國立臺灣大學理學院應用數學科學研究所

碩士論文

Institute of Applied Mathematical Sciences

College of Science

National Taiwan University

Master's Thesis

評估 Google 翻譯表現
—以 4 種語言翻譯 100 篇英文論文爲例

Evaluating the Performance of Google Translation
—For 4 languages on 100 English written papers

陳玟瑾

Wen-Chin Chen

指導教授: 杜憶萍 博士

Advisor: I-Ping Tu Ph.D.

中華民國 113 年 7 月

July, 2024

doi:10.6342/NTU202404301

摘要

隨著翻譯的需求與日俱增，各種翻譯軟體如雨後春筍般湧現。時逢論文抄襲爭議層出不窮，因此許多學校發布新的相關規定來因應，要求在提交論文前需通過指定的論文相似度檢測。若有人利用翻譯軟體將他人的論文經過多層翻譯後，當作自己的論文，則如何評估翻譯的表現也因此成爲了一個重要的議題。在眾多案例中，中國蘇州大學社會學院的學生邵寶在其博士論文《清末留日學生與日本社會》中，在未引用或標注來源的情況下將日本學者酒井順一郎之著作翻譯至中文後直接使用，相似比例達到其論文的 80 %，而我們決定採用其中翻譯相似的概念，以翻譯軟體生成不同版本的譯本後，再透過論文比對系統來評估該翻譯軟體的表現。

在本論文中，我們會瞭解檢測翻譯表現這個領域的發展現況，並以 100 篇英文論文爲我們的資料。我們將論文相似度比對系統 Turnitin 作爲檢測 Google 翻譯軟體表現的工具，來分析牽涉到中文、日文、法文和德文的翻譯中不同的翻譯方式對論文相似度的影響，包括單層翻譯和雙層翻譯。結果顯示出翻譯方式對論文相似度具有顯著影響，其中牽涉到日文的翻譯表現較差，而牽涉到中文的翻譯表現較佳。

關鍵字：翻譯表現、翻譯表現評估、Turnitin（相似度比對系統）

doi:10.6342/NTU202404301

Abstract

With the increasing demand for translation, various translation software has emerged rapidly. Amidst ongoing plagiarism controversies, many universities have implemented new regulations requiring theses to pass designated similarity checks before submission. If someone uses translation software to translate another person's paper multiple times and then presents it as their own, evaluating the performance of such translations becomes a critical issue. A notable case is that of Shao Bao, a student at the School of Sociology at Soochow University in China, who translated and used approximately 80% of Japanese scholar Sakai Junichiro's work "Qing Dynasty Chinese Students in Japan and Language and Culture Contact" in his doctoral dissertation without proper citation. Based on this case, we adopt the concept of translation similarity by using translation software to generate different versions of translations and then using a thesis similarity detection system to evaluate the accuracy of the translation software.

This study explores the current state of translation accuracy evaluation, using 100 English academic papers as the dataset. Turnitin, the similarity detection system is employed to assess the performance of Google Translate, analyzing the impact of different translation methods—single-layer and double-layer translations—on thesis similarity across Chinese, Japanese, French, and German translations. The results indicate that translation methods significantly affect thesis similarity, with translations involving Japanese

doi:10.6342/NTU202404301

performing worse and those involving Chinese performing better.

Keywords: Translation performance, Evaluation of translation performance, Turnitin (Similarity checking system)

doi:10.6342/NTU202404301

目次

doi:10.6342/NTU202404301

doi:10.6342/NTU202404301

圖次

doi:10.6342/NTU202404301

表次

doi:10.6342/NTU202404301

第一章　緒論

1.1　研究背景與動機

2022 年，台灣政壇爆發了一系列碩士論文抄襲事件，導致各大專院校對碩博士論文在繳交前有更嚴格的規範，如必須先上傳到相似度比對系統後，其報告結果無虞才能畢業。其後，教育部不只多次召開會議以商討如何精進論文品質和學術倫理審議機制，也投入了 2500 萬台幣與國家圖書館合作，並委託台灣師範大學的技術團隊建立「全國學位論文比對系統」，該系統將免費提供給大學使用。該系統的主要功能是比對學生論文與國內已有的碩博士論文的相似度，是目前民間系統所缺乏的。學生可以同時使用官方和民間的系統進行比對，從而提高比對結果的可靠性。由此可知，不管是抄襲概念的認知還是比對系統的使用，其重要性是與日俱增。

『抄襲』在這兩年已經是大眾耳熟能詳的字眼，根據 Fishman [13]，其定義爲在沒有適當承認來源的情況下使用他人的內容、結構或想法。若再將其聲稱爲自己原創的內容，則定義爲剽竊。而在此定義下，Foltýnek et al. [14] 又將抄襲行爲分爲以下五種：

1. 保字抄襲（Characters-preserving）：直接複製文本內容的抄襲，可能會註明來源
2. 保留語法的抄襲（Syntax-preserving）：使用同義詞或相似詞替換文本內容，但語法架構不變
3. 保留語義的抄襲（Semantics-preserving）：涉及改變單詞和句子結構，但保留段落的含義，如翻譯或釋義
4. 保留概念的抄襲（Idea-preserving）：使用他人提出的概念或結構，並以自己的文字描述代筆（Ghostwriting）

其中保留語義的抄襲引起了我們的注意，因爲翻譯在幫助我們理解國外論文

doi:10.6342/NTU202404301

上是非常有用的工具。但若是因爲投機取巧，將翻譯後的內容稱爲自己所創，進而將其當作自己的成果繳交的話，那此行爲其實就是違反了學術倫理，並構成抄襲與剽竊。儘管抄襲的定義已廣爲人知，但這類抄襲情況卻仍然時有耳聞，例如中國蘇州大學社會學院中國近現代史專業學生邵寶的博士論文《清末留日學生與日本社會》（2013 年 9 月）被舉報涉嫌抄襲日本學者酒井順一郎 [1]（2010 年 3 月），其論文內有大約 80% 的內容是透過翻譯該書至中文而來。經蘇州大學學術委員會認定其論文存在學術不端，最後撤銷邵寶歷史學博士學位。除此之外，劉繼仁等人 [17] 也曾訪問國立成功大學和國立中山大學的教授對於抄襲情事的瞭解，部分教授表示曾得知學生使用翻譯軟體將他人文章翻譯後，再將其放到自己的作業中並繳交。而該研究也透過訪問學生後發現，如此抄襲的動機爲投機和懶惰心態、語言能力不足或是對於學術規範及抄襲行爲的認識不足等等。不論他們動機爲何，這些案例中的行爲都不可取，但卻讓我們聯想到一個問題：既然使用翻譯軟體以任何翻譯方式都可以協助我們來產生一篇和原文非常相似（實際上即是抄襲）的文章，而比對系統又是一個可以檢測文章和其他文章的相似度的系統，那我們是否可以利用這個概念，將選定一個翻譯軟體爲主要研究對象，再選出一個我們所信任的比對系統來當成檢測翻譯軟體表現的工具呢？爲此，我們決定先挑選出我們認爲可信的比對系統。

1.2 Turnitin 之選取與簡介

如前面所述，教育部雖有推出全國學位論文比對系統，但因其在我們開始研究時尚未開放使用，所以並未將其納入考慮。而目前在台灣較知名的比對系統爲 Turnitin，Turnitin 是一家提供學術不誠信檢測和教學工具的公司，由 Chris Hables Gray 和 Scott Atkins 於 1997 年在美國創立。Turnitin 最初開發了一個名爲 OriginalityTester 的軟件，用於檢測學術論文中可能存在的抄襲行爲。隨著時間的推移，Turnitin 發展成爲一個全面的學術工具，提供多種功能，包括但不限於原創性檢查、教師評分和反饋、學術寫作指導、學術誠信教學和學術資源庫等。其中原創性檢查系統即是我們需要的相似度比對系統，它能夠檢測學生提交的論文、報告和其他學術作品與其資料庫內容的相似度。而 Turnitin 最爲人所知的便是其龐大的資料庫，其資料庫從 1988 年收錄至今，包括以下三個類別：

1. 網際網路來源：約 993 億篇來自現存和歸檔網頁的網際網路資料，
2. 出版物：約 8940 萬篇經常更新之專業雜誌、期刊和出版物內容，

doi:10.6342/NTU202404301

3. 學生文稿 - 約 18 億篇由全球 Turnitin 用户中已儲存至系統的學生文稿。

除了龐大的資料庫，Turnitin 更是支援 30 種語言的文章檢測，爲此它不只被全球多個教育機構所使用，國立臺灣大學、國立清華大學、國立陽明交通大學、國立成功大學以及臺灣其他多所學校皆是使用此系統作爲主要的論文比對系統，因此我們有理由相信其檢測相似度的能力，最終也決定將其選定爲我們檢測翻譯軟體準確度的工具。

1.3 Google 翻譯之選取與簡介

選定 Turnitin 爲我們檢測翻譯軟體表現的工具後，接下來就是決定要檢測哪個翻譯軟體。隨著翻譯的需求與重要性與日俱增，許多公司也紛紛推出翻譯軟體，而以下是目前較爲知名的翻譯軟體，他們在易用性、翻譯準確度和語言支持方面都受到廣泛的認可：

1. Google 翻譯（Google Translate）：由 Google 公司開發，支持超過 100 種語言，使用者包含了全球範圍內的廣大用户，其特點爲免費、廣泛的語言支持、實時翻譯、圖片翻譯、手寫翻譯等。
2. DeepL 翻譯器（DeepL Translator）：由 DeepL GmbH 公司開發，支持多種歐洲語言，特別是德語、西班牙語和法語的翻譯，使用者包含了專業翻譯者和普通用户，以高準確度和自然翻譯聞名。
3. 微軟翻譯器（Microsoft Translator）：由 Microsoft 公司開發，支持超過 60 種語言，使用者包含了全球用户，特別是微軟產品的用户，其特點爲與微軟的各種產品和服務集成，包括 Office 和 Bing 搜索。
4. Yandex 翻譯（Yandex.Translate）：由 Yandex 公司開發，是一家俄羅斯網際網路企業，旗下的搜尋引擎曾在俄國內擁有逾 60% 的市場占有率。其支持超過 90 種語言，使用者包含了俄羅斯及其他俄語使用國家的用户，特點爲俄語翻譯特別出色，同時提供語音和圖片翻譯。
5. Papago：由 Naver 公司開發，主要支持韓語和英語，但也支持其他語言。使用者主要在韓國和需要韓語翻譯的國家，其特點爲提供實時翻譯和語音翻譯。

上述這些翻譯軟體在不同的領域和用途中都有其特點和優勢，因此在全球範圍內都有大量的用户群。而在此之中我們選擇了 Google 翻譯做爲我們的分析對象，不只是因爲它是免費的且無使用次數及字數的上限，更因爲它是目前使用次

doi:10.6342/NTU202404301

數最多，且最爲人熟知的翻譯軟體，因此我們相信以 Google 翻譯爲我們的翻譯工具所得到的結果會具有一定的參考價值。以下我們將詳細介紹 Google 翻譯。

Google 翻譯是由谷歌公司開發的一個免費的網絡翻譯服務，它讓用户可以將文本、網頁、圖片和語音從一種語言翻譯成另一種語言。該服務於 2006 年推出，並迅速成爲全球最廣泛使用的翻譯工具之一。其特點如下：

1. 多語言支持：Google 翻譯支持超過 100 種語言的翻譯，涵蓋了全球大部分語言，且數量仍在增加。
2. 互動翻譯：除了文本翻譯，Google 翻譯還支持翻譯網頁、圖片、手寫文字等，增加了其使用範圍。
3. 語音翻譯：用户可以輸入語音，Google 翻譯不僅可以將其翻譯成文字，還提供了聽覺輸出。
4. 實時翻譯：Google 翻譯的實時翻譯功能允許用户在對話中即時翻譯語言，不需要等待過久的時間，非常適合旅行和跨國溝通。
5. 翻譯準確度：Google 翻譯由先進的 Pathways Language Model 2 (PaLM 2) 模型支援，這種技術能夠更好地理解語言的上下文和語法結構，因此能夠提供比傳統機器翻譯更自然、更準確的翻譯結果，尤其是在翻譯複雜句式和專業術語的部分。
6. 易用性：Google 翻譯界面簡單易用，無需專業知識或複雜的安裝過程，只需在網頁上輸入文本即可，任何人都可以輕鬆使用。
7. 免費使用：Google 翻譯是可以免費使用的，這使得它對廣大用户來説非常實惠。
8. 廣泛的互操作性：Google 翻譯可以與多種應用程序和設備集成，例如智能手機、平板電腦和電腦，這使得翻譯更加方便。
9. 廣泛的市場推廣：谷歌公司在全球範圍内進行了廣泛的市場推廣，使得 Google 翻譯成爲全球知名的品牌。
10. 數據和算法優化：Google 擁有大量的數據和強大的算法能力，不斷優化翻譯模型，提高翻譯品質。
11. 社區參與：Google 翻譯的用户可以參與翻譯社區，幫助改善翻譯品質，這種參與感也促使用户繼續使用該服務。

