

計畫編號：

國立中央大學前瞻科技研究中心 113 年度專案研究計畫

「人物關係擷取模型開發之研究」

期 中 報 告 書

※報告書請加入頁數，雙面列印裝訂，紅字請於印製時刪除。(期中期末審查會後請於兩星期內交付報告書及光碟其內含報告書、簡報、研發之相關程式軟體)

計畫主持人：張嘉惠教授

研究人員：洪閔昭

計畫期程：113 年 1 月 1 日 至 113 年 11 月 30 日

目 錄

- 一、 計畫概要
- 二、 計畫目標與預期成果
- 三、 研究程序步驟與方法
- 四、 執行成果
- 五、 成果交付項目
- 六、 檢討與建議
- 七、 結語
- 八、 附錄

一、計畫概要(背景及需求說明)

(研究動機、研究期程、研究人力)

1. 研究動機

隨著網際網路的迅速發展，大量的非結構化數據持續湧現，許多中國政商人物之間的關係也隱藏在這些網路資訊當中，因此從這些數據中擷取出有用的關係訊息變得愈發重要。我們希望能夠找出一套聯合實體關係擷取的架構，讓我們在能夠在真實的網路數據中，全面探索中國政商人物之間的關係，以大幅增進國家安全情報收集與分析的能力。

然而現今的關係擷取及命名實體類別任務資料集，通常來源都是單一特定的資料庫，例如：維基百科、特定新聞網站、文學作品集，因此，訓練出來的 AI 模型較難泛化到真實世界中五花八門的網路內容上。為了要讓聯合實體關係擷取的任務，能夠直接的使用在各式的網頁上，我們必須先從資料集著手。

另外，目前在關係擷取任務的主要資料集，例如:ACE05 [1]、CoNLL [2]、NYT [3] 等資料集，都是屬於 Sentence 級別的任務，模型只需要在較短的語句或段落中找出實體以及關係即可。但其實在現實的情況下，實體關係三元組中的實體以及關係，並不一定會同時出現在同一個句子，或是相鄰的句子中。根據統計[4]，超過 40.7%的關係只能在文章級別上被識別出來。因此模型必須閱讀完整的文章，才能找到這些跨句子或者跨段落的實體關係。

因此，在這個計畫中，我們將聚焦於開發高效的人物關係擷取模型，專注於從多來源文本中提取實體和關係資訊，並使用文章級別方式探索及優化實體識別和關係擷取技術。透過自動化的政商人物關係擷取和，大幅提高國安單位對於關鍵人物的了解和把握，極大提升情報搜集的效率與精確性，進而強化對潛在風險的預警能力，有效維護國家安全與社會穩定。

2. 研究期程

專案執行期程：自 113 年 1 月 1 日起至 113 年 11 月 30 日止，共 11 月。進度甘特圖執行如下表：

次 工作項目	第 1 月	第 2 月	第 3 月	第 4 月	第 5 月	第 6 月	第 7 月	第 8 月	第 9 月	第 10 月	第 11 月
專案需求規格確認	■	■									
大數據網路資料清洗		■	■								
聯合實體關係擷取研究			■	■							
大型語言模型標記架構設計				■	■						
實體擴充標記研究					■	■					
關係擷取模型訓練						■	■				
期中報告							■				
關係擷取模型優化研究							■	■			
系統整合									■	■	
教育訓練										■	
期末報告											■
預定進度累計百分比	10	22	34	46	58	70	82	94	97	99	100

3. 研究人力

類別/級別	姓 名	工作月數	在本研究計畫內擔任之具體 工作性質、項目及範圍
主持人	張嘉惠	11	進行計畫管理、演算法設計
兼任碩士研究助理	洪閔昭	11	進行系統設計及開發、文件撰寫
兼任碩士研究助理	葉季儒	11	進行系統設計及開發、文件撰寫

二、計畫目標與預期成果

本計畫旨在基於人物關係擷取的人工智慧模型建立，透過 AI 全面探索中國政商人物之間的關係，以大幅增進國家安全情報收集與分析的能力。政商人物之間的複雜關係對於國安的維護與情報預警至關重要，然而這些關係時常散布於大量非結構化文本資料中，如新聞報導、社群媒體等，因此需要運用自然語言處理技術，自動擷取並建構知識圖譜，以解析這些密集而多樣的人物關聯。

現今對於關係擷取的研究多侷限於特定數據庫來源，難以全面捕捉網路上的各類型資料。因此，本計畫將以真實網路資料為出發點，並通過大型語言模型的架構設計，提供自動化的擷取與標記機制，實現中文人物關係的自動化流程。

最終，我們將利用該資料集訓練自有的人物關係擷取模型，使其達到最佳評估效能，從而使人物關係擷取自動化技術能夠自主運行。

本計畫完成時各工作項目以及其分別之預期成果分述如下：

1. 針對人物關係擷取進行研究，說明具體實驗結果及相關評估指標。
2. 研究大型語言模型在關係擷取任務上的突破：探討其在關係擷取任務上的改進可能。
3. 建構一套自有模型系統雛型，展示相關功能及效能。
4. 本計畫之預期成果，將提供研究方法過程中所有研發之原始程式碼，並建構乙套系統雛型以展示相關功能成效，及乙份研究報告，主要內容說明如下：
 - 提供進行「人物關係擷取模型開發之研究」之軟體需求規格建議及系統建置應內含所需之各項 Open Source 安裝套件軟體之建議。
 - 研究大型語言模型在關係擷取任務上改善方式，提供各個方法之間精準度、時間、所需資料量等客觀量化指標差異性比較。
 - 提供人物關係擷取之效能基準（benchmark）評估，以作為後續硬體設備應用與擴充之參考依據。

三、研究程序步驟與方法

(含研討合作機制與事項)

1. Common Crawl 數據庫前處理

在關係擷取的領域中，現有的資料集往往存在著來源方面的限制，主要來自特定的資料庫。例如，NYT(New York Times Annotated Corpus) 的內容僅來自《紐約時報》的檔案，而 DocRED[5] 則是從維基百科和維基數據中彙編而來。此外，像是專門領域的資料集，如 GDA (基因-疾病關聯語料庫) 和 CDR (化學物質-疾病關聯)，則主要來自於 PubMed 等生物醫學文獻庫。

為了創建一個更具彈性和全面性的資料集，我們的研究旨在跨越特定領域的界限。我們的目標不僅僅是擷取，更希望為未來多元實體關係擷取任務奠定基礎。為了實現這一目標，我們從網絡的廣大的網頁文章中彙編我們的資料集，我們利用了 Common Crawl，這是一個包含各種領域和文章類別的網頁存檔。這些網頁涵蓋了多種寫作風格、觀點和主題，使我們的資料集在本質上具有多樣性，並能夠真實反映現實世界的文本數據。

通過從 Common Crawl 中獲取數據，我們確保了包含多種文章類型的包容性，從新聞文章到個人部落格、學術論文等各種文體。這種多樣性不僅豐富了資料集本身，還促進了關係擷取模型的發展，使其能夠應對領域變化，並適應於異質文章輸入。我們的方法有助於更全面地評估模型的泛化性和可擴展性，為跨領域和多樣化情境下的關係擷取提供了基礎。

Common Crawl 自 2007 年成立以來，已經累積了 17 年的網路爬蟲數據集，收集了約 2500 億筆網頁資料，這些網站資料橫跨 40 多種語言，包含原始網頁數據 (WARC)、元數據 (WAT) 和文本提取 (WET)。Common Crawl 的數據集廣泛應用於學習詞嵌入、文本挖掘和自然語言理解等領域。

