



(12) 发明专利申请

(10) 申请公布号 CN 114220417 A

(43) 申请公布日 2022. 03. 22

(21) 申请号 202111506578.0

G06N 3/08 (2006.01)

(22) 申请日 2021.12.10

(71) 申请人 京东科技信息技术有限公司

地址 100176 北京市大兴区北京经济技术
开发区科创十一街18号院2号楼6层
601

(72) 发明人 冯明超 陈蒙 乐雨泉 赵宇明
王泽勋

(74) 专利代理机构 北京路浩知识产权代理有限
公司 11002

代理人 任少瑞

(51) Int. Cl.

G10L 15/02 (2006.01)

G10L 15/06 (2013.01)

G10L 15/18 (2013.01)

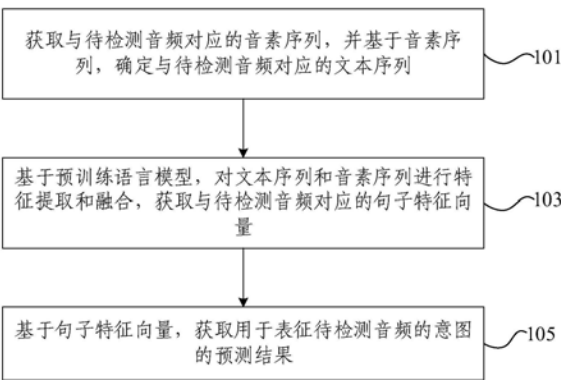
权利要求书2页 说明书11页 附图6页

(54) 发明名称

一种意图识别方法、装置及相关设备

(57) 摘要

本公开提供一种意图识别方法、装置及相关设备,其方法包括:获取与待检测音频对应的音素序列,并基于所述音素序列,确定与所述待检测音频对应的文本序列;基于预训练语言模型,对所述文本序列和所述音素序列进行特征提取和融合,获取与所述待检测音频对应的句子特征向量;基于所述句子特征向量,获取用于表征所述待检测音频的意图的预测结果,该方法能够有效提高口语意图识别准确率。



1. 一种意图识别方法,其特征在于,包括:

获取与待检测音频对应的音素序列,并基于所述音素序列,确定与所述待检测音频对应的文本序列;

基于预训练语言模型,对所述文本序列和所述音素序列进行特征提取和融合,获取与所述待检测音频对应的句子特征向量;

基于所述句子特征向量,获取用于表征所述待检测音频的意图的预测结果。

2. 根据权利要求1所述的意图识别方法,其特征在于,所述基于所述句子特征向量,获取用于表征所述待检测音频的意图的预测结果具体包括:

输入所述句子特征向量至全连接神经网络进行分类预测,获取所述预测结果。

3. 根据权利要求1所述的意图识别方法,其特征在于,

所述对所述文本序列和所述音素序列进行特征提取和融合,获取与所述待检测音频对应的句子特征向量包括:

输入所述文本序列和所述音素序列至所述预训练语言模型,获取与所述待检测音频对应的鲁棒性词向量表征;

基于所述鲁棒性词向量表征,获取所述预训练语言模型输出的并与所述待检测音频对应的句子特征向量。

4. 根据权利要求1所述的意图识别方法,其特征在于,所述预训练语言模型通过如下方式训练确定,包括:

将真实文本序列和真实音素序列输入至初始预训练语言模型中,调整并确定第一预训练语言模型网络参数;其中,所述真实文本序列为从互联网中获取的文本序列;所述真实音素序列根据所述真实文本序列获得;所述第一预训练语言模型网络参数用于表征所述真实文本序列与所述真实音素序列发音关系;

将样本文本序列和对应的样本音素序列,输入至网络参数为所述第一预训练语言模型网络参数的预训练语言模型,获得第二预训练语言模型网络参数;其中,所述样本文本序列和对应的样本音素序列通过对预设的多个样本音频进行语音识别得到;

基于所述第二预训练语言模型网络参数,构建具有泛化能力的所述预训练语言模型。

5. 根据权利要求4所述的意图识别方法,其特征在于,

所述将真实文本序列和真实音素序列输入至初始预训练语言模型中,调整并确定第一预训练语言模型网络参数进一步包括:

将所述真实文本序列和所述真实音素序列,输入至所述初始预训练语言模型,对所述真实文本序列和所述真实音素序列进行掩码,获取第一掩码文本序列和第一掩码音素序列,通过上下文预测被掩码的所述第一掩码文本序列和所述第一掩码音素序列,调整并确定所述第一预训练语言模型网络参数;

其中,所述第一掩码文本序列为从所述真实文本序列中随机选择的20%的文本序列;所述第一掩码音素序列为从所述真实音素序列随机选择的20%的音素序列。

6. 根据权利要求4或5所述的意图识别方法,其特征在于,

所述将样本文本序列和对应的样本音素序列输入至网络参数为第一预训练语言模型网络参数预训练语言模型,构建第二预训练语言模型网络参数进一步包括:

将所述样本文本序列和所述样本音素序列,输入至网络参数为第一预训练语言模型网

络参数预训练语言模型,对所述样本文本序列和所述样本音素序列进行掩码,得到第二掩码文本序列和第二掩码音素序列,通过上下文预测被掩码的所述第二掩码文本序列和所述第二掩码音素序列,获得所述第二预训练语言模型网络参数;

其中,所述第二掩码文本序列为从所述样本文本序列中随机选择的20%的文本序列;所述第二掩码音素序列为从所述样本音素序列随机选择的20%的音素序列。

7.根据权利要求4所述的意图识别方法,其特征在于,所述真实音素序列根据所述真实文本序列获得具体包括:

对所述真实文本序列进行分词,得到所述真实字;

根据音素对照表,获取所述真实字的音素;

基于所有所述真实字的音素,获得所述真实音素序列。

8.根据权利要求6所述的意图识别方法,其特征在于,所述第二预训练语言模型网络参数通过如下方式进行调整,具体包括:

确定由人工听取的所述样本音频转写的真实文本;

基于交叉熵损失函数,计算表征所述样本音频的意图的预测结果与所述真实文本的差异;

基于所述差异,调整所述第一预训练语言模型网络参数,得到第二预训练语言模型网络参数。

9.一种意图识别装置,其特征在于,包括:

语音识别模块,用于获取与待检测音频对应的音素序列,并基于所述音素序列,确定与所述待检测音频对应的文本序列;

预训练语言模块,基于预训练语言模型,用于对所述文本序列和所述音素序列进行特征提取和融合,获取与所述待检测音频对应的句子特征向量;

全连接神经网络分类模块,用于基于所述句子特征向量,获取用于表征所述待检测音频的意图的预测结果。

10.一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,其特征在于,所述处理器执行所述程序时实现如权利要求1至8任一项所述意图识别方法的步骤。

