(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118215920 A
(43)申请公布日 2024.06.18

(21)申请号 202280074232.4

(22)申请日 2022.11.07

(30)优先权数据
63/263,728 2021.11.08 US
18/052,694 2022.11.04 US

(85)PCT国际申请进入国家阶段日
2024.05.07

(86)PCT国际申请的申请数据
PCT/US2022/049164 2022.11.07

(87)PCT国际申请的公布数据
WO2023/081483 EN 2023.05.11

(71)申请人 甲骨文国际公司
地址 美国加利福尼亚州

(72)发明人 T·T·乌 P·扎雷穆迪 D·武 M·E·约翰逊 钟旭 V·布利诺夫 C·D·V·黄 洪宇衡 V·戈尔 P·V·奥格伦 S·P·K·加德 V·比什诺伊 T·L·杜翁

(74)专利代理机构 北京市汉坤律师事务所 11602
专利代理师 魏小薇 吴丽丽

(51)Int.Cl.
G06F 40/263(2020.01)
G06F 40/274(2020.01)
G06F 40/216(2020.01)
G06F 18/2415(2023.01)
G06N 3/0464(2023.01)
G06N 3/0499(2023.01)
G06N 3/08(2023.01)
H04L 51/02(2022.01)
G06F 16/332(2019.01)

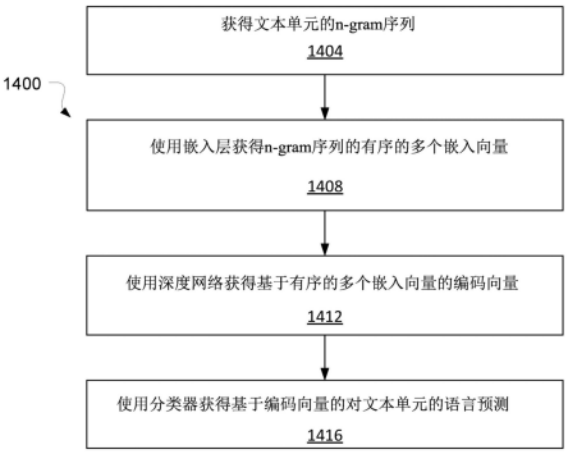
权利要求书3页 说明书31页 附图18页

(54)发明名称

用于使用散列嵌入进行语言检测的宽深网络

(57)摘要

本文所公开的技术总体上涉及语言检测。在一个特定方面,提供了一种方法,该方法包括:获得文本单元的n-gram序列;使用嵌入层获得n-gram序列的有序的多个嵌入向量;使用深度网络获得基于有序的多个嵌入向量的编码向量;以及使用分类器获得基于编码向量的对文本单元的语言预测。深度网络包括注意力机制,并且使用嵌入层获得有序的多个嵌入向量包括,对于n-gram序列中的每个n-gram:获得n-gram的散列值;基于散列值,从多个分量向量当中选择分量向量;以及获得n-gram的基于分量向量的嵌入向量。



CN 118215920 A

1.一种语言检测方法,所述方法包括:

获得文本单元的n-gram序列;

使用嵌入层获得所述n-gram序列的有序的多个嵌入向量;

使用深度网络获得基于所述有序的多个嵌入向量的编码向量;以及

使用分类器获得基于所述编码向量的对所述文本单元的语言预测,

其中,所述嵌入层包括具有多个分量向量的经训练的模型,并且

其中,所述深度网络包括注意力机制,并且

其中,使用所述嵌入层获得所述有序的多个嵌入向量包括,对于所述n-gram序列中的每个n-gram:

获得所述n-gram的第一散列值和所述n-gram的第二散列值;

基于所述第一散列值,从所述多个分量向量当中选择第一分量向量;

基于所述第二散列值,从所述多个分量向量当中选择第二分量向量;以及

获得所述n-gram的基于所述第一分量向量和所述第二分量向量的嵌入向量。

2.如权利要求1所述的方法,其中,所述n-gram序列包括多个字符级n-gram和多个单词级n-gram。

3.如权利要求2所述的方法,其中,所述多个字符级n-gram的n值不同于所述多个单词级n-gram的n值。

4.如权利要求1所述的方法,其中,所述深度网络包括经训练的卷积神经网络。

5.如权利要求1所述的方法,其中,对于所述n-gram序列中的每个n-gram:

获得所述n-gram的所述第一散列值包括对所述n-gram应用具有第一种子值的散列函数,并且

获得所述n-gram的所述第二散列值包括对所述n-gram应用具有第二种子值的所述散列函数,所述第二种子值不同于所述第一种子值。

6.如权利要求1所述的方法,其中,使用所述嵌入层获得所述有序的多个嵌入向量包括,对于所述n-gram序列中的每个n-gram:对所述第一散列值应用模函数以获得第一索引并且对所述第二散列值应用所述模函数以获得第二索引,其中,选择所述第一分量向量基于所述第一索引并且选择所述第二分量向量基于所述第二索引。

7.如权利要求1至6中任一项所述的方法,其中,对于所述n-gram序列中的每个n-gram,获得所述n-gram的所述嵌入向量包括级联所述第一分量向量和所述第二分量向量。

8.如权利要求1至6中任一项所述的方法,其中,对于所述n-gram序列中的每个n-gram,获得所述n-gram的所述嵌入向量包括:

对所述第一分量向量应用第一加权值以获得第一加权向量,以及

对所述第二分量向量应用第二加权值以获得第二加权向量,其中,所述嵌入向量基于所述第一加权向量和所述第二加权向量。

9.如权利要求1至6中任一项所述的方法,其中,所述分类器包括前馈神经网络。

10.如权利要求9所述的方法,其中,使用所述分类器包括对所述前馈神经网络的最终层的输出应用softmax函数。

11.一种系统,包括:

一个或多个数据处理器;以及

一个或多个非暂态计算机可读介质,所述一个或多个非暂态计算机可读介质存储有指令,所述指令在由所述一个或多个数据处理器执行时使所述一个或多个数据处理器执行包括以下各项的处理:

获得文本单元的n-gram序列;

使用嵌入层获得所述n-gram序列的有序的多个嵌入向量;

使用深度网络获得基于所述有序的多个嵌入向量的编码向量;以及

使用分类器获得基于所述编码向量的对所述文本单元的语言预测,

其中,所述嵌入层包括具有多个分量向量的经训练的模型,并且

其中,所述深度网络包括注意力机制,并且

其中,使用所述嵌入层获得所述有序的多个嵌入向量包括,对于所述n-gram序列中的每个n-gram:

获得所述n-gram的第一散列值和所述n-gram的第二散列值;

基于所述第一散列值,从所述多个分量向量当中选择第一分量向量;

基于所述第二散列值,从所述多个分量向量当中选择第二分量向量;以及获得所述n-gram的基于所述第一分量向量和所述第二分量向量的嵌入向量。

12.如权利要求11所述的系统,其中,对于所述n-gram序列中的每个n-gram:

获得所述n-gram的所述第一散列值包括对所述n-gram应用具有第一种子值的散列函数,并且

获得所述n-gram的所述第二散列值包括对所述n-gram应用具有第二种子值的所述散列函数,所述第二种子值不同于所述第一种子值。

13.如权利要求11所述的系统,其中,使用所述嵌入层获得所述有序的多个嵌入向量包括,对于所述n-gram序列中的每个n-gram:对所述第一散列值应用模函数以获得第一索引并且对所述第二散列值应用所述模函数以获得第二索引,其中,选择所述第一分量向量基于所述第一索引并且选择所述第二分量向量基于所述第二索引。

14.如权利要求11所述的系统,其中,对于所述n-gram序列中的每个n-gram,获得所述n-gram的所述嵌入向量包括:

对所述第一分量向量应用第一加权值以获得第一加权向量,以及

对所述第二分量向量应用第二加权值以获得第二加权向量,其中,所述嵌入向量基于所述第一加权向量和所述第二加权向量。

15.如权利要求11至14中任一项所述的系统,其中,训练所述深度网络包括根据对应输入文本单元的脚本信息来限制语言预测。

16.一种有形地体现在一个或多个非暂态机器可读介质中的计算机程序产品,所述计算机程序产品包括被配置为使一个或多个数据处理器执行包括以下各项的处理的指令:

获得文本单元的n-gram序列;

使用嵌入层获得所述n-gram序列的有序的多个嵌入向量;

使用深度网络获得基于所述有序的多个嵌入向量的编码向量;以及

使用分类器获得基于所述编码向量的对所述文本单元的语言预测,

其中,所述嵌入层包括具有多个分量向量的经训练的模型,并且

其中,所述深度网络包括注意力机制,并且

其中,使用所述嵌入层获得所述有序的多个嵌入向量包括,对于所述n-gram序列中的每个n-gram:

获得所述n-gram的第一散列值和所述n-gram的第二散列值;

基于所述第一散列值,从所述多个分量向量当中选择第一分量向量;

基于所述第二散列值,从所述多个分量向量当中选择第二分量向量;以及获得所述n-gram的基于所述第一分量向量和所述第二分量向量的嵌入向量。

17.如权利要求16所述的计算机程序产品,其中,对于所述n-gram序列中的每个n-gram:

获得所述n-gram的所述第一散列值包括对所述n-gram应用具有第一种子值的散列函数,并且

获得所述n-gram的所述第二散列值包括对所述n-gram应用具有第二种子值的所述散列函数,所述第二种子值不同于所述第一种子值。

18.如权利要求16所述的计算机程序产品,其中,使用所述嵌入层获得所述有序的多个嵌入向量包括,对于所述n-gram序列中的每个n-gram:对所述第一散列值应用模函数以获得第一索引并且对所述第二散列值应用所述模函数以获得第二索引,其中,选择所述第一分量向量基于所述第一索引并且选择所述第二分量向量基于所述第二索引。

19.如权利要求16所述的计算机程序产品,其中,对于所述n-gram序列中的每个n-gram,获得所述n-gram的所述嵌入向量包括:

对所述第一分量向量应用第一加权值以获得第一加权向量,以及

对所述第二分量向量应用第二加权值以获得第二加权向量,其中,所述嵌入向量基于所述第一加权向量和所述第二加权向量。

20.如权利要求16至19中任一项所述的计算机程序产品,其中,使用所述深度网络包括:

对与所述n-gram序列中的第一n-gram相对应的第一特征值应用第一注意力权重,以及

对与所述n-gram序列中的第二n-gram相对应的第二特征值应用不同于所述第一注意力权重的第二注意力权重。

用于使用散列嵌入进行语言检测的宽深网络

[0001] 相关申请的交叉引用

[0002] 本申请要求于2021年11月8日提交的名称为“WIDE AND DEEP NETWORK FOR LANGUAGE DETECTION USING HASHED EMBEDDINGS[用于使用散列嵌入进行语言检测的宽深网络]”的美国临时专利申请号63/263,728和于2022年11月4日提交的名称为“WIDE AND DEEP NETWORK FOR LANGUAGE DETECTION USING HASH EMBEDDINGS[用于使用散列嵌入进行语言检测的宽深网络]”的美国非临时申请号18/052,694的优先权,这些美国申请出于所有目的通过引用以其全文并入本文。

技术领域

[0003] 本公开总体上涉及自然语言处理,并且更具体地涉及用于语言检测的技术。

背景技术

[0004] 为了获得即时反应,世界各地的许多用户使用即时消息传递或聊天平台。组织经常使用这些即时消息传递或聊天平台与客户(或最终用户)进行实时会话。然而,雇用服务人员与客户或最终用户进行实时交流对于组织来说会是非常昂贵的。已经开始开发聊天机器人(或“机器人”)来模拟与最终用户的会话,尤其是通过因特网。最终用户可以通过最终用户已经安装并使用的消息传递应用程序与机器人交流。智能机器人(通常通过人工智能(AI)提供动力)可以在实时会话中更智能地且根据上下文进行交流,并且因此可以允许机器人与最终用户之间进行更加自然的会话以改善会话体验。不是最终用户去学习机器人知道如何作出回应的固定的一组关键词或命令,而是智能机器人可以能够基于自然语言的用户话语理解最终用户的意图并且相应地作出回应。

[0005] 语言检测是一项识别文本单元的语言的任务。文本单元的示例可以包括句子、电子邮件、帖子、文本消息、产品评论、段落或文档。文本单元可以由语音到文本模块响应于话语而产生。语言检测是如机器翻译、文本分类等许多文本处理任务的首要步骤之一。例如,准确的语言检测对于成功部署多语言机器人可能很重要。

发明内容

[0006] 本文所公开的技术总体上涉及语言检测(例如,在自然语言处理中)。可以用于执行语言检测的机器学习(ML)模型的示例包括宽网络。例如,语言检测的机器学习方法可以包括将输入文本作为字符序列或作为n-gram序列或子词(subword)序列呈现给宽网络。本文所公开的技术可以提供对文本单元的语言检测。

[0007] 在各种实施例中,一种用于语言检测的计算机实施的方法包括:获得文本单元的n-gram序列;使用嵌入层获得n-gram序列的有序的多个嵌入向量;使用深度网络获得基于有序的多个嵌入向量的编码向量;以及使用分类器获得基于编码向量的对文本单元的语言预测。嵌入层包括具有多个分量向量的经训练的模型,并且深度网络包括具有注意力机制(例如,一个或多个注意力层)的经训练的卷积神经网络。在这种方法中,使用嵌入层获得有

序的多个嵌入向量包括,对于n-gram序列中的每个n-gram:获得n-gram的第一散列值和n-gram的第二散列值;基于第一散列值,从多个分量向量当中选择第一分量向量;基于第二散列值,从多个分量向量当中选择第二分量向量;以及通过级联第一分量向量和第二分量向量来获得n-gram的嵌入向量。在一些实施例中,深度网络包括具有注意力机制的经训练的卷积神经网络。

[0008] 在一些实施例中,n-gram序列包括多个字符级n-gram和多个单词级n-gram。在一些实施例中,多个字符级n-gram的n值不同于多个单词级n-gram的n值。

[0009] 在一些实施例中,对于n-gram序列中的每个n-gram,获得n-gram的第一散列值包括对n-gram应用具有第一随机种子值的散列函数,并且获得n-gram的第二散列值包括对n-gram应用具有第二随机种子值的散列函数,第二种子值不同于第一种子值。

[0010] 在一些实施例中,使用嵌入层获得有序的多个嵌入向量包括,对于n-gram序列中的每个n-gram:对第一散列值应用模函数以获得第一索引并且对第二散列值应用模函数以获得第二索引,其中,选择第一分量向量基于第一索引并且选择第二分量向量基于第二索引。

[0011] 在一些实施例中,对于n-gram序列中的每个n-gram,获得n-gram的嵌入向量包括级联第一分量向量和第二分量向量。

[0012] 在一些实施例中,对n-gram序列的嵌入向量使用包括具有注意力机制的经训练的卷积神经网络的深度网络,以考虑n-gram在文本单元中出现的顺序并产生表示文本单元的最终编码向量。

[0013] 在一些实施例中,分类器包括前馈神经网络。在一些实施例中,对于文本单元的编码向量,使用分类器包括对前馈神经网络的最终层的输出应用softmax函数。

[0014] 在各种实施例中,提供了一种装置,该装置包括用于执行本文所公开的一种或多种方法中的部分或全部的处理电路系统,以及耦接到该处理电路系统并用于存储n-gram序列的存储器。

[0015] 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一个或多个非暂态计算机可读介质,该一个或多个非暂态计算机可读介质存储有指令,这些指令在由一个或多个数据处理器执行时使一个或多个数据处理器执行本文所公开的一种或多种方法中的部分或全部。

[0016] 在各种实施例中,提供了一种有形地体现在一个或多个非暂态机器可读介质中的计算机程序产品,该计算机程序产品包括被配置为使一个或多个数据处理器执行本文所公开的一种或多种方法中的部分或全部的指令。

[0017] 可以用多种方式并且在多种上下文中实施上文和下文所描述的技术。如下文更详细地描述的,参考以下附图提供了几种示例实施方式和上下文。然而,以下实施方式和上下文仅是许多实施方式和上下文中的一些。

附图说明

[0018] 图1是并入了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施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施技能机器人的计算系统的简化框图。

[0021] 图4示出了根据各种实施例的模型架构的一个示例。

[0022] 图5示出了根据各种实施例的模型架构的另一示例。

[0023] 图6示出了根据各种实施例的对图5的模型架构进行修改的示例。

[0024] 图7示出了根据各种实施例的对API的请求的示例。

[0025] 图8示出了根据各种实施例的来自API的回应的示例。

[0026] 图9示出了描述OPUS源数据集的表。

[0027] 图10示出了根据各种实施例的来自语言检测测试的结果。

[0028] 图11示出了根据各种实施例的装置的框图。

[0029] 图12示出了根据各种实施例的可以包括在装置中的、具有注意力机制的深度网络的示例。

[0030] 图13示出了根据各种实施例的可以由嵌入层执行的操作的示例。

[0031] 图14图示了根据各种实施例的用于语言检测的过程流程。

[0032] 图15图示了根据各种实施例的用于语言检测的过程流程。

[0033] 图16描绘了用于实施各种实施例的分布式系统的简化图。

[0034] 图17是根据各种实施例的系统环境的一个或多个部件的简化框图,通过该系统环境,由实施例系统的一个或多个部件提供的服务可以作为云服务被提供。

[0035] 图18图示了可以用于实施各种实施例的示例计算机系统。

具体实施方式

[0036] 在以下描述中,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,将显而易见的是,可以在没有这些具体细节的情况下实践各个实施例。附图和描述不旨在是限制性的。词语“示例性”在本文中用来意指“用作示例、实例或说明”。在本文中被描述为“示例性”的任何实施例或设计不必被解释为比其他实施例或设计更优选或有利。

[0037] 如本文所使用的,当动作“基于(based on)”某物时,这意味着该动作至少部分地基于该某物的至少一部分。“基于(based on)”的使用意指开放性的和包含性的,因为“基于”一个或多个叙述的条件、项或值的过程、步骤、计算或其他动作实际上可以基于所叙述的条件、项或值之外的附加条件、项或值。如本文所使用的,术语“基本上(substantially)”、“大约(approximately)”和“约(about)”被定义为如本领域普通技术人员所理解的那样,很大程度上但不一定完全是所指定的内容(并且完全包括所指定的内容)。在任何公开的实施例中,术语“基本上”、“大约”或“约”可以用“在所指定内容的[百分比]内”来代替,其中,该百分比包括0.1%、1%、5%和10%。

[0038] 介绍

[0039] 先前的工作已经将语言检测任务制定为文本分类任务。一种这样的方法利用传统的基于特征的机器学习(例如,具有n-gram特征的朴素贝叶斯)来处置该任务。另一种这样的方法利用了深度学习(例如,卷积神经网络(CNN)或长短期记忆(LSTM)网络)来处置该任务。

[0040] 本文描述的技术包括利用具有n-gram特征的注意力CNN(也就是说,具有注意力机制的CNN)来处置语言检测任务。例如,描述了一种用于利用深度学习为135种语言构建语言

检测应用编程接口(API)的方法。在公开可用的数据集上进行的实验表明,这种模型的工作效果与fastText、Google和Microsoft语言检测API相比极具竞争力或甚至更好。

[0041] 自然语言处理有许多应用。例如,数字助理是帮助用户使用自然语言会话完成各种任务的人工智能驱动接口。针对每个数字助理,客户可以组装一个或多个技能。技能(在本文中也被描述为聊天机器人、机器人或技能机器人)是重点关注如跟踪库存、提交时间卡、订购披萨、取得银行信息和创建费用报告等特定类型的任务的单独的计算机程序。为了执行任务,机器人可以与最终用户进行会话。机器人通常可以通过使用自然语言消息的消息传递应用程序对自然语言消息(例如,问题或评论)作出回应。企业可以使用一个或多个机器人系统通过消息传递应用程序与最终用户交流。消息传递应用程序(其可以被称为通道)可以是最终用户已经安装并且熟悉的最终用户优选的消息传递应用程序。因此,为了与机器人系统聊天,最终用户不需要下载并安装新的应用程序。消息传递应用程序可以包括例如过顶(OTT)消息传递通道(如Facebook Messenger、Facebook WhatsApp、微信、Line、Kik、Telegram、Talk、Skype、Slack或SMS)、虚拟私人助理(如Amazon Dot、Echo或Show、Google Home、Apple HomePod等)、本地或混合扩展的移动和web应用程序扩展/具有聊天功能的响应式移动应用程序或web应用程序、或基于声音的输入(如具有使用Siri、微软小娜(Cortana)、谷歌之音(Google Voice)或用于交互的其他语音输入的设备或应用程序)。

[0042] 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可以使用一串字符标识机器人系统。URI可以用作一个或多个消息传递应用程序系统的webhook(网络挂接)。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程序系统接收消息(例如,超文本传送协议(HTTP)post调用消息)。HTTP post调用消息可以涉及来自消息传递应用程序系统的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将机器人系统接收到的通信称作消息,但是应理解,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。

[0043] 最终用户可以通过会话式交互(有时称为会话式用户接口(UI))与机器人系统交互,正如人之间的交互那样。在一些情况下,交互可以包括最终用户对机器人说“Hello(你好)”以及机器人用“Hi(嗨)”回应并询问最终用户机器人可以如何提供帮助。在一些情况下,交互还可以是与例如银行业务机器人的交易式交互,如将钱从一个账户转移到另一个账户;与例如HR机器人的信息式交互,如查询假期结余;或与例如零售机器人的交互,如讨论退回已购商品或寻求技术支持。

[0044] 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交互的情况下智能地处置最终用户交互。例如,最终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。消息可以包括某种内容,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,利用适当的参数针对企业服务的表述性状态转移(REST)调用)并且生成自然语言回应。机器人系统还可以提示最终用户输入另外的输入参数或请求其他另外的信息。在一些实施例中,机器人系统还可以发起与最终用户的交流,而不是被动地回应最终用户话语。本文所描述的是用于识别机器人系统的显式调用并确定被调用的机器人系统的输入的各种技术。在某些实施例中,显式调用分析是由主机器人基于检测到话语中的调用名称而执行的。响应