Common Crawl 約 1 至 2 個月會提供一次快照，更新最新的網路爬蟲數據，每次內容約 20 到 40 億個 pages 不等。而本次研究是使用

2023-50 的快照來進行處理，由於檔案龐大，該快照共切分成 90000 個 Segments，我們擷取了其中的 990 個 Segments，並分成 11 個 shards 進行處理。而 Common Crawl 的數據處理流程繁瑣，包括多個步驟，我們採用 Wenzek 等人 [6] 所提出的做法進行處理，並將步驟分為：去重、語言辨識、品質篩選等三個 pipeline 步驟，如圖 1。

在預處理階段，首先對每個段落進行小寫轉換，將所有數字替換為佔位符，並消除所有 Unicode 標點符號和重音符號。此外，為了有效地儲存檔案，採用了 64 位 SHA-1 哈希算法。

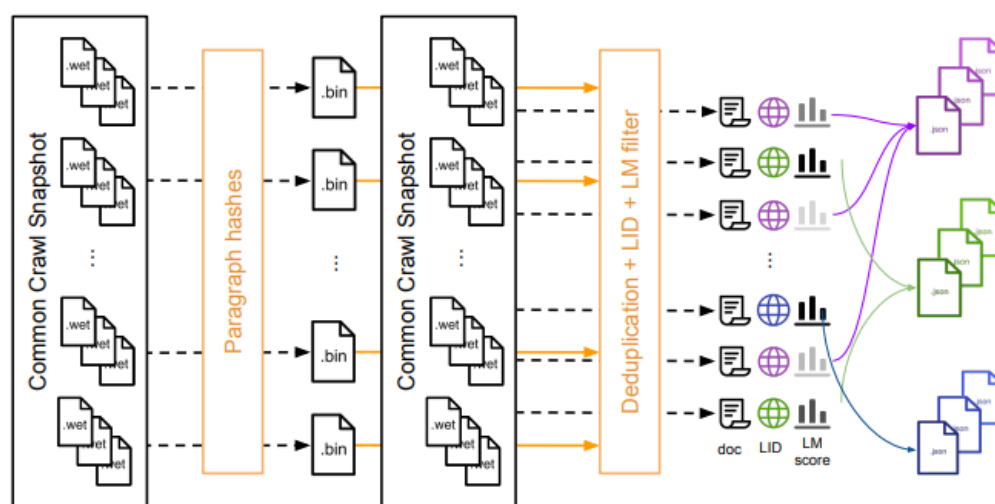


圖 1 CCNet 架構流程

去重過程通過計算每個段落的雜湊值來實現，從而消除重複的內容。我們使用 64-bits 的 SHA-1 作為雜湊值。去重的主要目標是確保網頁內容唯一性，避免網頁搬運所造成的大量重複內容。這一步可以有效地從其他語言的網頁中移除大量的英文文本，如網頁導覽列、cookie 警告、聯絡資訊等冗餘信息，進而降低後續語言辨識的難度。

語言辨識我們使用 fastText [40] 作為語言分類器，fastText 是 meta 的一個語言分類模型可以分類 294 種語言，主要在 Wikipedia、Tatoeba、SETimes 上面進行預訓練。fastText 會遍歷所有的網頁內容，並對所有語言分類打分數，最終保留最高且大於 0.5 的三個分類，如果沒有語言得分超過 0.5，那該筆資料會被丟棄，我們透過該方法篩選出所有的中文網頁內容。

品質篩選則在擷取文字資料、去重並確定了網頁資料的語言之後，仍然需要對文字進行一定的清洗來提高品質，稱為品質篩選。首先，我們使用 Sentence Piece tokenizer 把每一個網頁在句子層次做 tokenize，然後使

用評分語言模型來為每一個自然段做評分，我們用來進行品質篩選的評分語言模型為 KenLM3 庫裡面的 Kneser-Ney，該模型是在一些高品質的文本數據中訓練好的打分器，其所評的分數代表困惑度 (perplexity 簡稱 ppl) 分數，ppl 的分數越低，與高品質文本的目標分佈越接近，品質越高。代表其行文越流暢、越有邏輯，所以就更像是好的自然段。最後每個 shard 會依照 ppl 的分數分為頭部、中間、尾部，品質狀況為頭部 > 中間 > 尾部。

2. 前處理資料統計

我們針對 11 個 shards 進行去重、語言辨識及品質篩選的流程。在進行品質篩選後，每一個 shards 都可以得到頭部、中間、尾部三個部份的檔案。我們統計各個 shards 的網頁文章數量如表 1，可以看出該方法過濾掉了大量品質不佳的網頁資料。我們最終只取頭部的高品質資料作為我們的資料來源，並將 shard 0 的頭部 (head 0)，共 26,293 筆資料，作為測試資料集的來源，而 head 1 ~ head 10 共 260,469 筆資料作為訓練資料集的來源。

表 1: 清洗數據後，各個 shards 中頭部、中間、尾部三個部份的網頁數量統計

	head	middle	tail
shard 0	26,293	27,529	96,582
shard 1	26,305	27,708	99,120
shard 2	25,947	27,723	98,529
shard 3	26,090	27,441	98,466
shard 4	25,950	27,459	98,218
shard 5	26,007	27,861	98,217
shard 6	26,235	27,591	99,111
shard 7	26,156	27,533	98,065
shard 8	26,104	27,828	99,129
shard 9	25,906	27,407	98,378
shard 10	25,769	27,168	98,779
total	286,762	303,248	1,082,594

3. 大型語言模型標記流程

在這次標記實驗中，我們選擇以人物作為實體，並定義了 [親屬、師生、同事、其他] 作為我們標記的 4 種關係類型。

由於人工標記大量文章級別的資料既耗時又費力，我們使用 Gemini-1.0-pro(後稱 Gemini) 和 GPT-3.5-turbo(後稱 GPT) 作為標記工

具。而標記過程可以細分為四個步驟，分別為：三元組生成、關係分類、交叉驗證、合併資料。我們將會在本章各個小節詳細說明具體作法。但基於 GPT-3.5 的資源有限，因此，在測試資料集和訓練資料集的標記流程上有些許調整。

在測試資料集上，我們的標記流程如圖 2，我們在三元組生成階段讓 Gemini 和 GPT 都跑完所有資料，並過濾掉無關係和錯誤回覆的資料，並在關係分類階段，將關係分類為我們所限定的範圍，最後將模型生成不一致部分進行交叉驗證而合併為共識的人物關係。

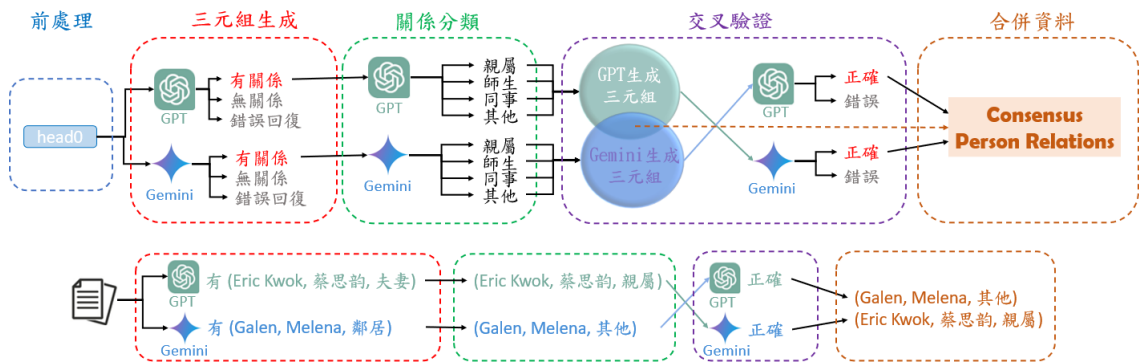


圖 2: 測試資料集的流程 (上半) 及範例 (下半)

