(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 113627160 A
(43)申请公布日 2021.11.09

(21)申请号 202111093910.5

(22)申请日 2021.09.17

(71)申请人 平安银行股份有限公司
地址 518000 广东省深圳市罗湖区深南东路5047号

(72)发明人 李骁 赖众程 王亮 高洪喜 许海金 吴鹏召 李会璟 李兴辉 周柱君

(74)专利代理机构 深圳市沃德知识产权代理事务所(普通合伙) 44347
代理人 高杰 于志光

(51)Int.Cl.
G06F 40/232(2020.01)
G06F 40/242(2020.01)
G06F 40/289(2020.01)

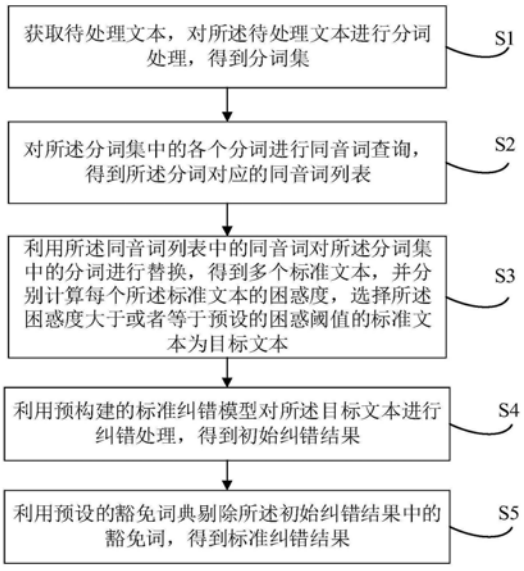
权利要求书2页 说明书12页 附图2页

(54)发明名称

文本纠错方法、装置、电子设备及存储介质

(57)摘要

本发明涉及人工智能及数字医疗技术，揭露了一种文本纠错方法，包括：对待处理文本分词，得到分词集，对分词集中的各个分词进行同音词查询，利用查询到的同音词替换对应的分词，得到多个标准文本，分别计算标准文本的困惑度，选择困惑度大于或者等于预设困惑阈值的标准文本为目标文本，利用标准纠错模型对所述目标文执行纠错处理，得到初始纠错结果，利用豁免词典对初始纠错结果进行剔除，得到标准纠错结果。此外，本发明还涉及区块链技术，分词集可存储于区块链的节点。本发明还提出一种文本纠错装置、电子设备以及存储介质。本发明可以提高文本纠错的准确度。



CN 113627160 A

1.一种文本纠错方法,其特征在于,所述方法包括:

获取待处理文本,对所述待处理文本进行分词处理,得到分词集;

对所述分词集中的各个分词进行同音词查询,得到所述分词对应的同音词列表;

利用所述同音词列表中的同音词对所述分词集中的分词进行替换,得到多个标准文本,并分别计算每个所述标准文本的困惑度,选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本;

利用预构建的标准纠错模型对所述目标文本进行纠错处理,得到初始纠错结果;

利用预设的豁免词典剔除所述初始纠错结果中的豁免词,得到标准纠错结果。

2.如权利要求1所述的文本纠错方法,其特征在于,所述对所述待处理文本进行分词处理,得到分词集,包括:

删除所述待处理文本中的特殊词和停用词,得到初始文本;

将所述初始文本输入预设的基准分词器中,得到多个分词并汇总得到分词集。

3.如权利要求1所述的文本纠错方法,其特征在于,所述利用预构建的标准纠错模型对所述目标文本进行纠错处理,得到初始纠错结果之前,所述方法还包括:

获取训练文本集,对所述训练文本集进行编码处理,得到训练向量集;

对所述训练向量集进行特征提取处理,得到特征向量集;

将所述特征向量集输入至所述预设文本纠错模型中的全连接层进行概率计算,得到预测概率值集;

计算所述预测概率值集中预测概率值和预设的真实概率值之间的交叉熵损失值;

根据所述交叉熵损失值调整所述预设文本纠错模型的内部参数,直到所述交叉熵损失值小于预设的损失阈值,得到标准纠错模型。

4.如权利要求3所述的文本纠错方法,其特征在于,所述计算所述预测概率值集中预测概率值和预设的真实概率值之间的交叉熵损失值,包括:

利用如下计算公式计算交叉熵损失值:

$$L=-\frac{1}{n}\sum_{x}[ylna+(1-y)\ln(1-a)]$$

其中,L为交叉熵损失值,x为训练文本集中的训练文本,y为预设的真实概率值,a为预测概率值,n表示训练文本集中训练文本的总数。

5.如权利要求1所述的文本纠错方法,其特征在于,所述利用预设的豁免词典剔除所述初始纠错结果中的豁免词,得到标准纠错结果,包括:

判断所述初始纠错结果中是否存在与所述豁免词典中一致的豁免词;

若所述初始纠错结果中存在与所述豁免词典中一致的豁免词,则将所述豁免词进行剔除,得到标准纠错结果。

6.如权利要求1所述的文本纠错方法,其特征在于,所述对所述分词集中的各个分词进行同音词查询,得到所述分词对应的同音词列表,包括:

获取预设的同音词词库;

