Computational Linguistics and Chinese Language Processing
Vol. 25, No.2, December 2020, pp. 37-54

改善詞彙對齊以擷取片語翻譯之方法

Improving Word Alignment for Extraction Phrasal Translation

陳怡君*、楊謦瑜+、張俊盛*

Yi-Jyun Chen, Ching-Yu Helen Yang and Jason S. Chang

摘要

本研究專注於從雙語語料庫中自動擷取英文名詞與介系詞搭配的中文翻譯及例句，其結果可用於改善機器翻譯或提供語言研究者撰寫文法規則之參考。本方法使用統計方法由雙語語料庫中的詞彙自動對齊，分別擷取名詞及介系詞的翻譯，再根據中文搭配詞，將名詞及介系詞的翻譯做適當調整，並產生例句。本研究的評估方式是隨機抽取三十組名詞及介系詞的搭配，人工評估本研究方法產生的翻譯。

Abstract

This thesis presents a method for extracting translations of noun-preposition collocations from bilingual parallel corpora. The results provide researchers a reference tool for generating grammar rules. In this paper, we use statistical methods to extract translations of nouns and prepositions from bilingual parallel corpora with sentence alignment, and then adjust the translations according to the Chinese collocations extracted from a Chinese corpus. Finally, we generate example sentences for the translations. The evaluation is done using randomly 30 selected phrases. We used human judge to assess the translations.

*國立清華大學資工系
Department of Computer Science, National Tsing Hua University
E-mail: {yijyun; chingyu; jason}@nlplab.cc
+國立中興大學外語系
Department of Foreign Languages and Literatures, National Chung Hsing University

Keywords: Word Alignment, Grammar Patterns, Collocations, Phrase Translation

關鍵詞：雙語詞彙對齊、文法規則、搭配詞、片語翻譯

1. 簡介 (Introduction)

介系詞在英文使用頻率很高，因爲中英文性質不同，翻譯的變化非常多元，所以當我們將英文實詞與介系詞的搭配翻譯至中文時，會有很複雜的情況，以下列句子為例：

(1) a. In her **speech on** the motion of thanks , the hon margaret ng touched upon ...

b.吳靄儀 議員 **就** 致謝 議案 **發言** 時 ， 曾 談及 …

(2) a. ... the extent of business via ec was still relatively limited , so was its **impact on** the statistical systems .

b. …涉及 電子 貿易 的 業務 範圍 仍 相對 有限 ， **對** 統計 系統 的 **衝擊** 也 有限 。

(3) a. **impact on** the financial market

b.**衝擊** 金融 市場

(4) a. Mr downer believed the close **relationship between** hong kong and australia would continue to strengthen .

b.唐納 相信 ， 澳洲 與 香港 **之間** 的 密切 **關係** 是 會 繼續 加強 的 。

(5) a. up to now , the change in **relationship between** china and hong kong can be divided into three stages .

b.直至 目前 ， 中港 **關係** 的 變化 ， 可 分為 三 個 階段 。

相同的介系詞在搭配不同名詞時，可能會有不同的翻譯，例如在例句(1a)中的「speech on」，翻譯至例句(1b)中的「就 … 發言」，此時「on」翻譯為「就」，而在例句(2a)中的「impact on」，翻譯至例句(2b)中的「對 … 衝擊」，此時「on」翻譯為「對」。有時介系詞也可能在翻譯時被省略，例如例句(4a)中的「relationship between」翻譯為例句(4b)中的「之間 … 關係」，但例句(5a)中的「relationship between」，在例句(5b)中則是僅翻譯成「關係」。另外，在某些情況下，介系詞也可能和名詞一起翻譯為一個動詞，例如例句(3a)中的「impact on」翻譯為例句(3b)中的動詞「衝擊」。這些不同翻譯變化對於語言學習者或是機器翻譯演算法來說，是一個難解的議題。

本研究的目標，是由英文名詞和介系詞搭配（例如「speech on」），產生其中文翻譯（例如「就...發言」）。現有的自動對齊演算法雖可自動產生雙語平行語料的詞彙對應(word alignment)，但部分對應不準確，若直接以自動對齊的結果，來提取英文搭配的中文翻譯，將會產生許多錯誤的結果，我們認為若同時考量英文對中文的對應與反向的對應，對應的準確度將會更高，另外，在自動對齊的結果中，有時英文搭配所對應到的中文翻譯，若演算法能同時考量其是否符合中文搭配的性質，就能做出更正確的英文搭配翻譯。因此在本研究中，我們除了使用雙語自動對齊的平行語料庫，也使用中文單語語料庫，在平行語料庫中，我們不僅統計英文對中文的對應，也統計中文對英文的對應，同時，我們也在中文單語語料庫中，統計並篩選中文的高頻搭配，藉由結合中英對應的結果與中文搭配的性質，取得更精確的翻譯。

