資料科學與社會研究跨域專長
實習心得分享

意藍資訊-研發二部 RD2

B06109022 呂紀廷

公司、部門、實習專案介紹

- 意藍資訊提供網路口碑觀測與分析服務
- 研發二部RD2主要運用機器學習做產品的研發
- 實習專案:機器學習自動生成-抽象式新聞摘要

專案困難點為:繁體中文的抽象式摘要

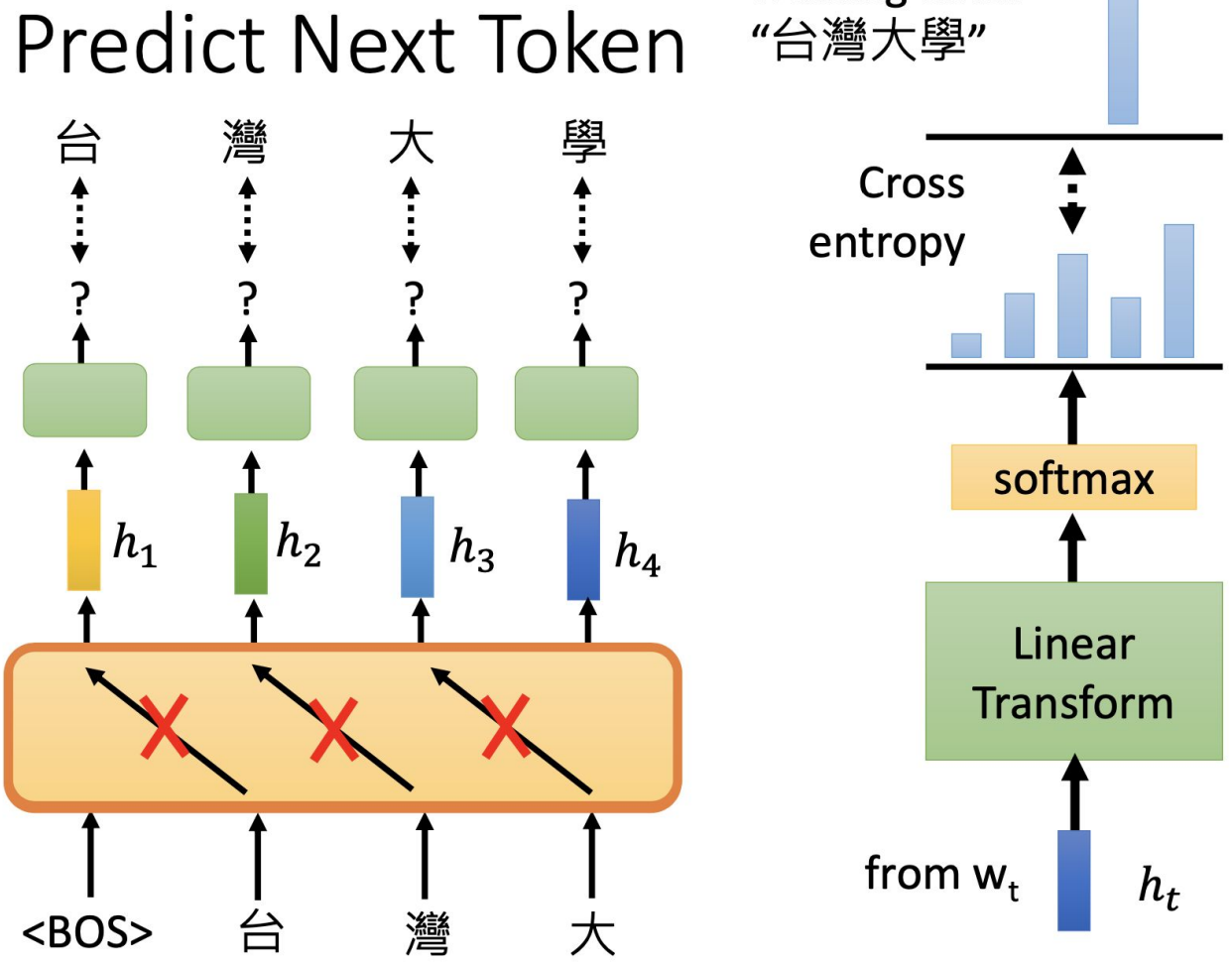
抽象式摘要

- 提取式摘要(Extraction-based summarization)
  - 選擇保留最重要觀點的單詞子集來總結文章
  - 統計PowerTerm分數，去抽取出較重要的句子，可能發生抓取上的錯誤
  - 為傳統的文章摘要方法
- 抽象式摘要(Abstraction-based summarization)
  - 透過機器學習抽象方法根據語義理解來總結文章
  - 能使用較短的句子來表達
  - 利用Natural Language Processing來生成新的文本摘要
  - 繁體中文的抽象式摘要在過往論文上仍不常見

