英文技術文獻中一般動詞與其受詞之中文翻譯的語境效用

Collocational Influences on the Chinese Translations of Non-Technical English Verbs and Their Objects in Technical Documents

莊怡軒 王瑞平 蔡家琦 劉昭麟
Yi-Hsuan Chuang Jui-Ping Wang Chia-Chi Tsai Chao-Lin Liu
國立政治大學資訊科學系
Department of Computer Science, National Chengchi University
{g9804,g9916,g9906,chaolin}@cs.nccu.edu.tw

摘要

本文探索英文動詞與英文名詞的英漢翻譯中，語境資訊對於翻譯品質的貢獻度。文獻常見的研究多集中於使用英文動詞本身的各項語言特徵，或者加上與該動詞搭配的英文名詞的相關資訊來推測英文動詞的翻譯。本文探索一極端假設下的翻譯成效：如果我們也能知道英文名詞的中譯時，是否有助於英文動詞的翻譯品質？我們利用 2011 年 NTCIR 的漢英翻譯工作坊的數萬句專利語料作為實驗資料來源，同時利用七年的科學人的語料進行實驗，目前實驗顯示在所假設之情形下，增加名詞中譯的資訊，固然有助於提高翻譯品質，但是效果暫不明顯，有待更精確的實驗設計來確認英文中譯詞對於英文動詞的翻譯貢獻度。

Abstract

We investigate the potential contribution of a very specific feature to the quality of Chinese translations of English verbs. Researchers have studied the effects of the linguistic information about the verbs being translated, and many have reported how considering the objects of the verbs will facilitate the quality of translations. In this paper, we take an extreme assumption and examine the results: How will the availability of the Chinese translations of the objects help the translations of the verbs. We explored the issue with thousands of samples that we extracted from 2011 NTCIR PatentMT workshop and Scientific American. The results indicated that the extra information improved the quality of the translations, but not quite significantly. We plan to refine and extend our experiments to achieve more decisive conclusions.

關鍵詞：機器翻譯、特徵評比、自然語言處理

1. 緒論

當今的社會可視為一個地球村，即使住在不同的國家、也使用不同的語言，無論是商業貿易或是文化交流，人們互相溝通的情形相當普遍；英文更因為其容易理解及表述的語言特質成為世界上不同語言使用者通用的溝通語言。因應世界文化潮流，除了自身國家的母語，英文成為最多人學習的語言。然而許多研究指出，將英文作為第一外語的學習

者 (EFL learners: English as a Foreign Language learners) 容易受到自己國家母語的文法影響，在英文動詞及名詞的搭配組合上會產生錯誤的用法。例如，「take pills」一詞若依照中文使用者的直覺，可能會翻譯解釋為「拿藥」而非正確對應至「吃藥」。因此，我們對於英文中常用的動詞與名詞組合，與中文的對應關係感到有趣，並想透過大量正確對應的英漢平行語料庫，找尋英漢動名詞組合 (V-N-collocation) 適切的對應關係。若提到大量的語料，我們首先聯想到了專利文書。

專利文書是一種宣示並提供專利保護的重要文件。世界社會持續地進步，許多發明與技術不斷創新並被撰寫成為專利文書。當發明一項專利時，專利發明者為了讓世界各國使用不同語言者可以共同瞭解這項專利，也同時向外擴張專利的保護領域，發明者可以提出多種語言版本的專利文書以保障自己的發明跟技術。專利文書的重要性更可以從 Google Patents beta[7]提供的英文專利文書檢索服務看出，Google[6]號稱他們的專利資料庫蒐集了七百萬篇以上的專利文書，以其豐富的收藏量宣示他們強大的檢索服務。既然單語言的專利文書數量如此龐大，那麼同時具有多種語言版本的專利文書也就不在少數。如果我們將專利文句作正確解析，排除技術名詞在外，剩餘的文句結構及內容不失為一個值得運用的語文使用參考資料；特別是許多專利文書具有英漢對應的語言版本，可以當作是雙語語料使用。因此，我們可以看待跨語言的專利文書為資料量豐富的平行語料庫。本研究利用專利文書豐富的英漢對應資料，並排除技術名詞的影響，試圖挖掘一般常用英漢動名詞組合對應的用法。

除了分析英漢專利平行語料庫[9]，本研究另外以相同方式分析科學人雜誌英漢對照電子書[16]，以比較不同語料間是否有不同的特性。本研究將中英文互為翻譯的文件視為一體，將英文及中文的動名詞組合作為我們的觀察對象，建構由真實世界語料反應的語言翻譯模型。本研究對於翻譯英文動詞及名詞皆有建立翻譯模型及測試其翻譯效能，不過因受限於篇幅關係，本篇論文僅會介紹翻譯英文動詞的部分；而翻譯英文名詞的成效與翻譯英文動詞相差不多。

關於專利文書的研究，田侃文[15]使用中英文互為翻譯關係的專利文書當作主要語料，並利用動態規劃演算法進行中英文句對列，設法將中文全文文章與英文全文文章的翻譯對應，拉抬至中文句子對列到英文句子的文句對列層級；本研究亦運用該文句對列系統找尋句對關係。Lu[8]提出如何建置漢英專利文句平行語料庫。該研究從網路上蒐集優良的中英專利文書平行語料，再根據專利文書的目次結構，將專利文書拆解成多個小單位。其集結了三種作法：使用雙語辭典比對詞彙、刪除過長的句子及使用 IBM M-1 為語言模型建立文句對列，其研究結果顯示準確率最高可達 97%。

關於語言輔助教學方面，Chang[1]則針對學習英文的中文使用者製作一套英文寫作校正系統。使用者將寫好的英文文章輸入至系統，系統便會偵測動名詞片語有無誤用之處。該研究蒐集學習英文的中文使用者之英文寫作文章當作學習者語料庫，從中發現常見的錯誤用法；另外同時蒐集正確的英文語料當作正確答案的參考語料庫。該系統將錯誤的動詞翻譯成中文詞彙，將這些中文詞彙重新翻譯回英文詞彙，再把這些英文動詞替換片語中原本的動詞成為新的片語，並重新查詢共現性分數，得分高者則為系統建議的校正答案。

關於動名詞組合方面的研究，Venkatapathy[12]首先介紹了 multi word expressions (MWEs) ，即為那些從字面上看不出實際表達意義的詞彙。有很大一部分的 MWEs 是具有文法結構性但沒有語義合成的關係，而其中一個子集就是動名詞共現性 (V-N collocations) ，也是該研究主要分析的目標。MWEs 很難區分是為組合性 (compositional) 或為非組合性 (non-compositional) ，早一些時間的相關研究不外乎是考慮頻率 (frequency) 、互信息或是使用 LSA 模型等相關數據作分類問題；該研究則將這些數據特色都加以考慮並列入使用。該研究聘請兩位人員進行人工標記詞彙是為組合性或是非組合性的程度，並將上述的數據當作特徵，作成向量再以 SVM 排序。最後發現合併特徵比起只單一考慮任一特徵都還要貼近人工標記的答案。

2. 語料來源介紹

2.1 專利文句

本研究使用 Patent Translation Task at NTCIR-9[9]的一百萬筆英漢對照的專利文句作為我們第一份研究語料，中文的部分為簡體中文。其使用編號標示英漢句對對應關係。由於專利文句的字數偏長、文句結構也較為複雜，如果直接使用長句進行英漢動名詞組合對列，不僅對列的時間加長，產生的對列效果也會較差。本研究認為，動名詞組合並不會跨過標點符號，因此我們把每一個長句視為一篇短文章，根據長句中暫停或結束的標點符號（逗號、分號、冒號、驚嘆號、問號及句號）作為短句的終點；一個長句可視為一篇由多句短句組合而成的短篇文章。本研究使用專利文句對列系統[15]得到英漢短句對應關係。我們設定門檻值為 0.3 取得較高對列品質的短句，作為我們的使用資料。原本一百萬組長句對中，超過本研究設定門檻值的句對有 338846 組；這三十三萬的長句對又被拆成 1148632 組短句對為本研究所使用。

