

經濟部 109 年度
《人工智慧產業關鍵技術拔尖計畫 (3/4)》
合作研究計畫

《中文自監督式學習語言表徵模型技術研究》
建議書徵求文件

財團法人資訊工業策進會

中華民國 109 年 01 月 30 日

109年度合作研究計畫建議書徵求文件

一、 簡介

監督式學習目前廣泛被使用於虛擬助理的問題回答，然而監督式學習必須是由人進行資料標記作為機器學習正確答案的依據。而資料標記往往需要花費大量人力與成本，因而造成虛擬助理優化與導入場域的重大瓶頸。自監督式學習可以預先訓練未標記文本以建立語言表徵模型，只需要微調即可針對不同虛擬助理任務建立訓練模型，將可以降低虛擬助理系統導入實際場域的人力與成本。

二、 計畫目標

利用自監督式學習，建立中文語言表徵預先訓練模型之技術研究。

三、 計畫範圍

- 中文自監督式學習語言表徵模型
 - 建構：中文自監督式學習語言表徵模型
 - 驗證：研發自監督式學習語言表徵訓練功能，可以預先訓練未標記文本以建立語言表徵模型，至少支援1個領域，如金融保險。

四、 預期成果

- 完成1個自監督式學習語言表徵訓練程式雛形
- 完成1個自監督式學習語言表徵模型（以既有公開語料）
- 完成1個中文自監督式學習語言表徵模型（使用收集未標註的中文語料）

※前述成果如有專利構想或專利申請產出時，需注意專利申請之新穎性(novelty)。因凡經公開發表之研發成果，如擬申請專利，須於公開發表後6個月內完成，前述成果如是以論文方式公開發表，將無法取得大陸與歐盟等國之專利。

五、 執行方式

- 合作計畫執行單位需定期舉行計畫控管會議，確保計畫的進度及執行方向一致性。
- 合作計畫執行單位需於6、12月繳交研究報告共二份。本單位要求合作計畫執行單位完成中文自監督式學習語言表徵訓練程式雛形。

六、 計畫期程及預估計畫總經費

計畫執行區間：109年01月01日至109年12月15日

總經費：800,000元

七、 驗收標準

- 2020/06 交付期中報告1份，內容應包括：研究動機、架構規劃、演算法設計與分析、使用技術、文獻搜尋
- 2020/12 交付期末報告1份，內容應包括：研究動機、架構規劃、使用技術、實驗設計、實驗結果、結論、文獻搜尋
- 2020/12 交付完成1個自監督式學習語言表徵訓練程式雛形、1個自監督式學習語言表徵模型（以既有公開語料）、1個中文自監督式學習語言表徵模型（使用收集未標註的中文語料）

八、 技術能力需求(請詳述所需要之技術能力或專長)

- 具中文自然語言處理研究經驗與成果之技術研究人員。

附件1：契約書格式

1-1：計畫書格式

1-2：經費動支報表

1-3：成果報告撰寫須知

1-4：報告格式

1-5：論文格式

1-6：保密聲明書

1-7：委託匯款同意書