于检测到调用名称,话语可以细化以供输入到与调用名称相关联的技能机器人。

[0045] 与机器人的会话可以遵循包括多个状态的特定会话流。所述流可以基于输入来定义接下来会发生什么。在一些实施例中,可以使用包括用户定义的状态(例如,最终用户意图)和在状态中或状态之间要采取的动作的状态机来实施机器人系统。会话可以基于最终用户输入来采用不同的路径,这可能会影响机器人针对所述流作出的决定。例如,在每个状态下,基于最终用户输入或话语,机器人可以确定最终用户的意图以便确定要采取的下一个适当动作。如本文所使用的且在话语的上下文中,术语“意图”是指提供话语的用户的意图。例如,用户可以打算让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“Order pizza(订购披萨)”来表示。用户意图可以涉及用户希望聊天机器人代表用户执行的特定任务。因此,话语可以表达为反映用户的意图的问题、命令、请求等。意图可以包括最终用户想要完成的目标。

[0046] 在聊天机器人的配置的上下文中,本文所使用的术语“意图”是指用于将用户的话语映射到聊天机器人可以执行的特定任务/动作或特定种类的任务/动作的配置信息。为了区分话语的意图(即,用户意图)与聊天机器人的意图,本文中有时将后者称为“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下订单购买披萨的期望的话语的各种排列。这些相关联的话语可以用于训练聊天机器人的意图分类器以使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相匹配。机器人意图可以与用于启动与用户在某个状态中的会话的一个或多个对话流相关联。例如,针对订购披萨意图的第一消息可以是问题“What kind of pizza would you like?(你想要哪种披萨?)”。除了相关联的话语之外,机器人意图可以进一步包括与意图有关的命名实体。例如,订购披萨意图可以包括用于执行订购披萨的任务的变量或参数,例如馅料1、馅料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交谈获得的。

[0047] 图1是并入了根据某些实施例的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,所述DABP使得DABP 102的用户能够创建并部署数字助理或聊天机器人系统。DABP 102可以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,代表特定企业的用户104可以使用DABP 102来创建并部署用于特定企业的用户的数字助理106。例如,DABP 102可以被银行用来创建供银行的客户使用的一个或多个数字助理。多个企业可以使用同一个DABP 102平台来创建数字助理。作为另一示例,餐馆(例如,披萨店)的所有者可以使用DABP 102来创建并部署使餐馆的客户能够订购食物(例如,订购披萨)的数字助理。

[0048] 出于本公开的目的,“数字助理”是通过自然语言会话帮助数字助理的用户完成各种任务的实体。可以仅使用软件(例如,数字助理是使用可由一个或多个处理器执行的程序、代码或指令实施的数字实体)、使用硬件、或使用硬件和软件的组合来实施数字助理。数字助理可以在如计算机、移动电话、手表、器具、车辆等各种物理系统或设备中体现或实施。数字助理有时也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。

[0049] 数字助理(如使用DABP 102构建的数字助理106)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获得返回的回应112。会话可以包括输入110和回应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执行一个或多个任务,并且作为回应,数字助理被配置为执行用户请求的任务并以适当的回应来回应用户。

[0050] 用户输入110通常呈自然语言的形式并且被称为话语。用户话语110可以是文本形式,如当用户键入句子、问题、文本片段或甚至单个单词并将文本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的形式,如当用户讲出或说出某些内容作为输入提供给数字助理106时。话语通常呈用户108说出的语言的形式。例如,话语可以是英语或某种其他语言。当话语呈语音形式时,将语音输入转换成该特定语言的文本形式的话语,并且然后由数字助理106来处理文本话语。可以使用各种语音到文本处理技术将语音或音频输入转换成文本话语,所述文本话语然后由数字助理106来处理。在一些实施例中,语音到文本的转换可以由数字助理106自身完成。

[0051] 话语(所述可以是文本话语或语音话语)可以是片段、一个句子、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。作为针对话语的NLU处理的一部分,数字助理106被配置为执行用于理解话语的含义的处理,所述处理涉及识别对应于话语的一个或多个意图和一个或多个实体。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,假设这些话语是已经由数字助理106的用户108直接提供的文本话语,或者是将输入语音话语转换为文本形式的结果。然而,这并不旨在以任何方式进行限制或约束。

[0052] 例如,用户108的输入可以通过提供如“I want to order a pizza(我想要订购披萨)”的话语来请求订购披萨。在接收到这种话语之后,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以请求关于用户期望订购的披萨的类型、披萨的大小、披萨的任何馅料等用户输入的问题来回应用户。由数字助理106提供的回应也可以是自然语言形式,并且通常以与输入话语相同的语言。作为生成这些回应的一部分,数字助理106可以执行自然语言生成(NLG)。为了用户能够经由用户与数字助理106之间的会话来订购披萨,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,并且然后在会话结束时使披萨被订购。数字助理106可以通过向用户输出指示已经订购披萨的信息来结束会话。

[0053] 在概念层级上,数字助理106响应于从用户接收的话语执行各种处理。在一些实施例中,该处理涉及一系列处理步骤或处理步骤流水线,包括例如理解输入话语的含义(有时称为自然语言理解(NLU)、确定响应于话语而要执行的动作、在适当的情况下使动作被执行、响应于用户话语生成要输出到用户的回应、向用户输出回应等。NLU处理可以包括对所接收的输入话语进行语法分析以理解话语的结构和含义、细化并重新形成所述话语以开发出所述话语的更好的可理解形式(例如,逻辑形式)或结构。生成回应可以包括使用NLG技术。

[0054] 数字助理(如数字助理106)执行的NLU处理可以包括如句子语法分析(例如,标记化、按屈折变化形式进行归类、识别句子的词性标签、识别句子中的命名实体、生成依存树来表示句子结构、将句子分成子句、分析单独的子句、解析指代、执行组块等)等各种NLP有关处理。在某些实施例中,NLU处理或其部分是由数字助理106自身执行的。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,可以通过使用语法分析器、词性标记器和/或命名实体识别器处理句子来识别输入话语句子的句法和结构。在一种实施方式中,针对英语语言,使用由斯坦福自然语言处理(NLP)小组提供的语法分析器、词性标记器和命名实体识别器来分析句子结构和句法。这些是作为斯坦福CoreNLP工具包的一部分提供的。

[0055] 虽然本公开中提供的各个示例示出了英语语言的话语,但是这仅意味着作为示例。在某些实施例中,数字助理106还能够处置除英语以外的语言的话语。数字助理106可以提供被配置用于针对不同语言执行处理的子系统(例如,实施NLU功能的部件)。这些子系统可以实施为可以使用服务调用从NLU核心服务器调用的可插单元。这使NLU处理对于每种语言而言是灵活且可扩展的,包括允许不同的处理顺序。可以为单独的语言提供语言包,其中,语言包可以登记可以从NLU核心服务器提供服务的子系统的列表。

[0056] 可以通过各种不同的通道(如但不限于经由某些应用程序、经由社交媒体平台、经由各种消息传递服务和应用程序、以及其他应用程序或通道)使数字助理(如图1中描绘的数字助理106)对其用户108而言是可用的或可访问的。单个数字助理可以为自身配置几个通道,使得单个数字助理可以同时在不同的服务上运行并通过不同的服务进行访问。

[0057] 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,这些技能是被配置为与用户交互并完成特定类型的任务(如跟踪库存、提交时间卡、创建费用报告、订购食物、查询银行账户、进行预约、购买小部件等)的单独的聊天机器人(称为技能机器人)。例如,针对图1所描绘的实施例,数字助理或聊天机器人系统106包括技能116-1、116-2等。出于本公开的目的,术语“技能(skill)”和“技能(skills)”分别与术语“技能机器人(skill bot)”和“技能机器人(skill bots)”同义地使用。

[0058] 与数字助理相关联的每个技能通过与用户的会话帮助数字助理的用户完成任务,其中,所述会话可以包括由用户提供的文本或音频输入与由技能机器人提供的回应的组合。这些回应可以呈以下形式:给用户的文本或音频消息和/或使用呈现给用户以供用户进行选择的简单的用户界面元素(例如,选择列表)。

[0059] 存在各种方式可以将技能或技能机器人与数字助理相关联或将其添加到数字助理。在一些实例中,可以由企业开发技能机器人,并且然后将其添加到使用DABP 102的数字助理。在其他实例中,可以使用DABP 102开发并创建技能机器人,并且然后将其添加到使用DABP 102创建的数字助理。在又其他实例中,DABP 102提供在线数字商店(称为“技能商店”),所述在线数字商店提供涉及各种各样的任务的多个技能。通过技能商店提供的技能还可以公开(expose)各种云服务。为了将技能添加到使用DABP 102生成的数字助理,DABP 102的用户可以经由DABP 102访问技能商店、选择所需的技能、并指示将所选技能添加到使用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添加到数字助理(例如,DABP 102的用户可以选择并复制技能商店提供的特定技能机器人、对所选技能机器人进行定制或修改并且然后将修改后的技能机器人添加到使用DABP 102创建的数字助理)。

[0060] 可以使用各种不同的架构来实施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创建并部署的数字助理可以使用主机器人/次(或子)机器人范式或架

构来实施。根据该范式,数字助理被实施为与是技能机器人的一个或多个次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的次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自身被认为充当主机器人。

[0061] 根据主-次机器人架构实施的数字助理使得数字助理的用户能够通过统一用户接口(即经由主机器人)与多种技能进行交互。当用户与数字助理互动时,主机器人会接收用户输入。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置,否则主机器人会选择适当的技能机器人来处置用户请求,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公共的单个接口与数字助理进行会话,并且仍然有能力使用被配置为执行特定任务的几个技能机器人。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特定功能的技能机器人(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)进行接口连接。以这种方式,数字助理的最终用户或消费者只需知道如何通过公共的主机器人接口访问数字助理,而在后台提供了多个技能机器人用于处置用户请求。

[0062] 在某些实施例中,在主机器人/次机器人基础设施中,主机器人被配置为了解可用技能机器人的列表。主机器人可以访问标识各种可用技能机器人的元数据,并且对于每个技能机器人,访问包括可以由技能机器人执行的任务的技能机器人的能力。在接收到话语形式的用户请求之后,主机器人被配置为从多个可用技能机器人中识别或预测可以最好地服务或处置用户请求的特定技能机器人。然后,主机器人将话语(或话语的一部分)路由到该特定技能机器人以进行进一步的处置。因此,控制从主机器人流动到技能机器人。主机器人可以支持多个输入通道和输出通道。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,如下文所讨论的,可以训练技能机器人以推断出话语的意图,并确定所推断出的意图是否与针对技能机器人所配置的意图相匹配。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指示,所述指示表明技能机器人是否已经配置有适合于处置话语的意图。

[0063] 虽然图1的实施例示出了数字助理106包括主机器人114以及技能机器人116-1、116-2和116-3,但这并不旨在是限制性的。数字助理可以包括提供数字助理的功能的各种其他部件(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)、仅以硬件或在使用软件和硬件的组合的实施方式中实施。

[0064] DABP 102提供了使DABP 102的用户能够创建数字助理(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。在一些实例中,可以通过克隆现有技能机器人来创建技能机器人,例如,克隆由技能商店提供的技能机器人。如前所述,DABP 102提供技能商店或技能目录,所述技能商店或技能目录提供了用于执行各种任务的多个技能机器人。DABP 102的用户可以从技能商店克隆技能机器人。根据需要,可以对克隆技能机器人进行修改或定制。在一些其他实例中,DABP 102的用户使用由DABP 102提供的工具和服务从头开始创建技能机器人。如前所述,DABP 102所提供的技能商店或技能目录可以提供用于执行各种任务的多个技能机器人。

[0065] 在某些实施例中,在高层级上,创建或定制技能机器人涉及以下步骤:

[0066] (1)为新技能机器人配置设置

[0067] (2)为技能机器人配置一个或多个意图

[0068] (3)针对一个或多个意图配置一个或多个实体

[0069] (4)训练技能机器人

[0070] (5)为技能机器人创建对话流

[0071] (6)根据需要将自定义部件添加到技能机器人

[0072] (7)测试并部署技能机器人

[0073] 下文简要描述了上述每个步骤。

[0074] (1)为新技能机器人配置设置——可以为技能机器人配置各种设置。例如,技能机器人设计者可以为正在创建的技能机器人指定一个或多个调用名称。然后,数字助理的用户可以使用这些调用名称来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。

[0075] (2)为技能机器人配置一个或多个意图和相关联的示例话语——技能机器人设计者为正在创建的技能机器人指定一个或多个意图(也称为机器人意图)。然后基于这些指定的意图来训练技能机器人。这些意图表示技能机器人被训练以针对输入话语推断出的类别或分类。在接收到话语之后,经训练的技能机器人推断话语的意图,其中,推断出的意图是从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来采取对该话语做出回应的适当的动作。在一些实例中,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。每个意图被赋予一个意图标识符或意图名称。例如,对于针对银行训练的技能机器人,为所述技能机器人指定的意图可以包括“CheckBalance(查询余额)”、“TransferMoney(转账)”、“DepositCheck(存款查询)”等。

[0076] 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一个或多个示例话语,这些示例话语表示并说明了意图。这些示例话语旨在表示用户可以针对该意图向技能机器人输入的话语。例如,针对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?(我的活期存款账户里有多少钱?)”、“How much money do I have in my account(我的账户里有多少钱)”等。因此,可以将典型的用户话语的各种排列指定为意图的示例话语。

[0077] 这些意图及其相关联的示例话语用作用于训练技能机器人的训练数据。可以使用各种不同的训练技术。作为该训练的结果,生成预测模型,所述预测模型被配置为将话语作为输入并输出所述预测模型针对所述话语推断出的意图。在一些实例中,将输入话语提供给意图分析引擎,所述引擎被配置为使用经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出的意图采取一个或多个动作。

[0078] (3)针对技能机器人的一个或多个意图配置实体——在一些实例中,可能需要另外的上下文来使技能机器人能够适当地回应用户话语。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,在以上示例中,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?(我的活期存款账户里有多少钱?)”均解析成相同的CheckBalance意图,但是这些

话语是请求不同事情的不同请求。为了阐明这种请求,将一个或多个实体添加到意图。使用银行业务技能机器人的示例,被称为账户类型(AccountType)的实体(所述实体定义了被称为“checking(活期存款)”和“saving(储蓄)”的值)可以使技能机器人能够对用户请求进行语法分析并适当地作出回应。在上面的示例中,虽然这些话语解析为相同的意图,但是这两个话语的与AccountType实体相关联的值是不同的。这使技能机器人能够针对这两个话语执行可能不同的动作,尽管这两个话语解析为相同的意图。可以针对为技能机器人配置的某些意图指定一个或多个实体。因此,实体用于为意图本身添加上下文。实体帮助更充分地描述意图并使技能机器人能够完成用户请求。

[0079] 在某些实施例中,有两种类型的实体:(a)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自定义实体。内置实体是可以与各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、循环时间段、货币、电话号码、URL等有关的实体。自定义实体用于更多定制的应用程序。例如,针对银行业务技能,AccountType实体可以被技能机器人设计者定义为通过检查用户输入的关键词(如活期存款、储蓄和信用卡等)来实现各种银行业务交易。

[0080] (4)训练技能机器人——技能机器人被配置为接收呈话语形式的用户输入,语法分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。如上文所指示的,为此,必须对技能机器人进行训练。在某些实施例中,基于为技能机器人配置的意图和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成其所配置的意图中的一个意图。在某些实施例中,技能机器人使用预测模型,所述预测模型是使用训练数据训练的并且允许技能机器人辨别用户说出(或者在一些情况下,正试图说出)的内容。DABP 102提供可以被技能机器人设计者用于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%)用于训练技能机器人模型并且另一部分(例如,其余20%)用于测试或验证模型。一旦进行了训练,经训练的模型(有时也称为经训练的技能机器人)便可以用于处置用户的话语并对用户的话语作出回应。在某些情况下,用户的话语可以是仅需要单一的回答并且无需另外的会话的问题。为了处置这种情况,可以为技能机器人定义Q&A(问题答案)意图。这使得技能机器人能够在不必更新对话定义的情况下输出对用户请求的回复。以与常规意图类似的方式创建Q&A意图。用于Q&A意图的对话流可能与用于常规意图的对话流不同。

[0081] (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了在响应于接收到的用户输入来解析技能机器人的不同意图时,技能机器人如何反应。对话流定义了技能机器人将采取的操作或动作,例如,技能机器人如何回应用户话语、技能机器人如何提示用户进行输入、技能机器人如何返回数据。对话流就像技能机器人所遵循的流程图。技能机器人设计者使用如markdown语言等语言来指定对话流。在某些实施例中,可以使用被称为OBotML的YAML版本来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,该模型是使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的模型。

[0082] 在某些实施例中,技能机器人的对话流定义包含以下三个部分:

[0083] (a)上下文部分

[0084] (b)默认转变部分

[0085] (c)状态部分

[0086] 上下文部分——技能机器人设计者可以在上下文部分中定义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:针对错误处置的变量、针对内置实体或自定义实体的变量、使技能机器人能够识别并保存用户偏好的用户变量等。

[0087] 默认转变部分——技能机器人的转变可以在对话流状态部分中或在默认转变部分中定义。在默认转变部分中定义的转变充当后备并且当状态内没有定义适用的转变或触发状态转变所需的条件无法得到满足时被触发。默认转变部分可以用于定义允许技能机器人得体地处置出乎意料的用户动作的路由。

[0088] 状态部分——对话流及其有关操作被定义为管理对话流内的逻辑的暂时状态的序列。对话流定义内的每个状态节点都命名一个部件,所述部件提供所述对话中的那个时候所需的功能。因此,状态是围绕部件构建的。状态包含特定于部件的性质并且定义了在部件执行之后被触发的向其他状态的转变。

[0089] 特殊情况场景可以使用状态部分来处置。例如,您有时可能想要为用户提供用于暂时让用户与之接洽的第一技能进行数字助理内的第二技能中的事情的选项。例如,如果用户忙于与购物技能进行会话(例如,用户已做出一些购买选择),则用户可能想要跳转至银行业务技能(例如,用户可能想要确保他/她有足够的钱用于购买)并且然后返回到购物技能以完成用户的订单。为了解决这一点,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并且然后返回到原始流。

[0090] (6)将自定义部件添加到技能机器人——如上文所描述的,技能机器人的对话流中指定的状态对与状态相对应的提供所需的功能的部件进行了命名。部件使技能机器人能够执行功能。在某些实施例中,DABP 102提供了用于执行各种各样的功能的一组预先配置的部件。技能机器人设计者可以选择这些预先配置的部件中的一个或多个并将其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DABP 102提供的工具来创建自定义或新的部件并将自定义部件与技能机器人的对话流中的一个或多个状态相关联。

[0091] (7)测试并部署技能机器人—DABP 102提供了使技能机器人设计者能够测试正在开发的技能机器人的几个特征。然后,可以将技能机器人部署并包括在数字助理中。

[0092] 虽然以上描述描述了如何创建技能机器人,但是还可以使用类似的技术来创建数字助理(或主机器人)。在主机器人或数字助理层级,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自身(即,主机器人)可以处置的一般任务。针对主机器人定义的系统意图的示例包括:(1)Exit(退出):当用户发出想要退出数字助理中的当前会话或上下文的信号时适用;(2)Help(帮助):当用户请求帮助或定向时适用;以及(3)UnresolvedIntent(未解析意图):适用于与退出意图和帮助意图不太匹配的用户输入。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。该信息使主机器人能够选择用于处置话语的特定技能机器人。

[0093] 在主机器人或数字助理层级,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用于确定如何路由话语和相关会话的处理。数字助理使用路由模型来确定这一点,所述路由模型可以是基于规则的、基于AI的或其组合。数字助理使用路由模型来确定对

应于用户输入话语的会话是要被路由到特定技能以进行处置、要由数字助理或主机器人自身按照内置系统意图进行处置、还是要被处置为当前会话流中的不同状态。

[0094] 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显式地识别技能机器人。如果调用名称存在于用户输入中,则所述调用名称被视为对应于调用名称的对技能机器人的显式调用。在这种场景下,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进行进一步处置。在某些实施例中,如果不存在特定调用或显式调用,则数字助理评估接收到的用户输入话语并针对与数字助理相关联的系统意图和技能机器人计算置信度得分。针对技能机器人或系统意图计算的得分表示用户输入表示技能机器人被配置为执行的任务或表示系统意图的可能性有多大。选择相关联的计算的置信度得分超过阈值(例如,置信度阈值路由参数)的任何系统意图或技能机器人作为候选以进行进一步的评估。然后,数字助理从所识别的候选中选择特定的系统意图或技能机器人用于对用户输入话语进行进一步的处置。在某些实施例中,在一个或多个技能机器人被识别为候选之后,对与那些候选技能相关联的意图进行评估(根据每个技能的意图模型)并且针对每个意图确定置信度得分。置信度得分超过阈值(例如,70%)的任何意图通常均被视为候选意图。如果选择了特定技能机器人,则将用户话语路由到该技能机器人以进行进一步的处理。如果选择了系统意图,则由主机器人自身根据所选系统意图执行一个或多个动作。

[0095] 图2是根据某些实施例的主机器人(MB)系统201的简化框图。MB系统201可以仅以软件、仅以硬件、或以硬件和软件的组合实施。MB系统201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和数据存储250。图2中描绘的MB系统201仅仅是主机器人中部件布置的示例。本领域普通技术人员将认识到许多可能的变化、替代方案和修改。例如,在一些实施方式中,MB系统201可以具有比图2所示的那些系统或部件更多或更少的系统或部件,可以组合两个或更多个子系统,或者可以具有不同的子系统配置或布置。

