EasyChair Preprint

№ 15202

Research on Document-Level Person Relation Extraction in Chinese

Min-Chao Hung, Chia-Hui Chang and Chi-Ju Yeh

October 6, 2024

中文文章級別人物關係擷取之研究
Research on Document-Level Person Relation Extraction in Chinese

洪閔昭
國立中央大學
111522155@cc.ncu.edu.tw

張嘉惠
國立中央大學
chia@csie.ncu.edu.tw

葉季儒
國立中央大學
tobyeider.ncu@g.ncu.edu.tw

摘要

本研究旨在構建一套可應用於眞實網路資料的聯合實體關係擷取架構。針對現有資料集來源單一且主要集中在句子級別的問題，我們利用大型語言模型（如 Gemini、GPT-3.5）對全篇文章進行標記，並使用中文 Common Crawl 數據構建更泛用的資料集。爲提高標記的可信度與實體對取樣的完整性，採用了交叉驗證與實體擴充方法。並通過微調預訓練模型來驗證與提升模型在眞實環境下進行實體關係擷取的性能。

關鍵字：命名實體識別、聯合實體關係擷取、文章級關係擷取

1 Introduction

關係擷取（Relation Extraction，簡稱 RE），旨在從文本中識別出實體間的關係。在傳統的關係擷取任務上，通常需要先進行命名實體識別 (Named Entity Recognition，簡稱 NER) 任務，即從文本中識別出具有指稱性的實體，如人物、地點、組織等。其次，需要擷取相關的特徵，包括詞的上下文、句法結構、實體間的相對位置等。最後再進行關係分類，通常使用機器學習或深度學習模型來將擷取的特徵與已知的關係標籤進行分類。而命名實體識別以及關係擷取兩個任務合在一起則稱爲聯合實體關係擷取 (Joint Entity and Relation Extraction)

近年來由於大型語言模型 (LLM) 的快速發展，聯合實體關係擷取的任務逐漸改由生成式模型來解決，而在先前的研究中也表明了，生成式模型可以在關係擷取達到 State-of-the-art (SOTA) 的效果 [1]。且在 [2] 的實驗中也發現像 GPT-3 規模的模型，即便不需經過 fine-tune，只需要給定良好的 instruction，即可達到 SOTA 的成效。

不過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、新聞網站，因此，訓練出來的模型較難泛化到眞實世界中多樣的網路數據上。再者，目前在關係擷取任務的主要資料集，例如:ACE05[3]、CoNLL[4]、NYT[5] 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 Yao 等人 [6] 統計，超過 40.7%的關係只能在文章級別上被識別出來。而少數像 DocRED[6] 這類的文章級別資料集，目前也都是以英文的資源爲主，缺乏相對應的中文資源。

在本篇論文中，我們利用當前最先進的大型語言模型（如 Gemini、GPT-3.5 等），對 Common Crawl 全網爬蟲資料內容進行標記，創建中文文章級別關係擷取資料集。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。而我們也在參數量較小的模型中實驗證實，該資料集所訓練出的模型，可泛化於眞實網路文章中的模型可行性，並爲未來的相關研究留下了基線做爲參考。

2 Related Work

關係擷取（RE）是自然語言處理（NLP）中的關鍵任務，廣泛應用於知識圖譜構建 [7, 8]、問答系統 [9, 10, 11]、對話系統 [12, 13] 等領域。傳統的 RE 方法將其視爲多類別分類問題，通過 LSTM 或 BERT 等模型進行監督訓練 [14, 15]，但這些方法依賴大量人工標記數據。爲減輕標記負擔，一些研究引入遠程監督 [16] 和半監督方法 [17]，但這可能導致不準確的標記。

傳統 RE 多使用 pipeline 方式，先抽取實體再進行關係分類，但這忽略了命名實體識別（NER）和關係擷取（RE）間的互動，導

致誤差傳遞問題。爲改善此問題，常用多任務學習 [18, 19] 和更換擷取策略的方式，如 CasRel[20]、ETL-Span[21] 等，通過共享參數或改變擷取步驟來減少誤差。然而，這些方法仍受限於誤差傳遞和暴露偏差問題 [22]。

爲解決這些問題，研究者提出了填表方法，例如 TPLinker[23] 和 UNIRE[24]，通過維護關係表來擷取三元組，避免了誤差傳遞和暴露偏差問題，但在處理長文本時效率較低。

