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.

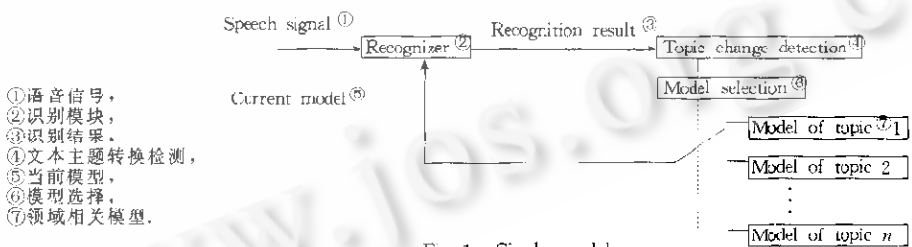

Fig. 1 Single model structure

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

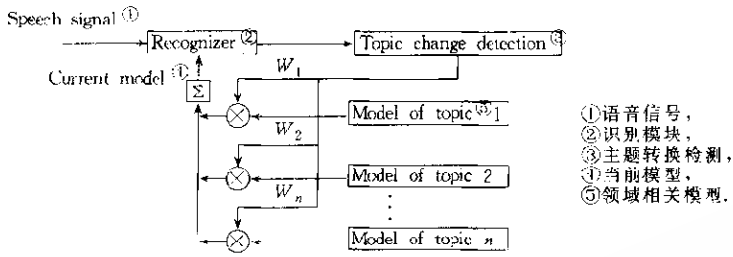

Fig. 2 Structure of mixture models
图2 混合模型工作方式

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,σ_1,σ_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

$$A_j(u)=\bigcup_{i=1,\cdots,m}(\lambda_i(A_j)_{\lambda_i}). \tag{10}$$

所以,我们可以重新将 $Topic_j$ 的支集划分如下:

$$SuppTopic_j=\bigcup_{i=1,\cdots,m}((A_j)_{\lambda_i}-(A_j)_{\lambda_{i+1}})=\bigcup_{i=1,\cdots,m}\Psi_i, \tag{11}$$

其中 $(A_j)_{\lambda_i}$ 为 $Topic_j$ 的 λ_i 截集.按上面的划分,每个子集 Ψ_i 中所包含的文章在 $Topic_j$ 中具有相同的隶属度.根据最大似然准则,用于集 Ψ_i 中的语料估计模型的参数,可得

$$P(w_2|w_1)=\frac{Num_i(w_1w_2)}{Num_i(w_1)}. \tag{12}$$

其中 $Num_i(*)$ 为事件"$*$"在子集 Ψ_i 中出现的次数.子集 Ψ_i 中所包含的文章在 $Topic_j$ 中的隶属度为 λ_i,因此我们也可以认为按式(12)估计的模型与领域 j 的相关模型的相似程度为 λ_i.因此,如果把从模糊集中估计模型参数的过程看作是一个清晰化计算过程,则领域 j 的相关模型按加权平均法估计为

$$P_j(w_2|w_1)=\sum_{i=1}^{m}\lambda_i*\frac{Num_i(w_1w_2)}{Num_i(w_1)}, \tag{13}$$

其中 λ_i 为归一化隶属度值.

按新方法构造的领域相关模型与传统模型相比,在以下几个方面具有明显的优势:

(1) 按传统方法训练领域相关模型只用到了那些被聚类到该领域语料的训练数据,而利用新方法,训练语料的范围被扩大到该领域模糊集的支集,训练数据大大增加,数据稀疏问题在很大程度上得到缓解.

(2) 训练数据中的每篇文章与各个领域都相互联系,联系强度取决与它在该领域的隶属度值,充分反映了一篇文章与多个领域有关的事实,而且避免了为一篇文章手工标注多个主题的繁重工作,也避免了这种人工标注所可能造成的不精确问题.

(3) 传统方法对语料库的标注是二值的,即一篇文章要么属于该领域,要么不属于该领域,而无法定量地表示文章与主题之间的联系强弱,而基于模糊训练集的方法用隶属度函数非常精确地描述了文章与主题之间的联系强弱,因而可以建立更加精确的模型.

