计算机工程与应用 2008 44 (19): 146-149 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

哈萨克语词法分析器的研究与实现

达吾勒·阿布都哈依尔, 古丽拉·阿东别克

新疆大学 信息科学与工程学院，乌鲁木齐 830046

收稿日期 2007-10-29 修回日期 2008-2-21 网络版发布日期 2008-6-26 接受日期

摘要 研究了哈萨克语自动词法分析中的附加成分的切分和词干提取问题，并实现了哈萨克语词法分析系统KazStemmer。系统首先对待切分词使用有限状态自动机进行分析。如果成功则将输出作为切分结果，否则再使用双向全切分和词法分析相结合的改进方法来进行切分。与最大匹配法相比，该方法提高了词干提取的正确率和切分速度。同时，在词干表的搜索中首次采用了改进的逐字母二分词典查询机制来提高了词干提取的效率。

关键词 附加成分切分 有限状态自动机 双向匹配 全切分

分类号

Study and implementation of Kazakh lexical scanner

DAWEL Abilhaye,GULILA Altenbek

College of Information Science and Engineering，Xinjiang University，Urumqi 830046，China

Abstract

This paper studies the problems of stem and affix segmentation in Kazakh automatic morphological analysis and develops a system called “KazStemmer”，which can automatically carry out the stem segmentation and tagging processes for Kazakh corpora.In this paper，the authors first use FSM to analyze the stemming words.IF the FSM does not work，then the combination of the bidirectional matching algorithm，omni-word segmentation algorithm and morphological analysis is used to implement the segmentation of stems and word affixes.Compared to the maximum matching algorithm，this method can get higher precision and processing speed.In addition，the authors use the improved binary-seek-by-character dictionary query mechanism.Its performance also influences the segmentation speed significantly.

Key words affixes segmentation FSM bidirectional matching algorithm omni-word segmentation algorithm

DOI:

通讯作者 达吾勒·阿布都哈依尔 dawel@xju.edu.cn