SAMSUNG

學習曲線 Part 3：使 AI 資料至臻完善

本報導探討越南三星研究院的研發歷程，一覽研發人員和創新科技如何使行動 AI 造福更多人的生活

三星正如火如荼開拓旗艦行動 AI 體驗。三星新聞中心專訪世界各地的三星研究院，深入了解 Galaxy AI 如何協助用戶發揮極致潛能。Galaxy AI 目前支援 16 種語言，憑藉通話即時翻譯、語音翻譯、筆記智慧助理和瀏覽助理等終端內建翻譯功能，使更多用戶即使在離線狀態也能拓展語言能力。團隊稍早造訪約旦，一窺擁有多種方言的阿拉伯語系，在開發 AI 模型時所面臨的複雜難題。本篇則前往越南，探索 AI 模型的資料建置過程。

在越南文中，「鬼魂」、「墳墓」和「母親」的發音差異微乎其微。全球有 9,700 萬人使用越南文，而上述三個詞分別對應為音譯「ma」、「mả」和「má」，區別只在於聲調，這顯示了 AI 在學習語言時的難度之高，因為 AI 不僅難以直接辨別語境和對話情緒，也難以理解語句背後的意圖。

三星越南研發中心（SRV）運用精細的資料，協助 AI 模型正確判讀語言中最細微的差異。

數據的品質將直接影響自動語音辨識（ASR）、神經機器翻譯（NMT）以及文字轉語音（TTS）的精準度。此三大技術應用於 Galaxy AI 通話即時翻譯、語音翻譯、訊息即時翻譯智慧助理和瀏覽助理等功能，致力破除語言藩籬。

災難級挑戰

SRV 的 NMT 負責人 Ngô Hồng Thái 表示：「越南文是一種複雜多樣的語言，表達方式相當豐富，許多細微之處難以捕捉。」在 Galaxy AI 支援的 16 種語言中，越南文的開發尤其困難。

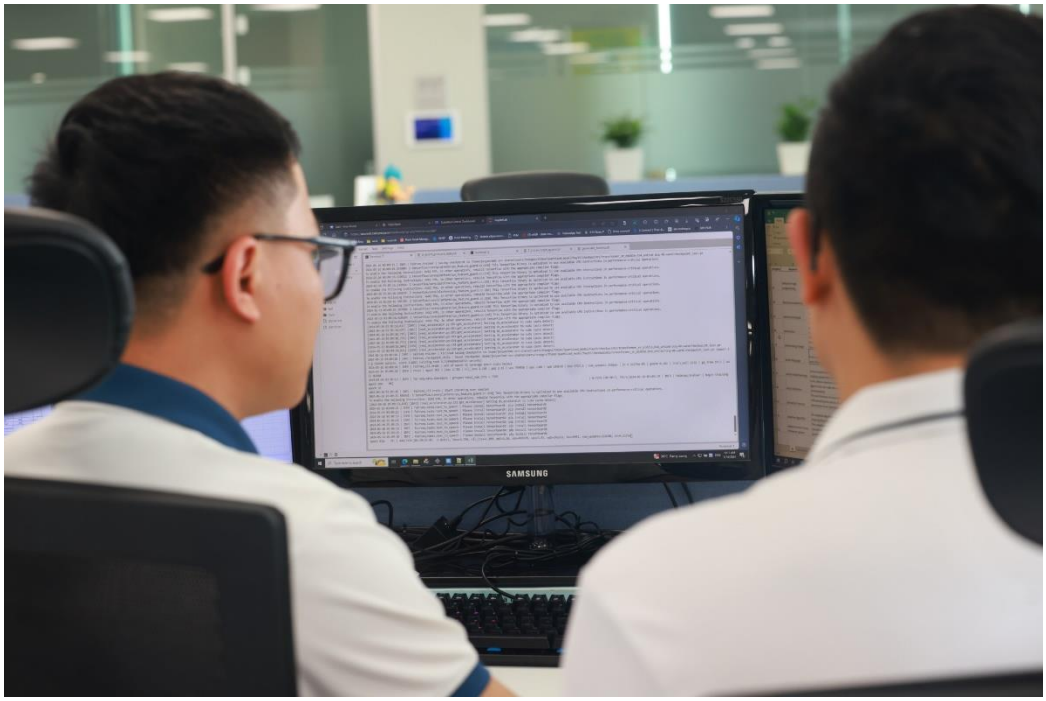
在繼續解釋開發過程所面臨的難題之前，Thái 補充：「對我來說，為越南文建立 AI 模型，比颱風還要更令人畏懼！」

