(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118235143 A
(43)申请公布日 2024.06.21

(21)申请号 202280071107.8
(22)申请日 2022.11.16
(30)优先权数据
63/280,580 2021.11.17 US
(85)PCT国际申请进入国家阶段日
2024.04.23
(86)PCT国际申请的申请数据
PCT/US2022/050076 2022.11.16
(87)PCT国际申请的公布数据
WO2023/091468 EN 2023.05.25
(71)申请人 甲骨文国际公司
地址 美国
(72)发明人 T·T·乌 T·Q·彭
M·E·约翰逊 T·L·董
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 罗亚男
(51)Int.Cl.
G06N 20/00(2006.01)
G06F 40/279(2006.01)
H04L 51/02(2006.01)

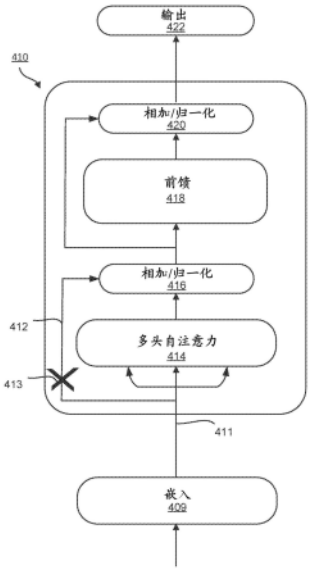
权利要求书3页 说明书33页 附图10页

(54)发明名称

自然语言处理的路径失活

(57)摘要

提供了用于机器学习模型的改进的训练的技术,该机器学习模型包括多个层并且被配置为处理文本语言输入。机器学习模型包括一个或多个块,其中每个块包括多头自注意力网络、用于向多头自注意力网络提供输入的第一连接、以及用于绕过多头自注意力网络向归一化层提供输入的第二(残差)连接;在训练期间,根据失活参数使第二连接失活。附加地或替代地,注意力权重矩阵用于通过阻塞注意力权重矩阵中的对角线项来失活。因此,机器学习模型越来越聚焦于上下文信息,这提供更准确的语言处理结果。



CN 118235143 A

1.一种用于训练机器学习模型以处理音频或文本语言输入的计算机实现的方法,所述方法包括:

访问机器学习模型,所述机器学习模型包括一个或多个块,其中每个块包括多头自注意力网络、用于向所述多头自注意力网络提供输入的第一连接、以及用于绕过所述多头自注意力网络向归一化层提供所述输入的第二连接;

访问训练数据集;

识别失活参数;

将所述失活参数应用于第二连接;

使用所述训练数据集训练所述机器学习模型以生成经过训练的机器学习模型,其中第二连接根据所述失活参数被失活;以及

促进对所述经过训练的机器学习模型的使用。

2.如权利要求1所述的方法,其中:

所述机器学习模型包括多个注意力头,每个注意力头提供被级联并乘以注意力权重矩阵的输出;以及

所述方法还包括使得所述注意力权重矩阵中的对角线项失活。

3.如权利要求1所述的方法,其中所述失活参数是所述机器学习模型的超参数,所述方法还包括:

执行超调优来识别所述失活参数。

4.如权利要求1所述的方法,其中,当第二连接被失活时,所述机器学习模型被致使通过多头自注意力来学习上下文信息。

5.如权利要求4所述的方法,其中:

所述训练数据集包括一个或多个话语,每个话语包括被标记为实体的一个或多个单词以及未被标记为实体的一个或多个单词;以及

所述上下文信息是基于未被标记为实体的所述一个或多个单词来确定的。

6.如权利要求1所述的方法,其中促进对所述经过训练的机器学习模型的使用包括:

将话语输入所述经过训练的机器学习模型以识别实体;以及

基于所识别的实体,准备并提供对所述话语的响应。

7.如权利要求1所述的方法,其中:

所述失活参数为失活率并且第二连接为残差连接;以及

应用所述失活参数包括根据所述失活率使所述残差连接失活。

8.一种系统,包括:

一个或多个处理器;

存储器,耦合到所述一个或多个处理器,所述存储器存储可由所述一个或多个处理器执行的多个指令,所述多个指令包括在由所述一个或多个处理器执行时致使所述一个或多个处理器执行包括以下的操作的指令:

访问机器学习模型,所述机器学习模型包括一个或多个块,其中每个块包括多头自注意力网络、用于向所述多头自注意力网络提供输入的第一连接、以及用于绕过所述多头自注意力网络向归一化层提供所述输入的第二连接;

访问训练数据集;

识别失活参数;

将所述失活参数应用于第二连接;

使用所述训练数据集训练所述机器学习模型以生成经过训练的机器学习模型,其中第二连接根据所述失活参数被失活;以及

促进对所述经过训练的机器学习模型的使用。

9.如权利要求8所述的系统,其中:

所述机器学习模型包括多个注意力头,每个注意力头提供被级联并乘以注意力权重矩阵的输出;以及

所述操作还包括使所述注意力权重矩阵中的对角线项失活。

10.如权利要求8所述的系统,其中:

所述失活参数是所述机器学习模型的超参数,以及

所述操作还包括执行超调优以识别所述失活参数。

11.如权利要求8所述的系统,其中:

当第二连接被失活时,所述机器学习模型被致使通过多头自注意力来学习上下文信息。

12.如权利要求11所述的系统,其中:

所述训练数据集包括一个或多个话语,每个话语包括被标记为实体的一个或多个单词以及未被标记为实体的一个或多个单词;以及

所述上下文信息是基于未被标记为实体的所述一个或多个单词来确定的。

13.如权利要求8所述的系统,其中促进对所述经过训练的机器学习模型的使用包括:

将话语输入所述经过训练的机器学习模型以识别实体;以及

基于所识别的实体,准备并提供对所述话语的响应。

14.如权利要求8所述的系统,其中:

所述失活参数为失活率并且第二连接为残差连接;以及

应用所述失活参数包括根据所述失活率使所述残差连接失活。

15.一种非暂时性机器可读存储器,存储多个指令,所述多个指令可由一个或多个处理器执行以致使所述一个或多个处理器执行包括以下的操作:

访问机器学习模型,所述机器学习模型包括一个或多个块,其中每个块包括多头自注意力网络、用于向所述多头自注意力网络提供输入的第一连接、以及用于绕过所述多头自注意力网络向归一化层提供所述输入的第二连接;

访问训练数据集;

识别失活参数;

将所述失活参数应用于第二连接;

使用所述训练数据集训练所述机器学习模型以生成经过训练的机器学习模型,其中第二连接根据所述失活参数被失活;以及

促进对所述经过训练的机器学习模型的使用。

16.如权利要求15所述的非暂时性计算机可读存储器,其中:

所述机器学习模型包括多个注意力头,每个注意力头提供被级联并乘以注意力权重矩阵的输出;以及

所述操作还包括失活所述注意力权重矩阵中的对角线项。

17.如权利要求15所述的非暂时性计算机可读存储器,其中:

所述失活参数是所述机器学习模型的超参数,以及

所述操作还包括执行超调优以识别所述失活参数。

18.如权利要求15所述的非暂时性计算机可读存储器,其中:

当第二连接被失活时,所述机器学习模型被致使通过多头自注意力来学习上下文信息。

19.如权利要求15所述的非暂时性计算机可读存储器,其中所述操作还包括

所述失活参数为失活率并且第二连接为残差连接;以及

应用所述失活参数包括根据所述失活率使所述残差连接失活。

20.如权利要求15所述的非暂时性计算机可读存储器,其中促进对所述经过训练的机器学习模型的使用包括:

将话语输入所述经过训练的机器学习模型以识别实体;以及

基于所识别的实体,准备并提供对所述话语的响应。

自然语言处理的路径失活

[0001] 对相关申请的交叉引用

[0002] 本申请要求2021年11月17日提交的第63/280,580号美国临时申请的优先权和权益,出于所有目的,该申请的公开内容通过引用整体并入本文。

发明领域

[0003] 本公开一般地涉及人工智能技术,并且更具体地,涉及用于使用路径失活(dropout)来训练用于自然语言处理的机器学习模型的技术。

[0004] 背景

[0005] 人工智能具有许多应用。举例来说,世界各地的许多用户都在即时消息传递或聊天平台上以获得即时反应。组织经常使用这些即时消息传递或聊天平台来与客户(或端用户)进行实时会话。然而,对于组织来说,雇用服务人员与客户或端用户进行实时通信的成本可能非常高。聊天机器人或机器人已经开始被开发来模拟与端用户的会话,尤其是通过互联网。端用户可以通过端用户已安装和使用的消息传递应用程序与机器人通信。智能机器人通常由人工智能(AI)提供动力,可以在实时会话中更智能、更根据上下文地通信,因此可以允许机器人和端用户之间更自然的会话,以用于改进的会话体验。代替端用户学习机器人知道如何响应的一组固定的关键词或命令,智能机器人可能能够基于自然语言的用户话语来理解端用户的意图并相应地进行响应。

[0006] 然而,诸如聊天机器人的基于人工智能的解决方案可能难以构建,因为这些自动化解决方案需要某些领域的具体知识以及某些技术的应用,而这些技术可能仅在专业开发人员的能力范围内。作为构建这样的聊天机器人的一部分,开发人员可能首先了解企业和端用户的需求。然后,开发人员可以分析并做出决策,这些决策例如与选择要被用于分析的数据集、准备用于分析的输入数据集(例如,在分析之前清理数据、提取、格式化和/或转换数据,执行数据特征工程等)、识别用于执行分析的适当的(一个或多个)机器学习(ML)技术或(一个或多个)模型以及改进技术或模型以基于反馈来改进结果/成果相关。识别适当模型的任务可以包括在识别供使用的一个特定模型(或多个模型)之前,可能并行地开发多个模型、迭代地测试和试验这些模型。此外,基于监督学习的解决方案通常涉及训练阶段、随后是应用(即,推断)阶段以及训练阶段和应用阶段之间的迭代循环。开发人员可以负责仔细实现和监控这些阶段,以实现最佳解决方案。例如,为了训练(一个或多个)ML技术或(一个或多个)模型,需要精确的训练数据来使得算法能够理解和学习某些模式或特征(例如,对于聊天机器人——意图提取和仔细的句法分析,而不仅仅是原始语言处理),(一个或多个)ML技术或(一个或多个)模型将使用这些模式或特征来预测所期望的成果(例如,从话语推断意图)。为了确保(一个或多个)ML技术或(一个或多个)模型正确学习这些模式和特征,开发人员可以负责选择、丰富和优化(一个或多个)ML技术或(一个或多个)模型的训练数据集。

发明内容

[0007] 本文公开的技术一般地涉及人工智能技术。更具体地且非限制性地,本文公开的技术涉及用于使用失活来教导机器学习模型聚焦于上下文特征的机器学习模型的改进的训练的技术。

[0008] 在各种实施例中,提供了一种用于训练机器学习模型以处理音频或文本语言输入的计算机实现的方法,该方法包括访问机器学习模型,该机器学习模型包括一个或多个块,其中每个块包括多头自注意力网络、用于向多头自注意力网络提供输入的第一连接、以及用于绕过多头自注意力网络向归一化层提供输入的第二连接;访问训练数据集;识别失活参数;将失活参数应用于第二连接;使用训练数据集训练机器学习模型以生成经过训练的机器学习模型,其中第二连接根据失活参数被失活;以及促进对经过训练的机器学习模型的使用。

[0009] 在一些实施例中,机器学习模型包括多个注意力头,每个注意力头提供被级联(concatenate)并乘以注意力权重矩阵的输出;并且该方法还包括使注意力权重矩阵中的对角线项失活。

[0010] 在一些实施例中,失活参数是机器学习模型的超参数,并且该方法还包括执行超调优以识别失活参数。

[0011] 在一些实施例中,当第二连接被失活时,机器学习模型被致使通过多头自注意力来学习上下文信息。在一些方面,训练数据集包括一个或多个话语,每个话语包括被标记为实体的一个或多个单词以及未被标记为实体的一个或多个单词;并且上下文信息是基于未被标记为实体的一个或多个单词确定的。

[0012] 在一些实施例中,促进对经过训练的机器学习模型的使用包括将话语输入到经过训练的机器学习模型以识别实体;以及基于所识别的实体,准备并提供对话语的响应。

[0013] 在一些实施例中,失活参数是失活率并且第二连接是残差连接,并且应用失活参数包括根据失活率使残差连接失活。

[0014] 在各种实施例中,提供了一种系统,该系统包括一个或多个数据处理器和非暂时性计算机可读存储介质,该非暂时性计算机可读存储介质包含指令,当所述指令在一个或多个数据处理器上执行时,致使该一个或多个数据处理器执行包括以下的操作:访问机器学习模型,该机器学习模型包括一个或多个块,其中每个块包括多头自注意力网络、用于向多头自注意力网络提供输入的第一连接、以及用于绕过多头自注意力网络向归一化层提供输入的第二连接;访问训练数据集;识别失活参数;将失活参数应用于第二连接;使用训练数据集训练机器学习模型以生成经过训练的机器学习模型,其中第二连接根据失活参数被失活;以及促进对经过训练的机器学习模型的使用。

[0015] 在各种实施例中,提供了一种计算机程序产品,该产品有形地体现在非暂时性机器可读存储介质中并且包括指令,该指令可由一个或多个处理器执行以致使该一个或多个处理器执行包括以下的操作:访问机器学习模型,该机器学习模型包括一个或多个块,其中每个块包括多头自注意力网络、用于向多头自注意力网络提供输入的第一连接、以及用于绕过多头自注意力网络向归一化层提供输入的第二连接;访问训练数据集;识别失活参数;将失活参数应用于第二连接;使用训练数据集训练机器学习模型以生成经过训练的机器学习模型,其中第二连接根据失活参数被失活;以及促进对经过训练的机器学习模型的使用。

[0016] 上面和下面描述的技术可以以多种方式并且在多个上下文中实现。参考以下附图提供了若干示例实现和上下文,如下面更详细地描述的。然而,以下实现和上下文只是众多实现和上下文中的一些。

[0017] 附图的简要说明

[0018] 图1是并入示例性实施例的分布式环境的简化框图。

[0019] 图2是根据某些实施例的实现主机器人的计算系统的简化框图。

[0020] 图3是根据某些实施例的实现技能机器人的计算系统的简化框图。

[0021] 图4A是根据各种实施例的聊天机器人训练和部署系统的简化框图。

[0022] 图4B是根据各种实施例的机器学习模型的简化框图。

[0023] 图5A是图示根据某些实施例的用于生成和使用具有路径失活的机器学习模型的过程的流程图。

[0024] 图5B是图示根据一些实施例的具有失活的分词配对的示例的图。

[0025] 图6描绘了用于实现各种实施例的分布式系统的简化图。

[0026] 图7是根据各种实施例的系统环境的一个或多个组件的简化框图,通过其实施例系统的一个或多个组件所提供的服务可以被供应为云服务。

[0027] 图8图示了可用于实现各种实施例的示例计算机系统。

[0028] 详细说明

[0029] 在下面的描述中,出于解释的目的,阐述了具体细节以便提供对某些实施例的透彻理解。然而,将显而易见,可以在没有这些具体细节的情况下实践各种实施例。附图和描述并不旨在是限制性的。本文使用的词语“示例性”意为“用作示例、实例或说明”。本文中描述为“示例性”的任何实施例或设计不一定被解释为比其他实施例或设计优选或有利。

[0030] 介绍

[0031] 人工智能具有许多应用。例如,数字助理是一种帮助用户在自然语言会话中完成各种各样任务的人工智能驱动的接口。对于每个数字助理,客户可以组合一个或多个技能。技能(本文也被描述为聊天机器人、机器人或技能机器人)是聚焦于具体类型任务的个别机器人,诸如跟踪库存、提交考勤卡和创建费用报告。当端用户与数字助理互动时,数字助理评估端用户的输入,并将会话路由到适当的聊天机器人以及从适当的聊天机器人路由。数字助理可以通过各种各样的渠道(诸如FACEBOOK® Messenger、SKYPE MOBILE® messenger或短信服务(SMS))提供给端用户。渠道将聊天在各种消息平台上的端用户和数字助理及其各种聊天机器人之间来回运送聊天。渠道还可以支持用户代理升级、事件发起的会话和测试。

[0032] 聊天机器人处理输入话语以辨别含义。如本文所使用的,话语或消息可以指在与聊天机器人会话期间交换的一组单词(例如,一个或多个句子)。

[0033] 作为由基于人工智能的技术(诸如聊天机器人系统)执行的处理的一部分,命名实体识别可以由命名实体识别器(NER)组件来执行。NER是被训练以识别事物名称的模型,这些名称可以是专有名称(例如,Florida(佛罗里达))、商品(例如,shoes(鞋子))、服务(例如,massage(按摩))等等。NER还可以被训练以识别描述性单词或短语,诸如颜色和大小。因此,本文使用的“命名实体”或“实体”可以包括机器学习模型已经被训练来识别的单词或短语,包括事物,以及与这样的事物相关联的单词或短语。在识别命名实体后,NER可以用命名

实体类型来标记对应的单词或短语。命名实体类型是命名实体的类别。例如,短语New York被用命名实体类型city标记,并且单词Seventy被用命名实体类型temperature标记。NER包括被训练以识别与所识别的单词或短语相对应的命名实体的机器学习模型(例如,神经网络)。

[0034] 构建基于人工智能的系统(诸如可以基于用户话语准确地识别命名实体的聊天机器人系统)是一个具有挑战性的任务,部分由于自然语言的微妙性和模糊性以及输入空间的维度(例如,可能的用户话语,小规模的训练数据)和输出空间的大小(命名实体的数量)。例如,在训练过程中,NER模型可能过度聚焦于实体值本身而不是周围的上下文信息。作为具体示例,在话语“I want to go to Central Park”(我想去中央公园)中,“Central Park”(中央公园)是命名实体,并且话语的其余部分“I want to go to”(我想去)提供了可用于命名实体识别和其他对话处理任务的上下文信息。已经观察到,当使用传统技术执行命名实体识别时,由于对上下文信息的不充分的关注,一些实体被错误分类。

[0035] 当执行命名实体识别时,已经发现模型通常不太关注上下文。例如,对于话语“Pay John$10using Evergreen account”(使用Evergreen帐户向John支付10美元),模型应将“Evergreen”标记为实体类型MerchantType,因为Evergreen后面跟着单词“account”(帐户),这是强烈的上下文信号。说明性结果在表1中示出。

[0036]

查询	预期实体	检测实体
what is the total amount I spent at W H Smiths（我在 W H Smiths 总共花了多少钱）	MERCHANT_NAME: W H Smiths	MERCHANT_NAME: Smiths
$40 for office supplies from Office Depot（从 Office Depot 购买办公用品花费 40 美元）	MERCHANT_NAME: Office Depot	MERCHANT_NAME: Office
Had drink with client at Boba Guys for $9.15 (在 Boba Guys 与客户喝酒花费 9.15 美元)	MERCHANT_NAME: Boba Guys	MERCHANT_NAME: Boba

[0037] 表1-无路径失活的示例NER结果

[0038] 在表1中,第一列是查询或输入话语,并且第二列示出要被识别的目标命名实体。第三列示出了使用先前技术(例如,无本文描述的失活技术)生成的不正确的命名实体标签的示例。例如,当目标命名实体是商家名称“Home Depot”时,由于对相关上下文数据的关注的缺乏,商家名称被错误地识别为Depot。可以解释模型为何表现得像这样的一个原因是它在训练时间期间过于聚焦于词汇特征。

