(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113627187 B
(45)授权公告日 2024.09.13

(21)申请号 202110926890.9

(22)申请日 2021.08.12

(65)同一申请的已公布的文献号
申请公布号 CN 113627187 A

(43)申请公布日 2021.11.09

(73)专利权人 平安国际智慧城市科技股份有限公司
地址 518000 广东省深圳市前海深港合作区妈湾兴海大道3048号前海自贸大厦1-34层

(72)发明人 冯豆豆

(74)专利代理机构 深圳市沃德知识产权代理事务所(普通合伙) 44347
专利代理师 高杰 于志光

(51)Int.Cl.
G06F 40/295(2020.01)

(56)对比文件
CN 108536679 A,2018.09.14
CN 112836046 A,2021.05.25

审查员 陈银兰

权利要求书2页 说明书11页 附图3页

(54)发明名称

命名实体识别方法、装置、电子设备及可读存储介质

(57)摘要

本发明涉及人工智能技术,揭露一种命名实体识别方法,包括:对待识别文本中的每个字符转换为字向量;对待识别文本进行分词处理,得到多个分词词语;将属性标签转化为标签向量;将字向量与标签向量拼接,得到拼接字向量,并将所有的拼接字向量进行组合,得到拼接字向量序列;对拼接字向量进行特征提取,得到特征向量;对特征向量进行识别,根据识别结果及拼接字向量序列对字符进行字识别,得到对应的字符属性;根据字符属性对待识别文本进行分词和属性标注,得到命名实体识别的结果。本发明还涉及区块链技术,所述待识别文本可以存储在区块链节点中。本发明还提出一种命名实体识别装置、设备以及介质。本发明可以提高命名实体识别的准确率。

CN 113627187 B

1.一种命名实体识别方法,其特征在于,所述方法包括:

接收待识别文本,对所述待识别文本中的每个字符进行向量转换,得到字向量;

利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语;

获取所述分词词语的属性标签,对所述属性标签进行向量转化,得到标签向量;

将每个所述字向量与所述标签向量进行向量拼接得到对应的拼接字向量,并将所有的所述拼接字向量进行组合得到拼接字向量序列,包括:选取所述字向量对应字符所属的分词词语得到字符分词词语,选取所述字符分词词语对应的所述标签向量得到目标标签向量,将所述字向量与对应的所述目标标签向量进行纵向拼接得到所述拼接字向量,及,将每个所述拼接字向量按照对应的字符在所述待识别文本中的先后顺序进行组合,得到所述拼接字向量序列;

对每个所述拼接字向量进行特征提取,得到对应的特征向量;

对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性,包括:利用预设的属性识别模型识别每个所述特征向量的属性得到向量属性,统计所述向量属性对应的拼接字向量在所述拼接字向量序列中的位置得到向量位置,及,将所述向量属性确定为所述待识别文本中与所述向量位置位置相同的字符的字符属性;

根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果。

2.如权利要求1所述的命名实体识别方法,其特征在于,所述利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语,包括:

利用预设的分词工具对所述待识别文本进行分词,得到初始分词结果;

根据所述标准词典对所述初始分词结果中的词语进行最长匹配,得到多个所述分词词语。

3.如权利要求1所述的命名实体识别方法,其特征在于,所述获取所述分词词语的属性标签,包括:

利用所述分词词语构建属性文本查询语句;

利用所述属性文本查询语句查询预设的词语属性表中所述分词词语对应的属性文本,得到属性标签。

4.如权利要求1所述的命名实体识别方法,其特征在于,所述对所述属性标签进行向量转化,得到标签向量,包括:

将所述属性标签中的每个字符转化为向量,得到标签字向量;

根据所有所述标签字向量进行计算,得到所述标签向量。

5.如权利要求1至4中任意一项所述的命名实体识别方法,其特征在于,所述根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果,包括:

遍历对比所述待识别文本中两个连续字符对应的所述字符属性是否相同;

将所述待识别文本中对应的所述字符属性不同的两个连续字符间添加预设的分词标记;

根据所述待识别文本中添加的所有分词标记对所述待识别文本进行分词,得到多个目标分词词语;

选取每个所述目标分词词语中任意字符对应的所述字符属性对对应的所述目标分词词语进行属性标记;

汇总所有属性标记的所述目标分词词语,得到所述命名实体识别结果。

6.一种命名实体识别装置,用于实现如权利要求1-5中任意一项所述的方法,其特征在于,该装置包括:

文本分词模块,用于接收待识别文本,对所述待识别文本中的每个字符进行向量转换,得到字向量;利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语;获取所述分词词语的属性标签,对所述属性标签进行向量转化,得到标签向量;将每个所述字向量与所述标签向量进行向量拼接,得到对应的拼接字向量,并将所有的所述拼接字向量进行组合,得到拼接字向量序列;对每个所述拼接字向量进行特征提取,得到对应的特征向量;

属性识别模块,用于对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性;

命名实体识别模块,用于根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果。

7.一种电子设备,其特征在于,所述电子设备包括:

至少一个处理器;以及,

与所述至少一个处理器通信连接的存储器;

其中,所述存储器存储有可被所述至少一个处理器执行的计算机程序,所述计算机程序被所述至少一个处理器执行,以使所述至少一个处理器能够执行如权利要求1至5中任一项所述的命名实体识别方法。

8.一种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5中任一项所述的命名实体识别方法。

命名实体识别方法、装置、电子设备及可读存储介质

技术领域

[0001] 本发明涉及人工智能技术,尤其涉及一种命名实体识别方法、装置、电子设备及可读存储介质。

背景技术

[0002] 随着互联网技术的发展,利用搜索引擎进行信息索引逐渐成为了人们生活中重要的一部分,但是搜索引擎进行信息索引时,为了保证搜索的准确率需要对待搜索文本进行命名实体识别,命名实体识别的准确度直接关乎搜索的准确度,因此,命名实体识别也越来越受到人们的重视。

[0003] 但是现有的命名实体识别大多利用分词工具直接对待识别文本进行分词,导致分词准确率较低,当分词准确率低时,属性标注的准确率更低,从而导致命名实体识别的准确率低。

发明内容

[0004] 本发明提供一种命名实体识别方法、装置、电子设备及计算机可读存储介质,其主要目的在于提高命名实体识别的准确率。

[0005] 为实现上述目的,本发明提供的一种命名实体识别方法,包括:

[0006] 接收待识别文本,对所述待识别文本中的每个字符进行向量转换,得到字向量;

[0007] 利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语;

[0008] 获取所述分词词语的属性标签,对所述属性标签进行向量转化,得到标签向量;

[0009] 将每个所述字向量与所述标签向量进行向量拼接,得到对应的拼接字向量,并将所有的所述拼接字向量进行组合,得到拼接字向量序列;

[0010] 对每个所述拼接字向量进行特征提取,得到对应的特征向量;

[0011] 对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性;

[0012] 根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果。

[0013] 可选地,所述利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语,包括:

[0014] 利用预设的分词工具对所述待识别文本进行分词,得到初始分词结果;

[0015] 根据所述标准词典对所述初始分词结果中的词语进行最长匹配,得到多个所述分词词语。

[0016] 可选地,所述获取所述分词词语的属性标签,包括:

[0017] 利用所述分词词语构建属性文本查询语句;

[0018] 利用所述属性文本查询语句查询预设的词语属性表中所述分词词语对应的属性文本,得到属性标签。

[0019] 可选地,所述对所述属性标签进行向量转化,得到标签向量,包括:

[0020] 将所述属性标签中的每个字符转化为向量,得到标签字向量;

[0021] 根据所有所述标签字向量进行计算,得到所述标签向量。

[0022] 可选地,所述将每个所述字向量与所述标签向量进行向量拼接,得到对应的拼接字向量,并将所有的所述拼接字向量进行组合,得到拼接字向量序列,包括:

[0023] 选取所述字向量对应字符所属的分词词语,得到字符分词词语;

[0024] 选取所述字符分词词语对应的所述标签向量,得到目标标签向量;

[0025] 将所述字向量与对应的所述目标标签向量进行纵向拼接,得到所述拼接字向量;

[0026] 将每个所述拼接字向量按照对应的字符在所述待识别文本中的先后顺序进行组合,得到所述拼接字向量序列。

[0027] 可选地,所述利用预设属性识别模型识别根据所述目标字向量序列中每个向量的属性,根据识别的向量的属性对所述待识别文本中的每个字符进行属性识别,得到对应的字符属性,包括:

[0028] 利用预设的属性识别模型识别每个所述特征向量的属性,得到向量属性;

[0029] 统计所述向量属性对应的拼接字向量在所述拼接字向量序列中的位置,得到向量位置;

[0030] 将所述向量属性确定为所述待识别文本中与所述向量位置位置相同的字符的字符属性。

[0031] 可选地,所述根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果,包括:

[0032] 遍历对比所述待识别文本中两个连续字符对应的所述字符属性是否相同;

[0033] 将所述待识别文本中对应的所述字符属性不同的两个连续字符间添加预设的分词标记;

