國立政治大學資訊科學系

Department of Computer Science
National Chengchi University

碩士論文

Master's Thesis

英文技術文獻中動詞與其受詞之中文翻譯的語境效用

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

研 究 生：莊怡軒

指導教授：劉昭麟

中華民國一百年七月

July 2011

英文技術文獻中動詞與其受詞之中文翻譯的語境效用

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

研 究 生：莊怡軒　　Student：Yi-Hsuan Chuang

指導教授：劉昭麟　　Advisor：Chao-Lin Liu

國立政治大學

資訊科學系

碩士論文

A Thesis

submitted to Department of Computer Science

National Chengchi University

in partial fulfillment of the requirements

for the degree of

Master

in

Computer Science

中華民國一百年七月

July 2011

英文技術文獻中動詞與其受詞之中文翻譯的語境效用

摘要

本研究使用英漢平行語料庫，試圖從中找尋英文與中文之間的翻譯情形，我們將英文及中文的動名詞組合 (V-N-collocation) 作為觀察對象。本研究各別分析英漢專利平行文句語料庫及科學人雜誌英漢對照電子書兩套語料庫，將中英文互為翻譯的文件視為一體，觀察英文及中文語言其中的特定結構及共現性 (collocation)，建構由真實世界的語料所反應的語言翻譯模型。

我們使用技術名詞表將平行語料庫進行技術名詞斷詞，再將句子進行結構剖析得到關係樹 (dependency tree)，並利用關係樹結構及近義詞典取得英漢動名詞組合。本研究運用英漢動名詞組合建立英文動詞與名詞的翻譯模型，我們的系統可以根據不同的模型推薦翻譯，並比較這些翻譯模型的成效；最後也加入中文語言使用者翻譯英文動詞的實驗與本研究的翻譯模型效果作比較，結果顯示本研究的翻譯模型比起受試者，可以有較好的推薦效果。

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

Abstract

In our investigation, we are interested in English Verb-Noun collocation (V-N collocation) and the corresponding usage in Chinese. To discover English-Chinese V-N collocation, a rich corpus is needed; therefore, we obtained one million English-Chinese parallel patent sentence pairs and seven years of bilingual *Scientific American* as two corpora to analyze. We trained translation models to find the usage of V-N collocations in English and Chinese. Given English V-N collocation and corresponding Chinese information, our system can recommend the proper translations of the English verb or object in collocation according to the translation models.

We experimented ten formulas to train our models using two corpora, and observed similar trends in the analyses. Preliminary comparisons of the translation quality of human subjects and our system indicated that our system could offer better recommendations for the translation tasks.

致謝

提起筆，準備寫下感謝。正要下筆的那一刻，卻在離紙張不到一公厘的距離凝結。只因為心情還沒準備好，要感謝的卻太多。

我珍惜研究生這個身分，特別當它發生在這一個科系，又是這一間實驗室。學習到的一切何以用文字就能闡述得明白。如果我不在這裡，也沒有這個身分，我想我永遠不知道，原來我可以有這麼多機會，完成這些自己從沒想過的事。那些機會，是讓我學習表達自己，讓我接觸研究的奧妙，讓我更接近這個世界，是讓我更茁壯一點。這些都是我最敬愛的指導教授　劉昭麟老師給予並教授於我的，劉老師是我最大的感謝；謝謝劉老師讓我更進步、邁向成為更好的人，我會十足地繼續努力！

親愛的家人絕對是我最大的支撐，謝謝你們。謝謝爸爸總是提醒我用功之餘要多多休息，謝謝媽媽總會聽我分享所有一切然後給我最棒的安全感，謝謝弟弟總在我忙碌時，貼心的端上一杯熱牛奶及熱宵，讓我能量百倍。最喜歡這個家，因為有你們。

兩年的歲月是成長中的一段過程，我要感謝機器智能實驗室成員的陪伴。即使實驗室的學長姐畢業了，我還是感覺到濃厚的關懷，學長姐關心著我們的研究，關心我們的生活；在我遇到困惑時，學長姐也總不吝惜地在百忙之中伸出雙手幫忙解答。學長姐的提攜，我好感謝。一同成長的同屆實驗室成員，建良與裕淇，謝謝你們一路來的扶持，與你們成為朋友，我的研究生活添加了好多色彩。最近我們總不自覺地聊到即將離別的話題，心裡都是一抹恬淡細長的惆悵。可愛的學弟妹，家琦、瑞平及柏廷，謝謝你們甜美的陪伴，乖巧的你們要努力加油，我相信你們都能做到，且做的很好。

不能常常待在身邊，但是卻都緊緊陪伴著我的朋友，謝謝你們，好多的鼓勵與支持，替彼此著想和高興，你們是我的活力來源。我喜歡我們在一起時的笑容，分享著彼此的世界，你們知道的，我是多麼地喜歡，我要繼續黏著你們。

我要感謝最親愛的侃文。謝謝你一路以來的陪伴，無時無刻的關懷；你是良師，是益友，是我生命中如此重要的人。謝謝你，豐富了我的人生。

能夠遇見美好的你們，是我最美好的幸福。

也謝謝口試委員　陳光華老師、　柯淑津老師及　張景新老師的指導。

莊怡軒　2011 年 8 月

目錄

圖目錄

表目錄

第一章　緒論

1.1 研究背景與動機

當今的社會可視為一個地球村，即使住在不同的國家、使用不同的語言，無論是商業貿易或是文化交流，人們相互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學習者 (EFL learners: English as a Foreign Language learners) 受到自身國家母語文法影響，容易在英文動詞及名詞的搭配組合上產生誤解及用法。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名詞組合與之對應至中文的關係感到有趣，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，許多不斷創新的發明與技術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，同時也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技術。專利文書的重要性更可以從 Google Patents beta[8]提供的英文專利文書檢索服務看出；Google[7]號稱其專利資料庫蒐集了七百萬篇以上的專利文書，以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正確解析、並排除技術名詞在外，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，可以作為雙語語料使用。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。由於我們希望有極豐沛的語料，能讓本研究統計並分析這些常見英文動名詞組合與中文動名詞之間的對應關係,因此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。

除了分析英漢專利文句平行語料庫[13]，為了比較不同語料是否有不同的分析結果，本研究另外以相同方式分析科學人雜誌英漢對照電子書[24]，以比較不同語料間是否有不同特性。本研究將英漢互為翻譯的文件視為一體，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。

1.2 研究方法

下頁圖 1.1 為本研究的系統流程圖。我們使用技術名詞表將英漢平行語料庫進行技術名詞斷詞,句子中剩餘未斷詞部分,我們使用 Stanford Chinese Segmenter[14]對於中文文句斷詞，英文文句則使用 Stanford Parser[15]及其字典模型進行詞幹還原。接著運用 Stanford Parser 將斷完詞後的句子進行結構剖析，得到關係樹結構 (dependency tree) ，再從關係樹結構取得句子中的動名詞組合。中文及英文文句都取得各自的動名詞組合後，本研究使用牛津現代英漢雙解詞典[4]、Dr.eye 譯典通線上字典[5]、E-HowNet[6]及一詞泛讀系統[21]製作成近義詞典，並使用近義詞典的資訊對列英漢動名詞組合。對列完成的英漢動名詞組合為本研究訓練及測試模型的資料，最後產生系統翻譯模型。

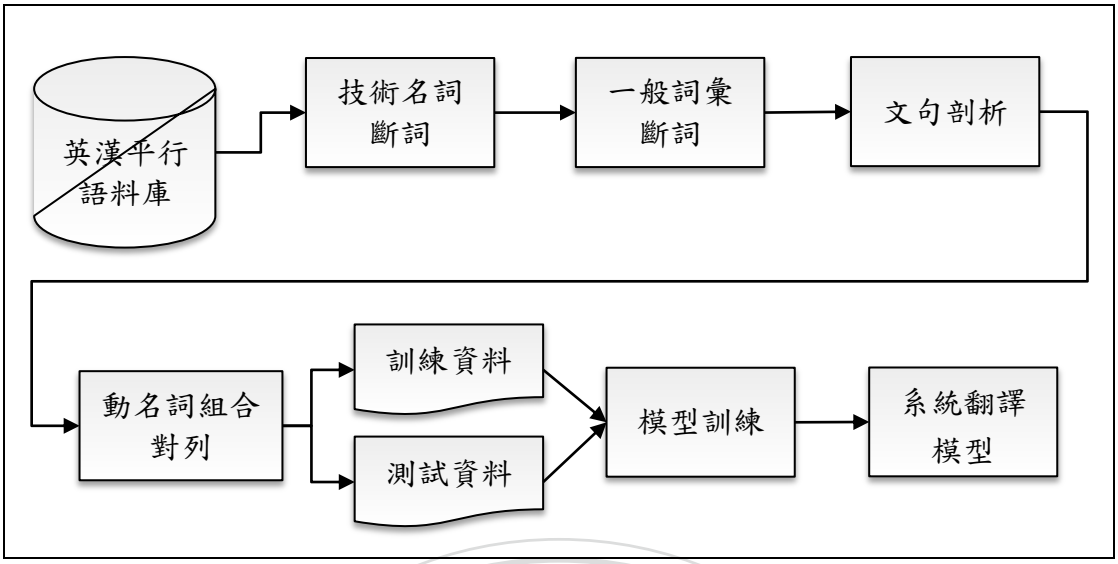


圖 1.1 系統流程圖

1.3 研究成果

本研究分析兩套同屬科技類但是不同性質的英漢平行語料庫：專利文句及科學人雜誌，以相同的方式處理語料、建置模型及評量翻譯效果。本研究分別針對英文動名詞組合中的動詞與名詞翻譯成中文，並設想加入中文對應的資訊是否能增進翻譯效能，因此進一步各別取出較難翻譯的動詞和名詞試探翻譯模型成效；實驗結果顯示本研究所提出公式組合翻譯模型能在提供五個答案時幾乎都能包含正確的翻譯答案，且經過我們的公式組合可以將正確答案往前排序。目前實驗顯示增加中文對應資訊時，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與名詞的翻譯貢獻度。

1.4 論文架構

在第一章的部分我們描述研究背景、研究方法成果及系統流程結構，第二章則介紹與專利文書、英文教學輔助翻譯、使用子結構輔助翻譯及英文動名詞組合等相

關研究。第三章交代本研究所使用的專利語料來源及技術名詞表的建置方法。第四章描述專利語料的前處理過程；第五章介紹本研究翻譯模型的原理公式。第六章與第七章個別使用專利文句和科學人雜誌語料建置翻譯模型並分析比較翻譯成效。另外在第八章設計了三樣實驗請具有資工背景的受試者參與，並比較受試者及本研究翻譯模型的表現。第九章為本研究的結論及未來展望。

第二章　文獻探討

文獻探討部分分為四個主題，2.1 小節介紹專利文書的相關研究；2.2 小節則為英文教學輔助翻譯的相關研究；2.3 小節描述了使用文句子結構資訊輔助翻譯的相關研究，2.4 小節則為針對英文動名詞組合的相關研究。

2.1　專利文書之相關研究

為了發掘專利文書的不同屬性以作參考，了解專利文書的相關研究相當重要；以語言的考量而言，專利文書除了作為保護智慧財產權的文件，其文件內容及架構其實可以作多面向的語言特性分析、系統的分析語料或是產生專利雙語對應的平行語料庫。以下是針對專利文書作相關研究的介紹。

同一篇專利文書可以發表不同語言的版本，而不同語言版本之間通常為全文篇幅的對應，文字細節部分的對應可能並不一致。田侃文[23]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英文句子的文句對列層級。本研究使用此系統，將英漢翻譯的專利長句視為一篇文章，由此系統產生短句之間的對列，提升對列文句的品質，在第三章會有更詳細的說明。

曾元顯[26]針對五十萬筆漢英專利平行語料文句，提出從語料中自動擷取中文與英文互為翻譯關係詞彙的系統。其使用了相互資訊 (mutual information) 、

相關分析 (correlation coefficient) 、可能性比例 (likelihood ratios) 、Dice 係數 (dice coefficient) 、分數累積 (fractional count) 及 EM 分析 (Expectation-Maximization analysis) 進行分析，發現使用 EM 的效果最佳。該研究亦將原本已有的中英技術名詞詞對組合加長比對，以擴充新的技術名詞詞對。

Lu [11]提出如何建置英漢專利文句對列的語料庫。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（例如：標題、摘要及專利範圍等）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%。

2.2 英文輔助翻譯教學之相關研究

如果跳脫出專利文書的世界，我們所注重的動名詞共現性或是其他詞彙間的關聯性是為真實世界生活中的問題。許多研究對於英文學習者容易共同犯錯的現象及特徵有不同的分析及統計方式；在教育目的上，如何增進英文學習者的英文能力已出現許多學習系統，以下是針對語言教育於詞彙特性的介紹。

Jian[10]使用 British NationalCorpus (BSC) 作為主要分析的英文語料，並運用其英文文句的子句結構 (clause parse) 及組塊 (chunking) ，提取出英文動名詞片語（包括 VN、VPN 及 VNP），計算動詞與名詞之間的共現性進而列表出英文語料中動名詞片語的共現性情況。該研究另外使用 Sinorama Parallel Corpus (SPC) 英漢平行語料庫，其運用詞彙對列技術 (word alignment) 來找尋中英文互為翻譯的動名詞片語。該研究將詞彙對列的方法為：首先判定英文名詞的中文翻譯，再依據中文翻譯句中離該中譯名詞最近的動詞，視為與英文動詞相對應的中文翻譯。

Chang [2]延續了 Jian[10]的基本做法。其針對把英文作為第一外語學習的中文使用者製作一套英文寫作校正系統。將英文作為第一外語學習者容易錯誤使用英文的動名詞片語組合，為了改善這個情形，該研究讓使用者能將寫好的英文文章輸入至該系統，系統便可偵測動名詞片語有無誤用之處，若有則提醒修正。該研究蒐集了正在學習英文之中文使用者的寫作文章當作學習者語料庫 (learner corpus) ，從中發現常見的錯誤用法；另外蒐集正確的英文語料當作正確答案的參考語料庫 (reference corpus) 。其主要方法為：依據參考語料庫中文句的子句結構 (clause parse) 及組塊 (chunking) ，找出相鄰的動詞片語 (VP) 及名詞 (NP) ，統計他們的共現性並輸出成結果。當系統使用者將寫好的英文文章輸入至系統，系統便找出當中的動名詞片語，查詢其共現性分數，若分數低於門檻值，則視為寫法錯誤；該系統將錯誤的動詞翻譯成中文詞彙，重新翻譯回英文詞彙，再將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。

Gamon[12]沒有像 Chang[2]去蒐集使用者語料庫、或是像 Jian[10]一樣運用英漢平行語料庫；Gamon 使用了 English Encarta encyclopedia 語料庫作為主要的英文語料。該研究利用決策樹及 5-grams 的資訊，針對介係詞及冠詞訓練語言模型。該研究也提出了一套系統，只要系統使用者輸入的英文句子有錯誤的冠詞或介係詞，系統便去計算冠詞或介系詞是否該出現或改變，才能接近真正答案的機率；如果將寫錯的冠詞或介係詞改變成系統推薦的詞彙的機率值超過門檻值，則將之作為系統的推薦修正答案。

2.3 運用文句子結構進行翻譯之相關研究

使用不同的語料會發現語料一些特別的屬性，多樣化的研究方法從不同角度觀看問題，都有不一樣的研究成果。如何找到不同語言之間的翻譯關係，有學者著重於利用文句子結構以限定翻譯範圍，依循文法規則尋找翻譯對應。以下介紹使用子結構幫助日英翻譯及使用中英翻譯結果改善英文剖析器的相關研究。

YOKOYAMA[20]針對專利文書的語料進行分析，該研究指出，專利文句的結構複雜且字數偏長，要進行分析及翻譯都是困難的。其使用日本專利局 (Japan Patent Office) 提供的公開專利文書並採用摘要部分，再利用人工翻譯得到日英的專利平行語料庫。該研究假設不同的 Japanese case frame 可能會對應到不同的英文翻譯，進而分析這樣的假想是否成立。如果不同的 case frame 組合會有不同的翻譯結果，則可以使用 case frame 資訊作為翻譯詞彙的挑選及限制條件。該研究發現，日文的傳統動詞並不容易從此方法得到對應的英文翻譯，如果是日本名詞常轉當動詞使用的動詞，則較有多義的情形，使用 case frame 有較好的翻譯效果。

英文的子句修飾問題 (prepositional phrase attachment problem) 一直是機器翻譯或是剖析器所欲解決的問題，Chen[3]便提出使用中文的語言特性輔助以解決這個問題的方法。英文及中文都是具備主謂賓結構 (SVO: subject verb object) 的語言，該研究認為，中文不論在前置詞、後置詞及所有格都有主謂賓結構的特性：即中文詞彙的出現順序有較明顯的修飾關係（前面的詞彙通常為修飾後面出現的詞彙）。其使用詞彙對列技術將中英文的關係樹進行對列，統計其對列關係及出現次數，並將高頻的出現關係當作規則，以此規則進行中英文的子句翻譯。該研究的實驗指出，採用中文語言的特性確實有助於提升英文剖析器解決介詞短語問題。

2.4 動名詞組合共現性之相關研究

動詞與名詞的組合現象是許多不同領域的學者都感興趣的議題;教育學者關心於如何教導及糾正學生正確使用片語,語言學家善於分析片語的特徵,心理學家分析挖掘人類使用片語的習慣及背景等等。在這裡,我們介紹資訊科學領域的學者對於動名詞的想法及相關研究。

Venkatapathy[16]首先介紹了 multi word expressions (MLEs) ,即為從字面上看不出實際表達意義的詞彙。有很大一部分的 MLEs 具有文法結構性但是沒有語義合成關係。MLEs 其中一個子集就是動名詞組合,也是該研究主要分析的目標。MLEs 很難區分是為組合性 (compositional) 或為非組合性 (non-compositional) ,在早一些時期的研究方式不外乎是考慮頻率 (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。

第三章　專利語料來源與技術名詞表建置

本章主要說明專利資料的處理，3.1 小節說明我們的專利語料來源及篩選方式取得高品質的專利句對;3.2 小節描述如何建置並過濾取得較高品質的技術名詞表。

3.1　專利語料來源

本研究使用 Patent Translation Task at NTCIR-9[13]一百萬筆英漢對照的專利文句作為研究語料，中文部分為簡體中文。該份語料分為兩個檔案，一為英文專利文句，另一則為對應英文句的中文專利文句，並使用編號末碼標示對應關係，如表 3.1 所示。3.1.1 小節敘述進行短句切割求得較高品質對應，3.1.2 小節則描述技術名詞斷詞問題。

表 3.1 英漢專利文句對應關係

英文專利文句	中文對應專利文句
WO9830090-2 First, the antimicrobial agent must be soluble or dispersible in the cyanoacrylate composition at the concentrations necessary to effect antimicrobial properties.	CN1246032-2 第一，抗微生物剂在腈基丙烯酸酯组合物内必须是可溶或可分散的，其浓度需要达到能产生抗微生物性质。

表 3.2 英漢專利短句對列範例

	英文專利文句	中文專利文句
原始長句	Accordingly, in one of its composition aspects, this invention is directs to an antimicrobial cyanoacrylate composition which comprises: (a) a polymerizable cyanoacrylate ester;	因此，在本发明组合物的其中一个方面，本发明涉及一种抗微生物组合物，它含：(a)可聚合的腈基丙烯酸酯；
短句對列	Accordingly,	因此，
	in one of its composition aspects,	在本发明组合物的其中一个方面，
	this invention is directs to an antimicrobial cyanoacrylate composition which comprises:	本发明涉及一种抗微生物组合物，它含：
	(a) a polymerizable cyanoacrylate ester;	(a)可聚合的腈基丙烯酸酯；

3.1.1 短句切割提升對列品質

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受到句長及結構影響而降低結果品質。為了改善對列品質可能下降的問題，我們提出這樣的觀點：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（例如：逗號、分號、冒號、驚嘆號、問號及句號）作為句子的終點；如此，一個長句即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[23]，將英文及中文的專利文句依據標點符號拆成短句組合，得到短句之間更細微的對應關係。短句不一定是一對一的對應關係，該系統可支援至四對一句的翻譯模組，因此我們相信該系統可以為本研究取得高品質的短句對列。短句對列的範例如表 3.2 所示。

專利文句對列系統會計算英漢文句對列的對應分數，因此我們設定值得信賴的門檻值取得較高對列品質的短句，作為我們的使用資料。在原本的一百萬組長句對中，超過本研究設定的門檻值有 338846 組長句對；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。這些短句對經過人工抽樣檢驗，我們相信是具有正確翻譯關係的英漢對列文句。

3.1.2 專利文句的斷詞問題

專利文書最大的特色，就是其內容包含許多技術名詞；而本研究為了排除技術名詞的資訊，以獲得常用的英漢動名詞組合，我們必須將技術名詞正確標記以便去除。技術名詞與一般詞彙（這裡所指稱的「一般詞彙」是指日常生活中對話、寫作或閱讀所習慣的用詞。）性質不同，不同專業領域有不同的技術名詞，而技術名詞通常含有知識性及專業意義；非專業領域、不熟悉技術名詞用法的人，如果要認知技術名詞的涵義有其困難性。就人類的閱讀上而言，我們需要有基本的詞彙單位判斷機制；例如，在英文專利文句中看到「adaptation level theory（適應水準理論）」這三個英文詞彙，如果閱讀者具有相關的專業背景，就不會把三個詞彙分開來閱讀，因為這三個詞彙的出現具有特定專業意義，是一個技術名詞，單位是一個複合詞。若是在中文專利文句出現「適應水準理論」，我們可能會誤解成「適應」為動詞、「水準理論」是一個詞彙，或是解讀成「適應」「水準」的「理論」，一樣需要有專業的知識才會知道這六個字是為一個技術名詞。

