(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116362249 A
(43)申请公布日 2023.06.30

(21)申请号 202310257606.2

(22)申请日 2023.03.09

(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 王泽勋 冯明超 陈蒙 覃杰

(74)专利代理机构 北京律智知识产权代理有限公司 11438
专利代理师 王辉

(51)Int.Cl.
G06F 40/295(2020.01)
G10L 15/26(2006.01)
G06F 16/35(2019.01)
G06F 16/332(2019.01)
G06N 20/00(2019.01)

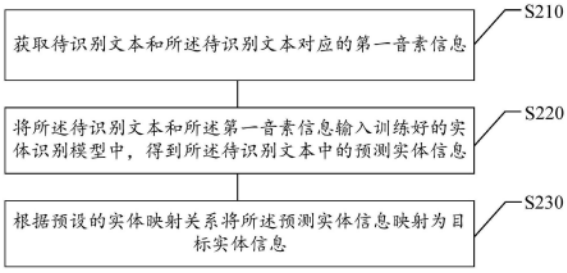
权利要求书2页 说明书11页 附图4页

(54)发明名称

实体识别方法、装置、存储介质及电子设备

(57)摘要

本公开提供一种实体识别方法、装置、存储介质及电子设备;涉及人工智能及自然语言处理技术领域。所述方法包括:获取待识别文本和所述待识别文本对应的第一音素信息;将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中,得到所述待识别文本中的预测实体信息;根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。本公开利用音素信息辅助文本实体识别,可以提高实体识别的准确率。



CN 116362249 A

1.一种实体识别方法,其特征在于,包括:

获取待识别文本和所述待识别文本对应的第一音素信息;

将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中,得到所述待识别文本中的预测实体信息;

根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。

2.根据权利要求1所述的实体识别方法,其特征在于,所述获取待识别文本和所述待识别文本对应的第一音素信息,包括:

获取待识别语音;

通过语音识别模型对所述待识别语音进行解码,得到所述待识别文本;

通过音素识别模型对所述待识别语音进行解码,得到所述第一音素信息。

3.根据权利要求1所述的实体识别方法,其特征在于,实体信息包括实体值和实体类型;

所述根据预设的实体映射关系将所述预测实体信息映射为目标实体信息,包括:

当所述预测实体信息中包括错误实体值时,确定所述错误实体值的实体类型;

根据所述实体映射关系确定与所述实体类型匹配的正确实体值,并将所述正确实体值作为所述目标实体信息。

4.根据权利要求3所述的实体识别方法,其特征在于,所述根据所述实体映射关系确定与所述实体类型匹配的正确实体值,包括:

根据所述实体映射关系确定与所述实体类型匹配的一个或多个候选实体值;

当匹配到唯一候选实体值时,将所述候选实体值作为所述正确实体值;

当匹配到多个候选实体值时,通过训练好的实体匹配模型计算各所述候选实体值与所述待识别文本之间的匹配得分,并根据所述匹配得分确定所述正确实体值。

5.根据权利要求1所述的实体识别方法,其特征在于,将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中之前,所述方法还包括:

获取训练文本和所述训练文本对应的第二音素信息;

根据所述训练文本和所述第二音素信息得到包含实体信息的四元组数据;

利用所述四元组数据对序列标注模型进行训练,得到所述实体识别模型。

6.根据权利要求5所述的实体识别方法,其特征在于,所述实体信息包括第一实体信息和第二实体信息;

所述根据所述训练文本和所述第二音素信息得到包含实体信息的四元组数据,包括:

根据所述第二音素信息对所述训练文本中的各个实体进行标注,得到所述第一实体信息;

当所述第一实体信息中包括错误实体值时,确定所述错误实体值的实体类型;

根据所述实体映射关系确定与所述实体类型匹配的正确实体值,并将所述正确实体值作为所述四元组数据中的所述第二实体信息。

7.根据权利要求5所述的实体识别方法,其特征在于,所述利用所述四元组数据对序列标注模型进行训练,得到所述实体识别模型,包括:

对所述四元组数据进行预处理,得到包含真实实体标签的训练数据集;

使用所述序列标注模型确定所述训练数据集中各个文本的预测实体标签;

利用损失函数对所述训练数据集中各个文本的预测实体标签和真实实体标签值进行计算，得到对应的损失值；

根据所述损失值对所述序列标注模型的模型参数进行迭代更新，当满足迭代终止条件时，得到所述实体识别模型。

8.一种实体识别装置，其特征在于，包括：

数据获取模块，用于获取待识别文本和所述待识别文本对应的第一音素信息；

实体识别模块，用于将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中，得到所述待识别文本中的预测实体信息；

实体映射模块，用于根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。

9.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1-7任一项所述的方法。

10.一种电子设备，其特征在于，包括：

处理器；以及

存储器，用于存储所述处理器的可执行指令；

其中，所述处理器配置为经由执行所述可执行指令来执行权利要求1-7任一项所述的方法。

实体识别方法、装置、存储介质及电子设备

技术领域

[0001] 本公开涉及人工智能及自然语言处理技术领域，具体而言，涉及一种实体识别方法、实体识别装置、计算机可读存储介质以及电子设备。

背景技术

[0002] 实体识别作为自然语言处理的一项基本任务，旨在从非结构化文本中识别出实体，其本质在于从一段文字描述中，识别出某一领域的专有名词或具有特殊意义的词或短语。

[0003] 目前，实体识别的过程通常是使用语音识别系统将音频转换为文本，再使用自然语言处理系统对文本中的实体进行识别。但是，由于背景音和环境音的干扰，增加了音频识别的难度，降低了语音识别系统对音频识别的准确率，进而降低了自然语言处理系统对文本中的实体识别的准确率。

[0004] 需要说明的是，在上述背景技术部分公开的信息仅用于加强对本公开的背景的理解，因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

[0005] 本公开的目的在于提供一种实体识别方法、实体识别装置、计算机可读存储介质以及电子设备，进而在一定程度上克服由于相关技术的限制和缺陷而导致实体识别的准确率较低的问题。

[0006] 根据本公开的第一方面，提供一种实体识别方法，包括：

[0007] 获取待识别文本和所述待识别文本对应的第一音素信息；

[0008] 将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中，得到所述待识别文本中的预测实体信息；