我們的訓練階段主要有四個步驟，依序是改善文本斷詞與對齊、產生實詞翻譯、產生介系詞翻譯、產生實詞介系詞搭配的翻譯。在改善文本斷詞與對齊這個步驟，我們統計雙語詞彙對應機率並計算各種不同斷詞及對應方式的機率，修正一些斷詞及對齊的錯誤；在產生實詞翻譯這個步驟，我們統計英文實詞所對應到的中文詞彙及次數，以及這些中文詞彙的反向對應，篩選中文詞彙做為實詞的翻譯；在產生介系詞翻譯這個步驟，我們計算當介系詞出現在實詞後面時對應至各中文詞性及詞彙的機率，並從中篩選中文詞彙做為介系詞的翻譯；在產生實詞介系詞搭配的翻譯這個步驟，我們由中文語料庫統計中文高頻搭配詞，用中文高頻搭配詞組合實詞的翻譯和介系詞的翻譯，產生比較符合搭配性質的翻譯，並再回到雙語語料庫中檢驗所產生的翻譯及篩選例句。

本研究所產生的翻譯，可用於協助語言學習者學習中文或英文，或為語言研究者提供撰寫文法規則的參考，也可輔助改進機器翻譯系統。

2. 相關研究 (Related Work)

機器翻譯一直是自然語言處理領域中的活躍研究領域，過去幾十年中，大量的雙語語料庫資源，使得統計式機器翻譯越來越可行，在 1990 年代，雙語句子對齊技術快速發展(Gale & Church, 1991a, 1991b, 1993; Brown, Lai & Mercer, 1991; Simard, Foster & Isabelle, 1992; Chen, 1993)。

除了找出相對應的雙語句子(Debili & Sammouda, 1992; Kay & Roscheisen, 1993)，有些研究使用統計模型以改善自動對齊所產生的詞彙對應，如隱藏式馬可夫模型(Hidden Markov Model, HMM) (Brown, Lai & Mercer, 1991) 、對數似然比(log-likelihood ratio) (Gale & Church, 1991a, 1993)，以及 K-Vec algorithm (Fung & Church, 1994)。奠基在前人基礎之上，Melamed (1999) 提出 the Smooth Injective Map Recognizer (SIMR)，將雙語片語對齊視為 x 軸與 y 軸在二維空間的最佳分佈，與前人研究不同之處在於，SIMR 採貪婪演算法，以最小可能運算出最佳分佈的二維空間為演算單位。SIMR 分為兩個階段，生成階段產生所有可能雙語對應（x 軸與 y 軸對應），以及辨識階段已選出最佳對應。然而，句子對齊技術仍有其限制，難免有許多錯誤的對應。

本研究是專注於使用統計方法擷取名詞介系詞搭配的翻譯，目的是獲得精確的翻譯。過去與本研究較相關的研究，主要集中於學習單詞翻譯並從現有語料庫中提取雙語單詞翻譯對(Catizone, Russell & Warwick, 1989; Brown *et al.*, 1990; Gale & Church, 1991a; Wu & Xia, 1994; Fung, 1995; Melamed, 1995; Moore, 2001)，計算平行句子中的單詞對之間的相互關係程度，從而推導出翻譯。我們提出了一種結合了雙語對齊統計和中文搭配詞統計的方法。

在 2006 年柯明憲(柯明憲，2006)提出的「雙語語料庫之多字詞語對應」論文，與本研究有類似目標，皆為擷取片語翻譯。該論文是針對特定動詞片段（例如：make a report to police），使用自動對齊技術計算雙語之間的搭配關係（例如：當「make」和「report」一起出現時，「report」常對應至「報案」），再於雙語句子中擷取對應（例如：make a report to police 的對應「向警察報案」），並統計翻譯。

本研究與上述論文同樣使用到自動對齊技術，與上述論文不同的是，本研究是針對名詞和介系詞的搭配，並且在計算完詞彙機率後，並非直接由句子擷取對應，而是計算出中文高頻搭配詞組合名詞與介系詞的翻譯後，再回到雙語語料庫中以句子篩選最合適的翻譯。本研究提出的新方法，成功達到較高的翻譯正確率，大量篩選掉錯誤的翻譯，減少查詢翻譯資訊的困難。

3. 方法 (Method)

在本章中，我們會說明如何進行資料前處理，以改善其斷詞和雙語詞彙對應（第（一）小節），接著分別詳細描述如何從已自動對齊的雙語平行語料庫中統計並篩選實詞和介系詞兩者個別的翻譯（第（二）、（三）小節），最後說明如何統計中文搭配詞以及使用中文搭配詞的統計結果來計算兩者精確翻譯（第（四）小節）。

3.1 改善斷詞及雙語詞對應 (Improving Word Segmentation and Word Alignment)

本階段的目標是改善資料的斷詞和詞彙對應，提升詞彙對應準確率。本階段的輸入為已標註中文斷詞及雙語自動對齊的雙語平行資料。此標註斷詞和詞彙對應的資料仍有不少錯誤，表 1 為句子範例，其中「中英詞對齊」代表中英詞彙的對應位置（由 0 起算），例如「4-0」代表中文句的第 4 個詞彙「反應熱」對應至英文句的第 0 個詞彙「enthusiastic」，在此句中，「反應熱烈」斷詞為「反應熱｜烈」，然而正確斷詞應為「反應｜熱烈」，也因為斷詞錯誤，導致詞彙對應的錯誤：

表 1. 英中詞對齊錯誤範例
[Table 1. An example of wrong word alignments of English and Chinese sentence]

中文例句	社區 投資 共享 基金 <u>反應熱 烈</u>
英文例句	**<u>enthusiastic response</u>** to community investment and inclusion fund
中英詞對齊	**<u>4-0 5-0 5-1</u>** 0-3 1-4 2-5 2-6 3-7

