(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 9 月 24 日 (24.09.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/186778 A1

(51) 国际专利分类号:
G06F 16/33 (2019.01)

(21) 国际申请号: PCT/CN2019/117237

(22) 国际申请日: 2019 年 11 月 11 日 (11.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910199221.9 2019年3月15日 (15.03.2019) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(72) 发明人: 解笑 (XIE, Xiao); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 徐国强 (XU, Guoqiang); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 邱寒 (QIU, Han); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市赛恩倍吉知识产权代理有限公司 (SHENZHEN SCIENBIZIP INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国广东省深圳市龙华新区龙观东路83号荣群大厦9楼, Guangdong 518109 (CN)。

(54) **Title:** ERROR WORD CORRECTION METHOD AND DEVICE, COMPUTER DEVICE, AND STORAGE MEDIUM

(54) **发明名称:** 错词纠正方法、装置、计算机装置及存储介质

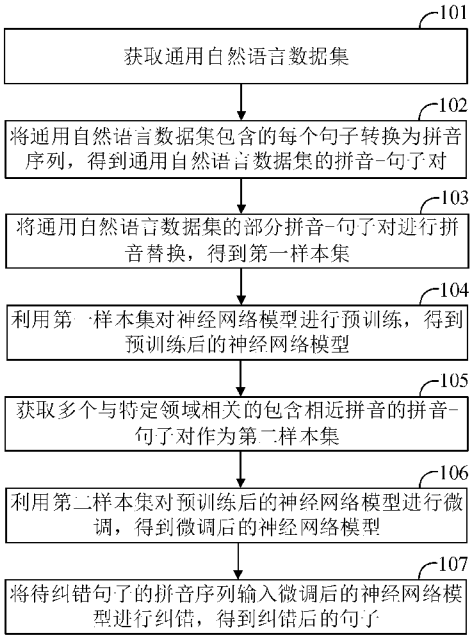

图 1

101 Obtain a general natural language data set
102 Convert each sentence comprised in the general natural language data set into a Pinyin sequence to obtain obtaining Pinyin-sentence pairs of the general natural language data set
103 Perform Pinyin replacement on some of the Pinyin-sentence pairs of the general natural language data set to obtain a first sample set
104 Pre-train a neural network model using the first sample set to obtain a pre-trained neural network model
105 Obtain a plurality of Pinyin-sentence pairs comprising similar Pinyin and related to a specific field as a second sample set
106 Perform fine tuning on the pre-trained neural network model using the second sample set to obtain a fine-tuned neural network model
107 Input a Pinyin sequence of a sentence to be corrected into the fine-tuned neural network model for correction to obtain a corrected sentence

(57) **Abstract:** An error word correction method and device, a computer device, and a storage medium. The error word correction method comprises: obtaining a general natural language data set (101); converting each sentence comprised in the natural language data set into a Pinyin sequence to obtain obtaining Pinyin-sentence pairs of the general natural language data set (102); performing Pinyin replacement on some of the Pinyin-sentence pairs of the general natural language data set to obtain a first sample set (103); pre-training a neural network model using the first sample set to obtain a pre-trained neural network model (104); obtaining a plurality of

WO 2020/186778 A1

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

Pinyin-sentence pairs comprising similar Pinyin and related to a specific field as a second sample set (105); performing fine tuning on the pre-trained neural network model using the second sample set to obtain a fine-tuned neural network model (106); and inputting a Pinyin sequence of a sentence to be corrected into the fine-tuned neural network model for correction to obtain a corrected sentence (107). By means of the method, error correction can be performed on special words identified as common words in language identification.

(57) 摘要：一种错词纠正方法、装置、计算机装置及存储介质。所述错词纠正方法包括：获取通用自然语言数据集（101）；将自然语言数据集包含的每个句子转换为拼音序列，得到通用自然语言数据集的拼音-句子对（102）；将通用自然语言数据集的部分拼音-句子对进行拼音替换，得到第一样本集（103）；利用第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型（104）；获取多个与特定领域相关的含相近拼音的拼音-句子对作为第二样本集（105）；利用第二样本集对预训练后的神经网络模型进行微调，得到微调后的神经网络模型（106）；将待纠错句子的拼音序列输入微调后的神经网络模型进行纠错，得到纠错后的句子（107）。该方法可以对语言识别中专有词语被识别为常用词进行纠错。

错词纠正方法、装置、计算机装置及存储介质

本申请要求于 2019 年 03 月 15 日提交中国专利局，申请号为 201910199221.9 申请名称为“错词纠正方法、装置、计算机装置及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及语音识别技术领域，具体涉及一种错词纠正方法、装置、计算机装置及非易失性可读存储介质。

背景技术

随着语音识别应用场景的迅猛拓宽，语音识别技术越来越成熟，市场对高准确度的语音识别需求越来越强烈。对于一些开发具有语音识别功能产品的公司，更多的情况是使用通用系统的语音识别模块，不针对其具体应用场景进行识别，就会很容易出现将某些专有词语识别为常用词。例如将“需要为谁投保”识别为“需要为谁淘宝”，由于其并没有明显的错误，现有错词纠正系统难以发现此类错误。

目前，对于如何提升语言识别在实际应用场景中的纠正效果并没有一个有效的解决方法。如何制定合适的方案，以减少语音识别的偏差，提升用户体验，是相关技术人员目前需要解决的技术问题。

发明内容

鉴于以上内容，有必要提出一种错词纠正方法、装置、计算机装置及非易失性可读存储介质，可以对语言识别中专有词语被识别为常用词进行纠错。

本申请的第一方面提供一种错词纠正方法，所述方法包括：

获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

本申请的第二方面提供一种错词纠正装置，所述装置包括：

第一获取模块，用于获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

转换模块，用于将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

生成模块，用于从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

预训练模块，用于用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

第二获取模块，用于获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

微调模块，用于利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

纠错模块，用于将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

本申请的第三方面提供一种计算机装置，所述计算机装置包括处理器，所述处理器用于执行存储器中存储的计算机可读指令时实现所述错词纠正方法。

本申请的第四方面提供一种非易失性可读存储介质，其上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现所述错词纠正方法。

