[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G10L 15/06 (2006.01)
G10L 15/00 (2006.01)
G10L 15/08 (2006.01)

[12] 发明专利申请公开说明书

[21] 申请号 200610005447.3

[43] 公开日 2006 年 8 月 16 日

[11] 公开号 CN 1819018A

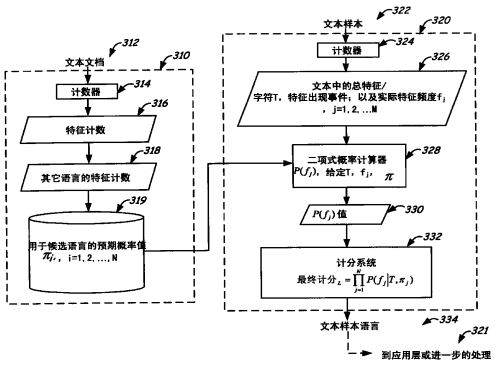
[22] 申请日 2006.1.16
[21] 申请号 200610005447.3
[30] 优先权
[32] 2005. 2.11 [33] US [31] 11/056,707
[71] 申请人 微软公司
地址 美国华盛顿州
[72] 发明人 K·R·鲍威尔 P·M·施密德
W·D·拉姆赛
[74] 专利代理机构 上海专利商标事务所有限公司
代理人 张政权
权利要求书 3 页 说明书 20 页 附图 13 页

[54] 发明名称

有效语言识别

[57] 摘要

提出一种对自然语言文本进行语言识别的系统和方法。 该系统包括用于自然语言中找到的一列字符的计分预期字符计数和方差。 在语言识别中存储预期字符计数和方差以用于待考虑的多个语言。 在运行时间中，基于对实际和预期字符计数的比较识别用于文本样本的一个或多个语言。 本方法可结合用于文本样本中字符的统一字符编码范围的上行分析以限制所考虑的语言的数目。 此外，可在下行处理中使用 n 元语法方法以从通过本系统和方法识别出的语言中选择最大可能语言。



1．一种识别文本的自然语言的方法，包括如下步骤：

接收以已知自然语言书写的文本文档；

对所述文本文档中的唯一特征的出现事件进行计数，以生成预期特征计数；以及

使用概率分布和所述预期特征计数，按照实际特征出现事件的函数来生成概率值。

2．如权利要求 1 所述的方法，其特征在于，使用概率分布包括使用离散或连续概率分布。

3．如权利要求 2 所述的方法，其特征在于，使用概率分布包括使用二项式或高斯分布。

4．如权利要求 1 所述的方法，其特征在于，还包括构建用于多个候选语言中的每一个的概率值表。

5．如权利要求 4 所述的方法，其特征在于，还包括：

接收以未知自然语言书写的文本样本；

确定所述文本样本中的某些特征的实际特征计数；以及

访问所述概率值表以基于所述实际特征计数识别用于所述文本样本的至少一个候选语言。

6．如权利要求 4 所述的方法，其特征在于，还包括通过将与所述实际特征计数相关联的概率值相乘以对每个候选语言计分。

7．一种识别文本中自然语言的方法，包括如下步骤：

接收以未知自然语言书写的文本样本；

确定所述文本样本中的至少一个字符窗口中的至少一个特征的当前计数；以及

对于多个候选语言取得所述至少一个特征的预期概率信息；

基于所述当前计数和已获得的预期概率信息，从多个候选语言中确定用于所述文本样本的至少一个语言。

8．如权利要求 7 所述的方法，其特征在于，获得预期概率信息包括基于二项式分布接收用于所述至少一个特征的概率值。

9．如权利要求 7 所述的方法，其特征在于，还包括通过选择性划分大小的样本，对训练全集进行采样以估计所述预期概率信息，所述预期概率信息包括所述至少一个特征的平均计数。

10．如权利要求 7 所述的方法，其特征在于，还包括使用所述至少一个已识别语言的 n 元语法语言档案以识别所述文本样本的最大可能语言。

11．如权利要求 7 所述的方法，其特征在于，还包括使用统一字符编码值以识别所述多个候选语言。

12．如权利要求 7 所述的方法，其特征在于，识别所述至少一个语言包括基于将所述至少一个特征的当前计数与所述已获得的预期概率信息进行比较，生成用于所述多个候选语言中的每一个的语言计分。

13．如权利要求 12 所述的方法，其特征在于，生成语言计分包括对所述文本样本中含有已确定当前计数的多个特征估计联合概率。

14．如权利要求 7 所述的方法，其特征在于，生成语言计分包括当所

述至少一个字符的当前计数属于所述已获得的预期概率信息的方差中时，对候选语言正向计分。

15．如权利要求 7 所述的方法，其特征在于，生成语言计分包括当所述至少一个字符的当前计数在所述已获得的预期概率信息的方差之外时，对候选语言负向计分。

16．如权利要求 7 所述的方法，其特征在于，生成计分包括对于所述文本样本中的预期特征的不出现事件，对候选语言进行负向计分。

17．如权利要求 7 所述的方法，其特征在于，还包括估计用于所述已识别的至少一个语言的置信计分。

18．一种计算机可读介质，包括指令，当执行所述指令时使计算机执行语言识别，所述指令包括：

用于构建并存储用于多个自然语言中每一个的特征列表以及与每一个所列出的特征相关联的预期概率值的模块；以及

用于对文本样本中的实际特征进行计数并访问与所述实际特征相关的已存储预期概率值，以识别用于所述文本样本的至少一个自然语言的模块。

19．如权利要求 18 所述的计算机可读介质，其特征在于，还包括用于确定用于所述已识别自然语言的置信计分并基于所述置信计分对自然语言进行排序的模块。

20．如权利要求 18 所述的计算机可读介质，其特征在于，还包括用于访问用于至少一个已识别自然语言中的每一个的 n 元语法语言档案以在所述文本样本上执行语言识别的模块。

有效语言识别

背景技术

虽然大量数据网络跨越全球使线上世界变为真正的多国社区，仍然没有单一的用来交流的人类语言。电子消息以及文档仍然用于特定人类语言来书写，诸如德语、西班牙语、葡萄牙语、希腊语、英语、汉语、日语、阿拉伯语、希伯来语、或印地语。

在许多情况下需要快速地识别特定文档的人类语言，以用于进一步的自然语言处理。例如，文档的人类或自然语言识别有助于对文档进行索引和分类。在其它情况中，文字处理需要识别文档语言以进行拼写检查、语法检查、使用语言翻译工具或库、或者启动合适的打印机字体。

先前的语言识别方法包括 n 元语法方法，特别是三元语法方法。在一些三元语法方法中，语言特定训练数据或文档已经被用于创建相应语言的表或档案，这些表或档案被称为三元语法语言档案。在一些实现中，一个三字母长的窗口滑过特定语言的训练文本。当三字母长的窗口滑过文本时，该方法对出现在窗口中的三字母长的序列进行计数以生成用于特定语言的三元语法语言档案。对各种语言的文本重复该处理以提供相应语言的三元语法语言档案集，其稍后被用于未知语言文本的语言识别。

在语言识别中，一个类似的三字母长的窗口滑过未知文档。对于未知文档中每个三字母长的序列，该方法寻求在每一个三元语法档案中查找匹配的三字母长的序列。如果查找到对于特定语言的匹配，则将用于已匹配的三字母长的序列的语言档案中的频度信息添加到该特定语言的累积计分上。这样，每个语言的累积计分随着窗口滑过整个未知文档而增加。也可使用其它计分方案，诸如将 n 元语法频度信息存储为概率值。在匹配时，可将这些概率值以生成累积语言计分。带有最高累积计分的语言被认定为是未知文档的语言。不幸的是，三元语法方法通常消耗大量计算。

语言识别的另一种方法包括改变 n 元语法序列的长度。在该语言识别系统中，n 元语法档案（更常见地被称为“语言档案”）包括各种长度的 n 元语法（如，二元语法、三元语法、或四元语法）的频度信息。然而，如同三元语法方法一样，其它 n 元语法方法也消耗大量计算，并因此是相对缓慢的。这样的速度不够通常当所考虑的语言数目增加时变得更成问题。此外，速度不够可在语言识别结合其它应用（诸如，文档索引）时尤为成问题。然而有利的是，当文档或文本样本相当简洁时（诸如单独的句子），三元语法和其它 n 元语法语言识别方法被认为是相当准确的。

考虑到现有技术语言识别方法和系统的问题的更快速和/或改进型的一种语言识别方法将是非常有用的。

发明内容

本发明包括建立用于各种自然语言的字符预期概率语言模型。在对文本样本进行语言识别时，访问语言模型以对各种语言进行计分和/或识别。文本样本的语言是基于计分而识别的。包括语言模型的本发明的语言识别可以集成在更大的语言学服务平台上，尤其是结合语言自动检测（LAD）功能。对输入文本的统一字符编码标准值的分析可与本方法或系统结合，尤其用于限制待考虑或计分的候选语言的数目。本发明可结合其它语言识别方法，诸如 n 元语法方法，以用于最优化性能。

附图说明

图 1 示出可使用本发明的一个示例性环境。

图 2 示出可使用本发明的自然语言处理系统的环境。

图 3 是本发明的广义流程图。

图 3A-3B 一起示出执行图 3 中所示方面的方法和系统，包括文本样本的语言识别。

图 4 是根据本发明用于增大词汇知识库的系统的框图。

图 5 示出通常对应于图 4 的系统的方法的步骤。

图 6 示出根据本发明用于执行语言识别的系统。

图 7 示出通常对应于图 6 的系统的方法的步骤。

图 8a-8b 示出在语言识别过程中已存储概率信息和已存储概率信息的使用的示例。

图 9 是示出根据本发明的计算机辅助训练过程的实施例的流程图。

图 10 是示出根据本发明的语言识别的实施例的流程图。

图 11 是示出根据本发明的确定文本的最相似语言的实施例的流程图。

具体实施方式

本发明涉及自然语言文本处理，尤其涉及输入或样本文本的自然语言识别。在一方面，构建了在各种自然语言中发现的字符概率语言模型。在另一方面，访问语言模型以执行自然语言文本的语言识别。在另一方面，本发明可结合其它识别语言的系统或方法，诸如通过分析字符统一编码范围或通过使用 n 元语法语言识别。

示例性环境

图 1 示出可实现本发明的示例性计算系统环境 100 的一个示例。计算系统环境 100 仅仅是合适的计算环境的一个示例且不旨在对本发明使用范围或功能提出任何限制。计算环境 100 也不能被解释是对示例性操作环境 100 中示出的任一组件或组合的有任何依赖或要求。

本发明可操作于多种其他通用功能或特定功能计算系统环境或配置。适合本发明使用的已知计算系统、环境、和/或配置的示例包括，但不限于，个人计算机、服务器计算机、手持或膝上设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费品电器、网络 PC、小型机、大型机、电话系统、包括任何上述系统或设备的分布式计算环境、以及诸如此类设备。

本发明可被描述在计算机可执行指令（诸如由一个计算机或其他设备执行的程序模块）的通常环境中。通常有，程序模块包括执行特定任务或实现特定抽象数据类型的例程、应用程序、对象、组件、数据结构、以及

诸如此类。本领域熟练技术人员可以将在此提供的描述和附图实现为处理器可执行的指令，这些指令可写入任何形式的计算机可读介质。

本发明还可被实现于分布式计算环境，后者中的任务被由通信网络连接在一起的远程处理设备所执行。在一个分布式计算环境中，程序模块可位于本地和/或远程计算机存储介质，包括内存存储器设备。

参照图 1，一个用于实现本发明的示例性系统包括一个如计算机 110 形式的通用功能计算设备。计算机 110 的组件可包括，但不限于，处理单元 120、系统存储器 130、以及将包括系统存储器在内的各种系统组件连接到处理单元 120 的系统总线 121。系统总线 121 可为多种类型的总线结构的任何一种，包括存储器总线或存储器控制器、外围设备总线、以及使用任何的多种总线体系结构中任何一种的局部总线。作为示例，而非限制，这些体系结构包括工业标准体系结构（ISA）总线、微通道体系结构（MCA）总线、增强型 ISA（EISA）总线、视频电子标准协会（VESA）局部总线以及外围设备组件互连（PCI）总线（也被称为 Mezzanine 总线）。