我們從資料中統計每個中文詞彙對應至每個英文詞彙的機率 P_{CtoE}，以及每個英文詞彙對應至每個中文詞彙的機率 P_{EtoC}，接著針對每一組中英詞彙的二對二對應，產生各種可能的斷詞及對應方式，並計算機率，取出機率最高者。以表 1 的句子為例，中英詞對齊中的「4-0 5-0 5-1」表示「反應熱」對應至「enthusiastic」、「烈」對應至「enthusiastic」和「response」，故「反應熱｜烈」共對應至「enthusiastic」和「 response」兩個英文詞彙，為一組二對二對應，表 2 即為產生之各種可能的斷詞及對應方式，經過計算機率後，取出機率最高的斷詞及對應方式，即斷為「反應｜熱烈」，且反應對應至「response」、「熱烈」對應至「enthusiastic」。

本階段的輸出為改善之後的資料，我們以上述方法將整份資料中的這些詞彙進行修正，有效減少部分詞彙的錯誤對應。

表 2. 斷詞及對應方式機率計算範例
[Table 2. An example of probability calculation for word segmentation and word alignment]

斷詞方式	對應方式	機率計算
反｜應熱烈	反 → enthusiastic 應熱烈 → response	P_{CtoE}(反,enthusiastic) × P_{CtoE}(應熱烈,response) × P_{EtoC}(enthusiastic,反) × P_{EtoC}(response,應熱烈)
	反 → response 應熱烈 → enthusiastic	P_{CtoE}(應熱烈,enthusiastic) × P_{CtoE}(反,response) × P_{EtoC}(enthusiastic,應熱烈) × P_{EtoC}(response,反)
反應｜熱烈	反應 → enthusiastic 熱烈 → response	P_{CtoE}(反應,enthusiastic) × P_{CtoE}(熱烈,response) × P_{EtoC}(enthusiastic,反應) × P_{EtoC}(response,熱烈)
	反應 → response **熱烈 → enthusiastic**	P_{CtoE}(熱烈,enthusiastic) × P_{CtoE}(反應,response) × P_{EtoC}(enthusiastic,熱烈) × P_{EtoC}(response,反應)
反應熱｜烈	反應熱 → enthusiastic 烈 → response	P_{CtoE}(反應熱,enthusiastic) × P_{CtoE}(烈,response) × P_{EtoC}(enthusiastic,反應熱) × P_{EtoC}(response,烈)
	反應熱 → response 烈 → enthusiastic	P_{CtoE}(烈,enthusiastic) × P_{CtoE}(反應熱,response) × P_{EtoC}(enthusiastic,烈) × P_{EtoC}(response,反應熱)

3.2 篩選實詞翻譯 (Extracting Translations of Content Words)

本階段的目標是產生實詞的翻譯。本階段的輸入為我們要查找的英文實詞（例如：「speech on」中的「speech」）和經前一階段（第三章第（一）節）改善後的資料。由於某些英文搭配經常被包含在更長的片語中（例如「connection with」經常被包含在「in connection with」中，且兩者中的實詞翻譯不同），因此在開始統計英文實詞所對應的中文詞彙之前，我們會先針對輸入的英文搭配，統計出現在其前面的詞是否高機率集中在某些詞，藉此排除英文搭配被包含在更長片語中的狀況。接著，我們統計資料中英文實詞對應至

的中文詞彙及次數，並從中選出次數較高者。由於翻譯後詞性經常出現變化，因此在統計時我們不限制中文詞彙的詞性。以「speech」為例，我們選出以下這些詞彙：

演辭	中	發言	致辭	全文	時
言	講話	發表	預算案	施政	報告
辭	演	演講	講	篇	演說

此步驟所選出的中文詞彙中，仍有許多並非原英文實詞正確的翻譯（如上例中的「預算案」、「施政」等），因此我們會再針對這些中文詞彙統計反向對應，也就是其對應至的英文詞彙及次數，並從中選出次數較高者，做為計算中文詞彙分數之用。以上例中的「演辭」為例，我們會選出以下這些詞彙：

speech	by	speeches
my	his	's

正確翻譯的反向對應時常也會對應到一些和原英文實詞的衍生詞，如複數、動詞變化形等，像是上例的「演辭」除了對應至原來的實詞 「speech」之外，也對應到「speech」的複數「speeches」。我們希望在以反向對應結果計算中文詞彙分數時，將這些情況也考慮進去，因此我們建立並結合了複數表、動詞時態表、動詞名詞變化型態表以及相似詞表，如表 3：

表 3.「discussion」衍生詞表
[Table 3. Derivatives of "discussion"]

與原名詞關係	詞彙
複數	discussions
動詞變化	discuss, discussed, discusses, discussed, discussing
近義詞	conference, argument, consideration, talk, consultation, session...

