(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
2022 年 10 月 6 日 (06.10.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/206091 A1

(51) 国际专利分类号:
G06F 40/20 (2020.01) *G06F 16/35* (2019.01)

(21) 国际申请号: PCT/CN2022/070250

(22) 国际申请日: 2022 年 1 月 5 日 (05.01.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110340188.4 2021年3月30日 (30.03.2021) CN

(71) 申请人: 京东科技控股股份有限公司 (JINGDONG TECHNOLOGY HOLDING CO., LTD.) [CN/CN]; 中国北京市北京经济技术开发区科创十一街18号C座2层221室, Beijing 100176 (CN)。

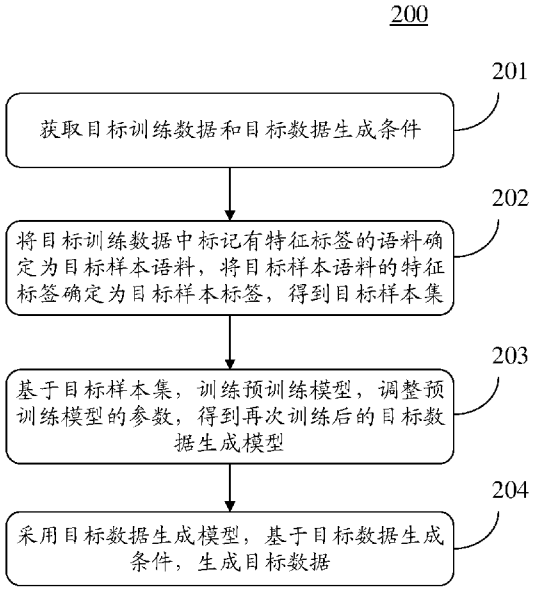
(72) 发明人: 刘瑞雪(LIU, Ruixue); 中国北京市北京经济技术开发区科创十一街18号C座2层221室, Beijing 100176 (CN)。 陈蒙(CHEN, Meng); 中国北京市北京经济技术开发区科创十一街18号C座2层221室, Beijing 100176 (CN)。

(74) 代理人: 北京英赛嘉华知识产权代理有限责任公司 (INSIGHT INTELLECTUAL PROPERTY LIMITED); 中国北京市朝阳区建外大街光华东里8号中海广场中楼7层, Beijing 100020 (CN)。

(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, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) **Title:** DATA GENERATION METHOD AND APPARATUS

(54) **发明名称:** 用于生成数据的方法和装置



201 Acquire target training data and a target data generation condition

202 Determine a language material, in the target training data, marked with a feature tag as a target sample language material, and determine a feature tag of the target sample language material as a target sample tag to obtain a target sample set

203 Train a pre-training model on the basis of the target sample set, and adjust parameters of the pre-training model to obtain a re-trained target data generation model

204 Use the target data generation model to generate target data on the basis of the target data generation condition

图 2

(57) **Abstract:** Disclosed in embodiments of the present disclosure are a data generation method and apparatus. A specific embodiment of the method comprises: acquiring target training data and a target data generation condition, wherein the target training data comprises language materials in the target field, and each language material is marked with a feature tag; constructing a target sample set on the basis of the target training data; training a pre-training model on the basis of the target sample set, and adjusting parameters of the pre-training model to obtain a re-trained target data generation model, wherein the pre-training model is obtained by the following

[见续页]

WO 2022/206091 A1

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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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))。

steps: constructing an initial model and training the initial model on the basis of a general sample set to obtain the pre-training model; and using the target data generation model to generate target data on the basis of the target data generation condition.

(57) 摘要：本公开的实施例公开了用于生成数据的方法和装置。该方法的一具体实施方式包括：获取目标训练数据和目标数据生成条件，其中，目标训练数据包括目标领域的语料，语料标记有特征标签；基于目标训练数据，构建目标样本集；基于目标样本集，训练预训练模型，调整预训练模型的参数，得到再次训练后的目标数据生成模型，其中，预训练模型经由如下步骤得到：构建初始模型并基于通用样本集训练初始模型，得到预训练模型；采用目标数据生成模型，基于目标数据生成条件，生成目标数据。

用于生成数据的方法和装置

交叉引用

本申请要求于2021年3月30日提交的、申请号为202110340188.4、发明名称为“用于生成数据的方法和装置”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

本公开的实施例涉及计算机技术领域，具体涉及人工智能领域，尤其涉及用于生成数据的方法和装置。

背景技术

数据增强（Data Augmentation）是一种让有限的数据产生更多的等价数据以扩展训练数据集的技术，是克服训练数据不足的有效手段。例如深度学习方法通常需要大量的训练数据才能避免过度拟合，然而实践中有时无法获得足够的数据，这就需要借助数据增强解决这类问题。

相关技术中，文本数据增强方法包括两种，一种是在保持句子原有结构的前提下，对句子的进行局部修改，生成新的句子。例如，利用简单的同义词替换、随机交换词、随机删词等方法生成新句子。再例如最近提出的Masked（掩模）语言模型，对单词进行掩蔽预测，同时在类标签上进行条件调整，从而实现数据扩充。另一种是通过大量的数据对文本生成模型进行预训练，然后通过文本生成模型采用数据生成的方法生成完整的句子，而不是做一些局部的改变。例如，反向翻译（back translation）,即先将语料翻译成另外一种语言，再将其翻译回源语言，以生成更多变化的句子。再例如，采用复述（paraphrasing）的方法，在文本生成模型的输入端增加噪音来生成更多语句。

发明内容

本公开的实施例提出了用于生成数据的方法和装置。

本公开的实施例提供了一种用于生成数据的方法，该方法包括：获取目标训练数据和目标数据生成条件，目标训练数据包括标记有特征标签的目标领域的语料；将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集；基于目标样本集，训练预训练模型，调整预训练模型的参数，得到再次训练后的目标数据生成模型，其中，预训练模型经由如下步骤得到：构建初始模型并基于通用样本集训练初始模型，得到预训练模型；采用目标数据生成模型，基于目标数据生成条件，生成目标数据。

在一些实施例中，基于目标样本集，训练预训练模型，包括：将目标样本标签输入预训练模型，将目标样本语料作为期望输出，训练预训练模型，得到目标数据生成模型。

在一些实施例中，目标数据生成条件包括目标特征标签；以及，采用目标数据生成模型，基于目标数据生成条件，生成目标数据，包括：将目标特征标签输入目标数据生成模型，得到目标语料；将目标语料确定为目标数据。