11.一种非暂态计算机可读存储介质,其上存储有计算机程序,其特征在于,该计算机程序被处理器执行时实现如权利要求1至8任一项所述意图识别方法的步骤。

一种意图识别方法、装置及相关设备

技术领域

[0001] 本公开涉及智能语音技术领域,尤其涉及一种意图识别方法、装置及相关设备。

背景技术

[0002] 当前的智能语音对话系统主要由语音识别、口语理解、对话管理、对话生成以及语音合成五部分生成,而口语理解主要由领域识别,意图识别,槽位填充三部分组成。意图识别通常取语音识别模块输出的top1文本进行分类预测,提升分类预测准确率的手段主要有扩充训练数据、采用更好的分类模型、加入预训练等。预训练作为增强文本表征的常用手段,主要是通过建模句子中某个字与其周围上下文之间的关系从而获取相应的词向量。然而,面对语音识别错误的同音字和近音字时,应用现有的意图识别方法中会识别出错误意图,口语意图识别准确率较低。如何提高口语意图识别准确率亟待解决。

发明内容

[0003] 本公开提供一种意图识别方法、装置及相关设备,用以解决现有技术中面对语音识别错误的同音字和近音字时,应用现有的意图识别方法中会识别出错误意图的缺陷,实现提高口语意图识别准确率。

[0004] 第一方面,本公开提供一种意图识别方法,包括:获取与待检测音频对应的音素序列,并基于所述音素序列,确定与所述待检测音频对应的文本序列;基于预训练语言模型,对所述文本序列和所述音素序列进行特征提取和融合,获取与所述待检测音频对应的句子特征向量;基于所述句子特征向量,获取用于表征所述待检测音频的意图的预测结果。

[0005] 根据本公开提供的一种意图识别方法,所述基于所述句子特征向量,获取用于表征所述待检测音频的意图的预测结果具体包括:输入所述句子特征向量至全连接神经网络进行分类预测,获取所述预测结果。

[0006] 根据本公开提供的一种意图识别方法,所述对所述文本序列和所述音素序列进行特征提取和融合,获取与所述待检测音频对应的句子特征向量包括:输入所述文本序列和所述音素序列至所述预训练语言模型,获取与所述待检测音频对应的鲁棒性词向量表征;基于所述鲁棒性词向量表征,获取所述预训练语言模型输出的并与所述待检测音频对应的句子特征向量。

[0007] 根据本公开提供的一种意图识别方法,所述预训练语言模型通过如下方式训练确定,包括:将真实文本序列和真实音素序列输入至初始预训练语言模型中,调整并确定第一预训练语言模型网络参数;其中,所述真实文本序列为从互联网中获取的文本序列;所述真实音素序列根据所述真实文本序列获得;所述第一预训练语言模型网络参数用于表征所述真实文本序列与所述真实音素序列发音关系;将样本文本序列和对应的样本音素序列,输入至网络参数为所述第一预训练语言模型网络参数的预训练语言模型,获得第二预训练语言模型网络参数;其中,所述样本文本序列和对应的样本音素序列通过对预设的多个样本音频进行语音识别得到;基于第二预训练语言模型网络参数,构建具有泛化能力的所述预

训练语言模型。

[0008] 根据本公开提供的一种意图识别方法,所述将真实文本序列和真实音素序列输入至初始预训练语言模型中,调整并确定第一预训练语言模型网络参数进一步包括:将所述真实文本序列和所述真实音素序列,输入至所述初始预训练语言模型,对所述真实文本序列和所述真实音素序列进行掩码,获取第一掩码文本序列和第一掩码音素序列,通过上下文预测被掩码的所述第一掩码文本序列和所述第一掩码音素序列,调整并确定所述第一预训练语言模型网络参数;其中,所述第一掩码文本序列为从所述真实文本序列中随机选择的20%的文本序列;所述第一掩码音素序列为从所述真实音素序列随机选择的20%的音素序列。

[0009] 根据本公开提供的一种意图识别方法,所述将样本文本序列和对应的样本音素序列输入至网络参数为第一预训练语言模型网络参数预训练语言模型,构建第二预训练语言模型网络参数进一步包括:将所述样本文本序列和所述样本音素序列,输入至网络参数为第一预训练语言模型网络参数预训练语言模型,对所述样本文本序列和所述样本音素序列进行掩码,得到第二掩码文本序列和第二掩码音素序列,通过上下文预测被掩码的所述第二掩码文本序列和所述第二掩码音素序列,获得所述第二预训练语言模型网络参数;其中,所述第二掩码文本序列为从所述样本文本序列中随机选择的20%的文本序列;所述第二掩码音素序列为从所述样本音素序列随机选择的20%的音素序列。

[0010] 根据本公开提供的一种意图识别方法,所述真实音素序列根据所述真实文本序列获得具体包括:对所述真实文本序列进行分词,得到所述真实字;根据音素对照表,获取所述真实字的音素;基于所有所述真实字的音素,获得所述真实音素序列。

[0011] 根据本公开提供的一种意图识别方法,所述第二预训练语言模型网络参数通过如下方式进行调整,具体包括:确定由人工听取的所述样本音频转写的真实文本;基于交叉熵损失函数,计算表征所述样本音频的意图的预测结果与所述真实文本的差异;基于所述差异,调整所述第一预训练语言模型网络参数,得到第二预训练语言模型网络参数。

[0012] 第二方面,本公开还提供一种意图识别装置,包括:语音识别模块、预训练语言模块和全连接神经网络分类模块。其中,语音识别模块,用于获取与待检测音频对应的音素序列,并基于所述音素序列,确定与所述待检测音频对应的文本序列;预训练语言模块,基于预训练语言模型,用于对所述文本序列和所述音素序列进行特征提取和融合,获取与所述待检测音频对应的句子特征向量;全连接神经网络分类模块,用于基于所述句子特征向量,获取用于表征所述待检测音频的意图的预测结果。

[0013] 第三方面,本公开还提供一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,所述处理器执行所述程序时实现如上述任一种所述意图识别方法的步骤。

[0014] 第四方面,本公开还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现如上述任一种所述意图识别方法的步骤。

[0015] 本公开提供一种意图识别方法、装置及相关设备,通过获取与待检测音频对应的音素序列,并基于所述音素序列,确定与所述待检测音频对应的文本序列;基于预训练语言模型,对所述文本序列和所述音素序列进行特征提取和融合,获取与所述待检测音频对应的句子特征向量;预训练语言模型可以将字对应的音素信息融入到预训练任务中,由此,在

对所述文本序列和所述音素序列进行特征提取和融合过程中,信息不但具有文字层面特征还具有发音特征,从而具有较强的抗噪能力的句子特征向量,基于所述句子特征向量,获取用于表征所述待检测音频的意图的预测结果会更为准确,如在面对用户回复中的相似发音错误如“是”和“四”的时候,虽然在文本层面,二者对应的句子特征向量表征相关性很低,句子特征向量相去甚远,但是具有融合音素信息的预训练模型获取的句子特征向量会考虑文本层面和发音特征两方面信息,也就是说能考虑到“是”和“四”之间的发音相似性,从而能将二者表征在一定程度上对齐,进而获得准确的用于表征所述待检测音频的意图的预测结果,提高口语意图识别准确率。

附图说明

[0016] 为了更清楚地说明本公开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一简单地介绍,显而易见地,下面描述中的附图是本公开的一些实施例,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,还可以根据这些附图获得其他的附图。

