同位詞夾子:主題式分類詞庫萃取演算法

謝育平 *

在資訊檢索、自然語言處理、數位典藏等眾多領域之中，文字處理始終是研究者面臨的第一個課題。對中文資料處理來說，自動斷詞、命名實體萃取、詞彙自動分類是前置工作的重點，傳統研究著重於各項自動化工作的精準率與召回率。本文提出半自動主題式詞庫萃取演算法，命名爲「同位詞夾子」，主要利用人工來保證精準率，利用機器速度來補足召回率，以達到極高的準確率與儘量高的召回率。

同類的詞彙具有很高的同位性；例如「台北」與「高雄」就有很高的同位性；所謂同位性係指在文件中所有出現「台北」的地方，幾乎都可以使用「高雄」來替代，且替代後文句仍是非常通順，所以我們稱「台北」與「高雄」互爲同位詞。「同位詞夾子」是由五個部件所組成（前文、前綴、中綴、後綴、後文），主要描述一個詞彙在文件某處的特徵，用以在文件中萃取該詞彙的同位詞。

本文演算法事先要求使用者提供該類詞彙的種子範例，演算法利用種子範例在文件中掃描以產生該類詞彙的詞夾子，再利用產生的詞夾子在文件中掃描並夾出該類詞彙的候選詞，依照候選詞數值化後的「同位性分數」排序供使用者人工決定是否符合該分類；再依人工幫助擴充種子範例、重啓演算法，如此互動循環到滿意爲止。

本文成功萃取台灣歷史數位圖書館中的人名、地名、官職名、事件名等，也成功在中國古典小說中萃取三國演義的武器名、西遊記的法術名、紅樓夢的衣飾名、金瓶梅的小吃名等非傳統命名實體研究的詞彙分類。平均而言，一個分類詞庫的萃取可以在兩個小時內完成。

關鍵字：同位詞、詞夾子、命名實體、詞彙萃取、文字探勘

* 銘傳大學資訊工程學系助理教授。

Appositional Term Clip: A Subject Oriented Appositional Term Extraction Algorithm

Yuh-pyng Shieh*

On Chinese text processing, automated term extraction, named entity recognition, and term classification are important. Traditional researches focus on recall and precision of these automated works. This paper provides a semi-automated subject-oriented term-extraction algorithm called “Appositional Term Clip”. We utilize “human-aid” to ensure extremely high precision and utilize "machine-aid" to improve recall as higher as possible.

Terms in the same category have high appositional similarity, for example, Taipei and Kaohsiung. Appositional similarity describes that almost all occurrences of “Taipei” can be substituted by “Kaohsiung” with high readability and fluency. Taipei and Kaohsiung are called appositional terms. An appositional term clip comprises 5 parts (preamble, prefix, infix, suffix, postamble) to describe a feature of some term's occurrence.

Applying the algorithm, users are asked to provide some terms of some category as seeds. Appositional term clips are produced by scanning all occurrences of seeds. Candidate (appositional) terms are produced by scanning all occurrences of the produced clips. An ordered list of candidate terms is provided for users to check whether a candidate is in the category. According to the users' work, seeds are enriched and the algorithm is applied again. The semi-automated procedure is applied again and again until the result is satisfied.

We successfully extract some categories such as person names, places, official positions and events from Taiwan History Digital Library, and some special ones such as weapons, magic, clothing, and snacks from Chinese ancient archives. In average, a category can be completed in two hours depending on the size of corpus.

Keywords: Appositional Term, Term Clip, Named Entity, Term Extraction, Text Mining

* Assistant Professor, Department of Computer Science and Information Engineering, Ming Chuan University.