(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111723791 A
(43)申请公布日 2020.09.29

(21)申请号 202010529548.0

(22)申请日 2020.06.11

(71)申请人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 杨震 李彦 亓超 马宇驰

(74)专利代理机构 北京市立方律师事务所 11330
代理人 张筱宁

(51)Int.Cl.
G06K 9/20(2006.01)
G06K 9/62(2006.01)
G06F 40/232(2020.01)

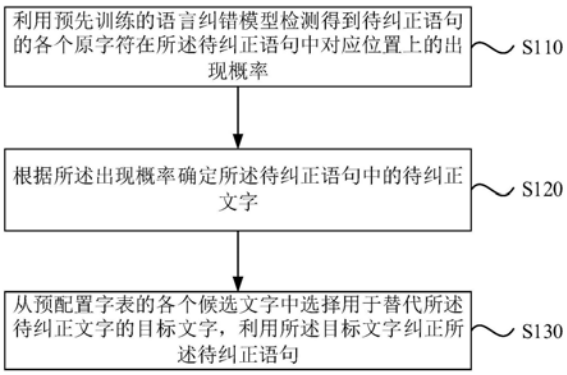
权利要求书2页 说明书12页 附图3页

(54)发明名称

文字纠错方法、装置、设备及存储介质

(57)摘要

本申请实施例提供了一种文字纠错方法、装置、设备及存储介质，涉及计算机领域，该方法通过利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在该待纠正语句中对应位置上的出现概率，并根据所述出现概率确定所述待纠正语句中的待纠正文字；从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，纠正该待纠正语句。本技术方案实现了高效而准确地识别出待纠正语句中的错别字并进行纠正。



CN 111723791 A

1.一种文字纠错方法,其特征在于,包括以下步骤:

利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率;

根据所述出现概率确定所述待纠正语句中的待纠正文字;

从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字,利用所述目标文字纠正所述待纠正语句。

2.根据权利要求1所述的文字纠错方法,其特征在于,所述利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率的步骤包括:

获取待纠正语句,将所述待纠正语句输入到预先训练的语言纠错模型,以通过所述语言纠错模型分析所述待纠正语句的各个原字符之间的语义关联关系;

基于所述语义关联关系得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率。

3.根据权利要求2所述的文字纠错方法,其特征在于,所述获取待纠正语句的步骤包括:

对视频帧图像进行光学符号识别得到视频字幕语句,从所述视频字幕语句中筛选出待纠正语句。

4.根据权利要求1所述的文字纠错方法,其特征在于,所述根据所述出现概率确定所述待纠正语句中的待纠正文字的步骤包括:

获取所述待纠正语句的各个原字符在所述待纠正语句中对应位置的出现概率;

将所述出现概率与预设阈值进行比较,若所述出现概率小于预设阈值,则将该位置确定为待纠正位置,将所述纠正位置上的原字符确定为待纠正文字。

5.根据权利要求1所述的文字纠错方法,其特征在于,所述从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字的步骤包括:

获取预配置字表的各个候选文字,利用所述预先训练的语言纠错模型分析所述候选文字在所述待纠正位置处对应的概率向量;

根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目标文字。

6.根据权利要求5所述的文字纠错方法,其特征在于,所述根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目标文字的步骤包括:

从所述候选文字中提取所述待纠正文字的形近字及其在所述概率向量中对应的概率值;

对所述形近字对应的概率值进行比较,根据比较结果选择概率值最大的形近字作为目标文字,以将所述目标文字替代所述待纠正文字。

7.根据权利要求1所述的文字纠错方法,其特征在于,所述从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字的步骤之前,还包括:

利用汉字的形近码进行编码构建出候选文字,集合所述候选文字生成预配置字表;其中,所述形近码包括汉字结构、笔画和四角码中的至少一者。

8.一种文字纠错装置,其特征在于,包括:

检测模块,用于利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所

述待纠正语句中对应位置上的出现概率；

确定模块，用于根据所述出现概率确定所述待纠正语句中的待纠正文字；

纠正模块，用于从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，利用所述目标文字纠正所述待纠正语句。

9.一种文字纠错设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求1-7任一项所述的文字纠错方法的步骤。

10.一种包含计算机可执行指令的存储介质，其特征在于，所述计算机可执行指令在由计算机处理器执行时用于执行如权利要求1-7任一项所述文字纠错方法的步骤。

文字纠错方法、装置、设备及存储介质

技术领域

[0001] 本申请涉及计算机软件领域，具体而言，本申请涉及一种文字纠错方法、装置、设备及存储介质。

背景技术

[0002] 随着数字媒体技术的高速发展，人们需要将一些其他媒介，如纸质或多媒体(如视频)等媒介的文字输入到计算机，以通过计算机对文字进行分析处理。

[0003] 在相关技术中，采用光学字符识别(Optical Character Recognition，OCR)技术能够快速识别出纸上或视频中的文字，得到可在计算机上编辑的计算机文本。然而，在对文字进行OCR时，往往会因为文字背景或文字字体等问题导致OCR识别出来的汉字出现错误，将其识别为它的形近字，以使得OCR识别出来的文字输出结果准确率低。