[0039] 在一些实施例中,为了致使模型更多地聚焦于上下文信息,路径失活被实现。路径失活可以通过失活从单词/子词嵌入到其表示输出的路径来实现。更具体地,系统将失活应用于(a)图4B中所示的第二连接412和/或(b)注意力权重矩阵中的对角线项。通过这样做,从单词/子词输入嵌入到其输出表示的残差路径被阻塞。因此,模型将需要依赖上下文信息(例如,通过多头自注意力网络,如本文所述)。

[0040] 本公开的失活技术用于训练机器学习模型以更多地聚焦于上下文信息,这提供了改进的结果。机器学习模型可以被训练以执行命名实体识别或其他自然语言处理任务。失

活涉及随机失活或忽略某些层输出和/或连接。在一些实施例中,机器学习模型中的特定残差连接被失活以强制上下文学习。

[0041] 在一些实施例中,机器学习模型是或者包括变换器块,其可以是预训练的语言模型的一部分。变换器块包括多头自注意力网络。多头自注意力网络经由(例如,到之前的编码器的)第一连接接收输入嵌入。机器学习模型包括残差连接,该残差连接用于绕过多头自注意力网络向归一化层提供输入嵌入。路径失活被应用以使残差连接失活,这增加了自注意力将被应用以辨别上下文信息的可能性。

[0042] 在一些实施例中,失活被进一步应用于注意力权重矩阵,这进一步强制上下文学习并改进命名实体识别性能。例如,机器学习模型包括多个注意力头,每个注意力头提供被级联并乘以注意力权重矩阵的输出。注意力权重矩阵中所选择的项被随机失活。在一些方面,注意力权重矩阵中的对角线项被失活,这对应于忽略将分词(token)与其本身配对的项。相反,将分词与其他上下文分词配对的项被保留,这也致使模型越来越聚焦于上下文信息。

[0043] 在一些实施例中,使用失活参数来应用失活,该失活参数是机器学习模型的超参数。超调优过程可以被执行以选择适当的失活率,该失活率控制路径被失活的频率。

[0044] 机器人和分析系统

[0045] 机器人(也称为技能、聊天机器人、闲聊机器人或谈话机器人)是可以与端用户执行会话的计算机程序。机器人通常可以通过使用自然语言消息的消息传递应用来响应自然语言消息(例如,问题或评论)。企业可以使用一个或多个机器人系统通过消息传递应用与端用户进行通信。可以被称为渠道的消息传递应用可以是端用户已经安装并且熟悉的端用户优选的消息传递应用。因此,端用户无需下载和安装新应用以便与机器人系统聊天。消息传递应用可以包括例如过顶(OTT)消息传递渠道(诸如Facebook Messenger、Facebook WhatsApp、微信、Line、Kik、Telegram、Talk、Skype、Slack或SMS)、虚拟私人助理(诸如Amazon Dot、Echo或Show、Google Home、Apple HomePod等)、扩展具有聊天能力的本机或混合/响应式移动应用程序或web应用的移动和web应用程序扩展,或基于语音的输入(诸如具有使用Siri、Cortana、Google Voice或其他语音输入以用于交互的接口的设备或应用程序)。

[0046] 在一些示例中,机器人系统可以与统一资源标识符(URI)相关联。URI可以使用字符串来标识机器人系统。URI可用作一个或多个消息传递应用系统的网络钩子。URI可以包括例如统一资源定位符(URL)或统一资源名称(URN)。机器人系统可以被设计成从消息传递应用系统接收消息(例如,超文本传输协议(HTTP)post调用消息)。HTTP post调用消息可以从消息传递应用系统定向到URI。在一些实施例中,消息可以不同于HTTP post调用消息。例如,机器人系统可以从短消息服务(SMS)接收消息。虽然本文的讨论可以指的是机器人系统作为消息接收的通信,但是应当理解,消息可以是HTTP post调用消息、SMS消息或两个系统之间的任何其他类型的通信。

[0047] 端用户可以通过会话交互(有时被称为会话用户界面(UI))与机器人系统交互,正如人之间的交互一样。在一些情况下,交互可能包括端用户向机器人说“Hello”并且机器人用“Hi”响应并询问端用户它可以如何帮助。在一些情况下,交互也可以是与例如银行机器人的交易交互,诸如将钱从一个账户转移到另一账户;与例如HR机器人的信息交互,诸如检

查假期余额;或与例如零售机器人的交互,诸如讨论退回购买的商品或寻求技术支持。

[0048] 在一些实施例中,机器人系统可以在没有与机器人系统的管理员或开发人员的交互情况下智能地处理端用户交互。例如,端用户可以向机器人系统发送一个或多个消息以便实现期望的目标。消息可以包括某些内容,诸如文本、表情符号、音频、图像、视频或传达消息的其他方法。在一些实施例中,机器人系统可以将内容转换成标准化形式(例如,具有适当参数的针对企业服务的表征状态转移(REST)调用)并生成自然语言响应。机器人系统还可以提示端用户附加的输入参数或请求其他附加信息。在一些实施例中,机器人系统还可以发起与端用户的通信,而不是被动地响应端用户话语。本文描述了用于识别机器人系统的显式调用并确定正在被调用的机器人系统的输入的各种技术。在某些实施例中,显式调用分析由主机器人基于检测话语中的调用名称来执行。响应于对调用名称的检测,话语可以被精炼以用于到与调用名称相关联的技能机器人的输入。

[0049] 与机器人的会话可以遵循包括多个状态的具体会话流。该流可以基于输入定义接下来将发生什么。在一些实施例中,包括用户定义的状态(例如,端用户意图)和在这些状态中或从一个状态到另一个状态采取的动作的状态机可以用于实现机器人系统。会话可能基于端用户输入来采取不同的路径,这可能影响机器人针对流做出的决策。例如,在每个状态下,基于端用户输入或话语,机器人可以确定端用户的意图以便确定要采取的适当的下一步动作。如本文中以及在话语的上下文中所使用的,术语“意图”是指提供话语的用户的意图。例如,用户可能旨在使机器人参与用于订购披萨的会话,使得用户的意图可以通过话语“订购披萨”来表示。用户意图可以针对用户希望聊天机器人代表用户执行的特定任务。因此,话语可以被表述为反映用户的意图的问题、命令、请求等。意图可以包括端用户想要完成的目标。

[0050] 在聊天机器人的配置的上下文中,术语“意图”在本文中用于指代用于将用户的话语映射到聊天机器人可以执行的具体任务/动作或任务/动作类别的配置信息。为了区分话语的意图(即,用户意图)和聊天机器人的意图,后者在本文中有时被称为“机器人意图”。机器人意图可以包括与意图相关联的一组一个或多个话语。例如,订购披萨的意图可以具有表达订购披萨的期望的各种话语排列。这些相关联的话语可被用于训练聊天机器人的意图分类器,以使得意图分类器能够随后确定来自用户的输入话语是否与订购披萨意图匹配。机器人意图可以与用于在某种状态下开始与用户的会话的一个或多个对话流相关联。例如,订购披萨意图的第一消息可以是问题“What kind of pizza would you like?(您想要哪种披萨?)”。除了相关联的话语之外,机器人意图还可以包括与意图相关的命名实体。例如,订购披萨意图可以包括用于执行订购披萨任务的变量或参数,例如,配料1、配料2、披萨类型、披萨大小、披萨数量等。实体的值通常是通过与用户会话获得的。

[0051] 图1是根据某些实施例的并入聊天机器人系统的环境100的简化框图。环境100包括数字助理构建器平台(DABP)102,该平台使得DABP 102的用户能够创建和部署数字助理或聊天机器人系统。DABP 102可用于创建一个或多个数字助理(或DA)或聊天机器人系统。例如,如图1所示,表示特定企业的用户104可以使用DABP 102来为特定企业的用户创建和部署数字助理106。例如,DABP 102可以由银行用来创建一个或多个数字助理以供银行的客户使用。相同的DABP 102平台可以由多个企业用来创建数字助理。作为另一示例,餐馆(例如,披萨店)的所有者可以使用DABP 102来创建和部署使得餐馆的顾客能够订购食物(例

如,订购披萨)的数字助理。

[0052] 出于本公开的目的,“数字助理”是帮助数字助理的用户通过自然语言会话完成各种任务的实体。数字助理可以仅使用软件(例如,数字助理是使用由一个或多个处理器可执行的程序、代码或指令实现的数字实体)、使用硬件或使用硬件和软件的组合来被实现。数字助理可以体现或实现在各种物理系统或设备中,例如在计算机、移动电话、手表、电器、车辆等中。数字助理有时也被称为聊天机器人系统。因此,出于本公开的目的,术语数字助理和聊天机器人系统是可互换的。

[0053] 诸如使用DABP 102构建的数字助理106的数字助理可以用于经由数字助理与其用户108之间基于自然语言的会话来执行各种任务。作为会话的一部分,用户可以向数字助理106提供一个或多个用户输入110并从数字助理106得到响应112。会话可以包括输入110和响应112中的一个或多个。经由这些会话,用户可以请求由数字助理106执行的一个或多个任务,并且作为响应,数字助理被配置为执行用户请求的任务并用适当的响应来响应用户。

[0054] 用户输入110通常是自然语言形式并且被称为话语。用户话语110可以是文本形式,诸如当用户键入句子、问题、文本片段、或者甚至单个单词并将其作为输入提供给数字助理106时。在一些实施例中,用户话语110可以是音频输入或语音形式,诸如当用户说出或讲出作为输入提供给数字助理106的某事时。话语通常是用户108所讲的语言。例如,话语可以是英语,或其他一些语言。当话语是语音形式时,语音输入被转换为该特定语言的文本形式话语并且文本话语然后由数字助理106处理。各种语音到文本处理技术可以被用来将语音或音频输入转换到文本话语,然后该文本话语由数字助理106处理。在一些实施例中,语音到文本转换可以由数字助理106本身完成。

[0055] 可以是文本话语或语音话语的话语,可以是片段、句子、多个句子、一个或多个单词、一个或多个问题、上述类型的组合等。数字助理106被配置为将自然语言理解(NLU)技术应用于话语以理解用户输入的含义。作为话语的NLU处理的一部分,数字助理106被配置为执行处理以理解话语的含义,这涉及识别与话语相对应的一个或多个意图以及一个或多个实体。在理解话语的含义后,数字助理106可以响应于所理解的含义或意图来执行一个或多个动作或操作。出于本公开的目的,假设话语是已经由数字助理106的用户108直接提供的文本话语或者是输入语音话语到文本形式的转换的结果。然而,这并不旨在以任何方式进行限制或约束。

[0056] 例如,用户108输入可以通过提供诸如“I want to order a pizza(我想订购披萨)”的话语来请求订购披萨。当接收到这样的话语后,数字助理106被配置为理解话语的含义并采取适当的动作。适当的动作可以涉及例如用请求用户输入关于用户期望订购的披萨类型、披萨的大小、披萨的任何配料等的问题来响应用户。由数字助理106提供的响应也可以是自然语言形式并且通常是与输入话语相同的语言。作为生成这些响应的一部分,数字助理106可以执行自然语言生成(NLG)。对于订购披萨的用户,经由用户和数字助理106之间的会话,数字助理可以引导用户提供用于披萨订购的所有必要信息,然后在会话结束致使披萨被订购。数字助理106可以通过向用户输出指示披萨已被订购的信息来结束会话。

[0057] 在概念级别上,数字助理106响应于从用户接收到的话语来执行各种处理。在一些实施例中,此处理涉及一系列或流水线的处理步骤,包括例如理解输入话语的含义(有时称为自然语言理解(NLU)、确定响应于该话语而要执行的动作、在适当的情况下致使该动作被

执行、生成响应于用户话语而要输出到用户的响应、将该响应输出到用户等等。NLU处理可以包括解析接收到的输入话语以理解话语的结构和含义,精炼和改造话语以开发可更好理解的话语的形式(例如,逻辑形式)或结构。生成响应可以包括使用NLG技术。

[0058] 由数字助理(诸如数字助理106)执行的NLU处理可以包括各种NLP相关处理,诸如句子解析(例如,分词化、词形还原、识别句子的词性标签、识别句子中的命名实体、生成依赖树以表示句子结构、将句子分割为子句、分析个别子句、解析指代、执行分块等)。在某些实施例中,NLU处理或其部分由数字助理106本身执行。在一些其他实施例中,数字助理106可以使用其他资源来执行NLU处理的部分。例如,输入话语句子的语法和结构可以通过使用解析器、词性标记器和/或命名实体识别器处理句子来识别。在一个实现中,对于英语语言,解析器、词性标记器和命名实体识别器(诸如由斯坦福自然语言处理(NLP)组提供的识别器)被用于分析句子结构和语法。这些作为斯坦福CoreNLP工具包的一部分被提供的。

[0059] 虽然本公开中提供的各种示例示出了英语语言的话语,但这仅意味着作为示例。在某些实施例中,数字助理106还能够处理英语以外的语言的话语。数字助理106可以提供被配置为用于执行不同语言的处理的子系统(例如,实现NLU功能的组件)。这些子系统可以被实现为可以使用来自NLU核心服务器的服务调用来被调用的可插入单元。这使得NLU处理对于每个语言都灵活且可扩展,包括允许不同的处理次序。可以为个别语言提供语言包,其中语言包可以注册可以从NLU核心服务器提供服务的子系统的列表。

[0060] 诸如图1中描绘的数字助理106的数字助理,可以通过各种各样的不同渠道(诸如但不限于,经由某些应用、经由社交媒体平台、经由各种消息传递服务和应用、以及其他应用或渠道)使其用户108可用或可访问。单个数字助理可以具有为其配置的若干渠道,以便它可以同时在不同的服务上运行并被不同的服务访问。

[0061] 数字助理或聊天机器人系统通常包含一个或多个技能或与一个或多个技能相关联。在某些实施例中,这些技能是个别的聊天机器人(被称为技能机器人),其被配置为与用户交互并完成具体类型的任务,诸如跟踪库存、提交考勤卡、创建费用报告、订购食物、检查银行账户、做出预订、购买小部件等。例如,对于图1中描绘的实施例,数字助理或聊天机器人系统106包括技能116-1、116-2等等。出于本公开的目的,术语“技能”和“多个技能”分别与术语“技能机器人”和“多个技能机器人”同义地使用。

[0062] 与数字助理相关联的每个技能帮助数字助理的用户通过与用户的会话来完成任务,其中会话可以包括由用户提供的文本或音频输入以及由技能机器人提供的响应的组合。这些响应可以是以给用户的文本或音频消息的形式和/或使用呈现给用户以用于用户做出选择的简单用户界面元素(例如,选择列表)。

[0063] 存在可以将技能或技能机器人关联或添加到数字助理的各种方式。在一些情况下,技能机器人可以由企业开发,然后使用DABP 102被添加到数字助理。在其他情况下,技能机器人可以使用DABP 102来开发和创建,然后添加到使用DABP 102创建的数字助理。在又另一些情况下,DABP 102提供在线数字商店(被称为“技能商店”),其提供针对广泛任务的多种技能。通过技能商店供应的技能还可以公开各种云服务。为了将技能添加到正在使用DABP 102而生成的数字助理,DABP 102的用户可以经由DABP 102访问技能商店,选择期望的技能,并且指示所选择的技能将被添加到使用DABP 102创建的数字助理。来自技能商店的技能可以按原样或以修改的形式被添加到数字助理(例如,DABP 102的用户可以选择

并克隆由技能商店提供的特定技能机器人、对所选择的技能机器人做出定制或修改,然后将被修改的技能机器人添加到使用DABP 102创建的数字助理中)。

[0064] 各种不同的架构可以被用来实现数字助理或聊天机器人系统。例如,在某些实施例中,使用DABP 102创建和部署的数字助理可以使用主机器人/子(或次)机器人范例或架构来实现。根据此范例,数字助理被实现为与作为技能机器人的一个或多个子机器人交互的主机器人。例如,在图1所描绘的实施例中,数字助理106包括主机器人114和作为主机器人114的子机器人的技能机器人116-1、116-2等。在某些实施例中,数字助理106本身被认为充当主机器人。

[0065] 根据主-子机器人架构实现的数字助理使得数字助理的用户能够通过统一的用户接口(即经由主机器人)与多个技能交互。当用户与数字助理互动时,用户输入由主机器人接收。然后,主机器人执行处理以确定用户输入话语的含义。然后,主机器人确定用户在话语中请求的任务是否可以由主机器人本身处理,否则主机器人选择适当的技能机器人以用于处理用户请求并将会话路由到所选择的技能机器人。这使得用户能够通过通用的单个接口与数字助理交谈,并且仍然提供使用被配置为执行具体任务的若干技能机器人的能力。例如,对于为企业开发的数字助理,数字助理的主机器人可以与具有具体功能的技能机器人(诸如用于执行与客户关系管理(CRM)相关的功能的CRM机器人、用于执行与企业资源规划(ERP)相关的功能的ERP机器人、用于执行与人力资本管理(HCM)相关的功能的HCM机器人等)接口。这样,数字助理的端用户或消费者只需要知道如何通过通用的主机器人接口访问数字助理,并且在幕后多个技能机器人被提供用于处理用户请求。

[0066] 在某些实施例中,在主机器人/子机器人基础设施中,主机器人被配置为知道技能机器人的可用列表。主机器人可以访问元数据,该元数据标识各种可用技能机器人,以及对于每个技能机器人,标识技能机器人的能力,包括可以由技能机器人执行的任务。在接收到话语形式的用户请求后,主机器人被配置为从多个可用技能机器人中识别或预测能够最好地服务或处理用户请求的具体技能机器人。然后,主机器人将话语(或话语的一部分)路由到该具体技能机器人以用于进一步处理。因此,控制从主机器人流向技能机器人。主机器人可以支持多个输入和输出渠道。在某些实施例中,可以借助于由一个或多个可用技能机器人执行的处理来执行路由。例如,如下面讨论的,技能机器人可以被训练来推断话语的意图并且确定所推断的意图是否与技能机器人被配置的意图匹配。因此,由主机器人执行的路由可以涉及技能机器人向主机器人传送对技能机器人是否已经被配置有适合于处理话语的意图的指示。

[0067] 虽然图1中的实施例图1示出了包括主机器人114和技能机器人116-1、116-2和116-3的数字助理106,但这并不旨在进行限制。数字助理可以包括提供数字助理的功能的各种其他组件(例如,其他系统和子系统)。这些系统和子系统可以仅以软件(例如,存储在计算机可读介质上并且可由一个或多个处理器执行的代码、指令)、仅以硬件、或以使用软件和硬件的组合的实现来实现。

[0068] DABP 102提供使得DABP 102的用户能够创建包括与数字助理相关联的一个或多个技能机器人的基础设施以及各种服务和特征。在一些情况下,技能机器人可以通过克隆现有的技能机器人(例如,克隆技能商店提供的技能机器人)来创建。如之前所指示的,DABP 102提供供应用于执行各种任务的多个技能机器人的技能存储或技能目录。DABP 102的用

