(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113095072 B
(45)授权公告日 2024.06.28

(21)申请号 201911335070.1

(22)申请日 2019.12.23

(65)同一申请的已公布的文献号
申请公布号 CN 113095072 A

(43)申请公布日 2021.07.09

(73)专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 刘杰 祝官文

(74)专利代理机构 北京龙双利达知识产权代理有限公司 11329
专利代理师 张振 王君

(51)Int.Cl.
G06F 40/284(2020.01)
G06F 16/35(2019.01)
G06N 3/06(2006.01)
G06N 3/048(2023.01)
G06N 3/0464(2023.01)
G06N 3/084(2023.01)

(56)对比文件
CN 105975625 A,2016.09.28

审查员 薛双双

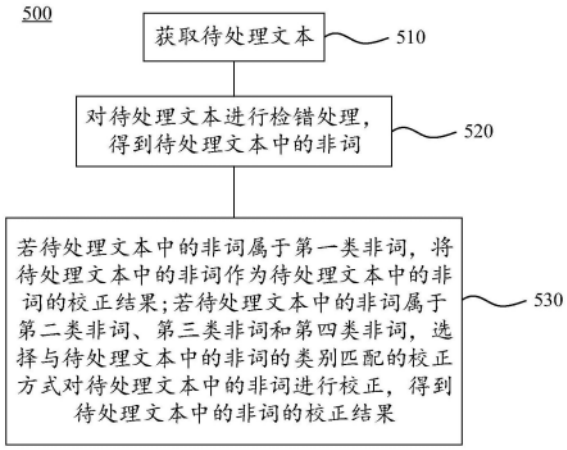
权利要求书3页 说明书27页 附图9页

(54)发明名称

文本处理方法及装置

(57)摘要

本申请公开了人工智能领域中自然语言处理领域的文本处理方法以及装置,该方法包括:获取待处理文本;对待处理文本进行检错处理,得到待处理文本中的非词;若待处理文本中的非词属于第一类非词,则不对待处理文本中的非词进行校正;若待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与待处理文本中的非词的类别匹配的校正方式对待处理文本中的非词进行校正,得到所述待处理文本中的非词的校正结果。本申请的技术方案能够对多种类型的字符串进行检测和过滤,提高了文本纠错的准确率。



CN 113095072 B

1.一种文本处理方法,其特征在于,包括:

获取待处理文本;

对所述待处理文本进行检错处理,得到所述待处理文本中的非词;

若所述待处理文本中的非词属于第一类非词,将所述待处理文本中的非词作为所述待处理文本中的非词的校正结果;

若所述待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与所述待处理文本中的非词的类别匹配的校正方式对所述待处理文本中的非词进行校正,得到所述待处理文本中的非词的校正结果;

若所述待处理文本中的非词属于所述第三类非词,将所述待处理文本中的非词中的非字母的字符修改为字母,将修改后的词作为所述待处理文本中的非词的校正结果;

若所述待处理文本中的非词属于所述第四类非词,基于编辑距离和/或公共字符串生成所述待处理文本中的非词对应的候选词;

根据所述待处理文本中的非词与所述待处理文本中的非词对应的候选词之间的相似性以及所述待处理文本中的非词对应的候选词的困惑度对所述待处理文本中的非词对应的候选词进行评分,其中,所述待处理文本中的非词与所述待处理文本中的非词对应的候选词之间的相似性包括所述待处理文本中的非词与所述待处理文本中的非词对应的候选词之间的编辑距离和/或公共字符串,所述待处理文本中的非词对应的候选词的困惑度用于指示所述待处理文本中的非词对应的候选词在所述待处理文本中出现的可能性;

将所述待处理文本中的非词对应的候选词中评分最高的候选词确定为所述待处理文本中的非词对应的目标候选词;

在所述待处理文本中的非词对应的目标候选词的困惑度低于或等于第一困惑度阈值的情况下,利用所述待处理文本中的非词对应的目标候选词替换所述待处理文本中的非词,作为所述待处理文本中的非词的校正结果;

其中,所述第一类非词包括全大写字母非词、字长在预设字长范围内的非词和属于第一预设词库的非词,所述第二类非词包括合并错误非词,所述第三类非词包括包含非字母的字符的非词,所述第四类非词包括除所述第一类非词、所述第二类非词、所述第三类非词之外的其他非词。

2.如权利要求1所述的方法,其特征在于,所述待处理文本中的非词与所述待处理文本中的非词对应的候选词之间的相似性满足第一预设条件。

3.如权利要求1所述的方法,其特征在于,所述合并错误非词为包括至少两个真词的非词,以及

所述若所述待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与所述待处理文本中的非词的类别匹配的校正方式对所述待处理文本中的非词进行校正,得到所述待处理文本中的非词的校正结果,包括:

若所述待处理文本中的非词属于所述第二类非词,对所述待处理文本中的非词进行校正,得到所述至少两个真词,作为所述待处理文本中的非词的校正结果。

4.如权利要求1所述的方法,其特征在于,所述方法还包括:

对所述待处理文本进行真词错误检测,得到待处理文本中的错误真词;

生成所述错误真词对应的候选词;

在所述错误真词对应的候选词中确定所述错误真词对应的目标候选词;

根据所述错误真词对应的目标候选词对所述错误真词进行校正,得到所述错误真词的校正结果。

5.如权利要求4所述的方法,其特征在于,所述在所述错误真词对应的候选词中确定所述错误真词对应的目标候选词,包括:

根据所述错误真词与所述错误真词对应的候选词之间的相似性以及所述错误真词对应的候选词的困惑度对所述错误真词对应的候选词进行评分,其中,所述错误真词对应的候选词的困惑度用于指示所述错误真词对应的候选词在所述待处理文本中出现的可能性;

将所述错误真词对应的候选词中评分最高的候选词确定为所述错误真词对应的目标候选词。

6.如权利要求4所述的方法,其特征在于,所述错误真词与所述错误真词对应的候选词之间的相似性满足第二预设条件。

7.如权利要求4所述的方法,其特征在于,所述根据所述错误真词对应的目标候选词对所述错误真词进行校正,得到所述错误真词的校正结果,包括:

在所述错误真词对应的目标候选词的困惑度低于或等于第二困惑度阈值的情况下,利用所述错误真词对应的目标候选词替换所述错误真词,作为所述错误真词的校正结果。

8.一种文本处理装置,其特征在于,包括:

获取单元,所述获取单元用于获取待处理文本;

处理单元,所述处理单元用于:

对所述待处理文本进行检错处理,得到所述待处理文本中的非词;

若所述待处理文本中的非词属于第一类非词,将所述待处理文本中的非词作为所述待处理文本中的非词的校正结果;

若所述待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与所述待处理文本中的非词的类别匹配的校正方式对所述待处理文本中的非词进行校正,得到所述待处理文本中的非词的校正结果;

若所述待处理文本中的非词属于所述第三类非词,将所述待处理文本中的非词中的非字母的字符修改为字母,将修改后的词作为所述待处理文本中的非词的校正结果;

若所述待处理文本中的非词属于所述第四类非词,基于编辑距离和/或公共字符串生成所述待处理文本中的非词对应的候选词;

根据所述待处理文本中的非词与所述待处理文本中的非词对应的候选词之间的相似性以及所述待处理文本中的非词对应的候选词的困惑度对所述待处理文本中的非词对应的候选词进行评分,其中,所述待处理文本中的非词与所述待处理文本中的非词对应的候选词之间的相似性包括所述待处理文本中的非词与所述待处理文本中的非词对应的候选词之间的编辑距离和/或公共字符串,所述待处理文本中的非词对应的候选词的困惑度用于指示所述待处理文本中的非词对应的候选词在所述待处理文本中出现的可能性;

将所述待处理文本中的非词对应的候选词中评分最高的候选词确定为所述待处理文本中的非词对应的目标候选词;

在所述待处理文本中的非词对应的目标候选词的困惑度低于或等于第一困惑度阈值的情况下,利用所述待处理文本中的非词对应的目标候选词替换所述待处理文本中的非

词,作为所述待处理文本中的非词的校正结果;

其中,所述第一类非词包括全大写字母非词、字长在预设字长范围内的非词和属于第一预设词库的非词,所述第二类非词包括合并错误非词,所述第三类非词包括包含非字母的字符的非词,所述第四类非词包括除所述第一类非词、所述第二类非词、所述第三类非词之外的其他非词。

9.如权利要求8所述的装置,其特征在于,所述待处理文本中的非词与所述待处理文本中的非词对应的候选词之间的相似性满足第一预设条件。

10.如权利要求8所述的装置,其特征在于,所述合并错误非词为包括至少两个真词的非词,以及所述处理单元用于:

若所述待处理文本中的非词属于所述第二类非词,对所述待处理文本中的非词进行校正,得到所述至少两个真词作为所述待处理文本中的非词的校正结果。

11.如权利要求8所述的装置,其特征在于,所述处理单元还用于:

对所述待处理文本进行真词错误检测,得到待处理文本中的错误真词;

生成所述错误真词对应的候选词;

在所述错误真词对应的候选词中确定所述错误真词对应的目标候选词;

根据所述错误真词对应的目标候选词对所述错误真词进行校正,得到所述错误真词的校正结果。

12.如权利要求11所述的装置,其特征在于,所述在所述错误真词对应的候选词中确定所述错误真词对应的目标候选词,包括:

根据所述错误真词与所述错误真词对应的候选词之间的相似性以及所述错误真词对应的候选词的困惑度对所述错误真词对应的候选词进行评分,其中,所述错误真词对应的候选词的困惑度用于指示所述错误真词对应的候选词在所述待处理文本中出现的可能性;

将所述错误真词对应的候选词中评分最高的候选词确定为所述错误真词对应的目标候选词。

13.如权利要求11所述的装置,其特征在于,所述错误真词与所述错误真词对应的候选词之间的相似性满足第二预设条件。

14.如权利要求11所述的装置,其特征在于,所述根据所述错误真词对应的目标候选词对所述错误真词进行校正,得到所述错误真词的校正结果,包括:

在所述错误真词对应的目标候选词的困惑度低于或等于第二困惑度阈值的情况下,利用所述错误真词对应的目标候选词替换所述错误真词,作为所述错误真词的校正结果。

15.一种文本处理装置,其特征在于,包括:

存储器,用于存储程序;

处理器,用于执行所述存储器存储的程序,当所述处理器执行所述存储器存储的程序时,所述处理器用于执行权利要求1至7中任一项所述的文本处理方法。

16.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质中存储有程序指令,当所述程序指令由处理器运行时,实现权利要求1至7中任一项所述的文本处理方法。

文本处理方法及装置

技术领域

[0001] 本申请涉及自然语言处理领域,并且更具体地,涉及一种文本处理方法及装置。

背景技术

[0002] 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。换句话说,人工智能是计算机科学的一个分支,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式作出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0003] 随着人工智能技术的不断发展,让人机之间能够通过自然语言进行交互的自然语言人机交互系统变的越来越重要。人机之间能够通过自然语言进行交互,就需要系统能够识别出人类自然语言的具体含义。通常,系统通过采用对自然语言的句子进行关键信息提取来识别句子的具体含义。

[0004] 文本纠错(text error correction)是对原始文本进行错误检测(error detection),并根据自然语言处理技术对错误进行纠正。目前,通常有两种方法可以实现文本纠错。一种是基于字典判断输入的查询词是否正确,实现错词的检测,然后对错词进行纠正。另一种是通过语言模型对上下文语义信息进行提取,对错词进行纠正。在原始文本中可能包括多种类型的字符,上述方法均只能对单一类型的字符进行检查和纠正,导致纠错准确率降低。

发明内容

[0005] 本申请提供一种文本处理方法及装置,能够对多种类型的字符进行检测和过滤,提高了文本纠错的准确率。

[0006] 第一方面,提供了一种文本处理方法,包括:获取待处理文本;对待处理文本进行检错处理,得到待处理文本中的非词;若待处理文本中的非词属于第一类非词,将待处理文本中的非词作为待处理文本中的非词的校正结果;若待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与待处理文本中的非词的类别匹配的校正方式对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果。其中,第一类非词包括全大写字母非词、字长在预设字长范围内的非词和属于第一预设词库的非词,第二类非词包括合并错误非词,第三类非词包括包含非字母的字符的非词,第四类非词包括除第一类非词、第二类非词、第三类非词之外的其他非词。

[0007] 待处理文本可以是光学字符识别(optical character recognition,OCR)输出文本,或者,可以是用户输入的文本。例如,用户输入的文本可以包括社交网络中发布的内容,或者可以是搜索引擎的搜索框中输入的内容等。应理解,待处理文本可以是任意需要进行纠错的文本,本申请对待处理文本的具体形式不做限定。

[0008] 可以基于第二预设词库对待处理文本进行非词错误检测,得到待处理文本中的非

词。非词指的是不存在于第二预设词库中的词。第一预设词库与第二预设词库不同。

[0009] 应理解,能够用于检测非词错误的词库均可以理解为第二预设词库。例如,在方法500应用于英文文本纠错的情况下,第二预设词库可以为英文词库。非词即为不存在于英文词库的词,例如,werld。

[0010] 在本申请的实施例中,能够对待处理文本中的多种类型的字符分别进行检测以及分别进行处理,降低了多种类型的字符对纠错过程的干扰,提升了文本纠错的准确率,提高了纠错方法对输入文本的鲁棒性。

[0011] 结合第一方面,在一种可能的实现方式中,若待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与待处理文本中的非词的类别匹配的校正方式对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果,包括:若待处理文本中的非词属于第四类非词,生成待处理文本中的非词对应的候选词;在待处理文本中的非词对应的候选词中确定待处理文本中的非词对应的目标候选词;根据待处理文本中的非词对应的目标候选词对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果。

[0012] 其中,待处理文本中的非词对应的候选词可以为一个,也可以为多个。

[0013] 结合第一方面,在一种可能的实现方式中,在待处理文本中的非词对应的多个候选词中确定待处理文本中的非词对应的目标候选词,包括:根据待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性以及待处理文本中的非词对应的候选词的困惑度对待处理文本中的非词对应的候选词进行评分,其中,待处理文本中的非词对应的候选词的困惑度用于指示待处理文本中的非词对应的候选词在待处理文本中出现的可能性;将待处理文本中的非词对应的候选词中评分最高的候选词确定为待处理文本中的非词对应的目标候选词。

[0014] 其中,非词对应的候选词的困惑度可以通过语言模型进行评分。

[0015] 每个候选词对应的评分可以由上述几项对应的评分进行加权得到,也就是为每一项对应的评分设置权重。该权重可以是预先设定的,也可以是训练得到的。

[0016] 在本申请实施例中,采用候选词与非词之间的相似性以及候选词的困惑度对候选词进行评分,同时考虑了非词与候选词的相似性以及待处理文本的语义信息,能够更符合输入文本的原意,能够得到较优的候选词,提高了文本纠错的准确性。

[0017] 结合第一方面,在一种可能的实现方式中,待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性满足第一预设条件。

[0018] 例如,待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性可以包括待处理文本中的非词与待处理文本中的非词对应的候选词之间的编辑距离和/或公共字符串。也就是可以基于编辑距离和/或公共字符串确定待处理文本中的非词对应的候选词。

