(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 114023306 A
(43)申请公布日 2022.02.08

(21)申请号 202210001331.1

(22)申请日 2022.01.04

(71)申请人 阿里云计算有限公司
地址 310024 浙江省杭州市西湖区转塘科技经济区块12号

(72)发明人 汪诚愚 邱明辉 黄俊

(74)专利代理机构 北京展翼知识产权代理事务所(特殊普通合伙) 11452
代理人 张阳

(51)Int.Cl.
G10L 15/01(2013.01)
G10L 15/06(2013.01)
G10L 15/16(2006.01)

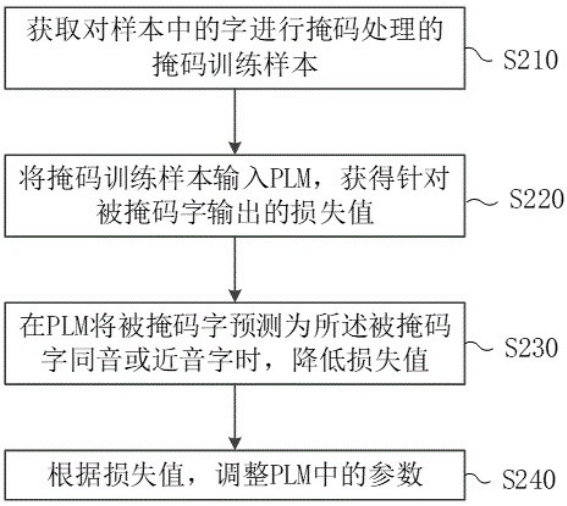
权利要求书2页 说明书12页 附图5页

(54)发明名称

用于预训练语言模型的处理方法和口语语言理解系统

(57)摘要

公开了一种用于预训练语言模型的处理方法和口语语言理解系统。所述方法包括:获取对样本中的字进行掩码处理的掩码训练样本;将掩码训练样本输入所述预训练语言模型,获得所述预训练语言模型针对被掩码字输出的损失值;在所述预训练语言模型将被掩码字预测为所述被掩码字同音或近音字时,降低所述损失值;以及根据所述损失值,调整所述预训练语言模型中神经网络模型的参数。由此,通过在预训练期间在输入文本的某些部分被同音字或近音字替换时减少语言表示的波动来得到对自动语音识别错误具有鲁棒性的预训练模型,即,得到能够容忍自动语音识别模型错误的预训练模型。



CN 114023306 A

1.一种用于预训练语言模型的处理方法,包括:

获取对样本中的字进行掩码处理的掩码训练样本;

将所述掩码训练样本输入所述预训练语言模型,获得所述预训练语言模型针对被掩码字输出的损失值;

在所述预训练语言模型将所述被掩码字预测为所述被掩码字的同音或近音字时,降低所述损失值;以及

根据所述损失值,调整所述预训练语言模型中的参数。

2.如权利要求1所述的方法,其中,在所述预训练语言模型将所述被掩码字预测为所述被掩码字的同音或近音字时,降低所述损失值包括:

将用于计算所述损失值的损失函数规定为预测目标向量与模型预测概率分布之间的交叉熵并据此求取所述损失值,其中,所述预测目标向量中与所述被掩码字同音或近音字对应的项的系数不为零。

3.如权利要求2所述的方法,其中,所述预测目标向量包括:

对应于被掩码字的第一预测目标系数;

分别对应于与所述被掩码字同音或近音的前k个字中每一个字的k个第二预测目标系数;

对应于字表中其他字的0,

其中,所述第一预测目标系数和所述k个第二预测目标系数都大于零,所述第一预测目标系数大于所述k个第二预测目标系数中的每一个系数,并且所述第一预测目标系数和所述k个第二预测目标系数之和为1。

4.如权利要求3所述的方法,其中,前k个字中每一个字各自的第二预测目标系数基于与所述被掩码字的声母、韵母和声调的近似程度确定。

5.如权利要求3所述的方法,还包括:

获得所述预训练语言模型针对被掩码字输出的第二损失值;以及

根据所述第二损失值,调整所述预训练语言模型中的参数,

其中,将第二损失函数规定为第二预测目标向量与模型预测概率分布之间的交叉熵并据此求取所述第二损失值,并且所述第二预测目标向量是根据上游自动语音识别模型的错误集调整所述预测目标向量中的第二预测目标系数的取值得到的。

6.如权利要求2所述的方法,其中,基于语音相似性预先计算字表中每个字的预测目标向量中同音或近音字对应的项的系数。

7.如权利要求1所述的方法,还包括:

获得所述预训练语言模型针对被掩码字输出的第三损失值,其中,将第三损失值函数规定为独热向量与模型预测概率分布之间的交叉熵并据此求取所述第三损失值;以及

根据所述第三损失值,调整所述预训练语言模型中的参数。

8.如权利要求7所述的方法,还包括:

求取所述损失值和所述第三损失值的加权和,以获取整体损失值;以及

根据所述整体损失值,调整所述预训练语言模型中的参数。

9.一种口语语言理解系统,包括:

自动语音识别系统,用于将获取的用户语言输入识别为文字;以及

根据如权利要求1-8中任一项所述的方法获取的预训练语言模型，用于在所述识别出的文字包含所述自动语音识别系统的识别错误的情况下，基于当前口语理解任务自动更正识别出的文字中包含的错误。

10.一种计算设备，包括：

处理器；以及

存储器，其上存储有可执行代码，当所述可执行代码被所述处理器执行时，使所述处理器执行如权利要求1-8中任一项所述的方法。

11.一种非暂时性机器可读存储介质，其上存储有可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器执行如权利要求1-8中任一项所述的方法。

用于预训练语言模型的处理方法和口语语言理解系统

技术领域

[0001] 本公开涉及一种深度学习领域，尤其涉及一种用于预训练语言模型的处理方法和相应的口语语言理解系统。

背景技术

[0002] 口语语言理解(SLU)作为任务型对话系统的核心组件，目的是为了获取用户询问语句的框架语义表示信息，即理解和解释人类语音的含义。SLU任务通常包含以下两个任务：意图识别和槽位填充。在进行SLU任务之前，需要利用自动语音识别(ASR)将人类语音转录为文本。通过结合ASR系统和SLU模型，可以提取语音信号中文本内容的含义，并将其用于具有丰富的人机交互的环境中。

[0003] 在现有技术中，已经存在利用神经网络结果实现的高性能ASR系统。但上述ASR系统在生成文本时通常不可避免的包含正确识别结果被同音或近音字替换的错误，例如，将用户语音输入“我说的是拖鞋”识别为文字“我说的是妥协”。ASR系统产生的上述错误极易传播至下游的SLU模型，从而降低SLU模型的鲁棒性。虽然现有技术中公开了ASR错误检测方案来识别转录文本中的ASR错误，但上述方案涉及对ASR系统的实质性修改，从而导致额外的系统修改开销。

[0004] 为此，需要一种能够解决ASR系统中错误识别对下游SLU模型影响的改进方法。

发明内容

[0005] 本公开要解决的一个技术问题是提供一种用于预训练语言模型的数据处理方案，该方案能够通过在预训练期间，在输入文本的某些部分被同音字或近音字替换时减少语言表示的波动来得到对ASR错误具有鲁棒性的预训练模型，即，得到能够容忍ASR模型错误的预训练模型。由此获取的SLU模型能够在ASR模型输入的语音识别结果有误的情况下，让人能够进行正确的意图识别或是槽位填充。

[0006] 根据本公开的第一个方面，提供了一种用于预训练语言模型(PLM)的处理方法，包括：获取对样本中的字进行掩码处理的掩码训练样本；将所述掩码训练样本输入所述PLM，获得所述PLM针对被掩码字输出的损失值；在所述PLM将被掩码字预测为所述被掩码字同音或近音字时，降低所述损失值；以及根据所述损失值，调整所述PLM中神经网络模型的参数。

