(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115859951 B
(45)授权公告日 2023.05.02

(21)申请号 202310175363.8

(22)申请日 2023.02.28

(65)同一申请的已公布的文献号
申请公布号 CN 115859951 A

(43)申请公布日 2023.03.28

(73)专利权人 环球数科集团有限公司
地址 518063 广东省深圳市南山区粤海街道高新南九道10号深圳湾科技生态园10栋B座17层01-03号

(72)发明人 张卫平 丁烨 张伟 米小武 刘顿 郑小龙

(74)专利代理机构 北京清控智云知识产权代理事务所(特殊普通合伙) 11919
专利代理师 马肃

(51)Int.Cl.
*G06F 40/232*(2020.01)
*G06F 40/284*(2020.01)
*G06F 40/289*(2020.01)

(56)对比文件
CN 115688748 A,2023.02.03

审查员 王永波

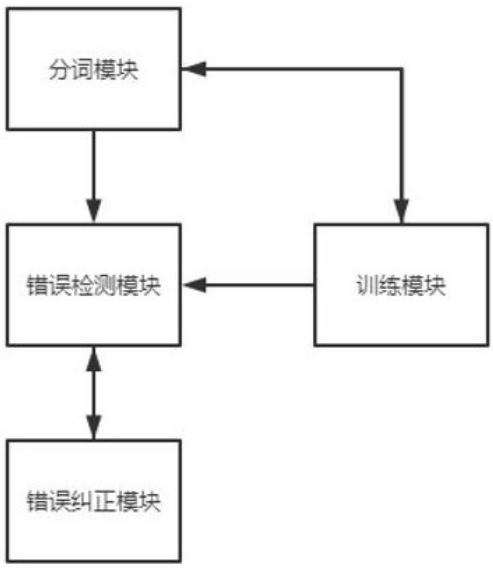
权利要求书2页 说明书6页 附图3页

(54)发明名称

一种用于AIGC的内容纠错系统

(57)摘要

本发明提供了一种用于AIGC的内容纠错系统,包括分词模块、错误检测模块、错误纠正模块和训练模块,所述分词模块将AIGC输出的内容拆分成多个词语,并将每个词语根据词性类别转换成词性编码,所述错误检测模块对词性编码之间的顺序进行检测,输出检测结果,所述错误纠正模块用于对内容进行修正,所述训练模块基于文本进行训练,得到用于所述错误检测模块的检测模型参数;本系统能够对内容的结构句式进行判断并进行纠错,使得AIGC最终输出的内容表达正确并能够被理解。




CN 115859951 B

1.一种用于AIGC的内容纠错系统,其特征在于,包括分词模块、错误检测模块、错误纠正模块和训练模块;

所述分词模块将AIGC输出的内容拆分成多个词语,并将每个词语根据词性类别转换成词性编码,所述错误检测模块对词性编码之间的顺序进行检测,输出检测结果,所述错误纠正模块用于对内容进行修正,所述训练模块基于文本进行训练,得到用于所述错误检测模块的检测模型参数;

所述错误检测模块包括预处理单元和执行单元,所述预处理单元用于对词性编码进行预处理得到初始值,所述执行单元包括节点,并由节点构成输入处理层、中间处理层和输出处理层,所述输入处理层对预处理后的初始值进行处理,中间处理层用于对数据进行缩减性处理,所述输出处理层用于输出检测结果,所述节点为一个计算处理器并包含两个输入端与一个输出端;

所述输入处理层、中间处理层和输出处理层共有n-1层,n为所述预处理单元处理后的初始值数量,第i层处理层包含n-i个节点,每个节点与上一层的两个节点的输出连接,并将输出值与下一层的一个节点连接,第一处理层为输入处理层,第n-1层处理层为输出处理层,所述节点根据两个输入值确定一个连接系数,并将同一层中最大的连接系数变更为1,其余连接系数变更为0;

所述输入处理层的节点根据下述公式处理得到输出值Out:

$$Out = in(1) + k \cdot in(2)/2^8;①$$

$$Out = k \cdot in(1)/2^8 + in(2);②$$

其中,①式为连接系数为1对应的节点以及左侧节点的执行公式,②式为连接系数为1对应的节点右侧的节点的执行公式,k为连接系数,in(1)为第一输入值,in(2)为第二输入值;

