(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 112101003 A
(43)申请公布日 2020.12.18

(21)申请号 202010963503.4

(22)申请日 2020.09.14

(71)申请人 深圳前海微众银行股份有限公司
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 汤耀华 周楠楠 杨海军 徐倩

(74)专利代理机构 深圳市世纪恒程知识产权代理事务所 44287
代理人 张志江

(51)Int.Cl.
G06F 40/211(2020.01)
G06F 40/151(2020.01)

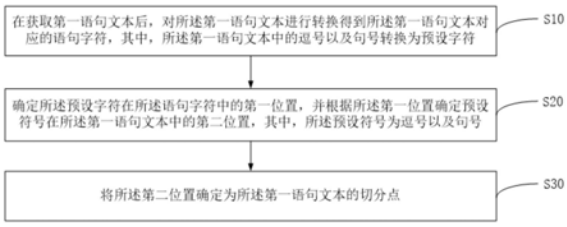
权利要求书2页 说明书10页 附图3页

(54)发明名称

语句文本的切分方法、装置、设备和计算机可读存储介质

(57)摘要

本发明涉及金融科技技术领域，公开了一种语句文本的切分方法、装置、设备和计算机可读存储介质。所述语句文本的切分方法包括：在获取第一语句文本后，对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符，其中，所述第一语句文本中的逗号以及句号转换为预设字符；确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号；将所述第二位置确定为所述第一语句文本的切分点。本发明提高了人工语音智能客户系统对客户意图识别的准确性。



CN 112101003 A

1.一种语句文本的切分方法,其特征在于,所述语句文本的切分方法包括:

获取第一语句文本,对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符,其中,所述第一语句文本中的逗号以及句号转换为预设字符;

确定所述预设字符在所述语句字符中的第一位置,并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置,其中,所述预设符号为逗号以及句号;

将所述第二位置确定为所述第一语句文本的切分点。

2.如权利要求1所述的语句文本的切分方法,其特征在于,所述获取第一语句文本的步骤之后,还包括:

确定所述第一语句文本是否满足校正条件;

在所述第一语句文本不满足校正条件时,执行所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤。

3.如权利要求2所述的语句文本的切分方法,其特征在于,所述确定所述第一语句文本是否满足校正条件的步骤之后,还包括:

在所述第一语句文本满足校正条件时,将所述第一语句文本发送至各个终端;

接收各个所述终端反馈的第二语句文本,并将相同的第二语句文本确定为目标语句文本,其中,所述第二语句文本为重新标注预设符号的所述第一语句文本;

对所述目标语句文本进行转换得到所述目标语句文本对应的语句字符,并确定所述预设字符在所述目标语句文本对应的语句字符中的第三位置;

根据所述第三位置确定预设符号在所述目标语句文本中的第四位置,并将所述第四位置确定为所述目标语句文本的切分点。

4.如权利要求2所述的语句文本的切分方法,其特征在于,所述校正条件包括以下至少一种:

所述第一语句文本的属性为预设属性,所述属性包括转换第一语句文本的语音数据的类型及/或所述第一语句文本的来源;

所述第一语句文本中预设符号的数量小于预设数量。

5.如权利要求1-4任一项所述的语句文本的切分方法,其特征在于,所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤包括:

确定所述第一语句文本对应的向量;

对所述向量进行特征提取得到所述第一语句文本中文字以及符号对应的特征值;

根据各个所述特征值确定所述第一语句文本对应的语句字符。

6.如权利要求1-4任一项所述的语句文本的切分方法,其特征在于,所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤包括:

根据转换模型对所述第一语句文本进行转换,得到所述第一语句文本对应的语句字符。

7.如权利要求6所述的语句文本的切分方法,其特征在于,所述根据转化模型对所述第一语句文本进行转换,得到所述第一语句文本对应的语句字符的步骤包括:

裁剪或补全所述第一语句文本,以使所述第一语句文本的长度为预设长度;

对预设长度的第一语句文本的句首设置句首标签,且对预设长度的第一语句文本的句尾设置句尾标签,得到第三语句文本;

将所述第三语句文本输入转换模型，得到所述转换模型输出的所述语句字符。

8.如权利要求1-4任一项所述的语句文本的切分方法，其特征在于，所述将所述第二位置确定为所述第一语句文本确定为切分点的步骤之后，还包括：

对所述第一语句文本的切分点进行标记，以得到第四语句文本；

将各个所述第四语句文本作为训练样本，并根据各个所述训练样本对预设模型进行训练得到语句切分模型。

9.一种语句文本的切分装置，其特征在于，所述语句文本的切分装置包括：

转换模块，用于获取第一语句文本，对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符，其中，所述第一语句文本中的逗号以及句号转换为预设字符；

确定模块，用于确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号；

所述确定模块，还用于将所述第二位置确定为所述第一语句文本的切分点。

10.一种语句文本的切分设备，其特征在于，所述语句文本的切分设备包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的切分程序，所述切分程序被所述处理器执行时实现如权利要求1至8中任一项所述的语句文本的切分方法的步骤。

11.一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上存储有切分程序，所述切分程序被处理器执行时实现如权利要求1至8中任一项所述的计算机可读存储方法的步骤。

语句文本的切分方法、装置、设备和计算机可读存储介质

技术领域

[0001] 本发明涉及金融科技(Fintech)技术领域，尤其涉及一种语句文本的切分方法、装置、设备和计算机可读存储介质。

背景技术

[0002] 随着计算机技术的发展，越来越多的技术应用在金融领域，传统金融业正在逐步向金融科技(Fintech)转变，但由于金融行业的安全性、实时性要求，也对技术提出了更高的要求。