户可以从技能商店克隆技能机器人。根据需要,可以对克隆的技能机器人做出修改或定制。在一些其他情况下,DABP 102的用户使用DABP 102供应的工具和服务从头开始创建技能机器人。如之前所指示的,DABP 102提供的技能存储或技能目录可以供应用于执行各种任务的多个技能机器人。

[0069] 在某些实施例中,在高级别上,创建或定制技能机器人涉及以下步骤:

[0070] (1)为新技能机器人配置设定

[0071] (2)为技能机器人配置一个或多个意图

[0072] (3)为一个或多个意图配置一个或多个实体

[0073] (4)训练技能机器人

[0074] (5)为技能机器人创建对话流

[0075] (6)根据需要向技能机器人添加定制组件

[0076] (7)测试和部署技能机器人

[0077] 下面简要描述上面步骤中的每一个。

[0078] (1)为新技能机器人配置设定——可以为技能机器人配置各种设定。例如,技能机器人设计者可以为正在被创建的技能机器人指定一个或多个调用名称。然后,这些调用名称可以由数字助理的用户用以显式调用技能机器人。例如,用户可以在用户的话语中输入调用名称来显式调用对应的技能机器人。

[0079] (2)为技能机器人配置一个或多个意图以及相关联的示例话语——技能机器人设计者为正在被创建的技能机器人指定一个或多个意图(也被称为机器人意图)。然后基于这些指定的意图训练技能机器人。这些意图表示技能机器人被训练以为输入话语推断的类别或类。在接收到话语后,经过训练的技能机器人推断话语的意图,其中所推断的意图是从被用于训练技能机器人的一组预定义意图中选择的。然后,技能机器人基于为话语推断的意图,采取适当的动作来响应该话语。在一些情况下,技能机器人的意图表示技能机器人可以为数字助理的用户执行的任务。每个意图被给与意图标识符或意图名称。例如,对于为银行训练的技能机器人,为技能机器人指定的意图可以包括“CheckBalance”、“TransferMoney”、“DepositCheck”等。

[0080] 对于为技能机器人定义的每个意图,技能机器人设计者还可以提供代表并说明意图的一个或多个示例话语。这些示例话语旨在表示用户可以出于该意图而输入到技能机器人的话语。例如,对于CheckBalance意图,示例话语可以包括“What's my savings account balance”(我的储蓄账户余额是多少?)、“How much is in my checking account?”(我的支票账户中有多少钱?)、“How much money do I have in my account”(我的账户中有多少钱)等。因此,典型用户话语的各种排列可以被指定为意图的示例话语。

[0081] 意图及其相关联的示例话语被用作训练数据来训练技能机器人。可以使用各种不同的训练技术。作为此训练的结果,生成预测模型,该预测模型被配置为将话语作为输入并输出由预测模型为该话语推断出的意图。在一些情况下,输入话语被提供给意图分析引擎,该意图分析引擎被配置为使用经过训练的模型来预测或推断输入话语的意图。然后,技能机器人可以基于推断的意图采取一个或多个动作。

[0082] (3)为技能机器人的一个或多个意图配置实体——在一些情况下,可能需要附加的上下文来使得技能机器人能够正确地响应用户话语。例如,可能存在用户输入话语在技

能机器人中解析为相同意图的情况。例如,在上面的示例中,话语“What's my savings account balance?”(我的储蓄账户余额是多少?)和“How much is in my checking account?”(我的支票账户中有多少钱?)两者都解析到相同的CheckBalance意图,但这些话语是请求不同事情的不同请求。为了澄清这样的请求,将一个或多个实体添加到意图中。使用银行技能机器人示例,被称为AccountType的实体定义被称为“checking”(支票)和“saving”(储蓄)的值,可以使得技能机器人能够解析用户请求并适当响应。在上面的示例中,虽然话语解析到相同的意图,但与AccountType实体关联的值对于两个话语是不同的。这使得技能机器人能够对于两个话语执行可能不同的动作,尽管它们解析到相同的意图。可以为为技能机器人配置的某些意图指定一个或多个实体。因此,实体用于向意图本身添加上下文。实体有助于更全面地描述意图并使得技能机器人能够完成用户请求。

[0083] 在某些实施例中,存在两种类型的实体:(a)由DABP 102提供的内置实体,以及(2)可以由技能机器人设计者指定的定制实体。内置实体是可与多种机器人一起使用的通用实体。内置实体的示例包括但不限于与时间、日期、地址、数字、电子邮件地址、持续时间、重复时间段、货币、电话号码、URL等相关的实体。定制实体用于更多定制的应用。例如,对于银行技能,AccountType实体可以由技能机器人设计者定义,该实体通过针对关键字(如支票、储蓄和信用卡等)来检查用户输入以使能各种银行交易。

[0084] (4)训练技能机器人——技能机器人被配置成接收话语形式的用户输入,解析或以其他方式处理接收到的输入,并且识别或选择与接收到的用户输入相关的意图。如上面所指示的,必须为此训练技能机器人。在某些实施例中,基于为技能机器人配置的意图以及与意图相关联的示例话语(统称为训练数据)来训练技能机器人,使得技能机器人可以将用户输入话语解析到其配置的意图之一。在某些实施例中,技能机器人使用使用训练数据训练并且允许技能机器人辨别用户所说的内容(或者在一些情况下,正在试图说出的内容)的预测模型。DABP 102提供可由技能机器人设计者用来训练技能机器人的各种不同的训练技术,包括各种基于机器学习的训练技术、基于规则的训练技术和/或其组合。在某些实施例中,训练数据的一部分(例如,80%)被用于训练技能机器人模型,并且另一部分(例如,剩余的20%)被用于测试或验证模型。一旦被训练,经过训练的模型(有时也被称为经过训练的技能机器人)就可以被用于处理和响应用户话语。在某些情况下,用户的话语可能是仅需要单个答案并且无进一步会话的问题。为了处理这样的情况,可以为技能机器人定义Q&A(问答)意图。这使得技能机器人能够输出对用户请求的回复而无需更新对话定义。以与常规意图类似的方式创建Q&A意图。Q&A意图的对话流可能与常规意图的对话流不同。

[0085] (5)为技能机器人创建对话流——为技能机器人指定的对话流描述当响应于接收到的用户输入为技能机器人解析不同意图时,技能机器人如何反应。对话流定义技能机器人将采取的操作或动作,例如,技能机器人如何响应用户话语、技能机器人如何针对输入来提示用户、技能机器人如何返回数据。对话流就像技能机器人遵循的流程图。技能机器人设计者使用诸如Markdown语言的语言来指定对话流。在某些实施例中,被称为OBotML的YAML版本可以被用于为技能机器人指定对话流。技能机器人的对话流定义充当会话本身的模型,让技能机器人设计者编排技能机器人与技能机器人所服务的用户之间的交互。

[0086] 在某些实施例中,技能机器人的对话流定义包含三个部分:

[0087] (a)上下文部分

[0088] (b)默认转换部分

[0089] (c)状态部分

[0090] 上下文部分——技能机器人设计者可以定义在上下文部分中的会话流中使用的变量。可以在上下文部分中命名的其他变量包括但不限于:用于错误处理的变量、用于内置或定制实体的变量、使得技能机器人能够识别并保留用户偏好的用户变量等。

[0091] 默认转换部分——技能机器人的转换可以在对话流状态部分中或默认转换部分中定义。默认转换部分中定义的转换充当回退,并且当在状态内不存在定义的适用的转换或者不能满足触发状态转换所需的条件时被触发。默认转换部分可用于定义路由,该路由允许技能机器人优雅地处理意外的用户操作。

[0092] 状态部分——对话流及其相关操作被定义为一系列暂时性状态,其管理对话流内的逻辑。对话流定义内的每个状态节点命名提供对话中该点处所需的功能的组件。因此,状态是围绕组件构建的。状态包含组件特定的属性,并定义组件执行后被触发的到其他状态的转换。

[0093] 特殊情况场景可以使用状态部分来处理。例如,有时您可能为用户提供暂时离开他们参与的第一技能以在数字助理内以第二技能做某事。例如,如果用户正在参与利用购物技能的会话(例如,用户已经做出一些购买选择),则用户可能想要跳到银行技能(例如,用户可能想要确保他/她有足够的钱用于购买),然后返回到购物技能以完成用户的订单。为了解决此问题,第一技能中的动作可以被配置为发起与相同数字助理中的第二不同技能的交互,然后返回到原始流。

[0094] (6)将定制组件添加到技能机器人——如上所述,在技能机器人的对话流中指定的状态命名提供与状态相对应的所需功能的组件。组件使得技能机器人能够执行功能。在某些实施例中,DABP 102提供用于执行广泛功能的一组预配置组件。技能机器人设计者可以选择这些预配置组件中的一个或多个,并将它们与技能机器人的对话流中的状态相关联。技能机器人设计者还可以使用由DABP 102提供的工具来创建定制的或新的组件,并将定制组件与技能机器人的对话流中的一个或多个状态相关联。

[0095] (7)测试和部署技能机器人——DABP 102提供了使得技能机器人设计者能够测试正在被开发的技能机器人的若干特征。然后可以部署技能机器人并将其包含在数字助理中。

[0096] 虽然上面的描述描述了如何创建技能机器人,但是类似的技术也可以被用来创建数字助理(或主机器人)。在主机器人或数字助理级别,可以为数字助理配置内置系统意图。这些内置系统意图用于识别在不调用与数字助理相关联的技能机器人的情况下数字助理本身(即主机器人)可以处理的一般任务。为主机器人定义的系统意图的示例包括:(1)退出:当用户发出期望退出数字助理中的当前会话或上下文的信号时应用;(2)帮助:当用户请求帮助或指导时应用;(3)UnresolvedIntent:应用于与退出和帮助意图不匹配的用户输入。数字助理还存储关于与数字助理相关联的一个或多个技能机器人的信息。此信息使得主机器人能够选择特定的技能机器人以用于处理话语。

[0097] 在主机器人或数字助理级别,当用户向数字助理输入短语或话语时,数字助理被配置为执行处理以确定如何路由话语和相关对话。数字助理使用路由模型来确定此问题,该模型可以是基于规则、基AI的或其组合。数字助理使用路由模型来确定与用户输入话语

相对应的会话是否要被路由到特定技能以用于处理、是否将由数字助理或主机器人本身根据内置系统意图来处理,或者将作为当前会话流中的不同状态来处理。

[0098] 在某些实施例中,作为此处理的一部分,数字助理确定用户输入话语是否使用其调用名称来显式地识别技能机器人。如果用户输入中存在调用名称,则将其视为对与该调用名称相对应的技能机器人的显式调用。在这样的场景下,数字助理可以将用户输入路由到显式调用的技能机器人以用于进一步处理。如果没有具体或显式的调用,则在某些实施例中,数字助理评估接收到的用户输入话语并计算与数字助理相关联的系统意图和技能机器人的置信度分数。为技能机器人或系统意图计算的分数表示用户输入代表技能机器人被配置为执行的任务或代表系统意图的可能性有多大。具有超过阈值(例如,置信度阈值路由参数)的相关联的计算置信度分数的任何系统意图或技能机器人被选择作为用于进一步评估的候选。然后,数字助理从被识别出的候选者中选择特定的系统意图或技能机器人以用于用户输入话语的进一步处理。在某些实施例中,在将一个或多个技能机器人识别为候选之后,(根据每个技能的意图模型)评估与那些候选技能相关联的意图并且为每个意图确定置信度分数。一般地,具有超过阈值(例如,70%)的置信度分数的任何意图都被视为候选意图。如果选择了特定的技能机器人,则将用户话语路由到该技能机器人以用于进一步处理。如果选择了系统意图,则主机器人本身根据所选择的系统意图执行一个或多个动作。

[0099] 图2是根据某些实施例的主机器人(MB)系统200的简化框图。MB系统200可以仅以软件、仅以硬件或硬件和软件的组合来实现。MB系统200包括预处理子系统210、多意图子系统(MIS)220、显式调用子系统(EIS)230、技能机器人调用器240和数据仓库250。图2中描绘的MB系统200仅仅是主机器人中的组件布置的示例。本领域普通技术人员将认识到许多可能的变化、替代和修改。例如,在一些实现中,MB系统200可以具有比图2所示的系统或组件更多或更少的系统或组件,可以组合两个或多个子系统,或者可以具有不同的子系统配置或布置。

[0100] 预处理子系统210从用户接收话语“A”202,并通过语言检测器212和语言解析器214处理话语。如上面所指示的,可以以包括音频或文本的各种方式提供话语。话语202可以是句子片段、完整句子、多个句子等。话语202可以包括标点。例如,如果话语202被提供为音频,则预处理子系统210可以使用语音到文本转换器(未示出)将音频转换为文本,该转换器将标点符号插入到所得到的文本中,例如逗号、分号、句号等。

[0101] 语言检测器212基于话语202的文本来检测话语202的语言。处理话语202的方式取决于语言,因为每个语言具有其自己的语法和语义。在分析话语的语法和结构时,考虑语言之间的差异。

[0102] 语言解析器214解析话语202以提取话语202中的个别语言单元(例如,单词)的词性(POS)标签。POS标签包括例如名词(NN)、代词(PN)、动词(VB)等。语言解析器214还可以对话语202的语言单元进行分词(例如,以将每个单词转换成单独的分词)并对单词进行词形还原。词元(lemma)是字典中表示的一组单词的主要形式(例如,“run”是run、runs、ran、running等的词元)。语言解析器214可以执行的其他类型的预处理包括复合表达式的分块,例如,将“credit”和“card”组合成单个表达式“credit_card”。语言解析器214还可以识别话语202中的单词之间的关系。例如,在一些实施例中,语言解析器214生成指示话语的哪部分(例如特定名词)是直接宾语、话语的哪部分是介词等的依赖树。由语言解析器214执行的

处理的结果形成提取的信息205并且与话语202本身一起作为输入提供给MIS220。

[0103] 如上所述,话语202可以包括多于一个句子。出于检测多个意图和显式调用的目的,话语202可以被视为单个单元,即使它包括多个句子。然而,在某些实施例中,预处理可以例如由预处理子系统210执行以识别多个句子中的单个句子以用于多个意图分析和显式调用分析。一般地,MIS220和EIS230产生的结果基本上相同而不管话语202是在个别句子的级别上还是作为包括多个句子的单个单元被处理。

[0104] MIS220确定话语202是否表示多个意图。尽管MIS220可以检测话语202中多个意图的存在,但是MIS220执行的处理不涉及确定话语202的意图是否与已经为机器人配置的任何意图匹配。代替地,确定话语202的意图是否与机器人意图匹配的处理可以由MB系统200的意图分类器242或由技能机器人的意图分类器(例如,如图3的实施例中所示)执行。由MIS220执行的处理假设存在可以处理话语202的机器人(例如,特定技能机器人或主机器人本身)。因此,MIS220执行的处理不需要知道聊天机器人系统中有什么机器人(例如,向主机器人注册的技能机器人的身份),也不需要知道已经为特定机器人配置了什么意图。

[0105] 为了确定话语202包括多个意图,MIS220应用来自数据仓库250中的一组规则252中的一个或多个规则。应用于话语202的规则取决于话语202的语言并且可以包括指示多个意图的存在的句子模式。例如,句子模式可以包括连接句子的两个部分(例如,连接副词)的并列连词,其中两个部分对应于单独的意图。如果话语202与句型匹配,则可以推断话语202表示多个意图。应该注意,具有多个意图的话语不一定具有不同的意图(例如,针对不同机器人的意图或相同机器人内的不同意图)。代替地,话语可以具有相同意图的单独实例,例如“Place a pizza order using payment account X,then place a pizza order using payment account Y”(使用支付账户X下披萨订单,然后使用支付账户Y下披萨订单)。

[0106] 作为确定话语202表示多个意图的一部分,MIS220还确定话语202的哪些部分与每个意图相关联。对于包含多个意图的话语中表示的每个意图,MIS220构建用于单独处理的新话语来代替原始话语,例如话语“B”206和话语“C”208,如图2中所描绘的。因此,原始话语202可以被分割成一次被处理一个的两个或更多个单独的话语。MIS220使用提取的信息205和/或从对话语202本身的分析来确定两个或更多个话语中的哪一个应当被首先处理。例如,MIS220可以确定话语202包含指示特定意图应当被首先处理的标记词。对应于此特定意图的新形成的话语(例如,话语206或话语208中的一个)将首先被发送以供EIS230进一步处理。在由第一话语触发的会话已经结束(或者已经暂时暂停)之后,然后下一个最高优先级话语(例如,话语206或话语208中的另一个)可以被发送到EIS230以用于处理。

[0107] EIS230确定其接收的话语(例如,话语206或话语208)是否包含技能机器人的调用名称。在某些实施例中,聊天机器人系统中的每个技能机器人被分配唯一的调用名称,该调用名称将该技能机器人与聊天机器人系统中的其他技能机器人区分开。调用名称的列表可以被维护为数据仓库250中的技能机器人信息254的一部分。当话语包含与调用名称匹配的单词时,该话语被认为是显式调用。如果机器人没有被显式调用,则EIS230接收到的话语被认为是非显式调用话语234,并且被输入到主机器人的意图分类器(例如,意图分类器242)以确定要使用哪个机器人用于处理话语。在一些情况下,意图分类器242将确定主机器人应当处理非显式调用话语。在其他情况下,意图分类器242将确定将话语路由到其以用于处理的技能机器人。

[0108] EIS230提供的显式调用功能具有若干优点。它可以减少主机器人必须执行的处理量。例如,当存在显式调用时,主机器人可能不必做任何意图分类分析(例如,使用意图分类器242),或者可能必须做简化的意图分类分析以用于选择技能机器人。因此,显式调用分析可以在不诉诸意图分类分析的情况下使能对特定技能机器人的选择。

[0109] 此外,可能存在多个技能机器人之间存在功能重叠的情况。例如,如果两个技能机器人处理的意图重叠或彼此非常接近,则这可能发生。在这样的情况下,主机器人可能难以仅基于意图分类分析来识别选择多个技能机器人中的哪一个。在这样的场景下,显式调用可以使要被使用的特定技能机器人消除歧义。

[0110] 除了确定话语是显式调用之外,EIS230还负责确定话语的任何部分是否应当被用作对正在被显式调用的技能机器人的输入。特别地,EIS230可以确定话语的一部分是否与调用不相关联。EIS230可以通过对话语的分析和/或对所提取的信息205的分析来执行此确定。EIS 230可以将与调用不相关联的话语的部分发送到被调用的技能机器人来代替发送由EIS230接收的整个话语。在一些情况下,简单地通过移除与调用相关联的话语的任何部分来形成对所调用的技能机器人的输入。例如,“I want to order pizza using Pizza Bot”(我想使用Pizza Bot订购披萨)可以被缩短为“I want to order pizza”(我想订购披萨),因为“using Pizza Bot”(使用Pizza Bot)与对披萨机器人的调用相关,但与由披萨机器人执行的任何处理无关。在一些情况下,EIS230可以重新格式化要被发送到被调用的机器人的部分,例如以形成完整的句子。因此,EIS230不仅确定存在显式调用,而且确定当存在显式调用时向技能机器人发送什么。在一些情况下,可能没有任何文本输入到正在被调用的机器人。例如,如果话语是“Pizza Bot”,则EIS230可以确定披萨机器人正在被调用,但是没有文本要由披萨机器人处理。在这样的场景中,EIS230可以向技能机器人调用器240指示没有要发送的东西。

