國立臺灣大學管理學院資訊管理學研究所

碩士論文

Department of Information Management

College of Management

National Taiwan University

Master Thesis

以中文新聞報導為主題之文本分析與改寫生成

Document Analysis and Paraphrase Generation for Chinese News Reports

陳恩穎

En-Ying Chen

指導教授: 莊裕澤 博士

Advisor: Yuh-Jzer Joung, Ph.D.

中華民國 112 年 8 月

August, 2023

doi:10.6342/NTU202303745

致謝

首先，非常感謝莊裕澤教授在研究所兩年的悉心指導。自碩一起，每週和莊老師的會議中，老師會用他自身的經驗給予我們非常寶貴的建議，讓我們了解到自己的不足之處，打下扎實的研究基礎，藉此不斷進步。不論是在報告論文亦或是呈現自己的研究成果時，莊老師最重視的不外乎就是邏輯和證據：任何研究方法都必須有非常清晰、能說服人的邏輯，而研究中的每個論點都必須有過往的文獻為證，如此才能確立自己研究的整體價值。莊老師在短短兩年內帶給我們的諄諄教誨，讓我得到了許多課堂與業界都難以獲取的知識，真的非常謝謝莊老師。

此外，我也非常感謝實驗室的同學晨瑋、俊易以及承翰，在每週的會議前後我們都會互相討論老師提出的問題，時常互相鼓勵，也因為大家一起對研究室的付出，才有良好的環境可以學習。祝福大家畢業後鵬程萬里，一切順利。

最後，必須特別感謝我的親人以及朋友一路以來的支持與鼓勵，讓我在遇到疫情的研究所可以安心完成學業，因為有你們的關心及協助才能讓我無後顧之憂的努力。祝福所有人都能得到屬於自己最美好的未來。

陳恩穎 謹誌

國立臺灣大學資訊管理研究所

中華民國一百一十二年七月

doi:10.6342/NTU202303745

摘要

近年來網路以及通訊裝置以驚人的速度發展，同時也帶動了社群媒體的蓬勃發展，網路新聞成為民眾獲取新知的主要媒介。而在這個新媒體時代，對於媒體從業人員來說，若希望自己能在各個社群平台的激烈競爭下都保有一席之地，勢必要能夠創造出多樣化版本的新聞，以滿足不同平台的觀眾。此時，若有一個系統能將現有新聞快速產生出改寫版本，產生出一篇新的報導，勢必能為媒體從業人員提供很大的幫助，由此可知改寫系統對新聞領域有相當重要且急迫的需求。

因此，本論文設計了一個針對新聞的文本改寫系統，能夠讓使用者輸入一篇原始新聞之後，透過模型得到一篇文章結構、用字相異的改寫版本新聞，希望藉由這樣的模型來協助媒體從業人員，並為整體新聞產業做出一定程度的貢獻。

本論文除了使用強大的預訓練模型 GPT-2，並使用了經過整理的 TaPaCo 以及 PAWS-X 資料集進行實驗，並結合規則以及語序重構模型，藉此讓模型可以產生出以篇章為單位且有明顯結構不同的改寫新聞文章，而非只是逐句的改寫或字詞置換。在最後自動評估以及人工評估兩種評估方式上，本論文所提出的方法對於新聞的改寫程度明顯優於 Baseline Model 的表現，說明了在以篇章為單位的改寫當中，本論文加入的語序重構技術相較於單純的逐句改寫可以獲得更好的效果，也更貼近人類既定印象中的改寫。

關鍵字：新聞、改寫生成、語序重構、預訓練模型、深度學習

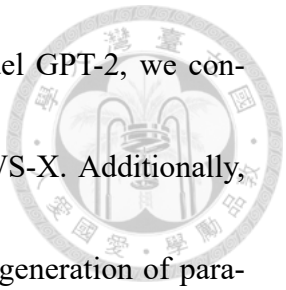
doi:10.6342/NTU202303745

Abstract

Recent years have witnessed astonishing advancements in internet and communication devices, which have also fueled the thriving growth of social media. Online news has become the primary medium for people to acquire new information. In this era of new media, media professionals strive to maintain a presence across various social platforms amidst fierce competition. It is imperative for them to create diverse versions of news to cater to different audiences on each platform. In such a scenario, a system capable of rapidly generating paraphrased versions of existing news, creating new reports, would undoubtedly provide significant assistance to media professionals. Hence, it is evident that a paraphrasing system is highly important and urgently needed in the field of news.

Consequently, this thesis presents a text paraphrasing system specifically designed for news articles. This system allows users to input an original news article and obtain a paraphrased version with different sentence structures and vocabulary through a model. The aim is to assist media professionals and contribute to the overall news industry.

doi:10.6342/NTU202303745



In this thesis, besides employing the powerful pre-trained model GPT-2, we conducted experiments using curated datasets such as TaPaCo and PAWS-X. Additionally, we incorporated rules and sentence reordering models to enable the generation of paraphrased news articles at the paragraph level, ensuring distinct structural differences rather than simple sentence rewrites or word replacements. In both automatic and manual evaluations, the proposed method outperformed the Baseline Model in terms of the extent of news paraphrasing. This indicates that the sentence reordering technique introduced in this thesis yields better results compared to merely rewriting sentences and aligns more closely with human perceptions of paraphrasing.

Keywords: News, Paraphrase generation, Sentence order reconstruction, Pre-trained model, Deep learning

doi:10.6342/NTU202303745

目錄

doi:10.6342/NTU202303745

doi:10.6342/NTU202303745

doi:10.6342/NTU202303745

圖目錄

doi:10.6342/NTU202303745

表目錄

doi:10.6342/NTU202303745

第一章　緒論

1.1　研究背景與動機

新聞的發展可以追溯到數百年前的時代，當時的新聞為了傳播技術和貿易資訊而迅速地發展，除了定期收集與傳播信息的功能之外，由於牽動著商人的經濟命脈，因此新聞一直是相當受到重視且和生活密不可分的一環。十八世紀隨著印刷術的蓬勃發展，新聞從原先的口耳相傳轉變為以報紙為主的傳播方式，而媒體記者這個行業也逐漸蓬勃，成為採訪名人、時事並撰寫新聞報導，供社會大眾掌握社會脈動的重要媒介。

二十世紀以來，網際網路的飛速發展劇烈改變了人類的生活習性，其中 2008 年起因智慧型手機的爆炸性成長徹底改變了這個世代獲取資訊的方式，網路新聞因應這樣的趨勢逐漸成為民眾掌握時事的主要媒介。特別在近年來各式各樣的社群媒體如雨後春筍般不斷出現，這樣的環境造成媒體從業者面臨越發重大的挑戰，除了過往的採訪、整理、查證以及編寫等耗時的工作之外，為了在各大社群平台都能有新聞發布，勢必需要有多樣化版本的新聞來滿足不同平台的閱眾，如此才能在這個極其競爭、人人都可以做為資訊傳播者的世代脫穎而出，博得更多的關注。

由上述背景可知，網際網路與社群媒體的發達無疑為這個世代創造出了更多

doi:10.6342/NTU202303745

采多姿的生活環境，然而這樣的環境對於媒體從業人員而言所需背負的壓力也更加龐大，除了需要趕著跑新聞並快速發表之外，創造出多樣化版本的新聞以滿足閱眾是這個世代記者額外需要達到的目標。有鑑於此，透過現今蓬勃發展的自然語言模型（Natural Language Model）協助便是一個良好的解決辦法。自然語言模型的發展是近年人工智慧當中非常具有代表性的熱門領域之一，其中自然語言處理與自然語言生成更是其中的重要核心，而自然語言處理領域中最重要也最廣為應用的預訓練語言模型（Pre-trained Language Model）則會是不可或缺的工具。預訓練語言模型乃是先運用非常巨量的文本來訓練一個模型，使其學習到人類日常通用的語言表示方式，爾後只需要相對小量的語料或資訊便可以讓該模型適用在特定的領域或下游任務（Downstream Task）有一定程度的表現，最重要的是可以作為人類的助手。以前述的新聞領域而言，媒體從業人員若能有語言模型的幫助，將現有的新聞或是自己過去發表過的新聞直接進行改寫，省去每次都要重新撰文的時間並同時獲得一定品質的改寫新聞，就能得到一篇嶄新的報導，相信這樣的流程無疑會對媒體從業人員帶來巨大的幫助。然而，在中文改寫領域的研究仍處在萌芽階段，而就我們截至目前的查閱結果，尚未有以中文新聞為主題的改寫任務研究。

因此，本論文將以中文新聞改寫為主題，結合許多技術與流程盡可能地建立一套改寫系統，達到為中文新聞改寫的任務，透過模型改寫的語句闡述與原新聞相同的客觀事實或數據，藉此創造出多樣化版本的新聞，使媒體從業者可以發表於各個不同的社群平台，減緩媒體從業者的壓力，為新聞產業做出貢獻。

doi:10.6342/NTU202303745

1.2 研究目的

本研究主要目的旨在將中文新聞報導透過自然語言模型進行內容改寫，由於新聞事件的客觀事實與資料來源應是開放且沒有受著作權、版權等規範的，然而當這些客觀的新聞事件被記者匯集並撰寫成新聞稿之後，該篇新聞稿就會受到著作權的規範。現在由於移動裝置與網路媒體的爆炸性的成長，新聞報導的型式及來源都來自四面八方，甚至原先身為閱眾的人們也都開始可以自由地在網路媒體或社群網站上撰寫與發表屬於自己觀點的新聞事件，因此多樣化版本的新聞報導對這個世代的記者顯得更為重要，必須在不同平台有不同版本的新聞供閱眾參考，才更有機會在這個極為競爭的新媒體世代存活，然而這對於媒體從業者來說是極大的挑戰。

有鑑於上述問題，本研究希望設計一個新聞報導改寫系統，這個系統並不僅限於和過往多數改寫研究 [21, 26] 一樣以句子為單位（Sentence Level）的字詞置換改寫，而是以整篇新聞報導為單位（Paragraph Level）的文本重新建構。本系統改寫的方式會先將一篇原始新聞報導輸入，首先將新聞翻譯為英文版本，接著先經過我們根據英語語言學所分析歸納出的規則保留住有明顯前後關係（Before-After Relationship）的相鄰語句，而後進入到整篇新聞的語句順序重構（Reordering），確保改寫前後的新聞是有整體結構上的改變。最後，重構後的新聞會以句為單位送入以 GPT-2（Generative Pre-trained Transformer 2）為核心之自建語句改寫模型，最後經過翻譯後得到中文的改寫版本新聞。在這個改寫系統中之所以會需要翻譯成英文主要是因為現今中文領域改寫尚未出現具指標性的資料集，而改寫資料的蒐集成本極高，品質也難以保證，因此我們決定採用英文資料集進行訓練。此外，在預訓練模型之所以會選擇 GPT-2 而非更高版本的 GPT-3 或 GPT-4 主要是因為第四代 GPT 目前尚未開放給一般研究人員使用，而目前硬體設備的發展也尚不足

doi:10.6342/NTU202303745

以支撐第三代 GPT 被廣為訓練及使用，一般等級的運算資源仍無法負荷在短時間內大量訓練及執行參數調整等動作，因此，我們最終選擇了 GPT-2 做為我們主要的預訓練模型。本研究結合了在過去改寫研究當中最常見的回譯法（Backward Translation）以及以句為單位的改寫（Sentence-Based Paraphrase），外加上句序重構的技術，目的在進行新聞改寫的過程中除了可以確保新聞當中想被傳達的客觀資訊與新知，包含人、事、時、地、物等等能夠被完整的保留，並同時做到對新聞改寫的最主要目的：為媒體從業者帶來協助，創造出多樣版本的新聞，確保不同平台的閱眾都可以被滿足。

1.3 論文架構

本篇論文將於第二章探討過往文本改寫議題中可學習之處與尚待精進之處，並且因應研究方法會同時探討過去關於自然語言處理、文本改寫、語序重構與新聞生成等領域的相關文獻，此外也會討論文本改寫任務應如何評估，以及使用預訓練模型來進行文本生成，在第三章則會描述整體實驗架構、資料集與實驗模型，在第四章說明實驗結果以及與其他模型之比較，並於最後的第五章說明結論以及有關本研究的未來展望。

doi:10.6342/NTU202303745

第二章　文獻探討

本論文旨在應用現今蓬勃發展的預訓練語言模型 (Language Pre-trained Model) 於中文文本生成，建立出以中文新聞報導為主題之文本改寫的下游任務，並在研究方法中結合許多技術包含語序保留的規則建立（Ruled-Based）、語序重構（Sentence Reordering）藉以提升改寫新聞的效果。在本章節主要目的是整理並探討與本研究密切相關的的文獻資料與類似領域之研究發展。以下將根據本研究需要探討的技術分為六個小節。

第一小節會從自然語言處理、自然語言生成為起始，首先會探討與自然語言處理領域密切相關的預訓練模型技術，接著向下探討其下游任務，並逐漸收斂至第二小節，也是本研究之主要探討領域: 文本改寫之應用，並在本小節點出現有研究之不足及可精進之方向，做為本研究價值之佐證。

再者，由於研究方法所採用技術之需求，因此在第三、四小節會分別整理語序重構與新聞文本生成之文獻與發展現況，評估兩者和文本改寫議題能夠如何做結合，最後綜合上述歸納出可協助提升文本改寫品質的模型與潛在方法。接著在第五小節將列出和文本改寫有關的模型評估方式，探討現有文獻中常見的自動評估與人工評估兩種方式之優劣，分析出最適合用於本研究的評估方式，提供用於評估本研究結果之參考。最後，於第六小節總結本研究所探討的文獻，整理出結論與可嘗試的方法。

doi:10.6342/NTU202303745

2.1 自然語言處理與預訓練模型

自然語言處理（Natural Language Processing）在人工智慧的議題受到眾人注目以來便是一個極為熱門的領域。語言是人類發展文化的基石，世界各地存在著各式各樣不同的語言，透過自己獨有的字詞、符號來傳達資訊，因此各國的學者無不投入大量的心力與資源進行自然語言領域的研究。然而人類的語言是複雜度極高並存在龐大非結構資料的集合，對於電腦而言要處理甚至理解這樣的訊息存在相當大的挑戰性。

