**探索20-4講座：鑒往知來：子非人，安知人之語？**

**—談自然語言處理 / 陳信希教授影片心得感想**

影片中一開始有提及翻譯的部分，到世界各地遊玩時，可以透過拍照進行文字分析，進而達到翻譯效果，但是每個國家的語言斷詞及片語等等，有不同的規則，所以翻譯有時候會有些語意上的差異，但大致上有一定程度都是可以理解的，裡面也提到問答系統，就像上課系上學長所教的問答系統，可以自己製作，但是背後需要很多手動輸入相關詞彙進行分類，但是如果匯入語言資料庫的話，問答系統的結果會比較完整，延伸出去還有我們日常所見語音問答如手機siri等，這些都基於針對不同資料型態「文本分析」，可以透過生活中新聞、論文等等進行文字探勘，其中主題很廣泛，能收集的資料系統都會盡力去收集，如市場產品資訊、熱門議題分析等等，如此一來可以看到大家對某產品的期待與反饋。

裡面有提及評論，文字分析之後可以針對一段文字裡面進行好與壞的意見分析，或是針對市場分析:標的追蹤，可以看到某個產品或服務的市場營收高低、預判市場危機等等，如此一來可以針對未來市場有更精確的判斷，還有提到有趣的機器人與人的辯論賽。岐義解析講解得很清楚，有分成詞彙層次、語法層次、語意層次，詞彙層次裡詞彙同時有動詞、名詞等不一樣的意思，那電腦會透過自然語言處理進行分析。

自然語言有很多任務要處理，包括中文斷詞、意見探勘、情感分析、立場偵測等等，詞彙跟類別都是以「符號」來呈現，匹配是基本符號計算，計算奇關連程度，後續還有提及上課所教的機器學習以及進階的深度學習。

讓我印象深刻的一句話是「意思來自使用。」意了解詞彙意斯，關鍵是那些伴隨出現的詞彙，就像前面提到的，一個詞彙單獨來看有很多意思，根據其後文字可以判斷是哪種意思，就如我們讀取英文文章遇到不會的單字，可以從上下文判斷他的意思大概是什麼的想法相同。詞彙用向量計算其關聯度，高維度計算，夾角越小，讓我覺得比較抽象，還要用矩陣分解來達到降維度在進行運算。關於句子相似性計算，如果有兩個句子相似度特別高，那會分別進入嵌入層再到編碼層，此時句子已經是用向量表示，然後計算出彼此的關聯程度，最後進入比較層進行比對，最後會出現他們兩個的關聯分數。

語言訓練裡面有個很特別的現象，再系統建置過程中沒有告訴機器日文和韓文有關係，但透過機器學習，和其他語言有關聯，最後他還會知道日文和漢文是有關聯性的。在意見探勘中有提到語言陳述可能含主觀、誇大不實等等的資訊，還有很大的空間可以走下去。語言理解不僅和語言相關，可能還需要世界知識等等的外部知識，希望未來可以建立人類常識、因果關係、動作過程等多模態知識庫，將其整合在一起，呈現更多知識表達和推理。