[0096] 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和语言语法分析器214来处理话语。如上文所指示的,可以以包括音频或文本的各种方式来提供话语。话语202可以是句子片段、完整句子、多个句子等。话语202可以包括标点符号。例如,如果话语202作为音频提供,则预处理子系统210可以使用将标点符号(例如,逗号、分号、句号等)插入到所得文本中的语音文本转换器(未示出)来将音频转换为文本。

[0097] 语言检测器212基于话语202的文本来检测话语202的语言。话语202的处置方式取决于语言,因为每种语言都有其自己的语法和语义。在分析话语的句法和结构时,会考虑语言之间的差异。

[0098] 语言语法分析器214对话语202进行语法分析以提取话语202中各个语言单元(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。语言语法分析器214还可以对话语202的语言单位进行标记化(例如,将每个单词转换为单独的记号)并对单词按屈折变化形式进行归类。词元是如在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)。语言语法分析器214可以执行的其他类型的预处理包括复合表达式的组块,例如,将“credit(信用)”和“card(卡)”组合成单个表达式“credit_card(信用卡)”。语言语法分析器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,语言语法分析器214生成依存树,该依存树指示话语的哪一部分(例如,特定名词)是直接宾语、话语的哪一部分是介词等。由语言语法分析器214执行的处理的结果形成提取的信息205,并与话语202本身一起作为输入提供给MIS220。

[0099] 如上文所指示的,话语202可以包括多于一个句子。为了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。通常,无论话语202是在单个句子的级别上处理还是作为包括多个句子的单个单元处理,MIS220和EIS230产生的结果都基本相同。

[0100] MIS220确定话语202是否表示多个意图。尽管MIS220可以检测到话语202中存在多个意图,但是由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相匹配。相反,用于确定话语202的意图是否与机器人意图相匹配的处理可以由MB系统201的意图分类器242或技能机器人的意图分类器来执行(例如,如图3的实施例所示)。MIS220执行的处理假设存在可以处置话语202的机器人(例如,特定技能机器人或主机器人本身)。因此,由MIS220执行的处理不需要知道聊天机器人系统中有哪些机器人(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。

[0101] 为了确定话语202包括多个意图,MIS220应用数据存储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存在多个意图的句子模式。例如,句子模式可以包括连接句子的两个部分(例如,连词)的并列连接词,其中,这两个部分对应于不同的意图。如果话语202与句子模式相匹配,则可以推断出话语202表示多个意图。应该注意的是,具有多个意图的话语不一定具有不同的意图(例如,涉及不同机器人的意图或同一机器人内的不同意图)。相反,话语可以有相同意图的不同实例,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)”。

[0102] 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。MIS220针对包含多个意图的话语中表示的每个意图构建用于单独处理的新话语来代替原始话语,例如,如图2中所描绘的话语“B”206和话语“C”208。因此,原始话语202可以被拆分成一次处置一个的两个或更多个单独话语。MIS220使用提取的信息205和/或根据对话语202本身的分析来确定应该首先处置两个或更多个话语中的哪一个。例如,MIS220可以确定话语202包含指示应该首先处置特定意图的标记词。与该特定意图相对应的新形成的话语(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。在由第一个话语触发的会话已经结束(或已被暂时暂停)之后,然后可以将下一个最高优先级的话语(例如,话语206或话语208中的另一个话语)发送到EIS230进行处理。

[0103] 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,给聊天机器人系统中的每个技能机器人指配将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的唯一调用名称。调用名称列表可以作为技能机器人信息254的一部分保存在数据存储250中。当话语包含与调用名称相匹配的单词时,话语被视为显式调用。如果机器人未被显式调用,则EIS230接收到的话语被视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)以确定使用哪个机器

人来处置话语。在一些实例中,意图分类器242将确定主机器人应该处置非显式调用话语。在其他实例中,意图分类器242将确定将话语路由到以进行处置的技能机器人。

[0104] EIS230所提供的显式调用功能有几个优点。它可以减少主机器人不得不执行的处理量。例如,当存在显式调用时,主机器人可能不必(例如,使用意图分类器242)进行任何意图分类分析,或者可能不得不进行简化的意图分类分析以选择技能机器人。因此,显式调用分析可以实现选择特定技能机器人而无需诉诸意图分类分析。

[0105] 而且,可能存在多个技能机器人之间的功能重叠的情况。例如,如果两个技能机器人处置的意图重叠或彼此非常接近,则可能会发生这种情况。在这种情况下,主机器人可能难以仅基于意图分类分析来识别要选择多个技能机器人中的哪一个。在这种场景下,显式调用使得对于要使用的特定技能机器人没有歧义。

[0106] 除了确定话语是显式调用之外,EIS230还负责确定是否应该将话语的任何部分用作对被显式调用的技能机器人的输入。具体地,EIS230可以确定话语的一部分是否与调用无关。EIS230可以通过分析话语和/或分析提取的信息205来执行该确定。EIS230可以将与调用无关的话语部分发送到被调用的技能机器人,而不是发送EIS230所接收到的整个话语。在一些实例中,被调用的技能机器人的输入简单地通过删除与调用相关联的任何话语部分来形成。例如,“I want to order pizza using Pizza Bot(我想要使用披萨机器人订购披萨)”可以缩短为“I want to order a pizza(我想要订购披萨)”,因为“using Pizza Bot(使用披萨机器人)”与披萨机器人的调用相关,但与要由披萨机器人执行的任何处理无关。在一些情况下,EIS230可以重新格式化要发送给被调用机器人的部分,例如以形成完整的句子。因此,EIS230不仅确定存在显式调用,而且当存在显式调用时确定要向技能机器人发送什么。在一些实例中,可能没有任何文本可以输入到被调用的机器人中。例如,如果话语是“Pizza Bot(披萨机器人)”,则EIS230可以确定披萨机器人正在被调用,但没有文本要由披萨机器人处理。在这种场景下,EIS230可以向技能机器人调用器240指示没有要发送的内容。

[0107] 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特定技能机器人的指示235来调用机器人。指示235可以由EIS230连同用于显式调用的技能机器人的输入一起发送。在这种场景下,技能机器人调用器240会将会话的控制权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来自EIS230的输入视为独立话语来确定对该输入的适当回应。例如,回应可以是执行特定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。

[0108] 技能机器人调用器240可以调用技能机器人的另一种方式是通过使用意图分类器242进行隐式调用。可以使用机器学习和/或基于规则的训练技术来训练意图分类器242,以确定话语表示特定技能机器人被配置为执行的任务的可能性。在不同的分类上训练意图分类器242,每个技能机器人一个分类。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列表可以用于训练意图分类器242以确定特定话语表示新技能机器人可以执行的任务的可能性。作为该训练的结果产生的参数(例如,机器学习模型的一组参数值)可以存储为技能机器人信息254的一部分。

[0109] 在某些实施例中,意图分类器242是使用机器学习模型实施的,如本文进一步详细

描述的。机器学习模型的训练可以涉及至少输入来自与各种技能机器人相关联的示例话语的话语子集,以生成关于哪个机器人是用于处置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事实信息。然后可以适配机器学习模型的行为(例如,通过反向传播)以最小化所生成的推断与基本事实信息之间的差异。

[0110] 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置话语(例如,从EIS230接收的非显式调用话语234)的可能性的置信度得分。意图分类器242还可以确定已配置的每个系统级意图(例如,帮助、退出)的置信度得分。如果特定置信度得分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得分相关联的机器人。例如,可能需要满足阈值置信度得分值。因此,意图分类器242的输出245是系统意图的识别或对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得分值之外,置信度得分还必须超过次高置信度得分一定的裕量。当多个技能机器人的置信度得分均超过阈值置信度得分值时,施加这样的条件将能够路由到特定技能机器人。

[0111] 在基于置信度得分的评估识别出机器人之后,技能机器人调用器240将处理移交给所识别的机器人。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。进一步地,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。如上所述,在显式调用的情况下,输入247可以基于与调用无关的话语部分,或者输入247可以是什么都没有(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。

[0112] 数据存储250包括一个或多个计算设备,该一个或多个计算设备存储由主机器人系统201的各种子系统使用的数据。如上文所解释的,数据存储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MIS220确定话语何时表示多个意图以及如何拆分表示多个意图的话语的规则。规则252进一步包括用于由EIS230确定要将显示调用技能机器人的话语的哪些部分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列表。技能机器人信息254还可以包括意图分类器242用来确定聊天机器人系统中每个技能机器人的置信度得分的信息,例如,机器学习模型的参数。

[0113] 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是可以仅以软件实施、仅以硬件实施、或以硬件和软件的组合实施的计算系统。在某些实施例中,如图1所描绘的实施例中,技能机器人系统300可以用于在数字助理内实施一个或多个技能机器人。

[0114] 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类似于图2中的MIS220并提供类似的功能,包括可操作地使用数据存储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)如何将话语拆分为针对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拆分话语的规则与MIS220所应用的规则相同。MIS 310接收话语302和提取的信息304。提取的信息304类似于图1中的提取的信息205,并且可以使用语言语法分析器214或技能机器人系统300本地的语言语法分析器来生成。

[0115] 意图分类器320可以以与上文结合图2的实施例讨论的意图分类器242类似的方式被训练,并在此进一步详细描述。例如,在某些实施例中,意图分类器320是使用机器学习模型来实施的。针对特定技能机器人,至少使用与该特定技能机器人相关联的示例话语的子集作为训练话语来训练意图分类器320的机器学习模型。每个训练话语的基本事实将是与训练话语相关联的特定机器人意图。

[0116] 话语302可以直接从用户接收或通过主机器人提供。当通过主机器人提供话语302时,例如,作为通过图2中描绘的实施例中的MIS220和EIS230处理的结果,可以绕过MIS 310以避免重复已经由MIS220执行的处理。然而,如果直接从用户接收到话语302,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以将话语302拆分为针对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以进行意图分类,而不拆分话语302。

[0117] 意图分类器320被配置为将接收到的话语(例如,话语306或308)与同技能机器人系统300相关联的意图相匹配。如上文所解释的,技能机器人可以配置有一个或多个意图,每个意图包括与所述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201的意图分类器242被训练以确定各个技能机器人的置信度得分和针对系统意图的置信度得分。类似地,可以训练意图分类器320以确定针对与技能机器人系统300相关联的每个意图的置信度得分。由意图分类器242执行的分类是在机器人级别,而由意图分类器320执行的分类是在意图级别并且因此更细粒度。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包括表示意图并说明意图的含义并且通常与可由该意图执行的任务相关联的话语列表。意图信息354可以进一步包括作为在该话语列表上进行训练的结果而产生的参数。

[0118] 会话管理器330接收特定意图的指示322作为意图分类器320的输出,所述意图由意图分类器320识别为与输入到意图分类器320的话语最佳匹配。在一些实例中,意图分类器320不能确定任何匹配。例如,如果话语涉及系统意图或不同技能机器人的意图,则由意图分类器320计算的置信度得分可能低于阈值置信度得分值。当这种情况发生时,技能机器人系统300可以将话语提交给主机器人进行处置,例如,路由到不同的技能机器人。然而,如果意图分类器320成功地识别出技能机器人内的意图,则会话管理器330将发起与用户的会话。

[0119] 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施。状态机可以包括默认起始状态(例如,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,每个状态与将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、回应)相关联。因此,会话管理器330可以在接收到识别到意图的指示322时确定动作/对话335,并且可以响应于在会话期间接收到的后续话语来确定附加动作或对话。

[0120] 数据存储350包括一个或多个计算设备,其存储由技能机器人系统300的各种子系统使用的数据。如图3所描绘的,数据存储350包括规则352和意图信息354。在某些实施例中,数据存储350可以集成到主机器人或数字助理的数据存储中,例如图2中的数据存储

250。

[0121] 用于语言检测的模型架构

[0122] 图4示出了根据各种实施例的模型架构的一个示例400。在该示例中,输入文本单元在单词级被拆分成n-gram,并且文本单元中的每个单词也被拆分成基于字符的n-gram,以产生文本单元的n-gram序列(例如,根据n-gram在输入文本单元中出现的顺序)。在单词级将文本单元拆分成n-gram和在字符级将文本单元拆分成n-gram的操作可以串行执行(例如,单词级拆分之后进行字符级拆分)或并行执行。

[0123] 单词级n-gram的n值可以与字符级n-gram的n值相同或不同。在图4的示例400中,单词级n-gram的n值为一(基于单词的unigram),而字符级n-gram的n值为二(基于字符的bi-gram)。使用该方案,将文本单元“hello there”转换(在语法分析器410处)为n-gram序列[hello,_h,he,el,ll,lo,o_,there,_t,th,he,er,re,e_](可替代地,n-gram序列[_h,he,el,ll,lo,o_,hello,_t,th,he,er,re,e_,there])。应当注意,在该示例中,为了指示单词边界,在将文本单元的每个单词拆分为字符级n-gram之前,在单词的开头和结尾处添加了特殊字符(例如,下划线字符“_”)。

[0124] 将n-gram序列中的每个n-gram馈送到嵌入层420中以生成n-gram的对应表示(例如,特征向量或“嵌入向量”)。嵌入层420包括经训练的嵌入模型(例如,将每个n-gram与对应嵌入向量相关联的嵌入矩阵)。采用具有注意力机制430的CNN来捕获n-gram特征之间的关系(如可以例如由如n-gram在序列中的相对顺序和/或相对权重等方面来指示)以产生文本单元的编码向量。使用前馈网络(FFN)440和softmax激活函数对编码向量进行分类以产生输出预测(例如,对输入文本单元的预测主导语言的识别)。

[0125] 在上文参考图4描述的示例400中,词汇量是十三:两个单词级uni-gram(“hello”、“there”)和十一个字符级bi-gram(“_h”、“he”、“el”、“ll”、“lo”、“o_”、“_t”、“th”、“er”、“re”、“e_”)。实际上,用于语言检测任务的一组训练数据很大,并且可以表示超过一百种可用语言,因此词汇表通常是巨大的(例如,内部数据集有1.05亿个句子,词汇量超过3000万)。例如,词汇表可以包括一整套日语或汉语n-gram、另一套越南语n-gram、另一套英语n-gram等。即使嵌入向量空间的维数相对较小(例如,几十或几百维),嵌入模型中的参数数量也可能非常大,并且查找可能变得非常慢。

[0126] 散列嵌入可以用于减少词汇量,从而减少嵌入模型中的参数数量。图5示出了图4的模型架构的实施例500,其中嵌入层520通过将来自语法分析器410的每个n-gram转换(在散列操作515处)成对应散列标识符(ID)来采用散列嵌入,该散列ID然后被输入到嵌入模型525以获得对应的n-gram表示(例如,嵌入向量)。该示例中的散列ID的范围是零到九,因此词汇量从图4中的13减少到本例中的固定数量10。与图4所示的示例400一样,可以采用具有注意力模型530的CNN来捕获n-gram特征之间的关系以产生文本单元的编码向量,并且可以使用FFN 540和softmax激活函数对编码向量进行分类以产生输出(语言)预测。

[0127] 在一个示例中,散列函数的输出范围的大小等于期望的词汇量,并且通过对n-gram应用散列函数来直接从对应的n-gram中获得每个散列ID。换句话说,n-gram的散列ID是通过对n-gram应用散列函数而产生的散列值。在另一示例中,通过对n-gram应用散列函数,然后对所得散列值应用模B函数来获得每个散列ID,其中,B是期望的词汇量。例如,在图5中,散列操作515可以通过以下方式来获得每个散列ID:对n-gram应用MurmurHash算法版

本(例如,MurmurHash1、MurmurHash2或MurmurHash3)以获得对应散列值(例如,32位散列值),然后对散列值应用模10函数以获得对应散列ID。

[0128] 因为词汇量小于唯一n-gram的数量,所以散列嵌入会导致冲突。例如,如图5所示,n-gram“hello”和“11”具有相同的散列ID 1。可以使用布隆嵌入算法来大大降低冲突的发生率。具体地,每个n-gram可以被映射到两个(或更多个)散列ID,而不是将n-gram映射到单个散列ID。任何两个n-gram的两个(或所有)散列ID都相同的概率比两个n-gram映射到同一散列ID的概率小得多。

[0129] 图6示出了示例600,其中图5的模型架构的散列操作515的实施方式615执行布隆嵌入来为每个n-gram产生两个散列ID。嵌入层520的该实施方式620还包括嵌入模型525的实施方式625,该实施方式为两个散列ID中的每个散列ID产生嵌入向量。将散列ID的嵌入向量组合起来以获得n-gram的嵌入向量,该嵌入向量被输入到深度学习编码器630。与图5中一样,词汇量被设为期望的词汇量B(在该示例中为10),但是在本例中,散列桶中没有冲突。与图4所示的示例400一样,可以采用具有注意力模型630的CNN来捕获n-gram特征之间的关系以产生文本单元的编码向量,并且可以使用FFN 640和softmax激活函数对编码向量进行分类以产生输出(语言)预测。

[0130] 据发现,与基于字符的宽模型(该宽模型也需要大得多的CNN)相比,如上文参考图4至图6所述的基于n-gram的宽模型表现得更好。n-gram模型可以被实施为在深度网络430(530,630)的输入处包括查找层(在例如log(n)时间内执行),然后在顶部包括小的CNN层,这导致运行时间比其中深度网络430(530,630)在顶部具有非常大的CNN层的基于字符的模型好得多。

[0131] 如图5或图6所示的模型架构可以包括几个可调超参数。例如,可能期望将基于单词的n-gram的n值设为1(unigram),并为基于字符的n-gram选择三个n值,即2、3和4。对于散列嵌入,可能期望将桶(bucket)的数量(B)设为三百万(3M)并将散列的数量设为二以处置冲突问题。可以执行网格搜索来确定CNN窗口大小和丢弃概率的值。可能会发现超参数设置之间的差异并不显著。为了加速训练,可能期望为每个句子设置n-gram的最大数量(例如,512)。

[0132] 在训练期间,可以应用输入字符的脚本信息(例如,拉丁文、梵文等)来限制预测的候选。例如,如果输入文本单元的编码仅为CJK(汉语、日语和韩语)脚本(例如,如文本单元的统一码编码所指示的),则可以阻止对该文本单元进行基于拉丁语的语言的任何预测。另外或可替代地,由于单词可以用于许多语言(例如,单词“estas”可以用于西班牙语和世界语),因此可能期望将语言的相对流行度整合到模型预测中。例如,可以将更高的权重应用于对更流行语言的预测。

[0133] 出于比较的目的,上文参考图5和图6描述的模型架构分别被指定为“ODA单一型API”和“ODA API”。为了演示和评估这些模型架构,使用FastAPI web框架来构建RESTful服务应用编程接口(API)(即,符合表述性状态转移(REST)约束条件的API)。图7示出了对API的请求的示例,而图8示出了来自API的对应回应的示例。

[0134] 训练和评估

[0135] 如上所述的模型架构的训练数据是从开放并行语料库项目(OPUS)和Common Crawl数据以及维基百科中导出的。图9示出了描述OPUS源数据集的表。从经清理的Common

Crawl数据中获得的数据集包括来自176种语言的文本,其中包括来自165种语言中的每种语言的超过一千(1K+)个词元、来自127种语言中的每种语言的超过一百万(1M+)个词元,以及来自40种语言中的每种语言的超过十亿(1B+)个词元。

[0136] 可以假设,如果该模型对于短文本单元的效果很好,则它对于长文本单元的效果也会很好。从OPUS数据集中提取了短文本单元进行训练,包括多达10M个短句(<15个单词)和多达1M个长句(>=15个单词且<30个单词)。由于一组Common Crawl数据很大,首先提取所有不包含数字或特殊字符的页面标题,然后从正文内容中提取句子,每种语言的句子限制为1.5M个。所得的一组训练数据包括来自135种语言的文本。因为由于缺少训练数据,最终移除了135种语言中的35种语言,所以在该示例中总共支持100种语言。

[0137] 选择以下系统进行比较:

[0138] 1)FastText支持超过170种语言,可免费访问,并且据发现比其他免费语言检测工具包(例如,langdetect、langid、谷歌的Compact Language Detector 2(cld2)、谷歌的Compact Language Detector 3(cld3))表现得更好。

[0139] 2)谷歌语言检测API(支持109种语言的版本)。

[0140] 3)微软语言检测API(支持92种语言的版本)。

[0141] 4)亚马逊语言检测API(支持104种语言的版本)。

[0142] 如图6所示的模型架构的两个变体(记为“ODA API”)被用作基线。第一变体(记为“CNN API”)仅使用CNN,没有注意力机制。第二变体(记为“AVGAPI”)省略了CNN层,而是使用平均池化层(例如,如在fastText中)。基于字符的CNN模型(“Char-CNN”)也被用作纯深度学习基线(例如,以确定深度神经网络(DNN)与宽特征的组合是否比纯DNN表现更好)。为了模拟宽特征,使用了具有多核窗口大小的CNN。选择了所有比较系统(包括fastText、谷歌API和微软API)都支持的一组69种重叠语言。

[0143] 图10示出了使用ODA数据集(包含335051(335K)个话语)进行验证和早停的语言检测测试的结果。该实验被视为消融研究(例如,移除系统部件的研究)。据发现,CNN和ODA API的表现优于AGV API,这可能表明CNN层对模型的成功很重要。据发现,与CNN API相比,ODA API(具有注意力层)在验证集上表现得更好。据发现,与ODA单一型API(如图5中)相比,ODA API(包括如图6中的布隆嵌入)在验证集上表现得更好。应当注意,模型之间的参数数量没有显著差异。

[0144] Chatterbot数据集上的结果显示,该API的表现与谷歌API、微软API和亚马逊API的表现相当。还获得EuroParl数据集(摘自欧洲议会会议录并包括21种欧洲语言的句子对齐文本)、wiLI-2018数据集(维基百科的短文本摘录的数据集,包含235种语言的235000个段落)、LanideNN数据集、第一次简单英语测试(ODA-10K个示例)和第二次简单英语测试(336个示例)上的结果。尽管据发现fastText在免费语言检测工具当中表现最佳,但是据发现它在公开可用数据集上的表现比所有商业API都差。在我们的API(ODA、CNN和AVG)当中,结合了CNN和注意力层的ODA在除了简单英语测试之外的所有数据集上的表现都是最佳的。我们的ODA API在Chatterbot数据集和我们内部的简单英语测试上的表现与谷歌API、微软API和亚马逊API的表现也相当。对于其他数据集(例如,LanideNN、EuroParl),据发现,我们的ODA API比谷歌API和微软API表现得更好。还发现,宽特征和DNN的组合比纯DNN表现得更好。