[0017] 图1是本公开提供的意图识别方法的流程示意图之一;

[0018] 图2是本公开提供的意图识别方法中,基于预训练语言模型,对文本序列和音素序列进行特征提取和融合,获取与待检测音频对应的句子特征向量的步骤流程示意图;

[0019] 图3是本公开意图识别方法实施例中通过训练确定预训练语言模型的流程示意图;

[0020] 图4为本公开意图识别方法中,将样本文本序列和对应的样本音素序列,输入至网络参数为第一预训练语言模型网络参数的预训练语言模型,获得第二预训练语言模型网络参数步骤流程示意图;

[0021] 图5是本公开提供的意图识别装置的结构示意图;

[0022] 图6为本公开提供的意图识别装置中,预训练语言模块的结构示意图;

[0023] 图7为本公开提供的意图识别方法中的预训练语言模型训练示意图;

[0024] 图8为本公开提供的意图识别方法的具体实施例流程示意图;

[0025] 图9是本公开提供的电子设备的结构示意图。

具体实施方式

[0026] 为使本公开实施例的目的、技术方案和优点更加清楚,下面将结合本公开实施例中的附图,对本公开实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本公开实施例一部分实施例,而不是全部的实施例。基于本公开实施例中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本公开实施例保护的范围。

[0027] 下面结合图1描述本公开实施例的意图识别方法的各种实施例。

[0028] 参照图1,图1是本公开提供的意图识别方法的流程示意图之一,本公开实施例提供的意图识别方法,包括:

[0029] 步骤S101、获取与待检测音频对应的音素序列,并基于音素序列,确定与待检测音频对应的文本序列;

[0030] 步骤S103、基于预训练语言模型,对文本序列和音素序列进行特征提取和融合,获取与待检测音频对应的句子特征向量;

[0031] 步骤S105、基于句子特征向量,获取用于表征待检测音频的意图的预测结果。

[0032] 本公开实施例提供的意图识别方法,通过获取与待检测音频对应的音素序列,并基于音素序列,确定与待检测音频对应的文本序列;基于预训练语言模型,对文本序列和音素序列进行特征提取和融合,获取与待检测音频对应的句子特征向量;预训练语言模型可以将字对应的音素信息融入到预训练任务中,由此,在对文本序列和音素序列进行特征提取和融合过程中,信息不但具有文字层面特征还具有发音特征,从而具有较强的抗噪能力的句子特征向量,基于句子特征向量,获取用于表征待检测音频的意图的预测结果会更为准确,如在面对用户回复中的相似发音错误如“是”和“四”的时候,虽然在文本层面,二者对应的句子特征向量表征相关性很低,句子特征向量相去甚远,但是具有融合音素信息的预训练模型获取的句子特征向量会考虑文本层面和发音特征两方面信息,也就是说能考虑到“是”和“四”之间的发音相似性,从而能将二者表征在一定程度上对齐,进而获得准确的用于表征所述待检测音频的意图的预测结果,提高口语意图识别准确率。

[0033] 下面对图1所示的实施例的各个步骤做进一步说明。

[0034] 步骤S101、获取与待检测音频对应的音素序列,并基于音素序列,确定与待检测音频对应的文本序列;

[0035] 该步骤中,音素为根据语音的自然属性划分出来的最小语音单位,依据音节里的发音动作来分析,一个动作构成一个音素。音素分为元音与辅音两大类。如汉语音节啊(a)只有一个音素,爱(ai)有两个音素,代(dai)有三个音素等。

[0036] 获取与待检测音频对应的音素序列可以通过语音识别模块中的声学模型接收音频信号进行特征提取,经过转换得到。在声学模型给出音素序列之后,从候选的文字序列中找出概率最大的字符串序列作为识别top1文本,作为文本序列。

[0037] 步骤S103、基于预训练语言模型,对文本序列和音素序列进行特征提取和融合,获取与待检测音频对应的句子特征向量;

[0038] 预训练语言模型一般选用BERT模型,利用大规模无标注语料训练、获得包含丰富语义信息的文本的语义表示,然后将文本的语义表示在特定NLP任务中作微调,最终应用于该NLP任务中。

[0039] 步骤S105、基于句子特征向量,获取用于表征待检测音频的意图的预测结果。

[0040] 该步骤中,待检测音频的意图的预测结果可以为分类结果,可以根据需要自己设置。

[0041] 在一个可选实施例中,步骤S105具体包括:

[0042] 步骤S1051、输入句子特征向量至全连接神经网络进行分类预测,获取预测结果。

[0043] 全连接神经网络(FFNN)将信息从前面反馈到后面(分别是输入和输出)。神经网络通常被描述为具有多个层,其中每个层由并行的输入、隐藏或输出神经元组成。单独的一层不会有连接,通常相邻的两层是全连接的(一层的每个神经元连接到另一层的每个神经元)。最简单实用的网络有两个输入神经元和一个输出神经元,可用来建模逻辑门。全连接神经网络起到映射句子特征向量的作用,得到用户可以读懂的预测结果。

[0044] 人们通常通过反向传播算法来训练全连接神经网络,给网络一对数据集(输入数

据集+期望的输出数据集),这叫做监督学习,而不是只给它输入,让网络来填补空白的非监督学习。反向传播的错误通常是输入和输出之间的差异的一些变体。假设网络有足够多的隐藏神经元,理论上它总是可以模拟输入和输出之间的关系。实际上,它们的使用非常有限,但是它们通常与其他网络结合在一起形成新的网络。

[0045] 在一个可选实施例中,参照图2,图2为本公开提供的意图识别方法中,基于预训练语言模型,对文本序列和音素序列进行特征提取和融合,获取与待检测音频对应的句子特征向量的步骤流程示意图,步骤S103具体包括:

[0046] 步骤S2031、输入文本序列和音素序列至预训练语言模型,获取与待检测音频对应的鲁棒性词向量表征;

[0047] 鲁棒性词向量表征,即是更强、更健壮的词向量表征。因为语音识别中的识别单词可能存在错误,但是错误有迹可循,即发音相似。通过为Bert模型增加音素的MLM任务,将音素信息赋予到字向量上,能获得更健壮的词向量表征,即鲁棒性词向量表征。

[0048] 步骤S2032、基于鲁棒性词向量表征,获取预训练语言模型输出的并与待检测音频对应的句子特征向量。

[0049] 句子特征向量就是Bert模型输出的[CLS] token对应的向量。

[0050] 参照图3,图3为本公开意图识别方法实施例中通过训练确定预训练语言模型的流程示意图之一,具体包括如下步骤:

[0051] 步骤S301、将真实文本序列和真实音素序列输入至初始预训练语言模型中,调整并确定第一预训练语言模型网络参数;

[0052] 步骤S303、将样本文本序列和对应的样本音素序列,输入至网络参数为第一预训练语言模型网络参数的预训练语言模型,获得第二预训练语言模型网络参数;

[0053] 步骤S305、基于第二预训练语言模型网络参数,构建具有泛化能力的所述预训练语言模型;

[0054] 下面对图3所示的实施例的各个步骤做进一步说明。

[0055] 步骤S301、将真实文本序列和真实音素序列输入至初始预训练语言模型中,调整并确定第一预训练语言模型网络参数;