[0007] 可选地，在所述PLM将被掩码字预测为所述被掩码字同音或近音字时，降低所述损失值包括：将损失函数规定为预测目标向量与模型预测概率分布之间的交叉熵并据此求取所述损失值，其中，所述预测目标向量中与所述被掩码字同音或近音字对应的项的系数不为零。

[0008] 可选地，所述预测目标向量包括：对应于被掩码字的第一预测目标系数；对应于与所述被掩码字同音或近音的前k个字中每一个字的k个第二预测目标系数；对应于字表中其他字的0，其中，所述第一预测目标系数和所述k个第二预测目标系数都大于零，所述第一预测目标系数大于所述k个第二预测目标系数中的每一个系数，并且所述第一预测目标系数

和所述k个第二预测目标系数之和为1。

[0009] 可选地，根据前k个字中每一个字各自的第二预测目标系数基于与所述被掩码字的声母、韵母和声调的近似程度确定。

[0010] 可选地，所述方法还包括：获得所述PLM针对被掩码字输出的第二损失值；以及根据所述第二损失值，调整所述PLM中神经网络模型的参数，其中，将第二损失函数规定为第二预测目标向量与模型预测概率分布之间的交叉熵并据此求取所述第二损失值，并且所述第二预测目标向量是根据上游ASR模型的错误集调整所述预测目标向量中的第二预测目标系数的取值得到的。

[0011] 可选地，字表中每个字的基于启发式的语音相似性被作为先验知识注入所述损失函数。换句话说，基于语音相似性预先计算字表中每个字的预测目标向量中同音或近音字对应的项的系数。

[0012] 可选地，所述方法还包括：获得所述PLM针对被掩码字输出的第三损失值，其中，将第三损失值函数规定为独热向量与模型预测概率分布之间的交叉熵并据此求取所述第三损失值；以及根据所述第三损失值，调整所述PLM中神经网络模型的参数。

[0013] 可选地，所述方法还包括：求取所述损失值和所述第三损失值的加权和，以获取整体损失值；以及根据所述整体损失值，调整所述PLM中的参数。

[0014] 根据本公开的第二个方面，提供了一种口语语言理解系统，包括：自动语音识别(ASR)系统，用于将获取的用户语言输入识别为文字；以及根据如上述第一方面所述的方法获取的预训练语言模型，用于在所述识别出的文字包含ASR系统的识别错误的情况下，自动更正识别出的文字中包含的错误并执行相应的口语理解任务。

[0015] 根据本公开的第三个方面，提供了一种计算设备，包括：处理器；以及存储器，其上存储有可执行代码，当可执行代码被处理器执行时，使处理器执行如上述第一方面所述的方法。

[0016] 根据本公开的第四个方面，提供了一种非暂时性机器可读存储介质，其上存储有可执行代码，当可执行代码被电子设备的处理器执行时，使处理器执行如上述第一方面所述的方法。

[0017] 由此，基于经典的掩码语言建模(MLM)任务，提出了语音感知掩码语言模型(PMLM)和ASR模型自适应掩码语言模型(AMLMM)任务来训练BERT模型，由此使得预训练模型能够学习到字表中汉字的语音相似性并能够适应特定的上游ASR模型，由此使得学习得到的预训练模型能够容忍ASR模型错误，并在上游ASR模型提供了包括同音或近音字错误的识别文字的情况下仍然能够争取地完成下游任务。

附图说明

[0018] 通过结合附图对本公开示例性实施方式进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施方式中，相同的参考标号通常代表相同部件。

[0019] 图1示出了对样本进行处理获取掩码训练样本的例子。

[0020] 图2示出了根据本发明一个实施例的用于预训练语言模型的数据处理方法的示意性流程图。

[0021] 图3示出了基于不同的预测目标向量来求取损失值的例子。

[0022] 图4示出了汉语拼音的组成例。

[0023] 图5示出了基于另一个不同的预测目标向量来求取损失值的例子。

[0024] 图6示出了基于ASR模型的预测错误计数来求取同音近音字错误出现概率的示意图。

[0025] 图7示出了根据本发明一个优选实施例的ARoBERT模型的预训练示意图。

[0026] 图8示出了根据本发明的一个口语语言理解系统的组成示意图。

[0027] 图9示出了根据本发明一实施例可用于实现上述用于预训练语言模型的数据处理方法的计算设备的结构示意图。

具体实施方式

[0028] 下面将参照附图更详细地描述本公开的优选实施方式。虽然附图中显示了本公开的优选实施方式，然而应该理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了使本公开更加透彻和完整，并且能够将本公开的范围完整地传达给本领域的技术人员。

[0029] SLU旨在解释人类语音的含义，用以支持各种人机交互系统。正确SLU的前提是能够将语音信号转录为正确文本内容的自动语音识别(ASR)系统。

[0030] 针对ASR模型的研究由来已久，并且已经实现了深度语音到文本神经网络。ASR系统的代表性神经网络架构包括DeepSpeech2、SpeechTransformer、wav2vec自监督系统等。尽管这些神经网络架构具有优良的识别虚拟，但现有的ASR系统生成的转录文本仍然不可避免地包含错误，例如将用户语音输入“我说的是拖鞋”识别为文字“我说的是妥协”，而引入的“拖鞋”被“妥协”替换的同音字或近音字错误。ASR系统产生的错误很容易传播到下游的SLU模型。由于因为ASR引入错误与真实输入在语义上通常相距甚远(例如，替换的同音或近音字“妥协”与真实输入的“拖鞋”语义上完全不相关)，因此在ASR系统转录的文本上进行训练或测试的主流SLU模型通常无法提供实际应用所需的错误鲁棒性。

[0031] 为了解决上述问题，显而易见的方法是检测ASR转录文本的错误，或是修改ASR系统中的编码器-解码器架构，以提高系统将潜在表示解码为离散文本输出时的语言正确性和流畅性。当这两种方法应用于SLU任务时，必须修改现有的ASR系统，从而导致额外的修改开销。

[0032] 由此，本发明的发明人想到如果设计一个端到端的SLU模型能够直接从ASR转录的(可能存在同音或近音字替换错误的)文本中生成可靠的预测，那将是更可取的。

[0033] 为此，本发明提出了一种用于预训练语言模型的数据处理方案，该方案能够通过在预训练期间，在输入文本的某些部分被同音字或近音字替换时减少语言表示的波动来得到对ASR错误具有鲁棒性的预训练模型，即，通过训练得到能够容忍ASR模型错误的预训练模型。由此获取的SLU模型能够在ASR模型输入的语音识别结果有误的情况下，让人能够进行正确的意图识别或是槽位填充。

[0034] 由于近年来大规模预训练语言模型(PLM)的出现显着提高了各种语言理解任务的性能，而采用PLM来构建端到端SLU模型非常简单。学习过程涉及PLM预训练阶段和下游SLU任务的PLM微调阶段。然而，如果PLM预训练期间学习的表示对ASR错误不具有鲁棒性，则PLM

微调的性能将在很大程度上受到影响。由于大多数ASR错误发生在单词的同音字或同音字上，本发明设计了一种能够减少当部分输入文本被其同音字或同音字替换时语言表示的波动的预训练模型。在一个优选实施例中，该模型是基于BERT实现的ARoBERT。具体地，在ARoBERT中，使用了一堆转换器编码器来学习输入字(token)的表示。然后，ARoBERT中的转换器编码器能够容忍在语音上相似的ASR错误。为了预训练该ARoBERT，本发明中使用了在PLM将被掩码字预测为所述被掩码字同音或近音字时，降低损失值的新的自监督任务(如下将描述的PMLM和AMMLM)来对神经网络的模型参数进行微调。

[0035] 随着深度学习的发展，模型参数的数量飞速增长。为了训练这些参数，需要更大的数据集来避免过拟合。然而，对于大部分NLP任务来说，构建大规模的标注数据集会因为成本过高而无法实现，特别是对于句法和语义相关的任务。相比之下，大规模的未标注语料库的构建则相对容易，为了利用这些数据，可以先从其中学习到一个好的表示，再将这些表示应用到其他任务中，这也是“预训练”的含义。最近的研究表明，基于大规模未标注语料库的预训练模型(PTM)在很多NLP(自然语言理解)任务上取得了很好的表现。