[0009] 根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。

[0010] 在本公开的一种示例性实施例中，所述获取待识别文本和所述待识别文本对应的第一音素信息，包括：

[0011] 获取待识别语音；

[0012] 通过语音识别模型对所述待识别语音进行解码，得到所述待识别文本；

[0013] 通过音素识别模型对所述待识别语音进行解码，得到所述第一音素信息。

[0014] 在本公开的一种示例性实施例中，实体信息包括实体值和实体类型；

[0015] 所述根据预设的实体映射关系将所述预测实体信息映射为目标实体信息，包括：

[0016] 当所述预测实体信息中包括错误实体值时，确定所述错误实体值的实体类型；

[0017] 根据所述实体映射关系确定与所述实体类型匹配的正确实体值，并将所述正确实体值作为所述目标实体信息。

[0018] 在本公开的一种示例性实施例中，所述根据所述实体映射关系确定与所述实体类型匹配的正确实体值，包括：

[0019] 根据所述实体映射关系确定与所述实体类型匹配的一个或多个候选实体值；

[0020] 当匹配到唯一候选实体值时，将所述候选实体值作为所述正确实体值；

[0021] 当匹配到多个候选实体值时，通过训练好的实体匹配模型计算各所述候选实体值与所述待识别文本之间的匹配得分，并根据所述匹配得分确定所述正确实体值。

[0022] 在本公开的一种示例性实施例中，将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中之前，所述方法还包括：

[0023] 获取训练文本和所述训练文本对应的第二音素信息；

[0024] 根据所述训练文本和所述第二音素信息得到包含实体信息的四元组数据；

[0025] 利用所述四元组数据对序列标注模型进行训练，得到所述实体识别模型。

[0026] 在本公开的一种示例性实施例中，所述实体信息包括第一实体信息和第二实体信息；

[0027] 所述根据所述训练文本和所述第二音素信息得到包含实体信息的四元组数据，包括：

[0028] 根据所述第二音素信息对所述训练文本中的各个实体进行标注，得到所述第一实体信息；

[0029] 当所述第一实体信息中包括错误实体值时，确定所述错误实体值的实体类型；

[0030] 根据所述实体映射关系确定与所述实体类型匹配的正确实体值，并将所述正确实体值作为所述四元组数据中的所述第二实体信息。

[0031] 在本公开的一种示例性实施例中，所述利用所述四元组数据对序列标注模型进行训练，得到所述实体识别模型，包括：

[0032] 对所述四元组数据进行预处理，得到包含真实实体标签的训练数据集；

[0033] 使用所述序列标注模型确定所述训练数据集中各个文本的预测实体标签；

[0034] 利用损失函数对所述训练数据集中各个文本的预测实体标签和真实实体标签值进行计算，得到对应的损失值；

[0035] 根据所述损失值对所述序列标注模型的模型参数进行迭代更新，当满足迭代终止条件时，得到所述实体识别模型。

[0036] 根据本公开的第二方面，提供一种实体识别装置，包括：

[0037] 数据获取模块，用于获取待识别文本和所述待识别文本对应的第一音素信息；

[0038] 实体识别模块，用于将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中，得到所述待识别文本中的预测实体信息；

[0039] 实体映射模块，用于根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。

[0040] 根据本公开的第三方面，提供一种计算机可读存储介质，其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上述任意一项所述的方法。

[0041] 根据本公开的第四方面，提供一种电子设备，包括：处理器；以及存储器，用于存储所述处理器的可执行指令；其中，所述处理器配置为经由执行所述可执行指令来执行上述任意一项所述的方法。

[0042] 本公开示例性实施例可以具有以下部分或全部有益效果：

[0043] 在本公开示例实施方式所提供的实体识别方法中，通过获取待识别文本和所述待

识别文本对应的第一音素信息;将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中,得到所述待识别文本中的预测实体信息;根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。一方面,利用音素信息辅助文本实体识别,可以提高实体识别的准确率;另一方面,利用实体映射关系调整识别出的实体信息,以将识别出的错误实体信息映射为正确实体信息,可以提高实体识别的鲁棒性。

[0044] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本公开。

附图说明

[0045] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本公开的实施例,并与说明书一起用于解释本公开的原理。显而易见地,下面描述中的附图仅仅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它的附图。

[0046] 图1示出了可以应用本公开实施例中的实体识别方法的系统架构图;

[0047] 图2示出了本公开实施例中的一种实体识别方法的流程图;

[0048] 图3示出了本公开实施例中的一种实体识别模型训练的流程图;

[0049] 图4示出了本公开实施例中的一种实体映射流程图;

[0050] 图5示出了本公开实施例中的另一种实体识别方法的流程图;

[0051] 图6示出了本公开的一个实施例的实体识别装置的框图;

[0052] 图7示出了适于用来实现本公开实施例的电子设备的结构示意图。

具体实施方式

[0053] 现在将参考附图更全面地描述示例实施方式。然而,示例实施方式能够以多种形式实施,且不应被理解为限于在此阐述的范例;相反,提供这些实施方式使得本公开将更加全面和完整,并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。在下面的描述中,提供许多具体细节从而给出对本公开的实施方式的充分理解。然而,本领域技术人员将意识到,可以实践本公开的技术方案而省略所述特定细节中的一个或更多,或者可以采用其它的方法、组元、装置、步骤等。在其它情况下,不详细示出或描述公知技术方案以避免喧宾夺主而使得本公开的各方面变得模糊。

[0054] 此外,附图仅为本公开的示意性图解,并非一定是按比例绘制。图中相同的附图标记表示相同或类似的部分,因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体,不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体,或在一个或多个硬件模块或集成电路中实现这些功能实体,或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0055] 图1示出了可以应用本公开实施例的一种实体识别方法的系统架构图。

[0056] 如图1所示,系统架构100可以包括终端设备101、102、103中的一个或多个,网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。终端设

备101、102、103可以是各种电子设备，包括但不限于台式计算机、便携式计算机、智能手机和平板电脑等等。应该理解，图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。比如服务器105可以是多个服务器组成的服务器集群等。

