(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112364649 B
(45)授权公告日 2022.07.19

(21)申请号 202010936314.8

(22)申请日 2020.09.08

(65)同一申请的已公布的文献号
申请公布号 CN 112364649 A

(43)申请公布日 2021.02.12

(73)专利权人 深圳平安医疗健康科技服务有限公司
地址 518000 广东省深圳市福田区华强北街道华航社区华富路1018号中航中心2901

(72)发明人 王伟印

(74)专利代理机构 深圳市世联合知识产权代理有限公司 44385
专利代理师 汪琳琳

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06N 3/08(2006.01)
G06N 3/04(2006.01)

(56)对比文件
CN 109635288 A,2019.04.16
CN 106503192 A,2017.03.15
US 2019129932 A1,2019.05.02
CN 111353311 A,2020.06.30
CN 109918680 A,2019.06.21
CN 106815293 A,2017.06.09
CN 110929119 A,2020.03.27
CN 111476034 A,2020.07.31

审查员 李腾飞

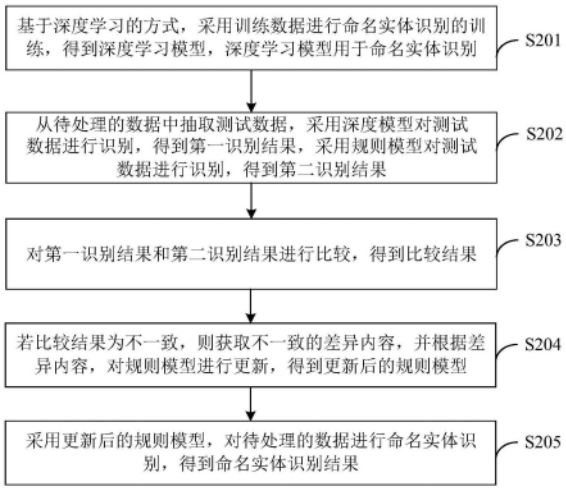
权利要求书2页 说明书11页 附图2页

(54)发明名称

命名实体的识别方法、装置、计算机设备及存储介质

(57)摘要

本发明涉及自然语言处理领域，公开了一种命名实体的识别方法、装置、计算机设备及存储介质，包括：基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型，从待处理的数据中抽取测试数据，采用深度学习模型对测试数据进行识别，得到第一识别结果，采用规则模型对测试数据进行识别，得到第二识别结果，对第一识别结果和第二识别结果进行比较，得到比较结果，若比较结果为不一致，则获取不一致的差异内容，并根据差异内容，对规则模型进行更新，采用更新后的规则模型，对待处理的数据快速命名实体识别，得到识别结果，本发明还涉及区块链技术，将得到的命名实体识别结果存储至区块链网络中，本发明提高命名实体识别的效率。



CN 112364649 B

1.一种命名实体的识别方法,其特征在于,包括:

基于深度学习的方式,采用训练数据进行命名实体识别的训练,得到深度学习模型,所述深度学习模型用于命名实体识别;

从待处理的数据中抽取测试数据,采用所述深度学习模型对测试数据进行识别,得到第一识别结果,采用规则模型对所述测试数据进行识别,得到第二识别结果;

对所述第一识别结果和所述第二识别结果进行比较,得到比较结果;

若所述比较结果为不一致,则获取不一致的差异内容,并根据所述差异内容,对所述规则模型进行更新,得到更新后的规则模型;

采用所述更新后的规则模型,对所述待处理的数据进行命名实体识别,得到命名实体识别结果。

2.如权利要求1所述的命名实体的识别方法,其特征在于,所述深度学习采用双向长短记忆神经网络,所述基于深度学习的方式,采用训练数据进行命名实体识别的训练,得到深度学习模型包括:

获取训练数据,其中,所述训练数据为标注好的语料数据;

将所述训练数据输入到初始双向长短时记忆神经网络模型中;

通过所述初始双向长短时记忆神经网络模型的预处理层,将所述训练数据转化为词向量;

采用所述词向量对所述初始双向长短时记忆神经网络模型进行训练,得到双向长短时记忆神经网络的输出矩阵;

使用所述输出矩阵的参数,对所述初始双向长短时记忆神经网络模型的参数进行更新,得到深度学习模型。

3.如权利要求2所述的命名实体的识别方法,其特征在于,所述标注好的语料数据采用BMES的标签标注,其中,B标签表示词首,M标签表示词中,E标签表示词尾,S标签表示单字。

4.如权利要求2所述的命名实体的识别方法,其特征在于,在所述使用所述输出矩阵的参数,对所述初始双向长短时记忆神经网络模型的参数进行更新,得到深度学习模型之后,所述命名实体的识别方法还包括:

获取新的标注样本数据作为验证数据,采用所述验证数据对深度学习模型是否达到预期结果;

若达到预期效果,则确认深度学习模型训练完成,若达不到预期效果,则重新选择深度学习的算法和参数,重新训练模型,直到深度学习模型达到预期效果。

5.如权利要求1所述的命名实体的识别方法,其特征在于,所述对所述第一识别结果和所述第二识别结果进行比较,得到比较结果包括:

通过预设规则,对第一识别结果和第二识别结果进行匹配,得到匹配结果;

若所述匹配结果中存在不匹配的命名实体,则分别从所述第一识别结果和所述第二识别结果中,获取所述不匹配的命名实体,作为待比较实体对;

对每对所述待比较实体对进行语义识别,得到语义识别结果,所述语义识别结果包括属于同一命名实体和不属于同一命名实体;

若每对所述待比较实体对的语义识别结果均为属于同一命名实体,则确认所述比较结果为一致;