而在訓練資料集上，由於總資料量約為測試資料的 10 倍，為節省 API 資源，在三元組生成階段我們先讓 Gemini 過濾掉大量無關係和錯誤回覆的資料後，只針對有關係的部分，讓 GPT 進行三元組生成，如圖 3，而後的關係分類、交叉驗證及合併資料流程則與測試資料集流程相同。

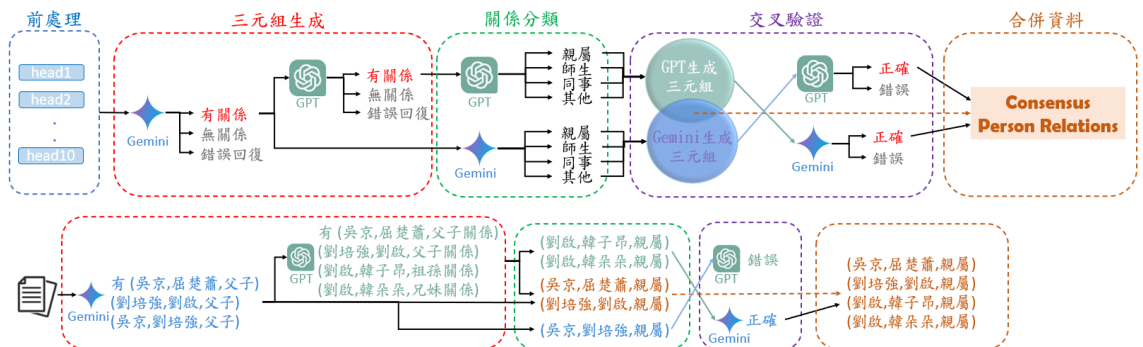


圖 3: 訓練資料集的流程 (上半) 及範例 (下半)

而由於測試資料集的標記流程是較為標準的，可以避免單一模型生成的盲點。因此在後續章節的數據分析中，若無特別註明，我們都會以測試資料集作為我們的統計資料。

4. 文章級別挑戰

在聯合實體關係擷取的生成式方法中，雖然有許多基於通用式生成模型的研究，使用 few-shot learning 進行實驗 [7,8]，但這些實驗都是在句子級別的資料集上進行。我們的網頁資料平均字數達到 1,502 字，希望在不切割文章的情況下，讓模型找出實體關係三元組，從而避免跨句的實體關係因切割而被遺漏。

雖然我們部屬的 Gemini-1.0 和 GPT-3.5 的 context window 可以達到 16k 個 tokens。然而若是在文章級別的資料集中，我們將長文本的文章當作範例加入 prompt，使用 few-shot learning，反而會造成整個 prompt 的過長，導致模型效能異常。

我們在 Gemini 上進行了小規模的測試，我們隨機取 100 筆在 zero-shot 下模型可以正常擷取關係三元組的資料。另外，我們從訓練資料中隨機抽取 2 筆資料做為範例，並加入人工標註的三元組做為範例答案，分別測試 1-shot 及 2-shots 效能，並且使用 Wadhwa 等人 [7] 的方法，人工針對這 2 筆資料加上 chain-of-thought(CoT) 的模板，具體的 prompt 內容詳見指令 1。

<p>1-shot</p> <p>請幫我找出以下文章中是否包含兩位具有明確姓名的人之間常見的人際關係 (例如: 親屬、師生、同事、同學...)? 且兩位關係人皆必須有明確名字, 只有稱謂的不算。</p> <p>若無關係直接回答: Relations: 無即可</p> <p>若有請依以下格式回答:</p> <p>Relations: 有 (人名, 人名, 關係), (人名, 人名, 關係)... 列舉出所有關係</p> <p>Explanation: 解釋原因</p> <p>範例如下:</p> <p>TEXT: 中国计划生育观察: 美国之音: 山东妇女怀孕 6 月, 被强迫堕胎</p> <p>下略 805 字....</p> <p>Relations: 有 (刘欣雯, 周国强, 夫妻)</p> <p>Explanation: 文章中提到刘欣雯和她的丈夫周国强在家中熟睡, 可見刘欣雯與周国强為夫妻關係</p> <p>文章如下:</p> <p>TEXT:{document}</p>
<p>2-shots</p> <p>請幫我找出以下文章中是否包含兩位具有明確姓名的人之間常見的人際關係 (例如: 親屬、師生、同事、同學...)? 且兩位關係人皆必須有明確名字, 只有稱謂的不算。</p> <p>若無關係直接回答: Relations: 無即可</p> <p>若有請依以下格式回答:</p> <p>Relations: 有 (人名, 人名, 關係), (人名, 人名, 關係)... 列舉出所有關係</p> <p>Explanation: 解釋原因</p> <p>範例如下:</p> <p>TEXT: 中国计划生育观察: 美国之音: 山东妇女怀孕 6 月, 被强迫堕胎</p> <p>下略 805 字....</p> <p>Relations: 有 (刘欣雯, 周国强, 夫妻)</p> <p>Explanation: 文章中提到刘欣雯和她的丈夫周国强在家中熟睡, 可見刘欣雯與周国强為夫妻關係</p> <p>TEXT: 成大材料系劉浩志團隊結合機器學習減少原子力顯微鏡量化量測誤差</p> <p>下略 2,843 字....</p> <p>Relations: 有 (劉浩志, 阮氏芳玲, 師生), (劉浩志, 張敬萱, 師生), (劉浩志, 簡錦樹, 同事), (劉浩志, 蔡佩珍, 同事)</p> <p>Explanation: 文章中提及劉浩志教授與當時的博士生張敬萱與阮氏芳玲發現, 可見劉浩志與阮氏芳玲為師生關係, 劉浩志與張敬萱也為師生關係</p> <p>另外文章中說到過去劉浩志教授曾與成大地科系簡錦樹教授研究嘉義布袋地底下抗砷的細菌, 還有他也曾與成大醫學檢驗生物技術系蔡佩珍教授對臨床腸病毒的病毒體進行物理特性研究, 所以可以得知劉浩志與簡錦樹為同事關係, 劉浩志與蔡佩珍也為同事關係</p> <p>文章如下:</p> <p>TEXT:{document}</p>

指令 1: few-shots 的 prompt 格式及內容

我們由實驗結果發現 (表 2), 模型在給定文章級別的範例後, 無論是 1-shot 或是 2-shot, 原本在 0-shot 可以正常生成關係三元組的網頁文章, 反而無法正常擷取出關係三元組, 所以我們在接下來的實驗都使用 zero-shot 進行。

表 2: Gemini 在 Document Level 的 few-shots 測試

	有關係文章數
0-shot	100
1-shot	0
2-shots	1

5. 三元組生成

我們希望透過通用式的生成方式，讓大型語言模型能夠直接生成關係三元組。由於我們定義的 NER 類型只有人名一種，找出人名實體的任務上相對單純，但是在關係類別上有四種分類（親屬、師生、同事、其他），若要在只能使用 zero-shot 的限制下一口氣讓模型完成 NER 和 RE 兩個任務是相對困難的。在我們的使用經驗上，沒辦法透過單純調整指令的方式，將模型的關係類型完全侷限在我們所定義的範圍內。

因此，在此階段我們我們就讓模型專注於找出有關係的人名即可，至於具體是哪種關係，我們只在 prompt 給定範例，但不強制侷限模型，讓模型將所有認為有關係的實體對，最大化的找出來。具體的 prompt 如指令 2。

請幫我找出以下文章中是否包含兩位具有明確姓名的人之間常見的人際關係（例如：親屬、師生、同事、同學...）？且兩位關係人皆必須有明確名字，只有稱謂的不算。
 若無關係直接回答：無 即可
 若有請依格式回答：有（人名，人名，關係），（人名，人名，關係）... 列舉出所有關係
 文章如下：
 {document}

