(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115309852 A
(43)申请公布日 2022.11.08

(21)申请号 202110494032.1

(22)申请日 2021.05.07

(71)申请人 京东科技控股股份有限公司
地址 100176 北京市大兴区经济技术开发区科创十一街18号C座2层221室

(72)发明人 赵宇明 乐雨泉 陈蒙

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
专利代理师 王达佐 马晓亚

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/332(2019.01)
G06F 16/35(2019.01)
G06F 40/216(2020.01)

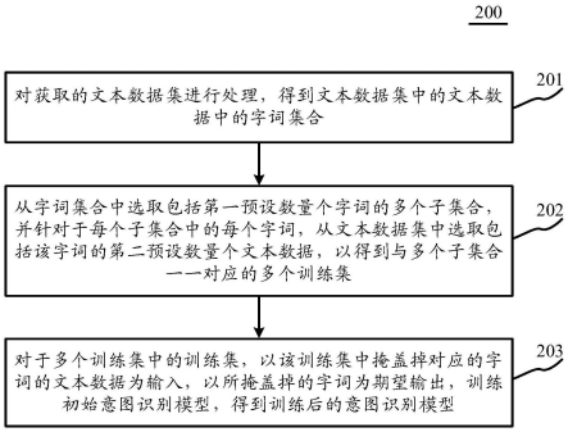
权利要求书3页 说明书13页 附图7页

(54)发明名称

意图识别模型的训练方法、装置及意图识别方法、装置

(57)摘要

本申请公开了一种意图识别模型的训练方法、装置及意图识别方法、装置。方法的一具体实施方式包括：对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合；从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集；对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。本申请提供了一种基于自监督的意图识别模型的训练方法，提高了训练任务的多样性和模型训练的灵活性。



CN 115309852 A

1.一种意图识别模型的训练方法,包括:

对获取的文本数据集进行处理,得到所述文本数据集中的文本数据中的字词集合;

从所述字词集合中选取包括第一预设数量个字词的多个子集合,并针对于每个子集合中的每个字词,从所述文本数据集中选取包括该字词的第二预设数量个文本数据,以得到与所述多个子集合一一对应的多个训练集;

对于所述多个训练集中的训练集,以该训练集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,训练初始意图识别模型,得到训练后的意图识别模型。

2.根据权利要求1所述的方法,其中,所述对于所述多个训练集中的训练集,以该训练集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,训练初始意图识别模型,得到训练后的意图识别模型,包括:

在初始语言模型的尾部添加分类层,得到所述初始意图识别模型;

对于所述多个训练集中的训练集,执行如下操作,直至得到所述意图识别模型:以该训练集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,优化表征所述初始意图识别模型的参数的第一参数集合;基于所述第一参数集合,更新第二参数集合,其中,所述第二参数集合表征根据所述初始语言模型的输出,生成所述分类层的函数中的参数。

3.根据权利要求2所述的方法,其中,

所述第一参数集合包括:

所述初始语言模型自注意力层之间的所有前馈神经网络层的参数;

所述初始意图识别模型中多层感知机的参数;

所述分类层中表征权重和偏置的参数;

所述第二参数集合包括:

所述语言模型中除前馈神经网络层外的网络结构的参数;

根据所述初始语言模型的输出,生成所述分类层的函数中的参数;

所述初始意图识别模型针对于训练集的训练过程的学习率。

4.根据权利要求1所述的方法,其中,还包括:

根据所接收的选取指令,确定从所述文本数据集选取出的监督集中的文本数据所要掩盖掉的字词;以及

所述对于所述多个训练集中的训练集,以该训练集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,训练初始意图识别模型,得到训练后的意图识别模型,包括:

在所述初始意图识别模型完成上一组合数据集的训练后,基于预设采样率,确定将对所述初始意图识别模型训练的、包括部分监督集和部分训练集的组合数据集,

以所述组合数据集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,训练所述初始意图识别模型,得到训练后的意图识别模型。

5.根据权利要求1所述的方法,其中,还包括:

根据待识别文本对应的具体场景,对所述意图识别模型进行微调,以通过微调后的意图识别模型,处理所述待识别文本,得到所述待识别文本所表达的意图。

6.根据权利要求1所述的方法,其中,所述对获取的文本数据集进行处理,得到所述文

本数据集中的文本数据中的字词集合，包括：

过滤掉所述文本数据集中的文本数据中词频低于预设词频阈值的字词，得到所述文本数据集中的文本数据中的字词集合。

7.一种针对自然语言的意图识别方法，包括：

获取表征自然语言的待识别文本；

通过意图识别模型，处理所述待识别文本，得到所述待识别文本所表达的意图，其中，所述意图识别模型根据权利要求1-6任一所述的训练方法训练得到。

8.一种意图识别模型的训练装置，包括：

处理单元，被配置成对获取的文本数据集进行处理，得到所述文本数据集中的文本数据中的字词集合；

选取单元，被配置成从所述字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从所述文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与所述多个子集合一一对应的多个训练集；

训练单元，被配置成对于所述多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。

9.根据权利要求8所述的装置，其中，所述训练单元，进一步被配置成：

在初始语言模型的尾部添加分类层，得到所述初始意图识别模型；对于所述多个训练集中的训练集，执行如下操作，直至得到所述意图识别模型：以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，优化表征所述初始意图识别模型的参数的第一参数集合；基于所述第一参数集合，更新第二参数集合，其中，所述第二参数集合表征根据所述初始语言模型的输出，生成所述分类层的函数中的参数。

10.根据权利要求9所述的装置，其中，

所述第一参数集合包括：

所述初始语言模型自注意力层之间的所有前馈神经网络层的参数；

所述初始意图识别模型中多层感知机的参数；

所述分类层中表征权重和偏置的参数；

所述第二参数集合包括：

所述语言模型中除前馈神经网络层外的网络结构的参数；

根据所述初始语言模型的输出，生成所述分类层的函数中的参数；

所述初始意图识别模型针对于训练集的训练过程的学习率。

11.根据权利要求8所述的装置，其中，还包括：

确定单元，被配置成根据所接收的选取指令，确定从所述文本数据集选取出的监督集中的文本数据所要掩盖掉的字词；以及

所述训练单元，进一步被配置成：

在所述初始意图识别模型完成上一组合数据集的训练后，基于预设采样率，确定将对所述初始意图识别模型训练的、包括部分监督集和部分训练集的组合数据集，以所述组合数据集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练所述初始意图识别模型，得到训练后的意图识别模型。

12.根据权利要求8所述的装置,其中,还包括:

微调单元,被配置成根据待识别文本对应的具体场景,对所述意图识别模型进行微调,以通过微调后的意图识别模型,处理所述待识别文本,得到所述待识别文本所表达的意图。

13.根据权利要求8所述的装置,其中,所述处理单元,进一步被配置成:

过滤掉所述文本数据集中的文本数据中词频低于预设词频阈值的字词,得到所述文本数据集中的文本数据中的字词集合。

14.一种针对自然语言的意图识别装置,包括:

获取单元,被配置成获取表征自然语言的待识别文本;

识别单元,被配置成通过意图识别模型,处理所述待识别文本,得到所述待识别文本所表达的意图,其中,所述意图识别模型根据权利要求1-6任一所述的训练方法训练得到。

15.一种计算机可读介质,其上存储有计算机程序,其中,所述程序被处理器执行时实现如权利要求1-7中任一所述的方法。

16.一种电子设备,包括:

一个或多个处理器;