[0003] 相比于文本客服系统，人工语音智能客服系统的成本较低。人工语音智能客服系统需要客户的输入语音，客户会用很长的一段文字语音描述客户遇到的问题。语音有可能较长，例如语音为“那个我之前我上个月找你们借了一万块钱啊那个当时办的是分十期还钱还给你们明天应该要还第一笔钱第一笔借款这个要怎么还你教一下我啊我不是很会弄主要是在哪里还啊没有头绪啊”。过长的语音句子包含了客户的多个意思，会给后续的自然语言理解带来很大的麻烦。所以在实际架构中非常有必要先对语音所识别的语句文本做语义切割，切成一段一段单一意思的短句。

[0004] 目前，语句文本按照固定长度进行切分得到多个短句，这种方式会切断一个完整的语义或者一个短句中包含多个语义，导致对客户的意图识别准确性较低。

发明内容

[0005] 本发明的主要目的在于提供一种语句文本的切分方法、装置、设备和计算机可读存储介质，旨在解决对客户的意图识别准确性较低的问题。

[0006] 为实现上述目的，本发明提供一种语句文本的切分方法，所述语句文本的切分方法包括：

[0007] 获取第一语句文本，对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符，其中，所述第一语句文本中的逗号以及句号转换为预设字符；

[0008] 确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号；

[0009] 将所述第二位置确定为所述第一语句文本的切分点。

[0010] 可选地，所述获取第一语句文本的步骤之后，还包括：

[0011] 确定所述第一语句文本是否满足校正条件；

[0012] 在所述第一语句文本不满足校正条件时，执行所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤。

[0013] 可选地，所述确定所述第一语句文本是否满足校正条件的步骤之后，还包括：

[0014] 在所述第一语句文本满足校正条件时，将所述第一语句文本发送至各个终端；

[0015] 接收各个所述终端反馈的第二语句文本，并将相同的第二语句文本确定为目标语句文本，其中，所述第二语句文本为重新标注预设符号的所述第一语句文本；

[0016] 对所述目标语句文本进行转换得到所述目标语句文本对应的语句字符，并确定所述预设字符在所述目标语句文本对应的语句字符中的第三位置；

[0017] 根据所述第三位置确定预设符号在所述目标语句文本中的第四位置，并将所述第四位置确定为所述目标语句文本的切分点。

[0018] 可选地，所述校正条件包括以下至少一种：

[0019] 所述第一语句文本的属性为预设属性，所述属性包括转换第一语句文本的语音数据的类型及/或所述第一语句文本的来源；

[0020] 所述第一语句文本中预设符号的数量小于预设数量。

[0021] 可选地，所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤包括：

[0022] 确定所述第一语句文本对应的向量；

[0023] 对所述向量进行特征提取得到所述第一语句文本中文字以及符号对应的特征值；

[0024] 根据各个所述特征值确定所述第一语句文本对应的语句字符。

[0025] 可选地，所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤包括：

[0026] 根据转换模型对所述第一语句文本进行转换，得到所述第一语句文本对应的语句字符。

[0027] 可选地，所述根据转化模型对所述第一语句文本进行转换，得到所述第一语句文本对应的语句字符的步骤包括：

[0028] 裁剪或补全所述第一语句文本，以使所述第一语句文本的长度为预设长度；

[0029] 对预设长度的第一语句文本的句首设置句首标签，且对预设长度的第一语句文本的句尾设置句尾标签，得到第三语句文本；

[0030] 将所述第三语句文本输入转换模型，得到所述转换模型输出的所述语句字符。

[0031] 可选地，所述将所述第二位置确定为所述第一语句文本确定为切分点的步骤之后，还包括：

[0032] 对所述第一语句文本的切分点进行标记，以得到第四语句文本；

[0033] 将各个所述第四语句文本作为训练样本，并根据各个所述训练样本对预设模型进行训练得到语句切分模型。

[0034] 为实现上述目的，本发明还提供一种语句文本的切分装置，所述语句文本的切分装置包括：

[0035] 转换模块，用于获取第一语句文本，对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符，其中，所述第一语句文本中的逗号以及句号转换为预设字符；

[0036] 确定模块，用于确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号；

[0037] 所述确定模块，还用于将所述第二位置确定为所述第一语句文本的切分点。

[0038] 为实现上述目的，本发明还提供一种语句文本的切分设备，所述语句文本的切分设备包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的切分程序，所述切分程序被所述处理器执行时实现如上所述的语句文本的切分方法的步骤。

[0039] 为实现上述目的，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上存储有切分程序，所述切分程序被处理器执行时实现如上所述的计算机可读存储方法的步骤。

[0040] 本发明提供一种语句文本的切分方法、装置、设备和计算机可读存储介质，语句文本的切分装置在获取到语句文本后，对语句文本进行转换得到语句文本对应的语句字符，且语句文本中的逗号以及句号转换为预设字符，装置再确定预设字符在语句字符中的第一位置，并根据第一位置确定句号以及逗号在文本语句中的第二位置，最后将第二位置作为切分点。本发明通过将语句文本中的句号以及逗号转换为预设字符，以识别出预设字符的位置，从而根据预设字符的位置确定句号以及逗号在文本语句中的位置，进而将逗号以及句号的位置作为切分点，以供装置根据切分点对语句文本进行切分，也即相对于现有技术中语句文本按照固定长度进行切分导致完整的语义被切断或切分的短句包含多个语义，本发明根据切分点对语句文本进行切分得到短句仅包括一个完整的语义，提高了人工语音智能客户系统对客户意图识别的准确性。