[0034] 根据所述待识别文本中添加的所有分词标记对所述待识别文本进行分词,得到多个目标分词词语;

[0035] 选取每个所述目标分词词语中任意字符对应的所述字符属性对对应的所述目标分词词语进行属性标记;

[0036] 汇总所有属性标记的所述目标分词词语,得到所述命名实体识别结果。

[0037] 为了解决上述问题,本发明还提供一种命名实体识别装置,所述装置包括:

[0038] 文本分词模块,用于接收待识别文本,对所述待识别文本中的每个字符进行向量转换,得到字向量;利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语;获取所述分词词语的属性标签,对所述属性标签进行向量转化,得到标签向量;将每个所述字向量与所述标签向量进行向量拼接,得到对应的拼接字向量,并将所有的所述拼接字向量进行组合,得到拼接字向量序列;对每个所述拼接字向量进行特征提取,得到对应的特征向量。

[0039] 属性识别模块,用于对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性;

[0040] 命名实体识别模块,用于根据所述字符属性对所述待识别文本进行分词和属性标

注,得到命名实体识别的结果。

[0041] 为了解决上述问题,本发明还提供一种电子设备,所述电子设备包括:

[0042] 存储器,存储至少一个计算机程序;及

[0043] 处理器,执行所述存储器中存储的计算机程序以实现上述所述的命名实体识别方法。

[0044] 为了解决上述问题,本发明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中存储有至少一个计算机程序,所述至少一个计算机程序被电子设备中的处理器执行以实现上述所述的命名实体识别方法。

[0045] 本发明实施例对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性,通过属性的二次识别,提高属性识别的准确率;根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果,利用属性识别结果进行二次分词提高分词的准确率,同时利用属性识别结果对二次分词后的词语进行属性标记,提高了命名实体识别的准确率;因此本发明实施例提出的命名实体识别方法、装置、电子设备及可读存储介质提高了命名实体识别的准确率。

附图说明

[0046] 图1为本发明一实施例提供的命名实体识别方法的流程示意图;

[0047] 图2为本发明一实施例提供的命名实体识别方法中得到拼接字向量序列的流程示意图;

[0048] 图3为本发明一实施例提供的命名实体识别装置的模块示意图;

[0049] 图4为本发明一实施例提供的实现命名实体识别方法的电子设备的内部结构示意图;

[0050] 本发明目的的实现、功能特点及优点将结合实施例,参照附图做进一步说明。

具体实施方式

[0051] 应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。

[0052] 本发明实施例提供一种命名实体识别方法。所述命名实体识别方法的执行主体包括但不限于服务端、终端等能够被配置为执行本申请实施例提供的该方法的电子设备中的至少一种。换言之,所述命名实体识别方法可以由安装在终端设备或服务端设备的软件或硬件来执行,所述软件可以是区块链平台。所述服务端包括但不限于:单台服务器、服务器集群、云端服务器或云端服务器集群等,服务器可以是独立的服务器,也可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network,CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0053] 参照图1所示的本发明一实施例提供的命名实体识别方法的流程示意图,在本发明实施例中,所述命名实体识别方法包括:

[0054] S1、接收待识别文本,对所述待识别文本中的每个字符进行向量转换,得到字向量;

[0055] 详细地,本发明实施例中所述待识别文本为用户输入的需要进行检索的文本。例如:所述待识别文本为“A公司投资的污水治理项目有哪些”。

[0056] 进一步地,本发明实施例中为了对所述待识别文本中进行的分词,需要所述待识别文本中的每个字符的特征进行量化,因此,本发明实施例将所述待识别文本中的每个字符的特征转换为向量,得到所述字向量。

[0057] 详细地,本发明实施例中将所述待识别文本输入至预设的自然语言模型中,利用预设的自然语言模型将所述待识别文本中的每个字符转化为向量,得到所述字向量。

[0058] 可选地,本发明实施例中所述预设的自然语言模型包括但不限于包括但不限于BERT、RoBerta、ALBert、PLM等。

[0059] 本发明另一实施例中所述待识别文本可以存储在区块链节点中,利用区块链节点高吞吐的特性,提高数据的存取效率。

[0060] S2、利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语;

[0061] 详细地,本发明实施例中,对所述待识别文本进行分词处理,得到多个所述分词词语,包括:

[0062] 利用预设的分词工具对所述待识别文本进行分词,得到初始分词结果;

[0063] 可选地,本发明实施例中所述分词工具为jieba分词工具,所述初始分词结果中包含多个分词后的词语。

[0064] 进一步地,由于使用jieba分词等分词工具进行分词后的初始分词结果存在不准确的情况,因此,为了提高初始初始分词结果的准确率,根据所述标准词典对所述初始分词结果进行最长匹配,得到多个分词词语;

