# 雲報專欄：雲端運算基於自然語言的巨量資料應用

技術專家委員會
英業達資深副總經理邱全成

## 一、前言

1994 年下雪的冬天，我第一次從我們上海公司來到了北京，並開始組成一個『自然語言』應用研發團隊。當時公司開發並出貨了許多文字處理機(Word Processor)，包括個人資料管理機(Data Bank, PDA 的前身)、電子字典和多國翻譯機。上述產品是在有限的記憶體裡塞入最多筆的個人資料(電話簿/行程表/備忘錄)和最多單字解釋例句，工程師要做的就是把這些資料正確無誤地預先處理並儲存到機器內，這些資料(準確說就是語料庫)通常是靜態不變的。基於下一代文字處理機的市場需求趨勢，公司開始研發人工智慧的自然語言應用，包括自動翻譯(Machine Translation)、語音合成(Text to Speech)和語音辨識(Automatic Speech Recognition)等技術。

今日(2013 年)隨著資訊爆增的智慧型搜尋、巨量資料發掘和雲端運算等技術發展與創新，現在許多時尚應用功能都已具備更成熟的自然語言處理引擎，讓人機交互的體驗感覺更進一大步；例如 Apple siri (提供自然語言輸入，並且可以調用系統自帶的天氣預報、排程、搜索資料等應用，還能夠不斷學習新的聲音和語調，提供對話式的應答。)、Google Now (與 Google 搜索功能的結合，使用者搜索的關鍵字被記錄下來，它智慧化讀取關鍵字後，為使用者提供相關的語音服務。同時它會全面瞭解你的各種習慣和正在進行的動作，並利用它所瞭解的來為你提供相關資訊。)、Samsung S Voice 和 Microsoft WP8 語音助理功能等等。

## 二、雲端運算來自搜索(Search)應用的需求，而自然語言處理與理解成為人機溝通的主要橋樑

自然語言處理（Natural Language Processing，簡稱 NLP）是人工智慧和語言學領域的分支學科。在這此領域中探討如何處理及運用自然語言；自然語言理解 (Natural Language Understanding，簡稱 NLU) 俗稱人機對話，則是指讓電腦「懂」人類的語言。


圖 1. 自然語言處理圖

自然語言處理的範圍涉及許多資通訊的應用，如語音的自動識別與合成、機器翻譯、人機對話、資訊檢索、文章分類、自動文摘等等。在語言學方面，最重要的是語法規則形式化和數學模型的建立。在資料處理方面，以前大都集中在語料庫的建設、各種機器可讀的電子辭典的開發，在網際網路盛行後則有更大規模的語料庫的相繼湧現。其它也包括在電腦上的人工智慧和認知應用的研發，這些都是屬於『人類語言工程技術』的應用範圍。


圖 2. 自然語言處理與理解在雲端搜索和巨量資料分析等應用圖

以自然語言理解為例，研究開發電子電腦類比人的語言交際過程，使電腦能理解和運用人類社會的自然語言如中文、英語等，實現人機之間的自然語言溝通，以代替人類的部分腦力勞動，包括查詢資料、解答問題、摘錄文獻、彙編資料以及一切有關自然語言資訊的代工處理。20 世紀 1960 年代初開始研究，在語音理解和文字理解兩個方面的應用已具一定成果，包括在電腦裡儲存某些單詞的聲學模式，用它來匹配輸入的語音信號，稱為語音辨識。但這只是一個初步的基礎，還不能達到語音理解或認知的應用目標。直到 1970 年代才有所突破，建立了一些實驗系統，能夠理解連續語音的內容，但是還限於少數簡單的語句。文字理解方面，則能在一定的辭彙、句型和題目範圍內查詢資料、解答問題、閱讀文章、解釋語句等等應用。而在 1990 年代開始，自然語言處理領域發生了巨大的變化，主要是真實語料庫的收集開發和巨量資料資訊的發掘分析。

