现代图书情报技术 » 2014, Vol. 30 » Issue (1) :22-27

面向文本挖掘的植物生长发育实体识别研究

汪润, 何琳, 王东波, 黄水清, 范远标

南京农业大学信息科学技术学院 南京 210095

Wang Run, He Lin, Wang Dongbo, Huang Shuiqing, Fan Yuanbiao

College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China

- 摘要
- 参考文献
- 相关文章

Download: PDF (471KB) HTML (1KB) Export: BibTeX or EndNote (RIS) Supporting Info

摘要 【目的】 研究从文本中识别植物生长发育实体（Plant Growth and Development Stage Named Entity，PDSE）的抽取。【应用背景】PDSE从本质上来说是一种命名实体。目前有关命名实体的识别已经成为自然语言处理领域最有价值的基础技术之一，被广泛应用于多种自然语言处理系统中。【方法】采用基于条件随机场和规则的混合策略，提出并实现针对PDSE特征的CRF特征模板、特征函数以及抽取规则的方法，并利用PubMed数据库收录的论文进行抽取效果测试。【结果】实验表明本文提出的混合策略能取得较高的准确率和召回率。【结论】本研究对生物学文本抽取具有一定的借鉴意义。

关键词： 植物生长发育时期 命名实体识别 条件随机场 特征选择

Abstract： [Objective] This paper researches in the extraction that identifies plant growth and development stage entity from text. [Context] PDSE is a kind of named entity essentially. Named entities recognition has become one of most valuable basic technologies in Natural Language Processing field,which is used widely in many Natural Language Processing systems. [Methods] It adopts multiple strategies based on conditional random field and rules,with putting forward and realizing a method of CRF template,characteristic function and extraction rules for the features of plant growth and development stage entity. Also,it tests the extraction effect by articles from the PubMed database. [Results] The experiment shows that the proposed hybrid strategies can obtain high accuracy and recall rate. [Conclusions] This research has a certain significant reference for biology text extraction.

Keywords： Plant growth and development stage, Named entity recognition, CRF, Feature selection

收稿日期: 2013-09-10;

基金资助:

本文系国家社会科学基金“面向知识服务的科学数据组织与应用研究”（项目编号：13CTQ035）、中央高校基本科研业务费资助项目“面向qRT-PCR实验的内参基因挖掘技术研究”（项目编号：KYZ201159）和南京农业大学SRT计划项目“基于混和策略的植物生长发育时期识别”（项目编号：1219A11）的研究成果之一。

通讯作者 何琳 Email: helin@njau.edu.cn

作者贡献：何琳，黄水清：提出研究思路，设计研究方案；汪润：进行实验；范远标：数据采集和清洗；汪润，何琳：论文起草；王东波：最终版本修订及数据结果评价分析。

引用本文:

汪润, 何琳, 王东波等 .面向文本挖掘的植物生长发育实体识别研究[J] 现代图书情报技术, 2014,V30(1): 22-27

Wang Run, He Lin, Wang Dongbo etc .Research on Plant Growth and Development Stage Named Entity Recognition for Text Mining[J] , 2014,V(1): 22-27

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2014/V30/I1/22

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
- 汪润
- 何琳
- 王东波
- 黄水清
- 范远标

[1] 宗萍,施水才,王涛,等. 基于条件随机场的英文地理行政实体识别[J]. 现代图书情报技术,2009（2）：51-55.（Zong Ping,Shi Shuicai,Wang Tao,et al. GP entity Recognition Based on Conditional Random Fields [J]. New Technology of Library and Information Service,2009（2）：51-55.）

[2] 周雅倩,郭以昆,黄萱菁,等. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展,2003,40（3）：440-446.（Zhou Yaqian,Guo Yikun,Huang Xuanjing,et al. Chinese and English BaseNP Recognition Based on a Maximum Entropy Model[J]. Journal of Computer Research and Development,2003,40（3）：440-446.）

[3] 张朝胜,郭剑毅,线岩团,等. 基于条件随机场的英文产品命名实体识别[J]. 计算机工程与科学,2010,32（6）：115-117.（Zhang Chaosheng,Guo Jianyi,Xia Yantuan,et al. Named Entity Recognition of the Products with English Based on Conditional Random Fields[J]. Computer Engineering and Science,2010,32（6）：115-117.）

[4] Ferro L,Gerber L,Mani I,et al.TIDES 2005 Standard for the Annotation of Temporal Expressions[R]. MITRE,2005：1-65.