[0019] 编辑距离是指将一个词转换为另一个词所需的编辑操作次数。编辑操作包括对词中的字符的插入、删除、易位和替换等操作。

[0020] 公共字符串是指两个词中所包含的连续相同的字符的数量。

[0021] 示例性地,第一预设条件可以为编辑距离小于第一预设值。

[0022] 结合第一方面,在一种可能的实现方式中,根据待处理文本中的非词对应的目标候选词对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果,包括:在待

处理文本中的非词对应的目标候选词的困惑度低于或等于第一困惑度阈值的情况下,利用待处理文本中的非词对应的目标候选词替换待处理文本中的非词,作为待处理文本中的非词的校正结果。

[0023] 在本申请实施例中,通过对目标候选词的困惑度的判断,利用困惑度低于或等于第一困惑度阈值的目标候选词替换待处理文本中的非词,能够充分利用文本的语义信息,进一步提升文本纠错的准确率。

[0024] 结合第一方面,在一种可能的实现方式中,合并错误非词为包括至少两个真词的非词,以及若待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与待处理文本中的非词的类别匹配的校正方式对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果,包括:若待处理文本中的非词属于第二类非词,对待处理文本中的非词进行校正,得到至少两个真词作为待处理文本中的非词的校正结果。

[0025] 具体地,可以在该非词中适当的位置添加空格,将该非词修改为至少两个真词。

[0026] 例如,该处理文本中的非词为inChina,该非词属于第二类非词,在该非词中添加空格,将该非词修改为in China。

[0027] 结合第一方面,在一种可能的实现方式中,将修改后的至少两个真词输入语言模型中,若该修改后的至少两个真词能够使得困惑度降低,则将修改后的至少两个真词作为待处理文本中的非词的校正结果。

[0028] 在本申请实施例中,通过语言模型对修改后的至少两个真词的困惑度进行进一步判断,利用了待处理文本的语义信息,能够更符合输入文本的原意,能够得到较优的候选词,提高了文本纠错的准确性。

[0029] 结合第一方面,在一种可能的实现方式中,若待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与待处理文本中的非词的类别匹配的校正方式对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果,包括:若待处理文本中的非词属于第三类非词,将待处理文本中的非词中的非字母的字符修改为字母,将修改后的词作为待处理文本中的非词的校正结果。

[0030] 结合第一方面,在一种可能的实现方式中,将修改后的词输入语言模型中,若该修改后的词能够使得困惑度降低,则将修改后的词作为待处理文本中的非词的校正结果。

[0031] 在本申请实施例中,通过语言模型对修改后的词的困惑度进行进一步判断,利用了待处理文本的语义信息,能够更符合输入文本的原意,能够得到较优的候选词,提高了文本纠错的准确性。

[0032] 结合第一方面,在一种可能的实现方式中,方法还包括:对待处理文本进行真词错误检测,得到待处理文本中的错误真词;生成错误真词对应的候选词;在错误真词对应的候选词中确定错误真词对应的目标候选词;根据错误真词对应的目标候选词对错误真词进行校正。

[0033] 具体地,可以基于语言模型对待处理文本进行真词错误检测,得到待处理文本中的错误真词。

[0034] 例如,该语言模型可以为统计语言模型。再如,该语言模型也可以为神经网络模型。

[0035] 错误真词对应的候选词可以为一个,也可以为多个。

[0036] 结合第一方面,在一种可能的实现方式中,在错误真词对应的候选词中确定错误真词对应的目标候选词,包括:根据错误真词与错误真词对应的候选词之间的相似性以及错误真词对应的候选词的困惑度对错误真词对应的候选词进行评分,其中,错误真词对应的候选词的困惑度用于指示错误真词对应的候选词在待处理文本中出现的可能性;将错误真词对应的候选词中评分最高的候选词确定为错误真词对应的目标候选词。

[0037] 其中,错误真词对应的候选词的困惑度可以通过语言模型进行评分。

[0038] 每个候选词对应的评分可以由上述几项对应的评分进行加权得到,也就是为每一项对应的评分设置权重。该权重可以是预先设定的,也可以是训练得到的。

[0039] 在本申请实施例中,采用候选词与错误真词之间的相似性以及候选词的困惑度对候选词进行评分,同时考虑了错误真词与候选词的相似性以及待处理文本的语义信息,能够更符合输入文本的原意,能够得到较优的候选词,提高了文本纠错的准确性。

[0040] 结合第一方面,在一种可能的实现方式中,待处理文本中的错误真词与待处理文本中的错误真词对应的候选词之间的相似性满足第二预设条件。

[0041] 例如,待处理文本中的错误真词与待处理文本中的错误真词对应的候选词之间的相似性可以包括待处理文本中的错误真词与待处理文本中的错误真词对应的候选词之间的编辑距离和/或公共字符串。也就是可以基于编辑距离和/或公共字符串确定待处理文本中的错误真词对应的候选词。

[0042] 示例性地,第二预设条件可以为编辑距离小于第二预设值。

[0043] 应理解,第一预设条件和第二预设条件可以相同,也可以不同。第一预设值与第二预设值可以相同,也可以不同。

[0044] 结合第一方面,在一种可能的实现方式中,根据错误非词对应的目标候选词对错误真词进行校正,得到错误真词的校正结果,包括:在错误真词对应的目标候选词的困惑度低于或等于第二困惑度阈值的情况下,利用错误真词对应的目标候选词替换错误真词,作为错误真词的校正结果。

[0045] 在本申请实施例中,通过对目标候选词的困惑度的判断,利用困惑度低于或等于第二困惑度阈值的目标候选词替换待处理文本中的非词,能够充分利用文本的语义信息,进一步提升文本纠错的准确率。

[0046] 应理解,第一困惑度阈值和第二困惑度阈值可以相同,也可以不同。

[0047] 第二方面,提供了一种文本处理装置,包括:获取单元和处理单元。获取单元用于获取待处理文本。处理单元用于:对待处理文本进行检错处理,得到待处理文本中的非词;若待处理文本中的非词属于第一类非词,将待处理文本中的非词作为待处理文本中的非词的校正结果;若待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与待处理文本中的非词的类别匹配的校正方式对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果。其中,第一类非词包括全大写字母非词、字长在预设字长范围内的非词和属于第一预设词库的非词,第二类非词包括合并错误非词,第三类非词包括包含非字母的字符的非词,第四类非词包括除第一类非词、第二类非词、第三类非词之外的其他非词。

[0048] 结合第二方面,在一种可能的实现方式中,处理单元用于:若待处理文本中的非词属于第四类非词,生成待处理文本中的非词对应的候选词;在待处理文本中的非词对应的

候选词中确定待处理文本中的非词对应的目标候选词;根据待处理文本中的非词对应的目标候选词对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果。

[0049] 结合第二方面,在一种可能的实现方式中,处理单元用于:根据待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性以及待处理文本中的非词对应的候选词的困惑度对待处理文本中的非词对应的候选词进行评分,其中,待处理文本中的非词对应的候选词的困惑度用于指示待处理文本中的非词对应的候选词在待处理文本中出现的可能性;将待处理文本中的非词对应的候选词中评分最高的候选词确定为待处理文本中的非词对应的目标候选词。

[0050] 结合第二方面,在一种可能的实现方式中,待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性满足第一预设条件。

[0051] 结合第二方面,在一种可能的实现方式中,处理单元用于:在待处理文本中的非词对应的目标候选词的困惑度低于或等于第一困惑度阈值的情况下,利用待处理文本中的非词对应的目标候选词替换待处理文本中的非词,作为待处理文本中的非词的校正结果。

[0052] 结合第二方面,在一种可能的实现方式中,合并错误非词为包括至少两个真词的非词,以及处理单元用于:若待处理文本中的非词属于第二类非词,对待处理文本中的非词进行校正,得到至少两个真词作为待处理文本中的非词的校正结果。

[0053] 结合第二方面,在一种可能的实现方式中,处理单元用于:若待处理文本中的非词属于第三类非词,将待处理文本中的非词中的非字母的字符修改为字母,将修改后的词作为待处理文本中的非词的校正结果。

[0054] 结合第二方面,在一种可能的实现方式中,处理单元还用于:对待处理文本进行真词错误检测,得到待处理文本中的错误真词;生成错误真词对应的候选词;在错误真词对应的候选词中确定错误真词对应的目标候选词;根据错误真词对应的目标候选词对错误真词进行校正。

[0055] 结合第二方面,在一种可能的实现方式中,处理单元用于:根据错误真词与错误真词对应的候选词之间的相似性以及错误真词对应的候选词的困惑度对错误真词对应的候选词进行评分,其中,错误真词对应的候选词的困惑度用于指示错误真词对应的候选词在待处理文本中出现的可能性;将错误真词对应的候选词中评分最高的候选词确定为错误真词对应的目标候选词。

[0056] 结合第二方面,在一种可能的实现方式中,错误真词与错误真词对应的候选词之间的相似性满足第二预设条件。

[0057] 结合第二方面,在一种可能的实现方式中,处理单元用于:在错误真词对应的目标候选词的困惑度低于或等于第二困惑度阈值的情况下,利用错误真词对应的目标候选词替换错误真词,作为错误真词的校正结果。

[0058] 应理解,在上述第一方面中对相关内容的扩展、限定、解释和说明也适用于第二方面中相同的内容。

[0059] 第三方面,提供了一种文本处理装置,该装置包括:存储器,用于存储程序;处理器,用于执行所述存储器存储的程序,当所述存储器存储的程序被所述处理器执行时所述处理器用于执行上述第一方面或第一方面中的任意一种实现方式中的文本处理方法。

[0060] 第四方面,提供一种计算机可读介质,该计算机可读介质存储用于设备执行的程

序代码,该程序代码包括用于执行第一方面或者第一方面中的任意一种实现方式中的文本处理方法。

[0061] 第五方面,提供了一种计算机程序产品,所述计算机程序产品包括:计算机程序代码,当所述计算机程序代码在计算机上运行时,使得计算机执行上述各方面中的方法。

[0062] 需要说明的是,上述计算机程序代码可以全部或者部分存储在第一存储介质上,其中第一存储介质可以与处理器封装在一起的,也可以与处理器单独封装,本申请实施例对此不作具体限定。

[0063] 第六方面,提供一种芯片,所述芯片包括处理器与数据接口,所述处理器通过所述数据接口读取存储器上存储的指令,执行上述第一方面或第一方面中的任意一种实现方式中的文本处理方法。

[0064] 可选地,作为一种实现方式,所述芯片还可以包括存储器,所述存储器中存储有指令,所述处理器用于执行所述存储器上存储的指令,当所述指令被执行时,所述处理器用于执行第一方面或者第一方面中的任意一种实现方式中的文本处理方法。

附图说明

[0065] 图1是本申请实施例提供的一种自然语言处理的应用场景示意图;

[0066] 图2是本申请实施例提供的另一种自然语言处理的应用场景示意图;

[0067] 图3是本申请实施例提供的自然语言处理的相关设备的示意图;

[0068] 图4是本申请实施例提供的一种系统架构的示意图;

[0069] 图5是本申请实施例提供的一种根据CNN模型进行文本处理的示意图;

[0070] 图6本申请实施例提供的另一种根据CNN模型进行文本处理的示意图;

[0071] 图7是本申请实施例提供的一种芯片的硬件结构的示意图;

[0072] 图8是本申请实施例提供的一种应用场景示意图;

[0073] 图9是本申请实施例提供的文本处理方法的示意性流程图;

[0074] 图10是本申请实施例提供的另一种文本处理方法的示意性流程图;

[0075] 图11是本申请实施例提供的又一种文本处理方法的示意性流程图;

[0076] 图12是本申请实施例提供的一种文本处理装置的示意性框图;

[0077] 图13是本申请实施例提供的另一种文本处理装置的示意性框图。

具体实施方式

[0078] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行描述,显然,所描述的实施例仅仅是本申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本申请保护的范围。

[0079] 为了更好地理解本申请实施例的方案,下面先结合图1至图3对本申请实施例可能的应用场景进行简单的介绍。

[0080] 图1示出了一种自然语言处理系统,该自然语言处理系统包括用户设备以及数据处理设备。其中,用户设备包括手机、个人电脑或者信息处理中心等智能终端。用户设备为自然语言数据处理的发起端,作为语言问答或者查询等请求的发起方,通常用户通过用户

设备发起请求。

[0081] 上述数据处理设备可以是云服务器、网络服务器、应用服务器以及管理服务器等具有数据处理功能的设备或服务器。数据处理设备通过交互接口接收来自智能终端的查询语句/语音/文本等问句,再通过存储数据的存储器以及数据处理的处理器环节进行机器学习,深度学习,搜索,推理,决策等方式的语言数据处理。数据处理设备中的存储器可以是一个统称,包括本地存储以及存储历史数据的数据库,数据库可以再数据处理设备上,也可以在其它网络服务器上。

[0082] 在图1所示的自然语言处理系统中,用户设备可以接收用户的指令,例如,用户设备可以接收用户输入的一段文本,然后向数据处理设备发起请求,使得数据处理设备针对用户设备得到的该一段文本执行自然语言处理应用(例如,文本分类、文本序列标注、翻译等),从而得到针对该一段文本的对应的自然语言处理应用的处理结果(例如,文本分类、文本序列标注、翻译等)。

[0083] 示例性地,用户设备可以接收用户输入的待处理文本,然后向数据处理设备发起请求,使得数据处理设备对该待处理文本进行分类,从而得到针对该待处理文本的分类结果。其中,分类结果可以是指该待处理文本所指示的用户语义意图,比如,用户用于指示放歌、设置时间、开启导航的意图;或者,分类结果还可以用于指示用户的情感分类结果,比如,分类结果可以指示待处理文本对应的用户情感分类为抑郁、开心或者生气等。

[0084] 例如,在图1中数据处理设备可以执行本申请实施例的文本处理方法。

[0085] 图2示出了另一种自然语言处理系统,在图2中,用户设备直接作为数据处理设备,该用户设备能够直接接收来自用户的输入并直接由用户设备本身的硬件进行处理,具体过程与图1相似,可参考上面的描述,在此不再赘述。

[0086] 在图2所示的自然语言处理系统中,用户设备可以接收用户的指令,由用户设备自身对待处理文本进行分类得到待处理文本的分类结果。

[0087] 在图2所示的自然语言处理系统中,用户设备可以接收用户的指令,例如用户设备可以接收用户输入的一段文本,然后再由用户设备自身针对该一段文本执行自然语言处理应用(例如,文本分类、文本序列标注、翻译等),从而得到针对该一段文本的对应的自然语言处理应用的处理结果(例如,文本分类、文本序列标注、翻译等)。

[0088] 在图2中,用户设备自身就可以执行本申请实施例的文本处理方法。

[0089] 图3是本申请实施例提供的自然语言处理的相关设备的示意图。

[0090] 上述图1和图2中的用户设备具体可以是图3中的本地设备130或者本地设备120,图1中的数据处理设备具体可以是图3中的执行设备110,其中,数据存储系统150可以存储执行设备110的待处理数据,数据存储系统150可以集成在执行设备110上,也可以设置在云上或其它网络服务器上。

[0091] 图1和图2中的处理器可以通过神经网络模型或者其它模型进行数据训练/机器学习/深度学习,并利用数据最终训练或者学习得到的模型对输入的待处理文本进行处理,从而得到待处理文本处理结果。

