Computational Linguistics and Chinese Language Processing
Vol. 9, No.1, February 2004, pp.97-112

基於術語抽取與術語叢集技術的主題抽取

Topic Extraction Based on Techniques of Term Extraction and Term Clustering

林頌堅*

Sung-Chen Lin*

摘要

本論文針對主題抽取的問題，提出一系列以自然語言處理爲基礎的技術，應用這些技術可以從學術論文抽取重要的術語，並將這些術語依據彼此間的共現關係進行叢集，以叢集所得到的術語集合表示領域中重要的主題，提供研究人員學術領域的梗概並釐清他們的資訊需求。我們將所提出的方法應用到ROCLING研討會的論文資料上，結果顯示這個方法可以同時抽取出計算語言學領域的中文和英文術語，所得到的術語叢集結果也可以表示領域中重要的主題。這個初步的研究驗證了本論文所提出方法的可行性。重要的主題包括機器翻譯、語音處理、資訊檢索、語法模式與剖析、斷詞和統計式語言模型等等。從研究結果中，我們也發現計算語言學研究與實務應用有密切的關係。

關鍵詞：主題抽取、術語抽取、術語叢集

Abstract

In this paper, we propose a series of natural language processing techniques to be used to extract important topics in a given research field. Topics as defined in this paper are important research problems, theories, and technical methods of the examined field, and we can represent them with groups of relevant terms. The terms are extracted from the texts of papers published in the field, including titles, abstracts, and bibliographies, because they convey important research information and are relevant to knowledge in that field. The topics can provide a clear outline of the field for researchers and are also useful for identifying users' information

*世新大學資訊傳播學系 Department of Information and Communications, Shih-Hsin University, Taipei, Taiwan, R.O.C.
Email: scl@cc.shu.edu.tw