计算机系统 110 通常包括多种计算机可读介质。计算机可读介质可为计算机 110 可访问的任何可用介质，并包括易失和非易失介质、可移动和不可移动介质。通过示例，而非限制，计算机可读介质可包括计算机存储介质和通信介质。计算机存储介质包括通过任何方法或技术实现的，用于存储诸如计算机可读指令、数据结构、程序模块或其他数据的信息的，易失性和非易失性、可移动和不可移动介质。计算机存储介质包括，但不限于，RAM、ROM、EEPROM、闪速存储器或其他存储器技术、CD-ROM、数字通用盘（DVD）或其他光学盘存储器、磁带盒、磁带、磁盘存储器或其他磁存储器设备、或任何其他可被用来存储所需信息并能够由计算机 110 访问的介质。通信介质通常以一个已调制的数据信号，诸如载波或其他传输机制的形式来体现计算机可读指令、数据结构、程序模块或其他的数据，并包括任何信息传递介质。术语“已调制的数据信号”表示为了在信号内编码信息而设置或改变其一个或多个特征的信号。通过示例，而非限制，通信介质包括有线介质，诸如有线网络或直线连接、以及无线介质诸如声学的、

FR、红外以及其他无线介质。任何上述的组合也被包括在计算机可读介质的范围内。

系统存储器 130 包括易失和/或非易失存储器形式的计算机存储器介质，诸如只读存储器(ROM)131 以及随机访问存储器(RAM)132。基本输入/输出系统 133（BIOS）一般被存储在 ROM 131 中，它包括诸如在起动过程中有助于计算机 110 内基本元件间传递信息的基本例程。RAM 132 通常包括可由处理单元 120 立即访问和/或当前进行操作的数据和/或程序模块。通过示例，而非限制，图 1 示出了操作系统 134、应用程序 135、其他程序模块 136、以及程序数据 137。

计算机 110 也可包括其他可移动/不可移动的易失/非易失计算机存储介质。仅仅作为示例，图 1 示出了从不可移动非易失磁性介质读取或写入的硬盘驱动器 151、从可移动非易失磁盘 152 读取或写入的磁盘驱动器 151、以及一个从诸如 CD-ROM 或其他光学介质的可移动非易失光盘 156 读取或写入的光盘驱动器 155。其他可被用于示例性操作环境的可移动/不可移动，易失/非易失计算机存储介质包括，但不限于，磁带盒、闪速存储卡、数字通用光盘、数字录影带、固态 RAM、固态 ROM、等等。硬盘驱动器 141 通常通过诸如接口 140 的不可移动存储器接口连接到系统总线 121，磁盘驱动器 151 以及光盘驱动器 155 通常通过一个诸如接口 150 的可移动存储器接口连接到系统总线 121。

以上讨论的且示于图 1 中的驱动器和它们相关的计算机存储介质为来自计算机 110 的计算机可读指令、数据结构、程序模块以及其他数据提供了存储。在图 1 中，例如，硬盘驱动器 141 被示为存储操作系统 144、应用程序 145、其他程序模块 146、以及程序数据 147。注意这些组件可与操作系统 134、应用程序 135、其他程序模块 136、以及程序数据 137 相同或相异。操作系统 144、应用程序 145、其他程序模块 146 以及程序数据 147 这里被给予了不同的标号用于表示在最小限度下，它们是不同的拷贝。

用户可以经由输入设备（诸如键盘 162、麦克风 163、以及通常为鼠标、轨迹球或触摸板的定位设备 161）输入命令和信息进入到计算机 110 中。没

有在图 1 中示出的其他输入设备可包括操纵杆、游戏垫、卫星天线、扫描仪等等。这些及其他输入设备经常经由连接到系统总线用户输入接口 160 连接到处理单元 120，但也可通过其他接口和总线结构进行连接，诸如并行端口、游戏端口或通用串行总线（USB）。一监视器 191 或其他类型的显示设备也经由诸如视频接口 190 的接口被连接到系统总线 121。除监视器以外，计算机也可包括其他外围输出设备，如扬声器 197 和打印机 186，它们可通过输出外围接口 190 连接。

计算机 110 可操作在一个网络环境下，该网络环境使用连接到一个或多个诸如远程计算机 180 的远程计算机的逻辑连接的。远程计算机 180 可为个人计算机、服务器、路由器、网络 PC、对等设备或其他公共网络节点，并通常包括许多或所有上述对应于计算机 110 的元件。图 1 描述的逻辑连接包括局域网（LAN）171 和广域网（WAN）173，但也包括其他网络。这些网络化环境在办公室、企业级计算机网络、内部网和因特网上是普通的。

当用于 LAN 网络环境时，计算机 110 通过网络接口或适配器 170 连接到 LAN 171 上。当用于 WAN 网络环境时，计算机 110 通常包括调制解调器 172 或其他装置，用于在诸如因特网的 WAN 173 上建立通信。调制解调器 172 可以是内置的或外置的，它通过用户输入接口 160 或其他合适的机制连接到系统总线 121 上。在网络环境中，与计算机 110 相关描述的程序模块或它的部分可存储在远程存储器设备中。通过示例，而非限制，图 1 示出了远程应用程序 185 驻留在存储器设备 181 上。可以被理解的是，所示网络连接是示例性的，且可以使用在计算机之间建立通信连接的其他装置。

图 2 是可实现本发明的另一种示例性环境的框图。更详细地，图 2 示出了带有自然语言识别能力的自然语言处理系统。在递交于 2004 年 3 月 30 日的美国专利申请 10/813.652 中已经详细描述了类似于图 2 的通用环境，其整体结合在此作为参考。

自然语言处理系统 200 包括自然语言编程接口 202、自然语言处理（NLP）引擎 204、以及相关词汇 206。图 2 也示出了系统 200 与包括应用程序的应用层 208 进行交互。这些应用程序可以是自然语言处理应用（诸

如，单词搜索、数据采集、或文档索引），其要求访问自然语言处理服务（可称为语言服务平台或“LSP”）。

编程接口 202 披露了可由应用层 208 调用的要素（方法、属性或接口）。编程接口 202 的元素由底层对象模型（在上述结合的专利申请中提供了它的详细资料）所支持，这样应用层 208 中的应用可以调用所披露的要素来获得自然语言处理服务。为了达到该目的，层 208 中的应用可首先访问披露接口 202 的对象模型以配置接口 202。术语“配置”应该包括选择所期望的自然语言处理特征或功能。例如，如 203 所示应用可以希望选择的语言自动检测（LAD）。

一旦配置了接口 202，应用层 208 可向接口 202 提供待处理的文本（诸如自然语言文本，样本）或文档。接口 202 接着可访问 NLP 引擎 204，NLP 引擎 204 可执行诸如包括根据本发明的语言识别的语言自动检测（LAD）205、单词分段、或其它自然语言处理。例如，所执行的自然语言处理的结果可以通过编程接口 202 被提供给应用层 208 中的应用，或者如下面描述的被用于更新词汇 206。

接口 202 或 NLP 引擎 204 也可利用词汇 206。词汇 206 是可更新或可修订的。系统 200 可提供核心词汇 206，这样就不需要额外的词汇。但是，接口 202 也披露了允许应用添加自定义词汇 206 的要素。例如，如果应用集中在文档索引或搜索上，则可添加或访问带有指定实体（如，人或公司名称）的自定义词汇。当然，也可以添加或访问其它词汇。

此外，接口 202 可披露允许应用向词汇添加符号的要素，这样当从词汇返回结果时，也可提供符号，例如，作为结果的属性。

二项式分布

二项式分布是众所周知的离散概率分布。例如，当弹起硬币时，结果可以是正面也可以是背面；当魔法师预测从牌叠中选出的纸牌时，魔法师可以是正确的或错误的；当婴儿降生时，婴儿可以是降生在四月也可以不在四月。在每个这些示例中，事件含有两个相互排斥的可能结果。这些结

果中的一个可以被标记为“成功”而另一个结果为“失败”。如果事件发生 T 次（例如，硬币被弹起 T 次或 T 次“试验”），则二项式分布可被用于确定在 T 次试验中恰好获得 C 次成功的概率。在 T 次试验中获得 C 次成功的二项式概率由下式给出：

$$P(c)=\frac{T!}{c!(T-c)!}\pi^{c}(1-\pi)^{T-c} \qquad \text{公式 1}$$

其中 $P(c)$ 在 c=C 时是恰好 C 次成功的概率，T 是事件的次数，π 是在任何一次试验中成功的概率或期望概率。该公式做出以下假设：

1．存在 T 次同样的试验，其中 T 是事先预定的；

2．每次试验含有两个可能结果，成功或失败；

3．试验是独立的，因此一次试验的结果对另一次试验的结果没有影响；以及

4．工程的概率从一次试验到另一次试验都是不变的。

对于二项式分布，x 的均值和方差相应地由下式给出：

$$E(c)=T\pi \qquad \text{公式 2}$$

$$Var(c)=T\pi(1-\pi) \qquad \text{公式 3}$$

这样，例如，假设碗中有 10 个球，3 个球是红色的 7 个球是蓝色的。成功被定义为摸到红球。如果随机摸球并且随后将球放回，则每次试验的成果的概率是 $\frac{3}{10}$ 或 $\pi=0.3$。如果摸 10 个球，则 T=10。这样，通过将 T、c、和 π 的值代入公式 1，可得在 10 次试验中摸到 5 个红球（即，c=5）的概率：

$$P(5)=\frac{10!}{5!(10-5)!}0.3^{5}(1-0.3)^{10-5}$$

$$=15504(0.3)^{5}(0.7)^{15}$$

$$=0.1789$$

这样，5 次成功（或摸 5 次红球）的概率是大约 18%。可为 0 到 10 之间（试验次数）c 的不同值计算和/或图解二项式分布。

此外，在上述示例中，二项式分布的均值 E(c)和方差 Var(c)可使用上述公式 2 和 3 来确定：

$$E(c)=T\pi=(10)(0.3)=3$$

$$Var(c)=T\pi(1-\pi)$$

$$=(10)(0.3)(1-0.3)=2.1$$

还需要朱德是当试验的次数增加时，作为所有试验的百分比的方差趋向于减小。这样，当试验次数增加时，预测精度提高。

通常，使用二项式分布的累积形式，这样由下式给出摸 5 次或更多次红球的概率：

$$P(\geq 5)=\sum_{i=5}^{10}P(c_i) \qquad \text{公式 4}$$

本发明的广义方面

在本发明中，二项式分布（或其它概率分布，诸如高斯分布）的含义可被用于语言识别。在语言 L 的全部 T 个特征中发现某个特征某些次数的概率可在给定语言 L 中特征 f 的期望概率后计算出。特征计数可被视为“成功”，特征总数可被视为“试验”次数。

此外，假定预期概率为π_1到π_N，特征总数为 T，可见特征 1 到 N 与计数 f_1 到 f_N 的联合概率可被估计或表示为：

$$P(f_1,...,f_N)|T,\pi_1,...,\pi_N)=\prod_{i=1}^{N}P(f_i|T,\pi_j) \qquad \text{公式 5}$$

其中可使用二项式分布或类似（离散或非离散）概率函数来获得每个$P(f_i|T,\pi_i)$值。在大多数实施例中，每个特征包括一个或多个在语言中找到的字符。例如，一特征可为类似“a”的单个字符或诸如“tr”或“and”的字符组合。同样，组成特征的一个或多个字符可以是连续的但不限于此。例如，一特征可为由第三个未知字符分割开的两个字符。特征也可包括一个或多个诸如“@”或“#”的符号。然而，在一个实施例中，每个特征表示诸如“a”或“b”的单个字符或字母。使用单个字符作为特征可以有提高计算速度的优点。

同样，在本发明的其它实施例中，可用数学方法$P(c_i|T,\pi_i)$的值获得（诸如，通过使用公式 1 中的二项式分布公式来计算离散概率）。在还有些实施例中，$P(f_i|T,\pi_i)$是通过物理方法获得或凭经验获得的（诸如通过对各种语言的训练全集中的特征或字符进行计数并标准化每个已选择的窗口或样

本的大小）。算术计算和物理计数的一些组合也可被用于确定$P(f_i|T,\pi_i)$值。

还需要注意的是在预期字符概率是通过物理方法确定的实施例中，使用已选择的产生数学整数的样本大小（如，对每 1000 个字符进行字符计数）来进行标准化是有利的。数学整数有利地提高了性能或速度。然而，数学整数是可选的并且可为了获得更精确的十进制值而省略，以获得更高的精确度。同样，注意 1000 个字符的样本大小合适于诸如英语的欧洲语言，在仅仅考虑单个字符或字母时，它们带有相对较少的特征。相反，对诸如汉语或日语的亚洲语言的预期特征概率将更可能使用大得多的样本大小来进行标准化（诸如对每 100000 个特征窗口进行预期特征计数），因为在这些语言的书写系统中使用更大量的特征或象形符号（相对于字母而言）。

图 3 是示出实现为单个方法 300 的本发明的两个广义方面或步骤 302、306 的整体流程图。图 4 和 6 是示出执行每个这些方面的模块的框图。步骤 302 包括使用带有用于多个语言的预期字符计数或概率信息或值的语言模型或表的信息来增大词汇知识库 418（在图 4 中示出），词汇知识库 418 之后被用于语言识别。

