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

needs when they are applied to information retrieval. To facilitate topic extraction, key terms in both Chinese and English are extracted from papers and are clustered into groups consisting of terms that frequently co-occur with each other. First, a PAT-tree is generated that stores all possible character strings appearing in the texts of papers. Character strings are retrieved from the PAT-tree as candidates of extracted terms and are tested using the statistical information of the string to filter out impossible candidates. The statistical information for a string includes (1) the total frequency count of the string in all the input papers, (2) the sum of the average frequency and the standard deviation of the string in each paper, and (3) the complexity of the front and rear adjacent character of the string. The total frequency count of the string and the sum of its average frequency and standard deviation are used to measure the importance of the corresponding term to the field. The complexity of adjacent characters is a criterion used to determine whether the string is a complete token of a term. The less complexity the adjacent characters, the more likely the string is a partial token of other terms. Finally, if the leftmost or rightmost part of a string is a stop word, the string is also filtered out. The extracted results are clustered to generate term groups according to their co-occurrences. Several techniques are used in the clustering algorithm to obtain multiple clustering results, including the clique algorithm and a group merging procedure. When the clique algorithm is performed, the latent semantic indexing technique is used to estimate the relevance between two terms to improve the deficiency of term co-occurrences in the papers. Two term groups are further merged into a new one when their members are similar because it is possible that the clusters represent the same topic. The above techniques were applied to the proceedings of ROCLING to uncover topics in the field of computational linguistics. The results show that the key terms in both Chinese and English were extracted successfully, and that the clustered groups represented the topics of computational linguistics. Therefore, the initial study proved the feasibility of the proposed techniques. The extracted topics included "machine translation," "speech processing," "information retrieval," "grammars and parsers," "Chinese word segmentation," and "statistical language models." From the results, we can observe that there is a close relation between basic research and applications in computational linguistics.

Keywords: Topic extraction, term extraction, term clustering

1. 緒論

本論文提出一個自動化的主題抽取方法，利用論文中的詞彙訊息來抽取學術領域的主題。論文的題名、摘要、本文，甚至所引用的參考文獻題名等文字資料表達了研究的問