存储装置,其上存储有一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-7中任一所述的方法。

意图识别模型的训练方法、装置及意图识别方法、装置

技术领域

[0001] 本申请实施例涉及计算机技术领域，具体涉及一种意图识别模型的训练方法、装置及意图识别方法、装置。

背景技术

[0002] 意图识别任务的本质为分类任务，其输入是对话上下文，包含当前问题，输出是当前问题的意图。意图识别任务可以通过分类模型完成，目前表现最好的分类模型主要是基于深度神经网络的分类模型，大致过程是用深度神经网络取提取特征，使用分类器对特征进行分类。深度神经网络通常需要大量的带标签数据去训练。在实际应用场景中，获取大量的带标签数据的成本较高，可以通过小样本学习方法使用少量样本进行模型训练。典型的，有监督元学习方式为，从一个带有大量类别数的固定任务的数据集中，采样一个类别子集的数据集，产生不同的元学习任务。

[0003] 现有的有监督元学习方法，仍需要一定数量的标注数据集，以从中采样数据，构建训练任务，并且存在训练任务多样性受限、训练缺乏灵活性的问题。

发明内容

[0004] 本申请实施例提出了一种意图识别模型的训练方法、装置及意图识别方法、装置。

[0005] 第一方面，本申请实施例提供了一种意图识别模型的训练方法，包括：对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合；从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集；对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。

[0006] 在一些实施例中，上述对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型，包括：在初始语言模型的尾部添加分类层，得到初始意图识别模型；对于多个训练集中的训练集，执行如下操作，直至得到意图识别模型：以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，优化表征初始意图识别模型的参数的第一参数集合；基于第一参数集合，更新第二参数集合，其中，第二参数集合表征根据初始语言模型的输出，生成分类层的函数中的参数。

[0007] 在一些实施例中，第一参数集合包括：初始语言模型自注意力层之间的所有前馈神经网络层的参数；初始意图识别模型中多层感知机的参数；分类层中表征权重和偏置的参数；第二参数集合包括：语言模型中除前馈神经网络层外的网络结构的参数；根据初始语言模型的输出，生成分类层的函数中的参数；初始意图识别模型针对于训练集的训练过程的学习率。

[0008] 在一些实施例中，上述方法还包括：根据所接收的选取指令，确定从文本数据集选

取出的监督集中的文本数据所要掩盖掉的字词;以及上述对于多个训练集中的训练集,以该训练集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,训练初始意图识别模型,得到训练后的意图识别模型,包括:在初始意图识别模型完成上一组合数据集的训练后,基于预设采样率,确定将对初始意图识别模型训练的、包括部分监督集和部分训练集的组合数据集,以组合数据集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,训练初始意图识别模型,得到训练后的意图识别模型。

[0009] 在一些实施例中,上述方法还包括:根据待识别文本对应的具体场景,对意图识别模型进行微调,以通过微调后的意图识别模型,处理待识别文本,得到待识别文本所表达的意图。

[0010] 在一些实施例中,上述对获取的文本数据集进行处理,得到文本数据集中的文本数据中的字词集合,包括:过滤掉文本数据集中的文本数据中词频低于预设词频阈值的字词,得到文本数据集中的文本数据中的字词集合。

[0011] 第二方面,本申请实施例提供了一种针对自然语言的意图识别方法,包括:获取表征自然语言的待识别文本;通过意图识别模型,处理待识别文本,得到待识别文本所表达的意图,其中,意图识别模型根据第一方面任一实现方式描述的训练方法训练得到。

[0012] 第三方面,本申请实施例提供了一种意图识别模型的训练装置,包括:处理单元,被配置成对获取的文本数据集进行处理,得到文本数据集中的文本数据中的字词集合;选取单元,被配置成从字词集合中选取包括第一预设数量个字词的多个子集合,并针对于每个子集合中的每个字词,从文本数据集中选取包括该字词的第二预设数量个文本数据,以得到与多个子集合一一对应的多个训练集;训练单元,被配置成对于多个训练集中的训练集,以该训练集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,训练初始意图识别模型,得到训练后的意图识别模型。

[0013] 在一些实施例中,训练单元,进一步被配置成:在初始语言模型的尾部添加分类层,得到初始意图识别模型;对于多个训练集中的训练集,执行如下操作,直至得到意图识别模型:以该训练集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,优化表征初始意图识别模型的参数的第一参数集合;基于第一参数集合,更新第二参数集合,其中,第二参数集合表征根据初始语言模型的输出,生成分类层的函数中的参数。

[0014] 在一些实施例中,第一参数集合包括:初始语言模型自注意力层之间的所有前馈神经网络层的参数;初始意图识别模型中多层感知机的参数;分类层中表征权重和偏置的参数;第二参数集合包括:语言模型中除前馈神经网络层外的网络结构的参数;根据初始语言模型的输出,生成分类层的函数中的参数;初始意图识别模型针对于训练集的训练过程的学习率。

[0015] 在一些实施例中,上述装置还包括:确定单元,被配置成根据所接收的选取指令,确定从文本数据集选取出的监督集中的文本数据所要掩盖掉的字词;以及训练单元,进一步被配置成:在初始意图识别模型完成上一组合数据集的训练后,基于预设采样率,确定将对初始意图识别模型训练的、包括部分监督集和部分训练集的组合数据集,以组合数据集中掩盖掉对应的字词的文本数据为输入,以所掩盖掉的字词为期望输出,训练初始意图识别模型,得到训练后的意图识别模型。

[0016] 在一些实施例中,上述装置还包括:微调单元,被配置成根据待识别文本对应的具

体场景，对意图识别模型进行微调，以通过微调后的意图识别模型，处理待识别文本，得到待识别文本所表达的意图。

[0017] 在一些实施例中，处理单元，进一步被配置成：过滤掉文本数据集中的文本数据中词频低于预设词频阈值的字词，得到文本数据集中的文本数据中的字词集合。

[0018] 第四方面，本申请实施例提供了一种针对自然语言的意图识别装置，包括：获取单元，被配置成获取表征自然语言的待识别文本；识别单元，被配置成通过意图识别模型，处理待识别文本，得到待识别文本所表达的意图，其中，意图识别模型根据第一方面任一实现方式描述的训练方法训练得到。

[0019] 第五方面，本申请实施例提供了一种计算机可读介质，其上存储有计算机程序，其中，程序被处理器执行时实现如第一方面任一实现方式描述的方法。

[0020] 第六方面，本申请实施例提供了一种电子设备，包括：一个或多个处理器；存储装置，其上存储有一个或多个程序，当一个或多个程序被一个或多个处理器执行，使得一个或多个处理器实现如第一方面任一实现方式描述的方法。

[0021] 本申请实施例提供的意图识别模型的训练方法及装置，通过对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合；从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集；对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型，从而提供了一种基于自监督的意图识别模型的训练方法，提高了训练任务的多样性和模型训练的灵活性。

附图说明

[0022] 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

[0023] 图1是本申请的一个实施例可以应用于其中的示例性系统架构图；

[0024] 图2是根据本申请意图识别模型的训练方法的一个实施例的流程图；

[0025] 图3是根据本实施例的意图识别模型的结构示意图；

[0026] 图4是根据本实施例的意图识别模型的训练方法的应用场景的示意图；

[0027] 图5是根据本申请的意图识别模型的训练方法的又一个实施例的流程图；

[0028] 图6是根据本申请的意图识别方法的一个实施例的流程图；

[0029] 图7是根据本申请的意图识别模型的训练装置的一个实施例的结构图；

[0030] 图8是根据本申请的意图识别装置的一个实施例的结构图；

[0031] 图9是适于用来实现本申请实施例的计算机系统的结构示意图。

具体实施方式

[0032] 下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

[0033] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相

互组合。下面将参考附图并结合实施例来详细说明本申请。

[0034] 图1示出了可以应用本申请的意图识别模型的训练方法及装置的示例性架构100。

[0035] 如图1所示，系统架构100可以包括终端设备101、102、103，网络104和服务器105。终端设备101、102、103之间通信连接构成拓扑网络，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。