如果人類要讀懂技術名詞需要有「知道這是技術名詞」的基本條件，那麼透過技術名詞表將專利文句中的技術名詞斷詞，就是讓剖析文句的系統能夠「知道」分割出來的是「技術名詞」，而不是當成一般詞彙處理。因此，得到較高品質的短句對後，如何將專利文句正確斷詞是我們接下來要解決的問題。

如果直接使用一般的方式斷詞，會造成技術名詞被錯誤切割、失去專利文句及技術名詞的原意，錯誤標記詞性，甚至造成文句結構被嚴重扭曲，再經過剖析器就會得到錯誤的剖析結果，對於我們想要尋找動名詞組合是很大的阻礙。因此，將技術名詞完整切割、並指定其詞性為名詞為最能幫助文句保持原意及被正確解析的方法。為了不讓技術名詞被錯誤斷詞，我們需要建立一個技術名詞資料表，以供技術名詞斷詞的比對；如果詞彙比對成功，便將專利文句中的技術名詞切割並標記之。我們以表 3.3 來說明英文技術名詞的斷詞問題。如果直接將未斷詞例句直接使用 Stanford Parser[15]進行剖析，Stanford Parser 會將該技術名詞斷為好幾個詞彙及詞性，使得技術名詞的特色消失，且剖析成不正確的結構樹。表 3.4 則為中文技術名詞的斷詞範例，文句中若包含化學合成物，通常會是關鍵的技術名詞。目前最常被用到的斷詞系統為中央研究院的中文斷詞系統[22]，但若直接將範例詞彙「羰基化戊烯腈」送至中研院斷詞系統，其回傳的斷詞結果不但有錯誤，甚至有罕見字「氰」的編碼錯誤的問題。由上述的兩則範例可知，正確切割技術名詞是基本且重要的步驟。為了求得更精確的技術名詞以增進斷詞效能，在下一節本研究將描述我們所蒐集的技術名詞表來源及技術名詞表的過濾方法。

表 3.3 英文技術名詞斷詞範例

原始詞彙	abbreviated address calling
正確斷詞	abbreviated address calling/NN
錯誤斷詞	abbreviated/NN address/NN calling/VBG

表 3.4 中文技術名詞斷詞範例

原始詞彙	羰基化戊烯腈
正確斷詞	羰基化戊烯腈/NN
錯誤斷詞	羰(FW) 基(Nc) 化(VG) 戊烯(Na) &#(FW) ３３０９６(Neu) ;

表 3.5 技術名詞表內容格式

英文技術名詞	對應的中文技術名詞
acceptable price range	可接受價格範圍
accessory olfactory bulb	副嗅球
accessibility heuristic **accessibility heuristic**	**易提取性捷思法** **易觸及性捷思法**
anamnia，Anamniota	***無羊膜動物***
densitometer; scanning	***掃描密度計***
demodulator; product; product detector ***demodulator; product; product detector***	***乘積解調器*** ***乘積檢波器***

3.2 技術名詞表建置

本研究使用國家教育研究院學術名詞資訊網[25]公開的技術名詞檔案整合為技術名詞表；我們取得 138 個不同領域的技術名詞 Excel 格式檔案，檔案大小共有 177MB，並統整成技術名詞表。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，如表 3.5 所示。3.2.1 與 3.2.2 小節分別描述如何使用 E-HowNet[6]及 WordNet[17]過濾技術名詞表，3.2.3 為小結論。

表 3.5 以粗框圈選的技術名詞代表同一個英文技術名詞對應到不只一個中文技術名詞翻譯。在這樣的情況下，我們把一對多的對應關係分列為一對一的對應模式，如「accessibility heuristic」對應到兩個不同的中文技術名詞，則在技術名詞表中會拆成兩筆紀錄儲存。以灰底及粗斜體標示的末四列，其英文技術名詞內含的標點符號具有不同的標示意義：倒數第四列的逗號表示前後詞彙是相等的（anamnia 等同 Anamniota）；而倒數第三列的分號表示分號後面的詞彙應搬到分

號前面的詞彙之前（scanning densitometer 等同於「掃描密度計」）；末兩列的分號意義就又不一樣了，第一個分號表示「product」應搬至「demodulator」之前形成「product demodulator」並對應至中文技術名詞的「乘積解調器」，而第二個分號表示「product detector」應對應至「乘積檢波器」。由上述的內容可以發現，英文技術名詞的相隔符號所代表的意義複雜，即使人為都不容易辨認其符號意義，我們亦無法精通各個專業領域以完全解讀符號帶有的實屬分隔意義對英文技術名詞作拆解。技術名詞表中具有分隔符號的英文技術名詞僅佔極少的比例，因此本研究不針對英文技術名詞的標點符號作拆解處理，僅將之簡單視為一筆技術名詞。我們的技術名詞表依照上述的規則，總共記錄了 804068 個英漢對應的技術名詞詞對。

我們發現，在技術名詞表當中，無論是英文或是中文，都有些許的技術名詞更常被當作一般用語詞彙。我們嘗試直接以 804068 個詞對將專利文句作斷詞，發現句子中幾乎每一個詞都被當作是技術名詞；許多詞彙為一般常用詞彙，卻被錯誤標記為技術名詞。探究其原因，發現從學術名詞資訊網取得的檔案含有不少一般常用詞彙。為了過濾這些詞彙，本研究提出使用 E-HowNet 及 WordNet 來幫助我們刪除一般詞彙，留下技術名詞於技術名詞表，以下兩小節作更多說明。

3.2.1 使用 E-HowNet 過濾技術名詞表

中央研究院所開發的 E-HowNet 是根據 HowNet[9]的語義義原知識本體架構修改建構而成， E-HowNet 內含 88075 個中文詞彙。本研究認為， E-HowNet 所收錄的中文詞彙可以代表我們日常生活中一般常用的詞彙，使用這些詞彙幫助過濾技術名詞表是可行的方式之一。如果技術名詞表中的中文技術名詞也有出現於 E-HowNet，我們相信該詞對應歸類為非技術名詞，當作一般詞彙使用的機率較

大，因此除去該詞對。E-HowNet 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。

3.2.2 使用 WordNet 過濾技術名詞表

使用 E-HowNet 過濾技術名詞表是從中文的角度發想，我們也須對稱地檢驗技術名詞表中是否內含英文的一般詞彙。我們相信 WordNet 包含的英文詞彙可以視為一般日常生活的英文用語代表，因此，除去技術名詞表與 E-HowNet 的交集後，我們改以英文詞彙的角度觀看，採用 WordNet 來幫助過濾技術名詞表。WordNet 中含有 154754 個英文詞彙及英文短片語。經過 WordNet 的比對，總共過濾了 80220 個詞對。雖然除去了八萬多個詞對，但是有許多詞對的英文詞彙是重複的，實際上並沒有真的除掉八萬多個英文詞彙，僅除去 29861 個英文詞彙。

3.2.3 小結

經過 E-HowNet 和 WordNet 的檢測，我們的技術名詞表約略除去了 14%的詞對，現存有 690640 個技術名詞詞對。我們相信這六十九萬個技術名詞詞對具有較高的品質，即為較準確的專業領域用語，降低與一般詞彙產生斷詞衝突的機率。

第四章　語料前處理及近義詞典建置

在本章我們介紹專利語料的前處理及本研究所使用的辭典建置方式。4.1 小節及 4.2 小節各自描述英文及中文專利文句的前處理，4.3 小節則描述建置近義詞典以進行英漢動名詞組合對列。本研究使用的 Stanford Parser[15]為版本 1.6.5，Stanford Chinese Segmenter[14]則為版本 2008-05-21。

4.1 英文專利文句前處理

在這一小節主要描述英文專利文句的前處理過程。4.1.1 小節進行技術名詞標記，4.1.2 描述詞幹還原，4.1.3 則為英文關係樹剖析。

4.1.1 英文技術名詞斷詞及標記

技術名詞多為複合詞彙，因此我們使用長詞優先的方式，從技術名詞表比對專利文句的詞彙，一經比對成功則將技術名詞標記並使用 Stanford Parser 的 TaggedWord() 函數指定詞性為名詞，以便除去該資訊。結果如下頁表 4.1 所示，以粗體標示並以「**<*** ***>**」前後標記者即為比對成功的技術名詞。

表 4.1 英文標記技術名詞前後對應範例

原始英文專利文句	技術名詞斷詞後文句
Such materials include, by way of example, inorganic materials such as Type 1 glass (including amber glass).	such material include , by way of example , **<***inorganic materials***>** such as type 1 glass (include **<***amber glass***>**) .

表 4.2 英文文句詞幹還原

原始文句	Such materials include, by way of example, inorganic materials such as Type 1 glass (including amber glass).
詞幹還原	such material include , by way of example , **<***inorganic materials***>** such as type 1 glass (include **<***amber glass***>**) .

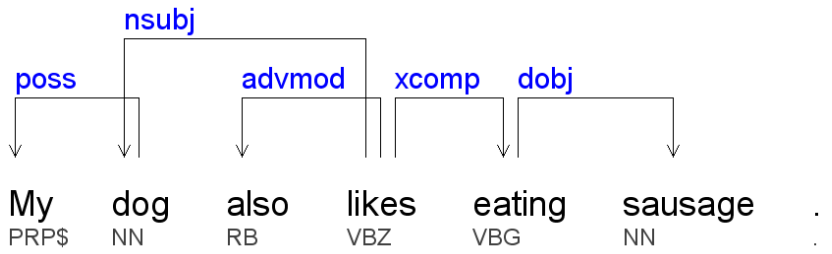
4.1.2 英文詞幹還原

完成技術名詞斷詞之後，根據 Stanford Parser FAQ 的建議，我們使用 englishPCFG.ser.gz 這部字典模型剖析英文專利文句，能夠較快速獲得效果不錯的剖析結果。本研究亦運用 Stanford Parser 的 Stemmer() 函數進行詞幹還原。在這個步驟，已經完成斷詞程序的技術名詞並不會被更動，且會維持其名詞詞性；剩下未斷詞的文句部分則會進行詞幹還原，且令 Stanford Parser 依據 englishPCFG.ser.gz 字典模型斷詞及標記詞性。結果如表 4.2 所示，粗體標示並以「**<*** ***>**」前後標記者為技術名詞，灰底標示者為詞幹還原的前後比較示意。

4.1.3 英文關係樹剖析

確認詞彙單位及詞性標記步驟之後，本研究繼續使用 Stanford Parser 剖析文句得到關係樹結構，Stanford Parser 的關係樹剖析總共含有 27 種文法關係的標記。一個句子經過剖析可以得知這個句子含有幾種文法關係，下頁表 4.3 即為關係樹

表 4.3 英文關係樹範例

輸入句	My dog also likes eating sausage.
關係樹樹狀圖	
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)**

範例。27 種文法關係中，「DIRECT_OBJECT」可以標記出文句中動詞片語的動詞及其述語對象，並以「dobj」為形式；例如表 4.3「My dog also likes eating sausage.」一句中，動詞「eat」的對象是名詞「sausage」，因此這兩個詞彙之間的關係會以「dobj(eating-5, sausage-6)」的形式標記表達關係；其中數字 5 與 6 代表詞彙在文句中出現的位置次序。

句子經剖析後，透過抽取其關係樹的「DIRECT_OBJECT」表示式，即可得到這個句子中互有描述關係的動名詞組合。

4.2 中文專利文句前處理

中文有斷詞的問題，與處理英文的方式不盡相同。英文的書寫方式為詞彙之間有空白相隔，但是中文的書寫方式則是字與字左右相接、其中並無間隔空隙。因此，中文除了技術名詞，一般詞彙也需要斷詞處理，將詞彙作為基本單位才能進行下一步的關係樹剖析。我們一樣使用技術名詞表比對中文專利文句，將技術名詞斷詞並標記詞性為名詞。剩下的文句部分大多為一般常見詞彙，本研究使用 Stanford

Chinese Segmenter[14]進行斷詞。以下各就兩小節 4.2.1 及 4.2.2 依序描述斷詞方法：技術名詞及一般詞彙斷詞，4.2.3 小節則描述中文關係樹剖析。

4.2.1 中文技術名詞斷詞及標記

中文技術名詞斷詞與標記和英文技術名詞的處理方式相同，以長詞優先的方式，將中文專利文句比對技術名詞表，並使用 Stanford Parser 的 TaggedWord()函數將技術名詞指定為名詞詞性。如表 4.4 所示，粗體並以「<*** ***>」標示者為比對成功的技術名詞。

4.2.2 使用 Stanford Chinese Segmenter 斷詞

標記完技術名詞之後，剩下未斷詞的文句部分，我們使用 Stanford Chinese Segmenter 進行斷詞。我們將斷好的詞彙以空白相隔，結果如表 4.5 所示。

表 4.4 中文技術名詞標記範例

初始文句	包括(但不限于)弹性蛋白酶的释放以及超氧化物的产生和活化的特性。
技術名詞標記	包括(但不限于)<***弹性蛋白酶***>的释放以及<***超氧化物***>的产生和活化的特性。

表 4.5 中文一般詞彙斷詞範例

斷詞結果	包括 (但 不 限于) <***弹性蛋白酶***> 的 释放 以及 <***超氧化物***> 的 产生 和 活化 的 特性

表 4.6 中文關係樹範例

輸入句	老師宣告了學生的成績。
關係樹結構	nsubj(宣告-2, 老師-1) asp(宣告-2, 了-3) assmod(成績-6, 學生-4) assm(學生-4, 的-5) **dobj(宣告-2, 成績-6)**

4.2.3 中文關係樹剖析

與 4.1.3 小節相同，我們使用 Stanford Parser 剖析文句得到關係樹結構，一樣使用「DIRECT_OBJECT」找出文句中動名詞組合的關係，如上頁表 4.6 所示，「老師宣告了學生的成績。」一句中，名詞「成績」就是動詞「宣告」的描述對象，因此這兩個詞彙之間的關係會以「dobj(宣告-2, 成績-6)」這樣的形式標記關係。有了英文與中文的動名詞組合，我們可以使用查詢辭典的方式，將英文與中文的動名詞組合翻譯對列，完成我們的英漢動名詞組合資料庫。

4.3 英漢動名詞組合對列

我們已經擁有英文及中文各自的動名詞組合，接下來就要把互為翻譯對照的動名詞組合對列產生翻譯結果。我們使用的方法是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[4]與 Dr.eye 譯典通線上字典[5]。但是只依靠英漢辭典的資訊是不足夠的，因為英漢辭典所列出與英文詞彙對應的中文翻譯詞彙有限；為了找尋更多與英文詞彙對應的中文翻譯詞彙，我們另外使用了一詞泛讀[21]及 E-HowNet [6]建立近義詞典，擴充我們的英漢詞彙對應，幫助英漢動名詞組合對列。4.3.1 小節為合併英漢辭典，4.3.2 說明近義詞典的建置過程。

4.3.1 英漢辭典合併

本研究使用兩部辭典，分別為牛津現代英漢雙解詞典（以下簡稱牛津詞典）與 Dr.eye 譯典通線上字典（以下簡稱譯典通字典）。如果查閱辭典的內容，英文詞彙的中文翻譯約略可以分為兩種翻譯情形：第一種為與英文詞彙相等對應的中文

詞彙，即為同一種意義的事物在不同語言中的詞彙使用對照，例如：「egg」與「蛋」的相等對應關係），這樣的詞彙本研究稱之為「對應詞彙」；第二種則為以中文片語解釋該英文詞彙的意義，是屬於語意上的理解說明，例如：「effusion」與「（尤指無約束的）思想和感情的流露；抒發感情」的註釋關係。我們需要英漢翻譯的詞彙可由第一種相等詞彙對應關係取得，第二種註釋關係的內容屬於語意解釋，主體對象為人類，因此不列入我們使用基於辭典資訊的機器翻譯方法。

4.3.1.1 牛津現代英漢雙解詞典

在牛津詞典中，並非每個英文詞彙都列有中文對應詞彙，中文翻譯部分亦混雜著兩種翻譯情形，辭典中的例句也一併出現於中文翻譯部分，而且沒有明顯的規則可以直接取出英文詞彙的中文相等詞彙。為了解決這個問題，我們將英文詞彙的中文翻譯根據標點符號為分割單位，分列出許多的中文候選字串。我們設定了門檻值：如果候選字串的長度不超過四個字，我們認為該字串是為中文對應詞彙的機會較大，予以保留；如果候選字串的長度過長，我們相信該字串屬於第二種語意說明的中文解釋的機會較大，便予以剔除。

表 4.7 牛津字典內容範例

英文詞彙：confusion	
中文對應詞彙	辭典中的語意解釋或例句
迷亂；惶惑	gazing in confusion at the strange sight 惶惑地凝視著這種奇怪的景象
混亂；雜亂	Her unexpected arrival threw us into total confusion. 她來得很突然，使我們完全不知所措。
混淆；混同	There has been some confusion of names. 有些名字弄混了。
~~不確定狀態~~	There is some confusion about what the right procedure should be. 對應該採取怎樣的步驟這一點還不太明確。

牛津詞典含有 39178 個英文詞彙，本研究依上述的規則運作，總共得到了 26896 個英文詞彙含有中文對應詞彙。見上頁表 4.7，以英文詞彙「confusion」為例，最後我們抽取出「迷亂、惶惑、混亂、雜亂、混淆及混同」作為我們的中文翻譯詞彙，最後一個候選字串「不確定狀態」因為長度超過四個字因此不列入中文對應詞彙內。

4.3.1.2 Dr. eye 譯典通線上字典

譯典通字典含有 106276 個英文詞彙，而且由 XML[19]格式撰寫而成，因此可以由格式標記取得英文詞彙的中文翻譯部分，不會採取到例句的部分；但是中文的翻譯部分仍然有上述的兩種翻譯情形，因此與牛津詞典相同，使用標點符號為分割單位來切割字串。由於譯典通字典的中文翻譯部分不含有例句，因此我們將詞彙的詞長條件放寬，將不超過五個字的候選字串視為中文對應詞彙，超出五個字的字串則視為語意解釋，不列入採用。我們一樣使用英文詞彙「confusion」作範例，表 4.8 中的「混亂、騷動、混亂狀況、混淆、困惑及慌亂」即是被我們認為的中文對應詞彙。字典中的十萬個英文詞彙，其中有 88507 個英文詞彙具有中文對應詞彙。

表 4.8 譯典通字典內容範例

英文詞彙：confusion	
中文對應詞彙	辭典中的例句
混亂；騷動； 混亂狀況	The room was in a state of confusion. 房間一片雜亂。
混淆	You can avoid confusion by speaking clearly. 你說得清楚些，這樣可以避免誤解。
困惑；慌亂	The old woman looked at him in confusion. 老婦人用迷茫的目光打量著他。

表 4.9 合併字典範例

英文詞彙：confusion	
辭典	辭典中的中文翻譯詞彙
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

4.3.1.3 合併牛津詞典及譯典通字典

由上頁表 4.8 可知，不同辭典對於英文詞彙所定義的中文對應詞彙並不完全相同；因此本研究將牛津詞典和譯典通字典的中文對應詞彙合併，以增加英文詞彙的中文對應詞彙數目，如表 4.9 所示。經合併之後，本研究的「英漢合併字典」總共含有 99805 個英文詞彙。

4.3.2 近義詞典建置

有了英漢合併字典，我們希望能再擴充多一點的中文對應詞彙。本研究設想，如果以英漢合併字典的中文對應詞彙為基礎，找尋與中文對應詞彙意義相近的詞彙，也就表示這些詞彙與該英文詞彙的意義也會近似。我們選擇透過兩種途徑來增加我們的中文對應詞彙：使用中央研究院現代漢語一詞泛讀[21]及E-HowNet[6]來找尋意義相近的近義詞彙，由於這些近義詞彙是經過第二個步驟擴充的詞彙，因此我們稱之為「次擴充詞彙」。

圖 4.1 一詞泛讀系統介面

4.3.2.1 一詞泛讀

圖 4.1 為現代漢語一詞泛讀系統（簡稱為一詞泛讀）的介面，按下「近義詞 Near Synonyms」的按鈕可以看到與輸入查詢字「混亂」相關的近義詞。我們將英漢合併字典中的中文對應詞彙輸入至一詞泛讀系統，最後回收系統所傳回的近義詞群（如圖中以粗框框起的詞彙群）。如果改用「混亂狀況」這一個片語輸入一詞泛讀系統，一詞泛讀系統會提醒我們這個查詢並不是一個詞彙。這樣的回傳結果有助於近義詞典的建構，即使在英漢合併字典中我們認定的對應詞彙其實並不是真正的詞彙，但是輸入一詞泛讀系統後，我們也不會得到錯誤的近義詞而擾亂近義詞集的構成。換句話說，一詞泛讀系統所回傳的結果是品質優良的近義詞群，且對於輸入的查詢詞彙有嚴謹的過濾作用。