指令 2: 三元組生成 Prompt，讓模型判斷內容是否具有人物關係，並限制模型的輸出格式

我們同時對 Gemini-1.0 和 GPT-3.5 給定以上的指令時，由於模型並不一定完全會依照我們所規定的格式回覆，我們會在每次取得 API 的回覆後，檢查模型的回覆格式。如果不符合我們所規定的格式，則會採用輪對話的方式，將強調格式規範的 prompt(如指令 3) 加入到之前的對話中，並重覆以上動作直到模型回覆符合正確格式，或是達到手動設定的上限值為止，在本次實驗中我們將上限值設定為 5 次，也就是模型如果連續 5 次無法依照格式回覆，我們則另外註記該筆資料。

請務必依照規定格式回答，若無關係直接回答：無
 若有請依格式回答：有（人名，人名，關係），（人名，人名，關係）.. 小括號中必須包含 2 個人名實體和 1 個關係

指令 3: 強調回覆格式 Prompt，會在收到模型錯誤回覆時加入多輪對話中，以增加模型格式的控制。

而我們將 Gemini 和 GPT 共同標記的測試資料集，共 26,293 筆資料，經過以上的標記方式，若能依照我們所設定的格式回覆三元組，或是判斷文章中不存在人名實體關係，我們都視為正確標記，反之，我們稱為錯誤標記。我們將錯誤標記的數量統計在表 3，並分為三種類型：

- 格式錯誤-模型回覆的關係格式並非三元組，可能出現四元組或是二元組等不符合格式規範的答案，例如：“(長瀨早瀨, 八王子直人, 學妹, 學長)”
- 無法識別-根據模型的回覆，我們無法判斷有無存在人物間的關係，通常是因模型受長文本的文章影響，導致回覆內容都和我們的指示無關，例如：“后句自由发挥，没有固定的下一句。”
- API 異常-指的是 API Error，通常都是因安全言論無法回答，由於我們的資料集是由 Common crawl 數據庫處理而來，內容可能包含色情、暴力、仇恨等言論，雖然在 Gemini 的 API 設置中我們可以調整安全言論的控制參數，但即便將安全言論控制參數調整為不封鎖，對於這些內容模型還是會封鎖回覆。

表 3: Gemini 和 GPT 正確及錯誤標記文章分析

	Gemini	GPT
正確標記	26,218	25,614
格式錯誤	29	226
無法識別	0	207
API 異常	46	246
總數	26,293	26,293

由表 3 中可以看出，Gemini 在錯誤回覆的數量上，相比 GPT 還要少許多，在同樣的 prompt 下，格式的可控性較為良好。而我們再進一步統計兩個模型所正確標記的資料如表 4，可以看出大部分的文章是不具有人名之間的關係的，多數網頁文章內容並不會存在人與人之間的關係，這也符合我們的認知預期。我們透過大型語言模型在此階段的生成標記，即可過濾掉 8 成以上的無關係網頁文章。

表 4: Gemini 和 GPT 正確標記中，具有人名關係文章占比

	Gemini	Gpt
有關係	2,268	3,576
無關係	23,950	22,038
有關係占比	8.65%	13.93%

6. 關係分類

在經過三元組生成的步驟後，我們可以分別得到 Gemini 和 GPT 所生成出來的三元組內容。然而，我們實際所生成的關係類型種類，並沒有在前一步驟中，將其侷限在我們所定義四種類型中。因此，產生的關係種類非常發散，我們具體統計在 Gemini 所生成的三元組中，關係的種類多達 1,142 種，在 GPT 所生成的三元組中，則多達 1,825 種。而在這步驟，我們需要將這些關係種類歸類為我所定義的四種分類（親屬、師生、同事、其他），例如：夫妻關係可被歸類為親屬。教練與學員關係可被歸類為師生等。因此，我們將三元組生成關係特別取出，並將其視為一個簡單的四元分類問題，透過 Gemini 和 GPT 以生成的方式，各自回答所生成的關係是屬於何種類別。具體如指令 4 所示。

我想將以下的關係進行分類成 [師生關係、同事關係、親屬關係、其他關係] 4 種類別
如果是師生關係：請回答 師生
如果是同事關係：請回答 同事
如果是親屬關係：請回答 親屬
如果是其他關係：請回答 其他
關係：
{博士生指導教授與博士生}
請問是 師生、同事、親屬、其他 哪一個？

指令 4: 關係分類 Prompt，該指令設計為簡單的四元分類問題

我們統計兩模型各自分類的數據結果，如表 5。可以看出無論 Gemini 或是 GPT 都是以其他類別的種類佔多數。代表我們生成出的人與人之間關係，多數被認定為我們所指定的親屬、師生、同事之外。而這一部分的資料，也可作為未來如果需要再細分更多關係所使用。

表 5: 模型標記的關係種類統計

	Gemini	GPT	佔比
親屬	135	105	8.09%
師生	61	108	5.70%
同事	194	93	9.67%
其他	752	1,519	76.54%
總關係數	1,142	1,825	100%

7. 交叉驗證

為了取得我們最終標記，我們取表 6 中，Gemini 和 GPT 所認為含有

人物關係的聯集資料，共 4,619 筆文章級別的網頁資料。並將其所生成出的三元組內容，進行關係分類。由於每筆文章可能會有多組實體關係三元組標記，我們將兩模型的資料筆數和三元組數量統計如表 6。

表 6: 有關係網頁資料筆數和生成三元組數量

	Gemini	Gpt	Intersection	Union
有關係網頁數	2,268	3,576	1,225	4,619
三元組數量	6,697	8,598	1,027	14,268
平均每筆網頁三元組數量	2.95	2.40	-	-

其中在比對交集的部份，我們採用的是嚴格的比對，即三元組中的文字必須完全相同才視為相同。但我們考慮我們所標示的親屬以及同事關係是沒有方向性的，而師生關係雖然具有方向性，但我們在標示時並沒有要求模型排序，因此，在實體順序上，若只是排序不同，我們視為相同的關係三元組。另外，由於我們的標記資料可能會有簡體及繁體中文，因此，比對時我們透過 OpenCC 翻譯為繁體中文後再進行比對，後續的所有比對流程也都相同。

我們可以看出，Gemini 和 GPT 所共同認定（交集）具有人物關係的網頁文章有 1,225 筆，但兩者共同認定的三元組 1,027 組，卻低於具有共識的文章數。而 Gemini 和 GPT 生成的三元組分別多達 6,697 及 8,598 組，但交集卻只有 1,027 組，佔比分別只有 15.34% 及 11.94%。代表兩者生成的三元組差異很大。可以看出我們要透過通用式生成方式，直接取得有共識的三元組是相對困難的任務。

因此，我們在此步驟簡化任務，將原本生成的任務改為相對容易的二元分類任務，透過是非題的方式，讓兩模型互相評估對方所生成出來的三元組是否正確。實作上，我們先將原本兩者本來就有共識 1,027 組，直接取出視為正確。只對無共識的三元組提問，且為了節省資源，我們會將同一筆文章的三元組編入題號，在同一次 request 一起詢問。另外，我們在指令中給定四種常見的標註錯誤，且皆給定範例讓模型對錯誤標記有所依據。錯誤範例分別為 A. 關係錯誤; B. 人名實體並非人的姓名; C. 人名實體沒有明確人名或是綽號，只有稱謂; D. 兩個人名相同。具體提問如指令 5。

我從以下文章中找出的 {3} 組人名和人際關係三元組 (人名, 人名, 關係), 關係共分為 4 種類別 [親屬、師生、同事、其他]。

文章如下:

{document}

關係如下:

{1.(邵智源, 林柏昇, 其他) 2.(邵智源, 泱泱, 其他) 3.(邵智源, 溫妮, 其他)}

請問以上 {3} 個人名關係三元組, 分別是正確或錯誤?