[0036] 终端设备101、102、103可以是支持网络连接从而进行数据交互和数据处理的硬件设备或软件。当终端设备101、102、103为硬件时，其可以是支持网络连接，信息获取、交互、显示、处理等功能的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、膝上型便携计算机和台式计算机等等。当终端设备101、102、103为软件时，可以安装在上述所列举的电子设备中。其可以实现成例如用来提供分布式服务的多个软件或软件模块，也可以实现成单个软件或软件模块。在此不做具体限定。

[0037] 服务器105可以是提供各种服务的服务器，例如获取用户通过终端设备101、102、103发送的待识别文本，通过意图识别模型确定待识别文本所表达的意图的后台处理服务器。此外，服务器还可以基于自监督方式训练得到意图识别模型。可选的，服务器可以根据所识别到的意图，进行信息处理，以确定用户所需了解的目标信息，并将目标信息反馈至终端设备。作为示例，服务器105可以是云端服务器。

[0038] 需要说明的是，服务器可以是硬件，也可以是软件。当服务器为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当服务器为软件时，可以实现成多个软件或软件模块(例如用来提供分布式服务的软件或软件模块)，也可以实现成单个软件或软件模块。在此不做具体限定。

[0039] 还需要说明的是，本申请的实施例所提供的意图识别模型的训练方法、意识识别方法可以由服务器执行，也可以由终端设备执行，还可以由服务器和终端设备彼此配合执行。相应地，意图识别模型的训练装置、意图识别装置包括的各个部分(例如各个单元)可以全部设置于服务器中，也可以全部设置于终端设备中，还可以分别设置于服务器和终端设备中。

[0040] 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。当意图识别模型的训练方法运行于其上的电子设备不需要与其他电子设备进行数据传输时，该系统架构可以仅包括意图识别模型的训练方法运行于其上的电子设备(例如服务器或终端设备)。

[0041] 继续参考图2，示出了意图识别模型的训练方法的一个实施例的流程200，包括以下步骤：

[0042] 步骤201，对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合。

[0043] 本实施例中，意图识别模型的训练方法的执行主体(例如图1中的服务器)可以基于有线连接或无线连接方式从远程，或从本地获取文本数据集，并对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合。

[0044] 文本数据集中所包括的文本数据可以是任意场景中自然语言所表达的文本信息。作为示例，在客服对话场景中，上述执行主体可以预先采集涉及真实客服对话的语料。客服

对话是以会话维度进行的，在一个会话中，用户和客服交替发出消息。使用序列{q_1,a_1,…,q_n,a_n}表示一个会话。其中，q_i表示用户消息，a_i表示客服消息，i表示当前轮次，为1-n之间的整数，n表示总的对话轮次数。上述执行主体提取每个会话中的用户消息，得到文本数据集，表示为{q_1,…,q_n}。

[0045] 在获取文本数据集后，上述执行主体可以确定文本数据集中的每个文本数据中的字词，进而得到文本数据集中的每个文本数据中的字词集合。

[0046] 在这里，字词集合中的字词具有唯一性，每个字词均拥有标识唯一性的ID(Identity document，身份标识号)。上述执行主体确定文本数据集中的每个文本数据中的字词后，过滤掉相同字词，得到字词集合。

[0047] 本实施例中，上述执行主体可以以单字为单位，确定文本数据集中的每个文本数据中的单字，进而得到包括文本数据集中的每个文本数据中的单字的字词集合；还可以以词语为单位，确定文本数据集中的每个文本数据中的词语，进而得到包括文本数据集中的每个文本数据中的词语的字词集合。可以理解，当以词语为单位时，词语包括单词、合成词和词组(又称短语)。

[0048] 在本实施例的一些可选的实现方式中，上述执行主体可以过滤掉文本数据集中的文本数据中词频低于预设词频阈值的字词，得到文本数据集中的文本数据中的字词集合。

[0049] 其中，预设词频阈值可以根据实际情况具体设置，在此不作限定。当字词在文本数据中的词频较低时，表征文本数据集中包括该字词的文本数据较少，对其进行过滤，可以提高后续步骤随机选取对应的文本数据的灵活性。

[0050] 步骤202，从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集。

[0051] 本实施例中，上述执行主体可以从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集。

[0052] 其中，第一预设数量和第二预设数量可以根据实际情况具体设置，在此不做限定。

[0053] 需要说明的是，子集合中的字词是从字词集合中随机选取的，不同的子集合中可能存在相同的字词。针对于每个子集合中的字词，在文本数据集中随机选取包括该字词的多个文本数据，以得到与多个子集合一一对应的多个训练集。可以理解，不同的训练集中可能存在相同的文本数据。

[0054] 本实施例中，可以将每个训练集中的文本数据划分为训练子集和验证子集。通过训练子集中的训练样本进行意图识别模型的拟合，通过验证子集中的训练样本调整意图识别模型的超参数，以及对意图识别模型的能力进行初步评估。

[0055] 步骤203，对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。

[0056] 本实施例中，对于多个训练集中的训练集，上述执行主体以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。

[0057] 其中，文本数据掩盖掉对应的字词表征：当该文本数据基于所包括的一个字词被选取为该训练集中的训练样本，则将该字词掩盖。

[0058] 本实施例中，意图识别模型可以是任意具有意图识别功能的神经网络语言模型，包括但不限于是BERT(Bidirectional Encoder Representations from Transformers，基于Transformers的双向编码器表示)模型、LSTM(Long Short Term Memory networks，长短期记忆网络)模型、RNN(Recurrent Neural Networks，循环神经网络)等。

[0059] 具体的，文本数据X的序列表示为$X=\{x_1,\cdots,x_l\}$。其中，文本序列长度为l，x_i表示文本数据X中的第i个字词，i为1-l之间的整数。根据字词集合，上述执行主体将X中的每个字词转化为字词集合中对应的ID，得到文本序列对应的编号表示：$W=\{w_1,\cdots,w_l\}$。以意图识别模型为BERT模型为例，上述执行主体会在序列头尾分别加入[CLS]、[SEP]字符，其最终的序列为$W=\{w_{CLS},w_1,\cdots,w_l,w_{SEP}\}$。

[0060] 针对于文本序列对应的编号表示，上述执行主体将对应的字词的编号通过预设标识符进行掩盖，得到初始意图识别模型的输入。其中，预设标识符可以是预先指定的标识符。作为示例，预设标识符为[m]，序列为$W=\{w_{CLS},w_1,\cdots,w_l,w_{SEP}\}$中的对应的掩盖字词为$w_1$对应的字词，则掩盖后的序列为$W=\{w_{CLS},m,\cdots,w_l,w_{SEP}\}$。

[0061] 本实施例中，上述执行主体可以基于初始意图识别模型输出的意图、所输入的文本数据中被掩盖掉的字词，确定损失值，进而使用随机梯度下降算法和反向传播算法优化初始意图别模型的网络参数。

[0062] 作为示例，当训练过程达到预设结束条件时，确定完成初始意图识别模型的训练过程，确定得到意图识别模型。其中，预设结束条件例如可以是，损失值小于预设损失值阈值，训练时间超过预设时间阈值，训练次数超过预设次数阈值。

[0063] 在本实施例的一些可选的实现方式中，上述执行主体可以通过如下方式执行上述步骤203：

[0064] 第一、在初始语言模型的尾部添加分类层，得到初始意图识别模型。

[0065] 本实现方式中，初始语言模型可以是任意的神经网络语言模型，包括但不限于是BERT模型、LSTM模型、RNN模型。分类层可以是由全连接层和softmax层构成的分类层。

[0066] 第二、对于多个训练集中的训练集，执行如下操作，直至得到意图识别模型：

[0067] 首先，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，优化表征初始意图识别模型的参数的第一参数集合。

[0068] 本实现方式中，初始意图识别模型的训练过程包括内循环过程和外循环过程，其中，内循环过程表征初始意图识别模型通过同一训练集中的训练样本进行训练的迭代过程，外循环过程表征初始意图识别模型通过不同的训练集进行训练的迭代过程。