由於上述多種因素的結合，Google 翻譯在翻譯市場上佔據了領先地位，並成爲了全球最多人使用的翻譯軟體。不過雖然 Google 翻譯具有許多優點，但也存在一些缺點是用户在使用時需要注意的：

1. 翻譯準確度問題：雖然 Google 翻譯的準確度很高，但對於一些專業或文化敏感

doi:10.6342/NTU202404301

的內容，翻譯可能會失眞或誤解。

2. 文化差異：Google 翻譯可能無法完全理解文化背景和語境，導致翻譯結果不夠準確或適應性差。

3. 隱私問題：Google 翻譯需要用户上傳文本或圖片，這可能會引發對用户數據隱私的擔憂。

4. 語音翻譯限制：Google 翻譯的語音翻譯功能可能會受到語音品質、語言風格和口音的影響。

5. 語法結構問題：對於某些語言的複雜語法結構，Google 翻譯可能無法正確翻譯。

6. 受制於谷歌服務：在一些受限的地區，由於谷歌服務可能被屏蔽，Google 翻譯的使用可能會受到影響。

Google 翻譯目前利用的模型是 PaLM 2（Anil et al. [7]），PaLM 2 是 Google 開發的新一代語言模型，旨在提升多語言能力、程式碼生成能力以及邏輯推理能力。PaLM 2 使用 Pathways 進行訓練。Pathways 是一種新的機器學習系統，可以在數千個加速器晶片（包括跨越多個 TPU v4 Pod 的加速器晶片）上進行高效訓練。PaLM 2 是 PaLM [10] 的後繼者，採用 Transformer 架構，Transformer 是一種主要用於序列轉換任務的神經網路架構。不同於傳統的循環神經網路 (RNN) 或卷積神經網路 (CNN)，Transformer 完全依賴注意力（Attention）機制來捕捉輸入和輸出序列中各個位置之間的關係，其目標爲減少序列計算，並允許更多平行化處理，從而提高訓練速度和翻譯品質。以下是 Transformer 主要工作原理的詳細說明（Vaswani et al. [21]）：

1. 自注意力機制（Self-Attention）：Transformer 的核心是自注意力機制，它允許模型在處理序列數據時關注不同位置的資訊，並學習它們之間的關係。以翻譯”The agreement on the table is important” 爲例，當模型處理”agreement” 時，自注意力機制可以讓模型關注”table”，從而理解”agreement” 的含義是指放在桌子上的協議。自注意力機制通過計算詞彙之間的點積並使用 softmax 函數來得到權重，從而學習詞彙之間的關係。

2. 多頭注意力機制（Multi-Head Attention）：Transformer 使用多頭注意力機制來捕捉詞彙之間更豐富的關係。多頭注意力機制會將詞彙的嵌入向量投影到多個不同的子空間中，並在每個子空間中分別進行自注意力計算，最後將結果合併起來。這樣可以讓模型從不同的角度來理解詞彙之間的關係。

3. 位置編碼（Positional Encoding）：由於 Transformer 沒有遞迴或卷積來處理詞

doi:10.6342/NTU202404301

彙的順序資訊，因此需要加入位置編碼來提供詞彙在序列中的位置資訊。位置編碼會將每個詞彙的位置資訊編碼成一個向量，並將其加到詞彙的嵌入向量中。Transformer 使用正弦和餘弦函數來生成位置編碼，這樣可以讓模型更容易學習到詞彙之間的相對位置關係。

4. 編碼器-解碼器注意力（Encoder-Decoder Attention）：編碼器-解碼器注意力機制允許解碼器在生成輸出序列時關注輸入序列中相關的部分。例如，在翻譯"The agreement on the table is important" 時，當模型生成" 協議" 時，編碼器-解碼器注意力機制可以讓模型關注 "agreement"，從而確保翻譯的準確性。

而 PaLM2 在 Transformer 架構中引入多項研究進展，例如：

1. 運算資源最佳化調整: 研究發現，資料量與模型大小同等重要。PaLM 2 在訓練時，資料量與模型大小的比例約爲 1:1，以達到最佳效能。

2. 架構和目標改進: PaLM 2 使用混合多種預訓練目標來訓練模型，以理解語言的不同面向。

3. 資料集改進: PaLM 2 的訓練資料集比 PaLM 更大，包含更多非英語資料，這有助於 PaLM 2 更好地處理多語言任務。此外，PaLM 2 訓練時加入了數百種語言的平行語料庫，進一步增強了模型理解和生成多語言文本的能力，並提升了翻譯能力。

4. 更長的上下文長度: PaLM 2 相較 PaLM 能夠處理更長的上下文資訊，這對於需要模型考慮大量上下文的任　（如長對話、長距離推理和理解、摘要等）至關重要。

綜上所述，PaLM2 是一種採用 Transformer 架構的大型語言模型，透過 Pathways 系統進行訓練，並使用高品質、多樣化的資料集及訓練目標，使得其在機器翻譯和其他自然語言處理任務中取得了很好的效果。

1.4　論文架構

在瞭解 Turnitin 和 Google 翻譯這兩大系統後，我們確立了他們作爲我們的研究工具，接下來說明本研究的完整架構。第一章爲緒論，其中整理了我們的研究動機與我們爲何選取 Turnitin 和 Google 翻譯爲我們的研究工具。第二章爲文獻回顧，會了解在檢測翻譯表現這個領域的發展現況。第三章爲資料選取與處理，會瞭解我們搜集資料的過程與 Turnitin 報告解讀。第四章爲資料分析，目的是瞭解

doi:10.6342/NTU202404301

不同的翻譯方式對於相似度的影響，包括單層翻譯和雙層翻譯。第五章爲結論，會整理我們的研究結果並總結。

doi:10.6342/NTU202404301

第二章　文獻回顧

一直以來對於翻譯的表現都有不少討論及研究，被最多人使用的 Google Translate (GT) 自然也不例外，而如何評估其翻譯表現則成爲了一個重要的議題。爲了分析 GT 在不同語言的翻譯表現，Aiken et al. [5] 將五個簡短英文句子透過 GT 翻成 40 種語言後再翻回英文，接著請兩位熟悉英文的評估員來對這些翻譯後的句子進行評分，而評分的標準參考 Guyon (2003) 如下：

理解（Comprehension）

1. 文字清晰、易於理解、文法正確，不需要任何更正。
2. 文本存在輕微錯誤，例如不正確的介詞或冠詞，但其他方面無可挑剔。
3. 文本中存在一些小錯誤和不正確的術語，但含義仍然可以理解。
4. 文本中混雜著小錯誤和不正確的術語，需要付出一定的努力才能理解其意義。
5. 文字是難以理解的亂碼。

可接受性（Acceptability）

1. 文字完全可以接受。
2. 讀者註意到文本中的輕微異常。
3. 讀者閱讀文本時感到有些不舒服。
4. 讀者覺得文字不太嚴肅。
5. 讀者因看到這樣的文本而感到受到侮辱。

意義（Meaning）

1. 譯文準確地表達了原文的意思。
2. 譯文缺乏細微差別。
3. 譯文或多或少傳達了原文的意思。
4. 譯文沒有準確表達原文的意思。
5. 譯文完全沒有表達原文的意思。

最後將所得結果平均並排名，獲得表2.1：（僅部分展示）

doi:10.6342/NTU202404301

Language	Comprehension		Acceptability		Meaning	
	Rank	Score	Rank	Score	Rank	Score
Dutch	1	1.3	1	1.3	3	1.5
Hungarian	1	1.3	1	1.3	3	1.5
Czech	3	1.4	3	1.4	1	1.4
Estonian	3	1.4	3	1.4	1	1.4
Chinese	5	1.5	5	1.5	9	1.8
Italian	5	1.5	5	1.5	9	1.8

Table 2.1: 各語言得分排名

隨著 GT 支援越來越多種語言，由人工評估也稍顯困難，因此 Papineni et al. [18] 提出雙語評估替補（bilingual evaluation understudy, BLEU）來解決此問題。BLEU 是一種自動評估指標，用於評估機器翻譯系統的性能，其運行方式是將機器翻譯的結果與人類翻譯或參考翻譯進行比較，從而計算出一個介於 0 至 1 的分數來表示機器翻譯的品質，分數越高表示翻譯越好。它的計算基於以下幾個步驟：

1. N-gram 匹配計算：首先，BLEU 將參考翻譯和機器翻譯的文本分割成 N-gram 序列（通常是 1-gram 到 4-gram）。然後，它計算機器翻譯中出現在參考翻譯中的 N-gram 的數量。
2. 短語匹配率計算：BLEU 將計算短語匹配率（Precision），即機器翻譯中匹配的 N-gram 數量與機器翻譯中 N-gram 總數的比率。
3. Brevity Penalty 計算：由於機器翻譯可能會產生比參考翻譯更短的結果，BLEU 引入了 Brevity Penalty 來懲罰這種情況，以確保較短的翻譯不會獲得過高的分數。
4. BLEU 分數計算：最後，BLEU 將短語匹配率和 Brevity Penalty 結合起來，計算出一個最終的 BLEU 分數。

Coughlin [11] 及 Culy and Riehemann [12] 皆展示了 BLEU 分數與人類的品質判斷具有高度相關性，因此其後發展爲較常見的指標。如 Aiken [4] 以 6 個短句在不同語言間翻譯後所獲得的 BLEU 分數來獲取兩種語言之間翻譯的可理解程度，結果表示歐洲語言之間的翻譯通常很好，而涉及亞洲語言的翻譯往往相對比較差。時隔八年，Aiken [3] 又將該實驗重做，實驗結果表明同樣以 BLEU 評估，GT 的準確度提高了 34%。

Ethan et al. [8] 的作者爲了評估使用 Google 翻譯將非英語的臨床醫學文章翻譯成英文的可行性和準確性，他們使用 GT 翻譯了 88 篇包括中文、法語、德語、

doi:10.6342/NTU202404301

希伯來語、義大利語、日語、韓語、葡萄牙語和西班牙語在內的 PDF 或 HTML 文件後，再請研究人員進行分析。值得注意的是，在使用 GT 時，作者發現翻譯的難易程度很大程度上取決於期刊使用的文件/文本類型以及 GT 是否可以直接閱讀它們。對於更具挑戰性的翻譯，需要花費額外的時間重複將文字區塊複製到 GT 網站，然後將翻譯後的文字複製到 Word 文件中。作者也發現，他們需要刪除亞洲語言文章中的錯誤換行符，以便進行有意義的翻譯。而表格的翻譯通常非常耗時，因爲它需要對各個行和列標題進行多次翻譯。對於許多文章，特別是希伯來語和亞洲語言的文章，由於列重疊，最終的翻譯文本無法閱讀，需要手動複製和貼上。該論文還強調了遇到的各種其他問題，例如 GT 有一天未能翻譯一篇意大利語文章，但第二天成功翻譯。除此之外，他們亦有符合資格標準但無法翻譯的 21 篇文章，其中法語、德語和日語各 1 篇，中文 2 篇，韓語 3 篇，義大利語 4 篇，希伯來文 9 篇，而故障的原因包括 GT 無法讀取 PDF 或 HTML 檔案、無法從 PDF 複製文字、光學字元辨識失敗或無法辨識某些字母或字元（尤其是亞洲語言）。獲得翻譯後的文本後，再請研究人員估從英語文章與其他語言文章之間的一致性程度。而實驗結果表明英文文章對 80% 的項目有很高的一致性，用作參考標準。其他語言的高一致性程度較低，包括法語、義大利語、日語和韓語文章（佔項目的 40-45%）、西班牙語文章（佔項目的 30%）、希伯來語文章（佔項目的 24%）和中文文章（佔項目的 40%）。與英語相比，所有語言的絕對一致性和高度一致性項目百分比的差異在統計上顯著較差。因此作者得到的結論是翻譯的準確性在很大程度上取決於文章的原始語言，與亞洲語言和希伯來語相比，羅曼語語言的一致性更高，且不同語言之間翻譯的準確性取決於所翻譯的單字和文件的數量。

在 Google 報告了其翻譯引擎的重大改進後，Jeffrey et al. [15] 以 Ethan et al. [8] 的概念重新進行實驗分析來確認 GT 在臨床醫學文章的表現。他們搜集了 2000 年到 2018 年間 9 種語言各 5 篇的隨機對照試驗（randomized controlled trial, RCT），包括中文、法語、德語、義大利語、日語、韓語、羅馬尼亞語、俄語和西班牙語，並請其中一位不會說這 9 種語言中的任何一種的作者評估翻譯成英文後的表現。透過計算一致性百分比，將翻譯後的資料與原始資料的準確性進行比較，將預先定義的良好一致性閾值設爲大於 80%。而研究中包含的 9 種語言的一致性從 85% 到 97% 不等，這表明翻譯後的資料具有很高的一致性，其中西班牙　的一致性最高，一致率高達 97%。

