(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 113962215 A
(43)申请公布日 2022.01.21

(21)申请号 202111215901.9

(22)申请日 2021.10.19

(71)申请人 平安普惠企业管理有限公司
地址 518000 广东省深圳市前海深港合作区前湾一路1号A栋201室(入驻深圳市前海商务秘书有限公司)

(72)发明人 郭丹丹

(74)专利代理机构 北京鸿元知识产权代理有限公司 11327
代理人 王迎 袁文婷

(51)Int.Cl.
G06F 40/232(2020.01)
G06F 40/237(2020.01)
G06F 40/289(2020.01)

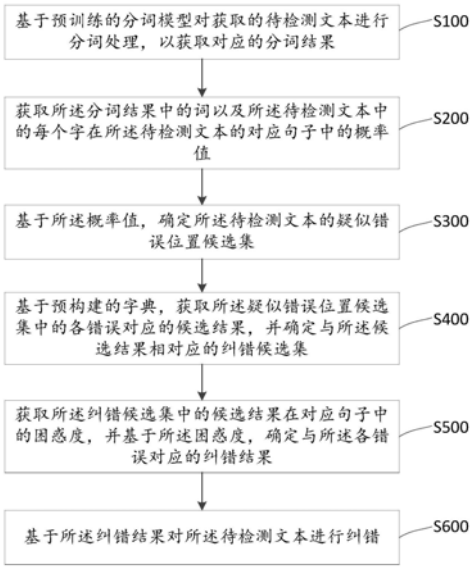
权利要求书2页 说明书14页 附图2页

(54)发明名称

基于人工智能的文本纠错方法、装置、设备及存储介质

(57)摘要

本发明涉及人工智能技术领域，揭露一种基于人工智能的文本纠错方法，包括：基于预训练的分词模型对获取的待检测文本进行分词处理，以获取对应的分词结果；获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值；基于所述概率值，确定所述待检测文本的疑似错误位置候选集；基于预构建的字典，获取所述疑似错误位置候选集中的各错误对应的候选结果，并确定与所述候选结果相对应的纠错候选集；获取所述纠错候选集中的候选结果在对应句子中的困惑度，并基于所述困惑度，确定与所述各错误对应的纠错结果；基于所述纠错结果对所述待检测文本进行纠错。通过本发明可以提高文本纠错的准确度。



CN 113962215 A

1.一种基于人工智能的文本纠错方法,其特征在于,所述方法包括:

基于预训练的分词模型对获取的待检测文本进行分词处理,以获取对应的分词结果;

获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值;

基于所述概率值,确定所述待检测文本的疑似错误位置候选集;

基于预构建的字典,获取所述疑似错误位置候选集中的各错误对应的候选结果,并确定与所述候选结果相对应的纠错候选集;

获取所述纠错候选集中的候选结果在对应句子中的困惑度,并基于所述困惑度,确定与所述各错误对应的纠错结果;

基于所述纠错结果对所述待检测文本进行纠错。

2.如权利要求1所述的基于人工智能的文本纠错方法,其特征在于,所述基于预训练的分词模型对获取的待检测文本进行分词处理,以获取对应的分词结果的步骤包括:

获取训练集语料库,并基于所述训练集语料库对初始化的N-gram模型进行训练,以获取训练完成的分词模型;

基于所述分词模型对所述待检测文本进行一次分词处理,并获取对应的第一分词结果;

基于前向最大匹配分词法,对所述第一分词结果进行二次分词处理,获取对应的第二分词结果;以及,基于后向最大匹配分词法,对所述第一分词结果进行二次分词处理,获取对应的第三分词结果;

基于预设规则,从所述第二分词结果和所述第三分词结果中选取目标文本作为所述分词结果。

3.如权利要求1所述的基于人工智能的文本纠错方法,其特征在于,所述获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值的步骤包括:

获取待检测文本中的每个字,确定对应的字集合;

对所述分词结果和所述字集合进行并集处理,以确定目标集合;

获取所述目标集合中的所有元素在对应句子中的概率值。

4.如权利要求1所述的基于人工智能的文本纠错方法,其特征在于,所述预构建的字典包括模糊音字典和形似字字典,所述确定与所述候选结果相对应的纠错候选集的步骤包括:

将所述各错误处的字和/或词转换为目标拼音;

在所述模糊音字典中,查找与所述目标拼音相对应的模糊音或相似音,以形成第一候选结果;同时,

对所述目标拼音的声母和韵母进行拆分,以获取拆分后的目标声母和目标韵母;

在所述模糊音字典中,查找与所述目标声母和所述目标韵母对应的模糊音或相似音,以形成第二候选结果;

在所述形似字字典中,查找与所述各错误相对应的所有形似字,以形成第三候选结果;

基于所述第一候选结果、所述第二候选结果和所述第三候选结果,形成所述纠错候选集。

5.如权利要求1至4中任意一项所述的基于人工智能的文本纠错方法，其特征在于，在获取所述纠错候选集中的候选结果在对应句子中的困惑度之前，还包括：

基于预训练的筛选模型对所述纠错候候选集中的候选结果进行初步筛选，以确定目标候选集。

6.如权利要求5所述的基于人工智能的文本纠错方法，其特征在于，所述基于预训练的筛选模型对所述纠错候候选集中的候选结果进行初步筛选，以确定目标候选集的步骤包括：

基于获取的训练数据训练逻辑回归模型；

基于所述逻辑回归模型对所述纠错候选集中的结果进行预测，并获取对应的预测分值；

基于预设范围，过滤所述预测分值小于预设范围的候选结果，以确定所述目标候选集。

7.如权利要求1所述的基于人工智能的文本纠错方法，其特征在于，所述困惑度的获取公式为：

$$PP\left(S\right) = 2^{-\frac{1}{N}\sum \log\left(P\left(w_i\right)\right)}$$

其中，w表示所述候选结果中的字或词，i表示w在对应句子中的序号，s表示替换该候选结果后的句子，N表示句子中所有字或词的个数，p表示w在对应句子中的概率值。

8.一种基于人工智能的文本纠错装置，其特征在于，所述装置包括：

分词结果获取单元，用于基于预训练的分词模型对获取的待检测文本进行分词处理，以获取对应的分词结果；

概率值获取单元，用于获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值；

疑似错误位置候选集确定单元，用于基于所述概率值，确定所述待检测文本的疑似错误位置候选集；

纠错候选集确定单元，用于基于预构建的字典，获取所述疑似错误位置候选集中的各错误对应的候选结果，并确定与所述候选结果相对应的纠错候选集；

纠错结果确定单元，用于获取所述纠错候选集中的候选结果在对应句子中的困惑度，并基于所述困惑度，确定与所述各错误对应的纠错结果；

文本纠错单元，用于基于所述纠错结果对所述待检测文本进行纠错。

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指令，所述指令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中任一所述的基于人工智能的文本纠错方法中的步骤。

