玉山人工智慧

公開挑戰賽

隊伍：Brainchild

成員：徐正憲，劉家達，游璿達，黃郁傑，莊子逵

★摘要[請簡單說明本次比賽所使用過之特徵、演算法以及訓練模型的方式]

用 Bert 分別訓練四階段模型

1. **犯罪模型（Bert + BiLSTM + Dense）**：將有犯罪事實標為 1，與犯罪無關標為 0 的資料訓練。初步篩選包含犯罪之新聞
2. **AML 犯罪模型（Bert(微調) + Dense）**：將犯罪且與 AML 有關標為 1，犯罪且與 AML 無關標為 0 的資料訓練。用以篩選有 AML 相關犯罪之新聞
3. **NER 模型（Bert + BiLSTM + CRF）**：用 CKIP 初步辨識並篩選出人名（包含三字、兩字簡稱及單名），以此訓練 NER 模型。辨識新聞中所有人名、簡稱及單名
4. **人名 AML 模型（Bert(微調) + Dense）**：取官方原始 331 筆包含 AML 人名新聞中所有人名的前後句訓練，將 AML 人名的前後句標為 1，非 AML 人名的標為 0。以前後句判斷是否為 AML 人名

★環境[請說明本次比賽所使用的系統平台、程式語言、函式庫]

虛擬環境：GCP + Ubuntu 18.04

程式語言：Python

函式庫：flask, hashlib, codes, os, re, collections, numpy, pandas, tensorflow, keras, keras_bert, keras_contrib

★特徵[請說明本次比賽所使用的特徵]

1. 資料前處理：

   (1) 將原始新聞刪除< >、【 】、( )、〔 〕中的字

   (2) 排除記者…報導及「 」中長度四以下的字，防止假名納入

2. 犯罪模型、AML 犯罪模型：

   (1) 若長度大於 512 則取首尾 256 字預測

3. NER 模型：

   (1) 用維基百科內百家姓協助判斷並篩選正確人名

4. 人名 AML 模型：

   (1) 取包含該人名的句子，並加上前後各一句（以。，？；切）

   (2) 若前後句遇到句點則取到該句為止，或從句點開始取

   (3) 若前後句包含其他人名則捨棄該句

   (4) 若中間句包含其他人名則置換成空字串

5. 規則：

   (1) 若中間句及後句含有：無罪定讞、確定無罪、無罪確定、罪嫌不足、罪證不足、不起訴則從預測結果排除

   (2) 若中間句及後句人名前含有：檢察官、員警等職稱，人名後含有：說、調查、辦理、偵訊、訊問、諭令等字眼則從預測結果排除（此為 CKIP 斷詞後取其動詞統計後之結果）

★訓練模型[請說明本次比賽所使用的訓練模型、參數]

1. 犯罪模型：
   (1) Bert + BiLSTM(128) + Dense(1)
   (2) maxlen = 512，batch_size = 8，epochs = 3，threshold = 0.4
2. AML 犯罪模型：
   (1) Bert 微調 Encoder-12 + Dense(64) + Dense(1)
   (2) maxlen = 512，batch_size = 8，epochs = 4，threshold = 0.3
3. NER 模型：
   (1) Bert + BiLSTM(128) + CRF(3)
   (2) maxlen = 512，batch_size = 8，epochs = 3
   (3) 若新聞長度大於 512 小於 1024 則以句點平分成兩句，若大於 1024 則以句點平分為三句
4. 人名 AML 模型：
   (1) Bert 微調 Encoder-12 + Dense(64) + Dense(1)
   (2) maxlen = 256，batch_size = 8，epochs = 3，threshold = 0.4

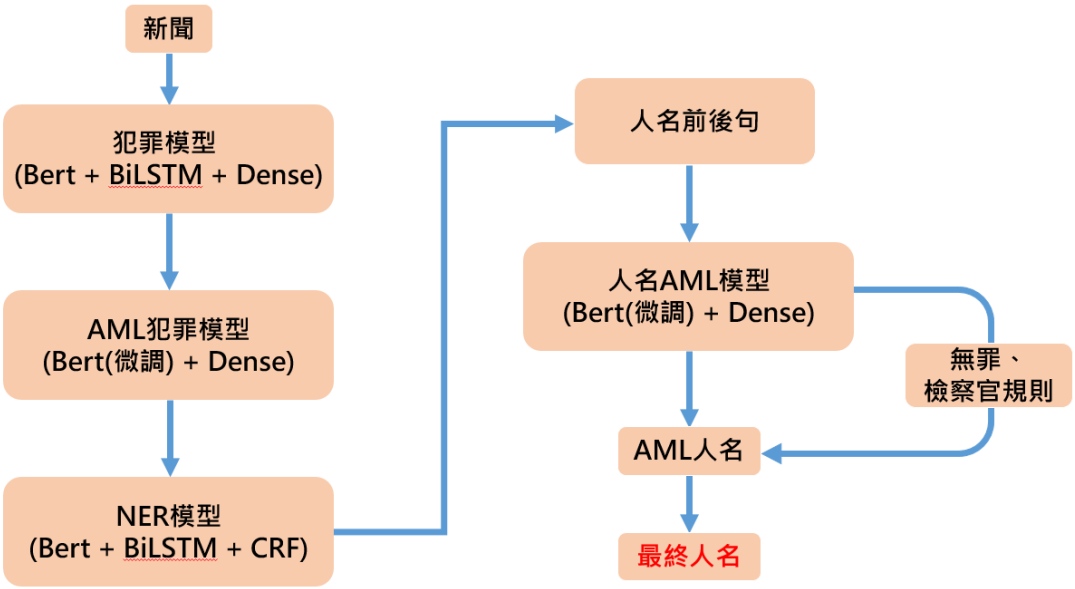
※以上模型 optimizer 均為 AdamWarmup，lr = 1e-3 ~ 1e-5

