MIT創大數據，母語非英語者的好消息

駐波士頓辦事處教育組

麻省理工學院（MIT）新聞辦公室表示，該校已完成一套針對母語非英語者的英文大數據資料庫，不但可分析出外國人表達英語文法各異之處，成為語言學者的研究資源；也希望能改進電腦軟體，使其更明白母語非英語者所表達或書寫的英文。

領導這項計畫的電腦暨電機系研究生柏沙克（Yevgeni Berzak）說，英語是全球約10億網路訪客最普遍的語言，而上網使用者的母語多半不是英語。目前手機和電腦程式，都是根據大量標準英文的語言數據模式做分析處理。他們建立新大數據庫的目的，是訓練電腦系統理解非規範式英語，更好地處理母語非英語者的介繫詞、動詞時態使用等，甚至還能修改文法錯誤。

研究員的數據集包含以英語為第二語言(ESL)作文測試的5,124個句子。這些句子是按比例取自全球約四成人口、十個非英語系國家的學生。每一句至少有一個文法錯誤，由英國劍橋大學的學生負責搜集並標註出錯誤。接著，柏沙克從該校電腦、電機、語言及機械系裡聚集一群大學部學生和研究生，經過8週辨識和分析英文語法錯誤的培訓後，分為三級來處理資料。

第一級是語法基本項目，如：錯用名詞、動詞、介詞等。第二級偏重細節，如：單複數、時態、比較級、最高級形容詞等。最後，研究者使用名為「普遍性依存（UD,Universal Dependency）形式」的新註釋法則，繪出詞語間的語法關係圖。句法關係包括動詞與名詞間、動詞與助動詞間、及名詞與形容詞間的關係，共完成20頁的註釋方案指南。他們表示，對這些語句的處理，有助於訓練機器學習系統正確辨識母語非英語者表達的真實語意。

使用圖表標示不合文法的句子，為該研究法提供了很好的證據。每一個句子，經過第一位評估員負責註記後，第二個評估員校正，若有不同意見則加註於上，第三位負責裁定分歧處。碰到有爭議時，非文法錯誤的句子和文法錯誤的句子，處理原則是相似的。MIT團隊將於8月7日至12日，在柏林舉行的電腦語言學協會（Association for Computational Linguistics）2016年年會中，提出這項研究報告。研究人員的數據，目前可通過組織系統取得。Berzak 為數據資料架構了一個網上交接點，便於研究人員尋找特定類型的錯誤。

瑞典Uppsala大學Joakim Nivre教授說，使用UD把ESL作系統性比較，

為研究數據開闢很多可能性，不僅可用於英語母語，也可用於其他的語言。希望其他的 ESL 研究員以他們為榜樣，可以從數個介面作進一步的比較，像是從 ESL 到 ESL，ESL 到英語母語等。註釋正確性和糾正判決的決定，是非常有價值的資料。例如，系統可以學習將 ESL 翻譯為標準英語，達成機器翻譯的功能。目前的語料庫，基本上是提供並行數據來訓練系統，若是句型註釋法可行，將使技術的開發更多樣化。

這項工作隸屬於麻省理工學院人腦及機器思維研究中心，由美國國家科學基金會部份資助。

譯稿人:趙維新
資料來源:2016 年 7 月 29 日 麻省理工學院校報

國家教育研究院 National Academy for Educational Research