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中任一所述的基于人工智能的文本纠错方法中的步骤。

基于人工智能的文本纠错方法、装置、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，尤其涉及一种基于人工智能的文本纠错的方法、装置、电子设备及计算机可读存储介质。

背景技术

[0002] 目前，在中文文本中不可避免的会存在各种错误，例如，形近字、同音字、方言等导致的各类错误；例如，在传统的语音识别方案中往往会由于一些客观原因，使得识别结果不能够很好的表达客户的真实意图，例如，客户说话时带有方言口音或者受到外界环境噪音的影响等，均会导致识别出的文本存在错误。

[0003] 针对上述问题，需要对相应的文本进行错误检查及纠正，以提升意图理解的准确性，进而提高用户体验。

[0004] 现有的文本纠错方案，主要是通过基于规则的模型或基于统计的模型生成针对待纠正文本的多个候选文本，然后从多个候选文本中筛选出最合理的文本。然而在基于规则的模型或者基于统计的模型进行文本纠错过程中，文本纠错的准确率较低，文本纠错效果并不能兼顾多种错误形式，例如，模糊音、口语、方言等，进而不能满足现阶段用户对文本纠错功能的需求。

[0005] 因此，目前亟需一种文本纠错方法，能够兼顾多种类型的文本错误，达到高效、全面、准确的错误发现及纠正效果。

发明内容

[0006] 本发明提供一种基于人工智能的文本纠错方法、装置、电子设备及计算机可读存储介质，其主要目的在于提高文本纠错的效率及准确度。

[0007] 为实现上述目的，本发明提供的一种基于人工智能的文本纠错方法，包括：基于预训练的分词模型对获取的待检测文本进行分词处理，以获取对应的分词结果；

[0008] 获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值；

[0009] 基于所述概率值，确定所述待检测文本的疑似错误位置候选集；

[0010] 基于预构建的字典，获取所述疑似错误位置候选集中的各错误对应的候选结果，并确定与所述候选结果相对应的纠错候选集；

[0011] 获取所述纠错候选集中的候选结果在对应句子中的困惑度，并基于所述困惑度，确定与所述各错误对应的纠错结果；

[0012] 基于所述纠错结果对所述待检测文本进行纠错。

[0013] 此外，可选的技术方案是，所述基于预训练的分词模型对获取的待检测文本进行分词处理，以获取对应的分词结果的步骤包括：

[0014] 获取训练集语料库，并基于所述训练集语料库对初始化的N-gram模型进行训练，以获取训练完成的分词模型；

[0015] 基于所述分词模型对所述待检测文本进行一次分词处理，并获取对应的第一分词结果；

[0016] 基于前向最大匹配分词法，对所述第一分词结果进行二次分词处理，获取对应的第二分词结果；以及，基于后向最大匹配分词法，对所述第一分词结果进行二次分词处理，获取对应的第三分词结果；

[0017] 基于预设规则，从所述第二分词结果和所述第三分词结果中选取目标文本作为所述分词结果。

[0018] 此外，可选的技术方案是，所述获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值的步骤包括：

[0019] 获取待检测文本中的每个字，确定对应的字集合；

[0020] 对所述分词结果和所述字集合进行并集处理，以确定目标集合；

[0021] 获取所述目标集合中的所有元素在对应句子中的概率值。

[0022] 此外，可选的技术方案是，所述预构建的字典包括模糊音字典和形似字字典，所述确定与所述候选结果相对应的纠错候选集的步骤包括：

[0023] 将所述各错误处的字和/或词转换为目标拼音；

[0024] 在所述模糊音字典中，查找与所述目标拼音相对应的模糊音或相似音，以形成第一候选结果；同时，

[0025] 对所述目标拼音的声母和韵母进行拆分，以获取拆分后的目标声母和目标韵母；

[0026] 在所述模糊音字典中，查找与所述目标声母和所述目标韵母对应的模糊音或相似音，以形成第二候选结果；

[0027] 在所述形似字字典中，查找与所述各错误相对应的所有形似字，以形成第三候选结果；

[0028] 基于所述第一候选结果、所述第二候选结果和所述第三候选结果，形成所述纠错候选集。

[0029] 此外，可选的技术方案是，在获取所述纠错候选集中的候选结果在对应句子中的困惑度之前，还包括：

[0030] 基于预训练的筛选模型对所述纠错候候选集中的候选结果进行初步筛选，以确定目标候选集。

[0031] 此外，可选的技术方案是，所述基于预训练的筛选模型对所述纠错候候选集中的候选结果进行初步筛选，以确定目标候选集的步骤包括：

[0032] 基于获取的训练数据训练逻辑回归模型；

[0033] 基于所述逻辑回归模型对所述纠错候选集中的结果进行预测，并获取对应的预测分值；

[0034] 基于预设范围，过滤所述预测分值小于预设范围的候选结果，以确定所述目标候选集。

[0035] 此外，可选的技术方案是，所述困惑度的获取公式为：

[0036] $$PP\left(S\right) = 2^{-\frac{1}{N}\sum \log\left(P\left(w_i\right)\right)}$$