越南文是一門聲調語言，具有六種不同音調。如同上述「ma」的例子，發音上的細微差異會大幅改變語意，因此在研發上必須非常注重細節。

SRV 的 ASR 負責人 Bui Ngoc Tung 指出：「發音相似的詞語在拆解後，單一字詞會產生數個短區段，或稱為『訊框集』。AI 模型會區分 20 毫秒左右的短音訊訊框，辨識字彙與特定連續訊框之間的關聯，因此團隊必須在 AI 學習的早期階段投入大量心力。」

此外，同音異義詞和同形異義詞在越南文中相當普遍，人們通常可依靠上下文和對話中的非語言元素，區別發音相同或拼字相同但意思不同的字。然而，AI 模型必須經過訓練，才能正確分辨語調和相似詞彙。

Thái 解釋：「這並不是一項簡單的任務。資料必須重質且重量，才能讓 AI 有能力判斷越南文中的細微差異。」



嚴謹的前置作業

資料精細化流程共分為三個步驟：首先，用於訓練 AI 模型的音訊和文字必須經過審閱和修正。接著，針對此資料集進行隨機抽查，確認整體品質。最後，資料集需進行標準化和淨化，才能用於訓練。

Nguyen Manh Duy 為 SRV 的 TTS 負責人，負責監督資料庫建立流程。他表示：「團隊仔細地執行了一系列的測試，確保資料庫的準確度。過程中遭遇了許多意外的問題，包括腳本中有拼字錯誤、錄製音訊時收錄了背景噪音或發音錯誤。我們花了大量的時間去完善並優化訓練資料。」

在數據精煉過程中，將 AI 數據從優良提升到卓越的關鍵要角，正是軟體品質工程（SQE）團隊。該團隊在測試和精進 AI 語言數據品質方面發揮重要作用，並與 AI 語言開發專案團隊密切合作，實現此目標。

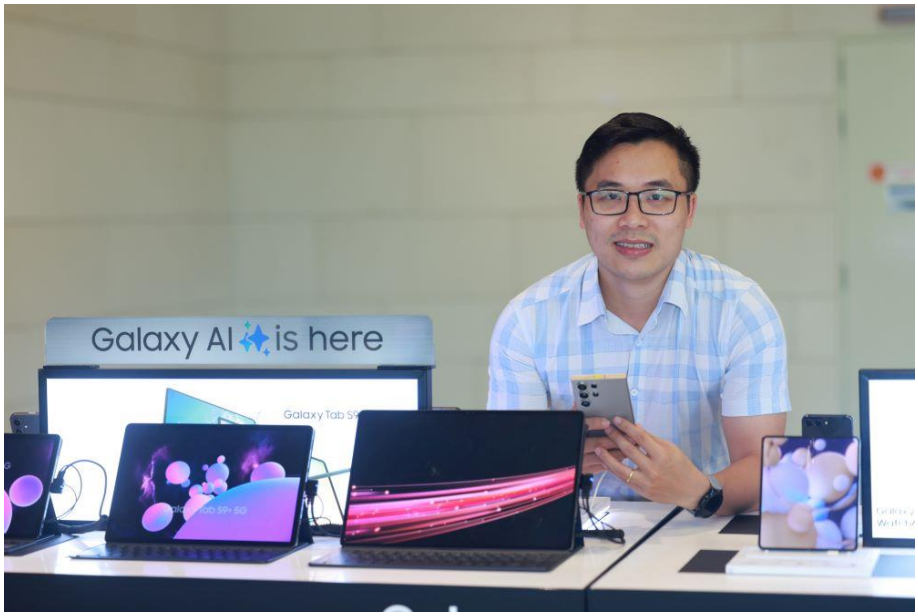
除了語言上的獨特挑戰，與其他廣泛使用的語言相比，越南文的開放性語料相當稀少。Duy 補充道：「這也是資料精細化階段如此重要的原因。由於來源有限，每一份資料均必須百分之百可靠，沒有出錯的餘地。」

此外，越南文的 AI 模型必須同時考量語調和地區差異。為了提升 AI 模型的精準度，團隊自越南北中南各地的口音收集大量資料，也因此產生了眾多需要精細化和驗證的資料。

精益求精

SRV 開發人員在歷經數月的努力後，成功使越南文成為 Galaxy AI 首批支援的語言。儘管取得了成功，團隊仍馬不停蹄地提升越南文的 Galaxy AI 體驗。

SRV 的 AI 語言開發專案負責人 Tran Tuan Minh 表示：「我們會採納用戶的回饋，改善 Galaxy AI 中的詞語與慣用語，持續不斷地優化 AI 模型。目前僅是向更開放的世界邁出第一步，未來還有諸多事物需要共同探索。」



於學習曲線下一篇報導中，三星新聞中心將前往中國，深入探討 AI 模型的訓練和校正方式。