语言识别步骤包括步骤 304，用于接收使用未知的或未识别的自然语言书写的输入文本。在步骤 306 中，访问语言模型来识别所接收的自然语言文本的语言。计分系统可被用于识别该文本的最大可能语言或最小不可能语言。或者，语言计分系统可识别最大不可能语言以排除低可能语言，例如，在导出可能语言的候选列表时作为过滤器。如上所述，步骤 306 可包括子步骤，诸如使用统一字符编码值或范围以及/或 n 元语法方法以最优化语言识别性能（如，提高速度和/或精确度）。尤其是，本发明可结合 n 元语法语言识别系统，诸如公布于 2001 年 8 月 7 日授予 de Campos 的美国专利 6,272,456 中所描述的系统，其整体结合在此作为参考。如箭头 308 所示，方法 300 可重复进行，因为根据本发明可接收并处理任何数目的输入文本样本。

图 3A－3B 一起示出了执行图 3 中步骤 302 和 306 的广义方法和以及同时所讨论的系统 310，320。系统 310 可执行步骤 302，系统 320 可执行步

骤 306。

在步骤 352 中，文本文档 312（书写为已知自然语言，如英语或汉语）由系统 310 接收。系统 310 包括计数器 314。在步骤 354 中，计数器 314 对以自然语言书写的文本文档 312 中唯一特征 1 到 N 的出现事件 316 的数目进行计数，并将这些特征计数 316 转换为预期概率或频度值π_i，其中如 316 所示$i=1,...,N$。

在步骤 356 中，对其它自然语言重复步骤 352 和 354 以生成预期特征概率或频度值 318。在步骤 358 中，存储用于所有候选语言的预期特征概率值 316，318 以用于之后在语言识别时的访问。

在步骤 360 中，系统 320 接收以未识别自然语言书写的文本样本 322。系统 320 包括计数器 324、二项式概率计算器 328、以及计分系统 332。在步骤 362 中，计数器 324 对文本样本 322 中的特征或字符 T 的总数目，以及文本样本 322 中唯一特征 1 到 M 的出现事件进行计数，如 326 中所表示的。在步骤 364 中，将观测得的、实际的、或当前的特征频度 $f_1,...f_M$ 计算为 326 中所表示的。在步骤 366 中，二项式概率计算器 328 计算给定文本样本 322 中 T 个总特征 326、已存储预期概率值π_i 319、以及实际特征频度f_i 326 时的概率值 330。在步骤 368 中，计分系统 332 使用例如上述公式 5 为各种候选语言计算语言计分。在步骤 370 中，系统 320 基于语言计分为文本样本 322 生成或确定语言列表 334。文本样本 322 和/或语言列表 334 可被返回给应用层或用于如 321 所示的进一步处理。

图 4 示出系统的另一个实施例，其可以根据本发明执行增加词汇知识库的步骤 302（在图 3 中所示的）。图 5 是通常对应于图 4 所示的模块的增加词汇知识库的步骤的流程图。如下文更详细描述的，根据本发明的词汇知识库可包括语言特定特征以及相关信息，诸如每个特征的预期计数和方差。需要注意的是图 4 和 5 中所示的模块和步骤仅仅是示例性的并且可以按照所期望而省略、组合和划分。同样，图 4 和 5 中的模块和步骤是为单个语言示出的，并可对语言识别步骤中考虑的每个自然语言进行重复。词汇知识库构建模块 404 可以是在计算机 110 中执行的或在 LAN 171 或 WAN 173

连接中的任何远程计算机中存储并执行的应用程序 135。类似地，词汇知识库 418 可安置计算机 110 上的任何本地存储设备中(诸如硬盘驱动器 141)，或可安置在任何光学 CD 上、或远程安置在 LAN 171 或 WAN 173 存储器设备中。

在步骤 502（图 5 中所示）中，词汇知识库构建模块 404 从上述任何输入设备以及结合图 1 所描述的任何存储设备中接收未处理的自然语言文本 402。此外，可在结合图 2 所描述的应用层 208 上接收未处理的文本 402。未处理的文本 402 可为来自书本、出版物、杂志、web 源、语音－文本引擎、以及类似物的自然语言文本。注意自然语言文本 402 通常使用一种自然语言输入。然而，如上所述，构建语言模型 420 以增加词汇知识库 418 的处理是重复的，因为出于语言识别目的构建了多个语言模型 420。

在步骤 504 中，预处理模块 406 可接收未处理的文本 402 以用于预处理，例如，通过移除诸如逗号和句号的语法特征或将诸如单个字母的字符从大写转换为小写。也可移除数字，因为在大部分情况中，数字不是某种语言专用的。然而，在一些实施例中，类似“1”或“2”的阿拉伯数字可以是某种语言专用的，诸如在考虑到类似技术领域的语言子集（诸如英语药学、德语工程学）时。在其它实施例中，数字可以是某种语言专用的，诸如当被考虑的自然语言使用不同的或双重的数字系统时。例如，汉语使用类似“1”和“2”的阿拉伯数字和象形文字来表示数字。

预处理模块 406 生成训练全集 408，其理想地以可代表自然语言的比率较佳地包含从特定语言中找到的字符（即、字母、符号、等等）以及其它特征。或者，有代表性的训练全集可被提供给词汇知识库构建模块 404 或由其来访问。

在步骤 506 中，识别或接收字符列表 412。在一些实施例中，通过字符或特征识别器 410 来接收训练全集 408，字符或特征识别器 410 识别训练全集 408 中的唯一字符以生成字符和/或特征列表 412。或者，用于特定自然语言的字符和/或特征列表可被提供给词汇知识库构建模块 404 或由其来访问。例如，英语的字符列表 412 可包括从字母表“a”到“z”的所有字母；

以及其它诸如“$”或“#”的字符、符号、或特征。然而，如上所述，使用汉字字符或象形符号的亚洲语言（诸如汉语或日语）的字符列表 412 可能是相当大的。

在步骤 508 中，概率计算模块 414 为字符列表 412 中的一些或所有字符生成字符计数概率值 P(c)，如上文详细讨论的。所生成的概率值的结果可被用于在成功数目上或标准化每个已选择的样本大小（如，1000 个字符）的出现事件上生成样本字符的概率分布。在其它实施例中，概率计算模块 414 包括计数器 415，其对每个字符的出现事件的平均次数进行计数，尤其是为多个等大小的带有已选择大小的样本窗口。

在步骤 510 中，概率计算模块 414 生成步骤 508 中采样的字符的“方差”416。在一些实施例中，可特别至少部分基于公式 3 来计算“方差”。例如，可通过获得公式 3 中给定的二项式方差值的平方根或分数值（如，1/10）来确定“方差”。在其它实施例中，可在数学上近似“方差”，诸如通过分析分布曲线的斜率或类似方法。也可通过比较一组等大小样本中字符的实际和预期计数来经验性地计算方差。在还有一些实施例中，“方差”是从人为选择的计数聚集的范围中通过物理方法生成的。

在步骤 512 中，使用带有预期计数或由概率计算模块 414 生成的概率信息和方差的语言模型或表 420 来增大词汇知识库 418。在步骤 514 中，来自字符列表 412 的另一个字符被按照上述方法处理以生成计数或概率信息和方差，以进一步增大词汇知识库 418。继续建立语言模型 420 的处理，直到字符列表 412 中的所有字符都被处理完。在另一实施例中，在计算预期计数和方差以增加语言模型 420 之前，对用于所有样本窗口的列表 412 中的所有字符进行计数。为在下述语言识别步骤的运行时间中待考虑的每个语言构建语言模型 420。

图 6 示出用于实现图 3 中所示的步骤 306 的语音识别系统或模块。语言识别模块 604 可类似于图 2 中的语言识别模块 205，如上所述将它集成到语言学服务平台。此外，图 7 包括与图 6 的语言识别系统有关的或对应于其的方法。因此，在下文一起描述图 6 和 7。同样，在图 6 和 7 中示出的模块

和方法仅仅是示例性的并且可以按照所期望而省略、组合和划分。例如，图 6 中的统一字符编码过滤器 606 和 n 元语法语言识别模块 619 是可选特征，如虚线所标识的。

在步骤 702 中，由语言识别模块 604 接收文本样本 602，语言识别模块 604 根据本发明执行文本样本 602 的语言识别以生成文本样本语言识别 620。在步骤 704，统一字符编码过滤器 606 分析文本样本 602 中的字符统一编码范围，以基于字符统一编码范围生成候选语言列表 608。这样，统一字符编码过滤器 606 可限制或“过滤”文本样本 602 中的待考虑语言的数量。

注意统一字符编码标准是国际字符编码系统，类似 ASCII，它为每一个已知的字符（包括符号）提供唯一的号码或值。因此，统一字符编码值可被确定识别而不考虑平台、程序、或语言。此外，每种人类语言的字符都属于特定统一字符编码范围。同样，人类语言通常被划分为围绕特定统一字符编码范围的族群。这样，欧洲语言（诸如英语、法语、或德语）的字符通常属于一个特定统一字符编码范围。亚洲语言（诸如汉语、日语和韩语）属于不同于欧洲语言的统一字符编码范围的另一个统一字符编码范围。关于统一字符编码标准的进一步信息可在该网站找到:http://www.Unicode.org/。

在步骤 705 中，计数器 611 对文本样本 602 中唯一特征或字符 j=1 到 M 的实际出现事件进行计数。计数器 611 可基于已选择的样本大小（诸如，1000 个字符）（对较短的文本样本可能有适当的缩放比例）来确定这些计数。在步骤 706 中，计分模块 612 接收文本样本 602 中计数的实际特征出现事件 f_j 以及候选语言列表 608 的预期概率或计数值 614，以识别或选择文本样本 602 的最大可能或最小不可能语言 618。

在给定预期概率 p_1 到 p_N 以及字符总数 T 下，计分系统 612 可通过计算文本样本 602 中的可见特征或字符 1 到 N 与观测得的或当前的计数 f_1 到 f_N 的联合概率来生成候选列表 608 中语言的计分。在这些实施例中，语言计分可按照上述公式 5，下面再次给出此式：

$$\text{最终计分}_L = P(f_1,...,f_N)|T,\pi_1,...,\pi_N) = \prod_{i=1}^{N} P(f_i|T,\pi_j) \qquad \text{公式 6}$$

其中每个 $P(f_i|T,\pi_i)$ 可通过访问候选语言的已存储特征概率信息（诸如图 3A 中 319 所示的）来获得。在这些实施例中，越好的语言计分是越高的，因为越高的计分表示文本语言是候选语言的可能性越大。计分模块 610 生成带有最好计分的语言 618。然而，在其它实施例中，计分系统 612 将低计分语言排除出考虑范围。

在一些实施例中，计分系统 612 将观测得的或当前的字符计数与正在考虑的多个候选语言的语言模型 614 中的预期计数或概率和方差进行比较。例如，参考图 8b，如果文本样本 602 对字符“i”有当前的或观测得的计数 75，该计数属于英语的方差，因此，将有利地对英语计分。如果当前的计数为 100 或 0，它属于方差之外很远，则对英语计分不会有利。计分系统 612 可包括将预期方差或范围之外的计数降低的算法。这样，负计分是可能的。注意在这些实施例中，越好的计分是越低的，因为越低的计分表示文本样本越接近于候选语言。换句话说，观测得的字符计数是“接近于”预期计数的。这样，最低计分语言可被识别为文本语言 618。或者，可移除较高计分语言，因为它不足够“接近”，这样保留“最接近”的一个或多个候选语言。

在一则实施例中，计分系统 612 实现下述计分算法：

$$\text{最终计分}_L = \sqrt{\sum_{i=1}^{N}(\text{已存储计数}_i - \text{当前计数}_i|\times\text{降低值})^2} \qquad \text{公式 7}$$

其中最终计分 $_L$ 是一个语言的最终计分；已存储计数 $_i$ 是字符 n 的每 1000 个的预期计数；当前计数 $_i$ 是文本样本中字符 n 的计数；N 是字符的数量；降低值是倍率，诸如，如果当前计数 $_i$ 是在方差内，则为 1，如果当前计数 $_i$ 是在方差外，则为 2。这样，计算了候选列表 608 中的每个语言的计分。可基于使用公式 6 生成的最低计分来选择最大可能或最小不可能语言 618。然而值得注意的是，公式 7 的计分函数是示例性的。属于语言识别的二项式或其它概率分布的其它计分系统可被使用于语言识别目的。

在步骤 710 中，计分模块 610 基于任何确定统计置信度的已知手段计算最大可能或最小不可能语言 618 的置信计分。下文的表格示出了一种根据

本发明计分并计算置信度的方法。

表格：选择获胜语言并确定置信度

例如，使用公式 7 来计算每个样本最佳计分语言的置信值。将带有最高置信度的语言作为样本的“获胜者”返回。在处理所有样本后，从获胜者中选择最相似语言并将其作为输入文本的语言返回。

