**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 | 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 | 聯絡方式：  0228891088  董先生 #55334  陳小姐 #55335  孫小姐 #55675  (上班時間無法使用手機，比較急時打電話聯絡) |