若存在语义识别结果均为不属于同一命名实体，则确认比较结果为不一致。

6.如权利要求1所述的命名实体的识别方法，其特征在于，所述对所述规则模型进行更新包括增加规则和修改规则。

7.一种命名实体的识别装置，其特征在于，包括：

模型训练模块，用于基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型，所述深度学习模型用于命名实体识别；

模型测试模块，用于从待处理的数据中抽取测试数据，采用所述深度学习模型对测试数据进行识别，得到第一识别结果，采用规则模型对所述测试数据进行识别，得到第二识别结果；

结果比较模块，用于对所述第一识别结果和所述第二识别结果进行比较，得到比较结果；

模型更新模块，用于若所述比较结果为不一致，则获取不一致的差异内容，并根据所述差异内容，对所述规则模型进行更新，得到更新后的规则模型；

命名实体识别模块，用于采用所述更新后的规则模型，对所述待处理的数据进行命名实体识别，得到命名实体识别结果。

8.如权利要求7所述的命名实体的识别装置，其特征在于，所述模型训练模块模块包括：

数据获取单元，用于获取训练数据，其中，所述训练数据为标注好的语料数据；

数据输入单元，用于将所述训练数据输入到初始双向长短时记忆神经网络模型中；

数据预处理单元，用于通过所述初始双向长短时记忆神经网络模型的预处理层，将所述训练数据转化为词向量；

迭代训练单元，用于采用所述词向量对所述初始双向长短时记忆神经网络模型进行训练，得到双向长短时记忆神经网络的输出矩阵；

参数更新单元，用于使用所述输出矩阵的参数，对所述初始双向长短时记忆神经网络模型的参数进行更新，得到深度学习模型。

9.一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至6任一项所述的命名实体的识别方法。

10.一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至6任一项所述的命名实体的识别方法。

命名实体的识别方法、装置、计算机设备及存储介质

技术领域

[0001] 本发明涉及人工智能领域，尤其涉及一种命名实体的识别方法、装置、计算机设备及存储介质。

背景技术

[0002] 人工智能(Artificial Intelligence)领域是一种新的能以人类智能相似的方式做出反应的智能处理领域，该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等，其中，在自然语言处理中，由于自然语言与机器语言区别较大，为了计算机可以更为准确了解自然语言中表达的意图，往往需要用到深度学习的方式。，例如在自然语言处理的命名实体的识别问题上，通过深度学习的方式，使得很多问题更容易处理，且准确率较高。

[0003] 通常采用深度学习模型进行命名实体识别，是从大量样本数据中学习得到的，要表征大量的样本数据，模型通常会有很多的权重参数，通过训练得到合适的权重参数模型即可实现命名实体识别的功能。

[0004] 在实现本申请的过程中，发明人意识到现有方式至少存入如下问题：深度学习模型先天的存在过拟合或者欠拟合的问题，而且模型由于存在大量的参数，使得算法执行效率比较低，在很多追求执行效率的工作场景，如大数据的场景，很难采用深度学习的方式来解决，但是在很多规则覆盖问题上，利用深度学习模型是可以很好的解决规则覆盖率的问题，但由于要处理的数据量极大，模型的效率问题是实际工作中不能接受的。因而，亟需一种能在数据量较大时实现快速进行命名实体的识别方法。

发明内容

[0005] 本发明实施例提供一种命名实体的识别方法、装置、计算机设备和存储介质，以实现在数据量较大时提高命名实体的识别效率。

[0006] 为了解决上述技术问题，本申请实施例提供一种命名实体的识别方法，包括：

[0007] 基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型，所述深度学习模型用于命名实体识别；

[0008] 从待处理的数据中抽取测试数据，采用所述深度学习模型对测试数据进行识别，得到第一识别结果，采用规则模型对所述测试数据进行识别，得到第二识别结果；

[0009] 对所述第一识别结果和所述第二识别结果进行比较，得到比较结果；

[0010] 若所述比较结果为不一致，则获取不一致的差异内容，并根据所述差异内容，对所述规则模型进行更新，得到更新后的规则模型；

[0011] 采用所述更新后的规则模型，对所述待处理的数据进行命名实体识别，得到命名实体识别结果。

[0012] 可选地，所述深度学习采用双向长短记忆神经网络，所述基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型包括：

[0013] 获取训练数据，其中，所述训练数据为标注好的语料数据；

[0014] 将所述训练数据输入到初始双向长短时记忆神经网络模型中；

[0015] 通过所述初始双向长短时记忆神经网络模型的预处理层，将所述训练数据转化为词向量；

[0016] 采用所述词向量对所述初始双向长短时记忆神经网络模型进行训练，得到双向长短时记忆神经网络的输出矩阵；

[0017] 使用所述输出矩阵的参数，对所述初始双向长短时记忆神经网络模型的参数进行更新，得到深度学习模型。

[0018] 可选地，所述标注好的语料数据采用BMES的标签标注，其中，B标签表示词首，M标签表示词中，E标签表示词尾，S标签表示单字。

[0019] 可选地，在所述使用所述输出矩阵的参数，对所述初始双向长短时记忆神经网络模型的参数进行更新，得到深度学习模型之后，所述命名实体的识别方法还包括：

[0020] 获取新的标注样本数据作为验证数据，采用所述验证数据对深度学习模型是否达到预期结果；

[0021] 若达到预期效果，则确认深度学习模型训练完成，若达不到预期效果，则重新选择深度学习的算法和参数，重新训练模型，直到深度学习模型达到预期效果。

[0022] 可选地，所述对所述第一识别结果和所述第二识别结果进行比较，得到比较结果包括：

[0023] 通过预设规则，对第一识别结果和第二识别结果进行匹配，得到匹配结果；

[0024] 若所述匹配结果中存在不匹配的命名实体，则分别从所述第一识别结果和所述第二识别结果中，获取所述不匹配的命名实体，作为待比较实体对；

[0025] 对每对所述待比较实体对进行语义识别，得到语义识别结果，所述语义识别结果包括属于同一命名实体和不属于同一命名实体；

[0026] 若每对所述待比较实体对的语义识别结果均为属于同一命名实体，则确认所述比较结果为一致；

[0027] 若存在语义识别结果均为不属于同一命名实体，则确认比较结果为不一致。

[0028] 可选地，所述对所述规则模型进行更新包括增加规则和修改规则。

[0029] 为了解决上述技术问题，本申请实施例还提供一种命名实体的识别装置，包括：

[0030] 模型训练模块，用于基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型，所述深度学习模型用于命名实体识别；