对于每个样本

计算所有语言的计分，并识别最低计分语言。将“获胜者阈值”设定为该语言的计分。该阈值等于最低计分加上该计分的一些百分比。该百分比仍然有待实验。目前，在仅仅运行字符 LAD 时，它是（最低计分＋最低计分的 10%）。当字符 LAD 运行为特里（trie-based）结构 LAD 的过滤器时，它是（最低计分＋最低计分的 25%）。这允许特里结构 LAD 考虑更多的潜在获胜者。过滤计分大于阈值的所有语言，仅仅留下最低计分者。

接下来计算每一个保留语言的置信程度。首先，将保留语言的计分相加以获得总组合计分（TotalCombinedScore）。这样，对于每个语言，查找它的计分与总组合计分之间的差值。这就是该语言的偏移量（Offset）。将所有语言的偏移量相加以获得总偏移量（TotalOffset）。这就是新的计分空间，其中越高的值越好。然后计算每个语言的偏移量作为总偏移量的百分比，即，每个语言的偏移量除以总偏移量。这提供了这些语言的置信百分比。

可能样本包含一些不使用字符分布方法而单独计分的唯一的统一编码字符。将实际处理的样本字符的百分比与每个置信值相乘。这通常是 100%，但对于不是的情况，缩放置信值以与唯一的统一字符编码语言进行比较，并已经对这些语言中的最佳计分者分配了置信值。待考虑的所有语言的置信值现在相加为 100。

- 现在存在一列带有置信值的一个或多个语言。对于不同理由调节置信值。
- 增大唯一的统一字符编码的置信值，因为期望将包含等量亚洲和西方语言（如，日语和英语）的文档识别为亚洲语言。
- 如果多个语言僵持，则增大最频繁语言的置信值。
- 基于在最近历史中发现的语言调节置信值。（也在此时更新该历史）

基于已调节的置信分数，挑选带有最高置信度的语言作为获胜者。如果存在僵持的获胜者，则根据如下顺序依次选择获胜者来打破僵持：（1）历史上最近的获胜者，（2）最频繁的语言，（3）直接挑选第一个。当仅仅运行字符 LAD 时（即，不作为过滤器），其中存在多个语言（当前运作值为 4）并且没有一个带有清楚的更高的置信度，则返回 LANG_NEUTRAL

来表示无法置信识别样本语言。返回一个语言，可能是 LANG_NEUTRAL。

对于整个文本

存储每个样本的返回语言和置信度。当完成所有样本的处理时，从这些样本中选择带有最高置信度的语言作为整体的获胜者。如果在置信值上存在僵持，则根据如下顺序依次选择获胜者来打破僵持：（1）最多的样本返回的语言，（2）历史上最近的获胜者，（3）最频繁的语言，（4）直接挑选第一个。LAD 将获胜者作为文本语言返回。也将置信值与阈值进行比较并相应地返回“可靠”（IsReliable）作为正确和错误。LAD 将获胜者和它的置信值存储为最近获胜者/置信度，以用作未来的参考。

在步骤 720 中，可选的 n 元语法语言识别模块 619 可接收文本样本 602 和已识别语言 618 以用作基于 n 元语法方法（诸如上述结合的专利中所述的）的进一步语言识别。注意步骤 720 可进一步增加精确度，尤其是在文本样本 602 相对短的情况下。在步骤 722 中，n 元语法语言识别模块 610 生成语言识别 620，它可以是基于置信度而返回的一种语言或一列语言。语言识别可被用于上述的文本样本 602 的进一步自然语言处理。

图 8a-8b 是根据本发明的概率计算模块 414 所生成的概率分布的示例。在一则实施例中，图 8a 示出在给定样本大小为 1000 个字符时的一个字符（诸如，英语中的字母“i”）的概率曲线 800。图 8a 示出最大概率 808 发生在大约计数＝72 处并且方差为±7。换句换说，当一个或多个 1000 个样本集中随机采样时，英语字符“I”很可能有 72 个字符，而不是其它计数。然而，值得注意的是当考虑英语或对英语计分时，随机采样通常会使字符计数限制在预期范围 806 之内。范围 806 包括所示的方差 802、804。

在本发明中，语言可负向计分和/或正向计分。正向计分包括使用相应与语言或与其相反的实际出现事件。负向计分包括使用相应与语言或与其相反的不出现事件。注意负向计分语言和/或正向计分语言的能力被确信相对于其它语言识别系统是有优势的，其它语言识别系统通常受限于正向证据计分系统。例如，n 元语法计分方法通常仅仅对正向证据进行计分。

如在此所使用的，“负向证据”是事件的不出现。例如，不出现事件可以是字符在文本样本中的任何地方都不出现。或者，不出现事件可以是字

符计数在预期范围内不出现，即，在字符预期范围之外发现字符计数。类似地，“正向证据”是事件的出现。该出现事件可以是字符在样本文本中出现或字符计数出现在预期范围内。还需注意的是在许多实施例中，计分方案可考虑相应与特定语言或与其相反的正向和负向证据。

进一步示例，葡萄牙语和西班牙语是非常类似的。然而，葡萄牙语包含字符“ç”但西班牙语不包含此字符。这样，如果文本样本包含字符“ç”，这就是对应于葡萄牙语的正向证据并且是与西班牙语相反的正向证据。如果文本样本不包含字符“ç”，这就是与葡萄牙语相反的负向证据并且也是对应于西班牙语的负向证据。

图 8b 类似于图 8a 但也示出可能数据置信 812、814、816。这样，字符计数 0 和 100（由数据指针 812 和 814 表示）的都属于预期范围 806 之外。这样，字符“i”在指针 814 处的英语预期范围内的不出现事件是与英语相反的负向证据。换句话说，因为“i”的预期字符计数的不出现事件对英语进行相反的负向计分。类似地，含有字符“i”的不出现事件的文本样本（如 814 所示）将导致与英语相反的负向证据计分。

相反，含有“i”计数 75 的文本样本（如 816 所示）将导致对应于英语的正向计分。换句话说，观测得的字符“i”的字符计数 75 是支持英语的正向证据，因为字符计数在预期范围内的出现事件。

训练处理

图 9 示出物理的或计算机辅助的训练处理的算法或实施例，诸如与图 4 中所示的计数器 415 相关描述的。所示的训练处理可被用于生成多种自然原的字符（如图 4 中字符列表 412 所示的）的概率或计数信息以及相关的方差（如图 4 中 416 所示的）。注意，下述变量名旨在是示例性的并且不一定是训练代码中所使用的实际变量名。对每个语言执行训练处理。训练处理的输出是字符计数的数列和每个语言的方差。

首先预计算文本中每个字符的概率。在训练中，考虑一系列等大小滑动窗口。训练可执行在 1000 个字符的窗口上，但理想地，只要窗口是等大小的就可使用其它窗口大小。窗口可以交迭或不交迭。可内部设置值来调节

窗口交迭的量，包括完全不交迭。

对于每个窗口，对每个字符出现事件的数目进行计数并存储为单独的总数。训练处理对字符计数循环，其被用于更新计数和窗口的运行总数。基于先前计算的字符概率检查每个计数以确定该计数是否高于或低于预期的每 1000 个计数。上方（即，正向）或下方（即，负向）方差总数是相应增加的。

在处理所有窗口后，多种总数值可被用于计算每个字符的每个窗口平均计数（AverageCountPerWindow）(Prob.)和平均方差（AverageVariance）。字符、字符的每个窗口平均计数和字符的平均方差可被打印到文件，该文件称为这些已存储值的运行时间资源。

注意对于每个字符，跟踪下述值：

• 对每个窗口重新计算的当前窗口计数（CharacterCount）；

• 全体值 TotalCharacterCount；

• 超过预期概率/计数的总计数超过（TotalCountAbove）；

• 总窗口超过（TotalWindowsAbove），即该字符计数超过预期概率/计数的窗口的总数；

• 低于预期概率/计数的总计数低于(TotalCountBelow)

• 总窗口低于（TotalWindowsBelow），即该字符计数低于预期概率/计数的窗口的总数；

此外，也跟踪整体的总窗口发现（TotalWindowsSeen）数。注意每个窗口平均计数值是与预计算的字符概率（预期的）近似相同的。

图 10 示出根据迸发明的语言识别的算法或实施例。对于每个输入文本，确定待考虑的样本数目。如果文本样本足够长，则可从文本中的多个点提取样本。如果文本为短，则选择仅仅一个样本，对于每个样本，首先通过去除空格和非语言专用的字符（诸如，URL 或数字）而对文本进行预处理。接下来，检查每个字符的字符范围，诸如拉丁文、西里尔文、阿拉伯文、等等。跟踪从每个范围中发现的字符的数目。在单独的模块（在此没有示出）中处理来自唯一的统一字符编码字符范围的字符。对于所有其它范围，对每个唯一的字符的出现事件的数目进行计数。取消选定任何的其字符范

围没有（或最低限度）表现在样本中的任何语言。

对于非唯一范围（即，由多个语言共享的范围）的字符而言，使用字符计数、存储数据、以及公式 6 的计分算法计算计分。确定最佳计分的语言。将这些获胜语言与来自唯一范围模块的任何获胜语言组合。最后，计算一个或多个获胜语言中的置信度，尤其是对标记获胜语言排序，获胜语言随后被返回。

图 11 是示出计算文本中每个候选语言的计分的特定实施例或算法的流程图。算法对没有被排除的语言循环。对于每个语言，访问预期字符计数和方差的队列。随后分离最小字符范围，最小字符范围包括文本中计数的所有字符以及该语言预期的所有字符。上述处理对这些字符循环。如果来自文本的计数（CurrentCount）或语言的预期计数（StoredCount）大于 0，或者如果都大于 0，使用下述公式对字符计算计分：

$$字符计分_i = (已存储计数_i - 当前计数_i) \times 降低值)^2 \qquad 公式\ 8$$

在当前计数（CurrentCount）不在字符的已存储方差中时，降低值大于 1。该计分被添加到语言总计分中。一旦对所有语言处理了所有字符，算法对总计分集进行循环并提取每个语言的总计分的平方根。由下述公式给出每个语言的最终计分：