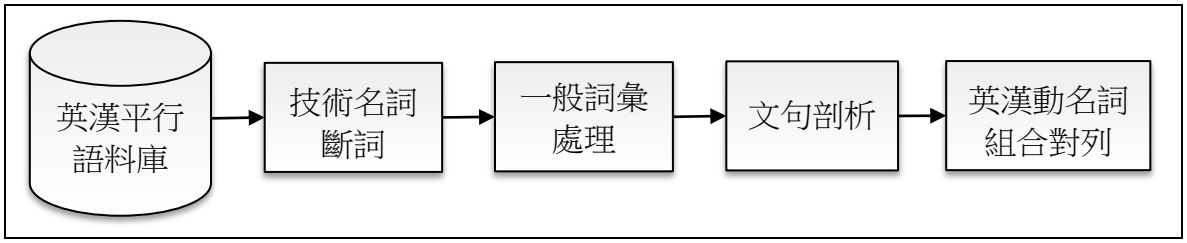
2.2 科學人雜誌

田侃文[15]將科學人雜誌英漢對照電子書[16]的 1745 篇文章使用該研究的文句對列系統產出 63256 個英漢對列的高品質句對。本研究沿用這 63256 個句對作為第二份分析語料。

3. 技術名詞表建置

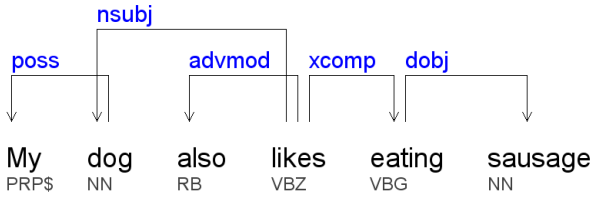
為了能順利排除技術名詞的資訊，我們需要有技術名詞表比對詞彙以便標記捨去。本研究從國家教育研究院學術名詞資訊網[17]取得公開的 138 個不同領域技術名詞 Excel 格式檔案，檔案大小共有 177MB 並整合為技術名詞表。在技術名詞表中，每一個英文技術名詞都有與其對應的中文技術名詞，且對應關係並不唯一，本研究將技術名詞表的翻譯詞對規列成一對一的形式。

我們發現在技術名詞表當中，英文及中文部分都有些許的技術名詞更常被當作一般用語詞彙；我們使用 E-HowNet[2][5]及 WordNet[13]來幫助刪除一般詞彙，留下技術名詞於技術名詞表。本研究認為，這兩部字典所收錄的詞彙可以代表生活中一般常用的詞彙，使用這些詞彙過濾技術名詞表是可行的方式。E-HowNet 內含 88075 個中文詞彙，共識別出技術名詞表中有 71333 個詞對更適合被當成一般詞彙而非技術名詞。我們也對稱檢驗技術名詞表中的一般英文詞彙，WordNet 內含 154754 個英文詞彙及英文短片語，



圖一、語料前處理流程圖

表一、英文及中文關係樹範例

英文句	My dog also likes eating sausage.
英文句關係樹 樹狀圖	
英文句關係樹 結構	poss(dog-2, My-1)、nsubj(likes-4, dog-2)、advmod(likes-4, also-3)、xcomp(likes-4, eating-5)、**dobj(eating-5, sausage-6)**
中文句	我的狗喜歡吃香腸。
中文句關係樹 結構	assmod(狗-3, 我-1)、assm(我-1, 的-2)、nsubj(吃-5, 狗-3)、advmod(吃-5, 喜歡-4)、**dobj(吃-5, 香腸-6)**

我們使用 WordNet 檢查共過濾了 80220 個詞對。經過以上檢測，我們的技術名詞表約略除去 14%的詞對，現存有 690640 組技術名詞詞對。我們相信這六十九萬組技術名詞詞對具有較高品質，可以降低與一般詞彙產生斷詞衝突的機率。

4. 語料前處理

本研究語料前處理的過程如圖一所示，以下逐一小節解釋各步驟流程。

4.1 技術名詞標記

技術名詞多為複合詞彙，因此我們使用長詞優先的方式，從技術名詞表比對英漢平行文句中的詞彙，一經比對成功則將技術名詞標記，並使用 Stanford Parser[11]的 TaggedWord() 函數指定詞性為名詞。本研究將技術名詞標記是為了提升文句剖析的準確率，以及處理排除技術名詞資訊。

4.2 英文詞幹還原及詞性標記

本研究使用 Stanford Parser 及其 englishPCFG.ser.gz 字典模型剖析英文文句，亦運用其 Stemmer() 函數進行詞幹還原。我們將技術名詞之外的文句部分進行詞幹還原，且令 Stanford Parser 依據字典模型斷詞及標記詞性。技術名詞在這個步驟不會被更動。

4.3 中文斷詞

標記完中文技術名詞之後，剩下的文句仍需進行斷詞，我們使用 Stanford Chinese Segmenter[10]進行斷詞，並將斷好的詞彙以空白相隔。同樣技術名詞在這個步驟不會被更動。

4.4 關係樹剖析

本研究使用 Stanford Parser 剖析文句得到關係樹結構，Stanford Parser 的關係樹剖析共含有 27 種文法關係標記。一個句子經過剖析可以得知這個句子含有幾種文法關係，上頁表一即為翻譯對應的英文及中文句關係樹範例。27 種文法關係標記中，「DIRECT_OBJECT」可以標記動詞片語的動詞及其述語對象，並以「dobj」為形式；以表一中的英文句為例，動詞「eat」的對象是名詞「sausage」，並以「dobj(eating-5, sausage-6)」標記，中文句的「dobj(吃-5, 香腸-6)」也是如此；其中數字 5 與 6 代表詞彙在文句中出現的位置次序。本研究將英文及中文的句子剖析，透過抽取「DIRECT_OBJECT」表示式得到句子中的動名詞組合。剖析英文的字典模型為 englishPCFG.ser.gz，中文剖析的部分，本研究使用 xinhuaFactored.ser.gz 字典模型處理簡體中文的專利文句，chineseFactored.ser.gz 則處理繁體中文的科學人雜誌。

5. 近義詞典建置

我們需要將互為翻譯對照的動名詞組合對列產生翻譯結果。本研究使用基於辭典資訊的機器翻譯 (dictionary-based machine translation)，採用的英漢辭典有兩部，分別為牛津現代英漢雙解詞典[3]與 Dr.eye 譯典通線上字典[4]。但是只依靠英漢辭典的資訊並不足夠，因為英漢辭典中列出與英文詞彙對應的中文翻譯詞彙有限；如果以英漢字典內的英文詞彙之中文對應詞彙為基礎找尋意義相近的中文詞彙，也就表示這些中文詞彙與該英文詞彙的意義也會近似，因此我們使用一詞泛讀[14]及 E-HowNet[5]建立近義詞典，擴充英文詞彙對應的中文翻譯詞彙，幫助英漢動名詞組合對列。

