计算机工程与应用 2009 45 (32): 111-113 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

潜在语义分析在连续语音识别中的应用

欧建林，林 茜，史晓东

厦门大学 计算机科学系，福建 厦门 361005

收稿日期 2008-12-4 修回日期 2009-2-19 网络版发布日期 2009-11-26 接受日期

摘要 研究了潜在语义分析（LSA）理论及其在连续语音识别中应用的相关技术，在此基础上利用WSJ0文本语料库上构建LSA模型，并将其与3-gram模型进行插值组合，构建了包含语义信息的统计语言模型；同时为了进一步优化混合模型的性能，提出了基于密度函数初始化质心的k-means聚类算法对LSA模型的向量空间进行聚类。WSJ0语料库上的连续语音识别实验结果表明：LSA+3-gram混合模型能够使识别的词错误率相比较于标准的3-gram下降13.3%。

关键词 潜在语义分析 N元文法 k均值聚类 连续语音识别

分类号 TP391

Application of latent semantic analysis in continuous speech recognition

OU Jian-lin，LIN Qian，SHI Xiao-dong

Department of Computer Science，Xiamen University，Xiamen，Fujian 361005，China

Abstract

The theory of Latent Semantic Analysis（LSA） for speech recognition is described，and the related techniques for implementing LSA-based language modeling in speech recognition systems are presented.An LSA-based semantic model is constructed on the WSJ0 text corpus.This paper uses the interpolation method to combine this semantic model with conventional 3-gram to form a hybrid language model（i.e.，LSA+3-gram）.To optimize the performance of the hybrid model，it applies k-means algorithm to perform vector clustering in the LSA vector space while the density function is used to initialize the centroid.The constructed hybrid language model outperforms the corresponding 3-gram baseline：Continuous speech recognition experiments conducted on the WSJ0 test corpus show a relative reduction in word error rate of about 13.3%.

Key words latent semantic analysis N-gram k-means clustering continuous speech recognition

DOI: 10.3778/j.issn.1002-8331.2009.32.035

通讯作者 欧建林 mandel@xmu.edu.cn