[0111] 技能机器人调用器240以各种方式调用技能机器人。例如,技能机器人调用器240可以响应于接收到作为显式调用的结果已经选择了特定技能机器人的指示235来调用机器人。指示235可以由EIS230与显式调用的技能机器人的输入一起发送。在此场景下,技能机器人调用者240将对会话的控制移交给显式调用的技能机器人。显式调用的技能机器人将通过将输入视为独立话语来确定对来自EIS230的输入的适当响应。例如,响应可以是执行具体动作或在特定状态下开始新会话,其中新会话的初始状态取决于从EIS230发送的输入。

[0112] 技能机器人调用器240可以调用技能机器人的另一方式是通过使用意图分类器242的隐式调用。意图分类器242可以使用机器学习和/或基于规则的训练技术来训练以确定话语代表特定技能机器人被配置为执行的任务的可能性。在不同的类别上训练意图分类器242,每个技能机器人一个类别。例如,每当向主机器人注册新技能机器人时,与新技能机器人相关联的示例话语的列表可用于训练意图分类器242以确定特定话语代表该新技能机器人可以执行的任务的可能性。作为此训练的结果产生的参数(例如,机器学习模型的参数的一组值)可以被存储为技能机器人信息254的一部分。

[0113] 在某些实施例中,意图分类器242使用机器学习模型来实现,如本文进一步详细描述的。机器学习模型的训练可以涉及输入来自与各种技能机器人相关联的示例话语的至少一个话语子集,生成关于哪个机器人是用于处理任何特定训练话语的正确机器人的推断,

作为机器学习模型的输出。对于每个训练话语,可以提供用于训练话语的正确机器人的指示作为基本真实信息。然后机器学习模型的行为可以被调整(例如,通过反向传播)以最小化生成的推断和基本真实信息之间的差异。

[0114] 在某些实施例中,意图分类器242针对向主机器人注册的每个技能机器人确定指示技能机器人可以处理话语(例如,从EIS230接收的非显式调用话语234)的可能性的置信度分数。意图分类器242还可以确定已经被配置的每个系统级别意图(例如,帮助、退出)的置信度分数。如果特定置信度分数满足一个或多个条件,则技能机器人调用器240将调用与特定置信度分数相关联的机器人。例如,可能需要满足阈值置信度分数值。因此,意图分类器242的输出245是系统意图的识别或者特定技能机器人的识别。在一些实施例中,除了满足阈值置信度分数值之外,置信度分数值还必须超过下一个最高置信度分数值一定的获胜裕度。当多个技能机器人的置信度分数每一个超过阈值置信度分数值时,施加这样的条件将使能路由到特定技能机器人。

[0115] 在基于置信度分数的评估来识别机器人之后,技能机器人调用器240将处理交接给所识别的机器人。在系统意图的情况下,所识别的机器人是主机器人。否则,所识别的机器人是技能机器人。此外,技能机器人调用器240将确定提供什么作为所识别的机器人的输入247。如上面所指示的,在显式调用的情况下,输入247可以基于与调用不相关联的话语的一部分,或者输入247可以是空(例如,空字符串)。在隐式调用的情况下,输入247可以是整个话语。

[0116] 数据仓库250包括存储由主机器人系统200的各种子系统使用的数据的一个或多个计算设备。如上面所解释的,数据仓库250包括规则252和技能机器人信息254。规则252包括:例如,用于由MIS 220确定话语何时表示多个意图以及如何分割表示多个意图的话语的规则。规则252还包括用于由EIS230确定明确调用技能机器人的话语的哪些部分以发送给技能机器人的规则。技能机器人信息254包括聊天机器人系统中的技能机器人的调用名称,例如向特定主机器人注册的所有技能机器人的调用名称的列表。技能机器人信息254还可以包括由意图分类器242用来确定聊天机器人系统中的每个技能机器人的置信度分数的信息,例如机器学习模型的参数。

[0117] 图3是根据某些实施例的技能机器人系统300的简化框图。技能机器人系统300是可以仅以软件、仅以硬件或硬件和软件的组合来实现的计算系统。在某些实施例中,诸如图1中所描绘的实施例,技能机器人系统300可用来在数字助理内实现一个或多个技能机器人。

[0118] 技能机器人系统300包括MIS 310、意图分类器320和会话管理器330。MIS 310类同于图2中的MIS220并且提供类似的功能,包括可操作来使用数据仓库350中的规则352来确定:(1)话语是否表示多个意图,以及如果是,(2)则如何将该话语分割成多个意图中的每个意图的单独话语。在某些实施例中,由MIS 310应用的用于检测多个意图和用于分割话语的规则与由MIS220应用的规则相同。MIS 310接收话语302和提取的信息304。提取的信息304类同于图1中的提取的信息205并且可以使用语言解析器214或技能机器人系统300本地的语言解析器来生成。

[0119] 意图分类器320可以以与上面结合图2的实施例讨论的意图分类器242类似的方式来训练并如本文进一步详细描述。例如,在某些实施例中,意图分类器320使用机器学习模

型来实现。使用与特定技能机器人相关联的示例话语的至少一个子集作为训练话语来针对特定技能机器人训练意图分类器320。每个训练话语的地面实况将是与训练话语相关联的特定机器人意图。

[0120] 话语302可以直接从用户接收或者通过主机器人供应。当话语302通过主机器人来供应时,例如,作为通过图2中所描绘的实施例中的MIS220和EIS230的处理的结果,MIS 310可以被绕过,以避免重复已经由MIS220执行的处理。然而,如果话语302直接从用户接收到,例如在路由到技能机器人之后发生的会话期间,则MIS 310可以处理话语302以确定话语302是否表示多个意图。如果是,则MIS 310应用一个或多个规则来将话语302分割成每个意图的单独话语,例如话语“D”306和话语“E”308。如果话语302不表示多个意图,则MIS 310将话语302转发到意图分类器320以用于意图分类并且不分割话语302。

[0121] 意图分类器320被配置成将接收到的话语(例如,话语306或308)与和技能机器人系统300相关联的意图相匹配。如上所述,技能机器人可以被配置有一个或多个意图,每个意图包括与意图相关联并用于训练分类器的至少一个示例话语。在图2的实施例中,主机器人系统200的意图分类器242被训练以确定个别技能机器人的置信度分数和系统意图的置信度分数。类似地,意图分类器320可以被训练以确定与技能机器人系统300相关联的每个意图的置信度分数。虽然意图分类器242执行的分类是在机器人级别,但是意图分类器320执行的分类是在意图级别因此是更细粒度的。意图分类器320可以访问意图信息354。对于与技能机器人系统300相关联的每个意图,意图信息354包括代表并说明意图的含义并且通常与该意图执行的任务相关联的话语的列表。意图信息354还可以包括作为在此话语列表上训练的结果而产生的参数。

[0122] 会话管理器330接收由意图分类器320识别为与输入到意图分类器320的话语最佳匹配的特定意图的指示322作为意图分类器320的输出。在一些情况下,意图分类器320无法确定任何匹配。例如,如果话语针对系统意图或不同技能机器人的意图,则由意图分类器320计算的置信度分数可能低于阈值置信度分数值。当这发生时,技能机器人系统300可以将话语提交给主机器人以用于处理,例如以路由到不同的技能机器人。然而,如果意图分类器320成功识别技能机器人内的意图,则会话管理器330将发起与用户的对话。

[0123] 由会话管理器330发起的会话是特定于由意图分类器320识别的意图的会话。例如,会话管理器330可以使用被配置为针对所识别的意图执行对话流的状态机来实现。状态机可以包括默认起始状态(例如,对于当在没有任何附加输入的情况下调用意图时)和一个或多个附加状态,其中每个状态与其要由技能机器人执行的动作(例如,执行购买交易)和/或要呈现给用户的对话(例如,问题、响应)相关联。因此,会话管理器330可以在接收到识别意图的指示322后确定动作/对话335,并且可以响应于在对话期间接收到的后续话语来确定附加动作或对话。

[0124] 数据仓库350包括存储由技能机器人系统300的各种子系统使用的数据的一个或多个计算设备。如图3中所描绘的,数据仓库350包括规则352和意图信息354。在某些实施例中,数据仓库350可以集成到主机器人或数字助理的数据仓库中,例如图2中的数据仓库250。

[0125] 聊天机器人训练和部署系统

[0126] 在一些实施例中,机器学习模型使用路径失活来训练以改进诸如命名实体识别的

语言处理任务的运行时性能。机器学习模型可以在聊天机器人系统中实现,如关于图1、2和3所描述的。

[0127] 图4A是根据各种实施例的聊天机器人训练和部署系统400的简化框图。聊天机器人训练和部署系统400将输入话语402作为输入,该输入话语402可以是与话语相对应的文本或口语语音。输入话语402由分词器404处理以产生分词406。分词406是文本单元,其可以包括由分词器404在输入话语402中识别的单词、单词的部分、字符或单词集。(例如,I-want-to-go-to-Jamaica,其中分词由破折号分开)。

[0128] 分词406被提供作为到编码器408的输入。编码器408可以例如是预训练语言模型(例如,来自变换器的双向编码器表示(BERT))或另一合适的编码器的一部分。BERT是包括编码器组件和解码器组件的基于变换器的网络。(参见Devlin et al.,"BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805(2018))。作为具体示例,编码器408是BERT的底部n(8)层。在一些示例中,编码器包括1个嵌入层和n-1(7)个变换器层。编码器408生成特征嵌入409。特征嵌入409可以是较低维空间中的分词(例如,子单词、单词等)的密集向量表示。特征嵌入409被提供作为机器学习模型410的组件的输入,诸如图4B中个所描绘的。

[0129] 具有失活路径的机器学习模型

[0130] 图4B示出了图示用于执行语言处理任务的机器学习模型410的组件的框图,包括对应于第二连接412的失活路径。机器学习模型410的组件可以是更大的模型或一组模型的一部分,如上面关于图4A和图1-3所描述的。

[0131] 机器学习模型410可以是被配置并用于(例如)识别话语中的一个或多个实体的命名实体识别器模型或者其一部分。在一些方面,机器学习模型410是诸如BERT的预训练语言模型或者是其一部分。在一些示例中,410中所示的组件对应于BERT解码器。

[0132] 在一些实施例中,机器学习模型410包括变换器编码器块或若干变换器编码器块。变换器编码器块的第一主要组件是级联多个自注意力网络的输出的多头自注意力组件。(参见Vaswani et al.,"Attention is All You Need,"31st Conf.on Neural Info.Processing Systs.(NPIS 2017),图2)。将残差连接和归一化应用于多头自注意力组件的输出,然后将其馈送到前馈神经网络,接着是残差连接和归一化步骤。

[0133] 机器学习模型410可以包括多个层、使用多头自注意力技术的多个层和/或包括一个或多个变换器块的多个层。

[0134] 机器学习模型410包括多头自注意力网络414。多头自注意力网络是多个层和多个注意力头的输出的级联。(参见Vaswani,上文)。每个注意力头可以提供被级联的输出。

[0135] 机器学习模型410包括第一连接411。第一连接411提供输入嵌入409(输入话语402的编码表示)作为多头自注意力网络414的输入。输入嵌入409经由第一连接411作为输入被提供给多头自注意力网络414。

[0136] 对于多头自注意力技术,多个头中的每一个使用不同的查询(Query)、键(Key)和值(Value)线性投影来执行自注意力技术。自注意力技术可以将给定的输入嵌入转换为(通过使用查询线性投影生成的)查询表示、(通过使用键线性投影生成的)键表示和(通过使用值线性投影生成的)值表示中的每一个。将查询表示乘以键表示(并潜在地应用激活函数,诸如softmax函数)以生成注意力分数,该分数指示在评估另一位置处的值时对给定位置处

的值给予多少注意力。然后可以将注意力分数乘以值表示以生成输出。这可以通过使用注意力权重矩阵来将注意力分数组织在矩阵中以利用批处理中并行化能力来完成。多头自注意力网络414的输出可以包括与每个分词的依赖性相关联的信息,诸如分词之间的句法或语义关系。(参见Vaswani等人,上文)。

[0137] 多头自注意力网络414的输出被提供给归一化组件416。归一化组件416可以对接收的输入归一化。例如,归一化组件基于输入x和Sublayer(x)(由子层实现的函数)计算并输出层归一化。

[0138] 机器学习模型410还包括第二连接412,其绕过多头自注意力网络414,将输入嵌入409直接传给归一化组件416。在一些实现中,第二连接412是残差连接,其是数据通过跳过一个或多个层到达机器学习模型后面的部分的路径。

[0139] 在一些实施例中,第二连接412被失活(如"X"413所指示)。例如,可以通过在第二连接412之前或之后包括失活层来失活第二连接412。通过失活第二连接412,输入嵌入将越来越多地由多头自注意力网络414处理。

[0140] 在一些实施例中,机器学习模型410还包括前馈组件418和用于相加和归一化前馈组件418的输出的第二归一化组件420。机器学习模型410可以包括通过前馈组件418的第三路径和将第一归一化组件416的输出连接到第二归一化组件420的第四路径。

[0141] 机器学习模型的最终输出422可以例如是与从输入识别的一个或多个单词相对应的命名实体类型。例如,对应于话语"What is the amount I spent at Joe's Shoe Depot"(我在Joe's Shoe Depot花了多少钱)的输出将Joe's Shoe Depot识别为MERCHANT_NAME类型的命名实体。在一些方面,多个机器学习模型410被包括在聊天机器人训练和部署系统400中。输出422可以被提供给下一个机器学习模型,等等,用于诸如变换器块的一组机器学习模型。

[0142] 在聊天机器人实现阶段,机器学习模型410可以与一个或多个其他模型(诸如用于确定话语代表特定技能机器人被配置为执行的任务的可能性的另一模型、用于从第一类型技能机器人的话语预测意图的另一模型、和/或用于从第二类型技能机器人的话语预测意图的另一模型)结合使用。

[0143] 生成和使用具有路径失活的机器学习模型的技术

[0144] 图5A是图示根据某些实施例的用于生成和使用具有路径失活的机器学习模型的过程500的流程图。如图5A中所描绘的处理可以以由相应系统的一个或多个处理单元(例如,处理器、核心)执行的软件(例如,代码、指令、程序)、硬件或其组合来实现。软件可以存储在非暂时性存储介质上(例如,存储器设备上)。图5A中呈现的和下面描述的方法旨在是说明性的且非限制性的。虽然图5A描绘了以特定顺序或次序发生的各种处理步骤,但这并不旨在进行限制。在某些替代实施例中,步骤可以以一些不同的次序执行,或者一些步骤也可以并行执行。

[0145] 在框502处,访问机器学习模型。机器学习模型包括一个或多个块。在一些实施例中,这些块是变换器块,其使用自注意力来处理输入数据,如上面关于图4B所描述的。每个块都包括多头自注意力网络。如上面关于图4B所描述的,每个多头自注意力网络包括被配置为实现多头自注意力技术的多个层。机器学习模型还包括用于向多头自注意力网络提供输入的第一连接(例如,图4B中所示的第一连接411)。机器学习模型还包括用于绕过多头自

注意力网络向归一化层提供输入的第二连接(例如,图4B中所示的第二连接412)。

[0146] 在框504处,访问训练数据集。训练数据集包括成对的输入数据(例如,样本话语)和对应的标签(例如,话语中的命名实体)。

[0147] 训练数据集可以包括与一个或多个技能机器人相关联的示例话语。如上所述,话语是文本格式的。话语可以是句子片段、完整句子、多个句子等。在一些情况下,示例话语由之前或现有的客户或顾客提供。在其他情况下,示例话语是从先前话语库自动生成的(例如,从库中识别特定于聊天机器人被指定学习的技能的话语)。

[0148] 训练数据还可以包括训练标签,其中训练标签中的每一个对应于个别的训练话语。训练标签可以存储在值的矩阵或表中或存储为值的矩阵或表。对于每个训练话语,可以提供对要推断的正确实体及其分类的指示作为用于训练标签的基本真实信息。然后可以(例如,通过反向传播)调整正在被训练的模型的行为以最小化为各种实体生成的推断与基本真实信息之间的差异。

[0149] 在框506处,识别失活参数。在一些实现中,失活参数是失活率。失活参数建立了在特定训练迭代上一些东西(例如,图4的第二连接412,或注意力权重,如下面所描述的)将被失活(即,阻塞或忽略)的可能性。例如,如果第二连接具有失活参数0,则它将根本不被失活,如果第二连接具有失活参数0.31,则第二连接将在训练迭代中的一些训练迭代被失活,如果第二连接具有失活参数0.9,它将在训练迭代中的许多(即,更高百分比的)训练迭代中被失活。

[0150] 在一些实施例中,失活参数是机器学习模型的超参数。超参数是可以被调优或优化以控制机器学习模型的行为的设定。

[0151] 可以通过执行超调优过程来识别失活参数。系统可以执行超参数优化过程来为失活参数选择合适的值。在一些实现中,失活参数使用合适的超参数优化技术(诸如网格搜索、随机搜索或贝叶斯优化)来识别。超参数可以使用超参数调优算法来建立,该算法为这些超参数选择值以优化或最小化反映模型性能的目标函数。模型的性能可以在与训练数据集分开的验证或测试数据集上测量。在合适的超调优过程的示例中,用户首先为每个超参数定义搜索空间(一组可能值)。存在多个过程,其中每个过程将运行一组固定的N个试验,其中N是可配置的。结果保存到共享数据库中。对于每个试验,过程将(使用过去的试验或随机搜索等)选择一组超参数,然后执行完整的端到端训练和评估。试验分数使用评估结果来计算。执行试验后,选择具有最佳分数的试验。

[0152] 在框508处,将失活参数应用于第二连接。根据在框506处识别的失活参数来失活第二连接。在一些实施例中,第二连接是残差连接并且失活参数是失活率。应用失活参数包括根据失活率失活残差连接。基于失活参数,第二连接不被使用并且输入流经第一连接。根据失活率,这可能发生在训练迭代的子集上。

[0153] 替代地或附加地,失活被应用到注意力权重矩阵。注意力权重矩阵由成对的分词的分数组成:模型尝试计算其上下文表示的位置处的分词以及周围的分词(包括其本身)。分词可以是例如单词或其他文本或口语单元。例如,一对句子s_0和s_1的输入表示特征图被匹配而得到注意力权重矩阵A,其中注意力矩阵中的每个单元A_{ij}表示s_0中的第i个分词和s_1中的第j个分词之间的注意力分数。(参见Mahendran Venkatachalam,"Attention in Neural Networks,"Towards Data Science,可在https://towardsdatascience.com/attention-

in-neural-networks-e66920838742(2019))获取)(第i个分词和第j个分词可以属于相同句子,因此“自注意力”中的“自”)。注意力权重矩阵可以包含从每个分词到每个其他分词(包括本身)的注意力权重。例如,对于100个分词,注意力权重矩阵的大小将为100×100。注意力权重矩阵中的对角线项表示从分词到本身的注意力。如果这些对角线项被丢掉,则这进一步迫使模型较少关注分词本身,并且更多地关注周围的分词。图5B图示了分词的配对的示例,如在注意力权重矩阵中。此示例中,失活被应用到对角线项,使得诸如“evergreen”的分词将关于本身具有0注意力分数,如下面关于图5B进一步详细描述的