[0031] 模型测试模块，用于从待处理的数据中抽取测试数据，采用所述深度学习模型对测试数据进行识别，得到第一识别结果，采用规则模型对所述测试数据进行识别，得到第二识别结果；

[0032] 结果比较模块，用于对所述第一识别结果和所述第二识别结果进行比较，得到比较结果；

[0033] 模型更新模块，用于若所述比较结果为不一致，则获取不一致的差异内容，并根据所述差异内容，对所述规则模型进行更新，得到更新后的规则模型；

[0034] 命名实体识别模块，用于采用所述更新后的规则模型，对所述待处理的数据进行命名实体识别，得到命名实体识别结果。

[0035] 可选地，所述模型训练模块包括：

[0036] 数据获取单元，用于获取训练数据，其中，所述训练数据为标注好的语料数据；

[0037] 数据输入单元，用于将所述训练数据输入到初始双向长短时记忆神经网络模型中；

[0038] 数据预处理单元，用于通过所述初始双向长短时记忆神经网络模型的预处理层，将所述训练数据转化为词向量；

[0039] 迭代训练单元，用于采用所述词向量对所述初始双向长短时记忆神经网络模型进行训练，得到双向长短时记忆神经网络的输出矩阵；

[0040] 参数更新单元，用于使用所述输出矩阵的参数，对所述初始双向长短时记忆神经网络模型的参数进行更新，得到深度学习模型。

[0041] 可选地，所述命名实体的识别装置还包括：

[0042] 模型验证模块，用于获取新的标注样本数据作为验证数据，采用所述验证数据对深度学习模型是否达到预期结果；

[0043] 训练结果判断模块，用于若达到预期效果，则确认深度学习模型训练完成，若达不到预期效果，则重新选择深度学习的算法和参数，重新训练模型，直到深度学习模型达到预期效果。

[0044] 可选地，所述结果比较模块包括：

[0045] 匹配单元，用于通过预设规则，对第一识别结果和第二识别结果进行匹配，得到匹配结果；

[0046] 待比较实体对确定单元，用于若所述匹配结果中存在不匹配的命名实体，则分别从所述第一识别结果和所述第二识别结果中，获取所述不匹配的命名实体，作为待比较实体对；

[0047] 语义识别单元，用于对每对所述待比较实体对进行语义识别，得到语义识别结果，所述语义识别结果包括属于同一命名实体和不属于同一命名实体；

[0048] 第一比较结果确定单元，用于若每对所述待比较实体对的语义识别结果均为属于同一命名实体，则确认所述比较结果为一致；

[0049] 第二比较结果确定单元，用于若存在语义识别结果均为不属于同一命名实体，则确认比较结果为不一致。

[0050] 为了解决上述技术问题，本申请实施例还提供一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现上述命名实体的识别方法的步骤。

[0051] 为了解决上述技术问题，本申请实施例还提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述命名实体的识别方法的步骤。

[0052] 本发明实施例提供的命名实体的识别方法、装置、计算机设备及存储介质，一方面，基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型，深度学习模型用于命名实体识别，再从待处理的数据中抽取测试数据，采用深度学习模型对测试数据进行识别，得到第一识别结果，采用规则模型对测试数据进行识别，得到第二识别结果，进而对第一识别结果和第二识别结果进行比较，得到比较结果，若比较结果为不一

致，则获取不一致的差异内容，并根据差异内容，对规则模型进行更新，得到更新后的规则模型，实现通过对未知的命名实体识别度高的深度学习模型来对规则模型进行更新，提高规则模型的识别准确率，另一方面，采用更新后的规则模型，对待处理的数据进行快速命名实体识别，得到命名实体识别结果，有利于提高命名实体识别的效率。

附图说明

[0053] 为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例的描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0054] 图1是本申请可以应用于其中的示例性系统架构图；

[0055] 图2是本申请的命名实体的识别方法的一个实施例的流程图；

[0056] 图3是根据本申请的命名实体的识别装置的一个实施例的结构示意图；

[0057] 图4是根据本申请的计算机设备的一个实施例的结构示意图。

具体实施方式

[0058] 除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同；本文中在申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请；本申请的说明书和权利要求书及上述附图说明中的术语“包括”和“具有”以及它们的任何变形，意图在于覆盖不排他的包含。本申请的说明书和权利要求书或上述附图中的术语“第一”、“第二”等是用于区别不同对象，而不是用于描述特定顺序。

[0059] 在本文中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

[0060] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0061] 请参阅图1，如图1所示，系统架构100可以包括终端设备101、102、103，网络104和服务器105。网络104用以在终端设备101、102、103和服务器105 之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0062] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。

[0063] 终端设备101、102、103可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、MP3播放器(Moving Picture E界面显示perts Group Audio Layer III，动态影像专家压缩标准音频层面3)、MP4(Moving Picture E界面显示perts Group Audio Layer IV，动态影像专家压缩标准音频层面4)播放器、膝上