[0036] 预训练任务对于学习语言的通用表示至关重要。通常，这些预训练任务应该具有一定的挑战性，并且有足够的训练数据支撑。原文将预训练任务包括三类:监督学习、无监督学习和自监督学习。监督学习(SL)基于包含输入输出对的训练数据学习一个将输入映射到输出的函数。无监督学习(UL)旨在从无标签数据中找到内在的知识，例如聚类、密度、潜在表示等。自监督学习(SSL)是监督学习和无监督学习的结合，其学习方式和监督学习一样，但是训练数据的标签是自动生成的。核心思想在于以某种形式预测输入的任意部分，基于该部分之外的其他部分。具体地，掩码语言模型(MLM)是一种自监督任务，其尝试去遮住句子中的一个词语，基于剩余的词语来预测它。

[0037] 图1示出了对样本进行处理获取掩码训练样本的例子。如图1所示，在基于MLM的训练方案中，PLM的训练样本是被掩码处理后的文本，即部分文字被随机替换成特殊的标记符号(例如，[MASK])的句子，例如，原文本是“我想买件礼物送你”，被掩码处理后的文本为“我想买件[MASK]物送你”。被掩码处理后的文本输入到PLM，PLM需要预测出被掩码的字分别是“礼”。PLM的训练样本可以称为掩码训练样本。在一个文本(例如，句子)中，对于被掩码处理的字，未被掩码处理的字是它的上下文信息，PLM通过预测被掩码处理的字，学习到了捕捉文字上下文信息的能力。因此基于MLM训练方案训练完成的PLM具有理解自然语言深度语义的能力，可用于一系列NLP相关的下游任务。

[0038] 预训练模型表示通过标签样本来学习(确定)所有权重和偏置的理想值。这些确定了的权重和偏置则能够在神经网络部署阶段对输入的特征值进行高准确率的推理，例如，基于上下文对掩码汉字的正确预测。

[0039] 在自监督式学习中，机器学习算法通过检查多个样本并尝试找出可最大限度地减少损失的模型来学习参数，这一过程称为经验风险最小化。

[0040] 损失是对不良预测的惩罚。即，损失可以是一个表示对于单个样本而言模型预测的准确程度的数值。如果模型的预测完全准确，则损失为零，否则损失会较大。训练模型的目标是从所有样本中找到一组平均损失“较小”的权重和偏差。

[0041] 在神经网络的训练和微调过程中，为了量化目前的权重和偏置是否能够让网络输入拟合所有的网络输入，需要定义一个损失函数。由此，训练网络的目的可以转变为最小化

权重和偏置的损失函数的过程。通常，使用梯度下降算法(多层神经网络训练中，使用反向传播算法)来实现上述最小化的过程。

[0042] 在反向传播算法中，涉及前向传播和反向传播的重复迭代过程。前向传播过程是层间神经元通过权值矩阵的连接使得刺激(特征值)经每一层的激励函数不断由前一层向下一层传递的过程。而在反向传播中，本层的误差需要由下一层的误差反向推导。由此通过上述正向和反向传播的迭代过程不断调整权重和偏置，使得损失函数逐渐接近最小值，从而完成对神经网络的训练。

[0043] 在MLM中，对于字表中的字，预测目标向量为one-hot向量。one-hot向量又被称为“独热向量”，即，在包含v个汉字的字表集合V中，只有与“礼”相对应的系数为1，其余v-1个汉字各自的系数都为0。因此在使用one-hot向量来构造损失函数时，只有模型预测到了被遮盖的汉字本身，例如模型输出为“礼”时，才不会引起损失，而当模型输出了“礼”之外的任何其他字，都会引起相同的损失。由上可知，使用MLM任务进行预训练，无法对字的语音知识进行编码，而只具备识别出正确的汉字本身的能力。

[0044] 相比之下，本发明的数据处理方法可以看作是针对MLM的修改和优化。具体地，本发明的数据处理方法同样是通过反向传播使得损失函数最小化来对预训练模型的参数进行调优。但不同之处在于，在本发明中，即便模型预测错误，但如果模型预测的是与被遮盖汉字语音相近的字(即，与被掩码字同音或近音的字)，那该预测错误的损失值也会小于其他完全无关的汉字。具体地，在预测到上述同音或近音字降低损失值，可以通过用特定的预测目标向量来代替MLM中的one-hot向量来实现。

[0045] 图2示出了根据本发明一个实施例的用于预训练语言模型的数据处理方法的示意性流程图。

[0046] 在本发明的训练任务中，会类似于经典的掩码语言模型(MLM)任务首先在输入句子中遮挡住部分的token(在中文中，一个token可以看作是“汉字”或“字”)，然后训练模型来基于剩下的词语预测被遮住的词语。但不同之处在于，本发明的预训练过程中，对于所用汉字表中所有的字，预测目标并非仅仅是MLM中使用的one-hot(独热)向量(即，只有被遮盖正确汉字所对应的系数为1，其他汉字的系数都为0)，而是可以使用与被遮盖汉字语音上类似的字都可以具有相应非零系数的预测目标向量，由此使得模型能够学习到汉字之间语音上的关联，由此提升后续SLU模型对ASR错误的鲁棒性。

[0047] 为此，在步骤S210，获取对样本中的字进行掩码处理的掩码训练样本。例如，可以与图1所示类似的，对样本句子中的字进行随机掩码，以获取带掩码的训练样本。例如，原文本是“我想买件礼物送你”，被掩码处理后的文本为“我想买件[MASK]物送你”。

[0048] 在步骤S220，将所述掩码训练样本输入所述预训练语言模型PLM，获得所述PLM针对被掩码字输出的损失值。在步骤S230，在所述PLM将被掩码字预测为所述被掩码字同音或近音字时，降低所述损失值。

[0049] 在步骤S240，根据所述损失值，调整所述PLM中神经网络模型的参数。

[0050] 具体地，被掩码处理后的文本输入到PLM，PLM需要预测出被掩码的字是“礼”。但不同之处在于，基本PLM预测输出为“礼”之外的其他字，只要这些字与“礼”在语音上近似，则预测损失就会降低。换句话说，同音和近音字引起的损失比预测到其他语音无关的字所引起的损失要小。

[0051] 在一个实施例中，可以将损失函数规定为预测目标向量与模型预测概率分布之间的交叉熵，并根据该损失函数来求取损失值。为此，在所述PLM将被掩码字预测为所述被掩码字同音或近音字时，降低所述损失值可以包括：使得预测目标向量中与所述被掩码字同音或近音字对应的项的系数不为零。此时，可以将步骤S230看作是步骤S220的子步骤，即通过对预测目标向量的构造，直接在基于损失函数的计算中降低模型预测出被掩码字同音或近音字时所造成的损失值。。

[0052] 图3示出了基于不同的预测目标向量来求取损失值的例子。如图所示，左侧的预测目标向量为one-hot(独热)向量，其仅在模型正确预测出被遮盖的“礼”字，才不会引起损失。而在模型将“礼”字错误的预测为其他字时，无论时同音近音字(例如，“李”或“离”或“底”)，还是其他语音无关字(例如，“废”或“怪”)，所引起的损失都是一样的。换句话说，在使用one-hot(独热)向量作为预测目标向量的MLM任务中，是无法习得各个汉字在语音上的相似性的。而在本发明中，可以使用右侧的预测目标向量来进行损失值的求取。在此，模型正确预测出被遮盖的“礼”字，仍然不引起损失(或是引起的损失最小，因为对应于“礼”的系数仍然最大)，但同音近音字(例如，“李”或“离”或“底”)由于其对应的系数不为零，因此模型预测出这些同音近音字所引起的损失，要小于其他语音无关字(例如，“废”或“怪”)所引起的损失。由此，该训练预测模型能够从一定程度上习得汉字的语音相关性。