[0154] 在一些方面,为第二连接识别第一失活参数,并且为注意力权重矩阵识别第二失活参数。取决于(一个或多个)失活参数(例如,如上面所描述的由超调优器选择的),可以在特定训练迭代上失活第二连接和/或注意力权重矩阵中的项。例如,在一次训练迭代上,失活第二连接和注意力权重矩阵中的对角线项两者;在第二训练迭代上,失活第二连接,并且不失活注意力权重矩阵中的对角线项;在第三训练迭代中,失活注意力权重矩阵中的对角线项,但不失活第二连接;等等。

[0155] 在框510处,机器学习模型使用训练数据集来被训练以生成经过训练的机器学习模型。在训练期间,根据在框506处识别并在框508处应用的失活参数来失活第二连接和/或注意力权重矩阵中的对角线项。

[0156] 训练可以包括使用模型的当前版本来处理一个或多个训练数据元素(例如,输入话语或对应的编码)中的每一个以生成预测,基于预测和与数据元素相对应的标签来计算损失,并基于损失更新模型的一个或多个参数。为了生成预测,可以为每个训练数据元素生成多个中间值。例如,中间值可以包括来自模型中的层的输出。每层可以生成多个输出(例如,超过100个、超过1,000个、超过10,000个或超过100,000个输出)。在一些情况下,只要输入相同、模型的超参数相同以及模型的参数相同,模型整体就被配置为生成一致的输出。一旦训练完成,经过训练的机器学习模型就是结果,其可以是实体识别模型,也可以是实体识别模型的一部分。

[0157] 通过应用失活(例如,当第二连接和/或注意力权重矩阵中的项被失活时),在训练过程中,机器学习模型被致使通过多头自注意力来学习上下文信息。如上所述,训练数据集可以包括一个或多个话语。话语可以包括被标记为实体的一个或多个单词(例如,California(加利福尼亚州)、Record Depot(记录库)等)。话语还可以包括未被标记为实体的一个或多个单词(例如,in(在)、the(该)、how(如何)、get(得到)、fast(快)等)。未被标记为实体的那些单词可用于确定上下文信息。例如,在“I need directions to Sydney”(我需要去悉尼的指引)的话语中,单词Sydney将被标记为实体(city),剩余的单词(I、need、directions、to)提供Sydney是地理位置的上下文。此类上下文信息可以帮助模型更准确地识别实体,特别是如果存在上下文可以解析的任何歧义。

[0158] 在框512处,促进对经过训练的机器学习模型的使用。例如,经过训练的机器学习模型被提供输入话语并识别输入话语中的实体。替代地或附加地,经过训练的机器学习模型可用于识别与话语相关联的意图和/或识别对话语的响应。在一些示例中,促进对经过训练的机器学习模型的使用包括将话语输入到经过训练的机器学习模型以识别实体。经过训练的机器学习模型处理输入话语以识别其中的一个或多个实体。执行机器学习模型的系统然后可以输出实体或执行进一步的处理,诸如确定意图、确定响应等。执行机器学习模型的

系统或者接收所识别的实体或其衍生物的下游系统然后可以基于所识别的实体准备并提供对话语的响应。例如,响应可以是回答问题的音频输出、显示的文本、显示的图像等。

[0159] 在实现时,经过训练的机器学习模型可被部署并用作经过训练的实体识别模型的一部分或用作经过训练的实体识别模型,该模型实现一个或多个聊天机器人。例如,一个或多个聊天机器人可以配置有经过训练的机器学习模型,以从一个或多个用户接收(例如,对应于话语的)文本数据,并且从一个或多个聊天机器人接收的各种话语中识别和提取实体。文本数据可以包括在与针对其来训练经过训练的机器学习模型的环境或上下文相对应的环境或上下文中接收的文本数据。实体可以是从文本数据获得的提取信息的一部分,并且可以用在下游处理(诸如意图分类)中。然后,基于所提取的实体生成的输出可以被传送到和/或呈现在与话语源相对应的设备处。

[0160] 图5B是图示根据一些实施例的具有失活的分词配对的示例的图550。示例分词是pay 552、john 554、$556、10 558、using 560、evergreen 562和account 564。在上述注意力权重矩阵中,每个分词将对每个其他分词(包括本身)都有一些确定程度的注意力。这样的配对由两列之间的线表示。

[0161] 如上面关于图5A描述的,在一些实施例中,失活注意力权重矩阵的对角线项。对角线项表示分词与其本身的配对。如图5B所示,当注意力权重矩阵中的对角线项被失活时,诸如evergreen 562的分词保留对话语中的其他分词(例如,本示例中的分词552-560和564)的注意力,这可以提供上下文信息。分词与本身的配对的注意力权重,诸如evergreen 562到evergreen 562,被失活或清零,如“X”580所示。当此失活被应用在训练中时,机器学习模型被致使学习越来越聚焦于上下文信息,诸如周围的形容词、动词等。

[0162] 使用本文描述的技术,通过将失活应用到第二连接和/或注意力权重矩阵的对角线项,机器学习模型被致使通过多头自注意力来学习上下文信息。因此,模型被迫更多地聚焦于上下文信息,这提供了改进的结果。下面的表2图示了具有和没有路径失活的示例结果。示出的值是F1分数,其是精度(正确预测的数量除以模型做出的总预测)和召回率(真阳性数除以真阳性和假阴性的总数)的加权平均。F1分数传达了精度和召回率之间的平衡,并且较高的F1分数指示更好(例如,更精确)的结果。

[0163]

配置	CoNLL2003	MIT-Movie	MIT-Restaurant	ATIS	Multi-WOZ	SNIPS	Wiki-Gold	Ex-pense	PPM	Calen-dar	平均
超调 + train_without_eval	0.892	0.717	0.787	0.957	0.885	0.943	0.769	0.867	0.993	0.825	0.8635

[0164]

+ *正则化*	0.893	0.718	0.793	0.955	0.888	0.946	0.780	0.880	0.993	0.849	0.8695
+ *失活路径*	0.893	0.725	0.793	0.959	0.889	0.948	0.772	0.885	0.993	0.984	0.8841

[0165] 表2-具有失活路径的示例结果

[0166] 说明性系统

[0167] 图6描绘了分布式系统600的简化图。在图示的实施例中,分布式系统600包括经由一个或多个通信网络610耦合到服务器612的一个或多个客户端计算设备602、604、606和608,客户端计算设备602、604、606和608可以被配置为执行一个或多个应用。

[0168] 在各种实施例中,服务器612可以适于运行使能本公开中所描述的一个或多个实施例的一个或多个服务或软件应用。在某些示例中,服务器612还可以提供可以包括非虚拟和虚拟环境的其他服务或软件应用。在一些示例中,这些服务可以作为基于web的服务或云服务,诸如在软件即服务(SaaS)模型下被供应给客户端计算设备602、604、606和/或608的用户。操作客户端计算设备602、604、606和/或608的用户继而可以利用一个或多个客户端应用来与服务器612交互以利用由这些组件提供的服务。

[0169] 在图6中描绘的配置中,服务器612可包括实现由服务器612执行的功能的一个或多个组件618、620和622。这些组件可包括可由一个或多个处理器执行的软件组件、硬件组件或其组合。应当理解,可以不同于分布式系统600的各种不同的系统配置是可能的。因此,图6中所示的示例是用于实现示例系统的分布式系统的一个示例并且不旨在进行限制。

[0170] 用户可以使用客户端计算设备602、604、606和/或608来执行一个或多个应用、模型或聊天机器人,这可以生成然后可以根据本公开的教导来实现或服务的一个或更多个事件或模型。客户端设备可以提供使得客户端设备的用户能够与客户端设备交互的接口。客户端设备还可以经由此接口向用户输出信息。尽管图6仅描绘了四个客户端计算设备,但是可以支持任何数量的客户端计算设备。

[0171] 客户端设备可以包括各种类型的计算系统,诸如便携式手持设备、通用计算机(诸如个人计算机和膝上型计算机)、工作站计算机、可穿戴设备、游戏系统、瘦客户端、各种消息传递设备、传感器或其他感测设备,等等。这些计算设备可以运行各种类型和版本的软件应用和操作系统(例如,Microsoft **Windows®** 、Apple **Macintosh®** 、**UNIX®**或类UNIX操作系统、Linux或类Linux操作系统(诸如Google Chrome™OS)),包括各种移动操作系统(例如,Microsoft Windows **Mobile®** 、 **iOS®** 、Windows **Phone®** 、Android™、**BlackBerry®** 、Palm **OS®**)。便携式手持设备可以包括蜂窝电话、智能手机(例如,**iPhone®**)、平板电脑(例如,**iPad®**)、个人数字助理(PDA)等。可穿戴设备可以包括Google **Glass®** 头戴式显示器和其他设备。游戏系统可以包括各种手持游戏设备、使能了互联网的游戏设备(例如,具有或不具有**Kinect®** 手势输入设备的Microsoft **Xbox®** 游戏控制台、Sony **PlayStation®** 系统、由**Nintendo®** 提供的各种游戏系统等),等等。客户端设备能够执行各种不同的应用,诸如各种互联网相关的应用程序、通信应用(例如,电子邮件应用、短消息服务(SMS)应用),并且可以使用各种通信协议。

[0172] (一个或多个)网络610可以是本领域技术人员熟悉的任何类型的网络,该网络可以支持使用各种各样可用协议中的任一种的数据通信,这些协议包括但不限于TCP/IP(传输控制协议/互联网协议)、SNA(系统网络架构)、IPX(互联网数据包交换)、**AppleTalk®** 等。仅举例来说,(一个或多个)网络610可以是局域网(LAN),基于以太网的网络、令牌环、广

域网(WAN)、互联网、虚拟网络、虚拟专用网(VPN)、内联网、外联网、公共交换电话(PSTN)、红外线网络、无线网络(例如,在任何电气和电子学会(IEEE)1002.11协议套件、Bluetooth®和/或任何其他无线协议中的任一个下操作的网络);和/或这些和/或其他网络的任何组合。

[0173] 服务器612可由一个或多个通用计算机、专用服务器计算机(举例来说,包括PC(个人计算机)服务器、UNIX®服务器、中型服务器、大型计算机、机架式服务器等)、服务器群、服务器集群或任何其他适当的布置和/或组合组成。服务器612可以包括运行虚拟操作系统的一个或多个虚拟机,或者涉及虚拟化的其他计算架构,诸如可以被虚拟化以维护服务器的虚拟存储设备的一个或多个灵活的逻辑存储设备池。在各种示例中,服务器612可以适合于运行提供前述公开中所描述的功能的一个或多个服务或软件应用。

[0174] 服务器612中的计算系统可以运行一个或多个操作系统,包括上面讨论的那些中任一个,以及任何商业上可获得的服务器操作系统。服务器612还可以运行各种各样附加服务器应用和/或中间层应用中的任一个,包括HTTP(超文本传输协议)服务器、FTP(文件传输协议)服务器、CGI(公共网关接口)服务器、JAVA®服务器、数据库服务器等。示例性数据库服务器包括但不限于可从Oracle®、Microsoft®、Sybase®、IBM®(国际商业机器)等商业上可获得的那些。

[0175] 在一些实现中,服务器612可以包括一个或多个应用以分析和整合从客户端计算设备602、604、606和608的用户接收的数据馈送和/或事件更新。作为示例,数据馈送和/或事件更新可包括但不限于Twitter®馈送、Facebook®更新或从一个或多个第三方信息源和连续数据流接收的实时更新,其可包括与传感器数据应用、金融行情、网络性能测量工具(例如,网络监控和流量管理应用)、点击流分析工具、汽车流量监控等相关的实时事件。服务器612还可以包括一个或多个应用以经由客户端计算设备602、604、606和608的一个或多个显示设备来显示数据馈送和/或实时事件。

[0176] 分布式系统600还可包括一个或多个数据储存库614、616。在某些示例中,这些数据储存库可用于存储数据和其他信息。例如,数据储存库614、616中的一个或多个可用于存储信息,例如与在执行根据各种实施例的各种功能时供由服务器612使用的聊天机器人使用的生成模型或聊天机器人性能相关的信息。数据储存库614、616可以驻留在各种各样的位置。例如,服务器612使用的数据储存库可以位于服务器612本地或者可以远离服务器612并且经由基于网络的连接或专用连接与服务器612通信。数据储存库614、616可以是不同类型的。在某些示例中,服务器612使用的数据储存库可以是数据库,例如关系数据库,诸如由Oracle Corporationn®和其他供应商提供的数据库。这些数据库中的一个或多个可以适合于响应于SQL格式的命令而使能向数据库以及从数据库的存储、更新和检索。

[0177] 在某些示例中,数据储存库614、616中的一个或多个也可以由应用用来存储应用数据。由应用使用的数据储存库可以是不同类型的,诸如,例如键值存储储存库、对象存储储存库或文件系统所支持的通用存储储存库。

[0178] 在某些示例中,可以经由云环境将本公开中描述的功能作为服务来供应。图7是根据某些示例的基于云的系统环境的简化框图,其中可以将各种服务作为云服务来供应。在

图7中所描绘的示例中,云基础设施系统702可提供可由用户使用一个或多个客户端计算设备704、706和708请求的一个或多个云服务。云基础设施系统702可以包括一个或多个计算机和/或服务器,其可以包括上面针对服务器612描述的那些。云基础设施系统702中的计算机可以被组织为通用计算机、专用服务器计算机、服务器群、服务器集群或任何其他适当的布置和/或组合。

[0179] (一个或多个)网络710可以促进客户端704、706和708与云基础设施系统702之间的数据通信和交换。(一个或多个)网络710可以包括一个或多个网络。网络可以是相同或不同类型的。(一个或多个)网络710可以支持一个或多个通信协议,包括有线和/或无线协议,用于促进通信。

[0180] 图7中描绘的示例仅是云基础设施系统的一个示例并且不旨在进行限制。应当理解,在一些其他实施例中,云基础设施系统702可以具有比图7中所描绘的那些更多或更少的组件,可以组合两个或更多个组件,或者可以具有不同的组件配置或布置。例如,尽管图7描绘了三个客户端计算设备,但在替代示例中任何数量的客户端计算设备可以被支持。

[0181] 术语云服务通常用于指由服务提供商的系统(例如,云基础设施系统702)按需并经由诸如互联网的通信网络提供给用户的服务。通常,在公共云环境中,构成云服务提供商系统的服务器和系统与客户自己的本地服务器和系统不同。云服务提供商的系统由云服务提供商管理。因此,客户可以利用云服务提供商提供的云服务,而无需为服务购买单独的许可、支持或硬件和软件资源。例如,云服务提供商的系统可以托管应用,并且用户可以经由互联网按需订购和使用应用而无需用户购买用于执行应用的基础设施资源。云服务被设计为提供对应用、资源和服务的轻松、可扩展的访问。若干提供商供应云服务。例如,由加利福尼亚州红木海岸的Oracle Corporation®供应若干云服务,诸如中间件服务、数据库服务、Java云服务等。

[0182] 在某些示例中,云基础设施系统702可以使用不同的模型来提供一个或多个云服务,诸如在软件即服务(SaaS)模型、平台即服务(PaaS)模型、基础设施即服务(基础设施即服务)模型等(包括混合服务模型)下。云基础设施系统702可以包括一套应用、中间件、数据库和使能对各种云服务的供应的其他资源。

[0183] SaaS模型使得应用或软件能够作为服务经由如互联网的通信网络递送给客户,而客户无需购买用于底层应用的硬件或软件。例如,SaaS模型可用于向客户提供对由云基础设施系统702托管的按需应用的访问。由Oracle Corporation®提供的SaaS服务的示例包括但不限于用于人力资源/资本管理的各种服务、客户关系管理(CRM)、企业资源规划(ERP)、供应链管理(SCM)、企业绩效管理(EPM)、分析服务、社交应用等。

[0184] IaaS模型通常用于将基础设施资源(例如,服务器、存储、硬件和联网资源)作为云服务提供给客户以提供弹性计算和存储能力。Oracle Corporation®提供各种IaaS服务。

[0185] PaaS模型通常用于提供作为服务的平台和环境资源,其使得客户能够开发、运行和管理应用和服务,而客户无需采购、构建或维护这样的资源。Oracle Corporation®提供的PaaS服务的示例包括但不限于Oracle Java云服务(JCS)、Oracle数据库云服务(DBCS)、数据管理云服务、各种应用开发解决方案服务等。

[0186] 云服务通常以按需自助服务为基础、基于订阅、可弹性扩展、可靠、高度可用且安全的方式提供。例如,客户可以经由订阅订单订购由云基础设施系统702提供的一个或多个服务。云基础设施系统702然后执行处理以提供客户的订阅订单中请求的服务。例如,如上所述,用户可以使用话语来请求云基础设施系统采取特定动作(例如,意图),和/或如本文所述为聊天机器人系统提供服务。云基础设施系统702可以被配置为提供一个或甚至多个云服务。

[0187] 云基础设施系统702可以经由不同的部署模型提供云服务。在公共云模型中,云基础设施系统702可以由第三方云服务提供商拥有,并且云服务被供应给任何一般公共客户,其中客户可以是个人或企业。在某些其他示例中,在私有云模型下,云基础设施系统702可以在组织内(例如,在企业组织内)操作并且服务可以被提供给组织内的客户。例如,客户可以是企业的各种部门,诸如人力资源部门、薪资部门等,甚至企业内的个人。在某些其他示例中,在社区云模型下,云基础设施系统702和所提供的服务可以由相关社区中的若干组织共享。也可以使用各种其他模型,诸如上述模型的混合。

[0188] 客户端计算设备704、706和708可以是不同类型的(诸如图6中描绘的客户端计算设备602、604、606和608)并且可以能够操作一个或多个客户端应用。用户可以使用客户端设备来与云基础设施系统702交互,诸如请求由云基础设施系统702提供的服务。例如,用户可以使用客户端设备来请求来自聊天机器人的信息或动作,如本公开中所描述的。

[0189] 在一些示例中,由云基础设施系统702执行的用于提供服务的处理可以涉及模型训练和部署。此分析可能涉及使用、分析和操纵数据集来训练和部署一个或多个模型。此分析可以由一个或多个处理器执行,可能并行处理数据、使用数据执行模拟等。例如,大数据分析可以由云基础设施系统702执行以用于生成和训练聊天机器人系统的一个或多个模型。用于此分析的数据可以包括结构化数据(例如,存储在数据库中的或根据结构化模型结构化的数据)和/或非结构化数据(例如,数据blob(二进制大对象))。

[0190] 如图7中的示例中所描绘的,云基础设施系统702可以包括用于促进由云基础设施系统702供应的各种云服务的供应的基础设施资源730。基础设施资源730可以包括例如处理资源、存储或存储器资源、联网资源等。在某些示例中,可用于服务从应用请求的存储的存储虚拟机可以是云基础设施系统702的一部分。在其他示例中,存储虚拟机可以是不同系统的一部分。

[0191] 在某些示例中,为了促进这些资源的有效供应以用于支持由云基础设施系统702为不同客户提供的各种云服务,资源可以被捆绑成资源集或资源模块(也称为“pod”)。每个资源模块或pod可以包括一种或多种类型的资源的预先集成和优化的组合。在某些示例中,可以为不同类型的云服务预先供应不同的容器。例如,第一组pod可以被供应用于数据库服务,第二组pod可以被供应用于Java服务,该第二组pod可以包括与第一组pod中的pod不同的资源组合。对于一些服务,为供应服务而分配的资源可以在服务之间共享。

