(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 111613214 A
(43)申请公布日 2020.09.01

(21)申请号 202010437113.3

(22)申请日 2020.05.21

(71)申请人 重庆农村商业银行股份有限公司
地址 400000 重庆市江北区金沙门路36号

(72)发明人 秦邱川 刘引 卢华玮 杨声春 徐欣欣 魏鑫 田成志 汪哲逸 王璇

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 薛娇

(51)Int.Cl.
G10L 15/18(2013.01)
G10L 15/26(2006.01)
G10L 15/06(2013.01)
G06F 16/31(2019.01)
G06F 16/33(2019.01)
G06F 40/211(2020.01)

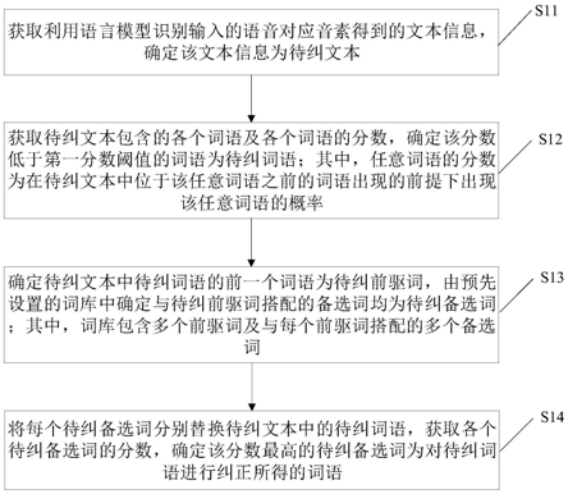
权利要求书2页 说明书10页 附图2页

(54)发明名称

一种用于提升语音识别能力的语言模型纠错方法

(57)摘要

本发明公开了一种用于提升语音识别能力的语言模型纠错方法、装置、设备及存储介质，该方法包括：获取识别语音得到的文本信息为待纠文本，获取待纠文本中各词语及各词语的分数，确定该分数低于对应阈值的词语为待纠词语；任意词语的分数为在待纠文本中位于该任意词语之前的词语出现前提下出现该任意词语的概率；确定待纠文本中待纠词语前一个词语为待纠前驱词，由词库中确定与待纠前驱词搭配的备选词为待纠备选词；词库包含多个前驱词及与各前驱词搭配的多个备选词；将各待纠备选词分别替换待纠文本中待纠词语，获取各待纠备选词的分数，确定该分数最高的待纠备选词为对待纠词语进行纠正所得词语。本申请能够提高语音识别准确性。



CN 111613214 A

1.一种用于提升语音识别能力的语言模型纠错方法,其特征在于,包括:

获取利用语言模型识别输入的语音对应音素得到的文本信息,确定该文本信息为待纠文本;

获取所述待纠文本包含的各个词语及各个词语的分数,确定该分数低于第一分数阈值的词语为待纠词语;其中,任意词语的分数为在所述待纠文本中位于该任意词语之前的词语出现的前提下出现该任意词语的概率;

确定所述待纠文本中所述待纠词语的前一个词语为待纠前驱词,由预先设置的词库中确定与所述待纠前驱词搭配的备选词均为待纠备选词;其中,所述词库包含多个前驱词及与每个所述前驱词搭配的多个备选词;

将每个所述待纠备选词分别替换所述待纠文本中的待纠词语,获取各个所述待纠备选词的分数,确定该分数最高的待纠备选词为对所述待纠词语进行纠正所得的词语。

2.根据权利要求1所述的方法,其特征在于,设置所述词库,包括:

通过依存句法分析在预先获取的语料中提取互相搭配的多对词语,确定每对词语中位于前面的词语为前驱词,每对词语中位于后面的词语为备选词;

获取每个所述备选词的分数,删除分数小于第二分数阈值的备选词,并将每个所述前驱词及每个所述前驱词对应的备选词均存储至预设的词库中;其中,任意备选词的分数为该任意备选词在对应的前驱词出现的前提下出现的概率。

3.根据权利要求2所述的方法,其特征在于,获取所述待纠文本包含的任意词语及任意所述待纠备选词及任意所述备选词的分数,包括:

确定所述待纠文本包含的任意词语或任意所述待纠备选词或任意所述备选词为待打分词语,将所述待打分词语所属的信息输入至预先训练得到的通用模型及定制模型中,将所述通用模型及所述定制模型输出的所述待打分词语的分数进行加权求和,得到所述待打分词语的分数;其中,所述待纠文本包含的任意词语及任意所述待纠备选词所属的信息为所述待纠文本,所述备选词所属的信息为所述备选词及对应的前驱词,所述通用模型为利用通用的文本信息训练得到的,所述定制模型为利用对应业务场景下符合该业务场景下用语规则的文本信息训练得到的,所述待纠文本为对在相应业务场景下输入的语音进行识别得到的文本信息。

4.根据权利要求3所述的方法,其特征在于,将每个所述前驱词及每个所述前驱词对应的备选词均存储至预设的词库中,包括:

将每个所述前驱词的备选词按照分数由高到底的顺序进行排列,并将每个所述前驱词及每个所述前驱词对应的备选词导入到哈希表中;其中,所述预设的词库为所述哈希表。

5.根据权利要求4所述的方法,其特征在于,如果所述待打分词语为所述待纠文本包含的任意词语或者任意待纠备选词,则所述通用模型及所述定制模型均包括2-gram模型及3-gram模型,如果所述待打分词语为任意备选词,则所述通用模型及所述定制模型均包括2-gram模型。