我們再以英文詞彙「confusion」為例，如下頁表 4.10 所示，「confusion」在我們的英漢合併字典中總共有十個中文對應詞彙，而這十個詞彙依據表格次序，第一（混亂）、三（騷動）、四（混淆）、六（慌亂）、八（惶恐）和第九個詞彙（雜

亂）都有從一詞泛讀系統得到回傳的近義詞群。我們認為這些近義詞群與「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經過一詞泛讀找到的次擴充詞彙。

表 4.10 一詞泛讀回傳結果

英文詞彙：confusion	
【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
【一詞泛讀】1.	混亂、糾紛、擾亂、紛亂、紛擾、亂糟糟、淆亂、亂哄哄、擾攘、攘攘、麻沸、絲紛棋布、亂麻麻、亂騰、搶攘、糜、糜沸蟻動、壤壤、憒憒
【一詞泛讀】3.	動亂、亂、騷擾、騷動、擾動、變亂
【一詞泛讀】4.	混淆、模糊、混為一談、歪曲、指鹿為馬、混淆是非、混淆黑白、習非成是、攪混
【一詞泛讀】6.	慌張、毛、不知所措、慌、驚慌、慌亂、手足無措、心慌、倉皇、心驚肉跳、發慌、手忙腳亂、驚惶、著慌、驚慌失措、失措、周章、失魂落魄、毛毛騰騰、毛咕、自相驚擾、周章失措、相驚伯有、茫然失措、張皇、惶遽、無所措手足、慌手慌腳、慌神兒、驚魂未定
【一詞泛讀】8.	害怕、怕、恐懼、恐怖、恐、懼怕、畏懼、畏、生怕、惶惑、提心吊膽、懼、疑懼、失色、悚然、心悸、心寒、噤若寒蟬、大驚失色、毛骨悚然、戒懼、魂飛魄散、望而卻步、膽顫心驚、畏怯、喪膽、面無人色、望而生畏、亡魂喪膽、不寒而栗、心膽俱裂、心驚膽顫、失容、生恐、忌憚、怵惕、怖、畏葸、面如土色、狼顧、脅肩累足、惕息、喪魂落魄、視為畏途、聞風喪膽、憚、震悚、魄散魂飛、懍懍、膽寒、懾、驚心掉膽、驚恐萬狀、撟舌
【一詞泛讀】9.	雜亂、亂、亂七八糟、紊亂、混雜、雜亂無章、忙亂、龐雜、間雜、橫生、雜七雜八、雜沓、駁雜、蓬亂、狼藉、亂套、紛雜、蕪雜、繚、夾七夾八、拉雜、凌雜、烏七八糟、紛披、紛綸、猥雜、亂營、亂雜、歷亂、蕪駁、錯落不齊、錯雜

```
<Word item = "和鳴">
  <WordFreq>0</WordFreq>
    <WordSense id="1">
      <English>harmonious</English>
      <Phone>ㄏㄜˊ　ㄇㄧㄥˊ</Phone>
      <PinYin>he2 ming2</PinYin>
      <SyntacticFunction>
        <POS>VA4</POS>
        <Freq>0</Freq>
      </SyntacticFunction>
      <TopLevelDefinition>{和諧:theme={聲音}}</TopLevelDefinition>
      <BottomLevelExpansion>
        {harmonious|和諧:theme={sound|聲}}
      </BottomLevelExpansion>
  </WordSense>
</Word>
```

圖 4.2 以「和鳴」一詞解釋 E-HowNet 詞彙架構

4.3.2.2 E-HowNet

除了從一詞泛讀得到次擴充詞彙，本研究也從 E-HowNet 中找尋近義詞；概念與 Budanitsky[1]相似，透過完整定義詞彙語意的架構尋找近義詞。首先介紹 E-HowNet 的結構，如圖 4.2 所示為「和鳴」一詞的內部定義。<WordFreq>代表該詞彙在中央研究院五百院詞語料庫中的詞頻統計數據，<WordSense> 則是以數字編號代表該詞彙有幾種語意，「和鳴」一詞在這裡只有一種語意，因此標記為 1。而在一個語意之下，「和鳴」可以對應到英文的「harmonious」一詞，<Phone> 及 <PinYin> 則說明了詞彙的發音方式，<POS> 標示詞彙的詞性。<TopLevelDefinition> 及 <BottomLevelExpansion> 則是本研究尋找近義詞最注重的兩個標記內容，因為這兩種標記含有定義詞彙的「義原」。「義原」就是定義

表 4.11 E-HowNet 之義原編寫情況一

類型一	<Word item = "混亂"> <TopLevelDefinition>**{chaotic|紛亂}**</TopLevelDefinition> <BottomLevelExpansion> **{chaotic|紛亂}** </BottomLevelExpansion> </WordSense>

及解釋詞彙的單位，在 E-HowNet 中以「英文|中文」的形式表示，例如上頁圖 4.2 中的「harmonious|和諧」及「sound|聲」。「和鳴」一詞的 <TopLevelDefinition> 定義了「和鳴」與「和諧」相關，而且主題是「聲音」的和諧；<BottomLevelExpansion> 則列出「和諧」的義原「harmonious|和諧」及「聲音」的義原「sound|聲」，因而可得知 <BottomLevelExpansion> 是針對 <TopLevelDefinition> 的內容作更細一步的意義拓展。

了解 E-HowNet 的架構及義原形式後，我們認為既然 E-HowNet 的每一個詞彙都有其定義義原，那麼就表示詞彙之間若具有相近的意思，則他們應該也享有相近的義原群；我們可以比對詞彙之間的義原群交集現象尋找近義詞，也就是利用英漢合併字典透過 E-HowNet 得到次擴充詞彙。我們發現在 E-HowNet 中的 <TopLevelDefinition> 及 <BottomLevelExpansion> 大略分為兩種編寫的情況，第一種類型如表 4.11 所示，以「混亂」一詞為例， <TopLevelDefinition> 即出現義原，且與 <BottomLevelExpansion> 的義原一模一樣；第二種類型則較為複雜，見下頁表 4.12，以「厚紙板」一詞為例， <TopLevelDefinition> 的敘述為「厚」的「紙板」兩個詞彙，而「厚」與「紙板」在 E-HowNet 中又有各自定義的義原；我們發現，詞彙「厚紙板」的 <BottomLevelExpansion> 即為「紙板」及「厚」兩個詞彙的 <BottomLevelExpansion> 內容聯集，也即是義原的聯集。因此，我們可以透過 <TopLevelDefinition> 及 <BottomLevelExpansion> 的義原內容來判斷詞彙之間是否為意義相近的近義詞。

表 4.12 E-HowNet 之義原編寫情況二

類型二	<Word item = "**厚紙板**"> <WordSense id="1"> <TopLevelDefinition> {紙板:qualification={厚}} </TopLevelDefinition> <BottomLevelExpansion> **{paper|紙張:telic={wrap|包紮:material={~}},** **attribute={hard|硬},qualification={thick|厚}}** </BottomLevelExpansion> </WordSense> </Word>
	<Word item = "**紙板**"> <TopLevelDefinition> {紙:telic={包裝:material={~}},attribute={硬}} </TopLevelDefinition> <BottomLevelExpansion> **{paper|紙張:telic={wrap|包紮:** **material={~}},attribute={hard|硬}}** </BottomLevelExpansion> </Word>
	<Word item = "**厚**"> <TopLevelDefinition>{thick|厚}</TopLevelDefinition> <BottomLevelExpansion>**{thick|厚}**</BottomLevelExpansion> </Word>

下頁圖 4.3 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。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各只有一種語意，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」。舉個

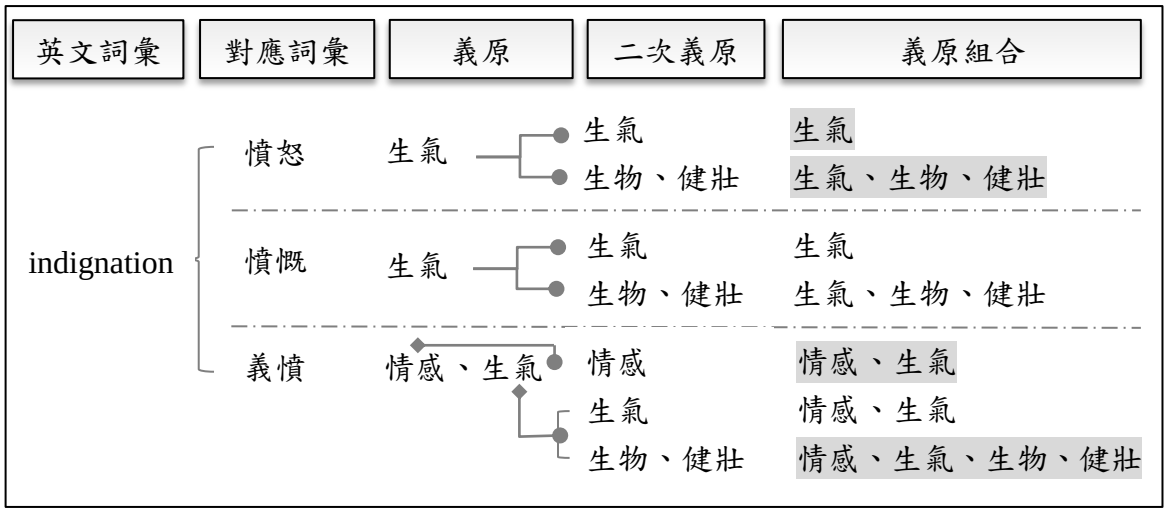


圖 4.3 E-HowNet 義原組合流程

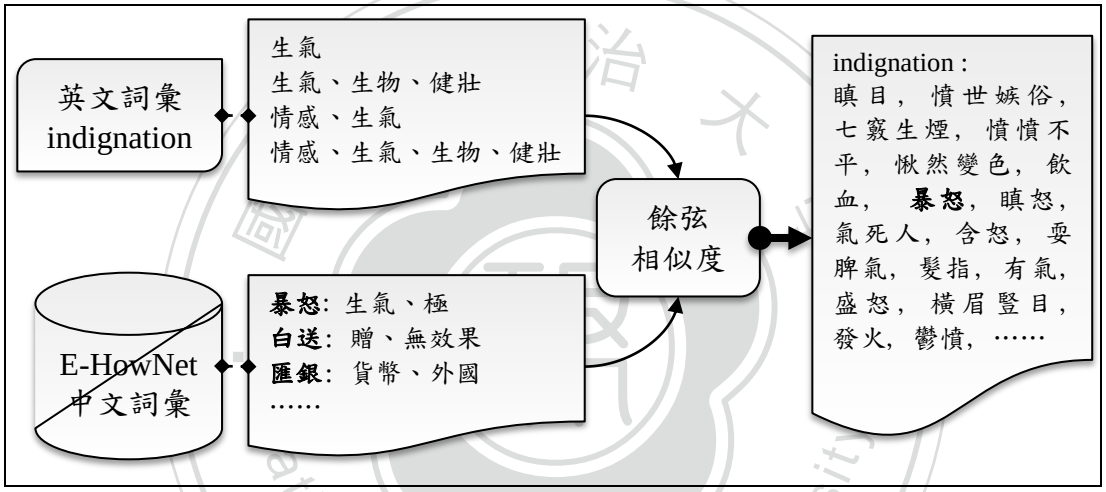


圖 4.4 使用義原組合找尋近義詞流程

例子說明二次義原，查詢「憤怒」的義原「生氣」這個詞彙，會發現它有兩種語意而有兩群義原群：第一群的義原群只有一個詞彙「生氣」，也就是自己定義自己的情形；第二群義原群則由兩個義原組成：「生物」及「健壯」。複雜一點的情況則如詞彙「義憤」，其義原群由兩個義原組成，義原「情感」是自己定義自己，「生氣」的義原則如之前描述過的由兩群義原群組成；因此「義憤」一詞有三群次義原群。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群組合起來，形成圖中的義原組合；排除重複的義原組合，就得到圖 4.3 以灰底標示的義原組合群，即為透過中文對應詞彙找到與英文詞彙意思相近的義原組合。

如上頁圖 4.4 所示，英文詞彙「indignation」有了義原組合群之後，本研究將 E-HowNet 中 88075 個中文詞彙都找出各自的義原及二次義原形成義原組合，然後從「indignation」的義原組合群逐一地把每條義原組合取出，與 E-HowNet 的 88075 個中文詞彙的義原組合作餘弦相似度 (cosine similarity) 比較，並設定門檻值為 0.7，取出相近的近義詞，成為我們從 E-HowNet 中得到的次擴充詞彙。最後，我們將從一詞泛讀系統及 E-HowNet 得到的次擴充詞彙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」。近義詞典的內容格式如表 4.13 所示，【Dictionary】標示的是英漢合併字典中的詞彙，【E-HowNet】則是取自 E-HowNet 的次擴充詞彙，【一詞泛讀】標示的次擴充詞彙則來自於一詞泛讀系統。

表 4.13 近義詞典內容格式範例

英文詞彙：indignation
【Dictionary】　憤怒，憤慨，義憤
【E-HowNet】　瞋目，憤世嫉俗，七竅生煙，憤憤不平，愀然變色，飲血，暴怒，瞋怒，氣死人，含怒，耍脾氣，髮指，有氣，盛怒，橫眉豎目，發火，鬱憤，發狠，負氣，賭氣，怒火中燒，發怒，嗔，掛火，忿，變色，嘔氣，悁，恚，怒，憤，慍，瞋，氣急攻心，火，鳥氣，氣急敗壞，憤恨不平，嗔怒，火冒千丈，戾氣，火冒三丈，慍色，憤然，滿面怒容，爆跳，惱羞成怒，狂怒，動氣，惱怒，惱恨，匿怨，怒潮，忿然，悻悻然，怒火，動火，悻悻，天怒人怨，恚怒，怒髮衝冠，息怒，動怒，怒氣，怒意，怨怒，忿忿不平，怨懟，怒斥，冒火，氣沖沖，忿鷙，氣極敗壞，憤怒，憤憤，憤慨，憤懣，懣憤，慍怒，火氣，惱火，勃然大怒，光火，憤世，怒色，無名火，滯憤，氣呼呼，氣不過，激憤，義憤，羞憤，老羞成怒，氣死，忿怒，忿忿，彆氣，氣憤憤，無明火，鬧脾氣，拂袖而去，氣走，氣頭上，怒不可遏，作怒，使性子，幽怨，幽恨，氣沖牛斗，發脾氣，震怒，止怒，怒沖沖，一肚子氣，氣昏，一肚子火，氣憤，氣惱，毆氣，氣忿，怒氣沖沖，氣咻咻，忿忿然，氣哼哼，好氣，氣嘟嘟，氣噓噓，怒氣衝天，氣憤難平，衝冠，惱，悲憤，退火，涼茶，洩忿，洩憤，一朝之忿，降火，雷霆
【一詞泛讀】　憤怒，氣，氣憤，憤，憤慨，氣惱，惱怒，惱羞成怒，激憤，憤然，惱，慍，憤激，氣沖沖，憤憤，義憤，含怒，怒沖沖，怒氣攻心，怒氣沖沖，氣乎乎，氣鼓鼓，氣囊囊，悻悻，恚，赧然

句對編號：54098		
英文動名詞組合	**對列關係**	**中文動名詞組合**
dobj(round-7, edge-10)		**dobj(清除-12, 部分-19)**
dobj(remove-15, portion-17)		dobj(使-24, 肩部-27)
		dobj(进-29, 圆滑-31)

圖 4.5 英漢動名詞組合對列範例

4.3.3 英漢動名詞組合對列流程

在第三章的部分，我們已經得到英文及中文專利文句的動名詞組合。經過統計，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。為了確保我們所使用的動名詞組合的品質，本研究使用英漢合併字典所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詞彙都有出現在字典中，我們才認定這個組合是正確的，在這個步驟也同時排除了那些含有技術名詞的動名詞組合；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了同等的檢驗，透過我們的近義詞典含有的中文詞彙過濾，最後有 249591 個組合通過檢測。為了檢視我們的近義詞典是否真的比起一般的中文字典能找到較多的中文動名詞組合，我們以 E-HowNet 收錄的中文詞彙來做測試，發現通過檢驗的動名詞組合只有 230492 個，比起近義詞典少找了 19099 個詞彙，證明我們的近義詞典確實有助於英漢文動名詞組合的對列。由於英漢專利平行文句語料庫有句對編號，我們可以透過編號得知英文及中文文句的對應關係；如圖 4.5 所示，編號第 54098 個句對中，英文句有兩個動名詞組合，中文句則有三個動名詞組合。我們的對列方式主要依賴近義詞典所提供的資訊，對列規則為：如果英文的動名詞組合，都能在各自的近義詞集中找到中文對應句中動名詞組合的動詞與名詞詞彙，才算對列成功。我們逐一地從英文句取出動名詞組合，以圖 4.5 為例，首先取出「round，edge」這一組動名詞組合，並

比對「round」及「edge」在近義詞典中所對應的近義詞集是否有出現中文句的動名詞組合，在這個例子當中「round，edge」無法找到與其對應的中文動名詞組合，因此接著取出「remove，portion」繼續比對，結果發現「remove，portion」可以透過近義詞典找到「清除，部分」這一個中文動名詞組合，如表 4.14 及下頁表 4.15 所示；「remove」從一詞泛讀的次擴充詞彙中比對到了「清除」一詞，「portion」則在英漢合併字典、E-HowNet 及一詞泛讀系統的次擴充詞彙都可以比對到「部分」一詞。完成對列的英漢文動名詞組合便以下頁表 4.16 的形式記錄，本研究對列成功的英漢動名詞組合共有 35811 組。

表 4.14 以圖 4.5 為例的對列說明：「remove」

英文詞彙：remove	
【Dictionary】	移動、搬開、調動、脫掉、去掉、消除、使離去、把...免職、撤去、殺死、殺害、移交、遷移、搬家、離開、距離、間隔、一步之差、英國學校中學校升級前被安排的班級、升級、移開、脫下、消除某物、移居、差距、間距
【E-HowNet】	動、移、動彈、躁進、運行、攜離、取走、播遷、拿開、挪開、搬移、搬給、抬走、拿走、搬、搬開、挪、搬走、挪動、挪移、搬動、搬到、提去、取去、卸下、遷往、帶走、帶出去、移置、移開、移靈、移挪、移山、領走、移交、徙、遷、遷徙、平移、移至、移樽、移位、喬遷、搬遷、搬家、遷居、寄籍、徙居、拆遷戶、相距、相隔、相間、相去、距、遠隔、移居、異質、異狀、鴻溝、分歧、另當別論、差距、此一時也彼一時也、差異
【一詞泛讀】	搬動、移動、搬、移、挪、挪動、挪移、騰挪、調動、調整、調、消除、除、破、排、解、排除、脫、免、散、消、**清除**、破除、打消、拔除、去掉、摒除、消弭、驅除、割除、摒、鏟除、紓、化除、祛、祛除、屏除、防除、屏退、弭、剪除、糞除、攘除、禳、禳解、擯除、舉行儀式驅除、殺死、結果、殺、誅、哈喇、戮、殛、殺害、殘殺、行凶、滅口、下毒手、凶殺、荼毒生靈、傷生害命、交卸、交接、交班、移交、遷移、遷、遷徙、搬遷、徙、搬家、移居、遷居、出谷遷喬、挪窩兒、喬遷、遷次、離開、去、走、撤離、背離、開走、起開、差別、別、差異、距離、差距、區別、出入、異樣、歧異、千差萬別、變異、間隔、晉升、升級、晉級、升遷、升官、鴻漸、升格、幸進、提任

表 4.15 以圖 4.5 為例的對列說明：「portion」

英文詞彙：portion	
【Dictionary】	**部分**、一份、一客、一份遺產、命運、定數、把...分成多份、分配、給...一份嫁妝、每人一份、嫁妝
【E-HowNet】	逢年過節、節日、節慶、撮、截、節、片段、些許、片斷、**部分**、部份、節錄、半政府、一部分、嫁妝費、添房、妝奩、陪嫁、嫁妝
【一詞泛讀】	**部分**、有、有的、片、有些、片段、局部、命運、數、命、運、運氣、天命、天數、天意、造化、定命、定數、氣運、氣數、大命、大數、世運、命數、運道、紫微斗數、分、分配、分發、分派、嫁妝、妝、陪嫁、陪送、陪奩

表 4.16 英漢動名詞對列格式

improve, efficiency：改善, 效率

第五章　翻譯模型公式

在本章節當中，5.1 小節介紹本研究使用的五個公式作為翻譯動詞模型；5.2 小節則使用與 5.1 小節對稱的公式作為翻譯名詞模型。5.3 小節介紹本研究如何運用公式建立模型，5.4 小節則說明翻譯模型效能的評量機制。

5.1 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。在公式的表示式中，我們以字母「E」代表英文、字母「C」代表中文，「V」代表動詞及「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，多為分析英文詞彙共現性再對應到中文翻譯的作法；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。接下來分別針對每個公式的意義進行解釋。

$$\underset{CV_i}{\operatorname{argmax}} \Pr(CV_i | EV, EN, CN) \tag{1}$$

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來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。

表 5.1 翻譯英文動詞公式於專利語料之英漢動名詞組合對應資訊

公式	英漢動名詞組合對應資訊		
公式(1)	**take**, [action:行动] : {**采取**=9}	**take**, [action : 功能] : {**执行**=1}	**take**, [action : 动作] : {采取=1, **执行**=2}
公式(2)	**take**, action : {(执行, 功能)=1; (采取, 动作)=1; (执行, 动作)=2; (**采取**, 行动)=9}		
公式(3)	**take**, action : {执行=3, **采取**=10}		
公式(4)	**take** : {产生=1, 作=1,…, 执行=5, 服用=5, 获得=5, …, 采用=42, **采取**=71}		
公式(5)	**take**, 行动 : {**采取**=9}		