[0053] 进一步地，为了在语音相关性的学习和计算量直接取得平衡，可以对同音近音字的范围进行限定，例如，仅规定与被掩码字读音最相近的k个字具有非零的预测目标系数，由此，可以为字表中的每一个汉字构造相应的预测目标向量。此时，预测目标向量包括：对应于被掩码字的第一预测目标系数；对应于与所述被掩码字同音或近音的前k个字中每一个字的k个第二预测目标系数；对应于字表中其他字的0。在此，第一预测目标系数和所述k个第二预测目标系数都大于零。

[0054] 例如，可以为字表中的“礼”字构造对应的预测目标向量，“礼”的系数可以如图3所示为0.4，而同音字“李”的系数可以如图3所示为0.1，近音字“离”和“底”的系数可以分别为0.05和0.025。显然，正确预测应该引起最小的损失，因此，所述第一预测目标系数大于所述k个第二预测目标系数中的每一个系数，并且为了归一化计算方便，第一预测目标系数和k个第二预测目标系数之和为1。为此，预测目标向量是一个v维向量，v为字表中包括的字的总数。在针对“礼”的预测目标向量中，“礼”字对应于第一预测目标系数y_0，“李”、“离”、“底”具有各自相同或不同的第二预测目标系数y_1、y_2和y_3，显然y_0要大于y_1、y_2和y_3，并且如果设k=30，则$y_0+y_1+y_2+y_3+\cdots+y_{30}=1$。可以为字表中的每个字求取上述预测目标向量，例如，求取字表中每一个字的30个同音或近音字的系数，并由此为每一个字确定其预测目标向量。由此，使得基于启发式的语音相似性被作为先验知识注入损失函数。

[0055] 在一个实施例中，可以利用字与字之间在拼音上的相似性来求取同音和近音字的系数。图4示出了汉语拼音的组成例。如图所示，汉语拼音包括声母、韵母和声调。为此，前k个字中每一个字各自的第二预测目标系数基于与所述被掩码字的声母、韵母和声调的近似程度确定(可以参见后续的应用例中的公式(6))。

[0056] 在一个实施例中，上述预测目标向量还可以根据上游的ASR模型的特性进行微调。图5示出了基于另一个不同的预测目标向量来求取损失值的例子。如图所示，可以根据上游ASR模型的错误集调整预测目标向量中的第二预测目标系数的取值(其实也可以相应调整

正确预测时的第一预测目标系数，也可以保持第一预测目标系数作为一个不变的常数，例如0.4)。为此，本发明的数据调整方法还可以包括获得所述PLM针对被掩码字输出的第二损失值；以及根据所述第二损失值，调整所述PLM中神经网络模型的参数，其中，将第二损失函数规定为第二预测目标向量与模型预测概率分布之间的交叉熵并据此求取所述第二损失值，并且所述第二预测目标向量是根据上游ASR模型的错误集调整预测目标向量中的第二预测目标系数的取值。

[0057] 具体地，可以根据上述ASR模型的预测错误计数来计算针对第二预测目标系数的修正系数，例如，根据ASR模型的同音近音字错误出现的次数来调整第二预测目标系数y_1，y_2，…y_{30}中每一个系数的取值。图6示出了基于ASR模型的预测错误计数来求取同音近音字错误出现概率的示意图。

[0058] 在一个实施例中，还可以将已有的MLM任务也考虑在内，由此确保模型正确预测的能力。此时，本发明的数据处理方法还可以包括：获得所述PLM针对被掩码字输出的第三损失值，其中，将第三损失值函数规定为独热向量与模型预测概率分布之间的交叉熵并据此求取所述第三损失值；以及根据所述第三损失值，调整所述PLM中神经网络模型的参数。

[0059] 在一个实施例中，可以求取所述损失值和所述第三损失值的加权和，以获取整体损失值，并且根据所述整体损失值，调整所述PLM中神经网络模型的参数。在此，可以将考虑语音相似性的预测目标向量所对应的损失函数表示为L_{PMLM}将在考虑语音相似性的基础上考虑ASR错误计数的第二预测目标向量所对应的损失函数表示为L_{AMMLM}，将用于已有MLM任务的独热向量所对应的损失函数表示为L_{MLM}。

[0060] 在一个实施例中，总损失函数为$L = L_{MLM} + \lambda_1 L_{PMLM}$。此时已经可以获得对ASR错误具有足够鲁棒性的预训练模型。而在一个优选实施例中，还可以加入对ASR模型错误本身的适配，因此，此时可以使得总损失函数为$L = L_{MLM} + \lambda_1 L_{PMLM} + \lambda_2 L_{AMMLM}$，并由此进行本发明的能够习得语音相似性并且能够保持模型正确预测能力的预训练模型。在该模型预训练结束后，还可以对此模型进行适用于下游文本理解任务的调优。

[0061] 如上已经结合图2-6描述了根据本发明的数据处理方案。如下将结合图7描述本发明的一个应用例。

[0062] 应用例

图7示出了根据本发明一个优选实施例的ARoBERT模型的预训练示意图。

[0063] 如前所述，现有技术都没有考虑预训练语言模型中的ASR错误鲁棒性以提高各种SLU任务性能的方案。为此，本发明提出了ARoBERT(ASRRobustBERT的缩写)。由于大多数ASR错误发生在单词的同音字或同音字上，我们设计了ARoBERT以减少当部分输入文本被其同音字或同音字替换时语言表示的波动。在ARoBERT中，使用了堆转换器编码器来学习输入标记表示(例如，汉字)，并且ARoBERT中的转换器编码器可以容忍在语音上与被掩码字相似的ASR错误。除了经典的掩码语言建模(MLM)任务之外，进一步提出了两个新的自监督任务用于预训练ARoBERT，即语音感知掩码语言模型(PMLM)和ASR模型自适应掩码语言模型(AMLMM)。

[0064] 具体地，PMLM任务可以看是MLM的扩展，因此当模型错误地将掩码标记为正确字的同音或同音字时，会遭受较小的损失。因此，同音和近音素具有相似的表示形式。在PMLM中，基于启发式的语音相似性作为知识先验被注入到损失函数中。

[0065] AMMLM任务基于PMLM任务的扩展得到。由于不同的ASR系统可能有不同类型的错误，简单的语音启发式可能对ASR错误的覆盖率较低。为此，通过进一步扩展PMLM使得损失函数可以适应特定的ASR模型。具体地，可以通过引入一种数据驱动算法来提取ASR错误作为种子错误集。然后将错误泛化并融合到新的损失函数AMMLM中。通过这种方式，预训练的ARoBERT模型可以拟合特定ASR系统产生的特定错误，而这些错误通常无法被PMLM中使用的启发式方法捕获。

[0066] ARoBERT使用与BERT相同的训练范式进行微调(finetune)，因此可以直接应用于各种基于PLM的SLU任务方法，并无需任何修改。

[0067] ARoBERT与BERT共享相同的转换器编码器架构来学习token(指示，在此例中为“字”)表示。它与BERT式模型的不同之处在于它在预训练期间结合了丰富的语音知识。具体来说，transformer编码器应该容忍ASR错误，这些错误在语音上与正确的抄本无错误相似。

[0068] 在图7中，我们给出了ARoBERT的三个预训练任务的说明性示例，即掩码语言建模(MLM)、语音感知MLM(PMLM)和可选的ASR模型自适应MLM(AMMLM)。将三个任务的损失函数分别表示为L_{MLM}、L_{PMLM}和L_{AMMLM}。ARoBERT的整体损失函数定义如下：

$$L = L_{MLM} + \lambda_1 L_{PMLM} + \lambda_2 L_{AMMLM} \quad (1)$$

其中λ_1和λ_2是平衡超参数。下面，我们将详细描述这三个任务。

[0069] 1)掩码语言建模(MLM)：与下一句预测相比，MLM对于BERT预训练更有效。因此，在ARoBERT中，可以采用MLM作为基本的预训练任务。在介绍PMLM和AMMLM之前，有必要仔细研究一下MLM的机制。让底层PLM由θ参数化。词汇集表示为V。假设任意字作为字表V中的索引)被屏蔽以用于模型预测。基于字的MLM损失$L_{MLM}(m)$定义为：

$$\mathcal{L}_{MLM}(m) = -\sum_{i=1}^{|V|} y_{i,m} \cdot \log \Pr(i, m|\theta), \quad (2)$$