根据所述同音词词库查询所述各个分词对应的同音词,并对所述同音词进行汇总,得到所述分词对应的同音词列表。

7.如权利要求1所述的文本纠错方法，其特征在于，所述分别计算每个所述标准文本的困惑度，包括：

利用如下计算公式分别计算每个所述标准文本的困惑度：

$$PP(S)=\sqrt[N]{\prod_{i=1}^{N}\frac{1}{p(w_i|w_1w_2\dots w_{i-1})}}$$

$$p(w_i|w_1w_2\dots w_{i-1})=\frac{p(w_iw_1w_2\dots w_{i-1})}{p(w_1w_2\dots w_{i-1})}$$

其中，PP(S)为所述困惑度，N为所述标准文本中的分词个数，w_i为第i个分词，$p(w_1w_2\cdots w_{i-1})$为分词$w_1w_2\cdots w_{i-1}$出现的概率。

8.一种文本纠错装置，其特征在于，所述装置包括：

文本分词模块，用于获取待处理文本，对所述待处理文本进行分词处理，得到分词集；

同音词查询模块，用于对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表；

困惑度计算模块，用于利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本；

文本纠错模块，用于利用预构建的标准纠错模型对所述目标文本进行纠错处理，得到初始纠错结果，利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果。

9.一种电子设备，其特征在于，所述电子设备包括：

至少一个处理器；以及，

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行如权利要求1至7中任意一项所述的文本纠错方法。

10.一种计算机可读存储介质，存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至7中任意一项所述的文本纠错方法。

文本纠错方法、装置、电子设备及存储介质

技术领域

[0001] 本发明涉及人工智能技术领域，尤其涉及一种文本纠错方法、装置、电子设备及计算机可读存储介质。

背景技术

[0002] 在银行客户投诉处理业务场景下，客户的投诉以投诉工单的形式在业务流程中进行流转处理。在投诉工单的整个生命周期中，在多个节点通常会产生大量文本，这些文本由人工通过键盘打字录入到系统中，因此会不可避免地录入错别字，大量错别字的存在影响了不同业务处理人员的工作效率，继而影响了客户满意度，甚至导致客户进行二次投诉。因需要对文本进行纠错。

[0003] 目前关于文本纠错的解决方案通常是构建并维护一个大型的纠错语料库，并结合预设规则进行文本纠错，这种方法需要总结一系列的业务规则，且初期构建时人力成本的投入大、后续维护的成本高、进行文本纠错的准确度较低。

发明内容

[0004] 本发明提供一种文本纠错方法、装置、电子设备及计算机可读存储介质，其主要目的在于解决进行文本纠错的准确度较低的问题。

[0005] 为实现上述目的，本发明提供的一种文本纠错方法，包括：

[0006] 获取待处理文本，对所述待处理文本进行分词处理，得到分词集；

[0007] 对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表；

[0008] 利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本；

[0009] 利用预构建的标准纠错模型对所述目标文本进行纠错处理，得到初始纠错结果；

[0010] 利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果。

[0011] 可选地，所述对所述待处理文本进行分词处理，得到分词集，包括：

[0012] 删除所述待处理文本中的特殊词和停用词，得到初始文本；

[0013] 将所述初始文本输入预设的基准分词器中，得到多个分词并汇总得到分词集。

[0014] 可选地，所述利用预构建的标准纠错模型对所述目标文本进行纠错处理，得到初始纠错结果之前，所述方法还包括：

[0015] 获取训练文本集，对所述训练文本集进行编码处理，得到训练向量集；

[0016] 对所述训练向量集进行特征提取处理，得到特征向量集；

[0017] 将所述特征向量集输入至所述预设文本纠错模型中的全连接层进行概率计算，得到预测概率值集；

[0018] 计算所述预测概率值集中预测概率值和预设的真实概率值之间的交叉熵损失值；

[0019] 根据所述交叉熵损失值调整所述预设文本纠错模型的内部参数，直到所述交叉熵

损失值小于预设的损失阈值，得到标准纠错模型。

[0020] 可选地，所述计算所述预测概率值集中预测概率值和预设的真实概率值之间的交叉熵损失值，包括：

[0021] 利用如下计算公式计算交叉熵损失值：

[0022] $$L = -\frac{1}{n}\sum_{x}[ylna + (1-y)\ln(1-a)]$$

[0023] 其中，L为交叉熵损失值，x为训练文本集中的训练文本，y为预设的真实概率值，a为预测概率值，n表示训练文本集中训练文本的总数。

[0024] 可选地，所述利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果，包括：

[0025] 判断所述初始纠错结果中是否存在与所述豁免词典中一致的豁免词；

[0026] 若所述初始纠错结果中存在与所述豁免词典中一致的豁免词，则将所述豁免词进行剔除，得到标准纠错结果。

[0027] 可选地，所述对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表，包括：

[0028] 获取预设的同音词词库；

[0029] 根据所述同音词词库查询所述各个分词对应的同音词，并对所述同音词进行汇总，得到所述分词对应的同音词列表。

[0030] 可选地，所述分别计算每个所述标准文本的困惑度，包括：

[0031] 利用如下计算公式分别计算每个所述标准文本的困惑度：

[0032] $$PP(S) = \sqrt[N]{\prod_{i=1}^{N}\frac{1}{p(w_i|w_1w_2\ldots w_{i-1})}}$$

[0033] $$p(w_i|w_1w_2\ldots w_{i-1}) = \frac{p(w_iw_1w_2\ldots w_{i-1})}{p(w_1w_2\ldots w_{i-1})}$$

[0034] 其中，PP(S)为所述困惑度，N为所述标准文本中的分词个数，w_i为第i个分词，$p(w_1w_2\cdots w_{i-1})$为分词$w_1w_2\cdots w_{i-1}$出现的概率。

