漢英機器翻譯系統研究與開發漫談

詹衛東

北京大學中文系 100871

zwd@pku.edu.cn

從20世紀40年代電子計算機問世開始﹐利用計算機來完成翻譯活動﹐就一直是人們力圖實現的一個夢想。盡管潮起潮落﹐這個夢想的實現時而被一些人吹噓為指日可待﹐時而又被一些人貼上“絕無可能”的封條打入冷宮﹐但曆史發展到今天﹐人們已經能夠比較客觀公正地來看待機器翻譯的研究及其存在與發展的價值。而且﹐隨著全球信息化步伐的加快﹐從專業人士到普通用戶﹐都對機器翻譯投以越來越多的關注﹐也有越來越多的研究人員涉足這一領域﹐同時翻譯的對象也拓展到越來越多的語言對﹐這些﹐都應該算是機器翻譯開始走向成熟的最好明証吧。

作為全球最早開始機器翻譯研究的國家之一﹐中國的機器翻譯研究跟世界其他國家的機器翻譯研究的一般發展曆程相似﹐也是從由外國語到本國語的機譯系統開始﹐繼而再發展出由本國語到外國語的機譯系統。20世紀80年代末、90年代初以來﹐隨著一批外漢機譯系統從實驗室走向巿場﹐漢外機器翻譯﹐尤其是漢英機器翻譯系統的研究工作﹐也在中國科學院、北京大學、哈爾濱工業大學等高校科研機構﹐以及中軟、桑夏等軟件公司逐漸開展起來﹐其中有的漢英機譯系統在90年代末也開始有初期版本推向巿場（比如中軟公司的譯星TranStar'99套裝機器翻譯軟件中就包含了一個漢英翻譯模塊）。

應該說﹐跟早期的外漢機器翻譯系統的開發相比﹐漢英機器翻譯系統的研發是有不少有利條件的。首先是機器翻譯的理論和技術都經過了近二三十年的發展進步﹐人們對機譯的認識已相對清晰﹐而且開發機器翻譯系統的實際經驗也有了很多積累；其次計算機軟硬件技術一日千里的發展速度使得機譯系統的開發平台也非昔日可比﹐研發人員有更多的選擇和開發空間可供自由馳騁。再次﹐語言學研究和形式語法理論在20世紀80年代以來的發展為自然語言的計算機處理提供了更多的語言學支撐。在這樣的大背景下﹐現有的漢英機譯系統一般都釆用的是相對比較成熟的機譯技術﹐即基于語言學規則的轉換式技術來實現自動翻譯過程。盡管不同的漢英機譯系統在具體實現時﹐算法上和程序以及語言知識表示的細節方面會有差別﹐但自動翻譯系統的基本流程大同小異。比如一個典型的漢英機器翻譯系統通常要經過這樣一系列步驟﹐首先要經過漢語的自動分詞（segmentation）與詞性識別（part-of-speech tagging）﹐然后對以詞串形式排列的句子進行自動句法分析（parsing）﹐得到漢語句子的結構樹（structure tree）﹐再將漢語的句法樹轉換成相應的英語句法樹﹐隨后根據已有的時態、性、數等句法信息對英語譯詞進行詞形變化﹐最終產生英語譯文句子結果。值得指出的是﹐上述過程中所需要的漢語分詞和詞性識別技術﹐漢語句法分析技術﹐以及英語生成技術等﹐中國學者自20世紀80年代中期以來都已做了不少規模較大的深度研究﹐特別是在一些基礎性的工作上進行了不懈努力（比如漢語語法信息詞典的建設﹐漢語短語功能分類的研究等等）﹐毫無疑問﹐這些工作都為漢英機器翻譯系統的研發打下了一個比較堅實的基礎。

不過我們也不能不正視這樣一個現實﹐那就是跟英漢機器翻譯相比﹐漢英機器翻譯系統的質量仍是略遜一籌﹐離真正進入實用還有一定距離。實際上這背后的原因也很自然平常。從上述我們對漢英機譯過程的諸多環節的簡要描述就不難了解﹐整個機譯過程中間涉及到的每一個環節得以順利通過的保証是機譯系統要有足夠的、准確的語言知識的支持。這一點對英漢機器翻譯也好﹐對漢英機器翻譯也好﹐都是一樣的。而特別針對漢英機器翻譯系統來講﹐英語譯文生成質量的好壞在很大程度上取決于漢語原文分析的深度如何。要想最終產生的英語譯文句子結果能夠准確傳達漢語原句的意思并且可讀性好﹐首先就要求對漢語句子的分析達到一個較高的水平（比如除了漢語句子的句法結構正確分析外﹐還要能夠分析出句中謂語動詞跟前后名詞之間的語義關系﹐以及漢語句子所負載的時態信息等）﹐這樣才能提供給英語譯文句子生成時所必需的各種信息。但眾所周知﹐相比英語等印歐語的語法研究而言﹐漢語的語法研究還有很大的待改進提高的余地。在中國几代學者的努力下﹐目前漢語詞法分析、以及短語結構的研究可以說已經有了相當的積累﹐達到了一個較高的水平﹐但漢語句子的分析﹐包括句子結構模式﹐特別是長句的結構模式﹐以及句型與其表意模式之間的關聯方面的研究﹐都還處在起步階段。這種情況使得目前的漢英機器翻譯系統在處理漢語中的簡單句的英譯時尚顯應付有度﹐但在碰到真實語料中大量出現的長流水句（結構松散﹐意合連結而成的句子）﹐以及帶有濃厚篇章色彩的句群的英譯問題時﹐不可避免地落了左支右絀的尷尬窘境。

盡管漢英機器翻譯系統研發的前進道路困難重重﹐但機譯研究經過半個多世紀的發展﹐人們已經能夠很理智地面對困難﹐積極思索解決之道。實際上﹐不獨漢英機器翻譯的研究如此﹐任何科學探索也是同樣的道理﹐在通往真理的道路上﹐荊棘遍地應該是家常便飯﹐一馬平川才是奇怪的事情呢。從20世紀90年代末以來﹐從事漢英機器翻譯系統研發工作的研究人員從各個方面尋求突破翻譯障礙的辦法﹐并且已有許多策略已經在付諸實施的積極實驗過程中。比如通過基于翻譯記憶體（Translation Memory）的、或者基于模板（Pattern-based）的翻譯引擎的加入來提高漢英機器翻譯系統的總體效果。此外也有越來越多的漢英機器翻譯系統的開發者更加關注漢英機器翻譯的實際應用﹐在這種認識背景下﹐就產生了面向領域（比如新聞消息類漢語文章的自動英譯）﹐通過語料庫統計方法來組織機器翻譯所用的漢英雙語詞典以及相應的規則知識庫﹐來提高譯文質量的機譯策略。不求一個系統能夠八面玲瓏﹐但求切中肯綮﹐找准一個應用點﹐能切實幫助人們獲得信息﹐或者幫助人類譯員減輕翻譯中的繁重工作。相信隨著信息科技的不斷發展以及由此推動的人類社會生活的廣泛進步﹐漢英機器翻譯的研究亦會不斷走向深入和多樣化﹐從而能夠在不同層面不同應用領域滿足人們的不同需求。