$L_{MLM}(m)$是one-hot向量$\vec{y}_m$(其中第m个元素为1，其余为0)与模型预测概率分布之间的交叉熵。$y_{i,m}$是$\vec{y}_m$的第i个元素，Pr(i,m|θ)是m是字表V中第i个字的概率，通过由θ参数化的PLM预测。整体损失函数$L_{MLM}(m)$是语料库中所有被屏蔽字的损失之和。

[0070] 2)语音感知MLM(PMLM)：如前所述，MLM无法对单词的语音知识进行编码。为此在ARoBERT中进一步定义了基于字的PMLM损失$L_{PMLM}(m)$：

$$\mathcal{L}_{PMLM}(m) = -\sum_{i=1}^{|V|} y'_{i,m} \cdot \log \Pr(i, m|\theta), \quad (3)$$

其中$y'_{i,m}$整合了需要模型去近似的语音知识。将字表V中第i个和第m个标记之间的语音相似度表示为sim(i,m)。定义$y'_{i,m}$的一种简单方法是设$y'_{i,m} \propto sim(i,m)$。但是，它忽略了两个词之间的语义关系。此外，这种做法显着扩大了预训练数据的规模，从而增加了计算复杂度。具体地，|V|的数值(即在i=1,…,|V|时的$y'_{i,m}$)需要在预训练期间针对每个屏蔽字提供给所述模型。

[0071] 在ARoBERT中，对于每个掩码字m，可以检索前k个在语音上最相似的字。将这些字的索引集合表示为C_m。于是$y'_{i,m}$值定义如下：

$$y'_{i,m}=\begin{cases}\mathcal{M} & m=i\\ \frac{(1-\mathcal{M})\cdot sim(i,m)}{\sum_{j\in C_m} sim(j,m)} & i\in C_m\\ 0 & \text{其他}\end{cases} \quad (4)$$

其中M是一个预定义的常数(0<M<1)。因此$L_{PMLM}(m)$以重写为以下公式：

$$\mathcal{L}_{PMLM}(m)=-\mathcal{M}\cdot\log\Pr(m,m|\theta)-\sum_{i\in C_m} y'_{i,m}\cdot\log\Pr(i,m|\theta), \quad (5)$$

由此得到优化的$L_{PMLM}(m)$。

[0072] 与$L_{MLM}(m)$相比，$L_{PMLM}(m)$对错误预测为同音词或近音词的字具有更高的容忍度，其程度与两个单词之间的语音相似度成线性比例。因此，ARoBERT学习的表示对ASR错误不太敏感。与MLM类似，损失函数L_{PMLM}是所有屏蔽字的损失($L_{PMLM}(m)$)的总和。

[0073] 剩下的问题是如何正确计算语音相似度sim(i,m)。在普通话中，一个汉字的读音可以用拼音来表示。与英语中元音和辅音构成单词发音不同，汉语的音标主要由三个部分组成：声母、韵母和声调。一个简单的例子如上图4所示。

[0074] 可以计算语音相似度sim(i,m)如下：

$$\begin{aligned}sim(i,m)=&\alpha_1\cdot\mathbf{1}\big(initial(i)=initial(m)\big)\\&+\alpha_2\cdot\mathbf{1}\big(final(i)=final(m)\big)\\&+(1-\alpha_1-\alpha_2)\cdot\mathbf{1}\big(tone(i)=tone(m)\big),\end{aligned} \quad (6)$$

其中$0<\alpha_1<1$，$0<\alpha_2<1$和$0<\alpha_1+\alpha_2<1$。1(•)是指示函数，如果输入布尔表达式为真则返回1，否则返回0。initial(•)、final(•)和tone(•)分别代表底层汉字的语音成分。上述方法是基于启发式的，并且作为预训练ARoBERT的先验知识。

[0075] 3)ASR模型自适应MLM(AMMLM)：基于启发式的PMLM任务有一个相对强烈的假设，即语音相似性与ASR错误直接相关。然而，在实际应用中并非总是如此。AMMLM预训练任务是对PMLM的补充，旨在学习能够拟合ASR模型生成的错误的鲁棒表示。

[0076] 形式上，设*S*是ASR模型从人类语音生成的文本序列，*S'*是已被人类注释者纠正的相应文本。基于*S*和*S'*之间的对齐，生成了一个种子错误集W={(m,m')}，这样字表V中的第m个字可以被底层ASR模型错误地替换为第m'个字。W的构造的一个缺点是它需要人工更正转录文本的繁琐工作。为了最大限度地减少人工量并使发现的错误更广泛地应用于看不见的文本，在此进一步提出了一种ASR错误扩展算法。

[0077] 如所见，汉字的读音很大程度上取决于它们的声母和韵母。因此，有必要发现在发生ASR错误时声母和韵母是如何被替换掉的。出于两个原因，在此可以不考虑音调的替代。i)在PMLM中，当用于基于启发式的语音相似度计算的1和α_2相对较大时，如果两个汉字共享相同的声母和韵母，则它们已经具有很高的相似度。因此，音调相似性的知识主要由PMLM捕获。ii)包含用于ASR错误扩展的音调可能会在很大程度上扩展概率分布的参数空间，使它们不那么泛化到看不见的情况。将I和F分别表示为所有声母和韵母的集合。对于任意两个声母p,q∈I，我们通过以下公式计算声母替换概率：

$$\Pr(q|p) = \frac{\#(p,q)+\epsilon}{\sum_{\tilde{p}\in\mathcal{I}}\#(\tilde{p},q)+|\mathcal{I}|\cdot\epsilon}, \qquad (7)$$

其中#(p,q)是字m的声母p被W中另一个字m′的声母q替换的频率计数，ε是预定义的平滑因子(默认设置ε=1e-3)。计算声母替代概率的示例可以在图6中找到。

[0078] 类似地，对于两个韵母r,s∈F，我们也有韵母替换概率，定义如下：

$$\Pr(s|r) = \frac{\#(r,s)+\epsilon}{\sum_{\tilde{r}\in\mathcal{F}}\#(\tilde{r},s)+|\mathcal{F}|\cdot\epsilon}, \qquad (8)$$

其中#(r,s)是字m的韵母r被W中另一个字m′的韵母s替换的频率计数。

[0079] 基于这两个概率分布，可以直接扩展种子误差集以计算AMMLM损失。我们定义从第m个字到第i个字的替代分数(表示为subs(i,m))如下：

$$subs(i,m) = \begin{cases} \Pr(final(i)|final(m)) & initial(i)=initial(m) \\ \Pr(initial(i)|initial(m)) & final(i)=final(m) \\ 0 & \text{其他} \end{cases} \qquad (9)$$

由此得到subs(i,m)的表示。

[0080] 在此不考虑声母和韵母都不同的情况，因为这种情况在普通话ASR系统中非常罕见。令$L_{AMMLM}(m)$为基于字的AMMLM损失：

$$\mathcal{L}_{APMLM}(m) = -\sum_{i=1}^{|V|} y''_{i,m}\cdot\log\Pr(i,m|\theta), \qquad (10)$$

其中$y''_{i,m}$融合了从ASR模型产生的种子错误集中挖掘和概括的知识。与PMLM类似，在计算$y''_{i,m}$时也考虑具有最高替换分数的前k个字：

$$y''_{i,m} = \begin{cases} \mathcal{M} & m=i \\ \frac{(1-\mathcal{M})\cdot subs(i,m)}{\sum_{j\in\tilde{C}_m} subs(j,m)} & i\in\tilde{C}_m \\ 0 & \text{其他} \end{cases} \qquad (11)$$

其中$\tilde{C}_m$是对应第m个字而言具有top-k替换分数的字集合。很容易看出：

$$\mathcal{L}_{AMMLM}(m) = -\mathcal{M}\cdot\log\Pr(m,m|\theta) - \sum_{i\in\tilde{C}_m} y''_{i,m}\cdot\log\Pr(i,m|\theta). \qquad (12)$$