在一些实施例中，目标特征标签为预先构建的分类模型基于待识别语料估计出的分类标签；以及，在将目标语料确定为目标数据之前，方法还包括：将目标语料输入分类模型，得到目标语料的分类标签；响应于确定分类模型的预设标签集中包括目标语料的分类标签，将目标语料确定为目标数据，目标数据用于构建分类模型的训练样本。

在一些实施例中，基于目标样本集，训练预训练模型，包括：将目标样本语料输入预训练模型，将目标样本标签作为期望输出，训练预训练模型，得到目标数据生成模型。

在一些实施例中，目标数据生成条件包括目标待识别语料；以及，采用目标数据生成模型，基于目标数据生成条件，生成目标数据，包括：将目标待识别语料输入目标数据生成模型，得到目标待识别语料的特征标签；将目标待识别语料的特征标签确定为目标数据。

本公开的实施例提供了一种用于生成数据的装置，装置包括：数据获取单元，被配置成获取目标训练数据和目标数据生成条件，目标训练数据包括标记有特征标签的目标领域的语料；样本构建单元，被配置成将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集；模型调整单元，被配置成基于目标样本集，训练预训练模型，调整预训练模型的参数，得到再次训练后的目标数据生成模型，其中，预训练模型经由如下步骤得到：构建初始模型并基于通用样本集训练初始模型，得到预训练模型；数据生成单元，被配置成采用目标数据生成模型，基于目标数据生成条件，生成目标数据。

在一些实施例中，模型调整单元被进一步配置成：将目标样本标签输入预训练模型，将目标样本语料作为期望输出，训练预训练模型，得到目标数据生成模型。

在一些实施例中，目标数据生成条件包括目标特征标签；以及，数据生成单元被进一步配置成：将目标特征标签输入目标数据生成模型，得到目标语料；将目标语料确定为目标数据。

在一些实施例中，目标特征标签为预先构建的分类模型基于待识别语料估计出的分类标签；以及，数据生成单元还包括数据验证模块，被配置成：将目标语料输入分类模型，得到目标语料的分类标签；响应于确定分类模型的预设标签集中包括目标语料的分类标签，将目标语料确定为目标数据，目标数据用于构建分类模型的训练样本。

在一些实施例中，模型调整单元还被进一步配置成：将目标样本语料输入预训练模型，将目标样本标签作为期望输出，训练预训练模型，得到目标数据生成模型。

在一些实施例中，目标数据生成条件包括目标待识别语料；以及，数据生成单元还被配置成：将目标待识别语料输入目标数据生成模型，得到目标待识别语料的特征标签；将目标待识别语料的特征标签确定为目标数据。

本公开的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上述任一实施例中的方法。

本公开的实施例还提供了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上述任一实施例中的方法。

附图说明

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，本公开的其它特征、目的和优点将会变得更明显：

图 1 是本公开的一些实施例可以应用于其中的示例性系统架构图；

图 2 是根据本公开的用于生成数据的方法的一个实施例的流程图；

图 3 是根据本公开的用于生成数据的方法的又一个实施例的流程图；

图 4 是根据本公开的用于生成数据的方法的又一个实施例的流程图；

图 5 是根据本公开的用于生成数据的装置的一个实施例的结构示意图；

图 6 是适于用来实现本公开的实施例的电子设备的结构示意图。

具体实施方式

下面结合附图和实施例对本公开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方案，而非对该方案的限定。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方案相关的部分。

需要说明的是，在不冲突的情况下，本公开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公开。

图 1 示出了可以应用本公开的实施例的用于生成数据的方法或用于生成数据的装置的示例性系统架构 100。

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。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。

用户可以使用终端设备 101、102、103 通过网络 104 与服务器 105 交互，以接收或发送数据等，例如用户可以将目标领域的原始数据发送至服务器，还可以从服务器接收目标数据生成模型生成的目标数据。

终端设备 101、102、103 可以是硬件，也可以是软件。当终端设备 101、102、103 为硬件时，可以是具备通信功能的电子设备，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。

服务器 105 可以是提供各种服务的服务器，例如对终端设备 101、102、103 上传的原始数据进行处理（例如基于目标训练数据构建训练样本）的后台数据服务器。后台数据服务器可以利用接收到的原始数据对预训练模型进行调整，得到的数据生成模型用于生成新的数据，并将处理结果（例如生成的目标数据）反馈给终端设备。

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。

需要说明的是，本公开的实施例所提供的用于生成数据的方法可以由终端设备 101、102、103 执行，也可以由服务器 105 执行。相应地，用于生成数据的装置可以设置于终端设备 101、102、103 中，也可以设置于服务器 105 中。在此不做具体限定。

继续参考图 2，示出了根据本公开的用于生成数据的方法的一个实施例的流程 200。该用于生成数据的方法，包括以下步骤：

步骤 201，获取目标训练数据和目标数据生成条件。

在本实施例中，目标训练数据包括标记有特征标签的目标领域的语料。特征标签表征语料的特征，可以包括多个维度，例如，结构特征标签可以表征语料的结构特征，意图标签可以表征语料的意图特征，语义标签可以表征语料的语义特征。目标数据生成条件表征用户对于生成的数据的期望，例如可以是包括目标领域中实体信息的数据，还可以是包含特定句法结构或语义信息的数据。

作为示例，操作人员接收到某个技术领域的数据生成任务时，可以通过执行主体（例如图 1 中所示的服务器 105）从业务方直接获取该领域的目标训练数据和目标数据生成条件，还可以从网络上获取该领域的真实语料，并对各真实语料标记对应的特征标签，得到目标训练数据。

需要说明的是，目标训练数据还可以包括未经标注的语料。

步骤 202，将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集。

在本实施例中，目标领域的语料可以包括目标领域中真实语料的多个维度的特征，例如句子结构特征、词特征、语义特征等。相应地，目标样本标签可以从多个维度表征目标样本语料的特征，例如可以从句子结构维度标记目标样本语料，还可以从关键词的维度标记目标样本语料，再例如，还可以从目标领域的实体命名的维度标记目标样本语料。

步骤 203，基于目标样本集，训练预训练模型，调整预训练模型的参数，得到再次训练后的目标数据生成模型。

在本实施例中，预训练模型经由如下步骤得到：构建初始模型并基于通用样本集训练初始模型，得到预训练模型。其中，通用样本集中的训练数据为容易获取到的、各个领域的训练数据。作为示例，可以选择 ELMo（Embeddings from Language Models，语言模型的嵌入）、BERT（Bidirectional Encoder Representation from Transformers，来自变压器的双向编码器表示）或 GPT（Generative Pre-Training，生成式预训练）等模型作为初始模型，之后执行主体可以通过网络获取公开数