型便携计算机和台式计算机等等。

[0064] 服务器105可以是提供各种服务的服务器，例如对终端设备101、102、103上显示的页面提供支持的后台服务器。

[0065] 需要说明的是，本申请实施例所提供的命名实体的识别方法由服务器执行，相应地，命名实体的识别装置设置于服务器中。

[0066] 应该理解，图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器，本申请实施例中的终端设备101、102、103具体可以对应的是实际生产中的应用系统。

[0067] 请参阅图2，图2示出本发明实施例提供的一种命名实体的识别方法，以该方法应用在图1中的服务端为例进行说明，详述如下：

[0068] S201：基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型，深度学习模型用于命名实体识别。

[0069] 具体地，通过深度学习的方式，采用大量的训练数据进行命名实体识别的训练，得到训练好的深度学习模型。

[0070] 应理解，本实施例中，采用深度学习的方式，进行命名实体识别训练，得到的深度学习模型，属于深度学习模型，其对面未预先定义的命名实体也具有较好的识别准确率。

[0071] 其中，深度学习(DL，Deep Learning)是学习样本数据的内在规律和表示层次，是利用深度神经网络来解决特征表达的一种学习过程。深度神经网络本身并不是一个全新的概念，可大致理解为包含多个隐含层的神经网络结构。为了提高深层神经网络的训练效果，人们对神经元的连接方法和激活函数等方面做出相应的调整，这些学习过程中获得的信息对诸如文字，图像和声音等数据的解释有很大的帮助。它的最终目标是让机器能够像人一样具有分析学习能力，能够识别文字、图像和声音等数据。深度学习是一个复杂的机器学习算法，在语音和图像识别方面取得的效果，远远超过先前相关技术。深度学习在搜索技术，数据挖掘，机器学习，机器翻译，自然语言处理，多媒体学习，语音，推荐和个性化技术，以及其他相关领域都取得了很多成果。

[0072] 深度学习最基本的做法，是使用算法来解析数据、从中学习，然后对真实世界中的事件做出决策和预测。与传统的为解决特定任务、硬编码的软件程序不同，深度学习是用大量的数据来“训练”，通过各种算法从数据中学习如何完成任务。举个简单的例子，当我们浏览网上商城时，经常会出现商品推荐的信息。这是商城根据你往期的购物记录和冗长的收藏清单，识别出这其中哪些是你真正感兴趣，并且愿意购买的产品。这样的决策模型，可以帮助商城为客户提供建议并鼓励产品消费。

[0073] 其中，训练数据可以是通过网络爬虫的方式，按照业务需求，爬取业务需要的语料数据，网络爬虫又称全网爬虫(Scalable Web Crawler)，爬行对象从一些种子URL(Uniform Resource Locator，统一资源定位符)扩充到整个Web (World Wide Web，全球广域网)，主要为门户站点搜索引擎和大型Web服务提供商采集数据。

[0074] S202：从待处理的数据中抽取测试数据，采用深度学习模型对测试数据进行识别，得到第一识别结果，采用规则模型对测试数据进行识别，得到第二识别结果。

[0075] 具体地，在接收到待处理的数据样本的数据量较大时，从该待处理的数据样本中，随机取出小部分数据，作为测试数据，并分别采用深度学习模型和规则模型对该测试数据

进行识别,得到第一识别结果和第二识别结果。

[0076] 其中,模型规则又称为规则依存模型(RDM Rule Dependence Model),是指专家系统或决策支持系统推理获得的知识基于或依赖系统已存在的规则库和推理机制。规则库(Rule Base)是指一个用规则来表达的知识集,包括执行推理所需要的知识。依赖关系表示两个活动(前导活动和后续活动)中一个活动的变更将会影响到另一个活动的关系。

[0077] 需要说明的是,本实施例中的规则模型可以是先前已经有部分规则,也可以是空的(也即,不存在任何依赖关系),在规则模型为空时,采用后续方法相当于建立一个与深度学习模型功能相等的规则模型。

[0078] 应理解,深度学习模型为经过深度学习后得到的模型,对一些未知的数据具有较好的识别准确率,因而,得到的第一识别结果较为准确,规则模型采用预设的规则来对测试数据进行判断,识别速度较快,但是对一些未知的数据(为命中任何规则的数据),很难实现准确识别,由此,第一识别结果和第二识别结果可能相同,也可能不同。

[0079] S203:对第一识别结果和第二识别结果进行比较,得到比较结果。

[0080] 具体地,第一识别结果中,包含多个命名实体的识别,第二识别结果中,也包含多个命名实体的识别,对第一识别结果和第二识别结果中识别到的命名实体进行对比,得到第一识别结果和第二识别结果的比较结果。

[0081] 容易理解地,比较结果包括两种情况:识别到的命名实体完全一致和不完全一致,在存在不一致时,以第一识别结果为准,因为深度学习得到的深度学习模型,对未知类型的数据具有较高的识别准确率,而规则模型在未预设对应规则时,对未知数据不易准确识别。

[0082] 值得说明的是,考虑到需要对第一识别结果和第二识别结果进行比较,为提高比较效率和准确度,本实施例在进行识别时,对每个测试数据添加唯一的数据标识,在进行比较时,直接获取第一识别结果和第二识别结果中,具有相同数据标识的识别结果来进行比较,避免部分测试数据在第二识别结果为空时,导致第一识别结果和第二识别结果的比较不准确。

[0083] S204:若比较结果为不一致,则获取不一致的差异内容,并根据差异内容,对规则模型进行更新,得到更新后的规则模型。

[0084] 具体地,在将深度学习训练得到的深度学习模型的识别结果(第一识别结果)与规则模型的识别结果(第二识别结果)逐步对比,当碰到规则模型识别不准确或不能识别的情况(即第二识别结果无法识别),则根据第一识别结果与第二识别结果不一致的内容,对规则模型进行更新。

