自动化学报 2011 37 (2): 188-195 ISSN: 0254-4156 CN: 11-2109/TP

论文与报告

置信度加权在线序列标注算法

汤步洲, 王晓龙, 王轩

1. 哈尔滨工业大学深圳研究生院计算机科学与技术学科部 深圳 518055

收稿日期 2010-4-29 **修回日期** 2010-10-8 **网络版发布日期** **接受日期**

摘要
序列标注问题是自然语言处理领域的基本问题之一. 序列标注任务是将连续输入的不定长序列, 标注成连续等长的标签序列. 在在线序列标注方法的基本框架下, 针对序列标注任务的特征稀疏特性, 采用置信度加权分类算法思想, 提出了一种新的线性判别式在线序列标注方法---置信度加权在线序列标注算法. 该方法对每个特征权值参数引入一个概率置信度, 取得了优于其他相关算法的性能. 在中文分词, 中文名实体识别以及英文组块分析等问题上, 验证了本文方法的有效性.

关键词 序列标注问题 自然语言处理 在线序列标注算法 置信度加权 概率置信度

分类号

Confidence-weighted Online Sequence Labeling Algorithm

TANG Bu-Zhou, WANG Xiao-Long, WANG Xuan

1. Institute of Computer Science and Technology, Shenzhen Graduate School of Harbin Institute of Technology, Shenzhen 518055

Abstract
Sequence labeling problem is a basic problem in natural language processing field. The task of sequence labeling is to label an input sequence with a label sequence of the same length. Under the fundamental framework of sequence labeling methods, a new online sequence labeling linear algorithm---confidence-weighted online sequence labeling algorithm---was presented for the characteristic of sequence labeling task with sparse features, based on confidence-weighted classification. This algorithm introduced a probabilistic measure of confidence for each parameter of features, and showed better performance than other relative algorithms. Experiments on Chinese segmentation, Chinese named entity recognition and English chunking validated the effectiveness of the proposed algorithm.

Key words Sequence labeling problem natural language processing online sequence labeling linear algorithm confidence-weighted probabilistic measure of confidence

DOI: 10.3724/SP.J.1004.2011.00188

通讯作者 汤步洲 tangbuzhou@gmail.com

作者个人主页 汤步洲; 王晓龙; 王轩