[0069] 内循环过程用于优化初始意图识别模型优化表征初始意图识别模型的参数的第一参数集合。外循环过程用于优化第二参数集合中的参数。

[0070] 然后，基于第一参数集合，更新第二参数集合。其中，第二参数集合表征根据初始语言模型的输出，生成分类层的函数中的参数。

[0071] 作为示例，对于参数为π的BERT模型，使用符号$f_\pi(X)$表示其最后一层的输出。对于一个训练集所对应的训练任务t，共有M个类别，每个类别对应于一个字词。其中类别m(M个类别中的一个类别)的支撑集中的k个样本表示为：$\{X_{1m},\cdots,X_{km}\}$，softmax层的生成过程如

下：

[0072] $$w_t^m, b_t^m = g_\psi(\{f_\pi(X_{1m}), \ldots, f_\pi(X_{km})\})$$

[0073] $p(y|X) = \mathrm{Softmax}\{W_t h_\phi(f_\pi(X)) + b_t\}$

[0074] 其中，$\mathbf{W}_t = [w_t^1; \ldots; w_t^M] \in \mathbb{R}^{M \times d}$，表征权重，$\mathbf{b}_t = [b_t^1; \ldots; b_t^M] \in \mathbb{R}^d$，表征偏置，它们都是由每个类别的向量$w_t^m$、$b_t^m$拼接而成，输出维度为d维；$g_\psi$表示根据初始语言模型的输出，生成分类层的函数，ψ表示该函数的参数；h_ϕ表示参数为Φ的多层感知机。初始语言模型的输出$f_\pi(X)$首先经过多层感知机进行变换，再通过分类层确定关于意图的分类信息。y表示输入的文本数据的标签，也即被掩盖掉的字词。

[0075] 在本实施例的一些可选的实现方式中，为了加快训练过程的模型的收敛速度，将初始语言模型的参数，即π，拆成两部分进行训练，将拆分后的两部分参数依次划分至上述第一参数集合、第二参数集合。一部分为初始语言模型自注意力层之间的所有前馈神经网络层的参数，另一部分为初始语言模型中剩余的参数，也即除前馈神经网络层外的网络结构的参数。

[0076] 本实现方式中，第一参数集合包括：初始语言模型自注意力层之间的所有前馈神经网络层的参数；初始意图识别模型中多层感知机的参数；分类层中表征权重和偏置的参数。第二参数集合包括：语言模型中除前馈神经网络层外的网络结构的参数；根据初始语言模型的输出，生成分类层的函数中的参数；初始意图识别模型针对于训练集的训练过程的学习率。

[0077] 初始意图识别模型的参数优化过程为：

[0078] $$\theta_i' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_i(D^{tr}, \theta)$$

[0079] $$\Theta \leftarrow \Theta - \beta \nabla_\Theta \mathbb{E}_{T_i \sim \mathcal{P}(\mathcal{T})}[\mathcal{L}_i(D^{val}, \theta_i')]$$

[0080] 其中，θ:＝{π1，Φ，W_t，b_t}表征第一参数集合，π1，Φ，W_t，b_t依次表示初始语言模型自注意力层之间的所有前馈神经网络层的参数、初始意图识别模型中多层感知机的参数、分类层中表征权重和偏置的参数；Θ:＝{π2，ψ，α}表征第二参数集合，π2，ψ，α依次表示语言模型中除前馈神经网络层外的网络结构的参数，根据初始语言模型的输出，生成分类层的函数中的参数，初始意图识别模型针对于训练集的训练过程的学习率。

[0081] 另外，D^{tr}、D^{val}依次表示训练子集、验证子集；$\mathcal{L}_i$表示任务T_i上的监督损失；β表示训练过程的学习率；$\mathbb{E}_{T_i \sim \mathcal{P}(\mathcal{T})}$表示求取平均值。

[0082] 继续参见图3，示出了意图识别模型的网络结构示意图300。其中，意图识别模型种包括Transformer网络301、参数生成器302、全连接层303和softmax层304。意图识别模型的训练过程包括内循环过程和外循环过程。每次的内循环过程例如可以通过一个批次的训练样本更新优化初始意图识别模型第一参数集合中的参数，在每次内循环完成后，外循环过程优化第二参数集合中的参数。在内循环之前，将一个批次的训练样本输入Transformer网络301，并通过参数生成器302生成softmax层304的网络参数。训练样本可以被灵活地划分为训练数据和验证数据，基于训练数据得到的损失为训练损失，基于验证数据得到的损失为验证损失。

[0083] 继续参见图4，图4是根据本实施例的意图识别模型的训练方法的应用场景的一个示意图400。在图4的应用场景中，服务器401首先采集得到文本数据集A＝{a_1，a_2，a_3……a_Q}，共有Q个文本数据。然后，对文本数据集A进行处理，得到文本数据集A中的文本数据中的字词集合B＝{b_1，b_2，b_3……b_P}。字词集合B中包括P个字词。然后，服务器401从字词集合B中选取包括5个字词的多个子集合B_1＝{b_1，b_5，b_{10}}，B_2＝{b_3，b_7，b_{30}}，B_3＝{b_5，b_7，b_{20}}；并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集L_1，L_2，L_3。以L_1为例，其中的文本数据包括：包括b_1的5个文本数据a_1、a_8、a_{14}、a_{32}、a_{43}，包括b_5的5个文本数据a_2、a_3、a_{15}、a_{23}、a_{25}，包括b_{10}的5个文本数据a_7、a_9、a_{16}、a_{29}、a_{37}，共15(3*5)个文本数据。最后，对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。作为示例，将训练集L_1中的文本数据a_2、a_3、a_{15}、a_{23}、a_{25}掩盖掉b_5，作为输入数据，以b_5作为期望输出，训练初始意图识别模型。

[0084] 本申请的上述实施例提供的方法，通过对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合；从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集；对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型，从而提供了一种基于自监督的意图识别模型的训练方法，提高了训练任务的多样性和模型训练的灵活性。

[0085] 在本实施例的一些可选的实现方式中，根据待识别文本对应的具体场景，对意图识别模型进行微调，以通过微调后的意图识别模型，处理待识别文本，得到待识别文本所表达的意图。

[0086] 继续参考图5，示出了根据本申请的意图识别模型的训练方法的一个实施例的示意性流程500，包括以下步骤：

[0087] 步骤501，对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合。

[0088] 步骤502，从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集。

[0089] 本实施例中，上述执行主体可以参照步骤201-202的方式执行上述步骤501-502，在此不再赘述。

[0090] 步骤503，根据所接收的选取指令，确定从文本数据集选取出的监督集中的文本数据所要掩盖掉的字词。

[0091] 本实施例中，意图识别模型的训练方法的执行主体(例如图1中的服务器)可以根据所接收的选取指令，确定从文本数据集选取出的监督集中的文本数据所要掩盖掉的字词。

[0092] 本实施例中，可以人工选取文本数据以及文本数据中所需掩盖掉的字词，形成监督集。

[0093] 步骤504，在初始意图识别模型完成上一组合数据集的训练后，基于预设采样率，

确定将对初始意图识别模型训练的、包括部分监督集和部分训练集的组合数据集，以组合数据集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。

[0094] 本实施例中，上述执行主体可以基于预设采样率，确定多个包括部分监督集和部分训练集的组合数据集。对于多个组合数据集中的组合数据集，以该组合数据集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。

[0095] 其中，预设采样率为0至1之间的数值，可以根据实际情况具体设置，在此不作限定。

[0096] 作为示例，预设采样率为0.5，则上述执行主体从训练集中以0.5的采样率获取部分(具体为二分之一)训练集，以0.5(1-0.5)的采样率从监督集中获取部分(具体为二分之一)监督集，得到组合数据集。