5.1 英漢辭典合併

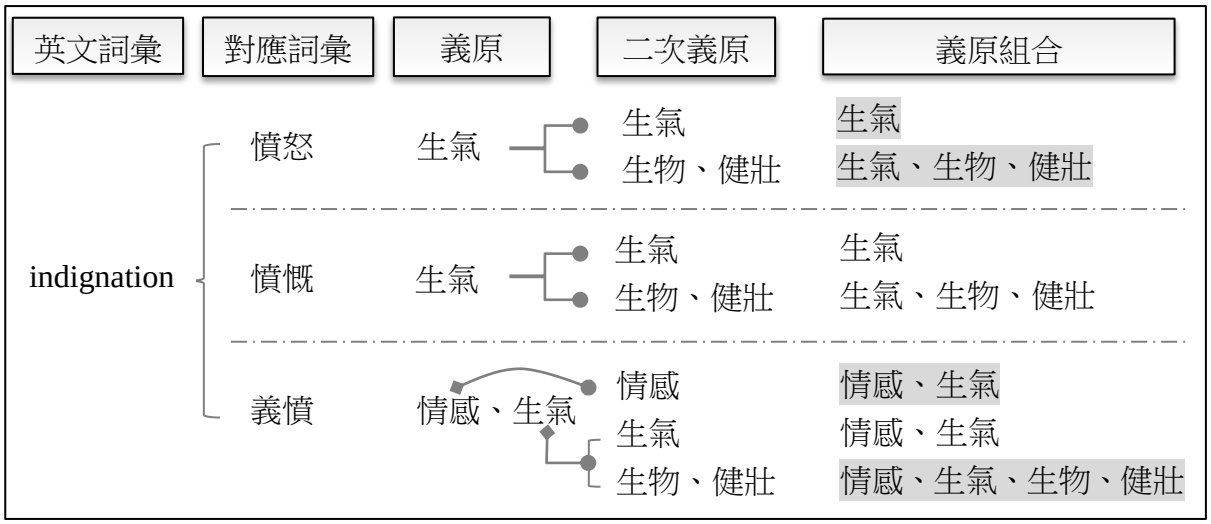
不同辭典對於同一個英文詞彙所定義的中文對應詞彙並不完全相同；因此本研究將牛津現代英漢雙解詞典和 Dr.eye 譯典通線上字典的中文對應詞彙合併，增加英文詞彙的中文對應詞彙數量。經合併之後，本研究的「英漢合併字典」內容格式如表二所示；合併之後英文詞彙「confusion」對應的中文詞彙數量明顯增加。

5.2 使用一詞泛讀尋找近義詞彙

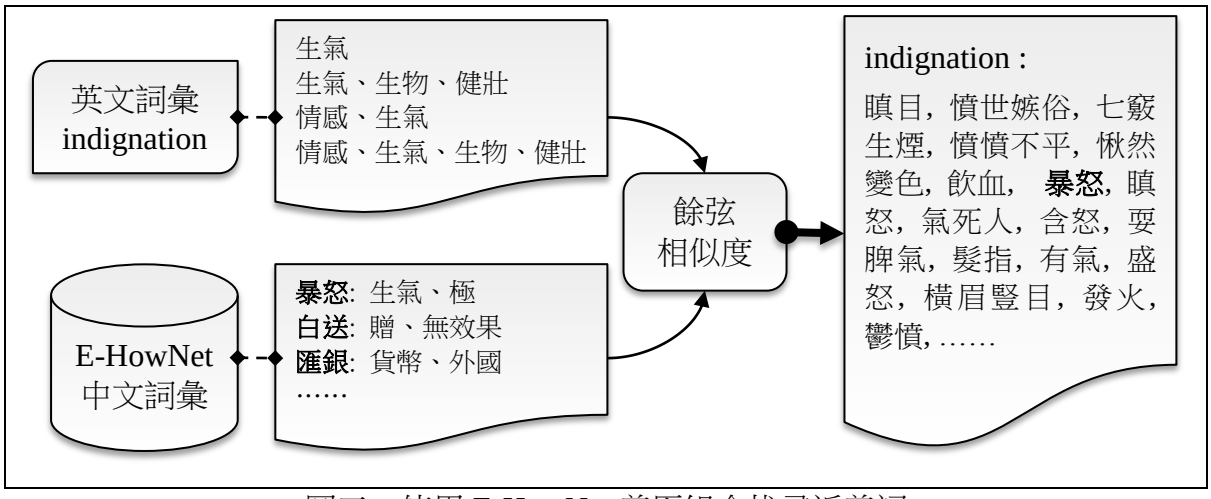
現代漢語一詞泛讀系統（簡稱為一詞泛讀）提供近義詞查詢服務。以表二的「confusion」為例，我們的做法為逐一將英漢合併字典一欄中的詞彙輸入至一詞泛讀，並聯集系統所傳回的近義詞群。我們認為回傳的近義詞群與「confusion」的中文對應詞彙意義相近，依照推理也與「confusion」的意思相近，因此這些近義詞群就是我們透過一詞泛讀找到的近義詞彙。

表二、英漢合併字典範例

英文詞彙：confusion	
辭典	辭典中的中文對應詞彙
牛津詞典	迷亂、惶惑、混亂、雜亂、混淆、混同
譯典通字典	混亂、騷動、混亂狀況、混淆、困惑、慌亂
英漢合併字典	混亂、混亂狀況、騷動、混淆、困惑、慌亂、迷亂、惶惑、雜亂、混同



圖二、E-HowNet 義原組合流程

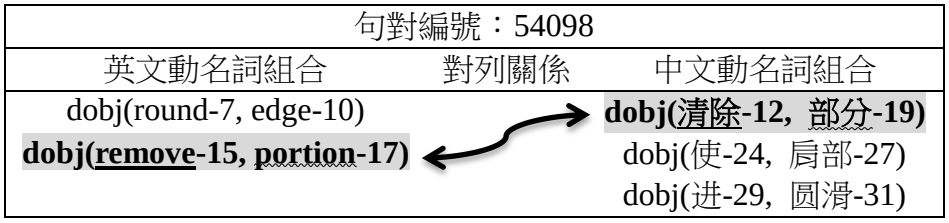


圖三、使用 E-HowNet 義原組合找尋近義詞

5.3　使用 E-HowNet 尋找近義詞彙

圖二為英文詞彙「indignation」透過中文對應詞彙至 E-HowNet 形成義原組合的過程範例。在我們的英漢合併字典中，「indignation」擁有三個中文對應詞彙，分別為「憤怒、憤慨及義憤」。而這三個中文詞彙恰巧都只有一種語意，在只有一種語意的情形之下中文詞彙的義原也只會有一群；「憤怒」及「憤慨」的義原只有「生氣」一個義原，「義憤」的義原群則由「情感」及「生氣」兩個義原組成。我們發現，E-HowNet 的義原本身同時也是一個詞彙，而且也有定義自己的義原。這種定義 E-HowNet 義原的義原，我們稱之為「二次義原」。找出中文對應詞彙的義原群及二次義原群之後，我們將義原以及各自的二次義原組合起來，形成圖二中的義原組合；排除掉重複的組合得到以灰底標示的義原組合群，即為透過中文對應詞彙找到英文詞彙可能的義原組合群。

如圖三所示，英文詞彙「indignation」有了義原組合群之後，本研究將 E-HowNet 中所有的中文詞彙依照同樣的流程組成義原組合，然後逐一取出「indignation」的每條義原組合與 E-HowNet 全部中文詞彙的義原組合作餘弦相似度 (cosine similarity) 計算並設定門檻值為 0.7。最後我們將從一詞泛讀及 E-HowNet 得到的近義詞與英漢合併字典整合，形成我們擴充英文詞彙的中文對應詞彙字典，稱之為「近義詞典」。



圖四、英漢動名詞組合示意圖

6. 英漢動名詞組合對列

表三、翻譯英文動詞模型公式

$\underset{CV_i}{\operatorname{argmax}} \Pr(CV_i \mid EV, EN, CN)$	(1)
$\underset{CV_i, CN_j}{\operatorname{argmax}} \Pr(CV_i, CN_j \mid EV, EN)$	(2)
$\underset{CV_i}{\operatorname{argmax}} \Pr(CV_i \mid EV, EN)$	(3)
$\underset{CV_i}{\operatorname{argmax}} \Pr(CV_i \mid EV)$	(4)
$\underset{CV_i}{\operatorname{argmax}} \Pr(CV_i \mid EV, CN)$	(5)

