(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101002198 B
(45)授权公告日 2013.10.23

(21)申请号 200580026350.4
(22)申请日 2005.06.21
(30)优先权数据
10/875,449 2004.06.23 US
(85)PCT申请进入国家阶段日
2007.02.02
(86)PCT申请的申请数据
PCT/US2005/022027 2005.06.21
(87)PCT申请的公布数据
WO2006/002219 EN 2006.01.05
(73)专利权人 GOOGLE 公司
地址 美国加利福尼亚州
(72)发明人 吴军 朱鸿隽 朱会灿 黄炜华 陈钊琪
(74)专利代理机构 北京康信知识产权代理有限责任公司 11240
代理人 余刚
(51)Int.Cl.
G06F 17/27(2006.01)
(56)对比文件
US 6167367 A,2000.12.26, 说明书第3栏第44行-第11栏第60行、图1,2.
US 2004/0024584 A1,2004.02.05, 全文.
CN 1143769 A,1997.02.26, 全文.
CN 1311881 A,2001.09.05, 全文.
CN 1223733 A,1999.07.21, 全文.
US 6401060 B1,2002.06.04, 说明书第5栏第63行-第8栏第9行, 第9栏第27行-第10栏第35行, 第12栏第12行-第14栏第14行、图2,3,7.
审查员 李文

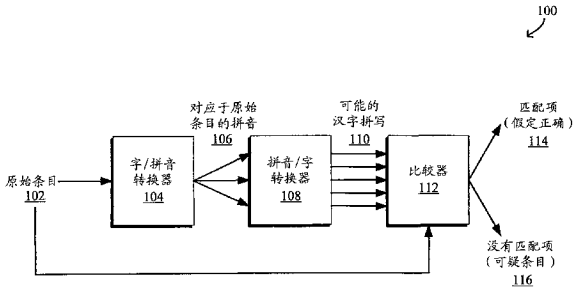
权利要求书2页 说明书7页 附图4页

(54)发明名称

用于非罗马字符和字的拼写校正系统和方法

(57)摘要

本发明披露了一种使用基于规则的分类器和隐马尔可夫模型来处理和校正诸如汉语、日语、和韩语的基于非罗马字体的字的拼写错误的系统和方法。该方法一般地包括将诸如汉语的第一语言的输入条目转换为不同于第一语言的诸如拼音的中间表示的至少一个中间条目,将中间条目转换为第一语言的输入的至少一个可能的可选拼写或形式,并且当输入条目和输入条目的所有可能可选拼写之间的匹配项分别被定位或没有定位时,确定输入条目是正确或可疑的输入条目。可以基于由变换规则发生器生成的变换规则,使用例如基于变换规则的分类器对可疑输入条目进行分类。



CN 101002198 B

1. 一种用于字的拼写校正方法，包括：

接收与第一语言相关的第一字符表示的多个输入条目；

确定一个或多个可疑条目，包括，对于每个输入条目：

生成一个或多个与所述第一语言相关的第二字符表示的中间条目，每个中间条目为所述输入条目的一个表示，其中，所述第二字符表示不同于所述第一字符表示；

从所述一个或多个中间条目生成所述输入条目的一个或多个可能的候选形式，其中，所述输入条目的所述一个或多个可能的候选形式为与所述第一语言相关的所述第一字符表示；

将所述输入条目与所述输入条目的所述一个或多个可能的候选形式中的每一个进行比较以定位匹配项；以及

当未从所述一个或多个可能的候选形式中定位到匹配项时，确定所述输入条目是可疑条目；

使用所确定的可疑条目和相应的一个或多个候选形式来生成和训练一组拼写校正变换规则，其中，每个拼写校正变换规则均与置信度测量结果相关联，使得具有较高置信度测量结果的拼写校正变换规则被应用在具有较低置信度测量结果的规则之后。

2. 根据权利要求 1 所述的方法，其中，所述第一字符表示是基于非罗马的字符表示。

3. 根据权利要求 1 所述的方法，其中，所述第一字符表示是汉字并且所述第二字符表示是拼音。

4. 根据权利要求 1 所述的方法，其中，每个输入条目是查询日志中的用户查询。

5. 根据权利要求 1 所述的方法，还包括：

基于所述一组拼写校正变换规则将每个可疑条目分类为正确拼写的条目或不正确拼写的条目。

6. 根据权利要求 5 所述的方法，其中，所述分类由基于变换规则的分类器执行。

7. 根据权利要求 5 所述的方法，还包括：

使用利用所述可疑条目和所述相应的一个或多个候选形式的变换规则发生器，生成和训练所述拼写校正变换规则。

8. 根据权利要求 7 所述的方法，其中，使用可疑条目的数据库自动执行所述生成和训练所述拼写校正变换规则。

9. 根据权利要求 5 所述的方法，其中，以自动或人工监控来执行所述分类。

10. 根据权利要求 5 所述的方法，还包括：

维持停止规则模式的用户可编辑表，所述停止规则模式禁止对用户输入和候选拼写的特定结合提供拼写校正建议或拼写校正。

11. 一种用于字的拼写校正系统，包括：

第一转换器，用于接收与第一语言相关的第一字符表示的输入条目以生成一个或多个与所述第一语言相关的第二字符表示的中间条目，每个中间条目为输入条目的表示，其中，所述第二字符表示不同于所述第一字符表示；

第二转换器，用于从所述一个或多个中间条目生成所述输入条目的一个或多个可能的候选形式，其中，所述输入条目的所述一个或多个可能的候选形式为与所述第一语言相关的所述第一字符表示；

第一比较器，用于将所述输入条目与所述输入条目的所述一个或多个可能的候选形式中的每个可能的候选形式进行比较以定位匹配项，所述比较器还被用于当未从所述一个或多个可能的候选形式定位到匹配项时确定所述输入条目是可疑输入条目；以及

变换规则生成器，用于使用所述可疑输入条目和相应的一个或多个候选形式以生成和训练拼写校正变换规则，其中，每个拼写校正变换规则均与置信度测量结果相关联，使得具有较高置信度测量结果的拼写校正变换规则被应用在具有较低置信度测量结果的规则之后。