[0057] 本公开实施例所提供的实体识别方法可以由服务器105执行，相应地，实体识别装置可以设置于服务器105中，服务器可以将待识别文本中的目标实体信息发送至终端设备，并由终端设备向用户进行展示。但本领域技术人员容易理解的是，本公开实施例所提供的实体识别方法也可以由终端设备101、102、103执行，相应的，实体识别装置也可以设置于终端设备101、102、103中，例如，由终端设备执行后可以将待识别文本中的目标实体信息直接显示在终端设备的显示屏上，以向用户进行展示，本示例性实施例中对此不做特殊限定。

[0058] 以下对本公开实施例的技术方案进行详细阐述：

[0059] 目前，实体识别的过程通常是使用语音识别系统将音频转换为文本，再使用自然语言处理系统对文本中的实体进行识别。但是，由于背景音和环境音的干扰，增加了音频识别的难度，降低了语音识别系统对音频识别的准确率，进而降低了自然语言处理系统对文本中的实体识别的准确率。

[0060] 基于上述问题，本示例实施方式提供了一种实体识别方法，利用语音到文本的中间信息(音素信息)来辅助文本进行实体识别，该方法利用了更多的语音信息，可以提升实体识别的准确率。参考图2所示，该实体识别方法可以包括以下步骤S210至步骤S230：

[0061] 步骤S210.获取待识别文本和所述待识别文本对应的第一音素信息；

[0062] 步骤S220.将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中，得到所述待识别文本中的预测实体信息；

[0063] 步骤S230.利用所述目标版本的布隆过滤器和所述当前最高版本的布隆过滤器对目标数据进行过滤。

[0064] 在本公开示例实施方式所提供的实体识别方法中，通过获取待识别文本和所述待识别文本对应的第一音素信息；将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中，得到所述待识别文本中的预测实体信息；根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。一方面，利用音素信息辅助文本实体识别，可以提高实体识别的准确率；另一方面，利用实体映射关系调整识别出的实体信息，以将识别出的错误实体信息映射为正确实体信息，可以提高实体识别的鲁棒性。

[0065] 下面，对于本示例实施方式的上述步骤进行更加详细的说明。

[0066] 在步骤S210中，获取待识别文本和所述待识别文本对应的第一音素信息。

[0067] 其中，文本可以是英文单词、汉字、不作为数值使用的数字和其他可输入的文本符，本公开对此不做限定。音素信息是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，一个发音动作构成一个音素，音素包括元音与辅音两大类。本公开示例实施方式中，可以通过对音频数据进行识别得到待识别文本和与待识别文本对应的第一音素信息。其中，音频数据为数字化的声音数据，对声音数据进行数字化即可得到音频数据，如以一定频率对来自麦克风等设备的声音数据进行模数转换，得到音频数据。

[0068] 以音频数据为语音为例，可以获取待识别语音，如收集线上日志中存储的语音。通过语音识别模型对待识别语音进行解码，得到待识别文本，通过音素识别模型对待识别语

音进行解码，得到第一音素信息。举例而言，在用户说出“我要去做核酸”的过程中进行录音，得到音频内容为“我要去做核酸”的待识别语音。可以通过语音识别模型对该待识别语音进行解码，将该待识别语音转换成文本，得到待识别文本，即“我要去做核酸”。同时，可以通过音素识别模型对该待识别语音进行解码，将该待识别语音转换成音素，得到与待识别文本对应的第一音素信息，即“uo3 iao4 qv4 zuo4 he2 suan1”。基于此，可以得到一个包含语音、文本和音素信息的三元组数据。

[0069] 由语音识别模型和音素识别模型两个不同的模型分别给出文本和与文本对应的音素信息，使得文本和音素信息之间的互补性更强，利用互补性强的文本和音素信息进行实体识别时，可以提高实体识别的准确率。

[0070] 可以理解的是，利用语音识别模型对音频数据进行解码时，可以携带出音素信息，因此，也可以通过语音识别模型对待识别语音进行解码，得到大量的文本和与各个文本对应的音素信息。可以看出，使用该方法可以简单直接地生成文本和音素信息。

[0071] 在步骤S220中，将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中，得到所述待识别文本中的预测实体信息。

[0072] 本公开示例实施方式中，可以先对实体识别模型进行预训练，以将将待识别文本和第一音素信息输入训练好的实体识别模型中，得到待识别文本中的预测实体信息。参考图3所示，可以根据步骤S310至步骤S330对实体识别模型进行训练。

[0073] 在步骤S310中，获取训练文本和所述训练文本对应的第二音素信息。

[0074] 示例性的，可以获取若干数目的训练语音，如收集线上日志中存储的大量语音，并通过语音识别模型将各个训练语音转换为训练文本，通过音素识别模型将各个训练语音转换为与训练文本对应的第二音素信息。

[0075] 在步骤S320中，根据所述训练文本和所述第二音素信息得到包含实体信息的四元组数据。

[0076] 得到训练文本和第二音素信息后，可以根据训练文本和第二音素信息对训练文本中的实体进行人工标注，训练语音作为辅助判断，以标注出训练文本中各个实体的实体信息，实体信息可以包括实体类型和实体值。其中，实体信息可以为第一实体信息和第二实体信息，第一实体信息中的实体值为错误实体值，第二实体信息中的实体值为正确实体值。

[0077] 示例性的，参考图4所示，步骤S320又可以包括步骤S410至步骤S430。

[0078] 步骤S410.根据所述第二音素信息对所述训练文本中的各个实体进行标注，得到所述第一实体信息。

[0079] 举例而言，对于训练语音“我明天要去窦店做核酸”，通过语音识别模型识别到的训练文本为“我明天要去到店做核酸”，通过音素识别模型识别到的与该训练文本对应的第二音素信息为“uo3 ming2 tian1iao4 qv4 dao4 dian4 zuo4 he2 suan1”。其中，可以根据第二音素信息标注出实体值“明天”，实体类型为时间。由于文本错误，无法标注出正确的实体值“窦店”，但是可以根据训练语音辅助识别，标注出实体值“到店”的实体类型，实体类型为地址。基于此，可以得到一个包含语音、文本、音素信息和第一实体信息的四元组数据。

[0080] 该示例中，即使通过语音识别模型识别到错误文本时，通过人工标注仍然能够提取到该错误文本的正确实体类型，以便于实体识别模型能够学习到实体类型，从而提高实体识别模型的实体识别性能。

