[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G10L 15/14
G10L 15/06

[12] 发明专利申请公开说明书

[21] 申请号 200410059511.7

[43] 公开日 2005 年 2 月 2 日

[11] 公开号 CN 1573926A

[22] 申请日 2004.6.3
[21] 申请号 200410059511.7
[30] 优先权
[32] 2003. 6. 3 [33] US [31] 10/453,349
[71] 申请人 微软公司
地址 美国华盛顿州
[72] 发明人 C·切尔巴 A·阿塞罗
M·马哈间

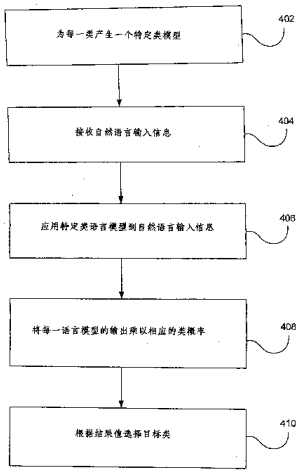
[74] 专利代理机构 上海专利商标事务所
代理人 李家麟

权利要求书 3 页 说明书 13 页 附图 5 页

[54] 发明名称 用于文本和语音分类的区别性语言模型训练

[57] 摘要

本发明公开了一种估计语言模型的方法，使得给定字串的类的条件似然最大化，其中该字串与分类精度非常相关。 该方法包括对所有类联合的调节统计语言模型参数，使得对于给定的训练句子或发声，分类器区分出正确的类和不正确的类。 本发明的特定实施例用于实现用于 n－字符列分类器的区别性训练方法的上下文中的有理函数增长变换。



ISSN1008-4274

1. 一种计算机实现的方法，其用于为对应于多个类的多个语言模型中的每一个估计一组参数，该方法包括：

为参数组设定初始值；和

联合的相对于另一个参数组调节参数组，以提高给定字串的类的条件似然。

2. 如权利要求1所述的方法，其中多个语言模型是多个n字符列语言模型。

3. 如权利要求2所述的方法，其中最大化条件似然。

4. 如权利要求3所述的方法，其中使用有理函数增长变换应用程序将条件似然最大化。

5. 如权利要求2所述的方法，其中字串是训练材料中的文本串。

6. 如权利要求5所述的方法，其中字串是从训练句子中导出的文本串。

7. 如权利要求5所述的方法，其中字串是从语音发声中导出的文本串。

8. 如权利要求2所述的方法，其中联合的调节参数组包括对于给定训练输入调节参数组以区分正确的类和不正确的类。

9. 如权利要求2所述的方法，其中调节参数组进一步包括训练参数组以适应未见过的数据。

10. 如权利要求9所述的方法，其中适应未见过的数据包括基于平滑调节参数。

11. 一种计算机实现的方法，其用来为多个n-字符列语言模型中的每一个估计一组参数，该方法包括：

联合的相对于另一个参数组产生至少两个参数组。

12. 如权利要求11所述的方法，其中联合的相对于另一个参数组产生至少两个参数组包括联合的相对于另一个参数组产生所有参数组。

13. 如权利要求11所述的方法，其中多个n-字符列语言模型中的每一个与一个类相关，并且其中联合的产生至少两个参数组包括产生至少两个参数组，使得第一特定类与特定字串关联的似然性会增加，并使得第二类与特定字串关联的似然性会减小。

14. 如权利要求13所述的方法，其中特定字串从训练句子中导出。

15. 如权利要求13所述的方法，其中特定字串从语音发声中导出。

16. 如权利要求11所述的方法，其中联合的产生至少两个参数组包括联合的产生多个参数组，使得对于给定训练输入n-字符列语言模型可以区分正确的类和不正确的类。

17. 如权利要求11所述的方法，进一步包括训练参数组以适应未见过的数据。

18. 一种计算机实现的对自然语言输入进行分类的方法，包括：

联合的相对于另一个组件训练多个统计分类组件，其对应于多个类，以增加给定字串的类的条件似然，多个统计分类组件为n-字符列语言模型分类器；

接收自然语言输入；

将该多个统计分类组件应用到自然语言输入，以将该自然语言输入划分到多个类中的一个中。

19. 如权利要求18所述的方法，其中最大化条件似然。

20. 如权利要求19所述的方法，其中使用有理函数增长变换应用程序将条件似然最大化。

21. 如权利要求20所述的方法，其中训练多个统计分类组件包括：

标识有理函数增长变换迭代的最优的次数以简化有理函数增长变换应用。

22. 如权利要求21所述的方法，其中标识有理函数增长变换迭代的最优的次数、和最优的CML加权β_{max}包括：

将训练数据集分割成主数据集和支持数据集；

使用主数据为该统计分类组件估计一系列相关频率；和

使用支持数据集调节有理函数增长变换迭代的最优的次数和最优的CML加权β_{max}。

23. 如权利要求22所述的方法，其中使用支持数据调节包括：

选定将要运行的有理函数增长变换迭代的预先确定的次数N；

选定为确定最优的CML加权β_{max}而研究的值的范围；

对于每一值β_{max}，运行尽可能多的RFGT迭代，多达N次，使得该主数据的条件似然在每次迭代都增加；和

把有理函数增长变换的迭代次数和使得该支持数据的条件似然最大化的β_{max}值标识为最优。

24. 如权利要求23所述的方法，其中训练多个统计分类组件进一步包括：

集中该主数据和支持数据以形成训练数据集的组合；和

使用该有理函数增长变换迭代的最优次数和最优的CML加权β_{max}，对训练数据的组合集合训练所述多个统计分类组件。

25. 如权利要求24所述的方法，其中多个统计分类组件是n-字符列语言模型。

用于文本和语音分类的区别性语言模型训练

发明背景

本发明涉及文本和语音分类。更具体的，本发明涉及语言模型的增强，以提高分类的精度。

自然语言理解包括使用计算机来确定用户产生的文本或者语音的意义。在确定输入自然语言意义中，其中一个步骤是将该输入划分到一组预定类中的一个类。例如，一特定输入诸如“I want to book a flight to Rome”可以被划分到旅行安排类。然后可以调用一个用于此类的应用，以从该输入中进一步解释信息并执行该输入所表示的用户目的。

这种分类在自然语言处理中是定义明确的问题。实际应用的特定范例包括对自动呼叫中心的呼叫选择路由和基于帮助系统的自然语言。

分类器可以被用来简化该分类处理。分类器的普通范例包括统计分类器，诸如n－字符列、Naive Bayes和最大熵值分类器。在n－字符列分类器中，统计语言模型被用来将自然语言字串（即句子）分配到类。具体的，分离的n－字符列语言模型是为每一类而构建的。在运行时，并行使用该语言模型以将概率分配到给定测试字串或语音发声。对该测试字串或发声表现出最高概率的与语言模型相关的类被指定为该串/发声所属的类。类分配不必是一对一的。对于给定的测试串或发声，根据每一类接收到的概率，可以将该测试句或发声分配到一组N－best候选类中。对于语音分类，n－字符列分类器具有的优势是，它们可以在单通场景中使用，其中集成了语音发声识别和分类。

