现代图书情报技术 » 2012, Vol. » Issue (12) :32-38

CSSCI语料中短语结构标注与自动识别

谢靖[1], 苏新宁[2], 沈思[2]

1. 南京中医药大学经贸管理学院 南京 210046;
2. 南京大学信息管理学院 南京 210093

Xie Jing[1], Su Xinning[2], Shen Si[2]

1. School of Economics and Management, Nanjing University of Chinese Medicine, Nanjing 210046, China;
2. School of Information Management, Nanjing University, Nanjing 210093, China

- 摘要
- 参考文献
- 相关文章

Download: PDF (583KB) HTML (KB) Export: BibTeX or EndNote (RIS) Supporting Info

摘要 将短语结构标注引入CSSCI期刊论文题录信息分析,在关键词、术语构成上从语法角度深度探讨各组成词汇之间的语法关系,力图通过语法功能分析揭示其所蕴含的语义知识。在进行一定规模语料标注基础上,通过短语词汇、词性统计及短语语法功能分析获取学术文献中短语结构构成特征,并将这部分特征与清华树库语料短语特征混合,提高短语结构在科技文献中的识别率。

关键词: 短语结构标记 CSSCI语料 混合特征 自动识别

Abstract: The paper introduces a new syntax method as the solution of term phrase identification on CSSCI corpus, and obtains the inter-relationship among terms in academic literature from the linguistic aspect based on phrase components, such as words, part-of-speech, grammar functions, etc. These linguistic features are mixed with phrase features which are extracted from Tsinghua Treebank so as to leverage the accuracy of phrase auto-identification in academic corpus.

Keywords: Phrase annotation, CSSCI corpus, Multi-feature, Auto-identification

收稿日期: 2012-11-14;

基金资助:本文系国家自然科学基金面上项目“面向知识服务的知识组织模式与应用研究”(项目编号:71273126)、高技术研究发展计划(863计划)项目“以科技文献服务为主的搜索引擎研制”(项目编号:2011AA01A206)和江苏省教育厅高校哲学社会科学研究基金项目“基于本体的高校突发事件网络舆情监控预警模式研究”(项目编号:2010SJB870003)的研究成果之一。

通讯作者 谢靖 Email: bmy_xj@163.com

Service
- 把本文推荐给朋友
- 加入我的书架
- 加入引用管理器
- Email Alert
- RSS

作者相关文章
- 谢靖
- 苏新宁
- 沈思

引用本文:

谢靖, 苏新宁, 沈思 .CSSCI语料中短语结构标注与自动识别[J] 现代图书情报技术, 2012,V(12): 32-38

Xie Jing, Su Xinning, Shen Si .Chinese Phrase Tagging and Automated Annotation Based on CSSCI Corpus[J] , 2012,V(12): 32-38

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2012/V/I12/32

[1] Chomsky N. Syntactic Structures[M].Berlin: Mouton de Gruyter, 1957.

[2] Abney S P. Parsing by Chunks[A]. // Berwick R C, Abney S P, Tenny C L. Principle-Based Parsing[M]. Springer, 1991.

[3] The Penn Treebank Project[EB/OL]. [2012-09-12]. http://www.cis.upenn.edu/~treebank/.

[4] 周强. 汉语句法树库标注体系[J]. 中文信息学报, 2004, 18(4):1-8. (Zhou Qiang. Annotation Scheme for Chinese Treebank[J]. *Journal of Chinese Information Processing*, 2004, 18(4):1-8.)

[5] 陈静, 王东波, 谢靖, 等. 基于条件随机场的兼语结构自动识别[J]. 情报科学, 2012, 30(3):439-443. (Chen Jing, Wang Dongbo, Xie Jing, et al. Auton Identification of Concurrent Structure Based on Conditional Random Field[J]. *Information Science*, 2012, 30(3):439-443.)

[6] 朱丹浩, 王东波, 谢靖. 基于条件随机场的介宾结构自动识别[J]. 现代图书情报技术, 2010(7-8):79-83. (Zhu Danhao, Wang Dongbo,Xie Jing. Automa Identification of Prepositional Phrase Based on Conditional Random Field[J]. *New Technology of Library and Information Service*, 2010(7-8):79-83.)

[7] Feng Z W. Analysis of Chinese Terms in Data Processing[R]. Report in Fraunhofer Institute, 1988.

[8] 冯志伟. 一个新兴的术语学科——计算术语学[J]. 术语标准化与信息技术, 2008(4):4-9.(Feng Zhiwei. A New Scientific Domain in Terminology——

Computational Terminology[J]. *Terminology Standardization & Information Technology*, 2008(4):4-9.)

[9] 冯志伟. 汉语单词型术语的结构[J]. 科技术语研究, 2004, 6(1):15-20. (Feng Zhiwei. Structure of Word Terms in Chinese Language[J]. *Chinese Science and Technology Terms Journal*, 2004, 6(1):15-20.)

[10] 冯志伟. 汉语词组型术语的结构[J]. 科技术语研究, 2004, 6(2):35-37. (Feng Zhiwei. Structure of Chinese Phrase Term[J]. *Chinese Science and Technology Terms Journal*, 2004, 6(2):35-37.)

[11] 冯志伟. 术语形成的经济律——FEL公式[J]. 中国科技术语, 2010, 12(2):9-15. (Feng Zhiwei. Economic Law of Term Formation— FEL Formula[J]. *Chinese Science and Technology Terms Journal*, 2010, 12(2):9-15.)

[12] CRF + +: Yet Another CRF Toolkit[EB/OL]. [2012-09-11]. http://crfpp.sourceforge.net/.

[1] 朱丹浩 王东波 谢靖.基于条件随机场的介宾结构自动识别*[J]. 现代图书情报技术, 2010,26(7/8): 79-83

[2] 李睿.Typereader—先进的扫描识别系统[J]. 现代图书情报技术, 1997,13(4): 43-45

[3] 万锦堃.SISAC期刊条码结构分析及性能评价[J]. 现代图书情报技术, 1994,10(6): 29-32