[0081] 步骤S420.当所述第一实体信息中包括错误实体值时,确定所述错误实体值的实体类型。

[0082] 可以理解的是,当四元组数据包括错误实体值时,对应的,使用包含错误实体值的四元组数据训练模型,并通过训练好的模型识别出的实体也会是错误实体。因此,当第一实体信息中包括错误实体值,可以确定错误实体值的实体类型,以根据实体类型将错误实体值映射为正确实体值。

[0083] 步骤S430.根据所述实体映射关系确定与所述实体类型匹配的正确实体值,并将所述正确实体值作为所述四元组数据中的所述第二实体信息。

[0084] 构建预设的实体映射关系时,可以获取大量语音识别文本,对各个语音识别文本中的所有实体进行准确标注,并提取标注文本中的错误实体值以及与错误实体值对应的正确实体值,由所有的错误实体值和对应的正确实体值构建实体映射关系。例如,可以构建一个实体映射表,在实体映射表中,同一个错误实体值可以对应多个正确实体值,同一个正确实体值也可以对应多个错误实体值。

[0085] 对应的,可以根据实体映射关系确定与实体类型匹配的一个或多个候选实体值,当匹配到唯一候选实体值时,可以将候选实体值作为正确实体值;而当匹配到多个候选实体值时,可以通过训练好的实体匹配模型计算各候选实体值与多个语音识别文本之间的匹配得分,并根据匹配得分确定正确实体值。

[0086] 举例而言,对于错误实体值“到店”,当识别到实体值的实体类型为地址时,在实体映射表中,“到店”对应的地址类型的正确实体值可能只有一个,如“窦店”,也可能对应多个地址类型的正确实体值,如“窦店”、“都店”等。当实体映射表中地址类型的正确实体值只有一个时,可以直接返回该正确实体值,并将错误实体值替换。当实体映射表中地址类型的正确实体值有多个时,将多个正确实体值作为候选实体值,并利用语音识别模型输出的N-Best数据,通过训练好的实体匹配模型将N-Best数据中的N个语音识别文本与各个候选实体值进行匹配,计算出对应的匹配度得分,返回匹配度得分最高的候选实体值。

[0087] 在步骤S330中,利用所述四元组数据对序列标注模型进行训练,得到所述实体识别模型。

[0088] 本公开示例实施方式中,序列标注模型可以是LSTM(Long Short-Term Memory,长短时记忆神经网络)、BiLSTM(Bi-directional Long Short-Term Memory,双向长短时记忆神经网络)、HMM(Hidden Markov Model,隐马尔可夫模型)、CRF(Conditional Random Field,条件随机场)等模型中的一种或多种,也可以是卷积神经网络模型或Transformer模型中的一种,还可以是由循环神经网络模型、卷积神经网络模型以及Transformer模型组成的架构模型,本公开对此不做具体限定。

[0089] 示例性的,利用四元组数据对序列标注模型进行训练时,可以先对四元组数据进行预处理,以得到包含真实实体标签的训练数据集。例如,可以为四元组数据的文本中包含的所有字符添加实体标签,如采用BIO标注方法,“B”表示字符所在位置处为实体开始,“I”表示字符所在位置处为实体继续,O表示字符所在位置处无实体。举例而言,对于文本“我明天要去到店做核酸”,该文本的真实实体标签为“O B time I O O B address I O O O”。

[0090] 由于序列标注模型是针对字符进行预测,而每个字符的文本和音素信息是相互补充的,因此,可以将字符和对应的音素信息进行拼接,如“我uo3在zai4到dao4店dian4”,并

利用拼接后的字符和音素信息进行模型训练。需要说明的是，中文每个字符一般对应两个音素，将字符和对应的音素信息进行拼接之前，可以使用对齐工具将文本中的每一个字符和对应的音素对齐。

[0091] 得到包含真实实体标签的训练数据集后，可以利用训练数据集对序列标注模型进行训练。示例性的，可以使用序列标注模型确定训练数据集中各个文本的预测实体标签，利用损失函数对训练数据集中各个文本的预测实体标签和真实实体标签值进行计算，得到对应的损失值，并根据损失值对序列标注模型的模型参数进行迭代更新，当满足迭代终止条件时，得到实体识别模型。

[0092] 参考图5所示，序列标注模型为BiLSTM+CRF模型，将大量训练语音501作为训练数据集，通过语音识别模型将训练语音转换为向量形式的训练文本502，记为W_1、W_2、W_3、…、W_n，通过音素识别模型将训练语音转换为向量形式的训练音素503，记为P_1、P_2、P_3、…、P_n，采用BIO标注方法为各个训练文本添加真实实体标签。将各个训练文本502和对应的训练音素503对齐并进行拼接，得到多个拼接向量504，并利用多个拼接向量504对BiLSTM+CRF模型505进行训练，输出预测实体标签506，如“O B time I O O B address I O O O”。

[0093] 示例性的，可以使用BiLSTM+CRF模型得到训练数据集中各个训练文本的预测实体标签，在模型的训练过程中，需要使得预测实体标签无限接近于真实实体标签，如可以选取交叉熵损失函数，利用该损失函数对训练数据集中各个文本的预测实体标签和真实实体标签值进行计算，得到对应的损失值，根据损失值对序列标注模型的模型参数进行迭代更新。具体地，可以利用随机梯度下降算法对模型参数进行更新，根据反向传播原理，不断计算交叉熵损失函数，并根据计算得到的损失值同时更新序列标注模型的模型参数。当交叉熵损失函数收敛至最小值时，完成对所有模型参数的训练。也可以反向迭代式更新模型参数，当满足预设的迭代次数时，完成对所有模型参数的训练，迭代完成后，可以得到优化后的模型参数。其他示例中，也可以使用交替最小二乘法、Adam优化算法等最小化目标函数，从后向前依次更新模型参数，对参数进行优化。

[0094] 对序列标注模型进行训练得到实体识别模型后，可以将待识别文本和第一音素信息输入训练好的实体识别模型中，以输出待识别文本中的预测实体信息。例如，将待识别文本“我明天要去到店做核酸”和对应的第一音素信息“uo3 ming2 tian1 iao4 qv4 dao4 dian4 zuo4 he2 suan1”输入训练好的实体识别模型中后，得到预测实体信息：实体值为“到店”，实体类型为地址。

