞和＜想起－comes to mind＞，故這些新詞組有助於當初無法對應的字詞進行對列，進而提升召回率和精確率的分數。這樣子的機制，可以從不同的訓練語料中得到不同的詞彙組合，再透過評估公式的篩選，從中獲取可靠度較高的詞彙組合，進而提升整體的對應結果。然而，這樣子的機制在科學人雜誌上，表現並非同樣亮眼。其主要的原因爲，科學人雜誌所提供的中英文平行語料難度較高，在進行第一階段的詞彙對列時，並無法有效地先將大多數詞彙進行過濾，產生如＜導致－strength＞、＜美國－it's＞和＜自尊－egotism＞這些錯誤的詞對組合，因而無法得到品質較高的詞彙組合，以輔助詞彙對列模組的進行，甚至還產生了錯誤的訊息，使得精確率因此下降。(我們並沒有另外計算 false positive 和 false alarm 兩項指標，目前只有計算召回率和精確率。)



圖八 召回率與精確率之綜合比較

5. 結論

本研究的目的是針對中英平行語料進行詞彙對列，系統可以讓使用者輸入平行句對，透過循序漸進的方式來完成詞彙對列的動作，更可透過詞對擷取分析公式來對遺留字詞產生的新詞組進行過濾，進而擴充原始的辭典。本研究最終期望在找出語言之間詞序的交互關係，以供後端翻譯系統進行詞序的調動。

由實驗數據顯示，對列的效能會因爲平行語料的難度而有所差異，針對國中語料，詞彙對列整體的召回率最高可達到八成，準確率則約有九成。但若對科學人雜誌進行詞彙對列，效果則不盡理想。事實上我們手邊並沒有正確的對列答案，僅只能透過人工檢

驗來進行評比，而我們必須對更多的對列結果進行檢測，才能得到更客觀公正的數據。我們也將改良過後的詞彙對列模組，整合至張智傑在 2007 年所建構的輔助式試題翻譯系統[11]，其 BLEU 和 NIST 僅微幅提升約 0.03。

總結整體的工作目的，是將原始詞彙對列技術進行改良，希望能獲得漢英之間詞序的交互關係，以輔助翻譯系統的效能，甚至是透過大量的平行語料，來獲得新的詞對組合，以擴充系統所依賴的漢英辭典，進而推廣至其餘需要新詞彙的研究工作。

致謝

本計畫承蒙國科會研究計畫案 NSC-97-2221-E-004-007 與 NSC-99-2221-E-004-007 補助，謹此致謝。我們感謝匿名評審的寶貴意見，雖然我們一時無法在有限的頁數之內回應所有意見(特別是關於相關研究的評比)，但是仍將在未來工作中，努力進行評審所提出的建議工作。

參考文獻

[1] 三民學習網，http://www.grandeast.com.tw/Englishsite/，[2010/08/15]。
[2] 中央研究院中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw/，[2010/08/15]。
[3] 呂明欣，*電腦輔助試題翻譯：以國際數學與科學教育成就調查爲例*，國立政治大學資訊科學所，碩士論文，2007。
[4] 哈爾濱工業大學訊息檢索實驗室同義詞詞林擴充版，http://www.nlp.org.cn/docs/doclist.php?cat_id=9&type=7，[2010/08/15]。
[5] 科學人雜誌中英對照電子書，http://edu2.wordpedia.com/taipei_sa/，[2010/08/15]。
[6] 旋元佑文法，http://tw.myblog.yahoo.com/jw!GFGhGimWHxN4wRWXG1UDIL_XSA—/，[2010/08/15] 。
[7] 基礎英文1200句，http://hk.geocities.com/cnlyhhp/eng.htm，[2010/02/25]。
[8] 國民中學學習資源網，http://140.111.34.172/teacool/new_page_2.htm，[2010/08/15]。
[9] 教育部委託宜蘭縣發展九年一貫課程教科書補充資料及題庫，http://140.111.66.37/english/，[2010/08/15]。
[10] 曾元顯、劉昭麟和莊則敬，專利雙語語料之中、英對照詞自動擷取，*第二十一屆自然語言與語音處理研討會論文集*，279-292，2009。
[11] 張智傑，以範例爲基礎之英漢TIMSS試題輔助翻譯，*第二十屆自然語言與語音處理研討會論文集*，308-322，2008。
[12] M. H. Bai, J. S. Chang, K. J. Chen and J. M. You, “Acquiring Translation Equivalences of Multiword Expressions by Normalized Correlation Frequencies,” *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 478-486, 2009.
[13] J. S. Chang and S. J. Ker, “A Class-Based Approach to Word Alignment,” *Computational Linguistics*, 23(2), 313-343, 1997.
[14] CEDICT漢英電子字典檔，http://us1.mdbg.net/chindict/chindict.php，[2010/08/15]。
[15] H. Isahara and M. Utiyama, “A Japanese–English Patent Parallel Corpus,” *Proceedings of the Eleventh Machine Translation Summit*, 475-482, 2007.
[16] C. D. Manning, P. Raghavan and H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
[17] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.
[18] R. Mihalcea and T. Pedersen, “An Evaluation Exercise for Word Alignment,” *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, 1-10, 2003.
[19] M. F. Porter, “An Alogrithm for Suffix Stripping,” *Program*, 130-137, 1980.
[20] The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex–parser.shtml, [2010/08/15].