如表 5.1 的公式(1)一欄所示，在專利語料中，當「take action」的「action」被翻譯成「行动」時，「take」會被翻譯為「采取」，且出現次數為 9 次；而當「action」被翻譯成「功能」時，「take」則被翻譯為「执行（出現次數為 1 次）」。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；所以當「take action」的「action」被翻譯成「动作」時，公式(1)便會優先推薦出現次數較多的「执行」。

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)可以透過這三個詞彙的資訊，在語料中觀察「take」跟「pills」一起使用且「pills」對應到「藥」時，「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)找到的中文翻譯中可以比對到「吃」這個詞彙，進而讓使用者知道使用「take pills」才是正確的用法。

公式(2)及公式(3)則為許多英漢翻譯使用的方法。公式(2)的推薦原理為：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過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。同樣以「take action」為例，見上頁表 5.1 公式(2)一欄，在語料中最常被翻譯成「采取行动（出現次數為 9 次）」，因此優先推薦「采取」為動詞的中文翻譯。

$$\underset{CV_i, CN_j}{\operatorname{argmax}} \Pr(CV_i, CN_j \mid EV, EN) \tag{2}$$

公式(3)的推薦原理為：如果看見特定英文動名詞組合 EV、EN，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 即為我們優先推薦的翻譯詞彙。以表 5.1 中「take action」為例，公式(3)一欄顯示語料中最常對應到的中文動詞為「采取（出現次數為 10 次）」。

$$\underset{CV_i}{\operatorname{argmax}} \Pr(CV_i \mid EV, EN) \tag{3}$$

公式(4)的原理為：如果看到一個特定的英文動詞 EV，則我們優先推薦的中文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。以表 5.1 中「take」為例，我們優先推薦中文動詞為「采取（出現次數為 71 次）」。

$$\underset{CV_i}{\operatorname{argmax}} \Pr(CV_i \mid EV) \tag{4}$$

而公式(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。以表 5.1 公式(5)一欄的「take, 行动」為例，語料中對應到的中文動詞為「采取」。

$$\underset{CV_i}{\operatorname{argmax}} \Pr(CV_i \mid EV, CN) \tag{5}$$

表 5.2 翻譯英文名詞公式於專利語料之英漢動名詞組合對應資訊

公式	英漢動名詞組合對應資訊	
公式(6)	[induce： 诱导], **response** : {反应=2, **应答**=12}	[induce： 引起], **response** : {**反应**=2}
公式(7)	induce, **response** : {(诱，应答)=1, (诱发，应答)=1, (引起，反应)=2, (诱导，反应)=2, (诱导，**应答**)=12}	
公式(8)	induce, **response** : {反应=4, **应答**=14}	
公式(9)	**response** : {响应=41, 反应=43, **应答**=68}	
公式(10)	引起, **response** : {应答=10, **反应**=11}	

5.2 翻譯英文名詞公式說明

與翻譯英文動詞的公式對稱，本研究提出五種公式訓練翻譯英文名詞模型。與公式(1)的翻譯原理對稱，公式(6)所考慮的資訊是最多的。公式(6)的原理為：假設如果同時看見英文的動詞、名詞及中文的動詞，在這些條件之下，出現機率最高的中文名詞 CN 即為英文名詞 EN 的中文優先推薦翻譯詞彙。如表 5.2 公式(6)一欄所示，當「induce」被翻譯為「诱导」且與「response」搭配時，我們會優先推薦「应答（出現次數為 12 次）」為「response」的對應翻譯。

$$\underset{CN_i}{\operatorname{argmax}} \Pr(CN_i | EV, EN, CV) \tag{6}$$

與公式(2)的翻譯原理對稱，公式(7)的原理為：如果看見該英文動名詞組合 EV、EN，我們的翻譯模型會從該動名詞組合對應過的中文動名詞組合中，取得出現機率最高的中文名詞當作我們的推薦翻譯詞彙。同樣以「induce response」為例，見表 5.2 公式(7)一欄，在語料中最常被翻譯成「诱导应答（出現次數為 12 次）」，因此優先推薦「应答」為名詞的中文翻譯。

$$\underset{CV_j, CN_i}{\operatorname{argmax}} \Pr(CV_j, CN_i | EV, EN) \tag{7}$$

與公式(3)的翻譯原理對稱，公式(8)的原理為：如果看見英文動名詞組合EV、EN，則該動名詞組合所對應到的中文名詞群中，出現機率最高的中文名詞CN 即為我們的推薦翻譯詞彙。如上頁表 5.2 所示，以公式(8)一欄中「induce response」為例，語料中最常對應到的中文動詞為「应答（出現次數為 14 次）」。

$$\underset{CN_i}{\operatorname{argmax}} \Pr(CN_i|EV, EN) \tag{8}$$

與公式(4)的翻譯原理對稱，公式(9)的原理為：如果看到一個特定的英文名詞 EN，則我們所推薦的中文翻譯詞彙即為與 EN 最常一起出現的中文名詞 CN。以表 5.2 公式(9)一欄的「response」為例，我們優先推薦中文名詞「应答（出現次數為 68 次）」。

$$\underset{CN_i}{\operatorname{argmax}} \Pr(CN_i|EN) \tag{9}$$

與公式(5)的翻譯原理對稱，公式(10)的原理為：假設看到一個英文名詞和英文動詞的中文翻譯，則我們推薦與這個組合最常一起出現的中文名詞做為推薦翻譯。以表 5.2 公式(10)一欄的「引起, response」為例，語料中對應到的中文名詞為「反应（出現次數為 11 次）」。

$$\underset{CV_i}{\operatorname{argmax}} \Pr(CN_i|CV, EN) \tag{10}$$

5.3 使用公式建立翻譯模型

除了評比翻譯英文動詞及名詞各自五個公式獨立運作的效果，本研究亦各別針對動詞及名詞翻譯，將公式搭配成十七種公式組合，因此動詞及名詞翻譯各有二十二種翻譯模型。我們讓公式組合的翻譯模型「協同推薦」英文的中文翻譯：組合

中的公式可以各自推薦它們認為的所有可能答案，答案順序根據答對的機率大小排列；而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・4」即為公式(1)、(2)及(4)的組合，各自推薦了一、二和五個答案；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(4)提供的五個答案中，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(4)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・4」就產生了三個候選答案。

5.4 翻譯模型評量方式

本研究將 F-measure 稍作變形，用以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「模型回答率」表示翻譯模型的作答數量，並使用答題正確率與模型回答率為新的評量參數，本研究以「***f***-measure」代表變形後的評量方式，如公式(11)所示。我們設定兩套 ***f***-measure 的係數值評量只推薦一個答案跟五個答案時翻譯模型的效果，「***f***1 score」將答題正確率和模型回答率的權重係數平分設定為 0.5，「***f***-measure, α=0.7」則設定答題正確率有較高的權重 0.7。

$$f\text{-measure} = \frac{1}{\frac{\propto}{\text{答題正確率}} + \frac{1-\propto}{\text{模型回答率}}} \tag{11}$$

第六章　使用專利文句語料建置翻譯模型

專利文句及科學人雜誌同屬科技類文章，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效果差異。專利文句語料庫[13]中，本研究對列成功的英漢動名詞組合共有 35811 組。我們利用亂數挑選的方式，將資料依據 8:2 的比例切割成訓練及測試資料。6.1 及 6.2 小節分別介紹翻譯英文動詞與名詞之相關分析，6.3 小節則為本章節的小結論。

6.1 翻譯英文動詞

本小節於 6.1.1、6.1.2 及 6.1.3 小節分別描述本研究的模型翻譯前一百名英文高頻動詞、前二十二名具競爭力候選人動詞及前十二和前六名具競爭力候選人動詞之相關表現評比。

6.1.1　前一百名英文高頻動詞分析

本研究對於頻繁出現的動詞有興趣，因此探究了前一百名在 35811 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，下頁表 6.1 即為前一百名動詞及其出現次

數列表。這一百個英文動詞總共出現於 30376 筆資料之中。本研究將資料切割成訓練及測試資料，訓練資料共有 24300 筆，測試資料則有 6076 筆。

圖 6.1 為翻譯模型在協同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。

表 6.1 專利前一百名英文高頻動詞及其出現次數

have=4530	cause=273	add=138	define=90	promote=60
provide=3345	require=266	limit=135	express=88	suppress=59
use=1993	control=265	employ=135	constitute=84	release=58
include=1954	inhibit=256	retain=131	overcome=82	extend=58
comprise=1588	carry=241	cover=127	supply=81	exert=58
contain=1080	prevent=237	affect=123	modify=81	stimulate=56
form=914	treat=233	enter=122	disclose=81	transmit=54
receive=863	generate=231	reach=119	satisfy=78	explain=54
reduce=774	utilize=212	eliminate=118	induce=77	exceed=54
perform=616	take=210	offer=114	depict=76	replace=53
increase=465	create=201	make=114	calculate=76	update=51
produce=453	support=193	meet=113	reflect=71	connect=51
maintain=397	select=190	identify=112	give=70	permit=50
determine=382	illustrate=185	decrease=112	alter=70	ensure=50
represent=373	implement=180	establish=105	execute=69	encode=50
show=352	enhance=178	exhibit=103	adjust=69	apply=50
obtain=329	avoid=163	complete=100	accept=65	facilitate=49
achieve=329	describe=159	process=96	hold=64	vary=48
improve=322	change=156	possess=96	leave=61	keep=48
allow=287	display=147	find=90	yield=60	lack=47

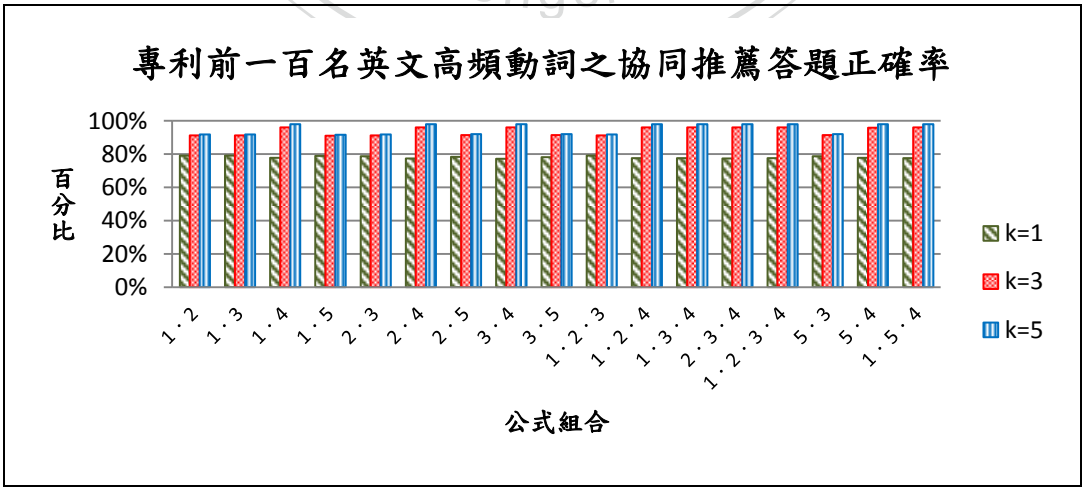


圖 6.1 專利前 100 名英文高頻動詞之協同推薦答題正確率

我們可以看到當推薦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協同推薦三個答案時，其中幾乎都包含了正確解答。圖 6.2 為公式獨立運作時的模型翻譯表現。

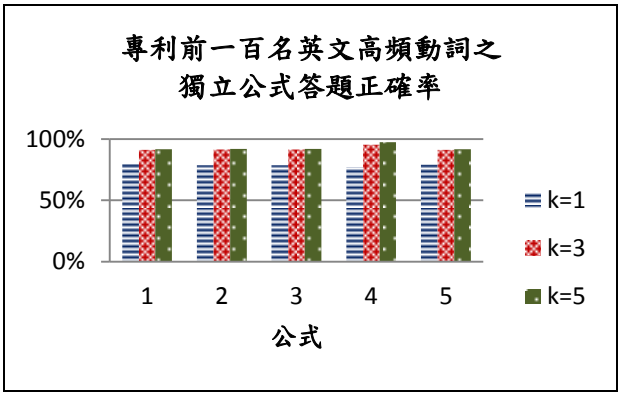


圖 6.2 專利前 100 名動詞公式答題正確率

下頁圖 6.3 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)都會因為測試資料中出現訓練資料所沒有的紀錄而無法作答，有回答率的問題，而如下頁圖 6.4 的答題拒絕率所示，公式(4)可以回答任何問題，只有答對與答錯的狀況，只要訓練語料中出現過的英文動詞都有與其對照的中文翻譯，因而拒絕率為零。雖然公式(1)、(2)、(3)及(5)在圖 6.3 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。

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，因為協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)搭配組合推薦答案中的排名，如圖 6.5 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；這可以證明雖然從圖 6.3 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。

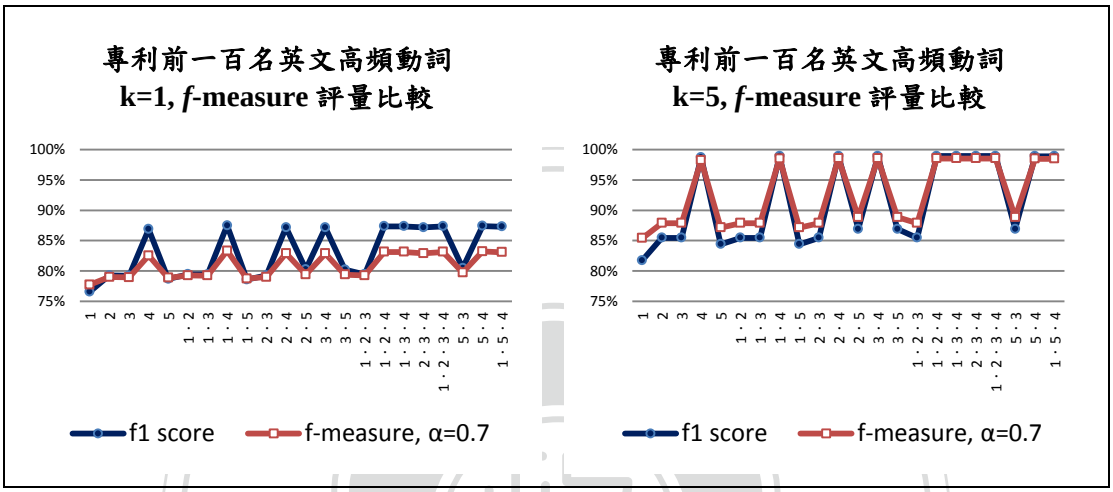


圖 6.3 翻譯模型在專利前 100 名動詞推薦一個及五個答案之 *f*-measure 成效

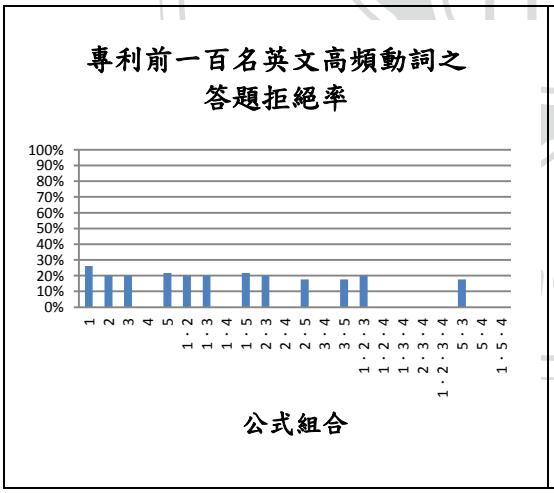


圖 6.4 專利前 100 名動詞答題拒絕率

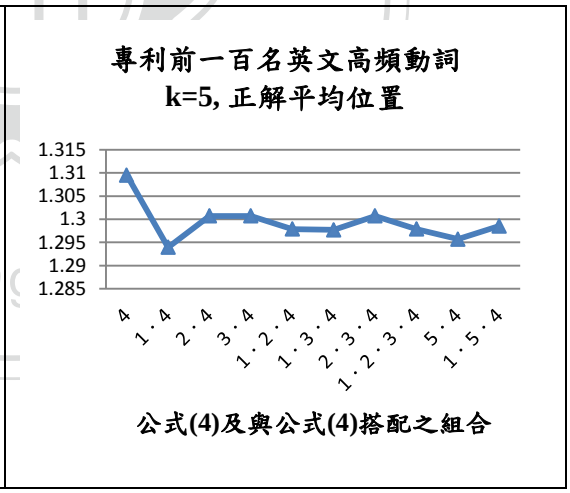


圖 6.5 正解位置於公式(4)組合比較

6.1.2 前二十二名具競爭力候選人之動詞分析

本研究由前一百名英文高頻動詞中選出一些動詞，這些動詞的特性為它們都不只對應到一個中文翻譯詞彙，而且出現次數最高的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其在語料中與 EV 一起出現的次數由少至多排列有 CV_1、CV_2 及 CV_3，則 CV_3 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。表 6.2 為前二十二名具競爭力候選人動詞列表及其對應的中文翻譯和對應次數。

表 6.2 專利前二十二名具競爭力候選人之動詞

carry={保持=1, 单元=1, 支撑=1, 运=1, 运载=1, 带=2, 搬运=2, 运送=2, 实行=3, 包含=5, 完成=7, 带有=8, 实现=30, 携带=33, 实施=37, 进行=38, 执行=69}
create={生=1, 制作=2, 制造=2, 引起=2, 有=2, 制备=4, 创造=8, 创建=72, 产生=108}
obtain={存在=1, 得出=1, 有=1, 取得=2, 获取=18, 得到=111, 获得=195}
retain={维持=3, 保留=58, 保持=70}
explain={阐明=1, 发明=2, 说明=17, 解释=34}
give={作=1, 发生=1, 带=1, 带来=1, 献出=1, 设置=1, 面向=1, 出=2, 给=6, 有=11, 赋予=11, 给予=13, 产生=20}
leave={离=1, 离开=27, 留下=33}
adjust={调=1, 调整=25, 调节=43}
employ={运用=1, 应用=3, 利用=6, 用=7, 使用=49, 采用=69}
represent={体现=1, 去=1, 视为=1, 表=2, 表现=2, 述=2, 代表=165, 表示=199}
exhibit={表=1, 展示=8, 表现=45, 显示=49}
replace={替换=4, 替代=6, 取代=16, 代替=27}
apply={使得=1, 利用=1, 施用=1, 运=1, 使=2, 实施=2, 采用=7, 使用=13, 应用=22}
lack={没有=15, 缺乏=15, 缺少=17}
reduce={减低=1, 削减=1, 压缩=1, 缩小=7, 降=9, 减=13, 减小=124, 减少=273, 降低=345}
make={到=1, 当作=1, 成=1, 打=1, 取得=2, 获得=2, 得到=3, 成为=3, 组成=3, 产生=6, 有=6, 制作=8, 制备=8, 作=9, 使得=9, 做=9, 构成=9, 做出=11, 制造=11, 形成=11}
achieve={上=1, 及=1, 获取=1, 达成=1, 达=2, 取得=3, 完成=3, 得到=20, 达到=77, 获得=92, 实现=128}
improve={使=1, 使用=1, 加强=1, 增=1, 增进=1, 采用=1, 增加=2, 改良=2, 增强=6, 改进=52, 提高=109, 改善=145}
add={增添=1, 加上=2, 加入=43, 添加=44, 增加=48}
enhance={增进=2, 加强=10, 增加=13, 提高=68, 增强=85}
reach={到=1, 抵达=1, 获得=1, 达成=1, 达=3, 达到=46, 到达=66}
take={产生=1, 作=1, 受到=1, 可=1, 吞服=1, 实施=1, 得出=1, 成=1, 拿走=1, 接受=1, 提取=1, 用=1, 要求=1, 需=1, 占=2, 发生=2, 可以=2, 获取=2, 使=3, 取得=3, 使用=4, 拍摄=4, 占据=5, 占用=5, 执行=5, 服用=5, 获得=5, 取出=6, 需要=6, 利用=7, 取=7, 花费=10, 采用=42, 采取=71}