[0192] 云基础设施系统702本身可以在内部使用由云基础设施系统702的不同组件共享并且促进由云基础设施系统702对服务的供应的服务732。这些内部共享服务可以包括但不限于安全和身份服务、集成服务、企业储存库服务、企业管理器服务、病毒扫描和白名单服务、高可用性、备份和恢复服务、用于使能云支持的服务、电子邮件服务、通知服务、文件传输服务等。

[0193] 云基础设施系统702可以包括多个子系统。这些子系统可以以软件、或硬件或其组合来实现。如图7中所描绘的,子系统可以包括用户接口子系统712,其使得云基础设施系统702的用户或客户能够与云基础设施系统702交互。用户接口子系统712可以包括各种不同的接口,诸如web接口714、其中由云基础设施系统702提供的云服务被广告并且可由消费者购买的在线商店接口716以及其他接口718。例如,消费者可以使用客户端设备来请求(服务请求734)由云基础设施系统702使用接口714、716和718中的一个或多个提供的一个或多个服务。例如,客户可以访问在线商店,浏览云基础设施系统702供应的云服务,并且对云基础设施系统供应的客户期望订阅的一个或多个服务下订阅订单。服务请求可以包括识别客户以及客户期望订阅的一个或多个服务的信息。例如,客户可以对由云基础设施系统702供应的服务下订阅订单。作为订单的一部分,客户可以提供识别要为其提供服务的聊天机器人系统的信息以及可选地提供聊天机器人系统的一个或多个凭证。

[0194] 在某些示例中,诸如图7中所描绘的示例,云基础设施系统702可以包括被配置为处理新订单的订单管理子系统(OMS)720。作为此处理的一部分,OMS 720可以被配置为:为客户创建账户,如果尚未完成;从客户接收计费和/或会计信息,该信息用于向客户计费以用于向客户提供所请求的服务;核实客户信息;核实后,为客户预订订单;以及编排各种工作流程来准备用于供应订单。

[0195] 一旦正确验证,OMS 720然后可以调用订单供应子系统(OPS)724,该子系统被配置为为订单供应资源,包括处理、存储器和网络资源。供应可以包括为订单分配资源以及配置资源以促进客户订单所请求的服务。为订单供应资源的方式以及所供应的资源的类型可以取决于客户已订购的云服务的类型。例如,根据一个工作流程,OPS 724可以被配置为确定正在被请求的特定云服务并识别可能已经为该特定云服务预先配置的pod的数量。为订单分配的pod数量可能取决于所请求的服务的大小/数量/级别/范围。例如,要被分配的pod的数量可以基于服务要支持的用户数量、请求服务的持续时间等来确定。然后所分配的pod可以针对特定请求客户来定制以用于提供所请求的服务。

[0196] 在某些示例中,如上所述,设置阶段处理可以由云基础设施系统702执行作为供应过程的一部分。云基础设施系统702可以生成应用ID,并从云基础设施系统702本身提供的存储虚拟机中或者从除云基础设施系统702之外的其他系统提供的存储虚拟机中选择用于应用的存储虚拟机。

[0197] 云基础设施系统702可以向请求客户发送响应或通知744以指示何时所请求的服务现在准备好被使用。在一些情况下,信息(例如,链接)可以被发送给客户,该信息使得客户能够开始使用并利用所请求的服务的好处。在某些示例中,对于请求服务的客户,响应可以包括由云基础设施系统702生成的聊天机器人系统ID以及识别由云基础设施系统702为与聊天机器人系统ID相对应的聊天机器人系统选择的聊天机器人系统的信息。

[0198] 云基础设施系统702可以向多个客户提供服务。对于每个客户,云基础设施系统702负责管理与从客户接收的一个或多个订阅订单相关的信息、维护与订单相关的客户数据、以及向客户提供所请求的服务。云基础设施系统702还可以收集关于客户对订阅的服务的使用的使用统计。例如,可以收集所使用的存储量、所传输的数据量、用户数量以及系统正常运行时间和系统停机时间量等的统计数据。此使用信息可用于向客户计费。例如,可以按月周期进行计费。

[0199] 云基础设施系统702可以并行地向多个客户提供服务。云基础设施系统702可以存储这些客户的信息,包括可能的专有信息。在某些示例中,云基础设施系统702包括身份管理子系统(IMS)728,该子系统被配置为管理客户信息并提供所管理的信息的分离,使得与一个客户相关的信息不可被另一客户访问。IMS 728可以被配置为提供各种与安全相关的服务,诸如身份服务,诸如信息访问管理、认证和授权服务、用于管理客户身份和角色以及相关能力的服务等。

[0200] 图8图示了计算机系统800的示例。在一些示例中,计算机系统800可以用于实现分布式环境内的数字助理或聊天机器人系统中的任一个以及上述的各种服务器和计算机系统。如图8中所示,计算机系统800包括各种子系统,包括经由总线子系统802与多个其他子系统通信的处理子系统804。这些其他子系统可以包括处理加速单元806、I/O子系统808、存储子系统818、和通信子系统824。存储子系统818可以包括非暂时性计算机可读存储介质,包括存储介质822和系统存储器810。

[0201] 总线子系统802提供了一种用于让计算机系统800的各种组件和子系统按预期彼此通信的机制。尽管总线子系统802被示意性地示为单个总线,但是总线子系统的替代示例可以利用多个总线。总线子系统802可以是若干类型的总线结构中的任何一种,包括存储器总线或存储器控制器、外围总线和使用各种各样的总线架构中的任何一种的局部总线等。例如,这样的架构可包括工业标准架构(ISA)总线、微通道架构(MCA)总线、增强型ISA(EISA)总线、视频电子标准协会(VESA)局部总线和可以被实现为按照IEEE P1386.1标准制造的夹层总线的外围组件互连(PCI)总线等。

[0202] 处理子系统804控制计算机系统800的操作并且可以包括一个或多个处理器、专用集成电路(ASIC)或现场可编程门阵列(FPGA)。处理器可以包括单核或多核处理器。计算机系统800的处理资源可以被组织成一个或多个处理单元832、834等。处理单元可以包括一个或多个处理器、来自相同或不同处理器的一个或多个核心、核心和处理器的组合、或核心和处理器的其他组合。在一些示例中,处理子系统804可以包括一个或多个专用协处理器,诸如图形处理器、数字信号处理器(DSP)等。在一些示例中,处理子系统804的处理单元中的一些或全部可以使用诸如专用集成电路(ASIC)或现场可编程门阵列(FPGA)的定制电路来实现。

[0203] 在一些示例中,处理单元804可以执行存储在系统存储器810中或计算机可读存储介质822上的指令。在各种示例中,处理单元可以执行多种多样的程序或代码指令并且可以维护多个同时执行的程序或过程。在任何给定时间,要被执行的程序代码中的一些或全部可以驻留在系统存储器810和/或计算机可读存储介质822上,包括潜在地在一个或多个存储设备上。通过合适的编程,处理子系统804可以提供上述各种功能。在计算机系统800正在执行一个或多个虚拟机的情况下,一个或多个处理单元可以被分配给每个虚拟机。

[0204] 在某些示例中,处理加速单元806可以可选地被提供用于执行定制处理或用于卸载由处理子系统804执行的处理中的一些处理以便加速由计算机系统800执行的整体处理。

[0205] I/O子系统808可以包括用于向计算机系统800输入信息和/或用于从或经由计算机系统800输出信息的设备和机构。通常,术语输入设备的使用旨在包括用于将信息输入到计算机系统800的设备和机构的所有可能的类型。用户接口输入设备可包括,例如,键盘、诸如鼠标或轨迹球的指点设备、并入显示器的触摸板或触摸屏、滚轮、点击轮、拨号盘、按钮、

开关、小键盘、带有语音命令识别系统的音频输入设备、麦克风和其他类型输入设备。用户接口输入设备还可以包括运动感测和/或手势识别设备，诸如使用户能够控制输入设备并与之交互的Microsoft Kinect®运动传感器、Microsoft Xbox®360游戏控制器，提供用于接收使用手势和口头命令的输入的接口的设备。用户接口输入设备还可以包括眼姿(eye gesture)识别设备，诸如Google Glass®眨眼检测器，其检测来自用户的眼睛活动(例如，拍照和/或做出菜单选择时“眨眼”)，并将眼姿转换为到输入设备(例如，Google Glass®)的输入。附加地，用户接口输入设备可以包括使用户能够通过语音命令与语音识别系统(例如，Siri®导航器)交互的语音识别感测设备。

[0206] 用户接口输入设备的其它示例还可以包括但不限于三维(3D)鼠标、操纵杆或指点杆、游戏手柄和图形输入板，以及音频/视频设备，诸如扬声器、数码相机、数码摄像机、便携式媒体播放器、网络摄像头、图像扫描仪、指纹扫描仪、条码读取器3D扫描仪、3D打印机、激光测距仪和眼睛注视跟踪设备。附加地，用户接口输入设备可以包括例如医学成像输入设备，诸如计算机断层扫描、磁共振成像、位置发射断层扫描以及医学超声设备。用户接口输入设备还可以包括例如音频输入设备，诸如MIDI键盘、数字乐器等。

[0207] 通常，术语输出设备的使用旨在包括用于将信息从计算机系统800输出到用户或其他计算机的所有可能类型的设备和机构。用户接口输出设备可包括显示子系统、指示灯或诸如音频输出设备的非视觉显示器等。显示子系统可以是阴极射线管(CRT)、诸如使用液晶显示器(LCD)或等离子显示器的平板设备、投影设备、触摸屏等。例如，用户接口输出设备可以包括但不限于以视觉方式传达文本、图形和音频/视频信息的各种各样的显示设备，诸如监视器、打印机、扬声器、耳机、汽车导航系统、绘图仪、语音输出设备和调制解调器。

[0208] 存储子系统818提供用于存储由计算机系统800使用的信息和数据的储存库或数据仓库。存储子系统818提供用于存储提供一些示例的功能的基本编程和数据构造的有形非暂时性计算机可读存储介质(例如，非暂时性计算机可读存储器)。存储子系统818可以存储当由处理子系统804执行时提供上述功能的软件(例如，程序、代码模块、指令)。软件可以由处理子系统804的一个或多个处理单元执行。存储子系统818还可以根据本公开的教导提供认证。

[0209] 存储子系统818可以包括一个或多个非暂时性存储器设备，包括易失性和非易失性存储器设备。如图8中所示，存储子系统818包括系统存储器810和计算机可读存储介质822。系统存储器810可以包括多个存储器，包括用于程序执行期间指令和数据的存储的易失性主随机存取存储器(RAM)和存储固定指令的非易失性只读存储器(ROM)或闪存。在一些实现中，包含有助于在计算机系统800内的元素之间传输信息(诸如在启动期间)的基本例程的基本输入/输出系统(BIOS)通常可以被存储在ROM中。RAM通常包含当前由处理子系统804操作和执行的数据和/或程序模块。在一些实现中，系统存储器810可以包括多种不同类型的存储器，诸如静态随机存取存储器(SRAM)或动态随机存取内存(DRAM)等。

[0210] 举例来说，而非限制，如图8中所描绘的，系统存储器810可以加载可以包括诸如Web浏览器、中间层应用、关系数据库管理系统(RDBMS)等的正在被执行的应用程序1212，程序数据814和操作系统816。举例来说，操作系统816可以包括各种版本的Microsoft Windows®、Apple Macintosh®和/或Linux操作系统、各种各样的商用UNIX®或类

UNIX操作系统(包括但不限于各种各样的GNU/Linux操作系统、Google Chrome® OS等)和/或移动操作系统,诸如iOS、Windows® Phone、Android® OS、BlackBerry® OS和Palm® OS操作系统等等。

[0211] 计算机可读存储介质922可以存储提供一些示例的功能的编程和数据构造。计算机可读介质822可以为计算机系统800提供计算机可读指令、数据结构、程序模块和其他数据的存储。当由处理子系统804执行时提供上述功能的软件(程序、代码模块、指令)可以被存储在存储子系统818中。举例来说,计算机可读存储介质822可以包括诸如硬盘驱动器、磁盘驱动器、诸如CD ROM、DVD和Blu-Ray® 光盘或其他光学介质的光盘驱动器的非易失性存储器。计算机可读存储介质822可以包括但不限于Zip®驱动器、闪存卡、通用串行总线(USB)闪存驱动器、安全数字(SD)卡、DVD盘、数字视频磁带等。计算机可读存储介质822还可以包括基于非易失性存储器的固态驱动(SSD)(诸如基于闪存的SSD、企业闪存驱动器、固态ROM等)、基于易失性存储器的SSD(诸如固态RAM、动态RAM、静态RAM、DRAM、磁阻RAM(MRAM)SSD)、以及使用DRAM和基于闪存的SSD组合的混合SSD。

[0212] 在某些示例中,存储子系统818还可以包括计算机可读存储介质读取器820,该计算机可读存储介质读取器820可以被进一步连接到计算机可读存储介质822。读取器820可以接收数据并且被配置为从诸如磁盘、闪存驱动器等的存储器设备读取数据。

[0213] 在某些示例中,计算机系统800可以支持虚拟化技术,包括但不限于处理和存储器资源的虚拟化。例如,计算机系统800可以为执行一个或多个虚拟机提供支持。在某些示例中,计算机系统800可以执行诸如促进虚拟机的配置和管理的管理程序的程序。每个虚拟机可被分配存储器、计算(例如,处理器、核心)、I/O和联网资源。每个虚拟机通常独立于其他虚拟机运行。虚拟机通常运行其自己的操作系统,该操作系统可以与由计算机系统800执行的其他虚拟机所执行的操作系统相同或不同。因此,多个操作系统可以潜在地由计算机系统800同时运行。

[0214] 通信子系统824提供到其他计算机系统和网络的接口。通信子系统824用作用于从计算机系统800接收来自其他系统的数据以及向其他系统发送数据的接口。例如,通信子系统824可以使得计算机系统800能够经由互联网建立到一个或多个客户端设备的通信信道,用于从客户端设备接收信息并将信息发送到客户端设备。例如,当计算机系统800用于实现图1中所描绘的机器人系统120时,通信子系统可用于与为应用而选择的聊天机器人系统通信。

[0215] 通信子系统824可以支持有线和/或无线通信协议两者。在某些示例中,通信子系统824可以包括用于访问无线语音和/或数据网络(例如,使用蜂窝电话技术、高级数据网络技术,诸如3G、4G或EDGE(增强数据速率全球演进)、WiFi(IEEE 902.11系列标准,或其他移动通信技术,或其任何组合)的射频(RF)收发器组件、全球定位系统(GPS)接收器组件和/或其他组件。在一些示例中,除了或替代无线接口,通信子系统824可以提供有线网络连接(例如,以太网)。

[0216] 通信子系统824可以接收和发送各种形式的数据。在一些示例中,除了其他形式之外,通信子系统824可以接收结构化和/或非结构化的数据馈送1326、事件流1328、事件更新

1330等形式的输入通信。例如,通信子系统824可以被配置为从社交媒体网络和/或其他通信服务的用户实时接收(或发送)数据馈送826,诸如Twitter®馈送、Facebook®更新、诸如丰富站点摘要(RSS)馈送的web馈送和/或来自一个或多个第三方信息源的实时更新。

[0217] 在某些示例中,通通信子系统824还可以被配置为接收连续数据流形式的数据,该数据可以包括实时事件的事件流828和/或事件更新830,其本质上可以是连续的或无界的,没有明确的结束。生成连续数据的应用的示例可以包括例如传感器数据应用、金融行情、网络性能测量工具(例如网络监控和流量管理应用)、点击流分析工具、汽车流量监控等。

[0218] 通信子系统824还可以被配置为将数据从计算机系统800传送到其他计算机系统或网络。数据可以以各种不同的形式(诸如结构化和/或非结构化数据馈送826、事件流828、事件更新830等)传送到一个或多个数据库,该一个或多个数据库可以与耦合到计算机系统1800的一个或多个流数据源计算机通信。

[0219] 计算机系统800可以是各种类型中的一种,包括手持便携式设备(例如,iPhone®蜂窝电话、iPad®计算平板、PDA)、可穿戴设备(例如,Google Glass®头戴式显示器)、个人计算机、工作站、大型机、信息亭、服务器机架或任何其他数据处理系统。由于计算机和网络的不断变化的性质,图8中所描绘的计算机系统800的描述仅旨在作为具体示例。具有比图8中描绘的系统更多或更少组件的许多其他配置是可能的。基于本文提供的公开和教导,应当理解存在其他方式和/或方法来实现各种示例。

[0220] 尽管已经描述了具体示例,但是各种修改、变更、替代构造和等同物是可能的。示例不限于某些具体数据处理环境内的操作,而是可以自由地在多个数据处理环境内操作。附加地,尽管已经使用特定系列的事务和步骤描述了某些示例,但是对于本领域技术人员来说应显而易见的是,这并不旨在进行限制。尽管一些流程图将操作描述为顺序过程,但是许多操作可以并行或同时执行。另外,操作的次序可以重新布置。过程可能具有未包含在图中的附加步骤。上述示例的各种特征和方面可以被个别地或联合使用。

[0221] 此外,虽然已经使用硬件和软件的特定组合描述了某些示例,但是应当认识到硬件和软件的其他组合也是可能的。某些示例可以仅以硬件、或仅以软件、或使用其组合来实现。本文描述的各种过程可以在相同处理器上或不同处理器上以任何组合实现。

[0222] 在设备、系统、组件或模块被描述为被配置为执行某些操作或功能的情况下,这样的配置可以例如通过设计电子电路来执行操作、通过对可编程电子电路(诸如微处理器)编程来执行操作(例如通过执行计算机指令或代码)或被编程为执行存储在非暂时性存储介质上的代码或指令的处理器或核心或其任何组合来实现。过程可以使用包括但不限于用于过程间通信的常规技术的各种各样的技术来通信,并且不同的过程对可以使用不同的技术,或者相同过程对可以在不同的时间使用不同的技术。

[0223] 本公开中给出了具体细节以提供对示例的透彻理解。然而,可以在没有这些具体细节的情况下实践示例。例如,已经在没有不必要的细节的情况下示出了众所周知的电路、过程、算法、结构和技术以避免模糊示例。本描述仅提供示例的示例,并且不旨在限制其他示例的范围、适用性或配置。相反,前面对示例的描述将为本领域技术人员提供用于实现各种示例的使能描述。可以对元件的功能和布置做出各种改变。

[0224] 因此,说明书和附图被认为是说明性的而不是限制性意义。然而,将显而易见,在

不偏离如权利要求中阐述的更广泛的精神和范围的情况下,可以对其做出添加、删减、删除和其他修改和改变。因此,尽管已经描述了具体示例,但是这些并不旨在进行限制。各种修改和等同物都在所附权利要求的范围内。

[0225] 在前述说明书中,参考本公开的具体示例描述了本公开的方面,但是本领域技术人员将认识到本公开不限于此。上述公开的各种特征和方面可以被个别地或联合使用。此外,可以在超出本文描述的那些环境和应用的任何数量的环境和应用中利用实施例,而不脱离本说明书的更广泛的精神和范围。因此,说明书和附图被认为是说明性的而不是限制性的。