Khanna et al. [16] 爲了評估 GT 在有關患者教育材料的準確性，找了一本有英文版和西班牙文版的健康手冊，再從中選取了 45 個句子，並用 GT 將英文版的翻

doi:10.6342/NTU202404301

譯成西班牙文。接著他們找了三位雙語使用者對這些 GT 翻譯的句子和西班牙版的進行評分，標準如下：

流暢性（Fluency）

1. 不流暢，沒有明顯的語法，無法理解
2. 勉強流暢，有幾個語法錯誤
3. 流暢性佳，有幾個語法錯誤，可以理解
4. 流暢度極佳；很少有語法錯誤
5. 完美流暢；就像讀報紙一樣

充分性（Adequacy）

1. 傳達了原文訊息量的 0%
2. 傳達了原文訊息量的 25%
3. 傳達了原文訊息量的 50%
4. 傳達了原文訊息量的 75%
5. 傳達了原文訊息量的 100%

意義（Meaning）

1. 和原文的意思完全不一樣
2. 與原始資訊相比添加/省略了誤導性訊息
3. 與原文部分意義相同
4. 和原文的意思幾乎一樣
5. 與原文意思相同

嚴重性（Severity）

1. 對病人有危險
2. 以某種方式損害照顧
3. 延遲必要的照顧
4. 對病患照顧的影響尚不清楚
5. 不影響患者照顧

除此之外，評估者還需要選擇他們更喜歡 GT 翻譯的句子還是專業翻譯的句子。與專業翻譯相比，GT 翻譯的句子的流暢度得分顯著較低（3.4 vs 4.7，$P < 0.001$），但充分性得分（4.2 vs 4.5，$P = 0.19$）和意義得分（4.5 vs 4.8，$P = 0.29$）相似。GT 翻譯的句子和專業翻譯相比更可能出現錯誤（39% vs 22%，$P = 0.05$），但從統計角度看，出現嚴重錯誤的可能性並不高（4% vs 2%，$P = 0.61$）。另外評

doi:10.6342/NTU202404301

估者更喜歡複雜句子的專業翻譯，但簡單句子則由 GT 勝出。

Chen et al. [9] 爲了評估 GT 網站將健康資訊從英文翻譯成西班牙文以及英文翻譯成中文時的準確性，從先一本針對糖尿病患者的健康教育手冊中選取 6 個句子，再請兩位專業翻譯人員將其從英文分別翻譯成西班牙文和中文。接著招募 6 位經過認證的翻譯人員（3 位西班牙語和 3 位中文）對以下版本進行盲評估：(1) GT 翻譯的句子，(2) 專業人工翻譯翻譯的句子。評估以參考 Jeffrey et al. [15] 的標準進行評分以獲取數據。得到結果後，作者先使用 Flesch-Kincaid 可讀性測驗來檢測六個原始英語句子的 Flesch-Kincaid 等級，具有較簡單詞彙的較短句子獲得較低分數，而包含更多醫學術語的較長句子獲得較高分數。此研究中句子的可讀性範圍爲 2.8 至 9.0（平均值 5.4，SD 2.7），且所有 GT 翻譯的句子的 Flesch-Kincaid 等級和翻譯準確性之間的相關係數都是負的，這表示 GT 爲簡單的句子提供了準確的翻譯。然而，當原始英語句子需要更高年級的程度才能理解時，翻譯錯誤的可能性就會增加。與 GT 相比，中文人工翻譯的更爲準確，而在西班牙文上人工翻譯並沒有提供明顯地比 GT 表現的更好。

爲了評估 GT 在常見英語醫療用語的準確性和實用性，Patil and Davies [19] 選擇了 10 個醫學短語，並將其用 GT 分別翻譯成 26 種語言後（8 種西歐語言、5 種東歐語言、11 種亞洲語言和 2 種非洲語言），再發送給每種語言的母語人士，讓他們將這些句子翻譯回英語。最後再對母語人士翻譯後的短語與原始短語進行比較並評估其含義。評估分爲正確和錯誤，若有輕微的語法錯誤是可以接受被分爲正確的，但若翻譯沒有意義或是不正確，則被視爲錯誤。在 260 個翻譯短語中，150 個（57.7%）正確，110 個（42.3%）錯誤。非洲語言得分最低（45%），其次是亞洲語言（46%），東歐語言次之，爲 62%，西歐語言最準確，爲 74%。斯瓦希里語得分最低，只有 10%，而葡萄牙語得分最高，達 90%。

Marcelo et al. [20] 亦發現了 GT 展現了機器偏見（machine bias）的現象，即訓練後的統計模型逐漸反映出一些有爭議的社會不對稱現象，例如性別或種族偏見。作者從美國勞工統計局 (Bureau of Labor Statistics, BLS) 取得一份完整的職位列表，並以 12 種不同的性別中立語言（例如匈牙利語、中文、約魯巴語和其他幾種語言）來造出” 他/她是工程師” 等的句子（其中” 工程師” 被替換爲感興趣的工作職位。接著使用 GT 將這些句子翻譯成英語，並收集翻譯輸出中女性、男性和中性代名詞出現頻率的統計資料。實驗結果表明，GT 表現出強烈的男性預設傾向，特別是在通常性別分佈不平衡或刻板印象相關的領域，例如 STEM（科學、

doi:10.6342/NTU202404301

技術、工程和數學）的工作。作者將這些統計數據與勞工統計局的數據進行對比後發現 GT 確實無法反映女性工人的眞實分佈。

爲了瞭解機器翻譯系統在翻譯文學作品時的表現，Agung et al. [2] 選擇了印尼短篇小說《Cerita-Cerita Jakarta》做爲研究資料來源。作者以 GT 和 DeepL 這兩個神經機器翻譯系統將小說翻譯後，使用了以下框架來將翻譯錯誤進行分類並評估：遺漏概念 (Omitted Concept) ：原文中的一個概念沒有傳達或翻譯到目標文本中。

1. 新增概念 (Added Concept) ：在目標文本中引入源文本中沒有的新訊息，從而可能改變意義。
2. 未翻譯概念 (Untranslated Concept) ：在目標文本中直接借用源語言詞彙，給目標讀者造成潛在的理解困難。
3. 誤譯概念 (Mistranslated Concep)：目標文本中的概念在特定情境中意義不正確。
4. 替代概念 (Substituted Concept)：在目標文本中使用一個並非直接對等但在語境中有效的概念。
5. 明示概念 (Explicitated Concept)：目標文本在不引入新細節的情況下明示了原文中隱含的訊息。分析結果顯示 GT 和 DeepL 在翻譯過程中產生的翻譯錯誤包括未翻譯概念、遺漏概念和誤譯概念，表示他們在翻譯文化術語、擬聲詞、縮寫、慣用表達、俚語和地址術語時仍面臨挑戰。

爲了研究機器翻譯系統處理阿拉伯語詞彙歧義的表現，Aldawsari [6] 選擇 16 個包含不同類型詞彙歧義的阿拉伯語句子，經過 GT 和 SYSTRAN（最早的線上機器翻譯系統之一）翻譯成英文後請四位評估人員對 (1) 英文翻譯的可理解性與 (2) 和阿拉伯文源文本相比，其翻譯的準確性進行評分。結果顯示，GT 和 SYSTRAN 都難以將模稜兩可的阿拉伯語單字翻譯成英文，但 GT 似乎更準確。GT 的可懂度得分爲 32.5%，準確度爲 24.75%；SYSTRAN 的可懂度得分爲 27.75%，準確度僅 11.25%。Google 翻譯在可懂度和準確度方面均優於 SYSTRAN，因此可以說 GT 產生的句子比 SYSTRAN 稍微更容易理解，但準確得多。

doi:10.6342/NTU202404301

第三章　資料選取與處理

有鑒於我們所瞭解到的論文大多將短句或短語作爲研究資料對象，所以在這篇論文中，我們決定擴展評估 GT 翻譯表現的樣本，包括文句長度及內容的複雜程度，因此我們選用了論文做爲我們的實驗樣本。論文的領域涵蓋數學、物理學、經濟學、生物學、電腦科學、資料科學、建築學、臨床醫學等。而因爲我們要將翻譯後的論文上傳至 Turnitin 以獲得相似度指數，所以我們所搜集的論文必須要在 Turnitin 的資料庫內，比對所得的結果才有其意義。因此我們去瞭解了 Turnitin 資料庫的具體範疇，並從中選擇了 ProQuest Dissertations & Theses (PQDT) Global、IEEE、Springer Nature 和 ScienceDirect，再從這些來源中選擇 100 篇英文文章，使用 GT 進行指定語言的翻譯後，再使用 Turnitin 來獲得相似度，最後再使用該數據來進行分析。

3.1　資料處理

生成原始文本：將原始文章的文字部分單獨取出。若文字內有特殊字符無法複製者（如 σ、Δ 等），爲保持句意完整不被破壞，將完整句子捨棄。值得一提的是，如同 Ethan et al. (2012) 所提到的，我們原本是將整份論文以 pdf 檔的形式上傳至 GT 以獲取譯本，但發現這樣所獲取的譯本會有很多錯誤及疏漏，如圖3.1及圖3.2中黃色部份所示，其原因在於 GT 在閱讀檔案時可能會受到圖片或無法讀取的字符影響。因此我們最後採取的方式是親手選取每個句子或段落，以確保其翻譯的完整。

doi:10.6342/NTU202404301

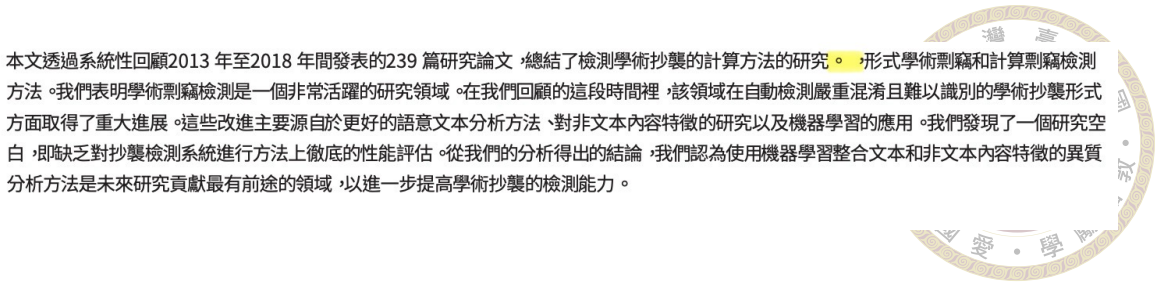

本文透過系統性回顧2013 年至2018 年間發表的239 篇研究論文，總結了檢測學術抄襲的計算方法的研究。，形式學術剽竊和計算剽竊檢測方法。我們表明學術剽竊檢測是一個非常活躍的研究領域。在我們回顧的這段時間裡，該領域在自動檢測嚴重混淆且難以識別的學術抄襲形式方面取得了重大進展。這些改進主要源自於更好的語意文本分析方法、對非文本內容特徵的研究以及機器學習的應用。我們發現了一個研究空白，即缺乏對抄襲檢測系統進行方法上徹底的性能評估。從我們的分析得出的結論，我們認為使用機器學習整合文本和非文本內容特徵的異質分析方法是未來研究貢獻最有前途的領域，以進一步提高學術抄襲的檢測能力。

Figure 3.1: 將論文以 pdf 檔上傳至 GT 可能獲得的錯誤結果

本文通過系統回顧 2013 年至 2018 年間發表的 239 篇研究論文，總結了檢測學術抄襲的計算方法的研究。為了構建研究貢獻的展示，我們提出了用於抄襲預防和檢測工作的新穎的技術導向類型，形式 學術剽竊和計算剽竊檢測方法。 我們表明學術剽竊檢測是一個非常活躍的研究領域。 在我們回顧的這段時間裡，該領域在自動檢測嚴重混淆且難以識別的學術抄襲形式方面取得了重大進展。 這些改進主要源於更好的語義文本分析方法、對非文本內容特徵的研究以及機器學習的應用。 我們發現了一個研究空白，即缺乏對抄襲檢測系統進行方法上徹底的性能評估。 從我們的分析得出的結論，我們認為使用機器學習整合文本和非文本內容特徵的異構分析方法是未來研究貢獻最有希望的領域，以進一步提高學術抄襲的檢測能力。

Figure 3.2: 從論文中只選取文字後以 GT 翻譯

3.2 資料生成

1. 翻譯原始文本：使用 GT 將原始文本進行指定語言間的翻譯，最後得到英文譯本。除了中文以外，我選擇了幾個在台大論文內較有可能被使用或參考的語言，分別是日文（日本語文系所）、德文和法文（領域專長）。以下使用代號來表示分別表示對應的語言：