[0065] 具体地,本发明实施例中根据所述标准词典对所述初始分词结果中的词语进行最长匹配,得到多个分词词语,包括:

[0066] 步骤I:将所述初始分词结果中每个词语及其之后的所有词语组合得到对应的匹配文本;

[0067] 例如:初始分词结果为“污水/治理/项目”,第一个词语“污水”对应的匹配文本为“污水治理项目”,第二个词语“治理”对应的匹配文本为“治理项目”,第三个词语“项目”对应的匹配文本为“项目”

[0068] 步骤II:将所述匹配文本所述标准词典中进行检索;

[0069] 步骤III:若所述标准词典中检索到与所述匹配文本相同的词语时,确定所述匹配文本为第一分词词语;

[0070] 步骤IV:若所述标准词典中检索不到与所述匹配文本相同的词语时,判断所述匹配文本中词语的个数,若所述匹配文本中词语的数量大于1,将所述匹配文本从右侧第一个词语切除,得到更新后的匹配文本,并返回所述将所述匹配文本所述标准词典中进行检索步骤,若所述匹配文本中词语的数量等于1,将所述匹配文本确定为第一分词词语;

[0071] 例如:匹配文本中右侧第一个词语为对应的所述初始分词结果中的词语,如初始分词结果为“污水/治理/项目”,匹配文本“污水治理项目”右侧第一个词语为“项目”。

[0072] 步骤V:汇总所述初始分词结果中每个词语对应的所述第一分词词语,得到第一分词词语集;

[0073] 步骤VI:依次判断所述第一分词词语集中每个第一分词词语是否与其他第一分词

词语有词语重叠,若有词语重叠,则删除两个词语重叠的第一分词词语中字符较少的第一分词词语,得到更新后的第一分词词语集,并返回所述依次判断所述第一分词词语集中每个第一分词词语是否与其他第一分词词语有词语重叠步骤;若无词语重叠,则将所述第一分词词语集中的每个第一分词词语确定为分词词语,得到多个所述分词词语。

[0074] S3、获取所述分词词语的属性标签,对所述属性标签进行向量转化,得到标签向量;

[0075] 详细地,本发明实施例中因为词典只能对词典中存在的的词进行属性标注,无法标注不在词典中的词。为了更全面的对所述待识别文本中的每个字符进行属性识别,先给出每个分词词语的属性,作为后续模型的输入。

[0076] 详细地,本发明实施例中对每个所述分词词语进行属性标注,得到每个所述分词词语对应的属性标签,包括:

[0077] 利用所述分词词语构建属性文本查询语句;

[0078] 利用所述属性文本查询语句查询预设的词语属性表的属性文本,得到属性标签;

[0079] 可选地,本发明实施例中所述词语属性表包含不同的词语及其对应的属性文本的数据表,其中所述属性文本包括:无意义词语、日常生活使用的地名、时间、业务属性等属性对应的文本,所述业务属性可以为社会资本名称、所述行业等。

[0080] 可选地,本发明实施例还可以将所述属性标签标记对应的所述分词词语。

[0081] 进一步地,本发明实施中将所述属性标签进行向量转化得到,所述标签向量。

[0082] 可选地,本发明实施例中对所述属性标签进行向量转化,得到标签向量。

[0083] 具体地,将所述属性标签中的每个字符转化为向量,得到标签字向量;根据所有所述标签字向量进行计算,得到所述标签向量。

[0084] 可选地,本发明实施例中利用预设的word2vec模型将所述属性标签中的每个字符转化为向量,得到标签字向量;将所有所述标签字向量进行算术平均计算,得到所述标签向量。

[0085] 例如:共有标签字向量$\begin{bmatrix}1\\2\end{bmatrix}$和$\begin{bmatrix}4\\3\end{bmatrix}$,那么(1+4)/2=2.5、(2+2)/2=2,那么所述标签向量为$\begin{bmatrix}2.5\\2\end{bmatrix}$。

[0086] S4、将每个所述字向量与所述标签向量进行向量拼接,得到对应的拼接字向量,并将所有的所述拼接字向量进行组合,得到拼接字向量序列;

[0087] 详细地,参阅图2所示,本发明实施例中将每个所述字向量与所述标签向量进行向量拼接,得到对应的拼接字向量,并将所有的所述拼接字向量进行组合,得到拼接字向量序列,包括:

[0088] S41、选取所述字向量对应字符所属的分词词语,得到字符分词词语;

[0089] S42、选取所述字符分词词语对应的所述标签向量,得到目标标签向量;