6.根据权利要求1所述的方法,其特征在于,将每个所述待纠备选词分别替换所述待纠文本中的待纠词语之前,还包括:

获取所述待纠词语及每个所述待纠备选词的拼音,对所述待纠词语的拼音及每个所述待纠备选词的拼音之间的编辑距离及最长公共子序列进行加权求和,得到每个所述待纠备

选词对应的分数，删除该分数小于第三分数阈值的待纠备选词。

7.根据权利要求6所述的方法，其特征在于，获取所述待纠文本包含的各个词语，包括：

去掉所述待纠文本中的标点，将所述待纠文本中的数字均用同一符号替换，对所述待纠文本进行断句分词处理，得到所述待纠文本包含的各个词语。

8.一种用于提升语音识别能力的语言模型纠错方法装置，其特征在于，包括：

第一确定模块，用于：获取利用语言模型识别输入的语音对应音素得到的文本信息，确定该文本信息为待纠文本；

第二确定模块，用于：获取所述待纠文本包含的各个词语及各个词语的分数，确定该分数低于第一分数阈值的词语为待纠词语；其中，任意词语的分数为在所述待纠文本中位于该任意词语之前的词语出现的前提下出现该任意词语的概率；

第三确定模块，用于：确定所述待纠文本中所述待纠词语的前一个词语为待纠前驱词，由预先设置的词库中确定与所述待纠前驱词搭配的备选词均为待纠备选词；其中，所述词库包含多个前驱词及与每个所述前驱词搭配的多个备选词；

获取模块，用于：将每个所述待纠备选词分别替换所述待纠文本中的待纠词语，获取各个所述待纠备选词的分数，确定该分数最高的待纠备选词为对所述待纠词语进行纠正所得的词语。

9.一种用于提升语音识别能力的语言模型纠错方法设备，其特征在于，包括：

存储器，用于存储计算机程序；

处理器，用于执行所述计算机程序时实现如权利要求1至7任一项所述用于提升语音识别能力的语言模型纠错方法的步骤。

10.一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计算机程序，所述计算机程序被处理器执行时实现如权利要求1至7任一项所述用于提升语音识别能力的语言模型纠错方法的步骤。

一种用于提升语音识别能力的语言模型纠错方法

技术领域

[0001] 本发明涉及语音识别技术领域，更具体地说，涉及一种用于提升语音识别能力的语言模型纠错方法、装置、设备及存储介质。

背景技术

[0002] 为了提升客户体验，当前许多行业都采用智能化设备响应客户发出的语音，实现相应的操作；在实现语音识别时，通常采用语音识别模型进行相应的语音识别，但是发明人发现，现有的技术方案在对语音进行识别得到相应的文本信息后，可能出现对语音识别得到的文本信息与语音所要表达的文本信息不一致的情况，进而导致语音识别的准确性较低。

发明内容

[0003] 本发明的目的是提供一种用于提升语音识别能力的语言模型纠错方法、装置、设备及存储介质，能够对语音识别得到的文本信息进行纠错，进而提高语音识别的准确性。

[0004] 为了实现上述目的，本发明提供如下技术方案：

[0005] 一种用于提升语音识别能力的语言模型纠错方法，包括：

[0006] 获取利用语言模型识别输入的语音对应音素得到的文本信息，确定该文本信息为待纠文本；

[0007] 获取所述待纠文本包含的各个词语及各个词语的分数，确定该分数低于第一分数阈值的词语为待纠词语；其中，任意词语的分数为在所述待纠文本中位于该任意词语之前的词语出现的前提下出现该任意词语的概率；

[0008] 确定所述待纠文本中所述待纠词语的前一个词语为待纠前驱词，由预先设置的词库中确定与所述待纠前驱词搭配的备选词均为待纠备选词；其中，所述词库包含多个前驱词及与每个所述前驱词搭配的多个备选词；

[0009] 将每个所述待纠备选词分别替换所述待纠文本中的待纠词语，获取各个所述待纠备选词的分数，确定该分数最高的待纠备选词为对所述待纠词语进行纠正所得的词语。

[0010] 优选的，设置所述词库，包括：

[0011] 通过依存句法分析在预先获取的语料中提取互相搭配的多对词语，确定每对词语中位于前面的词语为前驱词，每对词语中位于后面的词语为备选词；

[0012] 获取每个所述备选词的分数，删除分数小于第二分数阈值的备选词，并将每个所述前驱词及每个所述前驱词对应的备选词均存储至预设的词库中；其中，任意备选词的分数为该任意备选词在对应的前驱词出现的前提下出现的概率。

[0013] 优选的，获取所述待纠文本包含的任意词语及任意所述待纠备选词及任意所述备选词的分数，包括：

[0014] 确定所述待纠文本包含的任意词语或任意所述待纠备选词或任意所述备选词为待打分词语，将所述待打分词语所属的信息输入至预先训练得到的通用模型及定制模型

中，将所述通用模型及所述定制模型输出的所述待打分词语的分数进行加权求和，得到所述待打分词语的分数；其中，所述待纠文本包含的任意词语及任意所述待纠备选词所属的信息为所述待纠文本，所述备选词所属的信息为所述备选词及对应的前驱词，所述通用模型为利用通用的文本信息训练得到的，所述定制模型为利用对应业务场景下符合该业务场景下用语规则的文本信息训练得到的，所述待纠文本为对在相应业务场景下输入的语音进行识别得到的文本信息。