[0097] 需要说明的是，在不冲突的前提下，基于上述组合训练集对初始意图识别模型的训练过程，可以采用或结合步骤203中任一实施方式，以得到更多的实施方式。

[0098] 作为示例，基于上述组合训练集对初始意图识别模型的训练过程中，依然可以将所要优化的参数划分为第一数据集合、第二数据集合。采用内循环过程优化初始意图识别模型优化表征初始意图识别模型的参数的第一参数集合，外循环过程优化第二参数集合中的参数的方式，训练初始意图识别模型。

[0099] 从本实施例中可以看出，与图2对应的实施例相比，本实施例中的意图识别模型的训练方法的流程500具体说明了基于预设采样率得到的组合数据集对初始意图识别模型进行训练的过程，进一步提高了模型训练的灵活性。

[0100] 继续参考图6，示出了意图识别模型的训练方法的一个实施例的流程600，包括以下步骤：

[0101] 步骤601，获取表征自然语言的待识别文本。

[0102] 本实施例中，意图识别方法的执行主体(例如图1中的服务器)可以获取表征自然语言的待识别文本。

[0103] 其中，待识别文本可以是表征自然语言的任一文本。作为示例，在客服对话场景中，待识别文本为用户所发送的信息。

[0104] 步骤602，通过意图识别模型，处理待识别文本，得到待识别文本所表达的意图。

[0105] 本实施例中，上述执行主体可以通过意图识别模型，处理待识别文本，得到待识别文本所表达的意图。其中，意图识别模型根据实施例200、500中描述的训练方法训练得到。

[0106] 在本实施例的一些可选的实现方式中，在确定待识别文本所表达的意图后，上述执行主体可以进行信息处理，以确定用户所需了解的目标信息。

[0107] 继续以客服对话场景为例，用户的待识别文本为“身高175cm，体重70KG，什么尺码的裤子合适”，上述执行主体在确定用户的意图为“尺码”后，可以进行信息处理，确定对应于用户的身高、体重信息的裤子尺码，并生成关于尺码信息的目标信息，反馈至用户的终端设备。

[0108] 本申请的上述实施例提供的方法，通过获取表征自然语言的待识别文本；通过意图识别模型，处理待识别文本，得到待识别文本所表达的意图，从而提供了一种意图识别方

法，丰富了针对于自然语言的意图识别方式。

[0109] 继续参考图7，作为对上述各图所示方法的实现，本申请提供了一种意图识别模型的训练装置的一个实施例，该装置实施例与图2所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0110] 如图7所示，意图识别模型的训练装置包括：包括：处理单元701，被配置成对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合；选取单元702，被配置成从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集；训练单元703，被配置成对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。

[0111] 在本实施例的一些可选的实现方式中，训练单元703，进一步被配置成：在初始语言模型的尾部添加分类层，得到初始意图识别模型；对于多个训练集中的训练集，执行如下操作，直至得到意图识别模型：以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，优化表征初始意图识别模型的参数的第一参数集合；基于第一参数集合，更新第二参数集合，其中，第二参数集合表征根据初始语言模型的输出，生成分类层的函数中的参数。

[0112] 在本实施例的一些可选的实现方式中，第一参数集合包括：初始语言模型自注意力层之间的所有前馈神经网络层的参数；初始意图识别模型中多层感知机的参数；分类层中表征权重和偏置的参数；第二参数集合包括：语言模型中除前馈神经网络层外的网络结构的参数；根据初始语言模型的输出，生成分类层的函数中的参数；初始意图识别模型针对于训练集的训练过程的学习率。

[0113] 在本实施例的一些可选的实现方式中，上述装置还包括：确定单元(图中未示出)，被配置成根据所接收的选取指令，确定从文本数据集选取出的监督集中的文本数据所要掩盖掉的字词；以及训练单元，进一步被配置成：在初始意图识别模型完成上一组合数据集的训练后，基于预设采样率，确定将对初始意图识别模型训练的、包括部分监督集和部分训练集的组合数据集，以组合数据集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型。

[0114] 在本实施例的一些可选的实现方式中，上述装置还包括：微调单元(图中未示出)，被配置成根据待识别文本对应的具体场景，对意图识别模型进行微调，以通过微调后的意图识别模型，处理待识别文本，得到待识别文本所表达的意图。

[0115] 在一些实施例中，处理单元701，进一步被配置成：过滤掉文本数据集中的文本数据中词频低于预设词频阈值的字词，得到文本数据集中的文本数据中的字词集合。

[0116] 本实施例中，意图识别模型的训练装置中的处理单元对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合；选取单元从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集；训练单元对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型，从而提

供了一种基于自监督的意图识别模型的训练装置，提高了训练任务的多样性和模型训练的灵活性。

[0117] 继续参考图8，作为对上述各图所示方法的实现，本申请提供了一种意图识别模型的训练装置的一个实施例，该装置实施例与图6所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0118] 如图8所示，针对自然语言的意图识别装置，包括：获取单元801，被配置成获取表征自然语言的待识别文本；识别单元802，被配置成通过意图识别模型，处理待识别文本，得到待识别文本所表达的意图。其中，意图识别模型根据实施例200、500中描述的训练方法训练得到。

[0119] 本实施例中，意图识别装置中的获取单元获取表征自然语言的待识别文本；识别单元通过意图识别模型，处理待识别文本，得到待识别文本所表达的意图，从而提供了一种意图识别装置，丰富了针对于自然语言的意图识别方式。

[0120] 下面参考图9，其示出了适于用来实现本申请实施例的设备(例如图1所示的设备101、102、103、105)的计算机系统900的结构示意图。图9示出的设备仅仅是一个示例，不应对本申请实施例的功能和使用范围带来任何限制。

[0121] 如图9所示，计算机系统900包括处理器(例如CPU，中央处理器)901，其可以根据存储在只读存储器(ROM)902中的程序或者从存储部分908加载到随机访问存储器(RAM)903中的程序而执行各种适当的动作和处理。在RAM903中，还存储有系统900操作所需的各种程序和数据。处理器901、ROM902以及RAM903通过总线904彼此相连。输入/输出(I/O)接口905也连接至总线904。

[0122] 以下部件连接至I/O接口905：包括键盘、鼠标等的输入部分906；包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分907；包括硬盘等的存储部分908；以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分909。通信部分909经由诸如因特网的网络执行通信处理。驱动器910也根据需要连接至I/O接口905。可拆卸介质911，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器910上，以便于从其上读出的计算机程序根据需要被安装入存储部分908。

[0123] 特别地，根据本申请的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本申请的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分909从网络上被下载和安装，和/或从可拆卸介质911被安装。在该计算机程序被处理器901执行时，执行本申请的方法中限定的上述功能。

[0124] 需要说明的是，本申请的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——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。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中，计算机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。而在本申

请中，计算机可读的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，其中承载了计算机可读的程序代码。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。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、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。计算机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、电线、光缆、RF等等，或者上述的任意合适的组合。

[0125] 可以以一种或多种程序设计语言或其组合来编写用于执行本申请的操作的计算机程序代码，程序设计语言包括面向目标的程序设计语言—诸如Java、Smalltalk、C++，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客户计算机上执行、部分地在客户计算机上执行、作为一个独立的软件包执行、部分在客户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到客户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0126] 附图中的流程图和框图，图示了按照本申请各种实施例的装置、方法和计算机程序产品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，该模块、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，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示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。

[0127] 描述于本申请实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中，例如，可以描述为：一种处理器，包括处理单元、选取单元和训练单元；还可描述为：一种处理器，包括获取单元和识别单元。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定，例如，训练单元还可以被描述为"对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型的单元"。

[0128] 作为另一方面，本申请还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备中所包含的；也可以是单独存在，而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该装置执行时，使得该计算机设备：对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合；从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集；对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意

图识别模型。以及使得该计算机设备:获取表征自然语言的待识别文本;通过意图识别模型,处理待识别文本,得到待识别文本所表达的意图。

[0129] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解,本申请中所涉及的发明范围,并不限于上述技术特征的特定组合而成的技术方案,同时也应涵盖在不脱离上述发明构思的情况下,由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