經上述步驟，英文專利文句共產生 375041 個動名詞組合，中文專利文句則產生 465866 個動名詞組合。為了確保我們使用的動名詞組合品質，本研究使用英漢合併字典內收錄的英文詞彙檢驗英文動名詞組合，只有當組合中的動詞及名詞都有出現在字典中，我們才認定這個組合是正確的，這個步驟同時排除含有技術名詞的動名詞組合，因此不會有任何技術名詞的相關資訊。經過濾之後，有 254091 個英文動名詞組合通過檢測。我們對於中文的動名詞組合也進行同等檢驗，透過近義詞典含有的中文詞彙過濾，最後有 249591 個組合通過檢測，亦排除技術名詞資訊。透過句對編號及近義詞典，本研究的對列規則為：如果英文的動詞及名詞能在近義詞典中各自的中文對應詞彙集找到中文動名詞組合，才算對列成功，如圖四所示。對列成功的英漢動名詞組合會記錄成「remove, portion : 清除, 部分」，英文動名詞組合在前、中文動名詞組合在後的資料格式。

7. 翻譯模型建置

7.1 翻譯英文動詞公式說明

本研究提出了五種公式訓練模型翻譯英文動詞，如表三所示。我們以字母「E」代表英文、字母「C」代表中文，「V」代表動詞、「N」代表名詞；因此「EV」及「EN」各別代表英文動名詞組合中的動詞及名詞，「CV」及「CN」則為中文動名詞組合中的動詞及名詞。公式(1)至公式(4)為逐漸放寬條件的公式，公式(5)則是從另外一個觀點發想的公式。一般在考慮英漢翻譯問題時，分析英文內容的共現性 (collocation) 再對應到中文翻譯的作法較多，而本研究試想，除了考慮英文的部分，若加入中文對應翻譯的資訊是否能提升翻譯的效能。公式(1)即為我們這般考量下所提出的公式。

公式(1)除了考慮英文動名詞組合，也考慮了英文名詞的中文翻譯而推薦動詞的中文翻譯，我們想測試公式(1)會否蒐集的資訊最多而能翻譯的較為準確。對公式(1)最直覺的解釋為：若有一英文使用者在學習中文，他想把「take pills」翻譯成中文，但是他只確定「pills」可以翻譯為「藥」，則我們的公式(1)則可以透過這三個詞彙的資訊，觀察「take」跟「pills」一起使用且「pills」對應到「藥」時在語料中「take」容易被翻譯成什麼中文詞彙；如果從相反的角度解釋，則為一個中文使用者想練習英文，但是他不確定「吃藥」的「吃」該翻譯為「take」或是「eat」，但是他知道「藥」可以翻譯為「pills」，則公式(1)可以在語料中觀察「take pills」和「eat pills」跟「藥」組合在一起時哪一個的次數較多，且在公式(1)推薦的中文翻譯中可以找到「吃」這個詞彙，進而讓使用者知

道該使用「take pills」或是「eat pills」。公式(1)的原理為：如果同時看見英文的動詞、名詞及英文名詞的中文翻譯，則推薦與這三者一起出現機率最高的中文動詞 CV 為英文動詞 EV 的翻譯。公式(2)及公式(3)則為許多英漢翻譯使用的方法。公式(2)的原理為：如果看見特定英文動名詞組合 EV、EN，我們的翻譯模型會從該動名詞組合所對應的中文動名詞組合中，取得出現機率最高的組合，並推薦中文動名詞組合中的動詞當作我們的推薦翻譯詞彙。公式(3)的原理為：如果看見特定的英文動名詞組合，則該動名詞組合所對應到的中文動詞群中，出現機率最高的中文動詞 CV 即為我們的翻譯推薦詞彙。公式(4)的原理則最為寬鬆：如果看到一個英文動詞 EV，則我們所推薦的翻譯詞彙即為英漢動名詞組合當中與 EV 最常一起出現的中文動詞。公式(5)的原理較特別，我們假設如果看到一個英文動詞及其受詞的中文翻譯，則我們推薦與這個組合最常一起出現中文動詞做為推薦翻譯。

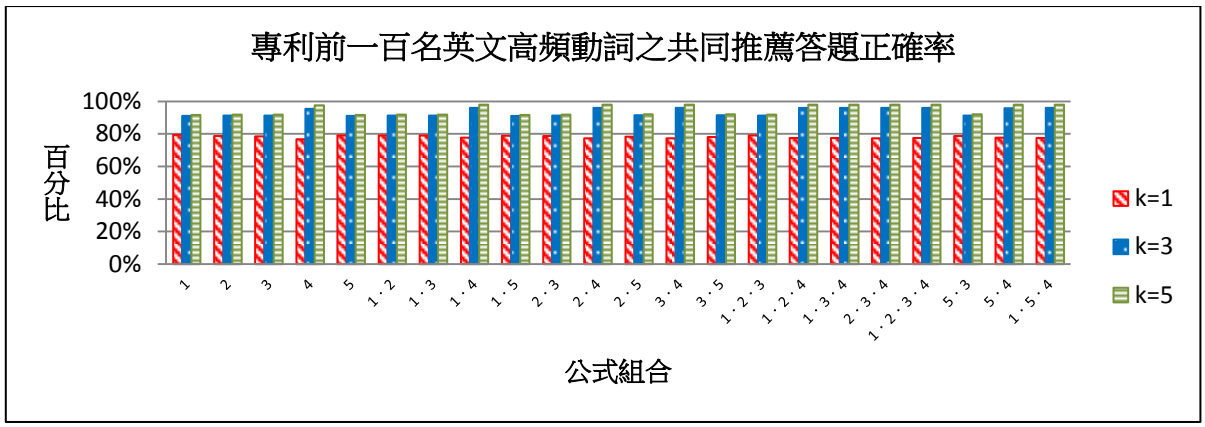
除了評比這五個公式獨立運作的效果，本研究亦將這五個公式搭配成十七種公式組合，共有二十二種翻譯模型。我們讓這些公式組合「共同推薦」英文動詞的中文翻譯：組合中的公式可以各自推薦它們認為的所有可能答案，且答案順序根據答對的機率大小排列。組合中公式的排列順序即為答題順序，且回答的答案不得重複。例如，我們設定翻譯模型最多可以回答三個答案，只要三個答案中包含正確解答即算答對；則公式組合「1・2・3」即為公式(1)、(2)及(3)的組合，各自推薦了一、二和四個答案；依照公式的排列順序，公式(1)擁有最高的回答優先權，因此公式(1)推薦的答案佔掉一個回答額度，而公式(2)提供兩個答案中最佳的答案跟公式(1)的答案重複，因此公式(2)只能回答次好的答案，公式(3)提供的四個答案中，它認為前兩好的答案恰好與公式(1)及公式(2)的答案相同，因此公式(3)只能回答第三好的答案，這時候回答的答案額度已滿，所以公式組合「1・2・3」就產生了三個可能的答案。