[0145] 用于语言检测的技术

[0146] 图11示出了根据各种实施例的装置1100的框图。图11中描绘的元件可以以由相应系统的处理电路系统(例如,一个或多个处理单元,如处理器和/或核)执行的软件(例如,代码、指令、模块、程序)、硬件或其组合实施并且耦接到存储器(以存储例如文本单元、n-gram序列和/或经训练的网络的参数)。装置1100包括语法分析器1110,该语法分析器接收文本单元作为输入,并产生文本单元的对应的n-gram序列(例如,如上文参考拆分410所述)。对应的n-gram序列可以包括单词级n-gram和/或字符级n-gram,并且多个字符级n-gram的n值可以与多个单词级n-gram的n值相同或不同。在一个示例中,对应的n-gram序列包括单词级uni-gram和字符级bi-gram。

[0147] 装置1100还包括嵌入层1120,该嵌入层接收n-gram序列并为n-gram序列产生对应的有序的多个嵌入向量。有序的多个嵌入向量可以是基于多个分量向量(例如,嵌入模型,如经训练的嵌入矩阵)。有序的多个嵌入向量中的嵌入向量的顺序可以指示或以其他方式对应于对应n-gram在文本单元中出现的顺序。

[0148] 装置1100还包括深度网络1130,该深度网络接收有序的多个嵌入向量并为文本单元产生编码向量。深度网络在其输入层与输出层之间包括至少一个隐藏层,并且可以包括经训练的CNN。深度网络可以包括注意力机制(例如,一个或多个注意力层),用于生成注意力权重,这些注意力权重指示例如在执行预测时哪些n-gram应得到更多的关注(例如,哪些n-gram应被赋予更高的权重)。

[0149] 图12示出了包括注意力机制的深度网络1130的另一示例1210。该机制包括注意力层1230,该注意力层被配置为向CNN层1220的输出指配注意力权重(用粗虚线表示)。输入文本的最终编码向量是CNN输出的加权和(例如,如使用注意力权重进行加权)。

[0150] 装置1100还包括分类器1140,该分类器接收编码向量并产生对文本单元的语言预测。该分类器可以包括前馈神经网络。在这种情况下,分类器可以被配置为对前馈神经网络的最终层的输出应用softmax函数(例如,如使用参考图12所述的注意力层进行加权)。

[0151] 如上所述,输入文本单元可以被语法分析成n-gram序列,该n-gram序列可以包括单词级n-gram和/或字符级n-gram,其中,n值在各自的情况下都可以是可调参数。图13示出了可以由嵌入层1120的实施方式1325执行以针对n-gram序列当中的每个n-gram获得有序的多个嵌入向量中的对应嵌入向量的操作的示例。在该示例中,对n-gram执行第一散列以获得第一散列值,并且对第一散列值应用模B运算以获得第一索引,其中,B是经训练的多个分量向量(例如,经训练的嵌入模型)中的分量向量的数量。同样,对n-gram执行第二散列以获得第二散列值,并且对第二散列值应用模B运算以获得第二索引。将由第一索引和第二索引指示的分量向量组合起来(例如,级联、加权和/或相加)以获得n-gram的嵌入向量,其中,组合操作的配置可以包括一个或多个可调参数(例如,分量向量是否被级联/加权/相加,如何确定权重等)。

[0152] 图14是图示了根据某些实施例的用于语言检测的过程1400的流程图。图14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)执行的软件(例如,代码、指令、程序)、硬件或其组合实施。软件可以存储在非暂态存储介质上(例如,存储器设备上)。图14所呈现的并且下文所描述的方法旨在是说明性的而非限制性的。尽管图14描绘了以特定次序或顺序发生的各种处理步骤,但这并不旨在是限制性的。在某些替代性实施例

中,步骤可以以某种不同的顺序执行或者一些步骤还可以并行地执行。在某些实施例中,如在图1至图3所描绘的实施例中,图14所描绘的处理可以由预处理子系统(例如,语言检测器212)执行,以生成提取信息以供一个或多个其他子系统(例如,多意图子系统220或310和/或显式调用子系统110或意图分类器320)使用。

[0153] 在框1404处,由数据处理系统(例如,分别关于图1至图3描述的聊天机器人系统106、201和/或300)获得文本单元的n-gram序列。获得n-gram序列可以包括接收文本单元作为输入,并对文本单元进行语法分析以产生n-gram序列。n-gram序列可以包括单词级n-gram和/或字符级n-gram,并且多个字符级n-gram的n值可以与多个单词级n-gram的n值相同或不同。在一个示例中,n-gram序列包括单词级uni-gram和字符级bi-gram。

[0154] 在框1408处,使用嵌入层获得n-gram序列的有序的多个嵌入向量。嵌入层包括具有多个分量向量的经训练的模型。

[0155] 在框1412处,使用深度网络获得基于有序的多个嵌入向量的编码向量。深度网络包括注意力机制(例如,一个或多个注意力层)。在各种实施例中,深度网络可以包括经训练的CNN。

[0156] 在框1416处,使用分类器获得基于编码向量的对文本单元的语言预测。在各种实施例中,分类器可以包括前馈神经网络。在这种情况下,使用分类器可以包括对前馈神经网络的最终层的输出应用softmax函数。

[0157] 图15是图示了根据某些实施例的用于语言检测的过程1500的流程图。图15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)执行的软件(例如,代码、指令、程序)、硬件或其组合实施。软件可以存储在非暂态存储介质上(例如,存储器设备上)。图15所呈现的并且下文所描述的方法旨在是说明性的而非限制性的。尽管图15描绘了以特定次序或顺序发生的各种处理步骤,但这并不旨在是限制性的。在某些替代性实施例中,步骤可以以某种不同的顺序执行或者一些步骤还可以并行地执行。在某些实施例中,如在图1至图3所描绘的实施例中,图15所描绘的处理可以由预处理子系统(例如,语言检测器212)执行,以生成提取信息以供一个或多个其他子系统(例如,多意图子系统220或310和/或显式调用子系统110或意图分类器320)使用。

[0158] 可以根据上文参考图14的对框1404、1412和1416的描述来实施框1504、1512和1516。在框1508处,使用嵌入层获得n-gram序列的有序的多个嵌入向量。嵌入层包括具有多个分量向量的经训练的模型。框1508包括框1508a至1508d,可以执行这些框以针对n-gram序列中的每个n-gram获得有序的多个嵌入向量中的对应嵌入向量。在框1508a处,获得n-gram的第一散列值和n-gram的第二散列值。例如,获得n-gram的第一散列值可以包括对n-gram应用具有第一种子值的散列函数,并且获得n-gram的第二散列值可以包括对n-gram应用具有第二种子值的散列函数,第二种子值不同于第一种子值。在框1508b处,基于第一散列值,从多个分量向量当中选择第一分量向量。在框1508c处,基于第二散列值,从多个分量向量当中选择第二分量向量。例如,过程1500可以包括对第一散列值应用模函数以获得第一索引并且对第二散列值应用模函数以获得第二索引,其中,选择第一分量向量可以是基于第一索引并且选择第二分量向量可以是基于第二索引。在框1508d处,获得n-gram的基于第一分量向量和第二分量向量的嵌入向量。例如,嵌入向量可以作为第一分量向量和第二分量向量的级联而获得。另外或可替代地,获得n-gram的嵌入向量可以包括:对第一分量向

量应用第一加权值以获得第一加权向量,以及对第二分量向量应用第二加权值以获得第二加权向量,其中,嵌入向量是基于第一加权向量和第二加权向量。

[0159] 说明性系统

[0160] 图16描绘了分布式系统1600的简化图。在图示的示例中,分布式系统1600包括经由一个或多个通信网络1610耦接到服务器1612的一个或多个客户端计算设备1602、1604、1606和1608。客户端计算设备1602、1604、1606和1608可以被配置为执行一个或多个应用程序。

[0161] 在各种示例中,服务器1612可以被适配成运行实现本公开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程序。在某些示例中,服务器1612还可以提供可以包括非虚拟环境和虚拟环境的其他服务或软件应用程序。在一些示例中,这些服务可以作为基于web的服务或云服务(如在软件即服务(SaaS)模型下)提供给客户端计算设备1602、1604、1606和/或1608的用户。操作客户端计算设备1602、1604、1606和/或1608的用户进而可以利用一个或多个客户端应用程序来与服务器1612交互以利用这些部件所提供的服务。

[0162] 在图16所描绘的配置中,服务器1612可以包括实施由服务器1612执行的功能的一个或多个部件1618、1620和1622。这些部件可以包括可以由一个或多个处理器、硬件部件或其组合执行的软件部件。应了解,可能与分布式系统1600不同的各种不同系统配置是可能的。因此,图16所示出的示例是用于实施示例系统的分布式系统的一个示例并且不旨在是限制性的。

[0163] 用户可以使用客户端计算设备1602、1604、1606和/或1608来执行一个或多个应用程序、模型或聊天机器人,该一个或多个应用程序、模型或聊天机器人可以生成然后可以根据本公开的传授内容来实施或服务的一个或多个事件或模型。客户端设备可以提供使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该接口向用户输出信息。尽管图16描绘了仅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。

[0164] 客户端设备可以包括各种类型的计算系统,如便携式手持设备、如个人计算机和膝上型计算机等通用计算机、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。这些计算设备可以运行各种类型和版本的软件应用程序和操作系统(例如,Microsoft Windows®、Apple Macintosh®、UNIX®或类UNIX操作系统、Linux或类Linux操作系统(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、智能电话(例如,iPhone®)、平板计算机(例如,iPad®)、个人数字助理(PDA)等。可穿戴设备可以包括Google Glass®头戴式显示器以及其他设备。游戏系统可以包括各种手持游戏设备、支持因特网的游戏设备(例如,有或没有Kinect®姿势输入设备的Microsoft Xbox®游戏控制台、Sony PlayStation®系统、由Nintendo®提供的各种游戏系统、以及其他)等。客户端设备可以能够执行各种不同的应用程序,如各种因特网相关应用程序、通信应用程序(例如,电子邮件应用程序、短消息服务(SMS)应用程序)并可以使用各种通信协议。

[0165] (多个)网络1610可以是本领域技术人员所熟悉的可以使用各种可用协议中的任何一种支持数据通信的任何类型的网络,这些可用协议包括但不限于TCP/IP(传输控制协

议/因特网协议)、SNA(系统网络架构)、IPX(因特网分组交换)、AppleTalk®等。仅通过示例的方式,(多个)网络1610可以是局域网(LAN)、基于以太网的网络、令牌环、广域网(WAN)、因特网、虚拟网络、虚拟私人网络(VPN)、内部网、外部网、公共交换电话网(PSTN)、红外线网、无线网(例如,根据电气与电子协会(IEEE)1002.11协议套件、Bluetooth®和/或任何其他无线协议中的任一种协议操作的网络)和/或这些网络和/或其他网络的任何组合。

[0166] 服务器1612可以由以下各项组成:一个或多个通用计算机、专用服务器计算机(通过示例的方式包括PC(个人计算机)服务器、UNIX®服务器、中档服务器、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。服务器1612可以包括运行虚拟操作系统的一个或多个虚拟机或涉及虚拟化的其他计算架构,比如逻辑存储设备的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池。在各种示例中,服务器1612可以被适配成运行提供前述公开内容所描述的功能的一个或多个服务或软件应用程序。

[0167] 服务器1612中的计算系统可以运行一个或多个操作系统,该一个或多个操作系统包括上文所讨论的那些操作系统中的任何一个操作系统以及任何可商购获得的服务器操作系统。服务器1612也可以运行各种附加服务器应用程序和/或中间层应用程序中的任何一种应用程序,包括HTTP(超文本运输协议)服务器、FTP(文件传送协议)服务器、CGI(通用网关接口)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®(国际商业机器公司)等商购获得的那些数据库服务器。

[0168] 在一些实施方式中,服务器1612可以包括一个或多个应用程序以分析并合并从客户端计算设备1602、1604、1606和1608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/或事件更新可以包括但不限于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,所述实时更新可以包括与传感器数据应用程序、财务收报机、网络性能测量工具(例如,网络监测和流量管理应用程序)、点击流分析工具、汽车交通监测等相关的实时事件。服务器1612还可以包括一个或多个应用程序以经由客户端计算设备1602、1604、1606和1608的一个或多个显示设备来显示数据馈送和/或实时事件。

[0169] 分布式系统1600还可以包括一个或多个数据储存库1614、1616。在某些示例中,这些数据储存库可以用于存储数据和其他信息。例如,数据储存库1614、1616中的一个或多个可以用于存储信息(如与聊天机器人性能或生成的模型有关的信息)以供服务器1612在执行根据各种实施例的各种功能时使用的聊天机器人使用。数据储存库1614、1616可以驻留在各种位置。例如,服务器1612所使用的数据储存库可以在服务器1612本地或者可以远离服务器1612并经由基于网络的或专用的连接与服务器1612通信。数据储存库1614、1616可以是不同类型。在某些示例中,服务器1612所使用的数据储存库可以是数据库,例如关系数据库,如由Oracle Corporation®和其他供应商提供的数据库。这些数据库中的一个或多个数据库可以被适配成响应于SQL格式的命令来实现数据到以及从数据库的存储、更新和取得。

[0170] 在某些示例中,数据储存库1614、1616中的一个或多个数据储存库还可以被应用

程序用来存储应用程序数据。应用程序所使用的数据储存库可以是不同类型,如例如键值存储储存库、对象存储储存库或由文件系统支持的通用存储储存库。

[0171] 在某些示例中,本公开所描述的功能可以经由云环境作为服务提供。图17是根据某些示例的各种服务在其中可以作为云服务被提供的基于云的系统环境的简化框图。在图17所描绘的示例中,云基础设施系统1702可以提供可以由用户使用一个或多个客户端计算设备1704、1706和1708请求的一个或多个云服务。云基础设施系统1702可以包括一个或多个计算机和/或服务器,该一个或多个计算机和/或服务器可以包括上文针对服务器1612所描述的那些计算机和/或服务器。云基础设施系统1702中的计算机可以被组织成通用计算机、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。

[0172] (多个)网络1710可以促进客户端1704、1706和1708与云基础设施系统1702之间的数据通信和交换。(多个)网络1710可以包括一个或多个网络。网络可以是相同或不同的类型。(多个)网络1710可以支持一个或多个通信协议(包括有线和/或无线协议)以促进通信。

[0173] 图17所描绘的示例仅是云基础设施系统的一个示例并且不旨在是限制性的。应了解,在一些其他示例中,云基础设施系统1702可以具有比图17所描绘的那些部件更多或更少的部件、可以组合两个或更多个部件或者可以具有不同的部件配置或布置。例如,尽管图17描绘了三个客户端计算设备,但是在替代性示例中,可以支持任何数量的客户端计算设备。

[0174] 术语云服务通常用于指代通过服务提供商的系统(例如,云基础设施系统1702)按需且经由如因特网等通信网络变得可用于用户的服务。通常,在公共云环境中,构成云服务提供商的系统的服务器和系统与客户自身的室内服务器和系统不同。云服务提供商的系统由云服务提供商管理。因此,客户可以使自身利用由云服务提供商提供的云服务,而不必针对服务购买单独的许可证、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程序,并且用户可以经由因特网按需订购并使用应用程序,而用户不必购买用于执行应用程序的基础设施资源。云服务被设计成提供对应用程序、资源和服务的容易的、可伸缩的访问。几个提供商提供云服务。例如,由加利福尼亚州红木海岸(Redwood Shores,California)的Oracle Corporation®提供如中间件服务、数据库服务、Java云服务以及其他服务等几种云服务。

[0175] 在某些示例中,云基础设施系统1702可以使用不同的模型(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(IaaS)模型和其他模型(包括混合服务模型)下)提供一个或多个云服务。云基础设施系统1702可以包括实现对各种云服务的供给的一套应用程序、中间件、数据库和其他资源。

[0176] SaaS模型使应用程序或软件能够作为服务通过如因特网等通信网络输送给客户,而客户不必购买底层应用程序的硬件或软件。例如,可以使用SaaS模型为客户提供对由云基础设施系统1702托管的按需应用程序的访问。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、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、社交应用程序以及其他等的各种服务。

[0177] 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。由Oracle Corporation®提供各种IaaS服务。

[0178] PaaS模型通常用于提供使客户能够开发、运行并管理应用程序和服务而客户不必

采购、构建或维护这种资源的平台和环境资源作为服务。由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、各种应用程序开发解决方案服务、以及其他服务。

[0179] 云服务通常是以按需自助服务基础、基于订阅、可弹性伸缩、可靠、高度可用且安全的方式提供的。例如,客户可以经由订阅订单来订购由云基础设施系统1702提供的一个或多个服务。然后,云基础设施系统1702执行处理以提供客户的订阅订单中请求的服务。例如,用户可以使用话语来请求云基础设施系统采取如上文所描述的某个动作(例如,意图)和/或为如本文所描述的聊天机器人系统提供服务。云基础设施系统1702可以被配置为提供一个或甚至多个云服务。

[0180] 云基础设施系统1702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1702可以由第三方云服务提供商拥有,并且云服务被提供给任何一般的公众客户,其中,该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1702可以在组织内(例如,在企业组织内)操作并且服务被提供给组织内的客户。例如,客户可以是如人力资源部门、薪资部门等企业的各个部门或甚至是企业内的个人。在某些其他示例中,在社区云模型下,云基础设施系统1702和所提供的服务可以由有关的社区中的几个组织共享。也可以使用如上文所提到的模型的混合等各种其他模型。

[0181] 客户端计算设备1704、1706和1708可以是不同类型的(如图16所描绘的客户端计算设备1602、1604、1606和1608)并且可以能够操作一个或多个客户端应用程序。用户可以使用客户端设备与云基础设施系统1702交互,如请求由云基础设施系统1702提供的服务。例如,用户可以使用客户端设备从如本公开所描述的聊天机器人请求信息或动作。

[0182] 在一些示例中,由云基础设施系统1702执行的用于提供服务的处理可以涉及模型训练和部署。此分析可以涉及使用、分析并操纵数据集来训练并部署一个或多个模型。该分析可以由一个或多个处理器执行,从而可能并行地处理数据、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1702执行以用于针对聊天机器人系统生成并训练一个或多个模型。用于该分析的数据可以包括结构化数据(例如,存储在数据库中或根据结构化模型结构化的数据)和/或非结构化数据(例如,数据块(二进制大对象))。

[0183] 如图17中的示例所描绘的,云基础设施系统1702可以包括被用于促进供给由云基础设施系统1702提供的各种云服务的基础设施资源1730。基础设施资源1730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,可用于服务从应用程序请求的存储的存储虚拟机可以是云基础设施系统1702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。

[0184] 在某些示例中,为了促进这些资源的高效供给以支持由云基础设施系统1702为不同客户提供的各种云服务,资源可以绑定到资源组或资源模块(也称为"群(pod)")中。每个资源模块或群可以包括一种或多种类型的资源的预先集成且优化的组合。在某些示例中,可以针对不同类型的云服务预先供给不同的群。例如,可以针对数据库服务供给第一组群、可以针对Java服务供给第二组群(所述第二组群可以包括与第一组群中的群不同的资源组合)等。对于一些服务,被分配用于供给服务的资源可以在服务之间共享。

[0185] 云基础设施系统1702自身可以在内部使用由云基础设施系统1702的不同部件共享并且促进云基础设施系统1702供给服务的服务1732。这些内部共享服务可以包括但不限

于安全和身份服务、整合服务、企业储存库服务、企业管理器服务、病毒扫描和白名单服务、高度可用性、备份和恢复服务、用于实现云支持的服务、电子邮件服务、通知服务、文件传送服务等。

[0186] 云基础设施系统1702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实施。如图17所描绘的,子系统可以包括使得云基础设施系统1702的用户或客户能够与云基础设施系统1702交互的用户接口子系统1712。用户接口子系统1712可以包括各种不同的接口,如web接口1714、在线商店接口1716(其中,广告并且消费者可购买由云基础设施系统1702提供的云服务)和其他接口1718。例如,客户可以使用客户端设备请求(服务请求1734)由云基础设施系统1702使用接口1714、1716和1718中的一个或多个接口提供的一个或多个服务。例如,客户可以访问在线商店、浏览由云基础设施系统1702提供的云服务并且针对客户希望订阅的由云基础设施系统1702提供的一个或多个服务下订阅订单。服务请求可以包括标识客户和客户期望订阅的一个或多个服务的信息。例如,客户可以针对由云基础设施系统1702提供的服务下订阅订单。作为订单的一部分,客户可以提供标识服务要针对其提供的聊天机器人系统的信息并且可选地为聊天机器人系统提供一个或多个凭证。

[0187] 在某些示例(如图17所描绘的示例)中,云基础设施系统1702可以包括被配置为处理新订单的订单管理子系统(OMS)1720。作为该处理的一部分,OMS1720可以被配置为:为客户创建账户(如果尚未创建的话);从客户接收要用于向客户开具账单以向客户提供所请求的服务的账单和/或计费信息;验证客户信息;在验证之后,为客户预订订单;并且策划各种工作流程来准备订单以进行供给。