[0092] 由于本申请实施例涉及大量神经网络的应用,为了便于理解,下面先对本申请实施例可能涉及的神经网络的相关术语和概念进行介绍。

[0093] (1)神经网络

[0094] 神经网络可以是由神经单元组成的,神经单元可以是指以x_s和截距1为输入的运算单元,该运算单元的输出可以为:

[0095] $$h_{W,b}(x)=f\left(W^{T}x\right)=f(\sum_{s=1}^{n}W_{s}x_{s}+b);$$

[0096] 其中,s=1、2、……n,n为大于1的自然数,W_s为x_s的权重,b为神经单元的偏置。f为神经单元的激活函数(activation functions),用于将非线性特性引入神经网络中,来将神经单元中的输入信号转换为输出信号。该激活函数的输出信号可以作为下一层卷积层的输入,激活函数可以是sigmoid函数。神经网络是将多个上述单一的神经单元联结在一起形成的网络,即一个神经单元的输出可以是另一个神经单元的输入。每个神经单元的输入可以与前一层的局部接受域相连,来提取局部接受域的特征,局部接受域可以是由若干个神经单元组成的区域。

[0097] (2)深度神经网络

[0098] 深度神经网络(deep neural network,DNN),也称多层神经网络,可以理解为具有多层隐含层的神经网络。按照不同层的位置对DNN进行划分,DNN内部的神经网络可以分为三类:输入层,隐含层,输出层。一般来说第一层是输入层,最后一层是输出层,中间的层数都是隐含层。层与层之间是全连接的,也就是说,第i层的任意一个神经元一定与第i+1层的任意一个神经元相连。

[0099] 虽然DNN看起来很复杂,但是就每一层的工作来说,其实并不复杂,简单来说就是如下线性关系表达式:$\vec{y}=\alpha(W\bullet\vec{x}+\vec{b})$,其中,$\vec{x}$是输入向量,$\vec{y}$是输出向量,$\vec{b}$是偏移向量,W是权重矩阵(也称系数),α()是激活函数。每一层仅仅是对输入向量$\vec{x}$经过如此简单的操作得到输出向量$\vec{y}$。由于DNN层数多,系数W和偏移向量$\vec{b}$的数量也比较多。这些参数在DNN中的定义如下所述:以系数W为例:假设在一个三层的DNN中,第二层的第4个神经元到第三层的第2个神经元的线性系数定义为W_{24}^{3}。上标3代表系数W所在的层数,而下标对应的是输出的第三层索引2和输入的第二层索引4。

[0100] 综上,第L-1层的第k个神经元到第L层的第j个神经元的系数定义为W_{jk}^{L}。

[0101] 需要注意的是,输入层是没有W参数的。在深度神经网络中,更多的隐含层让网络更能够刻画现实世界中的复杂情形。理论上而言,参数越多的模型复杂度越高,“容量”也就越大,也就意味着它能完成更复杂的学习任务。训练深度神经网络的也就是学习权重矩阵的过程,其最终目的是得到训练好的深度神经网络的所有层的权重矩阵(由很多层的向量W形成的权重矩阵)。

[0102] (3)卷积神经网络

[0103] 卷积神经网络(convolutional neuron network,CNN)是一种带有卷积结构的深度神经网络。卷积神经网络包含了一个由卷积层和子采样层构成的特征抽取器,该特征抽取器可以看作是滤波器。卷积层是指卷积神经网络中对输入信号进行卷积处理的神经元层。在卷积神经网络的卷积层中,一个神经元可以只与部分邻层神经元连接。一个卷积层中,通常包含若干个特征平面,每个特征平面可以由一些矩形排列的神经单元组成。同一特征平面的神经单元共享权重,这里共享的权重就是卷积核。共享权重可以理解为提取图像信息的方式与位置无关。卷积核可以以随机大小的矩阵的形式初始化,在卷积神经网络的

训练过程中卷积核可以通过学习得到合理的权重。另外,共享权重带来的直接好处是减少卷积神经网络各层之间的连接,同时又降低了过拟合的风险

[0104] (4)损失函数

[0105] 在训练深度神经网络的过程中,因为希望深度神经网络的输出尽可能的接近真正想要预测的值,所以可以通过比较当前网络的预测值和真正想要的目标值,再根据两者之间的差异情况来更新每一层神经网络的权重向量(当然,在第一次更新之前通常会有初始化的过程,即为深度神经网络中的各层预先配置参数),比如,如果网络的预测值高了,就调整权重向量让它预测低一些,不断地调整,直到深度神经网络能够预测出真正想要的目标值或与真正想要的目标值非常接近的值。因此,就需要预先定义“如何比较预测值和目标值之间的差异”,这便是损失函数(loss function)或目标函数(objective function),它们是用于衡量预测值和目标值的差异的重要方程。其中,以损失函数举例,损失函数的输出值(loss)越高表示差异越大,那么深度神经网络的训练就变成了尽可能缩小这个loss的过程。

[0106] (5)反向传播算法

[0107] 神经网络可以采用误差反向传播(back propagation,BP)算法在训练过程中修正初始的神经网络模型中参数的大小,使得神经网络模型的重建误差损失越来越小。具体地,前向传递输入信号直至输出会产生误差损失,通过反向传播误差损失信息来更新初始的神经网络模型中参数,从而使误差损失收敛。反向传播算法是以误差损失为主导的反向传播运动,旨在得到最优的神经网络模型的参数,例如,权重矩阵。

[0108] (6)自然语言处理(natural language processing,NLP)

[0109] 自然语言(natural language)即人类语言,自然语言处理(NLP)就是对人类语言的处理。自然语言处理是以一种智能与高效的方式,对文本数据进行系统化分析、理解与信息提取的过程。通过使用NLP及其组件,我们可以管理非常大块的文本数据,或者执行大量的自动化任务,并且解决各式各样的问题,如自动摘要(automatic summarization),机器翻译(machine translation,MT),命名实体识别(named entity recognition,NER),关系提取(relation extraction,RE),信息抽取(information extraction,IE),情感分析,语音识别(speech recognition),问答系统(question answering)以及主题分割等等。

[0110] (7)语言模型(language model,LM)

[0111] 语言模型是NPL中的基础模型,通过大量语料训练学习,使得LM能够根据已有的信息(例如上下文中已经出现过的词等文本信息)来推测未知词的概率,也可以理解为LM是用来计算一个句子的概率模型。

[0112] 换句话说,语言模型是自然语言文本序列的概率分布,表征特定长度特定序列文本存在的可能性。简而言之,语言模型即是根据上下文去预测下一个词是什么,由于不需要人工标注语料,因此语言模型能够从无限制的大规模语料中学习到丰富的语义知识。

[0113] 首先,介绍本申请实施例提供的文本处理模型的训练方法和文本处理方法的系统架构。参考图4,本申请实施例提供了一种系统架构200。如图4中的系统架构200所示,数据采集设备260用于采集训练数据。

[0114] 例如,本申请实施例中训练数据可以是训练文本处理模型的训练文本。

[0115] 在采集到训练数据之后,数据采集设备260将这些训练数据存入数据库230,训练

设备220基于数据库230中维护的训练数据训练得到目标模型/规则201(即本申请的文本处理模型)。

[0116] 另外,该目标模型/规则201能够用于实现本申请实施例提供的文本处理方法,即,将待处理文本通过相关预处理(可以采用预处理模块213和/或预处理模块214进行处理)后输入该目标模型/规则201中进行处理,即可得到与目标处理模型所执行的目标任务对应的处理结果。

[0117] 示例性地,目标任务是文本纠错,则目标处理模型可以文本纠错模型,待处理文本输入目标模型/规则201(即本申请的文本处理模型)中进行文本纠错处理,即可得到对待处理文本的纠错文本。

[0118] 示例性地,目标任务是文本翻译,则目标学生模型可以文本翻译模型,待处理文本输入目标模型/规则201(即本申请的文本处理模型)中进行翻译处理,即可得到对待处理文本的翻译文本。

[0119] 在本申请提供的实施例中,该目标模型/规则201是通过训练原始处理模型得到的。需要说明的是,在实际的应用中,所述数据库230中维护的训练数据不一定都来自于数据采集设备260的采集,也有可能是从其他设备接收得到的。

[0120] 另外需要说明的是,训练设备220也不一定完全基于数据库230维护的训练数据进行目标模型/规则201的训练,也有可能从云端或其他地方获取训练数据进行模型训练,上述描述不应该作为对本申请实施例的限定。还需要说明的是,数据库230中维护的训练数据中的至少部分数据也可以用于执行设210对待处理文本进行处理的过程。

[0121] 根据训练设备220训练得到的目标模型/规则201可以应用于不同的系统或设备中,如应用于图4所示的执行设备210,所述执行设备210可以是终端,如手机终端,平板电脑,笔记本电脑,增强现实(augmented reality,AR)/虚拟现实(virtual reality,VR),车载终端等,还可以是服务器或者云端等。

[0122] 在图4中,执行设备210配置输入/输出(input/output,I/O)接口212,用于与外部设备进行数据交互,用户可以通过客户设备240向I/O接口212输入数据,所述输入数据在本申请实施例中可以包括:待处理文本。

[0123] 预处理模块213和/或预处理模块214用于根据I/O接口212接收到的输入数据进行预处理,在本申请实施例中,也可以没有预处理模块213和预处理模块214(也可以只有其中的一个预处理模块),而直接采用计算模块211对输入数据进行处理。需要说明的是,预处理模块213或预处理模块214可以对全部的输入数据进行预处理,也可以对输入数据的部分数据进行预处理。

[0124] 需要说明的是,预处理模块113和/或预处理模块214也可以是在训练设备220中训练好的。计算模块211可以用于根据上述目标模型/规则201对来自预处理模块213或者I/O接口212的输入数据执行计算等相关的处理。

[0125] 在执行设备210对输入数据进行预处理,或者在执行设备210的计算模块211执行计算等相关的处理过程中,执行设备210可以调用数据存储系统250中的数据、代码等以用于相应的处理,也可以将相应处理得到的数据、指令等存入数据存储系统250中。

[0126] 最后,I/O接口212将处理结果(例如纠错结果、翻译结果等)反馈给客户设备240。应理解,对应于不同的自然语言处理任务,目标模型/规则201是不同的,其处理结果相应地

也是不同的。

[0127] 值得说明的是,训练设备220可以针对不同的下游系统,生成该下游系统对应的目标模型/规则201,该相应的目标模型/规则201即可以用于实现上述目标或完成上述任务,从而为用户提供所需的结果。需要说明的是,训练设备220还可以针对不同的下游系统对应的目标模型/规则201生成对应的预处理模型,例如预处理模块213和/或预处理模块214中对应的预处理模型等。

[0128] 在图4中所示情况下,用户可以手动给定输入数据(例如,待处理文本),该手动给定可以通过I/O接口212提供的界面进行操作。另一种情况下,客户设备240可以自动地向I/O接口212发送输入数据(例如,待处理文本),如果要求客户设备240自动发送输入数据需要获得用户的授权,则用户可以在客户设备240中设置相应权限。用户可以在客户设备240查看执行设备210输出的结果,具体的呈现形式可以是显示、声音、动作等具体方式。客户设备240也可以作为数据采集端,采集如图所示输入I/O接口212的输入数据及输出I/O接口212的输出结果作为新的样本数据,并存入数据库230。当然,也可以不经过客户设备240进行采集,而是由I/O接口212直接将如图所示输入I/O接口212的输入数据及输出I/O接口212的输出结果,作为新的样本数据存入数据库230。

[0129] 值得注意的是,图4仅是本申请实施例提供的一种系统架构的示意图,图中所示设备、器件、模块等之间的位置关系不构成任何限制。例如,在图4中,数据存储系统250相对执行设备210是外部存储器,在其它情况下,也可以将数据存储系统250置于执行设备210中。

[0130] 如图4所示,根据训练设备220训练得到目标模型/规则201,该目标模型/规则201可以是本申请实施例中的目标处理模型,具体的,本申请实施例提供的目标处理模型可以是神经网络模型。例如可以是CNN,深度卷积神经网络(deep convolutional neural network,DCNN)。

[0131] 由于CNN是一种非常常见的神经网络,下面结合图5重点对CNN的结构进行详细的介绍。如上文的基础概念介绍所述,卷积神经网络是一种带有卷积结构的深度神经网络,是一种深度学习(deep learning)架构,深度学习架构是指通过机器学习的算法,在不同的抽象层级上进行多个层次的学习。作为一种深度学习架构,CNN是一种前馈(feed-forward)人工神经网络,该前馈人工神经网络中的各个神经元可以对输入其中的图像作出响应。

[0132] 如图5所示,卷积神经网络(CNN)300可以包括输入层310,卷积层/池化层320(其中池化层为可选的),以及神经网络层330。下面对这些层的相关内容做详细介绍。

[0133] 卷积层/池化层320:

[0134] 卷积层:

[0135] 如图5所示卷积层/池化层320可以包括如示例321-326层,举例来说:在一种实现中,321层为卷积层,322层为池化层,323层为卷积层,324层为池化层,325为卷积层,326为池化层;在另一种实现方式中,321、322为卷积层,323为池化层,324、325为卷积层,326为池化层。即卷积层的输出可以作为随后的池化层的输入,也可以作为另一个卷积层的输入以继续进行卷积操作。

[0136] 下面将以卷积层321为例,介绍一层卷积层的内部工作原理。

[0137] 卷积层321可以包括很多个卷积算子,卷积算子也称为核,其在自然语言处理中的作用相当于一个从输入的语音或语义信息中提取特定信息的过滤器,卷积算子本质上可以

是一个权重矩阵,这个权重矩阵通常被预先定义。

[0138] 这些权重矩阵中的权重值在实际应用中需要经过大量的训练得到,通过训练得到的权重值形成的各个权重矩阵可以从输入数据中提取信息,从而帮助卷积神经网络300进行正确的预测。

[0139] 当卷积神经网络300有多个卷积层的时候,初始的卷积层(例如321)往往提取较多的一般特征,该一般特征也可以称之为低级别的特征;随着卷积神经网络300深度的加深,越往后的卷积层(例如326)提取到的特征越来越复杂,比如高级别的语义之类的特征,语义越高的特征越适用于待解决的问题。

[0140] 池化层:

[0141] 由于常常需要减少训练参数的数量,因此卷积层之后常常需要周期性的引入池化层,即如图5中320所示例的321-326各层,可以是一层卷积层后面跟一层池化层,也可以是多层卷积层后面接一层或多层池化层。在自然语言数据处理过程中,池化层的唯一目的就是减少数据的空间大小。

[0142] 神经网络层330:

[0143] 在经过卷积层/池化层320的处理后,卷积神经网络300还不足以输出所需要的输出信息。因为如前所述,卷积层/池化层320只会提取特征,并减少输入数据带来的参数。然而为了生成最终的输出信息(所需要的类信息或别的相关信息),卷积神经网络300需要利用神经网络层330来生成一个或者一组所需要的类的数量的输出。因此,在神经网络层330中可以包括多层隐含层(如图5所示的331、332至33n)以及输出层340,该多层隐含层中所包含的参数可以根据具体的任务类型的相关训练数据进行预先训练得到,例如该任务类型可以包括语音或语义识别、分类或生成等等。

[0144] 在神经网络层330中的多层隐含层之后,也就是整个卷积神经网络300的最后层为输出层340,该输出层340具有类似分类交叉熵的损失函数,具体用于计算预测误差,一旦整个卷积神经网络300的前向传播(如图5由310至340的传播为前向传播)完成,反向传播(如图5由340至310的传播为反向传播)就会开始更新前面提到的各层的权重值以及偏差,以减少卷积神经网络300的损失及卷积神经网络300通过输出层输出的结果和理想结果之间的误差。