E（English）：英文

C（Chinese）：中文

F（French）：法文

G（German）：德文

J（Japanese）：日文

爲了瞭解不同次數翻譯後對文本的影響，會對原始文本進行一層語言翻譯及二層語言翻譯。一層語言翻譯會將原始英文文本分別翻譯至中文、法文、德文、日文後再翻回英文，如圖3.3；二層語言翻譯翻譯則是將原始英文文本先分別翻譯至中文、法文、德文、日文，再翻譯至除了該語言的其他語言，最後翻回英文，如圖3.4。

doi:10.6342/NTU202404301

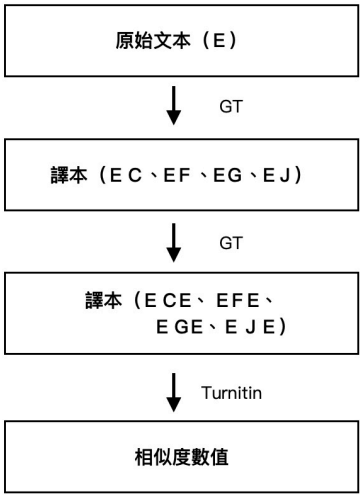


Figure 3.3: 一層語言翻譯流程圖

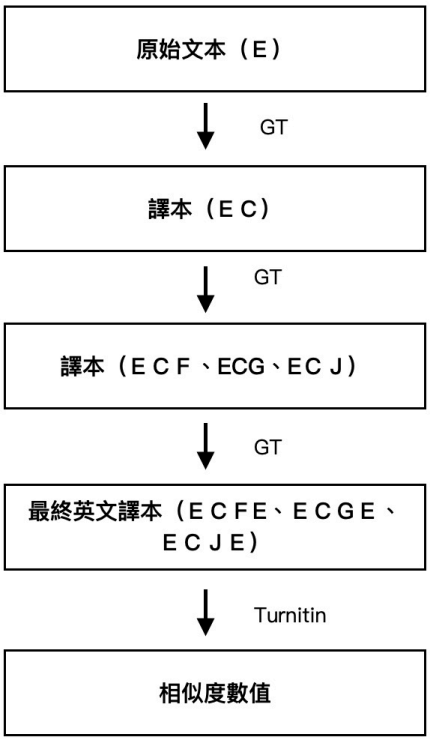


Figure 3.4: 雙層語言翻譯流程圖
（以中文爲第一層翻譯爲例）

2. 獲得 Turnitin 相似度報告：將所得的 1600 篇最終英文譯本上傳至 Turnitin 相似度比對系統並獲得相似度報告。值得一提的是，Turntin 相似度比對系統在前三次的使用只需要等待大約半個小時就可以得到相似度報告，但若超過三次以上則需要等 24 小時才會產生新的相似度報告。而每個人在 Turnitin 系統內可上傳的點也非常有限（如本系所提供的上傳點的數量爲 2），因此非常感謝其他朋友們願意讓我共用他們的 Turinitin 帳號，才讓我有辦法在有限的時間內搜集到完整數量的相似度報告。

3.3 報告結果與數值選取

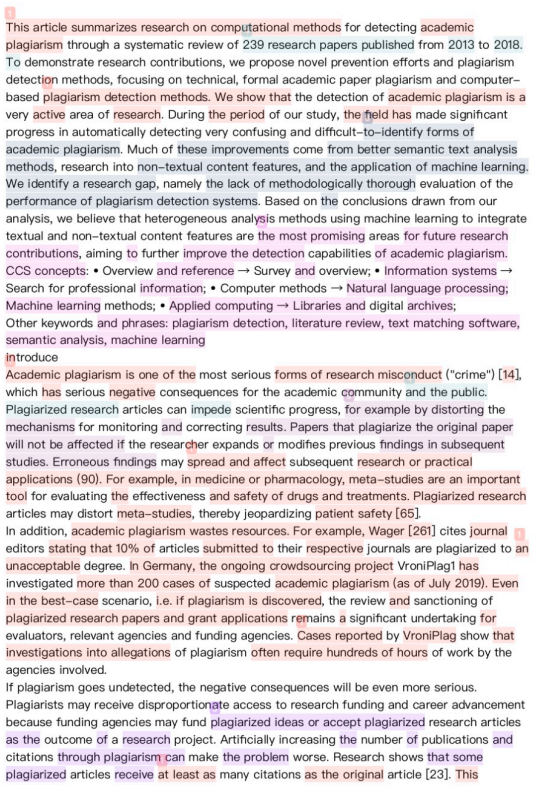
1. 相似度報告解讀：在系統顯示比對完成後，我們可以在開啓報告前簡單瞭解此份檔案的相似度結果。如圖3.5所示，每份檔案會有其對應的相似度數值，而顏色則分別代表上傳檔案與 Turnitin 資料庫收錄來源相似的程度，藍色代表相似度指數爲 0%，綠色代表相似指數介於 0% 至 24%，黃色代表相似度指數介於 25%

doi:10.6342/NTU202404301

至 49%，橘色代表相似度指數介於 50% 至 74%，紅色則代表相似度指數介於 75% 至 100%。

Figure 3.5: 相似度報告顏色意義

對相似度有大致了解後，接下來則是相似度報告的具體內容。由圖3.6可以看到，Turnitin 會將檢測到的相似處以不同顏色標出，其顏色對應到圖3.7的來源，且在圖3.7中亦提供該來源的相似百分比。而報告中相似度指數的定義則是文章內疑似與來源相似的文字比例，其中一個英文單字算一個字，非一個英文字母；中文字則是一個中文字算一個字。


This article summarizes research on computational methods for detecting academic plagiarism through a systematic review of 239 research papers published from 2013 to 2018. To demonstrate research contributions, we propose novel prevention efforts and plagiarism detection methods, focusing on technical, formal academic paper plagiarism and computer-based plagiarism detection methods. We show that the detection of academic plagiarism is a very active area of research. During the period of our study, the field has made significant progress in automatically detecting very confusing and difficult-to-identify forms of academic plagiarism. Much of these improvements come from better semantic text analysis methods, research into non-textual content features, and the application of machine learning. We identify a research gap, namely the lack of methodologically thorough evaluation of the performance of plagiarism detection systems. Based on the conclusions drawn from our analysis, we believe that heterogeneous analysis methods using machine learning to integrate textual and non-textual content features are the most promising areas for future research contributions, aiming to further improve the detection capabilities of academic plagiarism.
CCS concepts: • Overview and reference → Survey and overview; • Information systems → Search for professional information; • Computer methods → Natural language processing; Machine learning methods; • Applied computing → Libraries and digital archives;
Other keywords and phrases: plagiarism detection, literature review, text matching software, semantic analysis, machine learning
Introduce
Academic plagiarism is one of the most serious forms of research misconduct ("crime") [14], which has serious negative consequences for the academic community and the public. Plagiarized research articles can impede scientific progress, for example by distorting the mechanisms for monitoring and correcting results. Papers that plagiarize the original paper will not be affected if the researcher expands or modifies previous findings in subsequent studies. Erroneous findings may spread and affect subsequent research or practical applications (90). For example, in medicine or pharmacology, meta-studies are an important tool for evaluating the effectiveness and safety of drugs and treatments. Plagiarized research articles may distort meta-studies, thereby jeopardizing patient safety [65].
In addition, academic plagiarism wastes resources. For example, Wager [261] cites journal editors stating that 10% of articles submitted to their respective journals are plagiarized to an unacceptable degree. In Germany, the ongoing crowdsourcing project VroniPlag1 has investigated more than 200 cases of suspected academic plagiarism (as of July 2019). Even in the best-case scenario, i.e. if plagiarism is discovered, the review and sanctioning of plagiarized research papers and grant applications remains a significant undertaking for evaluators, relevant agencies and funding agencies. Cases reported by VroniPlag show that investigations into allegations of plagiarism often require hundreds of hours of work by the agencies involved.
If plagiarism goes undetected, the negative consequences will be even more serious. Plagiarists may receive disproportionate access to research funding and career advancement because funding agencies may fund plagiarized ideas or accept plagiarized research articles as the outcome of a research project. Artificially increasing the number of publications and citations through plagiarism can make the problem worse. Research shows that some plagiarized articles receive at least as many citations as the original article [23]. This

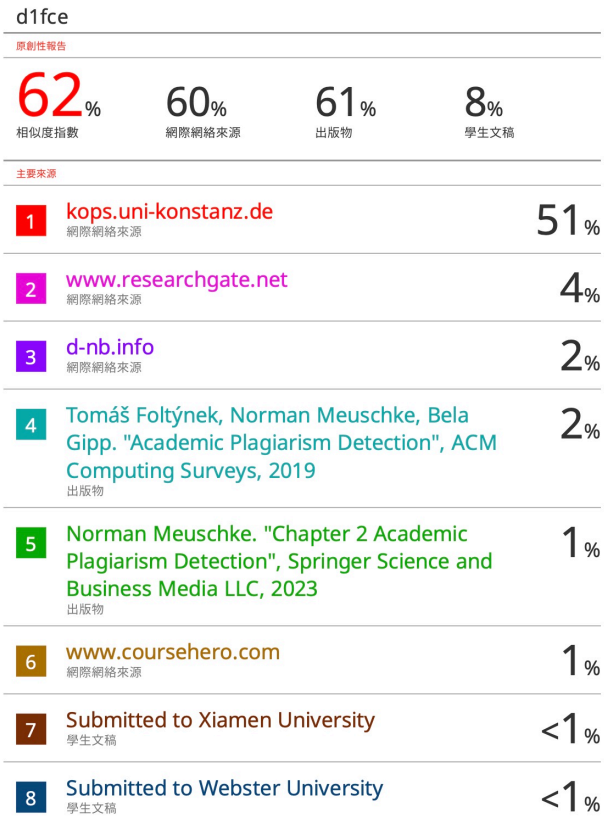


Figure 3.6: 相似度報告範例

Figure 3.7: 相似度報告範例（相似度指數）

接下來我們以圖3.7爲例，分兩部份來解釋報告，第一部分爲主要來源的定義。假設原文章內容總字數有 1000 字，我們可以看到相似度指數爲 62%，意思是

doi:10.6342/NTU202404301

原文內容 1000 字中，有 62%(假設是 620 字) 是疑似跟其他出處來源有相似地方，而第一個來源的數值是 51%，即第 1 個相似來源文字有 510 字和原文一樣；將原文 1000 字扣除第 1 個來源的 510 字，剩下無重複的 490 字後，再和第 2 個相似來源比對, 有 40 字和原文一樣，佔了 4%；接著，將原文 1000 字扣除第 1 個和第 2 個來源的 (510+40=590) 字，剩下無重複的 410 字後，再和第 3 個相似來源比對, 有 20 字和原文一樣，佔了 2%，以此類推。最後，將各相似來源百分比加總起來：(51%+4%+2%+2%+1%+1%+1%)=62% 即是總相似百分比，不會超過 100%。值得注意的是，在此範例中，雖然第一個來源是我們的原始文章，但此數值並不是我們要選取的和原文章的相似度（準確度），原因是系統會根據段落找最佳化的相似來源，且段落來源只會歸屬最相似的來源，正如圖3.6所示，第一來源中會夾雜其他來源，使得第一來源的百分比降低。

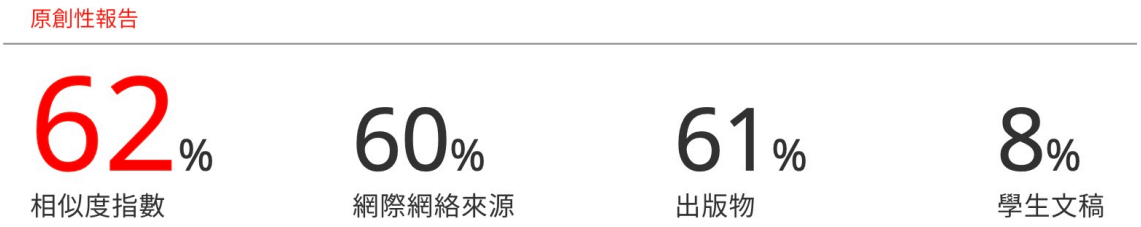


Figure 3.8: 相似度報告範例（來源）

第二部分爲不同來源的定義，如圖3.8所示，報告顯示相似總百分比爲 62%，共有三個主要相似來源，分別是 60% 網際網路來源、61% 出版物、8% 學生文稿。值得注意的是不能將各來源型態百分比加總 (60%+61%+8%=129%)，因爲學生文稿可能引用相同的網頁資訊，或網頁資訊包含該出版品的內容，所以同一個段落文字有可能彼此重疊，因此相似總百分比並不會等於各項數值總和。