以下 4 種情形視為錯誤:

A. 關係錯誤, 例如:(蔣中正, 蔣經國, 同事), 正確關係應為 (蔣中正, 蔣經國, 親屬)。

B. 人名實體並非人的姓名, 例如:(習近平, 共產黨, 同事), 因為"共產黨"並非人的姓名, 其他如單位、公司、組織、隊伍... 等名稱皆為錯誤。

C. 人名實體沒有明確人名或是綽號, 只有稱謂, 例如:(湯姆·克魯斯, 妻子, 親屬), 並沒有給出妻子姓名, 其他如老公、妻子、父親、母親、哥哥、姐姐、學生、某某... 等亦同。

D. 兩個人名相同, 例如:(徐志摩, 徐志摩, 其他), 兩個人名相同即視為錯誤。

請依格式回答:

{1. 正確/錯誤 2. 正確/錯誤 3. 正確/錯誤}

指令 5: 交叉詢問 Prompt, 設計成是非題的題組, 簡化任務難度, 且透過一次尋問多個三元組的方式, 減少 request 次數, 以節省資源

在兩模型互相交叉詢問後, 我們統計模型驗證的實體關係三元組, 通過驗證數量如表 7。可以看出雖然兩個模型在三元組生成的階段, 生成出來的三元組內容並不一致, 但是對於對方所生成出來的實體關係三元組, 卻又高度的表示認同, 兩個模型通過對方驗證的比例皆超過 9 成以上。對於這種情形, 我們透過研究兩模型標記高度不一致的實際範例。我們發現, 在這些網頁資料中如果出現大量的人名實體時, 如演員名單、出賽名單、員工名單等, 使用生成式方法, 模型通常無法找出所有人名的兩兩關係, 只會取樣部分的實體對並認定存在關係, 但兩模型的實體對取樣往往不相同, 導致兩模型標記一致性不佳。

表 7: 有關係網頁資料筆數和生成三元組數量

生成模型	驗證模型	通過交叉驗證	未通過交叉驗證	通過比例
Gemini	GPT	5,166	504	91.11%
GPT	Gemini	7,254	317	95.81%

8. 合併資料

我們將兩個模型原本就有共識的 1,027 組三元組, 以及 Gemini 通過 Gpt 驗證的 5,166 組、GPT 通過 Gemini 驗證的 7,254 組, 共 13,447 組實體關係三元組, 視為我們暫定的共識人物關係三元組。在這 13,447 組實體關係三元組, 共分佈在 4,515 筆網頁文章資料中, 即原本兩模型的聯集資料 4,619 筆, 經過交叉驗證後排除了未通過驗證的剩餘文章數。我們統計四種關係的實際分佈狀態如表 8。

表 8: 共識人物關係三元組中，不同類型關係分佈，由於一篇文章可能含有多種關係，因此含有四種關係的文章數加總後，會大於文章總數 4,515 筆

關係類型	三元組數量	三元組佔比	含有該關係文章數	文章佔比
親屬	1,168	8.69%	642	14.22%
師生	1,192	8.86%	743	16.46%
同事	6,172	45.90%	2,196	48.64%
其他	4,915	36.55%	2,255	49.94%
總數	13,447	100%	-	-

可以看出在我們所定義的四種關係類型，在隨機抽樣真實的網頁文章中，分布是相當不平衡的，其中同事和其他關係的三元組佔了 45.90% 及 36.55%。但我們對照表 5 數據，可以發現兩者佔比高的原因並不相同。我們可以看出在進行關係分類前，模型所生成的關係總類中，其他關係明顯多於另外三者，這是因為我們將其他我們未定義的人與人關係，如：朋友、同學..等，全部都歸類到” 其他” 關係所導致的資料不平衡。但同事關係的在種類占比中只有 9.67%，只略高於親屬和師生分別 1.58% ~ 3.97%，不過在三元組數量中卻多了超過 30% 以上。因此，我們可以推論，在我們所清洗的網頁文章中，包含同事關係的資料本身就遠超過親屬和師生的數量。

四、執行成果（請與計畫目標配合）

全案工作執行成效（含理論面、實務面）

1. NER 效能評估

我們使用傳統的序列標記模型，先將所有的人名找出來，我們使用的中文 SOTA-NER 模型為中研院所開源的 ckiplab/bert-base-chinese-ner (後稱 CKIP)。我們將 CKIP 所標記出來的人名視為標準答案，分別評估 Gemini、GPT，以及合併後共識人物關係三元組 (Consensus) 的 NER 效能，在 NER 效能的計算上，我們以每篇文章中的實體為單位，計算模型預估值和正確答案的實體是否一致，兩者在使用 OpenCC 翻譯為繁體中文後，在嚴格比對下需要完全一致才視為相同實體，最終效能如表 9。

表 9: 將 CKIP 所標的實體視為答案，評估模型 NER 效能

	Recall	Precision	Micro-f1
Gemini	11.53%	66.48%	19.65%
GPT	12.84%	51.92%	20.59%
Consensus	19.55%	55.47%	28.91%

由於我們最初的任務是實體關係三元組生成，而非找出所有人名實體。因此，如果人名之間並不存在關係，那該人名實體未被生成也實屬合理。但是我們可以看出即便是兩模型合併後的共識，Recall 也不到 20%。代表有過 80% 以上的人名之間都沒有關係，這較不符合常理。因此，我們將針對 CKIP 所標記出來的人名，進行實體關係三元組的擴充實驗。

2. 幻想評估

我們這次的實驗之所以使用文章級別的內容進行，就是希望能夠找出傳統句子級別的方法，所無法判斷的關係。因此，我們將具體統計所找到的實體關係三元組中，有多少比例的實體對是屬於跨句分佈。但由於我們使用的是生成式的方法，而非傳統序列標記方式，所以是有可能產生不存在於文章中的實體。因此，在查找實體位置分佈時，需要先確定我們所生成的實體並非模型所幻想出來的，我們以實體個數為單位，檢查 Gemini 和 GPT 所生成出來的人名實體，是否存在原本文章中，具體統計如表 10。可以發現約有 3% 左右的幻想實體存在。

表 10: 模型所幻想的人名實體統計，以實體個數為單位

	幻想實體數	總生成實體數	幻想實體比例
Gemini	173	9,687	1.79%
GPT	480	13,403	3.58%
Consensus	568	19,263	2.95%

3. 跨句評估

在檢查完幻想的實體後，我們針對模型所生成的實體對做分佈統計。首先，我們必須先將文章切割為句子級別。我們使用中文常用的全形、半形標點符號以及換行符號對句子進行切割，具體切割符為 `[\n。; ; ! ! ? ?]`。而平均每篇文章會被切割為 58.19 個句子。我們判斷實體對中的兩個人名是否有出現在相同句子，若未出現在同一個句子中，則視為跨句子的實體對。

表 11: 統計模型的跨句找出實體關係能力，以實體對個數為單位

	含有任一幻想實體	實體對存在相同句子	跨句實體對
Gemini	2.88%	63.42%	33.70%
GPT	6.59%	67.03%	26.38%
Consensus	4.91%	65.28%	29.81%

我們可以由表 11 中看出，模型所找出的實體關係約有 30% 左右，是屬於跨句子級別的關係。而我們進一步分析這些跨句子的實體關係中，具體能夠跨越的篇幅有多大。由於同一個實體在文章中可能出現不只一次，因此，我們計算實體對間的最小間隔字數，也就是兩實體的最短距離 (Shortest Distance)。如表 12 所示，我們計算了模型各自所生成的實體關係三元組中，實體對最小間隔字數的平均值，以及所有實體對最小間隔字數的最大值，代表模型最遠能發現的實體間隔字數可以到多遠，我們發現即便實體位置橫跨超過上千字以上，大型語言模型都還是有能力能夠找出兩者關係，這也證明了大型語言模型在跨句實體能力上傑出的表現。