[0145] 需要说明的是,如图5所示的卷积神经网络300仅作为一种卷积神经网络的示例,在具体的应用中,卷积神经网络还可以以其他网络模型的形式存在。

[0146] 如图6所示,卷积神经网络(CNN)300可以包括输入层310,卷积层/池化层320(其中池化层为可选的),以及神经网络层330,在图6中,卷积层/池化层320中的多个卷积层/池化层并行,将分别提取的特征均输入给全神经网络层330进行处理。

[0147] 图7为本申请实施例提供的一种芯片的硬件结构的示意图。该芯片包括神经网络处理器(neural processing unit,NPU)40。该芯片可以被设置在如图4所示的执行设备110中,用以完成计算模块111的计算工作。该芯片也可以被设置在如图4所示的训练设备120中,用以完成训练设备120的训练工作并输出目标模型/规则101。如图5和图6所示的卷积神经网络中各层的算法均可在如图7所示的芯片中得以实现。

[0148] NPU 40作为协处理器可以挂载到主CPU(host CPU)上,由主CPU分配任务。NPU 40的核心部分为运算电路403,在NPU 40工作时,NPU 40中的控制器404可以控制运算电路403

提取存储器(权重存储器或输入存储器)中的数据并进行运算。

[0149] 在一些实现中,运算电路403内部包括多个处理单元(process engine,PE)。在一些实现中,运算电路403是二维脉动阵列。运算电路403还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中,运算电路403是通用的矩阵处理器。

[0150] 举例来说,假设有输入矩阵A,权重矩阵B,输出矩阵C。运算电路从权重存储器402中取矩阵B相应的数据,并缓存在运算电路中每一个PE上。运算电路从输入存储器401中取矩阵A数据与矩阵B进行矩阵运算,得到的矩阵的部分结果或最终结果,保存在累加器(accumulator)408中。

[0151] 向量计算单元407可以对运算电路的输出做进一步处理,如向量乘,向量加,指数运算,对数运算,大小比较等等。例如,向量计算单元407可以用于神经网络中非卷积/非全连接层(fully connected layers,FC)层的网络计算,如池化(pooling),批归一化(batch normalization),局部响应归一化(local response normalization)等。

[0152] 在一些实现中,向量计算单元407能将经处理的输出的向量存储到统一缓存器406。例如,向量计算单元407可以将非线性函数应用到运算电路403的输出,例如累加值的向量,用以生成激活值。在一些实现中,向量计算单元407生成归一化的值、合并值,或二者均有。在一些实现中,处理过的输出的向量能够用作到运算电路403的激活输入,例如用于在神经网络中的后续层中的使用。

[0153] 统一存储器406用于存放输入数据以及输出数据。

[0154] 权重数据直接通过存储单元访问控制器405(direct memory access controller,DMAC)将外部存储器中的输入数据搬运到输入存储器401和/或统一存储器406、将外部存储器中的权重数据存入权重存储器402,以及将统一存储器406中的数据存入外部存储器。

[0155] 总线接口单元(bus interface unit,BIU)410,用于通过总线实现主CPU、DMAC和取指存储器409之间进行交互。

[0156] 与控制器404连接的取指存储器(instruction fetch buffer)409,用于存储控制器404使用的指令;

[0157] 控制器404,用于调用指存储器409中缓存的指令,实现控制该运算加速器的工作过程。

[0158] 一般地,统一存储器406,输入存储器401,权重存储器402以及取指存储器409均可以为片上(on-chip)存储器。NPU的外部存储器可以为该NPU外部的存储器,该外部存储器可以为双倍数据率同步动态随机存储器(double data rate synchronous dynamic random access memory,DDR SDRAM)、高带宽存储器(high bandwidth memory,HBM)或其他可读可写的存储器。应理解,图7示出的芯片硬件结果仅为示例性说明,本申请并未限定于此。

[0159] 图8为本申请实施例提供的一种翻译场景下的系统结构的示意图。如图8所示,本申请实施例中的文本处理方法可以由自然语言理解(natural language understand,NLU)云侧模块执行。

[0160] 该系统包括视觉模块、OCR引擎模块、OCR识别模块、NLU模块、NLU云侧模块、翻译模块和翻译云模块。

[0161] 视觉模块:用于采集图片。例如,视觉模块可以通过拍照等方式采集图片。

[0162] OCR引擎模块:用于OCR任务的调度。

[0163] OCR识别模块:用于基于OCR算法实现字符的识别。

[0164] NLU模块:用于NLU相关任务的调度。

[0165] NLU云侧模块:用于对接收到的文本中的错词/语法进行纠错。

[0166] 翻译模块:用于多种语言间的翻译任务的调度。

[0167] 翻译云模块:用于对接收到的文本进行翻译。

[0168] 下面结合图8对本申请实施例中的文本处理方法应用于翻译场景进行详细介绍。

[0169] S1:视觉模块将采集的图片传输至OCR引擎模块。

[0170] S2:OCR引擎模块通过调度将图片传输至OCR识别模块。

[0171] S3:OCR识别模块识别出图片中的文字,即原文,将原文返回至OCR引擎模块。

[0172] S4:OCR引擎模块将原文传输至NLU模块。

[0173] S5:NLU模块通过调度将原文传输至NLU云侧模块。

[0174] S6:NLU云侧模块对原文中的错词/语法进行纠错,得到纠错后的原文。

[0175] S7:NLU云侧模块将纠错后的原文返回至NLU模块。

[0176] S8:NLU模块将纠错后的原文回传至OCR引擎模块。

[0177] S9:OCR引擎模块将纠错后的原文传输至翻译模块。

[0178] S10:翻译模块通过调度将纠错后的原文传输至翻译云模块。

[0179] S11:翻译云模块进行翻译,得到译文,并回传至翻译模块。

[0180] S12:翻译模块将译文回传至OCR引擎模块。

[0181] 该系统中,文本处理方法用于文本纠错,也就是文本处理模型可以为文本纠错模型。将待处理文本输入文本处理模型中进行纠错处理,即可得到对待处理文本的纠正结果。应理解,图8仅为本申请实施例中的文本处理方法的一种示例,在另一种可能的实现方式中,文本处理模型可以文本翻译模型,将待处理文本输入文本翻译模型中进行纠错处理,并将纠错后的结果进行翻译处理,即可得到对待处理文本的翻译文本。

[0182] 应理解,以上仅为本申请实施例中的文本处理模型的示例。

[0183] 图8中的文本处理模型部署在云服务器上,应理解,文本处理模型也可以部署在智能终端设备上。智能终端可以是具有摄像头的电子设备,例如,智能终端可以是有图像处理功能的移动电话、平板个人电脑(tablet personal computer,TPC)、媒体播放器、智能电视、笔记本电脑(laptop computer,LC)、个人数字助理(personal digital assistant,PDA)、个人计算机(personal computer,PC)或者自动驾驶车辆中的车载终端等,本申请实施例对此不作限定。

[0184] 图9是本申请实施例提供的文本处理方法的示意性流程图。图9所示文本处理方法可以由文本处理装置执行,该装置具体可以是图1中的数据处理设备,也可以是图2中的用户设备,也可以是图3中的执行设备110或者本地设备,也可以是图4中的执行设备210。图8所示的方法包括步骤510至530,下面分别对步骤510至530进行详细的介绍。

[0185] 510,获取待处理文本。

[0186] 待处理文本可以是OCR输出文本,或者,可以是用户输入的文本。

[0187] 例如,OCR输出文本可以包括便携式文档格式(portable document format,pdf)

对应的OCR输出文本。OCR输出文本还可以包括演示文稿(power point,PPT)对应OCR输出文本。OCR输出文本还可以包括拍摄的图片对应的OCR输出文本。

[0188] 再如,用户输入的文本可以包括社交网络中发布的内容,或者可以是搜索引擎的搜索框中输入的内容等。应理解,待处理文本可以是任意需要进行纠错的文本,本申请实施例对待处理文本的具体形式不做限定。

[0189] 520,对待处理文本进行检错处理,得到待处理文本中的非词。

[0190] 待处理文本中的错误可以包括非词错误(non-word error)和真词错误(real-word error)。非词错误是指待处理文本中的词不在第二预设词库中。真词错误是指待处理文本中的词存在于第二预设词库中,但导致上下文语义出现问题,不是当前语境所需要的词。也就是说待处理文本中的错词可以包括非词和错误真词。

[0191] 示例性地,可以基于第二预设词库对待处理文本进行非词错误检测,得到待处理文本中的非词。第二预设词库能够用于区分真词和非词。真词指的是存在于第二预设词库中的词,相对应地,非词指的是不存在于第一预设词库中的词。

[0192] 能够用于检测非词错误的词库均可以理解为第二预设词库。例如,在方法500应用于英文文本纠错的情况下,第二预设词库可以为英文词库。非词即为不存在于英文词库的词,例如,werld。本申请实施例对第二预设词库的类型不做限定。

[0193] 应理解,在本申请实施例中,“词库”也可以称为“字典”或“词表”等。

[0194] 530,若待处理文本中的非词属于第一类非词,将待处理文本中的非词作为待处理文本中的非词的校正结果,即不对待处理文本中的非词进行处理。

[0195] 第一类非词包括全大写字母非词、字长在预设字长范围内的非词和属于预设词库的非词。

[0196] 若待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与待处理文本中的非词的类别匹配的校正方式对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果。

[0197] 第二类非词包括合并错误非词,第三类非词包括包含非字母的字符的非词。第四类非词包括除第一类非词、第二类非词、第三类非词之外的其他非词。在本申请实施例中,第四类非词也可以称为正常类型非词。

[0198] 如前所述,第一类非词包括字长在预设字长范围内的非词。字长在预设字长范围内的非词可以包括字长大于第一预设字长和/或字长小于第二预设字长的非词。

[0199] 也就是说字长在预设字长范围内的非词可以包括过长或者过短的非词。例如,过长的非词可能为网址。过短的非词可能仅包括一个或两个字符等。

[0200] 过长的非词通常是专有名词,一般不需要该类型的非词进行处理,通过将该类型的非词检测出来,避免在对该类型的非词进行处理,避免出现纠正错误。

[0201] 过短的非词中的有效信息过少,修正后的可信度不高,通过将该类型的非词检测出来,避免在对该类型的非词进行处理,能够提高文本处理的速度。

[0202] 如前所述,第一类非词包括属于第一预设词库的非词。

[0203] 其中,第一预设词库可以包括预设低频词库。预设低频词库可以根据应用需要设置。例如,预设低频词库可以包括人名、地名等。

[0204] 第一预设词库还可以包括其他语种词库。例如,方法500用于对英文的待处理文本

进行纠错,则其他语种词库可以包括俄法德意葡等语种词库和/或拼音词库。

[0205] 第二类非词包括合并错误非词。合并错误非词为包括至少两个真词的非词。

[0206] 若待处理文本中的非词属于第二类非词,对待处理文本中的非词进行校正,得到至少两个真词作为待处理文本中的非词的校正结果。也就是利用得到的至少两个真词替换该非词。

[0207] 具体地,可以在该非词中适当的位置添加空格,将该非词修改为至少两个真词。

[0208] 例如,该处理文本中的非词为inChina,该非词属于第二类非词,在该非词中添加空格,将该非词修改为in China。

[0209] 进一步地,可以将修改后的至少两个真词输入语言模型中,若该修改后的至少两个真词能够使得困惑度降低,则将修改后的至少两个真词作为待处理文本中的非词的校正结果。

[0210] 第三类非词包括包含非字母的字符的非词。

[0211] 示例性地,第三类非词包括包含数字的非词。例如,a5ses。

[0212] 若待处理文本中的非词属于第三类非词,将待处理文本中的非词中的非字母的字符修改为字母,将修改后的词作为待处理文本中的非词的校正结果。也就是用修改后的词替换该非词。

[0213] 具体地,可以通过预设的字符误判词库将待处理文本中的非词中除字母以外的其他字符改为字母,将修改后的词作为待处理文本中的非词的校正结果。

[0214] 可选地,还可以通过预设的字符误判词库将待处理文本中的非词作为待处理文本中的非词的校正结果,即不对该非词进行处理。

[0215] 其中,在待处理文本为OCR输出文本的情况下,字符误判词库可以是根据OCR识别错误的概率确定的。OCR识别错误的概率指的是将字母误识别为数字的概率。示例性地,字符误判词库可以是基于图8中的OCR识别模块的识别结果,通过历史经验确定的。

[0216] 例如,数字0和字母O相似,OCR识别错误的概率较大。当非词中出现0,可以将0替换为O。再如,数字5和字母s较相似,OCR识别错误的概率较大。当非词中出现5,可以将5替换为s,如a5sess可以替换为assess。再如,对于非词iphone4,数字4在字符误判词库中并没有对应的字母,则可以不对非词iphone4进行处理。

[0217] 进一步地,可以将修改后的词输入语言模型中,若该修改后的词能够使得困惑度降低,则将修改后的词作为待处理文本中的非词的校正结果。

[0218] 若待处理文本中的非词属于第四类非词,对待处理文本中的非词进行校正包括步骤A1至步骤A3。

[0219] A1,生成待处理文本中的非词对应的候选词。

[0220] 待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性满足第一预设条件。

[0221] 待处理文本中的非词对应的候选词可以为一个,也可以为多个。

[0222] 待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性可以包括待处理文本中的非词与待处理文本中的非词对应的候选词之间的编辑距离和/或公共字符串。也就是可以基于编辑距离和/或公共字符串确定待处理文本中的非词对应的候选词。

[0223] 编辑距离是指将一个词转换为另一个词所需的编辑操作次数。编辑操作包括对词

中的字符的插入、删除、易位和替换等操作。

[0224] 公共字符串是指两个词中所包含的连续相同的字符的数量。

[0225] 示例性地,可以通过BK树(Burkhard Keller tree)生成待处理文本中的非词对应的候选词。

[0226] 以待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性包括待处理文本中的非词与待处理文本中的非词对应的候选词之间的编辑距离为例对步骤A1进行说明。

[0227] 示例性地,第一预设条件可以为编辑距离小于第一预设值。

[0228] 例如,限制待处理文本中的非词与该待处理文本中的非词对应的候选词的编辑距离小于3,生成该待处理文本中的非词对应的候选词。也就是由非词生成该待处理文本中的非词对应的候选词的过程中操作次数小于三次。

[0229] 以待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性包括待处理文本中的非词与待处理文本中的非词对应的候选词之间的编辑距离和公共字符串为例对步骤A1进行说明。

[0230] 示例性地,第一预设条件可以包括预设条件一和预设条件二。预设条件一可以为编辑距离小于预设值。预设条件二可以为最大公共字符串的长度大于长度预设值。

[0231] 具体地,可以通过预设条件一和预设条件二分别生成待处理文本中的非词对应的候选词,即待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性满足预设条件一或预设条件二。

[0232] 可替换地,可以通过预设条件一和预设条件二同时生成待处理文本中的非词对应的候选词,即待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性满足预设条件一和预设条件二。