附图说明

[0041] 图1为本发明实施例方案涉及的硬件运行环境的语句文本的切分装置/设备的硬件结构示意图；

[0042] 图2为本发明语句文本的切分方法第一实施例的流程示意图；

[0043] 图3为本发明语句文本的切分方法第二实施例的流程示意图；

[0044] 图4为本发明语句文本的切分方法第三实施例的流程示意图；

[0045] 图5为本发明语句文本的切分方法第四实施例的流程示意图；

[0046] 图6为本发明语句文本的切分装置的功能模块示意图。

[0047] 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

[0048] 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。

[0049] 参照图1，图1为本发明实施例方案涉及语句文本的切分装置或语句文本的切分设备的硬件运行环境的硬件结构示意图。

[0050] 如图1所示，语句文本的切分设备/语句文本的切分装置可以包括：处理器1001，例如CPU，通信总线1002，用户接口1003，网络接口1004，存储器1005。其中，通信总线1002用于实现这些组件之间的连接通信。用户接口1003可以包括显示屏(Display)、输入单元比如键盘(Keyboard)，可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口(如Wi-Fi接口)。存储器1005可以是高速RAM存储器，也可以是稳定的存储器(non-volatile memory)，例如磁盘存储器。存储器1005可选的还可以是独立于前述处理器1001的存储装置。

[0051] 本领域技术人员可以理解，图1中示出的终端的结构并不构成对语句文本的切分装置或语句文本的切分设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

[0052] 如图1所示，作为一种计算机存储介质的存储器1005中可以包括操作系统、网络通

信模块、用户接口模块以及切分程序。

[0053] 在图1所示的语句文本的切分装置或语句文本的切分设备中，网络接口1004主要用于连接后台服务端，与后台服务端进行数据通信；用户接口1003主要用于连接客户端，与客户端进行数据通信；而处理器1001可以用于调用存储器1005中存储的切分程序，并执行以下操作：

[0054] 在获取第一语句文本后，对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符，其中，所述第一语句文本中的逗号以及句号转换为预设字符；

[0055] 确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号；

[0056] 将所述第二位置确定为所述第一语句文本的切分点。

[0057] 在一实施例中，处理器1001可以调用存储器1005中存储的切分程序，还执行以下操作：

[0058] 确定所述第一语句文本是否满足校正条件；

[0059] 在所述第一语句文本不满足校正条件时，执行所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤。

[0060] 在一实施例中，处理器1001可以调用存储器1005中存储的切分程序，还执行以下操作：

[0061] 在所述第一语句文本满足校正条件时，将所述第一语句文本发送至各个终端；

[0062] 接收各个所述终端反馈的第二语句文本，并将相同的第二语句文本确定为目标语句文本，其中，所述第二语句文本为重新标注预设符号的所述第一语句文本；

[0063] 对所述目标语句文本进行转换得到所述目标语句文本对应的语句字符，并确定所述预设字符在所述目标语句文本对应的语句字符中的第三位置；

[0064] 根据所述第三位置确定预设符号在所述目标语句文本中的第四位置，并将所述第四位置确定为所述目标语句文本的切分点。

[0065] 在一实施例中，处理器1001可以调用存储器1005中存储的切分程序，还执行以下操作：

[0066] 所述第一语句文本的属性为预设属性，所述属性包括转换第一语句文本的语音数据的类型及/或所述第一语句文本的来源；

[0067] 所述第一语句文本中预设符号的数量小于预设数量。

[0068] 在一实施例中，处理器1001可以调用存储器1005中存储的切分程序，还执行以下操作：

[0069] 确定所述第一语句文本对应的向量；

[0070] 对所述向量进行特征提取得到所述第一语句文本中文字以及符号对应的特征值；

[0071] 根据各个所述特征值确定所述第一语句文本对应的语句字符。

[0072] 在一实施例中，处理器1001可以调用存储器1005中存储的切分程序，还执行以下操作：

[0073] 根据转换模型对所述第一语句文本进行转换，得到所述第一语句文本对应的语句字符。

[0074] 在一实施例中，处理器1001可以调用存储器1005中存储的切分程序，还执行以下

操作：

[0075] 裁剪或补全所述第一语句文本，以使所述第一语句文本的长度为预设长度；

[0076] 对预设长度的第一语句文本的句首设置句首标签，且对预设长度的第一语句文本的句尾设置句尾标签，得到第三语句文本；

[0077] 将所述第三语句文本输入转换模型，得到所述转换模型输出的所述语句字符。

[0078] 在一实施例中，处理器1001可以调用存储器1005中存储的切分程序，还执行以下操作：

[0079] 对所述第一语句文本的切分点进行标记，得到第四语句文本；

[0080] 将各个所述第四语句文本作为训练样本，并根据各个所述训练样本对预设模型进行训练得到语句切分模型。

[0081] 基于上述语句文本的切分装置/语句文本的切分设备的硬件结构，提出本发明语句文本的切分方法的各实施例。

[0082] 本发明提供一种语句文本的切分方法。

[0083] 参照图2，图2为本发明语句文本的切分方法第一实施例，所述语句文本的切分方法包括：

[0084] 步骤S10，在获取第一语句文本后，对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符，其中，所述第一语句文本中的逗号以及句号转换为预设字符；

[0085] 在本实施中，执行主体为语句文本的切分装置，为了便于描述，以下采用装置指代语句文本的切分装置。装置可以获取第一语句文本。第一语句文本可以是文学作品或者新闻文本中的长句构成的文本，第一语句文本还可以是录音数据或者语音数据转换得到的文本。

[0086] 装置在得到第一语句文本后，对第一语句文本进行转换得到第一语句文本对应的语句字符。语句字符可以理解为：第一语句文本中每个文字以及每个标点转换为对应的字符所构成的字符串。