12. 根据权利要求 11 所述的系统，其中，所述第一字符表示是基于非罗马的字符表示。

13. 根据权利要求 11 所述的系统，其中，所述第一字符表示是汉字并且所述第二字符表示是拼音。

14. 根据权利要求 11 所述的系统，其中，由第一转换器接收的输入条目是查询日志中的用户查询。

15. 根据权利要求 11 所述的系统，其中：

分类器，包括所述变换规则生成器，用于基于所述拼写校正变换规则将所述可疑输入条目分类为正确拼写的条目或不正确拼写的条目。

16. 根据权利要求 15 所述的系统，其中，所述分类器是基于变换规则的分类器。

17. 根据权利要求 15 所述的系统，其中，包括所述变换规则生成器的所述分类器，用于使用所述可疑输入条目和所述相应的一个或多个候选形式来生成和训练所述拼写校正变换规则。

18. 根据权利要求 17 所述的系统，其中，所述变换规则生成器使用可疑输入条目的数据库自动生成所述变换规则。

19. 根据权利要求 15 所述的系统，其中，所述分类器以自动或人工监控来执行分类。

用于非罗马字符和字的拼写校正系统和方法

技术领域

[0001] 本发明一般地涉及处理基于非罗马字体的语言。更具体地，涉及用于使用基于规则的分类器和隐马尔可夫 (Markov) 模型处理和校正诸如汉语、日语、和韩语的基于非罗马字体的字的拼写错误的系统和方法。

背景技术

[0002] 拼写校正通常包括检测错字并且为错字确定适当的替换。在依字母顺序的即诸如英语的基于罗马字体的语言中，大部分拼写错误是词表外的词，例如，“thna”而不是“than”，或有效的字不适当地用在其上下文中，例如“stranger then”而不是“stranger than”。检测和校正基于罗马字体的语言中的词表外拼写错误的拼写检验器是众所周知的。

[0003] 然而，诸如汉语、日语、和韩语 (CJK) 的基于非罗马字体的语言在任何计算机字符集（例如，UTF-8 字符集）中都没有编码的有效字符，使得大部分拼写错误是不适当地用在上下文中的有效字符，而不是词汇拼写错误。在汉语中，字的正确使用通常只能在上下文中确定。因此，用于基于非罗马字体的语言的有效拼写检验器应该使用上下文信息来确定上下文中哪个字符和 / 或字是不合适的。

[0004] 对于诸如 CJK 语言的非罗马字体语言的拼写校正也是复杂的和具有挑战性的，由于这种语言没有标准词典，因为 CJK 字的定义 不清楚。例如，某些可能将汉语中的“Beijing city”看作一个词，而其他的可能将它们看作两个词。与此相反，在英文拼写校正中，英文词典 / 单词表查找是关键特征，因此，英文拼写校正方法不能很容易适用于 CJK 语言。另外，与英文的 26 个字母不相同，有几千个常用的汉字，从而使得由所有选择替换非法汉字中不正确的字符，然后确定新创建的词是否是适当的很不实际。而且，汉语具有大量同形词和同音字以及引起多义性的不可见的（或隐藏的）字边界，这也使得高效和有效的汉语拼写校正很复杂并且难于执行。很明显，由于汉语和英语之间的这种差别，英文拼写校正可用的许多有效技术不适于汉语拼写校正。

[0005] 从而需要一种用于有效、高效和准确的检测以及校正诸如汉语、日语和韩语的非罗马字体语言中的拼写错误的计算机系统和方法。

发明内容

[0006] 本发明披露了一种使用基于规则的分类器和隐马尔可夫模型处理和校正诸如汉语、日语、和韩语的基于非罗马字体的字中的拼写错误的系统和方法。特别地，该系统和方法使用变换规则、隐马尔可夫模型以及易混淆字符的相似性矩阵。在汉语拼写检查应用中，如果简体汉字或繁体汉字中的字符具有相同的发音和 / 或共享一些输入按键，则在一对易混淆字符之间的相似性 (similarity) 可为正数。否则，该值为零。在一个实施例中，相似性可具有布尔值，例如，1 用于一对易混淆字符，0 用于一对非易混淆字符。该系统和方法特别可应用于基于网络的搜索引擎和在客户端站点 (site) 可下载的应用程序，例如，在工具栏或桌面任务栏中执行，但是可应用于各种其他的应用。应该理解，可以通过许多方式实施

本发明，包括诸如过程、设备、系统、装置、方法或诸如计算机可读存储介质或计算机网络的计算机可读介质，其中，程序指令通过光或电子 通信线路被发送。术语“计算机”通常指的是诸如个人数字助理 (PDA)、移动电话、和网络交换机的具有计算能力的任何装置。下面描述本发明的几个独创性实施例。

[0007] 本方法一般地包括将诸如汉语的第一语言的输入条目 (entry) 转换为不同于第一语言（诸如拼音）的中间表示的至少一个中间条目，将中间条目转换为第一语言的输入的至少一种可能选择拼写，并且当分别定位或没有定位出输入条目和输入条目的所有可能选择拼写之间的匹配项时，确定输入条目是正确的或可疑的输入条目。如此处所使用的，“pinyin”指的是所有简体或繁体汉字的拼音符号，包括注音符号 (Bopomofo，汉语拼音字母)，即，“带注释的语音的符号”。可以根据中间表示中的公共标记限定成对的第一语言的易混淆字符之间的相似性。可以基于由变换规则发生器生成的变换规则，使用例如基于变换规则的分类器来分类可疑输入条目。可以类似地采用诸如决策树和神经网络分类器的各种其他分类器。

[0008] 转换可包括转换多个输入条目，例如查询日志中的用户查询。该方法还可以包括，:基于诸如拼写校正变换规则的一组规则，例如通过基于变换规则的分类器，将可疑条目分类为正确拼写或不正确拼写的条目。用户的表决 (vote)，例如，查询日志和 / 或网页，被优选地应用以生成变换规则。该方法还可以包括利用可疑输入条目和可能的选择拼写使用变换规则发生器来生成和训练拼写校正变换规则。该方法还包括接收第一语言的用户输入，确定是否有任何规则应用于用户输入，当确定至少一个规则应用于用户输入时，生成对应于用户输入的第一语言的至少一个候选拼写，比较用户输入的相似性与用户输入的至少一个候选拼写的相似性，并且使用用户输入（具有比用户输入更高的相似性）的至少一个候选拼写给出拼写校正建议和 / 或进行拼写校正。