[0233] 例如,限制待处理文本中的非词与该待处理文本中的非词对应的候选词的编辑距离小于3,生成该待处理文本中的非词对应的候选词A。也就是由非词生成该待处理文本中的非词对应的候选词的过程中操作次数小于三次。限制待处理文本中的非词与该待处理文本中的非词对应的候选词的最大公共字符串的长度大于3,生成该待处理文本中的非词对应的候选词B。也就是该待处理文本中的非词对应的候选词与该非词包括连续相同的字符的数量大于3。待处理文本中的非词对应的候选词可以包括候选词A和候选词B。可替换地,待处理文本中的非词对应的候选词可以包括候选词A和候选词B中相同的候选词。

[0234] 应理解,该相似性也可以是其他形式的相似性,例如字符相似性等,本申请对待处理文本中的非词与待处理文本中的非词对应的多个候选词之间的相似性的确定方式不作限定。

[0235] 例如,待处理文本中的非词为werld,基于最小编辑距离和/或最大公共字符串确定werld对应的候选词可以包括world、word和sword等。

[0236] A2,在待处理文本中的非词对应的候选词中确定待处理文本中的非词对应的目标候选词。

[0237] 具体地,步骤A2可以为,在待处理文本中的非词对应的候选词中随机确定待处理文本中的非词对应的目标候选词。

[0238] 可选地,步骤A2可以包括A21和步骤A22。

[0239] A21,根据待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性以及待处理文本中的非词对应的候选词的困惑度对待处理文本中的非词对应的候选词进行评分,其中,待处理文本中的非词对应的候选词的困惑度用于指示待处理文本中的非词对应的候选词在待处理文本中出现的可能性。

[0240] 示例性地,待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性可以包括:待处理文本中的非词与待处理文本中的非词对应的候选词之间的编辑距离。

[0241] 待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性可以基于编辑距离进行评分。

[0242] 待处理文本中的非词对应的候选词的困惑度可以通过语言模型进行评分。

[0243] 例如,该语言模型可以为统计语言模型,例如,n-gram模型。

[0244] 统计语言模型对中短文本语义信息的提取更占优势,适用于较少依赖长距离的语义信息的场景,例如,OCR场景下的文本纠错。

[0245] 再例如,该语言模型也可以为神经网络模型,例如,循环神经网络(recurrent neural network,RNN)模型。

[0246] 每个候选词对应的评分可以由上述几项对应的评分进行加权得到,也就是为每一项对应的评分设置权重。该权重可以是预先设定的,也可以是训练得到的。

[0247] 根据相似性和文本的困惑度进行评分,同时考虑了非词与候选词的相似性以及待处理文本的语义信息,能够得到更准确的评分结果。

[0248] A22,将待处理文本中的非词对应的候选词中评分最高的候选词确定为待处理文本中的非词对应的目标候选词。

[0249] 应理解,在步骤A1的“相似性”的确定方式和步骤A2中“相似性”的确定方式可以相同,也可以不同。

[0250] 在步骤A1中的待处理文本中的非词对应的候选词为一个时,步骤A2可以省略,即直接将该候选词作为待处理文本中的非词对应的目标候选词。

[0251] A3,根据待处理文本中的非词对应的目标候选词对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果。

[0252] 对待处理文本中的非词进行校正可以包括利用待处理文本中的非词对应的目标候选词替换待处理文本中的非词,也包括不对待处理文本中的非词进行处理,也就是不对待处理文本中的非词进行替换。

[0253] 可选地,步骤A3可以为步骤A31。

[0254] A31,直接利用待处理文本中的非词对应的目标候选词替换待处理文本中的非词,作为待处理文本中的非词的校正结果。

[0255] 可替换地,步骤A3可以为步骤A32。

[0256] A32,通过语言模型检测包含待处理文本中的非词对应的目标候选词的文本的困惑度,在困惑度低于或等于第一困惑度阈值的情况下,利用待处理文本中的非词对应的目标候选词替换待处理文本中的非词,作为待处理文本中的非词的校正结果。

[0257] 进一步地,在困惑度高于第一困惑度阈值的情况下,不对待处理文本中的非词进行纠正。也就是不利用待处理文本中的非词对应的目标候选词替代待处理文本中的非词。这样可以降低时间消耗,快速实现文本纠错。

[0258] 可替换地,在困惑度高于第一困惑度阈值的情况下,可以将步骤A21中评分第二高的候选词作为待处理文本中的非词对应的目标候选词,重复步骤A32,直至得到困惑度满足第一困惑度阈值的目标候选词,并利用待处理文本中的非词对应的目标候选词替换待处理文本中的非词。

[0259] 可选地,方法500还可以包括步骤540。具体地,步骤540包括步骤541至步骤543。

[0260] 541,对待处理文本中的真词进行真词错误检测,得到待处理文本中的错误真词。

[0261] 具体地,可以基于语言模型对待处理文本中的真词进行真词错误检测,得到待处理文本中的错误真词。例如,当一个词对应的文本困惑度(perplexity)高于设定阈值的情况下,判断该词为错误真词。

[0262] 542,生成错误真词对应的候选词。

[0263] 具体地,错误真词与错误真词对应的候选词之间的相似性满足第二预设条件。

[0264] 错误真词对应的候选词可以为一个,也可以为多个。

[0265] 错误真词与错误真词对应的候选词之间的相似性可以包括错误真词与错误真词对应的候选词之间的编辑距离和/或公共字符串。也就是可以基于编辑距离和/或公共字符串确定错误真词对应的候选词。

[0266] 示例性地,可以通过BK树生成错误真词对应的候选词。

[0267] 以错误真词与错误真词对应的候选词之间的相似性包括错误真词与错误真词对应的候选词之间的编辑距离为例对步骤542进行说明。

[0268] 第二预设条件可以为编辑距离小于第二预设值。例如,该第二预设值可以为2。

[0269] 具体地,限制错误真词与该错误真词对应的候选词的编辑距离小于3,生成该错误真词对应的候选词。也就是由错误真词生成该错误真词对应的候选词的过程中操作次数小于三次。

[0270] 以错误真词与错误真词对应的候选词之间的相似性包括错误真词与错误真词对应的候选词之间的编辑距离和公共字符串为例对步骤542进行说明。

[0271] 示例性地,第二预设条件可以包括预设条件三和预设条件四。预设条件三可以为编辑距离小于预设值。预设条件四可以为最大公共字符串的长度大于长度预设值。

[0272] 具体地,可以通过预设条件三和预设条四分别生成错误真词对应的候选词,即错误真词与错误真词对应的候选词之间的相似性满足预设条件三或预设条件四。

[0273] 可替换地,可以通过预设条件三和预设条件四同时生成错误真词对应的候选词,即错误真词与错误真词对应的候选词之间的相似性满足预设条件三和预设条件四。

[0274] 例如,限制错误真词与错误真词对应的候选词的编辑距离小于3,生成该错误真词对应的候选词C。也就是由错误真词生成该错误真词对应的候选词的过程中操作次数小于三次。限制错误真词与该错误真词对应的候选词的最大公共字符串的长度大于3,生成该错误真词对应的候选词D。也就是该错误真词对应的候选词与该错误真词包括连续相同的字符的数量大于3。错误真词对应的候选词可以包括候选词C和候选词D。可替换地,错误真词对应的候选词可以包括候选词C和候选词D中相同的候选词。

[0275] 第一预设条件和第二预设条件可以相同,也可以不同。第一预设值和第二预设值可以相同也可以不同。

[0276] 应理解,该相似性也可以是其他形式的相似性,例如字符相似性等,本申请对错误

真词与错误真词对应的候选词之间的相似性的确定方式不作限定。

[0277] 例如,word为错误真词,其对应的候选词可以包括:world、words和sword等。

[0278] 543,在错误真词对应的候选词中确定错误真词对应的目标候选词。

[0279] 具体地,步骤543可以为,在错误真词对应的候选词中随机确定错误真词对应的目标候选词。

[0280] 可选地,步骤543可以包括步骤543a和步骤543b。

[0281] 543a,根据错误真词与错误真词对应的候选词之间的相似性以及错误真词对应的候选词的困惑度对错误真词对应的候选词进行评分。其中,错误真词对应的候选词的困惑度用于指示错误真词对应的候选词在待处理文本中出现的可能性。

[0282] 示例性地,错误真词与错误真词对应的候选词之间的相似性可以包括:错误真词与错误真词对应的候选词之间的编辑距离。

[0283] 错误真词与错误真词对应的候选词之间的相似性可以基于编辑距离进行评分。

[0284] 错误真词对应的候选词的困惑度可以通过语言模型进行评分。

[0285] 每个候选词对应的评分可以由上述几项对应的评分进行加权得到,也就是为每一项对应的评分设置权重。该权重可以是预先设定的,也可以是训练得到的。

[0286] 根据相似性和文本的困惑度进行评分,同时考虑了错误真词与候选词的相似性以及待处理文本的语义信息,能够得到更准确的评分结果。

[0287] 543b,将错误真词对应的候选词中评分最高的候选词确定为错误真词对应的目标候选词。

[0288] 应理解,在步骤543a的“相似性”的确定方式和步骤543b中“相似性”的确定方式可以相同,也可以不同。

[0289] 544,根据错误真词对应的目标候选词对错误真词进行校正。

[0290] 对错误真词进行校正可以包括利用错误真词对应的目标候选词替换错误真词,也包括不对错误真词进行处理,也就是不对错误真词进行替换。

[0291] 可选地,步骤544可以为步骤544a。

[0292] 544a,直接利用错误真词对应的目标候选词替换错误真词,作为错误真词的校正结果。

[0293] 可替换地,步骤544可以为步骤544b。

[0294] 544b,通过语言模型检测包含错误真词对应的目标候选词的文本的困惑度,在困惑度低于或等于第二困惑度阈值的情况下,利用错误真词对应的目标候选词替换错误真词,作为错误真词的校正结果。

[0295] 第一困惑度阈值和第二困惑度阈值可以相同,也可以不同。

[0296] 进一步地,在困惑度高于第二困惑度阈值的情况下,不对错误真词进行纠正。也就是不利用错误真词对应的目标候选词替代错误真词。这样可以降低时间消耗,快速实现文本纠错。

[0297] 可替换地,在困惑度高于第二困惑度阈值的情况下,可以将步骤543a中评分第二高的候选词作为错误真词对应的目标候选词,重复步骤544b,直至得到困惑度满足第二困惑度阈值的目标候选词,并利用错误真词对应的目标候选词替换错误真词。

[0298] 在本申请的实施例中,能够对待处理文本中的多种类型的字符分别进行检测以及

分别进行处理,降低了多种类型的字符对纠错过程的干扰,提升了文本纠错的准确率,提高了纠错方法对输入文本的鲁棒性。

[0299] 此外,采用候选词与错词之间的相似性以及候选词的困惑度对候选词进行评分,同时考虑了错词与候选词的相似性以及待处理文本的语义信息,能够更符合输入文本的原意,能够得到较优的候选词,提高了文本纠错的准确性。

[0300] 图10是本申请实施例提供的文本处理方法600的流程示意图。方法600为通过方法500对正常类型非词和错误真词的处理方法的一种示例。方法600包括步骤610至步骤6120。下面对步骤610至步骤6120进行详细说明。

[0301] 610,获取待处理文本。

[0302] 620,对待处理文本进行非词错误检测。步骤620对应方法500中的步骤520。

[0303] 具体地,可以基于英文词库对待处理文本进行非词错误检测该英文词库为方法500中的第二预设词库的一例。

[0304] 步骤620用于得到待处理文本中的非词和真词。非词即为不存在于英文词库中的词。真词为存在于英文词库中的词。

[0305] 对待处理文本中的非词执行步骤630。对于待处理文本中的真词执行步骤680。

[0306] 630,判断待处理文本中的非词的类别,得到正常类型非词。该正常类型非词为方法500中的第四类非词的一例。

[0307] 为了便于描述,将正常类型非词称为非词1#。非词1#可以包括一个非词,也可以包括多个非词。

[0308] 640,生成非词1#对应的候选词。非词1#对应的候选词可以为一个,也可以为多个。

[0309] 非词1#与非词1#对应的候选词之间的相似性满足第一预设条件。

[0310] 示例性地,可以通过BK树生成非词1#对应的候选词。详细过程如方法500中的步骤A1所述,此处不再赘述。

[0311] 650,对非词1#对应的候选词进行评分。

[0312] 示例性地,可以通过语言模型和编辑距离等对非词1#对应的候选词进行评分。详细过程如方法500中的步骤A21所述,此处不再赘述。

[0313] 660,确定非词1#对应的目标候选词。

[0314] 具体地,可以将将非词1#对应的候选词中评分最高的候选词确定为非词1#对应的目标候选词。

[0315] 在非词1#对应的候选词为一个时,步骤650可以省略,即直接将非词1#对应的候选词作为非词1#对应的目标候选词。

[0316] 670,根据非词1#对应的目标候选词对非词1#进行校正,得到非词1#的校正结果。

[0317] 示例性地,可以通过语言模型检测包含非词1#对应的目标候选词的文本的困惑度,在困惑度低于或等于第一困惑度阈值的情况下,利用非词1#对应的目标候选词替换非词1#,作为非词1#的校正结果。详细过程如方法500中的步骤A32所述。

[0318] 步骤630至步骤670对应方法500中的步骤530。

[0319] 680,对待处理文本中的真词进行真词错误检测,得到待处理文本中的错误真词。

[0320] 具体地,基于语言模型对待处理文本中的真词进行真词错误检测,得到待处理文本中的错误真词。例如,当一个词对应的文本困惑度高于设定阈值,判断该词为错误真词。

步骤680对应方法500中的步骤541。

[0321] 对待处理文本中的错误真词执行步骤690。为了便于描述,将处理文本中的错误真词称为错误真词1#。错误真词1#可以包括一个错误真词,也可以包括多个错误真词。

[0322] 690,生成错误真词1#对应的候选词。错误真词1#对应的候选词可以为一个,也可以为多个。

[0323] 错误真词1#与错误真词1#对应的候选词之间的相似性满足第二预设条件。

[0324] 示例性地,可以通过BK树生成候选真词1#对应的候选词。详细过程如方法500中的步骤542所述,此处不再赘述。

[0325] 6100,对错误真词1#对应的候选词进行评分。

[0326] 示例性地,可以通过语言模型和编辑距离等对错误真词1#对应的候选词进行评分。详细过程如方法500中的步骤543a所述,此处不再赘述。

[0327] 6110,确定错误真词1#对应的目标候选词。

[0328] 具体地,将错误真词1#对应的候选词中评分最高的候选词确定为错误真词1#对应的目标候选词。

[0329] 在错误真词1#对应的候选词为一个时,步骤6110可以省略,即直接将错误真词1#对应的候选词作为错误真词1#对应的目标候选词。

[0330] 6120,根据错误真词1#对应的目标候选词对错误真词1#进行校正,得到错误真词1#的校正结果。

[0331] 示例性地,可以通过语言模型检测包含错误真词1#对应的目标候选词的文本的困惑度,在困惑度低于或等于第二困惑度阈值的情况下,利用错误真词1#对应的目标候选词替换错误真词1#,作为错误真词1#的校正结果。详细过程如方法500中的步骤544所述。

[0332] 在本申请实施例中,通过非词类别的判断,得到正常类型的非词,避免其他类别的非词对文本纠错的影响。对于正常类型的非词和错误真词,通过语言模型等方式对错词的候选词进行评分,充分利用了输入文本的语义信息,能够更符合输入文本的原意,进一步提高了文本纠错的准确率。