在瞭解 Turnitin 的運作方式後，我們將最終英文譯本和原始文本的翻譯表現指數定義爲

$$\text{翻譯表現指數} = \max\{\text{網際網路來源數值}, \text{出版物數值}\}。 \tag{3.1}$$

在本次研究中，我們選取了 100 篇論文作爲原始文本，翻譯語言爲中文、法文、德文和日文 4 種，則一層語言翻譯有 4 種，而二層語言翻譯有 12 種。因此每一份原始文本會有 16 種不同翻譯方式的最終英文譯本。我們都用式 (3.1) 來得到

doi:10.6342/NTU202404301

最終英文譯本和原始文本的翻譯表現指數，作爲我們的分析資料。

doi:10.6342/NTU202404301

第四章　資料分析

4.1　資料部分展示

值得一提的是，Turnitin 所產生的相似度數值皆爲整數，表4.1爲部分展示，而圖4.1爲原始文本的字數分佈圖。

	ECE	ECJE	ECFE	ECGE	EJE	EJCE	EJFE	EJGE	EFE	EFCE	EFJE	EFGE	EGE	EGCE	EGFE	EGJE
D1	71	60	65	65	60	57	54	55	65	61	55	61	64	59	57	55
D2	70	56	67	19	59	57	57	22	68	20	17	14	20	20	16	17
D3	72	58	67	66	63	53	55	57	68	62	57	63	66	56	60	57
D4	72	54	63	64	58	53	54	54	64	60	54	61	66	61	58	53
D5	67	50	63	60	50	47	48	46	61	56	45	55	61	54	52	48
⋮																
D100	5	3	5	1	3	3	4	3	4	5	5	3	5	5	5	3

Table 4.1: 資料部分展示

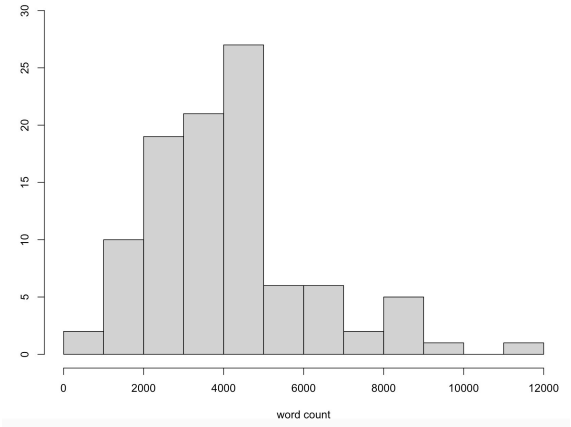


Figure 4.1: 相似度報告顏色意義

doi:10.6342/NTU202404301

4.2 資料分析

4.2.1 離群值

在整理資料時發現有兩篇原始文本的數據特別讓人注意，分別是 D2 和 D100，見表4.1。相較於其他文本的翻譯表現指數的一致性，D2 有 7 種最終英文譯本的翻譯表現指數超過 55%，而剩下譯本的翻譯表現指數卻都小於 22%，這是在其他文本上都沒有的現象，如圖4.2所示。而去查閱他們的相似度報告會發現其內容有些許相似的地方，但 Turnitin 卻並未將其標示，經過重測也如此，因此認定 Turnitn 在此篇的檢測上異常，將其當成離群值處理。

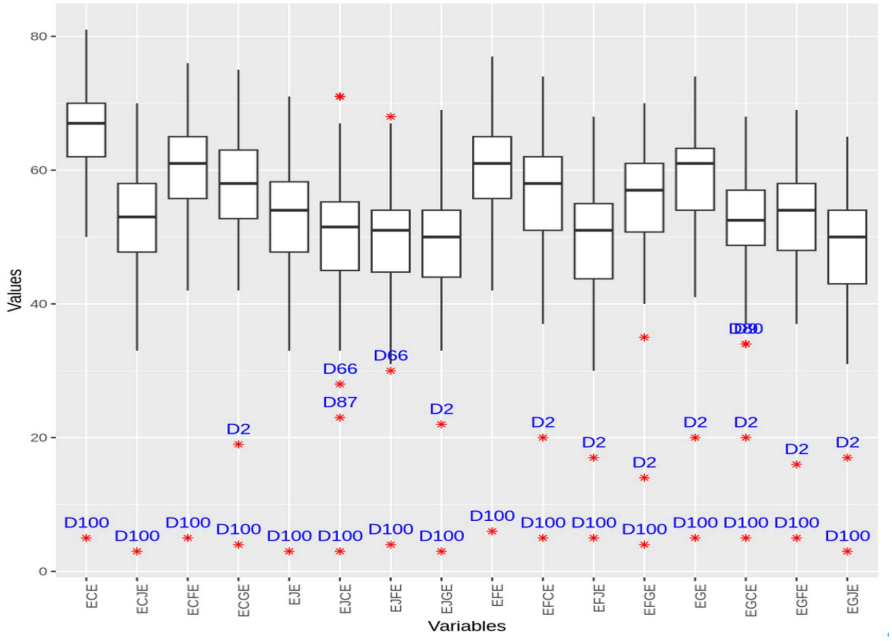


Figure 4.2: 離群值分佈（x 軸：翻譯方式，y 軸：翻譯表現指數）

至於 D100 的結果則是非常特別，我們可以觀察到其所有的最終英文譯本的翻譯表現指數皆小於 5%。由於該原始文章來自於 arXiv，我們推測因爲其尚未發表，所以不在 Turnitin 比對系統的資料庫內，使得翻譯表現指數異常的小。

爲了確認此類特殊資料的一致性，我們再另外找了 5 篇也不存在於 Turnitin 比對系統資料庫內的論文，發現其表現和 D100 一致，翻譯表現指數都小於 10%，如表4.2所示。因爲 D100 的原始文本具有不在 Turnitin 資料庫的特殊性，可以將 D100 視爲一個控制組，讓我們瞭解若是所上傳的文章沒有比對到和其高度相似的文章的話，其數據結果會是如何表現。綜上所述，最後決定將 D2 及 D100 當作離群值，分析時並不將其納入。

doi:10.6342/NTU202404301

	ECE	ECJE	ECFE	ECGE	EJE	EJCE	EJFE	EJGE	EFE	EFCE	EFJE	EFGE	EGE	EGCE	EGFE	EGJE
Out1	3	5	3	7	5	4	5	4	7	6	6	8	8	5	7	6
Out2	5	6	6	6	6	4	5	6	7	5	7	7	10	1	9	5
Out3	4	3	3	4	4	3	7	4	5	3	4	4	5	4	5	4
Out4	6	6	14	9	9	5	7	7	11	6	10	11	10	6	7	10
Out5	5	3	4	5	4	3	4	4	6	4	3	5	5	4	5	3

Table 4.2: 原文若不在 Turnitin 資料庫內之資料展示

4.2.2 箱形圖

圖4.3以第一個翻譯的語言爲組，可以看出在第一個翻譯語言固定下，翻譯兩層的翻譯表現指數確實較低。另外可以注意到日文（Ｊ）組整體分佈較低，在其他組中有涉及日文的翻譯表現指數分布也是最低的。

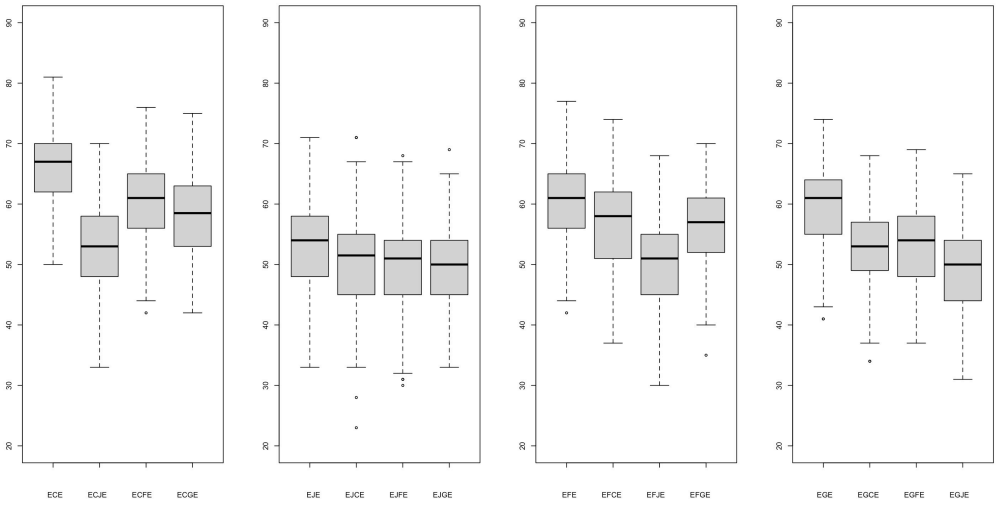


Figure 4.3: 箱型圖（以第一個翻譯的語言爲組）

doi:10.6342/NTU202404301

4.2.3 密度圖

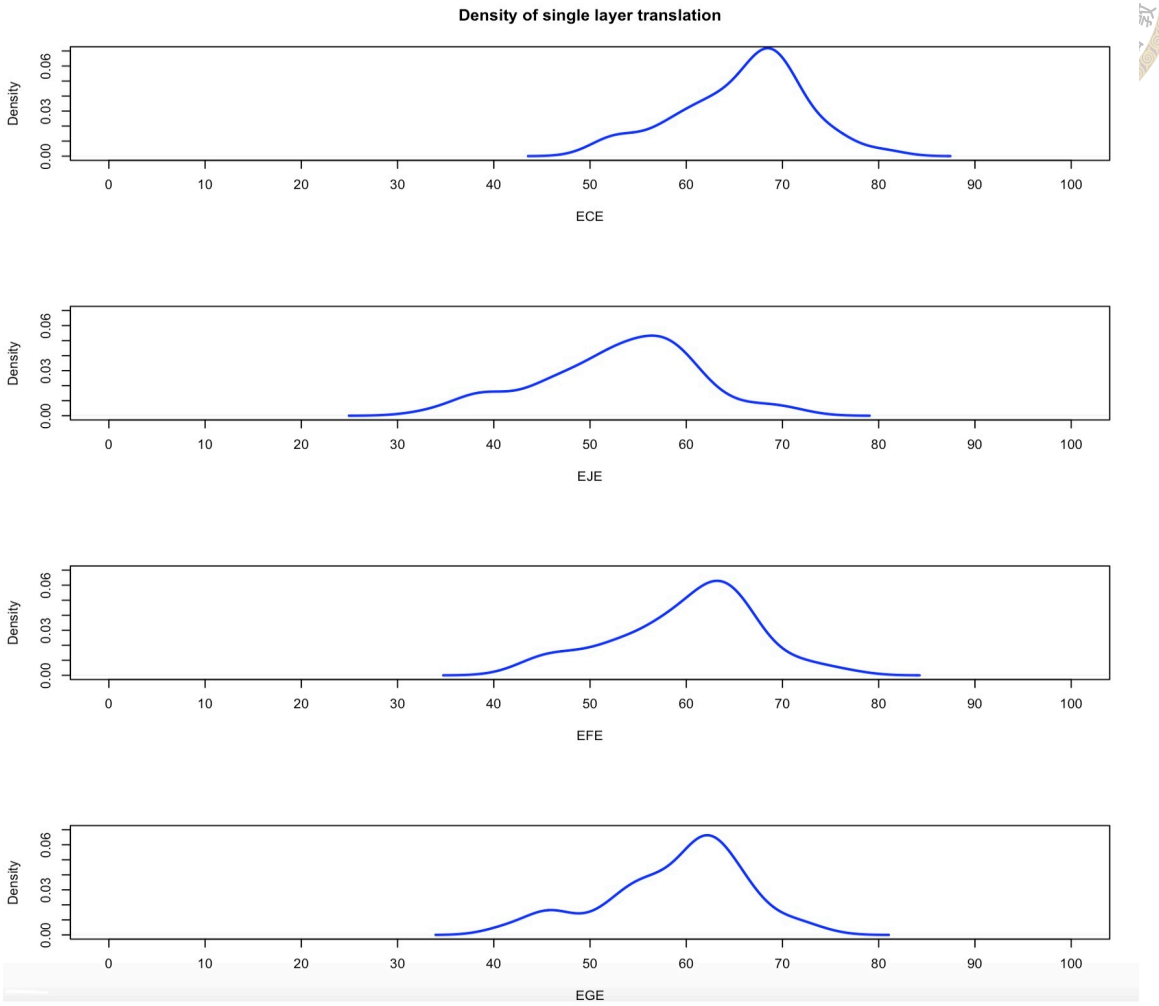


Figure 4.4: 單層翻譯密度圖

從圖4.4可觀察到大部分的語言翻譯表現指數皆爲單峰，而在單層翻譯中 EJE 的圖形高峰表現也是最靠左的，而 ECE 的高峰則是較其他語言靠右。對於日文的翻譯表現指數整體偏低的表現，我們猜測是否是因爲日文在語系上單獨屬於日本—琉球語系，其常被分類爲孤立語言；而英文和德文同屬日爾曼語系，法文屬於義大利語系，但此二語系都隸屬於印歐語系。因此涉及日文的翻譯時或許語意上會較大程度地被破壞，且不易還原，才使得涉及日文翻譯的表現指數整體偏低。但值得注意的是，中文亦是單獨屬於漢藏語系，但他的表現指數卻並沒有特別差，反而在單層翻譯中其表現還與和英文同語系的法文和德文至少持平，其原因可能在於隨著中國成爲全球僅次於美國的第二大經濟體，中文翻譯的使用需求也日益漸增，因此 GT 也投入大量資源來提升其翻譯表現指數。