[0009] 系统通常包括 :第一转换器，用于将第一语言的输入转换为输入条目的至少一个中间表示，中间表示不同于第一语言 ;第二转换器，用于将中间表示转换为第一语言的输入的至少一个可能的选择拼写，通过将可能的选择拼写与输入条目相比较来定位匹配项，并且如果没有从所有可能的选择拼写中定位出匹配项，则确定输入条目是可疑的输入条目，如果定位出匹配项，则输入条目是正确的输入条目。

[0010] 一种计算机程序产品，用于结合计算机系统使用，计算机程序产品具有其上存储有计算机处理器可执行的指令的计算机可读存储介质，指令通常包括 :接收第一语言的输入条目，将输入条目转换为输入条目的至少一个中间表示，中间表示不同于第一语言，将中间表示转换为第一语言的至少一种可能的选择拼写，通过将输入条目的至少一个可能的选择拼写与输入条目进行比较来定位匹配项，并且如果没有从所有可能的选择拼写中定位出匹配项，则确定输入条目是可疑的输入条目，如果定位出了匹配项，则输入条目是正确的输入条目。

[0011] 执行本系统和方法的应用程序可以在服务器站点上（例如在搜索引擎上）执行或在诸如用户计算机的客户端站点上执行（例如，下载的），以提供对输入到文档中的文本的拼写校正，或与诸如搜索引擎的远程服务器进行交互。客户端站点应用程序可选地包括停止规则模式的用户可编辑表，其允许用户通过指定某些拼写校正是被禁止的（例如，决不替换 X 和 Y，除非 X 在 Z 前或跟在 Z 之后）来定制应用程序。

[0012] 本发明的这些以及其他的特征和优点将在下列具体描述和通过实例示出了本发明的原理的附图中更详细地展示。

附图说明

[0013] 通过下列结合附图的详细描述，将很容易地理解本发明，其中附图中相同的参考标号表示相同的结构元件。

[0014] 图 1 是用于执行正向转换到基于非罗马字体语言的中间形式和从基于非罗马字体语言的中间形式反向转换，以确定可疑的原始输入的可能候选拼写的说明性的系统和方法的框图。

[0015] 图 2 是用于从条目集合生成拼写校正变换规则的说明性的系统和方法的框图。

[0016] 图 3 是示出用于自动生成拼写校正变换规则的过程的流程图。

[0017] 图 4 是示出利用用于处理条目以确定拼写校正建议（如果有的话）的变换规则的过程的流程图。

具体实施方式

[0018] 本发明披露了使用基于规则的分类器和隐马尔可夫模型处理和校正基于非罗马字体的字（例如汉语、日语、韩语）的系统和方法。应该注意，只是为了清楚起见，此处展示的实例可应用于汉语拼写错误检测和校正，更具体地是简体汉字拼写错误检测和校正。然而，用于拼写错误检测和校正的系统和方法可类似地应用于其他基于非罗马字体的语言（例如繁体汉语、日语、韩语、泰国语等）。提出下列描述以使得本领域任何技术人员都能够实现和使用本发明。提供具体实施例的描述和应用只是作为实例，对于本领域技术人员来说，各种修改将是显而易见的。在不脱离本发明的精神和范围的情况下，此处限定的一般原理可应用到其他实施例和应用。因此，本发明将被给予包括大量选择、修改以及与此处披露的原理和 特征一致的等同物的最宽的保护范围。为了清楚起见，没有详细描述关于在涉及本发明的技术领域中已知的技术材料，以免不必要地模糊本发明。

[0019] 此处描述的系统和方法一般地涉及使用由输入条目生成的拼写校正变换规则来处理和校正非罗马字体语言中的拼写错误的系统和方法。如此处所使用的，术语“拼写”指的是词表外的字符或字以及在上下文中不适当使用的有效字符或字。另外，术语输入的候选拼写或候选形式用在此处指的是不同于输入但是与输入相同语言的字符和 / 或字的集合，无论输入是单字符或字、字符和 / 或字的系列或集合、短语、句子等。从输入条目中识别可疑输入条目，并且由图 1 中示出的可疑输入条目检测器生成可能的候选拼写。使用可疑输入条目和由可疑输入条目检测器生成的可能的候选拼写作为输入，然后生成和训练拼写校正变换规则并且通过如图 2 中所示的变换规则生成器和分类器将可疑条目分类为正确的或不正确的。本系统和方法使用变换规则、隐马尔可夫模型和易混淆字符的相似性矩阵。在汉语应用中，如果在简体汉字或繁体汉字 (traditionalChinese) 中字符具有相同的发音和 / 或共享一些输入按键，则一对易混淆字符之间的相似性可为正数。否则，值为零。在一个实施例中，相似性可具有布尔值，例如，1 用于一对易混淆字符，0 用于一对非易混淆字符。图 4 的流程图中示出了使用拼写校正变换规则的训练集合来识别拼写错误和生成建议的拼写校正的过程。这样，通过使用输入集合来训练变换规则，最普通的拼写错误和校正可

以被确定和处理以增强拼写检查和校正系统的效率和有效性。

[0020] 图 1 是用于执行正向转换到简体汉字的中间形式（例如，拼音）和从简体汉字的中间形式反向转换，以识别可疑原始输入并确定可疑原始输入的可能的候选拼写的示意性可疑输入条目检测器 100 的框图。图 1 中示出的可疑输入条目检测器 100 利用拼音是用于简体 汉字的普遍使用的输入方法这个方便的事实。然而，可以实现和使用基于罗马字体或基于非罗马字体的任何其他中间形式。类似地，可疑输入条目检测器 100 适于使用各种其他基于非罗马字体的语言。

