ISSN1000-9825 Journal of Software 软件学报 2000,11(7):971~978

# 基于模糊训练集的领域相关统计语言模型*

陈浪舟 黄泰翼

(中国科学院自动化研究所 北京 100080)

E-mail: huang@nlpr.ia.ac.cn

**摘要** 统计语言模型在语音识别中具有重要作用.对于特定领域的识别系统来说,主题相关的语言模型效果远远优于领域无关的语言模型.传统方法在建立领域相关的语言模型时通常会遇到两个问题,一个是领域相关的语料不像普通语料那样充分,另一个是一篇特定的文章往往与好几个主题相关,而在模型的训练过程中,这种现象没有得到充分的考虑.为解决这两个问题,提出了一种新的领域相关训练语料的组织方法——基于模糊训练集的组织方法.领域相关的语言模型就建立在模糊训练集的基础上.同时,为了增强模型的预测能力,将自组织学习引入到模型的训练过程中,取得了良好的效果.

**关键词** 语音识别,统计语言模型,模糊,自组织学习.

**中图法分类号** TP391

统计 $n$-gram 语言模型在语音识别中为引导搜索过程到可能性最大的词串提供了重要的语言信息[1].但是,普通的语言模型,即主题无关的语言模型,不能很好地利用有关说话内容的领域知识.因此,对于一些特定的主题,普通模型的性能不可避免地会下降.为了解决这个问题,人们对特定的领域分别建立了相关的语言模型.领域相关的模型通常有单模型结构[2]和混合模型结构[3]两种类型.

单模型结构如图 1 所示.首先,语音识别模块根据当前的语言模型对输入的语音信号进行解码,然后对当前的识别结果进行文本主题转换检测,如果发现听写内容的主题发生了变化,则根据新的主题重新选择领域相关的语言模型.由于每次只加载一个领域相关模型,因而单模型结构的主要优点是系统资源的开销较小,缺点是预测性能比混合模型差.

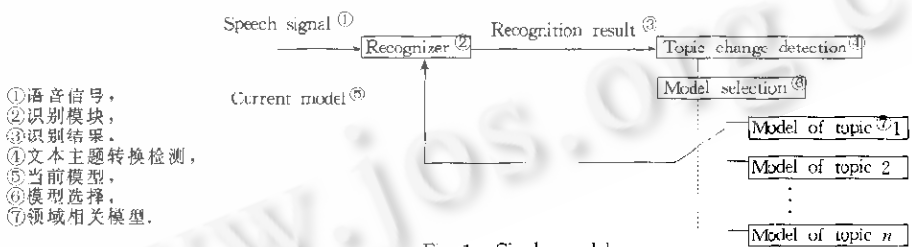


Fig. 1 Single model structure

图1 单模型工作方式示意图

混合模型的工作方式是另一种比较常见的领域相关语言模型工作方式.它的工作示意图为图 2.混合模型的主要特点是,不同领域的语言模型通过线性插值生成当前模型参与识别,但它们的权值不同,因此,对当前模型的贡献也不同.权值根据当前识别结果的主题变化动态地进行调整.混合模型的预测性能优于单模型,在实际系统中也多被采用,但由于要同时加载多套模型,因此系统资源开销较大.

* 本文研究得到国家自然科学基金(No. 69835003)资助.作者陈浪舟,1971 年生,博士,主要研究领域为语音识别,统计语言建模.黄泰翼,1934 年生,研究员,博士生导师,主要研究领域为语音识别,语音合成,自然语言口语处理及口语理解,语言信息处理.

本文通讯联系人:黄泰翼,北京 100080,中国科学院自动化研究所

本文 1999-02-08 收到原稿,1999-06-17 收到修改稿

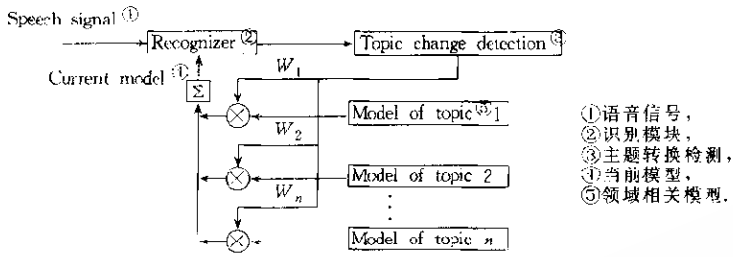


Fig. 2 Structure of mixture models
图2 混合模型工作方式

综上所述,领域相关语言模型用于语音识别必须解决3个问题:一个是领域相关模型的训练问题;另一个是识别结果主题转换的检测问题;最后一个问题主要针对混合模型,即混合权值的计算问题.