一种训练n－字符列分类器的简单的方法是使用最大似然（ML）估计分别为每一类训练语言模型。虽然这种训练方案容易实现，但是它们产生的分类器精度有限。

发明概述

本发明的实施例适合于一种训练语言模型的方法，其使得给定字串的类的条件似然最大化，其中该给定字串的类的条件似然与分类精度非常相关。该方

法包括为所有类共同调节统计语言模型参数，使得对于给定的训练句子或发声，分类器区分出正确的类和不正确的类。本发明的特定实施例适合于实现用于n一字符列分类器的区别性训练方法的文本中有理函数增长变换。

附图简述

图1中的方框图为在其中可以使用本发明的一个说明性环境。

图2中的方框图为自然语言分类系统的一部分。

图3中的方框图为另一自然语言分类系统的一部分。

图4中的流程图为与任务或分类鉴别相关的步骤。

图5中的流程图为与训练类别特定语言模型相关的步骤。

说明性实施例的详细描述

运行环境示例

本发明的各方面适合于语音模型的最大条件似然（ML）估计，其用于文本和语音发声分类。然而在更具体讨论本发明之前，首先要讨论实现本发明的示例环境的一个实施例。

图1所示是在其上能够实现本发明的合适的计算系统环境100的示例。该计算系统环境100只是一个合适的计算环境示例，并不是对本发明的使用或功能性范围做任何限制。也不应该认为该计算环境100具有对该示范操作环境100中所述的组件中的任何一个或其组合相关的依赖或需求。

该发明可以运行于多个其它通用或专用的计算系统环境或配置。所熟知的可以用于本发明的计算系统、环境、和/或配置范例包括：个人计算机、服务器计算机、手持或膝上器件、多处理器系统、基于微处理器的系统、机顶盒、可编程消费电子产品、网络PC、微电脑、大型计算机、电话系统、包括任何上述系统或器件的分布式计算环境，等等，但不限于此。

本发明可以在可执行计算机指令的一般环境中描述，诸如由计算机执行的程序模块。一般说来，程序模块包括例行程序、程序、对象、组件、数据结构等执行特定任务或实现特定抽象数据类型。设计本发明在分布计算环境中实现，其中由通过通信网络连接的远程处理装置来执行任务。在分布计算环境中，程序模块位于包括记忆存储装置的本地和远程计算机存储器媒体中。下面借助于附图描述由程序和模块执行的任务。本领域的熟练技术人员可以如处理器可执行指令那样实现该说明和附图，其中该指令可以记录在任何形式的计算机可读

媒体上。

参照图1，用于实现本发明的范例系统包括通用计算装置，其形式为计算机110。计算机110的组件可以包括处理单元120、系统存储器130、和将包括系统存储器的各种系统组件耦合到处理单元120的系统总线121，但并不限于此。该系统总线121可以是任何几个类型的总线结构，包括存储器总线或存储器控制器，周边总线、和使用任何多个总线结构的本地总线。作为范例，但并不限于此，这种结构包括工业标准结构（ISA）总线、微通道结构（MCA）总线、增强ISA（EISA）总线、视频电子标准协会（VESA）局部总线和也被称为Mezzanine总线的周边组件互联（PCI）总线。

计算机110典型的包括多个计算机可读媒体。计算机可读媒体可以是任何可以被计算机110访问的有效媒体，包括易失性和非易失性媒体，可拆卸和非可拆卸媒体。作为范例，但并不限于此，计算机可读媒体可以包括计算机存储媒体和通信媒体。计算机存储媒体包括可以用任何方法或技术存储诸如计算机可读指令、数据结构、程序模块或其它数据的信息的易失性和非易失性媒体。计算机存储媒体包括RAM、ROM、EEPROM、闪存或其它存储技术、CD-ROM、数字通用盘（DVD）或其它光盘存储器、磁盘、磁带、磁盘存储器或其它磁存储装置、或任何其它可以用来存储所想要的信息并可以被计算机110访问的媒体，但并不限于此。

通信媒体典型的包含在已调制的数据信号中的计算机可读指令、数据结构、程序模块或其它数据，该已调制的数据信号诸如载波或其它传输机制，并包括任何信息传送媒体。术语“已调制的数据信号”的意思是具有一个或多个特征集的信号，或者是按照在该信号中的编码信息的方式改变的信号。作为范例，但并不限于此，通信媒体包括诸如有线网络、直接有线连接的有线媒体，和诸如声音、RF、红外线和其它无线媒体的无线媒体。上述任何的组合也应该包括在计算机可读媒体的范围内。

系统存储器130包括易失性和/或非易失性存储器形式的计算机存储媒体，诸如只读存储器（ROM）131和随机访问存储器（RAM）132。ROM131中典型的存储有基本输入/输出系统（BIOS）133，其包含诸如在启动时帮助在计算机110中的组件之间传送信息的基本例行程序。RAM132典型的包含立即可以访问的和/或当前正在被处理单元120运行的数据和/或程序模块。作为范例，但并不限

于此，图1示出了操作系统134、应用程序135、其它程序模块136和程序数据137。

计算机110也可以包括其它可拆卸/非可拆卸的易失性和/或非易失性的计算机存储媒体。仅作为范例，图1示出了从非可拆卸的、非易失性磁性媒体中读出或写入其中的硬盘驱动141，从可拆卸的、非易失性磁盘152读出或写入其中的磁盘驱动151，和从可拆卸的、非易失性光盘156、诸如CDROM或其它光学媒体中读出或写入其中的光盘驱动155。其它可以用在该示例操作环境中的可拆卸/非可拆卸的、易失性和/或非易失性的计算机存储媒体包括磁带盒、闪存卡、数字通用盘、数字视频带、固态RAM、固态ROM等等，但并不限于此。硬盘驱动141典型的通过诸如接口140的非可拆卸存储器接口连接到系统总线121，磁盘驱动151和光盘驱动155典型的通过诸如接口150的可拆卸存储器接口连接到系统总线121。

上述讨论的和图1中示出的驱动以及它们相关的计算机存储媒体为计算机110提供计算机可读指令、数据结构、程序模块和其它数据的存储。例如在图1中，所示硬盘驱动141用作存储操作系统144、应用程序145、其它程序模块146和程序数据147。注意到这些组件可以与操作系统134、应用程序135、其它程序模块136和程序数据137相同，或者与其不同。这里对操作系统144、应用程序145、其它程序模块146和程序数据147给出不同的编号，以说明至少它们是不同的拷贝。