[0188] 一旦被正确地验证,那么,OMS1720就可以调用被配置成为订单供给资源(包括处理资源、存储器资源和联网资源)的订单供给子系统(OPS)1724。供给可以包括为订单分配资源并配置资源以促进客户订单所请求的服务。为订单供给资源的方式和所供给的资源的类型可以取决于客户已经订购的云服务的类型。例如,根据一个工作流程,OPS 1724可以被配置为确定正在请求的特定云服务并且识别可能已经为该特定云服务预先配置的群的数量。为订单分配的群的数量可以取决于所请求服务的大小/量/层级/范围。例如,要分配的群的数量可以基于服务所要支持的用户的数量、正在请求的服务的持续时间等来确定。然后,可以为特定的请求客户定制分配的群以用于提供所请求服务。

[0189] 在某些示例中,如上文所描述的设置阶段处理可以作为供给过程的一部分由云基础设施系统1702来执行。云基础设施系统1702可以生成应用程序ID并从云基础设施系统1702自身所提供的存储虚拟机中或从除云基础设施系统1702以外的其他系统所提供的存储虚拟机中为应用程序选择存储虚拟机。

[0190] 云基础设施系统1702可以向请求客户发送回应或通知1744,以指示何时所请求服务现在已准备好使用。在一些实例中,可以向客户发送使客户能够开始使用和利用所请求服务的益处的信息(例如,链接)。在某些示例中,对于请求服务的客户,回应可以包括由云基础设施系统1702生成的聊天机器人系统ID和标识由云基础设施系统1702对应于聊天机器人系统ID为聊天机器人系统选择的聊天机器人系统的信息。

[0191] 云基础设施系统1702可以向多个客户提供服务。对于每个客户,云基础设施系统1702负责管理与从客户接收的一个或多个订阅订单有关的信息、维护与订单相关的客户数据并且向客户提供所请求服务。云基础设施系统1702还可以收集关于客户使用所订阅服务

的使用情况统计数据。例如,可以收集所使用的存储量、所传送的数据量、用户的数量以及系统开机时间量和系统停机时间量等的统计数据。可以使用该使用情况信息向客户开具账单。可以例如以月为周期完成开具账单。

[0192] 云基础设施系统1702可以并行地向多个客户提供服务。云基础设施系统1702可以存储这些客户的信息(可能包括专有信息)。在某些示例中,云基础设施系统1702包括被配置为管理客户信息并提供对所管理信息的分离使得与一个客户有关的信息不能被另一个客户访问的身份管理子系统(IMS)1728。IMS1728可以被配置为提供如身份服务等各种安全有关服务,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功能的服务等。

[0193] 图18图示了计算机系统1800的示例。在一些示例中,计算机系统1800可以用于实施分布式环境内的任何数字助理或聊天机器人系统以及上文所描述的各种服务器和计算机系统。如图18所示,计算机系统1800包括各种子系统,包括经由总线子系统1802与多个其他子系统通信的处理子系统1804。这些其他子系统可以包括处理加速单元1806、I/O子系统1808、存储子系统1818和通信子系统1824。存储子系统1818可以包括非暂态计算机可读存储介质,包括存储介质1822和系统存储器1810。

[0194] 总线子系统1802提供用于允许计算机系统1800的各个部件和子系统按预期彼此通信的机构。虽然总线子系统1802被示意性地示出为单个总线,但是总线子系统的替代性示例可以利用多个总线。总线子系统1802可以是包括存储器总线或存储器控制器、外围总线、使用各种总线架构中的任何一种的局部总线等的若干种类型的总线结构中的任何一种。例如,这种架构可以包括工业标准架构(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)局部总线以及外围部件互连(PCI)总线(其可以被实施为被制造成IEEE P13156.1标准的夹层(Mezzanine)总线)等。

[0195] 处理子系统1804控制计算机系统1800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处理器或多核处理器。计算机系统1800的处理资源可以被组织成一个或多个处理单元1832、1834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、核与处理器的组合或核与处理器的其他组合。在一些示例中,处理子系统1804可以包括如图形处理器、数字信号处理器(DSP)等一个或多个专用协处理器。在一些示例中,处理子系统1804的处理单元中的一些或全部可以使用如专用集成电路(ASIC)或现场可编程门阵列(FPGA)等定制电路来实施。

[0196] 在一些示例中,处理子系统1804中的处理单元可以执行存储在系统存储器1810内或计算机可读存储介质1822上的指令。在各种示例中,处理单元可以执行各种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,要执行的程序代码中的一些或全部可以驻留在系统存储器1810中和/或计算机可读存储介质1822上(潜在地包括驻留在一个或多个存储设备上)。通过适合的编程,处理子系统1804可以提供上文所描述的各种功能。在计算机系统1800在执行一个或多个虚拟机的实例中,可以向每个虚拟机分配一个或多个处理单元。

[0197] 在某些示例中,可以可选地提供处理加速单元1806以用于执行定制处理或用于卸载处理子系统1804所执行的处理中的一些处理,从而加速计算机系统1800所执行的总体处

理。

[0198] I/O子系统1808可以包括用于向计算机系统1800输入信息和/或用于从或经由计算机系统1800输出信息的设备和机构。通常,使用术语输入设备旨在包括用于向计算机系统1800输入信息的所有可能类型的设备和机构。用户接口输入设备可以包括例如键盘、如鼠标或轨迹球等指向设备、合并到显示器中的触摸板或触摸屏、滚轮、点击轮、拨号盘、按钮、开关、小键盘、具有话音命令识别系统的音频输入设备、麦克风以及其他类型的输入设备。用户接口输入设备还可以包括运动感测和/或姿势识别设备,如使用户能够控制输入设备并与输入设备交互的Microsoft Kinect®运动传感器、Microsoft Xbox® 360游戏控制器、提供用于使用姿势和口述命令接收输入的界面的设备。用户接口输入设备还可以包括眼部姿势识别设备,如检测来自用户的眼部活动(例如,在拍照和/或进行菜单选择时“眨眼”)并将眼部姿势转化为到输入设备(如Google Glass®)的输入的Google Glass®眨眼检测器。另外,用户接口输入设备可以包括使用户能够通过声音命令与声音识别系统(例如,Siri®导航器)交互的声音识别感测设备。

[0199] 用户接口输入设备的其他示例包括但不限于三维(3D)鼠标、操纵杆或指向杆、游戏手柄和图形板、以及音频/视觉设备(如扬声器、数码相机、数码摄像机、便携式媒体播放器、网络摄像机、图像扫描仪、指纹扫描仪、条形码读取器3D扫描仪、3D打印机、激光测距仪、眼睛注视跟踪设备)。另外,用户接口输入设备可以包括例如医学成像输入设备,如计算机断层扫描、磁共振成像、正电子发射断层扫描和医学超声检查设备。用户接口输入设备还可以包括例如音频输入设备,如MIDI键盘、数码乐器等。

[0200] 通常,使用术语输出设备旨在包括用于从计算机系统1800向用户或其他计算机输出信息的所有可能类型的设备和机构。用户接口输出设备可以包括显示子系统、指示灯或如音频输出设备等非视觉显示器。显示子系统可以是阴极射线管(CRT)、平板设备(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在视觉上传达文本、图形和音频/视频信息的各种显示设备,如监视器、打印机、扬声器、头戴式耳机、汽车导航系统、绘图仪、话音输出设备和调制解调器。

[0201] 存储子系统1818提供用于存储计算机系统1800所使用的信息和数据的储存库或数据存储。存储子系统1818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态计算机可读存储介质。存储子系统1818可以存储当由处理子系统1804执行时提供上文所描述的功能的软件(例如,程序、代码模块、指令)。软件可以由处理子系统1804的一个或多个处理单元执行。存储子系统1818还可以根据本公开的传授内容提供认证。

[0202] 存储子系统1818可以包括一个或多个非暂态存储器设备,该一个或多个非暂态存储器设备包括易失性存储器设备和非易失性存储器设备。如图18所示,存储子系统1818包括系统存储器1810和计算机可读存储介质1822。系统存储器1810可以包括多个存储器,包括用于在程序执行期间存储指令和数据的易失性主随机存取存储器(RAM)和其中存储有固定指令的非易失性只读存储器(ROM)或闪速存储器。在一些实施方式中,包含在启动期间帮助计算机系统1800内的元件之间传送信息的基本例程的基本输入/输出系统(BIOS)通常可以存储在ROM中。RAM通常包含目前由处理子系统1804操作并执行的数据和/或程序模块。在一些实施方式中,系统存储器1810可以包括如静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)等多种不同类型的存储器。

[0203] 通过示例而非限制的方式,如图18所描绘的,系统存储器1810可以加载正在执行的应用程序1812(这些应用程序可以包括如Web浏览器、中间层应用程序、关系数据库管理系统(RDBMS)等各种应用程序)、程序数据1814和操作系统1816。通过示例的方式,操作系统1816可以包括各种版本的Microsoft Windows®、Apple Macintosh®和/或Linux操作系统、各种可商购获得的UNIX®或类UNIX操作系统(包括但不限于各种GNU/Linux操作系统、Google Chrome® OS等)和/或如iOS、Windows®电话、Android®OS、BlackBerry®OS、Palm®OS操作系统等移动操作系统、以及其他操作系统。

[0204] 计算机可读存储介质1822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1822可以为计算机系统1800提供对计算机可读指令、数据结构、程序模块和其他数据的存储。当由处理子系统1804执行时提供上文所描述的功能的软件(程序、代码模块、指令)可以存储在存储子系统1818中。通过示例的方式,计算机可读存储介质1822可以包括如硬盘驱动器、磁盘驱动器、光盘驱动器(如CD ROM、DVD、Blu-Ray®盘或其他光学介质)等非易失性存储器。计算机可读存储介质1822可以包括但不限于Zip®驱动器、闪速存储器卡、通用串行总线(USB)闪速存储器驱动器、安全数字(SD)卡、DVD盘、数字录像带等。计算机可读存储介质1822还可以包括如基于闪速存储器的固态驱动器(SSD)、企业级闪速存储器驱动器、固态ROM等基于非易失性存储器的SSD、基于如固态RAM、动态RAM、静态RAM等易失性存储器的SSD、基于DRAM的SSD、磁阻RAM(MRAM)SSD以及使用DRAM和基于闪速存储器的SSD的组合的混合SSD。

[0205] 在某些示例中,存储子系统1818还可以包括可以进一步连接到计算机可读存储介质1822的计算机可读存储介质读取器1820。读取器1820可以从如盘、闪速存储器驱动器等存储器设备接收数据并且被配置为从该存储器设备读取数据。

[0206] 在某些示例中,计算机系统1800可以支持虚拟化技术,包括但不限于对处理和存储器资源的虚拟化。例如,计算机系统1800可以提供对执行一个或多个虚拟机的支持。在某些示例中,计算机系统1800可以执行如促进对虚拟机的配置和管理的管理程序等程序。每个虚拟机可以分配有存储器、计算(例如,处理器、核)、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身的操作系统,该操作系统可以与由计算机系统1800所执行的其他虚拟机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1800同时运行。

[0207] 通信子系统1824提供到其他计算机系统和网络的接口。通信子系统1824用作用于从其他系统接收数据并且从计算机系统1800向其他系统传输数据的接口。例如,通信子系统1824可以使得计算机系统1800能够经由因特网构建到一个或多个客户端设备的通信通道以用于从客户端设备接收信息并向客户端设备发送信息。例如,当计算机系统1800用于实施图1所描绘的机器人系统120时,通信子系统可以用于与针对应用程序选择的聊天机器人系统通信。

[0208] 通信子系统1824可以支持有线通信协议和/或无线通信协议两者。在某些示例中,通信子系统1824可以包括用于访问无线声音和/或数据网络的射频(RF)收发器部件(例如,使用蜂窝电话技术、如3G、4G或EDGE(全球演进增强型数据速率)等先进的数据网络技术、WiFi(IEEE 1502.XX家庭标准、或其他移动通信技术、或其任何组合))、全球定位系统(GPS)

接收器部件和/或其他部件。在一些示例中,除了无线接口之外或替代无线接口,通信子系统1824可以提供有线网络连接性(例如,以太网)。

[0209] 通信子系统1824可以接收并发射各种形式的数据。在一些示例中,除了其他形式之外,通信子系统1824还可以接收结构化和/或非结构化数据馈送1826、事件流1828、事件更新1830等形式的输入通信。例如,通信子系统1824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地接收(或发送)数据馈送1826,如Twitter®馈送、Facebook®更新、web馈送(如丰富站点摘要(RSS)馈送)和/或来自一个或多个第三方信息源的实时更新。

[0210] 在某些示例中,通信子系统1824可以被配置为接收连续数据流形式的数据,这些连续数据流可以包括(可以没有显式结束的本质上连续的或无界的)实时事件的事件流1828和/或事件更新1830。生成连续数据的应用程序的示例可以包括例如传感器数据应用程序、财务收报机、网络性能测量工具(例如,网络监测和流量管理应用程序)、点击流分析工具、汽车交通监测等。

[0211] 通信子系统1824还可以被配置为将数据从计算机系统1800传送到其他计算机系统或网络。可以将数据以如结构化和/或非结构化数据馈送1826、事件流1828、事件更新1830等各种不同的形式传送到可以与耦接至计算机系统1800的一个或多个流数据源计算机通信的一个或多个数据库。

[0212] 计算机系统1800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®计算平板计算机、PDA)、可穿戴设备(例如,Google Glass®头戴式显示器)、个人计算机、工作站、主机、自助服务终端、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,对图18所描绘的计算机系统1800的描述旨在仅作为具体示例。具有比图18所描绘的系统更多或更少的部件的许多其他配置是可能的。基于本公开和本文所提供的传授内容,应了解,有其他方式和/或方法来实施各种示例。

[0213] 虽然已经描述了具体示例,但是各种修改、更改、替代性构造和等同物是可能的。示例不局限于在某些具体数据处理环境中的操作,而是自由地在多个数据处理环境中操作。另外,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而言应当显而易见的是,这不旨在是限制性的。虽然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地或同时地执行。另外,可以重新布置操作的顺序。过程可以具有图中未包括的另外的步骤。可以单独地或联合地使用上文所描述的示例的各种特征和方面。

[0214] 进一步地,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实施。本文所描述的各种过程可以在相同处理器或不同处理器上以任何组合实施。

[0215] 在将设备、系统、部件或模块描述为被配置为执行某些操作或功能的情况下,这种配置可以例如通过将电子电路设计成执行操作、通过对可编程电子电路(如微处理器)进行编程以执行操作(如通过执行计算机指令或代码)或通过被编程成执行存储在非暂态存储器介质上的代码或指令的处理器或核或其任何组合来完成。过程可以使用包括但不限于用于过程间通信的传统技术的各种技术进行通信,并且不同的过程对可以使用不同的技术,或者相同的过程对可以在不同的时间使用不同的技术。

[0216] 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,已经示出了公知的电路、过程、算法、结构和技术,而没有不

必要的细节,以避免模糊示例。此描述仅提供了示例性示例并且不旨在限制其他示例的范围、适用性或配置。而是,先前对示例的描述将为本领域技术人员提供用于实施各种示例的使能描述。可以对元件的功能和布置作出各种改变。

[0217] 因此,应当从说明性而非限制性意义上看待本说明书和附图。然而,将明显的是,在不脱离权利要求中阐述的更广泛的精神和范围的情况下,可以对其作出添加、减少、删除以及其他修改和改变。因此,虽然已经描述了具体示例,但是这些示例不旨在是限制性的。各种修改和等同物均在以下权利要求的范围内。

[0218] 在前述说明书中,参考本公开的具体示例描述了本公开的各方面,但是本领域技术人员将认识到,本公开并不限于此。可以单独地或联合地使用上文所描述的公开的各种特征和方面。进一步地,在不脱离说明书的更广泛的精神和范围的情况下,可以在除了本文所描述的那些环境和应用程序外的任何数量的环境和应用程序中利用示例。因此,说明书和附图应被视为说明性的而非限制性的。

[0219] 在前述描述中,出于说明的目的,以特定顺序描述了方法。应当了解,在替代性示例中,可以以与所描述的顺序不同的顺序来执行方法。还应当了解,上文所描述的方法可以由硬件部件执行或者可以体现在机器可执行指令的序列中,所述指令可以用于使机器(如用所述指令编程的通用或专用处理器或逻辑电路)执行所述方法。这些机器可执行指令可以存储在一个或多个机器可读介质上,如CD-ROM或其他类型的光盘、软盘、ROM、RAM、EPROM、EEPROM、磁卡或光卡、闪速存储器或适合于存储电子指令的其他类型的机器可读介质。可替代地,方法可以由硬件和软件的组合来执行。

[0220] 在将部件描述为被配置为执行某些操作的情况下,这种配置可以例如通过设计用于执行操作的电子电路或其他硬件、通过对用于执行操作的可编程电子电路(例如,微处理器或其他适合的电子电路)进行编程或其任何组合来完成。

[0221] 尽管本文已经详细描述了本申请的说明性示例,但是应当理解,可以以其他方式不同地体现并采用本发明概念,并且所附权利要求旨在被解释为包括这种变化,受现有技术限制的情况除外。

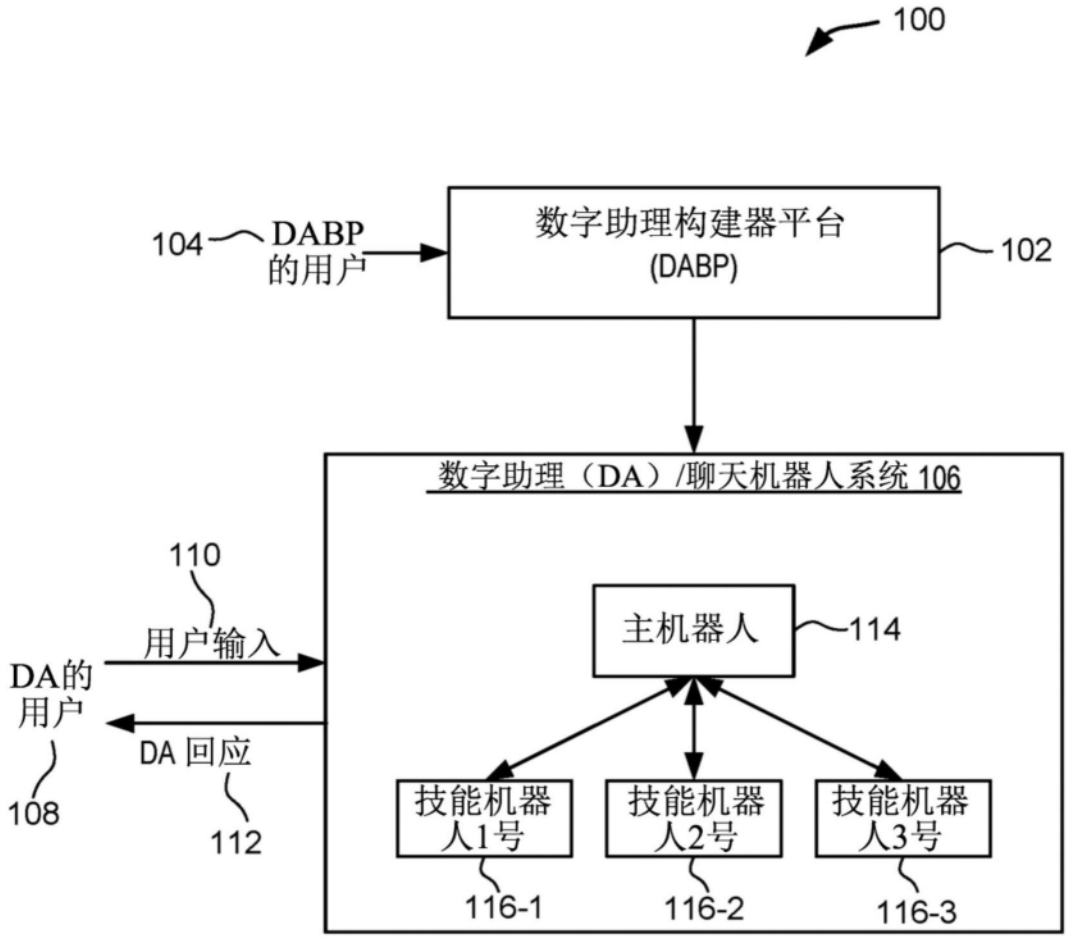

100
104
DABP
的用户
数字助理构建器平台
(DABP)
102
数字助理（DA）/聊天机器人系统 106
110
用户输入
DA的
用户
DA 回应
108
112
主机器人
114
技能机器
人1号
技能机器
人2号
技能机器
人3号
116-1
116-2
116-3