[0090] S43、将所述字向量与对应的所述目标标签向量进行纵向拼接,得到所述拼接字向量;

[0091] 例如:所述字向量为$\begin{bmatrix}1\\2\end{bmatrix}$,所述字向量对应的所述目标标签向量为$\begin{bmatrix}4\\3\end{bmatrix}$,将所述字向

量与对应的所述目标标签向量进行纵向拼接得到所述拼接字向量为$\begin{bmatrix}1\\2\\4\\3\end{bmatrix}$。

[0092] S44、将每个所述拼接字向量按照对应的字符在所述待识别文本中的先后顺序进行组合,得到所述拼接字向量序列。

[0093] 例如:所述待识别文本为“污水处理公司”,字符“污”对应的拼接字向量为$\begin{bmatrix}0\\1\\2\end{bmatrix}$,字符“水”对应的拼接字向量为$\begin{bmatrix}0\\2\\3\end{bmatrix}$,字符“处”对应的拼接字向量为$\begin{bmatrix}1\\1\\2\end{bmatrix}$,字符“理”对应的拼接字向量为$\begin{bmatrix}3\\1\\2\end{bmatrix}$,字符“公”对应的拼接字向量为$\begin{bmatrix}5\\1\\2\end{bmatrix}$,字符“司”对应的拼接字向量为$\begin{bmatrix}6\\8\\2\end{bmatrix}$,那么对应拼接字向量序列为$\begin{bmatrix}0&0&1&3&6\\1&2&1&1&8\\2&3&2&2&2\end{bmatrix}$。

[0094] S5、对每个所述拼接字向量进行特征提取,得到对应的特征向量;

[0095] 可选地,本发明实施例中将所述拼接字向量序列输入预设的深度学习模型,利用所述深度学习模型对所述拼接字向量序列中的每个拼接字向量进行特征提取,得到对应的特征向量。

[0096] 可选地,本发明实施例中所述预设的深度学习模型为bilstm模型。

[0097] S6、对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性;

[0098] 详细地,本发明实施例中利用预设属性识别模型对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性,包括:

[0099] 利用预设的属性识别模型识别每个所述特征向量的属性,得到向量属性;

[0100] 可选地,本发明实施例中所述属性识别模型为人工智能模型,可以为训练完成的CRF模型。

[0101] 统计所述向量属性对应的拼接字向量在所述拼接字向量序列中的位置,得到向量位置;

[0102] 将所述向量属性确定为所述待识别文本中与所述向量位置位置相同的字符的字符属性。

[0103] 进一步地,本发明实施例中利用预设的属性识别模型识别所述目标向量序列中每个向量的属性之前,所述方法还包括:

[0104] 步骤A:获取历史文本集,其中,所述历史文本集中每个历史文本的每个字符都有对应的属性标签;

[0105] 可选地,本发明实施例中所述历史文本集为多个历史文本的集合,所述历史文本为与所述待识别文本类型相同内容不同的自然语言文本,所述历史文本中每个字符都标注有对应的字符属性标签,其中,所述字符属性标签为字符对应的属性文本。

[0106] 步骤B:将所述历史文本中的每个字符进行向量转换,得到历史字向量;

[0107] 步骤C:利用预构建的标准词典对所述历史文本进行分词处理,得到多个历史分词词语;

[0108] 步骤D:获取所述历史分词词语的属性标签,得到历史属性标签;

[0109] 步骤E:对所述历史属性标签进行向量转化,得到历史标签向量;

[0110] 步骤F:将每个所述历史字向量及对应的历史标签向量进行向量拼接,得到对应的历史拼接字向量,并将所有的所述历史拼接字向量进行组合,得到历史拼接字向量序列;

[0111] 步骤G:对所述历史拼接字向量序列中每个历史拼接字向量进行特征提取,得到对应的历史特征向量;

[0112] 步骤H:利用预构建的初始识别模型识别所述目标历史字向量序列中的每个向量进行预测得到标签预测值;

[0113] 可选地,本发明实施例中所述初始识别模型为CRF模型。

[0114] 步骤I:根据所述目标历史字向量向量中每个向量的位置选取所述历史文本中对应位置的字符的字符属性标签,得到真实标签,根据所述真实标签确定标签真实值;

[0115] 例如:所述字符属性标签为公司机构,那么公司机构属性对应的标签真实值为1。

[0116] 步骤J:根据所述标签预测值与所述标签真实值,利用预设损失函数计算进行计算,得到目标损失值,当所述目标损失值大于或等于预设阈值时,更新所述初始识别模型的模型参数,并返回所述步骤B;当所述目标损失值小于预设阈值时,停止训练,得到所述属性识别模型。