[0035] 为了解决上述问题，本发明还提供一种文本纠错装置，所述装置包括：

[0036] 文本分词模块，用于获取待处理文本，对所述待处理文本进行分词处理，得到分词集；

[0037] 同音词查询模块，用于对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表；

[0038] 困惑度计算模块，用于利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本；

[0039] 文本纠错模块，用于利用预构建的标准纠错模型对所述目标文本进行纠错处理，得到初始纠错结果，利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠

错结果。

[0040] 为了解决上述问题,本发明还提供一种电子设备,所述电子设备包括:

[0041] 至少一个处理器;以及,

[0042] 与所述至少一个处理器通信连接的存储器;其中,

[0043] 所述存储器存储有可被所述至少一个处理器执行的计算机程序,所述计算机程序被所述至少一个处理器执行,以使所述至少一个处理器能够执行上述所述的文本纠错方法。

[0044] 为了解决上述问题,本发明还提供一种计算机可读存储介质,所述计算机可读存储介质中存储有至少一个计算机程序,所述至少一个计算机程序被电子设备中的处理器执行以实现上述所述的文本纠错方法。

[0045] 本发明实施例通过对待处理文本进行分词处理,得到分词集,将所述待处理文本拆分为一个一个单独的分词,便于后续进行同音词查询,所述同音词查询可以得到所述分词对应的同音词列表,增大了样本的数量,并利用所述同音词列表中的同音词对所述分词集中的分词进行替换,得到多个标准文本,并分别计算每个所述标准文本的困惑度,利用所述困惑度作为标准进行筛选,得到符合要求的目标文本。将所述目标文本输入至标准纠错模型中,得到初始纠错结果,所述标准纠错模型是利用训练文本集对文本纠错模型进行训练所得,进行纠错的效率较高。同时利用预设的豁免词典对所述初始纠错结果进行剔除,得到标准纠错结果。所述豁免词典可以起到二次纠错的效果,保证了标准纠错结果的准确度。因此本发明提出的文本纠错方法、装置、电子设备及计算机可读存储介质,可以解决文本纠错的准确度较低的问题。

附图说明

[0046] 图1为本发明一实施例提供的文本纠错方法的流程示意图;

[0047] 图2为本发明一实施例提供的文本纠错装置的功能模块图;

[0048] 图3为本发明一实施例提供的实现所述文本纠错方法的电子设备的结构示意图。

[0049] 本发明目的的实现、功能特点及优点将结合实施例,参照附图做进一步说明。

具体实施方式

[0050] 应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。

[0051] 本申请实施例提供一种文本纠错方法。所述文本纠错方法的执行主体包括但不限于服务端、终端等能够被配置为执行本申请实施例提供的该方法的电子设备中的至少一种。换言之,所述文本纠错方法可以由安装在终端设备或服务端设备的软件或硬件来执行,所述软件可以是区块链平台。所述服务端包括但不限于:单台服务器、服务器集群、云端服务器或云端服务器集群等。所述服务器可以是独立的服务器,也可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(ContentDelivery Network,CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0052] 参照图1所示,为本发明一实施例提供的文本纠错方法的流程示意图。在本实施例中,所述文本纠错方法包括:

[0053] S1、获取待处理文本，对所述待处理文本进行分词处理，得到分词集。

[0054] 本发明实施例中，所述待处理文本可以为银行客户投诉处理业务场景下的投诉工单相关的文本，例如，所述待处理文本可以为问题受理时的客户投诉记录文本，也可以为核实情况时的核实文本，或者处理意见时的处理意见文本。

[0055] 具体地，所述对所述待处理文本进行分词处理，得到分词集，包括：

[0056] 删除所述待处理文本中的特殊词和停用词，得到初始文本；

[0057] 将所述初始文本输入预设的基准分词器中，得到多个分词并汇总得到分词集。

[0058] 详细地，所述基准分词器的选择不受限制，可以选择基于字典的字符串匹配模型的分词器，也可以选择属于基于字符标注的机器学习模型的分词器，如stanford分词器。

[0059] 具体实施时，当选择不同的基准分词器，可能得到不同的分词集。在本方明实施例中，可以选择jieba分词器作为基准分词器进行分词处理。

[0060] 例如，所述待处理文本为“这个编号001的银行工作人员工作不认真”，利用jieba分词器进行分词处理的所述分词集为“这个/编号/001/的/银行/工作人员/工作/不认真”。

[0061] 本发明其中一个实施例中，所述待处理文本可以是数字医疗领域的文本，如医生开具的病历单等。

[0062] S2、对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表。

[0063] 本发明实施例中，所述对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表，包括：

[0064] 获取预设的同音词词库；

[0065] 根据所述同音词词库查询所述各个分词对应的同音词，并对所述同音词进行汇总，得到所述分词对应的同音词列表。

[0066] 详细地，所述同音词词库中包含目标词和所述目标词对应的多个同音词，其中，所述同音词包括：声、韵、调完全相同，分为同形同音词与异形同音词。

[0067] 例如，所述目标词为“工作人员”，所述目标词对应同音词可以为“龚作人员”，也可以为“公作人员”。

[0068] S3、利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本。

[0069] 本发明实施例中，利用所述同音词列表中的同音词对所述分词集中的分词进行替换，由于所述同音词列表中存在多个同音词，可以分别利用同音词对所述分词进行替换，得到多个标准文本。