图1

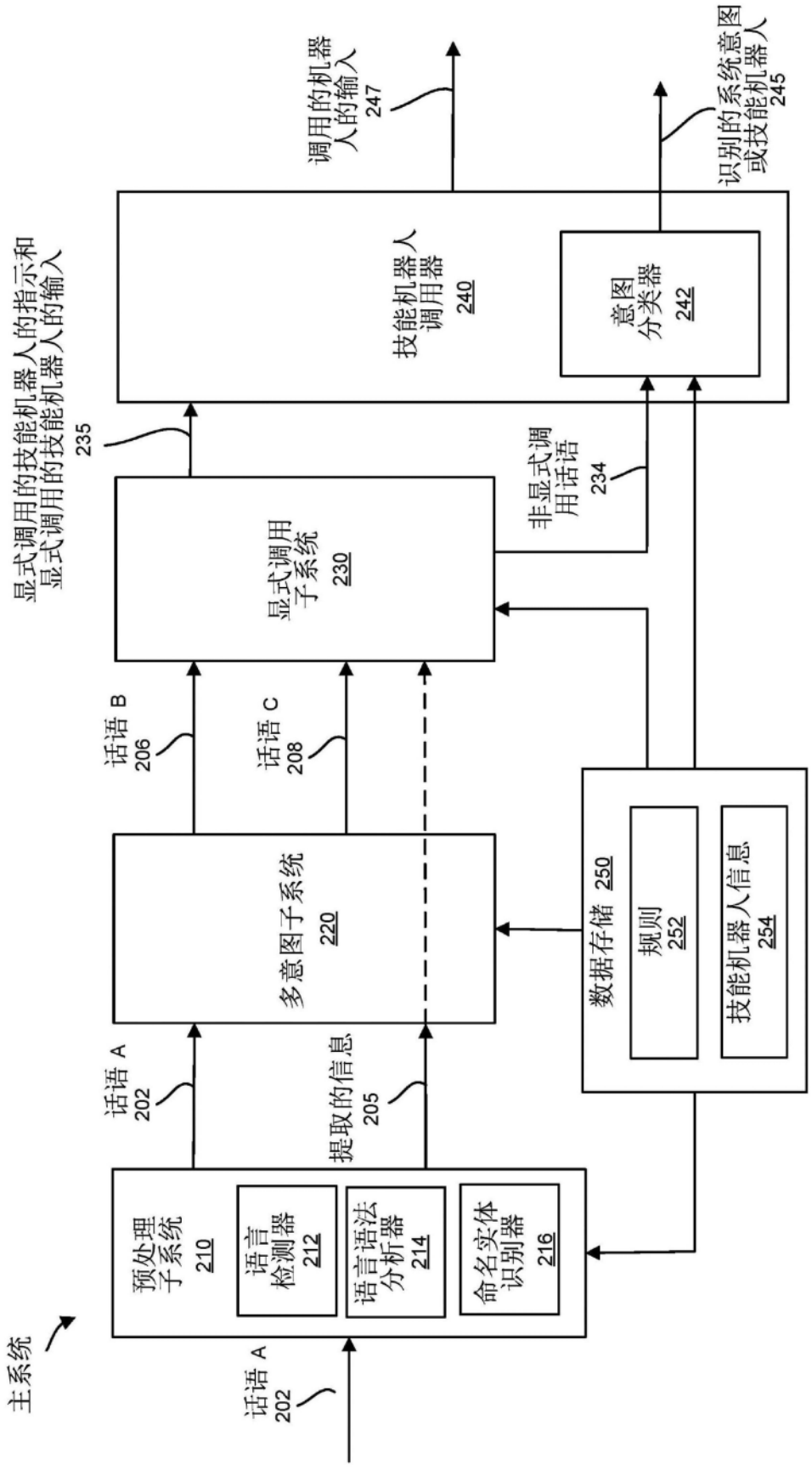

主系统
话语 A
202
预处理
子系统
210
语言
检测器
212
语言语法
分析器
214
命名实体
识别器
216
话语 A
202
提取的信息
205
多意图子系统
220
话语 B
206
话语 C
208
显式调用
子系统
230
显式调用的技能机器人的指示和
显式调用的技能机器人的输入
235
非显式调
用话语
234
技能机器人
调用器
240
意图
分类器
242
调用的机器
人的输入
247
识别的系统意图
或技能机器人
245
数据存储 250
规则
252
技能机器人信息
254

图2

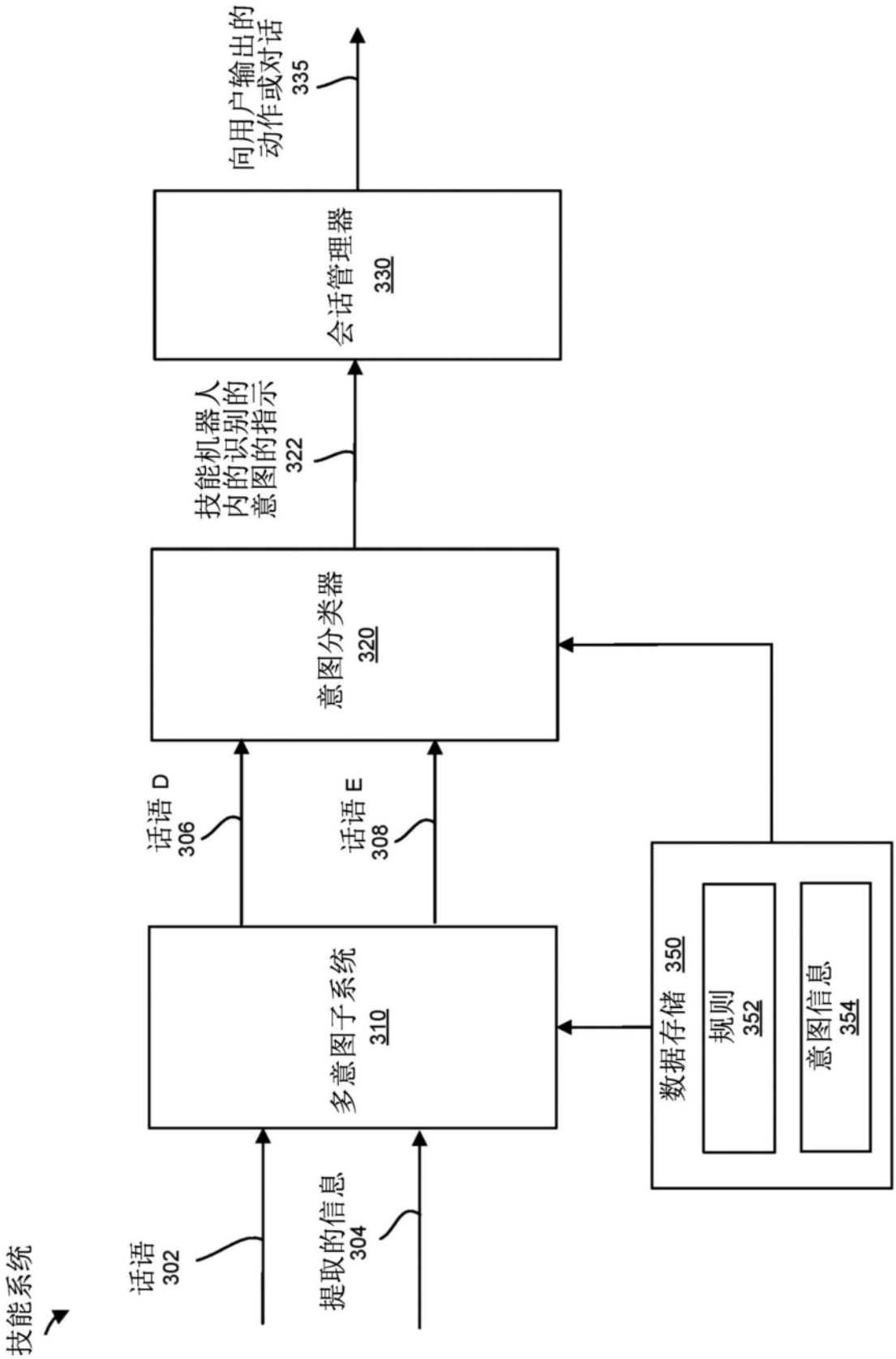

技能系统
话语
302
提取的信息
304
多意图子系统
310
话语 D
306
话语 E
308
意图分类器
320
技能机器人内的识别的意图的指示
322
会话管理器
330
向用户输出的动作或对话
335
数据存储 350
规则
352
意图信息
354

图3

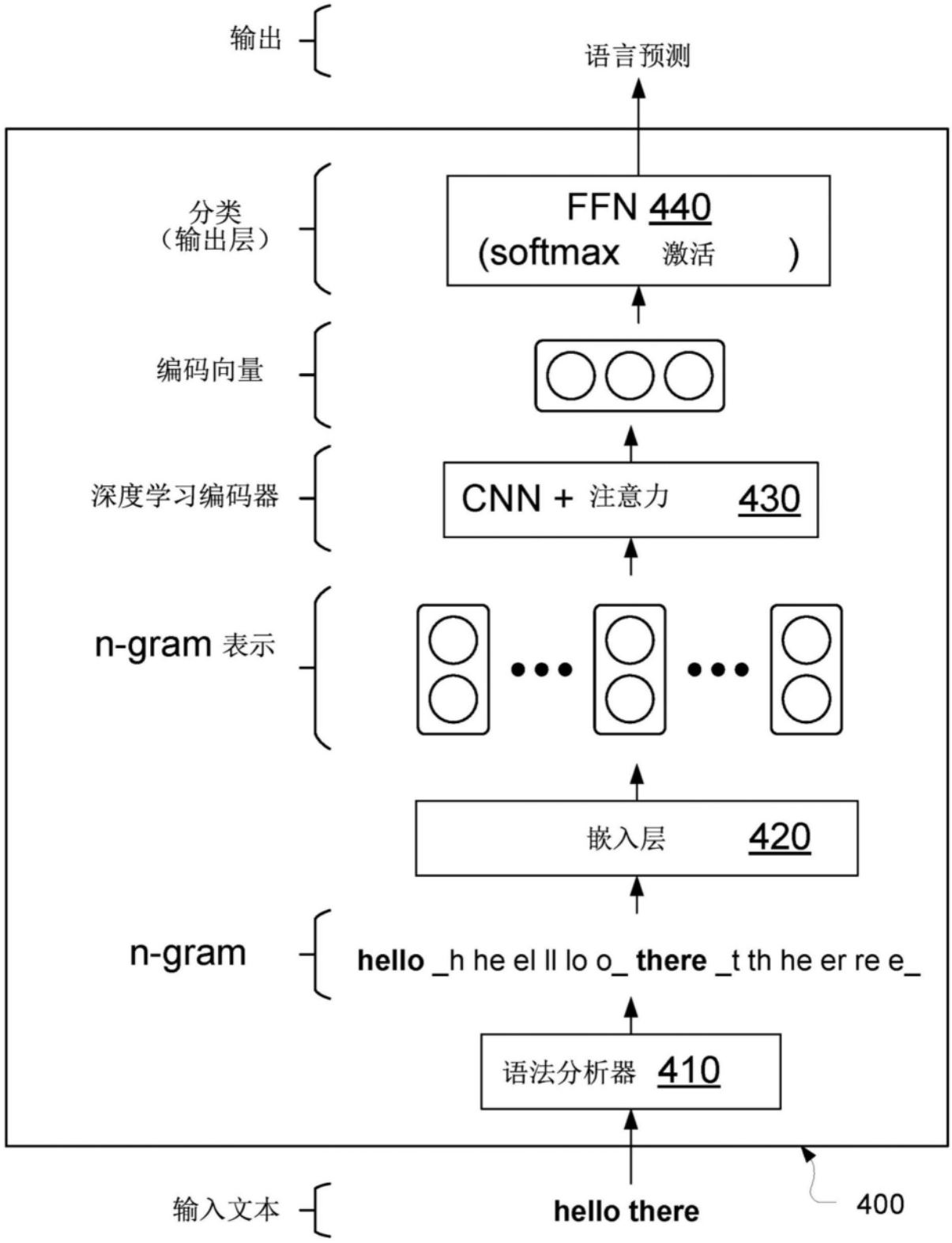

输出
语言预测
分类
（输出层）
FFN 440
(softmax 激活)
编码向量
深度学习编码器
CNN + 注意力 430
n-gram 表示
嵌入层 420
n-gram
hello _h he el ll lo o_ there _t th he er re e_
语法分析器 410
输入文本
hello there
400

图4

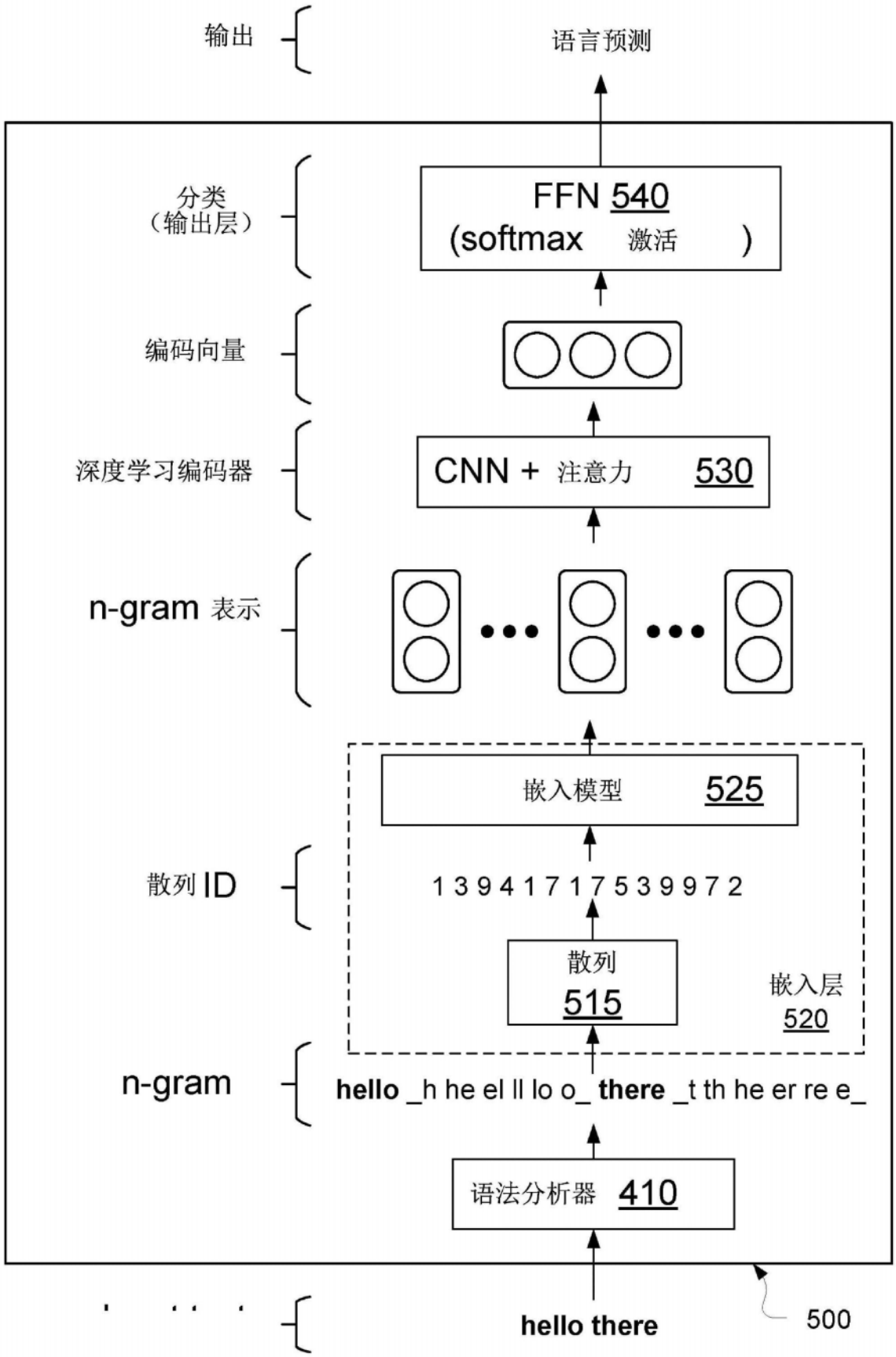

输出
语言预测
分类
（输出层）
FFN 540
(softmax 激活)
编码向量
深度学习编码器
CNN + 注意力 530
n-gram 表示
嵌入模型 525
散列 ID
1 3 9 4 1 7 1 7 5 3 9 9 7 2
散列
515
嵌入层
520
n-gram
hello _h he el ll lo o_ there _t th he er re e_
语法分析器 410
hello there
500

图5

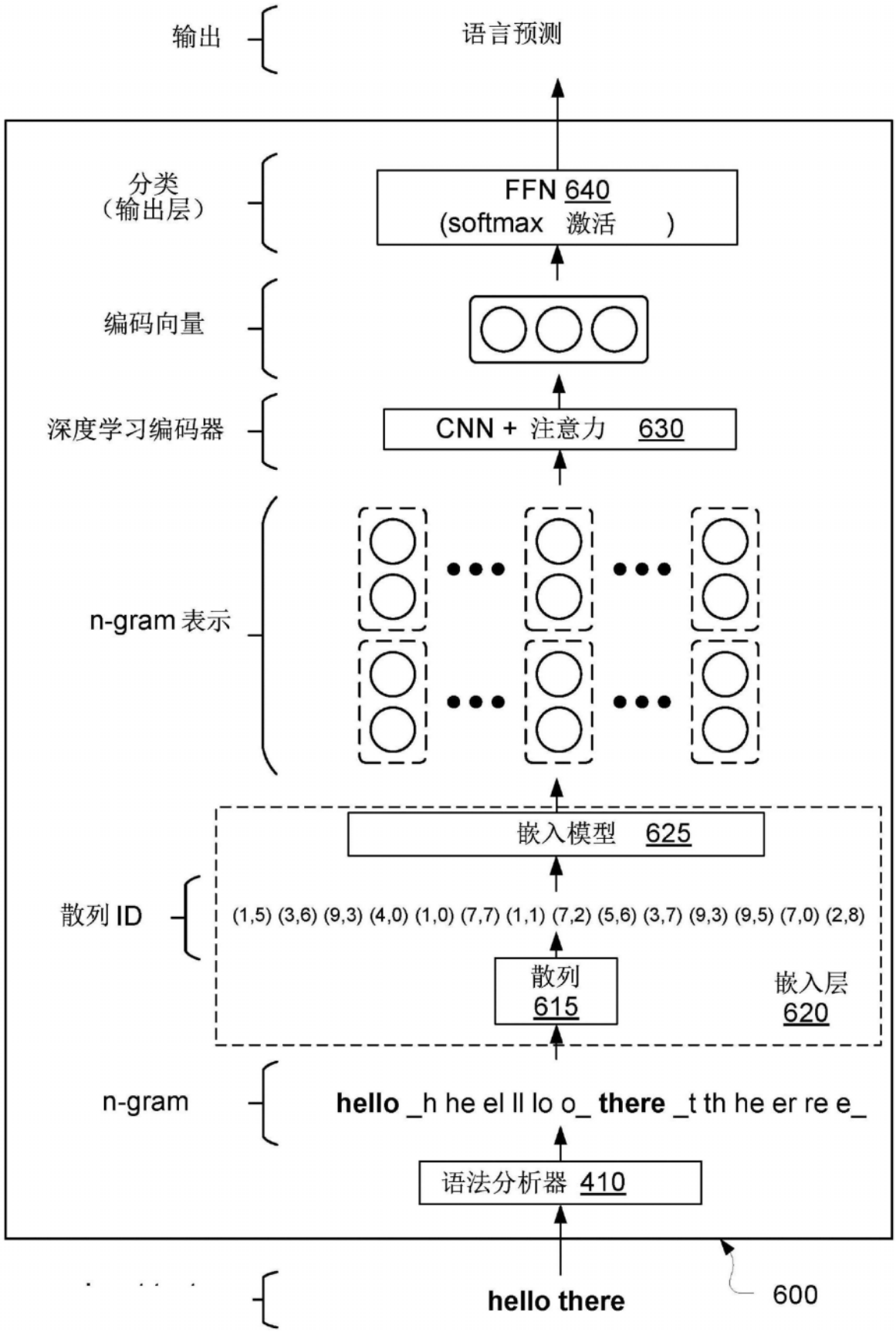

输出
语言预测
分类
（输出层）
FFN 640
(softmax 激活)
编码向量
深度学习编码器
CNN + 注意力 630
n-gram 表示
嵌入模型 625
散列 ID
(1,5) (3,6) (9,3) (4,0) (1,0) (7,7) (1,1) (7,2) (5,6) (3,7) (9,3) (9,5) (7,0) (2,8)
散列
615
嵌入层
620
n-gram
hello _h he el ll lo o_ there _t th he er re e_
语法分析器 410
hello there
600

图6

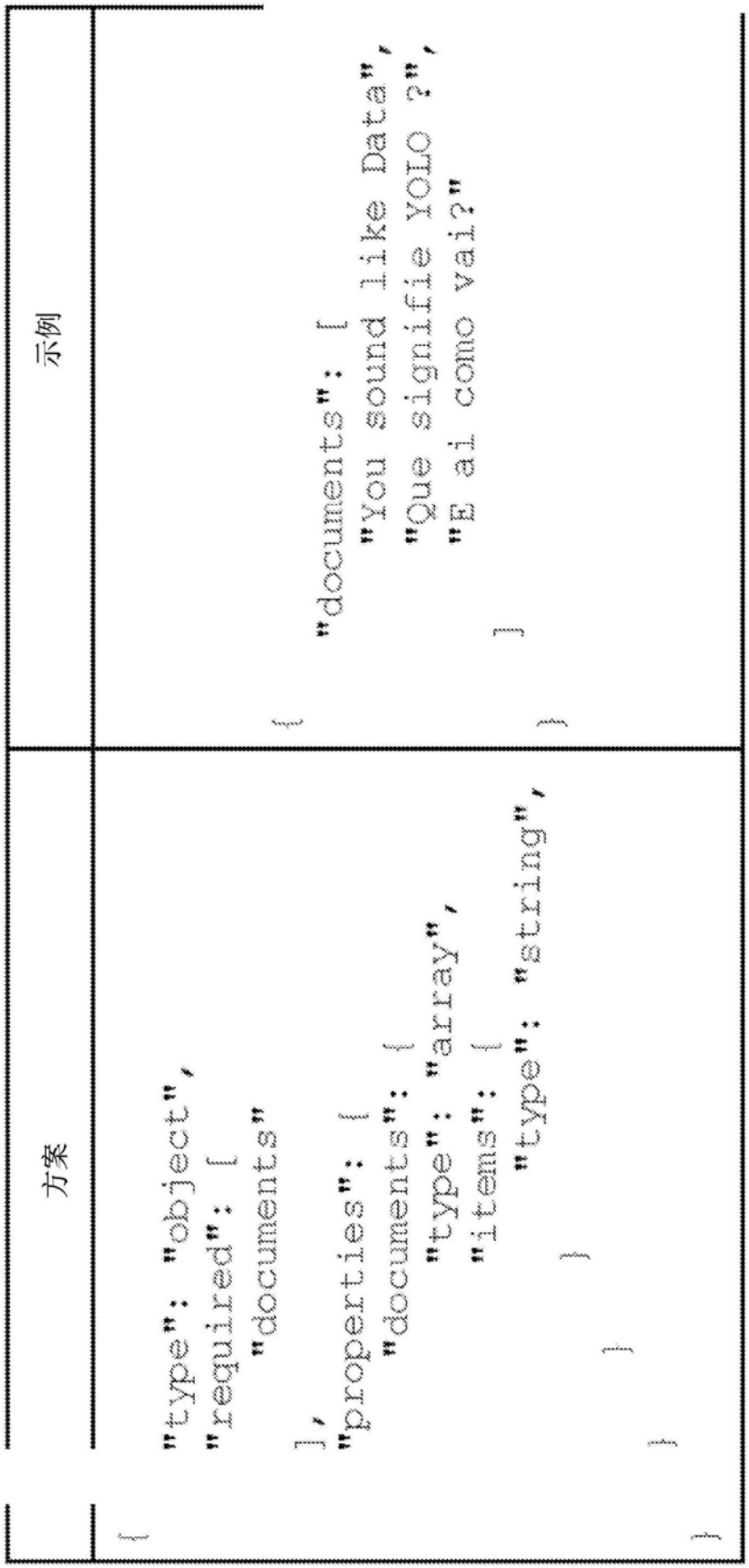

方案
示例
{
"type": "object",
"required": [
"documents"
],
"properties": {
"documents": {
"type": "array",
"items": {
"type": "string",
}
}
}
}
{
"documents": [
"You sound like Data",
"Que significa YOLO ?",
"E ai como vai?"
]
}