[0085] 进一步地,对规则模型进行更新,可以是对一条已有的规则进行更新,也可以是添加一条新的规则,这样可以逐步提升规则覆盖率,有利于提高采用规则模型对样本数据进行识别的准确率。

[0086] 应理解,在比较结果为一致时,直接采用现有的规则模型作为更新后的规则模型,来执行步骤S205的数据处理过程。

[0087] S205:采用更新后的规则模型,对待处理的数据进行命名实体识别,得到命名实体识别结果。

[0088] 具体地,经过更新后的规则模型,对待处理的数据具有较高的识别准确率,而规则模型相对于深度学习的模型,识别速度更快,因而,本实施例中,采用更新后的规则模型,对

待处理的数据进行命名实体识别，得到命名实体识别结果，提高识别效率。

[0089] 在本实施例中，一方面，基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型，深度学习模型用于命名实体识别，再从待处理的数据中抽取测试数据，采用深度学习模型对测试数据进行识别，得到第一识别结果，采用规则模型对测试数据进行识别，得到第二识别结果，进而对第一识别结果和第二识别结果进行比较，得到比较结果，若比较结果为不一致，则获取不一致的差异内容，并根据差异内容，对规则模型进行更新，得到更新后的规则模型，实现通过对未知的命名实体识别度高的深度学习模型来对规则模型进行更新，提高规则模型的识别准确率，另一方面，采用更新后的规则模型，对待处理的数据进行快速命名实体识别，得到命名实体识别结果，有利于提高命名实体识别的效率。

[0090] 在本实施例的一些可选的实现方式中，步骤S201中，深度学习采用双向长短记忆神经网络，基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型包括：

[0091] 获取训练数据，其中，训练数据为标注好的语料数据；

[0092] 将训练数据输入到初始双向长短时记忆神经网络模型中；

[0093] 通过初始双向长短时记忆神经网络模型的预处理层，将训练数据转化为词向量；

[0094] 采用词向量对初始双向长短时记忆神经网络模型进行训练，得到双向长短时记忆神经网络的输出矩阵；

[0095] 使用输出矩阵的参数，对初始双向长短时记忆神经网络模型的参数进行更新，得到深度学习模型。

[0096] 具体地，命名实体的识别，属于自然语言处理领域，本实施例通过采用双向长短记忆神经网络，对深度学习模型进行训练，提高语义识别的准确性，有利于提高深度学习模型的识别准确率。

[0097] 在本实施例的一些可选的实现方式中，标注好的语料数据采用BMES的标签标注，其中，B标签表示词首，M标签表示词中，E标签表示词尾，S标签表示单字。

[0098] 具体地，在一具体实施方式中，采用BMES的标注方式来标注语料数据中的每个字c对应的可能性，B(c)由语料数据中以c作为命名实体开头的词组组成，类似地，M(c)包括所有以c作为命名实体的中间部分的词组，E(c)包括可能以c作为命名实体的结尾的所有词组，而S(c)是c这个单字本身。如果词组集合为空，我们将在其中添加特殊单词“NONE”以指示这种情况。通过这种方式，我们现在可以引入预训练的单词嵌入，而且，我们可以从每个字符的词组集合中准确恢复相应的匹配结果。

[0099] 在本实施例中，通过采用BMES的标签标注，有利于提高命名实体识别准确性。

[0100] 在本实施例的一些可选的实现方式中，在使用输出矩阵的参数，对初始双向长短时记忆神经网络模型的参数进行更新，得到深度学习模型之后，命名实体的识别方法还包括：

[0101] 获取新的标注样本数据作为验证数据，采用验证数据对深度学习模型是否达到预期结果；

[0102] 若达到预期效果，则确认深度学习模型训练完成，若达不到预期效果，则重新选择深度学习的算法和参数，重新训练模型，直到深度学习模型达到预期效果。

[0103] 具体地，用于深度学习的算法模型越来越多，在不同场景中，不同模型所能达到的效果也有所不同，为确保深度学习模型的准确率，需要对深度学习模型进行验证，本实施例中，获取新的标注样本数据作为验证数据，采用验证数据对深度学习模型是否达到预期结果，若达到预期效果，则确认深度学习模型训练完成，若达不到预期效果，则重新选择深度学习的算法和参数，重新训练模型，直到深度学习模型达到预期效果。

[0104] 其中，预期效果的检验方式包括但不限于：准确率、召回率(Recall)或者 F1值(F1 measure)等，也可以根据实际业务场景需要进行制定，此处不做限制，例如，在一具体实施方式中，预期结果为准确率达到90％。

[0105] 在本实施例中，通过对训练好的深度学习模型进行校验，确保深度学习模型的准确率，有利于后面通过深度学习模型对规则模型进行更新时，提高规则模型的识别准确率。

[0106] 在本实施例的一些可选的实现方式中，步骤S203中，对第一识别结果和第二识别结果进行比较，得到比较结果包括：

[0107] 通过预设规则，对第一识别结果和第二识别结果进行匹配，得到匹配结果；

[0108] 若匹配结果中存在不匹配的命名实体，则分别从第一识别结果和第二识别结果中，获取不匹配的命名实体，作为待比较实体对；

[0109] 对每对待比较实体对进行语义识别，得到语义识别结果，语义识别结果包括属于同一命名实体和不属于同一命名实体；

[0110] 若每对待比较实体对的语义识别结果均为属于同一命名实体，则确认比较结果为一致；

[0111] 若存在语义识别结果均为不属于同一命名实体，则确认比较结果为不一致。

[0112] 具体地，通过预先设置的匹配规则，对第一识别结果和第二识别结果进行匹配，得到匹配结果，在匹配结果中存在不匹配的命名实体时，分别从第一识别结果和第二识别结果中，获取不匹配的命名实体，作为待比较实体对，通过语义识别的方式，对每对待比较实体对进行语义识别，得到语义识别结果，基于语义识别结果，确认是否属于同一命名实体。

