山东大学学报(工学版) 2010, 40(5) 117-122 DOI: ISSN: 0412-1961 CN: 21-1139

论文

基于字的词位标注汉语分词

于江德1,睢丹1,樊孝忠2

1. 安阳师范学院计算机与信息工程学院, 河南 安阳 455002;
2. 北京理工大学计算机科学技术学院, 北京 100081

摘要:

近年来基于字的词位标注方法极大地提高了汉语分词的性能,该方法将汉语分词转化为字的词位标注问题,借助于优秀的序列标注模型,基于字的词位标注汉语分词方法逐渐成为汉语分词的主要技术路线。该方法中特征模板选择至关重要,采用四词位标注集,使用条件随机场模型进一步研究基于字的词位标注汉语分词技术,在第三届和第四届国际汉语分词评测Bakeoff语料上进行封闭测试,并对比了不同特征模板集对分词性能的影响。实验表明采用的特征模板集:TMPT-10′ 较传统的特征模板集分词性能更好。

关键词: 汉语分词 条件随机场 词位标注 特征模板

Word-position-based tagging for Chinese word segmentation

YU Jiang-de1, SUI Dan1, FAN Xiao-zhong2

1. School of Computer and Information Engineering, Anyang Normal University, Anyang 455002, China;
2. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract:

The performance of Chinese word segmentation has been greatly improved by word-position-based approaches in recent years. This approach treats Chinese word segmentation as a word position tagging problem. With the help of powerful sequence tagging model, word-position-based method quickly rose as a mainstream technique in this field. Feature template selection is crucial in this method. We further studied this technique via using four word positions and conditional random fields. Closed evaluations are performed on corpus from the third and the fourth international Chinese word segmentation Bakeoff, and comparative experiments are performed on different feature templates. Experimental results show that the feature template set: TMPT-10′ is much better performance than the traditional template set.

Keywords: Chinese word segmentation conditional random fields word-position tagging feature template

收稿日期 2010-01-30 修回日期 网络版发布日期

DOI:

基金项目:

高等学校博士学科点专项科研基金资助项目(20050007023)

通讯作者:

作者简介: 于江德 (1971-),男,河南林州人,副教授,博士,研究方向为计算语言学、中文信息处理、文本信息抽取等.E-mail: jiangde-yu@tom.com

作者Email:

PDF Preview

参考文献:

本刊中的类似文章