

基于当代汉语流通语料库的 根词相关性知识挖掘研究

盛玉麒*

山东大学

提要

以往关于基本词汇和一般词汇的概括过于宏观。具有“稳定、能产、高频”等特点的“根词”集合是“基本词汇”的核心要素。“根词相关性”知识是解析所有短语乃至单句结构和生成模式的基础。本文运用语料库语言学的理论和方法，通过对 1400 万字符的当代汉语流通语料库中汉语根词相关性的充分描写和定量分析，揭示汉语临时短语结构模式和未登录词语辨识等所需知识的提取问题。¹ 本研究对于汉语本体和应用研究，都具有重要的理论意义和实践参考价值。

关键词

语料库 根词 相关性

KNOWLEDGE MINING ON ROOT WORD CORRELATION
BASED ON MODERN CHINESE CORPUS

Yuqi Sheng

Shandong University

ABSTRACT

In the past, the generalization of basic vocabulary and general vocabulary is too general. The core element of “basic vocabulary” is the “root word”, which is stable, productive and frequent. The knowledge of “root word correlation” is the basis to parse the structure and generative model of all phrases and sentences. This paper uses the corpus linguistics theory and method. Through the adequate description and quantitative analysis for the Chinese root word correlation based on the 14 million character Corpus of modern Chinese, this paper discovers the Chinese temporary phrase structure patterns and the knowledge extraction problems of unknown words identification. This study has important theory significance and the positive practical reference value for Chinese ontology and application research.

KEYWORDS

Corpus Root word Correlation