[0015] 优选的，将每个所述前驱词及每个所述前驱词对应的备选词均存储至预设的词库中，包括：

[0016] 将每个所述前驱词的备选词按照分数由高到底的顺序进行排列，并将每个所述前驱词及每个所述前驱词对应的备选词导入到哈希表中；其中，所述预设的词库为所述哈希表。

[0017] 优选的，如果所述待打分词语为所述待纠文本包含的任意词语或者任意待纠备选词，则所述通用模型及所述定制模型均包括2-gram模型及3-gram模型，如果所述待打分词语为任意备选词，则所述通用模型及所述定制模型均包括2-gram模型。

[0018] 优选的，将每个所述待纠备选词分别替换所述待纠文本中的待纠词语之前，还包括：

[0019] 获取所述待纠词语及每个所述待纠备选词的拼音，对所述待纠词语的拼音及每个所述待纠备选词的拼音之间的编辑距离及最长公共子序列进行加权求和，得到每个所述待纠备选词对应的分数，删除该分数小于第三分数阈值的待纠备选词。

[0020] 优选的，获取所述待纠文本包含的各个词语，包括：

[0021] 去掉所述待纠文本中的标点，将所述待纠文本中的数字均用同一符号替换，对所述待纠文本进行断句分词处理，得到所述待纠文本包含的各个词语。

[0022] 一种用于提升语音识别能力的语言模型纠错方法装置，包括：

[0023] 第一确定模块，用于：获取利用语言模型识别输入的语音对应音素得到的文本信息，确定该文本信息为待纠文本；

[0024] 第二确定模块，用于：获取所述待纠文本包含的各个词语及各个词语的分数，确定该分数低于第一分数阈值的词语为待纠词语；其中，任意词语的分数为在所述待纠文本中位于该任意词语之前的词语出现的前提下出现该任意词语的概率；

[0025] 第三确定模块，用于：确定所述待纠文本中所述待纠词语的前一个词语为待纠前驱词，由预先设置的词库中确定与所述待纠前驱词搭配的备选词均为待纠备选词；其中，所述词库包含多个前驱词及与每个所述前驱词搭配的多个备选词；

[0026] 获取模块，用于：将每个所述待纠备选词分别替换所述待纠文本中的待纠词语，获取各个所述待纠备选词的分数，确定该分数最高的待纠备选词为对所述待纠词语进行纠正所得的词语。

[0027] 一种用于提升语音识别能力的语言模型纠错方法设备，包括：

[0028] 存储器，用于存储计算机程序；

[0029] 处理器，用于执行所述计算机程序时实现如上任一项所述用于提升语音识别能力的语言模型纠错方法的步骤。

[0030] 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计算机程序，所述

计算机程序被处理器执行时实现如上任一项所述用于提升语音识别能力的语言模型纠错方法的步骤。

[0031] 本发明提供了一种用于提升语音识别能力的语言模型纠错方法、装置、设备及存储介质，该方法包括：获取利用语言模型识别输入的语音对应音素得到的文本信息，确定该文本信息为待纠文本获取所述待纠文本包含的各个词语及各个词语的分数，确定该分数低于第一分数阈值的词语为待纠词语；其中，任意词语的分数为在所述待纠文本中位于该任意词语之前的词语出现的前提下出现该任意词语的概率；确定所述待纠文本中所述待纠词语的前一个词语为待纠前驱词，由预先设置的词库中确定与所述待纠前驱词搭配的备选词均为待纠备选词；其中，所述词库包含多个前驱词及与每个所述前驱词搭配的多个备选词；将每个所述待纠备选词分别替换所述待纠文本中的待纠词语，获取各个所述待纠备选词的分数，确定该分数最高的待纠备选词为对所述待纠词语进行纠正所得的词语。本申请公开的技术方案中，在利用语言模型对输入的语音对应音素进行识别得到相应文本信息后，计算文本信息中包含的各词语在文本信息中位于其之前的各词语出现的前提下出现的概率，从而该概率较低的词语则为文本信息对应位置出现较不合理的词语，也即为需要进行纠错的词语；确定出在待纠文本中需要进行纠错的词语前一个词语，由词库中确定出可能位于该词语之后的全部备选词，进而将这些备选词分别替换文本信息后，再计算备选词在文本信息中位于其之前的各词语出现前提下出现的概率，从而该概率最高的词语则为文本信息对应位置出现最为合理的词语，因此利用该词语替换需要进行纠错的词语，实现对文本信息的纠错；可见，本申请通过上述技术方案能够有效实现对语音识别得到的文本信息的纠错，进而提高了语音识别的准确性。

附图说明

[0032] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据提供的附图获得其他的附图。

[0033] 图1为本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法的流程图；

[0034] 图2为本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法中依存句法分析树的示例图；

[0035] 图3为本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法装置的结构示意图。

具体实施方式

[0036] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，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，都属于本发明保护的范围。