在一些情況下，中文詞彙對應至某英文詞彙的次數雖然很高，但當該中文詞彙對應至該英文詞彙的時候，大多同時對應到不只一個詞彙。當這樣的狀況發生時，此對應很可能不是正確的翻譯，例如「重視」一詞對應至「importance」的次數相當高，但實際上「重視」並不適合做為「importance」的單詞翻譯，因為當「重視」對應至「importance」的時候，其完整對應多為「attaches great importance to」，而不是單獨對應至「importance」。因此，在統計中文詞的反向對應時，我們會計算當中文詞對應至某個英文詞時，同時對應至多個英文詞彙的機率，並排除此機率過高的對應。

本階段的輸出為英文實詞的翻譯篩選結果。我們以反向對應的統計結果，計算中文詞彙的分數，決定該中文詞彙是否做為原英文實詞的翻譯。若中文詞彙 u 對應至英文詞彙$v_1, v_2, v_3, \dots, v$ ，$Pro(u, v_i)$表示 u 對應至v_i的機率，則 u的分數為

$$Score(u) = \sum_{i=1}^{n} (Pro(u\ \ , v_i) \times f(v_i))$$

若v_i為原輸入實詞則$f(v_i) = 1$；若v_i為原輸入實詞的複數或動詞變化形，由於英文的複數變化及動詞名詞變形翻譯至中文時，經常翻譯成相同的中文詞彙，因此我們仍將$f(v_i)$設為 1 ；若v_i為原輸入實詞的相似詞，因相似詞在翻譯至中文時，可能會有些許差異，故我們將$f(v_i)$設為 0.5；另外由於實詞的正確翻譯也時常對應到與實詞搭配的介系詞，故若v_i為原輸入實詞所搭配的介系詞，我們也將$f(v_i)$設為 0.5 ；其他狀況則$f(v_i) = 0$。經過測試與觀察，我們訂定分數標準為 0.15，即若計算結果大於 0.15，則此中文詞彙入選為原英文實詞的翻譯。

以「speech」的翻譯「發言」為例，下表為「發言」所對應到的與原英文實詞「speech」相關的英文詞彙及機率，則發言的分數為 0.074×1 + (0.044 + 0.037 + 0.305 + ⋯) ×1 + (0.021 + 0.001 + 0.003 + ⋯)×0.5 = 0.531，大於 0.15，故入選為「speech」的翻譯。

表 4.「發言」所對應的英文詞彙及機率
[Table 4. Corresponding English words of "發言" fayan "speech" and the translation probability]

	詞彙	機率
原實詞	speech	0.074
原實詞的複數或動詞變化形	speeches	0.044
	speaking	0.037
	speak	0.305
	spoke	0.018
	speakers	0.003
	spoken	0.033
	speaks	0.002
	speaker	0.001

	詞彙	機率
原實詞的相似詞	address	0.021
	addressing	0.001
	addresses	0.003
	talked	0.0001
	addressed	0.0003
	articulate	4e-05
	voice	0.001
	talk	0.001
	voices	4e-05

3.3 篩選介系詞翻譯 (Extracting Translations of Prepositions)

本階段的目標是產生介系詞的翻譯。本階段的輸入為我們要查找的英文介系詞（例如：「speech on」中的「on」）和上一階段（第三章第（二）節）所輸出的實詞翻譯，以及上上階段（第三章第（一）節）改善後的資料。

我們首先統計英文介系詞出現在該英文實詞後面的時候，對應至每一種中文詞性的次數及機率，以及對應至該詞性的各種中文詞彙的次數。由於介系詞在翻譯時也時常被省略，因此我們除了統計介系詞對應至各詞性的機率，也計算介系詞「沒有對應」（記

為 NULL）的機率。由於在某些介系詞省略翻譯的狀況中，介系詞不會沒有對應，而是會對應至其所搭配的實詞所對應的中文實詞，例如「problem of」中，「of」可能和「problem」一起對應至「問題」，因此我們會將此類狀況也列入介系詞沒有對應的機率。以「problem of」為例，下表為「problem of」中的「of」的統計結果：

表 5.「problem of」的「of」翻譯結果
[Table 5. Chinese translations results of "of" in "problem of"]

詞性	機率	詞彙及次數
DE	0.69	的(3098) 之(14)
NULL	0.28	NULL(1247)
T	0.01	的(58)
Na	0.01	工作(7) 人數(6) 程度(5) 精神(5) 的(5) 過程(3) 成員(2) 問題(2) ...

介系詞的正確翻譯大多為特定詞性，因此許多錯誤的對應源自對應到錯誤的詞性，因此在統計之後，我們首先以詞性做篩選，我們認為較合理的詞性有「DE」（如「的」、「之」）、「P」（如「在」、「對」）「Ng」（如「上」、「之間」）、「Caa」（如「與」、「和」）。以上表的「problem of」為例，經過詞性篩選後，只會留下詞性「DE」。

然而在某些情況下，介系詞所對應的詞性會較為特殊，例如「action against」可翻譯為「反對...的行動」，其中介系詞「against」翻譯為「反對」，但「反對」是動詞，不屬於上述我們認為合理的詞性，因此若僅以上述的方法篩選，「反對」將不會被列入可能的翻譯。因此，我們人工整理了雙語辭典中的資料，並加入做為例外條件。

經過篩選後，刪除了許多不合理的對應，因此我們將將篩選後的對應的機率值做正規化，將機率值等比例放大至總和為 1。最後，我們以正規化後的機率，篩選詞性，再從選出的詞性中，以詞彙次數篩選詞彙。以「discussion on」中的「on」為例，表 6 為「on」的對應經過篩選並將機率值正規化後的結果，我們從詞性機率篩選出「NULL」及詞性「P」，並從詞彙次數篩選出「在」、「對」、「就」、「於」等詞彙。

表 6.「discussion on」中的「on」的對應機率
[Table 6. The translation probability of "on" and Chinese correspondences in "discussion on"]