[0070] 例如，所述分词集为“这个/编号/001/的/银行/工作人员/工作/不认真”，而所述分词“工作人员”的同音词可以为“龚作人员”或者“公作人员”，故利用所述同音词对所述分词进行替换可以得到“这个/编号/001/的/银行/龚作人员/工作/不认真”及“这个/编号/001/的/银行/公作人员/工作/不认真”两个标准文本。

[0071] 具体地，所述分别计算每个所述标准文本的困惑度，包括：

[0072] 利用如下计算公式分别计算每个所述标准文本的困惑度：

[0073] $$PP(S)=\sqrt[N]{\prod_{i=1}^{N}\frac{1}{p(w_i|w_1w_2\dots w_{i-1})}}$$

[0074] $$p(w_i|w_1w_2\dots w_{i-1})=\frac{p(w_iw_1w_2\dots w_{i-1})}{p(w_1w_2\dots w_{i-1})}$$

[0075] 其中，PP(S)为所述困惑度，N为所述标准文本中的分词个数，w_i为第i个分词，$p(w_1w_2\cdots w_{i-1})$为分词$w_1w_2\cdots w_{i-1}$出现的概率。

[0076] 详细地，所述困惑度用来度量一个概率分布或概率模型预测样本的好坏程度。

[0077] 具体地，将计算得到的困惑度与预设的困惑阈值进行比较，并将所述困惑度大于或者等于所述困惑阈值的标准文本作为目标文本，同时可以目标文本中的同音词用[MASK]代替。

[0078] S4、利用预构建的标准纠错模型对所述目标文本进行纠错处理，得到初始纠错结果。

[0079] 本发明实施例中，将所述目标文本输入至所述标准纠错模型中，所述标准纠错模型可以进行预测，即具有预测所述目标文本中[MASK]位置的能力，完成预测，取概率最大的字为预测结果，即初始纠错结果。

[0080] 本发明其中一个实施例中，S4之前还可以包括：获取训练文本集，利用所述训练文本集对预设的文本纠错模型进行训练，得到所述标准纠错模型。

[0081] 本发明实施例中，所述训练文本集可以为预设时间段内的文本数据，例如，某银行最近一年内的大概10万条投诉工单，通过多条投诉工单数据对预设的bert-base-chinese模型进行训练，得到所述标准纠错模型。

[0082] 详细地，基于bert-base-chinese进行投诉工单文本领域预训练，其中，batch_size＝64,learning_rate＝3e-5,num_train_steps＝50000,num_warmup_steps＝5000，得到预训练模型bert-private-complaint。

[0083] 具体地，所述利用所述训练文本集对预设的文本纠错模型进行训练，得到标准纠错模型，包括：

[0084] 对所述训练文本集进行编码处理，得到训练向量集；

[0085] 对所述训练向量集进行特征提取处理，得到特征向量集；

[0086] 将所述特征向量集输入至所述预设文本纠错模型中的全连接层进行概率计算，得到预测概率值集；

[0087] 计算所述预测概率值集中预测概率值和预设的真实概率值之间的交叉熵损失值；

[0088] 根据所述交叉熵损失值调整所述预设文本纠错模型的内部参数，直到所述交叉熵损失值小于预设的损失阈值，得到标准纠错模型。

[0089] 进一步地，所述计算所述预测概率值集中预测概率值和预设的真实概率值之间的交叉熵损失值，包括：

[0090] 利用如下计算公式计算交叉熵损失值：

[0091] $$L = -\frac{1}{n}\sum_{x}[ylna + (1-y)\ln(1-a)]$$

[0092] 其中，L为交叉熵损失值，x为训练文本集中的训练文本，y为预设的真实概率值，a为预测概率值，n表示训练文本集中训练文本的总数。

[0093] 具体地，根据所述交叉熵损失值调整所述预设文本纠错模型的内部参数，判断所述交叉熵损失值与所述损失阈值之间的大小，若所述交叉熵损失值大于或者等于预设的损失阈值，调整所述文本纠错模型的内部参数，将所述训练文本集输入至调整后的文本纠错模型中，得到预测概率值并重新计算交叉熵损失值，直到所述交叉熵损失值小于预设的损失阈值，得到标准纠错模型。

[0094] 优选地，内部参数可以为模型的梯度，模型的权重。

[0095] S5、利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果。

[0096] 本发明实施例中，所述利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果，包括：

[0097] 判断所述初始纠错结果中是否存在与所述豁免词典中一致的豁免词；

[0098] 若所述初始纠错结果中存在与所述豁免词典中一致的豁免词，则将所述豁免词进行剔除，得到标准纠错结果。

[0099] 其中，所述豁免词典是针对各种不同的应用场景下的专有词汇进行统计汇总形成的设定词典，在本发明实施例中，所述豁免词典中包含本场景下专有词汇，比如银行产品名称，银行专业术语词语等。

[0100] 本发明实施例通过对待处理文本进行分词处理，得到分词集，将所述待处理文本拆分为一个一个单独的分词，便于后续进行同音词查询，所述同音词查询可以得到所述分词对应的同音词列表，增大了样本的数量，并利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，利用所述困惑度作为标准进行筛选，得到符合要求的目标文本。将所述目标文本输入至标准纠错模型中，得到初始纠错结果，所述标准纠错模型是利用训练文本集对文本纠错模型进行训练所得，进行纠错的效率较高。同时利用预设的豁免词典对所述初始纠错结果进行剔除，得到标准纠错结果。所述豁免词典可以起到二次纠错的效果，保证了标准纠错结果的准确度。因此本发明提出的文本纠错方法可以解决文本纠错的准确度较低的问题。