3 自组织学习

基于模糊训练集的模型估计缓解了领域训练数据不足所带来的数据稀疏问题,同时精确地描述了文章与主题之间的耦合强度,但与此同时,不同主题相关模型之间的距离也随之缩小.由于传统方法中不同主题的训练之间彼此没有重合,因此,虽然数据稀疏问题比较严重,但不同模型之间区分度较大.我们希望基于模糊训练集的领域相关模型在拥有上一节所描述的优点的同时,不同模型之间依然能够保持较大的距离.为达到这一目的,我们在上一节参数估计的基础上引入了一个自组织学习[4]的过程.如图3所示的单层网络,我们把每个领域相关的模型看作是一种模式,以二元文法为例,我们把训练语料中的每一个句子转化为一个输入矢量,输入矢量的维数为 $N\times N$,其中 N 为词表大小,输入矢量的每一分量为相应词对在句子中出现的次数,而网络的初始权重设定为基于模糊训练集的参数估计所得到的模型参数取对数,因此,网络的输出为该句子的对数似然函数,如式(14)所示.

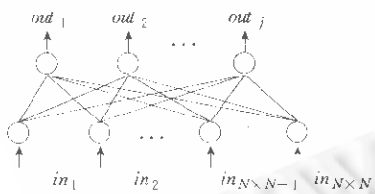

Fig. 3 Self organized learning
图3 自组织学习示意图

$$out_j=\sum_{i=1}^{N\times N}in_i*w_{ji}=\sum_{w_1,w_2}Num(w_1w_2)*\log P(w_2|w_1). \tag{14}$$

当训练语料依次输入到网络时,我们将输出按大小排序,与输出最大的节点相联系的权值将得到增强,与此同时,对排序较后即输出很小的那些节点,我们认为输入的训练数据与这些节点对应的主题关联很小,因此与这些输出相关的权重要削弱.在我们的系统中,对每个主题所对应的输出节点 out_j 都保留有两个集合,$Succ\ set_j$ 存放着使该节点输出最大的训练数据,$Fail\ set_j$ 存放的是该节点输出很小的训练数据.具体算法如下:

(1) 用基于模糊训练集的参数估计所得的模型参数初始化网络,$t=0$;

(2) 对所有的训练语料逐句处理:

(a) 将训练句子输入网络,按式(14)计算其输出.

(b) 对所有输出节点作如下处理:

如果当前节点 j 在竞争中胜出,将训练句子并入 $Succ\_set_j$ 集合.

如果当前节点 j 输出排序较后,将训练句子并入 $Fail\_set_j$ 集合;

(3) 经过步骤(2)的处理,每个主题输出都保留了使其排序靠前和靠后的训练数据.我们将利用这些数据更新模型参数.对所有节点作如下循环:

(a) 利用当前输出节点 out_j 的 $Succ\_set_j$ 集合中的训练数据训练模型,以 $P'(w_2|w_1)$ 更新领域 j 模型:

$$P(w_2|w_1)+=\eta(t)*P'(w_2|w_1). \tag{15}$$

其中 $\eta(t)$ 为学习常数随时间单调下降.

(b) 利用当前输出节点 out_j 的 $Fail\_set_j$ 集合中的训练数据训练模型,以 $P''(w_2|w_1)$ 更新领域 j 模型:

$$P(w_2|w_1)-=\alpha(t)*P''(w_2|w_1). \tag{16}$$

(c) 模型参数归一化,按新模型更新网络权值.

(4) $t++$;若满足结束条件,转(5);否则,转(2);

(5) 结束.

Kullback-Liebler 距离是一种分布之间非对称的距离度量方式,如式(17)所示.

$$D(Q(w_2|w_1),P(w_2|w_1))=\sum_{w_1,w_2}Q(w_2|w_1)\log\frac{Q(w_2|w_1)}{P(w_2|w_1)} \tag{17}$$

本文按式(18)获得对称的距离度量方式,检验自组织学习的效果.

$$\text{Distance}(Q,P)=D(Q(w_2|w_1),P(w_2|w_1))+D(P(w_2|w_1),Q(w_2|w_1)) \tag{18}$$

实验证明,经过自组织学习以后,模型间的距离明显增大.

4 实验及结果

我们的实验以《人民日报》的2万多篇文章,约1 000万词的语料作为训练数据.对训练语料的领域信息,我们首先进行了手工标注,结果见表1.

Table 1 Corpus information

表1 语料信息

The topic of corpus①	Size (word)②
International news③	1 059 795
National news④	1 503 522
Economy⑤	1 257 723
Sport⑥	265 718
Society⑦	1 324 261
Sciences and technology⑧	118 969

①语料主题,②词数,③国际新闻,④国内新闻,⑤经济,⑥体育,⑦人文,⑧科学技术.