$$最终计分_L = \sqrt{字符计分_1 + 字符计分_2 + \ldots + 字符计分_N} \qquad 公式\ 9$$

其中的术语已在上文中定义。

虽然参考特定实施例描述了本发明，但是本领域普通技术人员可以理解在形式和内容上可做出改动但不脱离本发明的精神和范围。

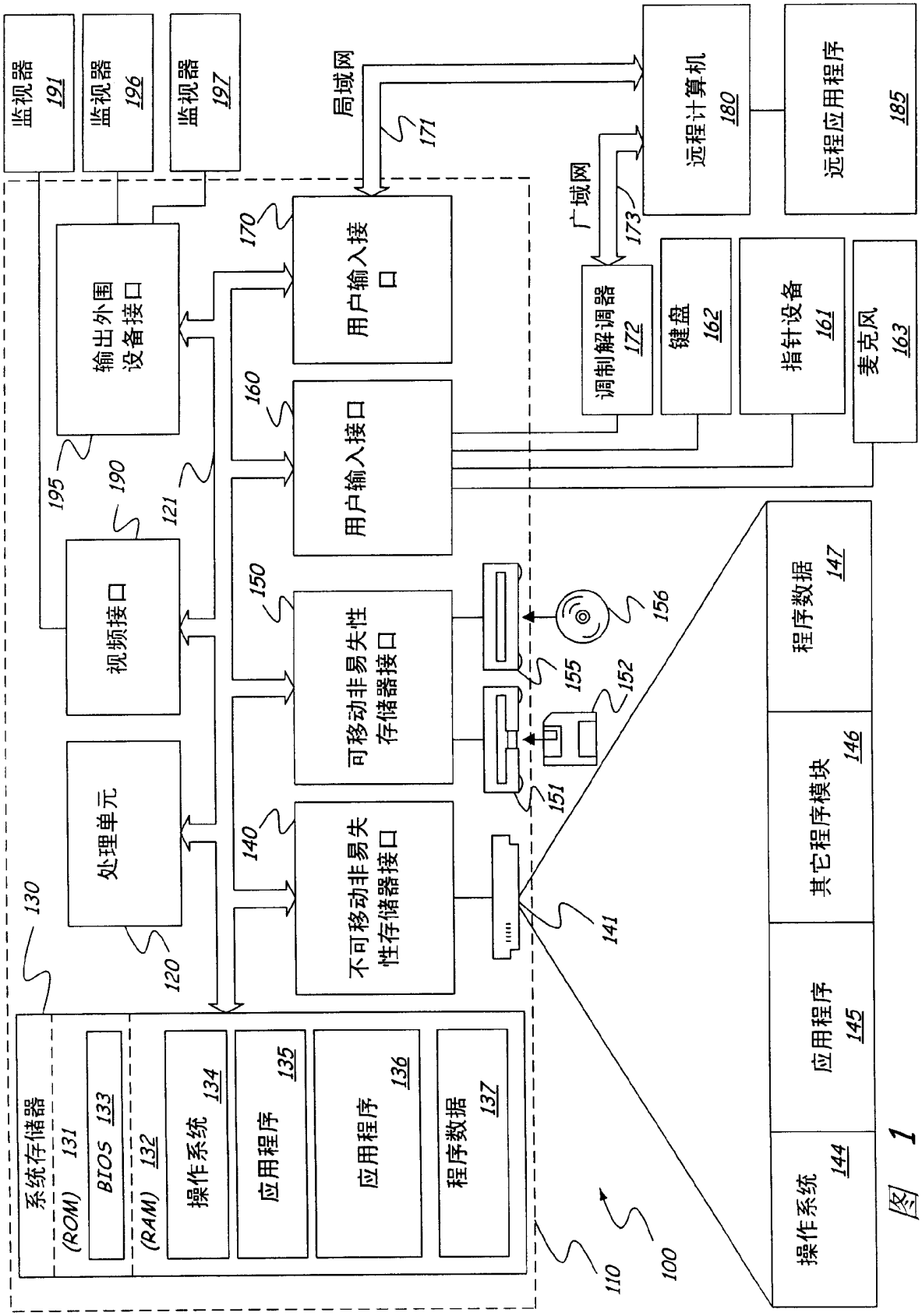

监视器 191
监视器 196
监视器 197
局域网
171
远程计算机 180
远程应用程序 185
广域网
173
输出外围设备接口
195
190
121
170
用户输入接口
160
用户输入接口
调制解调器 172
键盘 162
指针设备 161
麦克风 163
视频接口
150
可移动非易失性存储器接口
处理单元
120
140
不可移动非易失性存储器接口
130
系统存储器
(ROM) 131
BIOS 133
(RAM) 132
操作系统 134
应用程序 135
应用程序 136
程序数据 137
151
155
152
156
141
程序数据 147
其它程序模块 146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

图 1

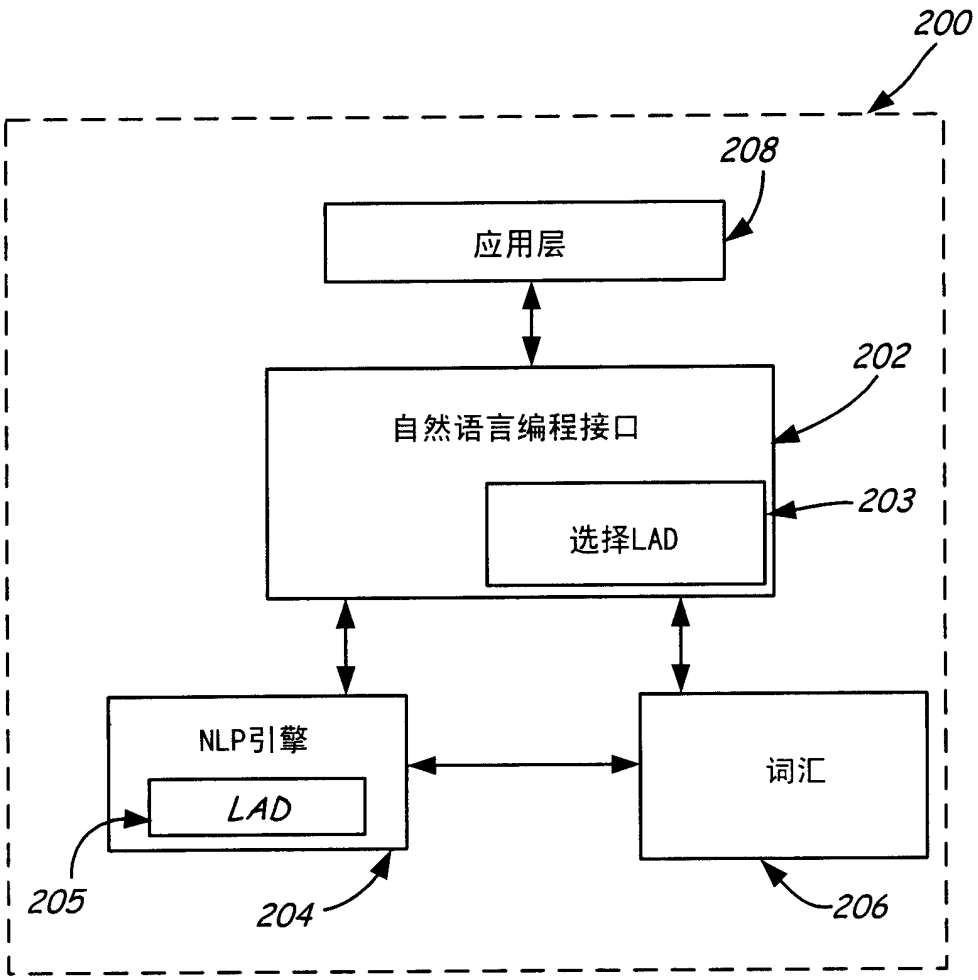


图 2

3/13

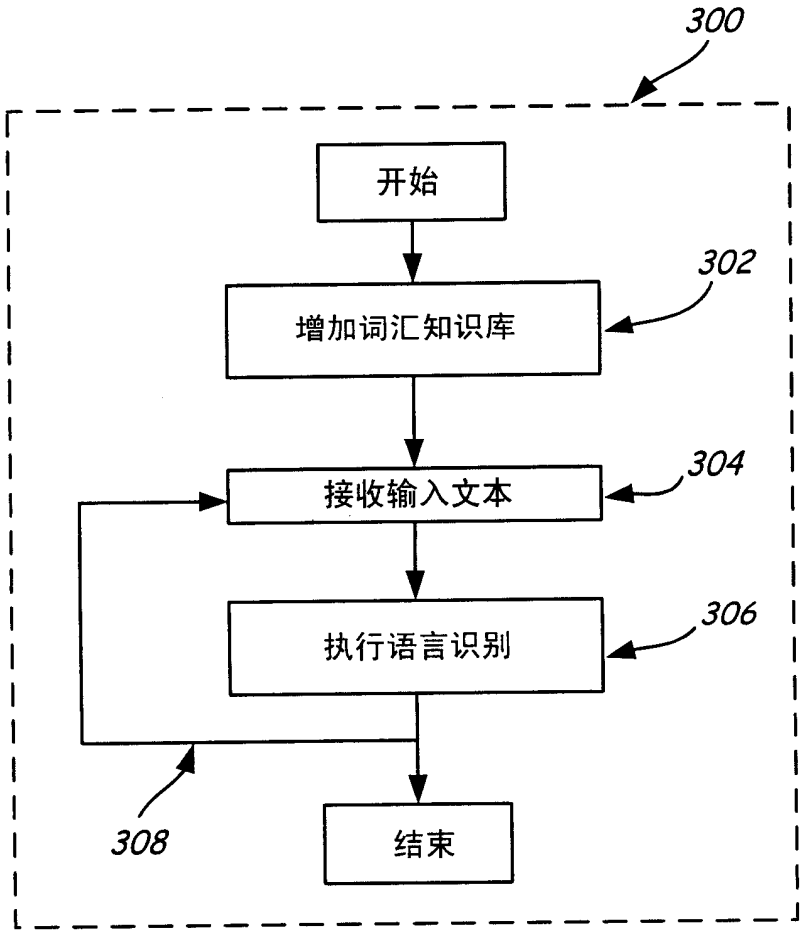


图 3

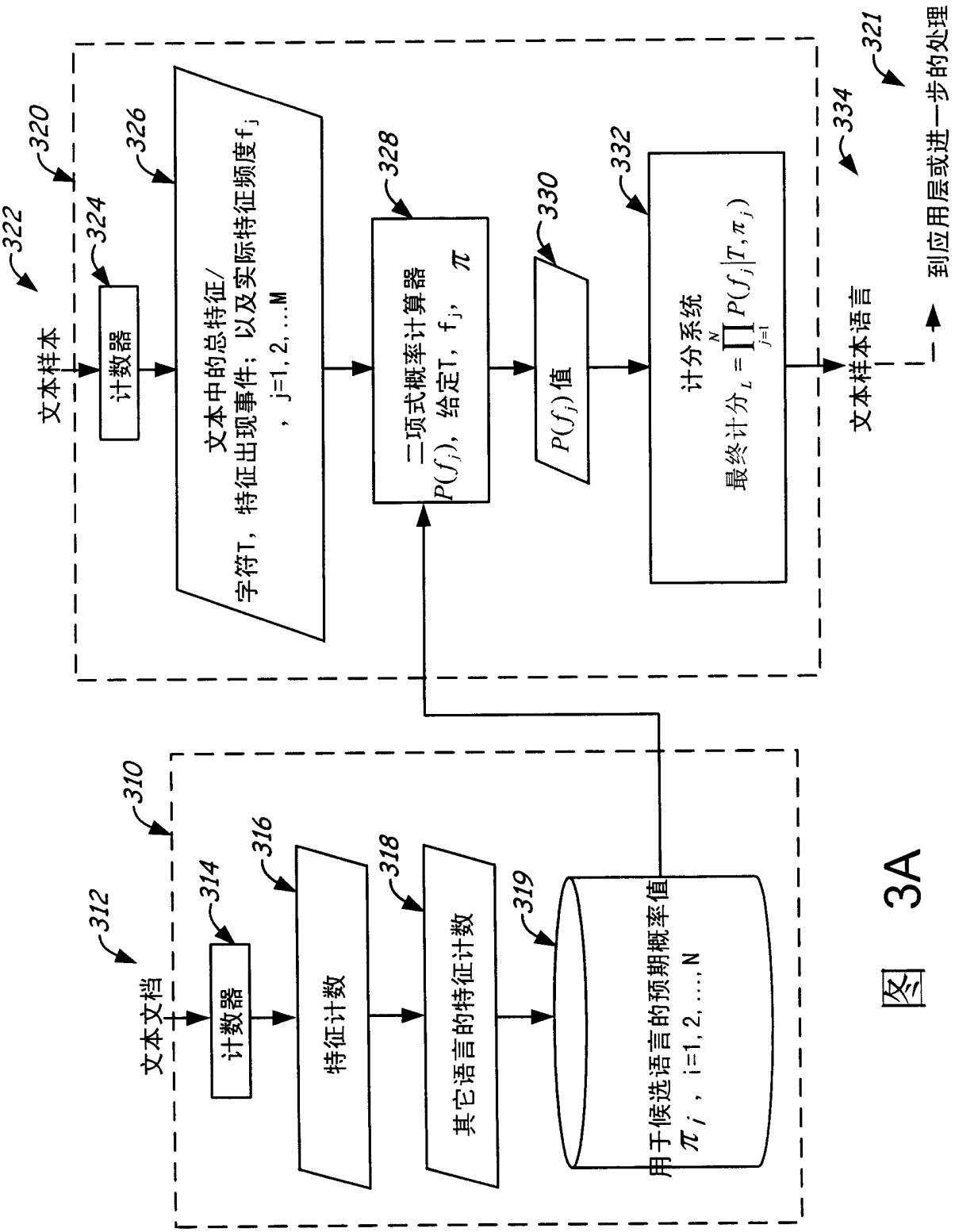

文本样本
320
322
计数器
324
文本中的总特征/
字符T，特征出现事件；以及实际特征频度 f_j
，j=1, 2, ...M
326
二项式概率计算器
$P(f_j)$，给定T，f_j，π
328
$P(f_j)$值
330
计分系统
最终计分 $L = \prod_{j=1}^{N} P(f_j|T,\pi_j)$
332
文本样本语言
334
321
到应用层或进一步的处理
文本文档
312
计数器
314
特征计数
316
其它语言的特征计数
318
用于候选语言的预期概率值
π_i，i=1, 2, ..., N
319
310