[0037] 请参阅图1，其示出了本发明实施例提供的一种用于提升语音识别能力的语言模

型纠错方法的流程图，可以包括：

[0038] S11：获取利用语言模型识别输入的语音对应音素得到的文本信息，确定该文本信息为待纠文本。

[0039] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法的执行主体可以为对应的用于提升语音识别能力的语言模型纠错方法装置。在对输入的语音进行识别得到相应的文本信息后，可以确定该文本信息则为需要实现纠错的文本信息，也即待纠文本。其中，实现语音识别包括两个步骤，分别为利用语音模型识别语音的音素，及利用语言模型将音素转换成相应的文本信息；因此，对输入的语音进行识别得到相应的文本信息可以是将语音输入到预先训练得到的语音模型中，将语音模型输出的音素输入到预先训练得到的语言模型中，从而得到语言模型输出的文本信息则为与输入的语音对应的文本信息，当然也可以根据实际需要进行其他设定，均在本发明的保护范围之内。

[0040] S12：获取待纠文本包含的各个词语及各个词语的分数，确定该分数低于第一分数阈值的词语为待纠词语；其中，任意词语的分数为在待纠文本中位于该任意词语之前的词语出现的前提下出现该任意词语的概率。

[0041] 对待纠文本进行预处理，可以得到待纠文本中包含的全部词语；求得每个词语在待纠文本中的后验概率作为每个词语的分数，具体来说，对于任意词语，在待纠文本中位于该任意词语之前的全部词语出现的前提下，在待纠文本中该任意词语的位置处出现该任意词语的概率，从而通过该概率可以确定每个词语在待纠文本中对应位置处出现的合理性，也即该概率越高，对应词语在待纠文本中对应位置处出现越合理，越可能对相应的部分语音的识别是正确的，反之则对应词语在待纠文本中对应位置处出现越不合理，越可能对相应的部分语音的识别是错误的。因此，本实施例中可以预先根据实际需要设定第一分数阈值(如默认-5)，如果任意词语的分数低于第一分数阈值，则该任意词语在待纠文本中对应位置处出现较不合理，对相应的部分语音的识别是错误的可能性较大，因此认为该任意词语为识别错误的词语，也即待纠词语，待纠词语的数量可能为1个，也可能为多个，如果待纠词语的数量为多个，则对于每个待纠词语均需按照本申请公开的技术方案进行纠错；如果待纠文本中不包含分数低于第一分数阈值的词语，则认为对输入的语音的识别所得文本信息是正确表达输入的语音所要表达信息的文本信息，因此可以确定无需进行后续纠错步骤。

[0042] S13：确定待纠文本中待纠词语的前一个词语为待纠前驱词，由预先设置的词库中确定与待纠前驱词搭配的备选词均为待纠备选词；其中，词库包含多个前驱词及与每个前驱词搭配的多个备选词。

[0043] 本实施例中可以预先配置有词库，词库中包含多个前驱词，每个前驱词具有一个备选词集，每个备选词集包含多个备选词；具体来说，每个前驱词中任意前驱词与该任意前驱词具有的备选词中的每个备选词均具有对应关系，每个前驱词与对应的每个备选词均可以组成一组互相搭配的词语，也即任意前驱词与该任意前驱词对应的任意备选词可以组成一组互相搭配的词语，且在组成互相搭配的词语后该任意前驱词位于前面，该任意前驱词对应的备选词位于后面。在确定出待纠词语后，可以确定待纠文本中位于待纠词语之前且距离待纠词语最近的一个词语为待纠词语的前驱词、即待纠前驱词，从词库中寻找与该待纠前驱词相同的前驱词，并且确定词库中与该待纠前驱词相同的前驱词具有的备选词集中

各备选词均为与该待纠前驱词对应的备选词，也即可能与该待纠前驱词组成一组互相搭配的词语的词语。

[0044] S14：将每个待纠备选词分别替换待纠文本中的待纠词语，获取各个待纠备选词的分数，确定该分数最高的待纠备选词为对待纠词语进行纠正所得的词语。

[0045] 在得到各个待纠备选词后，可以将每个待纠备选词分别替换待纠文本中的待纠词语，然后获取每个待纠备选词的分数，也即在待纠文本中位于任意待纠备选词之前的词语出现的前提下在该任意待纠备选词的位置处出现该任意待纠备选词的概率，则为该任意待纠备选词的分数；分数最高的待纠备选词则为在相应位置处出现的合理性及可能性最高的词语，也即为相应的部分语音所表达的真实的文本，因此分数最高的待纠备选词替换待纠词语后，所得的文本则为纠错完成后的文本信息。

[0046] 本申请公开的技术方案中，在利用语言模型对输入的语音对应音素进行识别得到相应文本信息后，计算文本信息中包含的各词语在文本信息中位于其之前的各词语出现的前提下出现的概率，从而该概率较低的词语则为文本信息对应位置出现较不合理的词语，也即为需要进行纠错的词语；确定出在待纠文本中需要进行纠错的词语前一个词语，由词库中确定出可能位于该词语之后的全部备选词，进而将这些备选词分别替换文本信息后，再计算备选词在文本信息中位于其之前的各词语出现前提下出现的概率，从而该概率最高的词语则为文本信息对应位置出现最为合理的词语，因此利用该词语替换需要进行纠错的词语，实现对文本信息的纠错；可见，本申请通过上述技术方案能够有效实现对语音识别得到的文本信息的纠错，进而提高了语音识别的准确性。