我们的实验按如图1所示的单模型结构和如图2所示的混合模型结构两种方式进行.首先,我们收集有关体育方面的语料约30 000词作为测试语料,测试语料与训练语料无重合,然后利用篇章聚类所得主题为体育的训练语料按传统方法训练以体育为主题的二元文法模型,然后按照本章所提出的方法生成有关体育的模糊训练集,生成二元文法模型.两种模型用同样的语料加以测试,结果见表2.我们可以看到,基于模糊训练集的方法其性能远远好于传统的领域相关模型.这种优越性主要来源于基于模糊训练集的方法充分运用了一篇文章与多个主题相联系的特性,极大地缓解了数据稀疏问题.传统方法在领域相关语料较少时,通常采用领域自适应的方法来解决数据稀疏问题,即利用较少的领域相关语料来对领域无关的普通模型作调整,使之对特殊领域有较好的性能.最具代表性的领域自适应方法是基于MAP(maximum a posteriori)原则[7]的方法,我们对这种方法也进行了实验,结果见表2.可以看到,基于模糊训练集的方法同样也优于基于MAP的领域自适应方法.这是因为新的方法对训练数据与领域之间的相关性有非常精确的定量描述,而MAP方法对训练数据与领域关系的描述是二

值的.换句话说,基于模糊训练集的方法比基于MAP的方法更加精确.

Table 2 The experiment of single model

表2 单模型结构实验

Model①	Perplexity②
Trained by traditional method③	165.8
MAP topic adaptation④	126.7
The method based on fuzzy traing subset⑤	121.1

①模型,②困惑度,③传统方法训练的模型,④基于MAP的领域自适应,⑤基于模糊训练集的方法.

我们的另一个实验是针对混合模型结构进行的.按照领域聚类所得到的主题,依照传统方法,生成如图2所示的混合模型结构.与此同时,又建立了上述主题的6个模糊训练集,按本文提出的新方法训练模型,同样也生成混合模型结构.我们的测试数据为选自《人民日报》的200 000词的语料,内容包括政治、经济、文化、体育、科学等各个方面,测试数据与训练数据无重叠,实验结果见表3.可以看到,在混合模型结构下,传统方法由于数据稀疏问题得到缓解,因此其性能也得到了极大的改善,但基于模糊训练集的混合模型依然优于传统方法.这是因为新方法在利用模糊训练集较精确地反映了文章与主题之间的相关程度的同时,又通过自组织学习而有效地拉开了不同模型之间的距离.

Table 3 The experiment of mixture models

表3 混合模型实验

Model①	Perplexity②
Traditional mixture models③	122.8
Mixture models based on fuzzy training subset④	112.3

①模型,②困惑度,③传统混合模型,④基于模糊训练集的混合模型.

第3个实验是对自组织学习的专门分析.首先我们比较了引入自组织学习和不引入自组织学习在性能上的差异,训练和测试数据如实验2,结果见表4.可以看出,未引入自组织学习以前虽然引入了对训练数据与主题之间的耦合强度信息,但由于同一时间不同领域相关模型之间的重合程度加大,因此,基于模糊训练集的方法的优势主要体现在,虽然在单模型结构中,但对于混合模型结构,其性能与传统方法相当.但是在引入自组织学习以后,模型性能有了大幅度的提高.

Table 4 The comparison of result before and after self organized learning

表4 自组织学习性能比较

Model①	Perplexity②
Before self organized learning③	122.1
After self organized learning④	112.3

①模型,②困惑度,③未引入自组织学习的混合模型,④引入自组织学习的混合模型.

Table 5 Distance between models before self organized learning

表5 自组织学习前模型间距离

Distance between models①	Model of topic② 1	Model of topic 2	Model of topic 3	Model of topic 4	Model of topic 5	Model of topic 6
Model of topic 1	0.0	12.2	7.9	10.3	17.5	11.2
Model of topic 2		0.0	14.1	12.2	10.5	12.4
Model of topic 3			0.0	8.2	10.3	13.2
Model of topic 4				0.0	8.6	11.3
Model of topic 5					0.0	9.3
Model of topic 6						0.0

①模型间距离,②领域模型.