[0056] 具体来说,因为Bert需要海量文本对初始预训练语言模型进行训练,但采集的样本音频没有那么多数据,所以需要先通过查表的方式进行预训练,得到第一预训练语言模型网络参数用于表征真实文本序列与真实音素序列发音关系。可以理解的是,真实文本序列为从互联网中获取的海量文本序列,然后根据真实文本序列获得真实音素序列。

[0057] 步骤S303、将样本文本序列和对应的样本音素序列,输入至网络参数为第一预训练语言模型网络参数的预训练语言模型,获得第二预训练语言模型网络参数;

[0058] 为了适应在真实场景下使用,真正的语音识别出的文本可能会真正意图的文本存在偏差,所以我们采用通过输入预设的多个样本音频进行语音识别得到样本文本序列和对应的样本音素序列预训练语言模型进行参数调整,即采用有噪声的训练数据进行参数调整,获得第二预训练语言模型网络参数,即真实使用时的模型网络参数。预训练语言模型训练如图7所示。

[0059] 步骤S305、基于第二预训练语言模型网络参数,构建具有泛化能力的所述预训练语言模型;

[0060] 基于训练好的网络参数,我们即可获得融合音素信息,可适应有噪声情况的BERT模型。

[0061] 对于步骤S301,在一个可选实施例中可以包括:

[0062] 步骤S3011、将真实文本序列和真实音素序列,输入至初始预训练语言模型,对真实文本序列和真实音素序列进行掩码,获取第一掩码文本序列和第一掩码音素序列,通过上下文预测被掩码的第一掩码文本和第一掩码音素序列,调整并确定第一预训练语言模型网络参数;

[0063] 步骤S3011是针对BERT模型的表征学习机制进行修改,即MLM(mask language model)。一般的MLM只需对文本序列的15%进行掩码,而我们由于融合了音素信息,相关性提高,需要提高掩码比例至20%,是以第一掩码文本序列为从真实文本序列中随机选择的20%的文本序列;第一掩码音素序列为从真实音素序列随机选择的20%的音素序列。在选取的字段中,10%的情况下字随机替换,10%的情况下字不做替换,80%的情况下字掩码,根据上下文即剩余词汇预测被掩码的第一掩码文本序列,同样的,在选取的音素段中,10%的情况下音素随机替换,10%的情况下音素不做替换,80%的情况下音素掩码,根据上下文即剩余音素预测被掩码的第一掩码音素序列。

[0064] 对于步骤S303,在一个可选实施例中可以包括:

[0065] 步骤S3031、将样本文本序列和样本音素序列,输入至网络参数为第一预训练语言模型网络参数预训练语言模型,对样本文本和样本音素序列进行掩码,得到第二掩码文本和第二掩码音素序列,通过上下文预测被掩码的第二掩码文本和第二掩码音素序列,获得第二预训练语言模型网络参数;

[0066] 对于步骤S3031,同样的,一般的MLM只需对文本序列的15%进行掩码,而我们由于融合了音素信息,相关性提高,需要提高掩码比例至20%,是以第二掩码文本序列为从样本文本序列中随机选择的20%的文本序列;第二掩码音素序列为从样本音素序列随机选择的20%的音素序列。在选取的字段中,10%的情况下字随机替换,10%的情况下字不做替换,80%的情况下字掩码,根据上下文即剩余词汇预测被掩码的第二掩码文本序列,同样的,在选取的音素段中,10%的情况下音素随机替换,10%的情况下音素不做替换,80%的情况下音素掩码,根据上下文即剩余音素预测被掩码的第二掩码音素序列。

[0067] 在一个优选实施例中,对于步骤S301之前,还包括:

[0068] 步骤S300:根据真实文本获得真实音素序列。

[0069] 步骤S300具体包括:

[0070] 步骤S3001、对所述真实文本进行分词,得到真实字;

[0071] 首先进行分词,是因为不同的词搭配的发音不同,因此需要先分词确定这个字在这个词搭配下的发音情况,中文的分词一般以字为单位。

[0072] 步骤S3002、根据音素对照表,获取真实字的音素;

[0073] 由于真实文本为海量互联网文本,不会经过语音识别系统,要确定这个字的发音情况,我们一般通过音素对照表,获取字的音素。面对具有多个音的字,我们选取最常用发音的音素作为其对应的音素。

[0074] 步骤S3003、基于所有真实字的音素,获得真实音素序列。

[0075] 参照图4,图4为本公开意图识别方法中,将样本文本序列和对应的样本音素序列,

输入至网络参数为第一预训练语言模型网络参数的预训练语言模型,获得第二预训练语言模型网络参数步骤流程示意图,即在一个具体实施例中,步骤S303具体包括如下步骤:

[0076] 步骤S401、将样本文本序列和对应的样本音素序列,输入至网络参数为第一预训练语言模型网络参数的预训练语言模型,输出句子特征向量;

[0077] 步骤S403、基于句子特征向量,获取用于表征音频样本的意图的预测结果;

[0078] 步骤S405、确定由人工听取的所述样本音频转写的真实文本;

[0079] 步骤S407、基于交叉熵损失函数,计算用于表征音频样本的意图的预测结果与真实文本的差异;

[0080] 步骤S409、基于差异,调整第一预训练语言模型网络参数,得到第二预训练语言模型网络参数。

[0081] 为获取更具有泛化能力的预训练语言模型的网络参数,我们往往需要进行多次调整来确定,针对具体应用进行优化,即BERT模型中的Finetune任务。而BERT模型所获得的句子特征向量,就和普通的CNN和LSTM在模型最后获取到的句子向量表征一样,都需要经过一个全连接神经网络向分类的标签空间映射,比如BERT模型拿到CLS向量768维度,需要向分类标签10个上映射,就需要一个全连接网络,参数w的shape位768*10进行连接,得到最终的10维度向量的logits,这个logits经过softmax进行归一化缩放,就可以获得各个意图的概率大小,一般取概率最大的值对应的标签为表征音频样本的意图的预测结果。交叉熵损失函数就是loss计算函数,计算表征音频样本的意图的预测结果的概率大小与真实意图0,1之间的差距大小,如果表征音频样本的意图的预测结果概率越高,且这个意图分类正确即1,则损失较小,反之,预测错误,损失越大。交叉熵损失比较的是经过softmax归一化概率缩放后的概率分布值与标签0,1之间的损失计算。具体计算公式:

$$[0082] \quad L = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})]$$

[0083] 当多次训练后的损失值小于我们预设的损失值,或达到我们预设的训练次数,训练结束,得到第二预训练语言模型网络参数。

[0084] 本公开提供的意图识别方法可以具体应用为例描述其实际应用时过程,参照图8,输入一段语音,对其进行语音识别,得到TOP1文本和对应音素,输入预训练语言模型,得到具有鲁棒性词向量特征,对鲁棒性词向量特征进行句子嵌入,即根据其位置属性进行句子拼接,又称模型变换,得到句子特征向量送入全连接网络,得到表征语音意图的预测结果。