21 世紀以來，由於全球網際網路的普及與流行，自然語言的電腦處理成為了從網際

網路上獲取知識的重要手段，生活在資訊網路時代的現代人，幾乎都要與網際網路打交道，都要或多或少地使用自然語言處理的研究成果來搜索或發掘在廣闊無邊的網際網路上的各種知識和資訊，因此，世界各國都非常重視相關的研究。在美、英、日、法、德等發達國家，自然語言處理如今不僅作為人工智慧的核心課題來研究：而且也作為新一代智慧電腦的核心課題來研究。從知識產業的角度來看，自然語言處理的軟體佔最重要地位，例如專家系統、資料庫、知識庫．電腦輔助設計系統(CAD)、電腦輔助教學系統(CAI)、電腦輔助決策系統，辦公室自動化管理系統、智慧型機器人等，無一不需要用自然語言做人機介面。從長遠看，具有文章認知能力的自然語言理解系統可用於機器自動翻譯、情報檢索、自動索引、自動文摘、自動寫作文/故事/小說等領域，具有更廣闊的智慧生活應用和令人鼓舞的應用前景。而隨著語料庫建設和語料庫語言規則的發展，在大規模真實巨量資料的處理和自動學習的統計分析技術等日漸成熟，自然語言處理已成為雲端運算的主要應用。

## 三、基於自然語言的巨量資料應用案例

雲端運算的核心是商業模式(Business Model)，本質是資料內容處理技術。資料內容是資產，雲端運算為資料內容資產提供了保管、搜索的平臺和通路服務。其中巨量資料則是雲端運算的主要核心之一。