對應中文詞性	機率	詞彙及次數
NULL	**0.724**	
P	**0.226**	**在|35 對|21 就|10 於|9** 關於|3 從|2 對於|2 以|1 自|1 針對|1 至於|1
Ng	0.039	上|7 時|7 後|1
DE	0.01	的|4

3.4 產生實詞與介系詞搭配後的翻譯 (Translating Content Words and Preposition Collocations)

本階段的目標是產生實詞和介系詞搭配後的翻譯。本階段的輸入是在上上階段（第三章第（二）節）所輸出的實詞翻譯和上階段（第三章第（三）節）所輸出的介系詞翻譯，我們也會使用在第一階段（第三章第（一）節）改善後的雙語對齊資料，以及中文語料庫。

首先我們要從中文語料庫中擷取中文高頻搭配。我們用中文斷詞與詞性標註系統，處理中文單語語料庫，將句子斷詞並標註詞性，然後使用 Smadja 於 1993 年(Smadja, 1993)提出的搭配詞提取方法「Retrieving Collocation from Text: Xtract」，在純中文語料庫中擷取每個詞彙的高頻搭配，建立高頻搭配表。由於英文介系詞所對應到的中文翻譯多為特定詞性，因此在計算高頻搭配詞時，我們將各種詞性分開計算，以「發言」為例，其搭配詞提取的部分結果如表 7：

表 7.「發言」的搭配詞
[Table 7. The collocation of "發言" fayan "speech"]

基本詞	搭配詞性	搭配詞	位置
發言	P	在	-3
發言	P	在	-2
發言	D	就	-2
發言	D	就	-3

接著，我們針對在前面階段中所擷取的實詞翻譯和介系詞翻譯，嘗試各種組合，檢查是否為中文高頻搭配，若是，則做為搭配後的翻譯。以上述提到的「speech on」為例，「speech」的翻譯有「發言」，而「on」的翻譯有「在」和「就」，則由上表我們可以找到「在 _ _ 發言 」、「在 _ 發言」、「就 _ _ 發言 」、「就 _ 發言」這幾組翻譯（此處以底線代表空格）。若介系詞翻譯包含「NULL」，代表此介系詞在翻譯時經常被省略，故我們會將實詞皆做為搭配後的翻譯。為了得到更精確的翻譯，產生翻譯後，我們會在回到平行語料庫中，檢查這些翻譯在平行語料庫中出現的次數，並篩除次數太少者。下表以「speech on」為例，展示得出搭配翻譯的過程。

表 8.「speech on」搭配翻譯的產生過程
[Table 8. The process of generating translation of "speech on"]

英文搭配	實詞翻譯	介系詞翻譯	搭配翻譯
speech on	演說、講、發言、演講、言論、辭、演辭、演、演詞	P:在 於 就 關於 NULL: NULL	在 _ _ 發言、在 _ 發言、就 _ _ 發言、就 _ 發言、演說、講、發言、演講、言論、辭、演辭、演、演詞

然而，有時某個介系詞翻譯雖然為某個實詞翻譯的高頻搭配詞，但當兩者搭配時，該介系詞翻譯卻經常不是對應到原輸入中與英文名詞搭配的英文介系詞，以「problem of」為例，在前述的篩選方法中，我們擷取了「problem」的翻譯「問題」及「of」的翻譯「的」，而在中文搭配的統計中，「問題 的」為一個相當高頻的搭配，因此在前一個步驟中，「問題 的」會被列為「problem of」的翻譯，但此時「的」通常並不會對應到「problem of」中的「of」，「問題 的」並不適合做為「problem of」的翻譯。為了改善這樣的狀況，我們針對每一組由中文搭配組合而成的搭配翻譯，計算當此組合出現時，介系詞翻譯對應正確（即對應至原輸入中的英文介系詞）的比例，並以此比例值篩選出更為精確的搭配翻譯。

產生搭配翻譯後，我們針對翻譯選取適合的例句。首先我們在平行語料庫中爲每一組英文搭配詞的每組中文搭配翻譯，抽取含有此搭配的句子，因為中文搭配在句子中時常跨越超過 1~3 個詞彙，因此我們在選取例句時我們放寬距離的限制，允許中間的空格填入較多詞彙。為了減少選取錯誤句子的機會，我們將句子原來的自動對齊納入考量，在此句子原來的自動對齊中，此中文搭配翻譯確實對應至此英文搭配，我們才會選取這個句子作為這個翻譯的例句。以「speech on」翻譯至「就 ... 發言」為例，表 9 呈現抽取例句中詞彙自動對齊：

表 9.「speech on」翻譯至「就 ... 發言」的例句
[Table 9. An example pair of sentences including translating "speech on" to "就 … 發言"]

中文例句	我(0) **在(1)** 二讀(2) **發言(3)** 時(4) ，(5) 已經(6) 頗為(7) 詳盡(8) 地(9) 講述(10) 這(11) 項(12) 動議(13) 。(14)
英文例句	i(0) have(1) dealt(2) with(3) this(4) at(5) some(6) length(7) in(8) my(9) **speech(10)** **on(11)** the(12) second(13) reading(14) .(15)
中英詞對齊	0-0 5-1 6-1 10-2 10-3 11-4 4-5 7-6 7-7 8-7 9-7 10-7 9-8 0-9 **3-10** 4-10 **1-11** 11-12 2-13 12-13 2-14 13-14 14-15