据，并基于公开数据构建通用样本集，如此可以确保初始模型在预训练阶段有足够的训练样本。

初始模型经通过样本集训练之后得到的预训练模型，可以学习基本的数据生成规则（例如可以生成连贯的、真实语料），但是对于某些数据获取难度较大的领域来说，预训练模型生成的数据与该领域的真实数据之间的相似度较低。此时，基于目标样本集对预训练模型进行再次训练，调整预训练模型中的参数，使预训练模型学习目标领域中数据的生成规则，以使经再次训练得到的目标数据模型生成的数据更加接近于真实数据。

步骤 204，采用目标数据生成模型，基于目标数据生成条件，生成目标数据。

在本实施例中，目标数据生成模型表征目标数据生成条件与目标数据之间的对应关系。

在一个具体的示例中，某个特定领域的数据量较小，为了增强该领域的数据，可以将该领域作为目标领域。执行主体可以基于网络上的公开数据（例如可以是中文小说或对话语料等数据）构建通用样本集，然后基于通用样本集训练初始的 GPT 模型，得到的预训练的 GPT 模型可以生成连贯的真实句子。之后，执行主体可以获取目标领域的语料作为目标训练数据，并构建目标样本集，然后基于目标样本集再次训练该 GPT 模型，调整 GPT 模型的参数，使其学习目标领域中真实语料的生成规则，训练完成后得到的 GPT 模型即为目标数据生成模型。再之后，执行主体可以获取目标数据生成条件（例如可以是关键词标签、句结构标签、语义标签等），并将目标数据标签输入 GPT 模型，即可由 GPT 生成新的语料，从而拓展了目标领域的数据量。表 1 示出了该示例中的目标训练数据（包括输入标签和训练语料）和 GPT 生成的目标语料。

表 1

输入标签	训练语料	目标语料
重发短信	能重新给我发个短信吗；我的购票短信不小心被删了；我在商店买的那个套餐的短信不见了，能给我重新发一下吗	我想问下我的短信什么时候能发、我有一个客户短信不小心把短信删了应该怎么办；我在商店买的这个套餐怎么还没有给我发兑换码呢
转赠亲友	我没有哪些地方的亲友我怎么转增；那怎么转增呢，我自己拿三个手机号买三次吗	我订了蛋糕的链接想转赠朋友如何操作；我都没用呢想转赠给朋友；那怎么转增呢我自己拿 2 次机会去使用吗
预约异常	不能预约了怎么处理呢；我查了预约日历根本约不了；我昨天去了他不给我打包	为什么预约时间还没有预约日期；我这边已经预约成功了为什么还没有预约；我这边预约了一个周末但是提前一天预约不了了

本公开的实施例提供的用于生成数据的方法和装置，通过目标领域的少量数据对预训练模型进行再次训练，以使得到的数据生成模型学习目标领域的数据生成规则，可以增强数据，并提高生成的数据的真实性和针对性。

继续参见图 3，图 3 根据本公开的用于生成数据的方法的又一个实施例的流程图，在图 3 所示的流程 300 中，包括以下步骤：

步骤 301，获取目标训练数据和目标数据生成条件。

步骤 302，将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集。步骤 301、步骤 302 与前述步骤 201、步骤 202 相近，此处不再赘述。

步骤 303，将目标样本标签输入预训练模型，将目标样本语料作为期望输出，训练预训练模型，得到目标数据生成模型。

在本实施例中，目标样本标签可以表征目标样本语料的特征。预训练模型以目标样本标签作为条件标签，以条件标签约束语料的生成过程，然后通过对比目标样本语料与生成的语料，确定损失函数，以此得到目标数据生成模型。目标数据生成模型表征条件标签与生成语料之间的对应关系。

步骤 304，将目标特征标签输入目标数据生成模型，得到目标语料。

在本实施例中，目标数据生成条件包括目标特征标签，表征用户在一个或多个维度上对生成的语料的期望。

执行主体将目标特征标签输入目标数据生成模型，作为目标数据生成模型的条件标签，约束其语料生成过程，以生成符合用户期望的目标语料。

步骤 305，将目标语料确定为目标数据。

从图 3 可以看出，本实施例中的用于生成数据的方法的流程 400 突出了利用目标数据生成模型生成语料数据的步骤，对于语料的特征标签数据量充足、语料数据量欠缺的应用场景，本实施例中的用于生成数据的方法仅需要少量目标领域的训练数据即可确保生成的语料更贴近于目标领域的真实预料，可以更有针对性地增强数据。

在本实施例的一些可选的实现方式中，目标特征标签为预先构建的分类模型基于待识别语料估计出的分类标签，以及，在将目标语料确定为目标数据（步骤 306）之前，上述流程 300 中的还可以进一步包括：将目标语料输入分类模型，得到目标语料的分类标签；响应于确定分类模型的预设标签集中包括目标语料的分类标签，将目标语料

确定为目标数据，目标数据用于构建分类模型的训练样本。

在本实现方式中，目标数据生成模型用于拓展分类模型的训练数据。若目标数据生成模型生成的语料可以被分类模型正确识别，则证明目标生成模型生成的语料的真实性符合分类模型的训练需求。

作为示例，需要构建用于识别某个特定领域的语料的分类模型。可以理解的是，训练样本的数据量与模型的准确度是正相关的，因而为了确保分类模型的准确度，需要足够的分类样本语料，而对于某些特定领域来说，语料的数据量获取难度较大。此时可以将该领域作为目标领域，由执行主体基于所获得的少量分类样本预料构建目标样本集，通过再次训练得到目标数据生成模型。之后将构建分类模型的样本分类标签输入目标数据生成模型，得到目标语料。然后将目标语料再输入该分类模型，若分类模型输出的分类标签与样本分类标签一致，则说明目标语料的真实性符合分类模型的训练需求。如此，得到的目标数据可以有效地拓展分类模型的样本数据。

需要说明的是，一个语料可以对应有多个特征标签，分别从多个维度表征语料的特征。在上述实施例的一些可选的实现方式中，基于目标样本集再次训练预训练模型时，可以同时将多个目标样本标签输入预训练模型，以使目标数据生成模型可以学习多个维度的数据生成规则。