图 3A

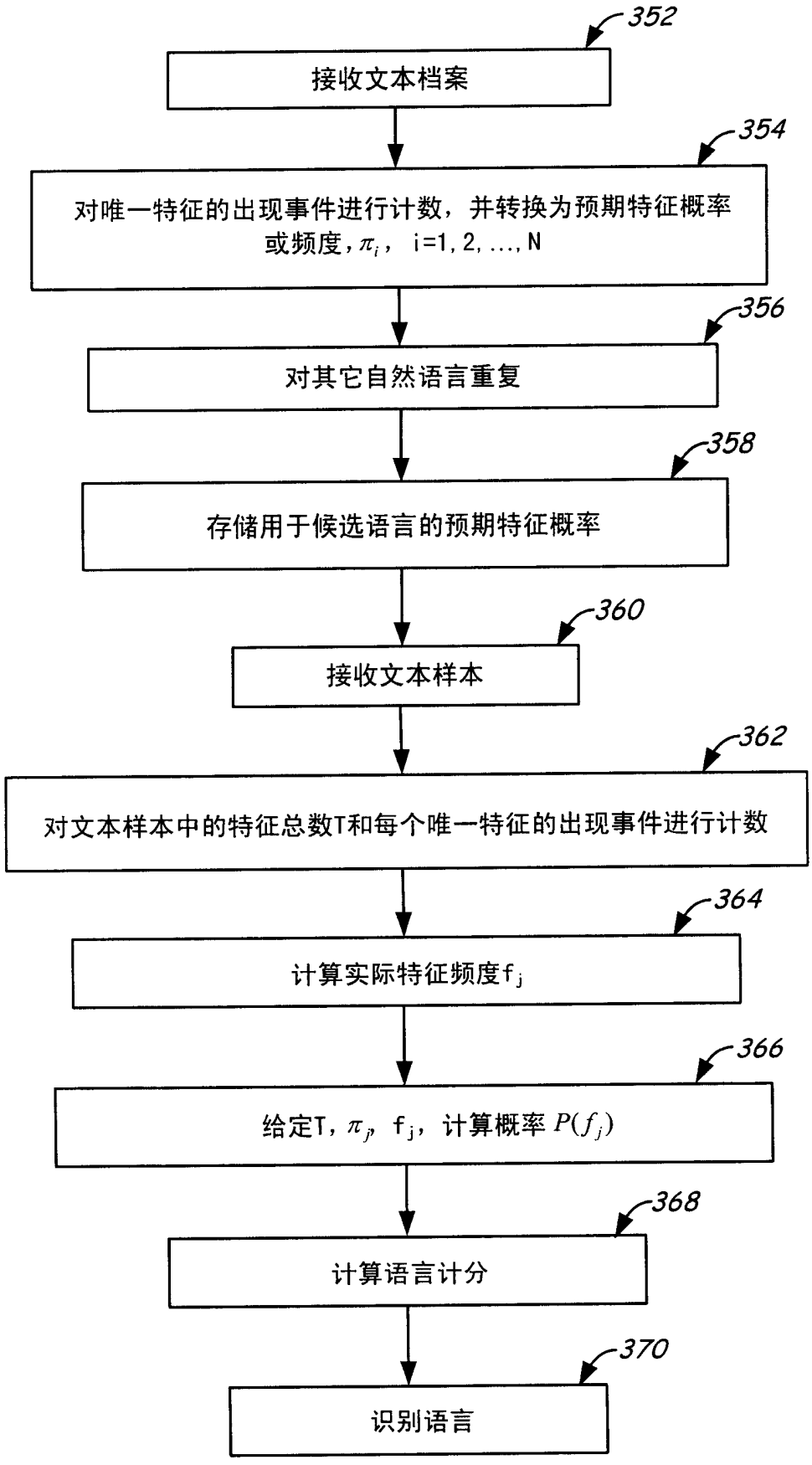

352
接收文本档案
354
对唯一特征的出现事件进行计数，并转换为预期特征概率或频度，π_i， i=1, 2, ..., N
356
对其它自然语言重复
358
存储用于候选语言的预期特征概率
360
接收文本样本
362
对文本样本中的特征总数T和每个唯一特征的出现事件进行计数
364
计算实际特征频度f_j
366
给定T，π_j，f_j，计算概率 $P(f_j)$
368
计算语言计分
370
识别语言

图 3B

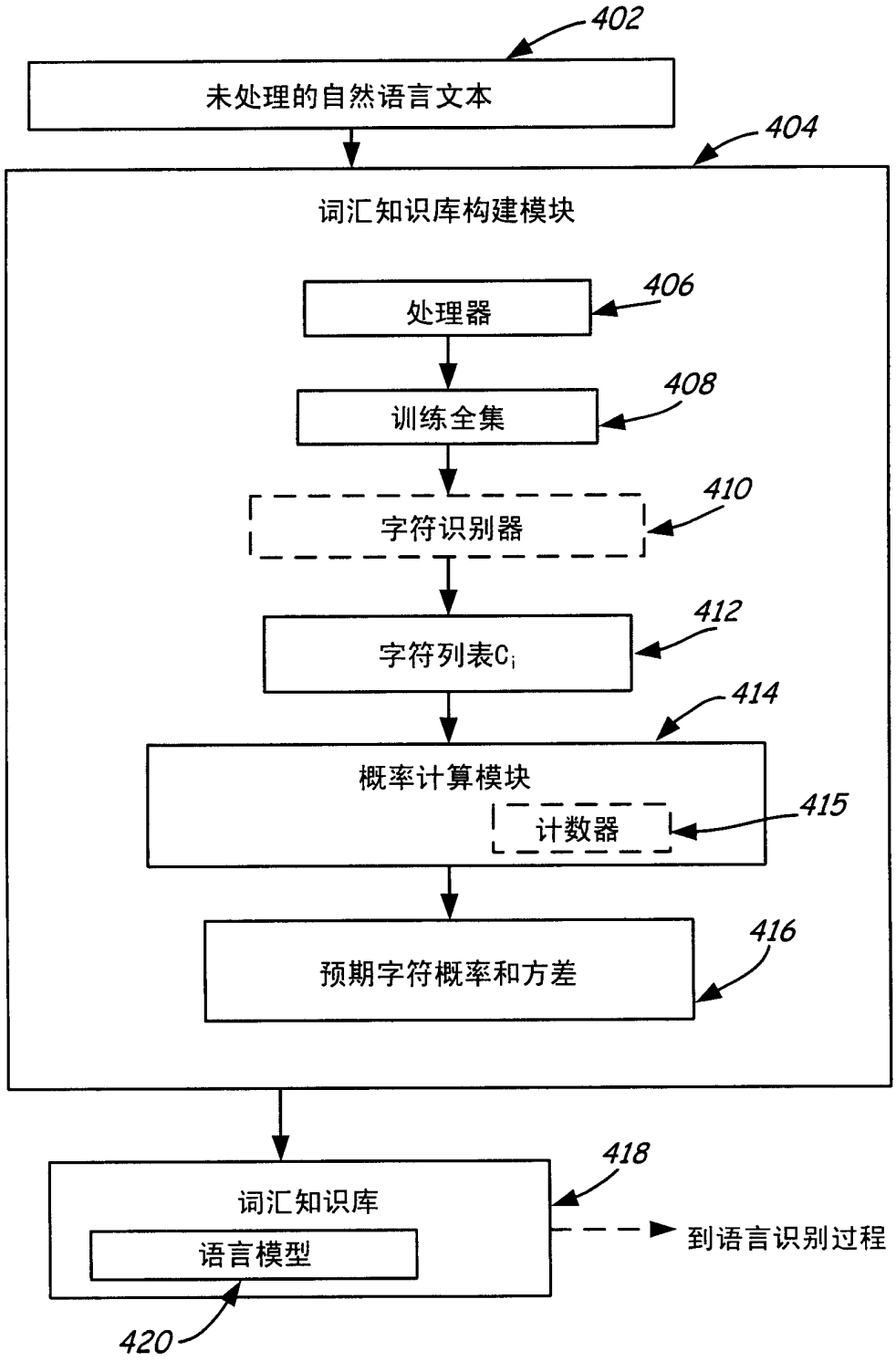

402
未处理的自然语言文本
404
词汇知识库构建模块
406
处理器
408
训练全集
410
字符识别器
412
字符列表C_i
414
概率计算模块
415
计数器
416
预期字符概率和方差
418
词汇知识库
语言模型
到语言识别过程
420

图 4

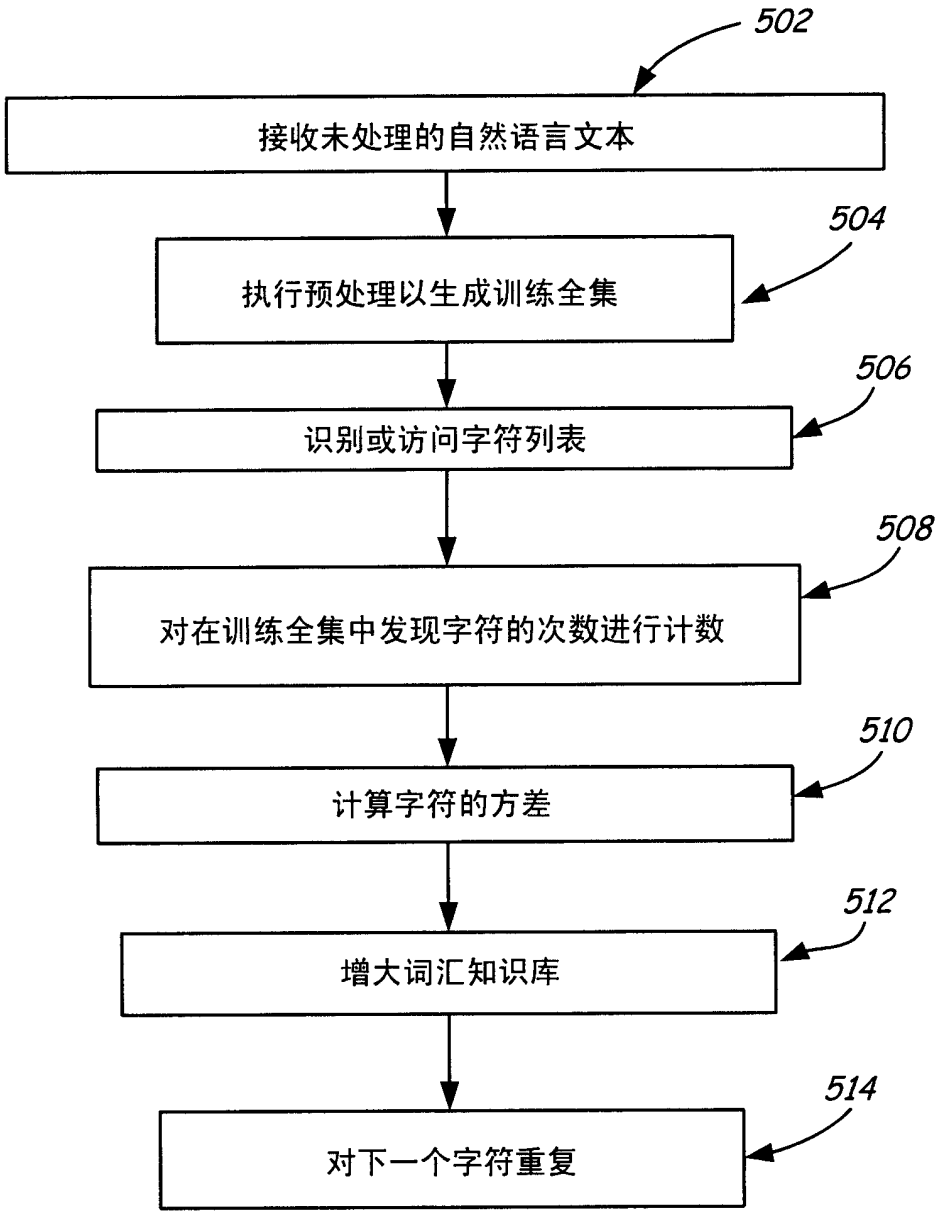

502
接收未处理的自然语言文本
504
执行预处理以生成训练全集
506
识别或访问字符列表
508
对在训练全集中发现字符的次数进行计数
510
计算字符的方差
512
增大词汇知识库
514
对下一个字符重复

