词汇化模板定量匹配——借助于搜索引擎的中文分析

孙茂松* **孙如颖**
清华大学，北京 四川大学，成都

提要
本文阐述了一种借助现有搜索引擎对中文进行辅助研究的思路。[1] 主要考量是本文所提出的“词汇化模板定量匹配”方法。这个方法的要点是期望设计一个针对中文的“词汇化模板体系”，依靠简单的字符串匹配技术，在语言的不同层次上实现对中文某种程度的分析。本文通过若干典型案例说明了所提方法的合理性，并讨论了若干相关的重要问题。这个思路还有待于大规模实验的检验。

关键词
词汇化模板定量匹配 **搜**索引擎 **互**联网语料库 **中**文分析
自然语言处理

LEXICALIZED STATISTICAL PATTERN MATCHING: SEARCH ENGINE-AIDED ANALYSIS FOR THE CHINESE LANGUAGE

Maosong Sun *Tsinghua University, Beijing*

Ruying Sun *Sichuan University, Chengdu*

ABSTRACT

This article presents an idea of search engine-aided analysis for the Chinese language. The core of the idea is the proposed concept "Lexicalized statistical pattern matching". The basic methodology is to perform some degree of Chinese analysis at different linguistic levels by designing and exploiting a lexicalized statistical pattern system, together with the simplest string matching technique search engines used. The rationality of the idea is discussed centering on several typical case studies and, some related key issues are also addressed. It should be noted that this idea is preliminary, needing further validation by large-scale experiments.

KEYWORDS

Lexicalized statistical pattern matching **S**earch engine **W**eb corpus **C**hinese analysis **N**atural language processing