[5] ACE（Automatic Content Extraction） Chinese Annotation Guidelines for TIMEX2（Summary）[EB/OL]. [2013-12-19]. http://www.ldc.upenn.edu/Projects/ACE/docs/Chinese-TIMEX2-Guideline-Summary_v1.2.pdf.

[6] Saquete E,Martínez-Barco P. Grammar Specification for the Recognition of Temporal Expressions[C]. In：Proceedings of Machir Translation and Multilingual Applications in the New Millennium.2000.

[7] Schilder F,Habel C. From Temporal Expressions to Temporal Information：Semantic Tagging of News Messages[C]. In：Proceedings of the Workshop on Temporal and Spatial Information Processing（TASIP’01）,Morristown,NJ. Stroudsburg：Association for Computati Linguistics,2001：Article No.9.

[8] Brill E. Transformation-based Error-driven Learning and Natural Language Processing：A Case Study in Part-of-Speech Tagging[J]. Computational Linguistics,1995,21（4）：543-565.

[9] 贺瑞芳,秦兵,潘越群,等. 基于启发式错误驱动学习的中文时间表达式识别[J]. 高技术通讯,2008,18（12）：1258-1262.（He Ruifang,Qin Bing,Pan Yuequn,et al. Recognizing Chinese Time Expressions Based on Heuristic Error-driven Learning[J]. High Technology Letters,2008,18（12）：1258-1262.）

[10] Hacioglu K,Chen Y,Douglas B. Automatic Time Expression Labeling for English and Chinese Text[C]. In：Proceedings of the 6th Internatior Conference on Computational Linguis- tics and Intelligent Text Processing（CICLing’05）. Berlin,Heidelberg：Springer-Verlag,20(548-559.

[11] Ahn D D,Adafre S F,De Rijke M. Towards Task-based Temporal Extraction and Recognition[C]. In：Proceedings of Dagstuhl Workshop on Annotating,Extracting,and Reasoning about Time and Events. 2005.

[12] 欧阳佑,李素建.条件随机域模型和实验分析[C]. 见：第三届学生计算语言学研讨会论文集,沈阳,辽宁,中国.中国中文信息学会,2006：134-139.（Ou Yangyou Sujian. Conditional Random Fields for Temporal Expression Recognition[C]. In: Proceedings of the SWCL-2006, Shenyang, Liaoning Provin China.Chinese Information Association of China, 2006：134-139.）

[13] 朱莎莎,刘宗田,付剑锋,等. 基于条件随机场的中文时间短语识别[J]. 计算机工程,2011,37（15）：164-167.（Zhu Shasha,Liu Zongtian,Fu Jianfeng,et Chinese Temporal Phrase Recognition Based on Conditional Random Fields[J]. Computer Engineering,2011,37（15）：164-167.）

[14] 许旭阳,李弼程,张先飞,等. 基于条件随机场与自定义规则的时间表达式识别[J]. 情报学报,2011,30（10）：1065-1071.（Xu Xuyang,Li Bicheng,Zhang Xianfei,et al. Recognition of Time Expressions Based on Conditional Random Fields and Rules[J]. Journal of the China Society for Scientific and Technical Information,2011,30（10）：1065-1071.）

[15] 孙镇,王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术,2010（6）：42-47.（Sun Zhen,Wang Huilin. Overview on the Advance of the Research on Named Entity Recognition[J]. New Technology of Library and Information Service,2010（6）：42-47.）

[16] Lafferty J D,McCallum A,Pereira F C N. Conditional Random Fields：Probabilistic Models for Segmenting and Labeling Sequence Data[C]. In Proceedings of the 18th International Conference on Machine Learning（ICML’01）. San Francisco：Morgan Kaufmann Publisher Inc.,2001：282-289.

[17] CRF++：Yet Another CRF Toolkit[EB/OL]. [2013-07-15]. http://crfpp.googlecode.com/svn/trunk/doc/index.html? source =navbar.

[1] 关晓炟, 吕学强, 李卓, 郑略省.用户查询日志中的中文机构名识别[J]. 现代图书情报技术, 2014,30(1): 72-78

[2] 胡昌平, 陈果.共词分析中的词语贡献度特征选择研究[J]. 现代图书情报技术, 2013,29(7/8): 89-93

[3] 何文静, 何琳.基于社会标签的文本聚类研究[J]. 现代图书情报技术, 2013,29(7/8): 49-54

[4] 王昊, 邹杰利, 邓三鸿.面向中文图书的自动标引模型构建及实验分析[J]. 现代图书情报技术, 2013,29(7/8): 55-62