[0117] S7、根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果。

[0118] 详细地,本发明实施例中根据所述字符属性对所述待识别文本进行分词处理和属性标注,得到命名实体识别的结果,包括:遍历对比所述待识别文本中两个连续字符对应的所述字符属性是否相同;将所述待识别文本中对应的所述字符属性不同的两个连续字符间添加预设的分词标记,可选地,本发明实施例中所述分词标记为“/”;根据所述待识别文本中添加的所有分词标记对所述待识别文本进行分词,得到多个目标分词词语,例如:添加了所有分词标记的待识别文本为“污水/处理/公司”,那么对应的目标分词词语为“污水”、“处理”、“公司”;选取每个所述目标分词词语中任意字符对应的所述字符属性对对应的所述目标分词词语进行属性标记,例如:目标分词词语“公司”中任意字符对应的字符属性为“企业”,那么将目标分词词语“公司”的属性标记为“企业”;汇总所有属性标记的所述目标分词词语,得到所述命名实体识别结果。

[0119] 本发明通过属性进行二次分词,分词结果更准确。业界常用基于词典的方式进行划分,分词及属性标注的灵活性较差,准确率较低。本发明在基于词典的基础上增加了属性向量,与字向量拼接后输入到bilstm+crf模型中,兼顾了语义信息、属性信息和上下文信息,大大提高命名实体识别的准确率。

[0120] 如图3所示,是本发明命名实体识别装置的功能模块图。

[0121] 本发明所述命名实体识别装置100可以安装于电子设备中。根据实现的功能,所述命名实体识别装置可以包括特征提取模块101、项目筛选模块102、命名实体识别模块103,本发所述模块也可以称之为单元,是指一种能够被电子设备处理器所执行,并且能够完成固定功能的一系列计算机程序段,其存储在电子设备的存储器中。

[0122] 在本实施例中,关于各模块/单元的功能如下:

[0123] 所述文本分词模块101用于接收待识别文本,对所述待识别文本中的每个字符进行向量转换,得到字向量;利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语;获取所述分词词语的属性标签,对所述属性标签进行向量转化,得到标签向量;将每个所述字向量与所述标签向量进行向量拼接,得到对应的拼接字向量,并将所有的所述拼接字向量进行组合,得到拼接字向量序列;对每个所述拼接字向量进行特征提取,得到对应的特征向量;

[0124] 所述属性识别模块102用于对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性;

[0125] 所述命名实体识别模块103用于根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果。

[0126] 详细地,本发明实施例中所述命名实体识别装置100中所述的各模块在使用时采用与上述图1中所述的命名实体识别方法一样的技术手段,并能够产生相同的技术效果,这里不再赘述。

[0127] 如图4所示,是本发明实现命名实体识别方法的电子设备的结构示意图。

[0128] 所述电子设备可以包括处理器10、存储器11、通信总线12和通信接口13,还可以包括存储在所述存储器11中并可在所述处理器10上运行的计算机程序,如命名实体识别程序。

[0129] 其中,所述存储器11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、移动硬盘、多媒体卡、卡型存储器(例如:SD或DX存储器等)、磁性存储器、磁盘、光盘等。所述存储器11在一些实施例中可以是电子设备的内部存储单元,例如该电子设备的移动硬盘。所述存储器11在另一些实施例中也可以是电子设备的外部存储设备,例如电子设备上配备的插接式移动硬盘、智能存储卡(Smart Media Card,SMC)、安全数字(Secure Digital,SD)卡、闪存卡(Flash Card)等。进一步地,所述存储器11还可以既包括电子设备的内部存储单元也包括外部存储设备。所述存储器11不仅可以用于存储安装于电子设备的应用软件及各类数据,例如命名实体识别程序的代码等,还可以用于暂时地存储已经输出或者将要输出的数据。

[0130] 所述处理器10在一些实施例中可以由集成电路组成,例如可以由单个封装的集成电路所组成,也可以是由多个相同功能或不同功能封装的集成电路所组成,包括一个或者多个中央处理器(Central Processing unit,CPU)、微处理器、数字处理芯片、图形处理器及各种控制芯片的组合等。所述处理器10是所述电子设备的控制核心(Control Unit),利用各种接口和线路连接整个电子设备的各个部件,通过运行或执行存储在所述存储器11内的程序或者模块(例如命名实体识别程序等),以及调用存储在所述存储器11内的数据,以执行电子设备的各种功能和处理数据。