[0087] 具体的，装置中存储有文字与字符的第一映射关系、标点符号与字符的第二映射关系。装置根据第一映射关系以及第二映射关系将语句文本中的每个文字以及每个标点符号转换为对应的字符，各个字符按照转换时间从早到晚进行排序得到语句字符，越接近于语句文本的句首的文字或者标点符号，文字或标点符号的转换时间越早。第二映射关系可以是预设符号与字符之间的映射关系，预设字符可以为句号以及逗号。本实施例中仅考虑语句文本的切分，因此，不同预设符号对应的字符可相同，且不同文字对应的字符也可相同。例如，句号以及逗号对应的字符均为1，不同的文字对应的字符均为0。

[0088] 步骤S20，确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号；

[0089] 装置在确定语句字符后，对语句字符中的预设字符进行识别。例如，预设字符为1，则对语句字符中的字符1进行识别。装置再确定预设字符在语句字符中的第一位置。预设字符的第一位置对应于预设符号在第一语句文本中的第二位置，因而装置可以根据第一位置确定预设符号在第一语句文本中的第二位置。预设符号为逗号以及句号。

[0090] 具体的，语句字符中每个字符具有对应的序号，序号通过语句文本中文字以及符号转换字符的转换时间排序得到，排序方式可以是按照转换时间从早到晚进行排序。语句

文本中的每个文字以及字符也具有对应的序号，序号指的是语句文本中文字或者符号所在的列。第一位置对应一个预设字符，因此，第一位置也具有对应的序号。逗号以及句号在语句文本中的列，因此，作为预设符号的逗号以及句号在语句文本的第二位置也具有序号。

[0091] 而语句文本中的每个文字以及每个符号均转换为对应的字符，因此，预设字符在语句字符的第一位置对应的序号实际上与预设符号在语句文本中的第二位置的序号是相同的，因而装置可以根据第一位置确定预设符号在语句文本中的第二位置。

[0092] 例如，第一语句文本为：“每个人都会经历这个阶段，看见一座山，就想知道山后面是什么。我很想告诉他，可能翻过去山后面，你会发觉没有什么特别，回头看会觉得这边更好。”，第一语句文本对应的语句字符为00000000000100001000000000100000100000001000000000010000000001，第一位置即为语句字符中的左数的第12个位置、第17个位置、第27个位置、第33个位置、第41个位置、第51个位置以及第61个位置，对应的，第二位置即为第一文本语句中的12个位置、第17个位置、第27个位置、第33个位置、第41个位置、第51个位置以及第61个位置。

[0093] 步骤S30，将所述第二位置确定为所述第一语句文本的切分点。

[0094] 装置在确定第二位置后，将每个第二位置作为切分点，也即将第一语句文本中的每个句号以及逗号作为切分点，再对切分点进行切分，从而将第一语句文本切分为多个短句，使得每个短句仅含有一个完整的语义。

[0095] 此外，装置在确定第一语句文本的第二位置后，将各个第二位置确定为切分点，确定为切分点的方式可以是对第二位置进行标记，并将标记的第二位置的第一语句文本作为标签(标记有标签的第一语句文本可为第四语句文本)。装置再将各个标记有标签的第一语句文本作为训练样本，最后根据各个训练样本对预设模型进行训练得到语句切分模型。语句切分模型可以放置于人工语音智能客户系统中自然语言理解流程的靠前位置，人工语音智能客户系统再获取到语句文本后，将语句文本输入值语句切分模型中，得到切分后的各个短句片段，再将各个短句片段进行后续的自然语言处理，从而识别出客户的意图。

[0096] 在本实施例提供的技术方案中，语句文本的切分装置在获取到语句文本后，对语句文本进行转换得到语句文本对应的语句字符，且语句文本中的逗号以及句号转换为预设字符，装置再确定预设字符在语句字符中的第一位置，并根据第一位置确定句号以及逗号在文本语句中的第二位置，最后将第二位置作为切分点。本发明通过将语句文本中的句号以及逗号转换为预设字符，以识别出预设字符的位置，从而根据预设字符的位置确定句号以及逗号在文本语句中的位置，进而将逗号以及句号的位置作为切分点，以供装置根据切分点对语句文本进行切分，也即相对于现有技术中语句文本按照固定长度进行切分导致完整的语义被切断或切分的短句包含多个语义，本发明根据切分点对语句文本进行切分得到短句仅包括一个完整的语义，提高了人工语音智能客户系统对客户意图识别的准确性。

[0097] 参照图3，图3为本发明语句文本的切分方法的第二实施例，基于第一实施例，所述语句文本的切分方法还包括：

[0098] 步骤S40，获取第一语句文本，确定所述第一语句文本是否满足校正条件；

[0099] 步骤S50，在所述第一语句文本不满足校正条件时，执行所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤；

[0100] 步骤S60，在所述第一语句文本满足校正条件时，将所述第一语句文本发送至各个

终端；

[0101] 步骤S70，接收各个所述终端反馈的第二语句文本，并将相同的第二语句文本确定为目标语句文本，其中，所述第二语句文本为重新标注预设符号的所述第一语句文本；

[0102] 步骤S80，对所述目标语句文本进行转换得到所述目标语句文本对应的语句字符，并确定所述预设字符在所述目标语句文本对应的语句字符中的第一位置；

[0103] 步骤S90，据所述第一位置确定预设符号在所述目标语句文本中的第三位置，并将所述第三位置确定为所述目标语句文本的切分点。

[0104] 在本实施例中，第一语句文本可能是文学作品或新闻文本中的句子，此种情况下，第一语句文本中的逗号以及句号是正确的断句符号，也即不存在对第一语句文本中的断句存在疑虑。第一语句文本还可能是特定领域的训练数据，例如，金融领域的客服产品的语句文本；此种情况下，不同客服对语句文本的断句方式不同。此外，第一语句文本还可能是录音数据或者语音数据转换的文本，第一语句文本中并未有标点符号。可以理解的是，某些语句文本需要重新标注逗号以及句号进行完整语义的断句，某些语句文本则无需重新批注逗号以及句号。对此，为需要重新标注预设符号(逗号以及句号)的语句文本设置校正条件，也即第一语句文本在满足校正条件时，需要对第一语句文本进行句号以及逗号的重新标注；若第一语句文本不满足校正条件，则无需对其标注预设符号。