解決方案:GPT-2模型

- 模型架構為 Transformers 裡的 Decoder
- 預訓練目標給定前 t 個在字典裡的詞彙, 語言模型要去估計第 t + 1 個詞彙的機率分佈 , 為一般的語言模型, 以此預測下個字
- 利用大量文本訓練出一個通用、具有高度自然語言理解能力的 NLP 模型, 再進行根據特定任務進行Fine tuning, 目前已釋出中文的GPT2模型
- 生成式模型, 適合文本生成、文本摘要、文本問答等生成任務

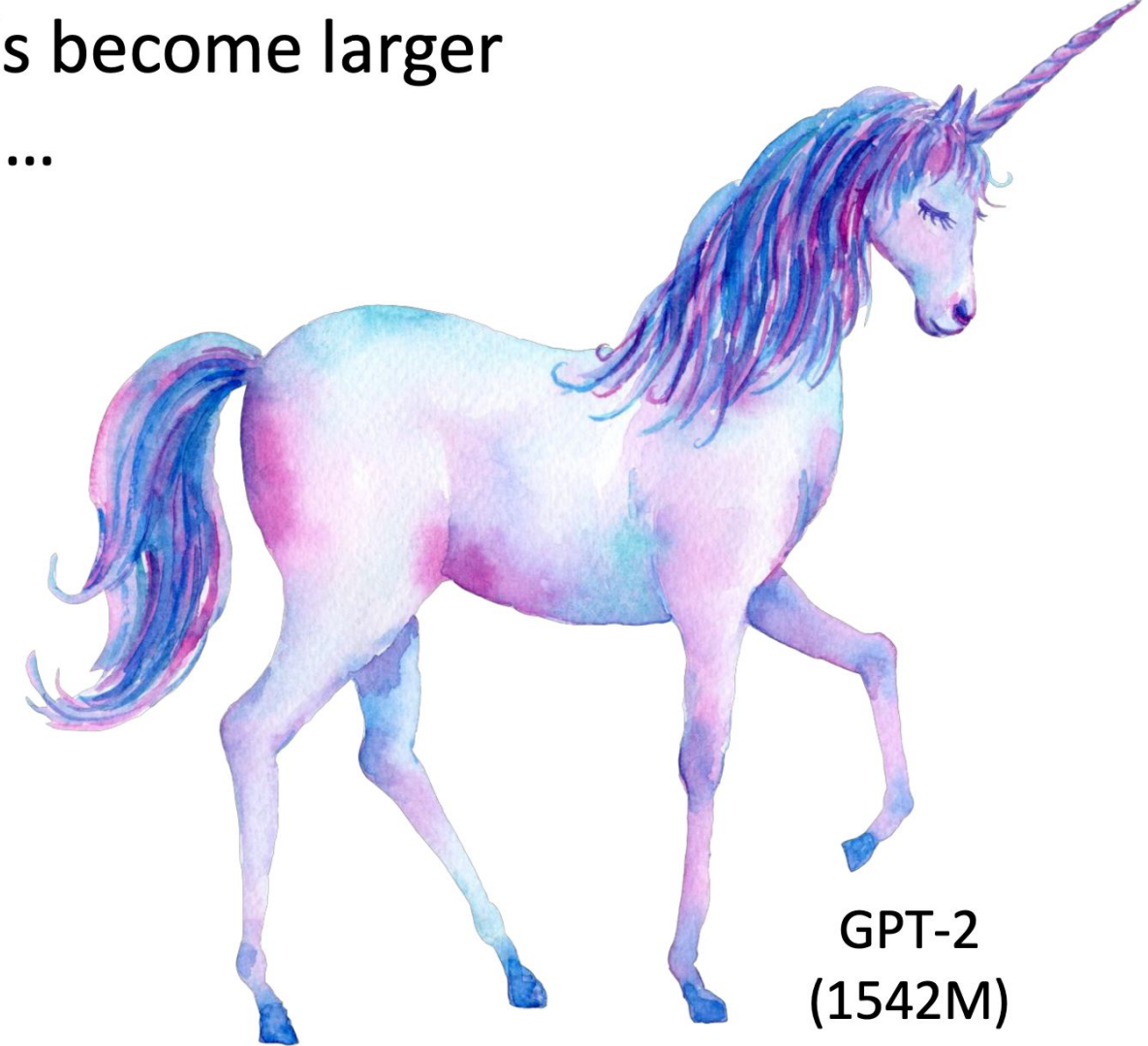
GPT-2預訓練任務


Predict Next Token
台
灣
大
學
?
?
?
?
h_1
h_2
h_3
h_4
<BOS>
台
灣
大
Training data:
“台灣大學”
w_{t+1}
Cross
entropy
softmax
Linear
Transform
from w_t
h_t