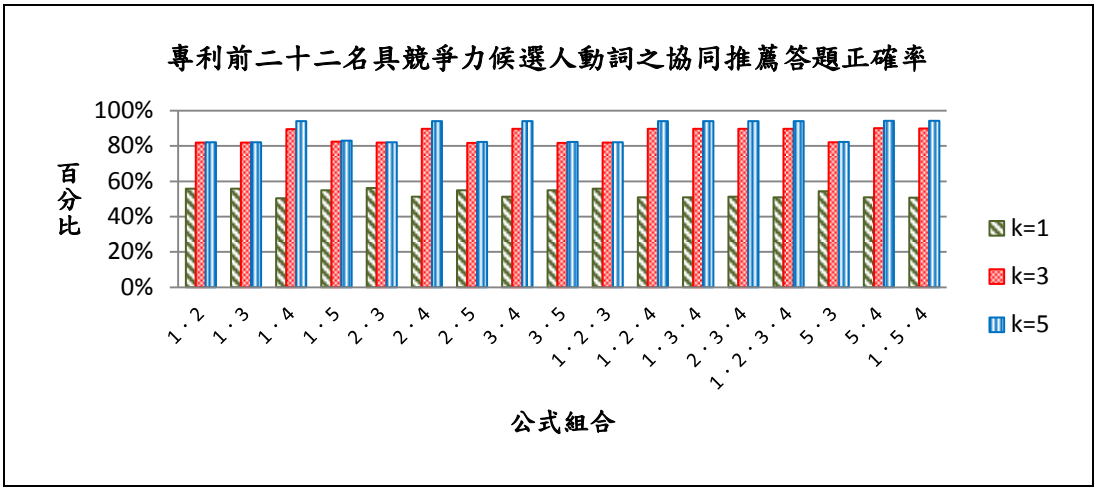


圖 6.6 專利前 22 名具競爭力候選人動詞之協同推薦答題正確率

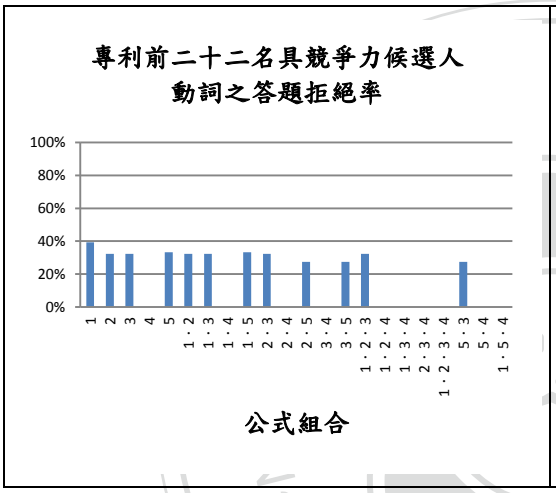


圖 6.7 專利前22名動詞之答題拒絕率

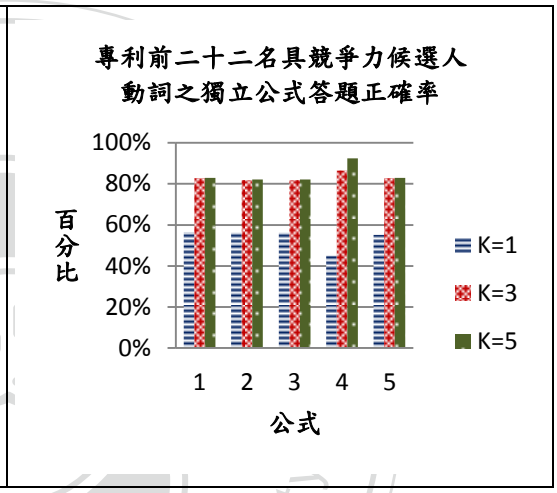


圖 6.8 專利前22名動詞之公式正確率

從圖 6.6 可以看到翻譯模型的表現與翻譯前一百名高頻動詞的趨勢大體相同，但是當推薦五個答案時，前二十二名具競爭力候選人動詞的平均答題正確率為 88%，低於前一百名高頻動詞的 95%；這是由於翻譯難度增高，以及訓練資料銳減的緣故，由圖 6.7 可看出答題拒絕率攀升，最高可達 40%的拒絕率。圖 6.8 為公式獨立運作時的翻譯表現。圖 6.9 為正確答案於與公式(4)搭配組合模型的平均答案位置，可以看到協同推薦的平均位置較為前面，具正面推薦效果。

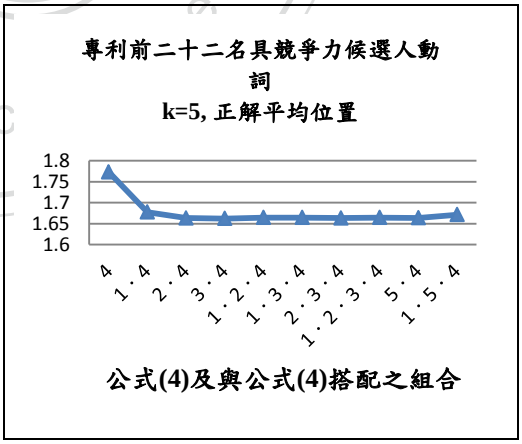


圖 6.9 正解位置於公式(4)組合比較

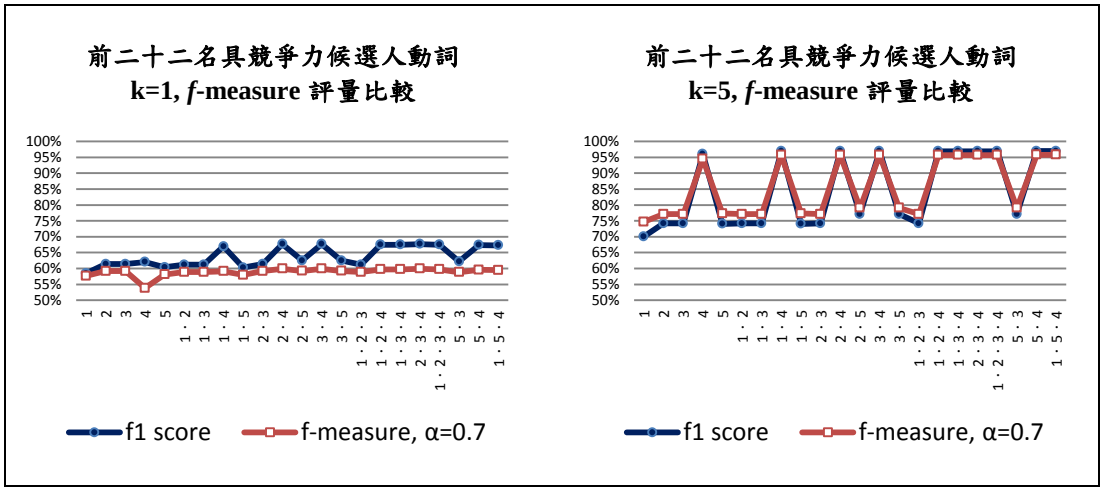


圖 6.10 翻譯模型在專利前 22 名動詞推薦一個及五個答案之 ***f***-measure 成效

由圖 6.10 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，翻譯模型只能推薦一個答案時，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下降，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。這是由於這二十二個動詞的翻譯活性較大，不容易透過統計猜出答案。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。

6.1.3 前十二及前六名具競爭力候選人之動詞分析

我們嚴選了翻譯競爭力更激烈的前十二個動詞以觀察翻譯模型的表現，競爭力在這裡的門檻調升為：第一名候選人出現的次數與第二名候選人的出現次數比值不得超過 1.42 倍。這十二個動詞總共出現於 2705 筆資料之中，訓練資料有 2164 筆，測試資料則有 541 筆。而為了觀察地更細微，我們再從前十二名具競爭力候選人之動詞中取出排名在前半段的六個動詞；競爭力在這裡的門檻調升為：第一

名候選人出現的次數與第二名候選人的出現次數比值不得超過 1.22 倍。前六名動詞總共出現於 906 筆資料之中，訓練資料有 724 筆，測試資料則有 182 筆。表 6.3 及表 6.4 分別為前十二名及前六名英文動詞於資料中對應到的翻譯詞彙及出現次數。

表 6.3 前十二名具競爭力候選人之動詞

retain={维持=3, 保留=58, 保持=70}
leave={离=1, 离开=27, 留下=33}
employ={运用=1, 应用=3, 利用=6, 用=7, 使用=49, 采用=69}
represent={体现=1, 去=1, 视为=1, 表=2, 表现=2, 述=2, 代表=165, 表示=199}
exhibit={表=1, 展示=8, 表现=45, 显示=49}
lack={没有=15, 缺乏=15, 缺少=17}
reduce={减低=1, 削减=1, 压缩=1, 缩小=7, 降=9, 减=13, 减小=124, 减少=273, 降低=345}
make={到=1, 当作=1, 成=1, 打=1, 取得=2, 获得=2, 得到=3, 成为=3, 组成=3, 产生=6, 有=6, 制作=8, 制备=8, 作=9, 使得=9, 做=9, 构成=9, 做出=11, 制造=11, 形成=11}
achieve={上=1, 及=1, 获取=1, 达成=1, 达=2, 取得=3, 完成=3, 得到=20, 达到=77, 获得=92, 实现=128}
improve={使=1, 使用=1, 加强=1, 增=1, 增进=1, 采用=1, 增加=2, 改良=2, 增强=6, 改进=52, 提高=109, 改善=145}
add={增添=1, 加上=2, 加入=43, 添加=44, 增加=48}
enhance={增进=2, 加强=10, 增加=13, 提高=68, 增强=85}

表 6.4 前六名具競爭力候選人之動詞

retain={维持=3, 保留=58, 保持=70}
represent={体现=1, 去=1, 视为=1, 表=2, 表现=2, 述=2, 代表=165, 表示=199}
exhibit={表=1, 展示=8, 表现=45, 显示=49}
lack={没有=15, 缺乏=15, 缺少=17}
make={到=1, 当作=1, 成=1, 打=1, 取得=2, 获得=2, 得到=3, 成为=3, 组成=3, 产生=6, 有=6, 制作=8, 制备=8, 作=9, 使得=9, 做=9, 构成=9, 做出=11, 制造=11, 形成=11}
add={增添=1, 加上=2, 加入=43, 添加=44, 增加=48}

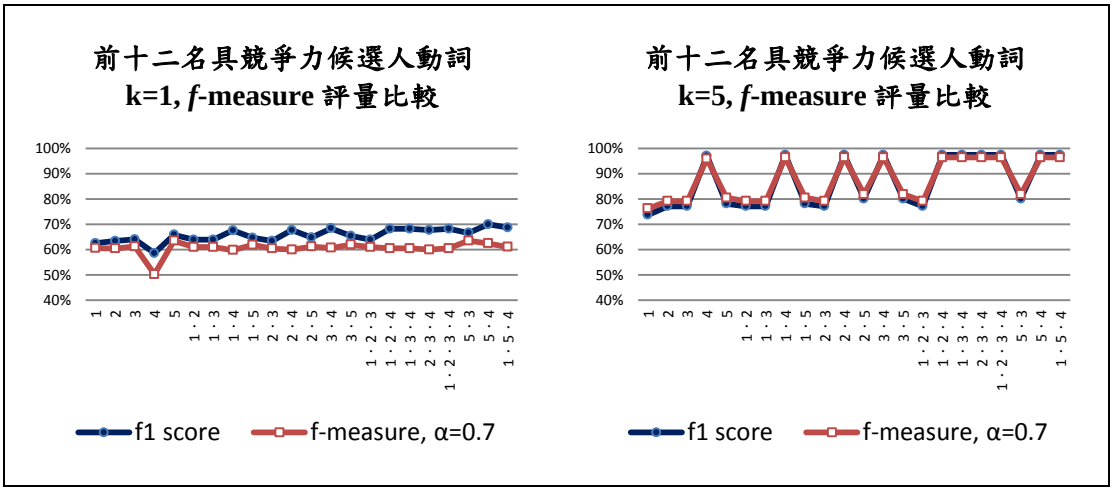


圖 6.11 翻譯模型在專利前 12 名動詞推薦一個及五個答案之 *f*-measure 成效

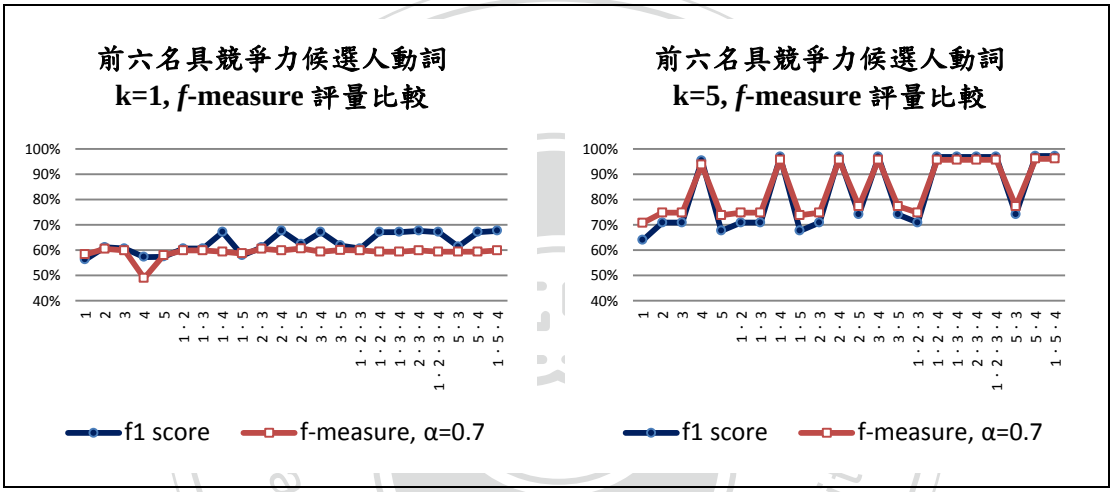


圖 6.12 翻譯模型在專利前 6 名動詞推薦一個及五個答案時之 *f*-measure 成效

圖 6.11 及圖 6.12 分別為翻譯模型針對前十二名及前六名具競爭力候選人動詞的表現評比。我們可以發現翻譯模型的表現趨勢在兩組資料中與前二十二名具競爭力候選人動詞相似，在推薦五個答案時協同推薦的翻譯模型有較好表現。

6.2 翻譯英文名詞

與 6.1 小節對稱，6.2.1 小節分析翻譯模型在翻譯前一百名英文高頻名詞的表現，6.2.2 小節分析前十九名具有競爭力候選人之名詞，6.2.3 小節分析前十名及前五名具有競爭力候選人之名詞。

表 6.5 前一百名英文高頻名詞及其出現次數

method=982	position=222	state=140	instruction=106
atom=777	condition=212	image=139	capacity=106
datum=503	invention=211	location=138	program=103
effect=501	result=203	power=137	procedure=103
function=493	problem=198	polymer=136	amount=101
signal=476	user=197	functionality=133	mode=100
portion=434	service=193	layer=131	solution=99
system=404	form=192	unit=130	pressure=99
step=400	ability=192	component=129	command=99
activity=394	advantage=188	communication=129	water=97
operation=370	technique=186	interface=127	size=97
structure=362	shape=185	gene=125	section=97
material=344	performance=185	element=124	particle=96
product=335	use=182	cost=123	moiety=94
surface=315	combination=175	cell=120	temperature=92
message=311	thickness=173	efficiency=118	movement=90
device=304	apparatus=169	column=114	growth=90
content=299	area=167	time=113	application=90
property=284	requirement=160	request=113	support=89
mixture=270	molecule=153	protocol=113	meaning=89
process=262	response=152	pattern=113	modification=87
part=260	capability=152	disease=113	energy=86
level=253	address=149	action=111	strength=85
number=248	stability=146	type=108	range=85
sequence=244	need=141	parameter=108	quality=83

6.2.1 前一百名英文高頻名詞分析

在先前的章節我們觀察了模型翻譯英文動詞的表現情形，為了保持研究的平衡，我們平行探究前一百名出現次數最多的英文名詞，這些名詞在我們 35811 筆資料中至少都出現過 83 次以上，最多的出現次數則為 982 次，表 6.5 為我們所使用的前一百名名詞及其出現次數。而這一百個名詞總共出現於 19756 筆資料之中，訓練資料有 15804 筆，測試資料則有 3952 筆。

如下頁圖 6.13 所示，翻譯模型推薦三個答案跟五個答案時表現相似，可見翻譯模型推薦三個答案時幾乎都包含了正確解答。下頁圖 6.14 為公式獨立運作時的模型翻譯表現。

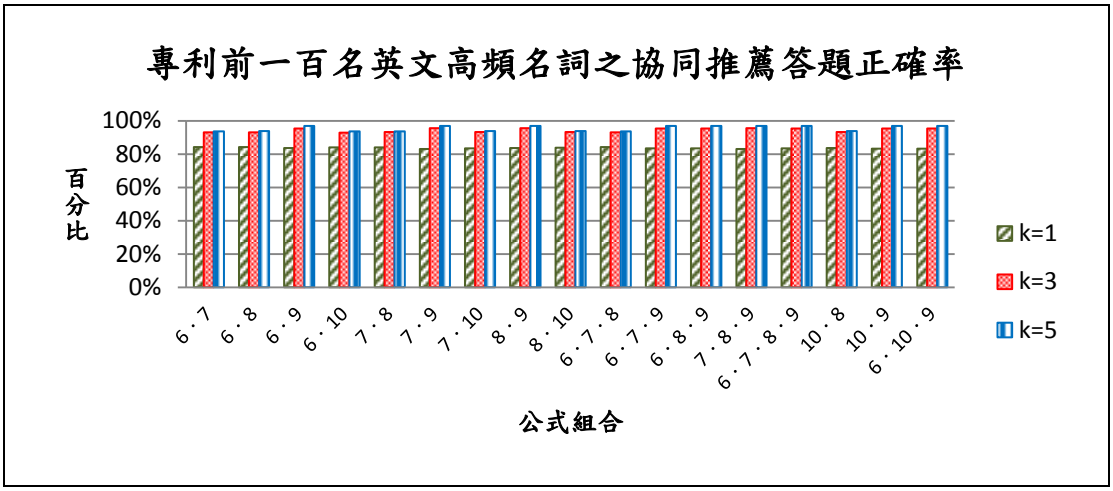


圖 6.13 專利前 100 名英文高頻名詞之協同推薦答題正確率

我們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。如下頁圖 6.15 所示，與公式(9)搭配的公式組合中正確解答的平均位置皆比在公式(9)的平均位置還要前面，特別是公式(6)效果特別明顯。如下頁圖 6.16 的答題拒絕率所示，翻譯前一百名名詞的拒絕率比起翻譯前一百名動詞相差不多。下頁圖 6.17 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名詞的效果比較。當翻譯模型只能推薦一個答案時(k=1)，公式(9)和那些與公式(9)搭配的公式組合在 *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 分數則往下降，其他沒有與公式(9)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(6)、(7)、(8)及(10)有無法回答的情形導致回答率得分較公式(9)低。雖然公式(6)、(7)、(8)及(10)在圖 6.17 只能推薦一個答案時的表現略差，但在兩種評分機制中都維持一樣的水準；相較之下公式(9)在精確率的表現較薄弱。翻譯

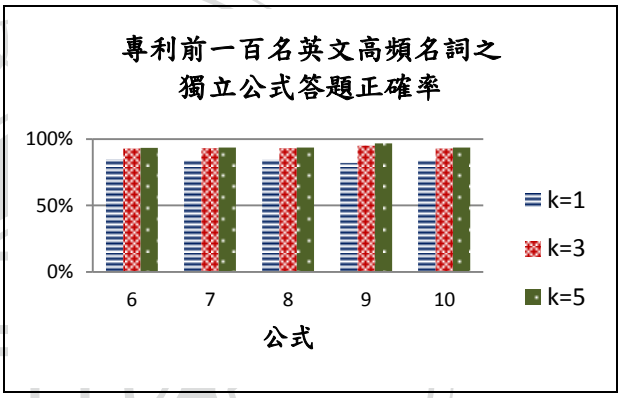


圖 6.14 專利前 100 名名詞公式答題正確率

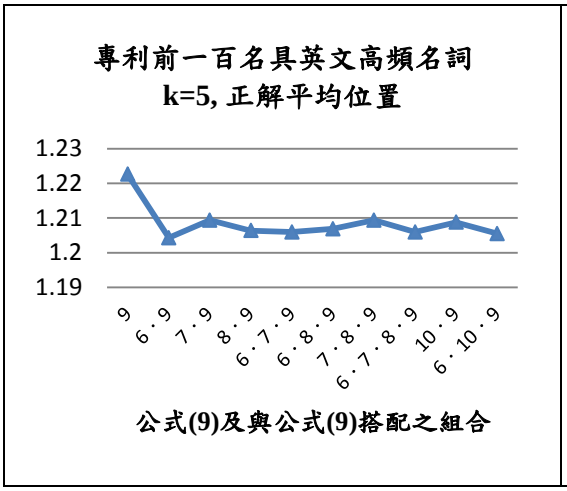


圖 6.15 正解位置於公式(9)組合比較

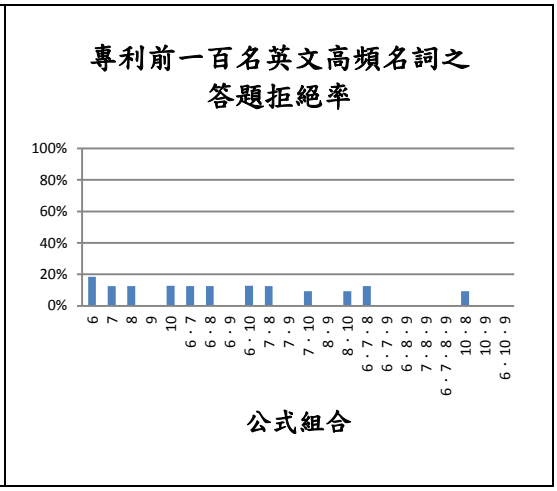


圖 6.16 專利前 100 名詞答題拒絕率

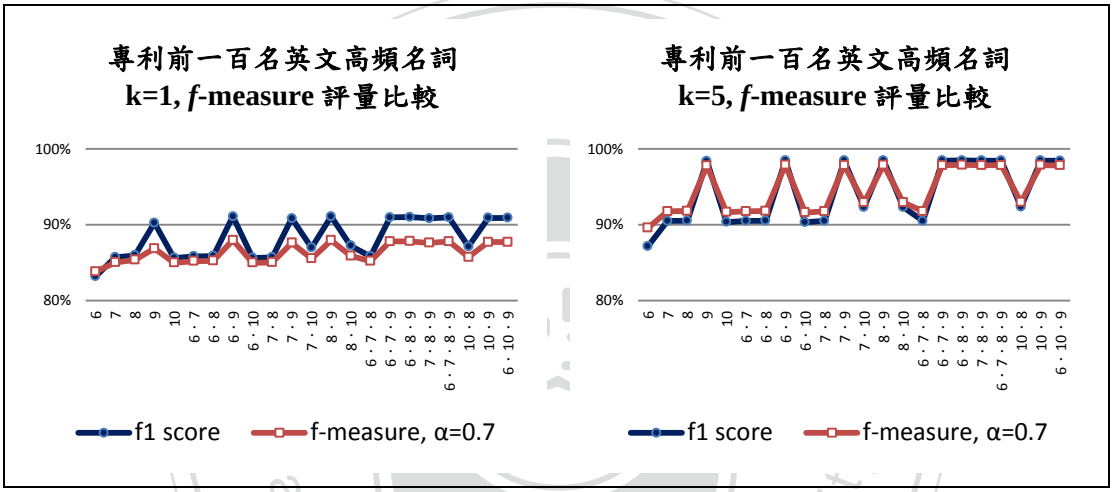


圖 6.17 翻譯模型在專利前 100 名名詞推薦一個及五個答案時之 *f*-measure 成效

模型能推薦五個答案 (k=5) 時，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(9)搭配的公式組合。