近年來隨著技術與運算資源的迅速演進，電腦對於自然語言處理領域的進步也達到了過往難以想像的地步。時至今日，自然語言處理領域又可以被細分為自然語言理解（Natural Language Understanding, NLU）以及自然語言生成（Natural Language Generation, NLG）兩個子領域，電腦首要任務必須先透過自然語言理解來讀懂自然語言的結構，了解語言當中想表達的意思，而後再透過自然語言生成創造出人類希望電腦協助產生的資訊，成為人類工作或生活的助手。在日常生活中存在著許多希望透過自然語言處理協助完成的任務，典型的問題包含了機器翻譯、聊天機器人以及文本生成等等，這些任務個別都存在著大量的文獻以及研究發展，由此可知運用電腦來對自然語言領域做處理已經是相當具有歷史的一個研究領域。

近年來由於硬體設備上限不斷突破，電腦運算速度進步到可以負荷巨大的運算量，使得需要巨量參數以及文本建構的預訓練模型得以迅速竄起。預訓練模型是近年來自然語言處理領域發展的重要推手，最具代表性的預訓練模型如 Google AI 提出的 BERT（Bidirectional Encoder Representations from Transformers）[6] 以及 OpenAI 團隊提出的 GPT-2（Generative Pre-trained Transformer 2）[24]，而預訓練模型在當時幾乎橫掃了所有自然語言處理相關任務的 state-of-the-art，可謂人工智

doi:10.6342/NTU202303745

慧領域一項劃時代的發展。其中 GPT-2 為 2018 年 GPT 模型的進階版本，模型的規模從原先的 1.1 億個參數升級到最大的 15 億個參數，而訓練文本則是蒐集了約 800 萬頁（約 40GB）的網頁文本所預訓練而成，使得 GPT-2 對於許多下游任務包含文本生成、機器翻譯、QA 等等，只需要取得和任務目標相關的少量訓練資料，將預訓練模型進行微調（fine-tuning）的動作，使強大的模型學習到現在專注的任務，便可以達到接近該領域 state-of-the-art 的表現。由此證明預訓練模型透過不斷提升模型容量與語料規模，模型的能力是可以不斷提升的，也因此近年來有無數自然語言處理領域的研究都是透過此類型的大型預訓練模型所完成的，在各項子任務的表現也逐年有突破性的發展。

BERT 以及 GPT-2 兩者皆是基於 Transformer 架構 [27]。Transformer 架構當中包含編碼器（Encoder）與解碼器（Decoder）兩大部分，並透過自注意力機制（self-attention mechanism）使整個序列可以平行運算，根據上下文的語意來計算出當前字詞應該填入哪個最合適的結果，產出合乎語意規則的資訊，而由於 Transformer 架構的輸入和輸出通常都是一長串的自然語言，因此也常被稱作 Seq2seq（Sequence-to-sequence）模型。本論文在研究架構方面會選擇使用 GPT-2 模型為主軸，而 GPT-2 的核心架構是 Transformer 的解碼器，且採用的是單向訓練，意指模型會不斷的透過前文並以機率的計算不斷預測下一個字，而非同時參考上下文的方式，這樣的序列生成方式相當適合文章這種有連貫性自然語言的資料型態。因此，GPT-2 在文本生成領域有相當卓越的效果，也是本研究使用 GPT-2 做為研究架構主軸的最主要原因。

doi:10.6342/NTU202303745

2.2 文本改寫

文本改寫（Paraphrasing）是自然語言處理的下游任務之一，也是發展時間相當長的一個領域，目的在於將原始輸入的語句或文章，利用不同的字詞結構來表達相同的意涵的內容。文本改寫可以被應用在非常多任務，例如 retrieval-based 的問答系統 [7, 32]、語意解析（semantic parsing）[3, 4] 以及對話系統的資訊增強（data augmentation）[8, 14] 等等都和文本改寫有直接的相關，由此可知文本改寫有相當大的價值與研究潛力。在深度學習尚未成為主流時就有許多學者嘗試解決過文本改寫的問題，例如將文本改寫視為語言翻譯的問題來研究 [1]，採用雙語平行語料庫（Bilingual Parallel Corpora）以及統計機率模型建構出文本改寫的方法。此外，改寫任務也時常被使用回譯法（Backward Translation）來實作，所謂的回譯法是指將原始的文本透過翻譯成不同語種的結果，再將這個結果翻譯回原本的語言，透過兩次語言的轉換差異來達到對原始文本改寫的方式，可以說是改寫任務當中最直觀也最基本的方式，卻也時常獲得不錯的結果。除了上述的方式之外，也有一派學者嘗試自己手工建立起改寫規則 [22]，並一脈相承至後續的研究 [2]，然而顯而易見的缺點便是手工建立起的規則太過複雜，同時也難以完全應付到所有情境。從上述眾方法以及面臨的問題便可以發掘改寫絕非一個容易的任務。

到了近年深度學習的概念快速發展且成熟後，則有了更進一步的文本改寫技術出現，如 Gupta 等人的研究 [11] 透過變分自動編碼器（Variational AutoEncoder, VAE）以及長短期記憶模型（Long Short-Term Memory）之結合成功建構出了 Seq2seq 的文本改寫模型。在 Li 等人的研究中 [18]，他們嘗試了使用強化學習（Reinforcement Learning）為基礎建構出 RbM（Reinforcement by Matching）架構進行文本改寫任務，生成器（Generator）生成出改寫語句後根據評估器（Evaluator）回饋的分數（Reward）做為依據調整自己生成的結果，如此反覆動作後目標是產

doi:10.6342/NTU202303745

出最大化 reward 的結果做為模型 output。

然而，根據文本改寫的 Survey 文獻 [31] 可知，由於應用情境著重在語意解析或是關鍵字重構等，因此目前的改寫研究多半是以句子為單位的重新排列組合或是單純的同義字詞置換，尚未有以文章為單位的改寫研究，也就是整篇文章的內容重構。此外，中文領域由於中文本身的文法、語句結構極為複雜，投入研究的時間與能量也相對較少，過去的研究 [26] 就有提到中文改寫生成（Chinese Paraphrase Generation）相較於其他 NLP 領域任務仍然處在起步階段。另外，與中文改寫相關的研究 [21] 提出了 LCQMC 資料集，主要是透過 Word Mover Distance 演算法 [15]，該演算法主要是基於 word-embedding 並計算兩個句話的 dissimilarity，並將每篇文章假設為 normalized bag-of-words (n-BOW)，並遍歷句子當中的每個字計算如公式2.1所示：

$$c(i,j) = \|x_i - x_j\|_2 \tag{2.1}$$

x_i 和 x_j 分別代表候選句當中的每個字，舉該篇論文當中的例子如圖2.1所示：

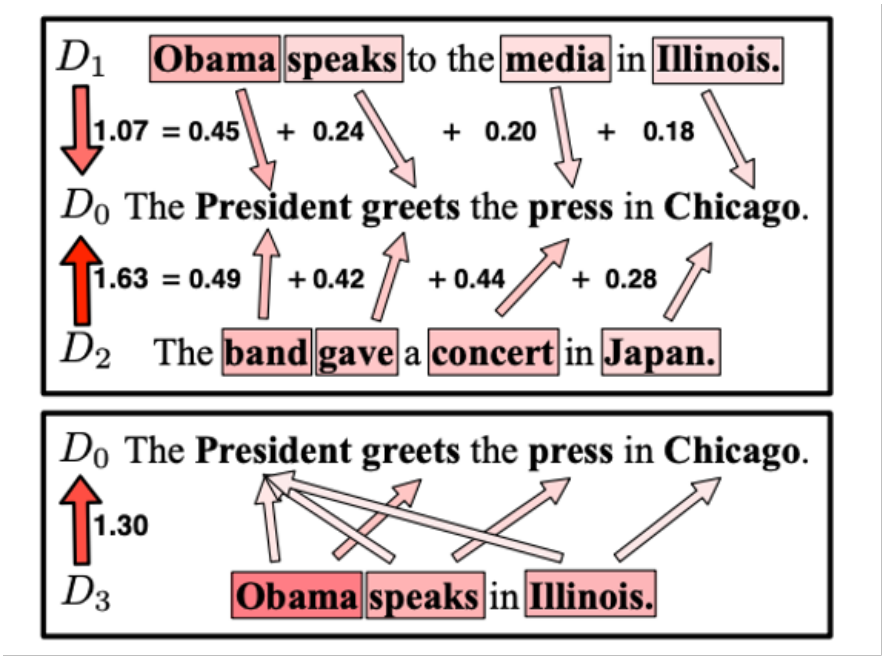


Figure 2.1: WMD 計算過程 [15]

doi:10.6342/NTU202303745

假設 D_0 中的任一字為 x_i，D_1 的任一字為 x_j，兩者之間的 word-embedding 為 $c(i,j)$，在遍歷了 D_0 和 D_1 並總和所有 c 以後，會得到兩句話之間的 distance，越小的即代表越相似的兩句話。LCQMC 資料集便是透過這樣的演算法，並定義出閾值來創造許多成對的語句，並為其標注兩句話是否互為改寫語句。

然而在 Liu 等人的研究 [21] 中也僅是停留在 sentence-level 的中文問句改寫配對，在中文改寫領域仍有相當大的發展空間。另外，就我們目前所知，在現有的文本改寫方式當中 [31]，也尚未有採取和本章開頭提及相同的文本改寫之研究方法。因此，本研究希望能按照前述的研究方法，以預訓練模型 GPT-2 做為基礎並建構出文本改寫模型，並以新聞報導的改寫為主題，期望在輸入原始新聞後，透過本研究建構之文本改寫模型產出一篇改寫版本的新聞，同時必須保留新聞重點以及文章必須邏輯通順、可讀，藉此達成研究目的的訴求。

2.3 語序重構和語言學分析

語序重構（Reordering）旨在對於原始輸入的文本進行結構上的重組，也就是將原始文本以句為單位分割，而後重新組合成新的一篇文章。在過去的研究當中，最著名的與語序重構相關的研究就是由 Google AI 提出的 BERT（Bidirectional Encoder Representations from Transformers）[6]。BERT 整體而言是一個自編碼模型（Autoencoder Language Model），研究人員設計了兩個子任務來預訓練出該模型，分別是遮罩（Masked LM）以及 Next Sentence Prediction（NSP）。其中 Next Sentence Prediction 所執行的任務便是和語序重構相關的任務，在使用巨量文本訓練的過程當中，作者將文本以「句」為單位分割，並不斷地將其中的兩句話輸入模型，希望模型判斷該兩句話在原文中是否為前後相鄰的兩句。在實際的預訓練過程當中，作者從文本語料庫當中隨機選擇了 50%的正確語句對以及 50%的錯誤

doi:10.6342/NTU202303745

語句對 BERT 進行 Next Sentence Prediction 的訓練，並在和 Masked LM 任務搭配之下使得模型對於語句甚至篇章等級的文本有理解與刻畫的能力，也因為有此基礎，BERT 對於每個字、詞向量都能夠有更全面地表示，為後續微調任務奠定了更好的參數初始值，也是 BERT 之所以能夠在當時稱霸各大自然語言處理相關任務的關鍵。

然而，從 BERT 發表以來，有許多研究人員都發表出了 BERT 的改良版甚至是加強版，使得各式各樣的模型在近年百花齊放，其中，Google AI 團隊提出的 ALBERT[16] 是其中一個改良版本。ALBERT 的全名為 A Lite BERT，顧名思義為一個輕量型的 BERT，比起 BERT 有更少的參數量。ALBERT 當中也如同 BERT 一樣有和語序重構相關連的任務，只是將 BERT 當中的 Next Sentence Prediction（NSP）改成了 Sentence Order Prediction（SOP），主要的想法是作者認為 NSP 任務太過簡單，模型只要判斷出兩個句子在闡述不一樣的主題，就可以輕易判別出兩句並不是相連的句子。而 Sentence Order Prediction 主要輸入給模型的同樣是兩個句子，但希望模型判斷的是哪一個句子應該被放置在前面，也就是模型必須能判斷兩句之間的順序關係。此外，輸入的訓練資料變成全部是文本中相鄰的兩個連續句子，而正樣本是正確的順序，負樣本則是調換的順序，使得模型無法單從句子的主題來判斷正確與否，而是必須對句子有更深入的學習和了解，也藉此學習到更多訊息。

和語序重構相關的任務在 BERT 以及 ALBERT 當中佔了舉足輕重的地位，也曾被後者的研究當作工具之一。在 Lin 等人的研究 [20]，由於需要訓練可以用來對文章做語序重構的模型，因此作者參考了 ALBERT 當中的 Sentence Order Prediction 來訓練一個語序重構的模型。假設輸入一篇文章 D，並將 D 拆成許多句子使 $D_p = S_p^1, S_p^2, ..., S_p^N$，共 N 句話，並且定義第 i 句話和第 j 句話之間的關係如公式2.2所示：

doi:10.6342/NTU202303745

$$\mathbb{G}(i,j) = \begin{cases} 1\left\{P_{SOP}\left(S_p^i, S_p^j\right) \geq \epsilon\right\}^1 & i \neq j \\ 0 & i = j \end{cases} \tag{2.2}$$

其中 $P_{SOP}(S_p^i, S_p^j) \geq \in$ 代表第 i 句話在第 j 句話之前的機率大於閥值，則判斷該情況可能會發生，則回傳 1。而當 $i = j$ 時，則表示 i 和 j 是同一句話，則回傳 0。作者使用大量的文章與上述方法訓練出了一個語序重構的模型，也是就我們所知極少數用以探討語序重構的研究之一。