部分句子可能同時含有兩組以上可能的翻譯，以表 10 的句子為例，在前述的方法中我們擷取了「speech on」的翻譯「就 … 發言」及「在 … 發言」，此句同時含有「就 … 發言」及「在 … 發言」，但僅適合做為「就 … 發言」的例句，不適合做為「在 … 發言」的例句。因此，在以自動對齊進行初步的例句篩選後，我們再針對同時含有兩組以上翻譯的例句，考量搭配翻譯在句子中所跨越的距離等因素，進行篩選，得到更精確的例句。最後，我們以篩選後例句數量，對搭配翻譯做最後一次篩選，得出本研究的翻譯結果。

表 10. 同時含有「speech on」兩組翻譯的例句
[Table 10. An example sentence containing two Chinese translations of "speech on"]

中文例句	… **在** 我 **就** 動議 議案 **發言** 後 …
英文例句	… after making the **speech on** my motion.

4. 實驗與評估 (Experiment and Evaluation)

4.1 資料集與工具 (Datasets and Tools)

4.1.1 香港立法局會議資料 (Minutes of Legislative Council of the Hong Kong Special Administrative)

我們採用香港立法局會議資料作為統計詞彙對應及擷取例句時使用的語料，此資料為中英雙語平行語料庫，共約 222 萬句，本研究實際使用約 164 萬句。（資料來源：catalog.ldc.upenn.edu）

4.1.2 聯合報 (United Daily News)

我們採用聯合報資料作為研究中文高頻搭時使用的語料，此資料為中文單語語料庫，涵蓋約 230 萬篇中文新聞，共約 7,118 萬句。（資料來源：udn.com）

4.1.3 CKIP中文斷詞系統 (Chinese Knowledge and Information Processing System)

我們使用 CKIP 中文斷詞系統處理中文句子，產生詞性標註。此系統是由中研院詞庫小組開發，提供中文的斷詞與詞性標註。（資料來源：ckipsvr.iis.sinica.edu.tw）

4.2 實驗設定 (Experimental Settings)

4.2.1 名詞介系詞搭配之片語翻譯 (Translations of Praises Including Nouns and Prepositions)

本實驗的輸入為 30 組英文名詞和介系詞的搭配。我們從香港立法局會議資料中，隨機抽取 30 組搭配，作為本實驗的輸入，抽取結果如下。

period in	scheme to	environment for	place at	emphasis on
development by	market for	scope of	time by	agreement between
extension of	help to	potential for	agreement with	pressure on
help from	return to	power in	industry in	communication with
view on	system at	success of	gap between	pressure from
link between	scheme by	satisfaction with	increase from	information about

我們使用第三章所介紹之方法，改善香港立法局會議資料的中文斷詞及中英對應，然後從資料中統計並擷取實詞翻譯及介系詞翻譯，最後將兩者結合產生翻譯及例句，即為本實驗的輸出。

本實驗評估分為翻譯正確率、翻譯召回率及例句正確率三個部份。翻譯正確率即本

實驗產生之翻譯的正確比例，由人工逐一評判得出。翻譯召回率則是指本實驗產生的正確翻譯在香港立法局會議資料全文所有正確翻譯中所佔的比例，由程式計算得出例如若原輸入為「disparity between」，產生之翻譯為「之間 … 差距」、「差距」、「懸殊」，則我們計算所有英文句包含「disparity between」的句子中，其中文句出現「之間 … 差距」、「差距」、「懸殊」的比例，作為翻譯召回率。例句正確率則是針對經人工評判為正確的翻譯，分別抽取 5 句例句，由人工逐一檢視，得出例句正確比例。

4.2.2 單詞翻譯 (Translations of Single Content Word)

因實詞翻譯為本研究中相當重要的一部分，因此我們另外設計了一組實驗評估實詞翻譯的效果。本實驗的輸入為 30 個英文名詞。我們從香港立法局會議資料中隨機抽取了 30 個名詞，作為本實驗的輸入，抽取結果如下。

super	seriousness	nurse	placing	inland
abuse	inject	final	designation	urgent
death	send	city	charge	pain
outlook	divorce	degree	adding	providing
signal	enhancement	cut	wisdom	auditor
position	hotel	identification	confirmation	administrator

我們使用第三章的二之一小節所介紹之方法，從香港立法局會議資料中統計並擷取翻譯，所得之翻譯即為本實驗的輸出。最後使用與評估名介搭配翻譯相同的方法，計算翻譯正確率、翻譯召回率及例句正確率，做為本實驗的評估結果。

4.3 實驗結果與討論 (Evaluation and Discussion)

名詞介系詞搭配翻譯的評估結果如下表：

表 11. 名詞介系詞搭配翻譯評估結果
[Table 11. Evaluation of translations of nouns and prepositions]

