

● EPO 如何提升其專利機器翻譯引擎之翻譯品質

歐洲專利局 (EPO) 於今 (2013) 年 6 月 26 日在其免費的「專利機器翻譯系統 (Patent Translate)」新增 6 種歐洲語言的機器翻譯服務，包括保加利亞語、捷克語、冰島語、羅馬尼亞語、斯洛伐克語，以及斯洛維尼亞語。該局另於今年 9 月 26 日再宣布新增俄文-英文機器翻譯，該系統目前共提供 22 種語言的即時互譯。然而，EPO 是如何達成此一成果，特別是某些語言，可用於訓練機器翻譯引擎的專利文獻數量有限，EPO 是如何訓練其機器翻譯的品質？

對此一疑問，EPO 在今年出版的第 3 期「專利資訊季刊 (Patent Information News)」中說明其如何獲得所需的翻譯品質水準並確保其持續提升。

一、為專利量身打造的翻譯引擎

專利語彙及書寫文體與大部分的文字不同，使得標準的機器翻譯引擎對專利文件的翻譯倍感困擾，因此需要改用不同的解決方法。

EPO 的局長 Battistelli 於 2010 年啟動該系統計畫時即認為，以往上百萬件經由人工翻譯的專利文件可以作為新系統的基礎。

此一概念是基於，假定在不同語言中的兩個專利家族成員是對等的，只需將此家族匯入機器翻譯引擎學習並等待其改善翻譯

品質。EPO 發現可與 Google 合作此項工作，並同意提供大量對等專利文件的雙語對列語料 (corpora) 給 Google，以訓練統計式的機器翻譯服務系統。

二、以主要的語言作為驗證

2012 年 2 月「專利機器翻譯系統」上線時，所提供的英文與法、德、義、葡及西文 5 種主要歐洲語言之間的互譯，即有大量的資料可供語料使用。

初步的測試結果令人印象深刻，顯現以往的系統已明顯的改善。評量翻譯品質的方法之一是採用 5 級的標準，其中品質最好的是接近人工翻譯的第 5 級。到了 2012 年 9 月，大部分的對列語言獲得的平均分數介於第 3 級和第 4 級之間，顯示「專利機器翻譯系統」可提供使用者對專利文件最基本的了解，並協助其決定是否需進一步訂購人工翻譯。

三、增加更多的語言，部分語言的翻譯品質令人驚訝

隨著時間的推移，歐洲專利局加入更多的專利語料，「專利翻譯系統」也出現更多的語言。然而，具有大量對等專利案件的語言有限，有時建立語料庫會較為困難。這些語料資源較少的語言納入系統後，經過品質分析，讓 EPO 的專家們驚喜，其評比僅略低於主要語言的一般水準，在某些情況下甚至更好。

分析師試圖找出結果如此好的原因，印證了「優進優出 (Quality in, quality out)」的俗語。簡而言之，如果語料庫中的人工翻譯語料非常的好，並以其所產生的最終語料訓練統計式翻譯引擎如何翻譯專利文件，就會產生正面的影響。

某些語言，包括芬蘭文、匈牙利文、土耳其文和日文，與英文的文法差異相當大，為一特殊的挑戰。對於這些語言中，語料庫的大小不是問題，且有許多的文件可供取得。況且，團隊中的語言學家能夠確認，這些語文的語法和結構與大多數的西歐語言非常不同。一般來說，「專利機器翻譯系統」已經能將詞彙翻譯的很好，並繼續處理以確保翻譯的文法能夠讓人了解專利文件。

四、一天翻譯 1 萬件，並持續增加

大部分的人並不看好「專利機器翻譯系統」會成功，該系統現在則每天處理 1 萬件翻譯文件。最常見的要求是將德文翻譯成英文，中譯英的需求亦快速地增加。翻譯品質將持續改善，並將會加入更多的語言。EPO 的目標是，到 2014 年年底，將所有 EPO 成員國的官方語言，再加上日文、簡體中文、俄文及韓文，不僅可與英文互譯，亦可與法文和德文互譯。

歐洲專利局

<http://www.epo.org/service-support/publications/patent-information/news/2013.html>