



(21) 申请号 202110279919.9

(22) 申请日 2021.03.16

(65) 同一申请的已公布的文献号

申请公布号 CN 113051894 A

(43) 申请公布日 2021.06.29

(73) 专利权人 京东科技控股股份有限公司

地址 100176 北京市北京经济技术开发区

科创十一街18号C座2层221室

(72) 发明人 王培英 陈蒙

(74) 专利代理机构 中原信达知识产权代理有限

责任公司 11219

专利代理师 杨倩 张效荣

(51) Int. Cl.

G06F 40/232 (2020.01)

G06F 40/289 (2020.01)

(56) 对比文件

CN 111523306 A, 2020.08.11

CN 111428494 A, 2020.07.17

汪权彬, 谭营. 基于数据增广和复制的中文语法错误纠正方法. 智能系统学报. 2020, 第15卷 (第1期), 第99-105页.

审查员 杨霜雪

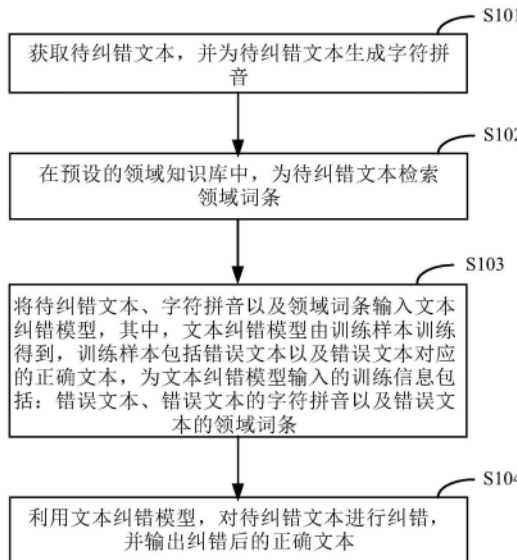
权利要求书2页 说明书11页 附图5页

(54) 发明名称

一种文本纠错的方法和装置

(57) 摘要

本发明公开了文本纠错的方法和装置,涉及计算机技术领域。该方法的具体实施方式包括:获取待纠错文本,并为待纠错文本生成字符拼音;在预设的领域知识库中,为待纠错文本检索领域词条;将待纠错文本、字符拼音以及领域词条输入文本纠错模型,其中,文本纠错模型由训练样本训练得到,训练样本包括错误文本以及对应的正确文本,为文本纠错模型输入的训练信息包括:错误文本、错误文本的字符拼音以及错误文本的领域词条;利用文本纠错模型,对待纠错文本进行纠错。该实施方式能够提高文本纠错的准确性以及效率。



1. 一种文本纠错的方法,其特征在于,包括:

获取待纠错文本,并为所述待纠错文本生成字符拼音;

在预设的领域知识库中,为所述待纠错文本检索领域词条;

将所述待纠错文本、所述字符拼音以及所述领域词条输入文本纠错模型,其中,所述文本纠错模型由训练样本训练得到,所述训练样本包括错误文本以及所述错误文本对应的正确文本,为所述文本纠错模型输入的训练信息包括:所述错误文本、所述错误文本的字符拼音以及所述错误文本的领域词条;

利用所述文本纠错模型,对所述待纠错文本进行纠错,并输出纠错后的正确文本;

为所述待纠错文本匹配领域词条,包括:

将所述待纠错文本划分为多个预设长度的字符片段;

根据所述字符片段对应的字符拼音,在预设的领域知识库中,为所述字符片段检索领域词条,其中,所述领域知识库是指包含了领域一系列专业词汇的集合。

2. 根据权利要求1所述的方法,其特征在于,进一步包括:

分别将所述待纠错文本、所述字符拼音以及所述领域词条转换为对应的向量表示;

对所述待纠错文本进行纠错,包括:

将所述待纠错文本的向量表示、所述字符拼音的向量表示以及所述领域词条的向量表示输入所述文本纠错模型;

所述文本纠错模型基于所述待纠错文本的向量表示、所述字符拼音的向量表示以及所述领域词条的向量,计算字符的输出概率分布;

根据所述字符的输出概率分布,确定所述正确文本包括的字符。

3. 根据权利要求2所述的方法,其特征在于,计算所述字符的输出概率分布,包括:

利用编码器分别对所述待纠错文本的向量表示、所述字符拼音的向量表示以及所述领域词条的向量表示进行编码;

将编码的结果输入所述文本纠错模型包括的解码器;

所述解码器根据编码的结果,计算所述字符的输出概率分布。

4. 根据权利要求3所述的方法,其特征在于,利用所述编码器分别对所述待纠错文本的向量表示、所述字符拼音的向量表示以及所述领域词条的向量表示进行编码,包括:

将所述领域词条的向量表示融入所述待纠错文本的向量表示和所述字符拼音的向量表示;

对融入后的结果进行编码。

5. 根据权利要求3所述的方法,其特征在于,进一步包括:

确定所述待纠错文本中的每一个所述字符的混淆集合,其中,所述混淆集合中包括多个近似字符;

基于所述混淆集合,所述解码器执行计算所述字符的输出概率分布的步骤。

6. 根据权利要求5所述的方法,其特征在于,计算所述字符的输出概率分布,包括:

基于所述待纠错文本以及为所述待纠错文本检索出的领域词条,计算所述待纠错文本包括的每一个字符的复制概率;

基于每一个所述字符对应的混淆集合,计算词表包括的字符的生成概率;

根据所述词表中包括的所述字符的生成概率和所述待纠错文本包括的每一个所述字

符的复制概率,计算每一个所述字符的输出概率。

7.根据权利要求6所述的方法,其特征在于,进一步包括:

根据每一个所述字符对应的混淆集合,为所述待纠错文本构建混淆集指示矩阵;

通过所述混淆集指示矩阵,将生成模式下的输出范围限制在所述混淆集合内,执行计算词表包括的每一个所述字符的生成概率的步骤。

8.根据权利要求5所述的方法,其特征在于,进一步包括:

利用每个所述训练样本的输出概率,构建损失函数;

通过最小化所述损失函数的值,训练模型参数,以得到所述文本纠错模型。

9.一种文本纠错的装置,其特征在于,包括:文本处理模块、领域匹配模块以及文本纠错模块,其中,

文本处理模块,用于获取待纠错文本,并为所述待纠错文本生成字符拼音;

领域匹配模块,用于将所述待纠错文本划分为多个预设长度的字符片段,并根据所述字符片段对应的字符拼音,在预设的领域知识库中,为所述字符片段匹配领域词条;

文本纠错模块,用于将所述待纠错文本、所述字符拼音以及所述领域词条输入文本纠错模型,利用文本纠错模型,对待纠错文本进行纠错,并输出纠错后的正确文本;其中,所述文本纠错模型由训练样本训练得到,所述文本纠错模型由训练样本训练得到,所述训练样本包括错误文本以及所述错误文本对应的正确文本,为所述文本纠错模型输入的训练信息包括:所述错误文本、所述错误文本的字符拼音以及所述错误文本的领域词条;

