

中文分詞語料庫

本院覽號

05T-890903

公告日期

智財權狀態

know-how

摘要

中文分詞語料庫為一包含兩百萬詞、不含詞類標記的語料庫，每個文句皆根據分詞原則來斷詞。而此分詞原則，乃是中華民國計算語言學學會在經濟部中央標準局委辦的「資訊處理用中文分詞規範調查研究及草案研擬」計畫中所訂定的。本語料庫來源包括書面語和口語兩部分，其中資訊類佔21%。

技術優勢

本與料庫每個文句皆根據分詞原則來斷詞。

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