Session 2.6 智能文本分析(AI text analytics)

Time & Location: 10:20-12:00, Dec. 1, L008

Chair: Chia-Hui Chang (張嘉惠)

(1) 適用於語音文件之有效的類神經摘要架構

劉慈恩(臺灣師範大學), 劉士弘(台達知識管理部), 陳柏琳(臺灣師範大學)

由於巨量資訊的傳播，有效率地瀏覽資訊是一個重要的課題，而自動摘要 (Automatic Summarization) 則被視為關鍵技術。單一文件節錄式摘要任務是根據主要的文字文件 (Text Document) 或語音文件 (Spoken Document) 內容，透過選擇與排序其重要語句，產生一個簡短的版本。本論文發展並改進以往的節錄式摘要模型，透過強化學習 (Reinforcement Learning) 將 ROUGE 評估指標作為目標優化架構，並探討如何有效地結合聲學特徵 (Acoustic Features) 產生摘要。我們採用中文廣播新聞語料 (MATBN) 進行實驗；實驗結果顯示，相較於目前最佳的摘要技術，我們所發展的摘要方法能提供明顯的效能改善。

(2) 基於上下文特徵萃取法之文章隱藏風格分析

劉庭維(清華大學), 李思叡(清華大學), 游雅雯(清華大學), 李婉瑄(清華大學), 黃彥皓(清華大學), 陳宜欣(清華大學)

近年來，由於網際網路資訊量的爆增以及影響範圍擴大，如何快速過濾不可信資料已經是備受關注的議題，若能以機器學習模型自動化識別資訊之可信度，將能夠大幅提升普羅大眾接收正確資訊之效率。基於此議題，本研究嘗試利用網路論壇文章，以圖論樣式特徵萃取方法進行樣式嵌入，加強模型對於語句關係之理解與泛化能力，並提出以弱標籤形式分類訓練搭配遷移學習架構，完成對論壇文章可信度之量化分析方法。經驗證，本方法可以成功在極少量具標籤的訓練資料中，學習出辨別廣告心得文章之能力，且能正確辨識達 70%。

(3) 韻律特徵與聲學特徵於錯誤發音檢測與診斷之研究

林奕儒(臺灣師範大學), 許曜麒(台達知識管理部), 楊明翰, (台達知識管理部) 陳柏琳(臺灣師範大學)

本論文探討韻律特徵應用多任務深層網路模型於錯誤發音檢測及診斷(mispronunciation detection and diagnosis, MDD)之研究。電腦輔助發音訓練(computer assisted pronunciation training, CAPT)之目的在於透過電腦自動地指正外語學習者的發音問題；其在程序上大致可分為錯誤發音檢測(mispronunciation detection)與錯誤發音診斷(mispronunciation diagnosis)等兩個階段。本論文主要探討 1.)韻律特徵與語音特徵結合後對於模型的幫助。 2.)希望利用多任務深層網路模型解決資料正例反例不平衡之問題。 3.)結合基於相似度的評分(Likelihood-based Scoring,GOP)以及基於分類器評分(classification-based Scoring)的方法達到更好的檢測結果以及診斷結果。

(4) 利用記憶增強條件隨機場域之深度學習及自動化詞彙特徵於中文命名實體辨識之研究

簡國峻(中央大學), 張嘉惠(中央大學)

近年來深度學習模型應用於命名實體辨識，已可達到使用客制化特徵的成果，然而對於社群媒體資料集中卻未能達到傳統條件隨機場域之基準值。如何有效地擷取文字中所隱含的資訊，使模型有較好的濾除雜訊之能力，是在應用上非常重要的一環。在本研究中，我們建構一個深度學習模型，使用門控卷積網路及雙向 GRU 網路來增強記憶條件隨機場域，以利模型抓取長距離的文章資訊。我們透過 Web 新聞人名實體辨識[1]資料集實驗發現，門控卷積網路及雙向 GRU 網路的設計，可以大幅改善基礎記憶模型 MECRF[8]的效能，將 F1 從 85.72% 提升至 90.75%。此外結合字元向量及詞向量，可再提升辨識效能至 91.10%。最後藉由特徵探勘擷取詞彙特徵，並使用模型自動訓練的參數，自動調整詞向量及詞彙特徵，獲得最佳 91.76%的 F1 效能。與傳統 CRF++的模型 86.30%相比大幅提升了 5.46%。

(5) 探究端對端混合模型架構於華語會議語音辨識

張修瑞(臺灣師範大學), 趙偉成(臺灣師範大學), 羅天宏(臺灣師範大學), 陳柏琳(臺灣師範大學)

近年來端對端(End-to-End)語音辨識的出現，簡化了不少語音辨識所需要的複雜步驟；其中最主要的模型架構分別為 Connectionist Temporal Classification (CTC)與 Attention 模型。本論嘗試結合上述兩種模型架構(即 CTC-Attention 混合模型)於華語會議語音辨識之使用，以期能進一步提升語音辨識的效能。為此，我們分析模型結合時混合權重調整的影響，並進一步探究 CTC-Attention 混和模型對於短句的辨識效果。實驗結果顯示 CTC-Attention 混合模型於中文會議語料辨識具有因應句子長度變化的彈性。

(6) 基於 Seq2Seq 深度學習模型之客服聊天語料庫開發

何應承(崑山科技大學), 鄭朝榮(崑山科技大學)

專業的客服雖然能做到細緻、貼心的服務，但人類不是鐵打的機器總有其極限，人工智慧客服可以將員工從重複性極高的日常工作中解放，讓他們去處理更需要複雜判斷能力的客服事務。雖然如此，目前機器人的對話情境設計，因為成本因素大多沒有屬於自己的語料資料庫，而直接套用軟體事先安排好的問答對話，當使用者詢問機器人相關的語音關鍵字，機器人才會做出回應，往往對消費者來說較無吸引力及實用性。若能結合用戶常用的通訊軟體，與顧客 24 小時隨時連結，店家將可以透過聊天機器人與顧客聊天探知喜好，並傳送更具個人化特色的商品推薦、優惠給顧客。本研究以電影語料庫詞句進行自然語言處理(Natural Language Processing)為例，使用網路爬蟲程式爬取 PTTbbs 電影討論版的網友討論串做為電影語料庫，先以 Jieba 斷詞演算法處理後，再透過 Seq2Seq 模型訓練，訓練好的模型即為聊天機器人的問答模組。進一步地，我們使用樹莓派結合 Line Messaging API，開發 LineBot 聊天機器人做為接收與回應使用者語音訊息的媒介。