与PMLM类似，AMMLM任务的损失函数是预训练语料库中所有屏蔽字的字损失总和($L_{AMMLM}(m)$)。在实现中，对于ARoBERT字表V中的所有汉字，我们在模型预训练前为例PMLM和AMMLM任务已经计算了所有的分数$y''_{i,m}$和$y'_{i,m}$。得分的总数($y''_{i,m}$和$y'_{i,m}$)为2k•|V|。因此，在预训练过程中，优化算法只需要访问内存中的相应值，这使得ARoBERT的预训练过程非常高效。

[0081] 4)通过时间更新ARoBERT：进一步分析当底层ASR系统发生变化时ARoBERT应该如

何更新。在ARoBERT中，我们的损失函数分为三个部分，分别是L_{MLM}、L_{PMLM}和L_{AMMLM}。可以看到，L_{MLM}和L_{PMLM}的优化不是特定于ASR模型的。因此，ASR系统的变化不会影响L_{MLM}和L_{PMLM}的值。相比之下，L_{AMMLM}与特定的ASR系统相关。

[0082] 由于ARoBERT使用transformer编码器来学习ASR鲁棒表示，因此很容易微调ARoBERT以解决各种SLU任务。例如，当ARoBERT应用于用户意图分类时，可以遵循BERT微调的相同过程。对于一些复杂的任务，例如槽填充，需要后处理步骤来生成完整的结果。

[0083] 提议的ARoBERT的主干主要基于BERT。因此，ARoBERT能够处理BERT可以处理的任何下游任务，并具有更高的ASR错误鲁棒性。

[0084] 为此，本发明还可以实现为一种口语语言理解系统。图8示出了根据本发明的一个口语语言理解系统的组成示意图。如图所示，该SLU系统可以包括自动语音识别ASR系统，用于将获取的用户语言输入识别为文字。例如，对用户输入的语音“粉底液可以退吗”进行识别，并且发生了近音字识别错误，将输入语音错位的识别为“分泌液可以退吗”。

[0085] 进一步地，该SLU系统还可以包括根据本发明如上所述获取的预训练语言模型。该PLM利用PMLM和可选地AMMLM任务能够识别语音相似性。因此，在识别出的文字包含ASR系统的识别错误的情况下，能够基于当前口语理解任务，自动更正识别出的文字中包含的错误，由此确保相应口语理解任务的正确执行。具体地，可以基于用户的上下文环境(例如，客服对话环境)，将“分泌液”更正为“粉底液”，并与用户在前购买的“XXX轻垫粉底液”相匹配。

[0086] 应该理解的是，如上给出的是SLU系统用于进行匹配任务的一个例子，在其他应用中，也可以利用本发明训练得到的预训练模型用于其他任务，例如主题归类任务。

[0087] 图9示出了根据本发明一实施例可用于实现上述用于预训练语言模型的数据处理方法的计算设备的结构示意图。

[0088] 参见图9，计算设备900包括存储器910和处理器920。

[0089] 处理器920可以是一个多核的处理器，也可以包含多个处理器。在一些实施例中，处理器920可以包含一个通用的主处理器以及一个或多个特殊的协处理器，例如图形处理器(GPU)、数字信号处理器(DSP)等等。在一些实施例中，处理器920可以使用定制的电路实现，例如特定用途集成电路(ASIC，ApplicationSpecificIntegratedCircuit)或者现场可编程逻辑门阵列(FPGA，FieldProgrammableGateArrays)。

[0090] 存储器910可以包括各种类型的存储单元，例如系统内存、只读存储器(ROM)，和永久存储装置。其中，ROM可以存储处理器920或者计算机的其他模块需要的静态数据或者指令。永久存储装置可以是可读写的存储装置。永久存储装置可以是即使计算机断电后也不会失去存储的指令和数据的非易失性存储设备。在一些实施方式中，永久性存储装置采用大容量存储装置(例如磁或光盘、闪存)作为永久存储装置。另外一些实施方式中，永久性存储装置可以是可移除的存储设备(例如软盘、光驱)。系统内存可以是可读写存储设备或者易失性可读写存储设备，例如动态随机访问内存。系统内存可以存储一些或者所有处理器在运行时需要的指令和数据。此外，存储器910可以包括任意计算机可读存储媒介的组合，包括各种类型的半导体存储芯片(DRAM，SRAM，SDRAM，闪存，可编程只读存储器)，磁盘和/或光盘也可以采用。在一些实施方式中，存储器910可以包括可读和/或写的可移除的存储设备，例如激光唱片(CD)、只读数字多功能光盘(例如DVD-ROM，双层DVD-ROM)、只读蓝光光盘、超密度光盘、闪存卡(例如SD卡、minSD卡、Micro-SD卡等等)、磁性软盘等等。计算机可读存

储媒介不包含载波和通过无线或有线传输的瞬间电子信号。

[0091] 存储器910上存储有可执行代码，当可执行代码被处理器920处理时，可以使处理器920执行上文述及的用于预训练语言模型的数据处理方法。

[0092] 上文中已经参考附图详细描述了根据本发明的用于预训练语言模型的数据处理以及相应的口语语音理解系统。

[0093] 此外，根据本发明的方法还可以实现为一种计算机程序或计算机程序产品，该计算机程序或计算机程序产品包括用于执行本发明的上述方法中限定的上述各步骤的计算机程序代码指令。

[0094] 或者，本发明还可以实施为一种非暂时性机器可读存储介质(或计算机可读存储介质、或机器可读存储介质)，其上存储有可执行代码(或计算机程序、或计算机指令代码)，当所述可执行代码(或计算机程序、或计算机指令代码)被电子设备(或计算设备、服务器等)的处理器执行时，使所述处理器执行根据本发明的上述方法的各个步骤。本领域技术人员还将明白的是，结合这里的公开所描述的各种示例性逻辑块、模块、电路和算法步骤可以被实现为电子硬件、计算机软件或两者的组合。

[0095] 附图中的流程图和框图显示了根据本发明的多个实施例的系统和方法的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分，所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标记的功能也可以以不同于附图中所标记的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0096] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

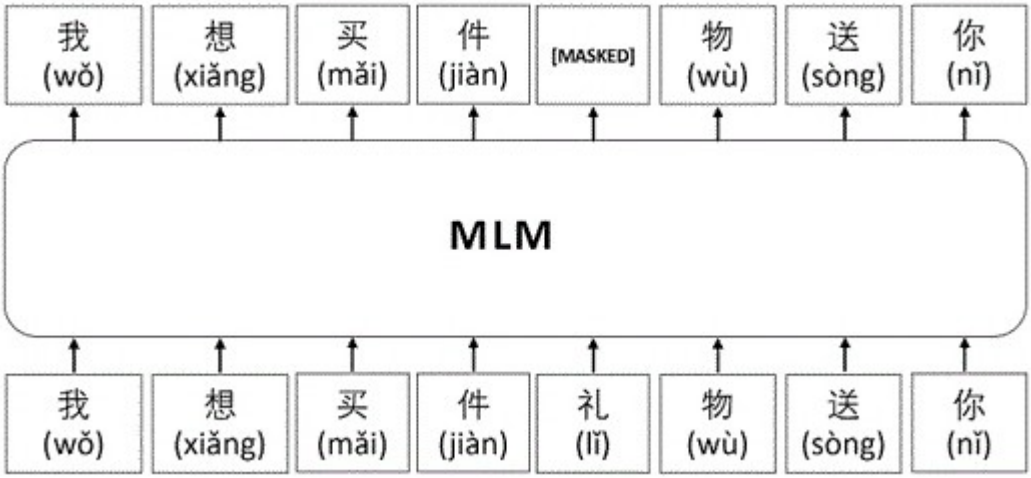

我 (wǒ)
想 (xiǎng)
买 (mǎi)
件 (jiàn)
[MASKED]
物 (wù)
送 (sòng)
你 (nǐ)
MLM
我 (wǒ)
想 (xiǎng)
买 (mǎi)
件 (jiàn)
礼 (lǐ)
物 (wù)
送 (sòng)
你 (nǐ)