然而，在上述的研究中，對於語序重構的探討並沒有加入對於一些有固定順序或是前後關係的語句做的額外處理，而是單純以學習過後的模型盡可能地保留住這些關係，因此，本研究會從語言學的觀點分析，並加入一些額外的規則，盡可能的使語序重構領域有更近一步的發展。在英語裡存在著一些規則足以明確的判斷語句的前後規則，其中連接副詞（Conjunctive Adverb）便是非常具代表性的一個例子。根據聖荷西州立大學寫作中心（San José State University Writing Center）曾經發表過對於連接副詞深入的研究 [29]，研究中曾經指出：連接副詞是被用來表達兩個分開的句子之間的前後關係，或是更直接的表達出兩句之間關聯為何。以文中整理出的不同種類為例，包含了因果、順序、時間、對比、強調、舉例、比較、總結等等，如圖2.2所示。

因此，本研究希望能參考圖2.2整理出來的連接副詞，根據其不同的功用列出不同的規則，並以 rule-based 的方式試圖保留住一些明顯應該被保留的句子順序，並且在確定這些順序有被保留後才進入本節初提到的語序重構模型，對於整篇輸入的原始新聞做出初步的文章結構重構，以確保改寫前後版本的文章並不只是一句一句的改寫，而是從整篇文章看起來有結構上的改變，為中文新聞改寫領域做出貢獻。

doi:10.6342/NTU202303745

Common Conjunctive Adverbs and Their Functions

Function	Examples			
Cause and Effect	accordingly	consequently	therefore	then
Sequence	first/next	finally	furthermore	in addition
Time	before	meanwhile	since	now
Contrast	however	instead	in spite of	rather
Emphasis	indeed	of course	certainly	definitely
Summarize	in conclusion	in summary	briefly	quickly
Illustrate	for example	for instance	namely	typically
Comparison	like/as	likewise	similarly	alternatively

Figure 2.2: Common Conjunctive Adverbs and Their Functions[29]

2.4 新聞之發展與自動生成

隨著人工智慧的技術演進，越來越多的人類行為可以被機器所學習甚至取代，新聞的自動生成便是一個很好的代表。根據德國顧問公司 Schickler 和世界報業協會 WAN-IFRA 於 2022 年共同撰寫的研究報告 [9] 指出，超過 77%的受調查出版商認為 AI 將會在未來三年內扮演企業能否順利發展的關鍵角色，調查統計圖如圖2.3所示。

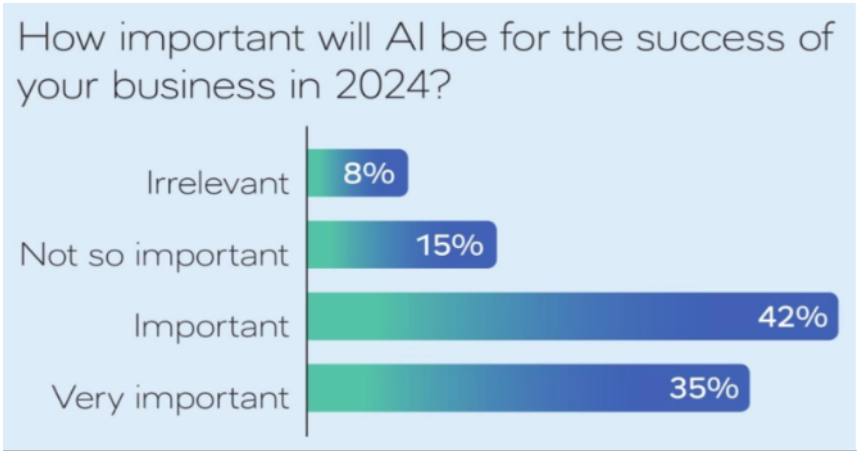


Figure 2.3: How important will AI be for the success of your business in 2024?[9]

另外，在報導中也有提到，世界報業協會認為由於 AI 日漸普及化，因此 AI

doi:10.6342/NTU202303745

在新聞業的應用將不會再是大財團或是技術先驅者的專利，許多規模較小的企業也可以發展屬於自己的人工智慧部門，推出自己的產品。由此可知，透過 AI 來生成新聞在國外已經不只是新鮮事，而是日常生活的一部份。然而，從研究數量的差異便可以發現，國外運用 AI 撰寫新聞的普及性以及實際應用的比例是遠高於中文領域的，除了模型發展時間的差異之外，中文由於訓練資料較為不足且本身語法結構極為複雜等緣故成為了限制。在外文領域有非常多 AI 撰寫新聞的系統以及方法的提出，而這些 AI 生成的新聞也實際會被應用並發布。如 Leppänen 等人 [17] 提出了 Data-driven 的新聞生成 NLG 系統，用於生成芬蘭的市政選舉新聞報導。另外 Kanerva 等人 [13] 曾經對體育新聞的自動生成研究出文本生成模型，用的是芬蘭的冰球運動新聞作為訓練資料集，由於體育新聞有很大一部份包含了客觀的數據統計，故此研究也被視為 Data-to-Text 的代表之一。相較而言，中文領域的新聞自動生成相對比較不普及，相關的研究有 Huang 等人 [12] 透過 SPOTSSUM dataset 作為模型訓練資料，並在 LSTM、Transformer 等框架建構足球中文新聞自動生成模型，為中文領域新聞自動生成的代表研究之一。然而在中文領域的新聞生成還並沒有被廣泛的應用於實際場域並發布供讀者閱讀，主要仍然受限於生成的新聞結果品質仍有進步的空間。

由上述的研究整理可知，中文新聞的生成尚未發展到可以完全開放並商業化的階段，也受限於此情況，在無法快速產生大量文本資料的情形下，改寫領域更是只有很粗淺的發展。因此，本研究希望能借助跨語種的方式，並結合自定義的語序規則以及部分現有的模型，最後再搭配上自行訓練的 Seq2seq 模型，為中文領域的改寫尋找出新的可能突破口，並為中文新聞做出改寫的，盡可能的達成研究目的與訴求。

doi:10.6342/NTU202303745

2.5 文本改寫評估方法

自從文本改寫受到學者的重視並投入大量研究能量後，各式各樣的研究方法與模型不斷的推出，因此用以評估各個模型產出結果的統一標準相當重要，有了這些標準才得以客觀地評斷改寫文本的優劣。下方將列出現有研究中常見的自動評估與人工評估方式，最後分析最適合用來評估本研究之實驗結果的指標。

2.5.1 自動評估

自動評估（Automated Evaluation）通常泛指透過預先建立的數學公式將生成出來的文本進行數值計算，以本研究的視角就是將改寫過後的新聞報導放進公式內計算，提供一個量化的數值作為衡量改寫新聞表現的重要依據，因此自動評估若使用得宜可以說是成本最低、最直觀也最有公信力的評估方法。以下內容會針對與文本改寫領域有相關聯的自動評估方式進行探討，從中歸納出可做為本研究評估方式之參考選項。

在 Witteveen 等人的研究中 [28] 列出了 USE 以及 ROUGE-L 等兩種模型評估指標，由於同樣是透過 GPT-2 預訓練模型來解決文本改寫的任務，與本研究之方法有相似之處，因此特別列出討論。

USE（Universal Sentence Encoder）是由 Google 團隊所提出的評估指標 [5]，透過計算生成出來的改寫語句和原始語句餘弦相似度（cosine similarity），判斷兩語句的語意相似程度。在文本改寫領域最強調的就是改寫前後版本所傳達的資訊一致性，而本研究聚焦的新聞報導則是希望在客觀新聞事件與數據被完整保留的前提下，利用不同的語句來重新建構一篇新聞報導，由此可知 USE 指標所顯示的文意相似度對於新聞改寫模型而言是相當重要的評估指標之一。

doi:10.6342/NTU202303745

ROUGE（Recall-Oriented Understudy for Gisting Evaluation）[19] 是文本摘要以及機器翻譯任務當中相當重要的指標，其概念是採用 N-gram 的方式計算模型產出語句和給定的標準答案之間的差距，並以混淆矩陣（confusion matrix）當中的召回度（recall）作為衡量依據。而 ROUGE-L 當中的 L 指的是 Longest Common Subsequence，其目標是在計算生成句和標準句裡頭最長的一串連續相同的字詞有多少的單位，再除以生成句的單位而得出一個數值。此指標之所以可以被應用於評估文本改寫任務的主要原因是對於以文章為單位的文本改寫而言，並不希望生成出來的改寫文章和原本的文章有太長的連續相同字句，也就是過大的 ROUGE-L 值，否則就沒有達成文章改寫的研究目標，而只是某些字詞的抽取置換，甚至有某些句子完全相同。因此，在 Witteveen 等人的研究中 [28]，評估的方式是剔除掉 ROUGE-L 大於 0.7 的結果，而在本研究中需要訂定什麼樣的 threshold 則同樣是需要透過實驗獲得的超參數（hyperparameter）。

此外，在 Gupta 等人 [11] 以及 Goyal 等人 [10] 兩篇研究當中，都是做出和改寫生成相關的研究，而兩者共同都使用了 self-WER 作為自動評估的指標之一。

WER（Word Error Rate）是改寫領域中最重要的評估指標之一，WER 代表的是字錯誤率的意思，最起初是被用於語音識別（ASR）的評估指標，為了判斷語音辨識的品質如何，如公式2.3所示：

$$WER = 100 * \frac{S + D + I}{N} \tag{2.3}$$

其中 N 代表的是正確文本的總字數，S 代表的是被替換掉的字數，D 代表的是被刪除掉的字數，I 代表的是插入的字數。而之所以被用來當作改寫領域的研究是因為改寫在做的工作也是對文章做出字詞上的置換或取捨，讓讀者有不同版本的新聞可以閱讀。因此，在本研究中，我們也會使用這個指標作為客觀的衡量標準，

doi:10.6342/NTU202303745

但我們會將分子部分的 I，也就是插入的字數拿掉，僅計算被替換以及被刪除的字數，因為若是改寫文章中被插入了許多無意義的字詞，會影響到此指標的客觀性，算出來的指數可能會被無限地放大。

此外，在文本生成領域當中還有一個相當常見的自動評估指標 BLEU[23]。BLEU（Bilingual evaluation understudy）是屬於 word-based 的通用性評估指標，最常被用來評估機器翻譯的結果，衡量生成語句的流暢性，主要是計算參照句（reference）和模型生成句（candidate）的 N-gram Precision，並會加入懲罰係數（Penality）控制生成句的長度。BLEU 因為計算成本小且快速、容易理解且和人類評價結果高度相關而被廣泛使用。

2.5.2 人工評估

人工評估意指由人類來對文本改寫的結果進行評估，這類的評估並不會有客觀的數值或是公式，而是完全交由人類以主觀的視角進行結果評估。由於自然語言本來就是人類日常生活學習並用來溝通的工具，因此在邏輯上透過人工的方式對文本改寫模型進行結果評估會相較於以數學公式為基礎的自動評估指標來的更有說服力也更具代表性。就本研究而言，在參考與本研究密切相關的文獻 [20] 後，歸納出在對改寫文本的人工評估階段共有三個指標，其一是 Relevancy，負責評估生成文章與原文表達的語意是否一致，是否通順可讀，其二是 Diversity，負責評估生成文章與原文之間被改寫的程度，最後是 Coherence，負責評估生成文章與原文之間是否有保留住文章的前後邏輯。因此，本研究也將參考這三個指標作為人工評估的一部份，分別是新聞可讀性，用讀者的觀點判斷模型改寫的新聞的用字語句及段落是否流暢，是否和人類撰寫的新聞一樣能讓使用者順利閱讀、新聞改寫程度則是用閱眾的角度來判斷改寫後的新聞與原始新聞在字詞上是否有足夠程度的改寫，以及新聞邏輯性，用讀者的角度判斷改寫過後的新聞是否有保持

doi:10.6342/NTU202303745

著原始新聞的語句語事件的邏輯性。此外，由於本研究有執行文本的語序重構，對輸入文章有做出結構上的改變，因此在人工評估的部分也特別加上新聞結構改變度作為指標之一，希望能透過讀者的觀點特別對判斷改寫後的文章與原文是否有明顯在文章結構上的差異，如此才能確保本研究對於以文章為單位的新聞改寫有確實達到研究目的之訴求。

2.5.3 適合本研究之評估指標

從上述列出的評估方式可知，不論是自動評估或人工評估都各自有其存在之必要，也各有其最佳的應用場域，卻也各自存在著缺陷。舉自動評估當中的 USE 指標為例，由於是以餘弦相似度公式為基礎之語句相似度計算，假設給定兩語句「我吃豬肉」與「豬肉吃我」，兩語句明顯是完全不同的含義，然而由於 USE 沒有考慮詞與詞順序的緣故，此兩句會被誤判為相同的含義，這樣的情況造成了 USE 指標無法單獨使用，需要透過其他指標的協助或是更進一步的驗證才能確保其可依賴性，特別是在新聞改寫領域當中可能會把張冠李戴的資訊誤判為兩者意義相同，造成評估上的錯誤。另外，在 BLEU 的部分，因為其公式計算的特性是不能接受替換字詞的出現，否則會影響評分結果，而且也會時常出現短句會獲得較高分的狀況。然而替換字詞的出現在改寫議題當中是非常重要且常出現的一環，且模型生成的語句長度也難以被控制，由以上缺點便可以將 BLEU 排除在本研究的評估方式之外。此外，由於文章的可讀性、邏輯性以及結構改變度三個面相是相當主觀且沒有標準答案的，自動評估指標在這個部分並無法提供協助，然而此三個面向對於本研究著重的新聞報導是相當重視的重點，種種因素導致新聞報導的改寫生成要完全使用自動評估指標作為評斷模型表現的想法是難以實現的。而人工評估固然可以解決上述問題，但是由於目前並沒有統一的研究單位執行評估，請到的評斷人員各自有自己主觀的判斷依據，且素質不一，無法利用客觀可衡量

doi:10.6342/NTU202303745

的數字為模型結果做出評斷，其中難以避免的存在主觀因素造成的影響。

然而，縱使兩者皆存在著各自的缺陷，然而由於目前並沒有約定俗成的評估方式，目前的研究也都是以這兩者交叉驗證作為自己模型輸出的評估方式。本研究目標是要對給定的新聞報導透過自建的改寫生成模型生成出其改寫版本的另一篇報導，由於希望能有別於以往改寫僅僅是字詞置換的改寫方式，因此本研究在自動評估方面會選擇 ROUGE-L 以及 WER 作為指標，透過足夠低的 ROUGE-L 值確保改寫前後的版本並沒有過大篇幅的重複用字，藉此驗證本研究的模型是真正以文章為單位的改寫生成，而透過計算出足夠高的 WER 值則可以確保本研究產出的改寫結果和原始文本確實有一定程度的改變。另外由於前述提及的自動評估指標限制，因此人工評估也會是本研究驗證的很大一部份，透過人工來判斷改寫新聞可讀性、新聞改寫程度、新聞邏輯性以及新聞結構改變度四個部分並進行評分，藉此確保改寫版本的新聞有一定的品質。

