具有新詞辨識能力的中文斷詞系統

本院覽號

05T-891002

公告日期

智財權狀態

know-how

摘要

詞是最小有意義且可以自由使用的語言單位。任何語言處理的系統都必須先能分辨文本中的詞才能進行進一步的處理，例如機器翻譯、語言分析、語言了解、資訊抽取。因此中文自動分詞的工作成了語言處理不可或缺的技術。 本系統整合了分詞及線上新詞辨識技術，為唯一具有新詞辨識能力並附加詞類標記的選擇性功能之中文斷詞系統。此一系統包含一個約拾萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。分詞依據為此一詞彙庫及定量詞、重疊詞等構詞規律及線上辨識的新詞，並解決分詞歧義問題。除了基本詞彙庫外，使用者可依需要附加領域專屬詞庫。一般文件若不考慮新詞平均切分正確率達99%以上。詞類標記為選擇性功能，可附加文本中切分詞的詞類解決詞類歧義，正確率在95%以上。分詞用詞典俱可擴充性，使用者可依據不同領域文件，補充領域詞典做為分詞之用。

技術優勢

由於中文詞集是一個開放集合，不存在任何一個詞典或方法可以盡列所有的中文詞。當處理不同領域的文件時，領域相關的特殊詞彙或專有名詞，常常造成分詞系統因為參考詞彙的不足而產生錯誤的切分。為了解決這個問題，最有效的方法是補充領域詞典加強詞彙的搜集。因此新的詞彙或關鍵詞的自動抽取成為分詞的先期準備步驟。領域關鍵詞彙多出現在該領域的文件中而少出現在其它領域，因此抽取關鍵詞時多利用此特性。高頻的關鍵詞比較容易抽取，少數低頻的新詞不容事先搜集，必須線上辨識。構詞律、詞素、詞彙及詞彙共現訊息，為線上新詞辨識依據。此一問題至今尚未有完整的解答，部分的研究成果包括人名、地名、公司名辨識，未知詞偵測，複合詞構詞律等。 本系統整合了分詞及線上新詞辨識技術，為唯一具有新詞辨識能力並附加詞類標記的選擇性功能之中文斷詞系統。

應用範圍

資訊檢索 機器翻譯 語言分析 語言了解 訊息抽取 自然語言人機介面

創作人

陳克健、馬偉雲

中央研究院
ACADEMIA SINICA