聞到有先後:基於多元文本的段落生成以及加值系統實作

ChronoNews: Multi-Source Textual Paragraph Generation and Value-Added System Implement

指導教授
高宏宇 Hung-Yu Kao 陳朝鈞 Chao-Chun Chen
專題生
陳冠廷 鄭宇辰

大綱

- 動機
- 核心技術
- 架構
- 結語

動機

動機

- 台灣的新聞較著重於在地事件報導，國外媒體有較多與國際事件相關的內容
- 想了解趨勢脈動與熱門事件
- 鎖定客群為愈接觸國際趨勢之非母語人士
- 報導繁多，希望可就同一事件的不同報導產生一段共同摘要
- 以懶人包格式，幫助使用者快速了解事件始末

核心技術

新聞分類

- Tech.
  - scikit-learn SVC model
- Implement
  - pickle 匯出 .pkl file

	Precision	Recall	F1-score	Support
Business	0.95	0.96	0.96	164
Entertainment	0.99	0.99	0.99	113
Politics	0.96	0.96	0.96	113
Sport	0.99	1.00	0.99	146
Tech	1.00	0.98	0.99	132
Macro Avg	0.98	0.98	0.98	668
Weighted Avg	0.98	0.98	0.98	668

圖一、SVC模型訓練數據

新聞標籤生成

- Tech.
  - NER
    - spaCy, en_core_web_sm
    - scikit-learn, TfidfVectorizer
  - Tfidf Vectorizer
    - 初步挑選 3 個與報導內文相似度最高的 Named Entity
- Implement
  - 從 trend keyword 的一篇至多篇報導中，建立Named Entity list
  - 隨機從 Named Entity list 選出 3 個 Named Entities 建立News Tag list

新聞內容問答

- Tech.
  - HuggingFace Transformers
- Implement
  - bert 模型可以接受的input 長度有限，故利用sliding window=5 概念將原始文章切出數個段落，每個step 為兩個句子，因此段落之間會有一定的重疊，以避免段落內容被破壞
    - 「問題表示」與「段落表示」並用 cosine 相似度最高者將被取出
  - 將取出的文章段落與問題放入QA 模型
  - （生成常見問題類似Q&A模板功能）

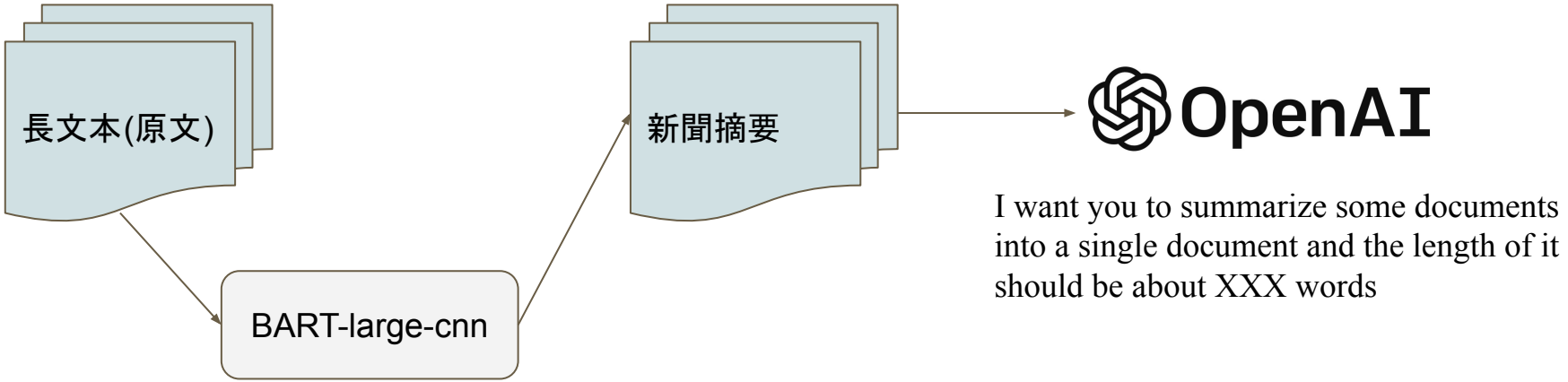
跨時間事件抽取與整合

- Tech.
  - NER, spaCy, en_core_web_sm
- Implement
  - 為每一個文章取出關鍵人物或事件的關鍵句
  - 確定句子中是否有關鍵人物或事件(字串比對)
  - 為比對成功的句子評分
    - 單位句子長度的 entity 個數越多，代表句子通常含有越豐富的語意，也就越重要
    - 不同 entity 會獲得不一樣的分數，特別著重於人物與組織
    - log2(句子長度) 為分母，以降低句子長度的影響效力
  - 取出最高分的句子在時間軸上代表該篇新聞的關鍵內容描述