2.6 總結

根據文獻探討的結果，本研究整理出以下結論：

1. 文本改寫的任務一直是自然語言處理當中非常受到重視的一個發展議題，然而目前所查閱的研究主要仍都是以句子為單位的，以文章為單位的文本改寫研究佔其中的非常少數。此外本研究欲使用的研究方法，也就是透過跨語種的轉換、自定義的語序保留規則、語序重構模型以及自行訓練的改寫模型等眾多方法結合的研究架構，截至目前並未看到與此相同的研究方法，也是本研究希望嘗試並實作的新型態文本改寫方式。

2. 本研究模型的選擇會以 Transformer 以及 Seq2seq 為基礎的 GPT-2，希望在

doi:10.6342/NTU202303745

本研究的後半段，也就是改寫新聞語句的部分能借助 GPT-2 隨機生成下文的特性獲得改寫後的結果，藉此達成文本改寫之目的。

3. 根據文獻的查閱，本研究認為現有的自動評估指標對於本研究表現之評估的幫助相當有限，再加上本研究著重在新聞領域的文本，而新聞報導的撰寫上是有非常專業且需要遵循的規則，特別是文章可讀性、邏輯性等等，因此在少數可參考的自動評估指標之外借助人工評估的協助，也就是前一小節提到的四種指標做出相對客觀的評估，如此才能有效的對模型的改寫結果做出具說服力的評估，確保改寫文章的品質。

doi:10.6342/NTU202303745

第三章　研究方法

本章節旨在詳細說明本論文之研究方法，包含整體研究架構，所使用之資料集、預訓練模型的完整介紹，此外也會說明本研究的驗證方式，包含了自動評估以及人工評估兩種，做為本研究品質之評判依據。

3.1　研究架構

本研究提出一個中文新聞文本的改寫生成系統架構，不同於以往類似主題的研究 [18][26]，比較專注的部分僅在以句子為單位（Sentence-Level）、架構也比較偏向直接從原始文本生成出改寫文本的一階段模型，本研究所採用的新聞改寫方式是先將原始新聞轉換成英文版本，首先經過透過參考 Wong 的研究中 [29] 所整理歸納的不同情境連接副詞構成的語序 rule-based，先對有明顯前後關係的語句做順序保留，第二階段會經過 Lin 等人的研究中 [20] 透過 ALBERT 的 SOP 訓練出來的語序重構模型，最後再以句為單位輸入到我們自行訓練的 GPT-2 文本改寫模型當中，得到最後的結果後轉譯回中文成為最後的改寫版本中文新聞。本實驗之所以會採用較多階段的複合式模型主要原因是一方面想借助改寫研究中常見的回譯法（Backward Translation）來加入在改寫的方法之中，另一方面是由於目前改寫領域的發展以及可供實驗用的資料多半還是以句為單位，且具有品質的資料目前仍是以英文為大多數，中文領域不僅品質多半參差不齊，更是多以簡體中文為

doi:10.6342/NTU202303745

主，繁體中文的研究以及資料幾乎寥寥無幾。而自行搜集則會受限於時間以及金錢成本的考量而有所卻步，且在沒有標準答案的改寫領域也會對搜集得到的資料有品質上的疑慮。有鑑於此，在本實驗所需要達到的以文章為單位的改寫之訴求中，勢必須要加入翻譯、語序重構甚至自定義的方法協助完成這個任務。以下為本研究的實驗架構所細分的三階段：

3.1.1 第一階段

第一階段我們首先將原始新聞報導翻譯成英文，而之所以需要翻譯成英文主要是考量到現階段中文改寫是相對不成熟的，特別是資料量多半是簡體中文，和台灣使用的繁體中文有著極大的用詞差異，同時品質也參差不齊，而在短時間內也非常難以蒐集到需要大量成本建構的優質改寫資料集。有鑑於此，本實驗決定使用相對較齊全的英文資料集做訓練，也才會需要將原始新聞翻譯成英文。在得到英文的新聞以後，我們會以抓取關鍵字的方式保留住明顯的語序前後關係，關鍵字的內容為根據 Wong 的研究中 [29] 所整理歸納的不同情境連接副詞，如圖3.1所示，並因應不同的連接副詞會有不同的規則，由此建立起自定義的 rule，為的是確保在後續語序重構時可以達到不破壞語序邏輯並同時做到文章結構變化。

3.1.2 第二階段

第二階段我們會將原始英文新聞切分成以句為單位，且以遍歷法兩兩送入文本語序重構模型 [20] 當中計算出該前後句組合的機率。在得到所有組合的機率以後，外加上第一階段根據 rule 保留下來的語序規則，在滿足規則的前提下重新組合出最大總合機率的新組合，最後得到出現結構變化的文章，藉此方式達成研究

doi:10.6342/NTU202303745

Function	Conjunctive Adverbs
Cause and Effect	therefore, thus, hence, as a result, consequently, as a consequence, accordingly
Sequence	also, in addition, moreover, furthermore, additionally, what's more, on the other hand
	besides
Emphasis	in other words, that is, namely
Time	after that, then, now
Contrast	however, nonetheless, nevertheless, still, otherwise, even so
Illustrate	for example, for instance, such as
Comparison	on the contrary, by contrast, instead, conversely, similarly
Summarize	Finally, Lastly, Eventually, In conclusion, Conclusively, To sum up, To conclude, In summary, Last but not least

Figure 3.1: 關鍵規則連接副詞與種類範例

目的所追求的篇章結構等級的變化。而第一階段的存在主要是為了能在規則之下避免第二階段重構時出現了非常明顯的語序邏輯錯誤，也藉此達到了優化前人模型的研究貢獻。

以下舉一個例子做第二階段的說明，假設某文章原始有 ABC 三個句子，順序為 $A \to B \to C$，且該文章在第一階段得到的語序關係規則為 B 必須在 C 前面。則在第二階段我們會先將語句兩兩一組放進模型中計算該組合的機率為何，得到語序 $A \to B$ 的機率 $P_{A \to B}$、$B \to A$ 的機率 $P_{B \to A}$、$A \to C$ 的機率 $P_{A \to C}$、$C \to A$ 的機率 $P_{C \to A}$、$B \to C$ 的機率 $P_{B \to C}$、$C \to B$ 的機率 $P_{C \to B}$ 共六個。而三個句子排列過後共有 $3! = 6$ 種結果，如表3.1所示。

在表3.1中，原新聞共可以排列出 6 種不同的語序結構，而其中僅有三種是符合語序關係規則，也就是 B 必須在 C 前面，此時，系統會從三個符合語序關係規則的語序結構當中挑選一個機率總和最大的組合，做為輸出結果，同時也是第二階段的結果。

doi:10.6342/NTU202303745

Table 3.1: 語序重組結果表

Permutation Results	Probability Combinations	Follow Enforced Orders
$A \to B \to C$	$P_{A\to B} + P_{B\to C}$	**True**
$A \to C \to B$	$P_{A\to C} + P_{C\to B}$	False
$B \to A \to C$	$P_{B\to A} + P_{A\to C}$	**True**
$B \to C \to A$	$P_{B\to C} + P_{C\to A}$	**True**
$C \to A \to B$	$P_{C\to A} + P_{A\to B}$	False
$C \to B \to A$	$P_{C\to B} + P_{B\to A}$	False

3.1.3 第三階段

最後階段我們會將前二階段所得到經過語序重構的英文文章以句為單位送入我們自行建構並訓練的文本改寫生成模型，此模型是以深度學習預訓練模型 GPT-2 為基礎所微調（fine-tuning）出的下游任務模型，並得到每句話的改寫版本，經過組合並翻譯回中文後得到最後最終的模型輸出，也就是改寫過後的中文新聞。此階段為本研究的核心實驗內容，我們將針對這個模型做各種參數上的調整，並適時地修正實驗步驟，為求盡可能的達到此模型最佳的表現。詳細的文本改寫系統流程如圖3.2所示。

本研究架構採取多階段複合式文本改寫的原因有二。首先，本研究著重的領域是以文章為單位（document-level）的文本改寫任務，目前並不存在也難以取得足量 document-level 的文章改寫前後版本做為模型訓練的輸入與輸出資料，因此並沒有辦法直接訓練一個單一階段的文本改寫模型提供使用者輸入想改寫的文章並直接獲得改寫過後的版本。此外，有鑒於單純地以句子為單位改寫多半只能為文句字詞做同義字詞的置換，在意義上並未真正達到符合人類理想中「改寫」的目的。因此，本研究提出結合回譯法以及改良版本的句序重構的技術，再加上透過 GPT-2 模型建構而成的語句改寫生成模型中，生成出以原新聞為基礎而重製出

doi:10.6342/NTU202303745

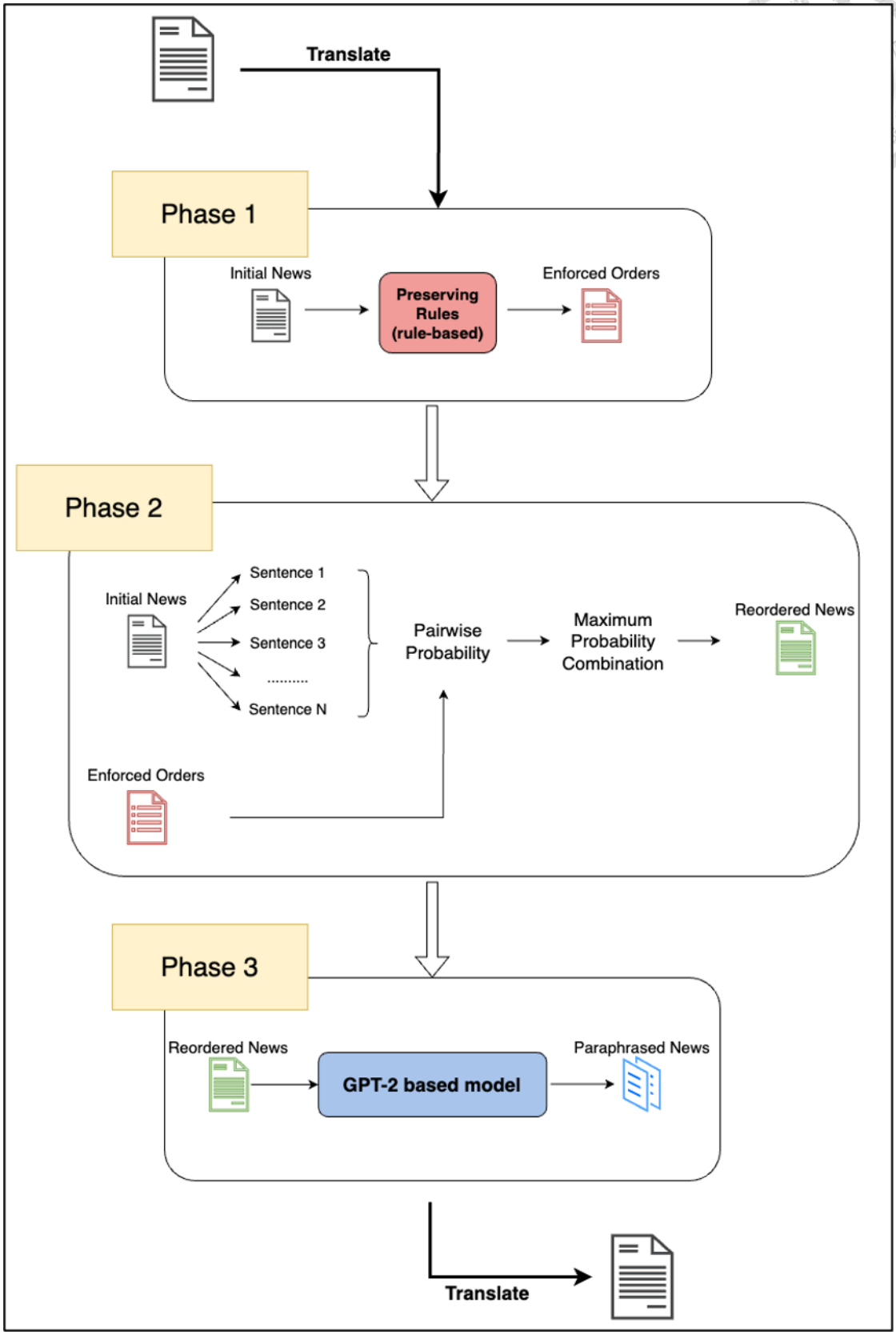


Figure 3.2: 新聞改寫系統流程圖

doi:10.6342/NTU202303745

的新版本文章，透過此複合式的模型重新產生一篇新聞報導，以達成研究目的所闡述的目標。

3.2 資料集

本研究所使用之資料集主要使用於第三階段之 GPT-2 Seq2seq 模型之訓練，由於本階段主要是要進行以句為單位的英文語句改寫，因此本研究整理了 TaPaCo[25] 與 PAWS-X[30] 兩個資料集所整合出的內容。此二者皆是公開發表於網路的開源內容，特色是兩者皆是大量以句為單位的語句改寫配對，且皆是多語言的版本的內容。又根據本研究之第三階段所需，因此蒐集目標是英文的版本，外加自行手動整理的方式歸納出用於微調 GPT-2 模型之訓練資料。

資料集當中最終被整理出來的版本由於是以句為單位，每句話落在 20 字以內，此長度用於微調（Fine-Tuning）GPT-2 模型是在可接受的範圍內。此外，從前人整理的文獻 [31] 可知，與數字相關之文本生成任務仍是尚待解決之議題，還沒有辦法很穩定的生成出正確的資訊以及內容，而對於文本改寫議題同樣不例外，此領域的研究人員皆是以文字內容作為改寫的目標，並沒有探討與數字或數據相關的內容，資料集也會將數字相關的內容屏除在外。因此，本研究也會在處理數據時做同樣的處理，只留下純文字相關的內容，以避免出現不可預期之錯誤與生成結果。而本研究整理過後所得之資料內容如圖3.3所示：

3.3 GPT-2 文本生成模型

在研究架構中 Phase 3 的英文語句改寫生成模型是本研究的主要實驗對象，在建構此模型主要是要對大型預訓練模型 GPT-2 做微調（fine-tuning），進而建構出

doi:10.6342/NTU202303745

Sentence-Based Paraphrasing Dataset	
Initial Sentence	**Paraphrased Sentence**
We had an early lunch.	We ate lunch early.
Tom has already decided where to go.	Tom already decided where he wants to go.
There cannot be progress without communication.	There is no progress without communication.
The old lady busied herself on her vegetable garden.	The elderly lady is busy in the garden.
Viviano Codazzi was an important contemporary artist of the genre, whose work was influenced by Alessandro Salucci.	Viviano Codazzi was an important contemporary practitioner of the genre whose work was influenced by Alessandro Salucci.