本申请获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集

的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。本实施例可以解决由于语音识别系统的通用性在特定领域内无法准确预测专有词语的问题，能够对语言识别中专有词语被识别为常用词进行纠错。

附图说明

图 1 是本申请实施例提供的错词纠正方法的流程图。

图 2 是本申请实施例提供的错词纠正装置的结构图。

图 3 是本申请实施例提供的计算机装置的示意图。

具体实施方式

为了能够更清楚地理解本申请的上述目的、特征和优点，下面结合附图和具体实施例对本申请进行详细描述。需要说明的是，在不冲突的情况下，本申请的实施例及实施例中的特征可以相互组合。

优选地，本申请的错词纠正方法应用在一个或者多个计算机装置中。所述计算机装置是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路(Application Specific Integrated Circuit，ASIC)、可编程门阵列(Field－Programmable Gate Array，FPGA)、数字处理器(Digital Signal Processor，DSP)、嵌入式设备等。

所述计算机装置可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述计算机装置可以与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互。

实施例一

图 1 是本申请实施例一提供的错词纠正方法的流程图。所述错词纠正方法应用于计算机装置。

本申请的错词纠正方法是对语言识别得到的句子进行纠错。所述错词纠正方法可以解决由于语音识别系统的通用性在特定领域内无法准确预测专有词语的问题，同时增强了纠错系统在专有词语被替换为常用词时的错词寻找能力，提升用户的使用体验。

如图 1 所示，所述错词纠正方法包括：

步骤 101，获取通用自然语言数据集，所述通用自然语言数据集包含多个句子。

所述通用自然语言数据集是包含日常用语的中文文本。

可以从书籍、新闻、网页（例如百度百科、维基百科等）等数据源中收集所述通用自然语言数据集。例如，可以对书籍中的文字进行文字识别，得到所述通用自然语言数据集。又如，可以对播报的新闻进行语言识别，得到所述通用自然语言数据集。再如，可以从网页中抓取文本，得到所述通用自然语言数据集。

或者，可以从预设数据库读取所述通用自然语言数据集。所述预设数据库可以预先存储大量的中文文本。

或者，可以接收用户输入的中文文本，将用户输入的中文文本作为所述通用自然语言数据集。

步骤 102，将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对。

在本实施例中，所述通用自然语言数据集可以包括多个中文文本，每个中文文本可以包括多个句子（即多句话）。这种情况下，可以根据标点符号（例如逗号、分号、句号等）将每个中文文本划分为多个句子，将划分得到的每个句子转换为拼音序列，即得到每个句子对应的拼音-句子对。

可以根据汉字的 ASCII 码将所述句子转换为拼音序列。由于汉字在计算机系统中以 ASCII 码表示，只需要利用计算机系统中已有的或用户建立的每个拼音与每个 ASCII 码对应关系，即可实现将句子转换成拼音序列。若句子含有多音字，可以列出多音字的多个拼音，接收用户选择的正确拼音。

或者，可以根据汉字的 Unicode 值将所述句子转换为拼音序列。具体步骤如下：

（1）建立拼音-编号对照表，对所有拼音进行编号并将所有拼音对应的编号添加到所述拼音-编号对照表中。所有汉字的拼音不超过 512 个，可以用两个字节对拼音进行编号。每个拼音对应一个编号。

（2）建立 Unicode 值-拼音编号对照表，将汉字对应拼音的编号按照汉字的 Unicode 值添加到所述 Unicode 值-拼音编号对照表中。

（3）逐一读取所述句子中的待转换汉字，确定所述待转换汉字的 Unicode 值，根据所述待转换汉字的 Unicode 值从所述 Unicode 值-拼音编号对照表中获取所述待转换汉字对应的拼音的编号，根据所述待转换汉字对应的拼音的编号从所述拼音-编号对照表获得所述待转换汉字对应的拼音，从而将所述句子中的每个汉字转换为拼音。

若所述句子中含有多音字，可以在上述步骤（2）中将所述多音字对应的多个拼音的编号按照所述多音字的Unicode值添加到所述Unicode值-拼音编号对照表中，在上述（3）中确定所述多音字的 Unicode 值，根据所述多音字的 Unicode 值从所述 Unicode 值-拼音编号对照表中获取所述多音字对应的多个拼音的编号，根据所述多音字对应的多个拼音的编号从所述拼音-编号对照表获得所述多音字对应的多个拼音。可以接收用户从所述多个拼音中选择的正确拼音，将用户选择的拼音作为所述多音字在所述句子中的正确拼音。

步骤 103，从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集。

可以从所述通用自然语言数据集的拼音-句子对中随机选择所述多个拼音-句子对，将选择的每个拼音-句子中的部分拼音替换为相近拼音。

可以按照预设比例从通用自然语言数据集的拼音-句子对中选择多个拼音-句子对。例如，可以从所述通用自然语言数据集的拼音-句子对中选择 20%的拼音-句子对进行拼音替换。举例来说，若所述通用自然语言数据集包括 100 个句子（即包括 100 个拼音-句子对），则选择 20 个拼音-句子对进行拼音替换。

所述第一样本集的训练样本包括未选择的拼音-句子对，即正确的拼音-句子对，还包括替换后的拼音-句子对，即将部分拼音替换为相近拼音的拼音-句子对。

本申请主要用于对语言识别得到的句子进行纠错。由于语音识别得到的句子错误大多是句子中的词语有意义而句子无意义，例如“需要为谁投保”有时会被识别成“需要为谁淘宝”。因此，不仅需要正确的拼音-句子对作为训练样本，还需要将部分拼音替换为相近拼音的拼音-句子对作为模型的训练样本。

步骤 104，利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型。

所述神经网络模型的输入为拼音序列，输出为对应的句子（即汉字序列），对拼音序列中的每一个拼音，预测其对应的汉字。

在对神经网络模型进行训练时，以每个未选择的拼音-句子对（即未替换的拼音-句子对）和每个替换后的拼音-句子对作为训练样本。拼音-句子对中的拼音序列为神经网络模型的输入，拼音-句子对中的句子为真实结果。