相应地，目标数据生成条件可以包括多个维度的目标特征标签，每个目标特征标签表征一个维度的数据生成条件。如此一来，执行主体可以从多个维度约束其语料生成过程，从而实现了融合多个维度的数据增强。

作为示例，目标数据生成条件可以同时包括意图标签、结构标签、实体标签、技术领域标签，分别表征用户从意图、结构、实体、技术领域等维度对生成的语料的期望。执行主体可以将上述多个特征标签同时输入目标数据生成模型，从以上多个维度约束语料的生成过程，以得到满足用户需求的目标语料。

在一个具体的示例中，用户利用目标生成模型拓展空调领域的语料数据，用户可以根据自身需求将目标数据生成条件设定为：“空调”、

“绿色”、“购买”，其中，“空调”为领域标签，“绿色”为实体标签，“购买”为意图标签，之后，执行主体将上述三个特征标签同时输入目标数据生成模型，以生成目标语料，例如目标语料可以是“我想买一个绿色的空调”、“如何购买绿色的空调”等。表 2 示出了本示例中多维度标签、训练语料以及目标语料的对应关系。

表 2

输入标签	训练语料	目标语料
Inform（空调=价格=1800 元；空调=使用面积=20 平米）	我想买个 1800 元的空调，我家面积 20 平米	有没有 1800 元的空调，适用面积 20 平米
Request（手机=价格=？）	请问这个手机多少钱？	手机价格是多少？
Inform（空调=能效=三级能效）#Select（空调=变频/定频=定频）	这两个三级能效的空调，哪个是变频的？	我想在这些三级能效的空调里找一个变频的
Inform（手机=颜色=绿色；手机=数量=0）	抱歉，我们没有绿色的手机，您还有其他想要的颜色吗？	抱歉，本店绿手机的库存为 0

接下来参考图 4，其示出了用于生成数据的方法的又一个实施例的流程 400。该用于生成数据的方法的流程 400，包括以下步骤：

步骤 401，获取目标训练数据和目标数据生成条件。

步骤 402，将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集。步骤 401、步骤 402 与前述步骤 201、步骤 202 相近，此处不再赘述。

步骤 403，将目标样本语料输入预训练模型，将目标样本标签作为期望输出，训练预训练模型，得到目标数据生成模型。

步骤 404，将目标待识别语料输入目标数据生成模型，得到目标待识别语料的特征标签。

在本实施例中，目标数据生成条件包括目标待识别语料。目标数据生成模型表征语料与标签之间的对应关系。执行主体将目标待识别语料输入目标数据生成模型，识别目标待识别语料的特征，并输出表征目标待识别语料特征的目标特征标签。

步骤 405，将目标待识别语料的特征标签确定为目标数据。

从图 4 中可以看出，本实施例中的用于生成数据的方法的流程 400 体现了通过目标数据生成模型识别语料的特征标签的步骤，对于某些语料数据量较大且只有少部分标记有特征标签的场合，本实施例中的用于生成数据的方法仅需要少量的目标领域的训练数据即可确保识别的准确度，可以更有效地增强数据。

进一步参考图 5，作为对上述各图所示方法的实现，本公开提供了一种用于生成数据的装置的一个实施例，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。

如图 5 所示，本实施例的用于生成数据的装置 500 包括：数据获取单元 501，被配置成获取目标训练数据和目标数据生成条件，目标训练数据包括标记有特征标签的目标领域的语料；样本构建单元 502，被配置成将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集；模型调整单元 503，被配置成基于目标样本集，训练预训练模型，调整预训练模型的参数，得到再次训练后的目标数据生成模型，其中，预训练模型经由如下步骤得到：构建初始模型并基于通用样本集训练初始模型，得到预训练模型；数据生成单元 504，被配置成采用目标数据生成模型，基于目标数据生成条件，生成目标数据。

在本实施例中，模型调整单元 503 被进一步配置成：将目标样本标签输入预训练模型，将目标样本语料作为期望输出，训练预训练模型，得到目标数据生成模型。

在本实施例中，目标数据生成条件包括目标特征标签；以及，数据生成单元 504 被进一步配置成：将目标特征标签输入目标数据生成

模型，得到目标语料；将目标语料确定为目标数据。

在本实施例中，目标特征标签为预先构建的分类模型基于待识别语料估计出的分类标签；以及，数据生成单元 504 还包括数据验证模块，被配置成：将目标语料输入分类模型，得到目标语料的分类标签；响应于确定分类模型的预设标签集中包括目标语料的分类标签，将目标语料确定为目标数据，目标数据用于构建分类模型的训练样本。

在本实施例中，模型调整单元 503 还被进一步配置成：将目标样本语料输入预训练模型，将目标样本标签作为期望输出，训练预训练模型，得到目标数据生成模型。

在本实施例中，目标数据生成条件包括目标待识别语料；以及，数据生成单元 504 还被配置成：将目标待识别语料输入目标数据生成模型，得到目标待识别语料的特征标签；将目标待识别语料的特征标签确定为目标数据。

下面参考图 6，其示出了适于用来实现本公开的实施例的电子设备（例如图 1 中的服务器或终端设备）600 的结构示意图。本公开的实施例中的终端设备可以包括但不限于诸如移动电话、笔记本电脑、数字广播接收器、PDA（个人数字助理）、PAD（平板电脑）等等的移动终端以及诸如数字 TV、台式计算机等等的固定终端。图 6 示出的终端设备仅仅是一个示例，不应对本公开的实施例的功能和使用范围带来任何限制。

如图 6 所示，电子设备 600 可以包括处理装置（例如中央处理器、图形处理器等）601，其可以根据存储在只读存储器（ROM）602 中的程序或者从存储装置 608 加载到随机访问存储器（RAM）603 中的程序而执行各种适当的动作和处理。在 RAM 603 中，还存储有电子设备 600 操作所需的各种程序和数据。处理装置 601、ROM 602 以及 RAM 603 通过总线 604 彼此相连。输入/输出（I/O）接口 605 也连接至总线 604。

通常，以下装置可以连接至 I/O 接口 605：包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置 606；包括例如液晶显示器（LCD）、扬声器、振动器等的输出装置 607；

包括例如磁带、硬盘等的存储装置 608；以及通信装置 609。通信装置 609 可以允许电子设备 600 与其他设备进行无线或有线通信以交换数据。虽然图 6 示出了具有各种装置的电子设备 600，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。图 6 中示出的每个方框可以代表一个装置，也可以根据需要代表多个装置。