领域匹配模块,用于将所述待纠错文本划分为多个预设长度的字符片段;根据所述字符片段对应的字符拼音,在预设的领域知识库中,为所述字符片段检索领域词条,其中,所述领域知识库是指包含了领域一系列专业词汇的集合。

10.一种电子设备,其特征在于,包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-8中任一所述的方法。

11.一种计算机可读介质,其上存储有计算机程序,其特征在于,所述程序被处理器执行时实现如权利要求1-8中任一所述的方法。

一种文本纠错的方法和装置

技术领域

[0001] 本发明涉及计算机技术领域,尤其涉及一种文本纠错的方法和装置。

背景技术

[0002] 在很多应用场景比如检索、文本转换、意图识别、智能客服等,均涉及对文本纠错(即对文本中的错误进行修正的过程),以下游处理过程能够准确地对文本进行词法分析、意图识别等,因此从自然语言处理整体技术角度来说,文本纠错起着保驾护航的作用。

[0003] 目前,文本纠错一般依赖于人工构建的错别字词典进行错误匹配和纠正。

[0004] 在实现本发明过程中,发明人发现现有技术中至少存在如下问题:

[0005] 由于错别字典的限制,对于一些比较稀有的专有名词等可能未被收录到错别字典,导致文本纠错的准确性以及效率较低。

发明内容

[0006] 有鉴于此,本发明实施例提供一种文本纠错的方法和装置,能够有效地提高文本纠错的准确性以及效率。

[0007] 为实现上述目的,根据本发明实施例的一个方面,提供了一种文本纠错的方法,包括:

[0008] 获取待纠错文本,并为所述待纠错文本生成字符拼音;

[0009] 在预设的领域知识库中,为所述待纠错文本检索领域词条;

[0010] 将所述待纠错文本、所述字符拼音以及所述领域词条输入文本纠错模型,其中,所述文本纠错模型由训练样本训练得到,所述训练样本包括错误文本以及所述错误文本对应的正确文本,为所述文本纠错模型输入的训练信息包括:所述错误文本、所述错误文本的字符拼音以及所述错误文本的领域词条;

[0011] 利用所述文本纠错模型,对所述待纠错文本进行纠错,并输出纠错后的正确文本。

[0012] 优选地,为所述待纠错文本匹配领域词条,包括:

[0013] 将所述待纠错文本划分为多个预设长度的字符片段;

[0014] 根据所述字符片段对应的字符拼音,在预设的领域知识库中,为所述字符片段检索领域词条。

[0015] 优选地,上述文本纠错的方法,进一步包括:

[0016] 分别将所述待纠错文本、所述字符拼音以及所述领域词条转换为对应的向量表示;

[0017] 对所述待纠错文本进行纠错,包括:

[0018] 将所述待纠错文本的向量表示、所述字符拼音的向量表示以及所述领域词条的向量表示输入所述文本纠错模型;

[0019] 所述文本纠错模型基于所述待纠错文本的向量表示、所述字符拼音的向量表示以及所述领域词条的向量,计算字符的输出概率分布;

- [0020] 根据所述字符的输出概率分布,确定所述正确文本包括的字符。
- [0021] 优选地,计算所述字符的输出概率分布,包括:
- [0022] 利用所述编码器分别对所述待纠错文本的向量表示、所述字符拼音的向量表示以及所述领域词条的向量表示进行编码;
- [0023] 将编码的结果输入所述文本纠错模型包括的解码器;
- [0024] 所述解码器根据编码的结果,计算所述字符的输出概率分布的步骤。
- [0025] 优选地,利用所述编码器分别对所述待纠错文本的向量表示、所述字符拼音的向量表示以及所述领域词条的向量表示进行编码,包括:
- [0026] 将所述领域词条的向量表示融入所述待纠错文本的向量表示和所述字符拼音的向量表示;
- [0027] 对融入后的结果进行编码。
- [0028] 优选地,上述文本纠错的方法,进一步包括:
- [0029] 确定所述待纠错文本中的每一个所述字符的混淆集合,其中,所述混淆集合中包括多个近似字符;
- [0030] 基于所述混淆集合,所述解码器执行计算所述字符的输出概率分布的步骤。
- [0031] 优选地,计算所述字符的输出概率分布,包括:
- [0032] 基于所述待纠错文本以及所述领域词条,计算所述待纠错文本包括的每一个字符的复制概率;
- [0033] 基于每一个所述字符对应的混淆集合,计算词表包括的字符的生成概率;
- [0034] 根据所述词表中包括的所述字符的生成概率和所述待纠错文本包括的每一个所述字符的复制概率,计算每一个所述字符的输出概率分布。
- [0035] 优选地,上述文本纠错的方法,进一步包括:
- [0036] 根据每一个所述字符对应的混淆集合,为所述待纠错文本构建混淆集指示矩阵;
- [0037] 通过所述混淆集指示矩阵,可将生成模式下的输出范围限制在所述混淆集合内,执行计算词表包括的每一个所述字符的生成概率的步骤。
- [0038] 优选地,上述文本纠错的方法,进一步包括:
- [0039] 利用每个所述训练样本的输出概率,构建损失函数;
- [0040] 通过最小化所述损失函数的值,训练模型参数,以得到所述文本纠错模型。
- [0041] 第二方面,本发明实施例提供一种文本纠错的装置,包括:文本处理模块、领域匹配模块以及文本纠错模块,其中,
- [0042] 文本处理模块,用于获取待纠错文本,并为所述待纠错文本生成字符拼音;
- [0043] 领域匹配模块,用于将所述待纠错文本划分为多个预设长度的字符片段,并根据所述字符片段对应的字符拼音,在预设的领域知识库中,为所述字符片段检索领域词条;
- [0044] 文本纠错模块,用于将所述待纠错文本、所述字符拼音以及所述领域词条输入文本纠错模型,利用文本纠错模型,对待纠错文本进行纠错,并输出纠错后的正确文本;其中,所述文本纠错模型由训练样本训练得到,所述训练样本包括错误文本以及所述错误文本对应的正确文本,为所述文本纠错模型输入的训练信息包括:所述错误文本、所述错误文本的字符拼音以及所述错误文本的领域词条。
- [0045] 上述发明中的一个实施例具有如下优点或有益效果:由于本申请提供的方案通过

为待纠错文本生成字符拼音以及将待纠错文本划分为多个预设长度的字符片段,并根据字符片段对应的字符拼音,在预设的领域知识库中,为所述字符片段检索领域词条,即为待纠错文本引入字符拼音以及领域词条,该字符拼音以及领域词条的引入,一方面可以为待纠错文本增加特征,另一方面可以缩小文本纠错模型复制或者生成正确文本的范围,从而有效地提高文本纠错的准确性以及文本纠错效率。

[0046] 上述的非惯用的可选方式所具有的进一步效果将在下文中结合具体实施方式加以说明。

附图说明

[0047] 附图用于更好地理解本发明,不构成对本发明的不当限定。其中:

[0048] 图1是根据本发明实施例的文本纠错的方法的主要流程的示意图;

[0049] 图2是根据本发明实施例的文本纠错的主要结构的示意图;