图 5

8/13

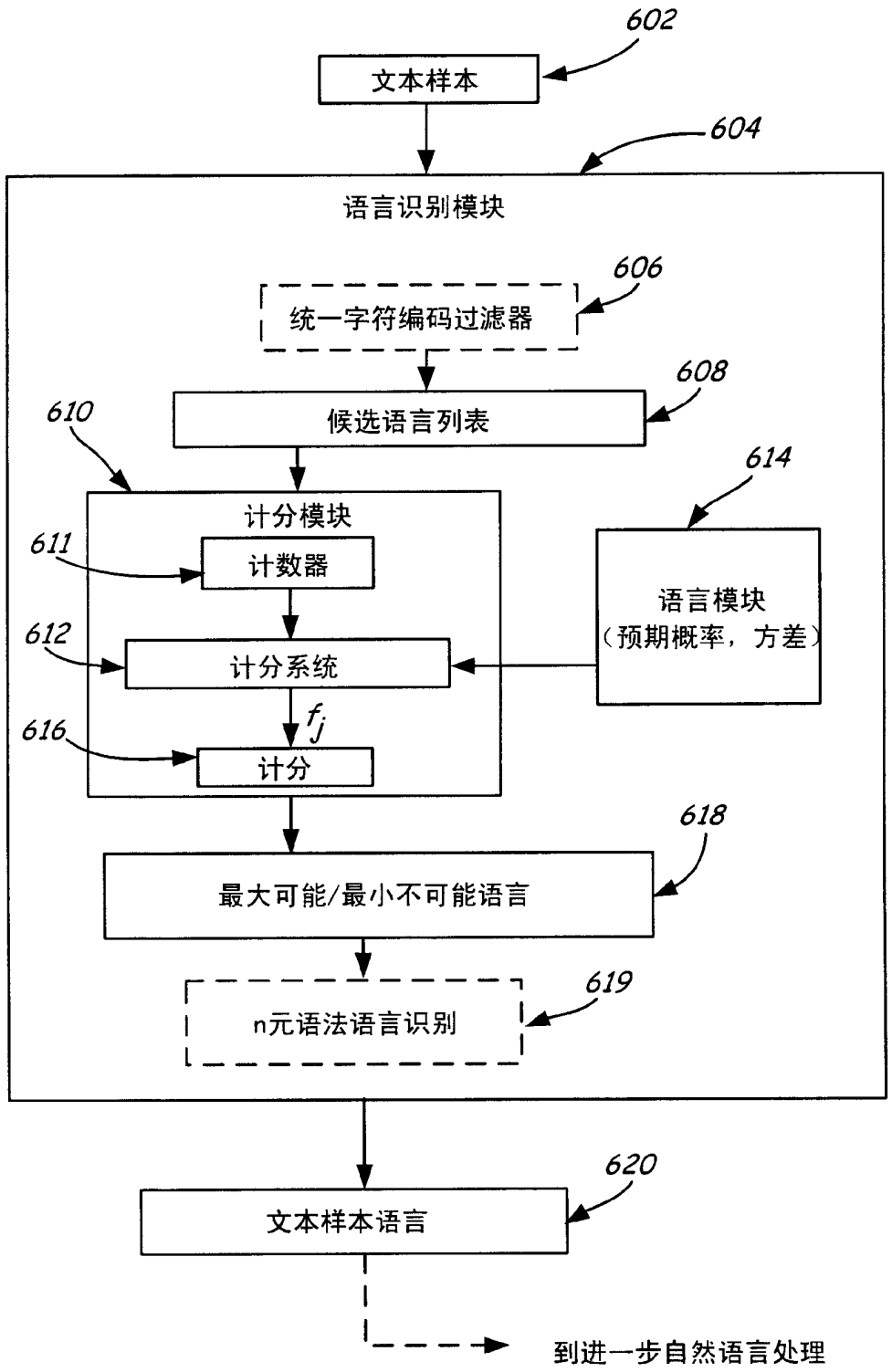

602
文本样本
604
语言识别模块
606
统一字符编码过滤器
608
候选语言列表
610
计分模块
611
计数器
612
计分系统
f_j
616
计分
614
语言模块
（预期概率，方差）
618
最大可能/最小不可能语言
619
n元语法语言识别
620
文本样本语言
到进一步自然语言处理

图 6

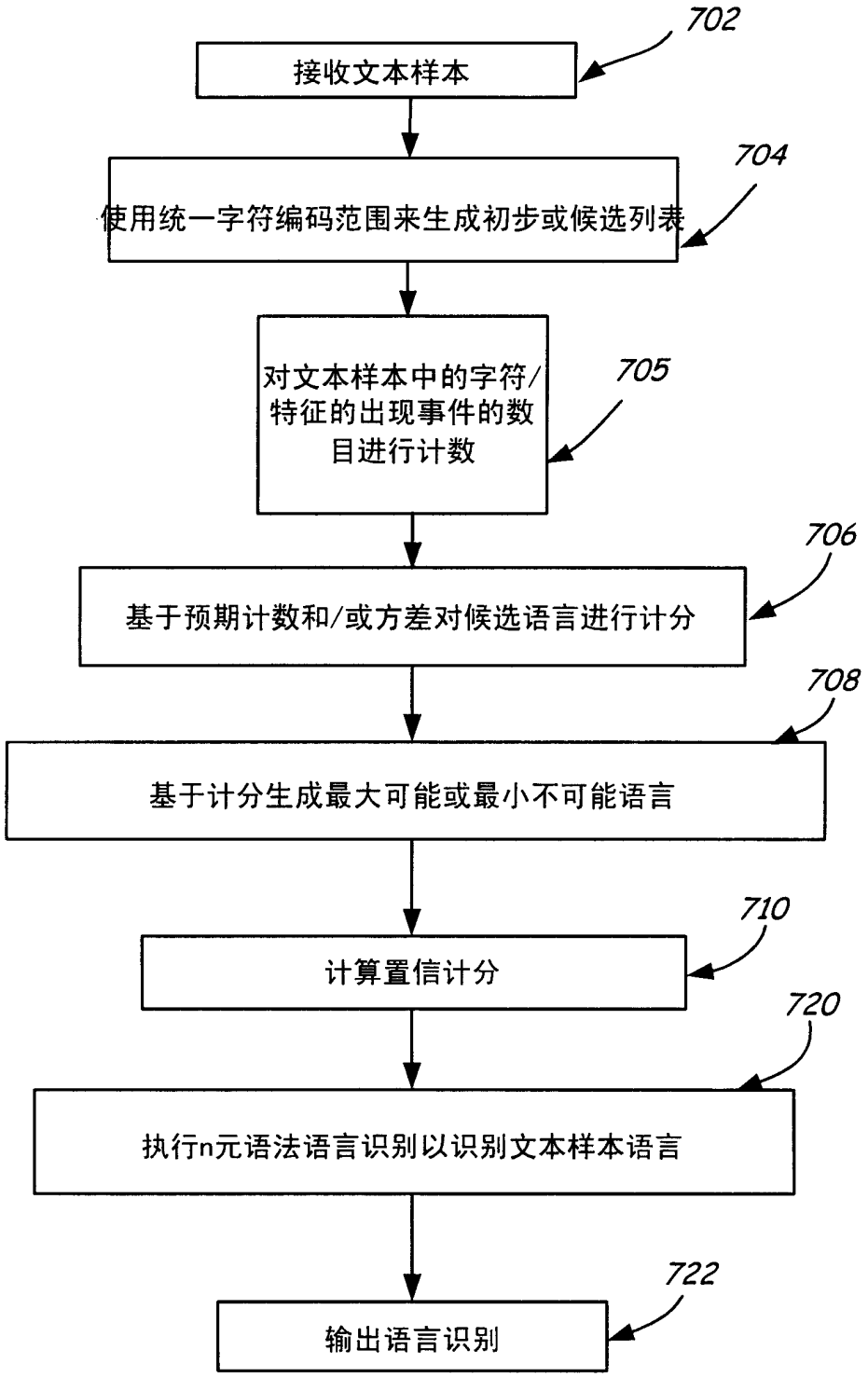

702
接收文本样本
704
使用统一字符编码范围来生成初步或候选列表
705
对文本样本中的字符/特征的出现事件的数目进行计数
706
基于预期计数和/或方差对候选语言进行计分
708
基于计分生成最大可能或最小不可能语言
710
计算置信计分
720
执行n元语法语言识别以识别文本样本语言
722
输出语言识别

图 7

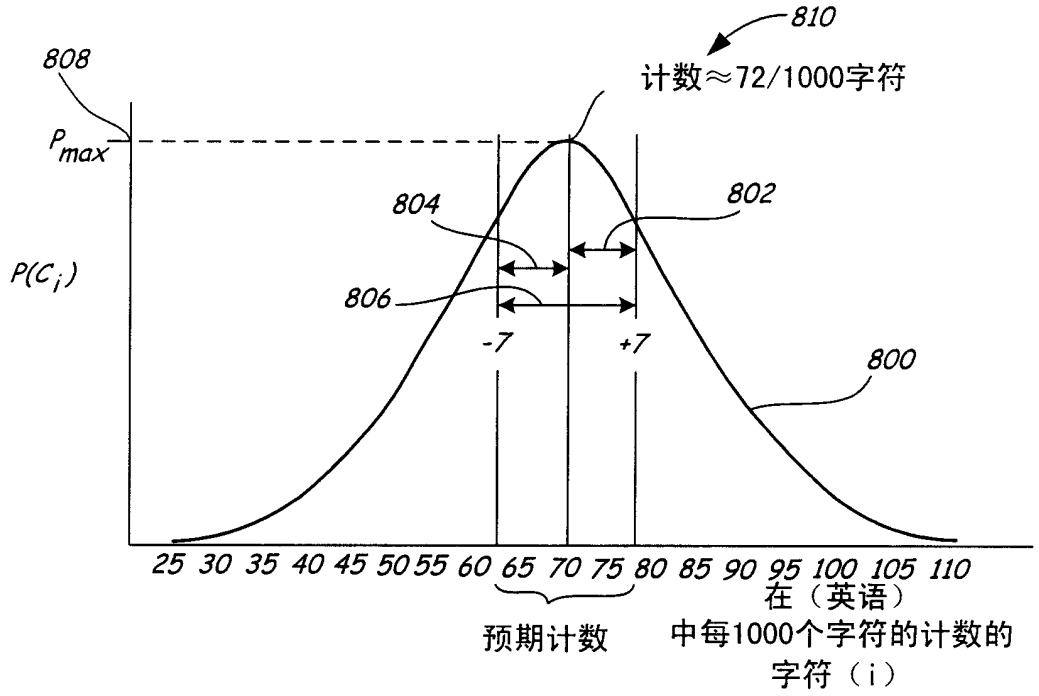

810
计数≈72/1000字符
808
P_{max}
804
802
$P(C_i)$
806
-7
+7
800
25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
预期计数
在（英语）
中每1000个字符的计数的
字符（i）

图 8a

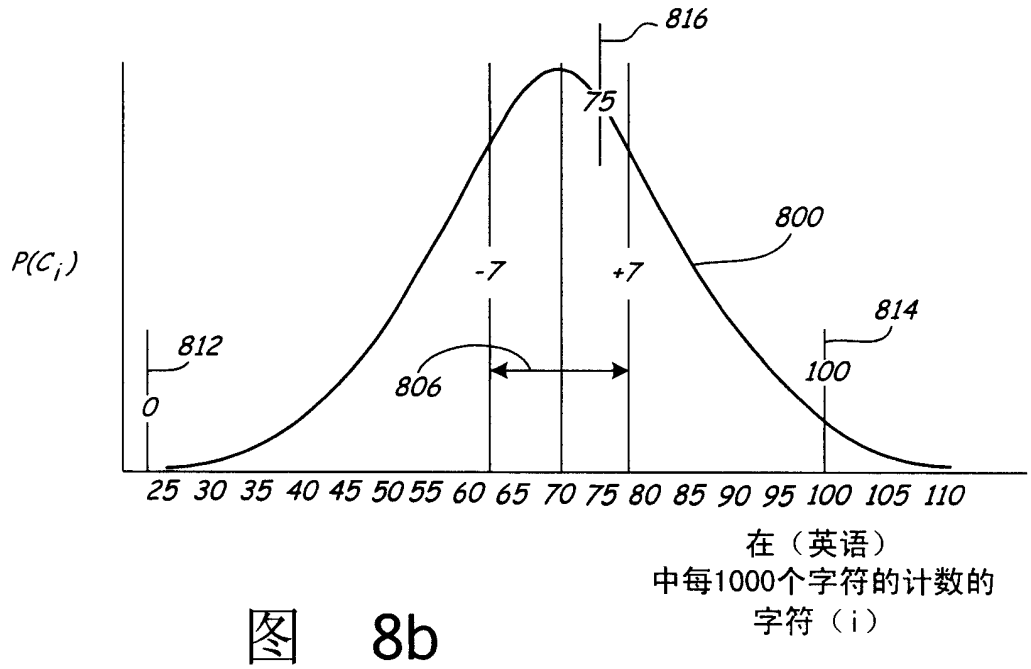

816
75
800
$P(C_i)$
-7
+7
814
812
806
100
0
25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
在（英语）
中每1000个字符的计数的
字符（i）