发明内容

[0004] 本申请的目的旨在至少解决上述技术缺陷之一，特别是文字识别的输出结果准确率低的问题。

[0005] 第一方面，本申请实施例提供了一种文字纠错方法，包括以下步骤：

[0006] 利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率；

[0007] 根据所述出现概率确定所述待纠正语句中的待纠正文字；

[0008] 从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，利用所述目标文字纠正所述待纠正语句。

[0009] 在一实施例中，所述利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率的步骤包括：

[0010] 获取待纠正语句，将所述待纠正语句输入到预先训练的语言纠错模型，以通过所述语言纠错模型分析所述待纠正语句的各个原字符之间的语义关联关系；

[0011] 基于所述语义关联关系得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率。

[0012] 在一实施例中，所述获取待纠正语句的步骤包括：

[0013] 对视频帧图像进行光学符号识别得到视频字幕语句，从所述视频字幕语句中筛选出待纠正语句。

[0014] 在一实施例中，所述根据所述出现概率确定所述待纠正语句中的待纠正文字的步骤包括：

[0015] 获取所述待纠正语句的各个原字符在所述待纠正语句中对应位置的出现概率；

[0016] 将所述出现概率与预设阈值进行比较，若所述出现概率小于预设阈值，则将该位置确定为待纠正位置，将所述纠正位置上的原字符确定为待纠正文字。

[0017] 在一实施例中，所述从预配置字表的各个候选文字中选择用于替代所述待纠正文

字的目标文字的步骤包括：

[0018] 获取预配置字表的各个候选文字，利用所述预先训练的语言纠错模型分析所述候选文字在所述待纠正位置处对应的概率向量；

[0019] 根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目标文字。

[0020] 在一实施例中，所述根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目标文字的步骤包括：

[0021] 从所述候选文字中提取所述待纠正文字的形近字及其在所述概率向量中对应的概率值；

[0022] 对所述形近字对应的概率值进行比较，根据比较结果选择概率值最大的形近字作为目标文字，以将所述目标文字替代所述待纠正文字。

[0023] 在一实施例中，所述从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字的步骤之前，还包括：

[0024] 利用汉字的形近码进行编码构建出候选文字，集合所述候选文字生成预配置字表；其中，所述形近码包括汉字结构、笔画和四角码中的至少一者。

[0025] 第二方面，本申请实施例还提供了一种文字纠错装置，包括：

[0026] 检测模块，用于利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率；

[0027] 确定模块，用于根据所述出现概率确定所述待纠正语句中的待纠正文字；

[0028] 纠正模块，用于从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，利用所述目标文字纠正所述待纠正语句。

[0029] 第三方面，本申请实施例还提供了一种文字纠错设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现如第一方面任一实施例提及的文字纠错方法的步骤。

[0030] 第四方面，本申请实施例还提供了一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行如第一方面任一实施例提及的文字纠错方法的步骤。

[0031] 上述实施例提供的文字纠错方法、装置、设备及存储介质，通过利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在该待纠正语句中对应位置上的出现概率，并根据所述出现概率确定所述待纠正语句中的待纠正文字；从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，纠正该待纠正语句，从而实现高效而准确地识别出待纠正语句中的错别字，并对待纠正语句中的错别字进行纠正。

[0032] 本申请附加的方面和优点将在下面的描述中部分给出，这些将从下面的描述中变得明显，或通过本申请的实践了解到。

附图说明

[0033] 本申请上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

[0034] 图1是一实施例提供的文字纠错方法流程图；

[0035] 图2是一实施例提供的待纠正语句的各原字符的出现概率的预测原理示意图；

[0036] 图3是一实施例提供的待纠正语句的各原字符的出现概率的另一预测原理示意图；

[0037] 图4是一实施例提供的待纠正语句的各原字符的出现概率的再一预测原理示意图；

[0038] 图5是一实施例提供的文字纠错装置的结构示意图；

[0039] 图6是一实施例提供的文字纠错设备的结构示意图。

具体实施方式

[0040] 下面详细描述本申请的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本申请，而不能解释为对本申请的限制。

[0041] 本技术领域技术人员可以理解，除非特意声明，这里使用的单数形式“一”、“一个”、“所述”和“该”也可包括复数形式。应该进一步理解的是，本申请的说明书中使用的措辞“包括”是指存在所述特征、整数、步骤、操作、元件和/或组件，但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、组件和/或它们的组。这里使用的措辞“和/或”包括一个或更多个相关联的列出项的全部或任一单元和全部组合。

[0042] 本技术领域技术人员可以理解，除非另外定义，这里使用的所有术语(包括技术术语和科学术语)，具有与本申请所属领域中的普通技术人员的一般理解相同的意义。还应该理解的是，诸如通用字典中定义的那些术语，应该被理解为具有与现有技术的上下文中的意义一致的意义，并且除非像这里一样被特定定义，否则不会用理想化或过于正式的含义来解释。

[0043] 图1是一实施例提供的文字纠错方法流程图，该文字纠错方法执行于计算机设备，如服务器、个人电脑、笔记本电脑、平板电脑、扫描机和智能手机等。

[0044] 具体的，如图1所示，该文字纠错方法可以包括以下步骤：

[0045] S110、利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率。

[0046] 其中，语言纠错模型可以是预先通过大量的词语样本训练得到。可选的，将词语训练样本集输入到预设的神经网络语言模型中进行训练，得到语音纠错模型。输入的词语训练样本可以是一个词，还是可以多个词组成的语句。

[0047] 对于语言模型，简单来说就是一串词序列的概率分布。具体来说，语言模型的作用是为一个长度为m的文本确定一个概率分布P，表示这段文本存在的可能性。一个预训练的词表示应该能够包含丰富的句法和语义信息，并且能够对多义词进行建模。利用语言模型来获得一个上下文相关的预训练表示。

[0048] 可选的，在本实施例中，语言纠错模型可以是BERT语言模型，该预先训练的BERT语言模型对输入的语句做拼写检查，通过softmax函数输出当前字符位置的下一字符位置上原字符的概率向量，从而得到该原字符在该输入语句中对应位置的出现概率。

[0049] 由于每个词语或语句的各个字符之间具有句法和语义关联关系，在本实施例中，利用语言纠错模型可以检测出待纠错语句中每个字符在该待纠正语句中对应位置上的出现概率。其中，各个原字符在待纠正语句中对应位置可以按照某一个或多个字符的位置为