6.2.2 前十九名具競爭力候選人之名詞分析

本研究從前一百名名詞中挑選了不只對應到一個中文翻譯的英文名詞，而且這些英文名詞前兩名的中文翻譯是具有競爭力的；本研究在這裡定義了「競爭力」：第一名候選人出現的次數與第二名候選人的出現次數的比值不得超過兩倍。這十九個名詞總共出現於 3447 筆資料之中，訓練資料有 2757 筆，測試資料則有 690

筆。表 6.6 為十九名名詞於資料中對應到的中文翻譯詞彙和出現次數。從下頁圖 6.18 可以看到翻譯模型的表現趨勢與翻譯前一百名高頻名詞的趨勢大體相同，但由於翻譯難度增高以及訓練資料銳減的緣故，推薦五個答案時，比起前一百名高頻名詞的答題正確率還要低。下頁圖 6.19 為公式獨立運作時的翻譯表現。

表 6.6 前十九名具有競爭力候選人之名詞

movement={变化=1, 导向=1, 情况=1, 装置=1, 移动=40, 运动=46}
apparatus={仪器=2, 装置=80, 设备=87}
support={柱=1, 支架=2, 支承=13, 支持=35, 支撑=38}
number={下面=1, 功率=1, 参数=1, 属性=1, 志=1, 次序=1, 温度=1, 物体=1, 物质=1, 特色=1, 系统=1, 计算=1, 质量=1, 重量=1, 顺序=1, 个数=2, 元素=2, 序列=2, 成分=2, 技术=2, 时间=2, 特性=2, 状态=2, 功能=3, 号=3, 形式=3, 形状=3, 性质=3, 总数=4, 编号=4, 序号=5, 数值=5, 数字=8, 量=11, 数=12, 号码=31, 数目=50, 数量=72}
content={参数=1, 形式=1, 成分=1, 状态=1, 级别=1, 内容=117, 含量=177}
use={优点=1, 功能=1, 效率=1, 用处=1, 运用=1, 作用=2, 利用=2, 技术=2, 应用=33, 使用=56, 用途=82}
modification={改动=3, 改进=8, 改变=10, 变化=17, 修饰=21, 修改=28}
capacity={产量=1, 能力=39, 容量=66}
procedure={序=1, 条件=1, 常规=2, 步骤=18, 程序=27, 方法=54}
meaning={意思=1, 含意=2, 意义=29, 含义=57}
process={变化=1, 法=1, 加工=2, 步骤=2, 工序=3, 程序=5, 进程=12, 处理=27, 过程=95, 方法=114}
amount={多=1, 效应=1, 数=1, 数目=1, 数额=1, 种=1, 种类=1, 系统=1, 质=1, 元素=2, 环境=2, 形式=4, 成分=4, 数量=29, 量=51}
action={功能=1, 活性=1, 活动=2, 行为=6, 行动=18, 作用=39, 动作=44}
property={效果=1, 性=2, 属性=14, 特性=79, 性质=86, 性能=102}
program={界面=1, 计划=1, 软件=1, 日程=2, 方法=3, 节目=33, 程序=62}
response={响应=41, 反应=43, 应答=68}
area={方向=1, 部位=1, 域=6, 表面积=7, 区=10, 领域=10, 面积=45, 区域=87}
effect={功能=1, 原因=1, 反应=1, 意义=1, 效力=1, 方法=1, 果=1, 深度=1, 效能=2, 结果=5, 效应=40, 影响=96, 作用=175, 效果=175}
device={图案=1, 性能=1, 条件=1, 源=1, 装备=1, 器=2, 器械=2, 手段=2, 程序=2, 缘故=2, 方法=17, 设备=106, 装置=166}

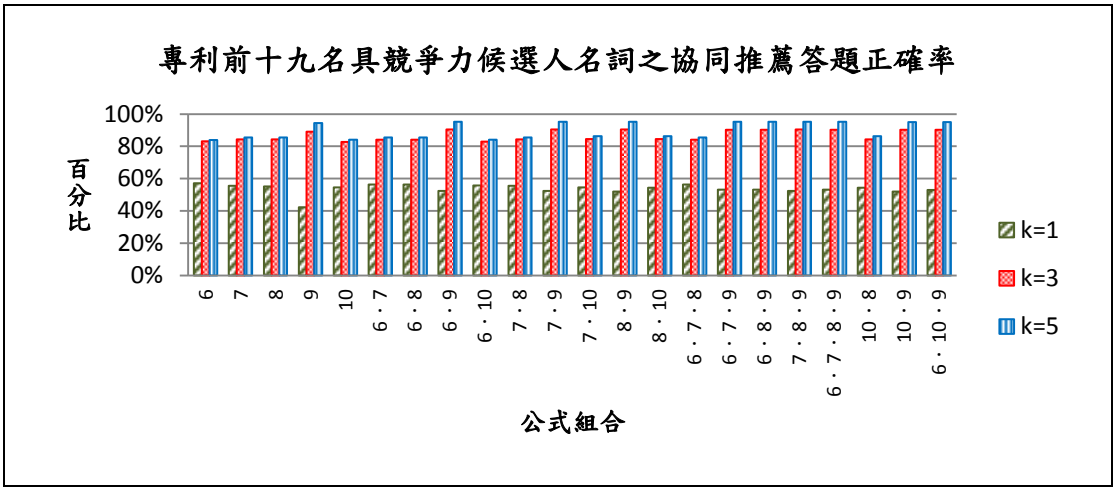


圖 6.18 專利前 19 名具競爭力候選人名詞之協同推薦答題正確率

下頁圖 6.20 為前十九名具競爭力候選人名詞之答題拒絕率，與前一百名高頻名詞的趨勢幾無相差。下頁圖 6.21 為正確答案於與公式(9)搭配組合模型的平均答案位置，可以看到協同推薦的平均位置較為前面，具正面推薦效果。

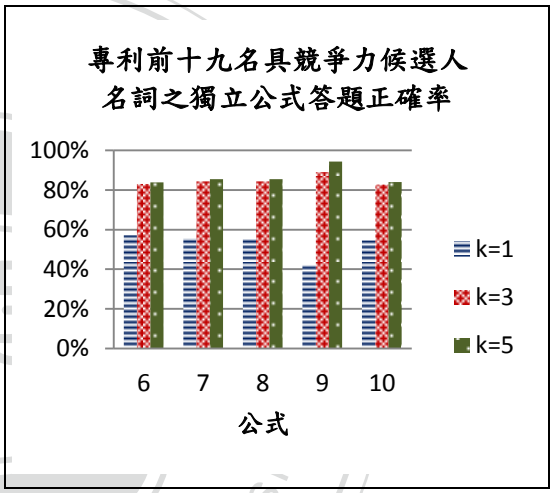


圖 6.19 專利前 19 名名詞答題正確率

由下頁圖 6.22 所示，當翻譯模型只能推薦一個答案時，前十九名具競爭力候選人的名詞與前一百名高頻名詞的 *f*-measure 趨勢並不完全相同。*f*1 score 除了公式(9)表現最不佳，其他的翻譯模型效果幾乎沒有太大的差異，在著重於精確率的 *f*-measure, α =0.7 也是如此，且每個翻譯模型在 *f*1 score 及 *f*-measure, α =0.7 兩套衡量標準之下的差距都相同。而在翻譯模型最多能推薦五個答案的情形下，*f*-measure 的趨勢走向則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(9)搭配的公式組合。

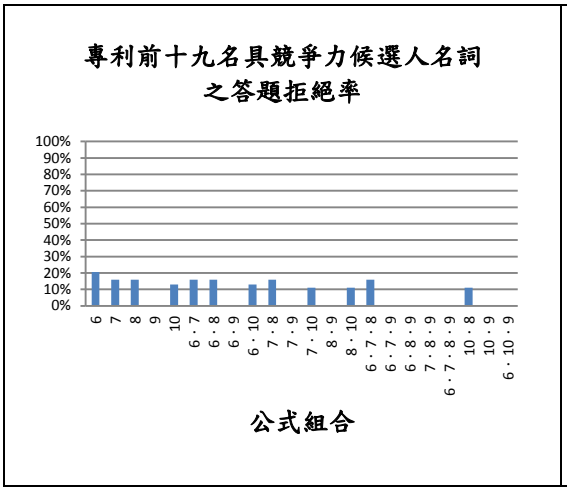


圖 6.20 專利前 19 名名詞答題拒絕率

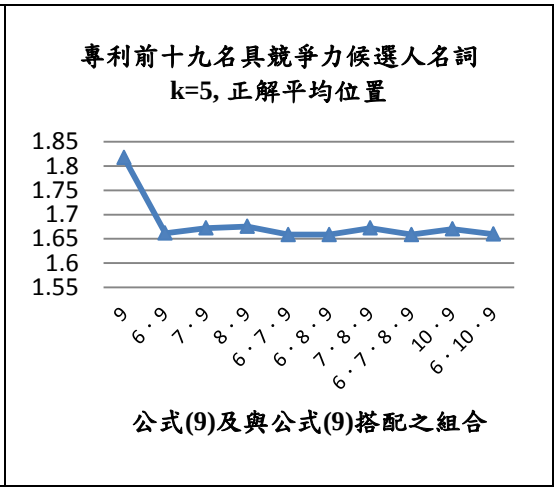


圖 6.21 正解位置於公式(9)組合比較

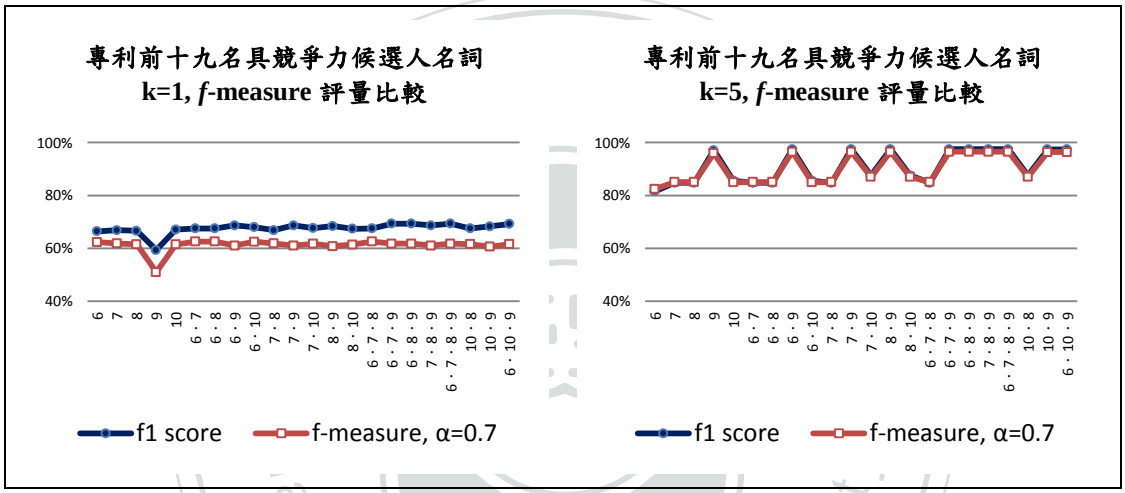


圖 6.22 翻譯模型在專利前 19 名名詞推薦一個及五個答案時之 *f*-measure 成效

6.2.3 前十及前五名具競爭力候選人之名詞分析

我們嚴選前十九名具競爭力候選人名詞中排名前半名次的十個名詞；競爭力在這裡的門檻調升為：第一名候選人出現的次數與第二名候選人的出現次數比值不得超過 1.5 倍。這前十名名詞總共出現於 2023 筆資料之中，訓練資料有 1618 筆，測試資料則有 405 筆。我們再從前十名具競爭力名詞中選出翻譯更具競爭力的五個名詞；競爭力在這裡的門檻調升為：第一名候選人出現的次數與第二名候選人的出現次數比值不得超過 1.15 倍。前五名名詞總共出現於 960 筆資料之中，訓

練資料有 768 筆，測試資料則有 192 筆。表 6.7 及表 6.8 各為前十名及前五名具競爭力之名詞及其翻譯列表。

下頁圖 6.23 及圖 6.24 分別為翻譯模型針對前十名及前五名具競爭力候選人名詞的表現評比。我們可以發現翻譯模型的表現趨勢在兩組資料中與前十九名具競爭力候選人之名詞的效果相似，推薦五個答案時，公式組合協同推薦的翻譯模型有較好的表現。

表 6.7 前十名具有競爭力候選人之名詞

movement={变化=1, 导向=1, 情况=1, 装置=1, 移动=40, 运动=46}
apparatus={仪器=2, 装置=80, 设备=87}
support={柱=1, 支架=2, 支承=13, 支持=35, 支撑=38}
number={下面=1, 功率=1, 参数=1, 属性=1, 志=1, 次序=1, 温度=1, 物体=1, 物质=1, 特色=1, 系统=1, 计算=1, 质量=1, 重量=1, 顺序=1, 个数=2, 元素=2, 序列=2, 成分=2, 技术=2, 时间=2, 特性=2, 状态=2, 功能=3, 号=3, 形式=3, 形状=3, 性质=3, 总数=4, 编号=4, 序号=5, 数值=5, 数字=8, 量=11, 数=12, 号码=31, 数目=50, 数量=72}
use={优点=1, 功能=1, 效率=1, 用处=1, 运用=1, 作用=2, 利用=2, 技术=2, 应用=33, 使用=56, 用途=82}
modification={改动=3, 改进=8, 改变=10, 变化=17, 修饰=21, 修改=28}
process={变化=1, 法=1, 加工=2, 步骤=2, 工序=3, 程序=5, 进程=12, 处理=27, 过程=95, 方法=114}
action={功能=1, 活性=1, 活动=2, 行为=6, 行动=18, 作用=39, 动作=44}
property={效果=1, 性=2, 属性=14, 特性=79, 性质=86, 性能=102}
effect={功能=1, 原因=1, 反应=1, 意义=1, 效力=1, 方法=1, 果=1, 深度=1, 效能=2, 结果=5, 效应=40, 影响=96, 作用=175, 效果=175}

表 6.8 前五名具有競爭力候選人之名詞

movement={变化=1, 导向=1, 情况=1, 装置=1, 移动=40, 运动=46}
apparatus={仪器=2, 装置=80, 设备=87}
support={柱=1, 支架=2, 支承=13, 支持=35, 支撑=38}
action={功能=1, 活性=1, 活动=2, 行为=6, 行动=18, 作用=39, 动作=44}
effect={功能=1, 原因=1, 反应=1, 意义=1, 效力=1, 方法=1, 果=1, 深度=1, 效能=2, 结果=5, 效应=40, 影响=96, 作用=175, 效果=175}

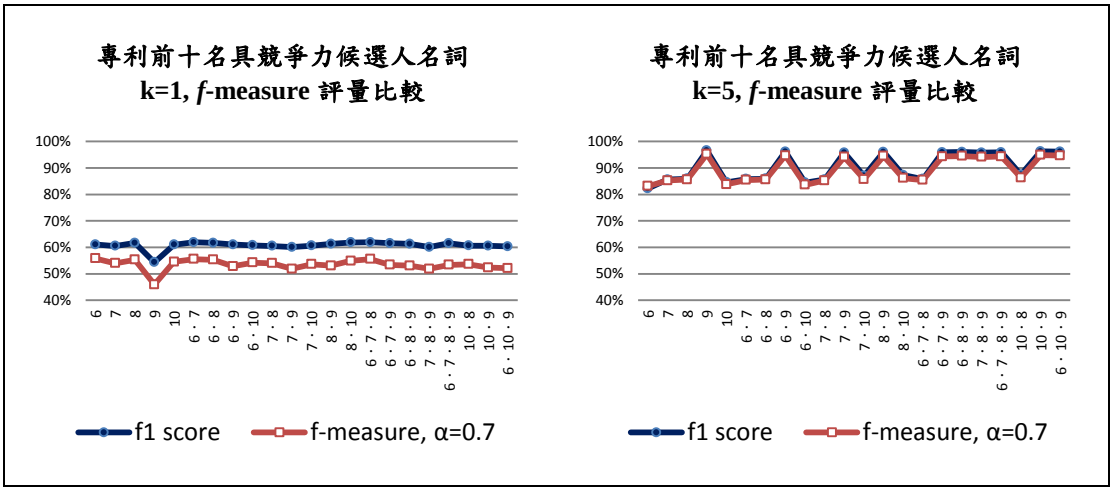


圖 6.23 翻譯模型在專利前 10 名名詞推薦一個及五個答案時之 *f*-measure 成效

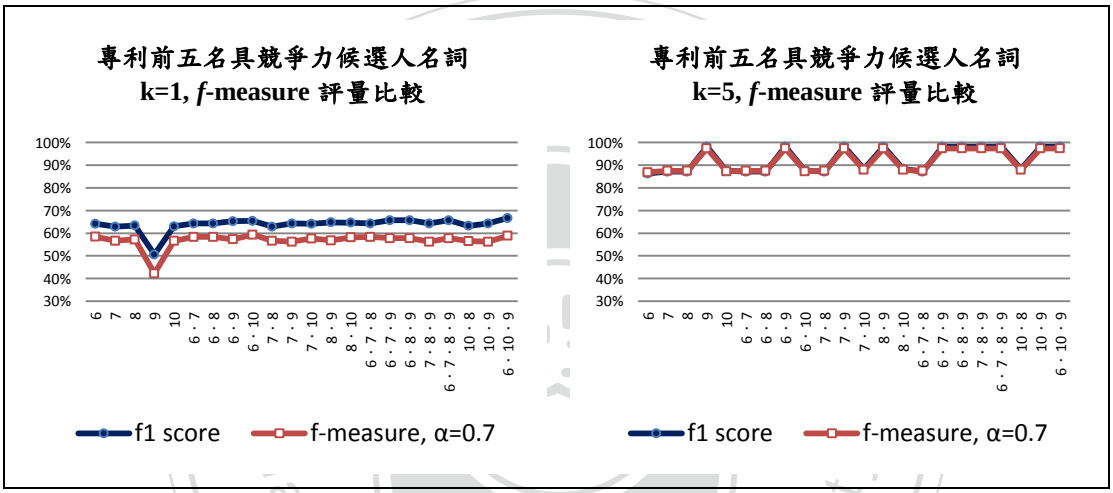


圖 6.24 翻譯模型在專利前 5 名名詞推薦一個及五個答案時之 *f*-measure 成效

6.3 小結

本研究針對翻譯英文動詞分析了翻譯模型於前一百名高頻動詞、前二十二、前十二和前六名具競爭力候選人動詞的翻譯成效；對於翻譯英文名詞的部分，本研究分析了前一百名高頻名詞、前十九、前十和前五名具競爭力候選人名詞使用翻譯模型的翻譯效果。實驗結果發現，無論是英文動詞或名詞，當推薦五個答案時，與公式(4)或公式(9)組合協同推薦的翻譯模型幾乎可包含正確解答，且正確解答在推薦答案中的位置會比單純使用公式(4)或公式(9)還要前面，另外在 *f*-measure 的兩套評量標準下都有較高的成效。

第七章　使用科學人雜誌語料建置翻譯模型

本研究在以上章節分析了專利語料的英漢動名詞組合，我們也以同樣的方式分析科學人雜誌英漢對照電子書[24]。本章節透過科學人語料觀察動詞和名詞的翻譯效果。田侃文[23]將科學人雜誌英漢對照電子書的 1745 篇文章，使用該文句對列系統產出 63256 個英漢對列的高品質句對。本研究沿用這 63256 個句對，與處理專利語料一樣的方式產生關係樹，再將英漢動名詞組合取出對列，最後對列產生了 4814 個英漢動名詞組合。7.1 小節探討動詞翻譯的相關分析，7.2 小節則探討名詞翻譯的分析，7.3 小節為本章節小結論。

