

【語文教育及編譯研究中心助理研究員 白明弘】

離合詞是指在中文裡具備離析與結合兩種表達形式之語言結構，為中文特有的語言現象之一。以「結婚」一詞為例，既可以結合的形式使用：「他們在法院裡公證**結婚**。」也可以離析的形式使用：「盧娜**結過三次婚**，每次都嫁給醫生。」在「結婚」的兩個語素中間，插入了時貌標記「過」表示經驗，以及數量「三次」表示經驗次數。由於離合詞是中文裡獨特的現象，對華語學習者來說極容易造成混淆。例如華語學習者可能產生如下列的偏誤：「*我媽媽叫我**結婚臺灣人**。」、「*我們已經各自**結婚過了**。」前者是由於學習者不瞭解「結婚」在中文語法屬不及物動詞，所以誤當及物動詞使用；後者是學習者不瞭解「結婚」屬於離合詞，誤將時貌標記附加在離合詞後面。

過去的研究發現，離合詞語法變化複雜，而離合形式間的語義變化細微，導致離合詞成為華語學習的困難點。但是，離合詞卻常常是日常生活中的重要用語，使得離合詞教學成為不可迴避的問題。根據學者的建議，在教材中提供典型的例句及使用情境，能讓學生清楚掌握離合詞的使用情境與時機，提高學習者對離合詞離合形式特徵的辨識能力，有效降低離合詞使用的偏誤。然而，要對離合詞做有系統的整理實為困難的工作，即使是母語者也很難全面掌握離合詞的性質。近年來，離合詞的研究逐漸藉助於語料庫的觀察。透過大量的語言真實使用情境記錄，學者得以全面考察離合詞豐富的性質。可惜的是，目前除了少數人工標記的語料庫外，中文語料庫大多缺乏離合詞訊息。究其原因在於人工標注離合詞的成本極高，而且數量也不易擴大。若能發展一套有效率的離合詞自動標記系統，對離合詞的教學研究將產生極大的幫助。

本計畫的主要目的是要建構一個可靠的離合詞自動識別與標記系統。藉由深入分析離合詞特性，結合機器學習理論，開發離合詞自動標記的工具，以做為語料庫建立離合詞標記的基礎。在機器學習的自動標記方法中，大致可分為規則式自動標記法與統計式自動標記法。規則式自動標記法是自動從訓練語料庫中抽取離合詞的規則，以建立離合詞辨識的規則知識庫；再以辨識規則知識庫作為大量離合詞辨識的依據。統計式自動標記法則是利用統計式機器學習技術，自動從訓練語料庫中抽取特徵與建立自動辨識模型。此自動辨識模型，即可作為大量離合詞辨識之應用。經過實驗比較發現，規則式自動標記法的優點是系統精確率高，但缺點是適應性較差，導致離合詞辨識的召回率低。另一方面，統計式自動辨識模型，雖然精確率較低，但適應性強，辨識結果可達極高的召回率，而精確率也只略低於規則式自動標記法而已。從整體來說，以統計式機器學習模型作為離合詞標記系統，可以達到令人滿意的離合詞辨識效果。

資料來源：

白明弘（2018）。中文離合詞自動識別與標記之研究。國家教育研究院整合研究計畫案成果報告（NAER-107-12-F-2-01-00-1-01）。新北市：國家教育研究院。