标准确定的位置，如“打篮球”这一语句中，“打”所对应位置可以理解为与“球”字所在位置相隔一个字符的位置，也可以是结合待纠正语句的语义所确定的其他位置。

[0050] 例如，训练得到的语言纠错模型基于大量的词语训练样本的训练，学习各个词语之间的语法、句法和语音关联关系，基于当前位置字符和下一位置字符之间的语义关联关系，计算当前位置字符的下一位置字符的各个可能字的出现概率，得到各个可能字的概率分布。

[0051] 又如，当前位置字符为“名”，语言纠错模型会计算得到下一字符“句”或其他输入的其他字符，如“包”的出现概率，并基于“名”与“句”，或者“名”与“包”之间的语义关联关系等，得到“句”的出现概率比“包”的出现概率高。

[0052] 同理，将待纠正语句“这是一个千古名包”输入到语言纠错模型中，语音纠错模型会根据待纠错语音之间的句法、语法和语音关联关系等计算出待纠正语句的各个原字符在待纠错语句上对应位置上的出现概率。例如，待纠正语句包括“这”、“是”、“一”、“个”、“千”、“古”、“名”和“包”8个原字符，进一步的，“这”在待纠正语句的对应位置为第一个字符位置、“是”在待纠正语句中的对应位置为第二个字符位置(或“这”的下一字符位置，或“一”的上一字符位置)，以此类推，“包”在待纠正语句中的对应位置为第八个字符位置(或“名”的下一字符位置)。

[0053] S120、根据所述出现概率确定所述待纠正语句中的待纠正文字。

[0054] 在本实施例中，原字符的出现概率与正确率相关，语言纠错模型输出的待纠正语句的某一原字符在该待纠正语句上对应位置出现的概率越高，说明该原字符的正确率越高，存在输入或识别错误的可能性越小。

[0055] 可选的，获取所述待纠正语句的各个原字符在所述待纠正语句中对应位置的出现概率；将所述出现概率与预设阈值进行比较，若所述出现概率小于预设阈值，则将该位置确定为待纠正位置，将所述纠正位置上的原字符确定为待纠正文字，其中，待纠正位置可以有一个或多个，对应的待纠正文字也可以有一个或多个。

[0056] 需要说明的是，由于语言纠错模型默认偏向于待纠正语句的各个原字符为正确字符，因此，各原字符在其对应位置上的出现概率会比其他文字高，即便该原字符是错误字符，也即是，原字符在待纠正语句的对应位置上的出现概率为该位置上的最高出现概率。在这种情况下，若当前位置上的各个字符的最高出现概率小于预设阈值，则该当前位置为待纠正位置；提取该待纠正位置上的原字符确定为待纠正文字。

[0057] 假设预设阈值为0.9，基于上述例子，待纠正语句的“这”、“是”、“一”、“个”、“千”、“古”、“名”和“包”8个原字符在该待纠正语句中对应位置出现的概率向量分别为0.99、0.99、0.98、0.99、0.99和0.7。由于该待纠正语句中的原字符“包”的出现概率为0.6，低于预设阈值0.9，则将待纠正语句的第8个字符对应位置，即“名”字符紧跟随的下一字符位置确定为待纠正位置，将该待纠正位置上的原字符，即“包”确定为待纠正文字。

[0058] S130、从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，利用所述目标文字纠正所述待纠正语句。

[0059] 在本实施例中，预配置字表中包括多个候选文字，该候选文字包括与待纠正语句的原字符相同的文字，也包括与待纠正语句的原字符的形近字。

[0060] 可选的，根据预配置字表中的各个候选字在该待纠正位置处的出现概率，根据各

个候选字的出现概率选择目标文字，该目标文字可以是除待纠正文字外的候选字中的出现概率最高的文字。可选的，还可以利用其它选择方式，如用户自定义从预配置字表中选择与待纠正文字的读音或字形相同或相似的其他文字作为目标字。利用该目标文字替代待纠正文字，从而对待纠正语句进行纠正。

[0061] 本实施例提供的文字纠错方法，通过利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在该待纠正语句中对应位置上的出现概率，并根据所述出现概率确定所述待纠正语句中的待纠正文字；从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，纠正该待纠正语句，从而实现高效而准确地识别出待纠正语句中的错别字，并对待纠正语句中的错别字进行纠正。

[0062] 为了使本申请的技术方案更为清晰，更为便于理解，下面对本技术方案中的多个步骤的具体的实现过程和方式加以详细的描述。

[0063] 在一实施例中，步骤S110利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率，可以包括以下步骤：

[0064] S1101、获取待纠正语句，将所述待纠正语句输入到预先训练的语言纠错模型，以通过所述语言纠错模型分析所述待纠正语句的各个原字符之间的语义关联关系。

[0065] 可选的，从本执行设备中获取到待纠正语句，也可以是从本地外接的设备中获取到待纠正语句，还可以从云端获取到待纠正语句。该待纠正语句可以是人工输入的语句，还可以是通过设备扫描识别等方式得到的语句。

[0066] 在一实施例中，对视频帧图像进行光学符号识别得到视频字幕语句，从所述视频字幕语句中筛选出待纠正语句。

[0067] 在本实施例中，截取多帧视频，对该视频帧图像进行光学符号识别。可选的，在进行光学符号识别之前可以从截取到的视频帧图像中筛选出符合要求的视频帧图像，如将不带字幕、字幕不全的或字幕清晰度不符合要求的视频帧图像删除，得到符合要求原始视频帧图像，进一步的，可以对原始视频帧图像进行二值化处理等处理后，进行光学字符识别，得到该原始视频帧图像对应的视频字幕语句。

[0068] 进一步的，从视频字幕语句中筛选出需要进行识别错误检测的待纠正语句，可选的，可以通过标注的方式筛选出待纠正语句，也可以是随机抽取的方式筛选出待纠正语句，还可以是将全部视频字幕语句做出待纠正语句进行检测。