所述中间处理层的节点根据下述公式处理得到输出值Out:

$$Out = in(1) + k \cdot CH(in(1),\ in(2))/2^8;③$$

$$Out = k \cdot CH(in(2),\ in(1))/2^8 + in(2);④$$

其中,CH(a,b)为选择函数,用于将第二入参b的整数部分或者小数部分作为一个字节输出,③式为连接系数为1对应的节点以及左侧节点的执行公式,④式为连接系数为1对应的节点右侧的节点的执行公式;

输出处理层的节点将两个输入中对应的词性编码结构与基础句式结构比较,若词性编码结构存在于基础句式结构中,则输出判断值1,若词性编码结构不存在于基础句式结构中,则输出判断值0;

所述词性编码的大小为一字节,包括主码和副码,主码占七位,副码占一位,所述预处理单元将获取的词性编码转换成初始值,转换公式如下:

$$V = C \bmod 128;$$

其中,V为初始值,C为词性编码。

2.如权利要求1所述的一种用于AIGC的内容纠错系统,其特征在于,当所述错误检测模块输出的判断值为0时,所述错误纠正模块将主码进行变更使得主码结构存在于基础句式

结构中，将变更后的主码反馈给所述错误检测模块，重复所述错误检测模块输出的判断值为0时，所述错误纠正模块将主码进行变更使得主码结构存在于基础句式结构中的过程，直至所述错误检测模块输出的判断值为1，所述错误纠正模块对每次的主码变更进行记录，当所述错误检测模块输出的判断值为1后，所述错误纠正模块将根据主码的变更记录将对应的词汇变更为具有相近意思且词性编码正确的词汇。

# 一种用于AIGC的内容纠错系统

## 技术领域

[0001] 本发明涉及自然语言分析领域,具体涉及一种用于AIGC的内容纠错系统。

## 背景技术

[0002] 随着人工智能的发展,各种AIGC应用会在人们的生活中受到普及并改善生活,现有的AIGC应用能够输出规范并能够被理解的内容,但当输出的内容句式结构边复杂时,往往会出现一些错误使得无法正确理解输出的内容,现需要一种内容纠错系统对AIGC生成的内容进行检测并纠错,最终输出正确的内容。

[0003] 背景技术的前述论述仅意图便于理解本发明。此论述并不认可或承认提及的材料中的任一种公共常识的一部分。

[0004] 现在已经开发出了很多内容纠错系统,经过我们大量的检索与参考,发现现有的纠错系统有如公开号为CN112183072A所公开的系统,这些系统一般包括获得待纠错文本,对待纠错文本依次进行形近字纠错和常用字纠错,得到第一校正文本,并对待纠错文本进行常用词纠错,得到第二校正文本。然后获得第一校正文本和第二校正文本的困惑度,将困惑度最低的校正文本确定为待纠错文本的校正文本。但该系统仅仅是对常用词的纠错,当用词无错,但组成的复杂句式结构错误时无法检测出,导致无法理解正确的表达意思。

## 发明内容

[0005] 本发明的目的在于,针对所存在的不足,提出了一种用于AIGC的内容纠错系统。

[0006] 本发明采用如下技术方案:

[0007] 一种用于AIGC的内容纠错系统,包括分词模块、错误检测模块、错误纠正模块和训练模块;

[0008] 所述分词模块将AIGC输出的内容拆分成多个词语,并将每个词语根据词性类别转换成词性编码,所述错误检测模块对词性编码之间的顺序进行检测,输出检测结果,所述错误纠正模块用于对内容进行修正,所述训练模块基于文本进行训练,得到用于所述错误检测模块的检测模型参数;

[0009] 所述错误检测模块包括预处理单元和执行单元,所述预处理单元用于对词性编码进行预处理得到初始值,所述执行单元包括节点,并由节点构成输入处理层、中间处理层和输出处理层,所述输入处理层对预处理后的初始值进行处理,中间处理层用于对数据进行缩减性处理,所述输出处理层用于输出检测结果,所述节点为一个计算处理器并包含两个输入端与一个输出端;