[0131] 所述通信总线12可以是外设部件互连标准(perIPheral component interconnect,简称PCI)总线或扩展工业标准结构(extended industry standard architecture,简称EISA)总线等。该总线可以分为地址总线、数据总线、控制总线等。所述通信总线12总线被设置为实现所述存储器11以及至少一个处理器10等之间的连接通信。为便于表示,图中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

[0132] 图4仅示出了具有部件的电子设备,本领域技术人员可以理解的是,图4示出的结构并不构成对所述电子设备的限定,可以包括比图示更少或者更多的部件,或者组合某些部件,或者不同的部件布置。

[0133] 例如,尽管未示出,所述电子设备还可以包括给各个部件供电的电源(比如电池),优选地,电源可以通过电源管理装置与所述至少一个处理器10逻辑相连,从而通过电源管理装置实现充电管理、放电管理、以及功耗管理等功能。电源还可以包括一个或一个以上的直流或交流电源、再充电装置、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。所述电子设备还可以包括多种传感器、蓝牙模块、Wi-Fi模块等,在此不再赘述。

[0134] 可选地,所述通信接口13可以包括有线接口和/或无线接口(如WI-FI接口、蓝牙接口等),通常用于在该电子设备与其他电子设备之间建立通信连接。

[0135] 可选地,所述通信接口13还可以包括用户接口,用户接口可以是显示器(Display)、输入单元(比如键盘(Keyboard)),可选地,用户接口还可以是标准的有线接口、无线接口。可选地,在一些实施例中,显示器可以是LED显示器、液晶显示器、触控式液晶显示器以及OLED(Organic Light-Emitting Diode,有机发光二极管)触摸器等。其中,显示器也可以适当的称为显示屏或显示单元,用于显示在电子设备中处理的信息以及用于显示可视化的用户界面。

[0136] 应该了解,所述实施例仅为说明之用,在专利申请范围上并不受此结构的限制。

[0137] 所述电子设备中的所述存储器11存储的命名实体识别程序是多个计算机程序的组合,在所述处理器10中运行时,可以实现:

[0138] 接收待识别文本,对所述待识别文本中的每个字符进行向量转换,得到字向量;

[0139] 利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语;

[0140] 获取所述分词词语的属性标签,对所述属性标签进行向量转化,得到标签向量;

[0141] 将每个所述字向量与所述标签向量进行向量拼接,得到对应的拼接字向量,并将所有的所述拼接字向量进行组合,得到拼接字向量序列;

[0142] 对每个所述拼接字向量进行特征提取,得到对应的特征向量;

[0143] 对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性;

[0144] 根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果。

[0145] 具体地,所述处理器10对上述计算机程序的具体实现方法可参考图1对应实施例中相关步骤的描述,在此不赘述。

[0146] 进一步地,所述电子设备集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。所述计算机可读

介质可以是非易失性的,也可以是易失性的。所述计算机可读介质可以包括:能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM,Read-Only Memory)。

[0147] 本发明实施例还可以提供一种计算机可读存储介质,所述可读存储介质存储有计算机程序,所述计算机程序在被电子设备的处理器所执行时,可以实现:

[0148] 接收待识别文本,对所述待识别文本中的每个字符进行向量转换,得到字向量;

[0149] 利用预构建的标准词典对所述待识别文本进行分词处理,得到多个分词词语;

[0150] 获取所述分词词语的属性标签,对所述属性标签进行向量转化,得到标签向量;

[0151] 将每个所述字向量与所述标签向量进行向量拼接,得到对应的拼接字向量,并将所有的所述拼接字向量进行组合,得到拼接字向量序列;

[0152] 对每个所述拼接字向量进行特征提取,得到对应的特征向量;

[0153] 对每个所述特征向量进行向量属性识别,根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别,得到对应的字符属性;

[0154] 根据所述字符属性对所述待识别文本进行分词和属性标注,得到命名实体识别的结果。

[0155] 进一步地,所述计算机可用存储介质可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序等;存储数据区可存储根据区块链节点的使用所创建的数据等。

[0156] 在本发明所提供的几个实施例中,应该理解到,所揭露的设备,装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述模块的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式。

[0157] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的,作为模块显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0158] 另外,在本发明各个实施例中的各功能模块可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用硬件加软件功能模块的形式实现。

[0159] 对于本领域技术人员而言,显然本发明不限于上述示范性实施例的细节,而且在不背离本发明的精神或基本特征的情况下,能够以其他的具体形式实现本发明。

[0160] 因此,无论从哪一点来看,均应将实施例看作是示范性的,而且是非限制性的,本发明的范围由所附权利要求而不是上述说明限定,因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。

