

二・項目簡介

(項目所屬科學技術領域、主要技術內容、授權專利情況、技術經濟指標及應用推廣情況)

機器翻譯是人工智能與自然語言處理的核心領域之一，可以實現世界上不同國家不同語言間的無障礙交流，協助打破語言壁壘。相較於人工翻譯，機器翻譯具有低成本、易把控和速度快的優勢。機器翻譯的研究範式歷經基於規則和統計的方法，到當前的基於神經網絡技術的神經機器翻譯，譯文質量不斷提升，實用化程度顯著增高。但由於神經網絡的低解釋性因素，當前神經機器翻譯技術停滯不前，模型已達性能天花板，語言學知識的引入可以提升模型對句子的建模能力並幫助理解機器翻譯的推理過程，從而達到改善譯文質量的目的。本項目從澳門實際的中文與葡語的互譯需求出發，創新性地提出基於語言學驅動的神經機器翻譯並應用於中葡機器翻譯系統。

當前中葡機器翻譯研究的核心瓶頸是雙語數據缺乏的問題。為克服此問題，一方面，項目成員在二十年來不斷地收集和標註大規模中葡平行語料庫。另一方面，引入了語言學中詞法與句法的結構化知識，從有限的數據中訓練出高性能的機器翻譯系統。在詞法上，提出了複雜詞法形態語言的機器翻譯模型，形成了以中樞詞嵌入和層次化解碼器的代表性方法，以克服中葡機器翻譯中葡文形態變化複雜的問題。在句法上，提出了引入句法語義特徵的機器翻譯模型，基於句法成分和句法依存結構對句子建模，並強化句子中上下文特徵，從而克服了中葡機器翻譯中葡語法結構迥異的問題。所研技術不僅能用於中葡機器翻譯，也可以用於其他類似的低資源語言機器翻譯，具有重要的學術價值。

項目獲得國家自然科學基金、澳門科學技術發展基金、以及澳門大學的科研基金等資助，共計資助金額 400 餘萬澳門元。所研技術發明成果包含 32 篇學術論文（23 篇為 CCF-A/B 類論文）、6 項國內外機器翻譯競賽（CWMT 和 WMT 競賽）冠軍、4 項國家知識產權局發明專利和 1 個澳門商標。基於所研成果，本項目開發出「UM2T 中葡在線神經機器翻譯系統」和「UM-CAT 在線中葡輔助翻譯平臺」，在中葡的互譯效果顯著超越了國內外主流的在線翻譯系統，被大量用戶使用並收到好評。除了提供實時、高質量的全文翻譯外，所研產品還配備澳門術語本土化、翻譯記憶、輔助提示、多領域翻譯等功能，可服務於科技、政治、文化、教育等行業和領域，協助澳門打造「溝通無障礙」的智慧城市和「中國與葡語系國家商貿合作服務平臺」的樞紐城市，具有重要的應用價值。

項目所研機器翻譯系統及平臺已有超過 50 家機構註冊使用，其中包括 21 個澳門政府單位（如行政長官辦公室和初級法院等），6 家高校（如香港大學和新里斯本大學等），7 家律師事務所（如 C&C Lawyers & Notaries 等），22 家中大型企業（如美高梅中國等）。另為抗擊疫情，項目所研系統於 2020 年免費開放於澳門中小企業以協助抗疫，並被多家企業註冊使用，具有顯著的社會效益。項目成果也被轉化於初創公司新譯信息科技（深圳）有限公司（創始人為項目組碩士畢業生），是一家機器翻譯提供商，基於項目所研機器翻譯技術開發的語言平臺銷售額已達四千多萬元人民幣，具有良好的經濟效益發展前景。

(字數不超過 1200 字)