[0101] 如图2所示，是本发明一实施例提供的文本纠错装置的功能模块图。

[0102] 本发明所述文本纠错装置100可以安装于电子设备中。根据实现的功能，所述文本纠错装置100可以包括文本分词模块101、同音词查询模块102、困惑度计算模块103及文本纠错模块104。本发明所述模块也可以称之为单元，是指一种能够被电子设备处理器所执行，并且能够完成固定功能的一系列计算机程序段，其存储在电子设备的存储器中。

[0103] 在本实施例中，关于各模块/单元的功能如下：

[0104] 所述文本分词模块101，用于获取待处理文本，对所述待处理文本进行分词处理，得到分词集；

[0105] 所述同音词查询模块102，用于对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表；

[0106] 所述困惑度计算模块103，用于利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本；

[0107] 所述文本纠错模块104，用于利用预构建的标准纠错模型对所述目标文本进行纠错处理，得到初始纠错结果，利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果。

[0108] 详细地，所述文本纠错装置100各模块的具体实施方式如下：

[0109] 步骤一、获取待处理文本，对所述待处理文本进行分词处理，得到分词集。

[0110] 本发明实施例中，所述待处理文本可以为银行客户投诉处理业务场景下的投诉工单相关的文本，例如，所述待处理文本可以为问题受理时的客户投诉记录文本，也可以为核实情况时的核实文本，或者处理意见时的处理意见文本。

[0111] 具体地，所述对所述待处理文本进行分词处理，得到分词集，包括：

[0112] 删除所述待处理文本中的特殊词和停用词，得到初始文本；

[0113] 将所述初始文本输入预设的基准分词器中，得到多个分词并汇总得到分词集。

[0114] 详细地，所述基准分词器的选择不受限制，可以选择基于字典的字符串匹配模型的分词器，也可以选择属于基于字符标注的机器学习模型的分词器，如stanford分词器。

[0115] 具体实施时，当选择不同的基准分词器，可能得到不同的分词集。在本方明实施例中，可以选择jieba分词器作为基准分词器进行分词处理。

[0116] 例如，所述待处理文本为“这个编号001的银行工作人员工作不认真”，利用jieba分词器进行分词处理的所述分词集为“这个/编号/001/的/银行/工作人员/工作/不认真”。

[0117] 本发明其中一个实施例中，所述待处理文本可以是数字医疗领域的文本，如医生开具的病历单等。

[0118] 步骤二、对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表。

[0119] 本发明实施例中，所述对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表，包括：

[0120] 获取预设的同音词词库；

[0121] 根据所述同音词词库查询所述各个分词对应的同音词，并对所述同音词进行汇总，得到所述分词对应的同音词列表。

[0122] 详细地，所述同音词词库中包含目标词和所述目标词对应的多个同音词，其中，所述同音词包括：声、韵、调完全相同，分为同形同音词与异形同音词。

[0123] 例如，所述目标词为“工作人员”，所述目标词对应同音词可以为“龚作人员”，也可以为“公作人员”。

[0124] 步骤三、利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本。

[0125] 本发明实施例中，利用所述同音词列表中的同音词对所述分词集中的分词进行替换，由于所述同音词列表中存在多个同音词，可以分别利用同音词对所述分词进行替换，得到多个标准文本。

[0126] 例如,所述分词集为“这个/编号/001/的/银行/工作人员/工作/不认真”,而所述分词“工作人员”的同音词可以为“龚作人员”或者“公作人员”,故利用所述同音词对所述分词进行替换可以得到“这个/编号/001/的/银行/龚作人员/工作/不认真”及“这个/编号/001/的/银行/公作人员/工作/不认真”两个标准文本。

[0127] 具体地,所述分别计算每个所述标准文本的困惑度,包括:

[0128] 利用如下计算公式分别计算每个所述标准文本的困惑度:

[0129] $$PP(S)=\sqrt[N]{\prod_{i=1}^{N}\frac{1}{p(w_i|w_1w_2\dots w_{i-1})}}$$

[0130] $$p(w_i|w_1w_2\dots w_{i-1})=\frac{p(w_iw_1w_2\dots w_{i-1})}{p(w_1w_2\dots w_{i-1})}$$

[0131] 其中,PP(S)为所述困惑度,N为所述标准文本中的分词个数,w_i为第i个分词,$p(w_1w_2\cdots w_{i-1})$为分词$w_1w_2\cdots w_{i-1}$出现的概率。

[0132] 详细地,所述困惑度用来度量一个概率分布或概率模型预测样本的好坏程度。

[0133] 具体地,将计算得到的困惑度与预设的困惑阈值进行比较,并将所述困惑度大于或者等于所述困惑阈值的标准文本作为目标文本,同时可以目标文本中的同音词用[MASK]代替。

[0134] 步骤四、利用预构建的标准纠错模型对所述目标文本进行纠错处理,得到初始纠错结果。

[0135] 本发明实施例中,将所述目标文本输入至所述标准纠错模型中,所述标准纠错模型可以进行预测,即具有预测所述目标文本中[MASK]位置的能力,完成预测,取概率最大的字为预测结果,即初始纠错结果。

[0136] 本发明其中一个实施例中,步骤四之前还可以包括:获取训练文本集,利用所述训练文本集对预设的文本纠错模型进行训练,得到所述标准纠错模型。

[0137] 本发明实施例中,所述训练文本集可以为预设时间段内的文本数据,例如,某银行最近一年内的大概10万条投诉工单,通过多条投诉工单数据对预设的bert-base-chinese模型进行训练,得到所述标准纠错模型。