[0037] 其中，w表示所述候选结果中的字或词，i表示w在对应句子中的序号，s表示替换该

候选结果后的句子,N表示句子中所有字或词的个数,p表示w在对应句子中的概率值。

[0038] 为了解决上述问题,本发明还提供一种基于人工智能的文本纠错装置,所述装置包括:

[0039] 分词结果获取单元,用于基于预训练的分词模型对获取的待检测文本进行分词处理,以获取对应的分词结果;

[0040] 概率值获取单元,用于获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值;

[0041] 疑似错误位置候选集确定单元,用于基于所述概率值,确定所述待检测文本的疑似错误位置候选集;

[0042] 纠错候选集确定单元,用于基于预构建的字典,获取所述疑似错误位置候选集中的各错误对应的候选结果,并确定与所述候选结果相对应的纠错候选集;

[0043] 纠错结果确定单元,用于获取所述纠错候选集中的候选结果在对应句子中的困惑度,并基于所述困惑度,确定与所述各错误对应的纠错结果;

[0044] 文本纠错单元,用于基于所述纠错结果对所述待检测文本进行纠错。

[0045] 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:

[0046] 存储器,存储至少一个指令;及

[0047] 处理器,执行所述存储器中存储的指令以实现上述所述的基于人工智能的文本纠错方法。

[0048] 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指令,所述至少一个指令被电子设备中的处理器执行以实现上述所述的基于人工智能的文本纠错方法。

[0049] 本发明实施例通过确定所述待检测文本的疑似错误位置候选集;基于预构建的字典,获取疑似错误位置候选集中的各错误对应的候选结果,并确定与候选结果相对应的纠错候选集,然后获取纠错候选集中的候选结果在对应句子中的困惑度,并基于所述困惑度,确定与所述各错误对应的纠错结果,能够从字粒度和词粒度两个角度综合考虑,对文本错误进行检测定位,同时在错误纠正过程中可考虑方言、口音、形似字等多种因素的影响,提高错误纠正的准确性,进而提高后期的意图识别的准确度,可适用于多种类型的文本纠错过程。

附图说明

[0050] 图1为本发明一实施例提供的基于人工智能的文本纠错方法的流程示意图;

[0051] 图2为本发明一实施例提供的基于人工智能的文本纠错装置的模块示意图;

[0052] 图3为本发明一实施例提供的实现基于人工智能的文本纠错方法的电子设备的内部结构示意图;

[0053] 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

[0054] 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。

[0055] 为解决现有文本纠错存在的各种问题,本发明提供一种基于人工智能的文本纠错

方法，能够从字粒度和词粒度两个维度进行错误检测，能够兼顾方言、口音、模糊音和形似字等多方面的错误原因，提高对文本的错误检测准确度，进而提高意图识别的准确率和用户的体验效果，可适用于多种类型的文本纠错过程中。

[0056] 本发明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。

[0057] 人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、机器人技术、生物识别技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0058] 本发明提供一种基于人工智能的文本纠错方法。参照图1所示，为本发明一实施例提供的基于人工智能的文本纠错方法的流程示意图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

[0059] 在本实施例中，基于人工智能的文本纠错方法，主要包括以下步骤：

[0060] S100：基于预训练的分词模型对获取的待检测文本进行分词处理，以获取对应的分词结果。

[0061] 其中，该步骤S100可进一步包括：

[0062] S110：获取训练集语料库，并基于所述训练集语料库对初始化的N-gram模型进行训练，以获取训练完成的分词模型；

[0063] S120：基于所述分词模型对所述待检测文本进行一次分词处理，并获取对应的第一分词结果；

[0064] S130：基于前向最大匹配分词法，对所述第一分词结果进行二次分词处理，获取应的第二分词结果；以及，基于后向最大匹配分词法，对所述第一分词结果进行二次分词处理，获取对应的第三分词结果。

[0065] 具体地，第一分词结果可以理解为对待检测文本进行第一次分词处理后，获取的第一分词文本，然后对该第一分词文本进行二次分词处理，即可获取对应的第二分词结果和第三分词结果，即第二分词文本和第三分词文本。

[0066] 其中，前向最大匹配分词法主要包括：1、如果待分词的句子长度大于词表的最大词长度，则在当前句子的句首截取n个已经分好的词，直至n个词的总词长度大于或等于词表的最大词长度；2、如果这n个词合并成的词在词表中，则输出这个合并的词作为分词结果，否则查找前n-1个词或前n-1个词与第n个词的前k个字合并成的词是否在词表中，同时第n个词的剩余部分是否也在词表中且不为单字词，直到查找到符合条件的合并词为止；3、输出合并词，并把句子的剩余部分作为新的待分句子重复上述过程。

[0067] 同理，后向最大匹配方法和前向最大匹配方向相类似，主要是从句子的句末截取n个已经分好的词，然后进行逐步判断。

[0068] S140：基于预设规则，从所述第二分词结果和所述第三分词结果中选取目标文本作为所述分词结果。

[0069] 在该步骤中，预设规则可以包括：当第二分词结果以及第三分词结果中选取词长

期望不相同时，可从第二分词结果中以及所述第三分词结果中选取词长期望大的文本作为分词结果；或者，当第二分词结果以及第三分词结果中选取词长期望相同时，从第二分词结果以及第三分词结果中选取词长方差小的文本作为分词结果。

[0070] 通过步骤S100的分词处理后，会获取与待检测文本相对应的一个分词结果，该结果可结合待检测文本中的每一个字，以获取分词后的词和每个字在待检测文本的对应句子中的概率，并据此筛选出疑似错位的位置，进而通过词粒度和字粒度这两个角度进行错误检查，防止错误遗漏，提高后期的纠错准确度。

[0071] 需要说明的是，为了提高分词的准确度，在对待检测文本处理之前，还可以包括对待检测文本的预处理过程，例如该预处理过程可包括：对待检测文本中的特殊字符和表情符号进行过滤，然后在句子的句首和句尾添加标识符，如：[CLS]和[SEP]等，以便对句子进行标记和区分，便于后续的困惑度的计算。