[0105] 校正条件包括第一语句文本的属性为预设属性、第一语句文本中预设符号的数量小于预设数量中的至少一种。属性包括转换第一语句文本的语音数据的类型以及第一语句文本的来源中的至少一个。预设属性为转换语句文本的数量为录音数据以及语音数据中的至少一个，语句文本的来源为特性领域的数据(特定领域可为金融领域)也可视为预设属性。可以理解的是，在当转换为第一语句文本的语音数据为录音数据时，判定第一语音文本满足校正条件；若第一语句文本中的句号以及逗号的总数量小于预设数量，判定第一语句文本满足校正条件；若第一语音文本的来源为特定领域时，判定第一语音文本满足校正条件。预设数量可为任意合数的数值，例如预设数量为零，或者预设数量可为2。

[0106] 装置在确定第一语句文本不满足校正条件，则直接对第一语句文本进行转换得到第一语句文本对应的语句字符，也即执行步骤S10-步骤S30。

[0107] 若第一语句文本满足校正条件，则将第一语句文本发送至各个终端。各个终端可以为预设终端，预设终端的用户即为校正人员，校正人员对第一语句文本进行逗号以及句号的重新批注得到第二语句文本。装置接收到各个终端反馈的第二语句文本，装置再从各个第二语句文本中确定相同的第二语句文本，相同的第二语句文本即可视为相同批注的逗号以及句号的语句文本，也即为多数人所认同的语句文本的断句方式。装置将相同的第二语句文本作为目标语句文本，再对目标语句文本进行转换得到目标语句文本对应的语句字符，然后，装置确定预设字符在目标语句文本对应的语句字符中的第三位置，再根据第三位置确定预设符号在目标语句文本中的第四位置，最后将第四位置确定为目标语句文本的切分点。第三位置以及第四位置的确定参照第一位置以及第二位置的确定，在此不再进行赘述。

[0108] 在本实施例提供的技术方案中，装置在得到第一语句文本后，判断第一语句文本是否满足校正条件，从而根据校正条件准确的确定语句文本的切分点。

[0109] 参照图4，图4为本发明语句文本的切分方法的第三实施例，基于第一或第二实施

例,所述步骤S10包括:

[0110] 步骤S11,确定所述第一语句文本对应的向量;

[0111] 步骤S12,对所述向量进行特征提取得到所述第一语句文本中文字以及符号对应的特征值;

[0112] 步骤S13,根据各个所述特征值确定所述第一语句文本对应的语句字符。

[0113] 在本实施例中,装置可将第一语句文本转化为向量表示,也即第一语句文本转换为向量。装置对向量进行特征提取,从而得到各个特征值,每个特征值表征语句文本中的文字或符号。符号即为标点符号,也即一个特征值对应一个文字或者标点符号,向量进行特征提取,即可得到第一语句文本中文字以及符号对应的特征值。装置根据将各个特征值按照提取时间从早到晚进行排序即可得到由特征值构成的语句字符。例如,特征值为1时,表征该特征值对应的是逗号或者句号。

[0114] 在本实施例提供的技术方案中,装置通过向量表示第一语句文本,再对向量进行特征提取,从而得到各个特征值,最后根据各个特征值准确的得到第一语句文本对应的语句字符。

[0115] 参照图5,图5为本发明语句文本的切分方法的第四实施例,基于第一至第三种任一实施例,所述步骤S10包括:

[0116] 步骤S14,在获取第一语句文本后,根据转换模型对所述第一语句文本进行转换,得到所述第一语句文本对应的语句字符。

[0117] 在本实施例中,装置中可以包括转换模型,转换模型可包括BERT(Bidirectional Encoder Representations from Transformers)模型。装置将第一语句文本输入至转换模型中,得到第一语句文本对应的向量,再将向量输入转换模型的双层双向的神经网络LSTM(Long Short-Term Memory,时间循环神经网络)进行提取特征,得到第一语句文本的各个特征向量,各个特征向量再经过CRF层得到各个特征值,转换模型再将各个特征值构成语句字符,再将语句字符输出,使得装置得到语句字符。转换模型可为BERT+LSTM+CRF,或者,BERT+Pointer Net。

[0118] 进一步的,转换模型需要对语句文本进行转换,语句文本的长度不同的话,转换模型需要不断调整参数以适应语句文本的长度。对此,装置裁剪或补全第一语句文本,使得第一语句文本的长度为预设长度,也即对超过预设长度的第一语句文本进行裁剪,对未超过预设长度的第一语句文本进行补全。装置再对预设长度的第一语句文本设置句首标签,例如,将CLS作为句首标签;且对预设长度的第一语句文本设置句尾标签,从而得到第三语句文本,例如,将SEP作为句尾标签。装置再将第三语句文本输入转换模型从而得到转换模型输出的语句字符。通过对预设长度的第一语句文本设置句首标签以及句尾标签,使得转换模型在同时对大量的语句文本进行转换时,能够将各个语句文本进行区分。

[0119] 在本实施例提供的技术方案中,装置根据转换模型对第一语句文本进行转换,从而快速得到第一语句文本对应的语句字符。

[0120] 本发明还提供一种语句文本的切分装置。

[0121] 参照图6,图6为本发明语句文本的切分装置的功能模块示意图。

[0122] 如图6所示,所述音频识别设备包括:

[0123] 转换模块10,用于获取第一语句文本,对所述第一语句文本进行转换得到所述第

一语句文本对应的语句字符，其中，所述第一语句文本中的逗号以及句号转换为预设字符；

[0124] 确定模块20，用于确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号；

[0125] 确定模块20，用于将所述第二位置确定为所述第一语句文本的切分点。

[0126] 在一实施例中，语句文本的切分装置还包括：

[0127] 确定模块20，用于确定所述第一语句文本是否满足校正条件；

[0128] 执行模块，用于在所述第一语句文本不满足校正条件时，执行所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤。

[0129] 在一实施例中，所述语句文本的切分装置还包括：

[0130] 发送模块，用于在所述第一语句文本满足校正条件时，将所述第一语句文本发送至各个终端；

[0131] 接收模块，用于接收各个所述终端反馈的第二语句文本，并将相同的第二语句文本确定为目标语句文本，其中，所述第二语句文本为重新标注预设符号的所述第一语句文本；

[0132] 转换模块10，用于对所述目标语句文本进行转换得到所述目标语句文本对应的语句字符，并确定所述预设字符在所述目标语句文本对应的语句字符中的第三位置；

[0133] 确定模块20，用于根据所述第三位置确定预设符号在所述目标语句文本中的第四位置，并将所述第四位置确定为所述目标语句文本的切分点。

[0134] 在一实施例中，所述语句文本的切分装置还包括：

[0135] 确定模块20，用于确定所述第一语句文本对应的向量；

[0136] 提取模块，用于对所述向量进行特征提取得到所述第一语句文本中文字以及符号对应的特征值；

[0137] 确定模块20，用于根据各个所述特征值确定所述第一语句文本对应的语句字符。

[0138] 在一实施例中，所述语句文本的切分装置还包括：

[0139] 转换模块10，用于根据转换模型对所述第一语句文本进行转换，得到所述第一语句文本对应的语句字符。

[0140] 在一实施例中，所述语句文本的切分装置还包括：

[0141] 修改模块，用于裁剪或补全所述第一语句文本，以使所述第一语句文本的长度为预设长度；

[0142] 设置模块，用于对预设长度的第一语句文本的句首设置句首标签，且对预设长度的第一语句文本的句尾设置句尾标签，得到第三语句文本；

[0143] 输入模块，用于将所述第三语句文本输入转换模型，得到所述转换模型输出的所述语句字符。

[0144] 在一实施例中，所述语句文本的切分装置还包括：

[0145] 标记模块，用于对所述第一语句文本的切分点进行标记，以得到第四语句文本；

[0146] 训练模块，用于将各个所述第四语句文本作为训练样本，并根据各个所述训练样本对预设模型进行训练得到语句切分模型。

[0147] 其中，上述语句文本的切分装置中各个模块的功能实现与上述语句文本的切分方

法实施例中各步骤相对应，其功能和实现过程在此处不再一一赘述。

[0148] 本发明还提供一种计算机可读存储介质，该计算机可读存储介质上存储有切分程序，所述切分程序被处理器执行时实现如以上任一项实施例所述的语句文本的切分方法的步骤。

[0149] 本发明计算机可读存储介质的具体实施例与上述语句文本的切分方法各实施例基本相同，在此不作赘述。

[0150] 需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者系统不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

[0151] 上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。

[0152] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献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如上所述的一个存储介质(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，空调器，或者网络设备等)执行本发明各个实施例所述的方法。

[0153] 以上仅为本发明的优选实施例，并非因此限制本发明的专利范围，凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