表 12: 統計跨句實體對中，實體對最小間隔字數

	跨句實體對數量	平均最小間隔字數	最遠的最小間隔字數
Gemini	2,257	246	3,376
GPT	2,268	123	1,803
Consensus	4,009	186	3,376

4. 實體關係擴充

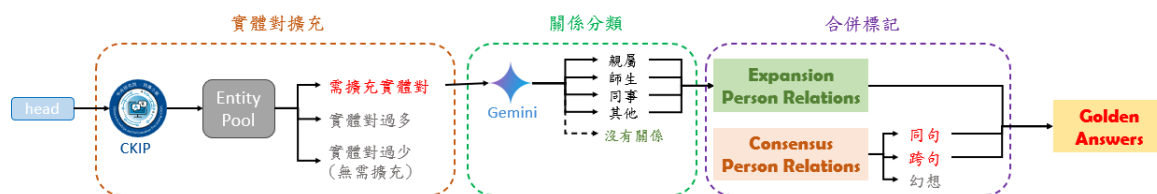


圖 4: 實體關係三元組擴充流程

我們在表 9 中看出，模型在實體對的 Recall 效能非常不足，有超過 80% 以上的人名都未被取出。因此，我們需要對每篇的網頁文章進行實體擴充，具體流程如圖 4。我們先透過 CKIP 找出文章中的所有人名實體，並將實體兩兩組成實體對，第二部分則針對需擴充的實體對，進行關係分類，最後合併原先共識的人物關係時，將含有幻想的實體關係給去除。

我們將 CKIP 標記中所得到的實體兩兩組成實體對時，會先比對該實體對是否存在原本 Gemini 和 GPT 最初生成的實體關係三元組中。但若實體數量為 n ，實體對數量則會增為 C_n^2 ，甚至出現如”畫廊會員名單”這種整篇文章的是人名的網頁時，最高達到 211,550 組實體對。考量到這類型的文章，並非我們最主要需要解決的類型。因此，我們設置了兩個條件來過濾掉如這種實體對過多的文章類型：

我們設置的第一個條件為人名密度，我們定義為：CKIP entity 數量/文章字數，我們計算所有有關係的文章平均人名密度為 0.95/100 字，我們取 2 倍的平均人名密度作為我們第一個篩選條件。另外，我們考量實驗成本，我們第二個條件直接設定實體對上限值，若實體對超過 C_{15}^2 ，也就是 105 組時，我們發現很難讓模型在一次 request 中完成。

除了實體對過多的文章類型，與之相反的就是實體對過少（無需擴充）的文章類型。如果 CKIP 所標記的實體小於 2 個則無法組成實體對，或是組成的實體對原本就存在 Gemini 或 GPT 生成的三元組之中，我們都視為實體對過少的文章。我們具體統計扣除實體對過多的文章類型後具體需要做實體擴充和無需做實體擴充的文章數量如表 13。

表 13: 實際需擴充佔比

	文章數	佔比
實體對過多	776	33.78%
無需擴充	1,524	17.19%
需擴充	2,215	49.03%
總數	4515	100%

在將實體兩兩組成實體對後，我們讓 Gemini 進行關係分類。然而和前述關係分為四種類型略為不同的是，實際上人名兩兩組成的實體對，並不一定具有關係。因此，我們新增一個類別為沒有關係，而若模型分類為沒有關係，我們則不會取用該實體對來新增標記。具體如指令 6。

根據以下文章，找出每組人名實體對中的人名之間的關係。關係分為：親屬關係、師生關係、同事關係、其他關係、沒有關係，共 5 種。
 人名實體對：
 {1.(丁淑君, 林慶芳) 2.(林慶芳, 梁作磊) 3.(丁淑君, 梁作磊)}
 文章如下：
 {document}
 回答格式：
 {1. 親屬/師生/同事/其他/沒有關係 2. 親屬/師生/同事/其他/沒有關係 3. 親屬/師生/同事/其他/沒有關係}
 請根據以上格式回答

指令 6: 實體對關係分類 Prompt，讓模型判斷每組人物實體對的關係

最後我們統計所擴充的三元組中，各個類別的分佈情形，如表 14。可以看出我們所擴充的類型中，同事的類別佔有極為懸殊的大量。而這也符合我們的預期，在真實網路文章中，包含同事關係的文章佔比確實較高，而如果將文章中人名都做兩兩配對，也確實讓同事關係數量暴增。

表 14: 加入擴充標記後，三元組的各類別佔比

	共識三元組	擴充三元組	Golden Answers	佔比
親屬	735	2,228	2,963	9.76%
師生	828	157	985	3.25%
同事	3,839	18,774	22,613	74.52%
其他	3,366	418	3,784	12.47%
總數	8,768	21,577	30,345	100%

5. 模型訓練成果

● 通用式生成:

我們考慮到文章中可能夾雜中英文或日文等人名實體，我們選定多語言的 mT5-base [9] 作為訓練的基底模型。我們實驗測試 mT5 模型是否能夠像 Gemini 和 GPT 一樣，在給定一整篇文章情形下，加上

三元組生成的 prompt(指令 2) 後，進行 full fine-tuning，讓模型直接將所有可能的實體關係三元組給生成出來。理論上依照 Gemini 和 GPT 的做法，在三元組生成步驟後，還需進行關係分類，將生成的關係收斂在我們所定義的 4 種類別，但實際上經過全微調後，模型已經可以直接在三元組生成的步驟就收斂到我們所定義的 4 種關係類別，因此我們即省略關係分類步驟。

mT5 的實際效能如表 15 所示，在二元組效能中，我們在小規模的測試資料集內進行訓練，Micro-f1 即可達到 23.41%，已相當接近 Gemini 效能。若使用完整的訓練資料集，增加訓練資料量，即可達到 30.61% 超越 Gemini 的 24.61%，達到接近 GPT 的 31.64%。而在三元組效能上，在小規模的測試資料集上訓練還未能超越 Gemini 和 GPT，但是在完整的訓練資料集上，即可達到 Micro-f1 25.00% 超越了 Gemini 效能的 23.38%，但還未能達到 GPT 的效能。

● Pipeline:

我們考量在較小參數量的模型上，通常還是以傳統 pipeline 的方式較具有優勢。因此，我們也實驗了傳統 pipeline 的方式來進行。我們就以 CKIP 所提供的 bert-base-chinese-ner 模型進行 NER 任務，由於我們的 Golden Answers 中並不包含沒有關係的人物實體，因此若以我們的 Golden Answers 中的實體去微調模型，反而會有可能遺漏人名實體，所以我們在 NER 任務並未進行微調訓練。

在訓練時，我們將 CKIP 所得到的人物實體兩兩組成配對，若其不存在 Golden Answers 中，我們會將該組合的關係設為”沒有”進行訓練，因此，在 RE 任務中，我們實際進行的是 5 分類任務，即 [親屬、師生、同事、其他、沒有]。另外，為了減少和通用式生成方式比較時的變因，我們一樣使用 mT5 來進行分類訓練。

而我們這進行 RE 任務時，由於不像 LLM 有 API 資源考量，因此，我們會將每一個實體對當做一筆資料進行訓練和推論，以降低任務複雜度。我們使用指令 7 的 prompt，詢問該人物實體對之間的關係。而在模型推論階段，若模型推論為”沒有”的關係時，我們則將該筆實體對刪除，不會放入我們的預測之中。

根據以下文章，找出 {person1} 與 {person2} 中之間的關係。關係分為：親屬關係、師生關係、同事關係、其他關係、沒有關係，共 5 種。
文章如下：
{document}

指令 7: RE 任務 Prompt，設計成 5 分類問題，每次只詢問一組實體對，避免增加問題複雜度

由表 15 可以看出，在 pipeline 方法中，二元組效能的在小規模的測試資料集內進行訓練，Micro-f1 即可達到 73.21%，遠高過於 Gemini 和 GPT 效能。而在完整的訓練資料集上，效能反而較小規模資料集上訓練的下降，由於簡單的分類問題，增加數據量並不一定會有更好效果。但還是可以遠超過 Gemini 和 GPT 效能。