特别地，根据本公开的实施例，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，本公开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装置 609 从网络上被下载和安装，或者从存储装置 608 被安装，或者从 ROM 602 被安装。在该计算机程序被处理装置 601 执行时，执行本公开的实施例的方法中限定的上述功能。需要说明的是，本公开的实施例所述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。在本公开的实施例中，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。而在本公开的实施例中，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，该计算机可读信号介质可以发送、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：电线、光缆、RF（射频）等等，或者上述的任意合适的组合。

上述计算机可读介质可以是上述电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。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电子设备执行时，使得该电子设备：获取目标训练数据和目标数据生成条件，目标训练数据包括标记有特征标签的目标领域的语料；将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集；基于目标样本集，训练预训练模型，调整预训练模型的参数，得到再次训练后的目标数据生成模型，其中，预训练模型经由如下步骤得到：构建初始模型并基于通用样本集训练初始模型，得到预训练模型；采用目标数据生成模型，基于目标数据生成条件，生成目标数据。

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公开的实施例的操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—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用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用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）——连接到用户计算机，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）。

附图中的流程图和框图，图示了按照本公开各种实施例的系统、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。

描述于本公开的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：一种处理器包括连数据获取单元、样本构建单元、模型调整单元和数据生成单元。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，数据获取单元还可以被描述为“获取目标训练数据和目标数据生成条件的单元”。

以上描述仅为本公开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，本公开的实施例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。例如上述特征与本公开的实施例中公开的（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。

权 利 要 求 书

1、一种用于生成数据的方法，其中，包括：

获取目标训练数据和目标数据生成条件，所述目标训练数据包括标记有特征标签的目标领域的语料；

将所述目标训练数据中标记有特征标签的语料确定为目标样本语料，将所述目标样本语料的特征标签确定为目标样本标签，得到目标样本集；

基于所述目标样本集，训练预训练模型，调整所述预训练模型的参数，得到再次训练后的目标数据生成模型，其中，所述预训练模型经由如下步骤得到：构建初始模型并基于通用样本集训练所述初始模型，得到所述预训练模型；

采用所述目标数据生成模型，基于所述目标数据生成条件，生成目标数据。

2、根据权利要求 1 所述的方法，其中，基于所述目标样本集，训练预训练模型，包括：将所述目标样本标签输入所述预训练模型，将所述目标样本语料作为期望输出，训练所述预训练模型，得到所述目标数据生成模型。

3、根据权利要求 1-2 任一项所述的方法，其中，所述目标数据生成条件包括目标特征标签；以及，

采用所述目标数据生成模型，基于所述目标数据生成条件，生成目标数据，包括：将所述目标特征标签输入所述目标数据生成模型，得到目标语料；将所述目标语料确定为目标数据。

4、根据权利要求 3 所述的方法，其中，所述目标特征标签为预先构建的分类模型基于待识别语料估计出的分类标签；以及，

在将所述目标语料确定为目标数据之前，所述方法还包括：将所述目标语料输入所述分类模型，得到所述目标语料的分类标签；响应

于确定所述分类模型的预设标签集中包括所述目标语料的分类标签，将所述目标语料确定为目标数据，所述目标数据用于构建所述分类模型的训练样本。

5、根据权利要求 1-4 任一项所述的方法，其中，基于所述目标样本集，训练预训练模型，包括：

将所述目标样本语料输入所述预训练模型，将所述目标样本标签作为期望输出，训练所述预训练模型，得到所述目标数据生成模型。

6、根据权利要求 1-5 任一项所述的方法，其中，所述目标数据生成条件包括目标待识别语料；以及，

采用所述目标数据生成模型，基于所述目标数据生成条件，生成目标数据，包括：将所述目标待识别语料输入所述目标数据生成模型，得到所述目标待识别语料的特征标签；将所述目标待识别语料的特征标签确定为目标数据。

7、一种用于生成数据的装置，其中，包括：

数据获取单元，被配置成获取目标训练数据和目标数据生成条件，所述目标训练数据包括标记有特征标签的目标领域的语料；

样本构建单元，被配置成将所述目标训练数据中标记有特征标签的语料确定为目标样本语料，将所述目标样本语料的特征标签确定为目标样本标签，得到目标样本集；

模型调整单元，被配置成基于所述目标样本集，训练预训练模型，调整所述预训练模型的参数，得到再次训练后的目标数据生成模型，其中，所述预训练模型经由如下步骤得到：构建初始模型并基于通用样本集训练所述初始模型，得到所述预训练模型；

数据生成单元，被配置成采用所述目标数据生成模型，基于所述目标数据生成条件，生成目标数据。

8、根据权利要求 7 所述的装置，其中，所述模型调整单元被进一

步配置成：将所述目标样本标签输入所述预训练模型，将所述目标样本语料作为期望输出，训练所述预训练模型，得到所述目标数据生成模型。

9、根据权利要求 7-8 任一项所述的装置，其中，所述目标数据生成条件包括目标特征标签；以及，

所述数据生成单元被进一步配置成：将所述目标特征标签输入所述目标数据生成模型，得到目标语料；将所述目标语料确定为目标数据。

10、根据权利要求 9 所述的装置，其中，所述目标特征标签为预先构建的分类模型基于待识别语料估计出的分类标签；以及，

所述数据生成单元还包括数据验证模块，被配置成：将所述目标语料输入所述分类模型，得到所述目标语料的分类标签；响应于确定所述分类模型的预设标签集中包括所述目标语料的分类标签，将所述目标语料确定为目标数据，所述目标数据用于构建所述分类模型的训练样本。

11、根据权利要求 7-10 任一项所述的装置，其中，所述模型调整单元还被进一步配置成：将所述目标样本语料输入所述预训练模型，将所述目标样本标签作为期望输出，训练所述预训练模型，得到所述目标数据生成模型。