图1

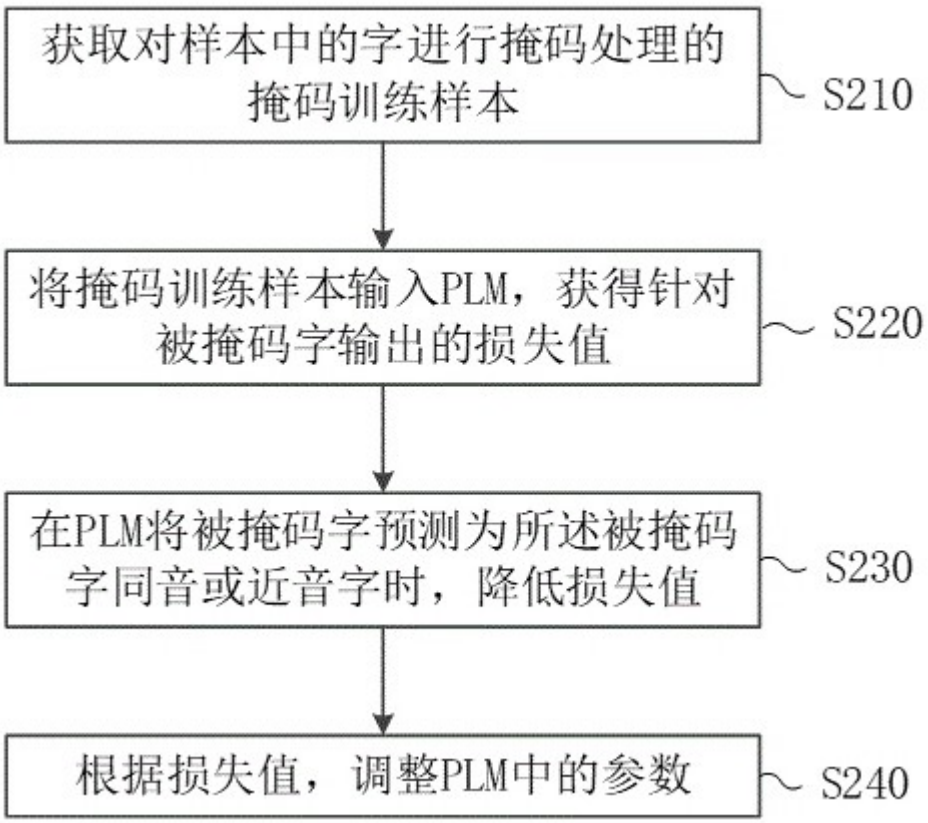

获取对样本中的字进行掩码处理的掩码训练样本
S210
将掩码训练样本输入PLM，获得针对被掩码字输出的损失值
S220
在PLM将被掩码字预测为所述被掩码字同音或近音字时，降低损失值
S230
根据损失值，调整PLM中的参数
S240

图2

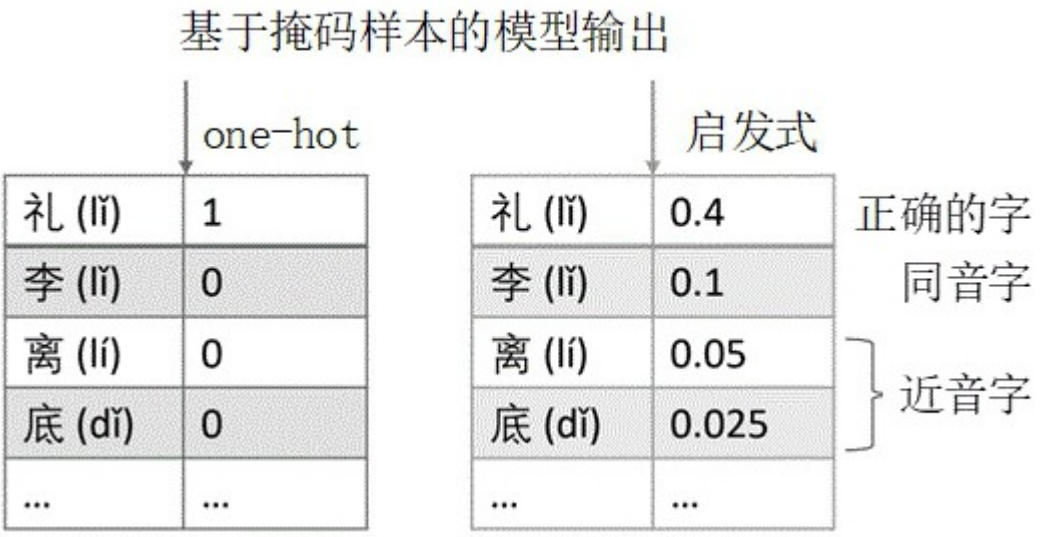

基于掩码样本的模型输出
one-hot
启发式
礼 (lǐ) 1
李 (lǐ) 0
离 (lí) 0
底 (dǐ) 0
... ...
礼 (lǐ) 0.4
李 (lǐ) 0.1
离 (lí) 0.05
底 (dǐ) 0.025
... ...
正确的字
同音字
近音字

图3

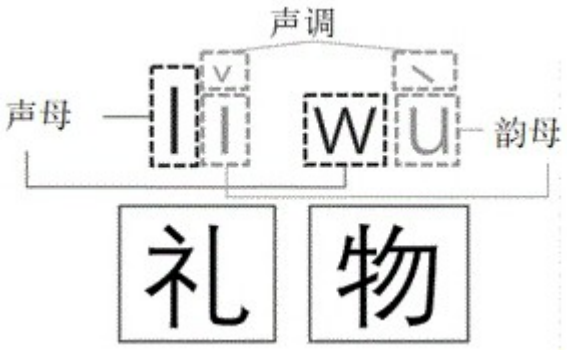

声调
声母
韵母
礼
物

图4

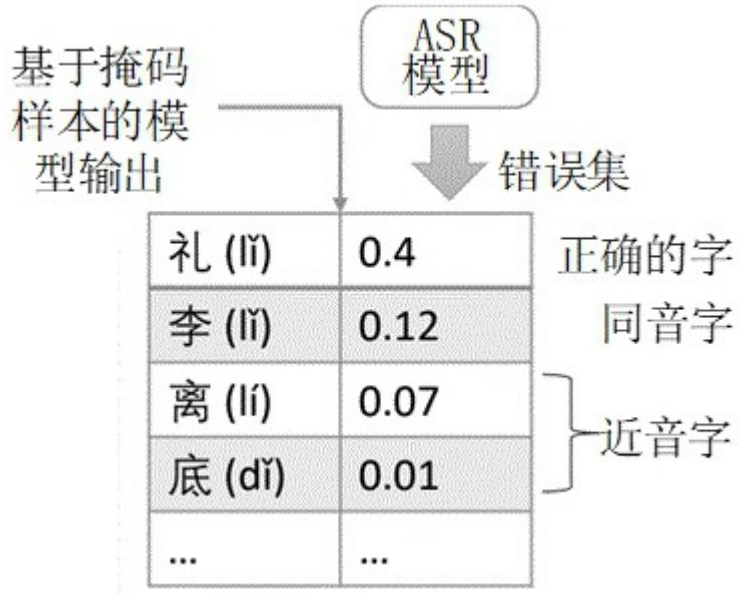

基于掩码
样本的模
型输出
ASR
模型
错误集
礼 (lǐ)
0.4
正确的字
李 (lǐ)
0.12
同音字
离 (lí)
0.07
近音字
底 (dǐ)
0.01
...
...