Figure 3.3: 英文改寫資料集範例

能夠生成改寫文本的的模型。而 GPT-2 預訓練模型由於是 Seq2Seq 為基礎的語言模型，對於輸入的長度有最多 1024 的限制，然而本實驗由於是以句為單位的改寫生成，資料集的每筆資料也是以一句話的長度做為 GPT-2 的輸入，訓練出 GPT-2 based 的文本改寫生成模型。前述提及的預訓練模型輸入長度規範在本研究並不會形成研究限制，因為若是將改寫前後語句串在一起輸入模型，字數約會落在 50 字以內，是 GPT-2 模型可以接受的範圍。

在微調階段，參照過往同樣使用 GPT-2 為基礎的研究，在對 GPT-2 進行語言模型建模時，微調（Fine-Tuning）階段採用的是監督式學習的方式，將成對，也就是改寫前後的文本資料，讓模型針對我們的資料集內容進行建模，不斷藉由上文來預測下一個字，讓模型針對輸入的語句去產生相對應的改寫結果，藉此方式

doi:10.6342/NTU202303745

來訓練模型內部的參數，建構出目標模型。就本研究而言，輸入的文本資料內容即是語句改寫前與改寫後的成對文本資訊，經過模型的訓練與參數微調之後得到文本改寫模型，也就是第三階段的任務目標。

3.4 研究驗證

本研究主要的任務為文本改寫，在研究驗證的部分主要會分成兩個部分，分別是自動評估與人工評估兩種方式，並將資料集切分成訓練、驗證與測試三部分資料，透過訓練集建構模型，驗證集微調模型，並用測試集來做模型評估。此外，由於本研究包含多個階段，因此在每個階段分別會有各自的評估方式或是指標，說明如下：

3.4.1 自動評估

在自動評估的部分，依據實驗架構可分為兩大部分，前半部分主要是針對成篇的文章做結構上的改變與重組，而後半段則是以句為單位的改寫生成，因此本小節將依序探討個別的自動評估指標。

首先是文章結構改變度的部分，由於本研究執行的是對於原始文章以語句為單位的位置重新排列與置換，為的就是讓原始文章與改寫後的文章有明顯結構上的變化，然而在語序置換的同時，也期望盡可能地保留住原文當中的前後語序關係，以維持改寫出來的文章有一定程度的品質以及邏輯性。然而就本研究目前的所知，尚未有探討和文章結構改變度以及文章前後邏輯順序保留程度的相關衡量指標，因此，本研究試圖提出一個評估指標作為參考指數，未來若有其他研究人員希望對語序重構領域有更深入的研究或優化，此指標也可供參考。

doi:10.6342/NTU202303745

就文章結構的改變程度，本研究提出的指標將被命名為 Reordering Ratio，也就是用於衡量在可接受調換的語序組合中，共會約有多少比例的組合被語序重構模型置換，如公式3.1所示：

$$ReorderingRatio = \frac{NumberofReversedPair}{C_2^N - NumberofEnforcedOrderPair} \tag{3.1}$$

其中 N 代表的是原始文章的總句數，因此 C_2^N 所代表的意思若任兩句話是一個組合，總共會有多少個組合。而 $NumberofEnforcedOrderPair$ 代表的是在原始新聞當中，共有多少對組合是有明顯前後關係，不應該被模型交換的語句對。目前由於尚未有約定俗成的判斷方式或公式，因此本研究使用的是用人工判讀的方式定義兩句之間是否存在必須遵守的語序關係（Enforced Order）。如此可以得到分母的值代表的是所有組合減去應該被保留的組合，留下可以被前後對調的語句組合數。

此外，假設原始文章中有 A 和 B 兩句話，原本 A 在 B 的前面，而經過語序重構模型後變成 B 在 A 的前面，則我們定義 A 和 B 這個組合是被反轉的組合（reversed pair），因此分子代表的就是共有幾組這樣子的組合。最後計算得出的結果就定義為重組比率（Reordering Ratio）。

另外，前述提及在語序重構的同時也希望能夠同時顧及到品質，而不單純只是追求大幅度的轉換，也就是希望可以同時顧及到語句之間的邏輯順序。因此，本研究定義出另一個指標如公式3.2所示：

$$EnforcedOrderPreservingRatio = \frac{NumberofPreservingPair}{NumberofEnforcedOrderPair} \tag{3.2}$$

其中分母代表的和 Reordering Ratio 中的分母一致，而分子代表的是經過模型的置換後，分母的這些組合有多少對組合被保留下來並維持原先的順序。如此計算可

doi:10.6342/NTU202303745

以對目前使用的語序重構模型做出客觀的品質依據。

以下用實際例子來說明 $ReorderingRatio$ 以及 $EnforcedOrderPreservingRatio$。假設某篇新聞共有 5 句話，依序為 $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$，且經過人工判定後，發現 A 與 B 有前後關係，D 與 E 有前後關係。接著，經過模型的語序重構以後得到新文章，5 句話依序為 $C \rightarrow B \rightarrow D \rightarrow A \rightarrow E$。此結果可以得到各項數值如下表3.2所示：

Table 3.2: 重構品質評估數值表

	Items	Value
N	A、B、C、D、E	5
C_2^N	AB、AC、AD、AE、BC、BD、BE、CD、CE、DE	10
Number of Enforced Order Pairs	AB、DE	2
Number of Reversed Pairs	AC、AD、BC	3
Number of Preserving Pairs	DE	1

根據表3.2中內容，可以計算 $ReorderingRatio$ 的分母 $10-2$ 得到 8，並且在這 8 個組合中，共有 3 個組合被反轉，因此得到 $ReorderingRatio$ 值為 0.375。此外，$EnforcedOrderPreservingRatio$ 的分母為 2，而在這兩個組合中有 1 個組合被保留，因此得到 $EnforcedOrderPreservingRatio$ 值為 0.5 。

實驗架構的後半段所做的是以句為單位的文句改寫，在評估指標的部分選擇的是 2.5.3 提到的 ROUGE-L 和 WER 兩個指標。首先，為了避免改寫過後的句子和原始句有過多的連續重複的用詞，確保有做到一定程度的「改寫」，因此本研究採用 ROUGE-L 指標來對改寫語句和原始語句進行最長連續相同字佔改寫後語句長度比例的計算，藉此評估模型對原始文本的改寫程度。ROUGE-L 指標的如公式3.3所示：

$$P_{lcs} = \frac{LCS(X,Y)}{N} \tag{3.3}$$

其中 X 代表原始語句，Y 代表改寫後的語句，$LCS(X,Y)$ 代表的是 Longest

doi:10.6342/NTU202303745

Common Subsequence，為原始語句和改寫語句當中最長的一串連續相同的字詞有多少的單位。N 則是代表改寫語句的長度。因此整個公式想計算最長重複字詞佔完整改寫語句的多少比例。我們會利用這個指標去計算本實驗的結果和 Baseline 模型的之間的比較結果，作為確認本實驗模型效果的客觀依據，而就客觀上而言，本指標應該是越低越好，代表生成句和改寫句有更大的變化，然而由於本指標僅依字詞的變化程度做數值的計算，並沒有辦法兼顧到改寫的品質，因此本研究必須同時借助人工評估來對改寫模型的品質做出客觀的衡量。

另外，本研究也將使用 WER 指標來對模型改寫的程度做出評估，本指標主要是要計算原始輸入語句在經過改寫模型的轉換後，有多少比例的字詞被修改或是刪除。WER 指標如公式3.4所示：

$$WordErrorRate = \frac{Substitution + Deletion}{N} \tag{3.4}$$

其中 N 代表的就是原始語句總共的字數，而 $Substitution$ 代表的就是被修改的字數，$Deletion$ 則是代表被刪除的字數。本指標的應用情境從原本的語音辨識結果移植到本研究的改寫領域，也因為運用場域的不同，因此本指標的數值也從原本應該要盡量的小，因為需要追求高準確率，變成希望追求本指數能夠越大越好，因為希望文本有更大幅度的改變。然而和 ROUGE-L 一樣，WER 並沒有辦法在計算改寫幅度的同時兼顧品質，因此同樣需要人工評估的協助，以確認改寫模型在各方面都有一定的品質。

3.4.2 人工評估

在人工評估的部分，我們會建立表單並請十名受測者對改寫版本的新聞進行分數評估，受測者的選擇標準是至少有大專以上的學歷，原因是因為本研究主要

doi:10.6342/NTU202303745

執行的是新聞改寫，且研究目的的一大部分是為了在生成或改寫新聞的同時可以兼顧到品質，而新聞的品質好壞應該是由接收到這些新聞內容的一般閱眾來決定和判斷，因此將大專學歷以上作為基準以代表一般會接收到新聞資訊的普羅大眾。

我們將從整理過的資料集取出部分當作測試資料集，並隨機整理出共 30 篇新聞，在經過我們的模型對這些新聞改寫後進行分數上的評估。此外，在基準模型（Baseline Model）的部分，其一是經典的方法 Backward Translation，用以代表早期的改寫方式，其二選擇的是在 2022 年由 OpenAI 團隊推出的劃時代自然語言聊天機器人 ChatGPT。ChatGPT 在推出之後可以說是對自然語言領域有相當大的衝擊，對於自然語言的任務幾乎可以說是無所不能，生成的結果在品質上也是有驚為天人的效果，而在改寫領域同樣也有這樣的情形。因此，本研究將以 ChatGPT 作為主要的比較對象，並參考與改良由 Lin 等人的研究中 [20] 針對人工評估所整理的人公評估指標，包含了新聞可讀性、新聞改寫程度、新聞邏輯性，外加在 2.5.2 小節討論過的新聞結構改變度共四個指標來做評估，讓受測者對改寫新聞進行評估，分數範圍為 1 至 5 分，1 分代表非常不同意，3 分代表普通，5 分代表非常同意，希望藉此評估本新聞改寫系統的結果對於一般閱眾而言是否流暢且可以順利閱讀、是否和原始新聞相較之下有明顯的改寫、是否有保留著原始新聞的語句與事件的邏輯性，以及是否和原新聞之間有明顯可見結構上的改變等效果。此外，也透過分數上的比較客觀的顯示和基準模型 ChatGPT 之間的比較，觀察生成後的自然語言結果在各指標上的表現之比較。透過上述的人工評估結果提供一個客觀的數值藉此評斷模型所產出的自然語言品質為何，同時也做為本研究驗證的重要依據之一。

doi:10.6342/NTU202303745

第四章　研究結果

本章節我們將對於本研究所提出的研究方法實際進行實驗，並詳細說明各階段之規則與訓練模型的超參數設定等等，除了分析我們的實驗結果以外，同時也會選擇 baseline model 和本研究之研究結果做比較，進行自動評估以及人工評估，最後針對評估的結果進行討論與小結。

4.1　語序規則建立和語序重構

4.1.1　規則詳細說明

本研究的前半部主要是使用語序重構模型 [20]，並額外參照 Wong 研究中 [29] 所整理歸納的不同情境連接副詞所自訂的語序邏輯規則，圖4.1中整理出的連接副詞是常被用來連接前句的承接型副詞：

圖4.1中整理了包含因果、順序、強調、時序、對比、舉例、比較等等功能中最常見的連接副詞，若是系統在拆解完所有句子後，發現其中某句話的開頭使用的是圖4.1中的某個連接副詞，舉 Wong 研究 [29] 當中的例子如圖4.2所示：遇見上述的狀況時，則第一句話稱為前導句，以連接副詞開頭的句子稱為承接句，我們會在語序重構時訂定一個強制的規則：重構後前導句必須要在承接句之前。藉由為語序重構訂定這個規則，可以有效的避免掉某些明顯的語序邏輯錯誤。

doi:10.6342/NTU202303745

Function	Conjunctive Adverbs
Cause and Effect	therefore, thus, hence, as a result, consequently, as a consequence, accordingly
Sequence	also, in addition, moreover, furthermore, additionally, what's more, on the other hand, besides
Emphasis	in other words, that is, namely
Time	after that, then, now
Contrast	however, nonetheless, nevertheless, still, otherwise, even so
Illustrate	for example, for instance, such as
Comparison	on the contrary, by contrast, instead, conversely, similarly

Figure 4.1: 承接型連結副詞表

Examples:	Several countries locked down during the pandemic. **Similarly,** other countries did the same. Few people believed the pandemic was a problem in the beginning. **Now,** they are facing the consequences.

Figure 4.2: 承接型連接副詞範例

此外，某些特殊的連接副詞會被專門用於文末收尾句的起始，用以對整篇文章做出結論或統整，如圖4.3所示，範例如圖4.4所示：

Function	Conjunctive Adverbs
Summarize	Finally, Lastly, Eventually, In conclusion, Conclusively, To sum up, To conclude, In summary, Last but not least

Figure 4.3: 統整型連接副詞表

Correction: The population practiced public safety. **Finally,** normalcy returned.