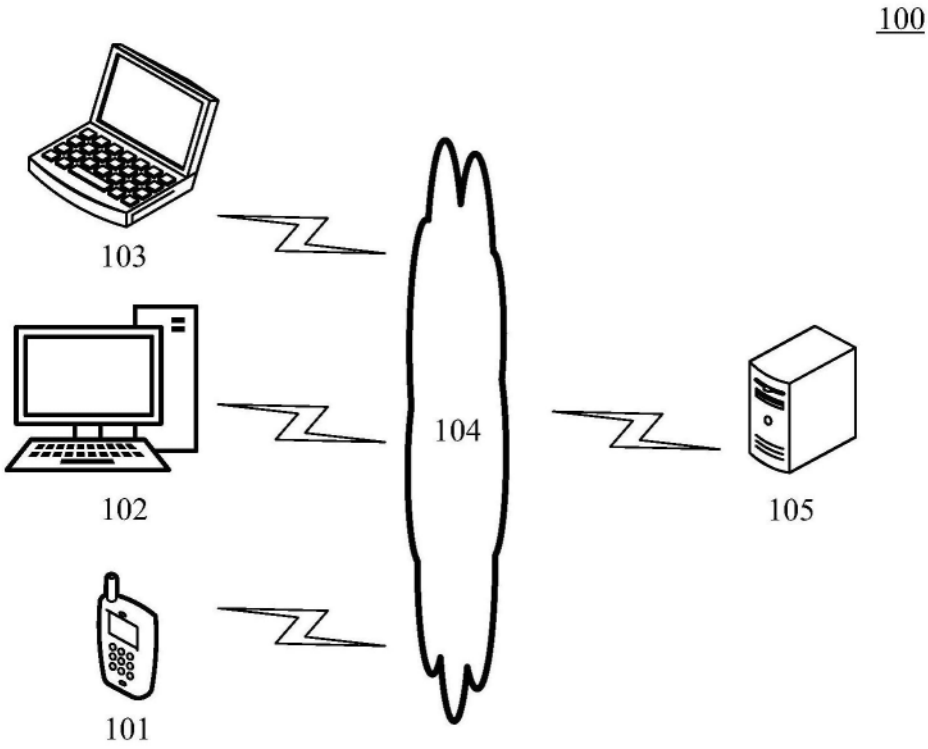


图1

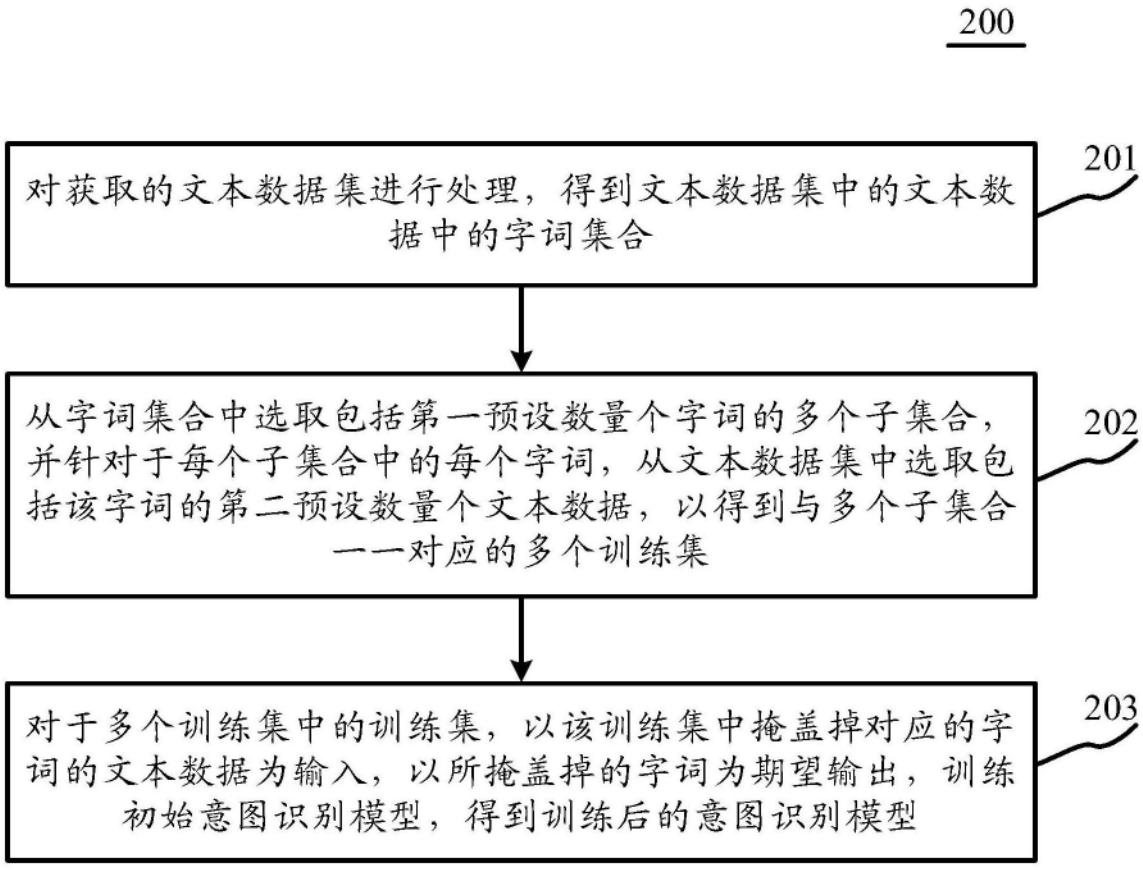

200
对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合
201
从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集
202
对于多个训练集中的训练集，以该训练集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型
203

图2

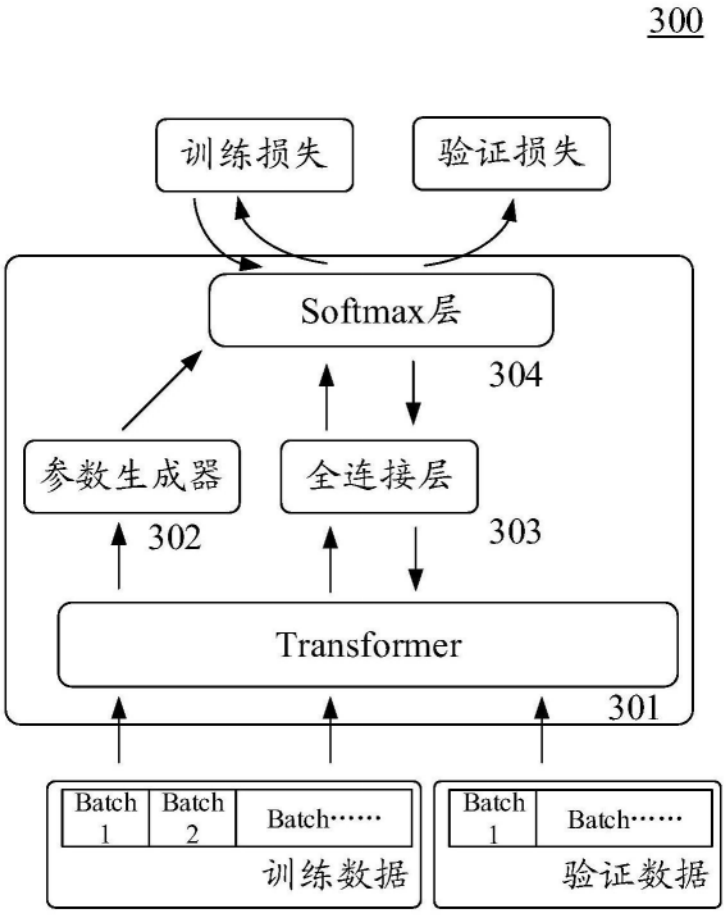

300
训练损失
验证损失
Softmax层
304
参数生成器
302
全连接层
303
Transformer
301
Batch 1
Batch 2
Batch……
训练数据
Batch 1
Batch……
验证数据