7.1 翻譯英文動詞

本小節分為兩個部分探討科學人雜誌的動詞翻譯情形，7.1.1 小節探討前二十五名高頻英文動詞及 7.1.2 小節描述前九名具競爭力候選人之動詞。

7.1.1　科學人前二十五名英文高頻動詞分析

由於科學人雜誌語料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，如下頁

表 7.1 科學人前二十五名英文高頻動詞及其出現次數

have=379 use=143 make=126 provide=119 find=105 take=86 play=83	reduce=65 offer=65 create=65 produce=60 require=59 cause=59	improve=50 form=49 increase=42 change=42 understand=40 develop=40	build=40 explain=37 give=34 study=33 solve=33 generate=31

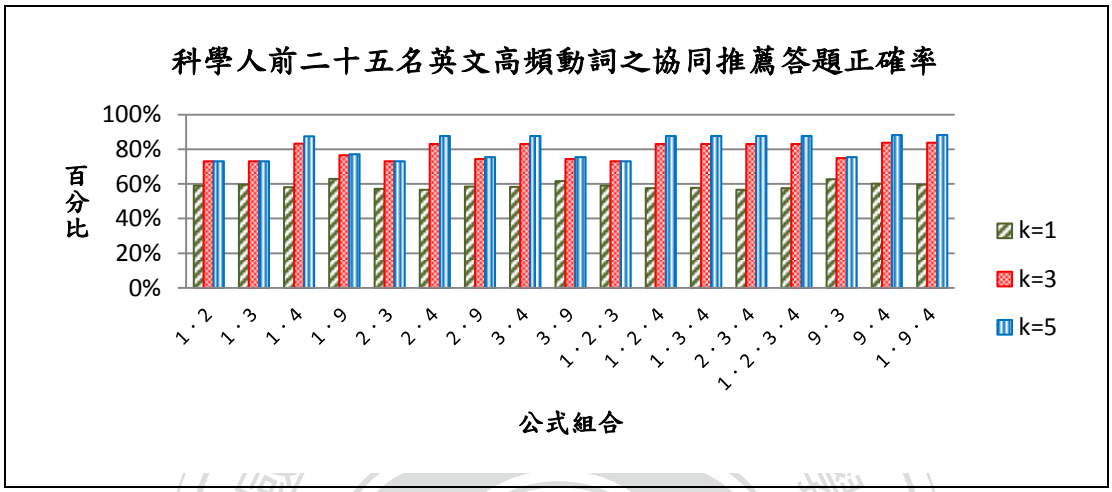


圖 7.1 科學人前 25 名英文高頻動詞之協同推薦答題正確率

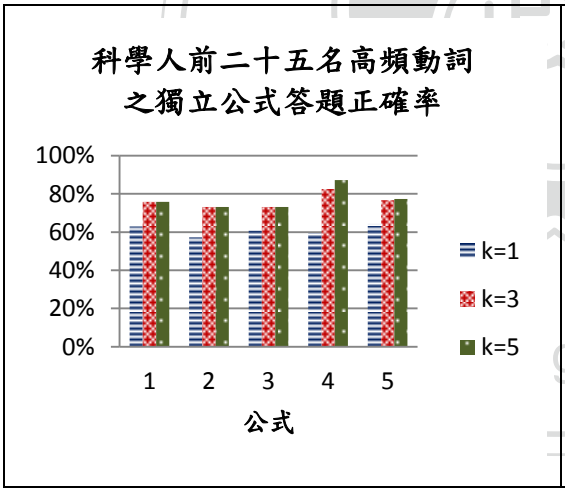


圖 7.2 科學人前 25 名動詞答題正確率

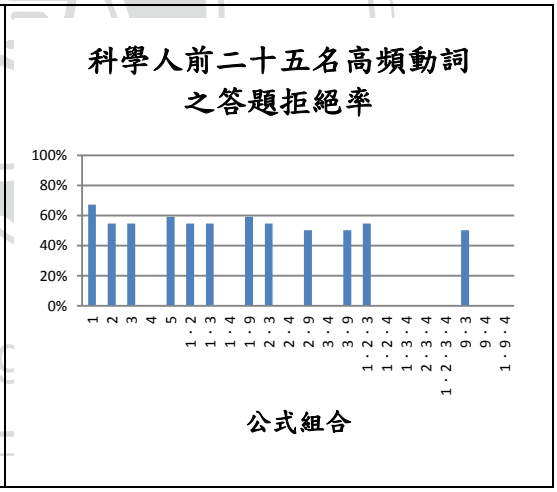


圖 7.3 科學人前 25 名動詞答題拒絕率

表 7.1 所示。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。從圖 7.1 及圖 7.2 可以看到科學人雜誌由於語料數量比起專利語料少許多（僅有專利語料的 13%），寫作風格又較專利文句活潑，因此在協同推薦的答題正確率不如在專利語料的表現，圖 7.3 的最高的答題拒絕率甚至逼近 70%。如下頁圖 7.4 所示，在翻譯模型推薦一個答案及五個答案

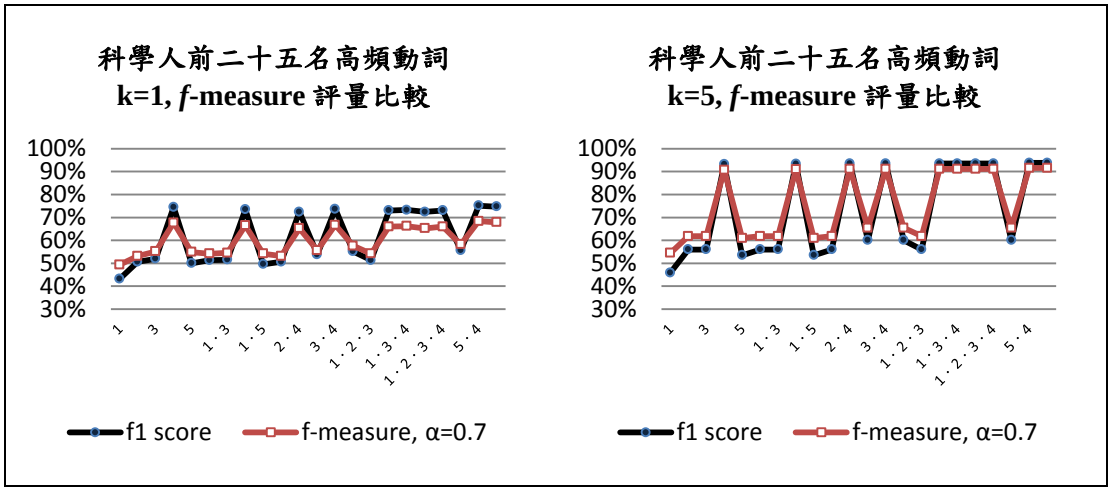


圖 7.4 翻譯模型在科學人前 25 名高頻動詞推薦一個及五個答案時之 ***f***-measure 成效

時的趨勢分布與專利語料的前一百名高頻動詞趨勢相同；因為語料數量較少而資料變化又較大（科學人文章風格較專利文句豐富），因此翻譯模型在推薦五個答案時 ***f***-measure 最高的成效落在 90%左右，低於在專利語料時的表現。

7.1.2 科學人前九名具競爭力候選人之動詞分析

本研究由前二十五名高頻動詞中選出了具翻譯競爭力的動詞，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。這九個動詞如下頁表 7.2 所示，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 7.5 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)搭配的公式組合翻譯模型仍是表現最為亮眼。

表 7.2 科學人前九名具有競爭力候選人之動詞

give={來=1, 供=1, 與=1, 賦予=2, 出=3, 做=4, 給予=4, 給=8, 有=10}
use={消耗=1, 應用=2, 採用=2, 運用=4, 用=7, 利用=60, 使用=67}
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}
build={增進=1, 提高=1, 立=1, 造=1, 興建=3, 建構=6, 建立=9, 建造=18}
develop={展開=1, 演化=1, 做=2, 成=2, 出現=5, 罹患=5, 開發=11, 發展=13}
find={到=1, 尋獲=1, 得到=1, 達成=1, 找尋=2, 搜尋=2, 尋找=5, 發現=24, 找到=33, 找=35}
create={任=1, 塑造=1, 引發=1, 戴=1, 來=2, 引起=2, 製作=2, 設計=2, 生=3, 製造=4, 造=4, 產生=5, 有=6, 創造=12, 出=19}
take={作=1, 使=1, 修=1, 入=1, 則=1, 化=1, 反應=1, 受=1, 可=1, 得到=1, 抱持=1, 接=1, 攻擊=1, 照=1, 產生=1, 看看=1, 耗費=1, 變成=1, 長=1, 需=1, 做=2, 出=2, 成=2, 拍=2, 服用=2, 利用=3, 取=3, 採用=3, 用=3, 取得=4, 拍攝=4, 與=4, 走=5, 需要=8, 採取=9, 有=10}
make={看=1, 組成=1, 行=1, 製=1, 形成=2, 得到=2, 構成=2, 獲得=2, 造=2, 產生=3, 化=4, 取得=4, 成=5, 製作=6, 製造=7, 出=9, 有=29, 做=45}

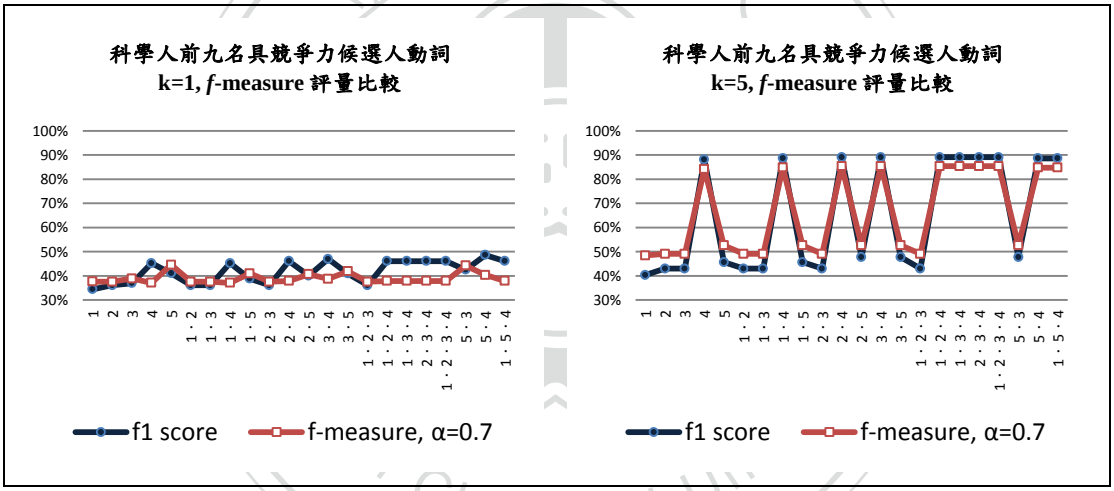


圖 7.5 翻譯模型在科學人前 9 名動詞推薦一個及五個答案時之 *f*-measure 成效

7.2 翻譯英文名詞

本小節分為兩個部分探討科學人雜誌的名詞翻譯情形，7.2.1 小節探討前二十五名高頻英文名詞及 7.2.2 小節描述前五名具競爭力候選人之名詞。

表 7.3 科學人前二十五名英文高頻名詞及其出現次數

role=80 way=67 effect=61 problem=55 risk=42 information=41 system=35	cell=35 property=34 energy=32 ability=32 disease=31 question=30	number=28 technique=26 gene=26 datum=26 behavior=26 activity=26	technology=25 structure=25 it=24 function=24 result=23 clue=23

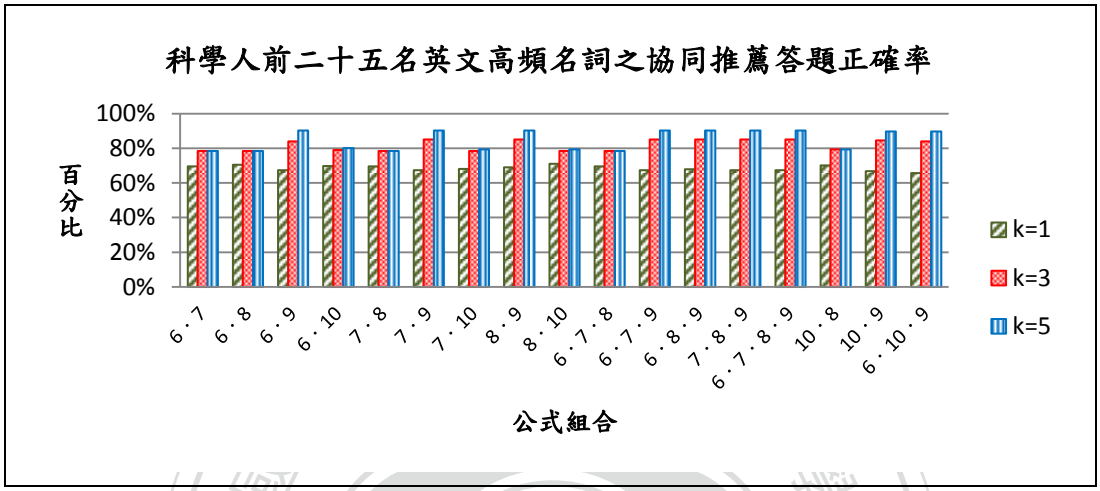


圖 7.6 科學人前 25 名英文高頻名詞之協同推薦答題正確率

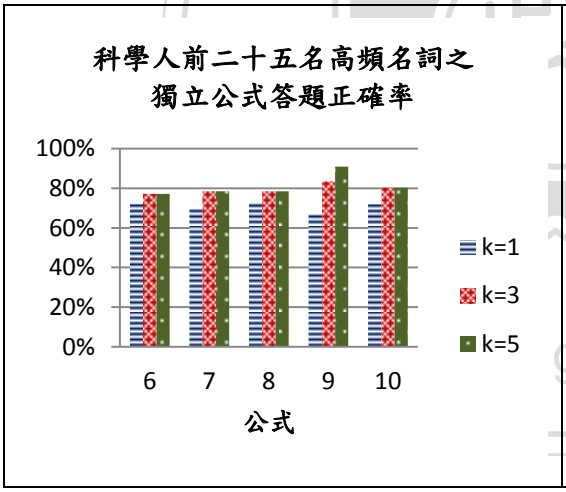


圖 7.7 科學人前 25 名名詞答題正確率

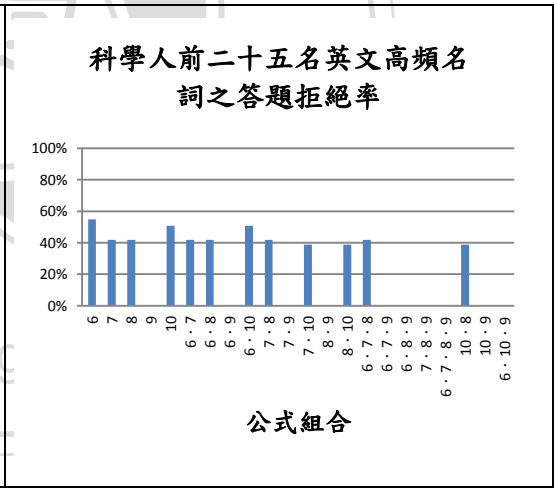


圖 7.8 科學人前 25 名名詞答題拒絕率

7.2.1 科學人前二十五名英文高頻名詞分析

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名詞，這些動詞在資料中至少出現過 23 次以上，最多的出現次數則為 80 次，

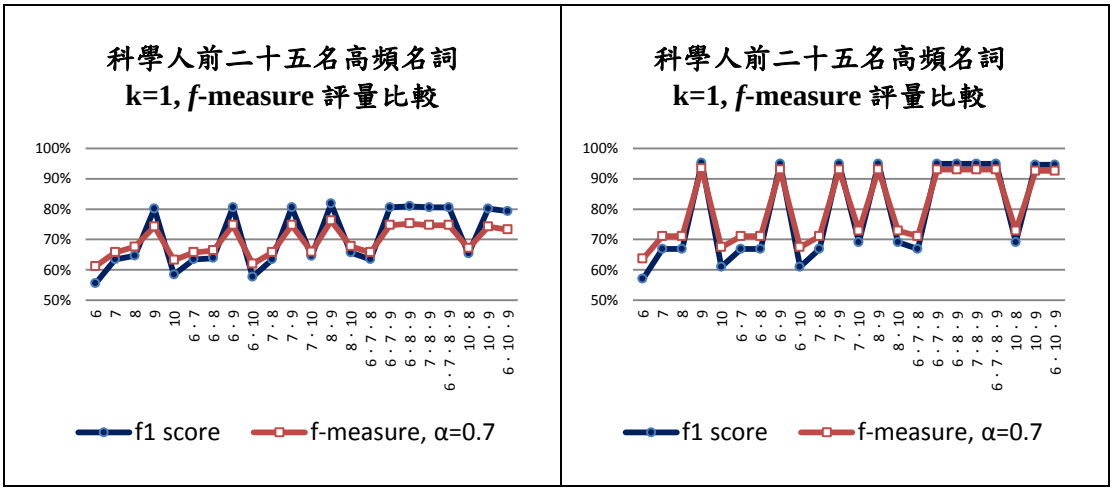


圖 7.9 翻譯模型在科學人前 25 名名詞推薦一個及五個答案時之 *f*-measure 成效

如上頁表 7.3 所示。這二十五個英文名詞總共出現於 877 筆資料之中，訓練資料共有 702 筆，測試資料則有 175 筆。從上頁圖 7.6 及圖 7.7 可以看到科學人語料量較少，寫作風格又較專利文句活潑，因此答題正確率不如在專利語料的表現，上頁圖 7.8 的最高的答題拒絕率接近 60%。如圖 7.9 所示，在翻譯模型推薦一個答案及五個答案時的趨勢分布與專利語料的前一百名高頻名詞趨勢相似；因為語料數量較少而資料變化又較大，在推薦五個答案時 *f*-measure 最高的成效落在 92%左右。

7.2.2 科學人前五名具競爭力候選人之名詞分析

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。這九個動詞如下頁表 7.4 所示，總共出現在 172 筆英漢動名詞組合，訓練資料有 138 筆，測試資料則有 34 筆。如下頁圖 7.10 所示，科學人前五名具競爭力候選人名詞在 *f*-measure 的趨勢較為曲折，可以特別注意到與翻譯動詞公式(5)對稱的公式(10)在只能推薦一個答案且注重於答題正確率時，表現相對於其他獨立運作的並無差異，而與公式(10)搭配的組合表現也沒有顯眼的表現，這與我們在動詞

表 7.4 科學人前五名具競爭力候選人之名詞及其翻譯對應

datum={訊號=1, 資訊=2, 數據=9, 資料=14}
activity={力=1, 活動=10, 活性=15}
structure={困難=1, 構造=10, 結構=14}
property={效果=1, 特性=16, 性質=17}
effect={功效=1, 反應=1, 後果=1, 結果=1, 影響=10, 效果=12, 作用=16, 效應=19}

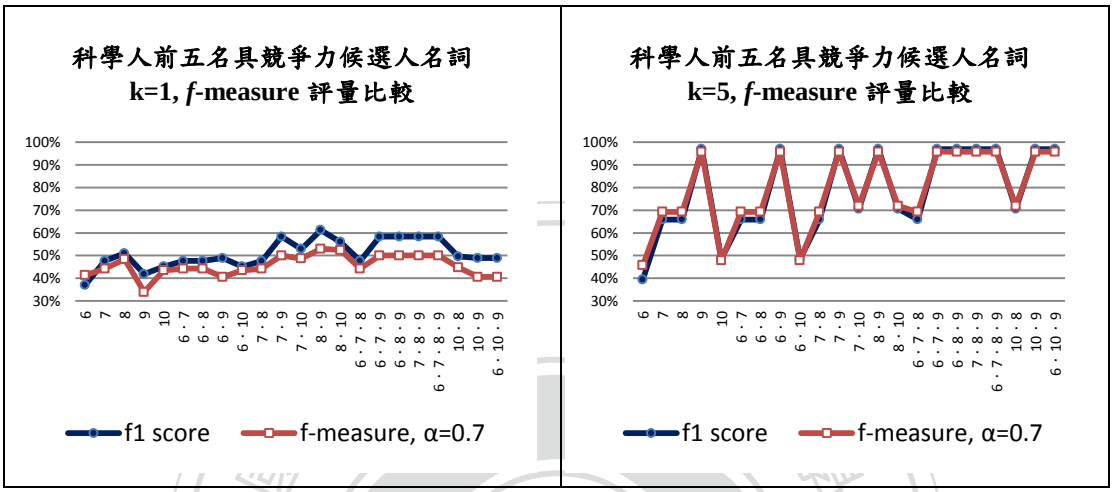


圖 7.10 翻譯模型在科學人前 5 名名詞推薦一個及五個答案時之 *f*-measure 成效

部分觀察到的現象不一樣。公式(9)單獨表現時效果最差，但是與其他公式搭配時因為加強了回答率的部份而使得分數提升。不過在推薦五個答案時與公式(9)搭配的公式組合仍是表現最為亮眼。

7.3 小結

本研究針對科學人語料中的英文動詞及名詞的翻譯成效進行分析。實驗結果發現，無論是英文動詞或名詞，當推薦五個答案時，使用公式組合協同推薦的翻譯模型幾乎可包含正確解答，且在 *f*-measure 的兩套評量標準下都有較高的成效；比較不同的是，在科學人語料中，動詞翻譯公式(5)在具競爭力動詞中可以有不錯的 f-measure 成效，但在翻譯名詞的部分則沒有對稱的成效表現。