[0333] 图11是本申请实施例提供的文本处理方法700的流程示意图。方法700为方法500的一种示例。方法700包括步骤710至步骤740。下面对步骤710至步骤740进行详细说明。

[0334] 710,获取待处理文本。

[0335] 进一步地,方法700还可以包括步骤711。

[0336] 711,判断待处理文本的长度,在待处理文本的长度大于或等于预设长度的情况下,执行步骤720。其中,待处理文本的长度指的是待处理文本中的词的数量。例如,预设长度可以为2。

[0337] 这样保证了待处理文本中的词的数量,能够更好地利用上下文语义信息,提高了文本纠错的准确率。

[0338] 720,基于英文词库对待处理文本进行非词错误检测。该英文词库为方法500中的第二预设词库的一例。步骤720对应方法500中的步骤520。

[0339] 步骤720用于得到待处理文本中的非词和真词。非词即为不存在于英文词库中的词。真词为存在于英文词库中的词。

[0340] 对待处理文本中的非词执行步骤730。对于待处理文本中的真词执行步骤740。

[0341] 730,判断待处理文本中的非词的类别。

[0342] 非词包括全字母非词和特殊非词。全字母非词指的是包含英文52个大小写字母的非词。全字母非词可以包括方法500中的第一类非词、第二类非词和第四类非词。特殊非词指的是包含非字母的字符的非词。特殊非词可以为方法500中的第三类非词的一例。

[0343] 示例性地,步骤730包括步骤731至步骤732。

[0344] 731,判断待处理文本中的非词属于全字母非词还是特殊非词。

[0345] 若待处理文本中的非词属于特殊非词,则可以对该非词进行针对性处理。例如将待处理文本中的非词中的非字母的字符修改为字母,将修改后的词作为待处理文本中的非词的校正结果。也就是用修改后的词替换该非词。

[0346] 具体地,可以通过预设的字符误判词库将待处理文本中的非词中除字母以外的其他字符改为字母,将修改后的词作为待处理文本中的非词的校正结果。

[0347] 进一步地,可以将修改后的词输入语言模型中,若该修改后的词能够使得困惑度降低,则将修改后的词作为待处理文本中的非词的校正结果。

[0348] 若待处理文本中的非词属于全字母非词,则执行步骤732。

[0349] 732,可以在全字母非词中进一步判断非词的类别。

[0350] 示例性地,可以对待处理文本中的非词进行大小写字母检测,若待处理文本中的非词为属于全大写非词,即待处理文本中的非词属于第一类非词,则将待处理文本中的非词作为待处理文本中的非词的校正结果,即不对待处理文本中的非词进行处理。

[0351] 示例性地,可以判断待处理文本中的非词的字长,若待处理文本中的非词属于字长在预设字长范围内的非词,即待处理文本中的非词属于第一类非词,则将待处理文本中的非词作为待处理文本中的非词的校正结果,即不对待处理文本中的非词进行处理。

[0352] 示例性地,可以判断待处理文本中的非词是否属于拼音词库,即待处理文本中的非词是否属于第一预设词库。属于拼音词库可以称为拼音非词。若待处理文本中的非词属于拼音非词,即待处理文本中的非词属于第一类非词,则将待处理文本中的非词作为待处理文本中的非词的校正结果,即不对待处理文本中的非词进行处理。

[0353] 示例性地,可以判断待处理文本中的非词是否属于预设低频词库,即待处理文本中的非词是否属于第一预设词库。属于预设低频词库的非词可以称为低频非词。若待处理文本中的非词属于低频非词,即待处理文本中的非词属于第一类非词,则将待处理文本中的非词作为待处理文本中的非词的校正结果,即不对待处理文本中的非词进行处理。

[0354] 示例性地,可以判断待处理文本中的非词是否属于其他语种词库,即待处理文本中的非词是否属于第一预设词库。属于其他语种词库的非词可以称为其他语种非词。若待处理文本中的非词属于其他语种非词,即待处理文本中的非词属于第一类非词,则将待处理文本中的非词作为待处理文本中的非词的校正结果,即不对待处理文本中的非词进行处理。

[0355] 示例性地,可以判断待处理文本中的非词是否为合并错误非词,即第二类非词。若待处理文本中的非词属于第二类非词,则可以在该非词进行针对性处理。例如,在该非词中适当的位置添加空格,将该非词修改为至少两个真词。进一步地,可以将修改后的至少两个真词输入语言模型中,若该修改后的至少两个真词能够使得困惑度降低,则将修改后的至少两个真词作为待处理文本中的非词的校正结果。

[0356] 示例性地,若待处理文本中的非词不属于上述第一类非词、第二类非词、第三类非词,则待处理文本中的非词属于正常类型非词,即方法500中的第四类非词。若待处理文本中的非词属于正常类型非词,生成该非词对应的候选词,根据该非词对应的候选词对该非词进行校正。详细过程可以按照方法500中的步骤A1至步骤A3对第四类非词进行校正。

[0357] 740,基于语言模型对待处理文本中的真词进行真词错误检测,得到待处理文本中的错误真词。生成错误真词对应的候选词,根据错误真词对应的候选词对错误真词进行校正。详细过程可以按照方法500中的步骤540对错误真词进行校正。

[0358] 在本申请的实施例中,能够对待处理文本中的多种类型的字符分别进行检测以及分别进行处理,降低了多种类型的字符对纠错过程的干扰,提升了文本纠错的准确率,提高了纠错方法对输入文本的鲁棒性。

[0359] 上文结合图1至图11,详细描述了本申请实施例文本处理方法,下面将结合图12和图13,详细描述本申请的装置实施例。应理解,方法实施例的描述与装置实施例的描述相互对应,因此,未详细描述的部分可以参见前面方法实施例。

[0360] 图12是本申请实施例提供的文本处理装置的示意性框图。应理解,文本处理装置1000可以执行图9、图10或图11所示的文本处理方法。该文本处理装置1000包括:获取单元1010和处理单元1020。

[0361] 其中,获取单元1010,用于获取待处理文本。处理单元1020,用于:对待处理文本进行检错处理,得到待处理文本中的非词;若待处理文本中的非词属于第一类非词,将待处理文本中的非词作为待处理文本中的非词的校正结果;若待处理文本中的非词属于第二类非词、第三类非词或第四类非词,选择与待处理文本中的非词的类别匹配的校正方式对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果。其中,第一类非词包括全大写字母非词、字长在预设字长范围内的非词和属于第一预设词库的非词,第二类非词包括合并错误非词,第三类非词包括包含非字母的字符的非词,第四类非词包括除第一类非词、第二类非词、第三类非词之外的其他非词。

[0362] 可选地,处理单元1020用于:若待处理文本中的非词属于第四类非词,生成待处理文本中的非词对应的候选词;在待处理文本中的非词对应的候选词中确定待处理文本中的非词对应的目标候选词;根据待处理文本中的非词对应的目标候选词对待处理文本中的非词进行校正,得到待处理文本中的非词的校正结果。

[0363] 可选地,处理单元1020用于:根据待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性以及待处理文本中的非词对应的候选词的困惑度对待处理文本中的非词对应的候选词进行评分,其中,待处理文本中的非词对应的候选词的困惑度用于指示待处理文本中的非词对应的候选词在待处理文本中出现的可能性;将待处理文本中的非词对应的候选词中评分最高的候选词确定为待处理文本中的非词对应的目标候选词。

[0364] 可选地,待处理文本中的非词与待处理文本中的非词对应的候选词之间的相似性满足第一预设条件。

[0365] 可选地,处理单元1020用于:在待处理文本中的非词对应的目标候选词的困惑度低于或等于第一困惑度阈值的情况下,利用待处理文本中的非词对应的目标候选词替换待处理文本中的非词,作为待处理文本中的非词的校正结果。

[0366] 可选地,合并错误非词为包括至少两个真词的非词,以及处理单元1020用于:若待

处理文本中的非词属于第二类非词,对待处理文本中的非词进行校正,得到至少两个真词作为待处理文本中的非词的校正结果。

[0367] 可选地,处理单元1020用于:若待处理文本中的非词属于第三类非词,将待处理文本中的非词中的非字母的字符修改为字母,将修改后的词作为待处理文本中的非词的校正结果。

[0368] 可选地,处理单元1020还用于:对待处理文本进行真词错误检测,得到待处理文本中的错误真词;生成错误真词对应的候选词;在错误真词对应的候选词中确定错误真词对应的目标候选词;根据错误真词对应的目标候选词对错误真词进行校正。

[0369] 可选地,处理单元1020用于:根据错误真词与错误真词对应的候选词之间的相似性以及错误真词对应的候选词的困惑度对错误真词对应的候选词进行评分,其中,错误真词对应的候选词的困惑度用于指示错误真词对应的候选词在待处理文本中出现的可能性;将错误真词对应的候选词中评分最高的候选词确定为错误真词对应的目标候选词。

[0370] 可选地,错误真词与错误真词对应的候选词之间的相似性满足第二预设条件。

[0371] 可选地,处理单元1020用于:在错误真词对应的目标候选词的困惑度低于或等于第二困惑度阈值的情况下,利用错误真词对应的目标候选词替换错误真词,作为错误真词的校正结果。

[0372] 需要说明的是,上述文本处理装置1000以功能单元的形式体现。这里的术语“单元”可以通过软件和/或硬件形式实现,对此不作具体限定。

[0373] 例如,“单元”可以是实现上述功能的软件程序、硬件电路或二者结合。所述硬件电路可能包括应用特有集成电路(application specific integrated circuit,ASIC)、电子电路、用于执行一个或多个软件或固件程序的处理器(例如共享处理器、专有处理器或组处理器等)和存储器、合并逻辑电路和/或其它支持所描述的功能的合适组件。

[0374] 因此,在本申请的实施例中描述的各示例的单元,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0375] 图13是本申请实施例提供的文本处理装置的硬件结构示意图。图13所示的文本处理装置1200(该文本处理装置1200具体可以是一种计算机设备)包括存储器1201、处理器1202、通信接口1203以及总线1204。其中,存储器1201、处理器1202、通信接口1203通过总线1204实现彼此之间的通信连接。

[0376] 存储器1201可以是只读存储器(read only memory,ROM),静态存储设备,动态存储设备或者随机存取存储器(random access memory,RAM)。存储器1201可以存储程序,当存储器1201中存储的程序被处理器1202执行时,处理器1202用于执行本申请实施例的文本处理方法的各个步骤,例如,执行图9、图10或图11所示的各个步骤。

[0377] 应理解,本申请实施例所示的文本处理装置可以是智能终端,也可以是配置于智能终端中的芯片。

[0378] 上述本申请实施例揭示的文本处理方法可以应用于处理器1202中,或者由处理器1202实现。处理器1202可能是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述文本处理方法的各步骤可以通过处理器1202中的硬件的集成逻辑电路或者软件形式的

指令完成。例如,处理器1202可以是包含图7所示的NPU的芯片。

[0379] 上述的处理器1202可以是中央处理器(central processing unit,CPU)、图形处理器(graphics processing unit,GPU)、通用处理器、数字信号处理器(digital signal processor,DSP)、专用集成电路(application specific integrated circuit,ASIC)、现成可编程门阵列(field programmable gate array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成,或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存取存储器(random access memory,RAM)、闪存、只读存储器(read-only memory,ROM)、可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器1201,处理器1202读取存储器1201中的信息,结合其硬件完成本申请实施中图12所示的文本处理装置中包括的单元所需执行的功能,或者,执行本申请方法实施例的图9、图10或图11所示的文本处理方法。

[0380] 通信接口1203使用例如但不限于收发器一类的收发装置,来实现装置1200与其他设备或通信网络之间的通信。

[0381] 总线1204可包括在文本处理装置1200各个部件(例如,存储器1201、处理器1202、通信接口1203)之间传送信息的通路。

[0382] 应注意,尽管上述文本处理装置1200仅仅示出了存储器、处理器、通信接口,但是在具体实现过程中,本领域的技术人员应当理解,文本处理装置1200还可以包括实现正常运行所必须的其他器件。同时,根据具体需要本领域的技术人员应当理解,上述文本处理装置1200还可包括实现其他附加功能的硬件器件。此外,本领域的技术人员应当理解,上述文本处理装置1200也可仅仅包括实现本申请实施例所必须的器件,而不必包括图13中所示的全部器件。

[0383] 本申请实施例还提供一种芯片,该芯片包括收发单元和处理单元。其中,收发单元可以是输入输出电路、通信接口;处理单元为该芯片上集成的处理器或者微处理器或者集成电路。该芯片可以执行上述方法实施例中的方法。

[0384] 本申请实施例还提供一种计算机可读存储介质,其上存储有指令,该指令被执行时执行上述方法实施例中的方法。

[0385] 本申请实施例还提供一种包含指令的计算机程序产品,该指令被执行时执行上述方法实施例中的方法。

[0386] 还应理解,本申请实施例中,该存储器可以包括只读存储器和随机存取存储器,并向处理器提供指令和数据。处理器的一部分还可以包括非易失性随机存取存储器。例如,处理器还可以存储设备类型的信息。

[0387] 应理解,本文中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系。

[0388] 应理解,在本申请的各种实施例中,上述各过程的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定,而不应对本申请实施例的实施

过程构成任何限定。

[0389] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0390] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0391] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、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,所述单元的划分,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,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0392] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元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单元来实现本实施例方案的目的。

[0393] 另外,在本申请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。

[0394] 所述功能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取存储介质中。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,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,服务器,或者网络设备等)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(read-only memory,ROM)、随机存取存储器(random access memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0395] 以上所述,仅为本申请的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以所述权利要求的保护范围为准。

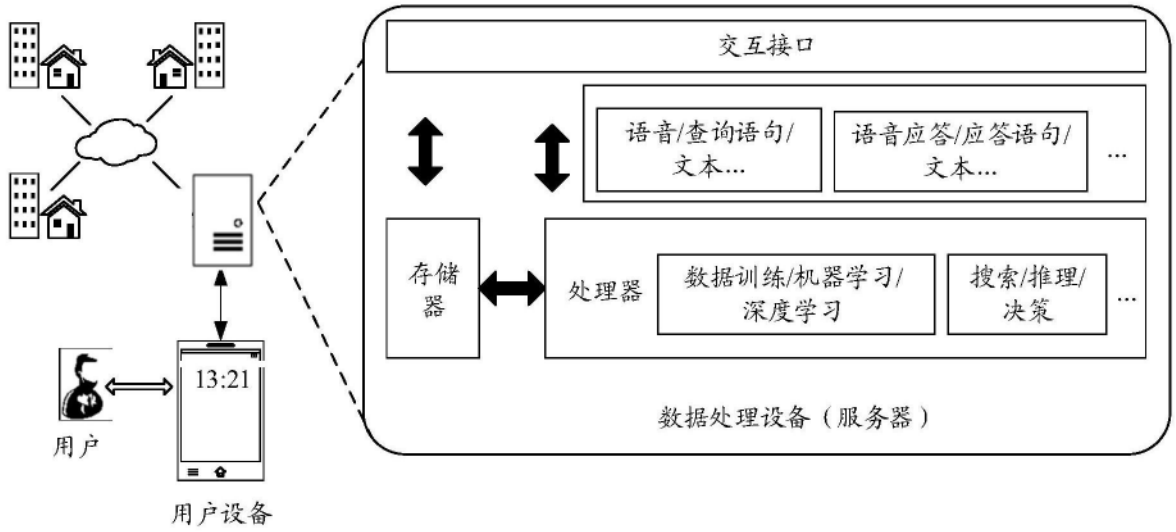

交互接口
语音/查询语句/文本...
语音应答/应答语句/文本...
...
存储器
处理器
数据训练/机器学习/深度学习
搜索/推理/决策
...
数据处理设备（服务器）
13:21
用户
用户设备