2.1 生成式聯合實體關係擷取

自 2018 年生成式聯合實體關係擷取方法如 CopyRE[25] 提出以來，研究逐漸轉向使用預訓練模型（如 UniLM[26]、Bart[27]、T5[28]）進行聯合實體關係擷取。例如，Ye 等人 [29] 提出基於 UniLM 的 CGT 對比學習方法，Cabot 等人 [30] 則以 BART 爲基礎提出 REBEL 架構。

生成式擷取可分爲多輪式生成和通用式生成。多輪式生成將實體和關係的擷取視爲多輪問答對話的問題，如 Li 等人 [31] 提出在前兩輪對話中抽取出實體對，後續的對話透由已知實體對抽取出關係。Wei 等人 [32] 則提出 ChatIE 這種逆向的做法，在第一輪對話中先抽取關係，藉由關係來尋找相關實體。然而，多輪式生成依賴於對話 schema 的設計，且需爲每種實體和關係設計專屬模板，隨著類別增多，模板設計變得更加複雜。

通用式生成希望模型直接輸出所需的答案結構，例如 Cabot 等人 [30] 將三元組表示爲文字序列。通用式生成方法更加直觀，不受 schema 限制，且能整合不同的 Information Extraction(IE) 任務，如 Paolini 等人 [33] 將不同 IE 任務視爲自然語言的翻譯任務處理。而 UIE[1]、LasUIE[34] 模型的提出，則可以將不同的 IE 任務整合成同一種答案結構。

綜合言之，隨著大型語言模型的發展，通用式生成方法在 RE 任務上展現出色表現，即使不經過微調，僅給定良好的 Instruction，也能達到 SOTA 效能 [2]。

3 Dataset Preparation

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，也就是資料來源單一的問題。例如，NYT[1](New York Times Annotated Corpus) 來自紐約時報，DocRED 來自維基百科。此外，像是專門領域的資料集，如 GDA[2]（基因-疾病關聯語料庫）和 CDR[3]（化學物質-疾病關聯），則來自 PubMed 等生物醫學文獻庫。

爲了創建一個更具彈性和全面性的資料集，能跨越特定領域的界限，我們利用了 Common Crawl[4]，這是一個包含各種領域和文章類別的網頁文章存檔，涵蓋了多種寫作風格、觀點和主題。使我們的資料集能夠反映現實世界多樣的文本數據。

3.1 Common Crawl 數據庫前處理

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，橫跨 160 多種語言。其中 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，我們擷取了其中的 990 個 Segments，並分成 11 個 shards 進行處理。數據處理流程採用 CCNet [35] 所提出的做法，將步驟分爲: 去重、語言辨識、品質篩選等三個 pipeline 步驟。

在預處理階段，首先對每個段落進行小寫轉換，將所有數字替換爲佔位符，並消除所有 Unicode 標點符號和重音符號。

1. 去重: 去重過程通過計算每個段落的 64-bits SHA-1 雜湊值來實現。主要目標是確保網頁內容唯一性，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、聯絡資訊等冗餘訊息，進而降低後續語言辨識的難度。

2. 語言辨識: 我們使用 fastText[36] 作爲語言分類器，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，我們透過該方法篩選出得分大於 0.5 的中文網頁內容。

3. 品質篩選: 最後對文件進行品質篩選。首先，我們使用 Sentence Piece tokenizer[5]把每一個網頁在句子層次做 tokenize，然後使用評分語言模型來爲每一個自然段做評分，我們使用 KenLM[6] 庫裡面的 Kneser-Ney 作爲評分語言模型，

[1] https://catalog.ldc.upenn.edu/LDC2008T19
[2] https://bitbucket.org/alexwuhkucs/gda-extraction/src/master/
[3] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/
[4] https://commoncrawl.org/
[5] https://github.com/google/sentencepiece
[6] https://github.com/kpu/kenlm

其所評的分數代表困惑度 (perplexity 簡稱 ppl) 分數，ppl 的分數越低，文本品質越高。代表其行文越流暢、越有邏輯。

總結來說，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，品質狀況爲頭部 > 中間 > 尾部。我們通過該方法過濾掉了大量品質不佳的網頁資料。只留取頭部約 17% 的高品質資料作爲我們的資料來源。其中 shard 0 的頭部 (head 0)，共 26,293 筆資料，作爲測試資料集的來源，而 head 1 ~ head 10 共 260,469 筆資料作爲訓練資料集的來源。

4 Data Annotation

在這個研究中，我們選擇以人物作爲實體，並定義了 [親屬、師生、同事、其他] 作爲我們標記的 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) 作爲標記工具。