7.2 翻譯模型評量方式

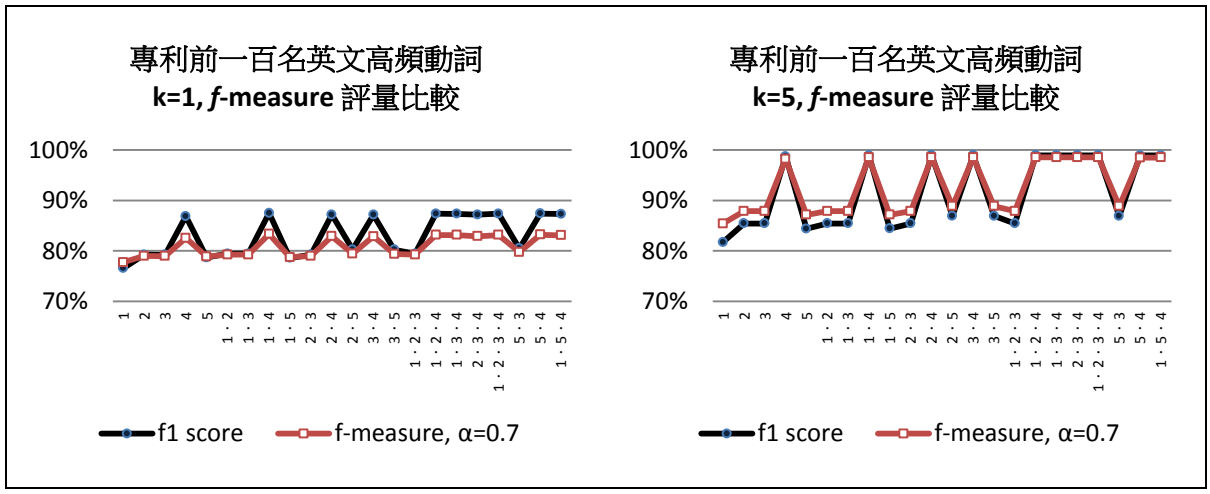
本研究將 F-measure 稍作變形，用來評量不同翻譯模型的翻譯效果。原始的 F-measure 為精確率 (precision) 和召回率 (recall) 的綜合評量。精確率可以對應為翻譯模型的答題正確率，而比起召回率，我們更著重於翻譯模型能夠回答的題目數量多寡，我們希望翻譯模型因為資訊不足而無法作答的情況越少越好，因此使用「回答率」表示翻譯模型的作答數量，並使用精確率與回答率為評量參數，本研究以「***f***-measure」代表變形後的評量方式。我們設定兩套 ***f***-measure 的係數值評量只推薦一個答案跟推薦五個答案時翻譯模型的效果，「***f***1 score」將精確率和回答率的係數平分設定為 0.5，「***f***-measure, α=0.7」則設定精確率有較高的權重 0.7，回答率的係數值為 0.3。以上說明本研究翻譯模型的原理及評量方式，接下來我們使用兩個語料庫來比較這二十二個翻譯模型的效果。

8. 使用專利語料及科學人雜誌建置翻譯模型

專利文句[9]及科學人雜誌[16]同屬於科技類文章，不過專利文句的寫作格式固定，而科學人雜誌風格較為活潑，因此本研究觀察類別相似但風格不同的兩套語料是否會造成翻譯模型的效果差異。我們利用亂數挑選的方式，將語料依據 8:2 的比例切割成訓練資料及測試資料。



圖五、專利前 100 名英文高頻動詞之共同推薦答題正確率



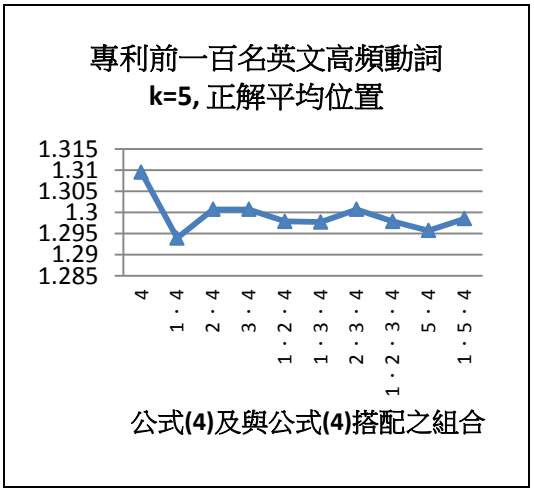
圖六、翻譯模型在專利前 100 名英文動詞推薦一個及五個答案時之 ***f***-measure 成效

8.1　使用專利文句語料建置翻譯模型

專利文句語料庫中，本研究對列成功的英漢動名詞組合共有 35811 組。

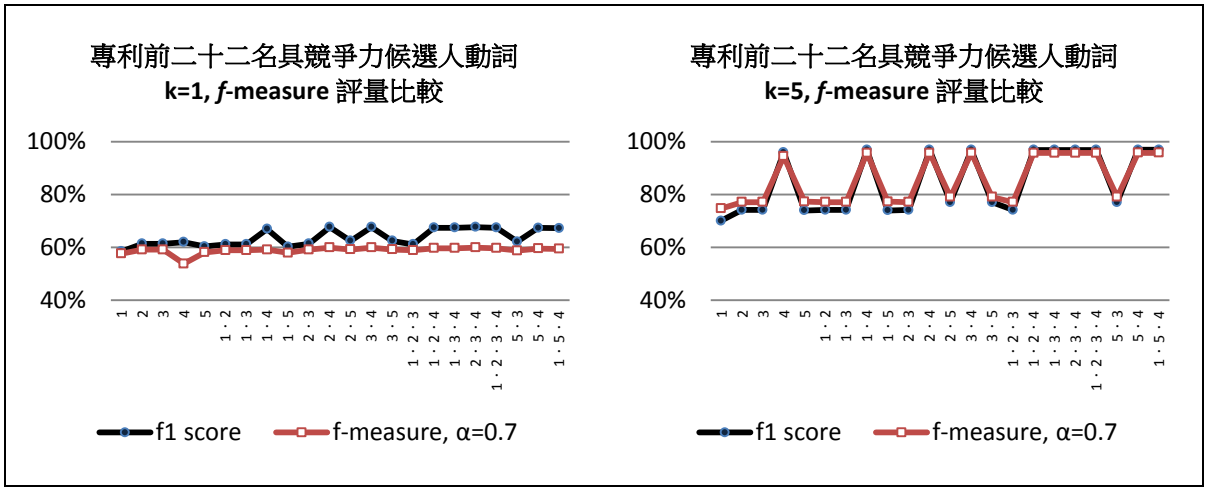
8.1.1　專利前一百名英文高頻動詞

本研究探究了在我們 35811 筆的動名詞組合資料當中，前一百名出現次數最多的英文動詞，這些動詞至少在資料中至少出現過 47 次以上，最多的出現次數則為 4530 次。這一百個英文動詞總共出現於 30376 筆資料之中，訓練資料共有 24300 筆，測試資料則有 6076 筆。



圖七、正解位置於公式(4)組合比較

圖五為翻譯模型在推薦不同數量答案時的答題正確率，圖中的 k 值為翻譯模型能夠推薦的答案數量，例如 k 設定成 5 表示翻譯模型最多可以推薦五個答案，且這五個推薦答案內如有包含正確答案即算答對。我們可以看到當推薦至三個答案跟五個答案時表現幾乎差不多，可見當我們的翻譯模型推薦三個答案時，其中幾乎都包含了正確解答。圖六為使用 ***f***-measure 評量二十二個模型翻譯專利語料中前一百名英文高頻動詞的效果比較。我們可以發現當翻譯模型只能推薦一個答案時 (k=1)，公式(4)和那些與公式(4)搭配的公式組合在 ***f***1 score 得到比較高的分數，但