[0010] 进一步的,所述输入处理层、中间处理层和输出处理层共有n-1层,n为所述预处理单元处理后的初始值数量,第i层处理层包含n-i个节点,每个节点与上一层的两个节点的输出连接,并将输出值与下一层的一个节点连接,第一处理层为输入处理层,第n-1层处理层为输出处理层,所述节点根据两个输入值确定一个连接系数,并将同一层中最大的连接系数变更为1,其余连接系数变更为0;

[0011] 进一步的，所述输入处理层的节点根据下述公式处理得到输出值Out：

[0012] $Out = in(1) + k \cdot in(2) / 2^8$；①

[0013] $Out = k \cdot in(1) / 2^8 + in(2)$；②

[0014] 其中，①式为连接系数为1对应的节点以及左侧节点的执行公式，②式为连接系数为1对应的节点右侧的节点的执行公式，k为连接系数，in(1)为第一输入值，in(2)为第二输入值；

[0015] 所述中间处理层的节点根据下述公式处理得到输出值Out：

[0016] $Out = in(1) + k \cdot CH(in(1), in(2)) / 2^8$；③

[0017] $Out = k \cdot CH(in(2), in(1)) / 2^8 + in(2)$；④

[0018] 其中，CH(a,b)为选择函数，用于将第二入参b的整数部分或者小数部分作为一个字节输出，③式为连接系数为1对应的节点以及左侧节点的执行公式，④式为连接系数为1对应的节点右侧的节点的执行公式；

[0019] 输出处理层的节点将两个输入中对应的词性编码结构与基础句式结构比较，若词性编码结构存在于基础句式结构中，则输出判断值1，若词性编码结构不存在于基础句式结构中，则输出判断值0；

[0020] 进一步的，所述词性编码的大小为一字节，包括主码和副码，主码占七位，副码占一位，所述预处理单元将获取的词性编码转换成初始值，转换公式如下：

[0021] $V = C \bmod 128$；

[0022] 其中，V为初始值，C为词性编码；

[0023] 进一步的，当所述错误检测模块输出的判断值为0时，所述错误纠正模块将主码进行变更使得主码结构存在于基础句式结构中，将变更后的主码反馈给所述错误检测模块，重复该过程直至所述错误检测模块输出的判断值为1，所述错误纠正模块对每次的主码变更进行记录，当所述错误检测模块输出的判断值为1后，所述错误纠正模块将根据主码的变更记录将对应的词汇变更为具有相近意思且词性编码正确的词汇。

[0024] 本发明所取得的有益效果是：

[0025] 本系统通过将文本内容切分成独立的词汇，将词汇转换成对应的编码，通过对编码进行分析来检测出内容中存在的问题，在检测的具体过程中，通过设置多层的节点对编码进行层层处理，降低内容量，最终对缩短后的编码内容进行比较判断，该方式能够对具有复杂结构的句式进行高准确率的纠错判断，而在处理过程中设计的参数，通过大量数据训练后获得，能够提高判断的准确性。

[0026] 为使能更进一步了解本发明的特征及技术内容，请参阅以下有关本发明的详细说明与附图，然而所提供的附图仅用于提供参考与说明，并非用来对本发明加以限制。

## 附图说明

[0027] 图1为本发明整体结构框架示意图；

[0028] 图2为本发明分词模块构成示意图；

[0029] 图3为本发明转换单元构成示意图；

[0030] 图4为本发明错误检测模块构成示意图；

[0031] 图5为本发明节点与处理层的关系示意图。

## 具体实施方式

[0032] 以下是通过特定的具体实施例来说明本发明的实施方式，本领域技术人员可由本说明书所公开的内容了解本发明的优点与效果。本发明可通过其他不同的具体实施例加以施行或应用，本说明书中的各项细节也可基于不同观点与应用，在不悖离本发明的精神下进行各种修饰与变更。另外，本发明的附图仅为简单示意说明，并非依实际尺寸的描绘，事先声明。以下的实施方式将进一步详细说明本发明的相关技术内容，但所公开的内容并非用以限制本发明的保护范围。