[0095] 在步骤S230中，根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。

[0096] 当预测实体信息中包括错误实体值时，可以确定错误实体值的实体类型，并根据实体映射关系确定与实体类型匹配的正确实体值，并将正确实体值作为目标实体信息。示例性的，可以根据实体映射关系确定与实体类型匹配的一个或多个候选实体值，当匹配到唯一候选实体值时，将候选实体值作为正确实体值；而当匹配到多个候选实体值时，可以通过训练好的实体匹配模型计算各候选实体值与待识别文本之间的匹配得分，并根据匹配得分确定正确实体值。步骤S230中根据预设的实体映射关系将预测实体信息映射为目标实体信息的过程与步骤S420和步骤S430中将错失实体值映射为正确实体值的过程类似，此处将不再赘述。

[0097] 在本公开示例实施方式所提供的实体识别方法中，通过获取待识别文本和所述待识别文本对应的第一音素信息；将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中，得到所述待识别文本中的预测实体信息；根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。一方面，利用音素信息辅助文本实体识别，可以提高实体识别的准确率；另一方面，利用实体映射关系调整识别出的实体信息，以将识别出的错误实体信息映射为正确实体信息，可以提高实体识别的鲁棒性。

[0098] 应当注意，尽管在附图中以特定顺序描述了本公开中方法的各个步骤，但是，这并非要求或者暗示必须按照该特定顺序来执行这些步骤，或是必须执行全部所示的步骤才能实现期望的结果。附加的或备选的，可以省略某些步骤，将多个步骤合并为一个步骤执行，以及/或者将一个步骤分解为多个步骤执行等。

[0099] 进一步的，本示例实施方式中，还提供了一种实体识别装置，该装置可以应用于一服务器或终端设备。参考图6所示，该实体识别装置600可以包括数据获取模块610、实体识别模块620和实体映射模块630，其中：

[0100] 数据获取模块610，用于获取待识别文本和所述待识别文本对应的第一音素信息；

[0101] 实体识别模块620，用于将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中，得到所述待识别文本中的预测实体信息；

[0102] 实体映射模块630，用于根据预设的实体映射关系将所述预测实体信息映射为目标实体信息。

[0103] 在一种可选的实施方式中，数据获取模块610包括：

[0104] 语音获取模块，用于获取待识别语音；

[0105] 文本获取模块，用于通过语音识别模型对所述待识别语音进行解码，得到所述待识别文本；

[0106] 音素获取模块，用于通过音素识别模型对所述待识别语音进行解码，得到所述第一音素信息。

[0107] 在一种可选的实施方式中，实体信息包括实体值和实体类型；实体映射模块630包括：

[0108] 实体类型确定模块，用于当所述预测实体信息中包括错误实体值时，确定所述错误实体值的实体类型；

[0109] 目标实体确定模块，用于根据所述实体映射关系确定与所述实体类型匹配的正确实体值，并将所述正确实体值作为所述目标实体信息。

[0110] 在一种可选的实施方式中，目标实体确定模块包括：

[0111] 候选实体值确定模块，用于根据所述实体映射关系确定与所述实体类型匹配的一个或多个候选实体值；

[0112] 第一实体值匹配模块，用于当匹配到唯一候选实体值时，将所述候选实体值作为所述正确实体值；

[0113] 第二实体值匹配模块，用于当匹配到多个候选实体值时，通过训练好的实体匹配模型计算各所述候选实体值与所述待识别文本之间的匹配得分，并根据所述匹配得分确定所述正确实体值。

[0114] 在一种可选的实施方式中，实体识别装置600还包括：

[0115] 训练文本获取模块，用于获取训练文本和所述训练文本对应的第二音素信息；

[0116] 四元组确定模块，用于根据所述训练文本和所述第二音素信息得到包含实体信息的四元组数据；

[0117] 模型训练模块，用于利用所述四元组数据对序列标注模型进行训练，得到所述实体识别模型。

[0118] 在一种可选的实施方式中，所述实体信息包括第一实体信息和第二实体信息；四元组确定模块包括：

[0119] 实体标注模块，用于根据所述第二音素信息对所述训练文本中的各个实体进行标注，得到所述第一实体信息；

[0120] 实体类型确定模块，用于当所述第一实体信息中包括错误实体值时，确定所述错误实体值的实体类型；

[0121] 实体值确定模块，用于根据所述实体映射关系确定与所述实体类型匹配的正确实体值，并将所述正确实体值作为所述四元组数据中的所述第二实体信息。

[0122] 在一种可选的实施方式中，模型训练模块被配置为用于对所述四元组数据进行预处理，得到包含真实实体标签的训练数据集；使用所述序列标注模型确定所述训练数据集中各个文本的预测实体标签；利用损失函数对所述训练数据集中各个文本的预测实体标签和真实实体标签值进行计算，得到对应的损失值；根据所述损失值对所述序列标注模型的模型参数进行迭代更新，当满足迭代终止条件时，得到所述实体识别模型。

[0123] 上述实体识别装置中各模块的具体细节已经在对应的实体识别方法中进行了详细的描述，因此此处不再赘述。

[0124] 上述装置中各模块可以是通用处理器，包括：中央处理器、网络处理器等；还可以是数字信号处理器、专用集成电路、现场可编程门阵列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。各模块也可以由软件、固件等形式来实现。上述装置中的各处理器可以是独立的处理器，也可以集成在一起。

[0125] 本公开的示例性实施方式还提供了一种计算机可读存储介质，其上存储有能够实现本说明书上述方法的程序产品。在一些可能的实施方式中，本公开的各个方面还可以实现为一种程序产品的形式，其包括程序代码，当程序产品在电子设备上运行时，程序代码用于使电子设备执行本说明书上述“示例性方法”部分中描述的根据本公开各种示例性实施方式的步骤。该程序产品可以采用便携式紧凑盘只读存储器(CD-ROM)并包括程序代码，并可以在电子设备，例如个人电脑上运行。然而，本公开的程序产品不限于此，在本文件中，可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0126] 程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0127] 计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其