標記流程主要分爲四個階段 (標記流程如圖1)：

三元組生成：讓兩個模型分別處理所有文本，生成潛在的人物關係三元組。關係分類：將生成的三元組分類至預定義的四種關係類型中。交叉驗證：對兩個模型的分類結果進行比對，以解決模型分歧。資料合併：將經過交叉驗證的結果合併，形成最終的標記資料集。

爲確保標記資料的可靠性，我們採用了雙模型交叉驗證的機制。如此一來，可有效避免單一模型可能產生的偏誤，提升標記結果的準確性。因此，除非另有說明，後續的分析皆以經過雙模型驗證的測試集爲基礎。

由於訓練資料約爲測試資料的 10 倍，爲節省 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2，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。

4.1 文章級別挑戰

現有的生成式模型多在句子級別資料上進行 few-shot 學習 [37, 2]，以實現聯合實體關係擷取。然而，我們的網頁資料篇幅較長 (平均字數 1,502 字)，若切割文章恐導致跨句實體關係的遺失。因此，我們希望模型能直接在整篇文章上進行關係擷取。

雖然我們部屬的 Gemini-1.0 和 GPT-3.5 的 context window 可以達到 16k 個 tokens。但當在長篇文本中引入 few-shot learning 範例時，過長的提示反而會影響模型性能。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，分別測試 1-shot 及 2-shots 效能，並且使用 Wadhwa 等人 [2] 的方法，人工針對這 2 筆測試資料加上 chain-of-thought(CoT) 的模板，具體的 prompt 內容詳見附錄A1。實驗結果發現，模型在給定文章級別的範例後，無論是 1-shot 或是 2-shot，其性能並未顯著提升。因此，我們決定在後續實驗中採用 zero-shot 的方式。

4.2 三元組生成

爲使大型語言模型能直接生成關係三元組，我們嘗試以通用方式引導模型。然而，在 zero-shot 設定下，同時完成命名實體識別（NER）和關係抽取（RE），並將關係類型限制在四種分類（親屬、師生、同事、其他）並不容易。因此，我們決定先讓模型聚焦於找出「有關係的人名」，暫不強求其分類關係。爲了確保模型產出符合預期格式，我們設計了 zero-shot 提示詞（如附錄A2），並採用輪對話方式，當輸出結果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(如附錄A3) 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。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。標註結果可分爲正確標註（符合格式的三元組或準確判斷無關係）和錯誤標註三類：格式錯誤（如出現四元組）、無法識別（模型回覆偏離主題）以及 API 異常（因內容不當導致模型無法回應）。

Table 1: Gemini 和 GPT 成功標記文章分析

	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

由上表1中可以看出，Gemini 在錯誤回覆的數量上，相比 GPT 還要少許多，對於格式的可控性較爲良好。再進一步統計兩個模型所正確標記的資料如表2，可以看出大部分的網頁文章是不具有人名之間的關係的，這也符合我們的認知預期。透過大型語言模型的生成標記，即可過濾掉 8 成以上的無關係網頁文章。

4.3 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，並沒有侷限在我們所定義四種類型中。具體統計在

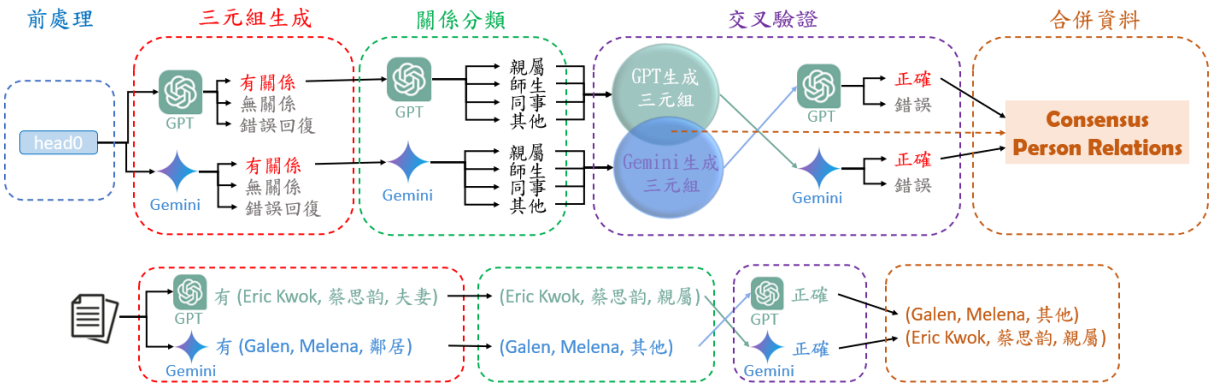

Figure 1: 測試資料集的流程 (上半) 及範例 (下半)