英文搭配	翻譯精確率	翻譯召回率	例句精確率	英文搭配	翻譯精確率	翻譯召回率	例句精確率
help from	1.0	0.42	1.0	gap between	1.0	0.08	1.0
power in	1.0	0.35	0.67	success of	0.86	0.65	1.0
help to	0.86	0.61	1.0	scope of	1.0	0.73	1.0
agreement with	0.75	0.51	1.0	period in	0.8	0.84	0.75
environment for	0.5	0.8	0.6	pressure from	1.0	0.07	1.0
agreement between	1.0	0.24	1.0	emphasis on	1.0	0.75	1.0

return to	1.0	0.38	0.6	communication with	1.0	0.71	0.8
extension of	0.78	0.63	1.0	view on	1.0	0.3	1.0
information about	1.0	0.77	1.0	scheme to	1.0	0.82	1.0
potential for	1.0	0.52	1.0	link between	1.0	0.03	1.0
market for	1.0	0.75	1.0	increase from	None	0	None
pressure on	1.0	0.44	0.4	place at	None	0	None

單詞翻譯的評估結果如下表：

表 12. 單詞翻譯評估結果
[Table 12. Evaluation of translations of single content words]

英文搭配	翻譯精確率	翻譯召回率	例句精確率	英文搭配	翻譯精確率	翻譯召回率	例句精確率
seriousness	1.0	0.42	1.0	degree	1.0	0.45	1.0
nurse	1.0	0.89	1.0	signal	1.0	0.49	1.0
inland	1.0	0.03	1.0	enhancement	1.0	0.44	1.0
abuse	1.0	0.69	1.0	cut	1.0	0.48	1.0
final	1.0	0.58	1.0	wisdom	1.0	0.54	1.0
designation	1.0	0.33	1.0	auditor	1.0	0.85	1.0
death	1.0	0.41	1.0	position	1.0	0.59	1.0
city	1.0	0.32	1.0	hotel	1.0	0.69	1.0
charge	1.0	0.39	1.0	identification	1.0	0.37	1.0
pain	1.0	0.52	1.0	confirmation	1.0	0.45	1.0
outlook	1.0	0.35	1.0	administrator	1.0	0.21	1.0

統計以上評估結果，得到兩組實驗的平均翻譯精確率、翻譯召回率及例句精確率，如下表所示：

表 13. 綜合評估結果
[Table 13. Evaluation]

	翻譯精確率	翻譯召回率	例句精確率
名介搭配翻譯	93%	47%	91%
名詞單詞翻譯	100%	49%	100%

我們嘗試直接擷取自動對齊結果做為翻譯，並以相同方式評估翻譯精確率及翻譯召回率，並與本論文方法比較，結果如下表：

表 14. 評估結果比較
[Table 14. Results and discussion]

	自動對齊 翻譯精確率	本論文方法 翻譯精確率	自動對齊 翻譯召回率	本論文方法 翻譯召回率
名介搭配翻譯	60%	93%	52%	47%
名詞單詞翻譯	73%	100%	54%	49%

本論文方法在翻譯精確率方面有大幅提升：於名介搭配翻譯較自動對齊方法精確率提升 33%，於名詞單詞翻譯方面，本論文方法提升 27%，達到 100%正確率。然而，本論文於翻譯召回率方面表現較差，分別為 47%與 49%。

我們深入探討未能擷取翻譯導致召回率較差的原因，原因大致可分為四類，分別為「無對應中文翻譯」、「翻譯為非獨立詞彙」、「斷詞錯誤」、「本方法未成功擷取翻譯」，說明如下表：

表 15. 錯誤類型說明
[Table 15. Descriptions of error types]

錯誤類型	說明
無對應中文翻譯	因雙語句子對齊錯誤或句法改寫，中文句不存在該英文詞彙的翻譯，或該英文詞彙翻譯至中文時被省略。
對應中文翻譯為非獨立詞彙	中文句中確實存在該英文詞彙的翻譯，但並不是一個獨立的詞彙，而是包含於某個中文詞彙中，導致找不到翻譯。例如「high degree of autonomy」譯為「高度 自治」，此時名詞「degree」的翻譯為「高度」中的「度」，而非一個獨立的詞彙。
斷詞錯誤	中文句中確實存在該英文詞彙的翻譯，但因斷詞錯誤導致無法找到翻譯。
本方法未成功擷取翻譯	中文句中確實存在該英文名詞的翻譯，亦沒有發生斷詞錯誤等問題，但本研究的方法無法成功擷取翻譯。

我們共抽取 50 個句子，人工觀察後，得到各類型錯誤的數量及比例，如下表：

表 16. 錯誤類型比例
[Table 16. Ratios of error types]

錯誤類型	數量	百分比
無對應中文翻譯	25	50%
對應中文翻譯為非獨立詞彙	19	38%
斷詞錯誤	2	4%
本方法未成功擷取翻譯	4	8%

由此可知未找到翻譯的句子中，約有 50%在原始資料即無相對應的中文翻譯，因此我們不將錯誤類型一（無對應中文翻譯）的資料納入統計，以此比例估計本方法真實召回率，以及直接擷取自動對齊結果做為翻譯的召回率，結果如下表：

表 17. 評估結果比較（修正召回率後）
[Table 17. Results and discussions with revised recall rates]

	自動對齊 翻譯精確率	本論文方法 翻譯精確率	自動對齊 翻譯召回率 （修正後）	本論文方法 翻譯召回率 （修正後）
名介搭配翻譯	60%	93%	68%	64%
名詞單詞翻譯	73%	100%	70%	66%