用户可以通过输入装置输入命令或信息到计算机110中，该输入装置诸如键盘162、麦克风163、指点设备161、如鼠标、轨迹球或触摸板。其它输入装置（未示出）可以包括游戏杆、游戏板、圆盘式卫星电视天线、扫描仪等。这些或其它输入装置通常通过耦合到系统总线的用户输入接口160连接到处理单元120，但是也可以通过其它接口和总线结构，诸如并行端口、游戏端口或通用串行总线（USB）连接。监视器191或其它类型的显示装置通过接口，诸如视频接口190也连接到系统总线121。除了监视器之外，计算机也可以包括其它外围输出装置，诸如可以通过输出外围接口195连接的扬声器197和打印机196。

计算机110运行在网络环境中，该计算机110使用逻辑连接到一个或多个远程计算机，诸如远程计算机180。远程计算机180可以是个人电脑、手持器件、服务器、路由器、网络PC、对等设备或其它普通网络节点，并典型的包括多个或所有上述与计算机110相关的组件。图1中所描述的逻辑连接包括局域网（LAN）

171和广域网（WAN）173，但也可以包括其它网络。这种网络环境在办公室、企业级计算机网络、内部网络和因特网中很普通。

当在LAN网络环境中使用时，计算机110通过网络接口和适配器170连接到LAN171。当在WAN网络环境中使用时，计算机110典型的包括调制解调器172或其它用于在WAN173，诸如因特网上建立连接的装置。调制解调器172可以是内置的或外置的，其可以通过用户输入接口160、或者其它适当的机制连接到系统总线121。在网络环境中，所述与计算机110相关的程序模块、或其中的部分可以存储在远程记忆存储装置中。作为范例，但并不限于此，图1示出了驻留在远程计算机180中的远程应用程序185。应该理解的是，所示的网络连接是示范性的，并且也可以使用其它在计算机之间建立通信连接的装置。

注意到本发明可以在诸如关于图1所描述的计算机系统中执行。然而本发明可以在服务器、用于信息传递的计算机上执行，或在分布式系统上执行，其中本发明的不同部分在该分布式计算系统的不同部分上执行。

任务分类系统总述

图2所示的方框图为自然语言分类系统200的一部分。系统200包括统计分类器204。系统200也可选择的包括语音识别引擎206。其中接口200接收语音信号作为输入，其包括该识别器206。然而在接口200接收文本输入的地方不需要识别器206。本讨论将按照存在有识别器206的实施例进行，但是需要理解的是它在其它实施例中并不必须。同样可以使用其它自然语言通信模式，诸如手写或其它模式。在这种情况下，使用适当的识别组件，诸如手写识别组件。

为了执行类或任务分类，系统200首先接收语音信号形式的发声208，该语音信号表示用户所说的自然语言语音。语音识别器206对发声208执行语音识别，并在其输出提供自然语言文本210。文本210是语音识别器206接收到的自然语言发声208的文本表示。语音识别器206可以是任何已知的对语音输入执行语音识别的语音识别系统。语音识别器206可以包括特定应用听写语言模式，但是语音识别器206识别语音的特定方式并不构成本发明的主体。相似的，在另一个实施例中，语音识别器206输出具有各自的概率的结果或解释列表。新的组件对每一解释操作，并在类或任务分类中使用该相关的概率。

根据一个实施例，将自然语言文本210的全部或部分提供到统计分类器204以用于分析和分类。在使用预处理从文本210中删除某些成分（即冠词a、an、the

等等）的情况下，可以提供部分。根据另一个实施例，将自然语言文本210的不太直接的表示（即向量表示）提供到统计分类器204，以用于分析和分类。

根据一个实施例，从该自然语言文本中提取一组特征，以提供到统计分类器204。该组特征说明性的作为最有助于执行任务分类的那些特征。或相反，这可以根据经验确定。

在一个实施例中，该提取的特征是一组字标识符，其标识词存在或不存在于该自然语言输入文本210中。例如，只有在为特定应用而设计的某一词汇表中的词可以被标记以使分类器204考虑，该词汇表之外的词被映射为奇异的词类型，诸如“未知”。

应该注意到，也可以选择更加复杂类型的特征用于考虑。例如词的共存可以是被选择的特征。例如为了更加清楚的标识要执行的任务，可以使用它。例如，词“send mail”的共存可以是被统计分类器204标记为对处理特别重要的特征。如果在该输入文本中发现这个顺序的这两个词，然后统计分类器204将会收到这一事实的通告。也可以选择其它非常多的特征，诸如双-字符列（bi-gram）、三-字符列（tri-gram）、其它n-字符列（n-gram）、以及任何其它理想的特征。

在特征提取处理之前，能够可选择的对自然语言文本210执行预处理，以简化对文本210的处理。例如理想的是，自然语言文本210只包括存在或不存在已经被预定带有某种类型内容的词的表示。

在特征提取之前也可以进行取词干。取词干是删除词中的形态变异以得到它们的词根形式的处理过程。形态变异的示例包括词尾变化（如复数、动词时态等）和改变词的语法作用的词源变化（如形容词到副词，slow到slowly的变化等）。取词干可以被用来将具有相同基本语义的多个特征精简为单个特征。这样可以帮助克服数据稀疏的问题，增强计算效率，并减小在统计分类方法中所使用的特征独立假定的影响。

在任何情况下，统计分类器204接收该自然语言文本210信息，并使用统计分类组件207来将该信息划分到多个预定类或任务的一个和多个中。组件207说明性的为多个统计语言模型中的任何一个，诸如与n-字符列（n-gram）、Naive Bayes或最大熵值分类器相关的模型。训练器209说明性的使用训练数据205和测试数据213的集合来训练该语言模型。测试数据213说明性的为有限量的数据，它抑制训练数据205用于简化该训练过程。

下面将更详细地说明分类器204执行任务或类标识的原理。为了固定标号，此后用A表示语音发声。引起发声A的字串表示为$W=w_1\ldots w_n$。发声A的类表示为C（A）。词的词汇表表示为v，类的词汇表表示为C。划分成分别为T和ε的训练数据和测试数据的文集由字节组（tuples）（或者样本）s组成，该字节组包含发声A、音标表示W和发声类C（A）。作为参考，对于给定的分类器，通过类误差率（CER）来检测其性能：

$$CER=\sum_{s\in\varepsilon}\delta\left(C(s.A),\hat{C}(s.A)\right) \qquad (1)$$

其中s.A表示样本s中的该发声，δ（.,.）是Kronecker-δ算子，当它的幅角彼此相等时，它等于1，否则它等于0。

在n-字符列语言模型分类器的文本中，假定双通场景，通过集中标记为类C的音标，为每一类C∈C构建n-字符列模型P（$w_i\mid w_{i-1}$，…，w_{i-n+1}，C）。除了该类特定语言模型P（·｜C）之外，从所有的训练音标中构建集中的n-字符列语言模型P（$w_i\mid w_{i-1}$，w_{i-n+1}）。然后通过使用该集中的语言模型，对1-best识别输出进行文本分类，每一测试发声被分配到类：

$$\text{最可能的类}=\hat{C}(A)=\arg\max_{C}\log P\left(\hat{W}|C\right)+\log P(C) \qquad (2)$$

$$\text{最可能的词序}=\hat{W}=\arg\max_{W}\log P\left(A|W\right)+\log P(W) \qquad (3)$$