12、根据权利要求 7-11 所述的装置，其中，所述目标数据生成条件包括目标待识别语料；以及，

所述数据生成单元还被配置成：将所述目标待识别语料输入所述目标数据生成模型，得到所述目标待识别语料的特征标签；将所述目标待识别语料的特征标签确定为目标数据。

13、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 1-6 中任一所述的方法。

14、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 1-6 中任一所述的方法。

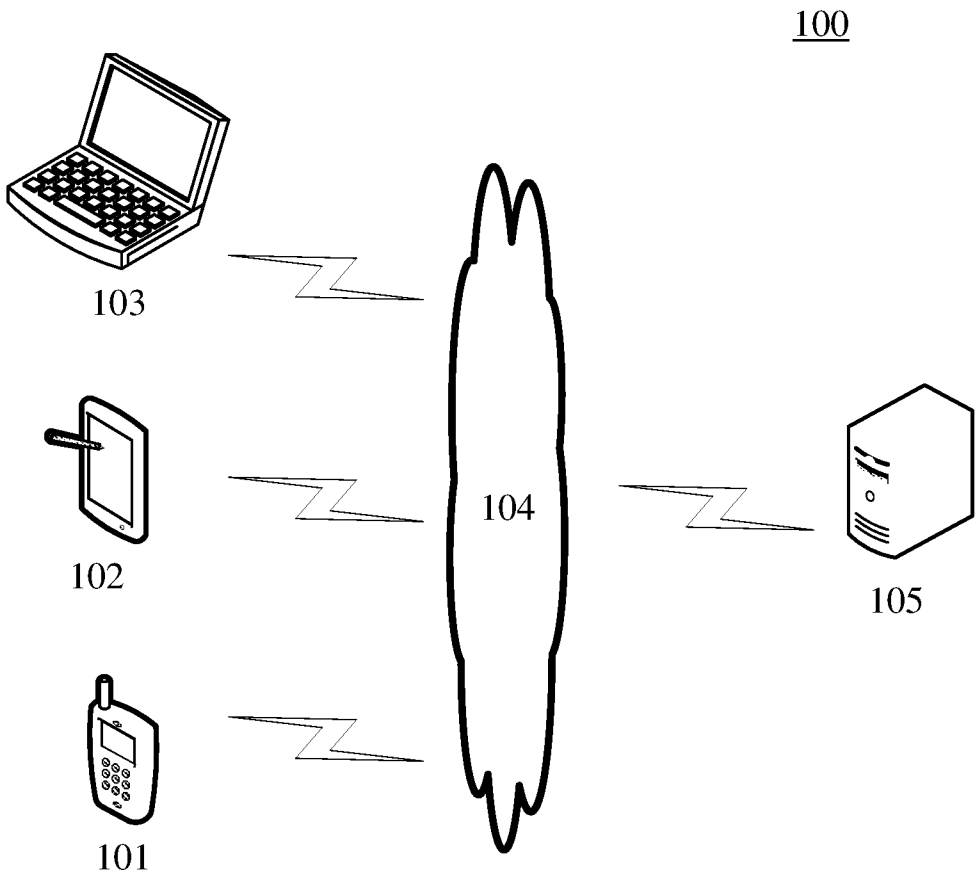


图 1

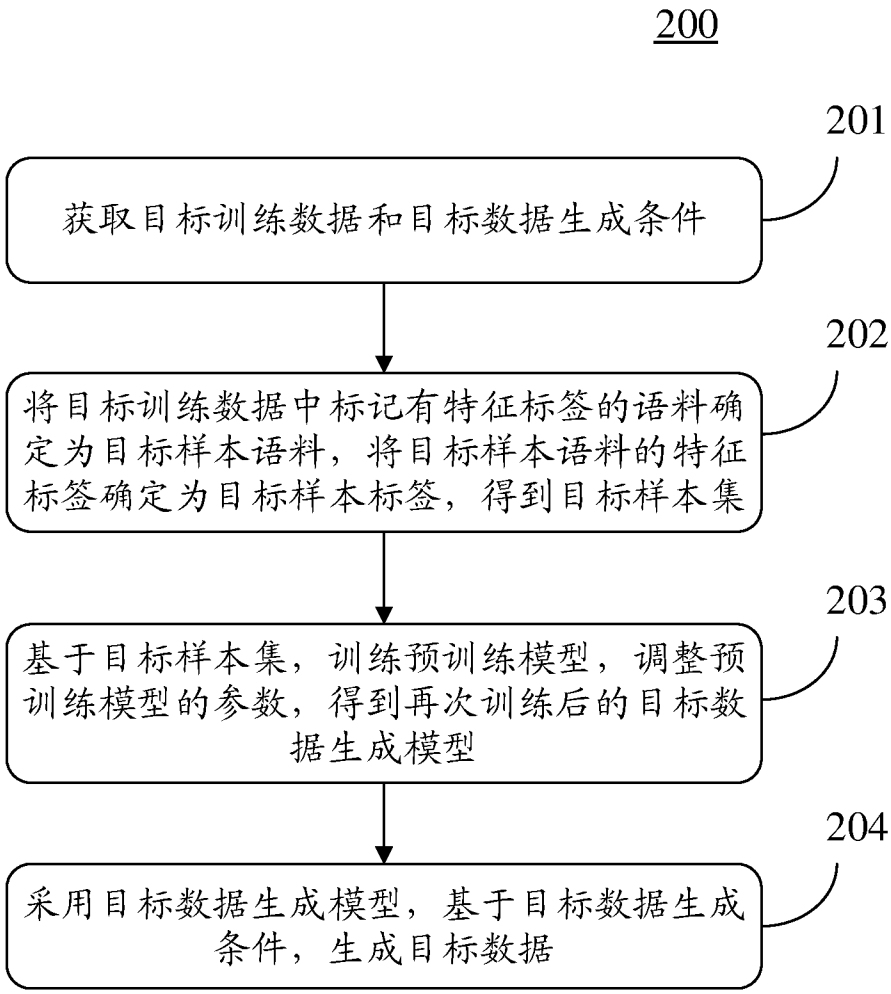

200
201
获取目标训练数据和目标数据生成条件
202
将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集
203
基于目标样本集，训练预训练模型，调整预训练模型的参数，得到再次训练后的目标数据生成模型
204
采用目标数据生成模型，基于目标数据生成条件，生成目标数据