[0138] 详细地,基于bert-base-chinese进行投诉工单文本领域预训练,其中,batch_size=64,learning_rate=3e-5,num_train_steps=50000,num_warmup_steps=5000,得到预训练模型bert-private-complaint。

[0139] 具体地,所述利用所述训练文本集对预设的文本纠错模型进行训练,得到标准纠错模型,包括:

[0140] 对所述训练文本集进行编码处理,得到训练向量集;

[0141] 对所述训练向量集进行特征提取处理,得到特征向量集;

[0142] 将所述特征向量集输入至所述预设文本纠错模型中的全连接层进行概率计算,得到预测概率值集;

[0143] 计算所述预测概率值集中预测概率值和预设的真实概率值之间的交叉熵损失值;

[0144] 根据所述交叉熵损失值调整所述预设文本纠错模型的内部参数,直到所述交叉熵

损失值小于预设的损失阈值，得到标准纠错模型。

[0145] 进一步地，所述计算所述预测概率值集中预测概率值和预设的真实概率值之间的交叉熵损失值，包括：

[0146] 利用如下计算公式计算交叉熵损失值：

[0147] $$L = -\frac{1}{n}\sum_{x}[ylna + (1-y)\ln(1-a)]$$

[0148] 其中，L为交叉熵损失值，x为训练文本集中的训练文本，y为预设的真实概率值，a为预测概率值，n表示训练文本集中训练文本的总数。

[0149] 具体地，根据所述交叉熵损失值调整所述预设文本纠错模型的内部参数，判断所述交叉熵损失值与所述损失阈值之间的大小，若所述交叉熵损失值大于或者等于预设的损失阈值，调整所述文本纠错模型的内部参数，将所述训练文本集输入至调整后的文本纠错模型中，得到预测概率值并重新计算交叉熵损失值，直到所述交叉熵损失值小于预设的损失阈值，得到标准纠错模型。

[0150] 优选地，内部参数可以为模型的梯度，模型的权重。

[0151] 步骤五、利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果。

[0152] 本发明实施例中，所述利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果，包括：

[0153] 判断所述初始纠错结果中是否存在与所述豁免词典中一致的豁免词；

[0154] 若所述初始纠错结果中存在与所述豁免词典中一致的豁免词，则将所述豁免词进行剔除，得到标准纠错结果。

[0155] 其中，所述豁免词典是针对各种不同的应用场景下的专有词汇进行统计汇总形成的设定词典，在本发明实施例中，所述豁免词典中包含本场景下专有词汇，比如银行产品名称，银行专业术语词语等。

[0156] 本发明实施例通过对待处理文本进行分词处理，得到分词集，将所述待处理文本拆分为一个一个单独的分词，便于后续进行同音词查询，所述同音词查询可以得到所述分词对应的同音词列表，增大了样本的数量，并利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，利用所述困惑度作为标准进行筛选，得到符合要求的目标文本。将所述目标文本输入至标准纠错模型中，得到初始纠错结果，所述标准纠错模型是利用训练文本集对文本纠错模型进行训练所得，进行纠错的效率较高。同时利用预设的豁免词典对所述初始纠错结果进行剔除，得到标准纠错结果。所述豁免词典可以起到二次纠错的效果，保证了标准纠错结果的准确度。因此本发明提出的文本纠错装置可以解决文本纠错的准确度较低的问题。

[0157] 如图3所示，是本发明一实施例提供的实现文本纠错方法的电子设备的结构示意图。

[0158] 所述电子设备1可以包括处理器10、存储器11、通信总线12以及通信接口13，还可以包括存储在所述存储器11中并可在所述处理器10上运行的计算机程序，如文本纠错程序。

[0159] 其中,所述处理器10在一些实施例中可以由集成电路组成,例如可以由单个封装的集成电路所组成,也可以是由多个相同功能或不同功能封装的集成电路所组成,包括一个或者多个中央处理器(Central Processing unit,CPU)、微处理器、数字处理芯片、图形处理器及各种控制芯片的组合等。所述处理器10是所述电子设备的控制核心(Control Unit),利用各种接口和线路连接整个电子设备的各个部件,通过运行或执行存储在所述存储器11内的程序或者模块(例如执行文本纠错程序等),以及调用存储在所述存储器11内的数据,以执行电子设备的各种功能和处理数据。

[0160] 所述存储器11至少包括一种类型的可读存储介质,所述可读存储介质包括闪存、移动硬盘、多媒体卡、卡型存储器(例如:SD或DX存储器等)、磁性存储器、磁盘、光盘等。所述存储器11在一些实施例中可以是电子设备的内部存储单元,例如该电子设备的移动硬盘。所述存储器11在另一些实施例中也可以是电子设备的外部存储设备,例如电子设备上配备的插接式移动硬盘、智能存储卡(Smart Media Card,SMC)、安全数字(Secure Digital,SD)卡、闪存卡(Flash Card)等。进一步地,所述存储器11还可以既包括电子设备的内部存储单元也包括外部存储设备。所述存储器11不仅可以用于存储安装于电子设备的应用软件及各类数据,例如文本纠错程序的代码等,还可以用于暂时地存储已经输出或者将要输出的数据。

[0161] 所述通信总线12可以是外设部件互连标准(peripheral component interconnect,简称PCI)总线或扩展工业标准结构(extended industry standard architecture,简称EISA)总线等。该总线可以分为地址总线、数据总线、控制总线等。所述总线被设置为实现所述存储器11以及至少一个处理器10等之间的连接通信。