文本摘要

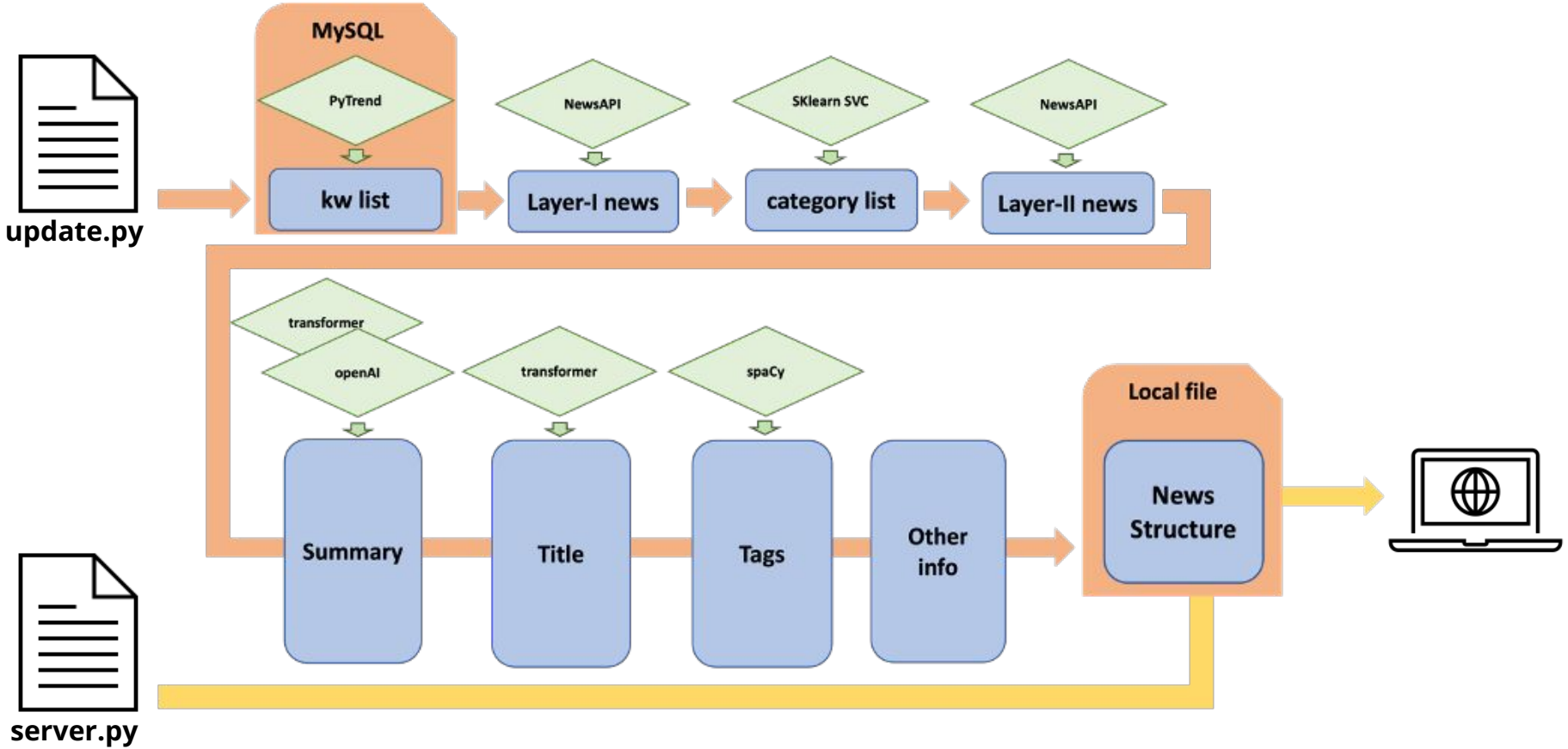
- Tech.
  - huggingface transformer
  - openai chatGPT-3.5
- Implement
  - 單篇文本摘要使用bart-large-cnn，finetune 於CNN Daily Mail 的模型
  - 使用 GPT-3.5 作為模型，並使用客製化的prompt 來產生跨文本摘要
  - 再跨文本摘要中因為各篇新聞內容冗長，放入模型中容易超過允許的輸入長度，同時為確保能夠最大地保留原文資訊，我們將原文先做初步摘要，再將其格式化後輸入chatGPT
  - 我們有嘗試使用multi-news SOTA 模型但發現 chatGPT 能夠生成更加通順且具有整合性的內容，而不只是拼貼式的整合

跨文本摘要架構



圖二、跨文本摘要流程示意圖

架構


update.py
MySQL
PyTrend
kw list
NewsAPI
Layer-I news
SKlearn SVC
category list
NewsAPI
Layer-II news
transformer
openAI
transformer
spaCy
Summary
Title
Tags
Other info
Local file
News Structure
server.py

圖三、系統架構示意圖

SWOT 分析

S

- 國際新聞與熱點趨勢
- 跨多媒體的整合
- 自然語言技術加值，以加速讀者獲取資訊的時間
- 良好的互動介面與視覺化

W

- 無法處理及時事件
- 沒辦法處理所有新聞媒體
- 資料處理速度有上限
- API token 額度有限
- 目前僅能處理英語新聞

O

近期大眾更傾向於快速且有效地獲取資訊(如:10分鐘看完電影)，整理大量資訊與懶人包功能需求增加

T

- 還有其他更具規模的類似平台

結語