[0069] 在本实施例中，将待纠正语句进行处理，将待纠正语句转换为对应的语句序列，为语句序列中的每个字符构造一个字符特征向量，并输入到预先训练的语言纠错模型中，以通过语言纠错模型结合待纠正语句的上下文词语之间的语义，分析待纠正语句中的各个原字符对应字符特征向量之间的语义关联关系。通常而言，若某一字符在某一词语或某一句子中出现的频次越多，则该字符在该词语或句子中的语义关联关系越大。

[0070] S1102、基于所述语义关联关系得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率。

[0071] 语义关联关系是根据对词语或句子的语义分析或该某一字符在某一词语或某一句子中出现的频次来确定，通常而言，常用的词语或词语组合得到的语句，该词语或语句中各个字符之间的语义关联关系越强。

[0072] 在本实施例中，若当前位置字符与其上一位置字符和/或下一位置字符之间的语

义关联关系越强，那么该字符在当前位置上出现的概率就越大；若当前位置字符与其上一位置字符和/或下一位置字符之间的语义关联关系越弱，则该字符在当前位置出现的概率越小。

[0073] 例如，待纠正语句为“打蓝球”，语言纠错模型基于该待纠正语句语义关联关系，得到“打”“蓝”“球”这三个原字符在待纠正语句中对应位置的概率。结合对“打蓝球”的语义关联关系，该词中的“蓝”字跟随在“打”字之前，或跟随在“球”字之后可能性比较低，得到“打”、“蓝”和“球”这三个字中“打”和“球”字在该待纠正语句中当前位置的出现概率较高，可以预测出“蓝”字可能为错别字。

[0074] 其中，各个原字符在待纠正语句中对应位置可以按照某一个或多个字符的位置为标准确定的位置，如“打”所对应位置可以理解为与“球”字相隔一个字符的位置，也可以是结合待纠正语句的语义所确定的其他位置。

[0075] 为了更清楚的阐述本申请的技术方案，下面以待纠正语句为“天气真晴朗”为例进行示例性说明。

[0076] 在一实施例中，可以根据上文的语义推测下文的语义，从而预测出当前字符的出现概率，如图2所示，图2是一实施例提供的待纠正语句的各原字符的出现概率的预测原理示意图，可以根据上文输入的“天”字符，预测当前字符“气”的出现概率，同理，根据上文输入的“天气”字符，预测当前字符“真”的出现概率等。

[0077] 在一实施例中，可以根据下文的语义推测上文的语义，从而预测出当前字符的出现概率，如图3所示，图3是一实施例提供的待纠正语句的各原字符的出现概率的另一预测原理示意图，可以根据下文输入的“朗”字符，预测当前字符“晴”的出现概率，同理，根据下文输入的“晴朗”字符，预测当前字符“真”的出现概率等。

[0078] 在一实施例中，可以根据上下文语义推测当前字符的出现概率，如图4所示，图4是一实施例提供的待纠正语句的各原字符的出现概率的再一预测原理示意图，可以根据上文输入的“天”字符和下文输入的“真晴朗”字符，预测当前字符“气”的出现概率等。

[0079] 本实施例提供的待纠正语句中各个原字符在该待纠正语句中对应位置的出现概率的预测，能够高效而准确地识别出待纠正语句中的错别字。

[0080] 在一实施例中，步骤S130中的从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，可以包括以下步骤：

[0081] S1301、获取预配置字表的各个候选文字，利用所述预先训练的语言纠错模型分析所述候选文字在所述待纠正位置处对应的概率向量。

[0082] 在本实施例中，预配置字表可以位于语言纠错模型中，也可以是外部输入的字表。

[0083] 当将待纠正语句输入到预先训练的语音纠错模型对待纠正语句的各个原字符进行检测时，语言纠错模型同时获取预配置字表的各个候选文字，以通过语言纠错模型检测出待纠正语句中各个原字符在其对应位置上的出现概率，同时检测出各个候选文字在各原字符对应位置上的出现概率。

[0084] 进一步的，将各个候选字在待纠正语句的原字符上对应位置上的出现概率利用softmax函数进行归一化处理得到候选字在该对应位置上的概率向量。其中，概率向量中的各个概率值与其对应的候选文字一一映射，各个候选文字对应的概率值之和等于一。

[0085] S1302、根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目

标文字。

[0086] 由于语言纠错模型默认偏向于待纠正语句的各个原字符为正确字符，因此，各原字符在其对应位置上的出现概率会比其他文字高，即便该原字符是错误字符，也即是，原字符在待纠正语句的对应位置上的出现概率为该位置上的最大出现概率。当该对应位置上各个文字的最大出现概率小于预设阈值，则说明，该原字符为错别字，将该原字符确定为待纠正文字，该对应位置确定为待纠正位置。

[0087] 在本实施例中，由于概率向量中的各个候选文字的概率值相对概率值，若概率向量中某一概率值的越高，则说明该概率值对应的候选文字的正确性越高。从候选文字中除待纠正文字(也即原字符)之外的其余候选字中选择概率值最大的候选文字作为目标文字，将目标文字代替待纠正文字以对待纠正语句进行纠正。

[0088] 进一步的，在一实施例中，步骤S1302根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目标文字，可以包括以下步骤：

[0089] S3021、从所述候选文字中提取所述待纠正文字的形近字及其在所述概率向量中对应的概率值。

[0090] 其中，形近字是指与待纠正文字汉字结构相近的字，可以是一个或多个。预配置字表的候选文字中包括待纠正文字的形近字，当将预配置字表的各个候选文字输入到语言纠错模型，得到各个候选字在待纠正语句的各个原字符所在的当前位置上的出现概率，将出现概率通过softmax函数做归一化处理，得到该待纠正文字的各个候选文字在待纠正位置上的概率向量。进行归一化处理后，各个候选文字对应的概率向量之和等于1。