上面所说的识别结果主题转换的检测问题是更换模型以及调整权值的依据.由于自然语言在每个不同的领域都有该领域特有的词汇,这些词汇在领域内频繁出现,而在领域以外出现的概率则很小,因此,文本在两个领域交接处词汇的变化很大.我们利用这一现象已非常成功地解决了这一问题,参见文献[4].至于混合权值的计算问题,可以通过著名的EM算法很好地得到解决[5].具体迭代公式如下:

$$x_{n+1}(i)=\frac{1}{M}\sum_{m=0}^{M}\frac{x_n(i)*P_i(w_{n-m}|h_{n-m})}{\sum_{j=1}^{I}x_n(j)*P_j(w_{n-m}|h_{n-m})}, \tag{1}$$

其中 $x_n(i)$ 为第 $I$ 个模型的第 $n$ 次迭代结果,$I$ 为模型个数,$M$ 为训练权值的语料长度.

上述3个问题中还有待改进的是领域相关模型的训练问题,这正是本文讨论的重点.传统的训练领域相关模型的方法是,首先对训练语料进行手工标注,然后对标注后的模型按领域进行分类,最后利用各领域的训练语料分别训练领域相关模型.但是,这种方法的缺点在于:(1) 领域相关的语料从数量上远远少于普通语料,因此数据稀疏问题会变得更加严重;(2) 一篇文章通常与几个主题相对应,因此将文章硬性规定为某一领域的语料并不是一种合理的利用语料的方法.同时,人工标注也很难正确地反映一篇文章与哪些主题相联系以及这种联系的强度.本文为了解决以上两个问题,提出了生成领域相关训练语料的新方法.新算法不再将训练语料划分为不同的主题,而是根据语料聚类的结果,以一种更加合理的方式定义相关训练集,即将领域相关的训练数据当作模糊集来考虑.假设我们的论域 $U=\{u_1,u_2,\ldots,u_n\}$ 为整个训练语料,$u_i,i=1,\ldots,n$ 为训练语料中的文章,传统的设计领域相关训练数据的方法是按篇章聚类的结果将 $U$ 划分为普通子集,每个子集之间有明确的界限.而在本文提出的方法中,每个领域的训练数据被定义为 $U$ 的一个模糊子集,即

$$U=\{u_1,u_2,\ldots,u_n\},$$
$$Topic_j=\sum_{i=1}^{n}\frac{A_j(u_i)}{u_i}, \tag{2}$$

其中 $Topic_j$ 表示主题 $j$ 的训练数据集,$A_j(u_i)$ 为隶属度函数.这样,在新的训练数据定义方法下,不同主题的训练数据之间不再有明确的界限,每个主题由它的隶属度函数所定义.在新的领域相关训练数据集的基础上,我们提出了相应的训练算法.为增加模型的预测能力,自组织竞争学习被引入到训练过程中.我们将新模型和传统模型相比较,无论是单模型还是混合模型,新模型均优于传统模型.

## 1 模糊训练集的构造

模糊训练集的构造是生成新模型的基础,其中主要的工作是为每个领域的模糊集构造一个隶属度函数.如前述,判别一篇文章的主题,主要取决于文中出现的领域关键词.这些关键词的特点是具有爆发特性,即在具有某些特点的文章中频繁出现,而在其他文章中则极少出现.每个主题的文本都有其特定的一些关键词,这些关键词的出现往往可以作为主题发生的标注,单个关键词有时会造成一些错误,但多个关键词在一篇文章中同时并发,通常能为文章的主题检测提供有力的依据.下面首先介绍关键词的选择问题.