[0047] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法，设置词库，可以包括：

[0048] 通过依存句法分析在预先获取的语料中提取互相搭配的多对词语，确定每对词语中位于前面的词语为前驱词，每对词语中位于后面的词语为备选词；

[0049] 获取每个备选词的分数，删除分数小于第二分数阈值的备选词，并将每个前驱词及每个前驱词对应的备选词均存储至预设的词库中；其中，任意备选词的分数为该任意备选词在对应的前驱词出现的前提下出现的概率；其中，任意备选词的分数为该任意备选词在对应的前驱词出现的前提下出现的概率。

[0050] 其中，预先获取的语料可以为由工作人员预先获取的语料，通过依存句法分析在这些语料中提取到词语搭配，也即互相搭配的多对词语(每对词语中的前驱词及备选词之间具有对应的关系)，确定每对互相搭配的词语中位于前面的为前驱词，位于后面的为备选词，获取到每个备选词在对应的前驱词之后出现的概率，如果该概率小于根据实际需要设定的第二分数阈值，则说明对应备选词在其前驱词后出现的概率是较小的，因此这种备选词则可以直接删除，从而保证备选词均为可靠性较高的备选词。

[0051] 其中，依存句法分析通过分析语料包含的句子中每一个语法成分之间的依存关系，分析出其句法结构，也即将句子中的“主谓宾”、“定状补”等语法成分之间的关系描述清楚。本实施例中可以使用stanford的nlp工具对句子进行语法标注，从而得到各个词语的搭配关系，也就是依存关系。例如句子“会议宣布了首批资深院士名单。”的依存句法分析树如图2所示，从图2可以看出，词语“宣布”支配“会议”、“了”和“名单”，因此可以将这些被支配的词语作为“宣布”的搭配词语，也即与“宣布”互相搭配的词语。

[0052] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法，获取待纠文本包含的任意词语及任意待纠备选词及任意备选词的分数，可以包括：

[0053] 确定待纠文本包含的任意词语或任意待纠备选词或任意备选词为待打分词语，将待打分词语所属的信息输入至预先训练得到的通用模型及定制模型中，将通用模型及定制模型输出的待打分词语的分数进行加权求和，得到待打分词语的分数；其中，待纠文本包含的任意词语及任意待纠备选词所属的信息为待纠文本，备选词所属的信息为备选词及对应的前驱词，通用模型为利用通用的文本信息训练得到的，定制模型为利用对应业务场景下符合该业务场景下用语规则的文本信息训练得到的，待纠文本为对在相应业务场景下输入的语音进行识别得到的文本信息。

[0054] 本实施例中的通用模型及定制模型均为语言模型，其可以利用相应的文本信息训练得到。其中，本申请实施例提供的技术方案可以应用于金融行业，当然也可以应用于其他需要办理业务且可支持语音交互实现业务办理的行业，如通信行业、采购行业等，均在本发明的保护范围之内。本申请实施例将对语音进行识别得到相应的文本信息所用的语音识别模型包括语音模型及语言模型，在训练该语音识别模型时，可以是先用在任意业务场景下客户输入的任意语音(通用的语音)及对应文本信息训练得到通用的语音识别模型，再利用每个业务场景下客户输入的任意语音及对应文本信息对通用的语音识别模型进行训练，得到与每个业务场景分别一一对应的定制的语音识别模型，进而利用定制的语音识别模型对相应业务场景下输入的语音进行识别得到相应文本信息，也即本申请实施例中的待纠文本；对应的，本申请实施例中用于识别词语打分的模型也包括用通用的文本信息(对应任意的业务场景)训练得到的通用模型及用每个业务场景下的文本信息训练得到的定制模型，其中，业务场景可以为具有自己的用于规则的场景，也可以为包含指定方言(如重庆话)的场景，还可以特指某领域(如金融)的场景；从而通过这两种模型综合实现词语打分，提高了词语打分的准确性。

[0055] 另外，如果通用模型所用语料为标准语言，而业务场景指指定方言的场景，则此时对业务场景下的语音识别得到的文本信息很可能因指定方言的发音而导致文本信息不通顺和覆盖范围不广泛，因此本实施例通过通用模型和定制模型多个角度评估识别文本信息的准确性，进而提高了语音识别准确性。

[0056] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法，将每个前驱词及每个前驱词对应的备选词均存储至预设的词库中，可以包括：

[0057] 将每个前驱词的备选词按照分数由高到底的顺序进行排列，并将每个前驱词及每个前驱词对应的备选词导入到哈希表中；预设的词库为哈希表。

[0058] 在得到每个前驱词的备选词后，可以将任意前驱词对应的多个备选词按照分数从高到低的顺序进行排列，将全部前驱词及对应的全部备选词保存为词语搭配文件，进而将该词语搭配文件导入到哈希表中，从而能够便于通过哈希表实现前驱词及对应备选词的查询。其中，将词语搭配文件导入到哈希表中后保存格式可以如(其中，词1至词n均为前驱词的编号)：

[0059] 【词1，前驱词】-【备选词11】【备选词12】….

[0060] 【词2，前驱词】-【备选词21】【备选词22】….

[0061] ...

[0062] 【词n,前驱词】-【备选词n1】【备选词n2】….