图7

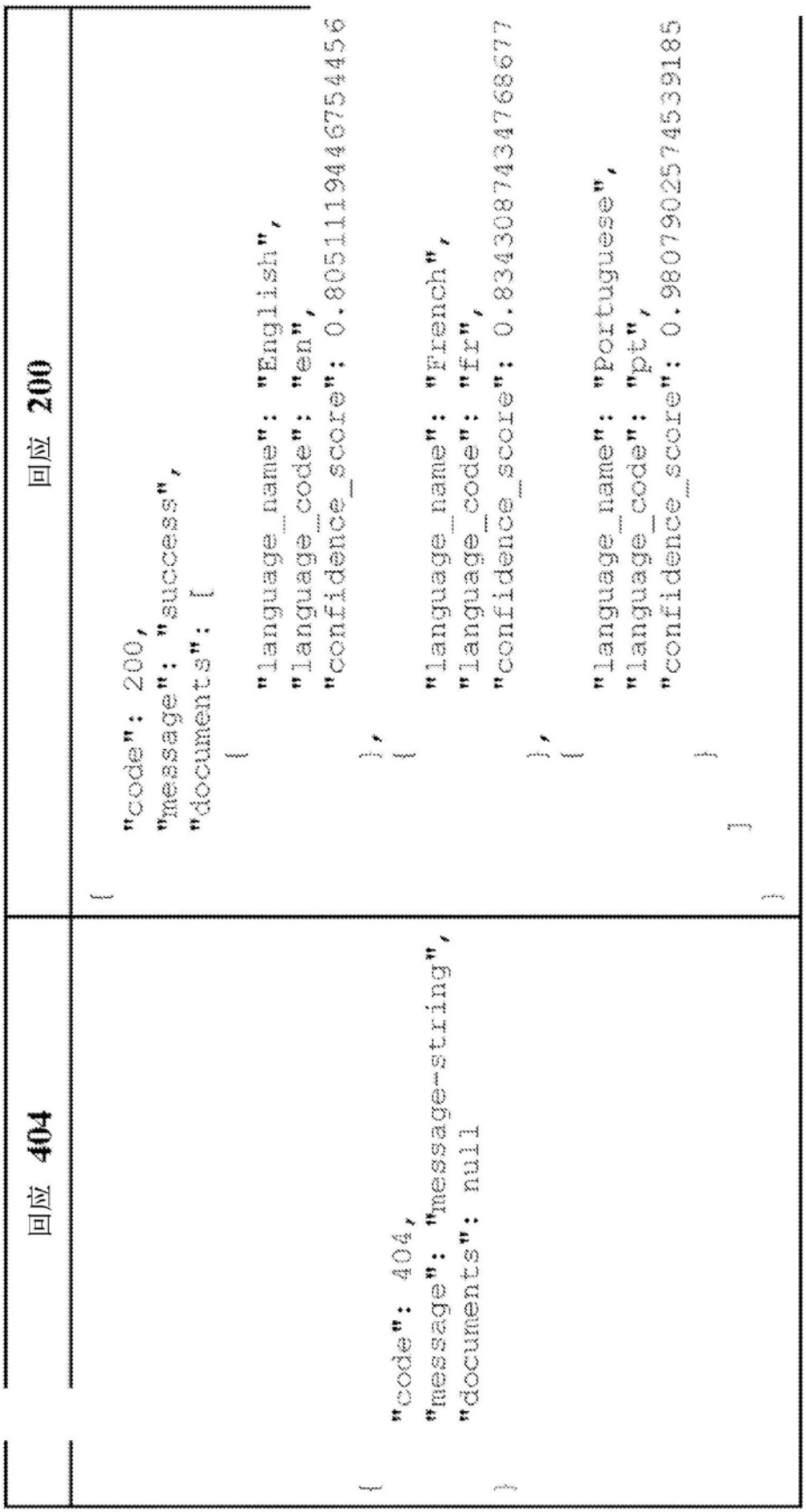
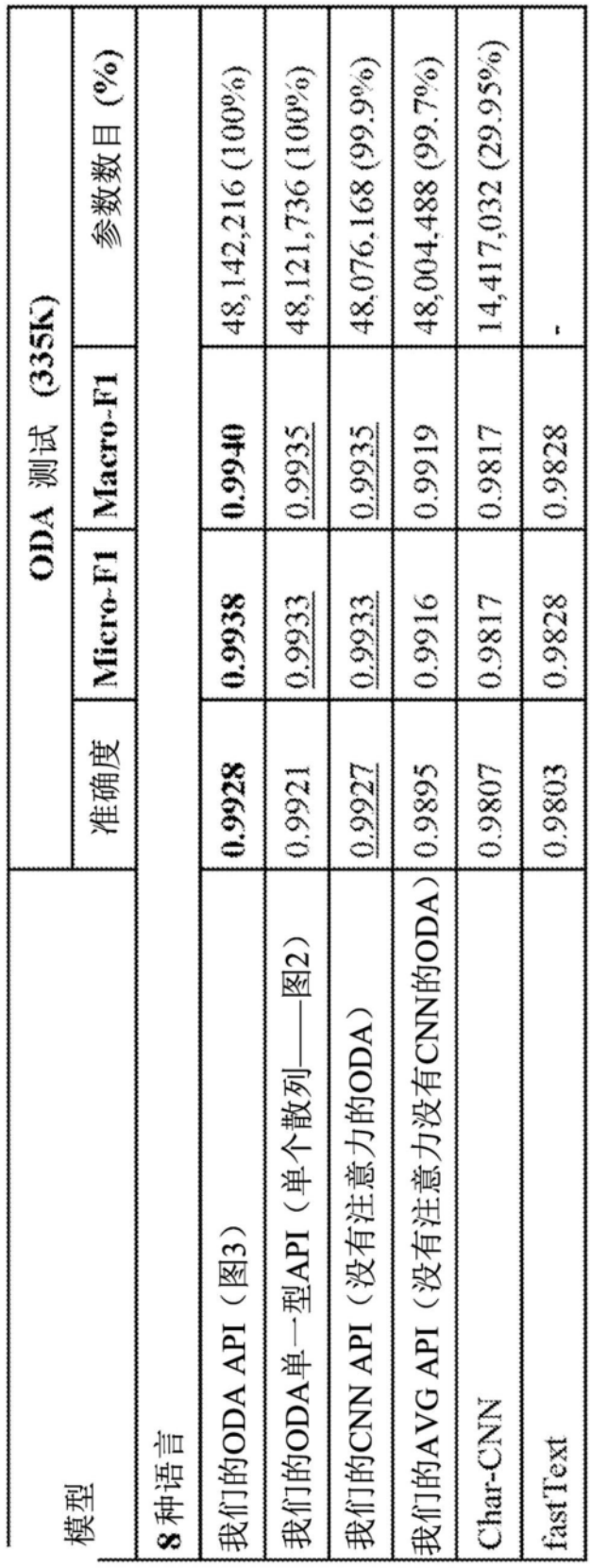

回应 404
{
"code": 404,
"message": "message-string",
"documents": null
}
回应 200
{
"code": 200,
"message": "success",
"documents": [
{
"language_name": "English",
"language_code": "en",
"confidence_score": 0.8051119446754456
},
{
"language_name": "French",
"language_code": "fr",
"confidence_score": 0.8343087434768677
},
{
"language_name": "Portuguese",
"language_code": "pt",
"confidence_score": 0.9807902574539185
}
]
}

图8

源和链接	说明	总数量		
		文件	词元	句子片段
开放并行语料库项目（OPUS）——维基百科 http://opus.nlpl.eu/Wikipedia.php	摘自维基百科的并列句子：20种语言，36个双语文本	114	610.13M	25.90M
开放并行语料库项目（OPUS）——Eubookshop http://opus.nlpl.eu/EUbookshop.php	来自欧盟书城的文档的语料库：48种语言，794个双语文本	135,785	3.60G	173.20M
开放并行语料库项目（OPUS）——EMEA http://opus.nlpl.eu/EMEA.php	由 European Medicines Agency [欧洲药品管理局]的PDF文档制作的平行语料库：22种语言，231个双语文本	41,957	311.65M	26.51M
开放并行语料库项目（OPUS）——GNOME http://opus.nlpl.eu/GNOME.php	GNOME本地化文件的平行语料库：187种语言，12,822个双语文本	113,344	267.27M	58.12M
开放并行语料库项目（OPUS）——Europarl http://opus.nlpl.eu/Europarl.php	21种语言，211个双语文本	207,775	759.05M	30.32M
开放并行语料库项目（OPUS）——Euconst http://opus.nlpl.eu/EUconst.php	21种语言，210个双语文本	986	3.01M	0.22M
开放并行语料库项目（OPUS）——OpenOffice http://opus.nlpl.eu/OpenOffice.php	8种语言，28个双语文本	18,120	3.56M	0.62M
开放并行语料库项目（OPUS）——infopankki-v1 http://opus.nlpl.eu/infopankki-v1.php	12种语言，66个双语文本	3,646	7.87M	0.79M
开放并行语料库项目（OPUS）——MultiUN http://opus.nlpl.eu/MultiUN.php	7种语言，21个双语文本	489,334	1.99G	81.41M
开放并行语料库项目（OPUS）——DGT http://opus.nlpl.eu/DGT.php	25种语言，299个双语文本	817,410	2.13G	113.52M

图9



模型	ODA 测试 (335K)			
	准确度	Micro-F1	Macro-F1	参数数目 (%)
8 种语言				
我们的ODA API（图3）	**0.9928**	**0.9938**	**0.9940**	48,142,216 (100%)
我们的ODA单一型API（单个散列——图2）	0.9921	0.9933	0.9935	48,121,736 (100%)
我们的CNN API（没有注意力的ODA）	0.9927	0.9933	0.9935	48,076,168 (99.9%)
我们的AVG API（没有注意力没有CNN的ODA）	0.9895	0.9916	0.9919	48,004,488 (99.7%)
Char-CNN	0.9807	0.9817	0.9817	14,417,032 (29.95%)
fastText	0.9803	0.9828	0.9828	-

图10

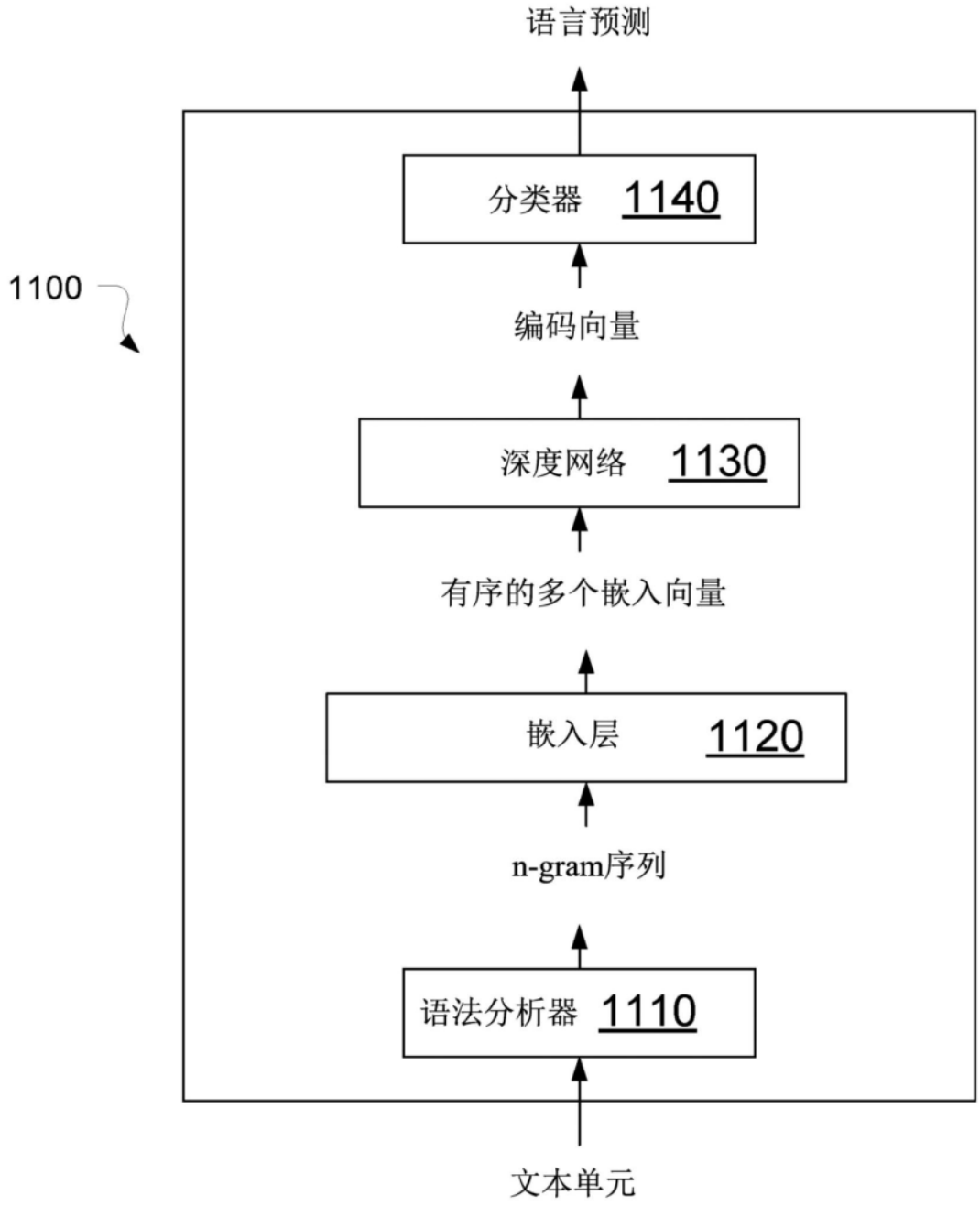

语言预测
分类器 1140
1100
编码向量
深度网络 1130
有序的多个嵌入向量
嵌入层 1120
n-gram序列
语法分析器 1110
文本单元

图11

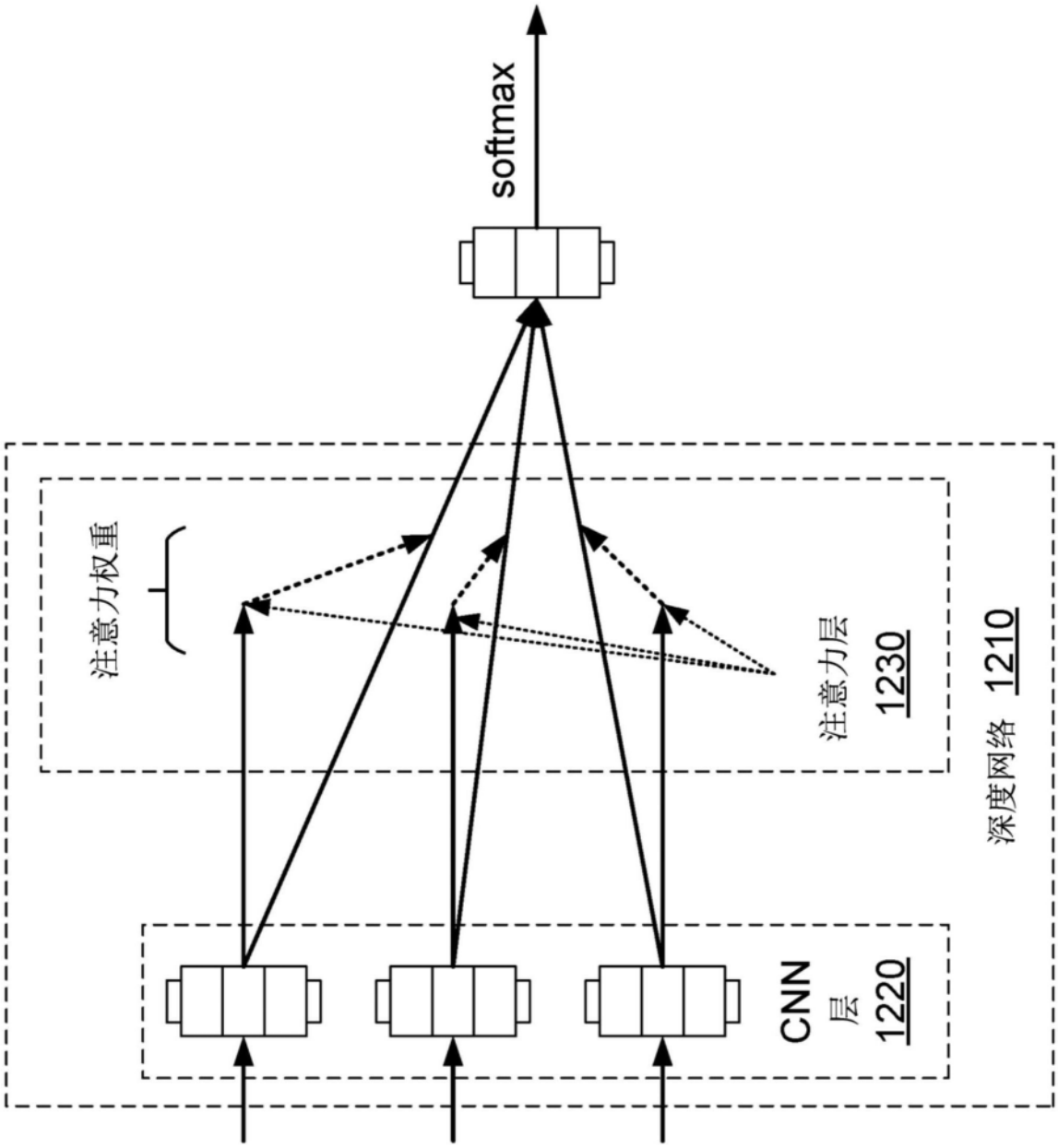


图12

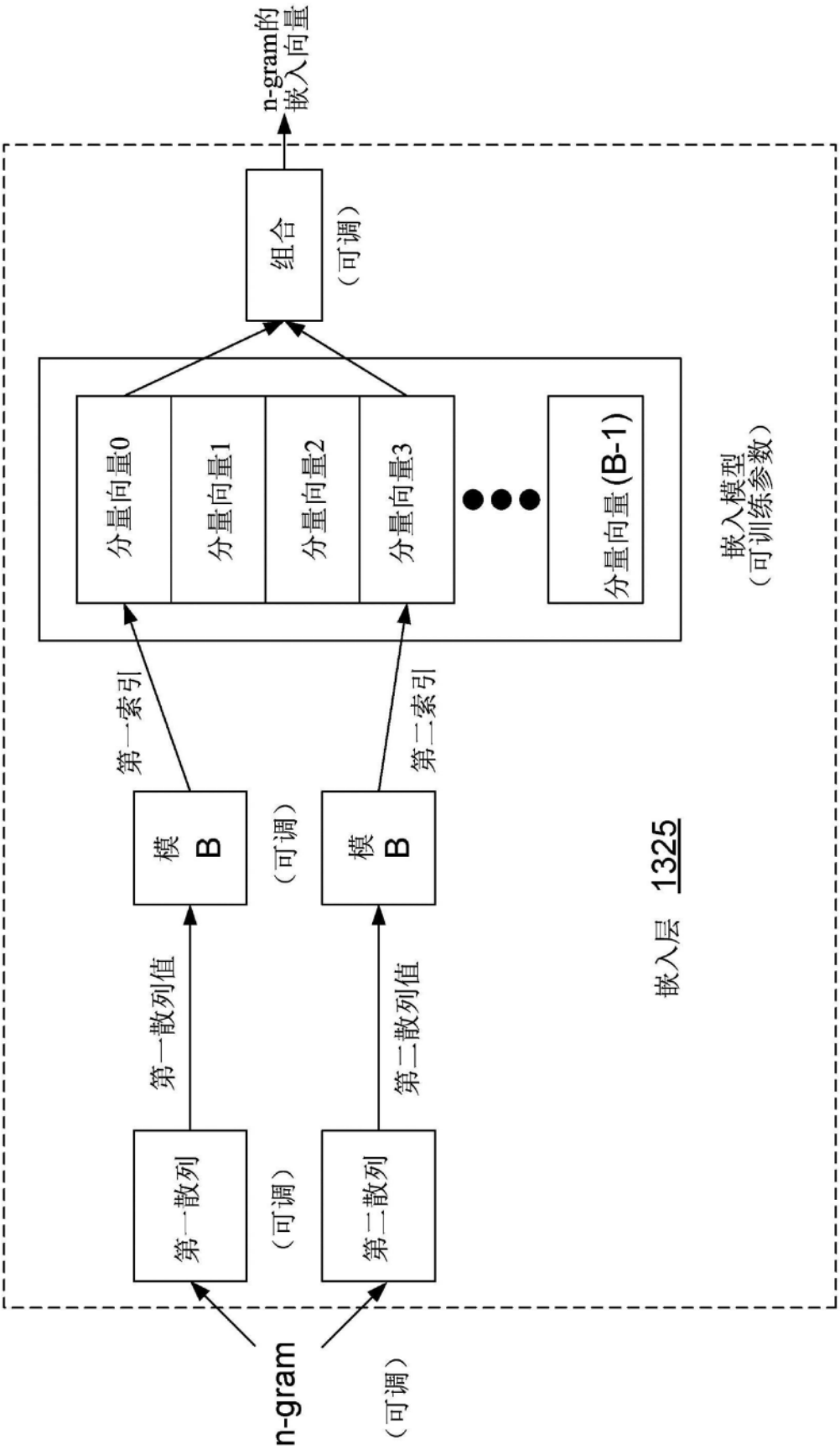

n-gram的嵌入向量
组合
（可调）
分量向量0
分量向量1
分量向量2
分量向量3
分量向量 (B-1)
嵌入模型（可训练参数）
第一索引
第二索引
模 B
（可调）
模 B
嵌入层 1325
第一散列值
第二散列值
第一散列
（可调）
第二散列
n-gram
（可调）