[0091] 在该概率向量中，各个候选文字与其在该待纠正语句的原字符的对应位置上的出现概率值一一映射，进一步的，在本实施例中，从在待纠正位置上进行检测的预配置字表的各个候选文字提取出待纠正文字的形近字，并从概率向量中提取各个形近字对应的概率值。

[0092] S3022、对所述形近字对应的概率值进行比较，根据比较结果选择概率值最大的形近字作为目标文字，以将所述目标文字替代所述待纠正文字。

[0093] 将各个待纠正文字的形近字的概率值进行比较，在本实施例中，某一形近字的概率值越大，该形近字作为目标文字的可能性越大，目标文字为正确文字，用于替代待纠正文字对待纠正语句进行纠正。

[0094] 例如，待纠正文字为“包”，从候选文字中提取出的“包”字的形近字为“句”、“匀”和“勺”，其在概率向量中对应的概率值分别为:0.2、0.1和0.05，由于“句”的概率向量中对应的概率值最大，所以将该“句”子作为目标文字，以将目标文字替代待纠正文字。

[0095] 在本实施例中，对各个候选文字(或形近字)的出现概率进行归一化处理，能够更准确地得出各个候选文字(或形近字)在待纠正语句各原字符所在的当前位置上的相对出现概率的大小，使得识别出目标文字的结果更准确。

[0096] 在一实施例中，步骤S130的从所述预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字之前，还可以包括以下步骤：

[0097] S100、利用汉字的形近码进行编码构建出候选文字，集合所述候选文字生成预配置字表。

[0098] 其中，所述形近码包括汉字结构、笔画和四角码中的至少一者。

[0099] 汉字结构是指构成汉字字形的各种特定的点和线，也是汉字的最小结构单位。笔画通常是指组成汉字且不间断的各种形状的点和线，如横(一)、竖(丨)、撇(丿)、点(丶)、折(乛)等，它是构成汉字字形的最小连笔单位。四角码一种计算机四角汉字输入法，它包括：包括对汉字的编码和取码，以及与键盘的对应关系，使用汉字的一些特定部首笔画作为代码，而这些代码分别与0、1、2、3……9十个数字相对应，然后，利用这些代码来拆解汉字的四个角，并以相应的数字作为编码来表示和区分汉字。

[0100] 在本实施例中，通过汉字的形近码，如汉字结构、笔画和四角码等进行编码组合，由于汉字的形近码是构成汉字的最小单位，通过对形近码的不同编码和组合方式，构建出不同的候选文字，该候选文字通常为常用文字，可以覆盖待纠正语句中可能出现的各个原字符及其对应的形近字。将候选文字集合起来生成预配置字表，并保存起来。

[0101] 下面对文字纠错装置的相关实施例进行详细阐述。

[0102] 图5是一实施例提供的文字纠错装置的结构示意图，如图5示，该文字纠错装置10适用于对输入的文字进行自动纠正，包括：检测模块110、确定模块120和纠正模块130。

[0103] 其中，检测模块110，用于利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率；

[0104] 确定模块120，用于根据所述出现概率确定所述待纠正语句中的待纠正文字；

[0105] 纠正模块130，用于从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，利用所述目标文字纠正所述待纠正语句。

[0106] 上述实施例提供的文字纠错装置，通过检测模块110利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在该待纠正语句中对应位置上的出现概率，确定模块120根据所述出现概率确定所述待纠正语句中的待纠正文字；纠正模块130从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，纠正该待纠正语句，从而实现高效而准确地识别出待纠正语句中的错别字，并对待纠正语句中的错别字进行纠正。

[0107] 在一实施例中，检测模块110包括：语义分析单元和概率得到单元；

[0108] 其中，语义分析单元，用于获取待纠正语句，将所述待纠正语句输入到预先训练的语言纠错模型，以通过所述语言纠错模型分析所述待纠正语句的各个原字符之间的语义关联关系；概率得到单元，用于基于所述语义关联关系得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率。

[0109] 在一实施例中，语义分析单元包括：语句筛选子单元，用于对视频帧图像进行光学符号识别得到视频字幕语句，从所述视频字幕语句中筛选出待纠正语句。

[0110] 在一实施例中，确定模块120包括：出现概率获取单元和待纠正文字确定单元；

[0111] 其中，出现概率获取单元，用于获取所述待纠正语句的各个原字符在所述待纠正语句中对应位置的出现概率；待纠正文字确定单元，用于将所述出现概率与预设阈值进行比较，若所述出现概率小于预设阈值，则将该位置确定为待纠正位置，将所述纠正位置上的原字符确定为待纠正文字。

[0112] 在一实施例中，纠正模块130包括：概率向量计算单元和目标文字确定单元；

[0113] 其中，概率向量计算单元，用于获取预配置字表的各个候选文字，利用所述预先训练的语言纠错模型分析所述候选文字在所述待纠正位置处对应的概率向量；目标文字确定单元，用于根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目标文

字。

[0114] 在一实施例中，目标文字确定单元包括：形近字提取子单元和目标文字选择子单元；

[0115] 其中，形近字提取子单元，用于从所述候选文字中提取所述待纠正文字的形近字及其在所述概率向量中对应的概率值；目标文字选择子单元，用于对所述形近字对应的概率值进行比较，根据比较结果选择概率值最大的形近字作为目标文字，以将所述目标文字替代所述待纠正文字。

[0116] 在一实施例中，文字纠错装置10还包括：字表生成模块，用于利用汉字的形近码进行编码构建出候选文字，集合所述候选文字生成预配置字表；其中，所述形近码包括汉字结构、笔画和四角码中的至少一者。

