计算机工程与应用 2007 43 (15): 166-169 ISSN: 1002-8331 CN: 11-2127/TP

数据库与信息处理

Web新闻语料分词和标注错误分析

张永奎[1，2]，张 彦[1，2]，安增波[3]，刘 睿[1，2]

1.山西大学 计算机与信息技术学院，太原 030006
2.计算智能与中文信息处理省部共建教育部重点实验室，太原 030006
3.中国人民解放军91708部队 自动化工作站，广州 510320

收稿日期 修回日期 网络版发布日期 2007-5-9 接受日期

摘要 通过分析Web突发事件语料库文本的加工统计得出11类错误类型，并对其中的一些错误提出了解决方案。研究结果不仅对语料库加工初期分词、标注方法的改进有启发作用，而且对中文的自动校对方法，提供一定的借鉴。

关键词 中文信息处理 分词 词性标注 错误类型 Web突发事件新闻语料库

分类号

Analysis of inaccurate style in processing Web true news text——about word segmentation and part of speech tagging

ZHANG Yong-kui[1，2]，ZHANG Yan[1，2]，AN Zeng-bo[3]，LIU Rui[1，2]

1.Department of Computer & Information Technology，Shanxi University，Taiyuan 030006，China
2.Key Laboratory of Ministry of Education for Computation Intelligence and Chinese Information Processing，Taiyuan 030006，China
3.Workstation Automation of 91708 PLA，Guangzhou 510320，China

Abstract

Eleven inaccurate styles are obtained through analyzing the processing of Web accidental news text，we propose resolvent for some styles.This not only illuminates the improvement of word segmentation and part of speech tagging methods in early process of corpora，but also provides references to automatic check，another branch of Chinese information processing.

Key words Chinese information processing word segmentation part of speech tagging inaccurate style Web accidental news corpora

DOI:

通讯作者 张永奎 E-mail：zjm1203@sxu.edu.cn