從專利文本建立大型平行語料庫：由雙語到多語

路斌[*] **鄒嘉彥 周嘉寶**
香港城市大學

提要
平行語料庫是很多自然語言處理(NLP)應用的關鍵性資源，比如機器翻譯(MT)或跨語言資訊檢索。本文探討一個新的、同時又很重要的領域，即利用可比多語專利(Comparable Multilingual Patents)建設大規模平行語料庫的可行性。其中，本文介紹我們已經建設的三個雙語平行語料庫以及一個三語平行語料庫，[1] 並涉及兩個問題：(1) 如何構建涉及多種語言的大規模可比專利語料庫； (2) 如何從這些可比語料中挖掘高品質的平行句對。另外，基於構建的平行專利語料，我們介紹一些初步的統計機器翻譯實驗。而且，我們進一步分析了構建涉及更多語言的大規模平行語料庫的可行性（例如中文、英文、日文、韓文、德語等），並對其規模做了初步的估計；這些基於專利的大規模平行語料庫將對多語言資訊處理起到促進作用。

關鍵詞
多語專利　**PCT**專利　**平**行語料庫　**機**器翻譯　**句**對

CULTIVATING LARGE-SCALE PARALLEL CORPORA FROM COMPARABLE PATENTS: FROM BILINGUAL TO TRILINGUAL, AND BEYOND

Bin Lu Benjamin K. Tsou Ka Po Chow
City University of Hong Kong

ABSTRACT

Parallel corpora are critical resources for many NLP applications, ranging from machine translation (MT) to cross-lingual information retrieval. In this chapter, we explore a new but important area involving patents by investigating the potential of comparable multilingual patents for building large-scale parallel corpora. Two major issues are investigated on multilingual patents: (1) How to build large-scale corpora of comparable patents involving many languages? (2) How to mine high-quality parallel sentences from these comparable patents? Three bilingual parallel corpora and one trilingual parallel corpus are presented as examples, and some preliminary SMT experiments are reported. Moreover, we investigate and show the considerable potential of getting large-scale parallel corpora from multilingual patents for a wide variety of languages, such as English, Chinese, Japanese, Korean, and German, which would to some extent reduce the parallel data acquisition bottleneck in multilingual information processing.

KEYWORDS

Multilingual patents **PCT** patents **P**arallel corpora
Machine translation **S**entence alignment