[0033] 实施例一：

[0034] 本实施例提供了一种用于AIGC的内容纠错系统，结合图1，包括分词模块、错误检测模块、错误纠正模块和训练模块；

[0035] 所述分词模块将AIGC输出的内容拆分成多个词语，并将每个词语根据词性类别转换成词性编码，所述错误检测模块对词性编码之间的顺序进行检测，输出检测结果，所述错误纠正模块用于对内容进行修正，所述训练模块基于文本进行训练，得到用于所述错误检测模块的检测模型参数；

[0036] 所述错误检测模块包括预处理单元和执行单元，所述预处理单元用于对词性编码进行预处理得到初始值，结合图5，所述执行单元包括节点，并由节点构成输入处理层、中间处理层和输出处理层，所述输入处理层对预处理后的初始值进行处理，中间处理层用于对数据进行缩减性处理，所述输出处理层用于输出检测结果，所述节点为一个计算处理器并包含两个输入端与一个输出端；

[0037] 所述输入处理层、中间处理层和输出处理层共有n-1层，n为所述预处理单元处理后的初始值数量，第i层处理层包含n-i个节点，每个节点与上一层的两个节点的输出连接，并将输出值与下一层的一个节点连接，第一处理层为输入处理层，第n-1层处理层为输出处理层，所述节点根据两个输入值确定一个连接系数，并将同一层中最大的连接系数变更为1，其余连接系数变更为0；

[0038] 所述输入处理层的节点根据下述公式处理得到输出值Out：

[0039] $$Out = in(1) + k \cdot in(2) / 2^8$$；①

[0040] $$Out = k \cdot in(1) / 2^8 + in(2)$$；②

[0041] 其中，①式为连接系数为1对应的节点以及左侧节点的执行公式，②式为连接系数为1对应的节点右侧的节点的执行公式，k为连接系数，in(1)为第一输入值，in(2)为第二输入值；

[0042] 所述中间处理层的节点根据下述公式处理得到输出值Out：

[0043] $$Out = in(1) + k \cdot CH(in(1),\ in(2)) / 2^8$$；③

[0044] $$Out = k \cdot CH(in(2),\ in(1)) / 2^8 + in(2)$$；④

[0045] 其中，CH(a,b)为选择函数，用于将第二入参b的整数部分或者小数部分作为一个字节输出，③式为连接系数为1对应的节点以及左侧节点的执行公式，④式为连接系数为1对应的节点右侧的节点的执行公式；

[0046] 输出处理层的节点将两个输入中对应的词性编码结构与基础句式结构比较，若词性编码结构存在于基础句式结构中，则输出判断值1，若词性编码结构不存在于基础句式结构中，则输出判断值0；

[0047] 所述词性编码的大小为一字节，包括主码和副码，主码占七位，副码占一位，所述预处理单元将获取的词性编码转换成初始值，转换公式如下：

[0048] $V=C \bmod 128$；

[0049] 其中，V为初始值，C为词性编码；

[0050] 当所述错误检测模块输出的判断值为0时，所述错误纠正模块将主码进行变更使得主码结构存在于基础句式结构中，将变更后的主码反馈给所述错误检测模块，重复该过程直至所述错误检测模块输出的判断值为1，所述错误纠正模块对每次的主码变更进行记录，当所述错误检测模块输出的判断值为1后，所述错误纠正模块将根据主码的变更记录将对应的词汇变更为具有相近意思且词性编码正确的词汇。

[0051] 实施例二：

[0052] 本实施例包含了实施例一中的全部内容，提供了一种用于AIGC的内容纠错系统，包括分词模块、错误检测模块、错误纠正模块和训练模块；

[0053] 所述分词模块将AIGC输出的内容拆分成多个词语，并将每个词语根据词性类别转换成编码，得到一串由编码构成的数组，所述AIGC输出的内容称为目标文本；

[0054] 所述错误检测模块对数组之间顺序进行检测，输出检测结果；

[0055] 所述错误纠正模块对内容进行修正；

[0056] 所述训练模块基于文本进行训练，得到用于所述错误检测模块的检测模型的参数；

[0057] 结合图2，所述分词模块包括单词库、转换单元和分词存储单元，所述单词库中记录了词汇以及每个词汇对应的词性编码，所述转换单元从目标文本中获取文字并在单词库中确定对应的词性编码，所述分词存储单元按照顺序存储词性编码；