这是双通方式，其中等式2的第一步骤由语音识别器206执行，等式2的第二步骤由统计分类器204执行。该双通方法的第二阶段说明性的实现n-字符列文本分类器。

该n-字符列型分类器具有特别可能的效率优势就是，它可以被用于单通系统，其中将给定的语音发声在语音识别的同时分配到库。另外，值得一提的是n-字符列型分类器的优势是，它能够允许具有相对高顺序的词共存的考虑。例如，对三-字符列（tri-grams）的考虑包括词的三元组（triplets）的检查。即使只考虑单-字符列（uni-grams），词的共存数目将在n-字符列分类器中考虑。

图3所示为单通分类系统220的一部分。图3中与图2中的组件功能相同或相似的组件标以相同或相似的编号。

系统220（图3）与系统200（图2）不同的地方在于它包括统计分类器/解码

器211。分类器/解码器211说明性的按单通方式解码和分类发声208。分类器/解码器211包括由训练器209训练的统计分类组件207（即语言模型）。组件207说明性的为用于单通系统构建的n-字符列语言模型，借此给定发声A在进行查找字串语音解码的同时，被分配到类 （A）。

根据图3中的单通系统，通过将所有标记为类C的训练音标集中，为每一类C∈C构建n-字符列模型P（$w_i \mid w_{i-1}$，…，w_{i-n+1}，C）。通过并行堆积为每一类带有相关标记的每一语言模型P（• | C）来构建识别网络。到每一语言模型P（• | C）的转换具有得分log P（C）。在单通中，如下标识最有可能的路径：

$$\left(\hat{C}(A),\hat{W}\right)= \arg\max_{(C,W)} \log P(A|W)+\log P(W|C)+\log P(C) \qquad (4)$$

因此，在接收到发声208时，分类器/解码器211在语音解码的同时能够分配类。被识别为最有可能的字串将会具有返回该串的类标签。该输出是任务或类ID214和解码串215。解码串215是说明性的表示发声208的字串。

在单通和双通系统的文本中，n-字符列语言模型可以被平滑，以适应未见过的训练数据。根据一个实施例，利用在不同的阶（诸如对于统一模型用0，对于n-字符列模型用n-1）的相关频率估计的线性插值，来估计类特定训练模型的n-字符列概率。根据上下文计数将不同阶的线性插值加权存储，并对交叉有效性数据使用最大概似法技术估计它们的值。然后将来自交叉有效性数据的n-字符列计数加到从主训练数据收集到的计数中，以增强相对频率估计的质量。在Jelinek and Mercer, Interpolated Estimation of Markov Source Parameters From Sparse Data, Pattern Recognition in Practice, Gelsema and Kanal editors, North-Holland (1980)中对这种平滑有更详细的陈述。这是平滑的一个示例。

有其它的方式定义平滑处理。根据另一个平滑实施例，为了估计n-字符列语言模型，以不同阶$f_k(\bullet)$，k=0…n，在相对频率估计之间使用回归删除插值：

$$P_n(w|h_n)=\lambda(h_n)\cdot P_{n-1}(w|h_{n-1})+\overline{\lambda(h_n)}\cdot f_n(w|h_n),$$
$$P_{-1}(w)=uniform(v) \qquad (5)$$

其中

$$\overline{\lambda(h_n)} = 1 - \lambda(h_n)$$

$$h_n = w_{-1}, \ldots, w_{-n}$$

分别从主数据和提供的数据中，利用最大似然法估计得到相对频率f_n（w/h_n）和插值加权λ（h_n），并且通过对给定语言模型有效的训练数据的70/30%的随机划分获得。根据上下文计数对插值加权进行存储：

$$\lambda(h_n) = \lambda(C_{ML}(h_n)) \qquad (6)$$

其中C_{ML}（h_n）为文本h_n在分配到该类的训练数据中出现的次数。

通常说来，配置分类器204和211以输出任务或类标识符214，其标识分配到对应的输入自然语言的特定任务或者类。标识符214可替换为任务或类标识符的分级表（或n-best表）。将标识符214提供给可以根据该标识的任务进行动作的应用程序或其它组件。例如，如果标识的任务是发送邮件，标识符214就被送到电子邮件应用程序，盖电子邮件应用程序能够依次显示用户所使用的电子邮件模板。当然，也可以考虑其它任何任务或类。相似的，如果输出标识符214的n-best表，表中的每一条目可以通过适当的用户接口显示，从而用户可以选择所需要的类或任务。注意到也将解码串215典型的提供到该应用程序。

任务或库标识

根据本发明的一个实施例，图4中的流程图为与任务或类标识相关的方法的步骤。应该注意到，该方法可以为单通或双通系统定制，如上参照图2和3所述。分类器采用分类组件207，其说明性的为一组独立于类的n-字符列统计语言模型分类器。根据块402，为每一类或任务产生一个类特定模型。因此当接收到自然语言输入210时（块404），为每一类在该自然语言输入信息上运行类特定语言模型（块406）。每一语言模型的输出乘以适于相应的类的先验概率（块408）。具有最高结果值的类说明性的对应于目标类（块410）。

根据一个实施例，图5中的流程图为与类特定语言模型的训练相关的的步骤。这种训练说明性的由训练器209执行（图2和3）。根据块502，在各种类中，通过将在训练文集中的句子划分到各种类中，对类特定n-字符列语言模型进行训练，其中各种类的n-字符列语言模型已经在过去得到训练。根据块506，对应于每一类的句子被用来为该类训练n-字符列分类器。这将产生给定数目的n-字符列语言

模型，其中该数目对应于被考虑的类的总数目。应该注意到，根据块504，在训练之前可选择的将特征提取技术应用到句子中，该特征提取技术诸如取词干或其它形式的预处理。该预处理可以在句子被划分到类之前或之后进行。步骤506也可以包括利用平滑定义语言模型以帮助减少如上所述的稀疏数据的影响。

N-字符列分类器的条件最大似然（CML）估计

根据本发明的某些实施例，利用条件最大似然（CML）估计方案来执行训练步骤506。该CML训练方案使得统计语言模型参数可以连同所有的类一起训练，从而该分类器对于给定的训练句子或发声区分出正确的类和不正确的类。