[0021] 如图 1 所示，字 - 拼音转换器 104 将每个汉字的原始条目 102 转换为一个或多个对应于原始条目 102 的发音或拼音 106。然后，拼音 - 字转换器 108 将拼音 106 转换为可能的拼写 110。可以采用用于将第一语言的文本转换为中间表示，然后转换回第一语言的其他的合适转换器 104、106。拼音仅是汉字或简体汉字的方便的中间表示。比较器 112 将都为第一语言的原始条目 102 和可能拼写 110 进行比较，以确定是否存在匹配项。如果原始条目 102 与由拼音 - 字转换器 108 输出的可能拼写 110 中的一个相匹配，则原始条目 102 匹配，假定被正确地拼写 114。然而，如果原始条目 102 不与由拼音 - 字转换器 108 输出的可能拼写 110 中的任何一个相匹配，则原始条目 102 是可疑条目 116，即，它可能是不正确的。

[0022] 拼音是主要用于输入简体汉字的语音输入方法。如此处所指出的，拼音通常指的是汉字的语音表达 (phonetic representation)，具有或没有与汉字相关的音调的表示。特别地，“拼音”指的是简体或繁体汉语的所有语音符号，包括注音符号 (Bopomofo，汉语拼音字母)，即，“带注释的语音的符号”。

[0023] 拼音使用罗马字体字符并且具有以多音节字形式列出的词汇。因为汉语具有大量同形字和同音字，所以每个原始条目 102 可以通过字 - 拼音转换器 104 转换为多个拼音 106，并且类似地，每个拼音 106 通过拼音 - 字转换器 108 可以被转换为汉字 110 的多个可能拼写。特别地，由于只有大约 1300 个具有音调的不同语音音节（这可以由拼音表示），以及大约 400 个没有表示好几万汉字 (Hanzi) 的语音音调的语音音节，一个语音音节（具有或没有音调）可对应于许 多不同汉字。例如，普通话中“yi”的发音可对应于超过 100 个汉字。因此，考虑到为同形字和 / 或同音字的汉字的巨大比例，由字 - 拼音转换器 104 和拼音 - 字转换器 108 执行的将每个原始条目 102 转换为拼音 106 然后转换回汉字 110 的过程可能并不是很平常的。

[0024] 此处描述的系统和方法使用变换规则、隐马尔可夫模型和易混淆字符的相似性矩阵。中文应用中，如果字符具有相似的发音、共享相似的输入按键、和 / 或拼写相似（即，视觉上相似），则一对易混淆字符之间的相似性可为正数。否则，值为零。在一个实施例中，相似性可具有布尔值，例如，1 用于一对易混淆字，0 用于一对非易混淆字。在第一语言的一对易混淆字符之间的相似性可根据中间表示中的公共标记而被限定。

[0025] 可以实施用于将汉字转换为拼音和用于将拼音转换为汉字的各种机制。例如，各种译码器可适于将拼音转换为汉字（中文字符）。在一个实施例中，可以实施使用隐马尔可夫模型的 Viterbi 译码器。例如，可以通过收集经验计数或通过计算期望值并且执行迭代最大化处理，来实现隐马尔可夫模型。Viterbi 算法是根据马尔可夫通信信道的输出观察来译码源输入的有用、高效的算法。Viterbi 算法已经成功地在用于自然语言处理（例如语音识别、光学字符识别、机器翻译、语音标记 (tagging)、解析和拼写检查）的各种应用中实

施。然而，应该理解，代替马尔可夫假设，在执行译码算法中可以做出其他各种适当的假设。另外，Viterbi 算法只是一种可以由译码器执行的适当译码算法，并且也可以执行各种其他适当的译码算法，例如有限状态机、Bayesian 网络、决策平面算法（高维 Viterbi 算法）或 Bahl-Cocke-Jelinek-Raviv(BCJR) 算法（两通道正向 / 反向 Viterbi 算法）。

[0026] 由可疑输入条目检测器 100 检测到的可疑条目通常基本包括所有拼写错误。然而，可疑条目也通常包括较高的假报警 / 假的正比率， 即，被标记为不正确的正确查询数量与不正确查询的数量的比率。这将在下面进行更详细地描述，然后，由可疑条目检测器 100 确定的可疑查询 116 可被分类为正确的或不正确的。分类器可为基于变换规则的分类器，这是优选的，或者可以为决策树分类器、神经网络分类器等。对于分类为正确的条目，不给出建议。对于分离为不正确的条目，可根据每个可能的选择拼写的相似性给出拼写校正建议。

[0027] 图 2 是用于从由可疑条目检测器 100 处理的原始条目集合 102 生成拼写校正变换规则的示意性系统和方法 120 的框图。特别地，原始条目集合 102 可包括诸如网络搜索引擎的查询日志的用户输入条目和 / 或例如从诸如那些互联网上可用的文档导出的条目。在用户输入条目的情况下，原始输入集合 102 可包括例如来自过去三周或两个月的用户查询集合。文档实例可包括诸如报纸、书籍、杂志、网页等的网络内容和各种出版物。原始输入集合 102 可从文档（例如，以互联网上可用的简体和 / 或繁体汉字写的文档）的组、集合或储存库导出。应该指出，此处所述的示意性系统和方法特别可应用于网络搜索引擎的上下文中，并且可应用于包括有组织的数据的数据库的搜索引擎中。然而，应该理解本系统和方法可以被修改和用于拼写错误检测和校正的各种其他应用，特别用于在非罗马字体语言中的条目。例如，本系统和方法适于 CJK 文本输入应用，例如，检测和校正拼写错误的字处理应用。

[0028] 变换规则发生器和分类器 120 由 Eric Brill 引入的基于变换的学习算法 (learning algorithm)，该算法在训练过程中，根据来自训练数据（例如，人注释的不正确拼写）的置信度测量结果自动提取（学习）和排列变换规则。这些变换规则由注释器 / 表决器 124 使用。注意，变换规则不同于语言学中使用的语法规则，因为变换规则是基于统计学而不是语言学知识。因此，例如，如果大部分条目以相同 的错误方式错误地拼写某些字，则错误的拼写将被分类为正确的。关于基于变换规则的方法的其他信息出现在 2004 年 1 月 27 日公布的 Eric Brill 的第 6684201 号，题为“Linguistic Disambiguation Systemand Method Using String-Based Pattern Training to Learn to ResolveAmbiguity sites”的美国专利，其全部内容结合与此作为参考。因此，变换规则发生器 120 通过利用用户的表决自动地（即，无人监控地）生成规则。换言之，字符模式的正确性是根据数据库中表决的多数来确定的，例如，查询日志，而不是人注释的数据。