中承载了可读程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0128] 可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于无线、有线、光缆、RF等等，或者上述的任意合适的组合。

[0129] 可以以一种或多种程序设计语言的任意组合来编写用于执行本公开操作的程序代码，程序设计语言包括面向对象的程序设计语言—诸如Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中，远程计算设备可以通过任意种类的网络，包括局域网(LAN)或广域网(WAN)，连接到用户计算设备，或者，可以连接到外部计算设备(例如利用因特网服务提供商来通过因特网连接)。

[0130] 本公开的示例性实施方式还提供了一种能够实现上述方法的电子设备。下面参照图7来描述根据本公开的这种示例性实施方式的电子设备700。图7显示的电子设备700仅仅是一个示例，不应对本公开实施方式的功能和使用范围带来任何限制。

[0131] 如图7所示，电子设备700可以以通用计算设备的形式表现。电子设备700的组件可以包括但不限于：至少一个处理单元710、至少一个存储单元720、连接不同系统组件(包括存储单元720和处理单元710)的总线730和显示单元740。

[0132] 存储单元720存储有程序代码，程序代码可以被处理单元710执行，使得处理单元710执行本说明书上述“示例性方法”部分中描述的根据本公开各种示例性实施方式的步骤。例如，处理单元710可以执行图2至图5中任意一个或多个方法步骤。

[0133] 存储单元720可以包括易失性存储单元形式的可读介质，例如随机存取存储单元(RAM)721和/或高速缓存存储单元722，还可以进一步包括只读存储单元(ROM)723。

[0134] 存储单元720还可以包括具有一组(至少一个)程序模块725的程序/实用工具724，这样的程序模块725包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0135] 总线730可以为表示几类总线结构中的一种或多种，包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

[0136] 电子设备700也可以与一个或多个外部设备800(例如键盘、指向设备、蓝牙设备等)通信，还可与一个或者多个使得用户能与该电子设备700交互的设备通信，和/或与使得该电子设备700能与一个或多个其它计算设备进行通信的任何设备(例如路由器、调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口750进行。并且，电子设备700还可以通过网络适配器770与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器770通过总线730与电子设备700的其它模块通信。应当明白，尽管图中未示出，可以结合电子设备700使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及

数据备份存储系统等。

[0137] 在一些实施例中，可以由电子设备的处理单元710执行本公开中所述的实体识别方法。在一些实施例中，可以通过输入接口750输入待识别语音，以将待识别语音转换为待识别文本和待识别文本对应的第一音素信息。在一些实施例中，可以通过输出接口750映射得到的目标实体信息输出至外部设备800以供用户查看。

[0138] 通过以上的实施方式的描述，本领域的技术人员易于理解，这里描述的示例实施方式可以通过软件实现，也可以通过软件结合必要的硬件的方式来实现。因此，根据本公开实施方式的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM，U盘，移动硬盘等)中或网络上，包括若干指令以使得一台计算设备(可以是个人计算机、服务器、终端装置、或者网络设备等)执行根据本公开示例性实施方式的方法。

[0139] 此外，上述附图仅是根据本公开示例性实施方式的方法所包括的处理的示意性说明，而不是限制目的。易于理解，上述附图所示的处理并不表明或限制这些处理的时间顺序。另外，也易于理解，这些处理可以是例如在多个模块中同步或异步执行的。

[0140] 应当注意，尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单元，但是这种划分并非强制性的。实际上，根据本公开的实施方式，上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之，上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

[0141] 应当理解的是，本公开并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本公开的范围仅由所附的权利要求来限制。

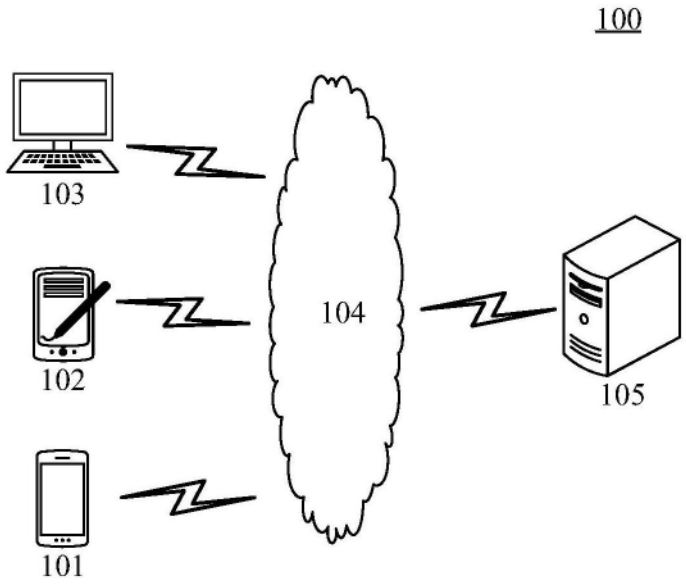

100
103
102
101
104
105

图1

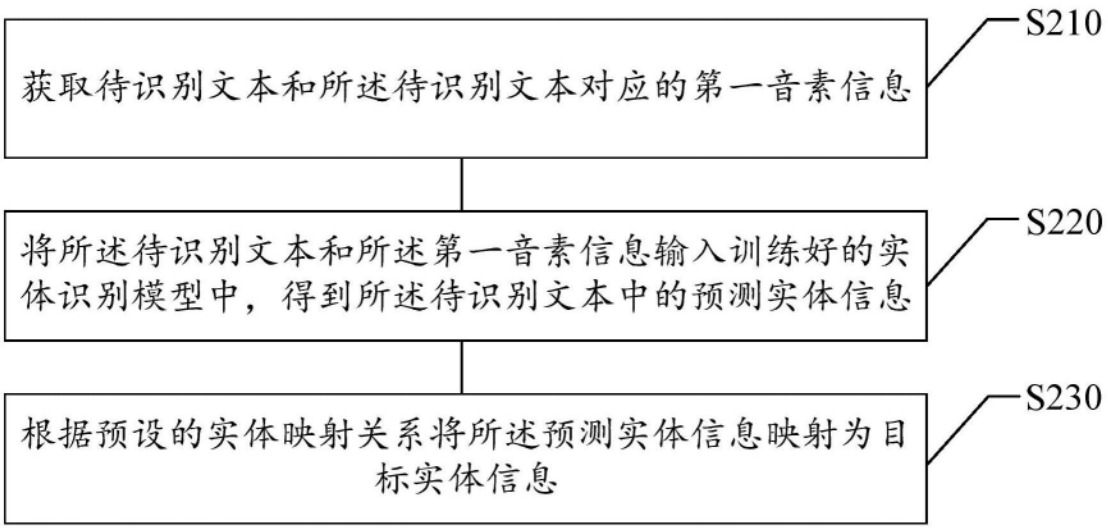

获取待识别文本和所述待识别文本对应的第一音素信息
S210
将所述待识别文本和所述第一音素信息输入训练好的实体识别模型中，得到所述待识别文本中的预测实体信息
S220
根据预设的实体映射关系将所述预测实体信息映射为目标实体信息
S230