[0063] 按照上述形式将前驱词及对应备选词一一保存,能够在获知前驱词后查询其对应的所有的备选词。

[0064] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法,如果待打分词语为待纠文本包含的任意词语或者任意待纠备选词,则通用模型及定制模型均可以包括2-gram模型及3-gram模型,如果待打分词语为任意备选词,则通用模型及定制模型均可以包括2-gram模型。

[0065] 其中,N-Gram是一种基于统计语言模型的算法,它的基本思想是将文本里的内容按照字节进行大小为N的滑动窗口操作,形成了长度是N的字节片段序列,每一个字节片段称为gram,对所有gram的出现频度进行统计,并且按照事先设定好的阈值进行过滤,形成关键gram列表,也即这个文本的向量特征空间,列表中的每一种gram就是一个特征向量维度;该模型基于这样一种假设,第N个词语的出现只与前面N-1个词语相关,而与其它任何词语都不相关,整句的概率就是各个词语出现概率的乘积,这些概率可以通过直接从语料中统计N个词语同时出现的次数得到;常用的是二阶的Bi-Gram和三阶的Tri-Gram,本实施例中采用的即为2-gram及3-gram。

[0066] 以待纠文本为:“系统提示查询密码不正确”进行说明,将待纠文本进行分词,需要输入至2-gram的词语对列表则是:[系统,提示]、[提示,查询]、[查询,密码]、[密码,不]、[不,正确],同理需要输入至3-gram的词语对列表是:[系统,提示,查询]、[提示,查询,密码]、[查询,密码,不],[密码,不,正确];分别通用模型2-gram、3-gram和定制模型2-gram、3-gram四个模型对分词后的文本进行打分,也即将一个分好词的文本整体输入到上述模型后即可得到该文本中每一个词语的分数。

[0067] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法,将每个待纠备选词分别替换待纠文本中的待纠词语之前,还可以包括:

[0068] 获取待纠词语及每个待纠备选词的拼音,对待纠词语的拼音及每个待纠备选词的拼音之间的编辑距离及最长公共子序列进行加权求和,得到每个待纠备选词对应的分数,删除该分数小于第三分数阈值的待纠备选词。

[0069] 其中,编辑距离是衡量两个拼音的字符串间的差异程度的一种方法,从一个字符串修改到另一个字符串时,编辑单个字符(比如修改、插入、删除)所需要的最少次数则为两个字符串之间的编辑哭距离;可以采用动态规划算法实现,具体来说,设字符串S、T的长度分别为m、n,记S(i)为S从第1个字符到第i个字符之间的子串,S(0)表示空串,S(m)表示S本身,因此,S和T间的编辑距离,可由S(i)和T(j)的编辑距离计算而来,取编辑距离值为S1;则递推公式如下:

[0070] $$Edit(i,j)=\begin{cases} i & \text{if } j=0 \\ j & \text{if } i=0 \\ \min\begin{cases} Edit(i-1,j)+1, \\ Edit(i,j-1)+1, \\ Edit(i-1,j-1)+[A[i]\neq B[j]] \end{cases} & \text{otherwise} \end{cases}$$

[0071] 例如从字符串“kitten”修改为字符串“sitting”只需3次单字符编辑操作,具体如

下：

[0072] sitten(k->s)

[0073] sittin(e->i)

[0074] sitting(_->g)；

[0075] 因此“kitten”和“sitting”的逻辑距离(Levenshtein)距离为3。

[0076] LCS(最长公共子序列)是指两个拼音的字符串的最长公共子序列，也即两个字符串中最长的有相同顺序的子序列。

[0077] 第三分数阈值可以根据实际需要进行设定，从而通过编辑距离和LCS的计算得到最大可能的正确文本序列，进而提高语音识别准确性。

[0078] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法，获取待纠文本包含的各个词语，可以包括：

[0079] 去掉待纠文本中的标点，将待纠文本中的数字均用同一符号替换，对待纠文本进行断句分词处理，得到待纠文本包含的各个词语。

[0080] 获取待纠文本中包含的各个词语时，具体可以包括去掉待纠文本中包含的各个标点符合，待纠文本中可能会包含各种数字，比如“2020年3月29日”，其中2020和3和29这种数字不需要考虑其具体是什么数值，可以是任意数字，只要是模式相同的数字，为了去除不同数字相同模式的影响，可以将待纠文本中的数字用同一符号(如星号等)一一对应的替换，具体用正则表达式匹配替换即可，最后再对待纠文本进行分居分词处理，得到其包含的全部词语，从而以这种方式快速方便的实现待纠文本中词语的提取。

[0081] 需要说明的是，在得到待纠词语后，如果待纠词语为多个，则可以将待纠词语均保存为JSON文件，在JSON文件以中括号[]嵌套的形式记录每个大句子(待纠文本)的标识(如id)，大句子分割成的小句子的错误词语索引及词语本身，以及词语的分数等，用于后续的纠错处理。

[0082] 在确定词库中与待纠前驱词相同的前驱词时，可以通过开源工具hanlp获取待纠词语以及词库中各词语的拼音，根据待纠文本以及待纠词语的索引获取待纠词语的前一个词语为待纠前驱词，查询词库得到待纠前驱词对应的备选词，其中，查询得到对应备选词时可以是通过哈希表，已知待纠词语的前一个词语的下标位置后，可以查询得到所有备选词的集合。另外，本申请实施例中加权求和的部分也可以是求加权平均值，当然也可以根据实际需要进行其他设定，均在本发明的保护范围之内。

[0083] 本申请通过通用模型和定制模型的2-gram、3-gram两个领域多个角度评估识别文本的准确性，并通过编辑距离和LCS的计算得到最大可能的正确文本序列，进而通过n-gram进行多重验证，最终选择最优备选词，将错误的词语替换成新的更可靠的词语，完成纠错，使得错误的词语所在的文本句子变的更加通顺，更加正确。