[0072] S200：获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值。

[0073] 其中，所述获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值的步骤可进一步包括：

[0074] S210：获取待检测文本中的每个字，确定对应的字集合；

[0075] S220：对所述分词结果和所述字集合进行并集处理，以确定目标集合；

[0076] S230：获取所述目标集合中的所有元素在对应句子中的概率值。

[0077] 其中，目标集合中的所有元素主要是分词结果和字集合，即目标集合中为分词后的词或字，当确定了所有元素在对应句子中的概率值后，即可根据各元素的概率值来初步判断其对应的位置是否存在错误。

[0078] S300：基于所述概率值，确定所述待检测文本的疑似错误位置候选集。

[0079] 在上述两步骤中，可基于训练好的N-gram语言模型获取分词结果中的词以及待检测文本的每个字在对应句子中的概率值，当对应的概率值小于预设阈值时，可判断当前位置的词或字属于疑似错误的位置，根据待检测文本中所有疑似错误的位置，即可获取最终的疑似错误位置候选集。

[0080] 具体地，N-gram语言模型是一个基于概率的判断模型，它的输入可以是分词结果的词和每个字的顺序序列，输出的为对应的词和字的概率。假设句子T是有词序列或字列w1，w_2，w_3...w_n组成，则N-gram语言模型输出的联合概率可表示为：$P(T)=P(w_1)*p(w_2)*p(w_3)*\cdots*p(wn)=p(w_1)*p(w_2|w_1)*p(w_3|w_1w_2)*\cdots*p(wn|w_1w_2w_3...)$。可见对于句子T中每个词和字出现的条件概率，可以通过在预设语料库中统计计数的方式得出。

[0081] 作为具体示例，对于n元的N-gram语言模型，第wn个字或词的概率可表示为：$p(wn|w_1w_2w_3...)=C(w_{i-n-1},\cdots,w_i)/C(w_{i-n-1},\cdots,w_{i-1})$；上公式中$C(w_{i-n-1},\cdots,w_i)$表示字符串$w_{i-n-1},\cdots,w_i$在预设语料库中出现的次数或频率。此外，上述预设阈值可根据适用场景或要求进行设置，也可取现有的经验值，在应用过程中，可设置样本文本和预设阈值进行错误检查，然后结合获取的错误检查结果以及样本文本中的错误位置，对预设阈值进行调整，以确保错误检查的精准度。

[0082] S400：基于预构建的字典，获取所述疑似错误位置候选集中的各错误对应的候选结果，并确定与所述候选结果相对应的纠错候选集。

[0083] 其中，字典可包括模糊音字典和形似字字典；其中模糊音字典的构建过程可根据拼音和模糊音规则来完成，例如，可根据不同地区的方言口音的习惯，统计对应的模糊音或相似音，例如，n和l，b和f，an和ang，en和eng等均可认为是相近的模糊音；此外，所构建的形似字字典中主要包括统计的形似字，例如，“己”、“已”和“巳”可认为是形似字，“夕”和“歹”等，以便能够考虑到各方面可能出现的错误，提高错误的纠正准确度。

[0084] 作为具体示例，上述步骤S400可进一步包括以下内容：

[0085] S410：将所述各错误处的字和/或词转换为目标拼音；

[0086] 其中，由于目标集合中的元素包括字和词，因此，在对应的疑似错误位置候选集中的错误位置可能是字也可能是词；待确定了错误位置后，可将该位置的字和/或词转换为拼音格式，以确定对应的目标拼音。

[0087] 此外，也可将错误位置所在的整个句子转换为拼音形式，然后基于预构建的字典对拼音形式的句子中的目标拼音(错误位置)进行模糊音或形似字的查找，后续可根据该查询结果进行替换处理，以确定最终的候选集。

[0088] S420：在所述模糊音字典中，查找与所述目标拼音相对应的模糊音或相似音，以形成第一候选结果；同时，

[0089] S430：对所述目标拼音的声母和韵母进行拆分，获取拆分后的目标声母和目标韵母；

[0090] S440：在所述模糊音字典中，查找与所述目标声母和所述目标韵母对应的模糊音或相似音，以形成第二候选结果；

[0091] 此外，还包括可以与上述各步骤可同时执行的步骤S450：在所述形似字字典中，查找与所述各错误相对应的所有形似字，以形成第三候选结果；

[0092] S460：基于所述第一候选结果、所述第二候选结果和所述第三候选结果，形成所述纠错候选集。

[0093] 其中，第一候选结果、第二候选结果和第三候选结果分别表示从不同角度和出发点考虑而获得的与各错误相对应的多个可能性结果，在这些结果中肯定会存在对错误的正确纠正结果，进而需要对纠错候选集中的多个可能性进行逐个判断筛选，直至确定最终的纠错结果。

[0094] 换言之，对上述三个候选结果取并集即可形成最终的纠错候选集，然后对该纠错候选集中的各结果再进行验证，即可筛选出最佳的纠错结果，并据此对对应的错误位置进行纠错处理。

[0095] S500：获取所述纠错候选集中的候选结果在对应句子中的困惑度，并基于所述困惑度，确定与所述各错误对应的纠错结果。

[0096] 此外，在执行步骤S500之前，还包括：基于预训练的筛选模型对纠错候选集中的候选结果进行初步筛选，确定目标候选集，然后获取所述目标候选集中的候选结果在对应句子中的困惑度。

[0097] 具体地，基于预训练的筛选模型对纠错候选集中的候选结果进行初步筛选，确定目标候选集的过程可进一步包括：

[0098] S510：基于获取的训练数据训练逻辑回归模型；

[0099] S520：基于所述逻辑回归模型对所述纠错候选集中的结果进行预测，并获取对应

的预测分值；

[0100] S530:基于预设范围,过滤所述预测分值小于预设范围的候选结果,以确定所述目标候选集。

[0101] 在该过程中,主要是通过逻辑回归模型对纠错候选集中的明显错误进行删除,选择预测分值较高的候选结果,以减小后续计算量的压力,完成对纠错候选集的初步筛选。

[0102] 进一步地,将目标候选集中的各候选结果替换至对应的句子中,并获取对应的替换结果后的句子的困惑度,然后选择困惑度最小的候选结果作为最终的纠错结果,并根据纠错结果对对应的错位位置进行替换,获取纠错后的正确文本。