图2

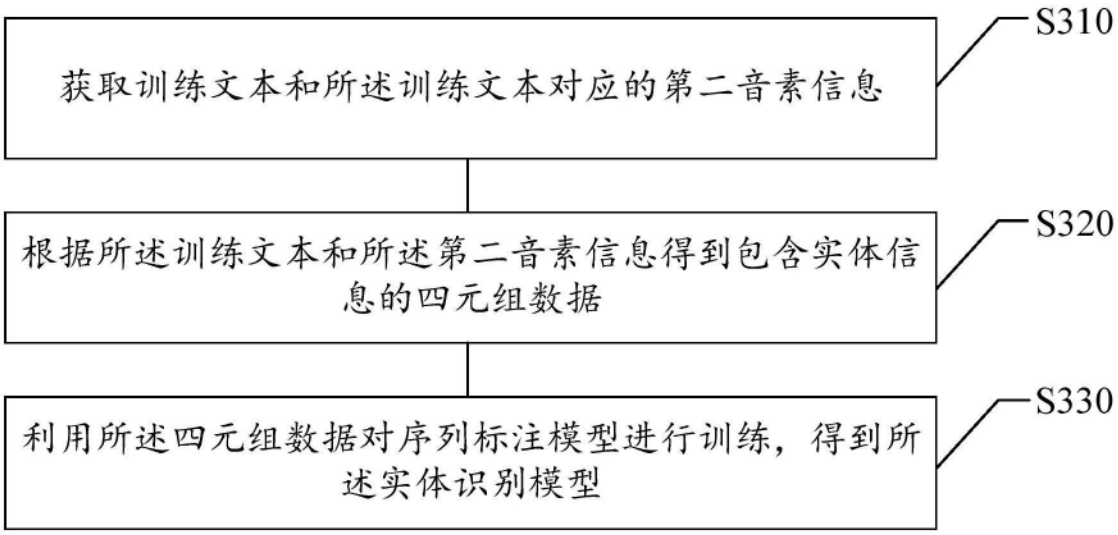

获取训练文本和所述训练文本对应的第二音素信息
S310
根据所述训练文本和所述第二音素信息得到包含实体信息的四元组数据
S320
利用所述四元组数据对序列标注模型进行训练，得到所述实体识别模型
S330

图3

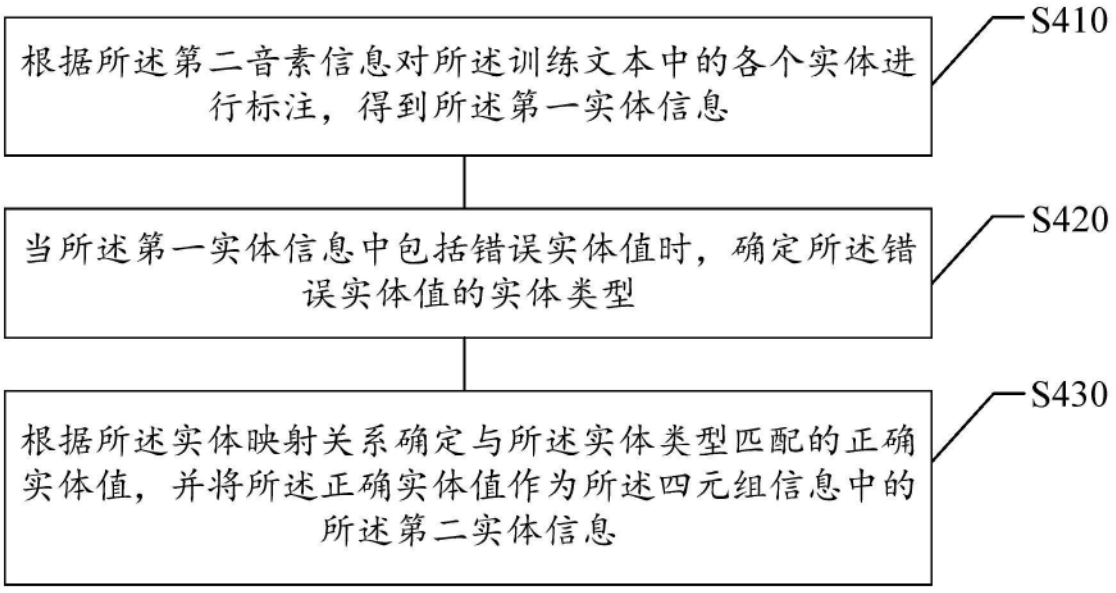

根据所述第二音素信息对所述训练文本中的各个实体进行标注，得到所述第一实体信息
S410
当所述第一实体信息中包括错误实体值时，确定所述错误实体值的实体类型
S420
根据所述实体映射关系确定与所述实体类型匹配的正确实体值，并将所述正确实体值作为所述四元组信息中的所述第二实体信息
S430