[0029] 每个变换规则均与置信度测量结果相关联，使得具有较高置信度测量结果的规则被应用在具有较低置信度测量结果的规则之后。例如，如果 B 在 X 之前，则第一变换规则可以指定用 Y 替换 X。如果 E 在 Y 之后，则具有较高置信度测量结果的第二变换规则可指定用用 X 替换 Y。因此，第一变换规则将首先被应用到条目 BXE 以生成 BYE。然后，第二变换规则将被应用到所得到的条目 BYE 以将该条目转换回 BXE。这是很明显的，变换规则被应用的顺序会影响结果。还应该注意，被替换的字符和替换字符可以为条目的任何成分，而不必

是字。类似地，条件可以基于任何上下文，词性 (part-of-speech) 标记或语法非末端标签（例如，NP 用于名词短语）。还要注意，虽然基于变换规则的分类器是优选的，然而，可以类似地实施简单 Bayesian 分类器、决策树分类器、神经网络分类器、或任何多种其他合适的分类器，以分类可疑条目 116。

[0030] 回到图 2，如所示，由可疑条目检测器 100 输出的每个可疑条目 116 和其相应的可能的候选拼写 110 通过拼写校正变换规则发生器 120 的注释器 (annotator)124 接收。注释器 124 最初基于初始变换规则 126 并且最终基于提取的和排列的变换规则 130 来分类条目 128。

[0031] 学习阶段可被监督（即，由人）和 / 或无人监督。在一个实施例中，少数普通人工创建的变换规则的初始集合被用于自动地注释可疑条目的小集合，一些人监控或通过利用用户的表决而没有人监控。在初始学习阶段之后，生成其他变换规则，优选地还有一些人监控，并且其他可疑条目被注释。所得到的规则（其例如使用较少的规则管理相当数量的用户通信量）可被看作非常可靠的，并且因此对应于高置信度测量结果。注意，由于具有较高置信度的规则通常比具有较低置信度的那些规则具有较小的覆盖率，所以具有高置信度的规则和具有较低置信度的规则都被使用。

[0032] 为了成本效率，可自动地生成较大数量的剩余可疑条目（例如占用户通信量的较小比例），无须人监督。一种用于自动生成这种规则的示意性过程 150 如图 3 的流程图所示。特别地，对于在循环 152 中的每个可疑查询 Q 以及对于在循环 154 中的每个相应候选拼写 Q'，在框 156 处，将 Q 和候选拼写 Q' 进行比较，以确定 Q 中的字符可能是不适当以及它们的替代 C'。在框 158 处，具有 C 的前 N 个字符以及后 N 个字符的宽度为 2N+1 的窗口被打开。注意，可以实施任何合适的上下文的长度，例如 2N+1，并且在有疑问的字符之前或之后的上下文的长度可以但不需要相等。所有 C_{-N},...C,...,C_{N} 的子序列 (pre-C, C, post-C) 的频率 F(pre-C, C, post-C) 均被计数以确定规则是有效的 (significant)，即，规则是否能够覆盖可疑条目中拼写错误的合理大比例。如果 $1 \leqslant s1 < s2... < sj < k$，则字符串 $S = x_{s1}, x_{s2}, ..., x_{sj}$ 是字符串 $X = x_1, x_2, ... x_k$ 的子序列。

[0033] 接下来，在框 160 处，通过替换 C 和 C'，确定相应的频率。然后决策框 162 通过使用查询日志和网页，即，用户表决，来确定规则是否是可靠的。如果确定规则是可靠的，则提取变换规则（即，用 C' 替换给定 pre-C 和 post-C 的 C)。特别地，如果满足下列条件则认为变换规则是可靠的：

[0034] $F(pre\text{-}C, C, post\text{-}C) > T1$ 以及

[0035] $F(pre\text{-}C, C, post\text{-}C)/F(pre\text{-}C, C, post\text{-}C) > T2$，

[0036] 其中，T1 是最小有效阈值，T2 是最小置信阈值。如上所述，通过利用用户表决使得根据数据库（即，查询日志）中的表决的多数而不是人注释的数据确定字符模式的正确性，由变换规则发生器执行的过程 150 自动（即，无人监督地）生成规则。

[0037] 因为最常见的变换规则将支配错误模式的非常大的一部分，规则集合的大小优选地不随可疑条目的数量迅速增加。也可设置每个规则的最小具体值 (minimum occurrence) 以限制变换规则集合的大小。

[0038] 实施此处描述的系统和方法的应用程序可以在诸如搜索引擎的服务器站点上执行，或可以在诸如终端用户的计算机的客户端站点（例如，下载的）上执行，以对输入到字

处理文档中的文本提供拼写校正，或与诸如搜索引擎的远程服务器相互作用。客户端站点应用程序可以在例如工具栏内执行，并且可以可选地包括停止规则模式的用户可编辑表，该表允许用户通过指定某些拼写校正是被禁止的（例如，决不替换 X 和 Y，除非当 X 在 Z 之前或 Z 之后）来定制应用程序。例如，一些汉字，例如“买”和“卖”，在该语言中具有相同的发音“mai”（但是音调不同）并且具有几乎相同的语法角色，然而具有完全不同的意义。许多自动拼写规则生成程序倾向于将“买”改为“卖”，反之亦然。终端用户可以在停止规则模式表中指定停止规则“(X,Y)”，以防止拼写校正应用程序用 Y 替换 X。

[0039] 图 4 是示出利用用于处理条目的变换规则以确定拼写校正建议（如果有的话）的过程 200 的流程图。决策框 202 确定是否有任何拼写校正规则适用于用户输入。为了执行决策框 202，可以检查拼写校正变换规则的哈希表，以确定是否有任何变换规则适用于用户输入。例如，对于给定汉语用户输入 ABCDE，如果变换规则指示如果 C 前面的字符是 AB，则用 C’替换字符 C，然后该特定规则可应用于用户输入。如果没有规则可应用于用户输入，不为用户输入做出拼写校正建议。可选地，在框 204 处，对于每个可应用于用户输入的拼写校正变换规则，生成对应于可应用的拼写变换规则的候选拼写。在上述实例中，为对应于可应用的拼写校正变换规则的用户输入 ABCDE 生成候选拼写 ABC’DE。