[0050] 图3是根据本发明实施例的对待纠错文本进行纠错的主要流程的示意图;

[0051] 图4是根据本发明实施例的计算融合领域词条的字符的输出概率的主要流程的示意图;

[0052] 图5是根据本发明实施例的基于混淆集合计算融合领域词条的字符的输出概率的主要流程的示意图;

[0053] 图6是根据本发明另一实施例的基于混淆集合计算融合领域词条的字符的输出概率的主要流程的示意图;

[0054] 图7是根据本发明实施例的文本纠错的装置的主要模块的示意图;

[0055] 图8是本发明实施例可以应用于其中的示例性系统架构图;

[0056] 图9是适于用来实现本发明实施例的终端设备或服务器的计算机系统的结构示意图。

具体实施方式

[0057] 以下结合附图对本发明的示范性实施例做出说明,其中包括本发明实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本发明的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0058] 图1是根据本发明实施例的一种文本纠错的方法,如图1所示,该文本纠错的方法可包括如下步骤:

[0059] 步骤S101:获取待纠错文本,并为待纠错文本生成字符拼音;

[0060] 由于文本中的字符通常具有多种同音字,通过该字符拼音能够比较好地扩展待纠错文本的纠错方向。比如,待纠错文本为“这款洗衣机是西虹一体的吗”,通过该步骤可以生成“zhe kuan xi yi ji shi xi hong yi ti de ma”。由于拼音可以对应多种同音词,因此,通过该步骤引入字符拼音,能够为后面提供更多的纠错特征。

[0061] 步骤S102:在预设的领域知识库中,为待纠错文本检索领域词条;

[0062] 步骤S103:将待纠错文本、字符拼音以及领域词条输入文本纠错模型,其中,文本纠错模型由训练样本训练得到,训练样本包括错误文本以及错误文本对应的正确文本,为

文本纠错模型输入的训练信息包括：错误文本、错误文本的字符拼音以及错误文本的领域词条；

[0063] 步骤S104：利用文本纠错模型，对待纠错文本进行纠错，并输出纠错后的正确文本。

[0064] 其中，待纠错文本可以来源于用户输入检索框的信息、用户在智能问答页面输入的信息、语音所转换成的初级文本等。该语音所转换成的初级文本是指，通过现有的语音文本转换技术直接转换出的文本，由于语音转换技术的限制，可能导致其转换出的初级文本也存在文字错误的情况，本申请提供的方案可针对该初级文本进一步纠错。

[0065] 其中，领域知识库是指包含了该领域一系列专业词汇的集合，例如商品的属性知识库（其中列出了商品的型号系列、属性名词等信息）。领域知识库可以包括各种常规领域知识以及小众的、处于长尾分布的数据（该处于长尾分布的数据是指某一领域中特有、非常用、稀有的词汇）等。

[0066] 其中，文本纠错模型结合了网络机器翻译（NMT）任务中的模型框架以及自注意力机制的Transformer框架。该模型框架采取一种结合领域知识的端到端文本纠错模型。其中，在NMT任务中的模型框架，主要采用编码器&解码器（encoder-decoder）的序列模型，其中encoder主要负责对源语言进行编码，得到一个最终的特征向量，而解码器则根据特征向量信息、生成目标语言序列，即纠错文本。本申请通过在编码器&解码器（encoder-decoder）的序列模型中，引入自注意力机制的Transformer框架，以确定纠正后的正确文本中的字符是来源于复制还是生成，从而提高文本纠错效率。

[0067] 在图1所示的实施例，由于本申请提供的方案通过为待纠错文本生成字符拼音以及将待纠错文本划分为多个预设长度的字符片段，并根据字符片段对应的字符拼音，在预设的领域知识库中，为所述字符片段匹配领域词条，即为待纠错文本引入字符拼音以及领域词条，该字符拼音以及领域词条的引入，一方面可以为待纠错文本增加特征，另一方面可以缩小文本纠错模型复制或者生成正确文本的范围，从而有效地提高文本纠错的准确性以及文本纠错效率。

[0068] 其中，上述步骤S102的具体实施方式可包括：将待纠错文本划分为多个预设长度的字符片段；根据字符片段对应的字符拼音，在预设的领域知识库中，为字符片段检索领域词条。该检索出的领域词条为与字符片段相关的领域词条。该预设长度的字符片段可以为2个字符长度的字符片段或3个字符长度的字符片段等，该预设长度可以根据用户需求进行相应的设置。比如，针对待纠错文本“这款洗衣机是西虹一体的吗”，划分为3个字符长度的字符片段：“洗衣机”、“西虹一”，在该待纠错文本中，将常规的无意义词滤除，比如“是”、“的”等无意义词。针对“西虹一”这一字符片段的拼音“xi hong yi”，在领域知识库中匹配领域词条“洗烘一体”。即通过上述过程将待纠错文本划分为1个字符片段，相应地，可得到领域词条集合：

[0069] $K = \{k_1, k_2, \dots, k_1\}$ 。

[0070] 其中，通过文本纠错模型进行文本纠错的主要架构可如图2所示。从图2可以看出，本发明实施例提供的方案，通过编码器对待纠错文本（比如“这款洗衣机是西虹一体的吗”）、待纠错文本的拼音（比如“zhe kuan xi yi ji shi xi hong yi ti de ma”）以及相关的领域词条（比如“洗烘一体”）进行编码，并通过交叉注意力机制的方式，将领域知识融

入到原文本的编码,计算待纠错文本中各个字符的复制概率(其中,cross attention是指交叉关注,即将领域知识融入到原文本的编码),并通过解码器计算待纠错文本中各个字符的生成概率,利用各个字符的复制概率和生成概率,得到各个字符的输出概率,根据输出概率输出纠错后的正确文本。

[0071] 针对图2所示的本发明实施例提供的方案的主体过程,本发明各个实施例分别对编码过程、解码过程、计算复制概率、计算生成概率以及计算输出概率分布等过程均进行了改进。

[0072] 在本发明实施例中,上述文本纠错的方法可进一步包括:分别将待纠错文本、字符拼音以及领域词条转换为对应的向量表示;相应地,如图3所示,对待纠错文本进行纠错的具体实施方式可包括如下步骤:

[0073] 步骤S301:将待纠错文本的向量表示、字符拼音的向量表示以及领域词条的向量表示输入文本纠错模型;

[0074] 该待纠错文本的向量表示、字符拼音的向量表示以及领域词条的向量表示可通过现有的文本转换向量的方式实现,比如直接通过编码器将待纠错文本和字符拼音转换为向量表示集合 $\{\vec{h}_1^s, \vec{h}_2^s, \dots, \vec{h}_n^s\}$,以及将领域词条 $K = \{k_1, k_2, \dots, k_l\}$ 的各个领域词条转换为向量表示,得到的领域词条的向量表示集合 $\{\vec{h}_1^k, \vec{h}_2^k, \dots, \vec{h}_l^k\}$ 。

[0075] 步骤S302:文本纠错模型基于待纠错文本的向量表示、字符拼音的向量表示以及领域词条的向量,计算字符的输出概率分布;