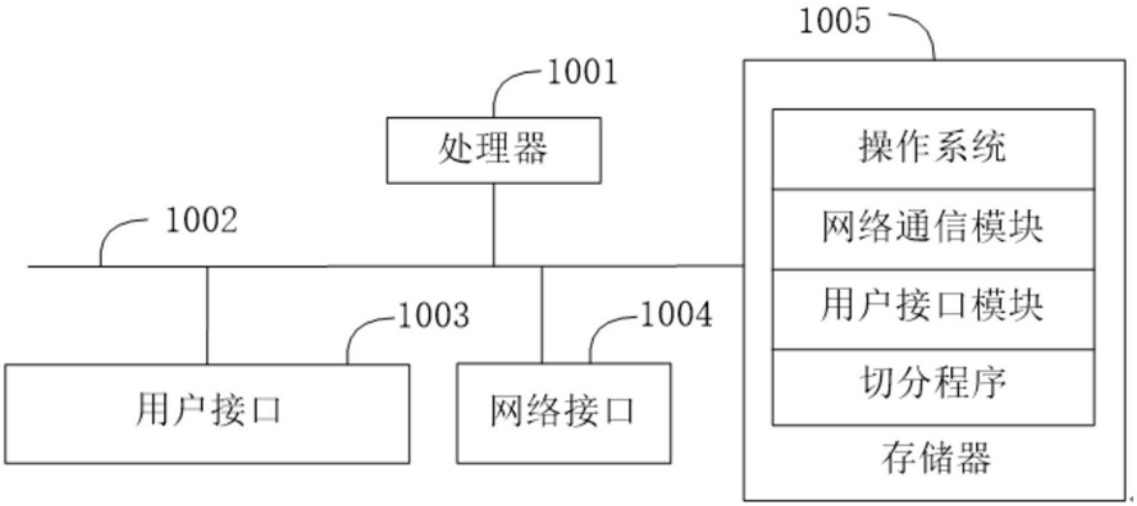

1005
1001
处理器
操作系统
网络通信模块
用户接口模块
切分程序
1002
存储器
1003
1004
用户接口
网络接口

图1

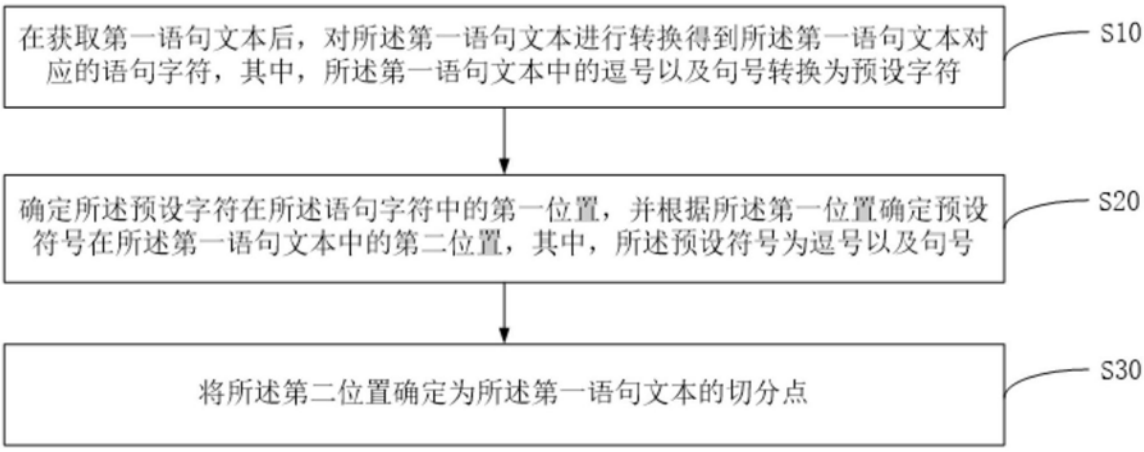

在获取第一语句文本后，对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符，其中，所述第一语句文本中的逗号以及句号转换为预设字符
S10
确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号
S20
将所述第二位置确定为所述第一语句文本的切分点
S30

图2

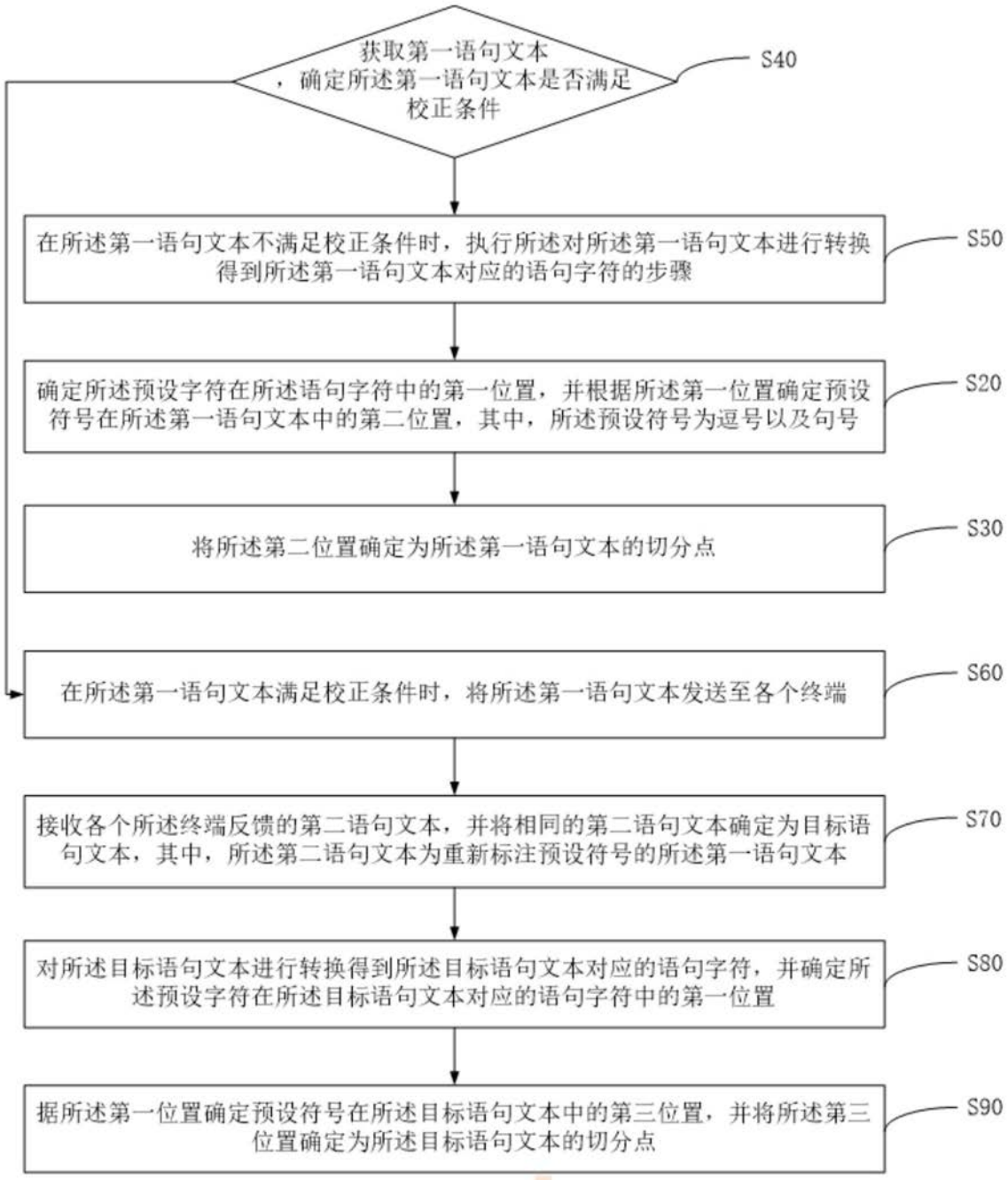

获取第一语句文本，确定所述第一语句文本是否满足校正条件
S40
在所述第一语句文本不满足校正条件时，执行所述对所述第一语句文本进行转换得到所述第一语句文本对应的语句字符的步骤
S50
确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号
S20
将所述第二位置确定为所述第一语句文本的切分点
S30
在所述第一语句文本满足校正条件时，将所述第一语句文本发送至各个终端
S60
接收各个所述终端反馈的第二语句文本，并将相同的第二语句文本确定为目标语句文本，其中，所述第二语句文本为重新标注预设符号的所述第一语句文本
S70
对所述目标语句文本进行转换得到所述目标语句文本对应的语句字符，并确定所述预设字符在所述目标语句文本对应的语句字符中的第一位置
S80
据所述第一位置确定预设符号在所述目标语句文本中的第三位置，并将所述第三位置确定为所述目标语句文本的切分点
S90

