語料庫建構技術研究

期末報告

壹、語料庫的定義與發展的歷史

語料庫是指經過抽樣選出具有某一種代表性的口語，書面語，或語音資料庫。這些語料通常以電腦儲存與分析。1947 年 Shannon 以統計的噪音通道模型(Noise Chanel Model)為基礎工具，所發展出的訊息理論(information theory)奠定了語料庫計算語言學的基礎。在實際語料的收集與語料庫的建構方面，1961 年美國布朗大學 Francis 與 Kucera 兩位學者開始建構 Brown Corpus，這個語料庫收集了各類文體的美式英文共一百萬字。Brown Corpus 的重要性在於它是第一個語言學家有計畫建構的大型平衡語料庫，為語料庫語言學的研究開啟了一個新紀元。

由於 Noam Chomsky 對語料庫的批評加上當時機讀語料十分缺乏，電腦價格昂貴且運算速度緩慢等許多因素，使得語料庫語言學的研究在 1960 到 1980 這 20 年僅限於少數學者。一直到 1980 年代中期語音辨認研究人員經過多年的努力證明以語料庫為主使用隱式馬可夫模型(Hidden Markov Model)的統計演算法明顯優於利用語言規則的方法。另一方面 John Sinclair，Geoffrey Leech, Sidney Greenbaum, Jan Svartvik, Randolph Quirk.等學者（其中大部分為英國語言學家）運用電腦與大量機讀語料從事英文詞彙，文法，辭典編纂與計算語言學的的研究，獲得相當好的成果（參看 Sinclair (1987)1, Quirk, Greenbaum, Leech, Svartvik (1985)2, Garside, Leech, Sampson (1987)3）英國伯明罕大學 John Sinclair 教授與 Harper Collins 出版社於 1980 年代合作，建立大型機讀語料庫，並以此語料庫的做為編纂 Collins Cobuild 英文辭典的基礎。Collins Cobuild 的成功促使包括牛津大學與劍橋大學在

[1] Looking Up: An Account of the Cobuild Project.
[2] A Comprehensive Grammar of the English Language.
[3] The Computational Analysis of English : A Corpus-Based Approach.

內的大出版社紛紛建構大型機讀語料庫來編纂英文辭典。以大型機讀語料庫來編纂辭典的好處是可以客觀且方便地檢視詞的頻率，搭配語，及語意，語法，與語用的功能，來判斷詞的用法。以語料庫為主的語言學與計算語言學研究在沈寂了近二十年後才漸漸復甦， 1990 年代從事自然語言處理的研究人員將原先為語音辨認所發展的統計演算法運用到自然語言剖析，詞彙知識自動習得(automatic lexical knowledge acquisition)，機器翻譯等以語料庫為主計算語言學的研究上，獲得豐盛的成果。在加上大型機讀語料因為網際網路的盛行而垂手可得，以及個人電腦功能日益強大，售價卻十分低廉，這些因素使得 1990 年中期以後，以語料庫為主的計算語言學研究成為主流。

貳 、語料庫的資源

料庫種類除了有口語，書面語，與語音資料庫還依是否為平衡語料庫，有否加標記，單語或多語等方式來區分。加標記的語料庫包括加註文章結構標記(如標題，句子，段落等)，詞類標記，語意標記，或語法樹等數種。未加標記的英文語料庫以 Brown Corpus, LOB Lancaster-Oslo-Bergen) Corpus, BNC (British National Corpus), 與 Project Gutenburg 最著名，後者收集了許多英文小說。加詞類標記的英文語料庫包括 Penn Corpus, Sussane Corpus 等，中文加詞類標記的語料庫目前有中研院平衡語料庫(約 500 萬詞)。加註語意標記的語料庫目前尚缺乏。英文語法樹庫則有 Penn Treebank。至於多語語料庫最著名的是加拿大國會以英法文記錄的的 Hansard Corpus。目前國內可以線上取得的中英，中日平行語料庫則有光華雜誌。

目前最常用的兩個句法樹庫資料,分別是中央研究院中文句結構樹資料庫（Sinica Treebank） (http://www.aclclp.org.tw/use_stb_c.php)，以及美國賓州大學