在本实施例中，所述神经网络模型可以是 transformer 模型。

transformer 模型可以接受一串序列作为输入，同时输出一串序列，在本申请中，Transformer 模型将拼音序列作为输入，输出汉字序列。

transformer 模型包含编码层、自注意力层、解码层。其中编码层和解码层分别对应拼音的编码和到汉字的解码。自注意力层则用于重复拼音的汉字预测。由于汉字拼音有大量重复，不同的汉字和词语对应于相同的拼音，例如“爆笑”和“报效”拥有同样的拼音和声调，因此在每一个拼音所在进行预测时，需要“关注”整个句子的拼音序列，而不是只看当前位置的拼音。自注意力机制可以使得某一位置的拼音获得其它所有位置的拼音表示，从而做出更符合该句子场景的汉字预测。

在经过大量样本的训练后，该 Ttransformer 模型可以通过输入拼音序列来输出对应的汉字序列。

步骤 105，获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集。

所述第二样本集中的每个训练样本是与特定领域相关的一个拼音-句子对，该拼音-句子对中包含与所述特定领域相关的相近拼音。

特定领域是本方法所要应用的专有领域，例如法律、保险等。

步骤 101 获得的语言数据集是通用自然语言数据集，主要包含一些日常用语，根据通用自然语言数据集得到的第一样本集是关于日常用语的训练样本，因此预训练得到的神经网络模型在当日常生活中的句子有明显的语音识别错误时，可以进行很好地纠错。但当遇到某些例如法律、保险等专有领域，则神经网络模型的纠错效果有所下降，会将很多专有词语识别为日常用语。例如将“需要为谁投保”中的“投保”识别为“淘宝”。因此要应用到特定领域进行错词纠错时，需要该特定领域的样本数据。

可以按照下述方法获取多个与特定领域相关的包含相近拼音的拼音-句子对：

获取所述特定领域的文本数据集，所述文本数据集包含多个句子；

将所述文本数据集包含的每个句子转换为拼音序列，得到所述文本数据集的拼音-句子对；

将所述文本数据集的拼音-句子对中所述特定领域的专有词语的拼音替换为相近拼音，得到与特定领域相关的包含相近拼音的拼音-句子对。例如，将“需要为谁投保”中的“投保”的拼音(tou，二声，bao，三声)替换为“淘宝”的拼音(tao，二声，bao，三声)。

或者，可以预先建立数据库，用于存储所述特定领域识别错误的拼音-句子对，从所述数据库获取多个与特定领域相关的包含相近拼音的拼音-句子对。

步骤 106，利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型。

利用所述第二样本集对所述神经网络模型进行微调的目的是使所述神经网络模型更适用于特定领域，提高特定领域的纠错准确率。

微调训练后的模型在拼音近似的情况下，更倾向于预测为该特定领域的专有词语，从而提高语音识别错误的错词纠正效果。

可以固定所述神经网络模型的前面几层神经元的权值，微调神经网络模型的后面几层神经元的权值。这样做主要是为了避免第二样本集过小出现过拟合现象，神经网络模型前几层神经元一般包含更多的一般特征，对于许多任务而言非常重要，但是后面几层神经元的特征学习注重高层特征，不同的数据集间差异较大。

步骤 107，将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

语言识别得到的结果可以包括多个中文文本，每个中文文本可以包括多个待纠错句子（即多句话）。这种情况下，可以根据标点符号（例如逗号、分号、句号等）将语言识别得到的中文文本划分为多个待纠错句子，将划分得到的每个待纠错句子转换为拼音序列。

可以根据汉字的 ASCII 码将所述待纠错句子转换为拼音序列。或者，可以根据汉字的 Unicode 值将所述待纠错句子转换为拼音序列。将待纠错句子转换为拼音序列的方法可以参考步骤 102。

或者，可以接收用户输入的待纠错句子，将所述待纠错句子转换为拼音序列。例如，可以生成用户界面，从所述用户界面接收用户输入的待纠错句子。也可以直接接收用户输入的待纠错句子的拼音序列。

实施例一的错词纠正方法获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。本实施例可以解决由于语音识别系统的通用性在特定领域内无法准确预测专有词语的问题，能够对语言识别中专有词语被识别为常用词进行纠错。

在另一实施例中，所述错词纠正方法还可以包括：对输入的语音进行识别，得到所述待纠错句子。可以采用各种语音识别技术，例如动态时间规整(Dynamic Time Warping，

DTW)、隐马尔可夫模型（Hidden Markov Model，HMM）、矢量量化（Vector Quantization，VQ）、人工神经网络（Artificial Neural Network，ANN）等技术对所述语音进行识别。

实施例二

图 2 是本申请实施例二提供的错词纠正装置的结构图。所述错词纠正装置 20 应用于计算机装置。如图 2 所示，所述错词纠正装置 20 可以包括第一获取模块 201、转换模块 202、生成模块 203、预训练模块 204、第二获取模块 205、微调模块 206、纠错模块 207。

第一获取模块 201，用于获取通用自然语言数据集，所述通用自然语言数据集包含多个句子。

所述通用自然语言数据集是包含日常用语的中文文本。

可以从书籍、新闻、网页（例如百度百科、维基百科等）等数据源中收集所述通用自然语言数据集。例如，可以对书籍中的文字进行文字识别，得到所述通用自然语言数据集。又如，可以对播报的新闻进行语言识别，得到所述通用自然语言数据集。再如，可以从网页中抓取文本，得到所述通用自然语言数据集。

或者，可以从预设数据库读取所述通用自然语言数据集。所述预设数据库可以预先存储大量的中文文本。

或者，可以接收用户输入的中文文本，将用户输入的中文文本作为所述通用自然语言数据集。

转换模块 202，用于将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对。

在本实施例中，所述通用自然语言数据集可以包括多个中文文本，每个中文文本可以包括多个句子（即多句话）。这种情况下，可以根据标点符号（例如逗号、分号、句号等）将每个中文文本划分为多个句子，将划分得到的每个句子转换为拼音序列，即得到每个句子对应的拼音-句子对。