[0058] 所述转换单元将目标文本转换成词性编码的过程包括如下步骤：

[0059] S1、从目标文本中获取一个汉字，将该汉字作为检索目标，令标记符k=0；

[0060] S2、查找检索目标的词性编码，若查找到，更新词性编码，并进入步骤S3，若查找不到且k=0，进入步骤S5，若查找不到且k=1，进入步骤S6；

[0061] S3、判断该词性编码是否为唯一性编码，若是，进入步骤S4，若不是进入步骤S5；

[0062] S4、将词性编码输出至分词存储单元，从目标文本中删除检索目标对应的汉字，回到步骤S1；

[0063] S5、从目标文本中获取下一个汉字，与原有的检索目标组成新的检索目标，令标记符k=1，回到步骤S2；

[0064] S6、删除检索目标中的最后一个汉字，回到步骤S4；

[0065] 结合图3，所述转换单元包括读取处理器、数据寄存器、检索处理器和逻辑控制器，所述读取处理器用于从目标文本中获取汉字，所述数据寄存器用于保存检索目标、词性编码和标记符，所述检索处理器用于查找检索目标的词性编码，所述逻辑控制器用于执行步骤S1至步骤S6的流程；

[0066] 需要注意的是，保存于数据寄存器中的检索目标和词性编码数量均为一个；

[0067] 步骤S3中的唯一性编码具有的特征为:唯一性编码对应的词汇在后面添加任意汉字后均不会出现在单词库中;

[0068] 所述词性编码由副码和主码构成,副码为1时词性编码为唯一性编码,副码为0时词性编码为非唯一性编码,主码用于表示“名词”、“动词”、“介词”等词性,词性编码的大小为一个字节,主码占用剩余的七个位数据;

[0069] 所述错误检测模块从所述分词存储单元中获取词性编码,并基于检测模型对词性编码进行计算处理;

[0070] 结合图4,所述错误检测模块包括预处理单元和执行单元;

[0071] 所述预处理单元将获取的词性编码转换成初始值,转换公式如下:

[0072] $V = C \bmod 128$;

[0073] 其中,V为初始值,C为词性编码;

[0074] 所述执行单元将初始值作为输入,经过处理后输出一个判断值,所述执行单元通过下述步骤对初始值进行处理:

[0075] S21、基于初始值的数量n生成n-1个处理层,第i个处理层包含n-i个处理节点每个节点与上一层的两个节点的输出连接,并将输出值与下一层的一个节点连接,第一层的节点将两个初始值作为输入,第n-1层的节点输出判断值,i用于表示处理层的层数;

[0076] S22、第一处理层的节点基于两个输入值生成一个连接系数,选择最大的一个连接系数,将该连接系数设置为1,其余连接系数设置为0;

[0077] S23、所述第一处理层的节点根据下述公式处理得到输出值Out:

[0078] $Out = in(1) + k \cdot in(2)/2^8$;①

[0079] $Out = k \cdot in(1)/2^8 + in(2)$;②

[0080] 其中,①式为连接系数为1对应的节点以及左侧节点的执行公式,②式为连接系数为1对应的节点右侧的节点的执行公式,k为连接系数,in(1)为第一输入值,in(2)为第二输入值,左右侧用于表示节点的顺序关系;

[0081] S24、第二处理层至第n-2处理层之一的节点基于两个输入值生成一个连接系数,每层处理层选择最大的一个连接系数,将该连接系数设置为1,其余连接系数设置为0;

[0082] S25、所述第二理层至第n-2处理层之一的节点根据下述公式处理得到输出值Out:

[0083] $Out = in(1) + k \cdot CH(in(1),\ in(2))/2^8$;③

[0084] $Out = k \cdot CH(in(2),\ in(1))/2^8 + in(2)$;④

[0085] 其中,CH(a,b)为选择函数,用于将第二入参b的整数部分或者小数部分作为一个字节输出,③式为连接系数为1对应的节点以及左侧节点的执行公式,④式为连接系数为1对应的节点右侧的节点的执行公式;

[0086] 当CH(a,b)中的第二入参b为整数时,将b直接作为结果输出;