Figure 4.4: 統整型連接副詞範例

若遇見上述情況時，則明顯代表該句話是做為文章的收尾句，因此，我們給予的規則就是在語序重構時，該句必須要維持在最後一句話。

doi:10.6342/NTU202303745

4.1.2 語序重構結果分析

綜合4.1.1的規則建立以及語序重構模型的使用後，便是本研究在實際為每個句子進行改寫前會對輸入文章做的篇章結構重組，如此操作是為了能確保改寫前後的新聞有結構上的改變，而非只有同義字詞上的字面抽換。以下我們實際用測試集的新聞資料來實際演示本研究在語序重構的流程，並說明結果。首先，我們先將原始的新聞輸入本模型，如圖4.5：

[1]In London, cybersecurity is one of the most critical challenges of the digital age . [2]Everyone from households to businesses to governments , has a stake in protecting our era 's most valuable assets, the data. [3]The question now is how to achieve this . [4]With attackers becoming ever more nimble and innovative , armed with an increasingly heterogeneous of weapons , cyber-attacks are occurring at an accelerating pace and with greater sophistication than ever before. [5]As a result, the scale of the challenge should not be underestimated .

Figure 4.5: 原始新聞輸入範例

透過模型的分句，原始新聞被切成 5 句話，每句話開頭前的中括號內的數字代表的是該句是第幾句話。此外，經過 rule-based 模型的判讀，發現第 5 句話的開頭為 As a result，為承接型連接副詞的範疇之一，因此可以給予規則為句 [4] 必須在語序重構後依然維持在句 [5] 之前。建立的規則會記錄如圖4.6所示：

Preserving Orders	
Conjunctive Adverbs	**Orders**
As a result	Sentence 4 must before Sentence 5

Figure 4.6: 語序規則保留結果範例

doi:10.6342/NTU202303745

緊接著是語序重構模型的部分，我們選擇的是過去研究 [20] 的模型，該模型使用了大量的英語文章做為訓練資料，並將每篇文章拆成許多句子，得到 $D_p = S_p^1, S_p^2, ..., S_p^N$，共 N 句話，並且定義第 i 句話和第 j 句話之間的關係如公式4.1所示：

$$\mathbb{G}(i,j) = \begin{cases} 1\left\{P_{SOP}\left(S_p^i, S_p^j\right) \geq \epsilon\right\}^1 & i \neq j \\ 0 & i = j \end{cases} \tag{4.1}$$

其中 $P_{SOP}(S_p^i, S_p^j) \geq \in$ 代表第 i 句話在第 j 句話之前的機率大於閥值，則判斷該情況可能會發生，則回傳 1。而當 $i = j$ 時，則表示 i 和 j 是同一句話，則回傳 0。如此訓練而得的模型會將原始新聞分為以句為單位，計算出所有任兩句話互為前後句的機率為何，最後將所有語句組合出一個最大機率的總和。而本研究則是在計算最大機率總和的同時加入了如表所示的規則，以求模型在進行語序重構時也必須滿足這些應該被保留的語序。根據上述的方法以及流程，在得到了圖4.5和圖4.6的規則後，我們放進上述的語序重構模型，得到如圖4.7的結果：

[1]In London, cybersecurity is one of the most critical challenges of the digital age. [3]The question now is how to achieve this. [4]With attackers becoming ever more nimble and innovative , armed with an increasingly heterogeneous of weapons , cyber-attacks are occurring at an accelerating pace and with greater sophistication than ever before. [5]As a result, the scale of the challenge should not be underestimated. [2]Everyone from households to businesses to governments, has a stake in protecting our era 's most valuable assets, the data.

Figure 4.7: 語序重構後結果範例

就本範例而言，原始新聞根據第一階段建立的規則，外加上語序重構模型的

doi:10.6342/NTU202303745

轉換，變成了另外一篇闡述內容相同但語序結構不同的新聞。本實驗就是以這樣的形式先初步的對原始新聞做出結構上的變化，而根據此研究架構相關之驗證與評估方式會在下個小節做討論。

4.1.3 語序重構評估結果

根據我們目前所了解，在語序重構的領域中尚未有客觀且具指標性的評估公式，特別是針對語序的前後邏輯關係保留程度尚未有詳細的討論與評估，因此我們在3.4.1小節當中提出的 Reordering Ratio 以及 Enforced Order Preserving Ratio 是本研究為了解決無法對語序重構品質做出評斷之問題所自行提出的兩個公式，不僅是要初步的對本研究的結果做客觀的數據評估，也提供給後續若有相關之研究，可以做為可能的參考指標之一。

在評估時，我們隨機選取了共 50 篇語序重構前後的結果，根據3.4.1小節介紹的評估方式做計算。此外，在公式4.1當中的 ϵ 則是模型中主要的超參數，透過調整 ϵ 可以讓模型獲得不同程度的 Reordering Ratio，也就是平均的語序重構比例。不同的 ϵ 分別所得到的數據如表4.1：

Table 4.1: 語序重構指標評估結果

Reordering Quality			
Indicator	**ϵ=0.4**	**ϵ=0.55**	**ϵ=0.65**
Reordering Ratio	0.79	0.676	0.427
Enforced Order Preserving Ratio	0.623	0.763	0.792

根據表4.1結果可知，當 ϵ 大至 0.65 時，在訓練時被判斷為可以置於前後句的條件較為嚴苛，導致 Reordering Ratio 較小，僅有 0.427，而較小的 Reordering Ratio 則會在 Enforced Order Preserving Ratio 有較佳的表現，達到 0.792。若將 ϵ 設成較小的值，以 0.4 為例，此時 Reordering Ratio 會上升至 0.79，而 Enforced Order Preserving Ratio 則會下降至 0.623。因此，ϵ 在模型中扮演的角色主要是決定文章

doi:10.6342/NTU202303745

平均而言的結構改變度，若使用者希望追求較大程度的結構改變而相對不在意語序邏輯保留與否，則可以設定較小的 ϵ 來達到這個目的，反之，若使用者較重視語序邏輯的保留而較不在意結構改變度，則可以設定較大的 ϵ。在以下所有針對模型的設定及討論討論我們都會將 ϵ 設定為中間值 0.55，而當 ϵ=0.55 時，Reordering Ratio 為 0.676，代表目前系統大約會對文章 67.6%的語句對做出前後對調。另外在 Enforced Order Preserving Ratio 則是 0.763，代表目前重構模型的能力再加上 rule-based 的協助大約可以保留住 76.3%應該具有前後關係的語序對。由此可知，採用目前 rule-based 的方式仍然有部分語句對會出現先後邏輯瑕疵，因此期望未來有致力於研究如何判別兩句話先後邏輯的研究出現，勢必也會對改寫領域帶來非常大的幫助。本小節之指標主要是要對目前的模型結果做客觀的數據呈現，以及供後進者可以在此領域有一個比較標準。

4.2 改寫系統訓練參數

本研究之研究架構的最後一個部分是句對句的改寫系統，在本小節將詳細說明實驗的模型以及相關參數設定。

本實驗所使用之資料集為英文句對句的改寫資料，而預訓練模型我們使用的是 GPT-2 Based Model 進行實驗，在實驗的過程中，我們將 learning rate 設為 1e-5，batch size 設為 8，共訓練 30 個 epoch。透過多次的訓練嘗試，我們整理出了上述的參數設定，並訓練出了本研究第三階段的句對句的改寫模型，使得輸入的新聞同時滿足結構上的改變以及語句上的改寫。

doi:10.6342/NTU202303745

4.3 改寫系統結果分析

在本小節，我們將針對整個研究模型從原始新聞產出改寫新聞的結果進行展示以及分析，整體的流程經歷了將原始新聞翻譯、ruled-base 尋找語序重構規則關鍵字、新聞語序重構、逐句放入改寫，最後轉譯回中文得到改寫過後版本的新聞。我們也將在此小節對目前的實驗結果做簡單分析，討論產出的新聞在改寫議題中的優缺點。此外，在 baseline model 的選擇說明如下：

1. Backward Translation : 回譯法是在2.2節中有提到的經典改寫方法之一，是一個在早期看似直觀卻非常主流的方法，因此本研究將此方法納入比較對象之一。在此方法我們會將原始中文新聞翻譯成英文版本，而後再將英文版本翻譯回中文，得到的結果便是此方法的改寫成果。

2. BT+ChatGPT : BT 代表的是 Backward Translation，本方法會將回譯法結合由 OpenAI 於 2022 年提出的 ChatGPT。整體的改寫流程會是先將原始新聞先翻譯成英文版本，而後將英文版本輸入至 ChatGPT 中進行改寫版本英文新聞，最後再將 ChatGPT 的改寫結果翻譯回中文得到這個最後的結果。

3. ChatGPT(ch2ch) : 本方法同樣是透過 ChatGPT 進行改寫，改寫的流程是直接將中文新聞輸入至 ChatGPT 並要求 ChatGPT 直接改寫，得到的結果就是本方法的改寫成果。

此外，在要求 ChatGPT 對一篇新聞改寫時，我們給予的指令為：「幫我改寫以下這篇新聞」，並輸入新聞內容。比較結果如圖4.8所示:

由圖4.8之結果可以發現，Backward Translation 不論在語句的改寫程度或是整體新聞的結構改變度都是明顯最小的，由於翻譯工具僅單純用於語言轉換，並沒

doi:10.6342/NTU202303745

<table>
<tr><th>原始新聞</th><th>Backward Translation</th><th>BT+ChatGPT</th></tr>
<tr><td rowspan="3">在倫敦，網路安全是數位時代最關鍵的挑戰之一。從家庭、企業再到政府，保護我們這個時代最有價值的資產，數據，與每個人都息息相關。目前的問題是該如何實現這一目標。隨著攻擊者變得越來越靈活和創新，擁有越來越多樣化的武器，網路攻擊正在以前所未有的速度和複雜度發生。因此，挑戰的程度不應被低估。</td><td>在倫敦，網路安全是數字時代最嚴峻的挑戰之一。從家庭到企業再到政府，保護我們這個時代最寶貴的資產—數據對每個人都很重要。現在的問題是如何實現這一目標。隨著攻擊者變得更加靈活和創新，使用更多樣化的武器，網路攻擊正以前所未有的速度和複雜性發生。因此，不應低估這一挑戰的嚴重性。</td><td>在數位技術的動態時代，保護網路安全是倫敦的首要挑戰。維護我們最寶貴的資產—數據，已成為家庭、企業和政府的集體責任。然而，緊迫的問題仍然存在：我們如何才能有效地完成這項任務？隨著網路領域的對手變得更加敏捷和富有創造力，配備了各種各樣的工具庫，網路攻擊的頻率和複雜性都激增。這種不斷升級的趨勢要求我們謹慎地承認眼前挑戰的嚴重性。</td></tr>
<tr><td>ChatGPT(ch2ch)</td><td>本研究改寫新聞</td></tr>
<tr><td>在倫敦，網路安全是數位時代最大的挑戰之一。數據已成為我們這個時代最有價值的資產，不論是家庭、企業或政府都必須與之相關聯。然而，實現這一目標卻面臨著重大的問題。網路攻擊者日益靈活創新，手中擁有越來越多樣化的武器，使網路攻擊以前所未有的速度和複雜度發生。因此，挑戰的難度絕不能被低估。</td><td>在倫敦，網路安全是數位時代的決定性挑戰之一。現在的問題是如何實現它。隨著攻擊者變得越來越敏捷和創新，擁有越來越多的武器，網路攻擊正在加速發生，並且比以往任何時候都更加複雜。因此，不應低估挑戰的規模。從家庭到企業再到政府，每個人都與保護我們這個時代的優質商品數據息息相關。</td></tr>
</table>

Figure 4.8: 改寫結果範例合併

有針對改寫任務做特別的訓練，因此在改寫程度上一定是相對有限的。

而在採用了目前最廣為人知也最強悍的 ChatGPT 模型，不論是 BT+ChatGPT 的方式或是 ChatGPT(ch2ch) 的方式，對於新聞改寫目前也是以和原新聞相同的敘事結構為基礎，並透過同義字詞的置換，或是單純的改變詞序來美化原始句，表達相同意義的新聞事件。相較之下，本研究之實驗結果由於加入了改良版的語序重構技術，在改寫前有先對新聞做出適度的結構重組，因此縱使若單純比較字句改寫上的品質或是用詞的優美度等方面可能因句而異，不一定可以每句話都超過 ChatGPT 的表現，然而將比較放大至以整篇新聞為單位，本研究之結果採用 ChatGPT 的方法所得之結果會更有結構上的差異，更像是一篇全新的新聞。

doi:10.6342/NTU202303745

4.4 自動評估結果

在自動評估的部分，我們使用的是 ROUGE-L 以及 WER 兩個指標，主要用於判斷以句為單位的改寫前後有多大程度的改變，用以衡量本研究結果和 Baseline Model 之間客觀上的差異。詳細的數據結果如表4.2所示。

Table 4.2: 自動評估結果表

Model	ROUGE-L	WER
Backward Translation	0.56	0.213
BT+ChatGPT	0.288	0.279
ChatGPT(ch2ch)	0.428	0.245
本研究	**0.313**	**0.287**