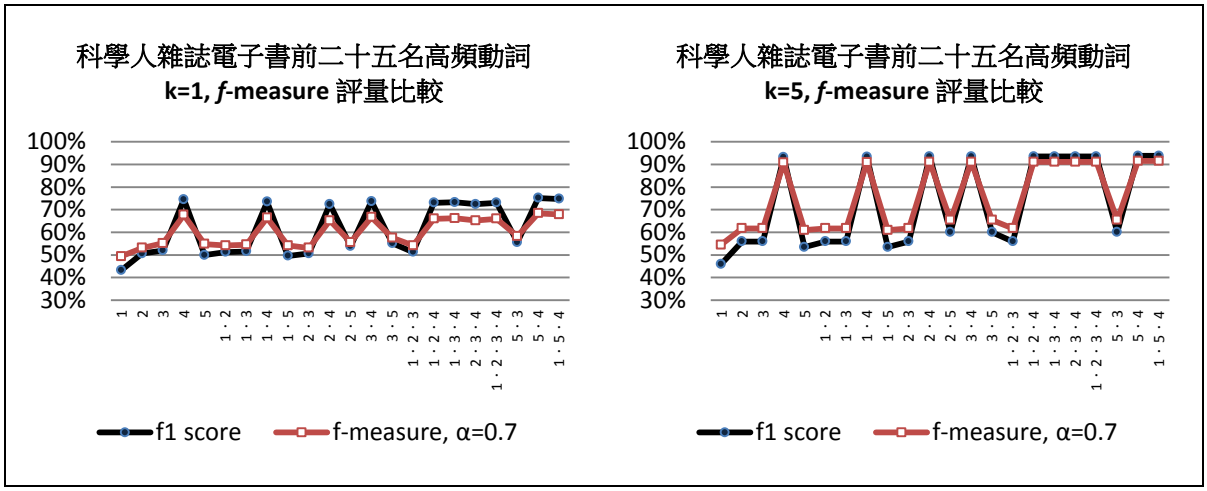
圖八、翻譯模型在專利前 22 名競爭動詞推薦一個及五個答案時之 ***f***-measure 成效

是在著重於精確率的 ***f***-measure, α =0.7 分數則往下降，其他沒有與公式(4)合作的組合及獨立運作的公式在這兩種評分機制則無差異，且分數分布略低；這是因為公式(1)、(2)、(3)及(5)都會因為測試語料中出現訓練語料所沒有的紀錄而無法作答，有回答率的問題，而公式(4)可以回答任何問題，只有答對與答錯的狀況，因為只要訓練語料有出現過的英文動詞都有其對照的中文翻譯。雖然公式(1)、(2)、(3)及(5)在圖六中只能推薦一個答案時的表現略差，但是在兩種評分機制中都維持一樣的水準；相較之下公式(4)在精確率的表現較薄弱，可以顯現出雖然公式(4)有很好的作答能力，但是僅靠著統計推薦答案效果較差，容易有答錯的情形。

翻譯模型最多能推薦五個答案 (k=5) 的情形下，每個公式組合在 ***f***1 score 及 ***f***-measure, α =0.7 的分數都有往上提升許多，特別是與公式(4)搭配的公式組合分數都相當的高；這是因為跟公式(4)搭配的公式如果有回答不出來的時候，公式(4)可以補上答案，或是當搭配的公式回答的並不是正確答案時，因為共同推薦答案不得重複的設定可以讓公式(4)更有機會補上正確解答。我們會希望當翻譯模型推薦多個答案時，正確解答能出現在推薦答案中越前面的位置越好，因此我們統計了正確答案在公式(4)及與公式(4)搭配的組合推薦答案中的排名，如上頁圖七所示。本研究發現與公式(4)搭配的公式組合中正確解答的平均位置皆比在公式(4)的平均位置還要前面；這可以證明雖然從上頁圖六公式(4)和其他與公式(4)搭配的公式組合效果近似，但是公式(1)、(2)、(3)及(5)具有把正確答案往前排名的拉提作用，特別是公式(1)效果特別明顯。

8.1.2 專利前二十二名具競爭力候選人之英文動詞

本研究由前一百名高頻動詞中選出一些英文動詞，這些動詞的特性為它們各自都不只對應到一個中文翻譯詞彙，而且出現次數最高前兩名候選人是具有競爭力的；本研究在這裡定義「競爭力」為：第一名候選人出現的次數不得多於第二名候選人出現次數的兩倍。假設英文動詞 EV 的中文翻譯候選人數依照在語料中與 EV 一起出現的次數多寡排列有 CV_1、CV_2 及 CV_3，則 CV_1 的出現次數不得多於 CV_2 的兩倍，EV 才會被我們挑選出來。根據這個門檻值的設定，我們總共找到二十二個動詞具有此特性，這二十二個動詞總共出現在 4101 筆英漢動名詞組合，訓練資料有 3280 筆，測試資料則有 821 筆。



圖九、翻譯模型在科學人前 25 名高頻動詞推薦一個及五個答案時之 ***f***-measure 成效

由上頁圖八所示，當翻譯模型只能推薦一個答案時，前二十二名具競爭力候選人的動詞與前一百名高頻動詞的趨勢並不完全相同。公式(4)和那些與公式(4)搭配的組合在 ***f***1 score 得到比較高的分數，但是在著重於精確率的 ***f***-measure, α =0.7，與公式(4)搭配的公式組合分數則往下與其他沒有與公式(4)合作的公式表現相同，特別可以注意到公式(4)在 ***f***-measure, α =0.7 的表現明顯低於其他獨立公式。這是因為訓練資料量銳減，但是資料的變化性仍然不小，因此其他考慮較多資訊的公式表現就超越了資訊考慮最少的公式(4)，這也證明公式(1)、(2)、(3)及(5)所考慮的資訊是有用的。而在翻譯模型最多能推薦五個答案的情形下，***f***-measure 的趨勢走向與前一百名高頻動詞雷同，比起只能推薦一個答案時，每個公式組合的分數都有所提升，特別是與公式(4)搭配的公式組合。

8.2 使用科學人雜誌語料建置翻譯模型

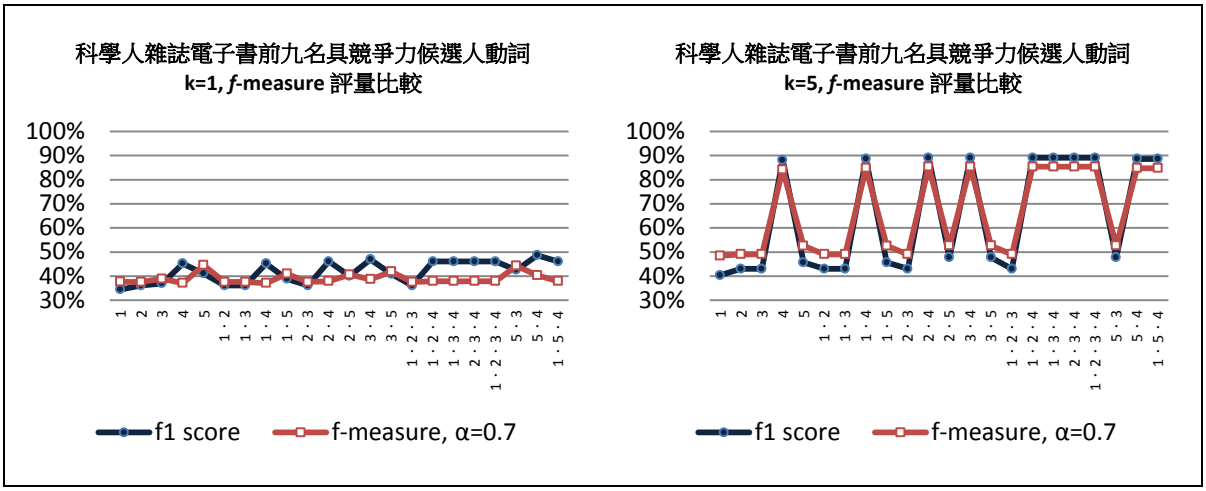
科學人雜誌語料庫中，本研究對列成功的英漢動名詞組合共有 4814 組。

8.2.1 科學人前二十五名英文高頻動詞

由於科學人雜誌語料所得的英漢動名詞組合數量比起專利語料少了許多，因此本研究探究前二十五名在我們 4814 筆的動名詞組合資料當中出現次數最多的英文動詞，這些動詞在資料中至少出現過 31 次以上，最多的出現次數則為 379 次。這二十五個英文動詞總共出現於 1885 筆資料之中，訓練資料共有 1508 筆，測試資料則有 377 筆。如圖九所示，在翻譯模型推薦一個答案及推薦五個答案時的趨勢分布與圖六專利語料的前一百名高頻動詞趨勢相同；因為語料數量較少（僅有專利語料的 13%）而資料變化又較大（科學人文章風格較專利文句豐富），因此在推薦五個答案時 ***f***-measure 最高的成效落在 90% 左右。