doi:10.6342/NTU202404301

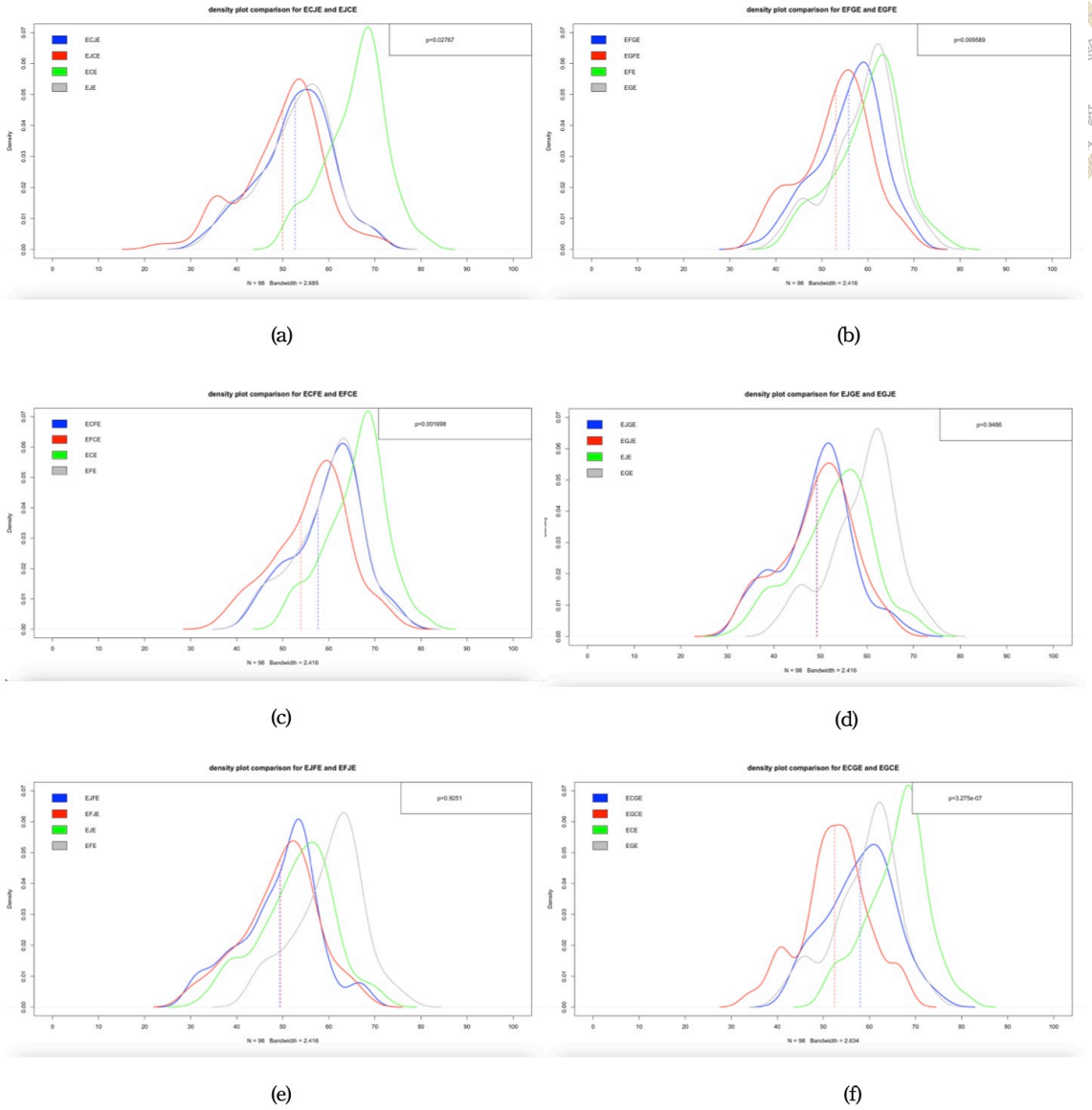
在確認涉及日文的翻譯對於整體翻譯表現指數的影響前，我們先瞭解二層翻譯中若是第一語言和第二語言翻譯順序交換（如 EFGE 和 EGFE），則翻譯結果是否有差。從最後的譯本來比較會發現兩者確實高度相似，但當然還是有些許地方不盡相同。我們透過 t 檢定來檢測兩組間平均值是否有顯著差異，其結果如表4.3所示，可以得知在 $\alpha = 0.05$ 的情況下只有 EJFE 與 EFJE 和 EJGE 與 EGJE 的平均值沒有顯著差異，其餘組別皆有顯著差異，括弧內則是平均值較高的組別。而在有涉及中文翻譯的組別中（ECFE 與 EFCE、ECGE 與 EGCE 和 ECJE 與 EJCE），我們可以觀察到先翻譯中文的小組平均值都較高，這也和我們前面的猜測相呼應，先翻譯至中文和先翻譯至其他語言相比可以更好地保留語意，使得最後整體的相似度較高。

T test	P value
ECFE & EFCE (ECFE)	0.001698
ECGE & EGCE (ECGE)	$3.275e{-07}$
ECJE & EJCE (ECJE)	0.02767
EFGE & EGFE (EFGE)	0.009589
EJFE & EFJE	0.9251
EJGE & EGJE	0.9486

Table 4.3: T 檢定結果（$\alpha = 0.05$）

在瞭解不同組別間的數值差異後，我們分別將上述組別和對應的單層翻譯進行密度圖的比較，如圖4.5所示。在圖4.5中展示了每一種翻譯方式的分佈，而每一組圖則讓我們可以更確切地瞭解到和其相關的分布差異。進行雙層翻譯後的分佈（藍色及紅色）都會比進行單層翻譯的分佈靠左，或至少重疊。其中 ECE 的分佈在所有翻譯方式中最靠右且最爲集中，再次呼應我們前述的猜測。而在圖 13(e) 和圖 13(d) 中我們可以觀察到他們的分佈非常類似，表示法文和德文在翻譯表現上是相差無幾的。

doi:10.6342/NTU202404301



(a) (b)

(c) (d)

(e) (f)

Figure 4.5: 翻譯順序交換的組別及其對應的單層翻譯之密度圖比較）

我們猜測第一個翻譯的語言的翻譯表現指數或許會較大程度的影響最後的翻譯表現指數結果，例如第一個翻譯的語言是日文的話，因爲語意及結構可能就先較大幅度破壞，使得最終翻譯表現指數也會較低；但若是第一個翻譯的語言是中文的話，其語意能被較好地保留，使最後翻譯表現指數較高。但若是觀察圖 12(b)後會發現，EFGE 的平均值是比 EGFE 高的，但雖然英文、德文和法文都隸屬於印歐語系，但英文和德文同屬日爾曼語系，法文卻屬於義大利語系，英文和德文在語系上的關係是較近的，照我們猜測應是先翻譯德文的表現較佳，結果卻不然。

doi:10.6342/NTU202404301

4.2.4 集群分析

而後我們進行資料探索，決定透過群聚分析來瞭解不同的分群因子可否使我們獲得有意義的分群結果。我們將 98 篇以相近的領域分成 4 類，括弧內爲領域的文章數：

1. Applied Math, Al, CS (53)
2. Econometrics, Economy (11)
3. Biology, Clinic, Medical (26)
4. Architecture, Geography (8)