根据一个实施例，联合的训练语言模型P（W｜C），∀C，从而最小化类误差率（等式1）。由于该CER并不是解析易处理的，对于语音发声分类，方便的替代是$\prod_{i=1}^{T} P(s_i \cdot C|s_i \cdot A)$，或对文本分类方便的替代是$\prod_{i=1}^{T} P(s_i \cdot C|s_i \cdot W)$，其中T是该训练数据中样本的数目，$|S_i \cdot A$表示训练样本i中的声音，$|S_i \cdot C$为与训练样本i相关的类，$|S_i \cdot W$是与训练样本i相关的句子（字串）。通过对该训练数据的CER的逆相关（等于该训练数据误差的期望概率）调整选择。

将注意限制到只有文本的情况，最好调节语言模型P（W｜C），∀C，以使该条件对数似然最大：

$$L(C|W) = \sum_{i=1}^{T} \log P(s_i \cdot C|s_i \cdot W) \qquad (7)$$

其中：

$$P(C|W) = P(C) \cdot P(W|C) / \sum_{L \in C} P(L) \cdot P(W|L) \qquad (8)$$

目标变为使等式7的目标函数最大。出于效率的原因，对于该类特定语言模型最好与最大似然（ML）情况（见等式5）保持相同的参数化。例如最好与ML情况（虽然为参数确定的值不同）保持相同的参数λ和f_n。该存储要求和运行时间应该说明性的与ML模式相同。这是可选的限制，并可以删除。

值得一提的是对于语音发声分类，可以调节语言模型P（W｜C），∀C，以使该条件似然L（C｜A）最大：

$$L(C|A)=\sum_{i=1}^{T}\log P(s_i\cdot C|s_i\cdot A) \quad (9)$$

其中

$$P(C|A)=P(C,A)/\sum_{L\in C}P(L,A) \quad (10)$$

$$P(C,A)=P(C)\sum_{w\in v^*}P(W|C)\cdot P(A|W) \quad (11)$$

为了最大化L（C｜A）（等式9）相对于L（C｜W）（等式7），调节语言模型之间的显著区别在于，前一种情况的语言模型会考虑字之间的声音模糊，并会试图降低高度容易混淆的词对分类结果中的作用。

最大化L（C｜W）需要用于语言模型训练的与类标识一起的词音标（该声音数据不是用于语言模型训练）。由于该结果模型是n-字符列，它们可以容易的使用于语音和文本分类。

用于CML N-字符列参数估计的有理函数增长变换

如上所述，根据等式5参数化每一类特定n-字符列模型。该CML估计处理的目标就是像插值加权$\lambda(w_{-1},\ldots,w_{-n})$一样在所有阶0...n-1调节相对频率值$f(w|w_{-1},\ldots,w_{-n})$，使得该训练数据的条件似然L（C｜W）（参见等式7）最大。

根据本发明的一方面，有理函数增长变换（RFGT）是一种高效并且有影响的技术。该熟知的数学应用RFGT算法在P.S Gopalakrishanan et al., An inequality for rational functions with applications to some statistical estimation problems, IEEE Transaction on Information Theory, Vol. 37, No. 1, pp. 107-113, January 1991中有详细描述。该RFGT处理特别适合当前所描述的应用，因为它对概率分布进行运算，并由此在每次迭代时对模型参数化下的概率分布执行适当的归一化。

根据等式5的参数化，k阶相对频率的再次估计等式为：

$$\hat{f}_k(w|h_k,c)=$$

$$\frac{f_k(w|h_k,c)+\beta(h_k,c)\frac{\Pi_{l=k+1}^{n}\lambda_l(h_l,c)\overline{\lambda_k(h_k,c)}}{P_n(w|h_n,c)}f_k(w|h_k,c)\cdot\frac{C_{CML}(w,h_k,c)}{C_{ML}(h_k,c)}}{norm(h_k,c)} \quad (12)$$

$$norm(h_k,c)=$$

$$1+\beta(h_k,c)\sum_{w\in v}\frac{\prod_{l=k+1}^{n}\lambda_l(h_l,c)\overline{\lambda_k(h_k,c)}}{P_n(w|h_n,c)}f_k(w|h_k,c)\cdot\frac{C_{CML}(w,h_k,c)}{C_{ML}(h_k,c)}$$

其中C_{ML}（w，h_k，c）表示类c的句子中该最大似然计数，该类c从下面的训练数据中产生：

$$C_{ML}(w,h_k,c)=\sum_{l=1}^{T}C((w,h_k)\in s_l\cdot W)\cdot\delta(c,s_l.C) \quad (13)$$

C_{ML}（w，h_k，c）表示类c的句子中（w，h_k）的“CML 计数（CML count）”

$$C_{CML}(w,h_k,c)=\sum_{l=1}^{T}C((w,h_k)\in s_l\cdot W)\cdot[\delta(c,s_l.C)-P(c|s_l\cdot W)] \quad (14)$$

关于使用该类特定n-字符列模型P_n（w｜h，c）和类P（c）的先验概率分配的给定的句子$W=w_1,\ldots,w_q$的类c的概率P（c｜W）：

$$P(c|W)=\frac{P(c)\cdot P(W|c)}{\sum_{d\in C}P(d)\cdot P(W|d)} \quad (15)$$

$$P(W|c)=\prod_{i=1}^{q}P_n(w_i|h_i,c) \quad (16)$$

在每次迭代，将上下文有关的“CML加权”β（h_k，c）设为$\beta_{max}>0$。在下面的推导，对于每一上下文β（h_k，c）它的值然后被分别降低，从而对于所有情况（w，h_k，c）等式12中的分子为非负：

(17)

$$f_k(w|h_k,c)+\beta(h_k,c)\frac{\prod_{l=k+1}^{n}\lambda_l(h_l,c)\overline{\lambda_k(h_k,c)}}{P_n(w|h_n,c)}f_k(w|h_k,c)\cdot\frac{C_{CML}(w,h_k,c)}{C_{ML}(h_k,c)}>\varepsilon$$

应该注意到对于大多数上下文（h_k, c），如果选择足够小的β_{max}值开始，这种调整是不必要的。选择小的β_{max}值的不利方面就是，每次迭代L（C｜W）的相对增加较小。

通过随机的将训练数据划分成主数据和支持数据，估计模型参数。使用与ML训练相同的部分。主数据用于估计相对频率，而支持数据用于调节RFGT迭代的次数和最优的CML加权β_{max}。不再次估计类的先验概率和插值加权。类的先验概率、相对频率和插值加权的初始值说明性的为ML的值。例如该插值加权为等式5中所计算的这些值。

RFGT迭代的次数和最优的CML加权β_{max}说明性的按照如下确定：

- 选定将要运行的RFGT迭代的预先确定的次数N；
- 选定研究的栅格（值的范围和步骤），以确定β_{max}参数值；
- 对于每一值β_{max}，运行尽可能多的RFGT迭代（不超过N），使得该主数据的条件似然L（C｜W）在每次迭代都增加；和
- 保留使得该支持数据的条件似然L（C｜W）最大的该对（迭代次数，β_{max}）作为理想值。