[0076] 由于计算字符的输出概率分布过程引入了字符拼音的向量表示以及领域词条的向量,即增加了待纠错文本的特征,可有效地提高计算输出概率的准确性。

[0077] 步骤S303:根据字符的输出概率分布,确定正确文本包括的字符。

[0078] 在本发明实施例中,如图4所示,上述文本纠错的方法可进一步包括如下步骤:

[0079] 步骤S401:利用编码器分别对待纠错文本的向量表示、字符拼音的向量表示以及领域词条的向量表示进行编码;

[0080] 其中,步骤S401的具体实施方式可包括:将领域词条的向量表示融入待纠错文本的向量表示和字符拼音的向量表示;对融入后的结果进行编码。

[0081] 该编码过程可采用下述计算公式(1)和计算公式(2)实现得到融合领域词条的字符向量表示。

$$[0082] \quad \alpha_{ij} = \text{softmax}((W_q \vec{h}_i^s)(W_k \vec{h}_j^k)^T / \sqrt{d}) \quad (1)$$

$$[0083] \quad h_i = \vec{h}_i^s + \sum_j \alpha_{ij} W_v \vec{h}_j^k \quad (2)$$

[0084] 其中, α_{ij} 表征待纠错文本中的第i个字符与第j个领域词条融合后的注意力权重;softmax()表征softmax函数; W_q 、 W_k 以及 W_v 表征编码器训练出的参数矩阵,d表征 $W_q \vec{h}_i^s$ 的维度; $\vec{h}_i^s$ 表征待纠错文本中的第i个字符的向量表示; $\vec{h}_j^k$ 表征为第i个字符检索出的领域词条中第j个字符的向量表示; h_i 表征融合知识的字符表示。

[0085] 即,上述计算公式(2)的结果为编码器的输出结果(融合知识的字符表示对应的编码)。

[0086] 步骤S402:将编码的结果输入文本纠错模型包括的解码器;

[0087] 步骤S403:解码器根据编码的结果,计算字符的输出概率分布的步骤。

[0088] 由于编码引入了领域词条,解码器在解码的过程中,能够更准确的计算出字符的输出概率分布。

[0089] 在本发明实施例中,如图5所示,上述文本纠错的方法可进一步包括如下步骤:

[0090] 步骤S501:确定待纠错文本中的每一个字符的混淆集合,其中,混淆集合中包括多个近似字符;

[0091] 该步骤S501的具体实现方式可以为:预先在数据库中存储各个字符的混淆集合,如下表1所示。通过查找的方式,查找到待纠错文本中的每一个字符的混淆集合。

[0092] 表1

字符	混淆集合
升	生牲胜笙甥声…乘绳剩圣盛省
文	免问紊蚊闻纹…吻瘟温稳
…	…
心	欣薪新辛锌芯…衅信寻

[0094] 该步骤S501的具体实现方式还可以为:为待纠错文本中的每一个字符查找近似字符、近音字符以及近形字符,并将查找到的近似字符、近音字符以及近形字符组合成为对应字符的混淆集合。

[0095] 步骤S502:基于混淆集合,执行计算待纠错文本中的字符的输出概率的步骤。

[0096] 其中,如图6所示,上述步骤S502执行计算待纠错文本中的字符的输出概率的过程,可包括如下步骤:

[0097] 步骤S601:基于待纠错文本以及领域词条,计算待纠错文本包括的每一个字符的复制概率;

[0098] 该步骤的计算过程采用下述计算公式(3)计算得到。

$$[0099] \quad p_t^{\text{copy}}(i) = \text{softmax}(q_t^T \times K) \quad (3)$$

[0100] 其中, $p_t^{\text{copy}}(i)$ 表征在解码的第 t 个时刻,针对待纠错文本中的第 i 个字符的复制概率(该复制概率是指,针对待纠错文本中的第 i 个字符,从源端(待纠错文本、领域词条)进行复制的概率); $q_t = s_t \times W_q^T$, $K = H^S \times W_k^T$, 其中, s_t 表征 t 时刻解码的隐状态; W_q 和 W_k 表征训练得到的参数矩阵; $H^S = [h_1, h_2, \dots, h_n; \hat{h}_1^k, \hat{h}_2^k, \dots, \hat{h}_m^k]$ 表征原输入文本(待纠错文本)表征以及领域知识的编码表示。

[0101] 步骤S602:基于每一个字符对应的混淆集合,计算词表包括的所述字符的生成概率;

[0102] 该词表是指的每一个字符对应的混淆集合中的字符所组成的词表。

[0103] 该步骤S602的具体实现方式可包括:根据每一个字符对应的混淆集合,为待纠错文本构建混淆集指示矩阵;通过该混淆集指示矩阵,可将生成模式下的输出范围限制在混

淆集合内,计算词表内每个字符的生成概率。

[0104] 混淆集指示矩阵 $M \in R^{g \times |V|}$,其中 $|V|$ 表征待纠错文本对应的所有混淆集合包括的字符个数, g 表征待纠错文本的长度; M 中的元素 M_{if} 取值0或1。其中, M 中的元素 M_{if} 取值可通过下述计算公式(4)计算得到。

$$[0105] \quad M_{if} = \begin{cases} 1, & \text{如果 } f \text{ 为 } i \text{ 的混淆字符} \\ 0, & f \text{ 不为 } i \text{ 的混淆字符} \end{cases} \quad (4)$$

[0106] 利用混淆集指示矩阵 M ,通过计算公式(5)计算每一个字符的生成概率,可以将生成模式下的输出范围限制在混淆集合内:

$$[0107] \quad p_t^{\text{gen}}(i) = p_t^{\text{gen}} \cdot M_i \quad (5)$$

[0108] p_t^{gen} 表征对待纠错文本进行纠错解码的 t 时刻混淆集合后的生成概率; p_t^{gen} 表征在解码器中预设的对待纠错文本纠错的 t 时刻的基准生成概率; M_i 表征从混淆集指示矩阵获取的与字符 i 相关的元素所组成的新的混淆集指示矩阵(也即混淆集指示矩阵中的第 i 行形成的新的混淆集指示矩阵)。

[0109] 步骤S603:根据词表中包括的每一个字符的生成概率和待纠错文本包括的每一个字符的复制概率,计算每一个字符的输出概率分布。

[0110] 该每一个字符为词表中包括的字符以及待纠错文本中包括的字符。

[0111] 该步骤中,计算输出概率可通过下述计算公式(6)计算得到。

$$[0112] \quad P_t(i) = (1 - \beta) \times p_t^{\text{gen}}(i) + \beta \times p_t^{\text{copy}}(i) \quad (6)$$

[0113] 其中, $P_t(i)$ 表征对待纠错文本进行纠错解码过程的 t 时刻,该待纠错文本中第 i 个字符的输出概率; β 表征训练得到的复制模式的权重,其作为解码时从混淆词表生成和从待纠错文本进行复制的一个平衡因子; $p_t^{\text{gen}}(i)$ 表征由上述计算公式(5)计算得到的结果;,该 $p_t^{\text{copy}}(i)$ 表征对待纠错文本进行纠错解码的第 t 时刻,该待纠错文本中第 i 个字符的复制概率。由上述计算公式(3)计算得到。该复制概率即从源端(带纠错文本、领域词条)进行复制的概率。