可以根据汉字的 ASCII 码将所述句子转换为拼音序列。由于汉字在计算机系统中以 ASCII 码表示，只需要利用计算机系统中已有的或用户建立的每个拼音与每个 ASCII 码对应关系，即可实现将句子转换成拼音序列。若句子含有多音字，可以列出多音字的多个拼音，接收用户选择的正确拼音。

或者，可以根据汉字的 Unicode 值将所述句子转换为拼音序列。具体步骤如下：

（1）建立拼音-编号对照表，对所有拼音进行编号并将所有拼音对应的编号添加到所述拼音-编号对照表中。所有汉字的拼音不超过 512 个，可以用两个字节对拼音进行编

号。每个拼音对应一个编号。

（2）建立 Unicode 值-拼音编号对照表，将汉字对应拼音的编号按照汉字的 Unicode 值添加到所述 Unicode 值-拼音编号对照表中。

（3）逐一读取所述句子中的待转换汉字，确定所述待转换汉字的 Unicode 值，根据所述待转换汉字的 Unicode 值从所述 Unicode 值-拼音编号对照表中获取所述待转换汉字对应的拼音的编号，根据所述待转换汉字对应的拼音的编号从所述拼音-编号对照表获得所述待转换汉字对应的拼音，从而将所述句子中的每个汉字转换为拼音。

若所述句子中含有多音字，可以在上述步骤（2）中将所述多音字对应的多个拼音的编号按照所述多音字的 Unicode 值添加到所述 Unicode 值-拼音编号对照表中，在上述（3）中确定所述多音字的 Unicode 值，根据所述多音字的 Unicode 值从所述 Unicode 值-拼音编号对照表中获取所述多音字对应的多个拼音的编号，根据所述多音字对应的多个拼音的编号从所述拼音-编号对照表获得所述多音字对应的多个拼音。可以接收用户从所述多个拼音中选择的正确拼音，将用户选择的拼音作为所述多音字在所述句子中的正确拼音。

生成模块 203，用于从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集。

可以从所述通用自然语言数据集的拼音-句子对中随机选择所述多个拼音-句子对，将选择的每个拼音-句子中的部分拼音替换为相近拼音。

可以按照预设比例从通用自然语言数据集的拼音-句子对中选择多个拼音-句子对。例如，可以从所述通用自然语言数据集的拼音-句子对中选择 20%的拼音-句子对进行拼音替换。举例来说，若所述通用自然语言数据集包括 100 个句子（即包括 100 个拼音-句子对），则选择 20 个拼音-句子对进行拼音替换。

所述第一样本集的训练样本包括未选择的拼音-句子对，即正确的拼音-句子对，还包括替换后的拼音-句子对，即将部分拼音替换为相近拼音的拼音-句子对。

本申请主要用于对语言识别得到的句子进行纠错。由于语音识别得到的句子错误大多是句子中的词语有意义而句子无意义，例如“需要为谁投保”有时会被识别成“需要为谁淘宝”。因此，不仅需要正确的拼音-句子对作为训练样本，还需要将部分拼音替换为相近拼音的拼音-句子对作为模型的训练样本。

预训练模块 204，用于利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型。

所述神经网络模型的输入为拼音序列，输出为对应的句子（即汉字序列），对拼音序列中的每一个拼音，预测其对应的汉字。

在对神经网络模型进行训练时，以每个未选择的拼音-句子对（即未替换的拼音-句子对）和每个替换后的拼音-句子对作为训练样本。拼音-句子对中的拼音序列为神经网络模型的输入，拼音-句子对中的句子为真实结果。

在本实施例中，所述神经网络模型可以是 transformer 模型。

transformer 模型可以接受一串序列作为输入，同时输出一串序列，在本申请中，Transformer 模型将拼音序列作为输入，输出汉字序列。

transformer 模型包含编码层、自注意力层、解码层。其中编码层和解码层分别对应拼音的编码和到汉字的解码。

自注意力层则用于重复拼音的汉字预测。由于汉字拼音有大量重复，不同的汉字和词语对应于相同的拼音，例如“爆笑”和“报效”拥有同样的拼音和声调，因此在每一个拼音所在进行预测时，需要“关注”整个句子的拼音序列，而不是只看当前位置的拼音。自注意力机制可以使得某一位置的拼音获得其它所有位置的拼音表示，从而做出更符合该句子场景的汉字预测。

在经过大量样本的训练后，该 Ttransformer 模型可以通过输入拼音序列来输出对应的汉字序列。

第二获取模块 205，用于获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集。

所述第二样本集中的每个训练样本是与特定领域相关的一个拼音-句子对，该拼音-句子对中包含与所述特定领域相关的相近拼音。

特定领域是本方法所要应用的专有领域，例如法律、保险等。

第一获取模块 201 获得的语言数据集是通用自然语言数据集，主要包含一些日常用语，根据通用自然语言数据集得到的第一样本集是关于日常用语的训练样本，因此预训练得到的神经网络模型在当日常生活中的句子有明显的语音识别错误时，可以进行很好地纠错。但当遇到某些例如法律、保险等专有领域，则神经网络模型的纠错效果有所下降，会将很多专有词语识别为日常用语。例如将“需要为谁投保”中的“投保”识别为“淘宝”。因此要应用到特定领域进行错词纠错时，需要该特定领域的样本数据。

可以按照下述方法获取多个与特定领域相关的包含相近拼音的拼音-句子对：

获取所述特定领域的文本数据集，所述文本数据集包含多个句子；

将所述文本数据集包含的每个句子转换为拼音序列，得到所述文本数据集的拼音-

句子对；

将所述文本数据集的拼音-句子对中所述特定领域的专有词语的拼音替换为相近拼音，得到与特定领域相关的包含相近拼音的拼音-句子对。例如，将“需要为谁投保”中的“投保”的拼音(tou，二声，bao，三声)替换为“淘宝”的拼音(tao，二声，bao，三声)。