为进一步探讨自组织学习对模型性能产生重大影响的原因,我们按式(18)对自组织学习前、后的领域相关模型之间的距离进行了度量,结果见表5和表6.由表5和表6可以看出,自组织学习成功地拉开了模型间的距

离,模型在充分反映了训练数据与主题的耦合关系的同时又充分拉开了彼此间的距离,因此性能得到较大的改善.

Table 6 Distance between models after organized learning

表 6 自组织学习后模型间距离

Distance between models①	Model of topic② 1	Model of topic 2	Model of topic 3	Model of topic 4	Model of topic 5	Model of topic 6
Model of topic 1	0.0	13.7	9.7	16.9	18.6	13.4
Model of topic 2		0.0	17.8	15.5	13.1	14.6
Model of topic 3			0.0	10.5	13.2	15.5
Model of topic 4				0.0	11.3	13.9
Model of topic 5					0.0	13.1
Model of topic 6						0.0

①模型间距离,②领域模型.

5 结束语

领域相关的统计语言模型与普通的 n-gram 模型相比,能够更好地把握说话内容的领域信息,但传统的建立领域相关模型的方法是基于人工标记语料和对各领域分别进行最大似然训练的.传统方法不能很好地反映训练数据与多主题的联系,不能高效地运用训练语料.本文提出了一种新的训练语料设计和参数估计方法.该方法将不同领域的训练数据表示为整个训练语料的模糊子集,而参数的估计也基于不同主题的模糊训练集.新方法在很大程度上缓解了训练语料不足所带来的数据稀疏问题,而且对训练数据与主题之间的联系给出了精确的定量描述,同时,本文还将自组织学习引入模型的训练过程,通过自组织学习,不同领域模型间的距离被拉大,预测能力明显加强.按本文提出的新方法训练的领域相关语言模型,无论是在单模型结构下,还是在混合模型结构下,性能均比传统方法有较大的改善.

参考文献

1 Jelinek F. Self-Organized Language Model for Speech Recognition. Readings in Speech Recognition. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1990

2 Lin Sung-chien, Lee Lin-shan. Chinese language model adaptation based on document classification and multiple domain-specific language models. In: Kokkinakis G, Fakotakis N, Dermatas E eds. Proceedings of European Conference of Speech Communication and Technology. Greece, European Speech Communication Association. 1997. 1463~1466

3 Clarkson P R, Robinson A J. Language model adaptation using mixtures and an exponentially decaying cache. In: Pango P A ed. Proceedings of the International Conference of Acoustics Speech and Signal Processing. Munich: IEEE Signal Processing Society, 1997. 799~802

4 Chen Lang-zhou, Huang Tai-yi. A new method for text segmenting based on neural network. In: Huang Chang-ning ed. Proceedings of the International Conference on Chinese Information Processing. Beijing: Tsinghua University Press, 1998. 125~129
(陈浪舟,黄泰翼. 一种基于神经网络的文本切分算法.见:黄昌宁编.中文信息处理国际会议论文集.北京:清华大学出版社,1998. 125~129)

5 Kneser R, Steinbiss V. On the dynamic adaptation of stochastic language modeling. In: Proceedings of the International Conference of Acoustics Speech and Signal Processing. Minneapolis: IEEE Signal Processing Society, 1993. 586~589

6 Huang De-shuang. Neural Network and Pattern Recognition System Theory. Beijing: Publishing House of Electronics Industry, 1996
(黄德双.神经网络模式识别理论.北京:电子工业出版社,1996)

7 Federico M. Bayesian estimation methods for n-gram language model adaptation. In: Bunnell T H ed. Proceedings of 1996 International Conference of Spoken Language Processing. Philadelphia: Press of University of Delaware, 1996. 240~243

Domain Dependent Language Model Based on Fuzzy Training Subset

CHEN Lang-zhou　HUANG Tai-yi

(*Institute of Automation　The Chinese Academy of Sciences　Beijing* 100080)

Abstract　Statistical language model is very important to speech recognition. To a system of special topic, domain dependent language model is much better than the general model. There are two problems in traditional method. (1) The corpus of special topic is not large enough as general corpus. (2) An article is always related to more than one topic, but these phenomena have not been considered during the process of model training. In this paper, the authors try to solve these two problems. They present a new method to organize the corpus——the method based on fuzzy training subset. And the training of domain dependent models is based on these fuzzy subsets. At the same time, self organized learning has been introduced in training process to improve the models' prediction ability. It can improve the performance of models evidently.

Key words　Speech recognition, statistical language model, fuzzy, self organized learning.