[0103] 具体地,获取困惑度的计算公式可表示为:

[0104] $$PP\left(S\right) = 2^{-\frac{1}{N}\sum \log\left(P\left(w_i\right)\right)}$$

[0105] 其中,w表示所述候选结果中的字或词,i表示w在对应句子中的序号,s表示替换该候选结果后的句子,N表示句子中所有字或词的个数,p表示w在对应句子中的概率值。

[0106] 通过上述公式即可获取出现错误的句子,在每个候选结果替换下的困惑度,然后基于该困惑度选取困惑度值最小的候选结果作为最终的纠错结果。

[0107] S600:基于所述纠错结果对所述待检测文本进行纠错。

[0108] 通过以上步骤,对于每一个错误位置均可获取对应的一个纠错结果,根据所有的纠错结果对对应位置的错误进行替换,即可完成对待检测文本的错误检测及纠正过程,并形成纠正后的文本,以便后续的意图识别等操作。

[0109] 根据上述基于人工智能的文本纠错方法,能够从字粒度和词粒度两个角度综合考虑,对文本错误进行检测定位,同时在错误纠正过程中可考虑方言、口音、形似字等多种因素的影响,提高错误纠正的准确性,进而提高后期的意图识别的准确度,可适用于多种类型的文本纠错过程。

[0110] 如图2所示,是本发明基于人工智能的文本纠错装置的功能模块图。

[0111] 本发明所述基于人工智能的文本纠错装置200可以安装于电子设备中。根据实现的功能,所述基于人工智能的文本纠错装置可以包括:分词结果获取单元210、概率值获取单元220、疑似错误位置候选集确定单元230、纠错候选集确定单元240、纠错结果确定单元250和文本纠错单元260。本发所述单元也可以称之为模块,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。

[0112] 在本实施例中,关于各模块/单元的功能如下：

[0113] 分词结果获取单元210,用于基于预训练的分词模型对获取的待检测文本进行分词处理,以获取对应的分词结果。

[0114] 其中,该单元210可进一步包括：

[0115] 分词模型训练模块,用于获取训练集语料库,并基于所述训练集语料库对初始化的N-gram模型进行训练,以获取训练完成的分词模型；

[0116] 第一分词结果获取模块,用于基于所述分词模型对所述待检测文本进行一次分词处理,并获取对应的第一分词结果；

[0117] 第二和第三分词结果获取模块,用于基于前向最大匹配分词法,对所述第一分词结果进行二次分词处理,获取对应的第二分词结果;以及,基于后向最大匹配分词法,对所

述第一分词结果进行二次分词处理，获取对应的第三分词结果。

[0118] 具体地，第一分词结果可以理解为对待检测文本进行第一次分词处理后，获取的第一分词文本，然后对该第一分词文本进行二次分词处理，即可获取对应的第二分词结果和第三分词结果，即第二分词文本和第三分词文本。

[0119] 其中，前向最大匹配分词法主要包括：1、如果待分词的句子长度大于词表的最大词长度，则在当前句子的句首截取n个已经分好的词，直至n个词的总词长度大于或等于词表的最大词长度；2、如果这n个词合并成的词在词表中，则输出这个合并的词作为分词结果，否则查找前n-1个词或前n-1个词与第n个词的前k个字合并成的词是否在词表中，同时第n个词的剩余部分是否也在词表中且不为单字词，直到查找到符合条件的合并词为止；3、输出合并词，并把句子的剩余部分作为新的待分句子重复上述过程。

[0120] 同理，后向最大匹配方法和前向最大匹配方向相类似，主要是从句子的句末截取n个已经分好的词，然后进行逐步判断。

[0121] 分词结果获取模块，用于基于预设规则，从所述第二分词结果和所述第三分词结果中选取目标文本作为所述分词结果。

[0122] 在该模块中，预设规则可以包括：当第二分词结果以及第三分词结果中选取词长期望不相同时，可从第二分词结果中以及所述第三分词结果中选取词长期望大的文本作为分词结果；或者，当第二分词结果以及第三分词结果中选取词长期望相同时，从第二分词结果以及第三分词结果中选取词长方差小的文本作为分词结果。

[0123] 通过分词结果获取单元210的分词处理后，会获取与待检测文本相对应的一个分词结果，该结果可结合待检测文本中的每一个字，以获取分词后的词和每个字在待检测文本的对应句子中的概率，并据此筛选出疑似错位的位置，进而通过词粒度和字粒度这两个角度进行错误检查，防止错误遗漏，提高后期的纠错准确度。

[0124] 需要说明的是，为了提高分词的准确度，在对待检测文本处理之前，还可以包括对待检测文本的预处理过程，例如该预处理过程可包括：对待检测文本中的特殊字符和表情符号进行过滤，然后在句子的句首和句尾添加标识符，如：[CLS]和[SEP]等，以便对句子进行标记和区分，便于后续的困惑度的计算。

[0125] 概率值获取单元220，用于获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值。

[0126] 其中，所述获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值的单元可进一步包括：

[0127] 字集合确定模块，用于获取待检测文本中的每个字，确定对应的字集合；

[0128] 目标集合确定模块，用于对所述分词结果和所述字集合进行并集处理，以确定目标集合；

[0129] 概率值获取模块，用于获取所述目标集合中的所有元素在对应句子中的概率值。

[0130] 其中，目标集合中的所有元素主要是分词结果和字集合，即目标集合中为分词后的词或字，当确定了所有元素在对应句子中的概率值后，即可根据各元素的概率值来初步判断其对应的位置是否存在错误。

[0131] 疑似错误位置候选集确定单元230，用于基于所述概率值，确定所述待检测文本的疑似错误位置候选集。

[0132] 在上述两单元中，可基于训练好的N-gram语言模型获取分词结果中的词以及待检测文本的每个字在对应句子中的概率值，当对应的概率值小于预设阈值时，可判断当前位置的词或字属于疑似错误的位置，根据待检测文本中所有疑似错误的位置，即可获取最终的疑似错误位置候选集。