[0162] 所述通信接口13用于上述电子设备与其他设备之间的通信,包括网络接口和用户接口。可选地,所述网络接口可以包括有线接口和/或无线接口(如WI-FI接口、蓝牙接口等),通常用于在该电子设备与其他电子设备之间建立通信连接。所述用户接口可以是显示器(Display)、输入单元(比如键盘(Keyboard)),可选地,用户接口还可以是标准的有线接口、无线接口。可选地,在一些实施例中,显示器可以是LED显示器、液晶显示器、触控式液晶显示器以及OLED(Organic Light-Emitting Diode,有机发光二极管)触摸器等。其中,显示器也可以适当的称为显示屏或显示单元,用于显示在电子设备中处理的信息以及用于显示可视化的用户界面。

[0163] 图3仅示出了具有部件的电子设备,本领域技术人员可以理解的是,图3示出的结构并不构成对所述电子设备1的限定,可以包括比图示更少或者更多的部件,或者组合某些部件,或者不同的部件布置。

[0164] 例如,尽管未示出,所述电子设备还可以包括给各个部件供电的电源(比如电池),优选地,电源可以通过电源管理装置与所述至少一个处理器10逻辑相连,从而通过电源管理装置实现充电管理、放电管理、以及功耗管理等功能。电源还可以包括一个或一个以上的直流或交流电源、再充电装置、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。所述电子设备还可以包括多种传感器、蓝牙模块、Wi-Fi模块等,在此不再赘述。

[0165] 应该了解,所述实施例仅为说明之用,在专利申请范围上并不受此结构的限制。

[0166] 所述电子设备1中的所述存储器11存储的文本纠错程序是多个指令的组合,在所

述处理器10中运行时，可以实现：

[0167] 获取待处理文本，对所述待处理文本进行分词处理，得到分词集；

[0168] 对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表；

[0169] 利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本；

[0170] 利用预构建的标准纠错模型对所述目标文本进行纠错处理，得到初始纠错结果；

[0171] 利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果。

[0172] 具体地，所述处理器10对上述指令的具体实现方法可参考附图对应实施例中相关步骤的描述，在此不赘述。

[0173] 进一步地，所述电子设备1集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读存储介质中。所述计算机可读存储介质可以是易失性的，也可以是非易失性的。例如，所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)。

[0174] 本发明还提供一种计算机可读存储介质，所述可读存储介质存储有计算机程序，所述计算机程序在被电子设备的处理器所执行时，可以实现：

[0175] 获取待处理文本，对所述待处理文本进行分词处理，得到分词集；

[0176] 对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表；

[0177] 利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本；

[0178] 利用预构建的标准纠错模型对所述目标文本进行纠错处理，得到初始纠错结果；

[0179] 利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果。

[0180] 在本发明所提供的几个实施例中，应该理解到，所揭露的设备，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

[0181] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0182] 另外，在本发明各个实施例中的各功能模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

[0183] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。

[0184] 因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图标记视为限

制所涉及的权利要求。

[0185] 本发明所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain),本质上是一个去中心化的数据库,是一串使用密码学方法相关联产生的数据块,每一个数据块中包含了一批次网络交易的信息,用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

[0186] 本申请实施例可以基于人工智能技术对相关的数据进行获取和处理。其中,人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

[0187] 此外,显然“包括”一词不排除其他单元或步骤,单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一、第二等词语用来表示名称,而并不表示任何特定的顺序。

[0188] 最后应说明的是,以上实施例仅用以说明本发明的技术方案而非限制,尽管参照较佳实施例对本发明进行了详细说明,本领域的普通技术人员应当理解,可以对本发明的技术方案进行修改或等同替换,而不脱离本发明技术方案的精神和范围。