[0117] 上述提供的文字纠错装置可用于执行上述任意实施例提供的文字纠错方法，具备相应的功能和有益效果。

[0118] 图6是一实施例提供的文字纠错设备的结构示意图，如图6所示，该文字纠错设备包括处理器60、存储器61、输入装置62以及输出装置63。存储器61上存储有可在处理器60上运行的计算机程序，处理器60执行所述程序时实现如上述任一实施例中的文字纠错方法。

[0119] 该文字纠错设备中处理器60的数量可以是一个或多个，图6以一个处理器60为例。该文字纠错设备中存储器61的数量可以是一个或者多个，图6中以一个存储器61为例。该文字纠错设备的处理器60和存储器61可以通过总线或者其他方式连接，图6中以通过总线连接为例。实施例中，文字纠错设备可以是电脑、扫描机和服务器等。

[0120] 存储器61作为一种计算机可读存储介质，可用于存储软件程序、计算机可执行程序以及模块，如本方案任意实施例所述的文字纠错方法对应的程序指令/模块(例如，检测模块110、确定模块120和纠正模块130)。存储器61可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序；存储数据区可存储根据设备的使用所创建的数据等。此外，存储器61可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实例中，存储器61可进一步包括相对于处理器60远程设置的存储器，这些远程存储器可以通过网络连接至设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0121] 输入装置62可用于接收输入的数字或者字符信息，以及产生与计算机设备的用户设置以及功能控制有关的键信号输入，还可以是用于获取图像的摄像头以及获取音频数据的拾音设备。输出装置63可以包括扬声器等音频设备或者打印机等文字输入设备。需要说明的是，输入装置62和输出装置63的具体组成可以根据实际情况设定。

[0122] 处理器60通过运行存储在存储器61中的软件程序、指令以及模块，从而执行设备的各种功能应用以及数据处理，即实现上述的文字纠错方法。

[0123] 上述提供的计算机设备执行上述任意实施例提供的文字纠错方法时，具备相应的功能和有益效果。

[0124] 本实施例还提供一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行一种文字纠错方法，包括：

[0125] 利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正

语句中对应位置上的出现概率；

[0126] 根据所述出现概率确定所述待纠正语句中的待纠正文字；

[0127] 从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，利用所述目标文字纠正所述待纠正语句。

[0128] 当然，本实施例所提供的一种包含计算机可执行指令的存储介质，其计算机可执行指令不限于如上所述的文字纠错方法操作，还可以执行任意实施例所提供的文字纠错方法中的相关操作，且具备相应的功能和有益效果。

[0129] 通过以上关于实施方式的描述，所属领域的技术人员可以清楚地了解到，本方案可借助软件及必需的通用硬件来实现，当然也可以通过硬件实现，但很多情况下前者是更佳的实施方式。基于这样的理解，本技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如计算机的软盘、只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、闪存(FLASH)、硬盘或光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务端，或者网络设备等)执行本方案任意实施例所述的文字纠错方法。

[0130] 本申请公开了A1、一种文字纠错方法，包括以下步骤：

[0131] 利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率；

[0132] 根据所述出现概率确定所述待纠正语句中的待纠正文字；

[0133] 从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，利用所述目标文字纠正所述待纠正语句。

[0134] A2.根据A1所述的文字纠错方法，所述利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率的步骤包括：

[0135] 获取待纠正语句，将所述待纠正语句输入到预先训练的语言纠错模型，以通过所述语言纠错模型分析所述待纠正语句的各个原字符之间的语义关联关系；

[0136] 基于所述语义关联关系得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率。

[0137] A3.根据A2所述的文字纠错方法，所述获取待纠正语句的步骤包括：

[0138] 对视频帧图像进行光学符号识别得到视频字幕语句，从所述视频字幕语句中筛选出待纠正语句。

[0139] A4.根据A1所述的文字纠错方法，所述根据所述出现概率确定所述待纠正语句中的待纠正文字的步骤包括：

[0140] 获取所述待纠正语句的各个原字符在所述待纠正语句中对应位置的出现概率；

[0141] 将所述出现概率与预设阈值进行比较，若所述出现概率小于预设阈值，则将该位置确定为待纠正位置，将所述纠正位置上的原字符确定为待纠正文字。

[0142] A5.根据A1所述的文字纠错方法，所述从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字的步骤包括：

[0143] 获取预配置字表的各个候选文字，利用所述预先训练的语言纠错模型分析所述候选文字在所述待纠正位置处对应的概率向量；

[0144] 根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目标文字。

[0145] A6.根据A5所述的文字纠错方法，所述根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目标文字的步骤包括：

[0146] 从所述候选文字中提取所述待纠正文字的形近字及其在所述概率向量中对应的概率值；

[0147] 对所述形近字对应的概率值进行比较，根据比较结果选择概率值最大的形近字作为目标文字，以将所述目标文字替代所述待纠正文字。

[0148] A7.根据A1所述的文字纠错方法，所述从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字的步骤之前，还包括：

[0149] 利用汉字的形近码进行编码构建出候选文字，集合所述候选文字生成预配置字表；其中，所述形近码包括汉字结构、笔画和四角码中的至少一者。

[0150] B8.一种文字纠错装置，包括：

[0151] 检测模块，用于利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率；

[0152] 确定模块，用于根据所述出现概率确定所述待纠正语句中的待纠正文字；

[0153] 纠正模块，用于从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，利用所述目标文字纠正所述待纠正语句。

[0154] B9.根据B8所述的文字纠错装置，所述检测模块包括：语义分析单元和概率得到单元；

[0155] 语义分析单元，用于获取待纠正语句，将所述待纠正语句输入到预先训练的语言纠错模型，以通过所述语言纠错模型分析所述待纠正语句的各个原字符之间的语义关联关系；

[0156] 概率得到单元，用于基于所述语义关联关系得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率。