图 8b

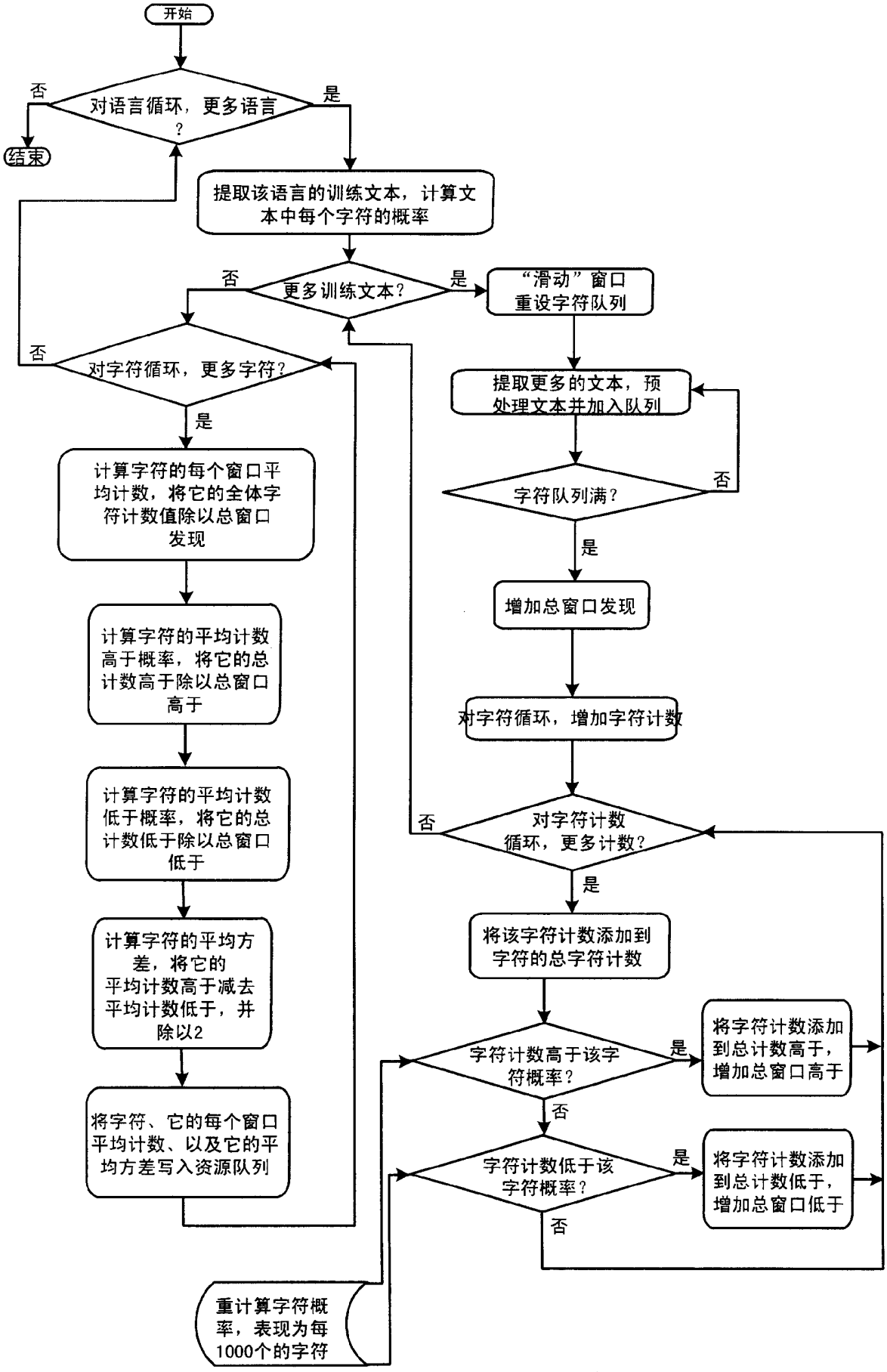

开始
对语言循环，更多语言？
否
是
结束
提取该语言的训练文本，计算文本中每个字符的概率
更多训练文本？
否
是
“滑动”窗口重设字符队列
对字符循环，更多字符？
否
是
提取更多的文本，预处理文本并加入队列
字符队列满？
否
是
计算字符的每个窗口平均计数，将它的全体字符计数值除以总窗口发现
增加总窗口发现
计算字符的平均计数高于概率，将它的总计数高于除以总窗口高于
对字符循环，增加字符计数
计算字符的平均计数低于概率，将它的总计数低于除以总窗口低于
对字符计数循环，更多计数？
否
是
计算字符的平均方差，将它的平均计数高于减去平均计数低于，并除以2
将该字符计数添加到字符的总字符计数
字符计数高于该字符概率？
是
将字符计数添加到总计数高于，增加总窗口高于
否
将字符、它的每个窗口平均计数、以及它的平均方差写入资源队列
字符计数低于该字符概率？
是
将字符计数添加到总计数低于，增加总窗口低于
否
重计算字符概率，表现为每1000个的字符

图 9

12/13

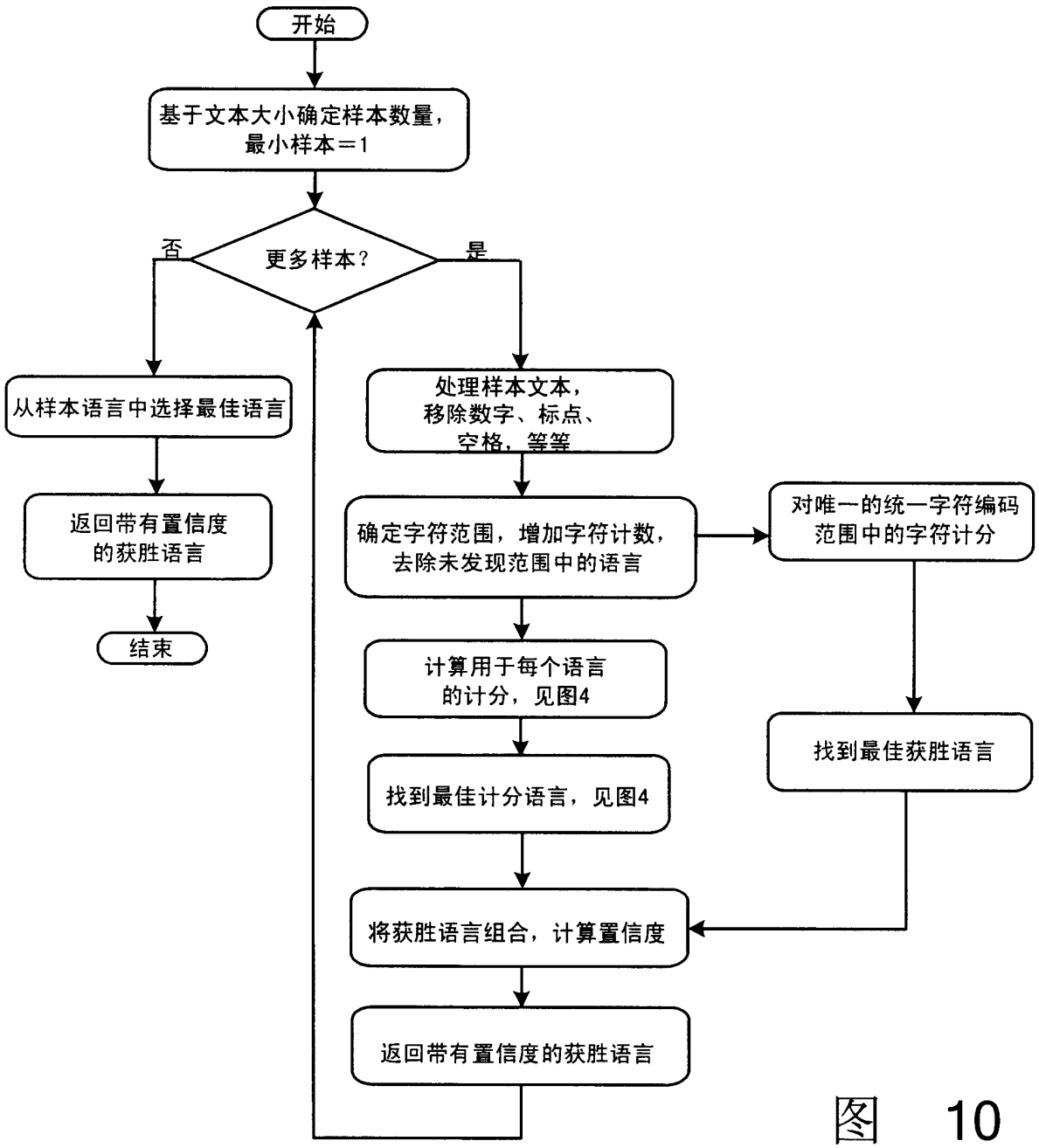

开始
基于文本大小确定样本数量，最小样本＝1
更多样本？
否
是
从样本语言中选择最佳语言
返回带有置信度的获胜语言
结束
处理样本文本，移除数字、标点、空格，等等
确定字符范围，增加字符计数，去除未发现范围中的语言
对唯一的统一字符编码范围中的字符计分
计算用于每个语言的计分，见图4
找到最佳获胜语言
找到最佳计分语言，见图4
将获胜语言组合，计算置信度
返回带有置信度的获胜语言

图 10

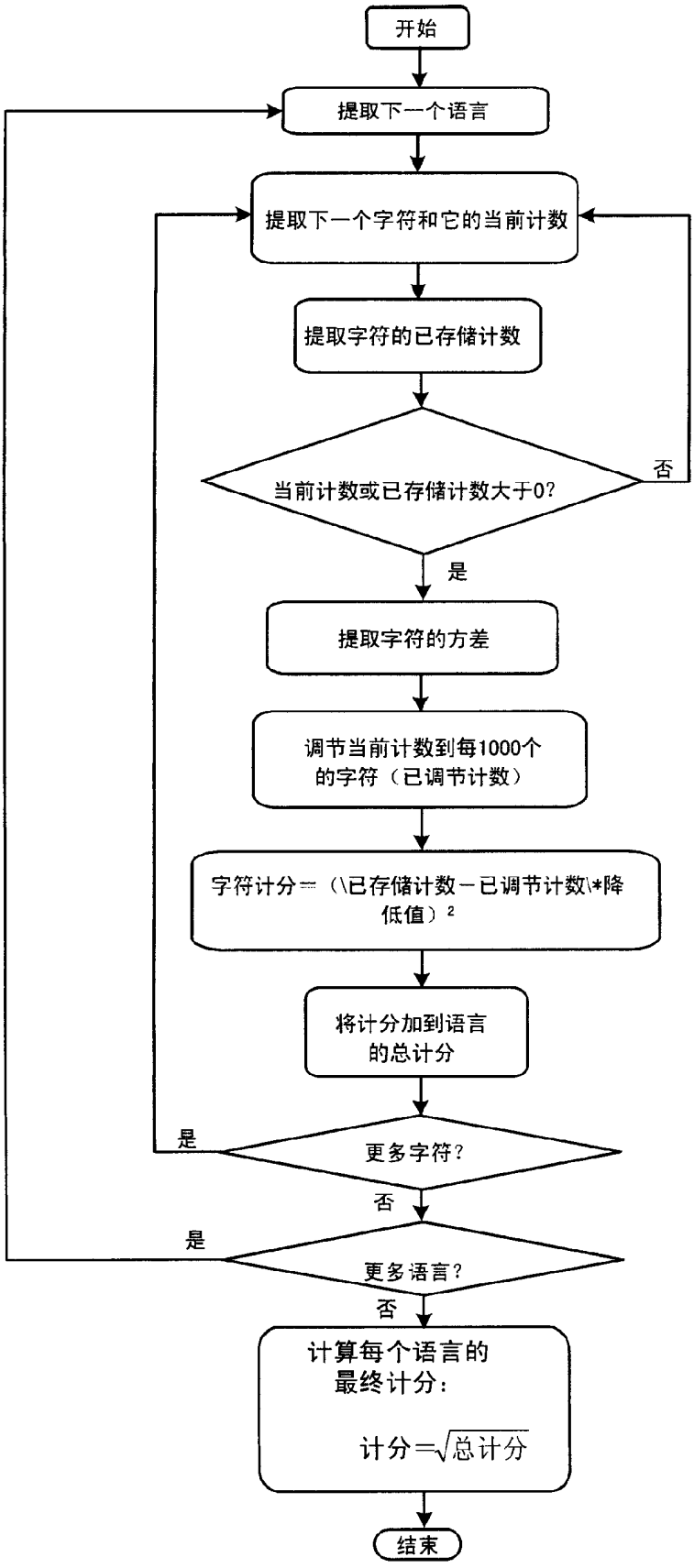

开始
提取下一个语言
提取下一个字符和它的当前计数
提取字符的已存储计数
当前计数或已存储计数大于0?
否
是
提取字符的方差
调节当前计数到每1000个的字符（已调节计数）
字符计分＝（\已存储计数－已调节计数*降低值）²
将计分加到语言的总计分
是
更多字符?
否
是
更多语言?
否
计算每个语言的最终计分：
计分＝√总计分
结束

图 11