



Total number of document(s): 1

文江學海
A23

恆管譯站

Wen Wei Po

昌博士
2017-02-08

準確評核機器翻譯利「自學」進化

近年機器翻譯（machine translation）發展蓬勃，市面上出現各種自動翻譯系統，如網上翻譯系統及手機翻譯程式等。特別是過去一兩年間，除了基於語言規則及統計方法的翻譯系統外，開發商陸續推出以深度學習（deep learning）及遞歸神經網絡（recurrent neural network）為基礎的神經機器翻譯系統（neural machine translation systems）。

如此一來，用家選擇更多，我們如何評鑑各家系統，加以比較，擇優而取，更形重要。為此本文概述若干評價方法，供各界參考。

「回譯」比較原文 未必可靠

最簡單的方法是「回譯」（round-trip translation），先利用機器翻譯系統把原文譯出，然後再經電腦把譯文譯回原來的語言，如此比較最初的原文及經回譯後的「原文」，若兩者相似，則系統表現較佳，否則未如理想。此法快捷簡便，用家即使不懂譯語，仍可據此估計電腦譯文是否穩妥。

不過，系統把譯文譯回原語時，譯文中的錯誤或有增減，故兩篇「原文」相似與否，未必在任何情況下均能如實反映翻譯系統之高下。

要更準確地評價翻譯系統，我們可設立評分準則，然後邀請評審員按之為電腦譯文評分。美國上世紀60年代設立語言自動處理諮詢委員會（Automatic Language Processing Advisory Committee），研究俄英自動翻譯水平，做法可供借鑑。

當局就譯文準確及可讀程度兩方面設立評分標準，分設9個至10個等級，描述各級譯文應有水平。當時還聘請了兩組評分人員，一組為英語人員，另一組則兼通英俄雙語；所有人員均須就兩方面評分，確保分數客觀準確。

人手評分需時耗費大

然而，人手評分往往需要不少時間及開支，自動的評鑑方法遂成潮流。自動評鑑即電腦自行評估系統質素，一般以參考譯文（如人工翻譯的譯文）為評分標準。電腦比較系統輸出及參考譯文，按相似度自動評分，兩者相似度高則給予較理想分數。常見計算方法有很多，包括精確率、召回率、F值、BLEU及NIST等，網上相關資源豐富，讀者可自行參考。

值得注意的是，翻譯系統各種評鑑方法，除了幫助用家挑選及善用系統外，對機器翻譯本身發展也有關鍵作用。更具體而言，新一代翻譯系統強調自行學習及提升翻譯能力，電腦怎樣準確評核翻譯水平，自動改進弱點，實在非常重要，值得深入探討。■ [恆生管理學院](#)商務翻譯課程總監 蕭世昌博士