[0040] 在决策框 206 处，确定每个候选拼写的相似性，并且与用户输入的相似性进行比较。在一个实施例中，决策框 206 可利用隐马尔可夫模型和 Viterbi 译码器以计算相似性。在当前实例中，ABCED 和 ABC’DE 的相对输出概率被确定和比较。如果下式成立，则候选拼写具有比用户输入更高的相似性，因此被看作有效校正：

[0041] P(ABC’DE)×P（变换规则）＞P(ABCDE)，

[0042] 其中，P（变换规则）可被限定为成功校正数量和校正总数量的比率。注意，P(ABCDE) 应该考虑分割的多义性。例如，如果 ABCDE 有两个可能的分割 AB-CDE 和 ABC-DE，则概率是 Bayesian 概率乘积的和：

[0043] P(ABCDE) ＝P（输入－结束 /CDE)×P(CDE/AB)×P(AB/ 输入－开始)+P（输入－结束 /DE)×P(DE/ABC)×P(ABC/ 输入－开始）

[0044] 注意，上面的方程是通过应用马尔可夫假设（其通过前面的字而不是通过整个历史来确定当前字）由原始 Bayesian 概率导出的 Bayesian 概率。可以类似地确定 P(ABC’DE)。

[0045] 如果给定的候选拼写不比在决策框 206 处所确定的用户输入更合适，则不给出特定拼写校正建议。然而，如果给定的候选拼写比在在决策框 206 处所确定的用户输入更合适，则在框 208 处建议和 / 或自动生成用户输入的相应候选拼写。

[0046] 此处描述的用于拼写校正的系统和方法特别适合用于基于非罗马字体的语言，并且在检测拼写错误和生成候选拼写建议或校正方面非常有效。另外，用于拼写校正的系统和方法还特别可应用在网络搜索引擎环境中以及应用到包括有组织的数据的数据库的搜索引擎中，执行各种用户输入或查询的拼写校正。

[0047] 虽然此处示出和描述了本发明的代表性实施例，然而应该理解它们只是说明性的，并且在不脱离本发明的精神和范围的情况下，可以对这些实施例进行修改。因此，本发明的范围将只由下列可以被修改的权利要求的术语所限定，其中每个权利要求均被特意地结合到该具体实施方式部分作为本发明的实施例。

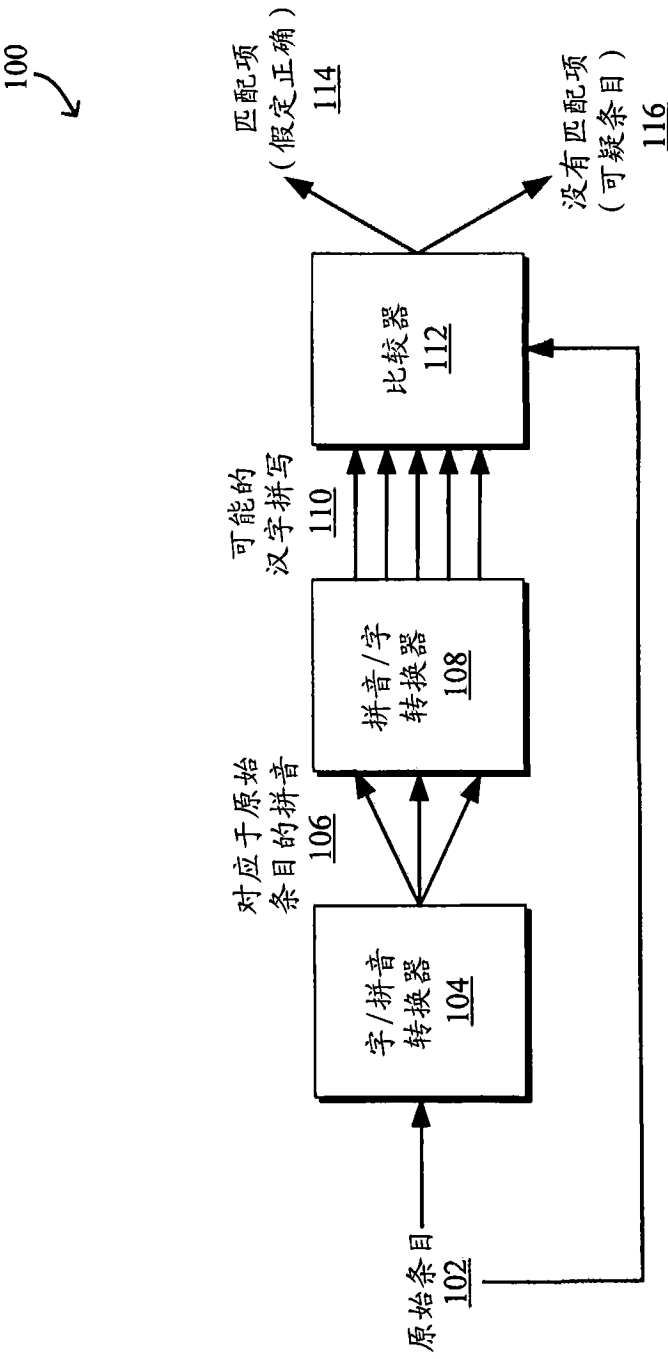

100
原始条目
102
字/拼音
转换器
104
对应于原始
条目的拼音
106
拼音/字
转换器
108
可能的
汉字拼写
110
比较器
112
匹配项
(假定正确)
114
没有匹配项
(可疑条目)
116