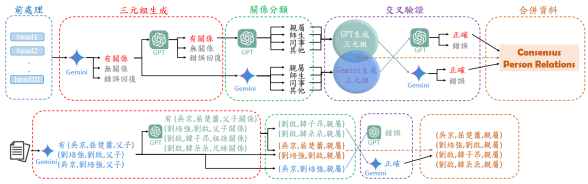

Figure 2: 訓練資料集的流程 (上半) 及範例 (下半)

Table 2: 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%

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爲我所定義的四種分類 (親屬、師生、同事、其他)。

因此，我們將三元組生成關係特別取出，並將其視爲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。具體指令如附錄A4所示。我們統計兩模型各自分類的數據結果，如表3。可以看出無論 Gemini 或是 GPT 都是以其他類別的種類佔多數。代表我們生成出的關係，多數被認定爲我們所指定的親屬、師生、同事之外。

Table 3: 模型標註的關係種類統計

	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%

4.4 交叉驗證

爲驗證 LLM 標記的準確性，我們選取 Gemini 和 GPT 認爲含人物關係的 4,619 筆網頁資料聯集 (見表2)。我們對其生成的三元組進行關係分類，並統計兩模型的資料筆數和三元組數量 (見表4)。

比對時，我們採用嚴格標準: 三元組文字須完全相同，但實體順序不拘。由於標記資料可能有簡繁體混用，我們使用 OpenCC[7]將所有文字轉爲繁體後再比對。

Table 4: 有關係網頁資料筆數和生成三元組數量

	Gemini	Gpt	Inter	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
三元組數量/doc	2.95	2.40	-	-

結果顯示，Gemini 和 GPT 共同認定 1,225 篇文章具人物關係，但共同認定的三元組僅 1,027 組，少於共識文章數。兩模型分別生成 6,697 和 8,595 組三元組，交集僅 1,027 組，佔比分別爲 15.34% 和 11.94%。這表明兩者生成的三元組差異較大，難以直接取得共識。

因此，我們簡化任務爲二元分類，讓兩模型互相評估對方生成的三元組是否正確。我們直接採用 1,027 組共識三元組，僅對無共識三元組提問 (具體指令見附錄A5)。指令中列出四種常見錯誤:A. 關係錯誤; B. 實體非人名; C. 實體僅爲稱謂; D. 兩實體相同。

交叉驗證結果見表5。儘管初始生成的三元組差異大，但兩模型對彼此生成的三元組認同度均超過 90%。進一步分析發現，當網頁含大量人名 (如演員名單、出賽名單等) 時，模型往往只取樣部分實體對並認定關係，導致標註不一致。關於模型識別人名實體的能力，我們將在章節5.1中深入探討並提出解決方案。

[7] https://github.com/BYVoid/OpenCC

Table 5: 兩模型交叉驗證通過統計

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

4.5 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，視爲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6。

Table 6: 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	佔比	# 文章數	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，在網頁文章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%。相比表3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，這是因爲我們將其他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。

5 Data Quality Evaluation

5.1 NER 效能評估

我們在章節4.4中透過分析 Gemini 和 GPT 的三元組交集，以及通過交叉驗證的比例，發現兩模型的所標記的三元組差異很大，但卻又很高比例的認同對方所標記爲正確。我們在前面只透過觀察實際案例，推測是兩模型所能找到的人名實體對過少，所導致的標記不一致。

爲了驗證我們的猜想，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 模型爲中研院所開源的 ckiplab/bert-base-chinese-ner[8](後稱 CKIP)。我們將 CKIP 所標記出來的人名視爲標準答案，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，我們以每筆文章中的實體爲單位，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 翻譯爲繁體中文後，在嚴格比對下需要完全一致才視爲相同實體，最終效能如表7。

Table 7: 將 CKIP 所標的實體視爲答案，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%。代表有超過 80% 以上的人名之間都沒有關係。因此，我們將在章節5.4中，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。

由表7還可以注意到一點，就是 Gemini 和 GPT 兩者所標記的人名實體 Precision 介於 51.92% ~ 66.48% 之間，可以看出大型語言模型找出了很多的人名實體是傳統的 SOTA-NER 所無法標記的。而這其實主要是由於三個原因所造成: 1. 嚴格比對落差: 由於我們的比對是使用嚴格比對，字符必須完全一致。但 LLM 在生成三元組時，常常會把如:" 先生"、" 小姐".. 等稱謂給生成出來。進而導致與序列標記的實體不一致。2. 外文實體: 我們的資料來源包含中英文或其他外語參雜的語法。但我們的 SOTA-NER 是使用 bert-base-chinese，對於英文、日文或其他外文的實體是無法辨認的。但是如 Gemini、GPT 等這些 LLM，在這些跨語言參雜的文章中找出實體是沒甚麼問題的。3. 幻覺生成: 目前幻覺問題一直存在於所有生成模型，Gemini 和 GPT 也不例外。因此，我們將在章節5.2中統計 Gemini 和 GPT 的實體幻覺比例。