确定RFGT迭代运行的次数和该β_{max}值之后，该主数据和支持数据被集中，并且使用这些值训练该模型。

尽管参照特定实施例描述了本发明，本领域的熟练技术人员会认识到在不脱离本发明的精神和范围的条件下可以对形式和细节作出改变。

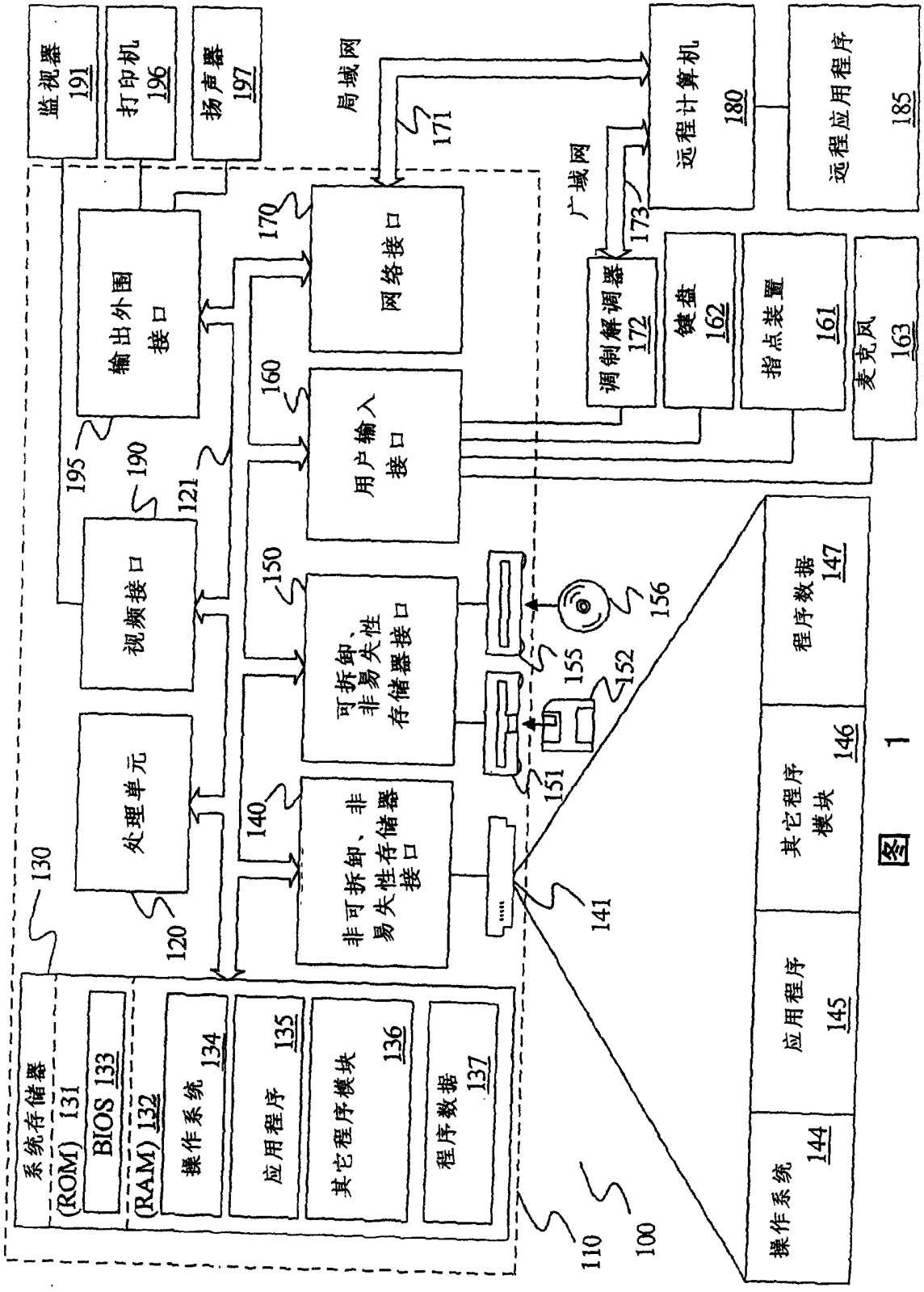

监视器 191
打印机 196
扬声器 197
局域网
171
远程计算机 180
远程应用程序 185
广域网
173
输出外围接口
网络接口
170
160
用户输入接口
调制解调器 172
键盘 162
指点装置 161
麦克风 163
195
190
121
视频接口
150
可拆卸、非易失性存储器接口
处理单元
120
140
非可拆卸、非易失性存储器接口
130
系统存储器 (ROM) 131
BIOS 133
(RAM) 132
操作系统 134
应用程序 135
其它程序模块 136
程序数据 137
110
100
141
151
155
152
156
程序数据 147
其它程序模块 146
应用程序 145
操作系统 144