[0114] 在本发明实施例中,上述文本纠错的方法可进一步包括:利用每个训练样本的输出概率,构建损失函数;通过最小化损失函数的值,训练模型参数,以得到文本纠错模型。

[0115] 上述步骤构建出的损失函数如下述计算公式(7)。

$$[0116] \quad \text{loss} = - \sum_{i'=1}^T \log(P_t(i')) \quad (7)$$

[0117] 其中, loss 表征损失值; $P_t(i')$ 表征训练样本包括的训练文本中第 i' 个字符的输出概率; T 表征训练样本包括的训练文本的总字数。

[0118] 综上可知,本发明实施例提供的方案,将领域知识融入到编码过程中,提高了文本纠错模型的检错纠错能力。另外,编码过程中加入了待纠错字符的拼音特征;同时在解码时,通过对文本纠错模型的生成范围引入了混淆集合的约束,减小了搜索空间,并提高了预测的准确性及计算效率。

[0119] 如图7所示,本发明实施例提供一种文本纠错的装置700,该文本纠错的装置700可包括:文本处理模块701、领域匹配模块702以及文本纠错模块703,其中,

[0120] 文本处理模块701,用于获取待纠错文本,并为待纠错文本生成字符拼音;

[0121] 领域匹配模块702,用于将待纠错文本划分为多个预设长度的字符片段,并根据字符片段对应的字符拼音,在预设的领域知识库中,为字符片段匹配领域词条;

[0122] 文本纠错模块703,用于将待纠错文本、字符拼音以及领域词条输入文本纠错模型,利用文本纠错模型,对待纠错文本进行纠错,并输出纠错后的正确文本;其中,文本纠错模型由训练样本训练得到,文本纠错模型由训练样本训练得到,训练样本包括错误文本以及错误文本对应的正确文本,为文本纠错模型输入的训练信息包括:错误文本、错误文本的字符拼音以及错误文本的领域词条。

[0123] 在本发明实施例中,领域匹配模块702,用于将待纠错文本划分为多个预设长度的字符片段;根据字符片段对应的字符拼音,在预设的领域知识库中,为字符片段检索领域词条。

[0124] 在本发明实施例中,文本纠错模块703,用于分别将待纠错文本、字符拼音以及领域词条转换为对应的向量表示;将待纠错文本的向量表示、字符拼音的向量表示以及领域词条的向量表示输入文本纠错模型;文本纠错模型基于待纠错文本的向量表示、字符拼音的向量表示以及领域词条的向量,计算字符的输出概率分布;根据字符的输出概率分布,确定正确文本包括的字符。

[0125] 在本发明实施例中,文本纠错模块703,进一步用于利用编码器分别对待纠错文本的向量表示、字符拼音的向量表示以及领域词条的向量表示进行编码;将编码的结果输入文本纠错模型包括的解码器;解码器根据编码的结果,计算字符的输出概率概率。

[0126] 在本发明实施例中,文本纠错模块703,用于将领域词条的向量表示融入待纠错文本的向量表示和字符拼音的向量表示;对融入后的结果进行编码。

[0127] 在本发明实施例中,文本纠错模块703,用于确定待纠错文本中的每一个字符的混淆集合,其中,混淆集合中包括多个近似字符;基于混淆集合,通过解码器执行计算字符的输出概率分布的步骤。

[0128] 在本发明实施例中,文本纠错模块703,用于基于待纠错文本以及为待纠错文本检索出的领域词条,计算待纠错文本包括的每一个字符的复制概率;基于每一个字符对应的混淆集合,计算词表内每一个字符的生成概率;根据词表内每一个字符的生成概率和待纠错文本中每一个字符的复制概率,计算每一个字符的输出概率分布。

[0129] 在本发明实施例中,文本纠错模块703,进一步用于根据每一个字符对应的混淆集合,为待纠错文本构建混淆集指示矩阵;通过所述混淆集指示矩阵,可将生成模式下的输出范围限制在所述混淆集合内,执行计算词表内的每一个字符的生成概率的步骤。

[0130] 在本发明实施例中,文本纠错模块703,进一步用于利用每个训练样本的输出概率,构建损失函数;通过最小化损失函数的值,训练模型参数,以得到文本纠错模型。

[0131] 上述文本纠错的装置可以以插件的方式安装在客户端,也可以安装在与客户端通信的服务端。

[0132] 图8示出了可以应用本发明实施例的文本纠错的方法或文本纠错的装置的示例性系统架构800。

[0133] 如图8所示,系统架构800可以包括终端设备801、802、803,网络804和服务器805。网络804用以在终端设备801、802、803和服务器805之间提供通信链路的介质。网络804可以

包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。

[0134] 用户可以使用终端设备801、802、803通过网络804与服务器805交互,以接收或发送消息等。终端设备801、802、803上可以安装有各种通讯客户端应用,例如购物类应用、网页浏览器应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等(仅为示例)。

[0135] 终端设备801、802、803可以是具有显示屏并且支持网页浏览的各种电子设备,包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

[0136] 服务器805可以是提供各种服务的服务器,例如服务端805将训练好的文本纠错模型封装到文本纠错的装置或者封装到插件中,并可以通过网络804将该文本纠错的装置或者插件发布到终端设备801、802、803安装的各种通讯客户端应用。服务端805还可将训练好的文本纠错模型封装到文本纠错的装置,并运行该文本纠错的装置。

[0137] 针对服务端805将训练好的文本纠错模型封装到文本纠错的装置或者封装到插件中,并可以通过网络804将该文本纠错的装置或者插件发布到终端设备801、802、803安装的各种通讯客户端应用的情况,在终端设备801、802、803上的通讯客户端应用接收到外部输入的文本时,将该文本作为待纠错文本,通过文本纠错的装置或者插件对待纠错文本进行纠错处理。

[0138] 针对服务端805将训练好的文本纠错模型封装到文本纠错的装置,并运行该文本纠错的装置的情况,获取用户通过终端设备801、802、803上的通讯客户端应用输入的文本,将该文本作为待纠错文本,通过文本纠错的装置或者插件对待纠错文本进行纠错处理,并将纠错处理后的正确文本输出给终端设备801、802、803上的通讯客户端应用。

[0139] 需要说明的是,本发明实施例所提供的文本纠错的方法可由终端设备801、802、803或者服务器805执行,相应地,文本纠错的装置可设置于终端设备801、802、803或者服务器805中。

[0140] 应该理解,图8中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器的。

[0141] 下面参考图9,其示出了适于用来实现本发明实施例的终端设备的计算机系统900的结构示意图。图9示出的终端设备仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0142] 如图9所示,计算机系统900包括中央处理单元(CPU)901,其可以根据存储在只读存储器(ROM)902中的程序或者从存储部分908加载到随机访问存储器(RAM)903中的程序而执行各种适当的动作和处理。在RAM 903中,还存储有系统900操作所需的各种程序和数据。CPU 901、ROM 902以及RAM 903通过总线904彼此相连。输入/输出(I/O)接口905也连接至总线904。