[0113] 其中，预设规则可以根据实际需求进行条件设置，例如，文本相似度达到 80％或者模糊匹配等，此处不做限定。

[0114] 其中，匹配结果包完全匹配和不完全匹配，完全匹配是指第一识别结果和第二识别结果中，每个相同数据标识对应的命名实体识别结果均一致。

[0115] 其中，不完全匹配是指第一识别结果和第二识别结果中，存在不匹配的命名实体，也即，存在相同数据标识对应的命名实体识别结果不一致的情况，不一致的命名实体识别结果数量，可以是一个，也可以是多个，或者是全部不一致，具体依实际识别结果而定，此处不作具体限制。

[0116] 本实施例中，通过预设规则，对第一识别结果和第二识别结果进行匹配，具体是从字面上进行匹配，但是自然语言中，涉及到一个词汇存在多种近义词、同义词和缩写词等，他们字面上不能匹配，但是实际上表达了相同的语义，因而，需要对匹配不一致的待比较实体对进行语义识别，根据语义识别结果来进一步确定实体识别出的实体，是否为同一实体。对每对所述待比较实体对进行语义识别，得到语义识别结果，是采用自然语义语义识别的方式实现。

[0117] 其中，自然语言语义识别(Natural Language Processing,NLP)是人工智能(AI)

的一个子领域,通过机器学习的方式,对自然语言进行理解解析,从而解决自然语言领域的一些问题,NLP主要应用范围包括但不限于:文本朗读(Text to speech)/语音合成(Speech synthesis)、语音识别(Speech recognition)、中文自动分词(Chinese word segmentation)、词性标注(Part-of-speech tagging)、句法分析(Parsing)、文本分类(Text categorization)、信息检索(Information retrieval)、自动摘要(Automatic summarization)和文字校对(Text-proofing) 等,进行自然语义语义识别的方式包括但不限于:马可夫模型(Markov models)、N-gram模型、fastText模型、基于注意力机制的TextRNN等。

[0118] 在本实施例中,通过文字匹配和语义识别,对识别结果的一致性进行判断,提高一致性识别结果判断的准确性。

[0119] 应理解,上述实施例中各步骤的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定,而不应对本发明实施例的实施过程构成任何限定。

[0120] 图3示出与上述实施例命名实体的识别方法一一对应的命名实体的识别装置的原理框图。如图3所示,该命名实体的识别装置包括模型训练模块31、模型测试模块32、结果比较模块33、模型更新模块34和命名实体识别模块 35。各功能模块详细说明如下:

[0121] 模型训练模块31,用于基于深度学习的方式,采用训练数据进行命名实体识别的训练,得到深度学习模型,深度学习模型用于命名实体识别;

[0122] 模型测试模块32,用于从待处理的数据中抽取测试数据,采用深度学习模型对测试数据进行识别,得到第一识别结果,采用规则模型对测试数据进行识别,得到第二识别结果;

[0123] 结果比较模块33,用于对第一识别结果和第二识别结果进行比较,得到比较结果;

[0124] 模型更新模块34,用于若比较结果为不一致,则获取不一致的差异内容,并根据差异内容,对规则模型进行更新,得到更新后的规则模型;

[0125] 命名实体识别模块35,用于采用更新后的规则模型,对待处理的数据进行命名实体识别,得到命名实体识别结果。

[0126] 可选地,模型训练模块31包括:

[0127] 数据获取单元,用于获取训练数据,其中,训练数据为标注好的语料数据;

[0128] 数据输入单元,用于将训练数据输入到初始双向长短时记忆神经网络模型中;

[0129] 数据预处理单元,用于通过初始双向长短时记忆神经网络模型的预处理层,将训练数据转化为词向量;

[0130] 迭代训练单元,用于采用词向量对初始双向长短时记忆神经网络模型进行训练,得到双向长短时记忆神经网络的输出矩阵;

[0131] 参数更新单元,用于使用输出矩阵的参数,对初始双向长短时记忆神经网络模型的参数进行更新,得到深度学习模型。

[0132] 可选地,命名实体的识别装置还包括:

[0133] 模型验证模块,用于获取新的标注样本数据作为验证数据,采用验证数据对深度学习模型是否达到预期结果;

[0134] 训练结果判断模块,用于若达到预期效果,则确认深度学习模型训练完成,若达不

到预期效果，则重新选择深度学习的算法和参数，重新训练模型，直到深度学习模型达到预期效果。

[0135] 可选地，结果比较模块33包括：

[0136] 匹配单元，用于通过预设规则，对第一识别结果和第二识别结果进行匹配，得到匹配结果；

[0137] 待比较实体对确定单元，用于若匹配结果中存在不匹配的命名实体，则分别从第一识别结果和第二识别结果中，获取不匹配的命名实体，作为待比较实体对；

[0138] 语义识别单元，用于对每对待比较实体对进行语义识别，得到语义识别结果，语义识别结果包括属于同一命名实体和不属于同一命名实体；

[0139] 第一比较结果确定单元，用于若每对待比较实体对的语义识别结果均为属于同一命名实体，则确认比较结果为一致；

[0140] 第二比较结果确定单元，用于若存在语义识别结果均为不属于同一命名实体，则确认比较结果为不一致。

[0141] 关于命名实体的识别装置的具体限定可以参见上文中对于命名实体的识别方法的限定，在此不再赘述。上述命名实体的识别装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