接著採用階層式群聚方法，以不同的聚合方式來確認最終分類結果會否和論文的領域有所關聯。結果如圖4.6至圖4.10所示，並未觀察到相關的趨勢。

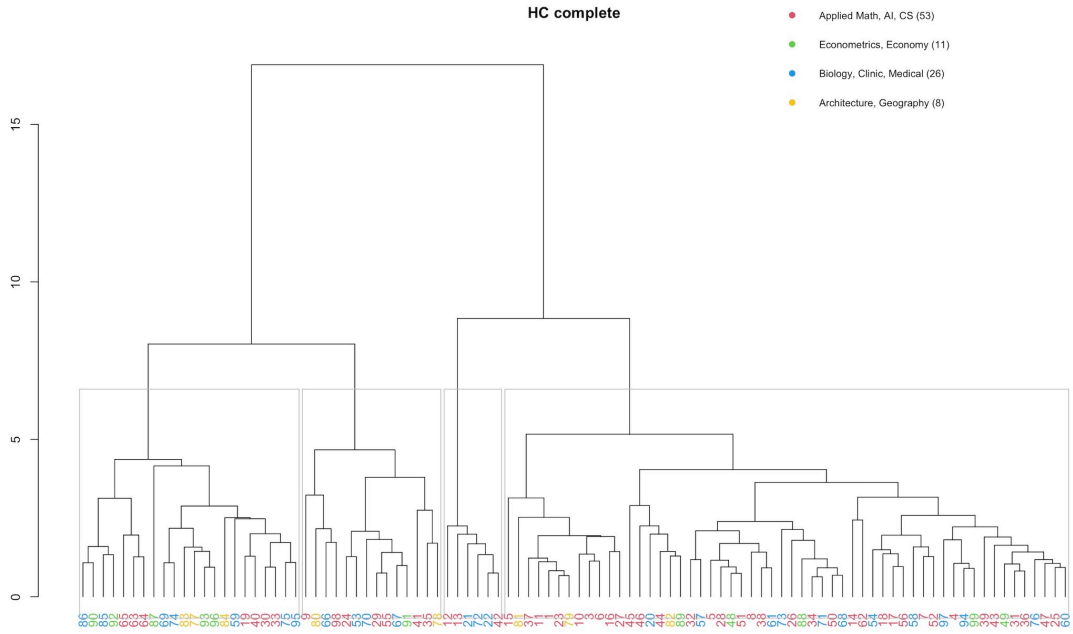


Figure 4.6: 以完整法聚合，顏色爲不同領域

doi:10.6342/NTU202404301

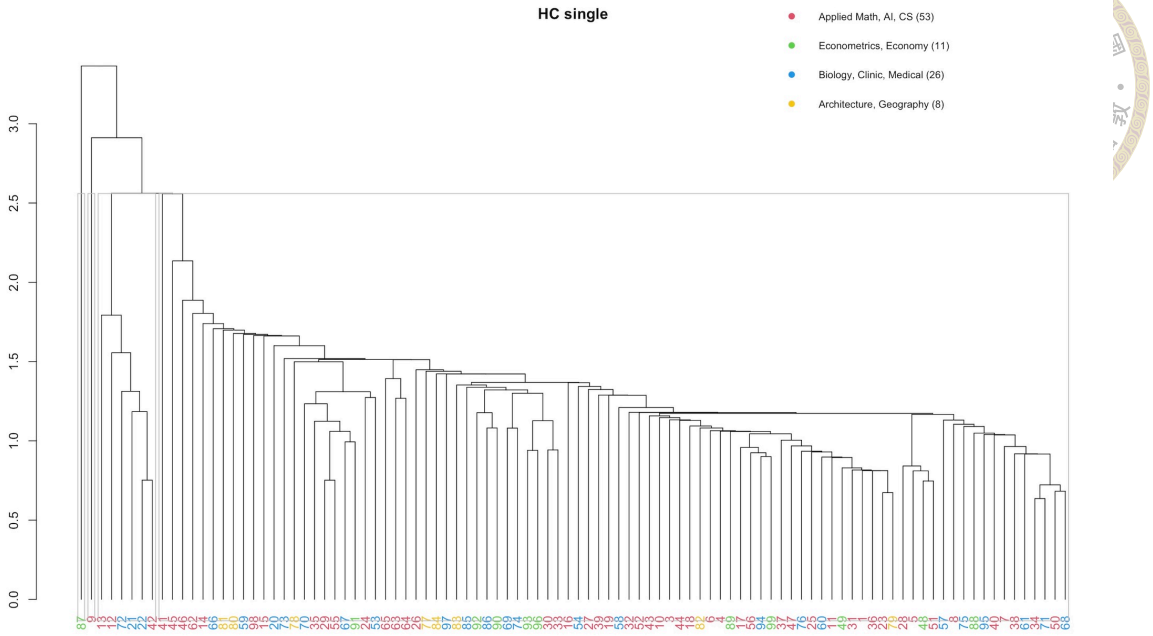


Figure 4.7: 以單一法聚合，顏色爲不同領域

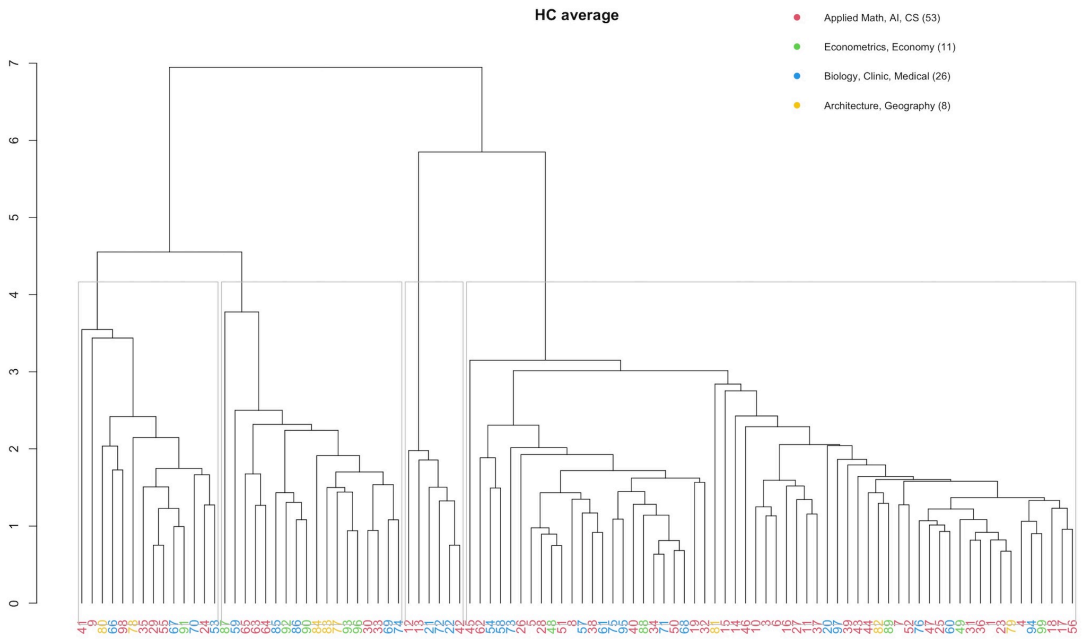


Figure 4.8: 以平均法聚合，顏色爲不同領域

doi:10.6342/NTU202404301

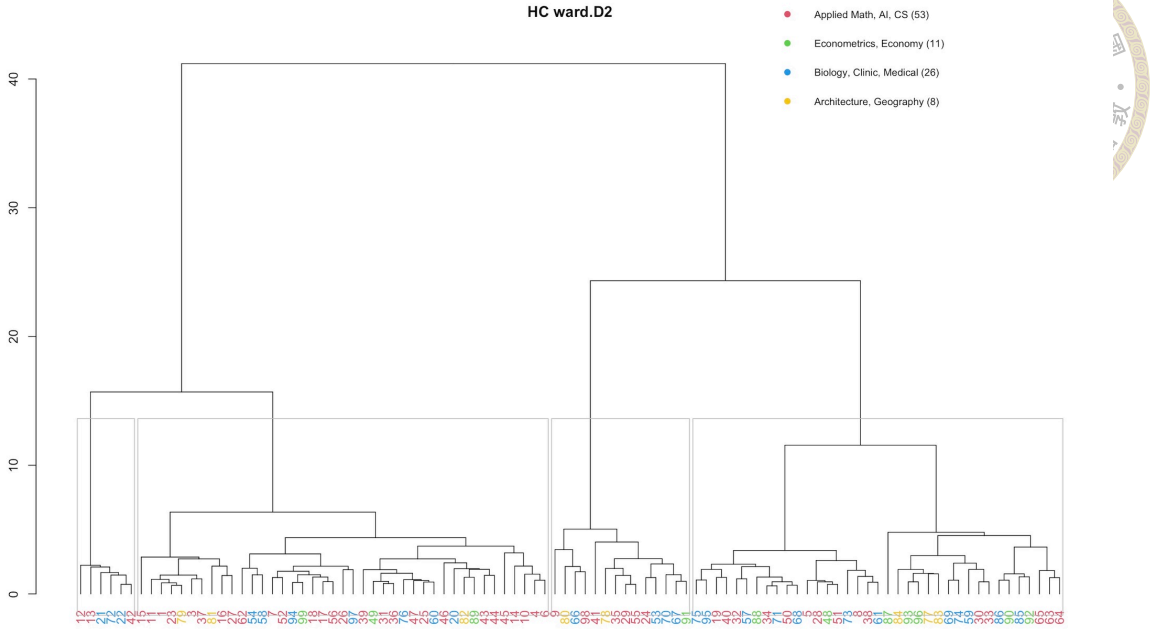


Figure 4.9: 以 Ward 法聚合，顏色爲不同領域

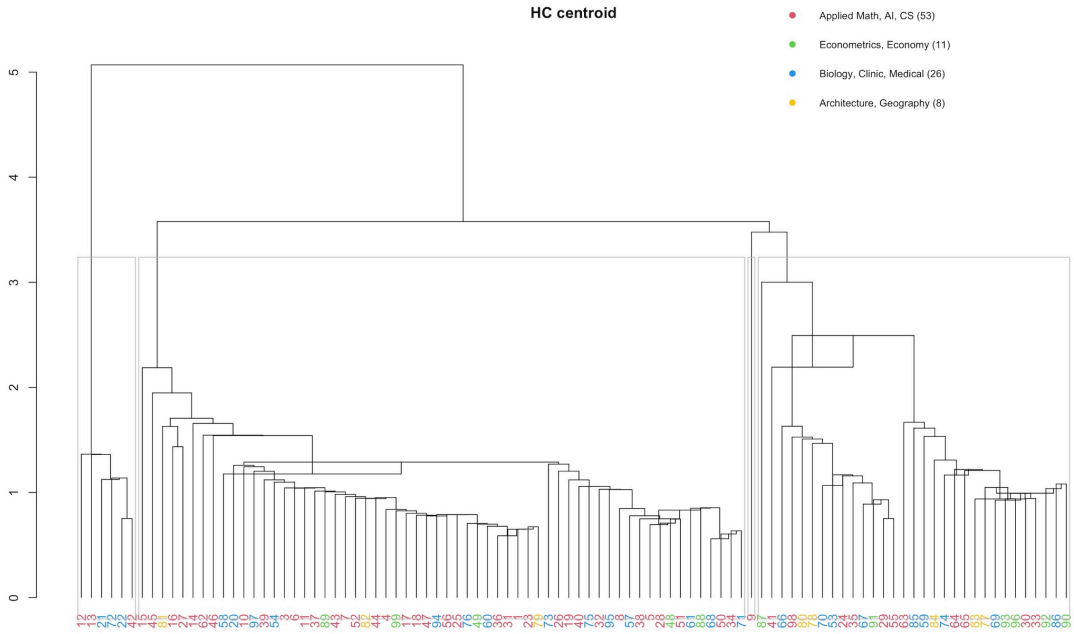


Figure 4.10: 以中心法聚合，顏色爲不同領域

doi:10.6342/NTU202404301

另外我們又觀察到，在 98 篇論文中有 27 篇來自阿拉伯出版的刊物，因此我們也想確認翻譯表現指數會不會和出版物的國家有關，一樣對其做了階層式群聚分析後，並未得到我們認爲有意義的結果，如圖4.11至圖4.15所示。