The models become larger and larger …

ELMO
(94M)

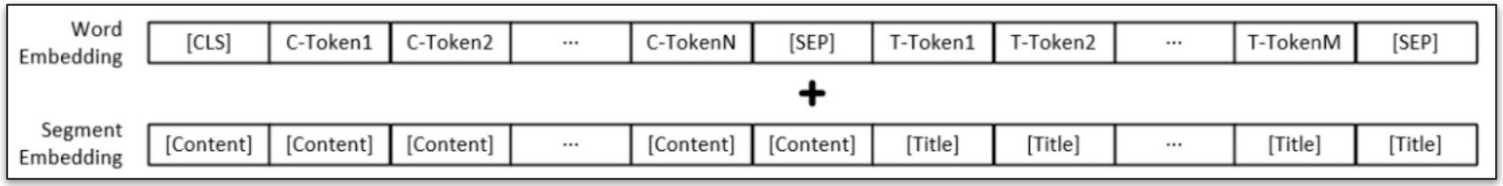
BERT
(340M)



GPT-2
(1542M)

GPT-2 Fine tuning

- 載入GPT-2 Chinese pretrained-model
- 搜集資料:意藍資訊新聞資料庫, 150000篇新聞
- 將訓練資料做處理:每筆資料皆有一篇文章內容以及對應的文章摘要
- 轉換成如下形式: [CLS] 文章內容 [SEP] 文章摘要 [SEP]



- 在原本GPT-2模型架構上, 加上一個全連接層, 使輸出維度為中文字典的大小
- 重新定義Loss函數:將資料丟入GPT-2生成摘要, 並只根據文章摘要部分來計算Loss
- 訓練過程中持續優化並下降Loss

GPT-2 摘要結果

印度神童新預言！「這4國」慎防新疫情　全球解封還要2年

被稱為「印度神童」的少年占星預言家阿南德（ Abhigya Anand），因預言武漢肺炎（新型冠狀病毒病， COVID-19）疫情爆發成真，受到各界矚目。日前他預測英國本月 20日會有大事發生，之後英首相強森宣布該國 6月21日延後解封，被外界認為阿南德的預言再次成真。16日他又再公布新預言，透露「 6月20日後可能發生的事」，曝光下一波疫情爆發的國家和時間點。阿南德 16日於YouTube頻道「Conscience」發布新影片，指稱本月 20日，木星將逆行回到土星，此現象對全球而言並非好事。他表示，木星逆行會從20日開始， 9月14日位置會落在摩羯座， 9月21日離開，預測該現象會影響全球疫情、經濟及股市。疫情方面，阿南德大膽指出，許多國家將會在本月 20日到9月21日爆發新一波疫情，受影響的國家包括英國、多數歐洲國家、印度、美國和日本。但阿南德也曖昧表示，下波疫情不會一夕之間爆發讓人措手不及，不過 6月20日是「某種改變」的開始。接著他又 說，等到7月火星和木星處於相對位置，全球動盪情勢會短暫回穩，但 9月至11月情況又會急轉直下；武肺疫情也得等到 2023年4月後才會真正好轉。國際經濟和股市方面，阿南德表示，本月 20日到9月21日會面臨困境， 11月後會改善，不過明年 4月可能還會再有一波低潮。

模[illegible]