[0142] 为解决上述技术问题，本申请实施例还提供计算机设备。具体请参阅图4，图4为本实施例计算机设备基本结构框图。

[0143] 所述计算机设备4包括通过系统总线相互通信连接存储器41、处理器42、网络接口43。需要指出的是，图中仅示出了具有组件连接存储器41、处理器 42、网络接口43的计算机设备4，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。其中，本技术领域技术人员可以理解，这里的计算机设备是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路(Application Specific Integrated Circuit，ASIC)、可编程门阵列(Field－Programmable Gate Array，FPGA)、数字处理器(Digital Signal Processor，DSP)、嵌入式设备等。

[0144] 所述计算机设备可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述计算机设备可以与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互。

[0145] 所述存储器41至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器(例如，SD或D界面显示存储器等)、随机访问存储器(RAM)、静态随机访问存储器(SRAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、可编程只读存储器(PROM)、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器41可以是所述计算机设备4的内部存储单元，例如该计算机设备4的硬盘或内存。在另一些实施例中，所述存储器41也可以是所述计算机设备4的外部存储设备，例如该计算机设备4上配备的插接式硬盘，智能存储卡(Smart Media Card，SMC)，安全数字(Secure Digital，SD)卡，闪存卡(Flash Card)等。当然，所述存储器 41还可以既包括所述计算机设备4的内部存储单元也

包括其外部存储设备。本实施例中，所述存储器41通常用于存储安装于所述计算机设备4的操作系统和各类应用软件，例如电子文件的控制的程序代码等。此外，所述存储器 41还可以用于暂时地存储已经输出或者将要输出的各类数据。

[0146] 所述处理器42在一些实施例中可以是中央处理器(Central Processing Unit，CPU)、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器42通常用于控制所述计算机设备4的总体操作。本实施例中，所述处理器42用于运行所述存储器41中存储的程序代码或者处理数据，例如运行电子文件的控制的程序代码。

[0147] 所述网络接口43可包括无线网络接口或有线网络接口，该网络接口43 通常用于在所述计算机设备4与其他电子设备之间建立通信连接。

[0148] 本申请还提供了另一种实施方式，即提供一种计算机可读存储介质，所述计算机可读存储介质存储有界面显示程序，所述界面显示程序可被至少一个处理器执行，以使所述至少一个处理器执行如上述的命名实体的识别方法的步骤。

[0149] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，空调器，或者网络设备等)执行本申请各个实施例所述的方法。

[0150] 显然，以上所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例，附图中给出了本申请的较佳实施例，但并不限制本申请的专利范围。本申请可以以许多不同的形式来实现，相反地，提供这些实施例的目的是使对本申请的公开内容的理解更加透彻全面。尽管参照前述实施例对本申请进行了详细的说明，对于本领域的技术人员来而言，其依然可以对前述各具体实施方式所记载的技术方案进行修改，或者对其中部分技术特征进行等效替换。凡是利用本申请说明书及附图内容所做的等效结构，直接或间接运用在其他相关的技术领域，均同理在本申请专利保护范围之内。