[0085] 下面对本公开实施例提供的意图识别装置进行描述,下文描述的意图识别装置与上文描述的意图识别方法可相互对应参照。

[0086] 参照图5,图5是本公开提供的意图识别装置的结构示意图,本公开实施例还提供了一种意图识别装置,包括:语音识别模块50、预训练语言模块52和全连接神经网络分类模块54。其中,语音识别模块50,用于获取与待检测音频对应的音素序列,并基于音素序列,确定与待检测音频对应的文本序列;预训练语言模块52,基于预训练语言模型,用于对文本序列和音素序列进行特征提取和融合,获取与待检测音频对应的句子特征向量;全连接神经网络分类模块54,用于基于句子特征向量,获取用于表征待检测音频的意图的预测结果。

[0087] 本公开实施例提供的意图识别装置,通过语音识别模块50获取与待检测音频对应的音素序列,并基于音素序列,确定与待检测音频对应的文本序列;紧接着预训练语言模块52基于预训练语言模型,用于对文本序列和音素序列进行特征提取和融合,获取与待检测

音频对应的句子特征向量;预训练语言模块52中的预训练语言模型可以将字对应的音素信息融入到预训练任务中,由此,在对文本序列和音素序列进行特征提取和融合过程中,信息不但具有文字层面特征还具有发音特征,从而具有较强的抗噪能力的句子特征向量,在通过全连接神经网络分类模块54得到的用于表征待检测音频的意图的预测结果会更为准确,如在面对用户回复中的相似发音错误如“是”和“四”的时候,虽然在文本层面,二者对应的句子特征向量表征相关性很低,句子特征向量相去甚远,但是具有融合音素信息的预训练模型获取的句子特征向量会考虑文本层面和发音特征两方面信息,也就是说能考虑到“是”和“四”之间的发音相似性,从而能将二者表征在一定程度上对齐,进而获得准确的用于表征所述待检测音频的意图的预测结果,提高口语意图识别准确率。

[0088] 在一个可选实施例中,全连接神经网络分类模块54进一步用于:输入句子特征向量至全连接神经网络进行分类预测,获取预测结果。

[0089] 全连接神经网络(FFNN)将信息从前面反馈到后面(分别是输入和输出)。神经网络通常被描述为具有多个层,其中每个层由并行的输入、隐藏或输出神经元组成。单独的一层不会有连接,通常相邻的两层是全连接的(一层的每个神经元连接到另一层的每个神经元)。最简单实用的网络有两个输入神经元和一个输出神经元,可用来建模逻辑门。全连接神经网络起到映射句子特征向量的作用,得到用户可以读懂的预测结果。

[0090] 人们通常通过反向传播算法来训练全连接神经网络,给网络一对数据集(输入数据集+期望的输出数据集),这叫做监督学习,而不是只给它输入,让网络来填补空白的非监督学习。反向传播的错误通常是输入和输出之间的差异的一些变体。假设网络有足够多的隐藏神经元,理论上它总是可以模拟输入和输出之间的关系。实际上,它们的使用非常有限,但是它们通常与其他网络结合在一起形成新的网络。

[0091] 参照图6,图6为本公开提供的意图识别装置中,预训练语言模块52的结构示意图,预训练语言模块52进一步包括:

[0092] 获取鲁棒性词向量表征单元520,用于输入文本序列和音素序列至预训练语言模型,获取与待检测音频对应的鲁棒性词向量表征;

[0093] 鲁棒性词向量表征,即是更强、更健壮的词向量表征。因为语音识别中的识别单词可能存在错误,但是错误有迹可循,即发音相似。通过为Bert模型增加音素的MLM任务,将音素信息赋予到字向量上,能获得更健壮的词向量表征,即鲁棒性词向量表征。

[0094] 获取句子特征向量单元522,用于基于鲁棒性词向量表征,获取预训练语言模型输出的并与待检测音频对应的句子特征向量。

[0095] 句子特征向量就是Bert模型输出的[CLS] token对应的向量。

[0096] 在一个可选实施例中,预训练语言模块52还用于训练确定预训练语言模型,具体包括:

[0097] 建模单元,用于将真实文本序列和真实音素序列输入至初始预训练语言模型中,调整并确定第一预训练语言模型网络参数;

[0098] 其中,真实文本序列为从互联网中获取的文本序列;真实音素序列根据真实文本获得;第一预训练语言模型网络参数用于表征真实文本序列与真实音素序列发音关系;

[0099] 获取网络参数单元,将样本文本序列和对应的样本音素序列,输入至网络参数为第一预训练语言模型网络参数的预训练语言模型,获得第二预训练语言模型网络参数;

[0100] 为了适应在真实场景下使用,真正的语音识别出的文本可能会真正意图的文本存在偏差,所以我们采用通过输入预设的多个样本音频进行语音识别得到样本文本序列和对应的样本音素序列预训练语言模型进行参数调整,即采用有噪声的训练数据进行参数调整,获得第二预训练语言模型网络参数,即真实使用时的模型网络参数。

[0101] 构建模型单元,基于第二预训练语言模型网络参数,构建具有泛化能力的所述预训练语言模型;

[0102] 在具体实施例中,建模单元进一步用于:

[0103] 将真实文本序列和真实音素序列,输入至初始预训练语言模型,对真实文本序列和真实音素序列进行掩码,获取第一掩码文本序列和第一掩码音素序列,通过上下文预测被掩码的第一掩码文本和第一掩码音素序列,调整并确定第一预训练语言模型网络参数;

[0104] 具体来说,因为BERT模型需要海量文本对初始预训练语言模型进行训练,但采集的样本音频没有那么多数据,所以需要先通过查表的方式进行预训练,得到第一预训练语言模型网络参数用于表征真实文本序列与真实音素序列发音关系。可以理解的是,真实文本序列为从互联网中获取的海量文本序列,然后根据真实文本序列获得真实音素序列。

[0105] 在一个具体实施例中,获取网络参数单元进一步用于:

[0106] 将样本文本序列和样本音素序列,输入至网络参数为第一预训练语言模型网络参数预训练语言模型,对样本文本和样本音素序列进行掩码,得到第二掩码文本和第二掩码音素序列,通过上下文预测被掩码的第二掩码文本和第二掩码音素序列,获得第二预训练语言模型网络参数;

[0107] 同样的,一般的MLM只需对文本序列的15%进行掩码,而我们由于融合了音素信息,相关性提高,需要提高掩码比例至20%,是以第二掩码文本序列为从样本文本序列中随机选择的20%的文本序列;第二掩码音素序列为从样本音素序列随机选择的20%的音素序列。在选取的字段中,10%的情况下字随机替换,10%的情况下字不做替换,80%的情况下字掩码,根据上下文即剩余词汇预测被掩码的第二掩码文本序列,同样的,在选取的音素段中,10%的情况下音素随机替换,10%的情况下音素不做替换,80%的情况下音素掩码,根据上下文即剩余音素预测被掩码的第二掩码音素序列。

[0108] 在一个可选实施例中,建模单元之前还包括获取真实音素序列单元,用于根据所述真实文本获得真实音素序列,具体包括:

[0109] 分词子单元,用于对真实文本进行分词,得到真实字;

[0110] 使用分词子单元进行分词,是因为不同的词搭配的发音不同,因此需要先分词确定这个字在这个词搭配下的发音情况,中文的分词一般以字为单位。

[0111] 获取音素子单元,用于根据音素对照表,获取真实字的音素;

[0112] 由于真实文本为海量互联网文本,不会经过语音识别系统,要确定真实文本中每个真实字的发音情况,我们一般通过音素对照表,获取字的音素。面对具有多个音的真实字,我们选取最常用发音的音素作为其对应的音素。

[0113] 获取音素子单元,用于基于所有真实字的音素,获得真实音素序列。

[0114] 在一个具体实施例中,获取网络参数单元具体包括:

[0115] 输出向量子单元,将样本文本序列和对应的样本音素序列,输入至网络参数为第一预训练语言模型网络参数的预训练语言模型,输出句子特征向量;

[0116] 获取结果子单元,用于基于句子特征向量,获取用于表征音频样本的意图的预测结果;

[0117] 确定文本子单元,用于确定由人工听取的所述样本音频转写的真实文本;

[0118] 计算子单元,用于基于交叉熵损失函数,计算用于表征音频样本的意图的预测结果与真实文本的差异;

[0119] 调整子单元,用于基于差异,调整第一预训练语言模型网络参数,得到第二预训练语言模型网络参数。

[0120] 为获取更具有泛化能力的预训练语言模型的网络参数,我们往往需要进行多次调整来确定,针对具体应用进行优化,即BERT模型中的Finetune任务。而BERT模型所获得的句子特征向量,就和普通的CNN和LSTM在模型最后获取到的句子向量表征一样,都需要经过一个全连接神经网络向分类的标签空间映射,比如BERT模型拿到CLS向量768维度,需要向分类标签10个上映射,就需要一个全连接网络,参数w的shape位768*10进行连接,得到最终的10维度向量的logits,这个logits经过softmax进行归一化缩放,就可以获得各个意图的概率大小,一般取概率最大的值对应的标签为表征音频样本的意图的预测结果。交叉熵损失函数就是loss计算函数,计算表征音频样本的意图的预测结果的概率大小与真实意图0,1之间的差距大小,如果表征音频样本的意图的预测结果概率越高,且这个意图分类正确即1,则损失较小,反之,预测错误,损失越大。交叉熵损失比较的是经过softmax归一化概率缩放后的概率分布值与标签0,1之间的损失计算。具体计算公式:

$$[0121] \quad L = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})]$$

[0122] 当多次训练后的损失值小于我们预设的损失值,或达到我们预设的训练次数,训练结束,得到第二预训练语言模型网络参数。

[0123] 图9示例了一种电子设备的实体结构示意图,如图9所示,该电子设备可以包括:处理器(processor)910、通信接口(Communications Interface)920、存储器(memory)930和通信总线940,其中,处理器910,通信接口920,存储器930通过通信总线940完成相互间的通信。处理器910可以调用存储器930中的逻辑指令,以执行上述意图识别方法。

[0124] 此外,上述的存储器930中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用,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,本公开实施例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)执行本公开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(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0125] 另一方面,本公开还提供一种计算机程序产品,所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序,所述计算机程序包括程序指令,当所述程序指令被计算机执行时,计算机能够执行上述各方法所提供的意图识别方法。

[0126] 又一方面,本公开还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现以执行上述各提供的意图识别方法。

[0127] 以上所描述的装置实施例仅仅是示意性的,其中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模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0128] 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各实施方式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。基于这样的理解,上述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,该计算机软件产品可以存储在计算机可读存储介质中,如ROM/RAM、磁碟、光盘等,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)执行各个实施例或者实施例的某些部分所述的方法。

[0129] 最后应说明的是:以上实施例仅用以说明本公开的技术方案,而非对其限制;尽管参照前述实施例对本公开进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本公开各实施例技术方案的精神和范围。