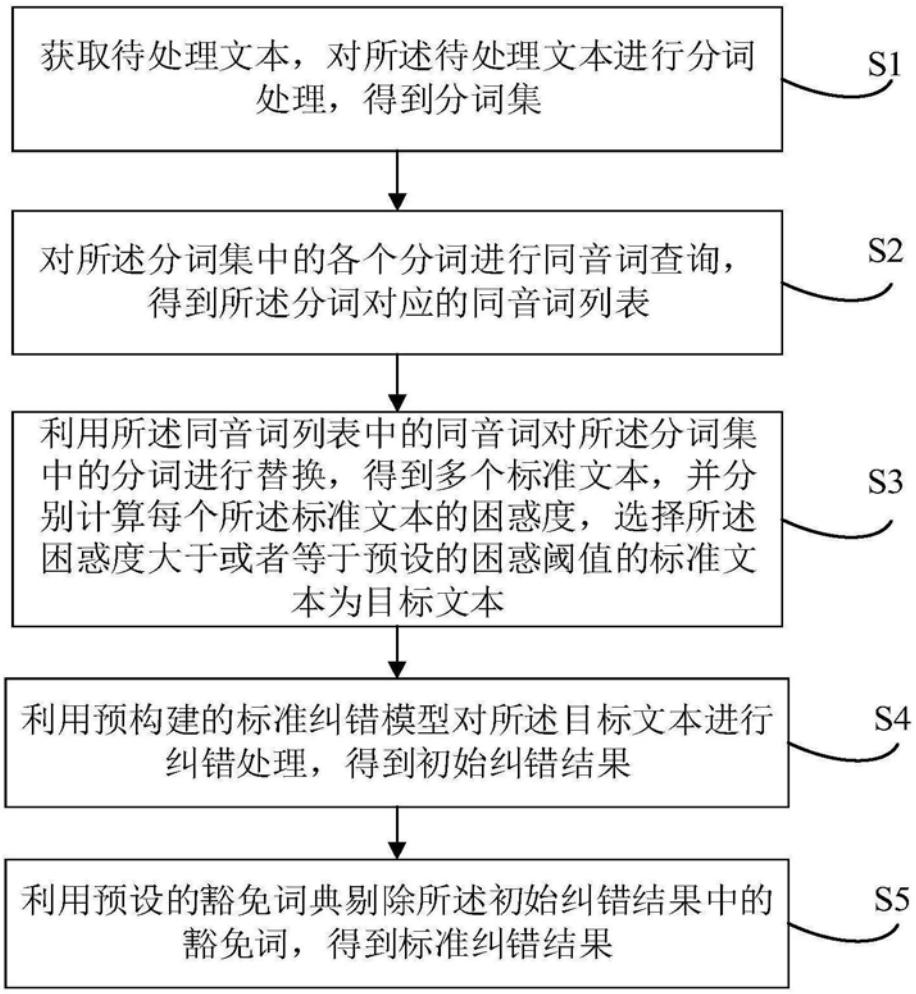

获取待处理文本，对所述待处理文本进行分词处理，得到分词集
S1
对所述分词集中的各个分词进行同音词查询，得到所述分词对应的同音词列表
S2
利用所述同音词列表中的同音词对所述分词集中的分词进行替换，得到多个标准文本，并分别计算每个所述标准文本的困惑度，选择所述困惑度大于或者等于预设的困惑阈值的标准文本为目标文本
S3
利用预构建的标准纠错模型对所述目标文本进行纠错处理，得到初始纠错结果
S4
利用预设的豁免词典剔除所述初始纠错结果中的豁免词，得到标准纠错结果
S5

图1

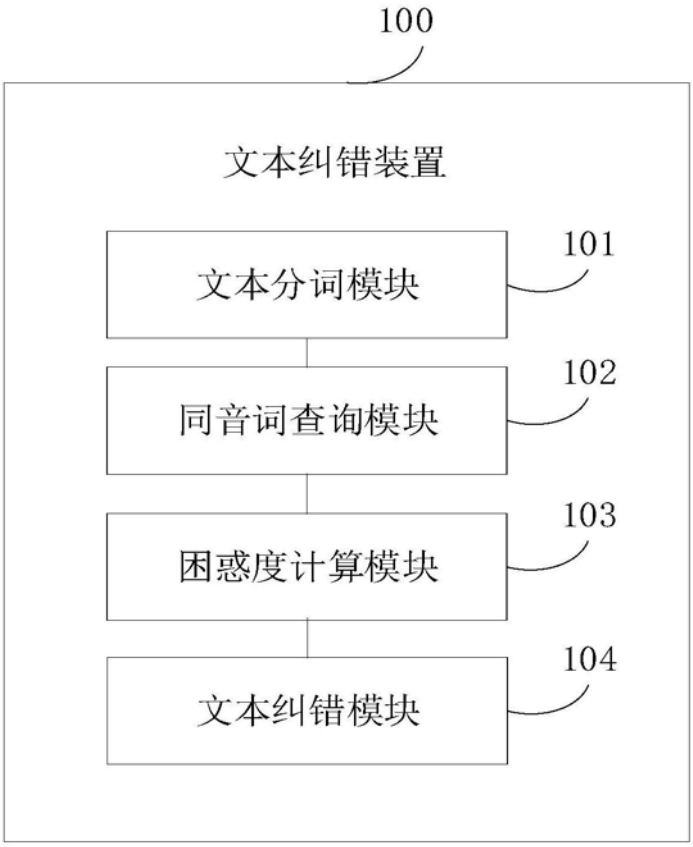

100
文本纠错装置
文本分词模块
101
同音词查询模块
102
困惑度计算模块
103
文本纠错模块
104

图2

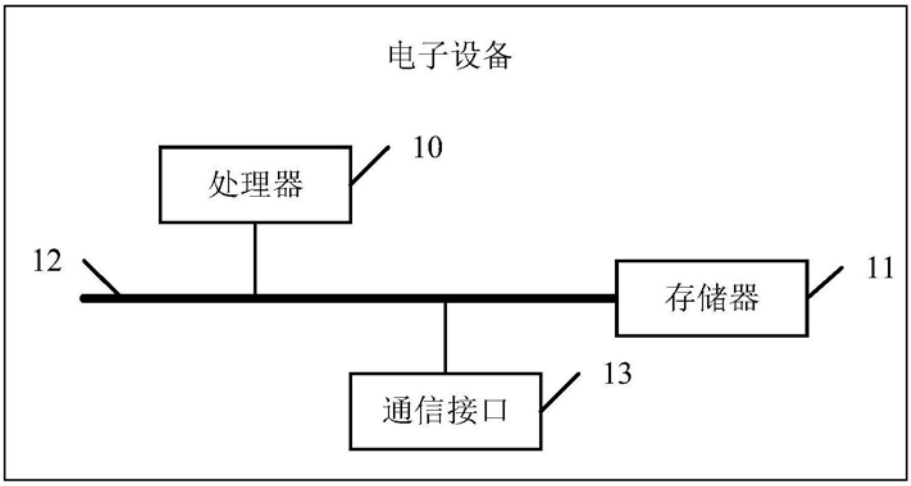

电子设备
处理器
10
12
存储器
11
通信接口
13

图3