图13

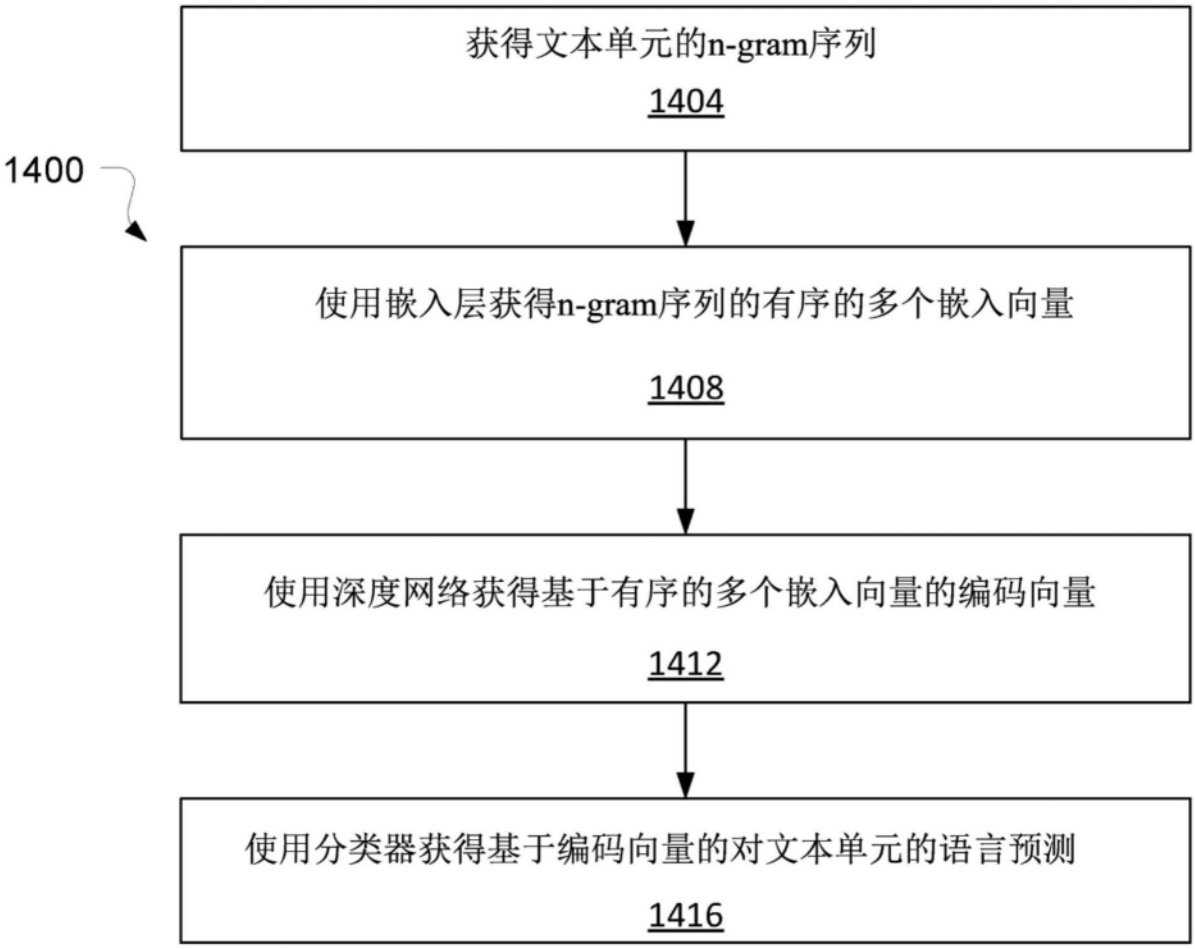

1400
获得文本单元的n-gram序列
1404
使用嵌入层获得n-gram序列的有序的多个嵌入向量
1408
使用深度网络获得基于有序的多个嵌入向量的编码向量
1412
使用分类器获得基于编码向量的对文本单元的语言预测
1416

图14

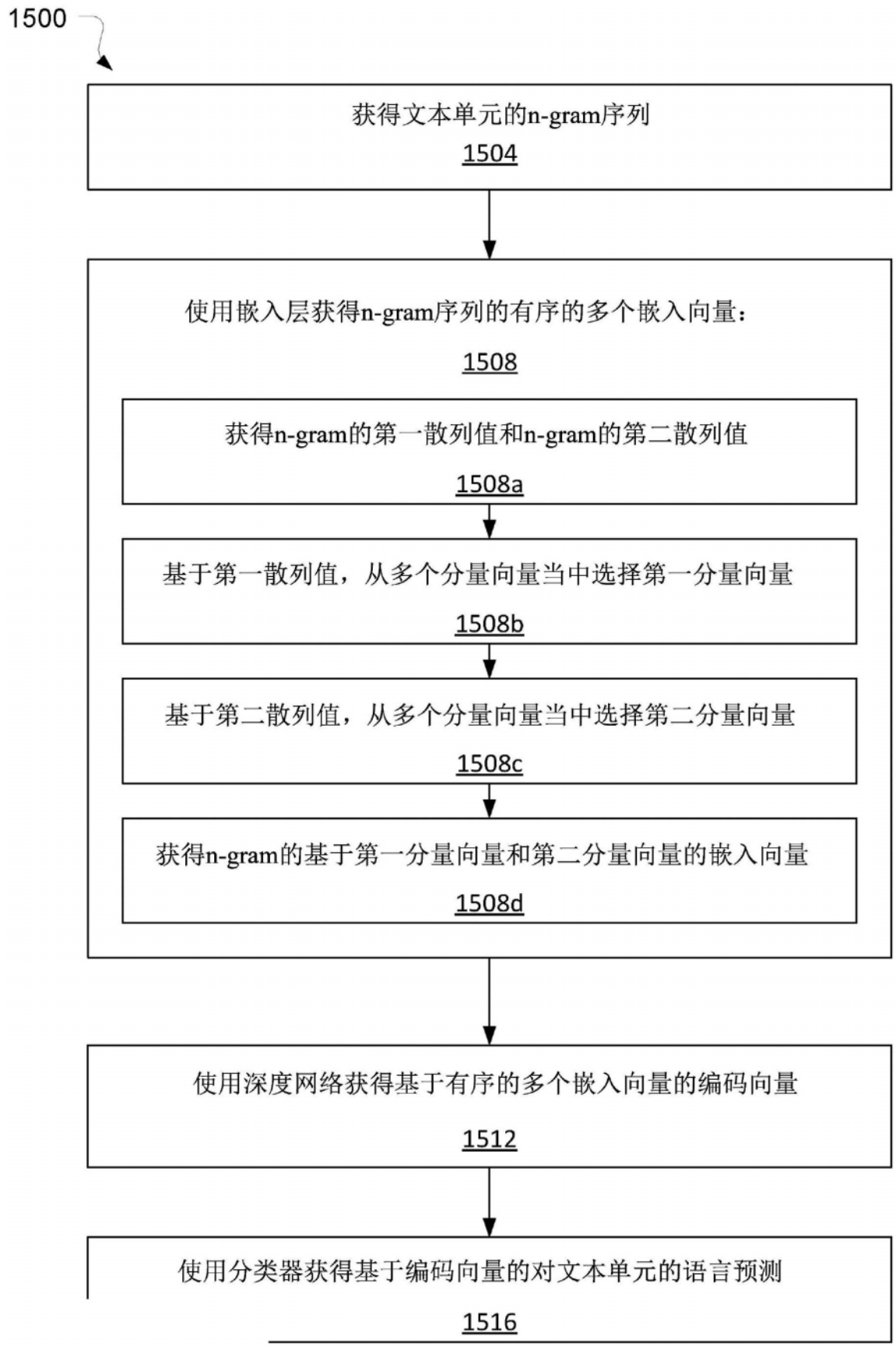

1500
获得文本单元的n-gram序列
1504
使用嵌入层获得n-gram序列的有序的多个嵌入向量：
1508
获得n-gram的第一散列值和n-gram的第二散列值
1508a
基于第一散列值，从多个分量向量当中选择第一分量向量
1508b
基于第二散列值，从多个分量向量当中选择第二分量向量
1508c
获得n-gram的基于第一分量向量和第二分量向量的嵌入向量
1508d
使用深度网络获得基于有序的多个嵌入向量的编码向量
1512
使用分类器获得基于编码向量的对文本单元的语言预测
1516

图15

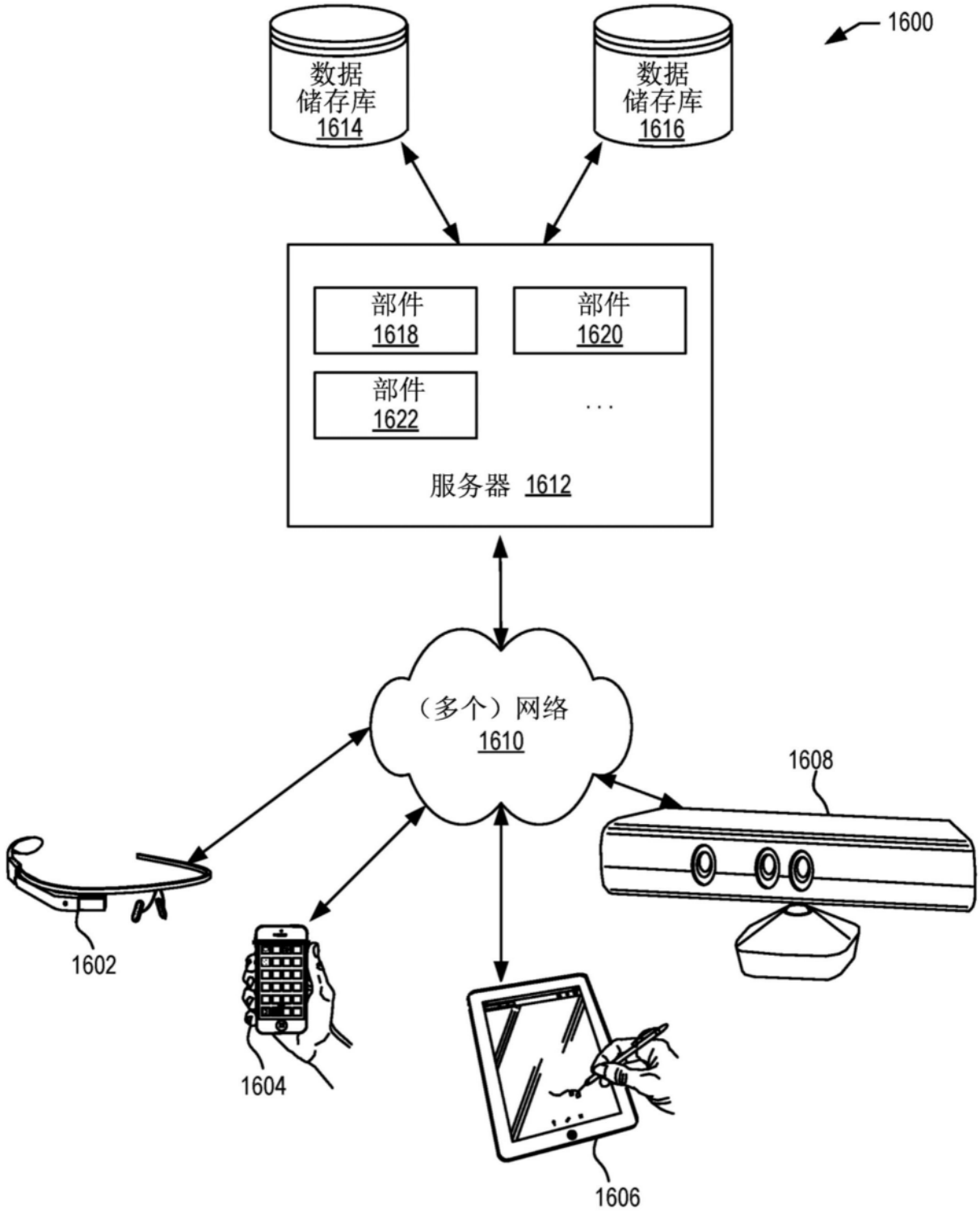

1600
数据
储存库
1614
数据
储存库
1616
部件
1618
部件
1620
部件
1622
...
服务器 1612
（多个）网络
1610
1608
1602
1604
1606

图16

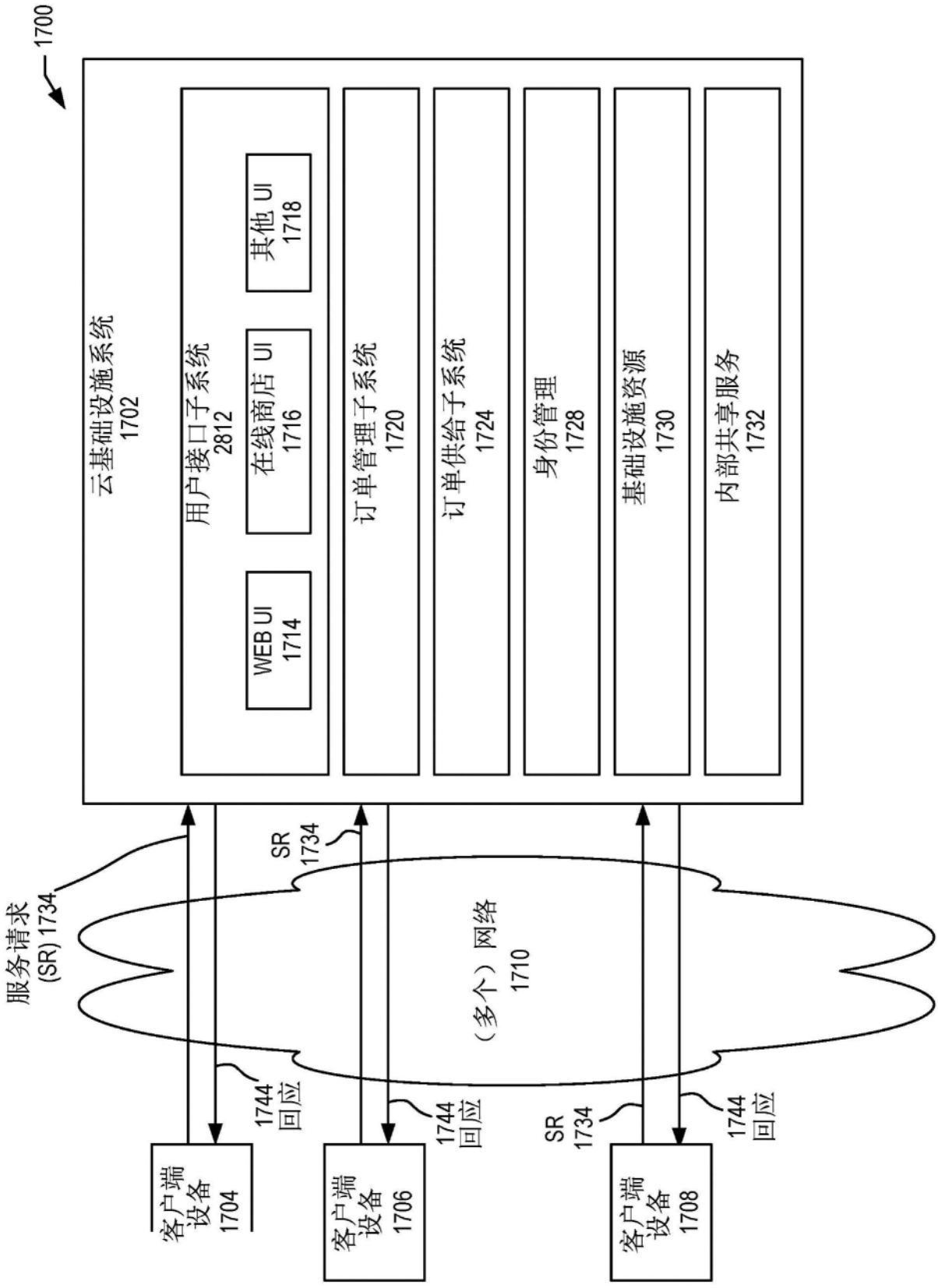

1700
云基础设施系统
1702
用户接口子系统
2812
WEB UI
1714
在线商店 UI
1716
其他 UI
1718
订单管理子系统
1720
订单供给子系统
1724
身份管理
1728
基础设施资源
1730
内部共享服务
1732
服务请求
(SR) 1734
SR
1734
（多个）网络
1710
1744
回应
1744
回应
SR
1734
1744
回应
客户端
设备
1704
客户端
设备
1706
客户端
设备
1708

图17

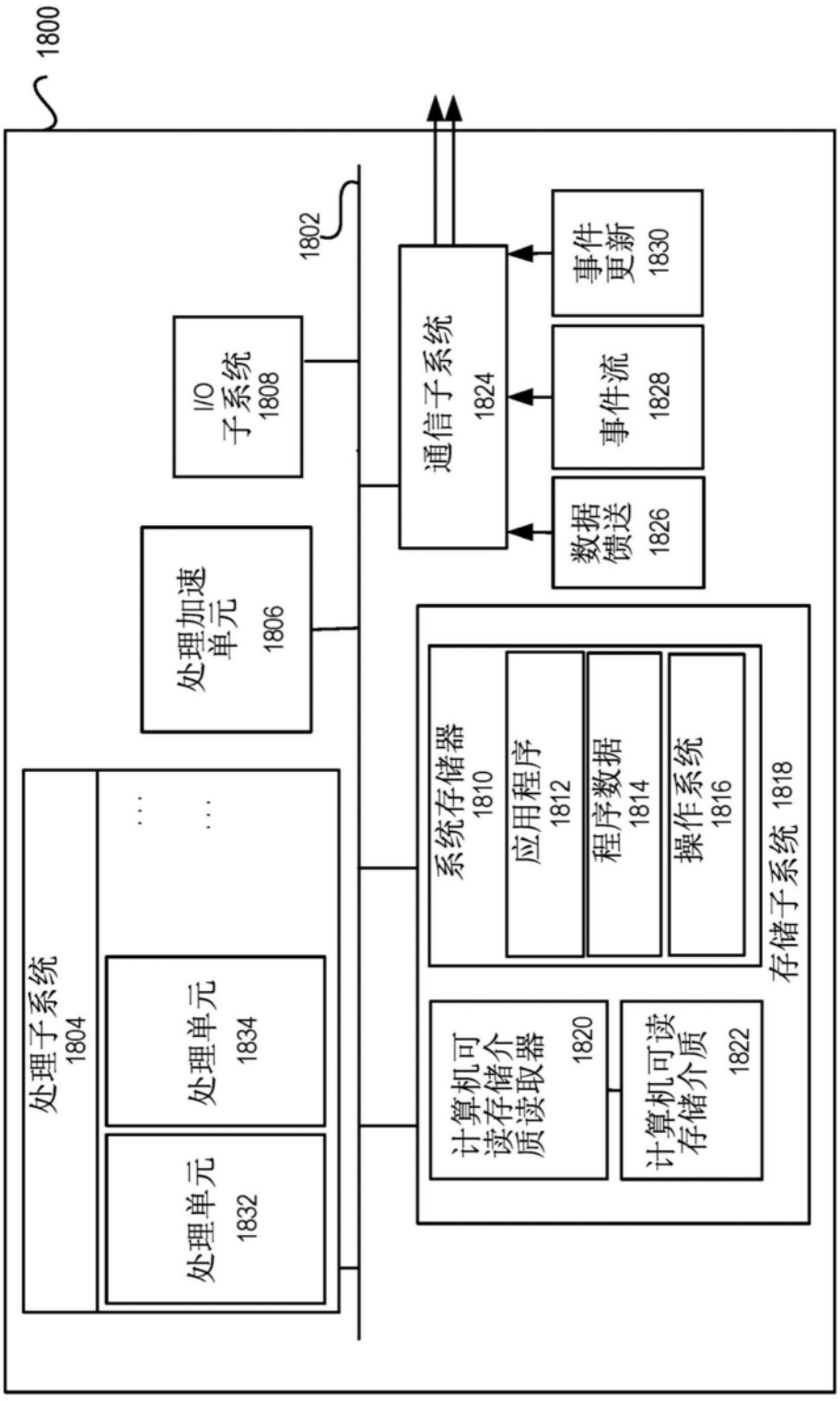

1800
1802
处理子系统
1804
处理单元
1832
处理单元
1834
...
...
处理加速单元
1806
I/O子系统
1808
通信子系统
1824
事件更新
1830
事件流
1828
数据馈送
1826
系统存储器
1810
应用程序
1812
程序数据
1814
操作系统
1816
存储子系统 1818
计算机可读存储介质读取器
1820
计算机可读存储介质
1822

图18