[0143] 以下部件连接至I/O接口905:包括键盘、鼠标等的输入部分906;包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分907;包括硬盘等的存储部分908;以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分909。通信部分909经由诸如因特网的网络执行通信处理。驱动器910也根据需要连接至I/O接口905。可拆卸介质911,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器910上,以便于从其上读出的计算机程序根据需要被安装入存储部分908。

[0144] 特别地,根据本发明公开的实施例,上文参考流程图描述的过程可以被实现为计

计算机软件程序。例如,本发明公开的实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分909从网络上被下载和安装,和/或从可拆卸介质911被安装。在该计算机程序被中央处理单元(CPU)901执行时,执行本发明的系统中限定的上述功能。

[0145] 需要说明的是,本发明所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是一——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本发明中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本发明中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、电线、光缆、RF等等,或者上述的任意合适的组合。

[0146] 附图中的流程图和框图,图示了按照本发明各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0147] 描述于本发明实施例中所涉及到的模块可以通过软件的方式实现,也可以通过硬件的方式来实现。所描述的模块也可以设置在处理器中,例如,可以描述为:一种处理器包括文本处理模块、领域匹配模块以及文本纠错模块。其中,这些模块的名称在某种情况下并不构成对该模块本身的限定,例如,文本处理模块还可以被描述为“获取待纠错文本,并为待纠错文本生成字符拼的模块”。

[0148] 作为另一方面,本发明还提供了一种计算机可读介质,该计算机可读介质可以是上述实施例中描述的设备中所包含的;也可以是单独存在,而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被一个该设备执行时,使得该设备包括:获取待纠错文本,并为待纠错文本生成字符拼音;在预设的领域知识库中,为待纠错文本检索领域词条;将待纠错文本、字符拼音以及领域词条输入文本纠错模型,其

中,文本纠错模型由训练样本训练得到,训练样本包括错误文本以及错误文本对应的正确文本,为文本纠错模型输入的训练信息包括:错误文本、错误文本的字符拼音以及错误文本的领域词条;利用文本纠错模型,对待纠错文本进行纠错,并输出纠错后的正确文本。

[0149] 根据本发明实施例的技术方案,由于本申请提供的方案通过为待纠错文本生成字符拼音以及将待纠错文本划分为多个预设长度的字符片段,并根据字符片段对应的字符拼音,在预设的领域知识库中,为所述字符片段匹配领域词条,即为待纠错文本引入字符拼音以及领域词条,该字符拼音以及领域词条的引入,一方面可以为待纠错文本增加特征,另一方面可以缩小文本纠错模型复制或者生成正确文本的范围,从而有效地提高文本纠错的准确性以及文本纠错效率。

[0150] 上述具体实施方式,并不构成对本发明保护范围的限制。本领域技术人员应该明白的是,取决于设计要求和因素,可以发生各种各样的修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等,均应包含在本发明保护范围之内。