※NER 模型 loss 為 crf_loss，其餘為 binary_crossentropy

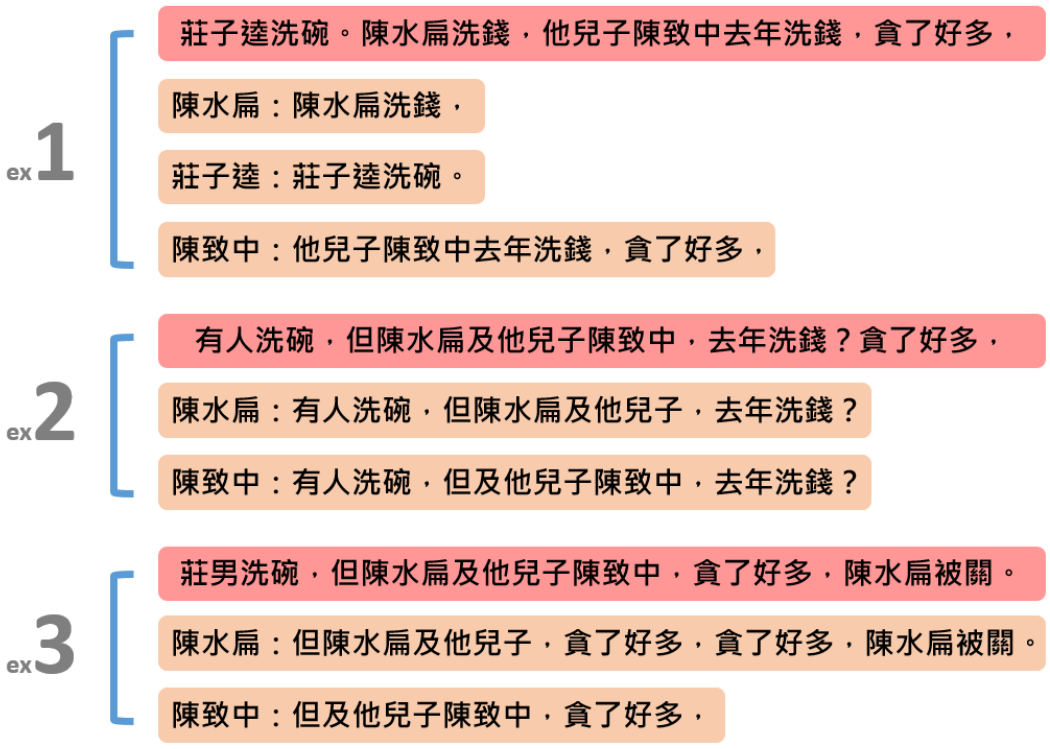
★訓練方式及原始碼 [請說明本次比賽答案的產出方式並提供有效之原始碼(連結亦可)]

原始碼：https://github.com/jasonliu1990/esun_summer_game_2020

答案產出流程：



人名前後句取法範例：



★ 結論[請簡易說明本次競賽後所得的結論]

此競賽主要勝負關鍵在於能否準確提出 AML 相關、並排除其他無關人名，本組在多次嘗試之後提取 AML 人名的 f1_score 約介於總分 88%～92%間，預測錯誤的大部分為犯罪事實與人名相距過遠，導致前後句接與 AML 無關的情況；另外少部分為由於模型架構、句子取法、資料少的缺陷，目前無法完美解決的父子問題（ex：陳其邁的父親陳哲男貪汙）、檢察官問題（ex：檢察官陳英俊表示，嫌犯於本月遭逮捕）及無罪問題（ex：陳水扁一審無罪，經再審後改判有罪），僅能透過簡單規則篩選。另外由於不確定官方 AML 標準為何，無法自行增加資料，因此若要從根解決可能需要引入指代消解的方法。