8.2.2 科學人前九名具競爭力候選人之英文動詞

本研究由前二十五名高頻動詞中選出了具競爭力候選人的動詞，這裡的「競爭力」意義相同：第一名候選人出現的次數不得多於第二名候選人出現次數的兩倍。這九個動詞總共出現在 689 筆英漢動名詞組合，訓練資料有 552 筆，測試資料則有 137 筆。如下頁圖十所示，科學人前九名具競爭力候選人動詞在 ***f***-measure 的趨勢大致上與專利前二十二名具競爭力候選人動詞的分布相同，不過可以特別注意到公式(5)在只能推薦一個答案



圖十、翻譯模型在科學人前 9 名競爭動詞推薦一個及五個答案時之 ***f***-measure 成效

且注重於答題正確率時，表現相對於其他獨立運作的公式突出，而與公式(5)搭配的組合也有較亮眼的表現；我們認為這是因為資料數量少而資料型態卻又豐富時，公式(5)反而可以用其獨特的觀點去猜到答案。在推薦五個答案時與公式(4)搭配的公式組合仍是表現最為亮眼。

8.3 小結

透過以上分析翻譯模型的表現，本研究提出的公式組合「共同推薦」不僅可以在推薦三個答案時幾乎就能找到正確解答，且可以透過蒐集資訊較多的公式把正確答案在推薦答案中的位置往前拉提，這對於翻譯效果都有正面的影響。

9. 受試者實驗評比

為了評比我們的翻譯模型是否能跟人類的翻譯能力競爭，我們從科學人語料中取出十句英漢對照的句子當作實驗題目，並設定三種翻譯英文動詞的實驗，邀請以中文為母語並具有資工背景的受試者參加。我們規定三種實驗的受試者不得重複跨實驗參加，每位受試者為獨立進行實驗。實驗一有 17 位受試者參與、實驗二有 19 位，實驗三則有 16 位受試者，共 52 位受試者參與實驗。我們使用公式(1)建置的翻譯模型作為參賽者，實驗題目則以公式(1)所能得到的資訊為基準，即受試者至少知道英文的動詞、名詞及中文的名詞這些資訊，不同實驗會附加其他不同程度的資訊以測試受試者會否因為附加資訊的多寡影響其答題效果。我們透過受試者的答題情況與我們的公式(1)翻譯模型作比較，驗證模型的翻譯效能。

9.1 三種實驗提供的題目資訊說明

在第一個實驗中，我們提供受試者英文及其中文翻譯的題目資訊，將題目中的英文目標動詞以灰底粗體標示，並將中文題目對應的動詞翻譯位置挖空，如下頁表四所示。為了不讓受試者只注意到英文的目標動詞及名詞而不完整閱讀題目，我們因此不將目標名詞特別標示。我們將正確答案藏在四個選項中，以表四為例，非正確答案的三個選項是從目標動詞「improving」在科學人語料對應的中文詞彙群中挑選出較高頻的三個詞彙當作誤導選項。這個實驗的目的為讓受試者在接收完整題目的資訊之下，要求受試者將目標動詞翻譯成中文詞彙，並提供選項作答。

表四、實驗一及實驗二題目範例

英文題目	Investigators are, of course, also exploring additional avenues for **improving** efficiency; as far as we know, though, those other approaches generally extend existing methods.
中文題目	當然，研究人員也在尋找其他可_____效率的方法，但就我們目前所知，其他方法一般只是延伸現有的途徑罷了。
答案選項	(1) 增進 (2) 提高 (3) 改進 (4) 改善
目標的中文翻譯群	improve={利用=1, 增加=1, 改良=1, 運用=1, 使=2, 加強=3, 提高=4, 改進=4, 增進=11, 改善=22}

表五、實驗三題目範例

題目	**improve** efficiency : _____ 效率
答案選項	(1) 增進 (2) 提高 (3) 改進 (4) 改善

關於第二個實驗，我們提供與實驗一相同的題目資訊，唯一不同的地方在於實驗一提供了四個選項讓受試者選擇，如表四所示，而實驗二不提供虛線框起的答案選項，直接要求受試者填寫他們心目中的詞彙。

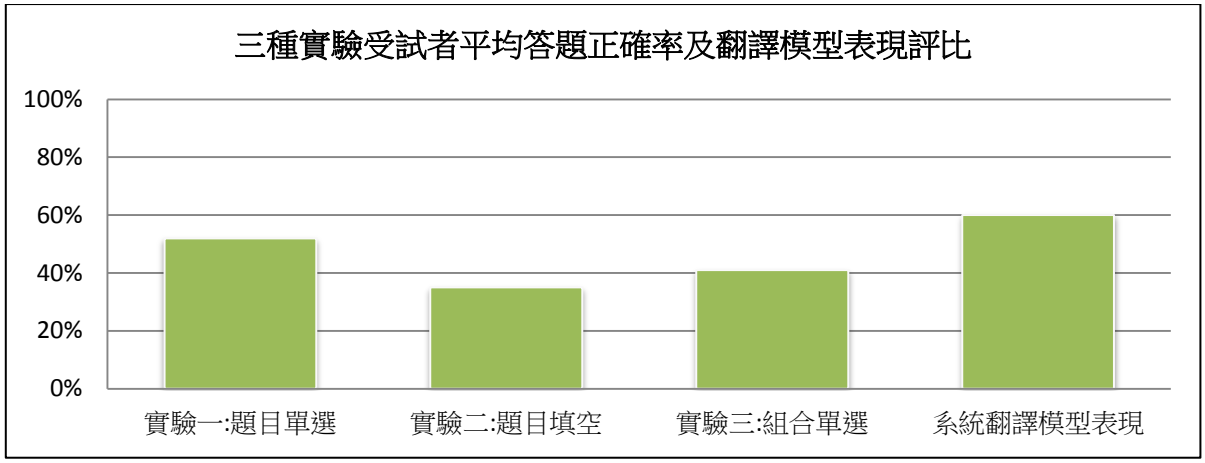
在第三個實驗中，我們不提供受試者題目的環境及提示，僅提供公式(1)翻譯模型所能得到的資訊給受試者，但是我們附加了答案選項提供選擇，如表五所示：我們僅提供英文動名詞組合及中文名詞，並將英文目標動詞以灰底粗體標記，要求受試者從我們答案選項中選出一個最適合的詞彙作答，這四個答案選項與實驗一的選項相同。