[0084] 本发明实施例还提供了一种用于提升语音识别能力的语言模型纠错方法装置，如图3所示，可以包括：

[0085] 第一确定模块11，用于：获取利用语言模型识别输入的语音对应音素得到的文本信息，确定该文本信息为待纠文本；

[0086] 第二确定模块12，用于：获取待纠文本包含的各个词语及各个词语的分数，确定该分数低于第一分数阈值的词语为待纠词语；其中，任意词语的分数为在待纠文本中位于该

任意词语之前的词语出现的前提下出现该任意词语的概率；

[0087] 第三确定模块13，用于：确定待纠文本中待纠词语的前一个词语为待纠前驱词，由预先设置的词库中确定与待纠前驱词搭配的备选词均为待纠备选词；其中，词库包含多个前驱词及与每个前驱词搭配的多个备选词；

[0088] 获取模块14，用于：将每个待纠备选词分别替换待纠文本中的待纠词语，获取各个待纠备选词的分数，确定该分数最高的待纠备选词为对待纠词语进行纠正所得的词语。

[0089] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法装置，还可以包括：

[0090] 设置模块，用于：通过依存句法分析在预先获取的语料中提取互相搭配的多对词语，确定每对词语中位于前面的词语为前驱词，每对词语中位于后面的词语为备选词；获取每个备选词的分数，删除分数小于第二分数阈值的备选词，并将每个前驱词及每个前驱词对应的备选词均存储至预设的词库中；其中，任意备选词的分数为该任意备选词在对应的前驱词出现的前提下出现的概率。

[0091] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法装置，第二确定模块、获取模块及设置模块均可以包括：

[0092] 确定单元，用于：确定待纠文本包含的任意词语或任意待纠备选词或任意备选词为待打分词语，将待打分词语所属的信息输入至预先训练得到的通用模型及定制模型中，将通用模型及定制模型输出的待打分词语的分数进行加权求和，得到待打分词语的分数；其中，待纠文本包含的任意词语及任意待纠备选词所属的信息为待纠文本，备选词所属的信息为备选词及对应的前驱词，通用模型为利用通用的文本信息训练得到的，定制模型为利用对应业务场景下符合该业务场景下用语规则的文本信息训练得到的，待纠文本为对在相应业务场景下输入的语音进行识别得到的文本信息。

[0093] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法装置，设置模块可以包括：

[0094] 导入单元，用于：将每个前驱词的备选词按照分数由高到底的顺序进行排列，并将每个前驱词及每个前驱词对应的备选词导入到哈希表中；其中，预设的词库为哈希表。

[0095] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法装置，还可以包括：

[0096] 筛选模块，用于：将每个待纠备选词分别替换待纠文本中的待纠词语之前，获取待纠词语及每个待纠备选词的拼音，对待纠词语的拼音及每个待纠备选词的拼音之间的编辑距离及最长公共子序列进行加权求和，得到每个待纠备选词对应的分数，删除该分数小于第三分数阈值的待纠备选词。

[0097] 本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法装置，第二确定模块可以包括：

[0098] 预处理单元，用于：去掉待纠文本中的标点，将待纠文本中的数字均用同一符号替换，对待纠文本进行断句分词处理，得到待纠文本包含的各个词语。

[0099] 本发明实施例还提供了一种用于提升语音识别能力的语言模型纠错方法设备，可以包括：

[0100] 存储器，用于存储计算机程序；

[0101] 处理器，用于执行计算机程序时实现如上任一项用于提升语音识别能力的语言模型纠错方法的步骤。

[0102] 本发明实施例还提供了一种计算机可读存储介质，计算机可读存储介质上存储有计算机程序，计算机程序被处理器执行时可以实现如上任一项用于提升语音识别能力的语言模型纠错方法的步骤。

[0103] 需要说明的是，本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法装置、设备及存储介质中相关部分的说明请参见本发明实施例提供的一种用于提升语音识别能力的语言模型纠错方法中对应部分的详细说明，在此不再赘述。另外，本发明实施例提供的上述技术方案中与现有技术中对应技术方案实现原理一致的部分并未详细说明，以免过多赘述。