图4

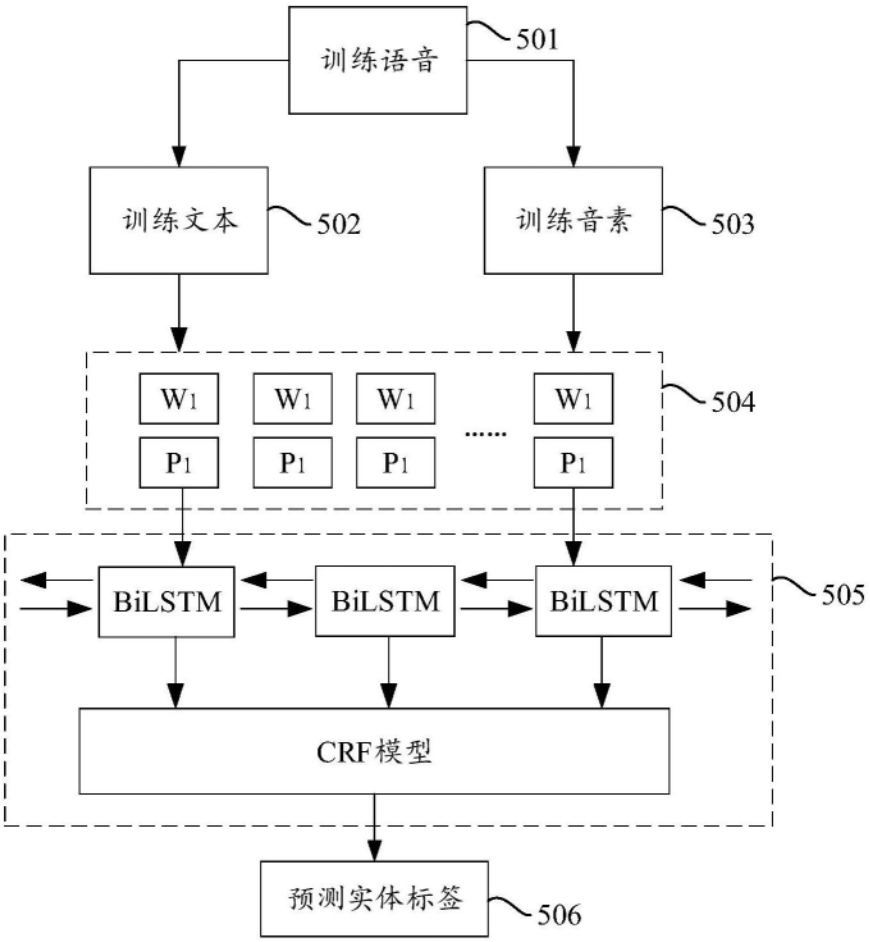

训练语音
501
训练文本
502
训练音素
503
W1
P1
504
BiLSTM
505
CRF模型
预测实体标签
506

图5

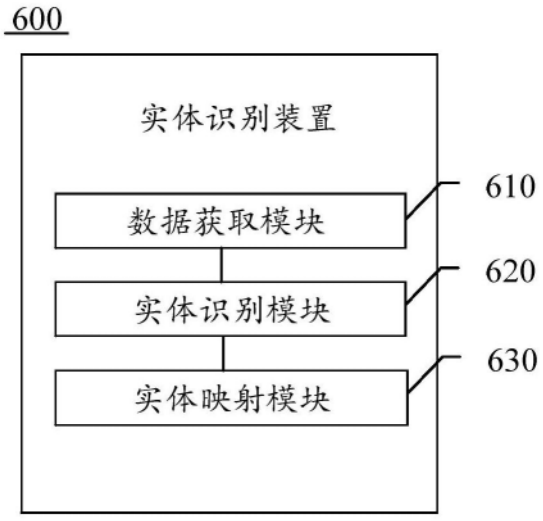

600
实体识别装置
数据获取模块
610
实体识别模块
620
实体映射模块
630

图6

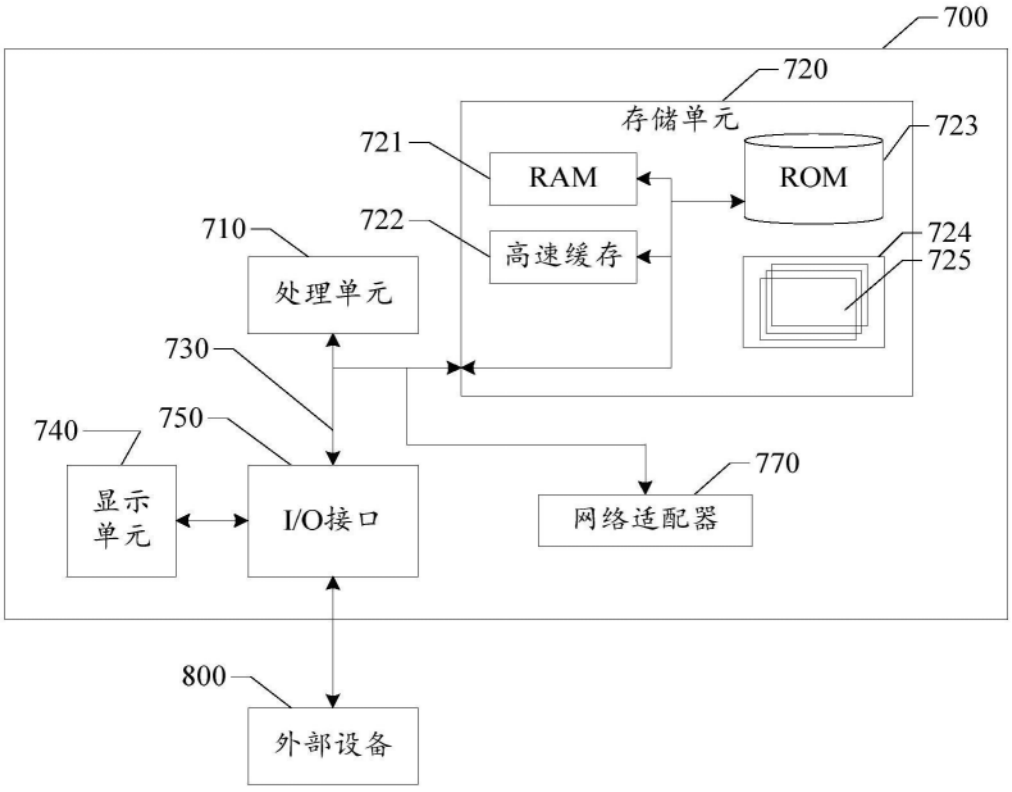

700
720
存储单元
721
RAM
ROM
723
722
高速缓存
710
处理单元
724
725
730
740
750
显示
单元
I/O接口
770
网络适配器
800
外部设备

图7