图 1

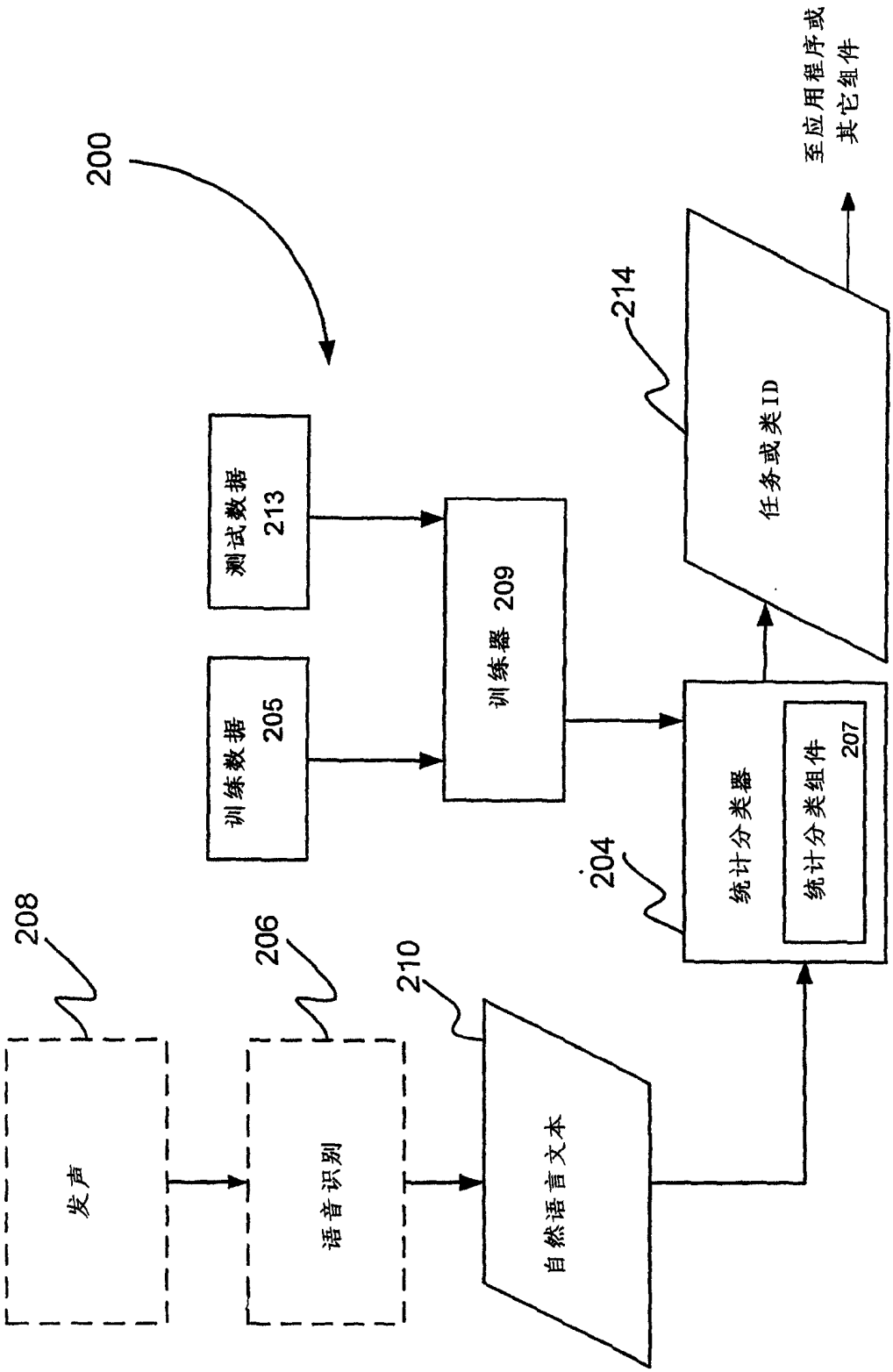

200
至应用程序或
其它组件
214
任务或类ID
测试数据
213
训练数据
205
训练器 209
统计分类器
204
统计分类组件
207
208
发声
206
语音识别
210
自然语言文本

图 2

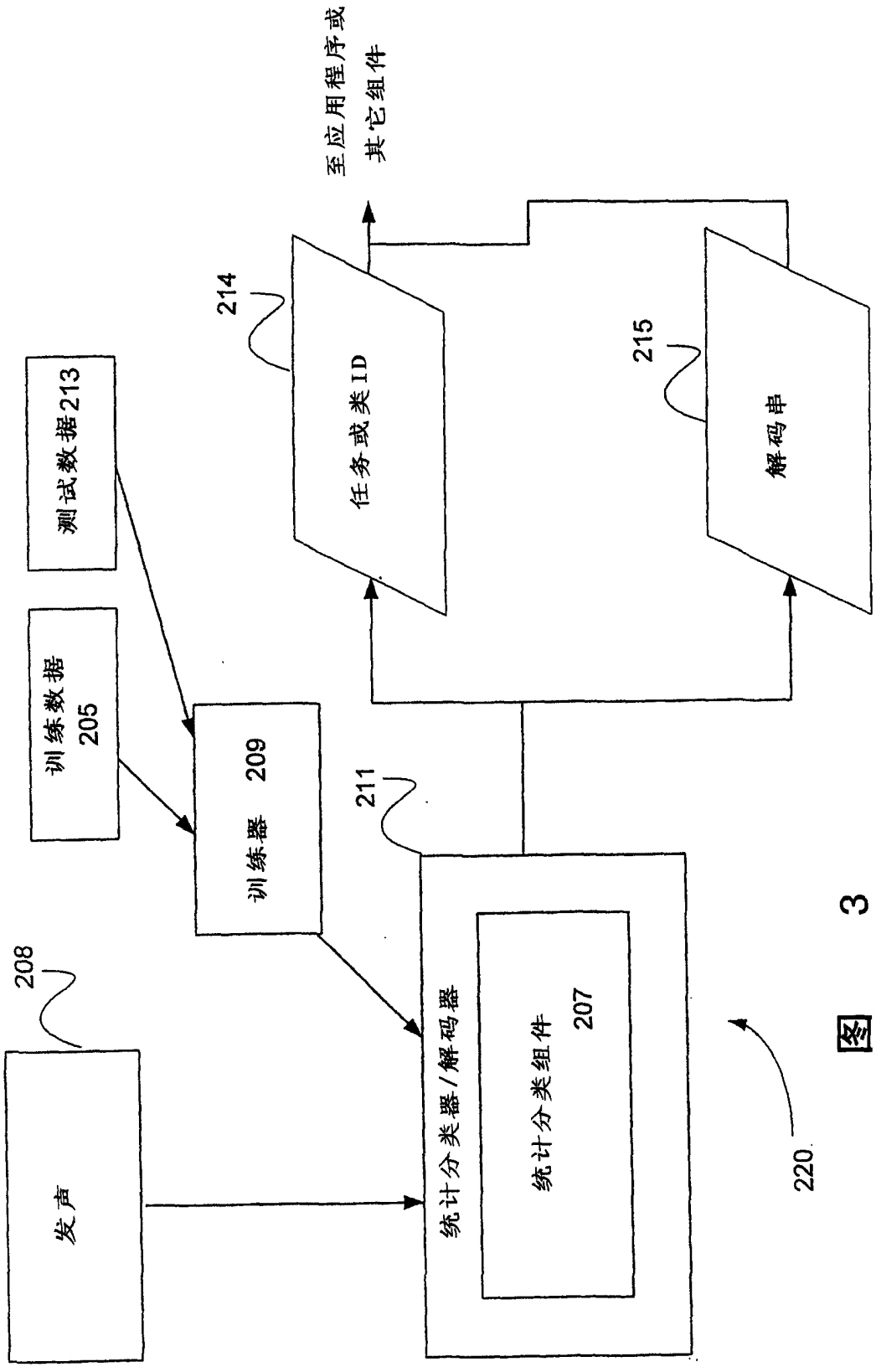

至应用程序或
其它组件
214
任务或类 ID
215
解码串
测试数据213
训练数据
205
训练器 209
211
208
发声
统计分类器/解码器
统计分类组件
207
220