[0104] 对所公开的实施例的上述说明，使本领域技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

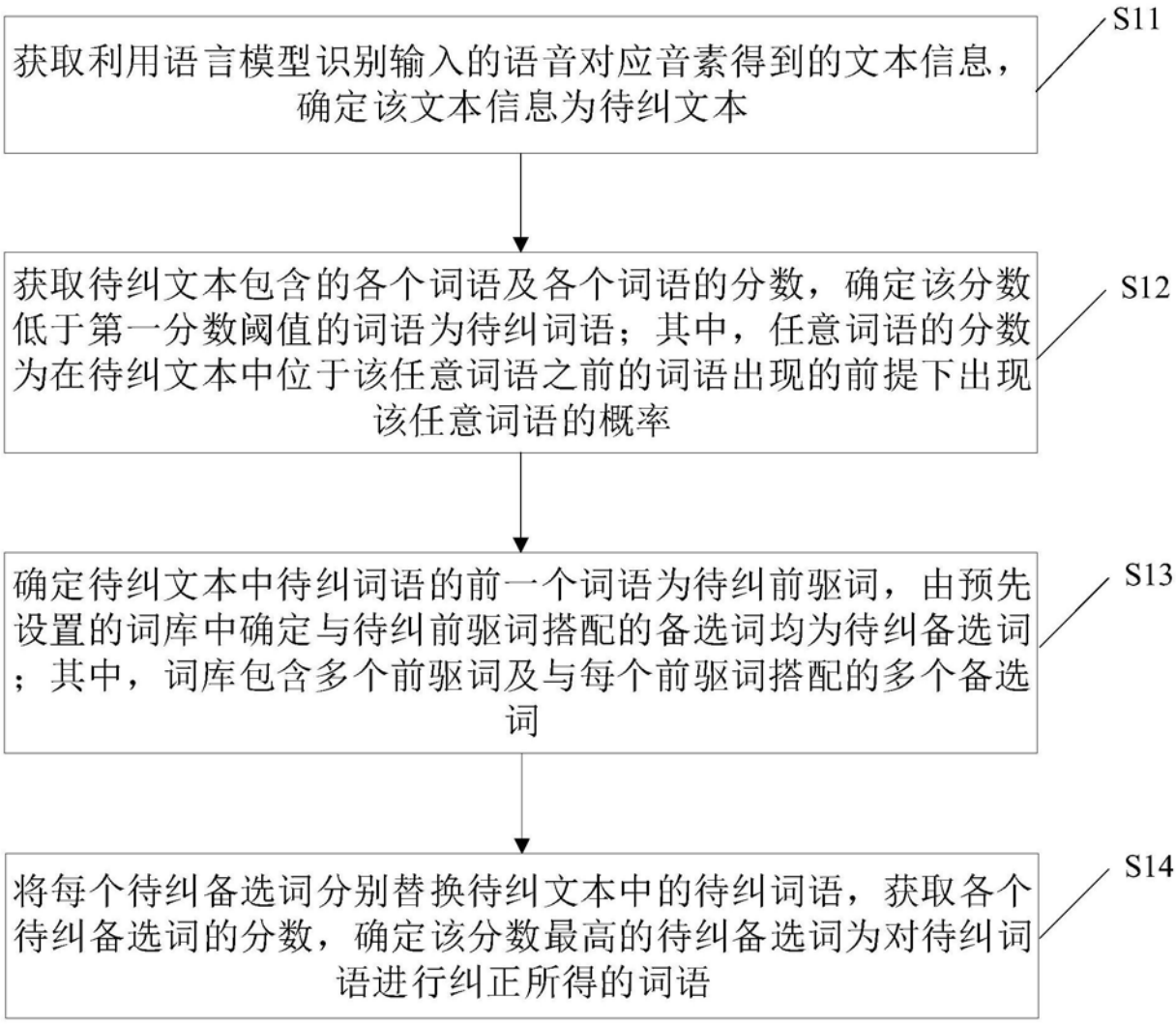

获取利用语言模型识别输入的语音对应音素得到的文本信息，确定该文本信息为待纠文本
S11
获取待纠文本包含的各个词语及各个词语的分数，确定该分数低于第一分数阈值的词语为待纠词语；其中，任意词语的分数为在待纠文本中位于该任意词语之前的词语出现的前提下出现该任意词语的概率
S12
确定待纠文本中待纠词语的前一个词语为待纠前驱词，由预先设置的词库中确定与待纠前驱词搭配的备选词均为待纠备选词；其中，词库包含多个前驱词及与每个前驱词搭配的多个备选词
S13
将每个待纠备选词分别替换待纠文本中的待纠词语，获取各个待纠备选词的分数，确定该分数最高的待纠备选词为对待纠词语进行纠正所得的词语
S14

图1

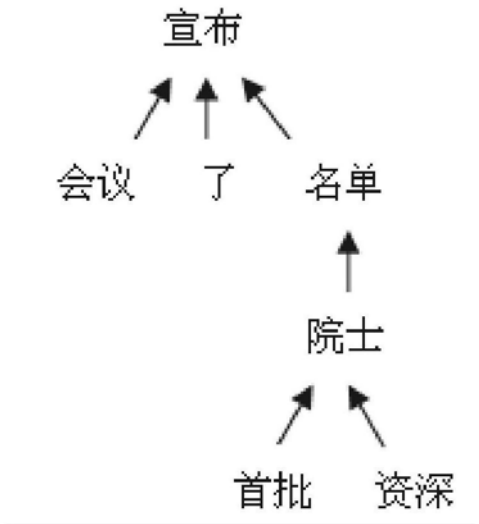

宣布
会议
了
名单
院士
首批
资深

图2

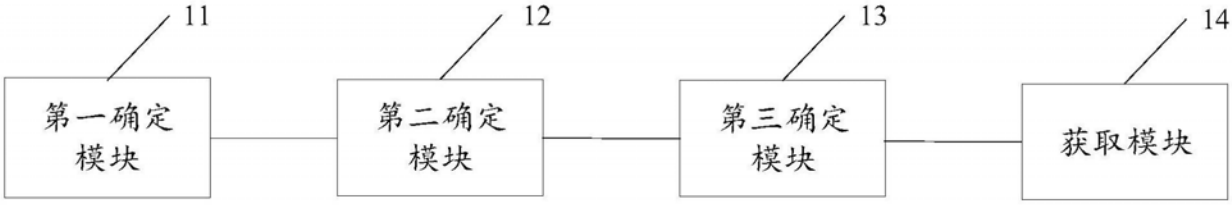


图3