9.2 受試者與翻譯模型效能評比

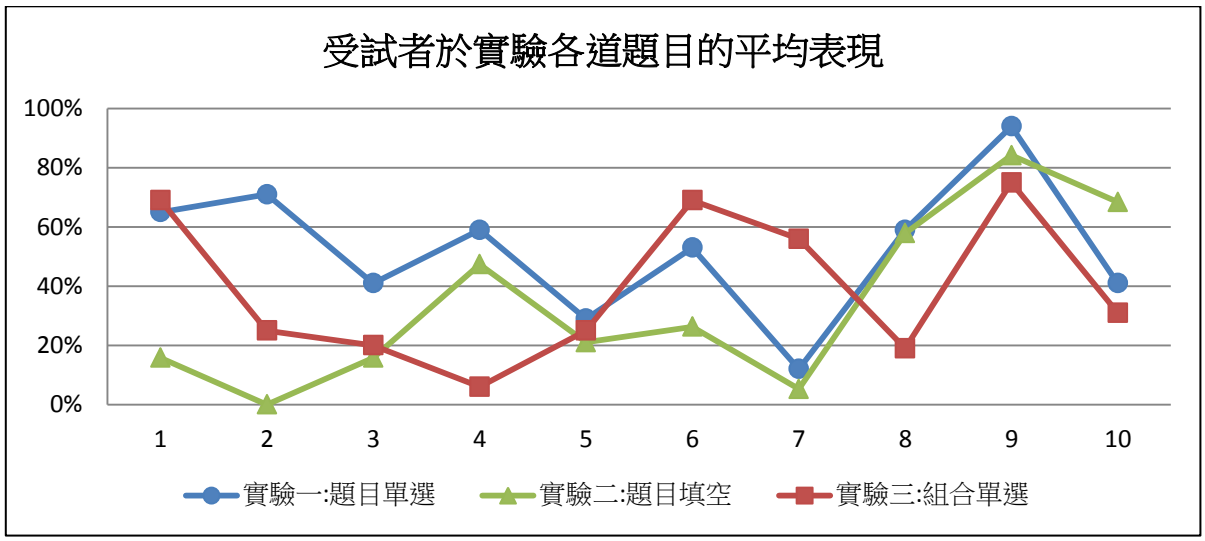
下頁圖十一為三項實驗受試者平均答題正確率及本研究翻譯模型的表現比較。實驗一提供最多的資訊，受試者平均答對的題數最多，約答對 50%；實驗二雖然提供英漢的題目資訊，但是沒有提供答案選項，受試者平均答對的題數最少，約答對 30%；實驗三的受試者則平均答對了 40%。本研究的翻譯模型答對六題，因此答題正確率為 60%，贏過三項實驗受試者的平均表現。這三個實驗讓我們發現，受試者在提供答案選項的實驗表現較為良好，即使我們提供了完整題目的資訊讓受試者填空，受試者還是很難猜出正確答案；這也就代表即使是人類來答題，在只能回答一個答案時都很難答出正確解答，而我們的翻譯模型則有較好的表現。

下頁圖十二為進一步觀察三群受試者的答題情形，本研究有有趣的發現。第一題的題目在提供題目語意及答案選項的實驗一及只有提供動名詞組合及答案選項的實驗三的答題效果相似，與實驗二的填空題則有很大的差距；第三題題目的實驗二及實驗三答題效果相似；第四題則是實驗三的答題效果最差；第五題確是三個實驗的效果都相似；第六及第七題反而是實驗三效果最好，但在第八題情形卻倒轉；第九題三個實驗的表現也接近，第十題卻是實驗二的填空題效果最好。

這些作答的現象讓我們認為人類在作答的時候，在題目提供的附加資訊多寡之外，人類在閱讀到動名詞組合時應該有其特定的直覺，而且直覺的影響力可能大過實驗所提供的附加資訊，不會根據附加資訊多寡而有固定的表現，因而產生這些有趣的曲線變化。而本實驗未蒐集受試者的個人資訊失為一考量，因此沒有受試者個人特質的相關統計評估，為本實驗待改進之處。



圖十一、三種實驗受試者平均答題正確率及翻譯模型表現評比



圖十二、受試者於實驗各道題目的平均表現

10. 結論

本研究使用了兩套科技技術類的英漢平行語料庫，並針對英漢動名詞組合進行英文動詞的推薦翻譯。我們分別使用了資訊蒐集程度不同的五種公式，建立針對英漢動名詞組合翻譯英文動詞的模型。我們的實驗結果顯示，將公式組合起來共同推薦能提供不錯的翻譯效果，且在 ***f***-measure 的評量下，與公式(4)搭配的公式組合效果最佳，其建置的翻譯模型推薦到三個答案時幾乎就能包含正確解答在內。蒐集資訊較多的公式(1)、(2)、(3)及(5)在與公式(4)一同搭配時，會將正確解答往前排在推薦答案中，特別是公式(1)的效果最為明顯，符合本研究對於公式(1)的期望。本研究對於英文名詞也建置了對稱的公式及翻譯模型並測試翻譯效果，因受限於篇幅的關係，我們僅描述翻譯結果，結果顯示成效與翻譯英文動詞差異不大，公式共同推薦的翻譯模型有良好的表現。

除了建置翻譯模型並比較成效，我們也設計了三項實驗讓受試者參與，並將受試者的答題正確率與我們使用公式(1)建立的翻譯模型比較表現。結果顯示三項實驗相比，本研究的翻譯模型都能贏過受試者的平均表現。

本研究透過翻譯模型的評量與分析，以及和受試者的翻譯表現作比較，可以驗證我們的翻譯模型具有不錯的推薦翻譯能力及表現。

致謝

本研究承蒙國科會研究計畫 NSC-100-2221-E-004-014 及 NSC-99-2221-E-004-007 的部份補助，僅此致謝。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] Yu-Chia Chang, Jason S. Chang, Hao-Jan Chen and Hsien-Chin Liou, An Automatic Collocation Writing Assistant for Taiwanese EFL Learners: A Case of Corpus-based NLP Technology. *Computer Assisted Language Learning*, 21(3), 283-299, 2008.

[2] Wei-Te Chen, Su-Chu Lin, Shu-Ling Huang, You-Shan Chung and Keh-Jiann Chen, E-HowNet and Automatic Construction of a Lexical Ontology, *Proceedings of the Twenty Third International Conference on Computational Linguistics*, 2010.

[3] Concise Oxford English Dictionary。 http://startdict.sourceforge.net/Dictionaries_zh_TW.php [連結已失效。]

[4] Dr.eye 譯典通。http://ajds.nsysu.edu.tw/learn/dict/ [Last visited on 15 June 2011]

[5] E-HowNet。http://ckip.iis.sinica.edu.tw/taxonomy/taxonomy-doc.htm [Last visited on 15 June 2011]

[6] Google。http://www.google.com.tw/ [Last visited on 15 June 2011]

[7] Google Patents beta。http://www.google.com/patents [Last visited on 15 June 2011]

[8] Bin Lu, Benjamin K. Tsou, Tao Jiang, Oi Yee Kwong and Jingbo Zhu, Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese Example and Its Application to SMT. *Proceedings of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 2010.

[9] Patent Translation Task at NTCIR-9。http://ntcir.nii.ac.jp/PatentMT/ [Last visited on 15 June 2011]

[10] Stanford Chinese Segmenter。http://nlp.stanford.edu/software/segmenter.shtml [Last visited on 15 June 2011]

[11] Stanford Parser。http://nlp.stanford.edu/software/lex-parser.shtml [Last visited on 15 June 2011]

[12] Sriam Venkatapathy and Aravind K. Joshi, Measuring the Relative Compositionality of Verb-noun (V-N) Collocations by Integrating Features. *Proceeding of Human Language Technology Conference on Empirical Methods in Natural Language Processing*, 899-906, 2005.

[13] WordNet。http://wordnet.princeton.edu/ [Last visited on 15 June 2011]

[14] 一詞泛讀。http://elearning.ling.sinica.edu.tw/c_help.html [Last visited on 15 June 2011]

[15] 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。

[16] 科學人雜誌英漢對照電子書。http://edu2.wordpedia.com/taipei_sa/ [Last visited on 15 June 2011]

[17] 國家教育研究院學術名詞資訊網。http://terms.nict.gov.tw/download_main.php [Last visited on 15 June 2011]