图 1

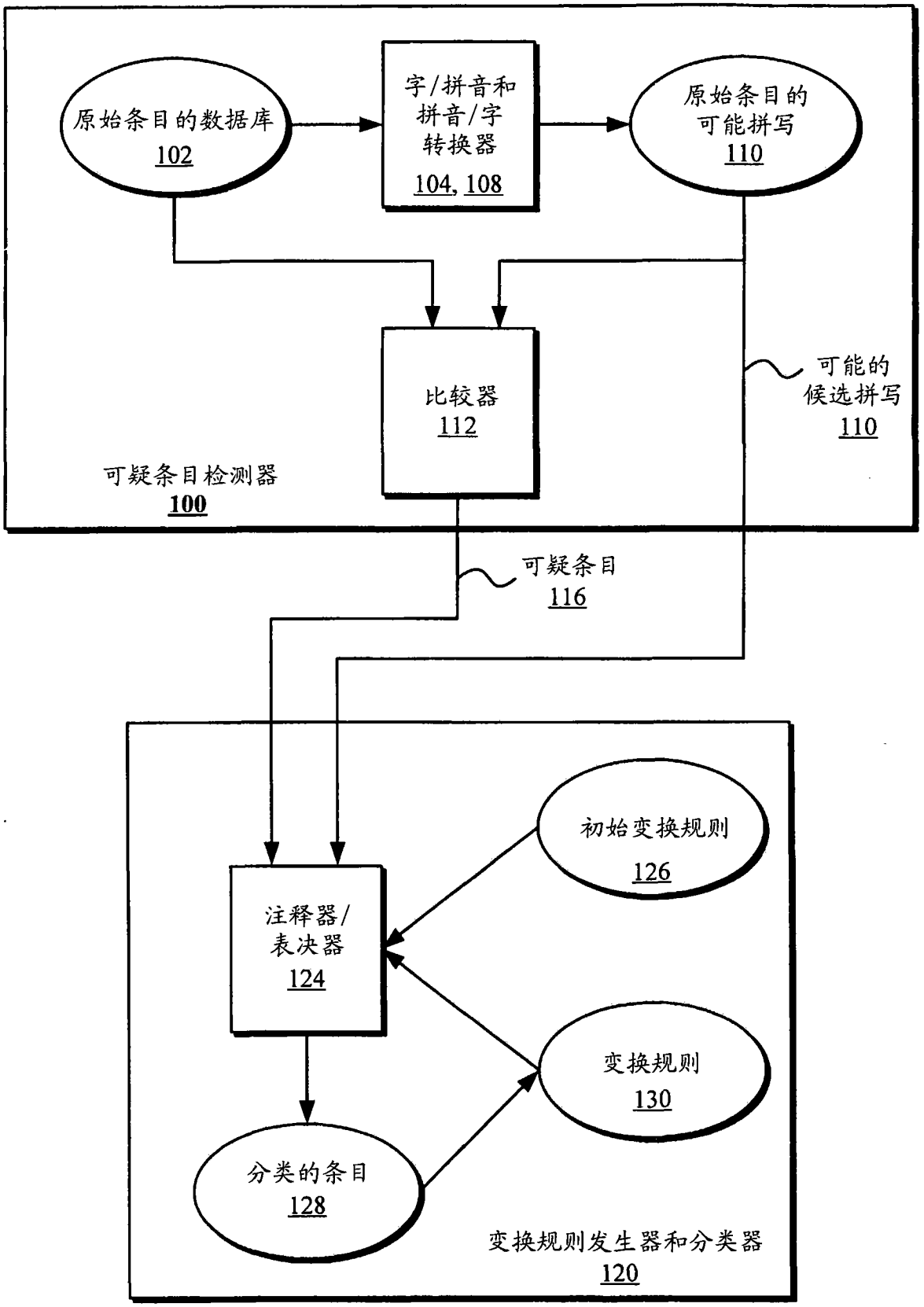

原始条目的数据库
102
字/拼音和
拼音/字
转换器
104, 108
原始条目的
可能拼写
110
比较器
112
可能的
候选拼写
110
可疑条目检测器
100
可疑条目
116
初始变换规则
126
注释器/
表决器
124
变换规则
130
分类的条目
128
变换规则发生器和分类器
120