[0157] B10.根据B9所述的文字纠错装置，所述语义分析单元包括：语句筛选子单元，用于对视频帧图像进行光学符号识别得到视频字幕语句，从所述视频字幕语句中筛选出待纠正语句。

[0158] B11.根据B8所述的文字纠错装置，所述确定模块包括：出现概率获取单元和待纠正文字确定单元；

[0159] 出现概率获取单元，用于获取所述待纠正语句的各个原字符在所述待纠正语句中对应位置的出现概率；

[0160] 待纠正文字确定单元，用于将所述出现概率与预设阈值进行比较，若所述出现概率小于预设阈值，则将该位置确定为待纠正位置，将所述纠正位置上的原字符确定为待纠正文字。

[0161] B12.根据B8所述的文字纠错装置，所述纠正模块包括：概率向量计算单元和目标文字确定单元；

[0162] 概率向量计算单元，用于获取预配置字表的各个候选文字，利用所述预先训练的语言纠错模型分析所述候选文字在所述待纠正位置处对应的概率向量；

[0163] 目标文字确定单元，用于根据所述概率向量从所述候选文字中确定用于替代所述待纠正文字的目标文字。

[0164] B13.根据B12所述的文字纠错装置，所述目标文字确定单元包括：形近字提取子单元和目标文字选择子单元；

[0165] 其中，形近字提取子单元，用于从所述候选文字中提取所述待纠正文字的形近字及其在所述概率向量中对应的概率值；

[0166] 目标文字选择子单元，用于对所述形近字对应的概率值进行比较，根据比较结果选择概率值最大的形近字作为目标文字，以将所述目标文字替代所述待纠正文字。

[0167] B14.根据B8所述的文字纠错装置，所述文字纠错装置10还包括：字表生成模块，用于利用汉字的形近码进行编码构建出候选文字，集合所述候选文字生成预配置字表；其中，所述形近码包括汉字结构、笔画和四角码中的至少一者。

[0168] C15.一种文字纠错设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现如权利要求A1至A7任一项所述的文字纠错方法的步骤。

[0169] D16.一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行如权利要求A1至A7任一项所述文字纠错方法的步骤。

[0170] 应该理解的是，虽然附图的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，其可以以其他的顺序执行。而且，附图的流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，其执行顺序也不必然是依次进行，而是可以与其他步骤或者其他步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0171] 以上所述仅是本申请的部分实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本申请的保护范围。

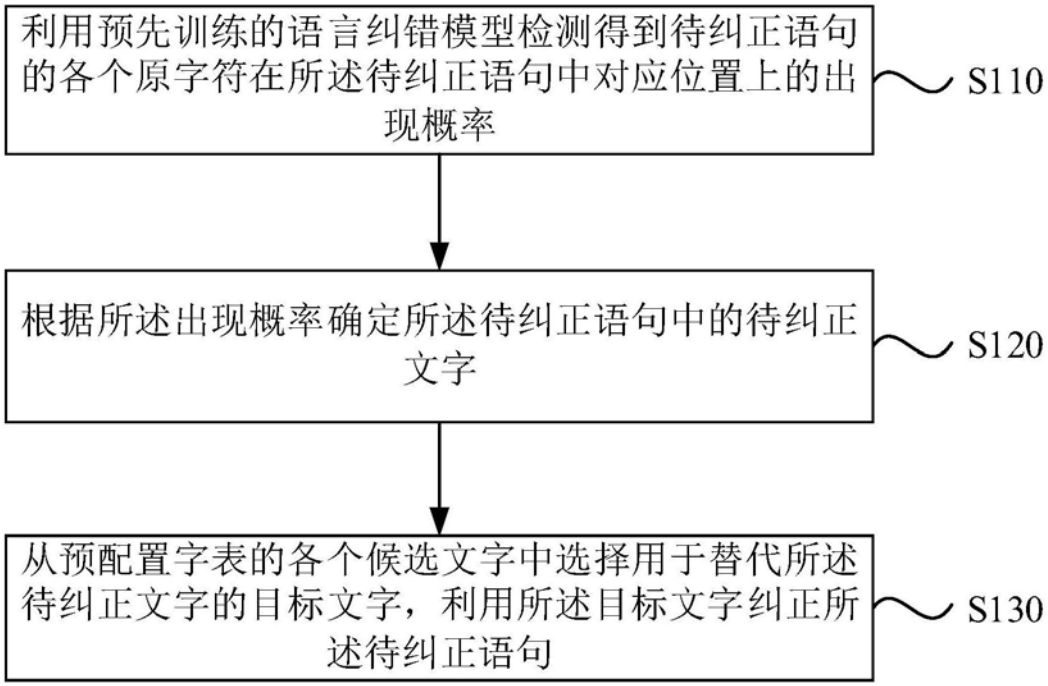
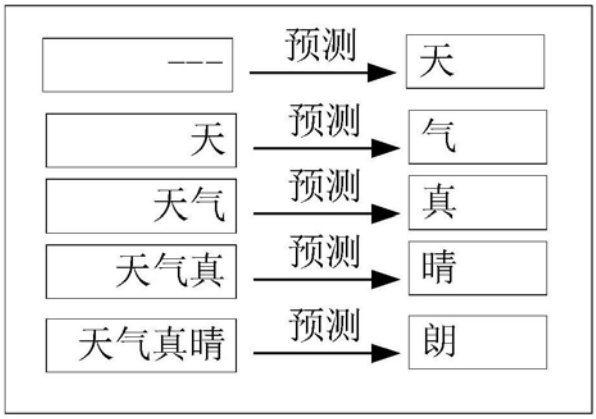

利用预先训练的语言纠错模型检测得到待纠正语句的各个原字符在所述待纠正语句中对应位置上的出现概率
S110
根据所述出现概率确定所述待纠正语句中的待纠正文字
S120
从预配置字表的各个候选文字中选择用于替代所述待纠正文字的目标文字，利用所述目标文字纠正所述待纠正语句
S130