在計算過後我們發現，抽樣超過 150 句話計算出來的結果便會收斂並不再有大幅度的增減，因此我們選擇抽樣 200 句話並計算出如表4.2的結果。從比較結果可知，本實驗所提出的研究方法在針對中文新聞的改寫有一定程度的能力，從 ROUGE-L 指標來看，也就是平均每一句話在改寫前後有多少比例的連續相同字，我們的結果和最小值，也就是 BT+ChatGPT 之間僅差距約 2.5%。此外，從 WER 指標來看，也就是計算平均每句話當中有多少比例的字詞被改變或是刪除，本研究之結果更是獲得最高的分數。由此可知單從字詞上的改變程度來說而言，本研究透過語種轉換以及訓練後的改寫模型搭配所得，對於句對句的中文新聞語句之改寫結果在兩項指標上皆有和目前最主流的 ChatGPT 模型相去不遠的效果，藉此結果確保本研究有對原始新聞語句做到一定程度的改寫，也證明本研究之實驗方法現階段對於中文改寫領域有實質上的提升。

doi:10.6342/NTU202303745

4.5 人工評估結果

在前一小節我們對於本實驗在改寫程度方面用現有的指標做出了自動評估的驗證，也分析了和 Baseline Model 之間的比較，然而目前對於改寫品質方面尚未做出衡量與比較，故在本小節，也就是人工評估的部分，我們將會用系統化的方式對改寫品質做評估，做為本實驗結果的評斷依據。

於人工評估階段，我們總共徵求了十位受測者，每位受測者至少都具備大專院校以上的學歷，確保其有一定程度的媒體識讀能力以及對於評估問題的思辨能力，足夠代表一般民眾在閱讀網路新聞時會有的想法以及感受。此外，我們從測試資料集中隨機選擇了 30 篇新聞，並使用了本實驗的模型以及各個 Baseline Model 分別生成改寫後的新聞，每個人會被分配到 15 篇新聞，需要受測者針對各個改寫結果在新聞可讀性、新聞改寫程度、新聞邏輯性，新聞結構改變度這四個指標進行評分，一分為最不符合指標之描述，三分為普通，而五分為最符合指標之描述。此外，在評估的過程中全程採用的是 Blind Test，也就是受測者僅會對各個改寫新聞評分，並不會知道該結果是由哪個方法產生的新聞。各項指標的問題描述方式如表4.3所示，最後將分數結果呈現在表4.4。

Table 4.3: 人工評估指標描述方式

	詢問方式
可讀性	本篇改寫新聞敘述流暢且可以順利閱讀。
改寫程度	本篇改寫新聞與原新聞在字句上有明顯的差異。
結構改變度	本篇改寫新聞和原始新聞在整體結構上有明顯差異，而非只是逐句改寫。
邏輯性	本篇改寫新聞有保留住原始新聞事件的前後邏輯關係。

從表4.4的人工評估結果可以發現，本研究的結果在改寫程度以及結構改變度兩個方面獲得了明顯高於 Baseline Model 的分數，因此可以推斷本研究採用的語序重構技術不僅能使文章有整體結構上的改變，而因為結構的改變可能使受測

doi:10.6342/NTU202303745

Table 4.4: 人工評估結果表平均分數

Model	可讀性	改寫程度	結構改變度	邏輯性
Backward Translation	4.007	2.447	2.007	3.927
BT+ChatGPT	4.207	2.94	2.193	3.933
ChatGPT(ch2ch)	4.213	2.887	2.153	4.027
本研究	**3.353**	**3.34**	**3.46**	**3.233**

者在閱讀的過程感覺到字詞上的改變較大，也因此這兩個指標會獲得明顯優於 Baseline Model 的分數。然而，受制於目前對於語序重構時的語序前後邏輯僅能以偵測有無連接副詞的方式保留，因此勢必還是會有出現本研究無法覆蓋到的情況，使本研究重構後的新聞出現可讀性以及邏輯性的瑕疵，因此沒有對文章結構重組的其他 Baseline Model 整體而言就會在可讀性以及邏輯性上獲得較佳的分數。

此外，表4.5呈現了各個模型在四個指標當中所獲得分數的標準差。

Table 4.5: 人工評估結果標準差

Model	可讀性	改寫程度	結構改變度	邏輯性
Backward Translation	0.815	0.863	0.823	0.778
BT+ChatGPT	0.678	0.876	0.800	0.774
ChatGPT(ch2ch)	0.774	0.863	0.757	0.802
本研究	**0.998**	**1.048**	**1.202**	**1.026**

從表4.5結果可以觀察到，本研究的模型在各項指標的標準差都是最大的，可以推斷本研究的模型所產出的新聞結果在各項指標所獲得的分數彼此差異較大，此情況乃是由於每一篇文章被語序重構的程度不盡相同，因此在改寫程度和結構改變度兩指標的得分會有較大的差異，而不同的改寫程度也間接影響到了可讀性以及邏輯性的得分差異較大，從而導致此情況。

另外，我們也用了四項指標的得分各別做了單因子變數分析 (One-Way ANOVA)，藉此驗證各個模型在該指標的得分是否有顯著不同，分析結果如

doi:10.6342/NTU202303745

表4.6。

Table 4.6: 人工評估指標 ANOVA

Indicator	P-Value
可讀性	1.3525E-21
改寫程度	1.3258E-14
結構改變度	1.7147E-44
邏輯性	6.80066E-17

由表4.6的結果可以發現，四項指標的顯著性 P 值皆小於 0.05。舉可讀性為例，由表中資訊可以發現可讀性的 ANOVA 檢測結果 P-Value 為 1.3525E-21，小於 0.05，表示四個模型在可讀性的得分是有顯著不同的，其他的指標則以此類推有一樣的結果。

最後，為驗證上述對於人工評估結果之推測，我們額外對於分數做了相關性的檢測。在數據的部分，我們結合了十位受測者的數據，由於每位受測者會負責 15 篇新聞的評分，且每篇新聞共會有 4 種改寫結果需要評分，因此結合這全部總計 600 筆的資料，每筆資料皆有可讀性、改寫程度、結構改變度以及邏輯性 4 個評估指標的分數。我們對於 4 個指標計算彼此的皮爾森積差相關係數，藉此確認各個指標的結果是否存在相關性，結果如下圖4.9所示。從圖4.9的結果可觀察到以下幾點：

1. 改寫程度的結構改變度的得分有最大的正相關，這個結果也呼應了前述對表4.4結果的推斷，當新聞出現了較大的結構改變時，因為改寫新聞的語句順序已經和原始新聞出現明顯落差，會使受測者閱讀時感受到較大的字詞改變。反之，當新聞結構沒有做出改變時，則閱眾對於改寫程度的感受度也會下降，分數也隨之較低。

2. 可讀性和邏輯性的得分有較大的正相關，這個結果主要顯示若改寫過程沒有

doi:10.6342/NTU202303745

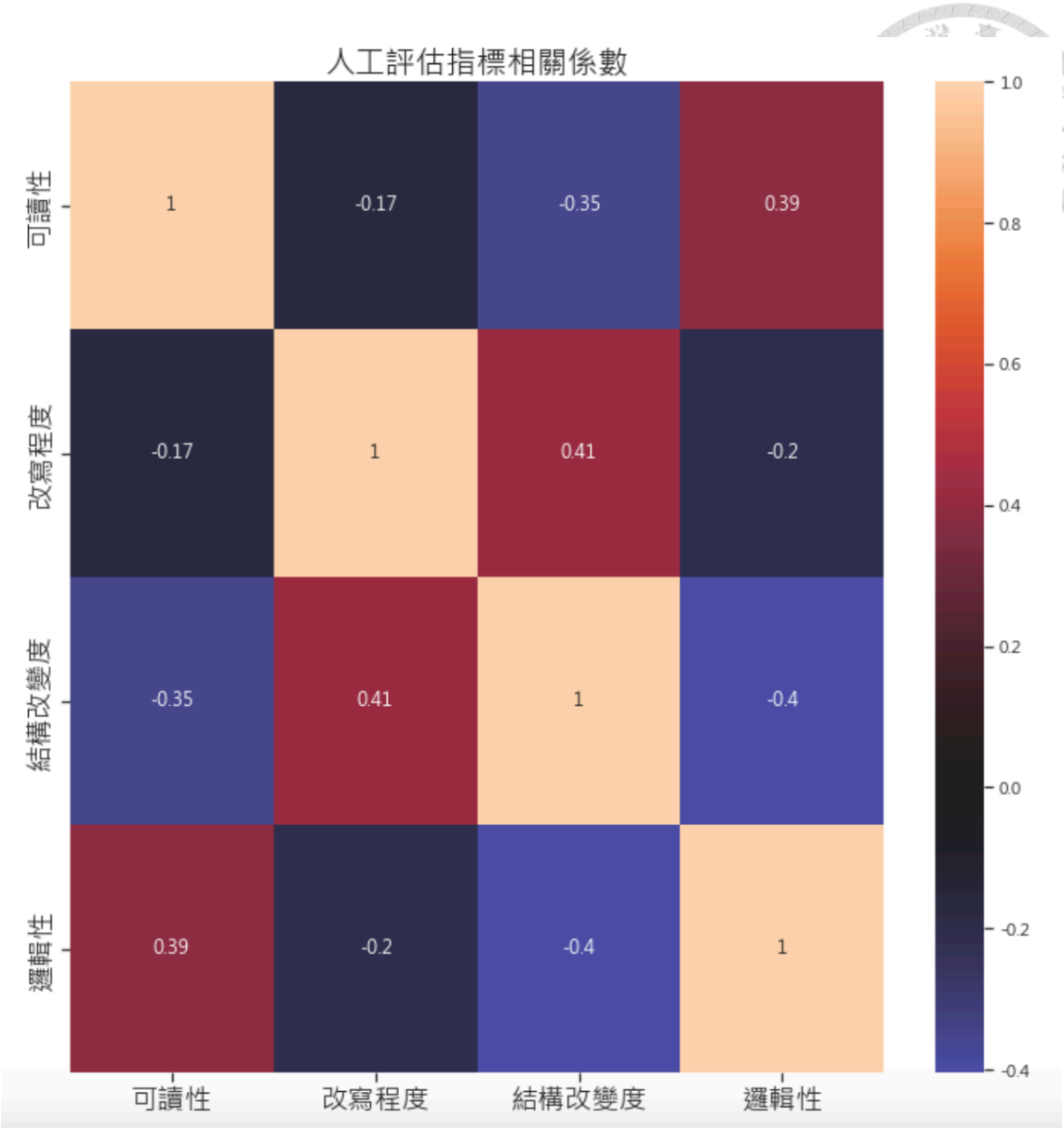


Figure 4.9: 人工指標相關係數 Heatmap

做結構改變僅字詞轉換，則可讀性以及邏輯性皆會獲得較高的分數。反之，若加上了對新聞結構的改變，則某些新聞可能會在改寫後出現前後事件的錯置，導致邏輯性分數較低，進一步影響新聞的可讀性表現。

3. 結構改變度和可讀性、邏輯性有較大的負相關，這個狀況也呼應了第二點提到的原因，解釋並驗證了結構改變對於改寫新聞的可讀性和邏輯性的影響。同時，我們也可以解釋為：結構改變和可讀性與邏輯性彼此呈現 Trade-off 關係，每當有其中一遍提升，另一邊就會隨之下降，反之亦然。

doi:10.6342/NTU202303745

在4.1.3小節當中，曾經提到公式4.1中 ϵ 的改變對於文章結構的改變程度的影響，越大的 ϵ 會對應到越小的 Reordering Ratio，同時 Enforced Order Preserving Ratio 會提升。因此，使用者可以根據自己的需求調整 ϵ，若是追求大幅度的結構改變而較不在意邏輯順序的保留，則可以設定較小的 ϵ，而若是強調邏輯順序的保留而較不追求結構改變，則可以設定較大的 ϵ。而無論 ϵ 如何調整，呈現的 Reordering Ratio 和 Enforced Order Preserving Ratio 的變化皆符合上方第三點的結果：結構改變度和可讀性、邏輯性有明顯的負相關。

4.6 小結

本研究採取了複合型的研究方法，首先對翻譯過後的新聞進行 rule-based 的語序規則保留以及語序重構，之後再以句為單位輸入到我們使用整理過後的英文句對句改寫資料集，並用 GPT-2 模型所訓練出來的改寫模型當中，最後將改寫後的結果組合成一篇新的新聞，最後再翻譯回中文得到整個改寫後的新聞。在本章節當中除了顯示我們的實驗結果外，我們也透過許多自動評估指標、包含了部分我們自行提出的指標，外加上人工評估做為我們研究之佐證，藉此驗證本研究方法在中文新聞改寫的效果在客觀數值上的表現以及人類評估的結果。

在4.1節當中詳細的說明了我們對現有的語序重構模型加入的規則，為的就是希望可以在語序重構的同時盡可能的保留住語序之間的前後邏輯關係，而我們也特別在4.1.3小節當中計算了 Reordering Ratio 以及 Enforced Order Preserving Ratio 兩個自定義的衡量指標，提出了現階段對語序重構程度以及語序邏輯關係保留程度兩個部分的衡量方式，提供後進者一個衡量實驗結果之標準。

於4.2節中，我們詳細列出了改寫模型的訓練參數，並於4.3節中呈現了整個模型的輸入新聞以及輸出的改寫新聞，並同時列出了 Baseline Model 的改寫結果，

doi:10.6342/NTU202303745

相比較後可以發現到：本研究的結果相較於 Baseline Model 明顯多了文章結構上的變化，由此驗證本研究之方法可以在新聞改寫時對新聞做出結構上的重組使改寫效果更好。此外，若是單純就字詞上的改寫程度去做比較，從4.4節的自動評估指標結果可以發現本研究的結果在平均字詞改寫程度和我們主要的比較對象 ChatGPT 有差不多的結果。最後的4.5小節呈現的結果說明了目前文章改寫的困境，現階段對於文章改寫的方法若是和本研究一樣採用語序重構技術，相對於 Baseline Model 的一句一句改寫會有更好的改寫程度表現，然而在語序前後邏輯的保留上則會是一個重大的討論議題，也可能在某些結果上間接造成了文章邏輯的瑕疵。根據對改寫結果的觀察以及在人工評估分數的統計數據，目前可以得出文章整體的改寫程度和文章的邏輯性互為 Trade-off 關係，在追求高改寫程度的同時勢必會犧牲掉邏輯性的表現，反之亦然。因此，如何在語序重構時兼顧語序前後邏輯便會是這個領域重要的未來研究方向。

doi:10.6342/NTU202303745

第五章　結論

本章節我們將綜合前四章所提出的內容以及研究結果，統整出整體研究成果，並列點說明本研究主要的貢獻以及所面臨的研究限制，最後將會提出本研究的未來展望，為後進研究者提供研究方向。

