现代图书情报技术 2006 (5): 13-17 ISSN: 1003-3513 CN: 11-2856/G2

数字图书馆

基于既定词表的自适应汉语分词技术研究

黄水清[1];程冲[1;2]

南京农业大学信息科技学院[1]

收稿日期 2005-12-1 修回日期 2006-2-7 网络版发布日期 2006-6-13 接受日期

摘要 提出一种汉语分词算法，在给定的分词词表的基础上进行汉语分词时，不但能成功切分出分词词表中已有的词，而且能同时自动识别出分词词表中没有的词，即未登录词。与逆向最长匹配法以及其他未登录词识别算法进行的测试比较表明，该分词算法可以有效地解决大多数未登录词的识别问题，并且能减少分词错误，同时对分词算法的效率基本没有影响。

Abstract This paper presents an algorithm of self-adaptive matching method in Chinese segmentation. This algorithm not only identifies Chinese words in vocabulary successfully but also identifies unlisted words which are not in vocabulary on basis of decided vocabulary automatically. The test which compares this algorithm with Reverse Maximum Matching Method and some methods which identify unlisted words proves that it can resolve unknown words segmentation effectively, decreases mistakes of Chinese segmentation and has no effect on the efficiency of Chinese segmentation largely.

关键词 自动分词 新词识别 未登录词

Key words Automatic segmentation;New word identification;Unlisted words

分类号 TP391

DOI:

通讯作者:
黄水清 sqhuang@njau.edu.cn
作者个人主页: 黄水清 程冲