或者，可以预先建立数据库，用于存储所述特定领域识别错误的拼音-句子对，从所述数据库获取多个与特定领域相关的包含相近拼音的拼音-句子对。

微调模块 206，用于利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型。

利用所述第二样本集对所述神经网络模型进行微调的目的是使所述神经网络模型更适用于特定领域，提高特定领域的纠错准确率。

微调训练后的模型在拼音近似的情况下，更倾向于预测为该特定领域的专有词语，从而提高语音识别错误的错词纠正效果。

可以固定所述神经网络模型的前面几层神经元的权值，微调神经网络模型的后面几层神经元的权值。这样做主要是为了避免第二样本集过小出现过拟合现象，神经网络模型前几层神经元一般包含更多的一般特征，对于许多任务而言非常重要，但是后面几层神经元的特征学习注重高层特征，不同的数据集间差异较大。

纠错模块 207，用于将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

语言识别得到的结果可以包括多个中文文本，每个中文文本可以包括多个待纠错句子（即多句话）。这种情况下，可以根据标点符号（例如逗号、分号、句号等）将语言识别得到的中文文本划分为多个待纠错句子，将划分得到的每个待纠错句子转换为拼音序列。

可以根据汉字的 ASCII 码将所述待纠错句子转换为拼音序列。或者，可以根据汉字的 Unicode 值将所述待纠错句子转换为拼音序列。将待纠错句子转换为拼音序列的方法可以参考转换模块 202 的描述。

或者，可以接收用户输入的待纠错句子，将所述待纠错句子转换为拼音序列。例如，可以生成用户界面，从所述用户界面接收用户输入的待纠错句子。也可以直接接收用户输入的待纠错句子的拼音序列。

本实施例的错词纠正装置 20 获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；从所述通用自然语言数据集的拼音-句子对中选择多

个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。本实施例可以解决由于语音识别系统的通用性在特定领域内无法准确预测专有词语的问题，能够对语言识别中专有词语被识别为常用词进行纠错。

在另一实施例中，所述错词纠正装置 20 还可以包括：识别模块，对输入的语音进行识别，得到所述待纠错句子。可以采用各种语音识别技术，例如动态时间规整(Dynamic Time Warping，DTW)、隐马尔可夫模型（Hidden Markov Model，HMM）、矢量量化（Vector Quantization，VQ）、人工神经网络（Artificial Neural Network，ANN）等技术对所述语音进行识别。

实施例三

本实施例提供一种非易失性可读存储介质，该非易失性可读存储介质上存储有计算机可读指令，该计算机可读指令被处理器执行时实现上述错词纠正方法实施例中的步骤，例如图 1 所示的步骤 101-107：

步骤 101，获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

步骤 102，将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

步骤 103，从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

步骤 104，利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

步骤 105，获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

步骤 106，利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

步骤 107，将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

或者，该计算机可读指令被处理器执行时实现上述装置实施例中各模块的功能，例

如图 2 中的模块 201-207：

第一获取模块 201，用于获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

转换模块 202，用于将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

生成模块 203，用于从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

预训练模块 204，用于利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

第二获取模块 205，用于获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

微调模块 206，用于利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

纠错模块 207，用于将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

实施例四

图 3 为本申请实施例四提供的计算机装置的示意图。所述计算机装置 30 包括存储器 301、处理器 302 以及存储在所述存储器 301 中并可在所述处理器 302 上运行的计算机可读指令 303，例如错词纠正程序。所述处理器 302 执行所述计算机可读指令 303 时实现上述错词纠正方法实施例中的步骤，例如图 1 所示的步骤 101-107：

步骤 101，获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

步骤 102，将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

步骤 103，从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

步骤 104，利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

步骤 105，获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

步骤 106，利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

步骤 107，将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

或者，该计算机可读指令被处理器执行时实现上述装置实施例中各模块的功能，例如图 2 中的模块 201-207：

第一获取模块 201，用于获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

转换模块 202，用于将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

生成模块 203，用于从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

预训练模块 204，用于利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

第二获取模块 205，用于获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

微调模块 206，用于利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

纠错模块 207，用于将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

示例性的，所述计算机可读指令 303 可以被分割成一个或多个模块，所述一个或者多个模块被存储在所述存储器 301 中，并由所述处理器 302 执行，以完成本方法。例如，所述计算机可读指令 303 可以被分割成图 2 中的第一获取模块 201、转换模块 202、生成 203、预训练模块 204、第二获取模块 205、微调模块 206、纠错模块 207，各模块具体功能参见实施例二。

所述计算机装置 30 可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。本领域技术人员可以理解，所述示意图 3 仅仅是计算机装置 30 的示例，并不构成对计算机装置 30 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，例如所述计算机装置 30 还可以包括输入输出设备、网络接入设备、总线等。

所称处理器 302 可以是中央处理单元(Central Processing Unit，CPU)，还可以是其他通用处理器、数字信号处理器(Digital Signal Processor，DSP)、专用集成电路(Application Specific Integrated Circuit，ASIC)、现场可编程门阵列(Field-Programmable Gate Array，FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器 302 也可以是任何常规的处理器等，所述处理器 302 是所述计算机装置 30 的控制中心，利用各种接口和线路连接整个计算机装置 30 的各个部分。

所述存储器 301 可用于存储所述计算机可读指令 303，所述处理器 302 通过运行或执行存储在所述存储器 301 内的计算机可读指令或模块，以及调用存储在存储器 301 内的数据，实现所述计算机装置 30 的各种功能。所述存储器 301 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如声音播放功能、图像播放功能等）等；存储数据区可存储根据计算机装置 30 的使用所创建的数据。此外，存储器 301 可以包括非易失性存储器，例如硬盘、内存、插接式硬盘，智能存储卡（Smart Media Card， SMC），安全数字（Secure Digital， SD）卡，闪存卡（Flash Card）、至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。

所述计算机装置 30 集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，可以存储在一个非易失性可读存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，也可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一非易失性可读存储介质中，该计算机可读指令在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机可读指令可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机可读指令代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、只读存储器（ROM，Read-Only Memory）。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

最后应说明的是，以上实施例仅用以说明本申请的技术方案而非限制，尽管参照较佳实施例对本申请进行了详细说明，本领域的普通技术人员应当理解，可以对本申请的技术方案进行修改或等同替换，而不脱离本申请技术方案的精神和范围。