5.1　研究成果

本研究的目的主要是為減輕現今媒體工作者的工作壓力，透過模型的幫助快速的產生多樣化版本的改寫新聞，為記者提供可以發布於不同平台上的新聞內容，藉此提升記者在現今新媒體世代的競爭力。針對過往的研究進行探討後發現，在現有的文獻中，改寫領域的發展絕大多數都是以外文為主，中文領域、特別是繁體中文鮮少有文獻曾討論過。此外，在改寫的探討領域與可用資料目前也絕大多數是以句為單位的改寫，而沒有以整篇文章為單位的文章改寫，因此，本論文探討中文新聞篇章之改寫。本研究透過跨語種的轉換來突破中文改寫資料不足，以及沒有以中文新聞為主題之困境，並利用擁有強大能力的 GPT-2 預訓練模型，嘗試為中文改寫做出貢獻。此外，為了讓改寫並不只是停留在一句話，我們加入了現有的語序重構技術，使得改寫前後的文章足以呈現出結構上的明顯變化，藉此提升新聞改寫的品質。

此外，在套用語序重構模型時，我們也額外處理了該模型並不能穩定的保留

doi:10.6342/NTU202303745

著各個語句之間的前後邏輯順序的問題，嘗試用語言學的角度分析歸納出關鍵字，並透過 ruled-based 的方式精進了語序重構前後的表現，對於改寫文章的品質有相對更大的保障。另外，我們也提出了嶄新的指標，用以衡量文章改寫前後的語句重構程度，以及對於語序前後邏輯的保留程度，根據我們目前的了解，這兩個面向在過去並無約定俗成的指標，因此我們決定自行訂定標準，以利未來若有研究人員希望能對語序重構這個領域做出新的研究時，可以做為客觀的研究結果評斷依據之一。

由上述可知，本論文採用的是複合式的研究架構，而透過自動評估指標以及人工評估的結果可知，在改寫程度的部分，本研究是可以達到和 Baseline Model 相當的程度，而在篇章結構的改變上是明顯優於 Baseline Model，由此可知本研究之結果有一定的實用性與可參考性，而本研究的架構也可為後續相關研究參考。若未來語言模型發展得更為成熟與強大，為改寫領域提供更便利以及更省成本的蒐集資料方式，相信改寫領域一定能有更多發展方向與進步空間值得後人深入研究探討。

5.2 研究貢獻

在本研究的主要研究貢獻如下：

1. 中文新聞領域改寫：就我們目前對於過往的文獻所知，本論文是首篇結合自然語言模型並以中文新聞改寫為主題之研究，在使用複合式的模型為不只是中文新聞，甚至是為其他領域中文文章之改寫提供新的參考研究方向。根據研究結果可以發現，我們提出的方法在人工評估和自動評估的各項指標上都有相當接近甚至超越 Baseline Model 的得分。也說明了在改寫領域當中，本研究所提出的研究方法是可以有與當今最強悍也最火紅的語言模型相等甚至

doi:10.6342/NTU202303745

更好的效果。

2. 以篇章為單位之改寫：在中文改寫領域當中，除了本身就面臨資料不足的窘境之外，在資料的型態也是問題之一。由於以句為單位的資料就已經相當稀少且難以蒐集，更遑論以篇章為單位的改寫前後資料，現實的情況也和本研究目前所知一樣，中文領域尚無探討以篇章為單位之改寫研究。而本研究透過跨語種的轉換以及結合語序重構的技術，首次為中文、篇章為單位的改寫領域提供了一個可能的研究方法，有效為現階段受制於資料量與型態之改寫研究過渡期提供了暫時的解決方法。

3. 精進語序重構並提出評估指標：根據前述提及的研究方法，本研究不僅僅是借助了現有的語序重構技術來對原始新聞做出篇章改寫，同時針對了模型對語序重構時無法保留語意前後關係的問題進行探討，從語言學的角度分析並訂出規則，嘗試更加改善語序重構時的品質。此外，本研究也同時提出了自定義的指標來對目前的語序重構品質做出客觀的衡量，同時也提供未來有興趣更精進語序重構的後進研究者提供一個可以評斷研究結果的衡量指標。

5.3 研究限制

本研究的過程主要包含了兩大研究限制，分別是與新聞相關的改寫資料集數量相當有限，特別是中文領域的改寫資料更是如此。此外，改寫資料本身就非常難以蒐集且成本非常高。以下將分成兩點進行深入的探討：

doi:10.6342/NTU202303745

5.3.1 資料量之限制

本研究之主軸雖然是放在中文新聞之改寫，然而在實驗架構的部分仍是以英文資料集做訓練模型，並利用跨語言轉換的方式達到中文新聞之改寫。主要原因是中文領域當中與改寫相關的資料數量非常少，且多半是以簡體中文為主，每筆資料之長度與品質也多半難以掌控，若是再加上需要與新聞相關這個條件的中文改寫資料，就本研究目前為止的了解並沒有大量且高品質的相關資料可供訓練大型語言模型之用，也因此本研究才會採取使用跨語種的方式達成為中文新聞改寫的研究目的。此外，由於新聞之主題與種類千變萬化，而在訓練階段給予模型之訓練資料相對來說還是比較有限，因此沒辦法讓模型學到所有面向之新聞以及新聞當中的用語等等，因此目前可還無法對與訓練資料差異過大的用詞或語句做出相對應的改寫，亦或是改寫的效果會不甚理想。在未來研究方向更加多元，或是 ChatGPT 以及 GPT-4 等資源開放程度更高，提供研究人員可以快速擴增資料集的同時，相信改寫領域受到資料量不足之限制，以及各類別、領域之新聞不夠豐富齊全等問題也會隨之獲得解方，成為未來研究方向之一。

5.3.2 改寫資料難以大量蒐集

在改寫領域裡，最大的痛點除了資料量不足以外，另一個問題就是改寫資料並非可以在短時間內大量蒐集而得。首先，「改寫」本身就是一個沒有標準答案的內容，也沒有客觀的指標說明對或錯，每個人對於改寫的標準都不盡相同，也因此，改寫的資料本身就是難以被制式化定義的內容，因此要以統一的標準蒐集到大量的資料本身就相當困難。此外，若是真的要花時間進行資料的蒐集，假設是用人工的方式蒐集到與中文新聞相關的改寫內容，所花費的時間或是金錢成本會不容小覷，所得到的改寫結果之品質也難以統一。而若是要以現有模型之協助蒐

doi:10.6342/NTU202303745

集，以現今最火紅的 ChatGPT 為例，因為改寫資料內容本身就是非常巨量連續的自然語言，由於 ChatGPT 目前尚未開放研究人員可以大量連續的使用，仍有一定的用量與流量限制，因此若是需要蒐集到足以訓練大型語言模型的資料量，所需花費的金額會相當可觀，而若是不使用 ChatGPT，目前也沒有開源且具公信力的改寫模型可用以蒐集改寫文本。由上述的討論可知，改寫資料所面臨的除了沒有標準答案，短時間內無從大量蒐集的管道也是造成改寫資料非常稀少的主要原因。因此在未來類似 ChatGPT 的大型語言模型更為蓬勃發展，開放程度更高後，中文改寫領域的研究一定會有更豐富的資料集可以使用，研究方向與方法也一定會有更大的突破。

5.4 未來研究方向

本研究的目的在於協助現今臺灣媒體從業者面對新媒體時代所面臨的強烈競爭所建立出的新聞改寫系統，透過複合式的研究架構讓模型能夠在取得原始文本的情況下改寫出另一個版本的新聞，以下將敘述在未來可能可以探討的研究方向，使改寫領域可以有更好的發展：

1. 語序重構模型的精進：由於目前尚未有可以明確判別出文章中哪些語句之間有明顯前後關係需要保留的模型或方法，也因此語序重構模型沒有辦法根據系統化的方法被精進。在本研究中採取的是建立規則（ruled-based）的方法試圖提升語序重構的效果與品質，未來若有針對語序前後邏輯關係判讀的研究或模型釋出時，現階段的語序重構模型便可以藉此得到更進一步的改善，確保在將一篇文章語序重構的同時可以兼顧到句子之間的前因後果，使改寫領域再向前走。

2. 篇對篇的資料與訓練方式：誠如5.3小節所述，中文改寫領域目前遇到的最

doi:10.6342/NTU202303745

大瓶頸絕對是資料量的不足以及蒐集成本過高。然而隨著目前 AI 時代的發展，越來越多強悍的語言模型或聊天機器人問世，而在激烈的競爭之下，各大企業的產品勢必會開放程度越來越高，生成語料的成本也會越降越低，屆時中文改寫領域資料量不足的問題或許就可以獲得解決。此外，若這些強大的語言模型可以提供研究人員在短時間內大量蒐集自然語料，在中文改寫的領域甚至可以不用單純停留在一句話為單位的等級，而是可以不斷地給模型成篇的文字內容，透過模型生成出篇章等級的改寫結果，並在大量蒐集資料之後訓練出更聚焦於改寫主題、且成本更低的語言模型。

3. 其他主題之文章改寫：在本研究中主要探討的是以新聞為主題的改寫，然而如2.2小節中所提到，改寫的應用場域非常的廣泛，如問答系統、語意解析、搜尋引擎的關鍵字重構等等常見的狀況都可能會需要改寫系統的協助，由此可知改寫系統的泛用性與重要性十足。然而這些領域的改寫發展目前所遇到的困難同樣是資料量的不足以及難以獲取，因此當大型語言模型更為開放解決資料蒐集的問題之後，上述的這些改寫主題勢必也可以投入更多的研究能量，解決相對應的問題。

4. 更高階的預訓練模型使用：在本研究中選擇的大型預訓練模型為 GPT-2，乃是因為本研究主要的 baseline 是同樣是以 GPT 系列模型訓練而得的 ChatGPT，希望能用相對低成本版本的模型搭配其他方法試圖達到與之相似的效果。然而在未來 GPT-3、GPT-3.5 甚至是 GPT-4 開放程度更高，且硬體設備也發展到足以支撐這些極大規模的預訓練模型時，或許在改寫領域的問題就會受到更多的討論，獲得更好的研究結果。

doi:10.6342/NTU202303745

參考文獻

[1] C. Bannard and C. Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL' 05)*, pages 597–604, 2005.

[2] R. Barzilay and L. Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *arXiv preprint cs/0304006*, 2003.

[3] J. Berant and P. Liang. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, 2014.

[4] R. Cao, S. Zhu, C. Yang, C. Liu, R. Ma, Y. Zhao, L. Chen, and K. Yu. Unsupervised dual paraphrasing for two-stage semantic parsing. *arXiv preprint arXiv:2005.13485*, 2020.

[5] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174, 2018.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of

doi:10.6342/NTU202303745

deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] L. Dong, J. Mallinson, S. Reddy, and M. Lapata. Learning to paraphrase for question answering. *arXiv preprint arXiv:1708.06022*, 2017.

[8] S. Gao, Y. Zhang, Z. Ou, and Z. Yu. Paraphrase augmented task-oriented dialog generation. *arXiv preprint arXiv:2004.07462*, 2020.

[9] W. A. o. N. P. W.-I. Germany-based consulting group Schickler. Ai' s rising role with editing and reader revenue, 2022.

[10] T. Goyal and G. Durrett. Neural syntactic preordering for controlled paraphrase generation. *arXiv preprint arXiv:2005.02013*, 2020.

[11] A. Gupta, A. Agarwal, P. Singh, and P. Rai. A deep generative framework for paraphrase generation. In *Proceedings of the aaai conference on artificial intelligence*, volume 32, 2018.

[12] K.-H. Huang, C. Li, and K.-W. Chang. Generating sports news from live commentary: A chinese dataset for sports game summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 609–615, 2020.

[13] J. Kanerva, S. Rönnqvist, R. Kekki, T. Salakoski, and F. Ginter. Template-free data-to-text generation of finnish sports news. *arXiv preprint arXiv:1910.01863*, 2019.

[14] A. Kumar, S. Bhattamishra, M. Bhandari, and P. Talukdar. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation.

doi:10.6342/NTU202303745

In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619, 2019.

[15] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

[16] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[17] L. Leppänen, M. Munezero, M. Granroth-Wilding, and H. Toivonen. Data-driven news generation for automated journalism. In *Proceedings of the 10th international conference on natural language generation*, pages 188–197, 2017.

[18] Z. Li, X. Jiang, L. Shang, and H. Li. Paraphrase generation with deep reinforcement learning. *arXiv preprint arXiv:1711.00279*, 2017.

[19] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[20] Z. Lin, Y. Cai, and X. Wan. Towards document-level paraphrase generation with sentence rewriting and reordering. *arXiv preprint arXiv:2109.07095*, 2021.

[21] X. Liu, Q. Chen, C. Deng, H. Zeng, J. Chen, D. Li, and B. Tang. Lcqmc: A large-scale chinese question matching corpus. In *Proceedings of the 27th international conference on computational linguistics*, pages 1952–1962, 2018.

doi:10.6342/NTU202303745

[22] K. McKeown. Paraphrasing questions using given and new information. *American Journal of Computational Linguistics*, 9(1):1–10, 1983.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[25] Y. Scherrer. Tapaco: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA), 2020.

[26] G. H. Song and Y. Wang. Paraphrase generation with chinese short text dataset. In *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*, pages 60–64. IEEE, 2020.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] S. Witteveen and M. Andrews. Paraphrasing with large language models. *arXiv preprint arXiv:1911.09661*, 2019.

[29] M. Wong. Conjunctive adverbs, 2021. `https://reurl.cc/GAaWE3`.

[30] Y. Yang, Y. Zhang, C. Tar, and J. Baldridge. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*, 2019.

doi:10.6342/NTU202303745

[31] J. Zhou and S. Bhat. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5075–5086, 2021.

[32] S. Zhu, X. Cheng, S. Su, and S. Lang. Knowledge-based question answering by jointly generating, copying and paraphrasing. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2439–2442, 2017.

doi:10.6342/NTU202303745