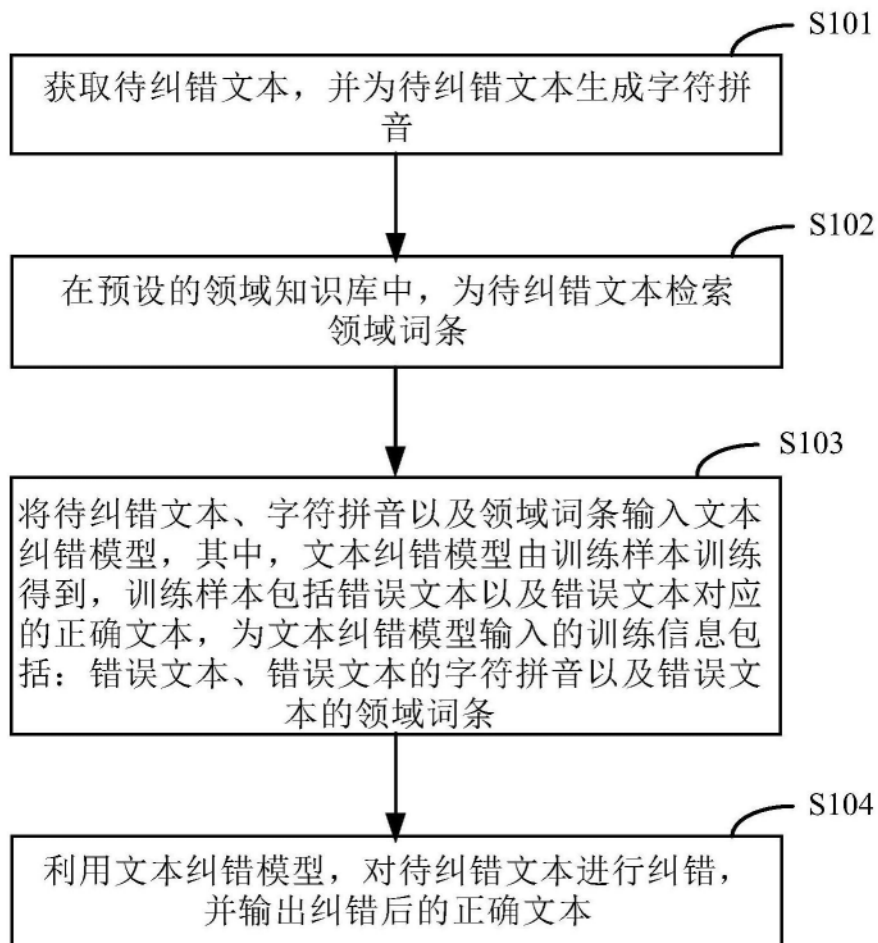


图1

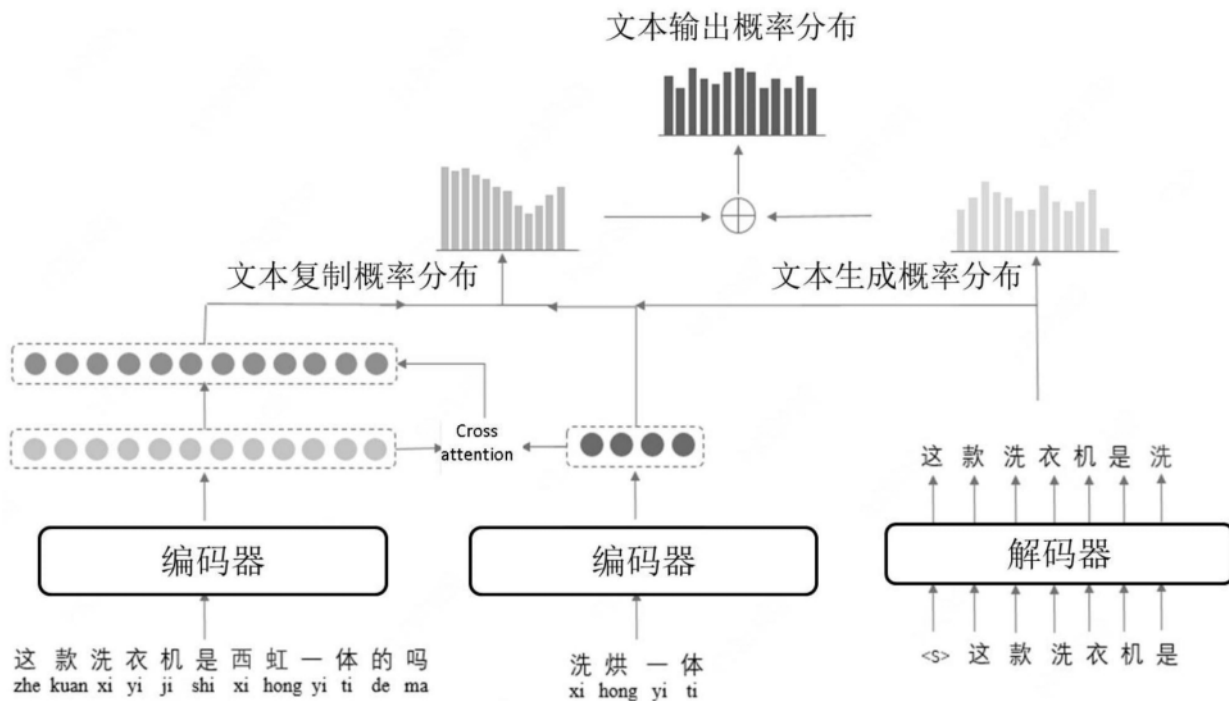


图2

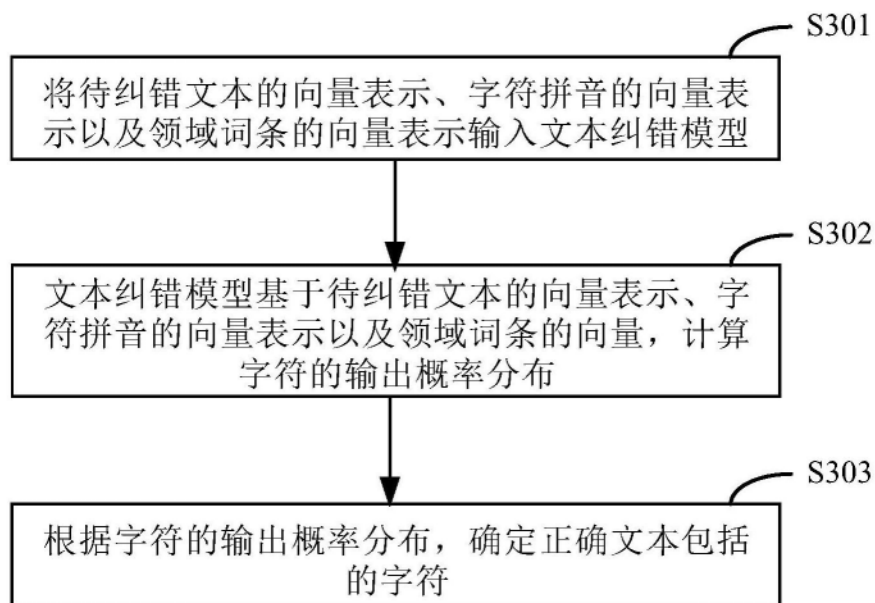


图3

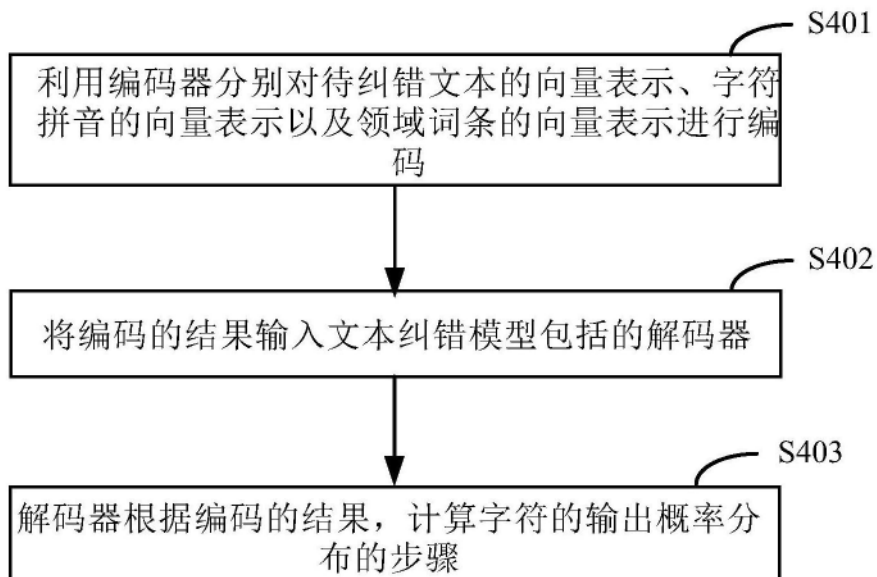


图4

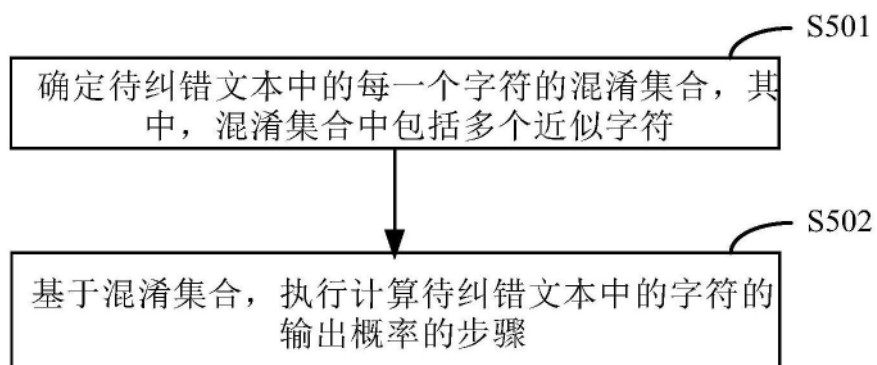