表 15: 二元組效能與三元組效能

模型	方法	訓練資料	二元組效能			三元組效能		
			Recall	Precision	Micro-f1	Recall	Precision	Micro-f1
Gemini	通用式生成	NO	14.18%	93.06%	24.61%	13.39%	92.27%	23.38%
GPT	通用式生成	NO	19.06%	93.06%	31.64%	18.27%	92.42%	30.50%
mT5	通用式生成	5-fold test	18.07 ± 1.00%	33.26 ± 1.57%	23.41 ± 1.19%	14.55±0.81%	26.43±1.16%	18.76±0.94%
mT5	通用式生成	train data	24.93%	39.65%	30.61%	20.47%	32.12%	25.00%
CKIP+mT5	pipeline	5-fold test	71.68 ± 2.51%	74.94 ± 3.63%	73.21 ± 2.86%	64.11±2.95%	67.52±3.98%	66.80±3.30%
CKIP+mT5	pipeline	train data	59.50%	71.29%	64.87%	53.26%	64.29%	58.26%

五、成果交付項目

- 期中(末)報告書
- 其他如系統軟硬體、操作手冊、教育訓練、技術轉移等(※ 請條列式列出)

.....

- 期中報告書
- 期中報告投影片
- 每月進度紀錄
- 專案程式碼 (含相關使用說明文件)

六、 檢討與建議

● 「期中審查會議決議事項改進說明」

(※請依據期中審查會議紀錄填寫後續處理情形)

項次	期中審查意見與建議	處 理 情 形
一		
二		
三		
四		
五		
六		

● 「經費使用情形」 (※請依據計畫書核定金額填寫)

計畫經費	計畫核定金額 A	已結報金額 B	待結報金額 C	餘額(A-B-C)	經費支用是/否 符合原簽訂之計畫書規劃 (如有變更,請於 本欄詳細說明)
人事費	539,143	262,011	61,265	215,867	
其他	100,857	39,476	0	61,381	
設備費	0	0	0	0	
管理費	44,800	0	0	44,800	
合 計	684,800	301,487	61,265	322,048	

七、結語

我們的研究成果在多個方面做出突破：

我們提出了一個基於生成式語言模型的標記流程，利用當前最先進的大型語言模型（如 Gemini、GPT-3.5 等），協助我們對未標記的文章級別內容進行標記。這種方法大幅度減少了對人工標記的依賴，節省了大量的人力和時間資源。通過設計和實施我們的標記流程，我們能夠提高標記過程的效率，並且能夠有效地補足單一模型在標記過程中的盲點，確保數據的質量和完整性。

我們採用 Common Crawl 這類全網爬蟲資料庫作為我們標記資料集的來源，這些資料來自於真實的網頁，具有廣泛的內容和多樣性，避免了傳統資料集內容過於侷限和狹隘的問題。我們通過從多樣化的網頁內容中擷取資料，構建了一個更具泛用性的資料集，這使得我們的訓練出來的模型能夠更好地適應不同的文本環境和應用場景，提升了模型在真實世界中的表現能力。

針對目前中文文章級別資料集不足的現狀，我們專注於利用先進的大型語言模型來協助標記中文資料，推動中文關係擷取任務的研究進展。通過我們的研究和實踐，我們成功創建了一個具有代表性的中文文章級別資料集，這對於我們探索中國政商人物之間的關係研究具有重要意義。

有別於傳統句子級別的關係擷取任務，我們的方法讓模型在整篇文章中進行跨句子、跨段落的關係擷取，也順利找出了大約 30% 只能在文章級別上被識別出來的關係。這克服了過去因文章長度限制而必須截斷文本或進行證據檢索的局限。

總之，我們的研究不僅在技術方法上有所創新，也為未來的國家安全情報領域提供極具價值的技術支援，透過自動化的人物關係擷取，大幅提高國安單位對於關鍵人物的了解和把握，極大提升情報搜集的效率與精確性，進而強化對潛在風險的預警能力，有效維護國家安全與社會穩定。

八、附錄

附件：研究本計畫參考文獻等相關資料

1. 參考文獻

- [1] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. ACE 2005 multilingual training corpus. In Linguistic Data Consortium, 2006.
- [2] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLTNAACL 2004, pages 1–8, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics.
- [3] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21, pages 148–163. Springer, 2010.
- [4] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 764–777, Florence, Italy, July 2019. Association for Computational linguistics.
- [5] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A largescale document-level relation extraction dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 764–777, Florence, Italy, July 2019. Association for Computational Linguistics.
- [6] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association.
- [7] Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics.

- [8] Hui Wu, Yuting He, Yidong Chen, Yu Bai, and Xiaodong Shi. Improving few-shot relation extraction through semantics-guided learning. *Neural Networks*, 169:453–461, 2024.
- [9] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.

※報告格式內容請依需求斟酌增加或刪減,紅色字型請於印製時刪除

人物關係擷取模型開發之研究 會議記錄

會議	專案啟動會議
日期時間	2023/12/08 14:00~15:30
地點	中央大學工程館 B323
出席人員	國安局: 董先生、陳小姐、孫小姐 WIDM 實驗室: 張嘉惠老師、洪閔昭、葉季儒
議程	討論專案合作模式細節內容

<p>結論事項</p>	<p>資料來源: 先從 Common Crawl 資料集，單一月份的中文部分著手。</p> <p>提交格式：不需要 DB，使用 CSV 檔案即可，欄位包含 entity1, entity2, relation, url</p> <p>需要萃取的關係：親屬、師生、部屬</p> <p>訓練方法：訓練自有模型（如果使用 LLM,可以跳過 Entity Extraction, 直接找 relation</p> <p>可參考網站:</p> <ol style="list-style-type: none"> 1. 赵家人俱乐部 2. 温家帝国 3. 周永康的家族
<p>備註</p>	<p>聯絡方式：</p> <p>0228891088</p> <p>董先生 #55334</p> <p>陳小姐 #55335</p> <p>孫小姐 #55675</p> <p>(上班時間無法使用手機，比較急時打電話聯絡)</p>

會議簽到表

會議主題：人物關係擷取模型開發之研究會議

時間：112 年 12 月 08 日 14:00~15:30

地點：中央大學工程五館三樓 (B323 室)

出席人員簽到處：

編號	姓名	單位	簽名
1	董先生	國家安全局	董木
2	陳小姐	國家安全局	陳言瑾
3	孫小姐	國家安全局	孫唯菴
4	張嘉惠	國立中央大學 WIDM 實驗室計畫主持人	張嘉惠
5	洪閔昭	國立中央大學 WIDM 實驗室研究生	洪閔昭
6	葉季儒	國立中央大學 WIDM 實驗室研究生	葉季儒

人物關係擷取模型開發之研究 會議記錄

會議	專案啟動會議
日期時間	2024/1/23 09:30~10:30
地點	中央大學工程館 B323
出席人員	國安局: 董先生、孫小姐 WIDM 實驗室: 張嘉惠老師、洪閔昭、葉季儒
議程	第一月份進度討論-Common Crawl 資料處理與檢索
結論事項	<ol style="list-style-type: none"> 1. 主要想要關注的對象會是中国经济网 (http://district.ce.cn/zt/rwk/sf/bj/index_21098.shtml)，地方領導人等中階幹部的名單。 2. 下一步先從 Common Crawl 數據中挑選出含有親屬關係的 Document。 3. 先使用 Taide 或 chatGPT 標記資料，做一個親屬關係分類器，並使用關鍵字方式實作作為效能比較。
備註	<p>聯絡方式：</p> <p>0228891088</p> <p>董先生 #55334</p> <p>陳小姐 #55335</p> <p>孫小姐 #55675</p> <p>(上班時間無法使用手機，比較急時打電話聯絡)</p>