权 利 要 求 书

1. 一种错词纠正方法，其特征在于，所述方法包括：

获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

2. 如权利要求 1 所述的方法，其特征在于，所述将所述通用自然语言数据集包含的每个句子转换为拼音序列包括：

根据汉字的 ASCII 码将所述句子转换为拼音序列；或

根据汉字的 Unicode 值将所述句子转换为拼音序列。

3. 如权利要求 2 所述的方法，其特征在于，所述根据汉字的 Unicode 值将所述句子转换为拼音序列包括：

建立拼音-编号对照表，对所有拼音进行编号并将所有拼音对应的编号添加到所述拼音-编号对照表中；

建立 Unicode 值-拼音编号对照表，将汉字对应拼音的编号按照汉字的 Unicode 值添加到所述 Unicode 值-拼音编号对照表中；

逐一读取所述句子中的待转换汉字，确定所述待转换汉字的 Unicode 值，根据所述待转换汉字的 Unicode 值从所述 Unicode 值-拼音编号对照表中获取所述待转换汉字对应的拼音的编号，根据所述待转换汉字对应的拼音的编号从所述拼音-编号对照表获得所述待转换汉字对应的拼音，从而将所述句子中的每个汉字转换为拼音。

4. 如权利要求 1 所述的方法，其特征在于，所述从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对包括：

从所述通用自然语言数据集的拼音-句子对中随机选择所述多个拼音-句子对；和/或

按照预设比例从所述通用自然语言数据集的拼音-句子对中选择所述多个拼音-句子对。

5. 如权利要求 1 所述的方法，其特征在于，所述神经网络模型是 transformer 模型。

6. 如权利要求 1 所述的方法，其特征在于，所述对所述预训练后的神经网络模型进行微调包括：

固定所述神经网络模型的前面几层神经元的权值，微调所述神经网络模型的后面几层神经元的权值。

7. 如权利要求 1-6 中任一项所述的方法，其特征在于，所述方法还包括：

对输入的语音进行识别，得到所述待纠错句子。

8. 一种错词纠正装置，其特征在于，所述装置包括：

第一获取模块，用于获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

转换模块，用于将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

生成模块，用于从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

预训练模块，用于用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

第二获取模块，用于获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

微调模块，用于利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

纠错模块，用于将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

9. 一种计算机装置，其特征在于，所述计算机装置包括处理器和存储器，所述处理器用于执行所述存储器中存储的计算机可读指令以实现以下步骤：

获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

10. 如权利要求 9 所述的计算机装置，其特征在于，所述处理器执行所述存储器中存储的计算机可读指令以实现所述将所述通用自然语言数据集包含的每个句子转换为拼音序列时，包括：

根据汉字的 ASCII 码将所述句子转换为拼音序列；或

根据汉字的 Unicode 值将所述句子转换为拼音序列。

11. 如权利要求 10 所述的计算机装置，其特征在于，所述处理器执行所述存储器中存储的计算机可读指令以实现所述根据汉字的 Unicode 值将所述句子转换为拼音序列时，包括：

建立拼音-编号对照表，对所有拼音进行编号并将所有拼音对应的编号添加到所述拼音-编号对照表中；

建立 Unicode 值-拼音编号对照表，将汉字对应拼音的编号按照汉字的 Unicode 值添加到所述 Unicode 值-拼音编号对照表中；

逐一读取所述句子中的待转换汉字，确定所述待转换汉字的 Unicode 值，根据所述待转换汉字的 Unicode 值从所述 Unicode 值-拼音编号对照表中获取所述待转换汉字对应的拼音的编号，根据所述待转换汉字对应的拼音的编号从所述拼音-编号对照表获得所述待转换汉字对应的拼音，从而将所述句子中的每个汉字转换为拼音。

12. 如权利要求 9 所述的计算机装置，其特征在于，所述处理器执行所述存储器中存储的计算机可读指令以实现所述从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对时，包括：

从所述通用自然语言数据集的拼音-句子对中随机选择所述多个拼音-句子对；和/或

按照预设比例从所述通用自然语言数据集的拼音-句子对中选择所述多个拼音-句子对。

13. 如权利要求 9 所述的计算机装置，其特征在于，所述处理器执行所述存储器中存储的计算机可读指令以实现所述对所述预训练后的神经网络模型进行微调时，包括：

固定所述神经网络模型的前面几层神经元的权值，微调所述神经网络模型的后面几层神经元的权值。

14. 如权利要求 9-13 中任一项所述的计算机装置，其特征在于，所述处理器执行所述存储器中存储的计算机可读指令还用以实现以下步骤：

对输入的语音进行识别，得到所述待纠错句子。

15. 一种非易失性可读存储介质，所述非易失性可读存储介质上存储有计算机可读指令，其特征在于，所述计算机可读指令被处理器执行时实现以下步骤：

获取通用自然语言数据集，所述通用自然语言数据集包含多个句子；

将所述通用自然语言数据集包含的每个句子转换为拼音序列，得到所述通用自然语言数据集的拼音-句子对；

从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对，将选择的每个拼音-句子对的部分拼音替换为相近拼音，得到替换后的拼音-句子对，将所述通用自然语言数据集的未选择的拼音-句子对和所述替换后的拼音-句子对组成第一样本集；

利用所述第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型；

获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集；

利用所述第二样本集对所述预训练后的神经网络模型进行微调，得到微调后的神经网络模型；

将待纠错句子的拼音序列输入所述微调后的神经网络模型进行纠错，得到纠错后的句子。

16. 如权利要求 15 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行以实现所述将所述通用自然语言数据集包含的每个句子转换为拼音序列时，包括：

根据汉字的 ASCII 码将所述句子转换为拼音序列；或

根据汉字的 Unicode 值将所述句子转换为拼音序列。

17. 如权利要求 16 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行以实现所述根据汉字的 Unicode 值将所述句子转换为拼音序列时，包括：