[0133] 具体地，N-gram语言模型是一个基于概率的判断模型，它的输入可以是分词结果的词和每个字的顺序序列，输出的为对应的词和字的概率。假设句子T是有词序列或字列w1，w_2，w_3...w_n组成，则N-gram语言模型输出的联合概率可表示为：$P(T)=P(w_1)*p(w_2)*p(w_3)*\cdots*p(wn)=p(w_1)*p(w_2|w_1)*p(w_3|w_1w_2)*\cdots*p(wn|w_1w_2w_3...)$。可见对于句子T中每个词和字出现的条件概率，可以通过在预设语料库中统计计数的方式得出。

[0134] 作为具体示例，对于n元的N-gram语言模型，第wn个字或词的概率可表示为：$p(wn|w_1w_2w_3...)=C(w_{i-n-1},\cdots,w_i)/C(w_{i-n-1},\cdots,w_{i-1})$；上公式中$C(w_{i-n-1},\cdots,w_i)$表示字符串$w_{i-n-1},\cdots,w_i$在预设语料库中出现的次数或频率。此外，上述预设阈值可根据适用场景或要求进行设置，也可取现有的经验值，在应用过程中，可设置样本文本和预设阈值进行错误检查，然后结合获取的错误检查结果以及样本文本中的错误位置，对预设阈值进行调整，以确保错误检查的精准度。

[0135] 纠错候选集确定单元240，用于基于预构建的字典，获取所述疑似错误位置候选集中的各错误对应的候选结果，并确定与所述候选结果相对应的纠错候选集。

[0136] 其中，字典可包括模糊音字典和形似字字典；其中模糊音字典的构建过程可根据拼音和模糊音规则来完成，例如，可根据不同地区的方言口音的习惯，统计对应的模糊音或相似音，例如，n和l，b和f，an和ang，en和eng等均可认为是相近的模糊音；此外，所构建的形似字字典中主要包括统计的形似字，例如，“己”、“已”和“巳”可认为是形似字，“夕”和“歹”等，以便能够考虑到各方面可能出现的错误，提高错误的纠正准确度。

[0137] 作为具体示例，上述纠错候选集确定单元240可进一步包括以下内容：

[0138] 目标拼音转换模块，用于将所述各错误处的字和/或词转换为目标拼音；

[0139] 其中，由于目标集合中的元素包括字和词，因此，在对应的疑似错误位置候选集中的错误位置可能是字也可能是词；待确定了错误位置后，可将该位置的字和/或词转换为拼音格式，以确定对应的目标拼音。

[0140] 此外，也可将错误位置所在的整个句子转换为拼音形式，然后基于预构建的字典对拼音形式的句子中的目标拼音(错误位置)进行模糊音或形似字的查找，后续可根据该查询结果进行替换处理，以确定最终的候选集。

[0141] 第一候选结果形成模块，用于在所述模糊音字典中，查找与所述目标拼音相对应的模糊音或相似音，以形成第一候选结果；同时，

[0142] 拼音拆分模块，用于对所述目标拼音的声母和韵母进行拆分，获取拆分后的目标声母和目标韵母；

[0143] 第二候选结果形成模块，用于在所述模糊音字典中，查找与所述目标声母和所述目标韵母对应的模糊音或相似音，以形成第二候选结果；

[0144] 此外，还包括可以与上述各步骤可同时执行的步骤S450：在所述形似字字典中，查找与所述各错误相对应的所有形似字，以形成第三候选结果；

[0145] 纠错候选集形成模块，用于基于所述第一候选结果、所述第二候选结果和所述第

三候选结果，形成所述纠错候选集。

[0146] 其中，第一候选结果、第二候选结果和第三候选结果分别表示从不同角度和出发点考虑而获得的与各错误相对应的多个可能性结果，在这些结果中肯定会存在对错误的正确纠正结果，进而需要对纠错候选集中的多个可能性进行逐个判断筛选，直至确定最终的纠错结果。

[0147] 换言之，对上述三个候选结果取并集即可形成最终的纠错候选集，然后对该纠错候选集中的各结果再进行验证，即可筛选出最佳的纠错结果，并据此对对应的错误位置进行纠错处理。

[0148] 纠错结果确定单元250，用于获取所述纠错候选集中的候选结果在对应句子中的困惑度，并基于所述困惑度，确定与所述各错误对应的纠错结果。

[0149] 其中，在执行该单元之前还包括：基于预训练的筛选模型对纠错候选集中的候选结果进行初步筛选，确定目标候选集，然后获取所述目标候选集中的候选结果在对应句子中的困惑度。

[0150] 具体地，基于预训练的筛选模型对纠错候选集中的候选结果进行初步筛选，确定目标候选集的过程可进一步包括：

[0151] 1、基于获取的训练数据训练逻辑回归模型；

[0152] 2、基于所述逻辑回归模型对所述纠错候选集中的结果进行预测，并获取对应的预测分值；

[0153] 3、基于预设范围，过滤所述预测分值小于预设范围的候选结果，以确定所述目标候选集。

[0154] 在该过程中，主要是通过逻辑回归模型对纠错候选集中的明显错误进行删除，选择预测分值较高的候选结果，以减小后续计算量的压力，完成对纠错候选集的初步筛选。

[0155] 进一步地，将目标候选集中的各候选结果替换至对应的句子中，并获取对应的替换结果后的句子的困惑度，然后选择困惑度最小的候选结果作为最终的纠错结果，并根据纠错结果对对应的错位位置进行替换，获取纠错后的正确文本。

[0156] 具体地，获取困惑度的计算公式可表示为：

[0157] $$PP\left(S\right) = 2^{-\frac{1}{N}\sum \log\left(P\left(w_i\right)\right)}$$

[0158] 其中，w表示所述候选结果中的字或词，i表示w在对应句子中的序号，s表示替换该候选结果后的句子，N表示句子中所有字或词的个数，p表示w在对应句子中的概率值。