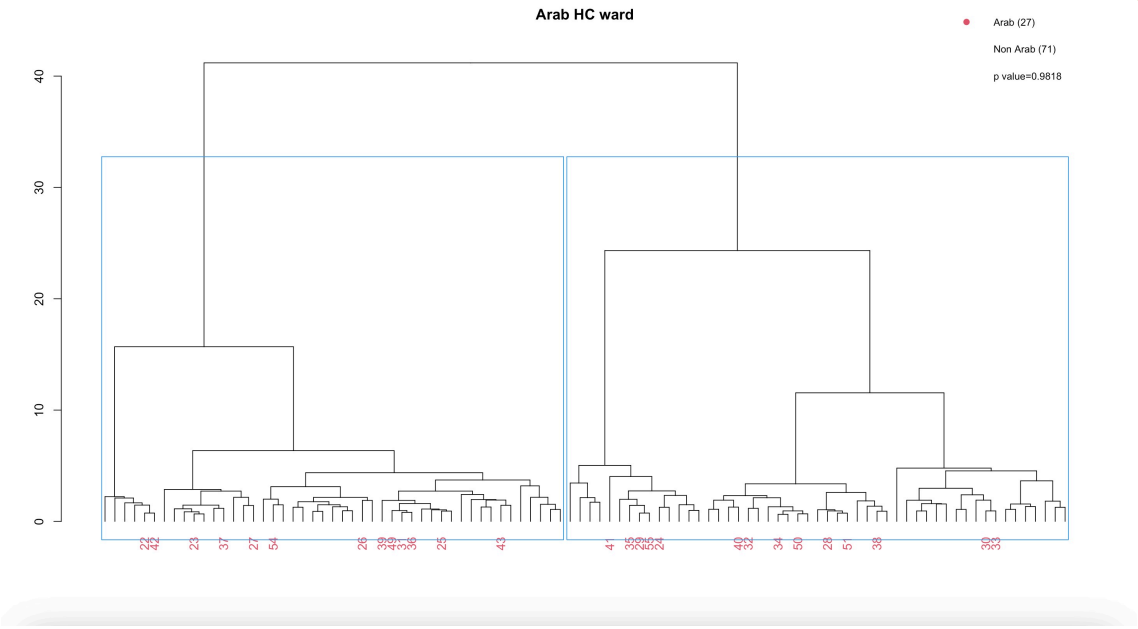


Figure 4.11: 以 Ward 法聚合，只標出阿拉伯出版刊物

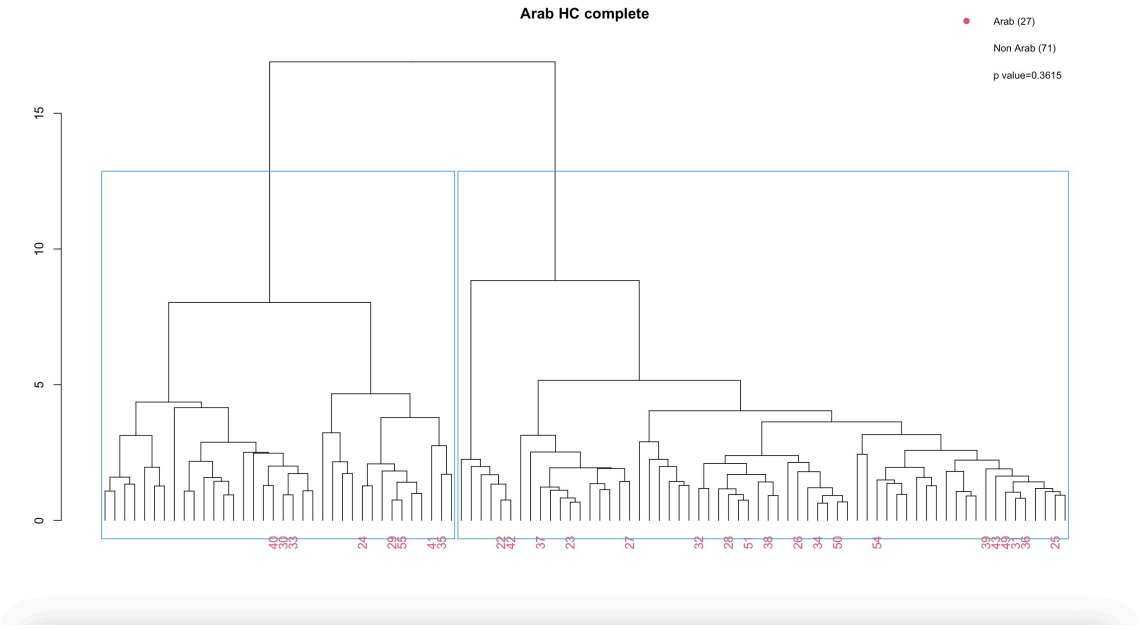


Figure 4.12: 以完整法聚合，只標出阿拉伯出版刊物

doi:10.6342/NTU202404301

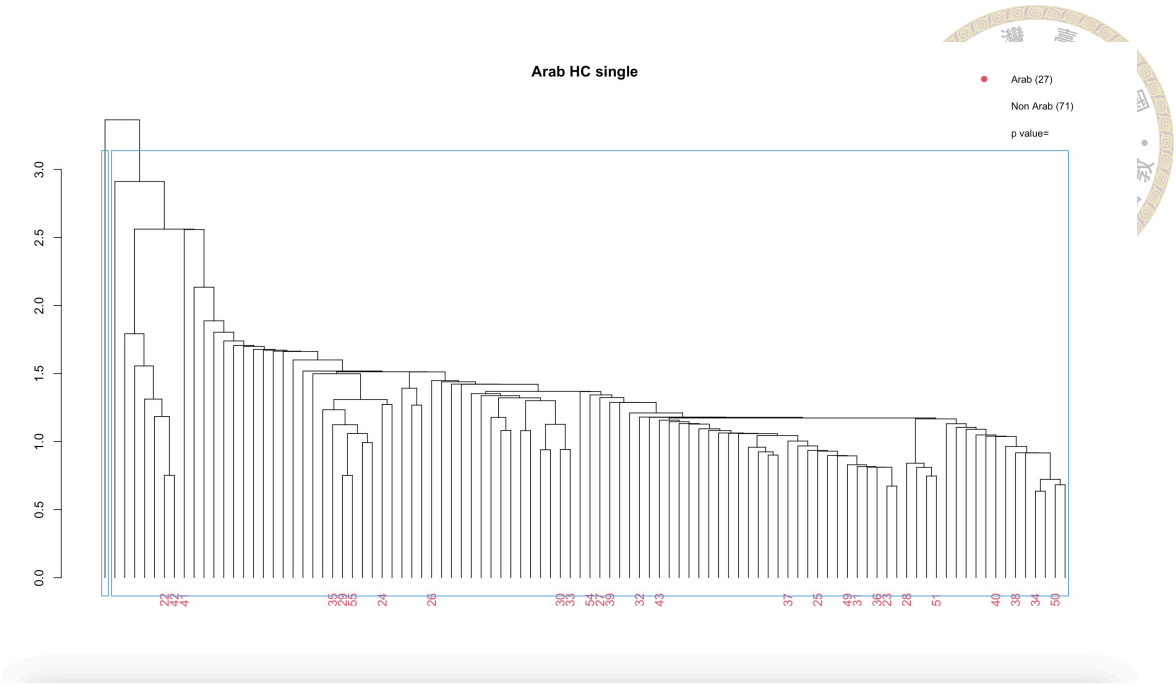


Figure 4.13: 以單一法聚合，只標出阿拉伯出版刊物

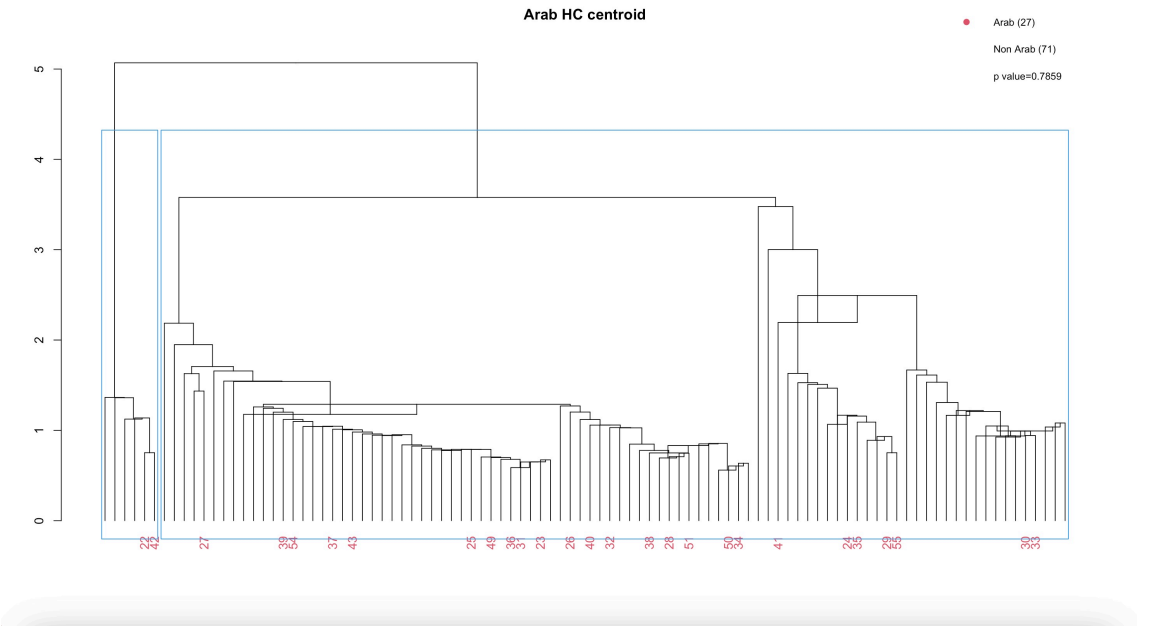


Figure 4.14: 以中心法聚合，只標出阿拉伯出版刊物

doi:10.6342/NTU202404301

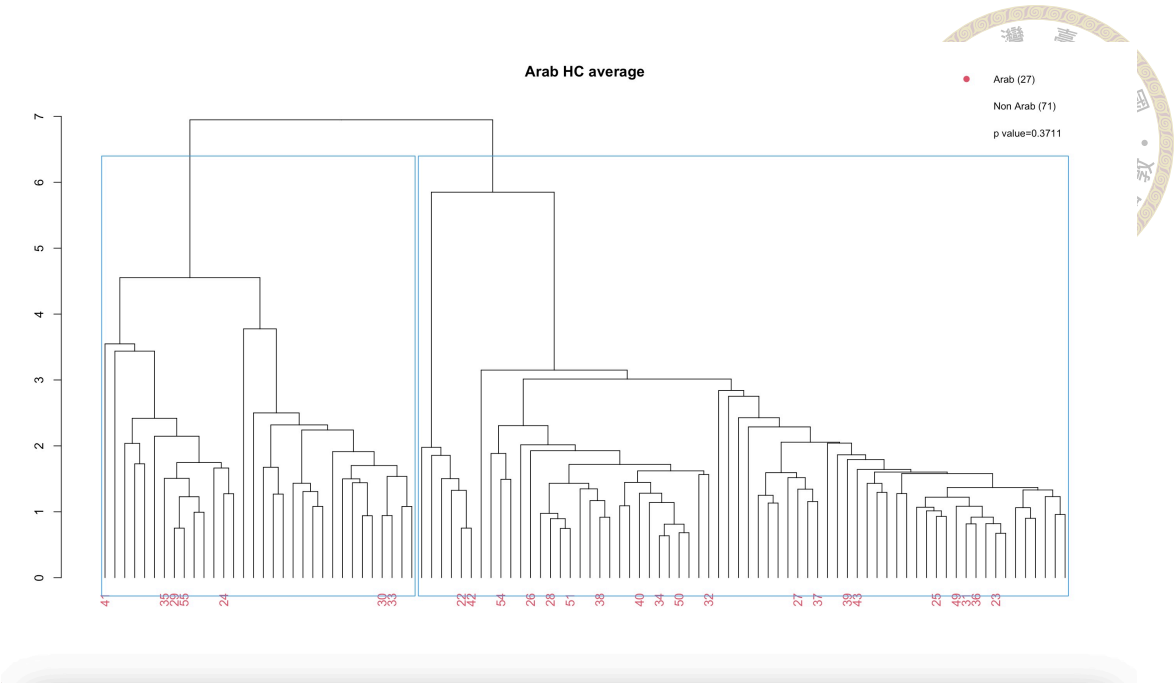


Figure 4.15: 以平均法聚合，只標出阿拉伯出版刊物

最後，考慮到 Google 翻譯針對同一內容的翻譯結果並不會每次都相同，我們想確認此種隨機性對我們的翻譯表現指數是否有影響，因此我們從所蒐集到的論文中隨機抽取 5 篇後，再將其重新翻譯，結果顯示其平均並未有顯著差異。

doi:10.6342/NTU202404301

第五章　結論

在本研究中，我們深入探討了翻譯表現的評估方法，特別針對 Google 翻譯在不同語言之間的表現進行了詳細分析。我們選取了 100 篇英文論文，通過單層和雙層翻譯的方式將其翻譯成中文、日文、法文和德文，並利用 Turnitin 相似度比對系系統來測量翻譯後的相似度，從而評估翻譯的準確性。

研究結果顯示，雙層翻譯的翻譯表現指數明顯低於單層翻譯，這可能是因爲在雙層翻譯過程中，訊息經過多次轉換，導致累積的誤差增多。因此，在實際應用中，若是爲了確保較高的翻譯表現指數，應優先考慮單層翻譯的方式。

此外，我們發現翻譯的語言對論文相似度具有顯著影響。特別是日文的翻譯結果表現較差，而中文的翻譯表現則相對較佳。這一發現與我們的預期一致，因爲日文的語法和句子結構與英文有較大差異，導致翻譯過程中訊息丟失和誤譯的概率較高，而中文的表現我們則是猜測因爲其使用需求持續增加，使得 Google 翻譯投入大量資源來提升其準確度。

本研究的結論對於學術界具有意義。根據現行規定，臺灣大學對於論文相似度的通過標準爲 20%，我們的數據結果表明若是採取翻譯抄襲的方法來撰寫論文，是沒有辦法通過 Turnitin 的檢測。換言之，我們可以相信 Turnitin 在檢測翻譯抄襲的能力。

然而，本研究也存在一些局限性。首先，我們只選取了 100 篇英文論文，樣本數量相對有限，未來可以擴大樣本範圍以獲得更具代表性的結果。其次，本研究僅考慮了四種語言的翻譯，未來可以增加更多語言，進行更廣泛的比較分析。

總結來說，本研究通過對論文進行不同翻譯方式和語言的詳細分析，揭示了翻譯表現的影響因素及 Turnitin 系統的檢測表現。

doi:10.6342/NTU202404301

參考文獻

[1] 清　人日本留　生　言語文化接触: 相互誤解　日中教育文化交流. 書房, 2010.

[2] I. G. A. M. Agung, P. Budiartha, and N. Suryani. Translation performance of google translate and deepl in translating indonesian short stories into english. 01 2024.

[3] M. Aiken. An updated evaluation of google translate accuracy. Studies in Linguistics and Literature, 3:p253, 07 2019.

[4] M. W. Aiken. An analysis of google translate accuracy. 2012.

[5] M. W. Aiken, M. Park, L. L. Simmons, and T. Lindblom. Automatic translation in multilingual electronic meetings. 2010.

[6] H. Aldawsari. Comparing the performance of google translate and systran on arabic lexical ambiguity. Arab World English Journal For Translation and Literary Studies, 7:19–34, 08 2023.

[7] R. Anil, A. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. Clark, L. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, and K. Robinson. Palm 2 technical report, 05 2023.

[8] E. M. Balk, M. Chung, N. Hadar, K. Patel, W. W. Yu, T. A. Trikalinos, and L. K. W. Chang. Accuracy of data extraction of non-english language trials with google translate. 2012.

[9] X. Chen, S. Acosta, and A. Barry. Evaluating the accuracy of google translate for diabetes education material. JMIR Diabetes, 1:e3, 06 2016.

doi:10.6342/NTU202404301

[10] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways. J. Mach. Learn. Res., 24:240:1–240:113, 2023.

[11] D. Coughlin. Correlating automated and human assessments of machine translation quality. Proceedings of MT Summit IX, 01 2001.

[12] C. Culy and S. Riehemann. The limits of n-gram translation evaluation metrics. 11 2003.

[13] T. Fishman. “we know it when we see it” is not good enough: toward a standard definition of plagiarism that transcends theft, fraud, and copyright. 2009.

[14] T. Foltýnek, N. Meuschke, and B. Gipp. Academic plagiarism detection: A systematic literature review. ACM Comput. Surv., 52(6), oct 2019.

[15] J. Jackson, A. Kuriyama, A. Anton, A. Choi, J. Fournier, A. Geier, F. Jacquerioz, D. Kogan, C. Scholcoff, and R. Sun. The accuracy of google translate for abstracting data from non-english-language trials for systematic reviews. Annals of internal medicine, 171(9):677–679, Nov. 2019.

[16] R. R. Khanna, L. S. Karliner, M. Eck, E. Vittinghoff, C. J. Koenig, and M. C. Fang. Performance of an online translation tool when applied to patient educational material. Journal of Hospital Medicine, 6(9):519—525, Oct 2011.

[17] G.-Z. Liu. 防治學生英文寫作抄襲之認知與方法探究：以二所研究型大學爲例. 英語教學期刊 (THCI Core), 36(4), Dec. 2012. THCI Core.

doi:10.6342/NTU202404301

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[19] S. Patil and P. Davies. Use of google translate in medical communication: Evaluation of accuracy. BMJ: British medical journal, 349:g7392, 12 2014.

[20] M. Prates, P. Avelar, and L. Lamb. Assessing gender bias in machine translation: a case study with google translate. Neural Computing and Applications, 32, 05 2020.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, Dec 2017.

doi:10.6342/NTU202404301