图 2

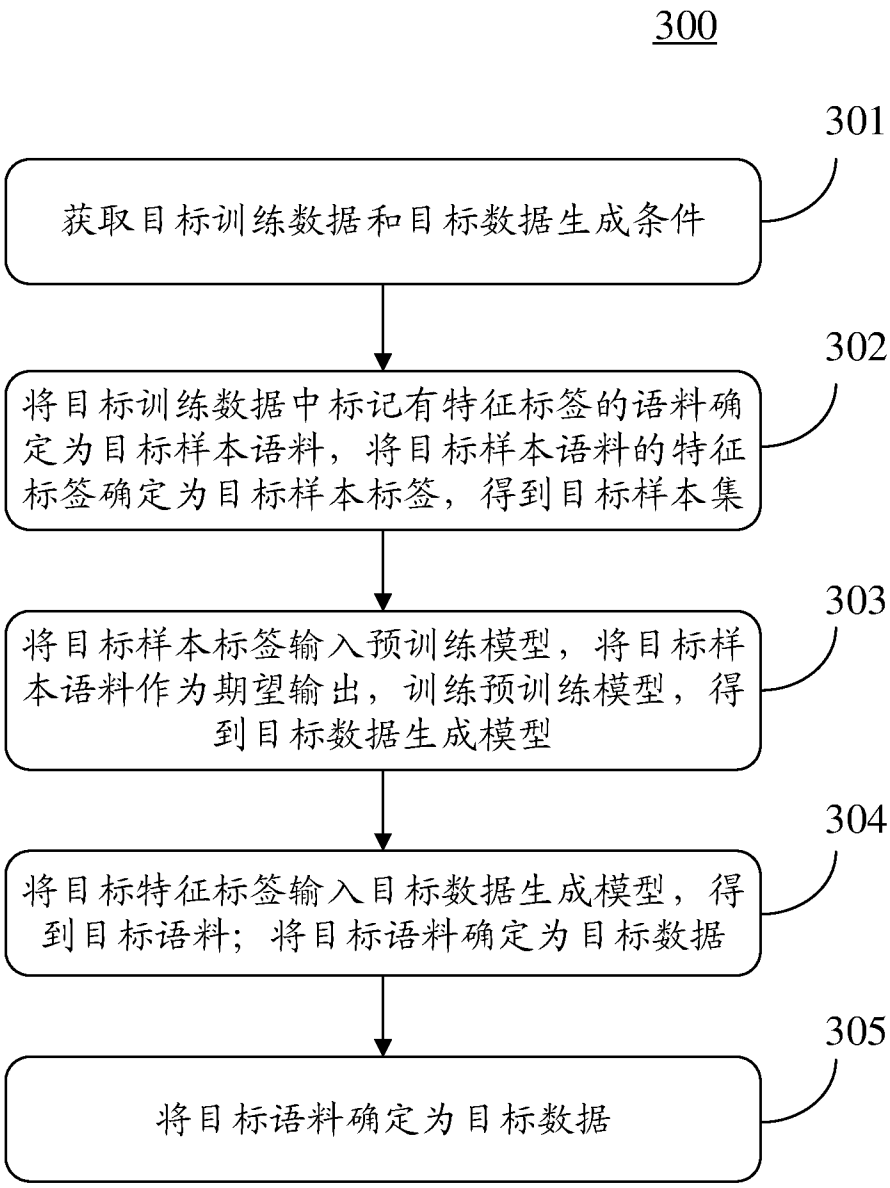

300
301
获取目标训练数据和目标数据生成条件
302
将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集
303
将目标样本标签输入预训练模型，将目标样本语料作为期望输出，训练预训练模型，得到目标数据生成模型
304
将目标特征标签输入目标数据生成模型，得到目标语料；将目标语料确定为目标数据
305
将目标语料确定为目标数据

图 3

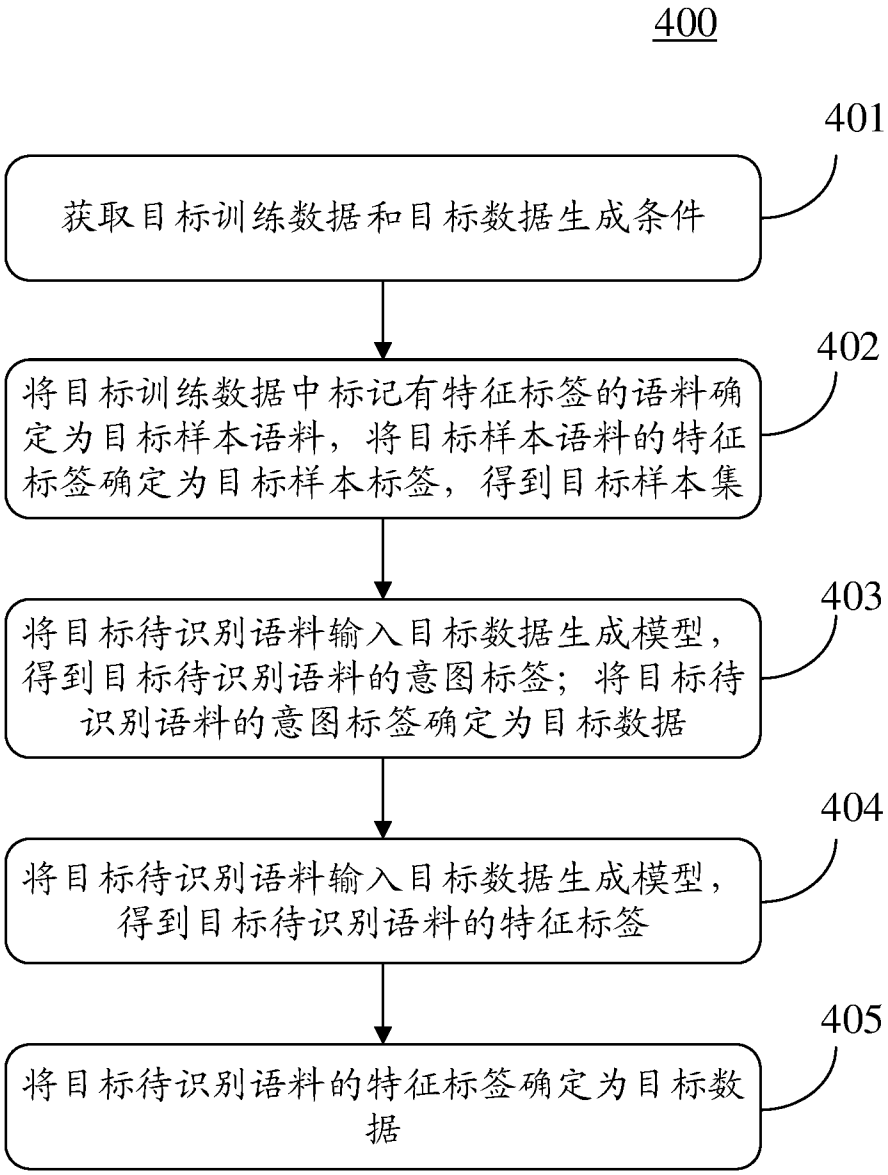

400
401
获取目标训练数据和目标数据生成条件
402
将目标训练数据中标记有特征标签的语料确定为目标样本语料，将目标样本语料的特征标签确定为目标样本标签，得到目标样本集
403
将目标待识别语料输入目标数据生成模型，得到目标待识别语料的意图标签；将目标待识别语料的意图标签确定为目标数据
404
将目标待识别语料输入目标数据生成模型，得到目标待识别语料的特征标签
405
将目标待识别语料的特征标签确定为目标数据