5.2 幻覺評估

由於我們使用的生成式方法，可能產生不存在於文章中的實體。因此，我們以實體個數爲單位，檢查 Gemini 和 GPT 所生成出來的

[8] https://github.com/ckiplab/ckip-transformers

人名實體，是否存在原本文章中，具體統計如表8。可以發現約有 3% 左右的幻覺實體存在，而這些幻覺實體和其他實體組成實體對時，會讓含有幻覺實體的實體對比例增加到約 5% 左右，如表9所示。

Table 8: 模型所幻覺的人名實體統計，以實體個數爲單位

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

5.3 跨句評估

在跨句評估上，我們先將文章切割爲句子級別，使用中文常用的標點符號 [\n。；;！!？?] 和換行符號進行切割。平均每個文章會被切割爲 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，則視爲跨句子的實體對。

Table 9: 統計模型的跨句找出實體關係能力，以實體對個數爲單位

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

我們可以由表9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。而我們進一步分析實體跨度，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10所示，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。

Table 10: 跨句實體對中，實體對最小間隔字數

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

5.4 實體關係擴充

我們在章節5.1中看出，模型在 NER 任務的 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3。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，再對需擴充的實體對

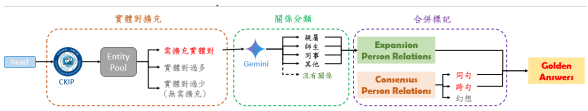

Figure 3: 實體關係三元組擴充流程

進行關係分類，最後合併時，去除幻覺實體關係。

由於實體對擴充時，若實體數量爲 n，實體對數量則會增爲 $\binom{n}{2}$。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: 第一個條件爲人名密度，我們定義爲: 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爲 0.95/100 字，我們取 2 倍的平均人名密度作爲我們第一個篩選條件。第二個條件則是實體對上限值，若實體對超過 $\binom{15}{2}$ 也就是 105 組時，我們發現很難讓模型在一次 request 中完成。

相對的，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，我們都視爲實體對過少的文章。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11。

Table 11: 實際需擴充佔比

	文章數	佔比
實體對過多	776	17.19%
無需擴充	1,525	33.78%
需擴充	2,214	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，並新增一個沒有關係的類別，而若模型分類爲沒有關係，則不會取用該實體。具體指令如表A6

在完成關係分類後，我們將會獲得擴充的人物關係三元組。接著，我們會在資料集中剔除實體對過多的 776 筆文章，使剩餘網頁文章數量來到 3,739 筆。隨後在去除共識人物關係中含有幻覺實體的三元組。在最終我們留下了 **3,583** 筆網頁文章，並得到共 **30,407** 組 (表12) 三元組，爲我們最終標記結果 (Golden Answers)，其中包含擴充的 21,632 組三元組，以及 Gemini、GPT 共識的 9,164 組三元組。

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12。可以看出我們所擴充的類型中，同事的類別佔有極爲懸殊的大量。符合我們在章節4.5中的預期，在眞實網路文章

中，包含同事關係的文章佔比確實較高，進行實體對擴充會讓同事關係數量暴增。

Table 12: 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	佔比
親屬	783	2,231	2,967	9.76%
師生	853	157	985	3.24%
同事	3,988	18,826	22,667	74.54%
其他	3,540	418	3,788	12.46%
總數	9,164	21,632	30,407	100%

5.5 實體對評估

在得到我們的 Golden Answers 後，我們可以回頭評估當初 Gemini 和 GPT 兩者所生成的實體關係二元組、三元組的效能。在比對二元組時，只要相同的關係實體對，我們視爲相同二元組。在表13中，我們將 Golden Answers 所取得二元組，視爲我們的 label 值。而測試 Gemini 和 GPT 在經過章節4.2三元組生成步驟後，所產生的實體對效能。

我們可以發現，由於該 label 是包含了 Gemini 和 GPT 的共識，兩者對於對方的標記大都表示認同，所以兩者的 Precision 都超過 90% 以上。但是兩者的 recall 則分別爲 14.18% 及 19.06%，可以看出 Golden Answers 所涵蓋的實體關係三元組範圍，較單一模型所能標記出的效能還要多出許多。