图1

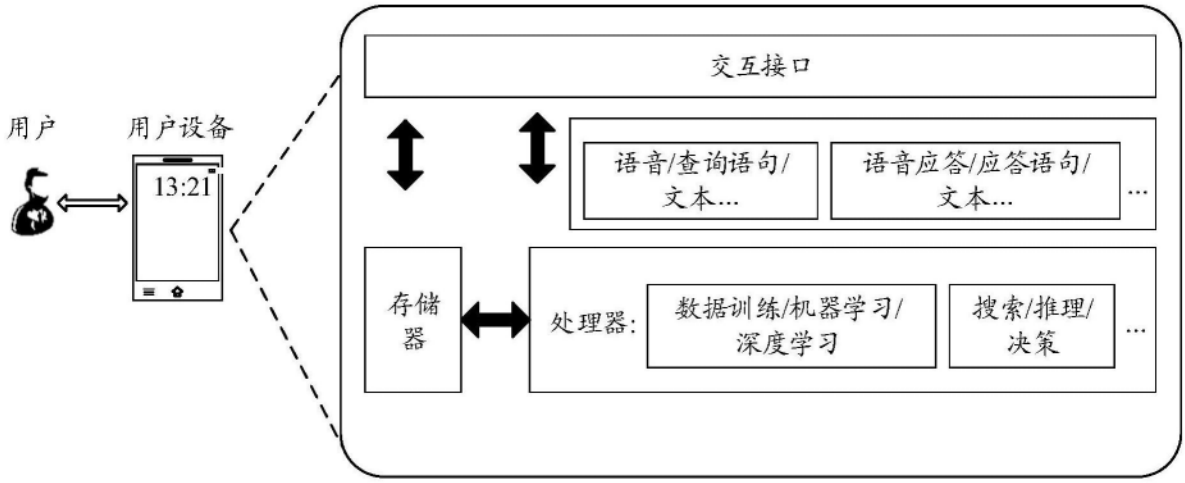

交互接口
语音/查询语句/文本...
语音应答/应答语句/文本...
...
存储器
处理器:
数据训练/机器学习/深度学习
搜索/推理/决策
...
用户
用户设备
13:21

图2

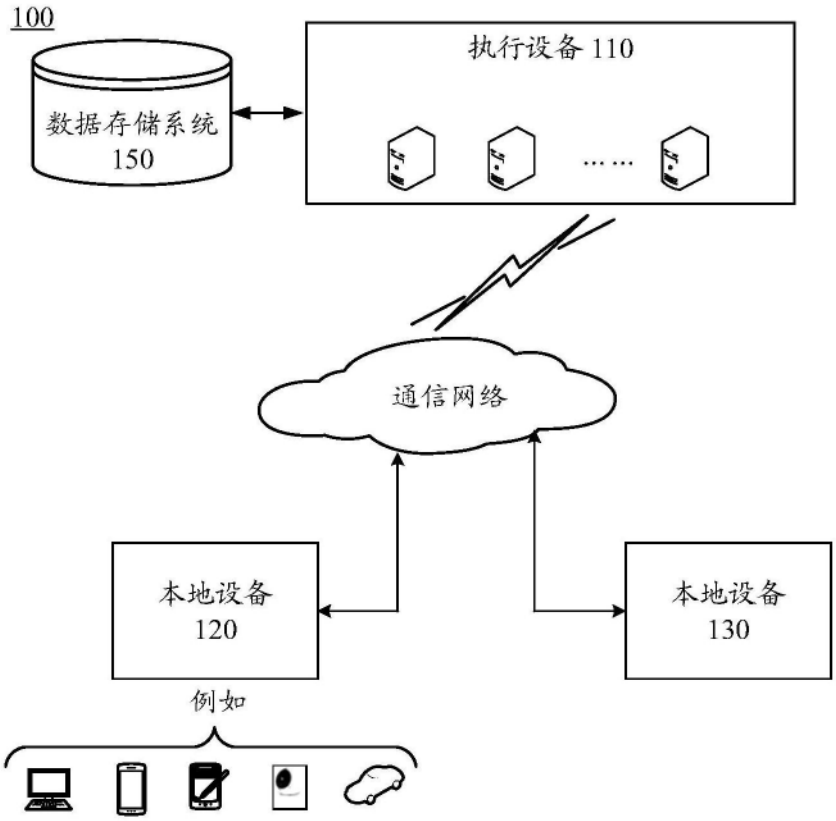


图3

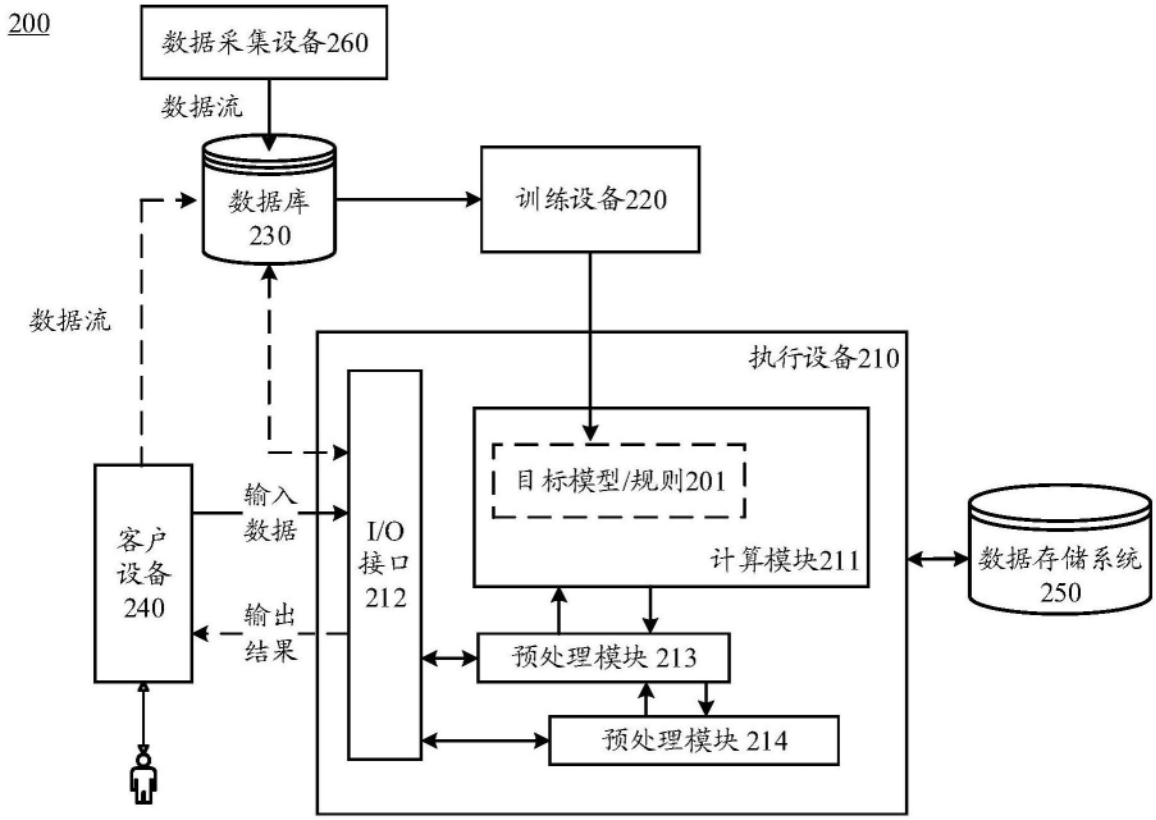

200
数据采集设备260
数据流
数据库
230
训练设备220
数据流
执行设备210
目标模型/规则201
计算模块211
数据存储系统
250
客户
设备
240
输入
数据
输出
结果
I/O
接口
212
预处理模块 213
预处理模块 214

图4

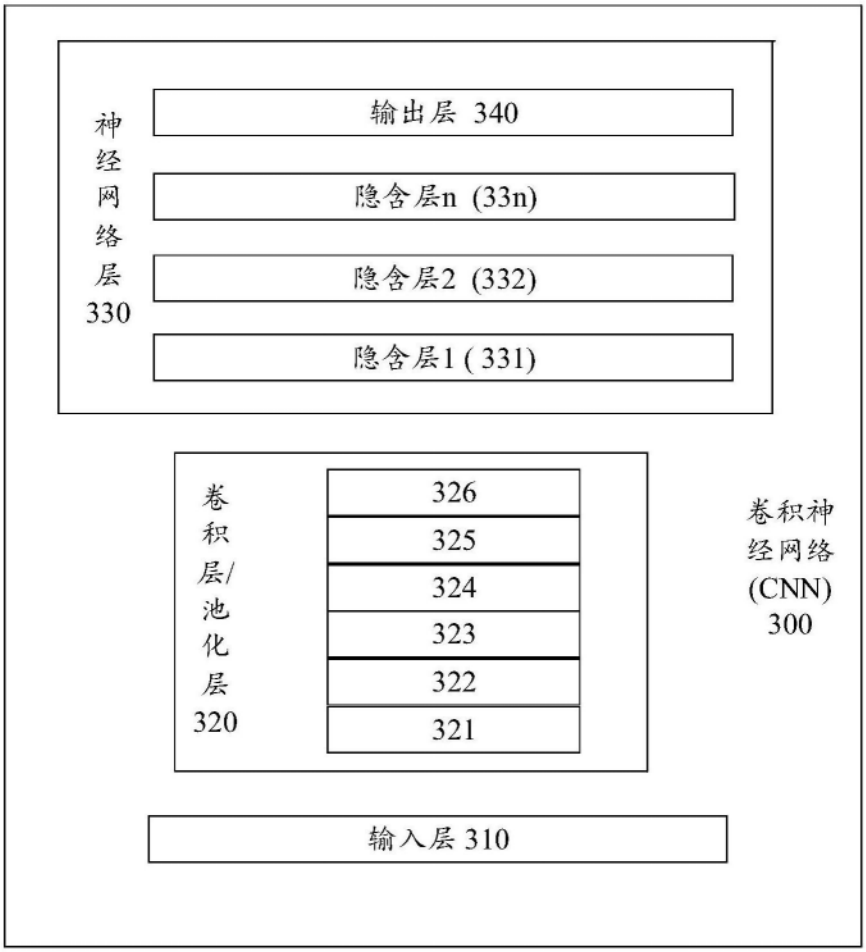

神经网络层330
输出层 340
隐含层n (33n)
隐含层2 (332)
隐含层1 (331)
卷积层/池化层320
326
325
324
323
322
321
卷积神经网络(CNN)300
输入层 310

图5

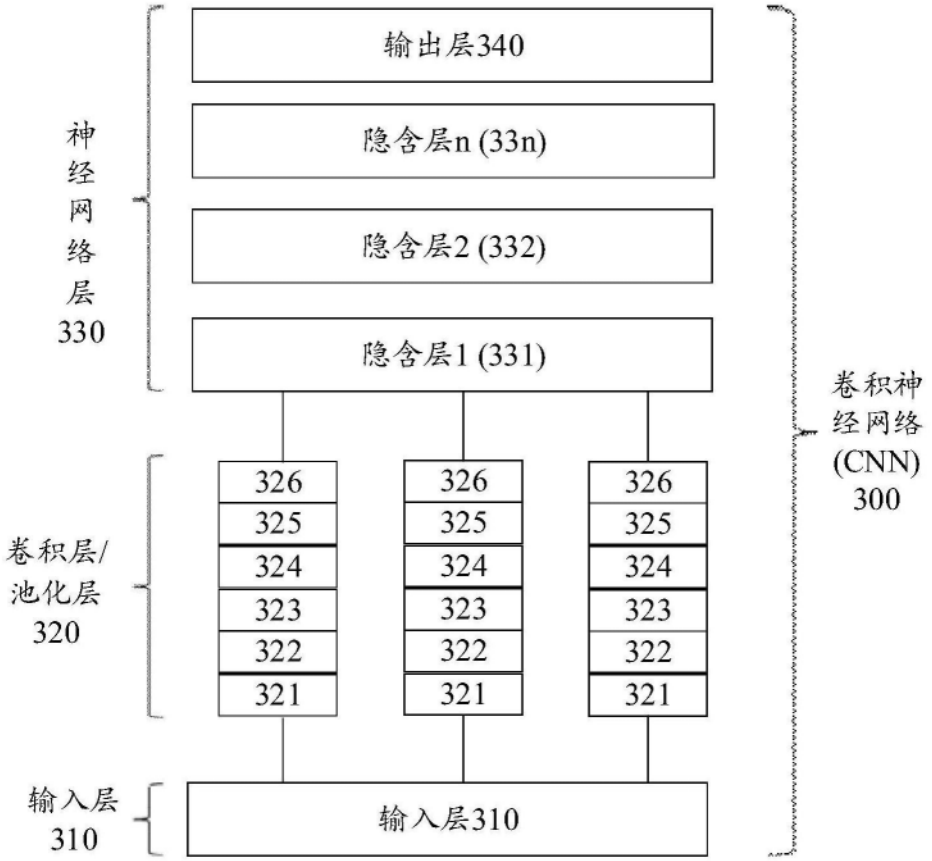

输出层340
隐含层n (33n)
隐含层2 (332)
隐含层1 (331)
神经网络层330
卷积层/池化层320
326
325
324
323
322
321
输入层310
卷积神经网络(CNN)300

图6

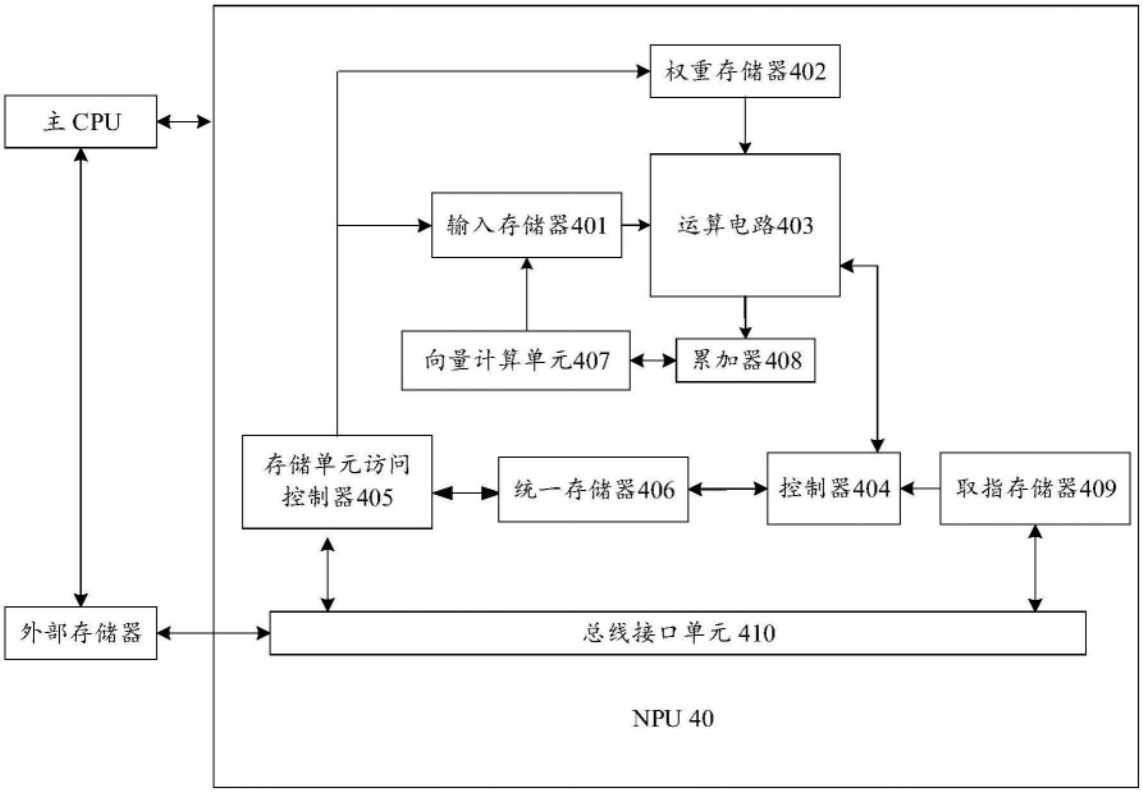

主 CPU
权重存储器402
输入存储器401
运算电路403
向量计算单元407
累加器408
存储单元访问控制器405
统一存储器406
控制器404
取指存储器409
外部存储器
总线接口单元 410
NPU 40