會議簽到表

會議主題：人物關係擷取模型開發之研究會議

時間：113 年 1 月 23 日 9:30am~10:30am

地點：中央大學工程五館三樓 (B323 室)

出席人員簽到處：

編號	姓名	單位	簽名
1	董先生	國家安全局	董木
2	陳小姐	國家安全局	陳言理
3	孫小姐	國家安全局	孫唯茹
4	張嘉惠	國立中央大學 WIDM 實驗室計畫主持人	張嘉惠
5	洪閔昭	國立中央大學 WIDM 實驗室研究生	洪閔昭
6	葉季儒	國立中央大學 WIDM 實驗室研究生	葉季儒

人物關係擷取模型開發之研究 會議記錄

會議	專案啟動會議
日期時間	2024/03/08 14:00~15:30
地點	中央大學工程館 B323
出席人員	國安局: 董先生、王先生 WIDM 實驗室: 張嘉惠老師、洪閔昭、葉季儒
議程	二月份進度討論-親屬關係生成實測結果討論
結論事項	<ol style="list-style-type: none"> 1. 再從 Common Crawl 數據中清洗出足夠的 positive data 訓練模型，若需要 10,000 筆 positive 資料，約需要原本資料量的 10 倍，也就是約 common crawl 2023-50 的 1%。 2. 並將原本模型標記 union 的 1098 筆作為 test data 的 positive 資料，原本的 26,293 筆則為 test data 總數。 3. (訓練)將新的清洗出來的數據，交給 Gemini 標記，標記完成後，切割出和 test data 相同數量作為 valid data。 4. 將 LLM 生成的各種關係，用於微調本地端的關係生成模型(嘗試 mt0 和 gemma)，實做測試將所有關係不過濾直接訓練。 5. (訓練&測試)做一個關鍵字篩選器，去除只有稱謂的情形，比較在拉除稱謂前的效能以及拉除稱謂後的效能。 6. 暫不考慮訓練及測試的速度。 7. 對於輸出格式不可控的部分，可以參考使用 guidance 來控制模型的自定義格式。 8. 如果 GPU 設備資源不足，可以參考唐鳳 github 上的 GGUF 方式。
備註	<p>聯絡方式：</p> <p>0228891088</p> <p>董先生 #55334</p> <p>陳小姐 #55335</p> <p>孫小姐 #55675</p> <p>(上班時間無法使用手機，比較急時打電話聯絡)</p>

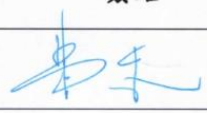
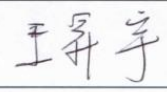



會議簽到表

會議主題：人物關係擷取模型開發之研究會議

時間：113 年 3 月 8 日 14:00~15:30

地點：中央大學工程五館三樓 (B323 室)

出席人員簽到處：

編號	姓名	單位	簽名
1	董先生	國家安全局	
2	陳小姐	國家安全局	
3	孫小姐	國家安全局	
4		國家安全局	
5	張嘉惠	國立中央大學 WIDM 實驗室計畫主持人	
6	洪閔昭	國立中央大學 WIDM 實驗室研究生	
7	葉季儒	國立中央大學 WIDM 實驗室研究生	

人物關係擷取模型開發之研究 會議記錄

會議	專案啟動會議
日期時間	2024/04/12 10:30~12:00
地點	中央大學工程館 B323
出席人員	國安局: 董先生、王先生 WIDM 實驗室: 張嘉惠老師、洪閔昭、葉季儒
議程	四月份進度討論-關係生成微調討論
結論事項	<ol style="list-style-type: none"> 1. Test data 部分讓 Gemini 和 chatGPT 大型語言模型標記，做交叉驗證，有衝突才人工看。 2. 小模型部分第一種方法:先做分類器、再做 seq2seq 3. 第二種方法使用 decompose 方式:先找出 entity、再找 relation。 4. 定義 relation:師生、同事、親屬 5. 實驗拆解文章成段落，測試 context window 多大，效果如何。 6. 找一下目前的 sentenxe level dataset，有沒有原始的 document，例如 ACE、CoNLL 7. 使用 ACE 的中文資料，抽出人物關係部分，做訓練和 test，檢視成效。
備註	<p>聯絡方式：</p> <p>0228891088</p> <p>董先生 #55334</p> <p>陳小姐 #55335</p> <p>孫小姐 #55675</p> <p>(上班時間無法使用手機，比較急時打電話聯絡)</p>

會議簽到表

會議主題：人物關係擷取模型開發之研究會議

時間：113 年 4 月 12 日 10:30~12:00

地點：中央大學工程五館三樓 (B323 室)

出席人員簽到處：

編號	姓名	單位	簽名
1	董先生	國家安全局	董先生
2	陳小姐	國家安全局	
3	孫小姐	國家安全局	
4		國家安全局	王昇宇
5	張嘉惠	國立中央大學 WIDM 實驗室計畫主持人	
6	洪閔昭	國立中央大學 WIDM 實驗室研究生	洪閔昭
7	葉季儒	國立中央大學 WIDM 實驗室研究生	葉季儒

人物關係擷取模型開發之研究 會議記錄

會議	專案啟動會議
日期時間	2024/5/10 10:30~12:00
地點	中央大學工程館 B326
出席人員	國安局: 董先生、王先生 WIDM 實驗室: 張嘉惠老師、洪閔昭、葉季儒
議程	五月份進度討論-關係生成微調討論
結論事項	<ol style="list-style-type: none"> 1. 微調模型先使用較小如 T5 seq2seq 模型執行。 2. 關係分類，列舉一些範例 3. Taide 8B 模型標註資料，計算評估指標 4. ACE 代名詞部分還原成實體，allen nlp，原本 ACE 的 coreference。 5. ACE 找出段落文章。
備註	<p>聯絡方式：</p> <p>0228891088</p> <p>董先生 #55334</p> <p>陳小姐 #55335</p> <p>孫小姐 #55675</p> <p>(上班時間無法使用手機，比較急時打電話聯絡)</p>

會議簽到表

會議主題：人物關係擷取模型開發之研究會議

時間：113 年 5 月 10 日 10:30~12:00

地點：中央大學工程五館三樓 (B326 室)

出席人員簽到處：

編號	姓名	單位	簽名
1	董先生	國家安全局	董
2	王樂先生	國家安全局	王樂
3	陳小姐	國家安全局	
4	孫小姐	國家安全局	
5	張嘉惠	國立中央大學 WIDM 實驗室計畫主持人	張嘉惠
6	洪閔昭	國立中央大學 WIDM 實驗室研究生	洪閔昭
7	葉季儒	國立中央大學 WIDM 實驗室研究生	葉季儒

人物關係擷取模型開發之研究 會議記錄

會議	專案會議
日期時間	2024/06/14 09:00~10:30
地點	中央大學工程館 B326
出席人員	國安局: 董先生、王先生 WIDM 實驗室: 張嘉惠老師、洪閔昭、葉季儒
議程	6 月份進度討論-實體擴充任務討論
結論事項	<ul style="list-style-type: none"> 4. 將訓練資料集 Gemini 篩出有關係的部分，再用 GPT 生成。 5. 計算人名密度的標準差。 6. 計算擴充後，Gemini 和 GPT 的效能。 7. 使用 MT5 訓練結果，並計算效能。 8. 使用 ckip 模型做 NER + 分類器，並計算效能。
備註	<p>聯絡方式：</p> <p>0228891088</p> <p>董先生 #55334</p> <p>陳小姐 #55335</p> <p>孫小姐 #55675</p> <p>(上班時間無法使用手機，比較急時打電話聯絡)</p>