另外，我們也將 CKIP 所生成的實體兩兩組成對後，計算二元組效能。我們發現與 Gemini 和 GPT 不同的是，CKIP 組成的實體對 Recall 達到 82%，可以涵蓋大多數的實體對，但是由於會產生許多沒有關係的實體對，導致其 Precision 相對不佳。

Table 13: 將 Golden Answers 中實體對二元組視爲 labels，評估實體對生成效能

	Recall	Precision	Micro-f1
Gemini	14.18%	93.06%	24.61%
GPT	19.06%	93.06%	31.64%
CKIP	82.21%	41.37%	55.04%

5.6 人工標記

除了自動標記系統外，我們也進行少量人工標記。由志願者標註共 20 筆測試集資料，以此評估 GPT, Gemini, mT5 系統的效能。結果如表15，在表現最優的 CKIP+mT5 的 pipeline 系統，可以達到 24.79% recall 及 50.00% precision。觀察人工標記的結果，如同段落5.1所示，嚴格比對落差、外文實體、幻覺生成，是導致標註錯誤的重要原因，另外人名選取不完全、分類關係不同也會導致效能降低。

6 Model Training

在取得了由 Common Crawl 中清洗，以及透過我們的標記流程、擴充方法得到 Golden Answers 的資料集以後，我們想驗證該資料集是否能夠在參數量較小的預訓練模型上進行微調訓練，達到通用實體關係擷取的目標。

由於考慮到設備記憶體資源有限，我們將文章給截斷至最長爲 1,024 字，因此在我們最終的 Golden Answers 中，我們也把實體未出現在前 1,024 字中的 label 給去除，所以我們的測試資料筆數由 3,583 筆減少爲 **3,392** 筆，三元組數量由 30,407 組，減少爲 **21,255** 組，訓練集和驗證集也有小幅減少。

另外，爲了評估資料集規模對於模型效能的影響，我們另外進行了小規模數據的實驗。我們將 test 資料集中的 3,392 筆文章再切割成五份，進行 5-fold 交叉驗證，降低模型訓練對於資料集的偏差。

而在評估模型的效能時，我們將分別評估二元組效能與三元組效能。其中二元組效能，即爲章節5.5的實體對評估，我們評估最終生成的三元組中實體對部分，且不考慮實體排序，而三元組效能則需包含分類的類別也需要正確。

6.1 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[38]作爲訓練的基底模型。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(如圖A2) 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。經過全微調實驗後，我們發現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，因此可省略關係分類步驟。

mT5 的實際效能如表14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。

Table 14: 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 %	33.26%	23.41	14.55%	26.43%	18.76
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 %	74.94 %	73.21	64.11%	67.52	66.80
CKIP+mT5	pipeline	train data	59.50%	71.29%	64.87%	53.26%	64.29%	58.26%

Table 15: 人工標記二元組效能與三元組效能

模型	方法	訓練資料	人工標記二元組效能			人工標記三元組效能		
			Recall	Precision	Micro-f1	Recall	Precision	Micro-f1
Gemini	通用式生成	NO	2.56%	100%	5.00%	1.71%	66.67%	3.33%
GPT	通用式生成	NO	12.82%	55.56%	20.83%	12.82%	55.56%	20.83%
mT5	通用式生成	train data	26.50%	40.79%	32.13%	23.08%	34.18%	27.55%
CKIP+mT5	pipeline	train data	25.64%	51.72%	34.29%	24.79%	50.00%	33.14%

6.2 pipeline

我們實驗了傳統 Pipeline 方法，使用 CKIP 的 bert-base-chinese-ner 模型進行 NER 任務。由於 Golden Answers 不包含無關的人物實體，因此未對 NER 模型進行微調。訓練時，我們將 CKIP 的人物實體兩兩配對，若該配對不在 Golden Answers 中，則將該關係標記爲" 沒有" 進行訓練。RE 任務爲 5 分類：[親屬、師生、同事、其他、沒有]，並使用 mT5 進行訓練。推論階段中，若預測關係爲" 沒有"，則刪除該實體對。由表14 可見，Pipeline 方法在小規模測試集上達到 73.21% 的 Micro-F1，遠高於 Gemini 和 GPT 效能。相較之下，在完整訓練集效能有所下降。因此對於簡單的分類問題，增加數據量不一定會有更好效果。