### 1.1 关键词的选择

由于关键词通常只在其相关领域内大量出现,而在其他领域出现的概率很小,因此它的检测也是利用这种爆发特性.对于词表中的任何一个词 $w_i$,它在文章 $u_j$ 中出现的概率可以表示为

$$P(w_i \in u_j)=\frac{Count(w_i,u_j)}{Count(w_i)}. \tag{3}$$

对词 $w_i$ 计算它在所有文章中出现的概率 $P(w_i \in u_j)$,$j=1,2,\ldots,n$,这些概率值形成一个一维数组.如果我们将这个一维数组聚为两类,使两类的均值之差最大,那么,我们可以按下式来衡量词汇的爆发特性:

$$d(w_i)=\frac{|m_2-m_1|}{\sigma_1+\sigma_2}, \tag{4}$$

其中 $m_1$,$m_2$ 为两类的均值,$\sigma_1$,$\sigma_1$ 为两类的方差.$d(w_i)$ 越大,$w_i$ 在不同语料之间分布的差别越大,越有可能是一个关键词;如果 $d(w_i)$ 很小,则说明 $w_i$ 在语料中分布较均匀,不是我们所需要的关键词.

我们对词表中的每一个词按上述方法逐一处理,从中选出了 3 000 个关键词.

### 1.2 隶属度函数的生成

由上一节可知,由于文章的领域信息主要包含在关键词的分布中,尤其是多个关键词在文章中的联合分布更包含了极为可靠的领域信息,因此,在计算训练语料关于某一领域的模糊集的隶属度函数时,我们以关键词的分布矢量作为特征.训练语料中的每一篇文章都被转化为一个关键词分布矢量,它的维数就是关键词的个数(在我们的系统中为 3 000 维),而每一个分量则代表了一个关键词在文章中出现的次数.

首先,我们按照传统方法对训练语料进行人工标注,这样,语料库中的所有文章都按其领域标注被划分到各个领域相关的子集.对每个领域的文章,以其关键词分布矢量的均值作为该领域的核,即该领域的核矢量:

$$Ker(Topic_j)=\frac{1}{n_j}*\sum_{article\in Topic_j} vector(article). \tag{5}$$

其中 $n_j$ 为人工标注领域 $j$ 中文章的数目,$vector(*)$ 表示文章的关键词分布矢量.

由于文章的主题信息存在于关键词分布矢量中,因此我们可以认为一篇文章与某个主题的关联程度取决于文章的关键词分布矢量与该主题的核矢量之间的相似程度,当两个矢量完全重合时,说明文章的内容完全符合该主题,若两个矢量垂直,说明文章的内容与该主题无关.某文章对于一个主题的隶属度值取决于文章的关键词分布矢量与该主题的核矢量之间夹角的余弦值.由此可得关于一个主题的模糊集的隶属度函数如下:

$$A_j(u_i)=f\left(\frac{vector(u_i)\cdot Ker(Topic_j)}{|vector(u_i)|*|Ker(Topic_j)|}\right). \tag{6}$$

由式(6)可得,一个主题的支集可表示为

$$SuppTopic_j=\left\{u\left|\frac{vector(u_i)\cdot Ker(Topic_j)}{|vector(u_i)|*|Ker(Topic_j)|}>0\right.\right\}. \tag{7}$$

在我们的系统中,函数 $f(*)$ 被设置为阶梯函数,假设我们将区间[0,1]分为 $m$ 级,由小到大分别为 $\{\lambda_1,\lambda_2,\ldots,\lambda_m\}$,令 $\Phi_j(u_i)=\frac{vector(u_i)\cdot Ker(Topic_j)}{|vector(u_i)|*|Ker(Topic_j)|}$,则主题 $j$ 的隶属度函数可表示为

$$A_j(u_i)=\left(\bigcup_{i=1,\ldots,m}\lambda_i\Phi_j\right)(u), \tag{8}$$

其中 $\lambda_i\Phi_j(u)=\lambda_i\wedge\Phi_j(u)$.

## 2 模型参数估计

从模糊集中估计模型参数的过程可以看作是一个清晰化的计算过程,普通训练集下的参数估计是采用最大似然准则,即

$$P(w_2|w_1)=\frac{Num(w_1w_2)}{Num(w_1)}. \tag{9}$$

在模糊训练集 $Topic_j$ 下,假设隶属度函数共有 $m$ 个取值 $\{\lambda_1,\lambda_2,\ldots,\lambda_m\}$(取决于我们所引入的阶梯函数),由分解定理得