阿南德再爆新預言 下波疫情恐再爆發

GPT-2 摘要結果

台積電陷疫苗爭議！遭外資降評等開盤股價跌破600元

受到上周五美股大跌影響，加上台積電陷入採購疫苗的爭議，遭到外資降評等，開盤跌出 600元大關，也拖累台股今天走勢，早盤一度重挫超過260點，摔出17300點關卡，分析師提醒，隨著美元走強，以及美股下殺，外資操盤也趨於保守，短期可能會面臨短期震盪。台股周一，出現殺盤賣壓，早盤一度跳水大跌 265點，摔出17100點關卡，面臨萬七保衛戰，晶圓雙雄台積電、聯電殺聲隆隆，就是受到上周五美股，道瓊指重挫超過 500點影響。儘管6/20全台確診數107例，創下三級警戒以來的新低，不過似乎無法為台股，加強走升力道。分析師王榮旭：「後續我認為還是要看美股的動向，因為最近美國股市的走勢是回檔之外，另外美元走強，這個會影響到外資，對於台股的操作，如果美股沒有辦法止跌，美元持續走強的話，恐怕外資還會再度調節台股。」除了外資操作轉趨於保守，市場也關注， 250萬劑莫德納抵台後，國 產疫苗需求是否暫緩，甚至在股價方面，高端疫苗先開低才走高，市場觀望氣氛濃。分析師王榮旭：「高端的影響，我認為莫德納的數量增加，應該不是最主要的股價影響變數，現在高端的股價走勢，就是比較呈現整理，應該就是在等待疫苗抗體效價的一個對比。」反觀代表政府，採購 BNT疫苗的台積電惹出爭議，一開盤失守600元大關，評等更被外資降至「中立」，也有股東質疑， 500萬劑BNT疫苗要花50到60億元，等於拿大筆資金用於非本業使用，一旦採購成功，將會影響第三季每股純益約 0.2元，而台積電闢謠，捐疫苗在 6月9日就獲得董事會支持。另外聯亞生技旗下的聯亞藥，也宣布將在 6/23再次登錄興櫃買賣，參考價為每股 30元，似乎也是趁， 7月國產疫苗施打前，搶攻多頭格局。

模型生成摘要

台積電陷入採購疫苗爭議 外資重挫逾百點

GPT-2 摘要結果

不要中國疫苗、不會外交轉向　外交部感謝瓜地馬拉總統堅持台瓜邦誼

我國中美洲友邦宏都拉斯急需疫苗緩和 COVID-19疫情，曾說為了取得疫苗而願在中國開設外交代表機構，而宏都拉斯鄰國、也是台灣邦交國的瓜地馬拉總統賈麥岱 2日接受路透專訪，直言沒興趣取得中國研發的疫苗，也不會捨棄建交 88年的台灣，轉投中國懷抱。不會轉投中國懷抱 對於賈麥岱（Alejandro Giammattei）表明維繫與台灣邦交，我國外交部發言人歐江安 3日強調，外交部誠摯感謝賈麥岱對台灣與瓜地馬拉邦誼的堅定支持，並稱瓜地馬拉是我國在中美地區的重要友邦，關係深厚友好，會在既有基礎上持續深化合作，攜手抗疫及促進雙邊永續發展。 路透3日登出賈麥岱的專訪 內容，提到瓜地馬拉有向美國尋求協助提供疫苗，「看起來他們是會給（疫苗）」，賈麥岱表示，美國提供的應是現有的阿斯特捷利康（ AstraZeneca，AZ）疫苗，但不知道數量和送達時間。他直言，瓜地馬拉不會跟宏都拉斯、薩爾瓦多一樣想要中國疫苗，因為疫苗有效率不高。 賀錦麗將訪瓜地馬拉賈麥岱亦稱，基於對長期盟邦台灣的忠誠，他領導的政府不會尋求與中國建立外交關係。我國和瓜地馬拉 1933年建立領事關係，1960年升格為大使級關係，雙方邦誼至今 88年。65歲的賈麥岱曾參選總統 4次，2019年勝選，隔年1月14日就任，當時我國外交部回應，賈麥岱是台灣長期友人。另外，美國副總統賀錦麗（ Kamala Harris）將於6日抵達z瓜地馬拉，8日再飛往墨西哥訪問，這是她就任副總統後的首次出訪，且她 3月被美國總統拜登交付處理中美洲「北三角」非法移民的重大任務。賈麥岱告訴路透，要解決移民問題根源，美國應把重心放在打擊毒品走私。「當我們看移民新地圖，大部分都是來自偏 鄉地區」，賈麥岱說，「瓜地馬拉與美國結盟，可藉由制度性方案來幫忙解決結構性原因」。他表示，會要求賀錦麗透過世界糧食計畫署（ WFP）或政府，建立與偏鄉地區之間的管道。美國大部分的援助是交給非政府組織（ NGO），並計畫投入40億美元援助「北三角」國家。

瓜地馬拉總統：不會捨棄建交88年的台灣

模型生成摘要

實習心得&收穫

- 非本科系學生能夠實際到研發部門做工程師實習
- 實作深度學習模型，並學會相關理論和程式工具操作來訓練模型
- 將繼續與部門討論使用者情境，未來朝向多文件摘要方向邁進

Thank You!