實驗結果顯示，相較於直接擷取自動對齊結果做為翻譯，本方法可在翻譯召回率僅小幅下降（兩組實驗分別下降 4%）的情況下，精確率大幅提升（兩組實驗分別提升 33%及 27%），表示本方法能在僅犧牲少數正確翻譯的情況下，篩選掉大量的錯誤翻譯。本方法所擷取的搭配翻譯及單詞翻譯對使用者來說是相當可信的，但召回率仍有改善的空間，且有少數搭配未能找到翻譯（例如「increase from」），顯示仍有不少正確的翻譯本方法尚無法成功擷取。

我們觀察實驗結果較不佳的搭配，發現兩個效果不佳的可能原因，其一，和實詞及介系詞的中文翻譯在句子中的位置距離有關，實詞與介系詞距離太遠、位置不定，導致搭配組合抽取困難，因而未能篩選出可做為翻譯的組合，例如「increase from」這組搭配，在擷取實詞翻譯時，成功擷取「增加」、「提高」、「增長」等翻譯，在擷取介系詞翻譯時，也成功擷取「由」、「從」等翻譯，但在最後根據中文高頻搭配組合的階段，卻未能篩選出可做為翻譯的組合，例如聯合報資料中的句子「**從** 去年 底 的 十一 人 **增加** 到 十四 人」雖包含「從...增加」，但「從」和「增加」的位置距離較遠，導致在計算中文高頻搭配時，未能擷取這樣的搭配。另一個召回率較低的原因為：單詞翻譯中英文的中文翻譯不是一個詞而是語素（例如「death penalty」翻譯至「死刑」，單詞「death」翻譯至「死」，單詞「penalty」翻譯至「刑」），然而本方法無法處理詞彙非一對一翻譯的狀況，因此無法擷取這些翻譯，這可能是導致召回率較低的重要原因。

5. 結論與未來展望(Conclusion and Future Work)

從我們的實驗結果能觀察到，我們的方法所擷取出的翻譯，已能達到不錯的精確率，但在召回率的部分仍有改善的空間，顯示仍有部分翻譯無法由我們的方法找到。

目前有許多方向可以繼續研究。計算高頻搭配時可嘗試考慮更大的距離，以應付介系詞與實詞的對應距離較遠的狀況（例如「increase from」對應至「從...增加」），提升召回率。目前的方法無法處理一詞翻譯至多詞的情況（例如「partnership」對應至「夥伴關係」、「死刑」對應至「death penalty」）若能將這些情況加入考慮，就能更精準擷取翻譯。目前僅限制在擷取名詞及名詞介系詞搭配的翻譯，未來可以嘗試用類似的方法，

來擷取其他詞性（如動詞及動詞介系詞搭配）的翻譯，或擴充至更長片語的翻譯。

綜上所述，我們的研究提出了一套方法，從已做雙語自動對齊的雙語語料庫中，擷取名詞單詞及名詞和介系詞搭配時的翻譯，使用的方法包含統計雙語語料庫中的正反向對應，以及統計單語語料庫中搭配詞，並結合以上兩者。經過評估後，證實我們的方法找到的翻譯已能達到較佳的精確率，並大多能找到正確的例句。

參考文獻 (References)

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D. … Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, *16*(2), 79-85.

Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of 29th Annual Meeting of the ACL*, 169-176. doi: 10.3115/981344.981366

Catizone, R., Russell, G., & Warwick, S. (1989). Deriving Translation Data from Bilingual Texts. In *Proceedings of the First International Lexical Acquisition Workshop*, 15-21.

Chen, S. F. (1993). Aligning Sentences in Bilingual Corpora Using Lexical Information. In *Proceedings of 31st Annual Meeting of the ACL*, 9-16. doi: 10.3115/981574.981576

Debili, F., & Sammouda, E. (1992). Aligning sentences in bilingual texts french-english and french-arabic. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING 1992)*, *2*, 5178524. doi: 10.3115/992133.992151

Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceedings of ACL-1995*, 236-243

Fung, P., & Church, K. (1994). K-vec: A new approach for aligning parallel texts. In arXiv preprint arXiv: cmp-lg/9407021

Gale, W. A., & Church, K. W. (1991a). Identifying Word Correspondences in Parallel Texts. In *Proceedings of the workshop on Speech and Natural Language*, 152-157. doi: 10.3115/112405.112428

Gale, W. A., & Church, K. W. (1991b). A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of 29th Annual Meeting of the ACL*, 177-184. doi: 10.3115/981344.981367.

Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, *19*(1), 75-102. doi: 10.5555/972450.972455

Kay, M., & Roscheisen, M. (1993). Text-translation alignment. *Computational linguistics*, *19*(1), 121-142. doi: 10.5555/972450.972457

Melamed, I. D. (1995). Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, 184-198.

Melamed, I. D. (1999). Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics*, *25*(1), 107-130.

Moore, R. C. (2001). Towards a Simple and Accurate Statistical Approach to Learning Translational Relationships Among Words. In *Proceedings of ACL-2001 Workshop on Data-Driven Methods in Machine Translation*, 79-86.

Simard, M., Foster, G. F., & Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, 67-81.

Smadja, F. (1993). Retrieving Collocation from Text: Xtract. *Computational Linguistics*, *19*(1), 143-177.

Wu, D. & Xia, X. (1994). Learning an English-Chinese Lexicon from a Paarallel Corpus. In *Proceedings of AMTA-94*, 206-213.

柯明憲(2006)。**雙語語料庫之多字詞語對應**(碩士論文)。[Ko, M. H. (2006). *Alignment of Multi-word Expressions in Parallel Corpora* (Master's thesis).