图3

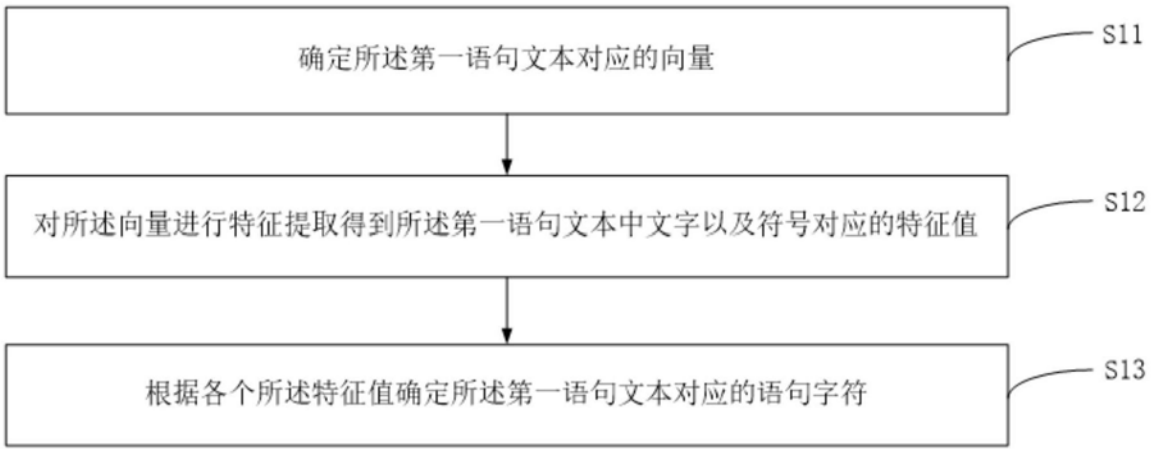

确定所述第一语句文本对应的向量
S11
对所述向量进行特征提取得到所述第一语句文本中文字以及符号对应的特征值
S12
根据各个所述特征值确定所述第一语句文本对应的语句字符
S13

图4

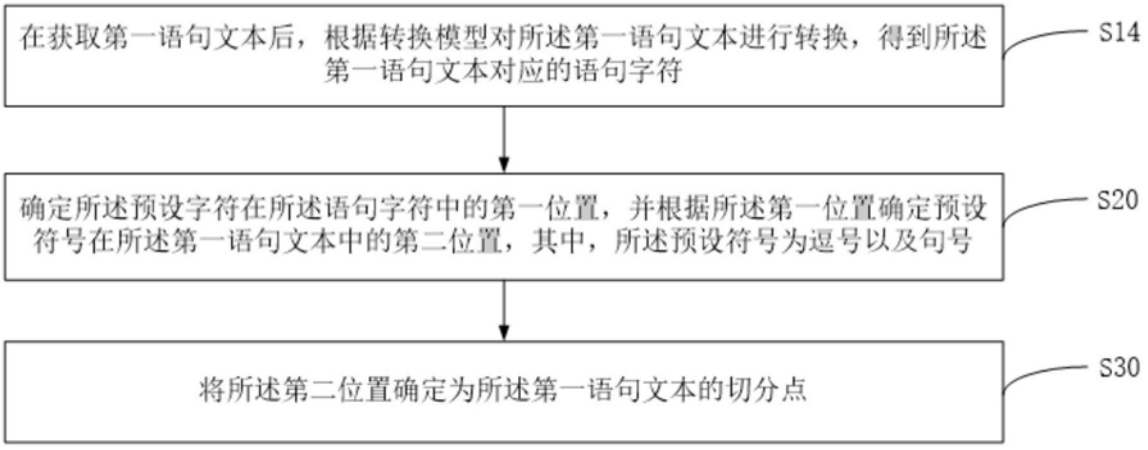

在获取第一语句文本后，根据转换模型对所述第一语句文本进行转换，得到所述第一语句文本对应的语句字符
S14
确定所述预设字符在所述语句字符中的第一位置，并根据所述第一位置确定预设符号在所述第一语句文本中的第二位置，其中，所述预设符号为逗号以及句号
S20
将所述第二位置确定为所述第一语句文本的切分点
S30

图5

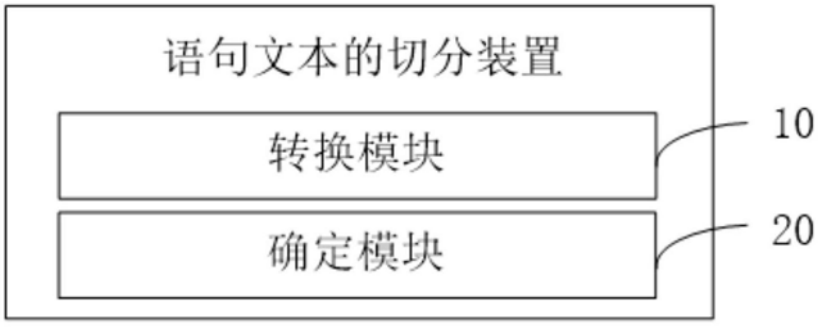

语句文本的切分装置
转换模块
10
确定模块
20

图6