图3

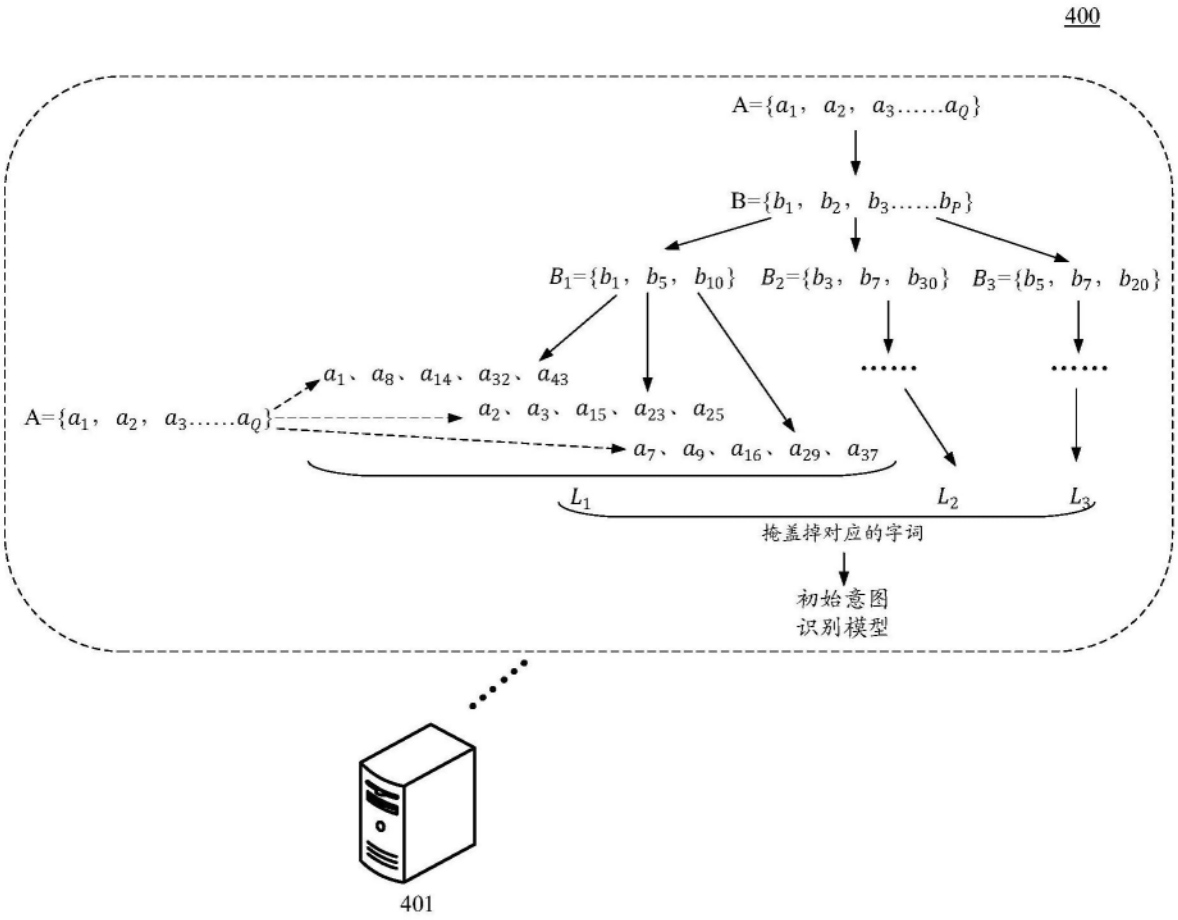

400
A={a1， a2， a3……aQ}
B={b1， b2， b3……bP}
B1={b1， b5， b10} B2={b3， b7， b30} B3={b5， b7， b20}
a1、 a8、 a14、 a32、 a43
A={a1， a2， a3……aQ}
a2、 a3、 a15、 a23、 a25
a7、 a9、 a16、 a29、 a37
L1
L2
L3
掩盖掉对应的字词
初始意图
识别模型
401

图4

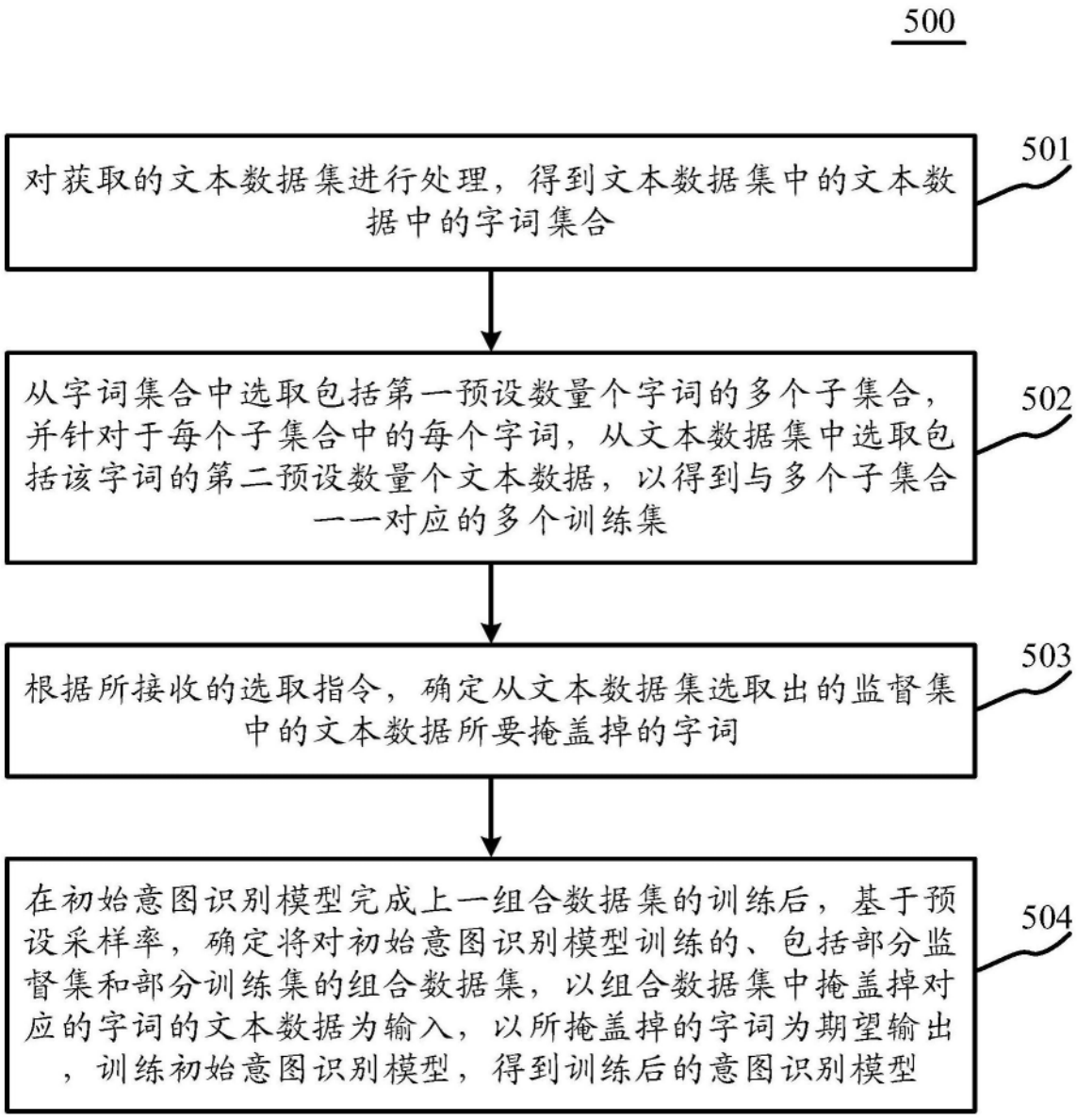

500
对获取的文本数据集进行处理，得到文本数据集中的文本数据中的字词集合
501
从字词集合中选取包括第一预设数量个字词的多个子集合，并针对于每个子集合中的每个字词，从文本数据集中选取包括该字词的第二预设数量个文本数据，以得到与多个子集合一一对应的多个训练集
502
根据所接收的选取指令，确定从文本数据集选取出的监督集中的文本数据所要掩盖掉的字词
503
在初始意图识别模型完成上一组合数据集的训练后，基于预设采样率，确定将对初始意图识别模型训练的、包括部分监督集和部分训练集的组合数据集，以组合数据集中掩盖掉对应的字词的文本数据为输入，以所掩盖掉的字词为期望输出，训练初始意图识别模型，得到训练后的意图识别模型
504