图7

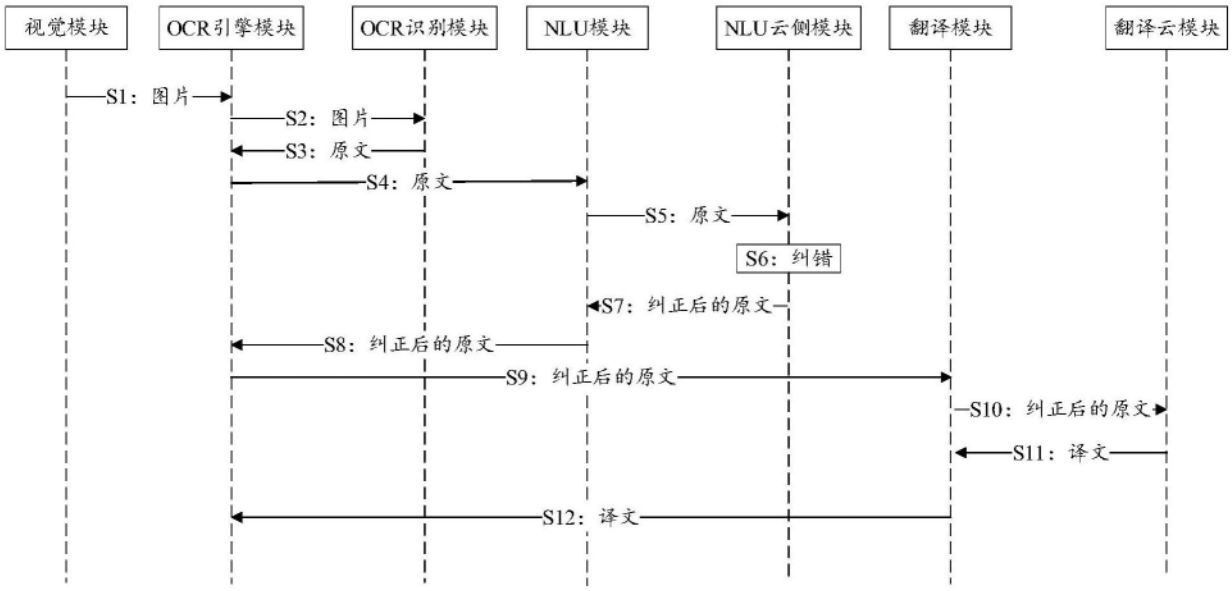

视觉模块
OCR引擎模块
OCR识别模块
NLU模块
NLU云侧模块
翻译模块
翻译云模块
S1：图片
S2：图片
S3：原文
S4：原文
S5：原文
S6：纠错
S7：纠正后的原文
S8：纠正后的原文
S9：纠正后的原文
S10：纠正后的原文
S11：译文
S12：译文

图8

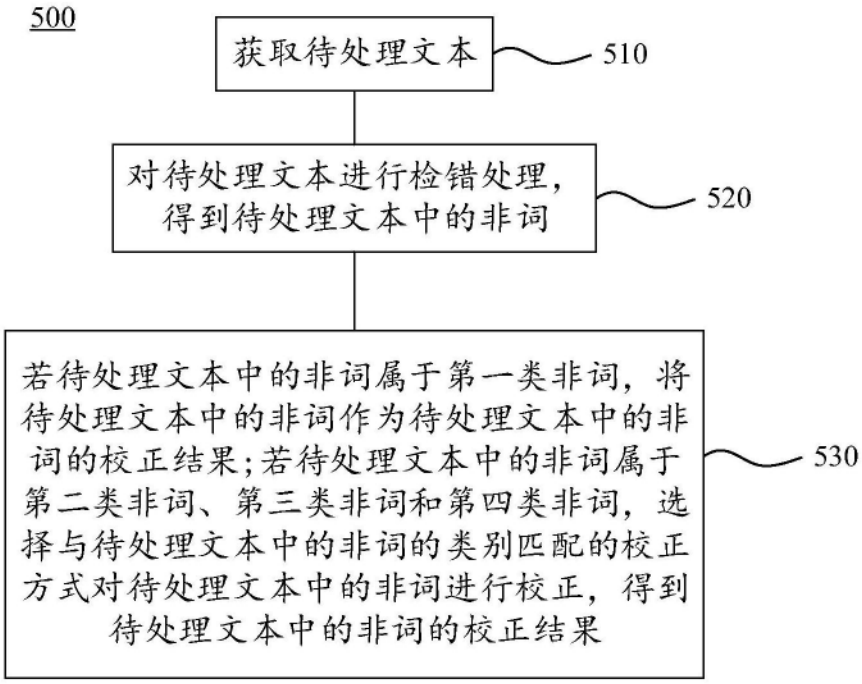

500
获取待处理文本
510
对待处理文本进行检错处理，得到待处理文本中的非词
520
若待处理文本中的非词属于第一类非词，将待处理文本中的非词作为待处理文本中的非词的校正结果；若待处理文本中的非词属于第二类非词、第三类非词和第四类非词，选择与待处理文本中的非词的类别匹配的校正方式对待处理文本中的非词进行校正，得到待处理文本中的非词的校正结果
530

图9

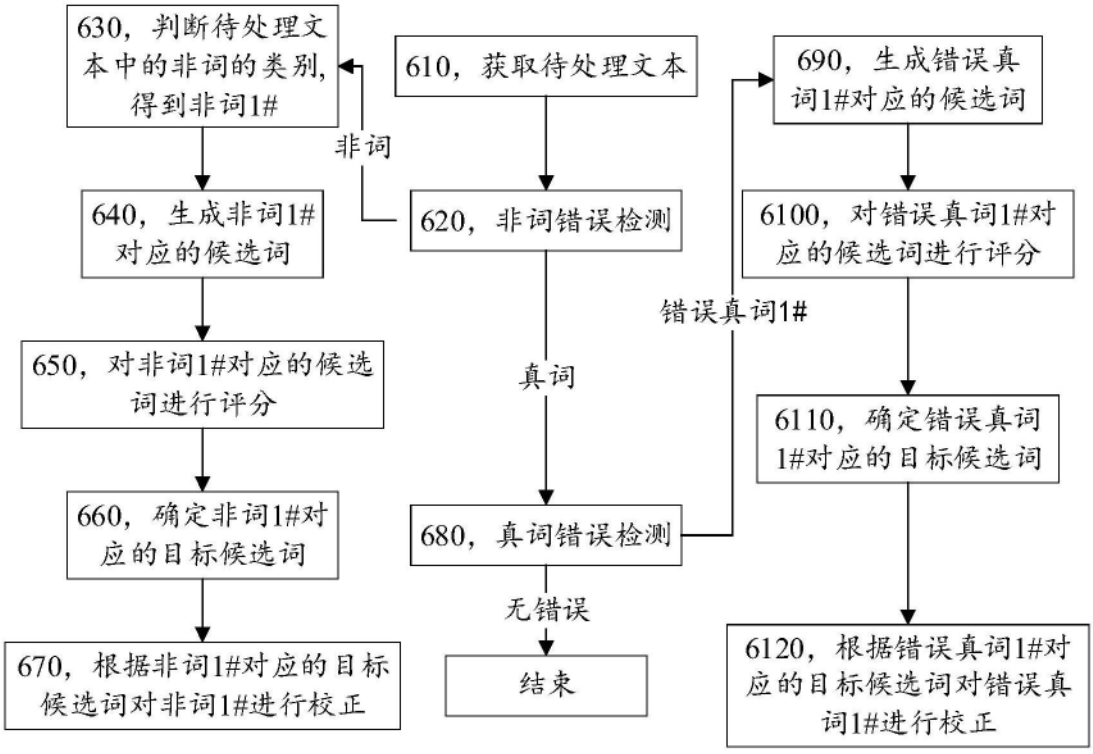

630，判断待处理文本中的非词的类别，得到非词1#
610，获取待处理文本
690，生成错误真词1#对应的候选词
非词
640，生成非词1#对应的候选词
620，非词错误检测
6100，对错误真词1#对应的候选词进行评分
错误真词1#
650，对非词1#对应的候选词进行评分
真词
6110，确定错误真词1#对应的目标候选词
660，确定非词1#对应的目标候选词
680，真词错误检测
无错误
670，根据非词1#对应的目标候选词对非词1#进行校正
结束
6120，根据错误真词1#对应的目标候选词对错误真词1#进行校正

图10

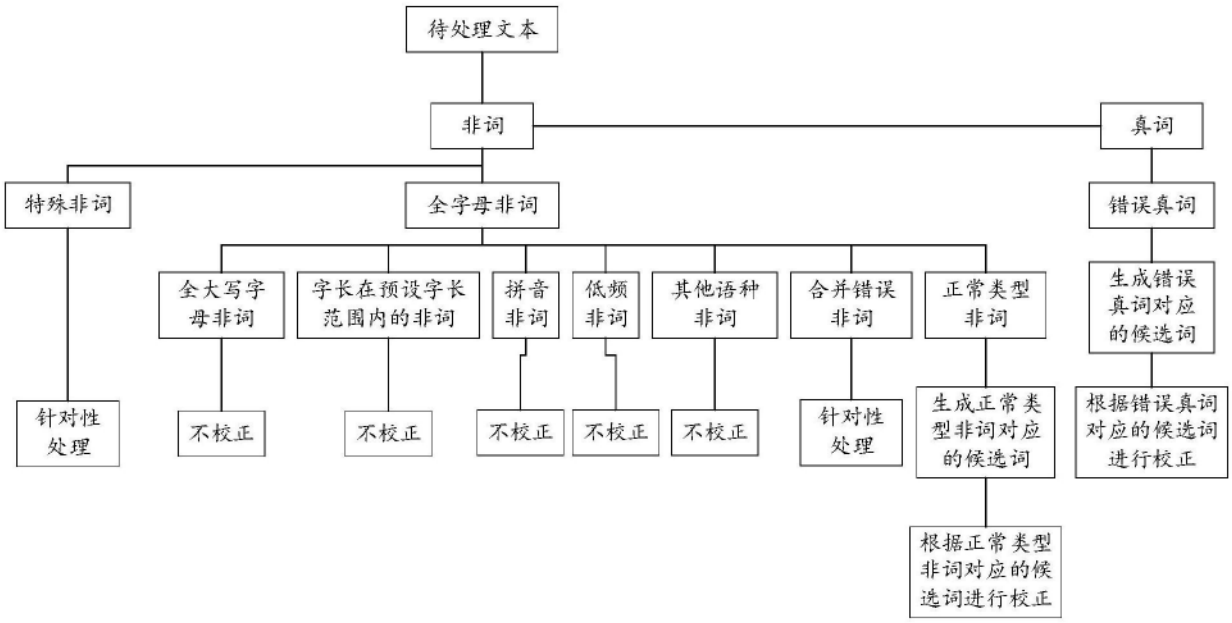

待处理文本
非词
真词
特殊非词
全字母非词
错误真词
全大写字母非词
字长在预设字长范围内的非词
拼音非词
低频非词
其他语种非词
合并错误非词
正常类型非词
生成错误真词对应的候选词
针对性处理
不校正
不校正
不校正
不校正
不校正
针对性处理
生成正常类型非词对应的候选词
根据错误真词对应的候选词进行校正
根据正常类型非词对应的候选词进行校正

图11

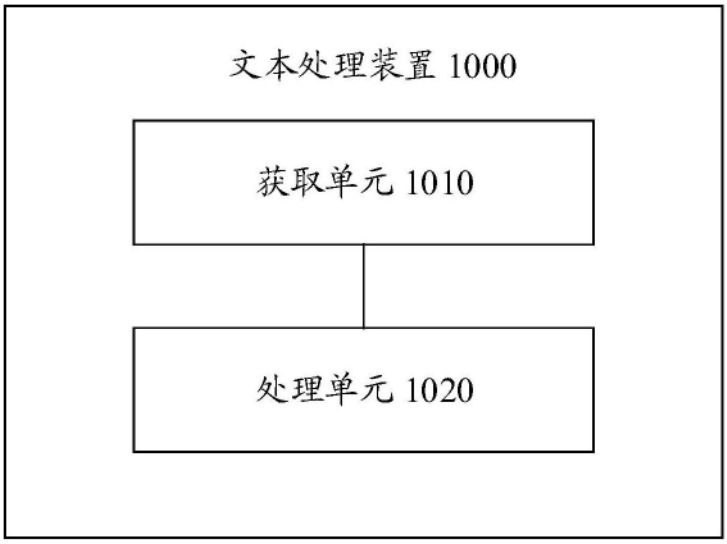

文本处理装置 1000
获取单元 1010
处理单元 1020

图12

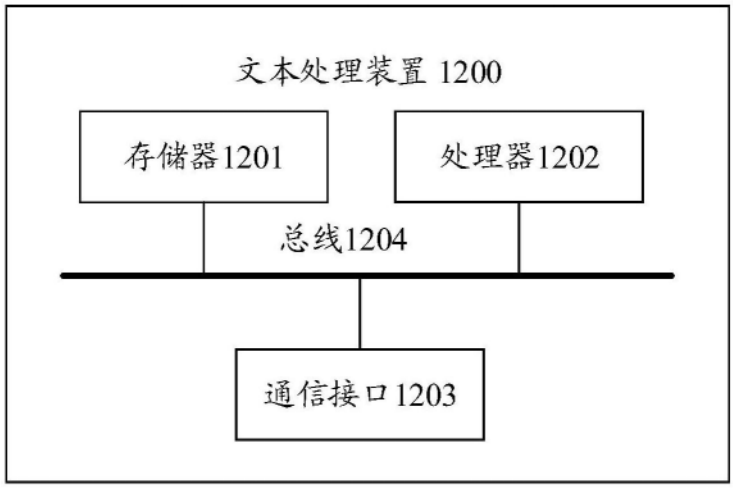


图13