图5

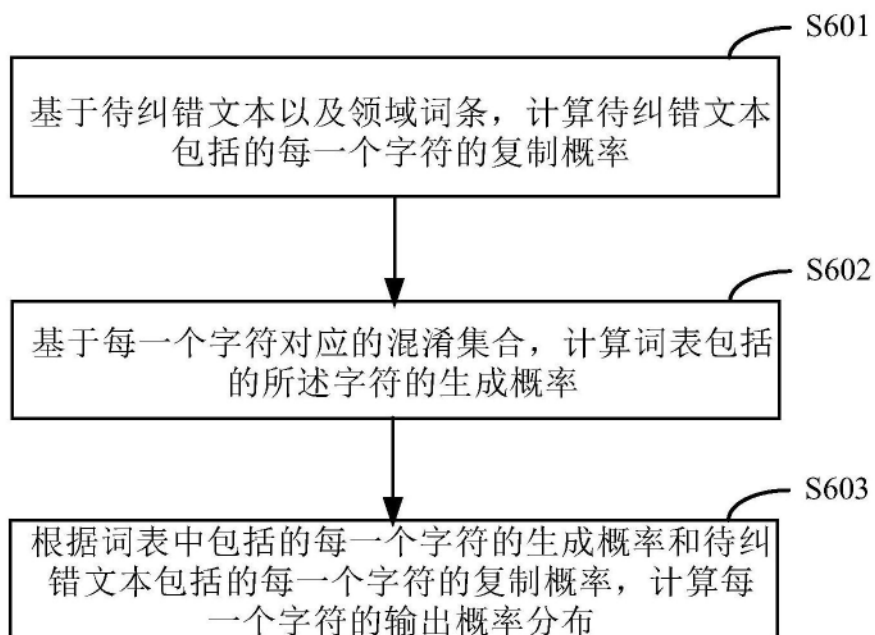


图6

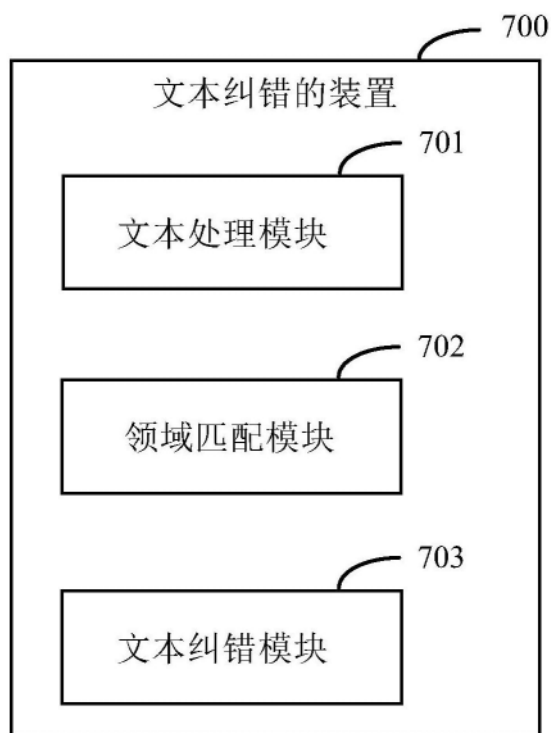


图7

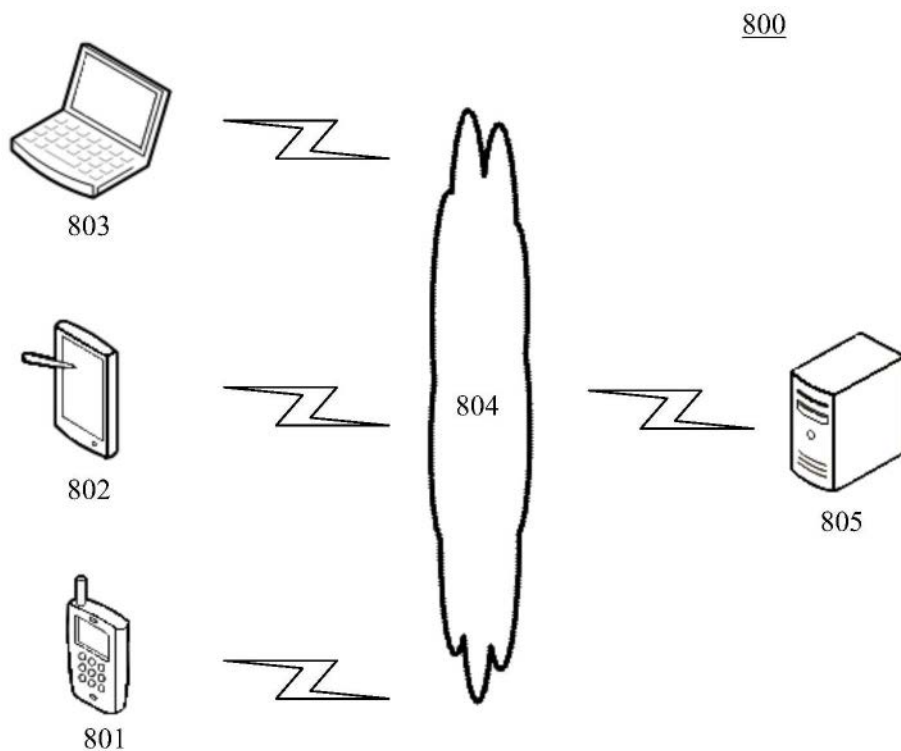


图8

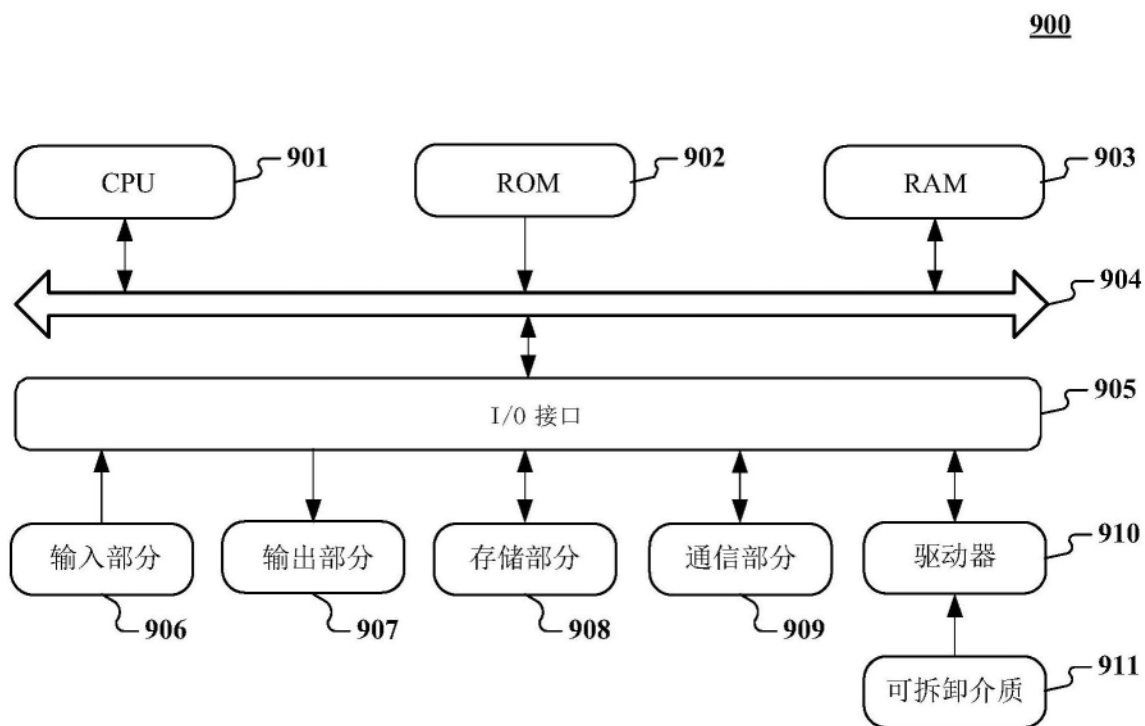


图9