[0226] 在前面的描述中,出于说明的目的,以特定次序描述了方法。应当理解,在替代示例中,可以以与所描述的次序不同的次序来执行这些方法。还应当理解,上述方法可以通过硬件组件来执行,或者可以体现在机器可执行指令的序列中,这些指令可以用于使机器(诸如通用或专用处理器或用指令编程的逻辑电路)来执行这些方法。这些机器可执行指令可以被存储在一个或多个机器可读介质上,诸如CD-ROM或其他类型的光盘、软盘、ROM、RAM、EPROM、EEPROM、磁卡或光卡、闪存或适合存储电子指令的其他类型的机器可读介质。替代地,方法可以通过硬件和软件的组合来执行。

[0227] 在组件或模块被描述为被配置为执行某些操作的情况下,这样的配置可以通过例如设计电子电路或其他硬件以执行操作、通过对可编程电子电路(诸如微处理器或其他合适的电子电路)进行编程以执行操作或其任何组合来实现。

[0228] 虽然本文已经详细描述了本申请的说明性示例,但是应当理解,本发明概念可以以其他方式不同地体现和采用,并且所附权利要求旨在被解释为包括这样的变化,除非受现有技术的限制。

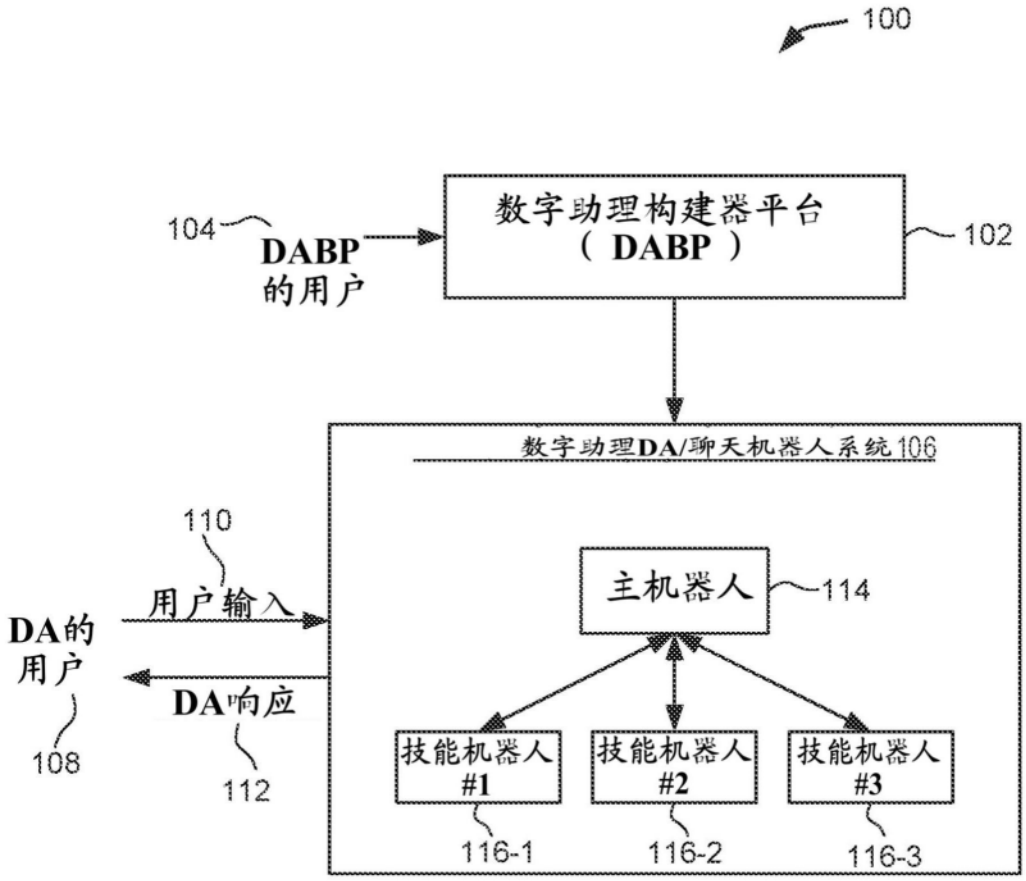

100
104
DABP
的用户
数字助理构建器平台
（ DABP ）
102
数字助理DA/聊天机器人系统106
110
用户输入
DA的
用户
DA响应
108
112
主机器人
114
技能机器人
#1
技能机器人
#2
技能机器人
#3
116-1
116-2
116-3

图1

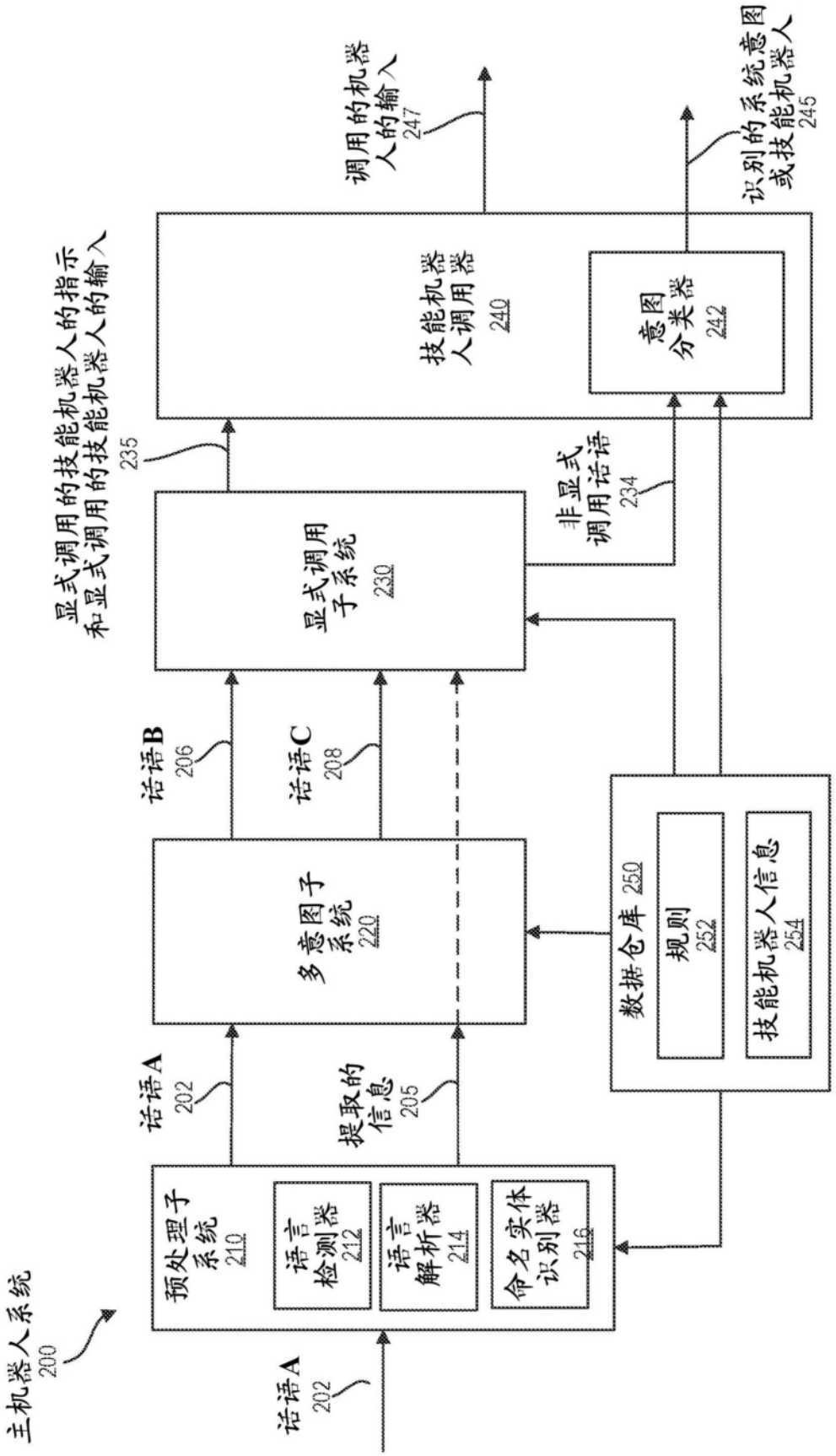

主机器人系统
200
话语A
202
预处理子系统
210
语言检测器
212
语言解析器
214
命名实体识别器
216
话语A
202
提取的信息
205
多意图子系统
220
话语B
206
话语C
208
显式调用子系统
230
显式调用的技能机器人的指示和显式调用的技能机器人的输入
235
非显式调用话语
234
数据仓库
250
规则
252
技能机器人信息
254
技能机器人调用器
240
意图分类器
242
调用的机器人的输入
247
识别的系统意图或技能机器人
245

图2

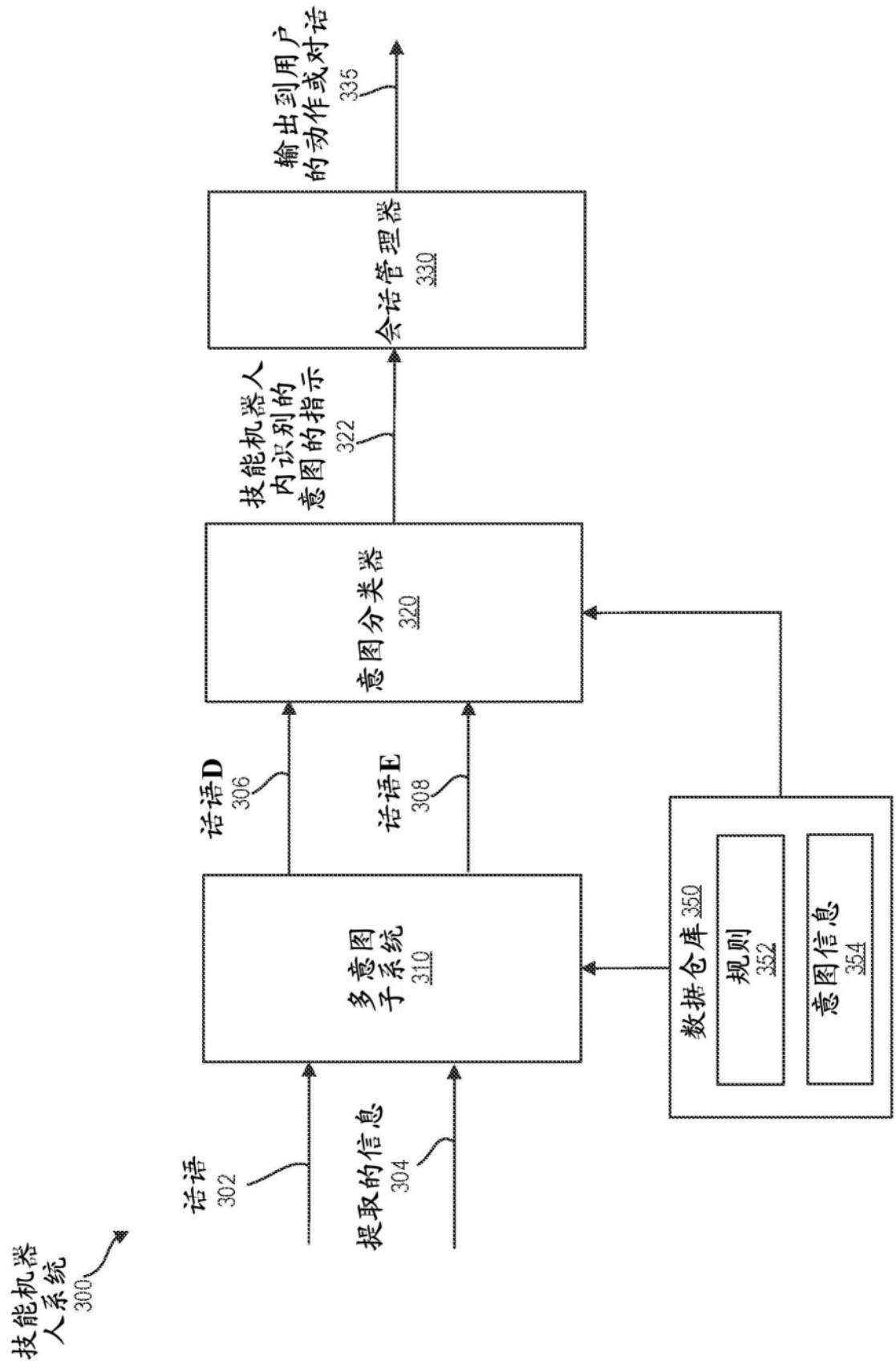

技能机器人系统
300
话语
302
提取的信息
304
话语D
306
话语E
308
多意图子系统
310
意图分类器
320
技能机器人内识别的意图的指示
322
会话管理器
330
输出到用户的动作或对话
335
数据仓库 350
规则
352
意图信息
354