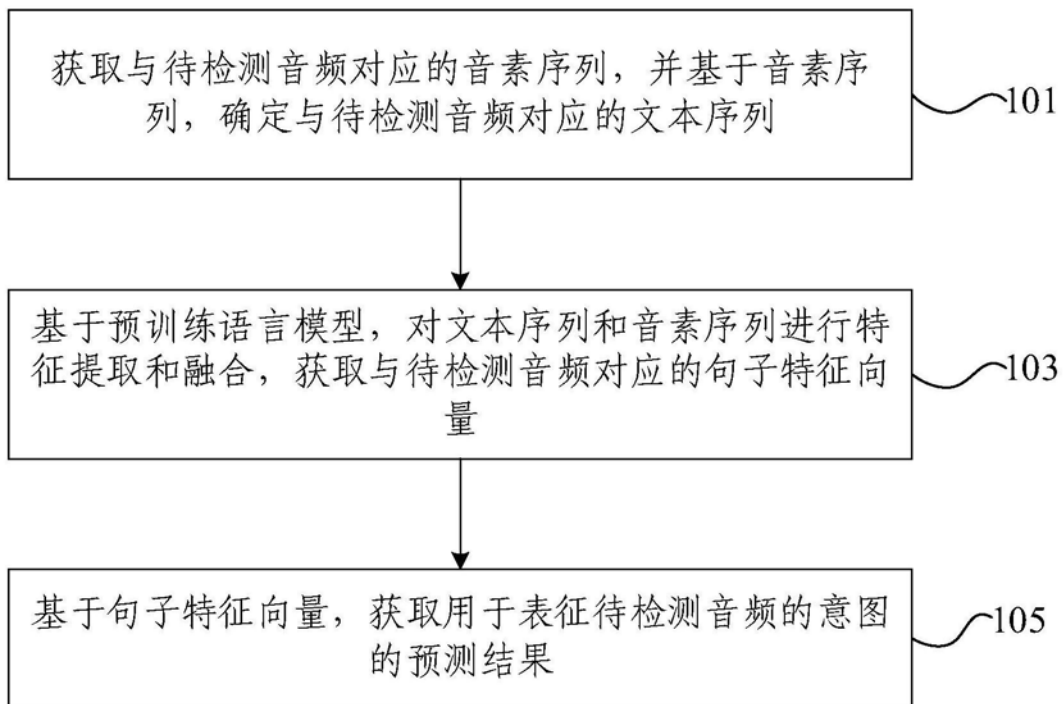


图1

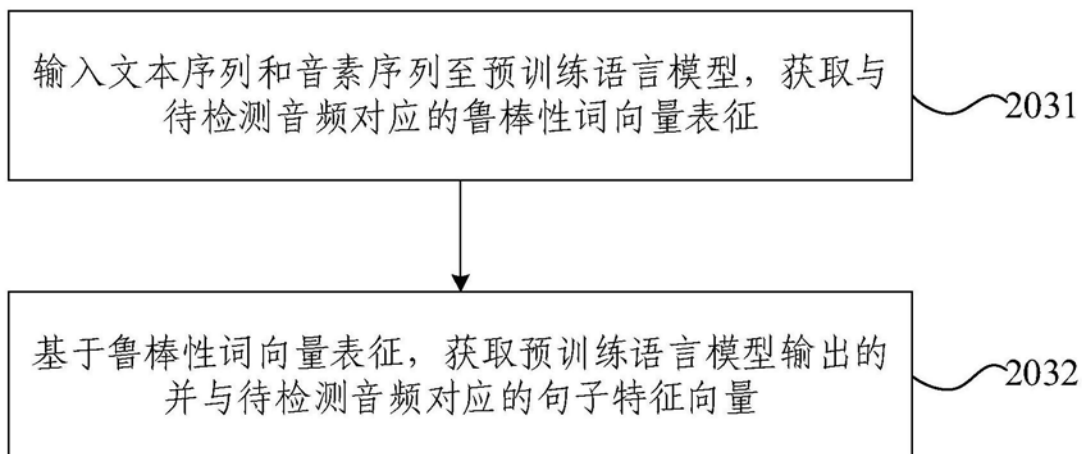


图2

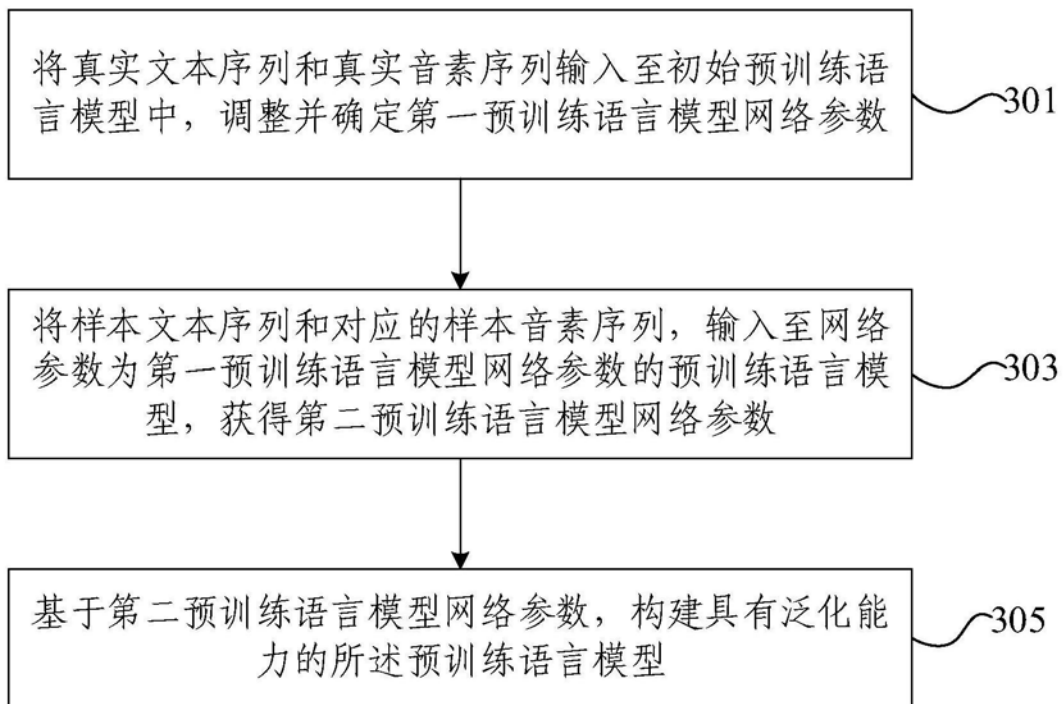


图3

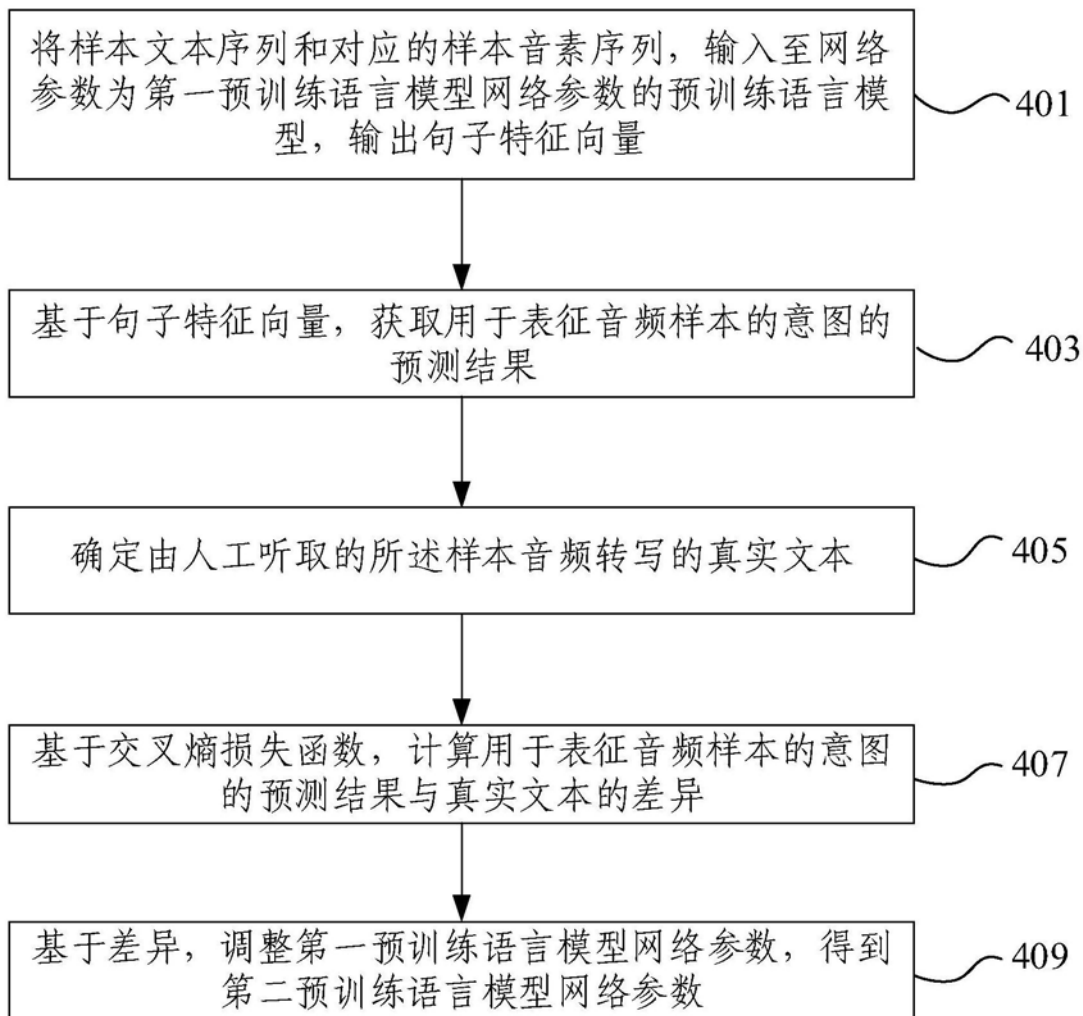


图4

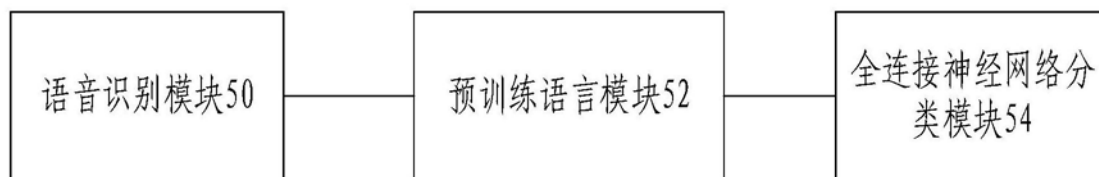