[0159] 通过上述公式即可获取出现错误的句子，在每个候选结果替换下的困惑度，然后基于该困惑度选取困惑度值最小的候选结果作为最终的纠错结果。

[0160] 文本纠错单元260，用于基于所述纠错结果对所述待检测文本进行纠错。

[0161] 通过以上各单元的处理后，对于每一个错误位置均可获取对应的一个纠错结果，根据所有的纠错结果对对应位置的错误进行替换，即可完成对待检测文本的错误检测及纠正过程，并形成纠正后的文本，以便后续的意图识别等操作。

[0162] 如图3所示，是本发明实现基于人工智能的文本纠错方法的电子设备的结构示意图。

[0163] 所述电子设备1可以包括处理器10、存储器11和总线，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，如基于人工智能的文本纠错程序12。

[0164] 其中，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1的内部存储单元，例如该电子设备1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1的外部存储设备，例如电子设备1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1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1的应用软件及各类数据，例如基于人工智能的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。

[0165] 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(例如基于人工智能的文本纠错程序等)，以及调用存储在所述存储器11内的数据，以执行电子设备1的各种功能和处理数据。

[0166] 所述总线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。

[0167] 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。

[0168] 例如，尽管未示出，所述电子设备1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1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。

[0169] 进一步地，所述电子设备1还可以包括网络接口，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1与其他电子设备之间建立通信连接。

[0170] 可选地，该电子设备1还可以包括用户接口，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1中处理的信息以及用于显示可视化的用户界面。

[0171] 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。

[0172] 所述电子设备1中的所述存储器11存储的基于人工智能的文本纠错程序12是多个指令的组合，在所述处理器10中运行时，可以实现：

[0173] 基于预训练的分词模型对获取的待检测文本进行分词处理，以获取对应的分词结果；

[0174] 获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值；

[0175] 基于所述概率值，确定所述待检测文本的疑似错误位置候选集；

[0176] 基于预构建的字典，获取所述疑似错误位置候选集中的各错误对应的候选结果，并确定与所述候选结果相对应的纠错候选集；

[0177] 获取所述纠错候选集中的候选结果在对应句子中的困惑度，并基于所述困惑度，确定与所述各错误对应的纠错结果；

[0178] 基于所述纠错结果对所述待检测文本进行纠错。

[0179] 此外，可选的技术方案是，所述基于预训练的分词模型对获取的待检测文本进行分词处理，以获取对应的分词结果的步骤包括：

[0180] 获取训练集语料库，并基于所述训练集语料库对初始化的N-gram模型进行训练，以获取训练完成的分词模型；

[0181] 基于所述分词模型对所述待检测文本进行一次分词处理，并获取对应的第一分词结果；

[0182] 基于前向最大匹配分词法，对所述第一分词结果进行二次分词处理，获取对应的第二分词结果；以及，基于后向最大匹配分词法，对所述第一分词结果进行二次分词处理，获取对应的第三分词结果；

[0183] 基于预设规则，从所述第二分词结果和所述第三分词结果中选取目标文本作为所述分词结果。

[0184] 此外，可选的技术方案是，所述获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值的步骤包括：

[0185] 获取待检测文本中的每个字，确定对应的字集合；

[0186] 对所述分词结果和所述字集合进行并集处理，以确定目标集合；

[0187] 获取所述目标集合中的所有元素在对应句子中的概率值。

[0188] 此外，可选的技术方案是，所述预构建的字典包括模糊音字典和形似字字典，所述确定与所述候选结果相对应的纠错候选集的步骤包括：

[0189] 将所述各错误处的字和/或词转换为目标拼音；

[0190] 在所述模糊音字典中，查找与所述目标拼音相对应的模糊音或相似音，以形成第一候选结果；同时，

[0191] 对所述目标拼音的声母和韵母进行拆分，以获取拆分后的目标声母和目标韵母；

[0192] 在所述模糊音字典中，查找与所述目标声母和所述目标韵母对应的模糊音或相似音，以形成第二候选结果；

[0193] 在所述形似字字典中，查找与所述各错误相对应的所有形似字，以形成第三候选结果；

[0194] 基于所述第一候选结果、所述第二候选结果和所述第三候选结果，形成所述纠错候选集。

[0195] 此外，可选的技术方案是，在获取所述纠错候选集中的候选结果在对应句子中的困惑度之前，还包括：

[0196] 基于预训练的筛选模型对所述纠错候候选集中的候选结果进行初步筛选，以确定目标候选集。

[0197] 此外，可选的技术方案是，所述基于预训练的筛选模型对所述纠错候候选集中的候选结果进行初步筛选，以确定目标候选集的步骤包括：

[0198] 基于获取的训练数据训练逻辑回归模型；

[0199] 基于所述逻辑回归模型对所述纠错候选集中的结果进行预测，并获取对应的预测分值；

[0200] 基于预设范围，过滤所述预测分值小于预设范围的候选结果，以确定所述目标候选集。

[0201] 此外，可选的技术方案是，所述困惑度的获取公式为：

[0202] $$PP\left(S\right) = 2^{-\frac{1}{N}\sum \log\left(P\left(w_i\right)\right)}$$

[0203] 其中，w表示所述候选结果中的字或词，i表示w在对应句子中的序号，s表示替换该候选结果后的句子，N表示句子中所有字或词的个数，p表示w在对应句子中的概率值。

[0204] 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取存储介质中。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)。

[0205] 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。

[0206] 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。

[0207] 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。

[0208] 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。

[0209] 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。

[0210] 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。第二等词语用来表示名称，而并不表示任何特定的顺序。