图 2

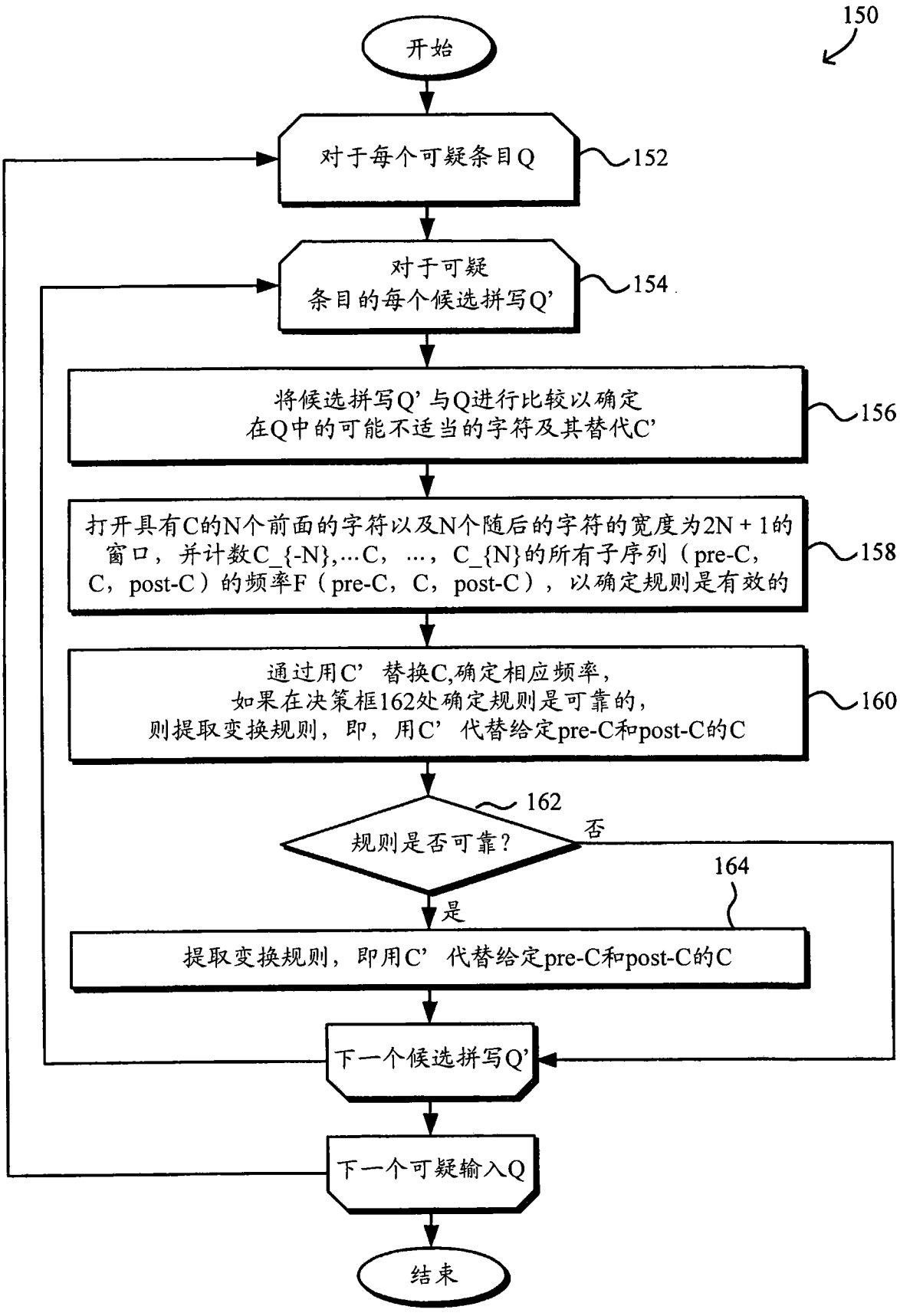

150
开始
对于每个可疑条目Q
152
对于可疑
条目的每个候选拼写Q'
154
将候选拼写Q'与Q进行比较以确定
在Q中的可能不适当的字符及其替代C'
156
打开具有C的N个前面的字符以及N个随后的字符的宽度为2N + 1的窗口，并计数C_{-N},…C，…，C_{N}的所有子序列（pre-C，C，post-C）的频率F（pre-C，C，post-C），以确定规则是有效的
158
通过用C' 替换C,确定相应频率，
如果在决策框162处确定规则是可靠的，
则提取变换规则，即，用C' 代替给定pre-C和post-C的C
160
162
规则是否可靠？
否
是
164
提取变换规则，即用C' 代替给定pre-C和post-C的C
下一个候选拼写Q'
下一个可疑输入Q
结束

图 3

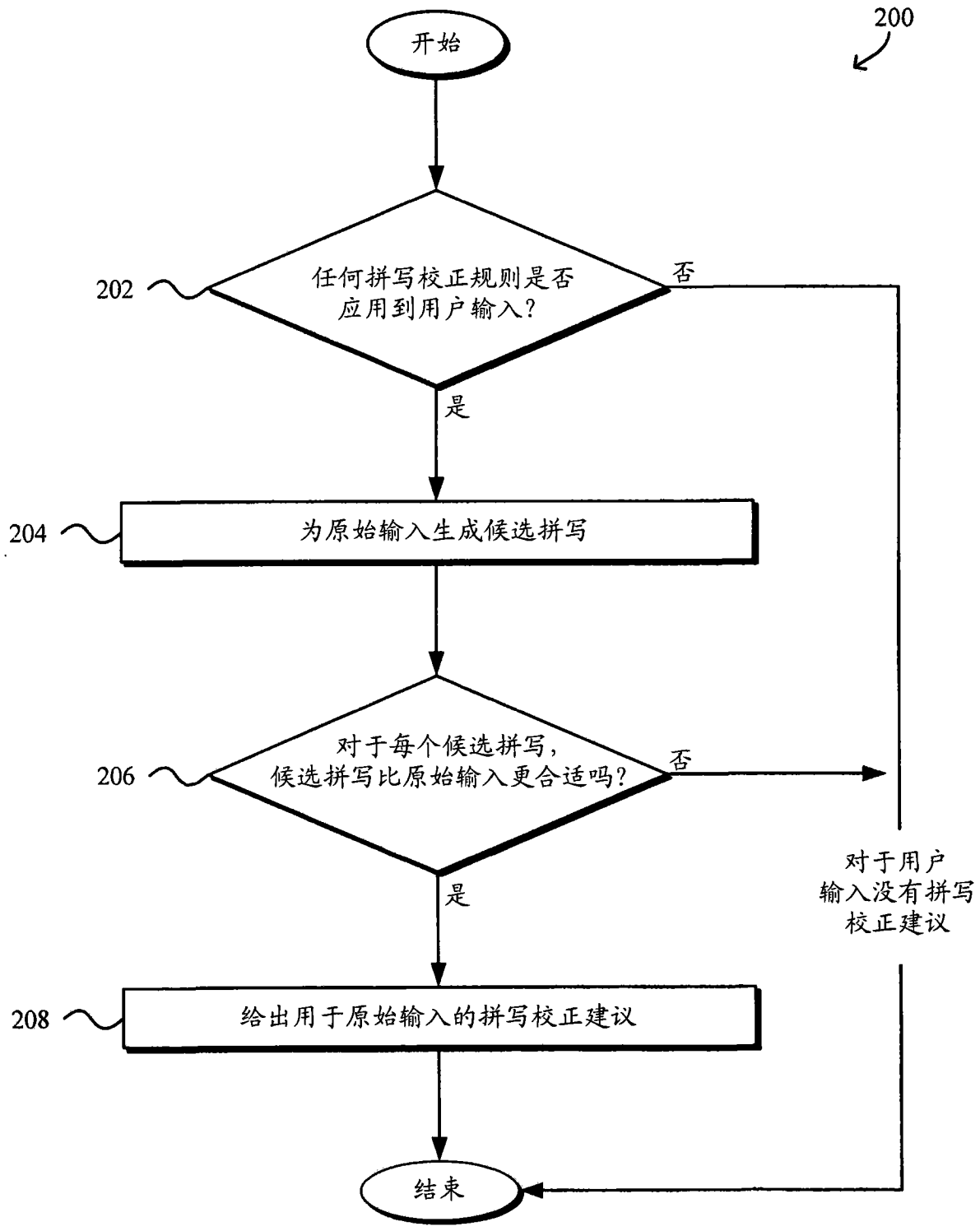

200
开始
202
任何拼写校正规则是否
应用到用户输入?
否
是
204
为原始输入生成候选拼写
206
对于每个候选拼写,
候选拼写比原始输入更合适吗?
否
是
对于用户
输入没有拼写
校正建议
208
给出用于原始输入的拼写校正建议
结束

图 4