图5

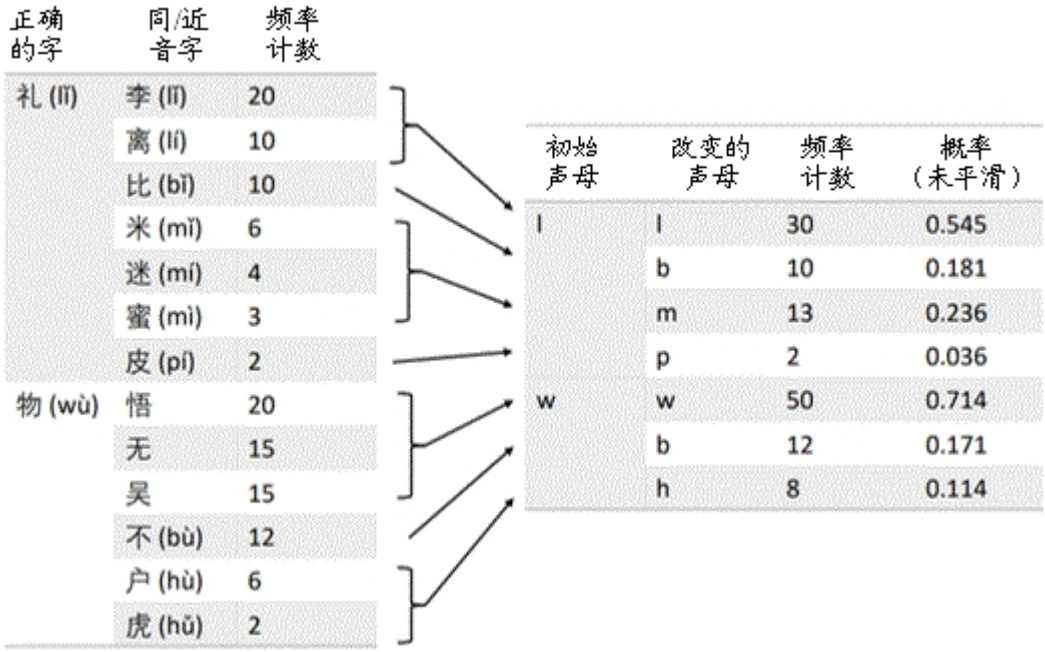

正确的字
同/近音字
频率计数
礼 (lǐ)
李 (lǐ) 20
离 (lí) 10
比 (bǐ) 10
米 (mǐ) 6
迷 (mí) 4
蜜 (mì) 3
皮 (pí) 2
物 (wù)
悟 20
无 15
吴 15
不 (bù) 12
户 (hù) 6
虎 (hǔ) 2
初始声母
改变的声母
频率计数
概率（未平滑）
l
l 30 0.545
b 10 0.181
m 13 0.236
p 2 0.036
w
w 50 0.714
b 12 0.171
h 8 0.114

图6

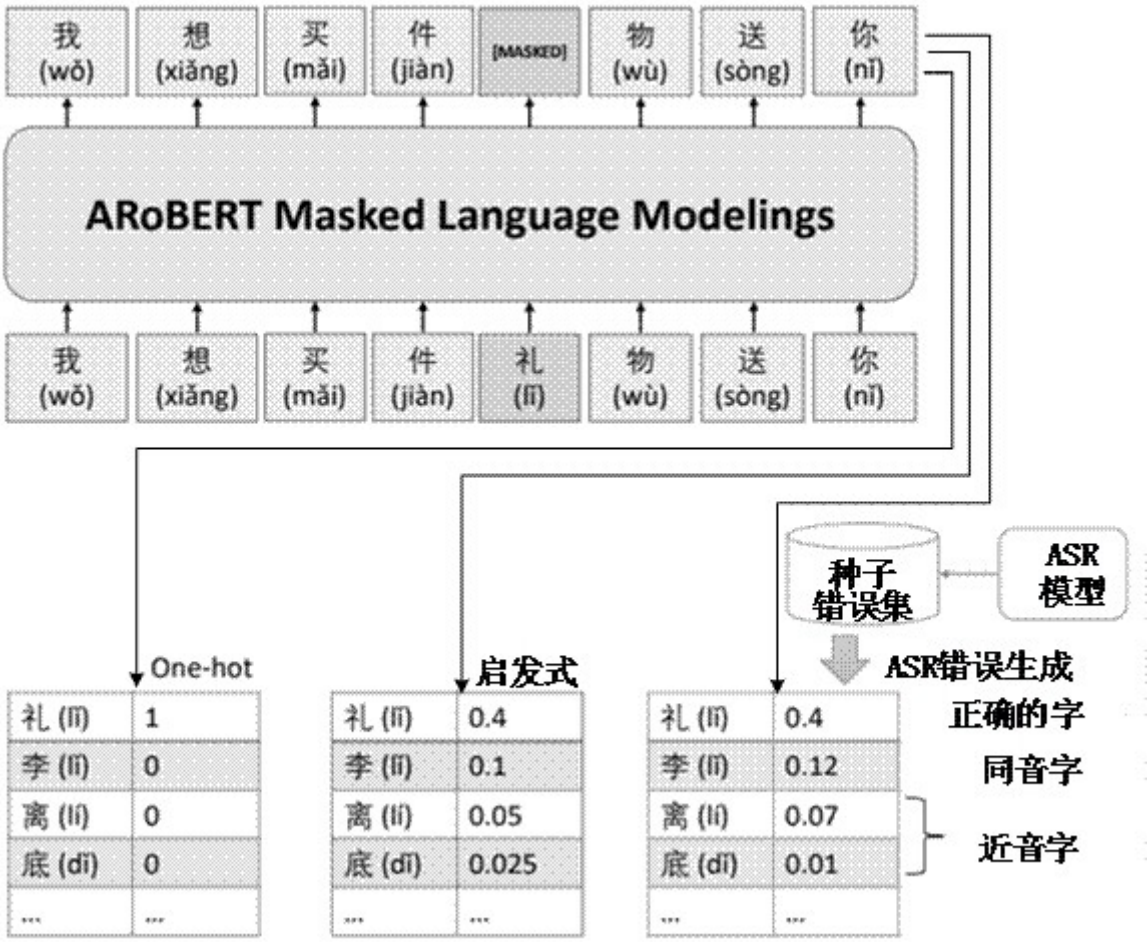

我 (wǒ)
想 (xiǎng)
买 (mǎi)
件 (jiàn)
[MASKED]
物 (wù)
送 (sòng)
你 (nǐ)
ARoBERT Masked Language Modelings
我 (wǒ)
想 (xiǎng)
买 (mǎi)
件 (jiàn)
礼 (lǐ)
物 (wù)
送 (sòng)
你 (nǐ)
种子错误集
ASR模型
ASR错误生成
One-hot
启发式
礼 (lǐ) 1
李 (lǐ) 0
离 (lí) 0
底 (dǐ) 0
礼 (lǐ) 0.4
李 (lǐ) 0.1
离 (lí) 0.05
底 (dǐ) 0.025
礼 (lǐ) 0.4
李 (lǐ) 0.12
离 (lí) 0.07
底 (dǐ) 0.01
正确的字
同音字
近音字

图7

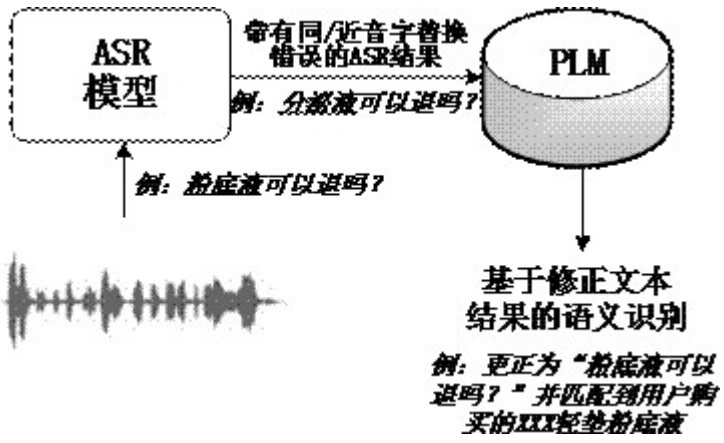

ASR模型
带有同/近音字替换错误的ASR结果
例：分泌液可以退吗？
PLM
例：粉底液可以退吗？
基于修正文本结果的语义识别
例：更正为“粉底液可以退吗？”并匹配到用户购买的xxx轻垫粉底液

图8

图9