第八章　受試者實驗

8.1 小節說明我們實驗設計，8.2 至 8.4 小節描述本研究設計的三組實驗，8.5 小節則是我們的實驗結論。

8.1 實驗說明

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，並設定三項翻譯英文動詞的實驗，邀請具有資工背景的受試者參加。我們規定三項實驗的受試者不得重複跨實驗參加，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。實驗題目以公式(1)所能得到的資訊為基準，即受試者至少知道英文動名詞組合及英文名詞的中文翻譯這些資訊，不同實驗附加不同程度資訊以測試受試者會否因為附加資訊的多寡影響答題效果。本研究將使用公式(1)建立的翻譯模型作為參賽者，以比較受試者的答題情況與使用公式(1)建立的翻譯模型表現，驗證模型的翻譯效能。

8.2 實驗一：提供題目英漢資訊的選擇題

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，並將中文題目中對應的動詞翻譯位置挖空，如表 8.1 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 8.1 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，挑選出較高頻出現的三個詞彙作為誤導選項。實驗一主要目的是讓受試者在接收完整題目資訊之下，要求受試者將目標動詞翻譯成中文詞彙，並提供選項選答。下頁圖 8.1 為受試者及本研究翻譯模型的答題正確率比較。

以公式(1)建立的翻譯模型為參賽者，本研究的翻譯模型答對了六題。我們可以發現翻譯模型與六位受試者的表現平手，並贏過了八位受試者的表現。雖然這個實驗提供了受試者最多的資訊（全題皆有英漢翻譯及四個答案選項供選擇），但是平均的表現而言，我們的翻譯模型還是略勝一籌。

表 8.1 實驗一題目範例

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}

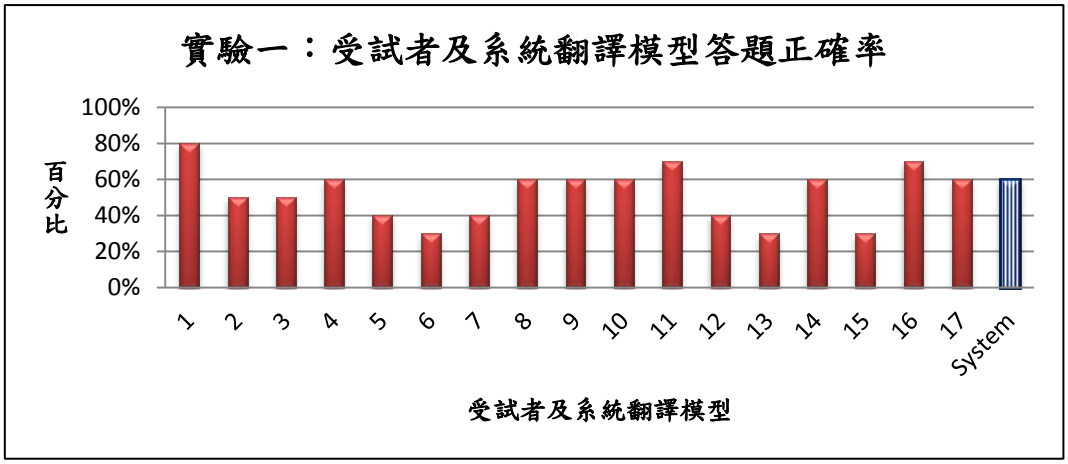


圖 8.1 實驗一：受試者及系統翻譯模型答題正確率

表 8.2 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。

8.3 實驗二：提供題目英漢資訊的填空題

與實驗一相同，我們提供受試者英文及其中文翻譯的題目資訊，並將題目的英文目標動詞以灰底粗體標示，並將中文題目對應的翻譯位置挖空；與實驗一不同的地方在於實驗一提供了四個選項讓受試者選擇，而在實驗二我們不提供選項，直接要求受試者填寫他們心目中的詞彙，如表 8.2 所示。

由下頁圖 8.2 可以看到 19 位受試者的答題表現。在不提供答案選項的條件下，與實驗一雖是同樣題目但是答題的困難度增加，受試者最多答對五題；比較起實驗一，實驗二雖然同樣提供了題目語意，但是在答案選項上沒有任何提示，答案的可能範圍增大，因此受試者的答案容易偏離題目的真正答案，導致平均答題正確率下降。

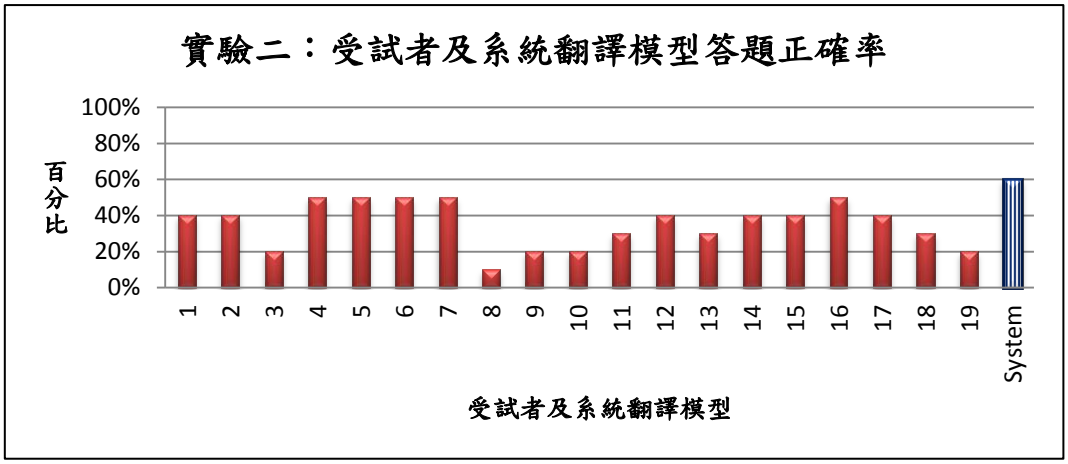


圖 8.2 實驗二：受試者及系統翻譯模型答題正確率

表 8.3 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**

8.4 實驗三：提供英漢資訊的動名詞組合選擇題

在實驗三中，我們不提供受試者完整題目的環境，僅提供翻譯模型所能得到的資訊給受試者並提供答案選項；如表 8.3 所示，我們僅提供英文動名詞組合及英文名詞的中文翻譯，並將英文目標動詞以灰底粗體標記，要求受試者從我們提供的四個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。

從下頁圖 8.3 可以看到除了編號 8 號的受試者答對 7 題，表現贏過我們的翻譯模型，其他受試者最多只能答對 5 題。從結果上來看，實驗三答對 5 題以上的受試者比實驗二的填空實驗人數多上 2 位，這是個有趣的現象。

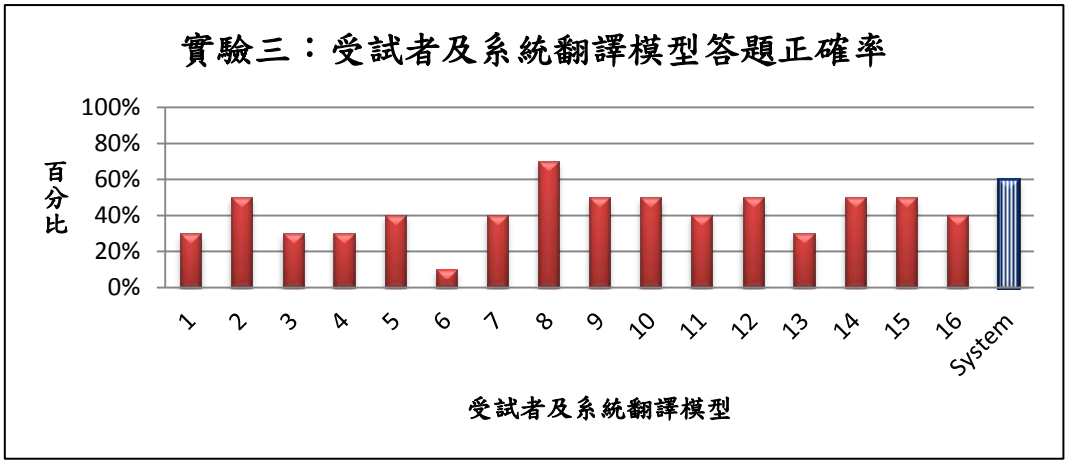


圖 8.3 實驗三：受試者及系統翻譯模型答題正確率

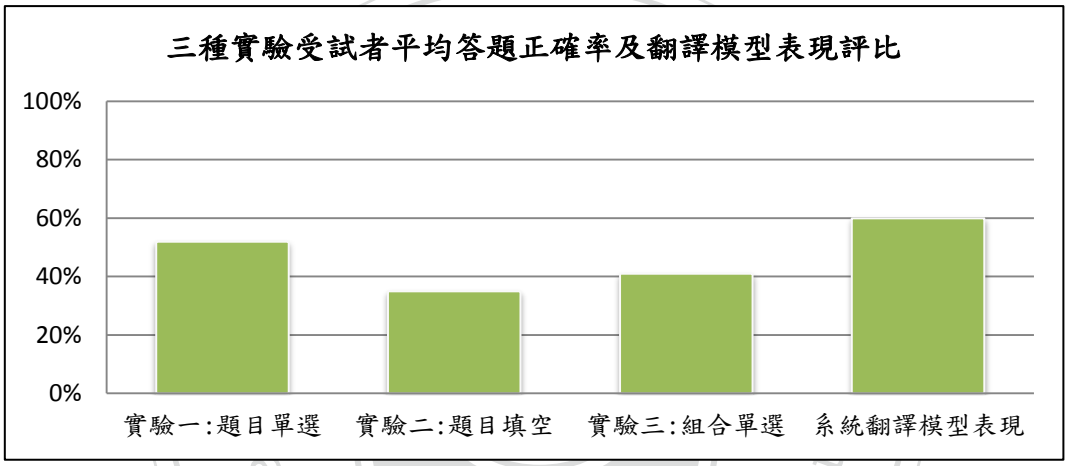


圖 8.4 三種實驗受試者平均答題正確率及翻譯模型表現評比

8.5 小結

圖 8.4 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 52%；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 35%；實驗三的受試者則平均答對了 41%。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。透過這三項實驗，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

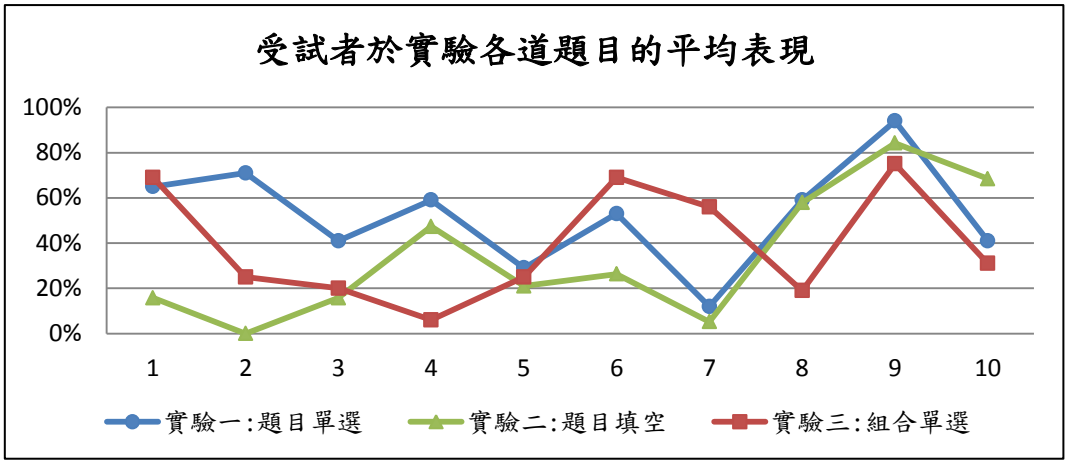


圖 8.5 三組實驗之答題情形比較

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翻譯答題，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。

圖 8.5 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完整語意環境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；第五題卻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受試者作答的時候，在題目提供的附加資訊多寡之外，受試者閱讀到動名詞組合時可能有其特定的直覺，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。

第九章　結論與未來展望

本章節分為兩個小節描述，9.1 小節為本研究的結論，9.2 小節為研究的未來展望。

9.1 結論

許多教學與研究指出英文的詞彙組合有其特定的共現性，而非任何同義的詞彙可以替換。而我們將焦點放在英文的動詞及名詞組合上，其特定的聯結關係為某個名詞適合當作特定動詞的對象，英文使用者在動詞及名詞組合的使用上也有習慣的一定用法。如果將英文翻譯成中文，我們認為英漢互為翻譯的動名詞組合應該也有特定的翻譯關係；因此本研究透過英漢平行語料庫從中找尋英漢對列的動名詞組合，以探究英文的動詞及名詞組合之下，有無特定的中文翻譯用法。

本研究對於英漢專利平行文句語料庫[13]及科學人雜誌英漢對照電子書[24]進行了平行且對稱的分析。我們從一百萬句專利平行句對中挑選出翻譯對列品質較高的 33 萬組長句對，再根據本研究完成的技術名詞表做技術名詞斷詞標記，接著把英文的語料使用 Stanford Parser[15]進行詞幹還原、中文的語料則使用 Stanford Chinese Segmenter[14]將一般詞彙斷詞。我們使用 Stanford Parser 分別剖析英文及中文文句得到句子的關係樹，並根據「DIRECT_OBJECT」的關係特徵取得動詞及名詞組合，再使用本研究的近義詞典進行英漢動名詞組合對列。

本研究分別使用資訊考慮程度不同的十種公式，建立針對英漢對列動名詞組合中的英文動詞與名詞推薦中文翻譯詞彙的模型。我們的實驗結果顯示，在專利平行文句語料庫和科學人雜誌英漢對照電子書中，考慮資訊較多的公式翻譯模型所提供的推薦翻譯較為準確，但是推薦答案平均只有辦法推薦一個，且有答題拒絕率的問題；而資訊考慮條件最為寬鬆的公式(4)及公式(9)都能推薦超過五位以上的推薦翻譯人選，因此公式組合協同推薦的翻譯模型能結合不同公式的優勢而有不錯的翻譯表現。

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)建立的翻譯模型表現比較。三項實驗分別提供了十道相同題目，但是不同實驗的題目含有不同的資訊程度，並要求受試者將題目中的英文動詞根據我們提供的選項挑選出合適的中文翻譯，或是不提供選項要求受試者直接填寫答案。本研究控制這三項實驗的受試者只能參加其中一項實驗而不得重複參加其他實驗，因此三項實驗共有 52 個受試者參與。根據三項實驗發現，提供作答選項的實驗一及實驗三受試者的平均答題正確率比較高，而當我們不提供答案選項而要求受試者直接填寫答案時，實驗二的平均答題正確率為三項實驗中最低，可見即便是人為也很難猜到真正的答案。三項實驗相比，我們的翻譯模型都能贏過受試者的平均表現。

9.2 未來與展望

在處理語料的部分，我們的技術名詞表需要作更多改善得到更精確的技術名詞，降低文句被錯誤斷詞而導致被錯誤剖析的機會。而在剖析文句得到的關係樹結構，我們也認為需要加強，增加正確剖析關係樹結構的數量，並擴張英漢動名詞組合數。由於我們使用的兩套語料都含有技術名詞，因此若採用不含技術名詞的

一般文本作類似的分析，我們也好奇翻譯模型的翻譯效果會不會有所差異，可以更進一步比較。另外，本研究著重於英漢動名詞組合之間詞彙翻譯關係，除了考慮英文及中文，如果有翻譯品質良好、數量豐沛的平行雙語語料，我們也可以觀察各國語言與中文的對應用法。我們的研究提供了對於英漢專利平行文句語料庫及科學人雜誌英漢對照電子書的分析及相關實驗結果，嘗試發掘語言之間的特定關係；我們的系統可以根據不同的翻譯模型推薦英文動詞或名詞的中文翻譯，可用於輔助教學用途，期望對於語言相關學習能有所幫助及成效。

參考文獻

[1] Alexander Budanitsky and Graeme Hirst, Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Association for Computational Linguistics*, 32(1), 13-47, 2006.

[2] 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.

[3] Wenliang Chen, Jun'chi Kazama and Kentaro Torisawa, Bitext Dependency Parsing with Bilingual Subtree Constraints. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 21-29, 2010.

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

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

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

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

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

[9] HowNet。http://www.keenage.com/html/c_index.html [Last visited on 15 June 2011]

[10] Jia-Yan Jian, Yu-Chia Chang, and Jason S. Chang, TANGO: Bilingual Collocational Concordancer. *Proceedings of ACL on Interactive poster and demonstration sessions*, 2004.

[11] 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.

[12] Michael Gamon, Jianfeng Gao, Chris Brockett, Alexander Klementiev, William Dolan, Dmitriy Belenko, and Lucy Vanderwende, Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. *Proceedings of the International Joint Conference on Natural Language Processing*, 2008.

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

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

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

[16] 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.

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

[18] Xing Yi, Jianfeng Gao and William B. Dolan, A Web-based English Proofing System for English as a Second Language Users. *Proceedings of the Third International Joint Conference on Natural Language Processing*, 619-624, 2008.

[19] XML。http://www.w3schools.com/xml/default.asp [Last visited on 15 June 2011]

[20] Shoichi YOKOAMA and Masumi OKUYAMA, Translation Disambiguation of Patent Sentences using Case Frames. *Machine Translation Summit XII WS7: Third Workshop on Patent Translation*, 33-36, 2009

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

[22] 中央研究院中文斷詞系統。http://ckipsvr.iis.sinica.edu.tw/ [Last visited on 15 June 2011]

[23] 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。

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

[25] 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]

[26] 曾元顯，劉昭麟，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*，279-292，2009。

附錄Ⅰ　口試問題紀錄

問題與答案紀錄

Q1.	圖 4.2 中講解 E-HowNet 架構是否改以「混亂」等詞說明，以和前面的講述範例一致？
A1.	E-HowNet 中「混亂」等詞的定義都沒有「和鳴」一詞的定義來的完整，為了能詳細描述 E-HowNet 架構，因此本研究選擇使用「和鳴」一詞講解。
Q2.	動名詞組合已有包含技術名詞，直接進行 pattern matching 即可，為何還需要針對名詞建置翻譯模型？
A2.	本研究使用的動名詞組合皆已排除技術名詞在外。本研究主要利用專利文句排除技術名詞之剩餘部分，探討一般常用的英漢動名詞組合翻譯情形。第三章技術名詞標記是為了能將技術名詞排除，以讓剖析器能正確剖析文句。本研究所列出的前一百名高頻名詞皆為一般名詞，沒有技術名詞在內。
Q3.	抽取關係樹中的動名詞組合如何能確定動詞與名詞之間的位置？例如「pill taking」會否表示成 dobj(pill, taking)？
A3.	Stanford Parser 對於關係樹的標示是固定的，一定是動詞在前、名詞在後，形成 dobj(verb-number1, noun-number2)的形式；number1 和 number2 即為記錄動詞與名詞在句子中的出現位置。
Q4.	公式(1)其實可以比擬成 n 比較大的 n-gram，是否有探討資料 sparse 的部分？
A4.	對於每個公式本研究都有列出其答題拒絕率，而公式(1)有最高的拒絕率，能解釋資料稀疏的問題。
Q5.	為何沒有記錄學生錯誤的資料庫，以校正學生錯誤？
A5.	本研究不是針對學生寫作錯誤提出的校正系統。本研究主旨為從大量正確對應的英漢語料中嘗試挖掘常被運用的正確用法及其常用的翻譯對照。

Q6.	公式(1)到公式(4)是為 smoothing，為何不使用係數調整各公式權重？
A6.	公式(1)到公式(4)並非為 smoothing 過程，它們是獨立的公式。本研究旨在評比各翻譯模型的翻譯成效，公式組合主要是讓公式能協同推薦。
Q7.	訓練與測試資料分割的比重為何？論文是否有提及？
A7.	8:2，在論文內文有描寫清楚。
Q8.	受試者實驗中，提供給受試者的答案選項都很像，是否因為這個原因才造成圖 8.5 的受試者表現趨勢？
A8.	本研究提出讓受試者回答的十道題目皆從科學人語料抽取而來，而每道題目皆只有一個答案，其他誤導答案也是針對該英文動詞在科學人語料中對應到的其他中文翻譯。本研究的翻譯模型便是面對如此的情況，因此設計讓受試者面臨一樣的狀況以比較翻譯成效。BLEU 指標也是如此運算，只有正確答案是唯一的答案，這是我們設計實驗的目的。
Q9.	公式(1)可用於何處？
A9.	已於論文第 36 頁內容補充。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)可以透過這三個詞彙的資訊，在語料中觀察「take」跟「pills」一起使用且「pills」對應到「藥」時，「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)找到的中文翻譯中可以比對到「吃」這個詞彙，進而讓使用者知道使用「take pills」才是正確的用法。

Q10.	公式(4)的翻譯原理並不合常理？
A10.	公式(4)的主要功能為與其他公式進行協同推薦。如論文第 43 頁所描述：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，因為協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。
Q11.	為何不使用其他不含技術名詞的語料？專利文句寫作模式是固定的。
A11.	其他語料數量並不夠多。我們想驗證專利文句排除技術名詞部分是否能有良好的參考價值，這是目前還沒有人嘗試過的。