计算机工程与应用 2010 46 (4): 139-141 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

分词语料库中的并列式四字格识别

徐润华，陈小荷，李 斌

南京师范大学 文学院，南京 210097

收稿日期 2008-9-12 修回日期 2008-12-11 网络版发布日期 2010-2-2 接受日期

摘要 并列式四字格是一种特殊却数量众多的四字格。介绍了在有词性标注语料库中基于条件随机场模型的四字格抽取工作，并在此基础上分析了并列式四字格的结构特点，提出了一种基于分词语料库环境的并列式四字格识别方法。通过不同语料库间的对比实验，结果表明该识别方法具有比较好的精确度和一定的适应性。

关键词 四字格 条件随机场模型 分词碎片 并列式四字格

分类号 TP391.1

Recognition of parallel four-character idioms in word-segmented corpora

XU Run-hua，CHEN Xiao-he，LI Bin

College of Liberal Arts，Nanjing Normal University，Nanjing 210097，China

Abstract

Among all kinds of Chinese four-character idioms，the Parallel Four-Character Idiom（PFCI） is special and numerous.This paper introduces the research based on Conditional Random Fields（CRF） model which can retrieve PFCI from a POS-tagged corpus.The paper then analyzes the structural characteristics of PFCI and proposes an approach on recognizing PFCI in word-segmented corpora.By comparing its application on different corpora，the evaluation results show that this recognition approach maintains relatively high precision and good adaptability.

Key words four-character idioms conditional random fields segmented fragments parallel four-character idioms

DOI: 10.3778/j.issn.1002-8331.2010.04.045

通讯作者 徐润华 runhuaxu@hotmail.com