图5

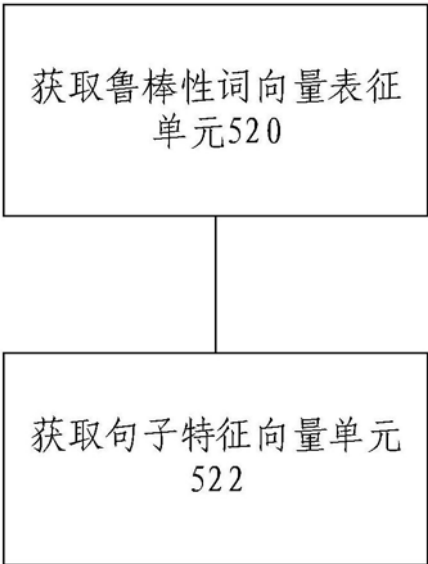


图6

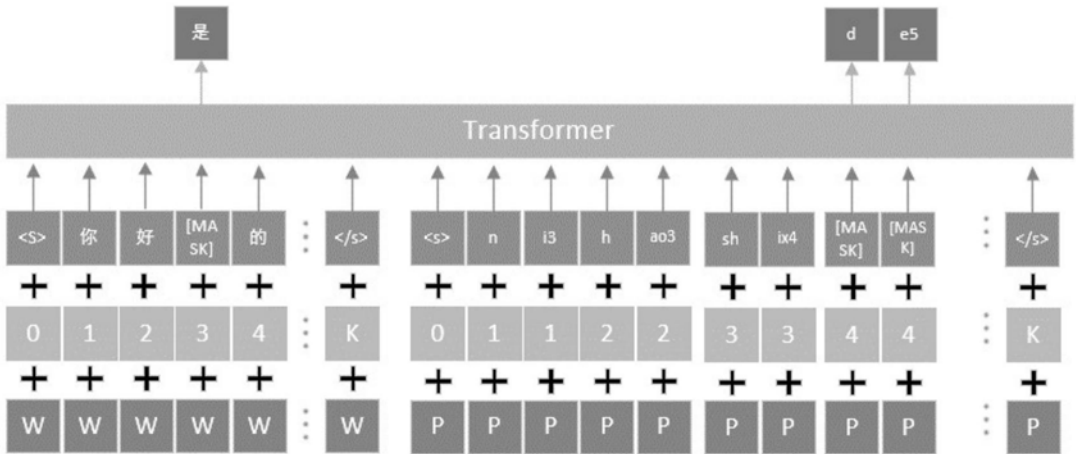


图7

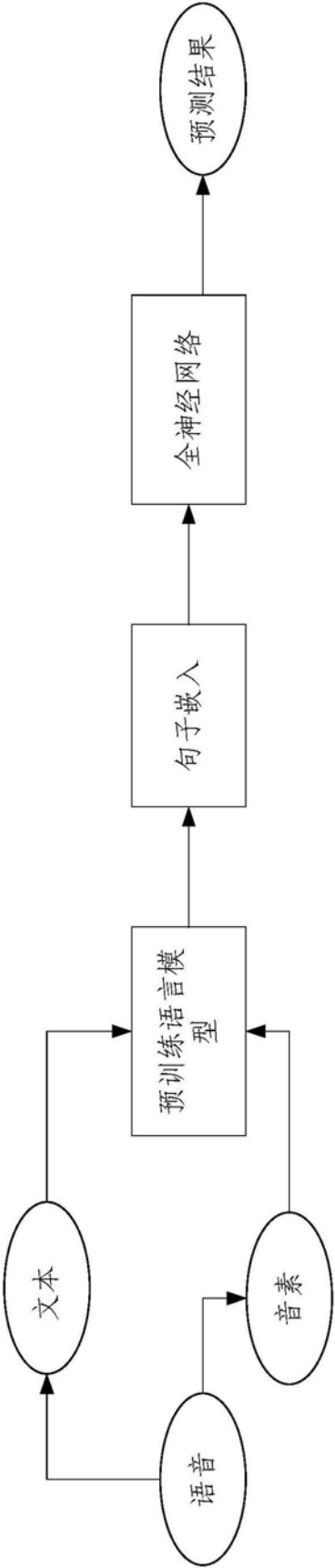


图8

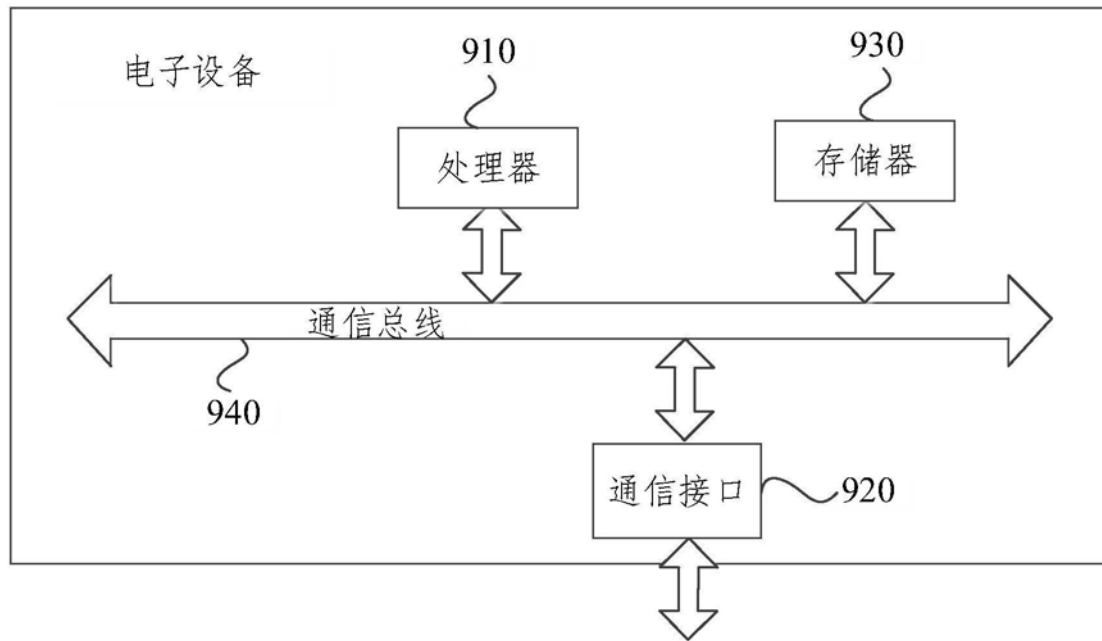


图9