[0211] 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。

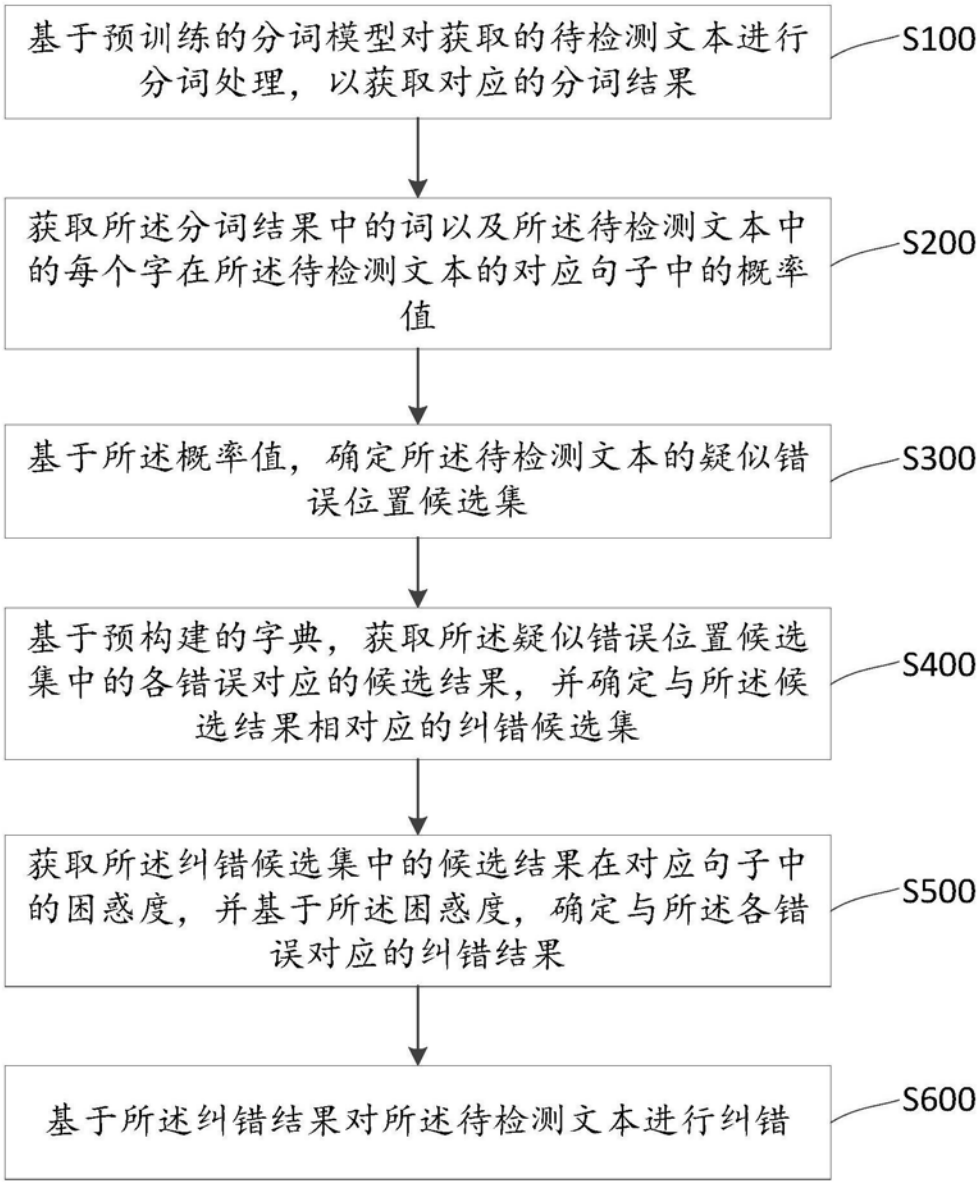

基于预训练的分词模型对获取的待检测文本进行分词处理，以获取对应的分词结果
S100
获取所述分词结果中的词以及所述待检测文本中的每个字在所述待检测文本的对应句子中的概率值
S200
基于所述概率值，确定所述待检测文本的疑似错误位置候选集
S300
基于预构建的字典，获取所述疑似错误位置候选集中的各错误对应的候选结果，并确定与所述候选结果相对应的纠错候选集
S400
获取所述纠错候选集中的候选结果在对应句子中的困惑度，并基于所述困惑度，确定与所述各错误对应的纠错结果
S500
基于所述纠错结果对所述待检测文本进行纠错
S600

图1


基于人工智能的文本纠错装置200
分词结果获取单元210
概率值获取单元220
疑似错误位置候选集确定单元230
纠错候选集确定单元240
纠错结果确定单元250
文本纠错单元260

图2

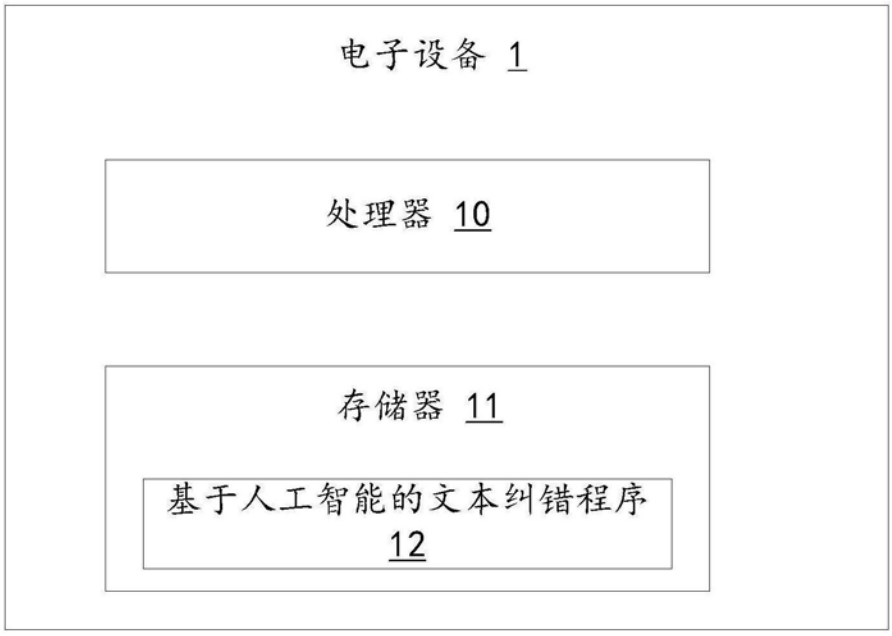

电子设备 1
处理器 10
存储器 11
基于人工智能的文本纠错程序 12

图3