建立拼音-编号对照表，对所有拼音进行编号并将所有拼音对应的编号添加到所述拼音-编号对照表中；

建立 Unicode 值-拼音编号对照表，将汉字对应拼音的编号按照汉字的 Unicode 值添加到所述 Unicode 值-拼音编号对照表中；

逐一读取所述句子中的待转换汉字，确定所述待转换汉字的 Unicode 值，根据所述待转换汉字的 Unicode 值从所述 Unicode 值-拼音编号对照表中获取所述待转换汉字对应的拼音的编号，根据所述待转换汉字对应的拼音的编号从所述拼音-编号对照表获得所述待转换汉字对应的拼音，从而将所述句子中的每个汉字转换为拼音。

18. 如权利要求 15 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行以实现所述从所述通用自然语言数据集的拼音-句子对中选择多个拼音-句子对时，包括：

从所述通用自然语言数据集的拼音-句子对中随机选择所述多个拼音-句子对；和/或

按照预设比例从所述通用自然语言数据集的拼音-句子对中选择所述多个拼音-句子对。

19. 如权利要求 15 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行以实现所述对所述预训练后的神经网络模型进行微调时，包括：

固定所述神经网络模型的前面几层神经元的权值，微调所述神经网络模型的后面几层神经元的权值。

20. 如权利要求 15-18 中任一项所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行还用以实现以下步骤：

对输入的语音进行识别，得到所述待纠错句子。

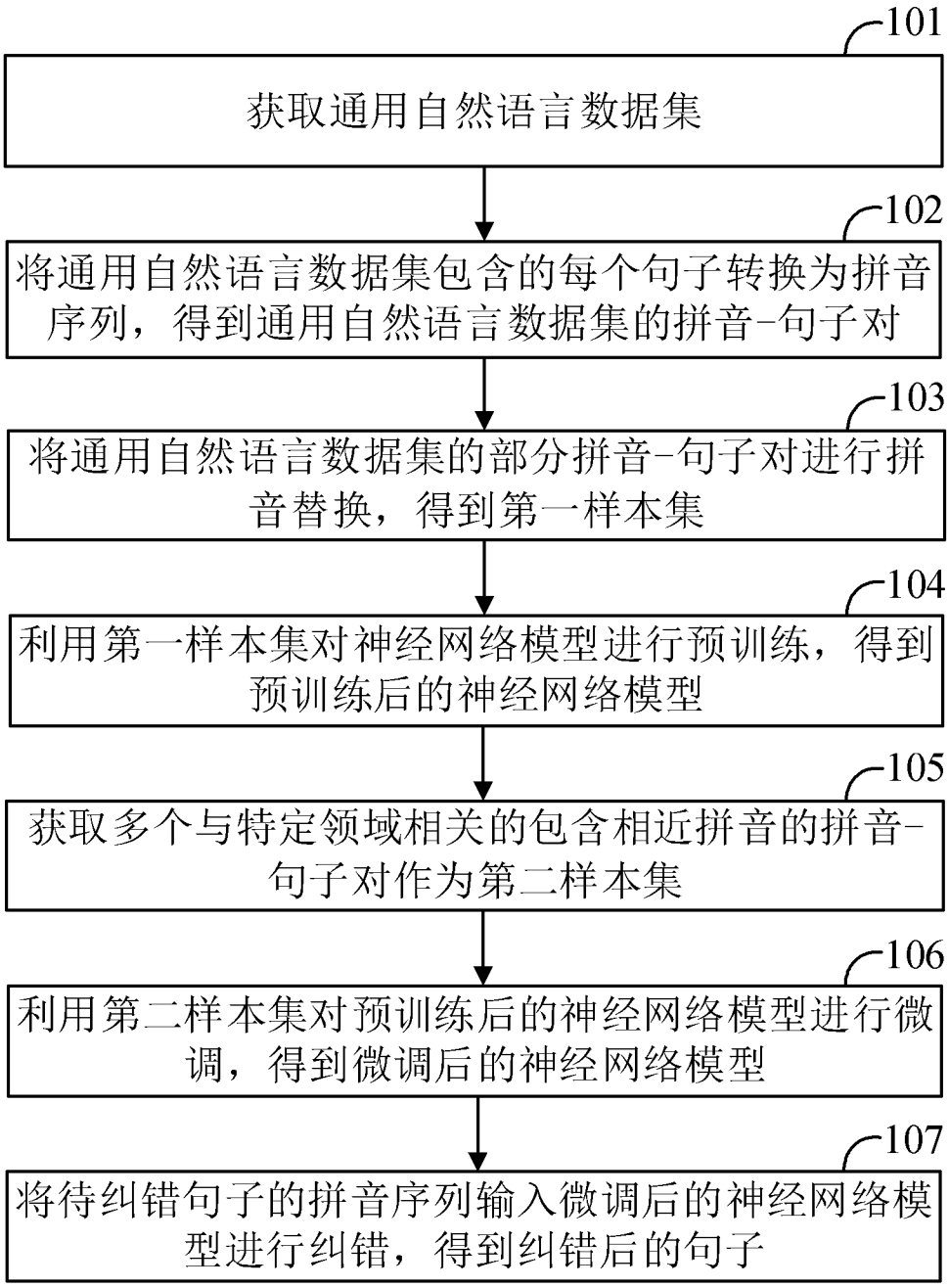
101
获取通用自然语言数据集
102
将通用自然语言数据集包含的每个句子转换为拼音序列，得到通用自然语言数据集的拼音-句子对
103
将通用自然语言数据集的部分拼音-句子对进行拼音替换，得到第一样本集
104
利用第一样本集对神经网络模型进行预训练，得到预训练后的神经网络模型
105
获取多个与特定领域相关的包含相近拼音的拼音-句子对作为第二样本集
106
利用第二样本集对预训练后的神经网络模型进行微调，得到微调后的神经网络模型
107
将待纠错句子的拼音序列输入微调后的神经网络模型进行纠错，得到纠错后的句子