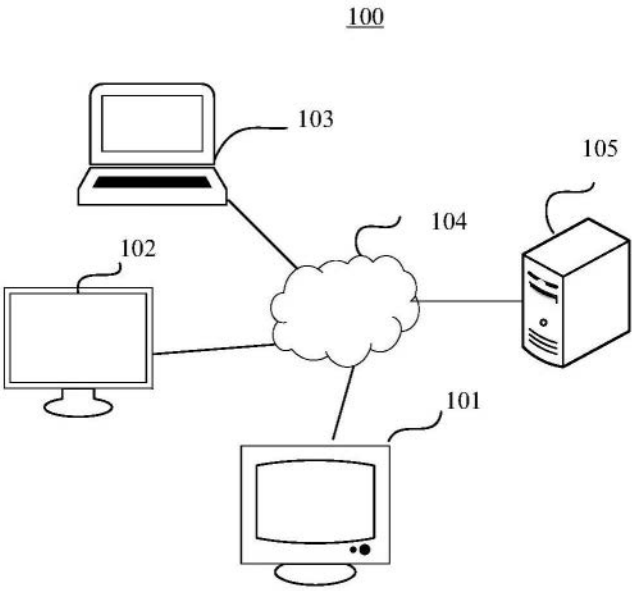

100
103
105
104
102
101

图1

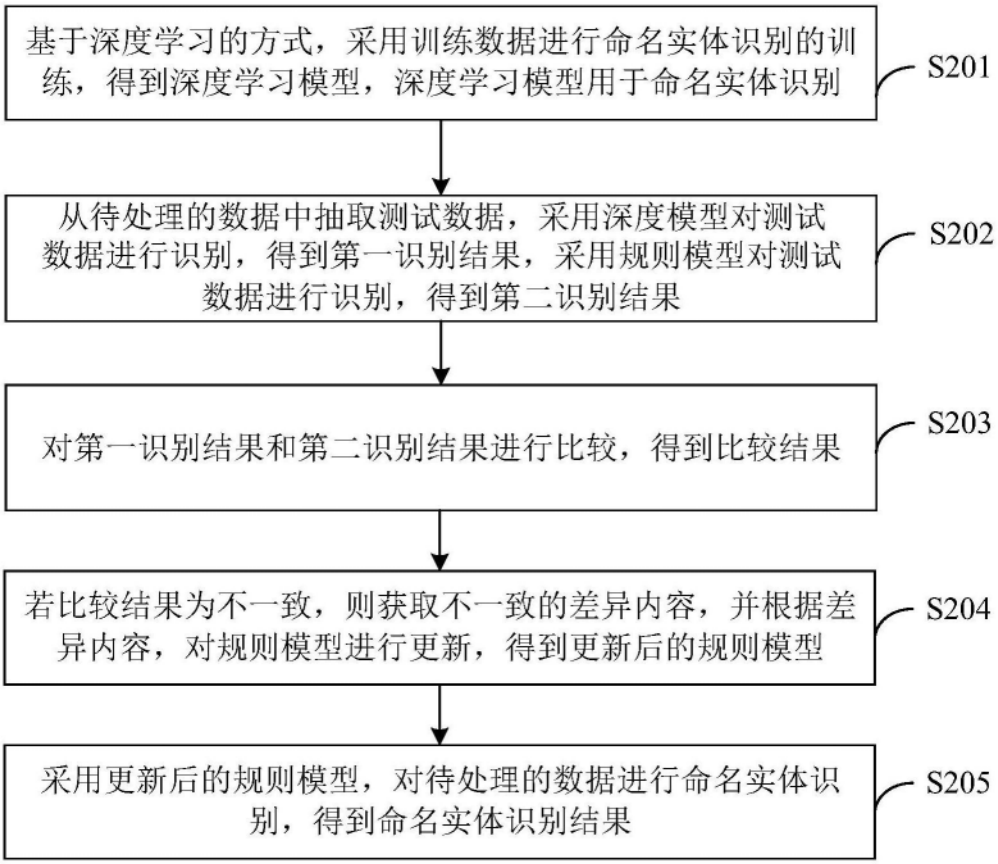

基于深度学习的方式，采用训练数据进行命名实体识别的训练，得到深度学习模型，深度学习模型用于命名实体识别
S201
从待处理的数据中抽取测试数据，采用深度模型对测试数据进行识别，得到第一识别结果，采用规则模型对测试数据进行识别，得到第二识别结果
S202
对第一识别结果和第二识别结果进行比较，得到比较结果
S203
若比较结果为不一致，则获取不一致的差异内容，并根据差异内容，对规则模型进行更新，得到更新后的规则模型
S204
采用更新后的规则模型，对待处理的数据进行命名实体识别，得到命名实体识别结果
S205

图2

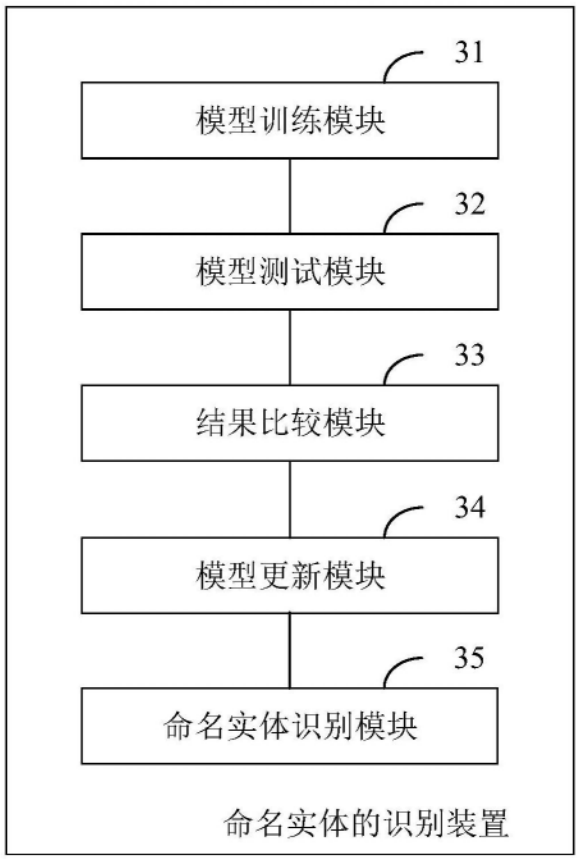

31
模型训练模块
32
模型测试模块
33
结果比较模块
34
模型更新模块
35
命名实体识别模块
命名实体的识别装置

图3

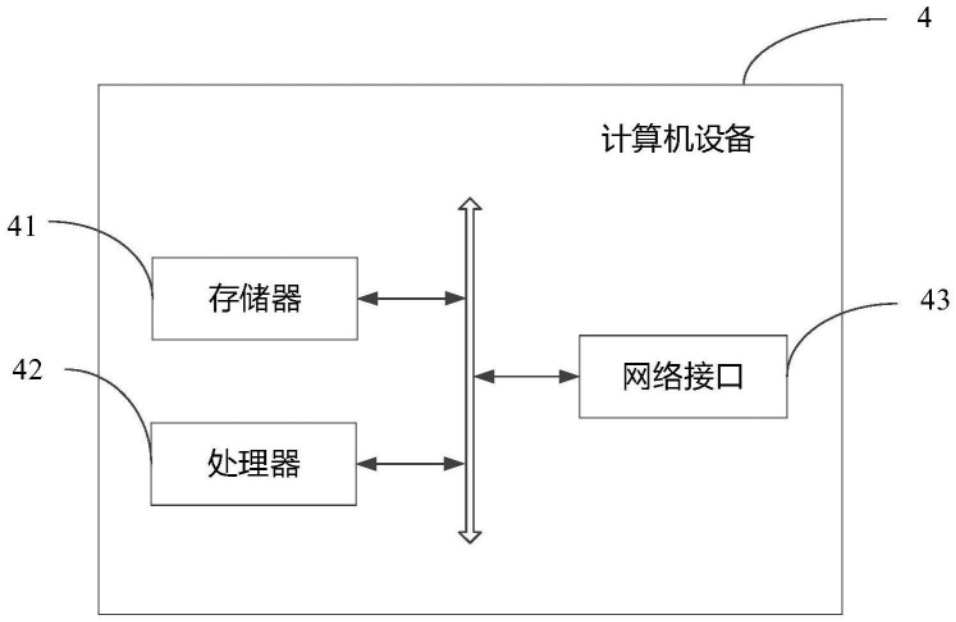

4
计算机设备
41
存储器
42
处理器
43
网络接口

图4