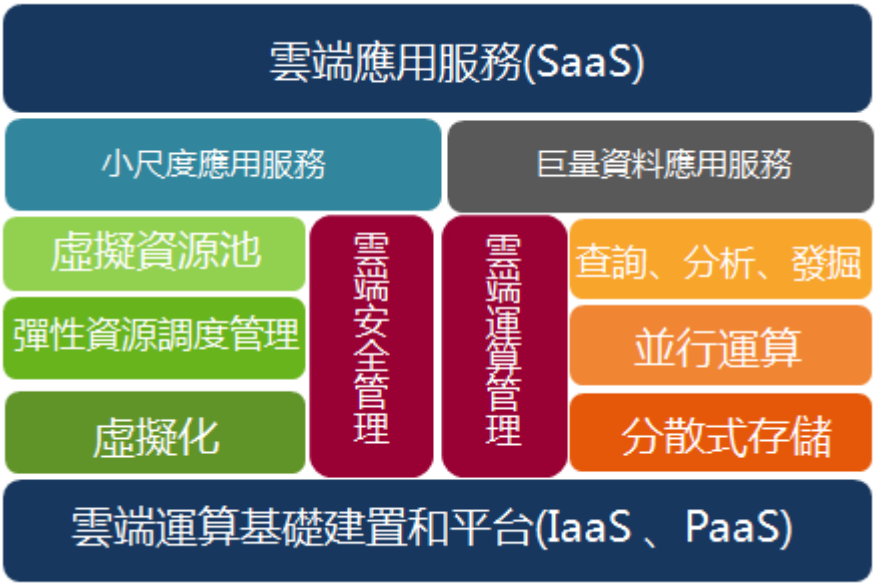


圖 3. 雲端運算應用服務架構圖

近年來應用資料規模急劇增加，傳統運算面臨嚴重挑戰。在大陸的中國移動公司一個省的電話通話記錄資料每月可達 0.5~1 PB (1 PB=$10^{15}$ B)，而整個中國移動公司每月則高達 7-15 PB 資料；如此巨大的資料量使得目前一些資料庫系統已經難以支撐和應付；例如大陸百度公司儲存數百 PB 巨量資料，每天處理資料高達 10 PB、大陸淘寶公司儲存 14 PB 交易的巨量資料，每天新增資料 40~50 TB (1 TB=$10^{12}$ B)等等。未來急劇增長的巨量資料迫切需要尋求新的處理技術方法。據 IDC 報告稱全世界資料總量未來 10 年將從 2009 年的 0.8 ZB(1 ZB=$10^{21}$ B)增長到 2020 年的 35 ZB，增長 44 倍！年均增長率大於 40%！在社交商務備受重視的今天，企業希望能從大量的影視、郵件、即時通訊、社交網站等獲取有價值的資訊，瞭解消費者的購買傾向，預測市場需求，進行相關商業活動分析，因此巨量資料在應用演算法的研究開發正是最熱門話題：

| 巨量資料在應用演算法的研究開發 | |
|---|---|
| 社會網絡 | 社群發現、網路建模、分類推薦、隱私安全等 |
| 排名與推薦系統 | 常規排名、多樣性排名、基於內容的推薦、基於標籤的推薦、協同過濾推薦等 |
| 商業智慧 | 不同產業模式的智慧統計分析模型 |
| 媒體分析檢索 | 大規模影像檢索、大規模影像分類、目標檢測、影視內容異常行為檢測等 |
| WEB搜索與資料發掘 | 深度Web搜索（精確化，智慧化，綜合化資訊搜索）、页面分类、网页摘要等 |
| 3D建模与運算 | 地質建模與分析、電影渲染、大規模資料的運算與分析等 |
| 生物科技資訊處理 | 基因序列拼序、基因序列對比、生物網路建模與分析等 |
| 自然语言处理 | 机器翻译、情感分析、智慧输入系统、人機问答系统 |

圖 4. 巨量資料在應用演算法的研究開發圖

由於雲端運算在巨量資料應用上需求日趨成熟，我們也看到國際雲端服務大企業正積極併購發展相關 NLP(自然語言處理)的核心技術。例如 Google 收購自然語言技術公司 Wavii，主要是 Wavii 適合 Google Now，其中 Wavii 的語義搜索功能將對 Google Knowledge 產生有效精準的幫助作用。另外，Wavii 的 NLP(自然語言處理)和理解認知技術也將會被

廣泛應用到 Google 的多個平臺之中，包括 Google News 和 Google Glass 等平臺在內。這是一個自然語言處理技術與雲端服務(搜索平臺)的有效整合案例。另外一個自然語言處理的巨量資料應用案例是有關『Yahoo 已經收購了自動新聞摘要應用 Summly，創始人 Nick D'Alosio 年僅 17 歲。』在這是個資訊爆炸、媒體革命的時代，社交軟體、網路日誌、網站等資訊來源如潮水般湧來，現代人卻愈來愈無法迅速或即時消化。因為傳統的網路日誌、網站文章篇幅往往很長，無法在行動電話上快速流覽，許多內容可能是一大串文章內容的連結。嗅到其中商機的德國青少年 Nick D'Alosio 就乾脆輟學組建團隊，研發出這款酷勁十足、獨樹一幟的 iOS 自動新聞摘要應用：Summly。這也是個基於自然語言處理的『自動摘要(Automatic Summarization)』技術的巨量資料應用之經典案例。

## 四、結語

自然語言工程發展至今，已從單純的自動翻譯、語音輸入、語音合成、語音識別等等資料處理應用，到目前依靠許多用戶習慣，使用時間、用戶年齡、用戶性別、地理位置等等資料發掘應用，包括對於用戶職業、用戶活躍度、用戶行為喜好等各種各樣的巨量資料進行分析。2012 年 3 月美國政府宣佈投資 2 億美元啟動「巨量資料研究和發展計畫」。這是繼 1993 年美國宣佈「資訊高速公路計畫」後又一次重大科技發展部署。巨量資料將基於包括自然語言處理在內的大尺度分析和知識發掘等應用，帶來新一輪的商業發展機會。