第一章　緒論

1.1 研究主題及動機

資訊技術不斷地突進，一直是資訊生活的最大推手。然而，找尋更寬闊的資訊系統應用領域，以及人機介面(Human Computer Interface, HCI)的充實、改良，卻是資訊生活普及所不可忽視的兩大環節。除了技術的精益求精，如何有效運用現有資源、激發創意靈感，便成為資訊發展的另個重要課題。

近年伴隨多媒體網路的發光，音頻訊號處理(Audio Signal Processing)成為資訊領域的另一研究焦點，無論是語音辨識或合成，都成為資訊生活化，以及改善人機介面的重要核心技術。以語音合成為例，為達成貼近自然人聲的遠程目標，如何讓機器學習、模擬人類說話的方式[1]，如語調的變化、韻律節奏的起頓乃至文意的表達等，都是語音合成研究的重點。而其中，斷詞在中文語音合成又極具影響。

此外，在許多自然語言的研究和應用中，如中文的輸入、文件檢索、OCR辨識與機器翻譯中，中文斷詞器都是不可少的前置處理。因此，本文嘗試藉由語言學的基礎，建立一套適合中文語音合成的斷詞器，除修正過往通適的斷詞方法，以找尋、建造適用於中文語音合成的前處理器，提供後續合成所需的各種可能訊息，如連讀變音、變調以及韻律生成器倚重的斷詞結果外，更期能融合語言與資訊兩大學科領域，在資訊系統應用下產生火花，作一初步的嘗試。

[1] 個體間雖有不同的發音“習慣”，但每種語言皆有其獨特的發音“方式”。

1.2 中文斷詞系統簡介

因中文方塊字及詞潔意深多樣組合等特性，使得中文斷詞的結果常面臨嚴重的歧異性問題，即一連續的中文字串，可能有許多不同的詞組變化，如“新年好不快樂”的斷詞結果將可能是“新年好　不快樂”，或可成為“新年　好不快樂”，如此詞彙混淆(Lexical Ambiguous)的現象，在語音合成的應用上，不僅可能造成韻律生成的不自然，對語意的傳達更是模糊。此外綜觀目前的斷詞系統，雖說中文斷詞為自然語言應用中不可少的前置處理，然而對「詞」的定義標準不一，不同的應用系統對「詞」的需求也有所不同，使得斷詞系統成效評估不易。故此，本文將嘗試針對語音合成的特性，建立一套中文語音合成前處理系統。

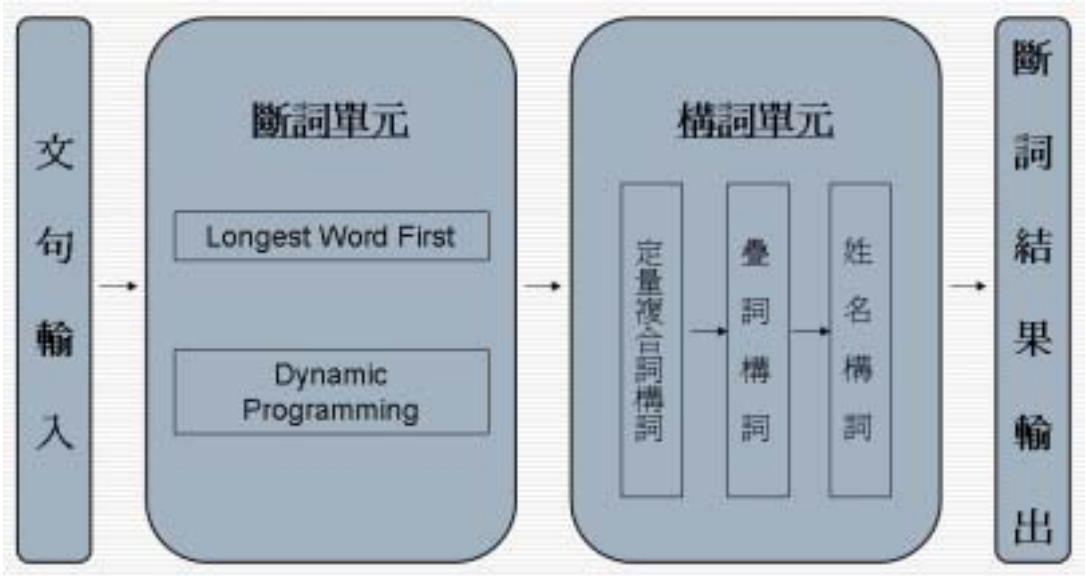

圖 1-1 斷詞系統基本架構及斷詞流程示意圖

本文的斷詞系統可略分兩大單元。其一的斷詞單元中，我們分別以長詞優先法(Longest Word First)及動態規劃演算法(Dynamic Programming)兩種方法，搭配

中央研究院漢語平衡語料庫[2]，以及敝實驗室歷年收集、整理的語文資料庫[3]，實作兩套系統分別測試斷詞的結果。然而詞庫的匯集耗費相當地時間與人力，且龐大的詞庫對斷詞系統的整體效能，如記憶體的使用與斷詞效率等也是一大負擔。因而考量中文構詞的特性，我們又加入構詞單元，其中包含定量複合詞構詞、疊詞構詞以及姓名構詞，參酌增減中央研究院詞庫小組所訂 34 條定量複合詞構詞規則、中文常見的六型疊詞[4]，以及最後利用百家姓的姓名構詞。

1.3 章節概要

本文第一章緒論簡介論文的研究主題與動機，並對斷詞系統作一整體的概觀。第二章則從語言學的基礎找尋中文語音合成中可用的資源，如連讀變調，「一」、「不」變調，疊詞變調及疑問句尾升調等，作為中文語音合成前處理的礎石。第三章為斷詞系統的詳細介紹，包含詞庫的建置與整理、Longest Word First 與 Dynamic Programming 兩大斷詞方法，以及三大構詞。第四章實驗過程與結果，第五章的錯誤分析與未來展望提出幾個未來可行的改進方向。

[2] http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/wordlist.htm，詳見 3.1.3 節。

[3] 詳見 3.1.2 節。

[4] 詳見 3.3.2 節。