[0087] 当CH(a,b)中的第二入参b包含整数部分b1和小数部分b2,且第一入参为整数时,比较a与b1的连接系数以及a与b2的连接系数,较大的连接系数对应的b1或者b2作为一个字节输出;

[0088] 当CH(a,b)中的第二入参b包含整数部分b1和小数部分b2,且第一入参a包含整数部分a1和小数部分a2时,比较a1与b1的连接系数、a1与b2的连接系数、a2与b1的连接系数以

及a2与b2的连接系数，最大的连接系数对应的b1或者b2作为一个字节输出；

[0089] S26、重复步骤S24和步骤S25，按照处理层的顺序依次处理对应的节点；

[0090] S27、第n-1层处理层的节点将两个输入中对应的3个或4个主码结构与基础句式结构比较，若主码结构存在于基础句式结构中，则输出判断值1，若主码结构不存在于基础句式结构中，则输出判断值0，当判断值为0时，该节点将两个输入值输出至错误纠错模块；

[0091] 所述训练模块用于获得任意两个主码之间连接系数，所述训练模块包括调用单元、统计单元和归一单元，所述调用单元用于调用所述分词模块将训练的文本转换成只含有主码的词性编码，所述统计单元将相邻的词性编码作为对象进行统计，得到两个对应主码的出现频次，所述归一单元将出现频次根据下式进行处理得到连接系数：

[0092] $$k=\log_{\left(10^{\left|\lg\sqrt{M}\right|}\right)}m$$；

[0093] 其中，m为两个相邻主码的出现频次，M为统计的出现频次总和；

[0094] 在步骤S24中，当只有一个输入值包含小数部分时，将该输入值的整数部分作为第一主码，小数部分作为第二主码，另一个输入值作为第三主码，将第一主码与第三主码的连接系数以及第二主码与第三主码的连接系数的较大值作为两个输入值的连接系数，当两个输入值均包含小数部分时，将一个输入值的整数部分作为第一主码，小数部分作为第二主码，将另一个输入值的整数部分作为第三主码，小数部分作为第四主码，将第一主码与第三主码的连接系数、第二主码与第三主码的连接系数、第一主码与第四主码的连接系数以及第二主码与第四主码的连接系数中的最大值作为两个输入值的连接系数；

[0095] 所述错误纠正模块将接收的两个输入值转换成对应的主码，对主码进行变更使得主码结构存在于基础句式结构中，将变更后的主码反馈给所述错误检测模块，重复该过程直至所述错误检测模块输出的判断值为1，所述错误纠正模块对每次的主码变更进行记录，当所述错误检测模块输出的判断值为1后，所述错误纠正模块将根据主码的变更记录将对应的词汇变更为具有相近意思且词性编码正确的词汇。

[0096] 以上所公开的内容仅为本发明的优选可行实施例，并非因此局限本发明的保护范围，所以凡是运用本发明说明书及附图内容所做的等效技术变化，均包含于本发明的保护范围内，此外，随着技术发展其中的元素可以更新的。