6.3 Gemini 實體擴充

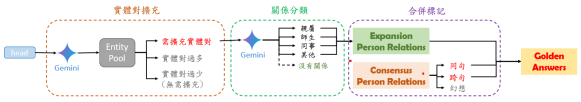

Figure 4: 使用 Gemini 取代 CKIP 實體關係三元組擴充流程

由於我們在章節5.4進行實體關係擴充，合併產生 Golden Answer 時，部分採用了 CKIP 產生的人物實體。而 pipeline 方法的實體也來自 CKIP，使得 Golden Answer 有些許不客觀。因此，我們針對實體關係三元組擴充進行些許調整。

在測試資料集中，我們將實體對擴充所使用的模型由 CKIP 替換爲 Gemini。也就是先由 Gemini 找出人物實體，再標記關係，如圖4流程。我們給 Gemini 如表A8的指令，讓其進行 NER 任務，找出所有的人物實體，並進行兩兩配對成實體對後執行關係分類。在更新 Golden Answer 後我們重新訓練通用式生成以及 pepeline 方法實驗，並使用更新後的 Golden Answer 作爲評估指標，重新評估所有的模型。

若由 Gemini 來進行 NER 任務，會讓擴充的實體關係和大型語言模型生成的結果雷同度提高，所以 Gemini 和 GPT 的效能會因此提高。而在通用式生成的方法中，整體的 Micro-f1 效能則並未有大幅度的改變。在 pipeline 的方法中，則可以二元組 micro-f1 由 64.87% 下降至 52.39%，三元組 micro-f1 也由 58.26% 下降至 46.21%。但 pipeline 的方法還是能有優於通用式生成的整體表現。

7 Conclusion

本研究在聯合實體關係擷取領域引入了基於通用式生成語言模型的自動化標記流程，利用 Gemini、GPT-3.5 等大型語言模型，取代了人工標記的需求，提高了標記效率。且對於文章級別的標記，克服了過去因文長限制而必須截斷文本或進行證據檢索的局限性。透過我們本次的中文文章級聯合實體關係擷取之研究的進展，我們利用 Common Crawl 的全網爬蟲資料庫，創建了多樣性和廣泛性的中文文章級別資料集，爲泛化的中文關係擷取研究提供了新資源。

最後，我們在參數量較小的 mT5 模型中實驗證實，該資料集所訓練出的模型，可用於泛化的眞實網路文章。總結來說，本研究不僅在技術方法上進行了創新，還爲關係擷取和命名實體識別研究提供了新的思路和資源。未來，我們的方法和成果有望進一步推動 NLP 技術

在實際應用中的應用和發展。

References

[1] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2] 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.

[3] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. ACE 2005 multilingual training corpus. In *Linguistic Data Consortium*, 2006.

[4] 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 HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics.

[5] 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.

[6] 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.

[7] Guoquan Dai, Xizhao Wang, Xiaoying Zou, Chao Liu, and Si Cen. Mrgat: Multi-relational graph attention network for knowledge graph completion. *Neural Networks*, 154:234–245, 2022.

[8] Linfeng Li, Peng Wang, Jun Yan, Yao Wang, Simin Li, Jinpeng Jiang, Zhe Sun, Buzhou Tang, Tsung-Hui Chang, Shenghui Wang, and Yuting Liu. Real-world data medical knowledge graph: construction and applications. *Artificial Intelligence in Medicine*, 103:101817, 2020.

[9] Jung-Jun Kim, Dong-Gyu Lee, Jialin Wu, Hong-Gyu Jung, and Seong-Whan Lee. Visual question answering based on local-scene-aware referring expression generation. *Neural Networks*, 139:158–167, 2021.

[10] Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online, July 2020. Association for Computational Linguistics.

[11] Marco Antonio Calijorne Soares and Fernando Silva Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 32(6):635–646, 2020.

[12] Weizhao Li, Feng Ge, Yi Cai, and Da Ren. A conversational model for eliciting new chatting topics in open-domain conversation. *Neural Networks*, 144:540–552, 2021.

[13] Yunyi Yang, Yunhao Li, and Xiaojun Quan. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14230–14238, May 2021.

[14] Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng, and Ying Shen. Chinese relation extraction with multi-grained information and external linguistic knowledge. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4377–4386, Florence, Italy, July 2019. Association for Computational Linguistics.

[15] Jiaqi Hou, Xin Li, Haipeng Yao, Haichun Sun, Tianle Mai, and Rongchen Zhu. Bert-based chinese relation extraction for public security. *IEEE Access*, 8:132367–132375, 2020.

[16] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[17] Ang Sun, Ralph Grishman, and Satoshi Sekine. Semi-supervised relation extraction with large-scale word clustering. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 521–529, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[18] Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy, July 2019. Association for Computational Linguistics.