图5

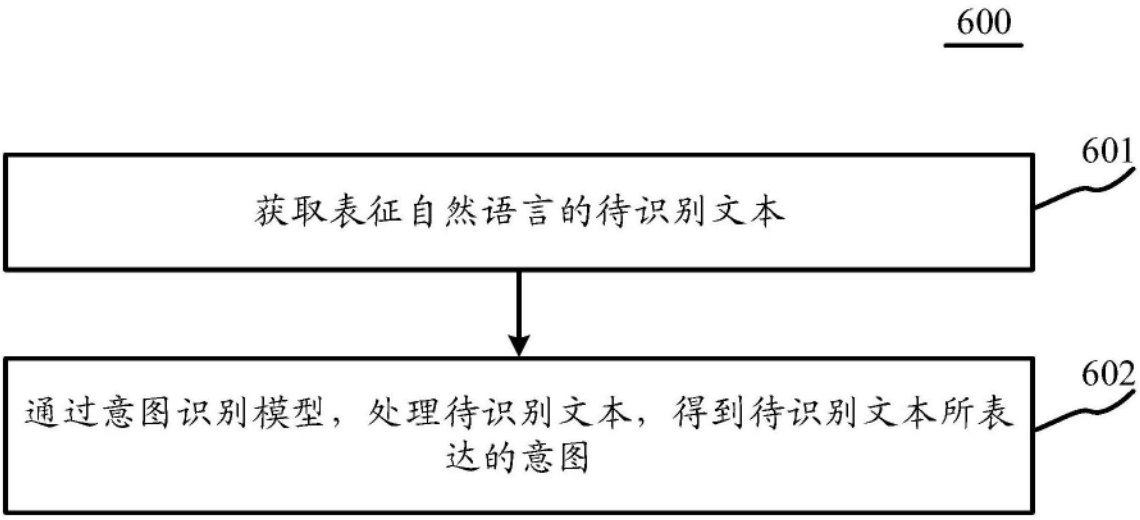

600
601
获取表征自然语言的待识别文本
602
通过意图识别模型，处理待识别文本，得到待识别文本所表达的意图

图6

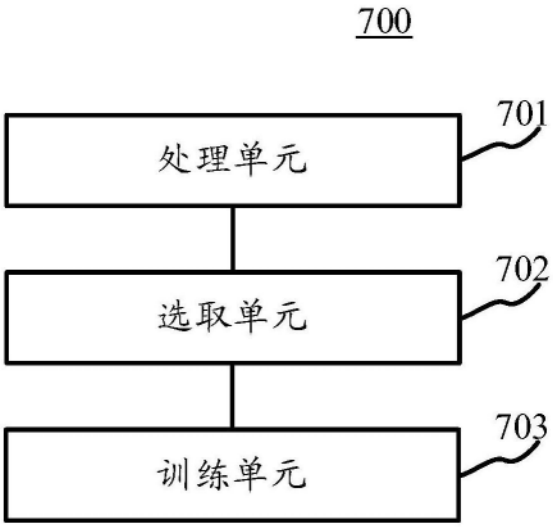

700
701
处理单元
702
选取单元
703
训练单元

图7

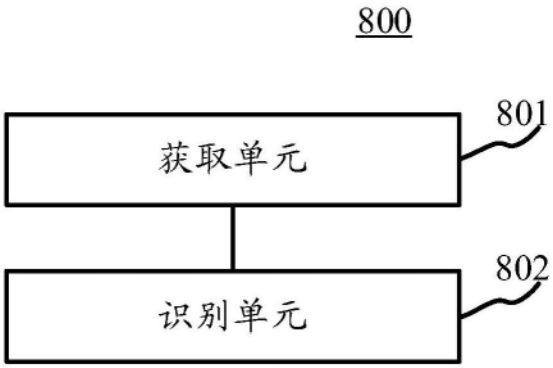

800
801
获取单元
802
识别单元

图8

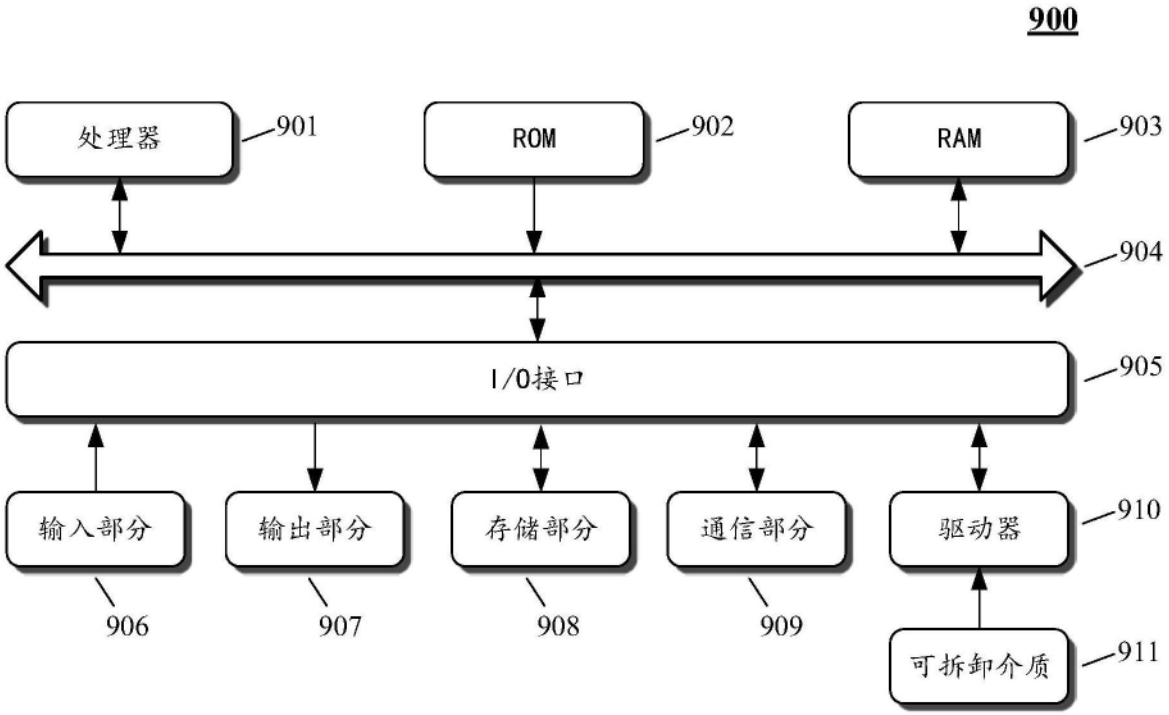

900
处理器
901
ROM
902
RAM
903
904
I/O接口
905
输入部分
输出部分
存储部分
通信部分
驱动器
910
906
907
908
909
可拆卸介质
911

图9