图 3

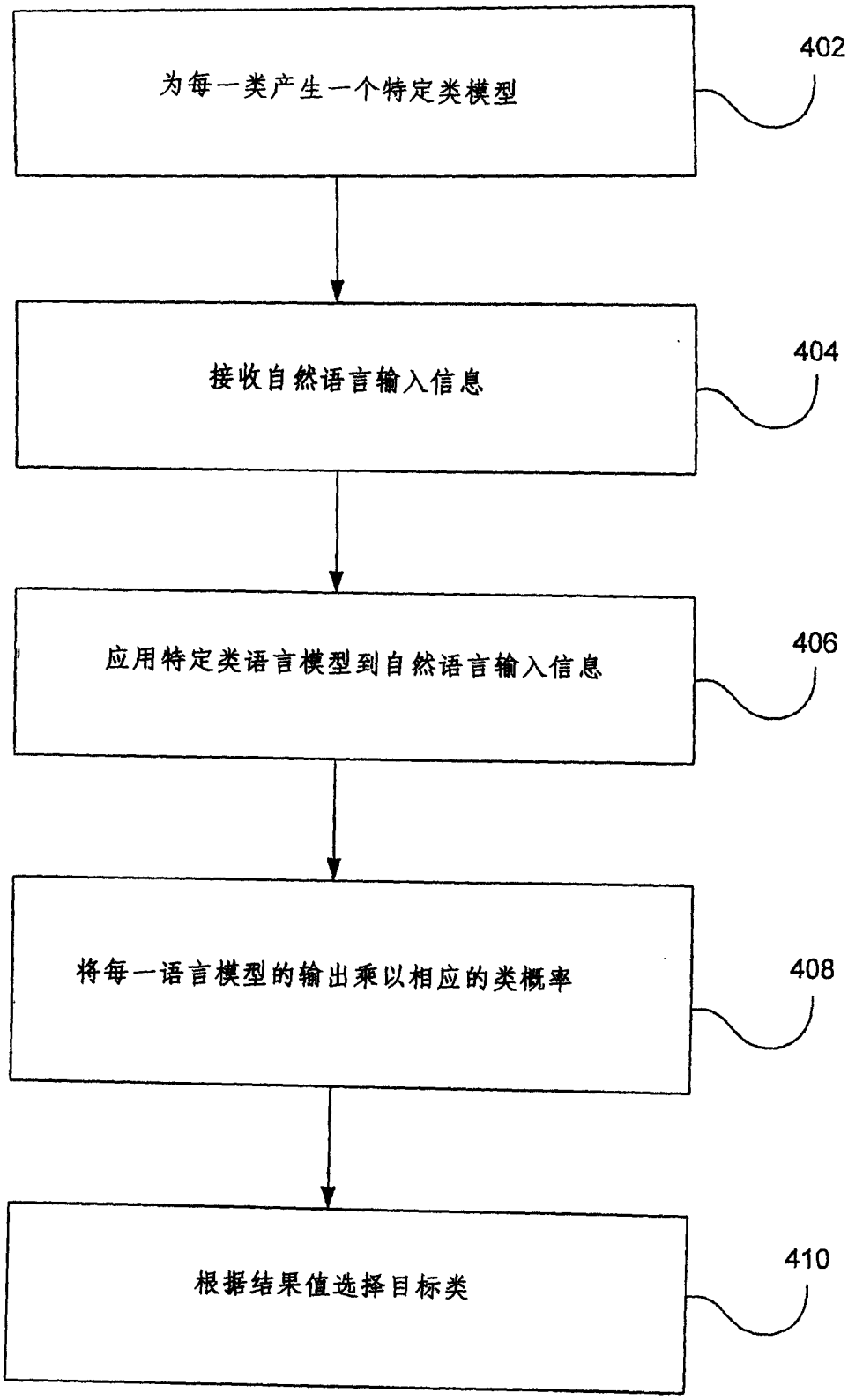

为每一类产生一个特定类模型
402
接收自然语言输入信息
404
应用特定类语言模型到自然语言输入信息
406
将每一语言模型的输出乘以相应的类概率
408
根据结果值选择目标类
410

图 4

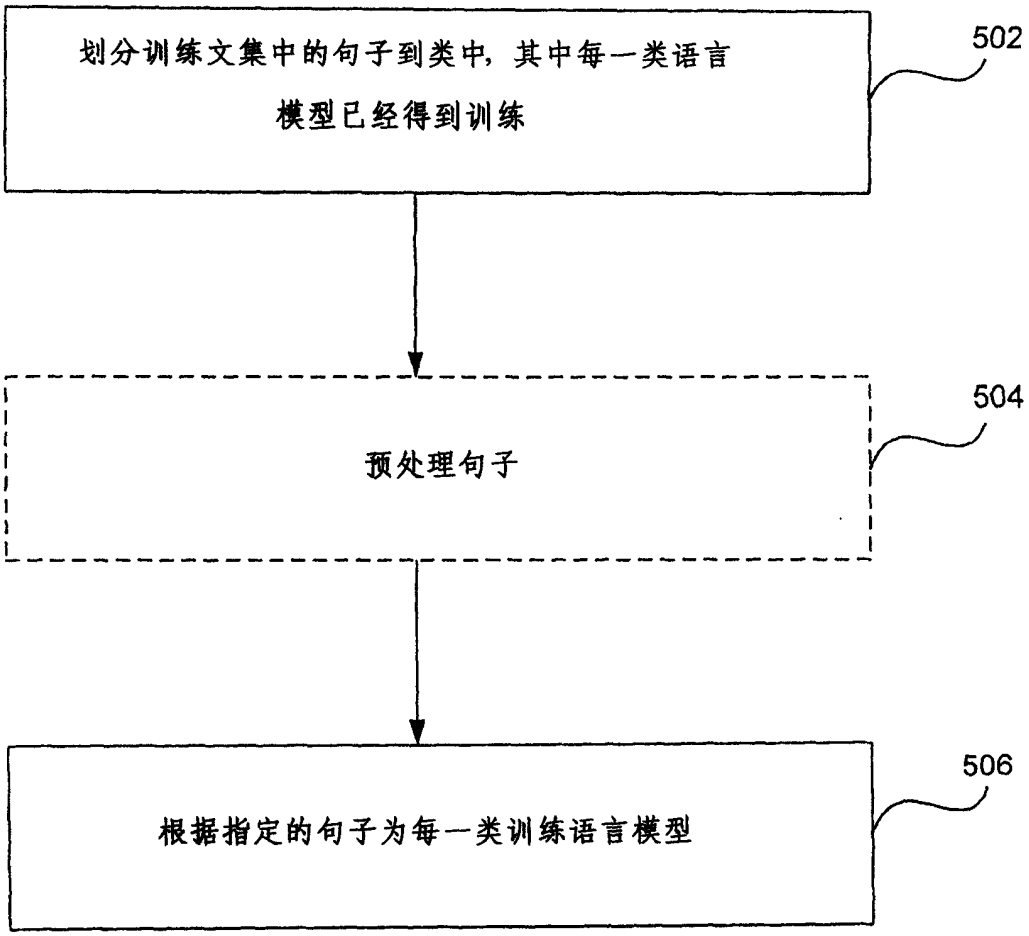

划分训练文集中的句子到类中，其中每一类语言模型已经得到训练
502
预处理句子
504
根据指定的句子为每一类训练语言模型
506

图 5