[19] Changzhi Sun, Yeyun Gong, Yuanbin Wu, Ming Gong, Daxin Jiang, Man Lan, Shiliang Sun, and Nan Duan. Joint type inference on entities and relations via graph convolutional networks. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1361–1370, Florence, Italy, July 2019. Association for Computational Linguistics.

[20] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. A novel cascade binary tagging framework for relational triple extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online, July 2020. Association for Computational Linguistics.

[21] Bowen Yu, Zhenyu Zhang, Xiaobo Shu, Yubin Wang, Tingwen Liu, Bin Wang, and Sujian Li. Joint extraction of entities and relations based on a novel decomposition strategy. In *Proc. of ECAI*, 2020.

[22] Shuai Zhang, Yongliang Shen, Zeqi Tan, Yiquan Wu, and Weiming Lu. De-bias for generative extraction in unified NER task. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 808–818, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[23] Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[24] Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. UniRE: A unified label space for entity relation extraction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 220–231, Online, August 2021. Association for Computational Linguistics.

[25] Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[26] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019.

[27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[29] Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. Contrastive triple extraction with generative transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14257–14265, 2021.

[30] Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[31] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. Entity-relation extraction as multi-turn question answering. In Anna Korhonen, David

Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy, July 2019. Association for Computational Linguistics.

[32] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*, 2023.

[33] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[34] Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. *Advances in Neural Information Processing Systems*, 35:15460–15475, 2022.

[35] 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.

[36] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[37] 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.

[38] 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.

A Appendix: Prompts for guiding LLM annotation

Table A1: few-shots 的 promt 格式及內容

1-shot
請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:Relations: 無即可 若有請依以下格式回答: Relations: 有 (人名, 人名, 關係),(人名, 人名, 關係)... 列舉出所有關係 Explanation: 解釋原因 範例如下: 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 下略 **805** 字**....** Relations: 有 (刘欣雯, 周国强, 夫妻) 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爲夫妻關係 文章如下: TEXT:{document}
2-shots
請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:Relations: 無即可 若有請依以下格式回答: Relations: 有 (人名, 人名, 關係),(人名, 人名, 關係)... 列舉出所有關係 Explanation: 解釋原因 範例如下: 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 下略 **805** 字**....** Relations: 有 (刘欣雯, 周国强, 夫妻) 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爲夫妻關係 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 下略 **2,843** 字**....** 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) 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爲師生關係，劉浩志與張敬萱也爲師生關係 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爲同事關係，劉浩志與蔡佩珍也爲同事關係 文章如下: TEXT:{document}

Table A2: 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

```
請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
若有請列舉出所有關係並依格式回答: 有 (人名, 人名, 關係),(人名, 人名, 關係)
文章如下: {document}
```

Table A3: 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。

```
請務必依照規定格式回答，若無關係直接回答: 無，
若有請依 2 個人名實體和 1 個關係格式回答: 有 (人名, 人名, 關係),(人名, 人名, 關係)
```

Table A4: 關係分類 Prompt，該指令設計爲簡單的四元分類問題

```
請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?
```

Table A5: 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

```
分析以下文章中的人名關係三元組 (人名, 人名, 關係)。找出親屬、師生、同事等三種關係，其餘標爲其他，即類別：[親屬、師生、同事、其他]。
文章如下: {document}
關係如下: {1.(邰智源, 林柏昇, 其他) 2.(邰智源, 泱泱, 其他) 3.(邰智源, 温妮, 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 種情形視爲錯誤:
    A. 關係錯誤，例如:(蔣中正, 蔣經國, 同事)，正確關係應爲 (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)，因爲" 共產黨" 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，例如:(湯姆·克魯斯, 妻子, 親屬)，並沒有給出妻子姓名。
    D. 兩個人名相同，例如:(徐志摩, 徐志摩, 其他)，兩個人名相同即視爲錯誤。
請依格式回答: {1. 正確/錯誤 2. 正確/錯誤 3. 正確/錯誤}
```

Table A6: 實體對關係分類 Prompt，讓模型判斷每組人物實體對的關係

```
根據以下文章，判斷文中每組人名實體對的人物關係。人物關係分爲親屬、師生、同事、其他、沒有，共 5 種類型。

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 2. 親屬/師生/同事/其他/沒有 3. 親屬/師生/同事/其他/沒有 }
```

Table A7: 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

```
根據以下文章，找出 {person1} 與 {person2} 中之間的關係。關係分爲: 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}
```

Table A8: NER 任務 Prompt，讓模型找出所有的人名實體

```
請找出以下文章中所有的人名，並依格式回答:(人名 1, 人名 2, 人名 3...)，若文章中沒有具體人名，
則回答: 無
文章如下:
{document}
```