图1



预测
天
天
预测
气
天气
预测
真
天气真
预测
晴
天气真晴
预测
朗

图2

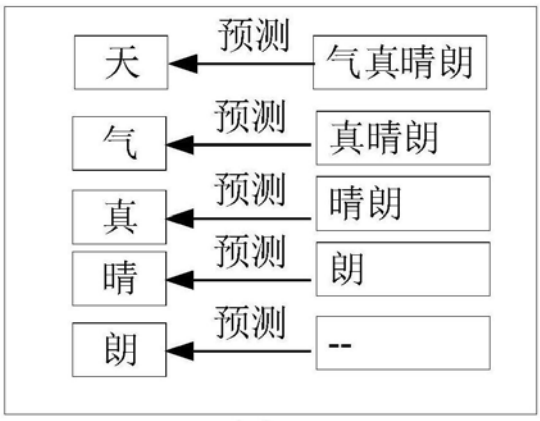

天
预测
气真晴朗
气
预测
真晴朗
真
预测
晴朗
晴
预测
朗
朗
预测
--

图3

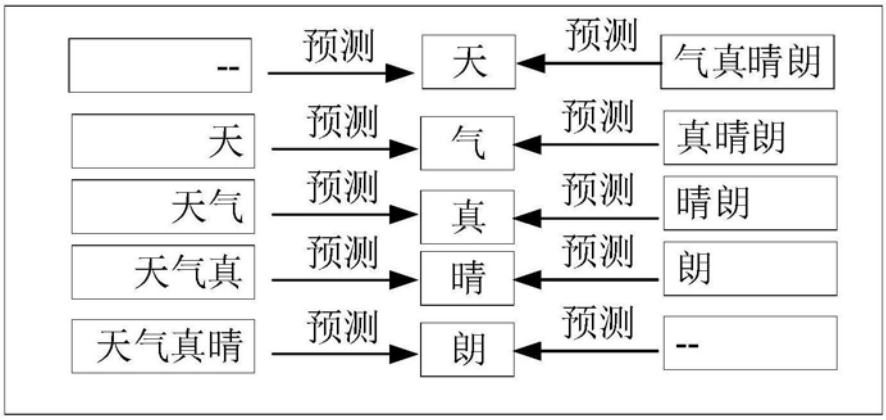

--
预测
天
预测
气真晴朗
天
预测
气
预测
真晴朗
天气
预测
真
预测
晴朗
天气真
预测
晴
预测
朗
天气真晴
预测
朗
预测
--

图4

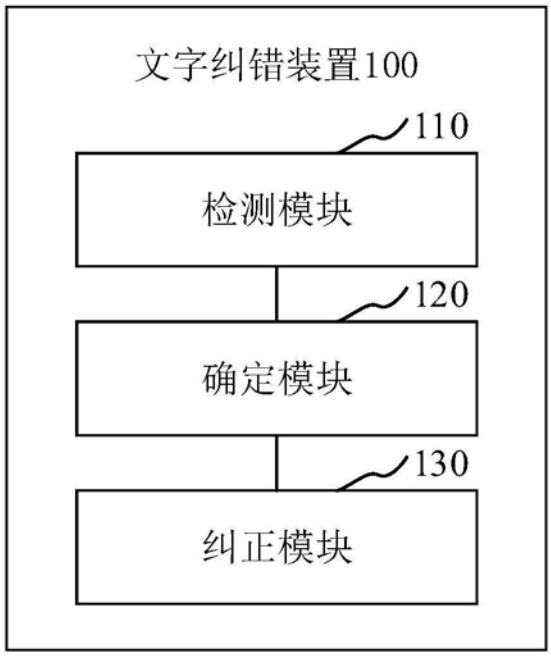

文字纠错装置100
110
检测模块
120
确定模块
130
纠正模块

图5

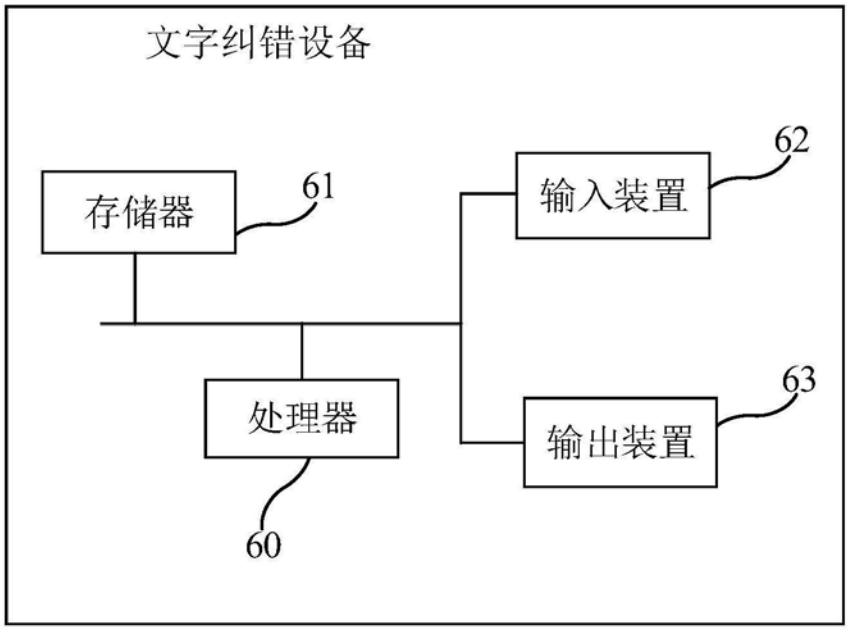

文字纠错设备
存储器
61
输入装置
62
处理器
60
输出装置
63

图6