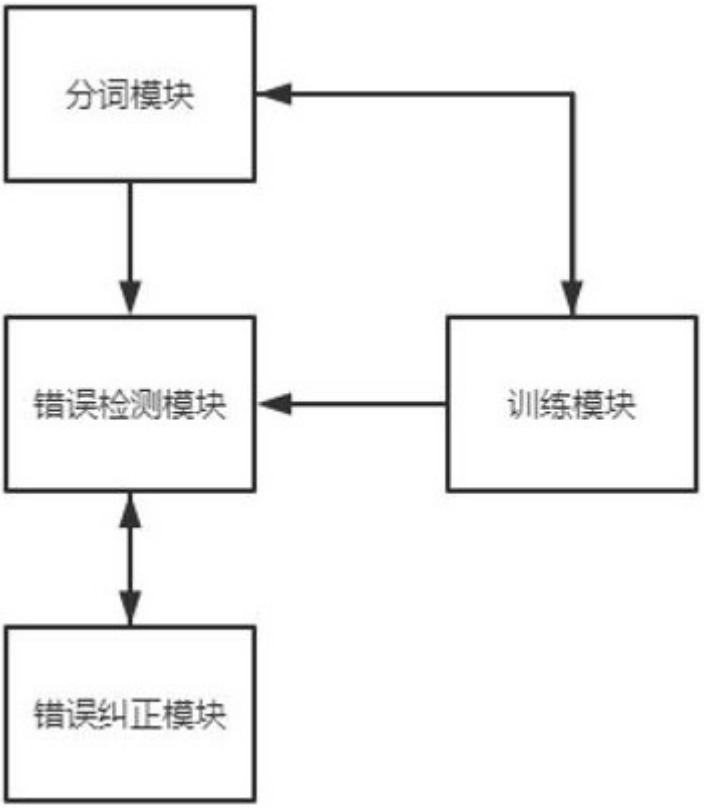



图 1


图 2

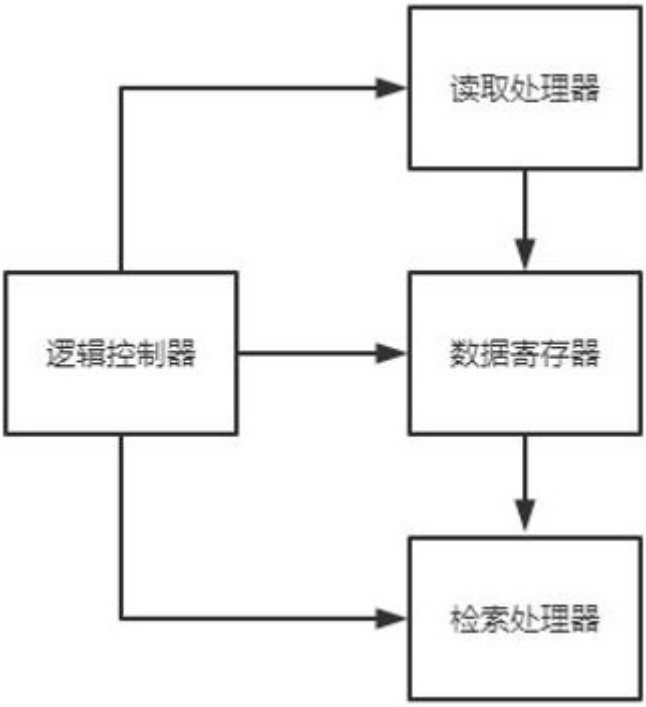



图 3

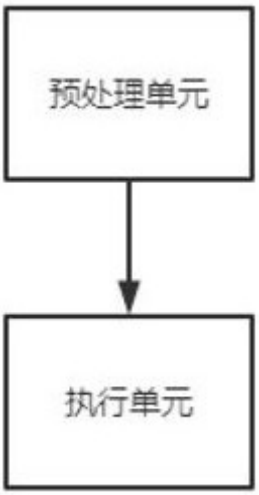



图 4

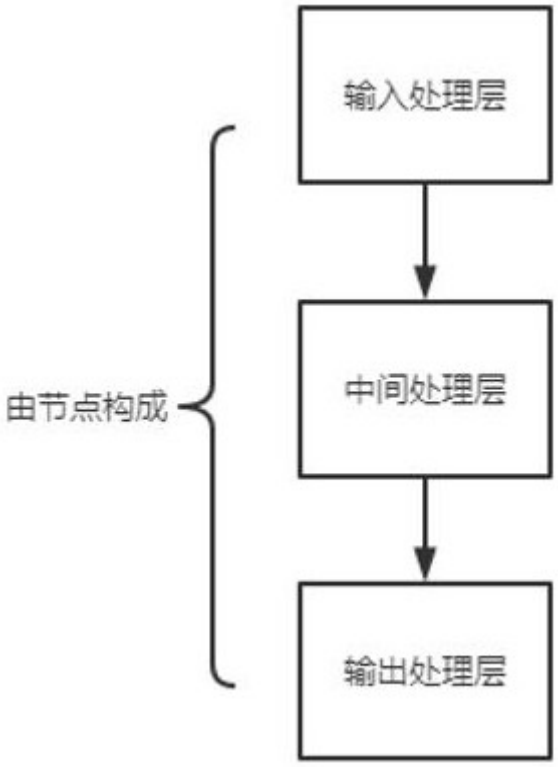



图 5