图 1

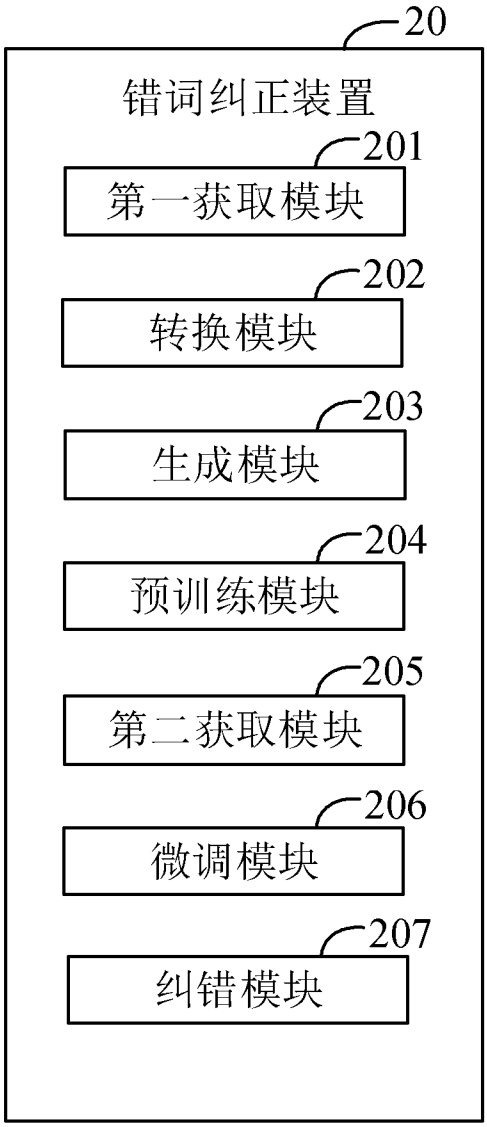
20
错词纠正装置
201
第一获取模块
202
转换模块
203
生成模块
204
预训练模块
205
第二获取模块
206
微调模块
207
纠错模块

图 2

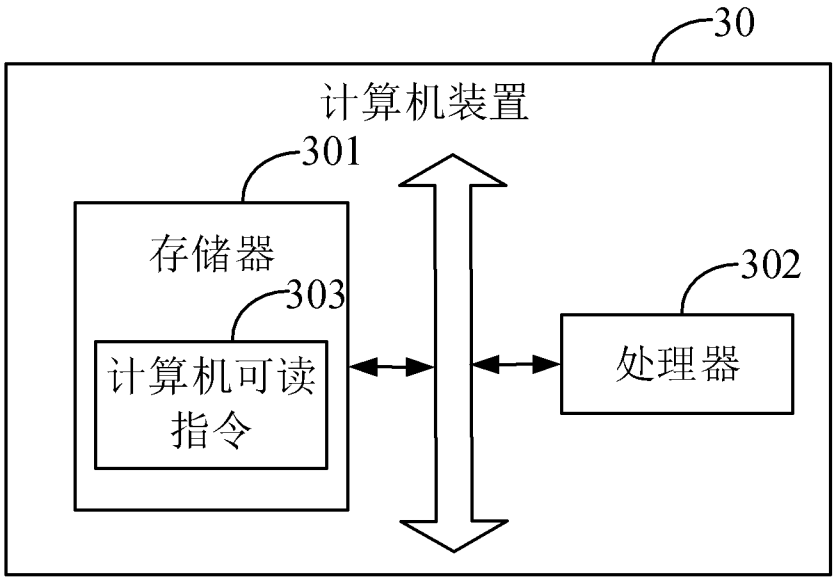

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/117237

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/33(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC, GOOGLE: 纠正, 纠错, 语音识别, 通用, 专用, 领域, 句子, 拼音, 相近, 相似, 近似, 文本, 文字, 训练, 神经网络, 替换, 代替, 注意力, correct+, speech recognition, common, general, specific, domain, field, sentence, pinyin, similar, text, train+, neural network, replace, transformer model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110110041 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 09 August 2019 (2019-08-09) claims 1-10	1-20
A	CN 108874174 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 23 November 2018 (2018-11-23) description, paragraphs [0083]-[0098], and figure 1c	1-20
A	CN 107357775 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 17 November 2017 (2017-11-17) entire document	1-20
A	CN 105869634 A (CHONGQING UNIVERSITY) 17 August 2016 (2016-08-17) entire document	1-20
A	CN 108021554 A (WUXI LITTLE SWAN CO., LTD.) 11 May 2018 (2018-05-11) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 January 2020	**07 February 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/117237

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 110110041 A	09 August 2019	None	
CN 108874174 A	23 November 2018	None	
CN 107357775 A	17 November 2017	US 2018349327 A1	06 December 2018
CN 105869634 A	17 August 2016	CN 105869634 B	19 November 2019
CN 108021554 A	11 May 2018	None	

国际检索报告

国际申请号
PCT/CN2019/117237

A. 主题的分类

G06F 16/33(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT, CNKI, WPI, EPODOC, GOOGLE: 纠正，纠错，语音识别，通用，专用，领域，句子，拼音，相近，相似，近似，文本，文字，训练，神经网络，替换，代替，注意力，correct+, speech recognition, common, general, specific, domain, field, sentence, pinyin, similar, text, train+, neural network, replace, transformer model

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 110110041 A (平安科技深圳有限公司) 2019年 8月 9日 (2019-08-09) 权利要求1-10	1-20
A	CN 108874174 A (腾讯科技深圳有限公司) 2018年 11月 23日 (2018-11-23) 说明书第[0083]-[0098]段，附图1c	1-20
A	CN 107357775 A (百度在线网络技术北京有限公司) 2017年 11月 17日 (2017-11-17) 全文	1-20
A	CN 105869634 A (重庆大学) 2016年 8月 17日 (2016-08-17) 全文	1-20
A	CN 108021554 A (无锡小天鹅股份有限公司) 2018年 5月 11日 (2018-05-11) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 1月 11日	2020年 2月 7日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 王晓燕 电话号码 86-(10)-53961384

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/117237

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 110110041 A	2019年 8月 9日	无	
CN 108874174 A	2018年 11月 23日	无	
CN 107357775 A	2017年 11月 17日	US 2018349327 A1	2018年 12月 6日
CN 105869634 A	2016年 8月 17日	CN 105869634 B	2019年 11月 19日
CN 108021554 A	2018年 5月 11日	无	