图 4

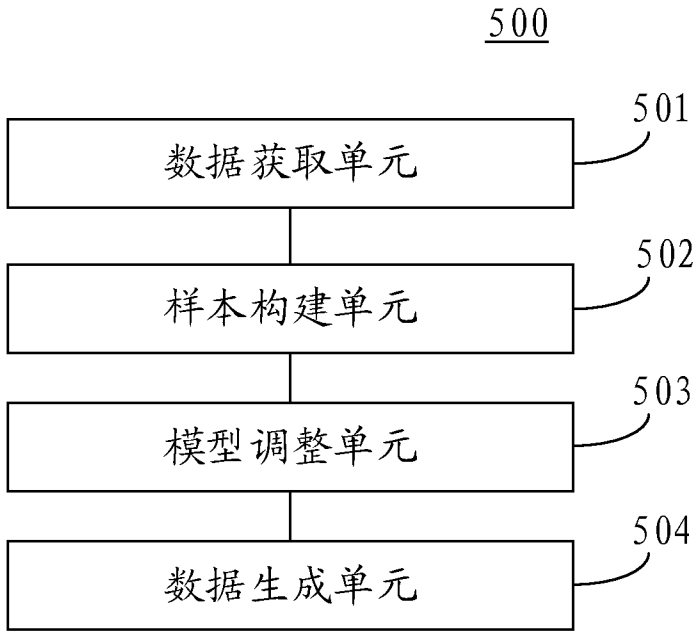

500
501
数据获取单元
502
样本构建单元
503
模型调整单元
504
数据生成单元

图 5

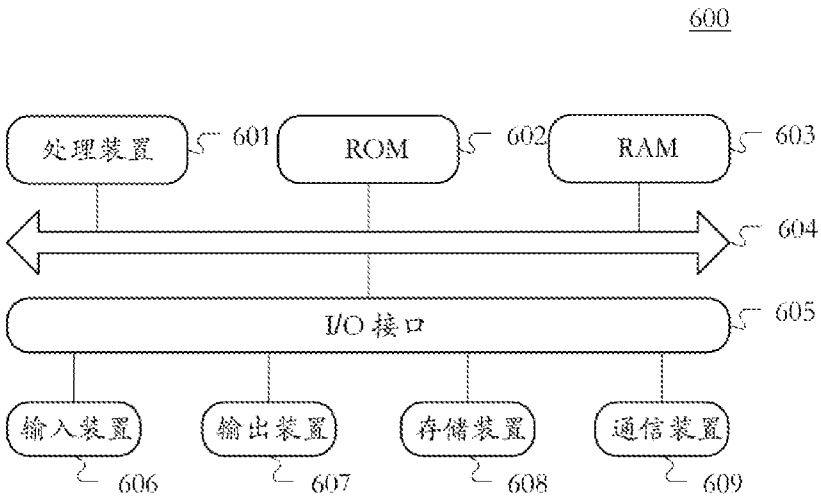

600
处理装置
601
ROM
602
RAM
603
604
I/O 接口
605
输入装置
606
输出装置
607
存储装置
608
通信装置
609

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/070250

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/20(2020.01)i; G06F 16/35(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)

CNKI, CNAPT, EPODOC, WPI, IEEE: 语言, 文本, 数据, 生成模型, 预训练, 通用, 特定领域, 目标领域, 样本, 语料, 分类, 标签, language, text, data, generat+ model, GPT, pre?train+, general, specific domain, sample, corpus, classify, sort, label

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 112182210 A (SICHUAN HUAKONG TIANXING TECHNOLOGY CO., LTD.) 05 January 2021 (2021-01-05) description, paragraphs 60-89	1-14
A	CN 111339278 A (ZHIFUBAO (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) 26 June 2020 (2020-06-26) entire document	1-14
A	CN 111898369 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 06 November 2020 (2020-11-06) entire document	1-14
A	CN 112541346 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 23 March 2021 (2021-03-23) entire document	1-14
A	US 2017358295 A1 (CONDUENT BUSINESS SERVICES, LLC) 14 December 2017 (2017-12-14) entire document	1-14

☐ 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
17 March 2022	**07 April 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/070250

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 112182210 A	05 January 2021	None	
CN 111339278 A	26 June 2020	None	
CN 111898369 A	06 November 2020	None	
CN 112541346 A	23 March 2021	None	
US 2017358295 A1	14 December 2017	None	

国际检索报告

国际申请号
PCT/CN2022/070250

A. 主题的分类

G06F 40/20(2020.01)i; G06F 16/35(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））

CNKI, CNAPT, EPODOC, WPI, IEEE: 语言，文本，数据，生成模型，预训练，通用，特定领域，目标领域，样本，语料，分类，标签，language, text, data, generat+ model, GPT, pre?train+, general, specific domain, sample, corpus, classify, sort, label

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
X	CN 112182210 A (四川华空天行科技有限公司) 2021年1月5日 (2021-01-05) 说明书第60-89段	1-14
A	CN 111339278 A (支付宝杭州信息技术有限公司) 2020年6月26日 (2020-06-26) 全文	1-14
A	CN 111898369 A (腾讯科技深圳有限公司) 2020年11月6日 (2020-11-06) 全文	1-14
A	CN 112541346 A (北京百度网讯科技有限公司) 2021年3月23日 (2021-03-23) 全文	1-14
A	US 2017358295 A1 (CONDUENT BUSINESS SERVICES, LLC) 2017年12月14日 (2017-12-14) 全文	1-14

☐ 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
2022年3月17日	2022年4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 宋朝 电话号码 86-(10)-53961349

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/070250

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 112182210 A	2021年1月5日	无	
CN 111339278 A	2020年6月26日	无	
CN 111898369 A	2020年11月6日	无	
CN 112541346 A	2021年3月23日	无	
US 2017358295 A1	2017年12月14日	无	