图3

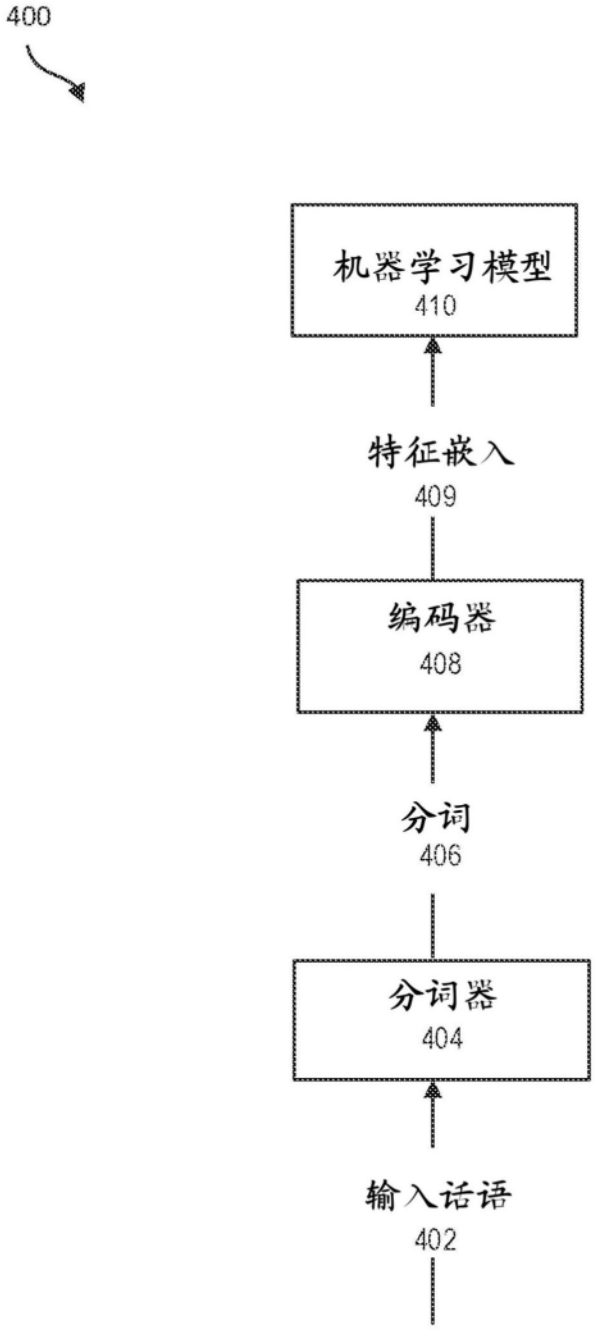


图4A

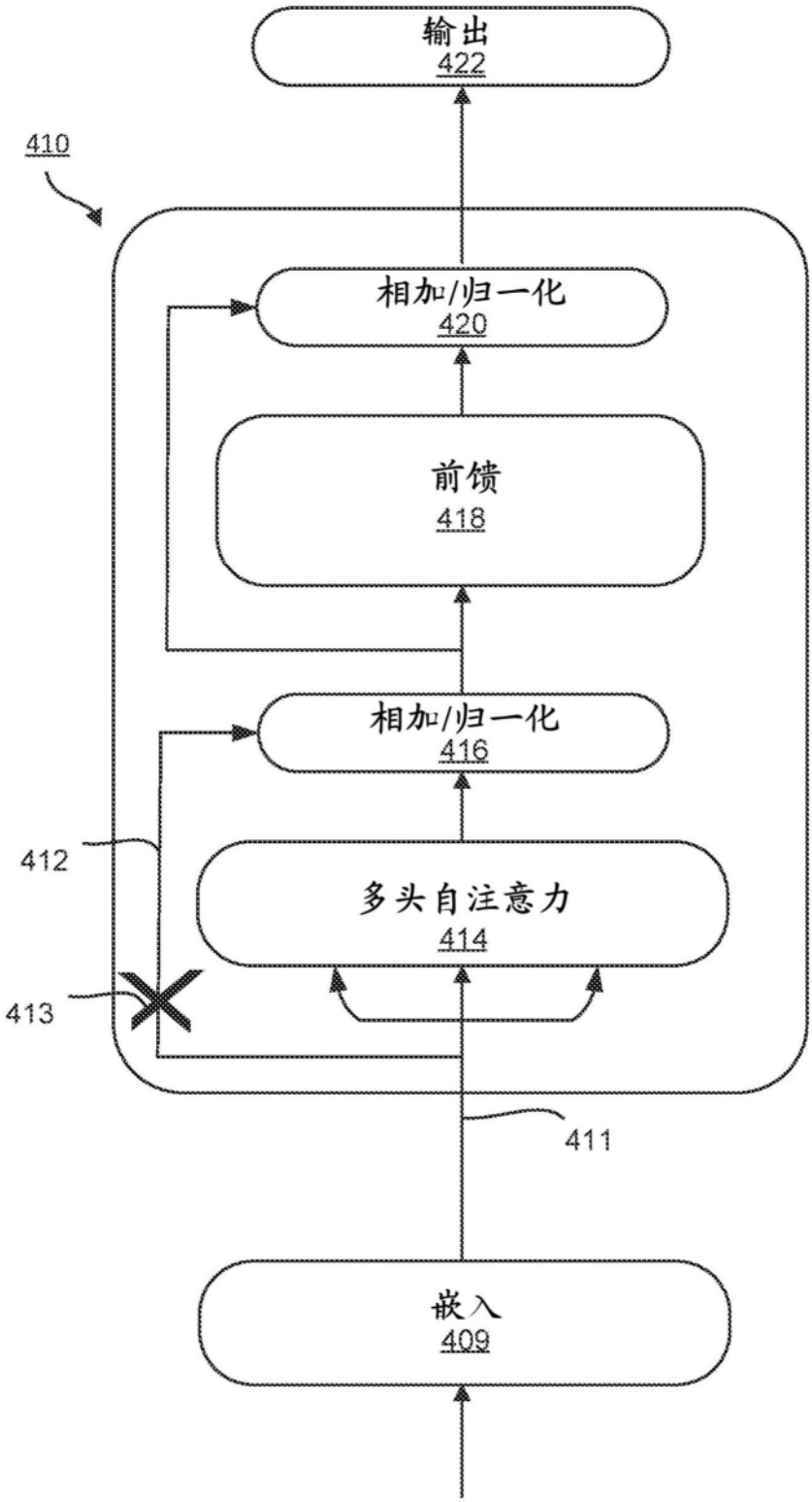

输出
422
410
相加/归一化
420
前馈
418
相加/归一化
416
412
多头自注意力
414
413
411
嵌入
409

图4B

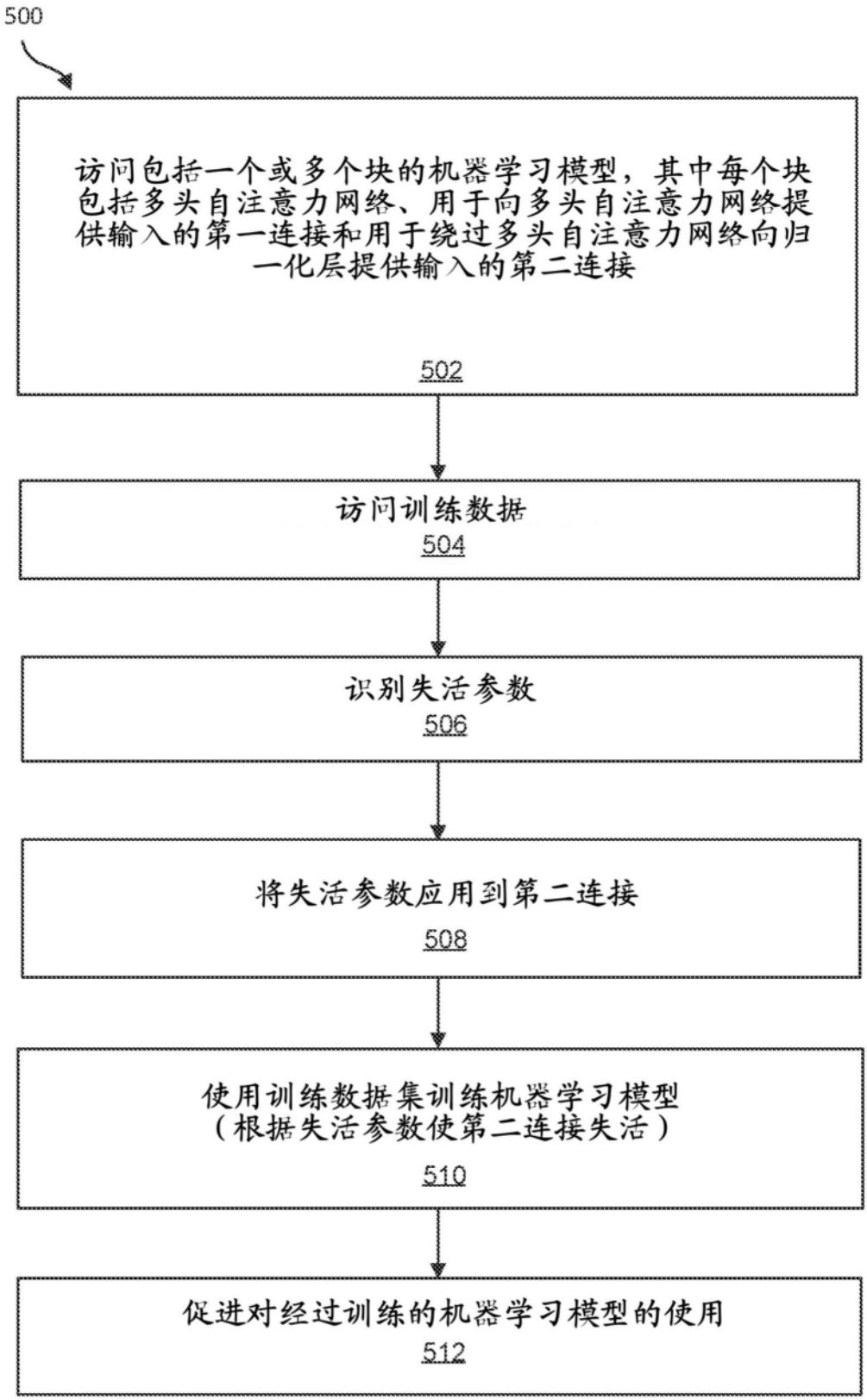

500
访问包括一个或多个块的机器学习模型，其中每个块包括多头自注意力网络、用于向多头自注意力网络提供输入的第一连接和用于绕过多头自注意力网络向归一化层提供输入的第二连接
502
访问训练数据
504
识别失活参数
506
将失活参数应用到第二连接
508
使用训练数据集训练机器学习模型（根据失活参数使第二连接失活）
510
促进对经过训练的机器学习模型的使用
512

图5A

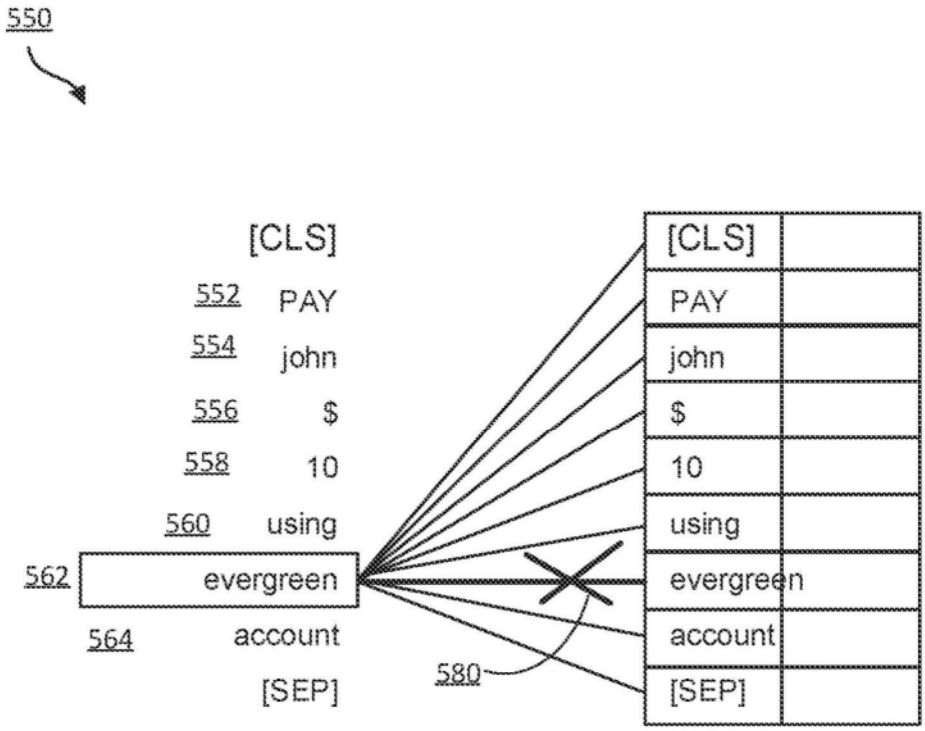

550
[CLS]
552
PAY
554
john
556
$
558
10
560
using
562
evergreen
564
account
[SEP]
580
[CLS]
PAY
john
$
10
using
evergreen
account
[SEP]

图5B

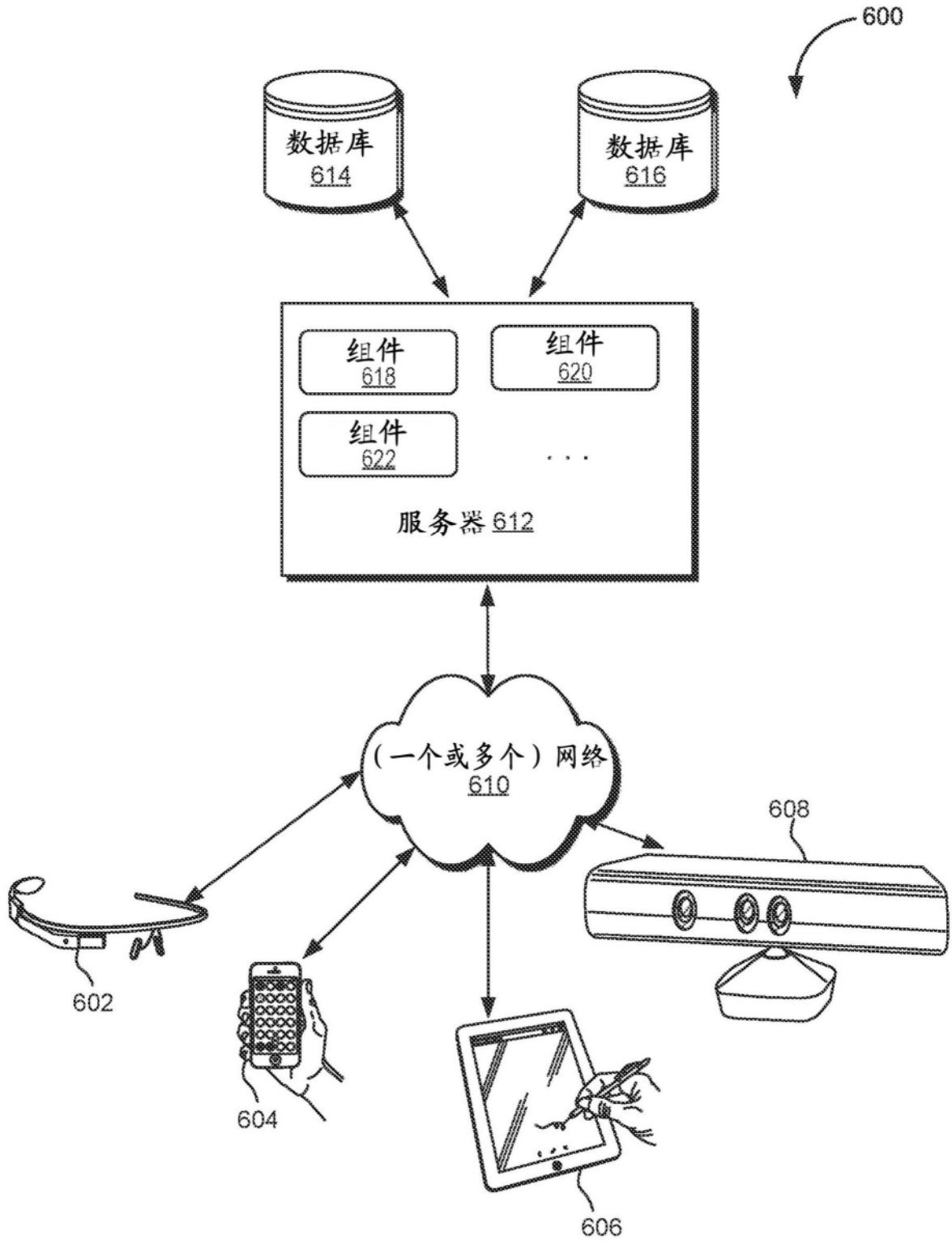

600
数据库
614
数据库
616
组件
618
组件
620
组件
622
. . .
服务器 612
(一个或多个)网络
610
608
602
604
606

图6

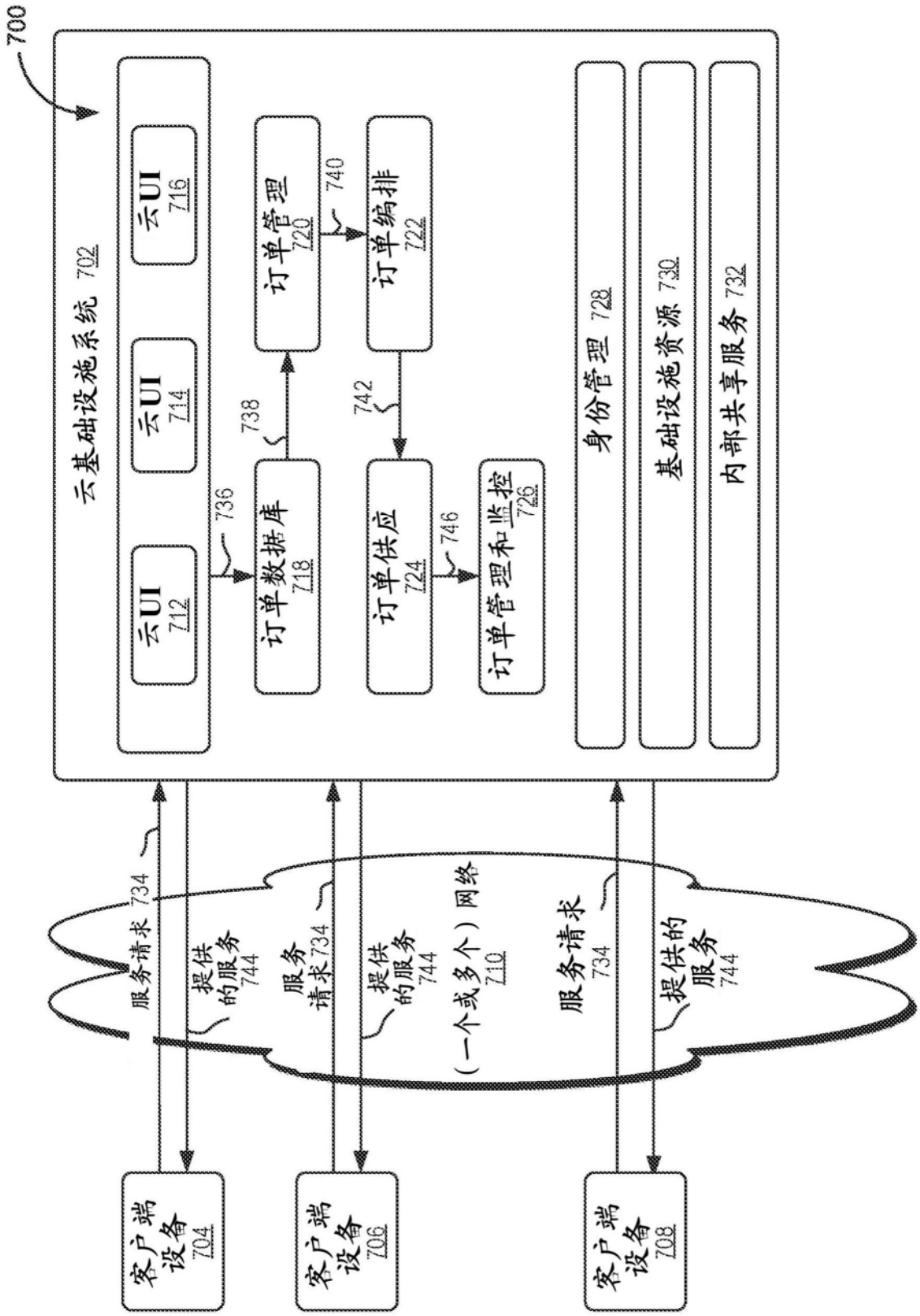

700
云基础设施系统 702
云UI 712
云UI 714
云UI 716
订单数据库 718
订单管理 720
订单编排 722
订单供应 724
订单管理和监控 726
身份管理 728
基础设施资源 730
内部共享服务 732
736
738
740
742
746
服务请求 734
提供的服务 744
服务请求 734
提供的服务 744
（一个或多个）网络 710
服务请求 734
提供的服务 744
客户端设备 704
客户端设备 706
客户端设备 708

图7

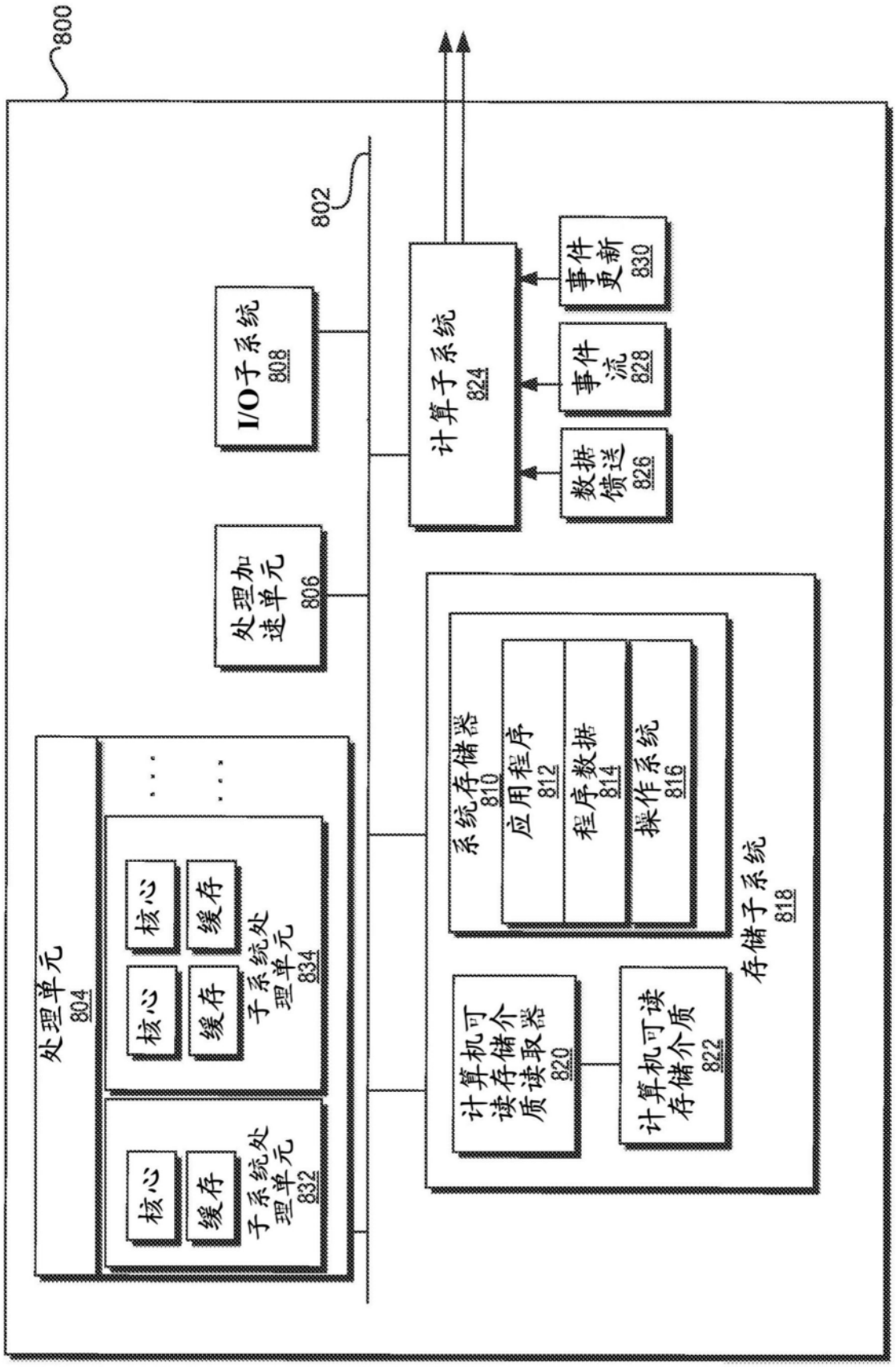

800
802
处理单元
804
核心
缓存
子系统处理单元
832
核心
缓存
子系统处理单元
834
处理加速单元
806
I/O子系统
808
计算子系统
824
事件更新
830
事件流
828
数据馈送
826
系统存储器
810
应用程序
812
程序数据
814
操作系统
816
存储子系统
818
计算机可读存储介质读取器
820
计算机可读存储介质
822

图8