[0161] 本申请实施例可以基于人工智能技术对相关的数据进行获取和处理。其中,人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

[0162] 人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、

大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、机器人技术、生物识别技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0163] 本发明所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain),本质上是一个去中心化的数据库,是一串使用密码学方法相关联产生的数据块,每一个数据块中包含了一批次网络交易的信息,用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

[0164] 此外,显然“包括”一词不排除其他单元或步骤,单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第二等词语用来表示名称,而并不表示任何特定的顺序。

[0165] 最后应说明的是,以上实施例仅用以说明本发明的技术方案而非限制,尽管参照较佳实施例对本发明进行了详细说明,本领域的普通技术人员应当理解,可以对本发明的技术方案进行修改或等同替换,而不脱离本发明技术方案的精神和范围。

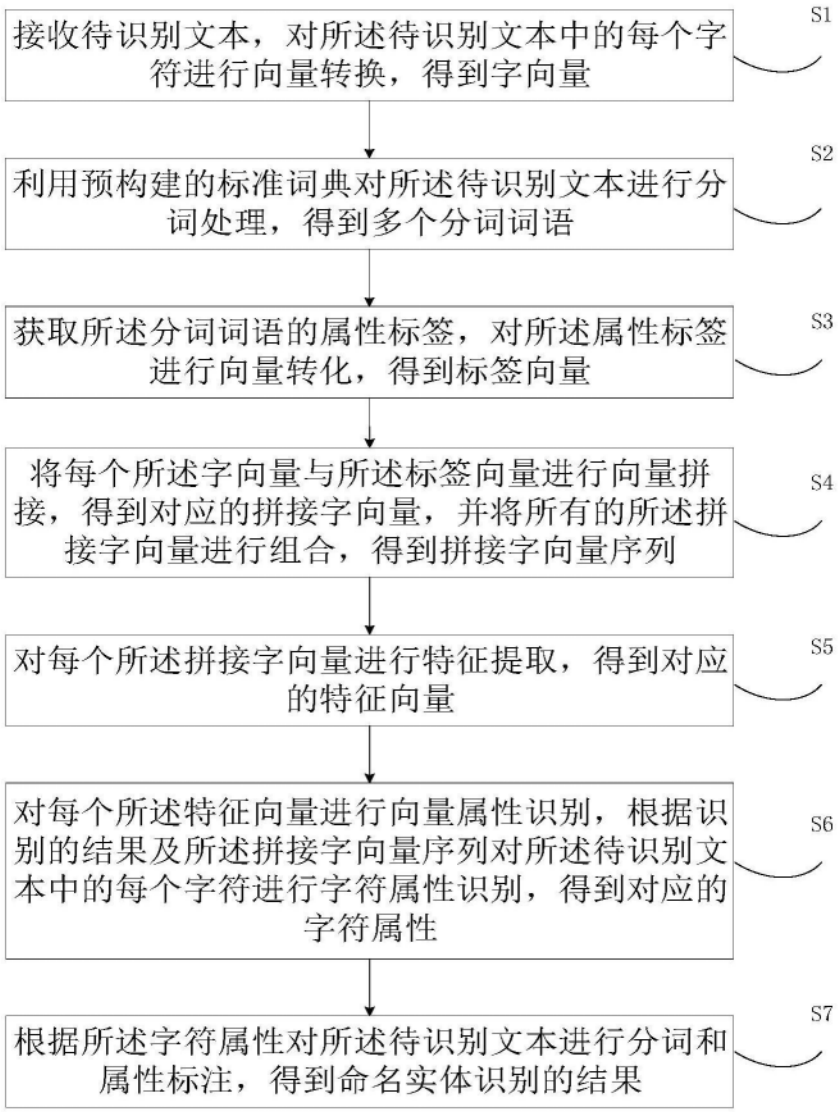
S1
接收待识别文本，对所述待识别文本中的每个字符进行向量转换，得到字向量
S2
利用预构建的标准词典对所述待识别文本进行分词处理，得到多个分词词语
S3
获取所述分词词语的属性标签，对所述属性标签进行向量转化，得到标签向量
S4
将每个所述字向量与所述标签向量进行向量拼接，得到对应的拼接字向量，并将所有的所述拼接字向量进行组合，得到拼接字向量序列
S5
对每个所述拼接字向量进行特征提取，得到对应的特征向量
S6
对每个所述特征向量进行向量属性识别，根据识别的结果及所述拼接字向量序列对所述待识别文本中的每个字符进行字符属性识别，得到对应的字符属性
S7
根据所述字符属性对所述待识别文本进行分词和属性标注，得到命名实体识别的结果

图1

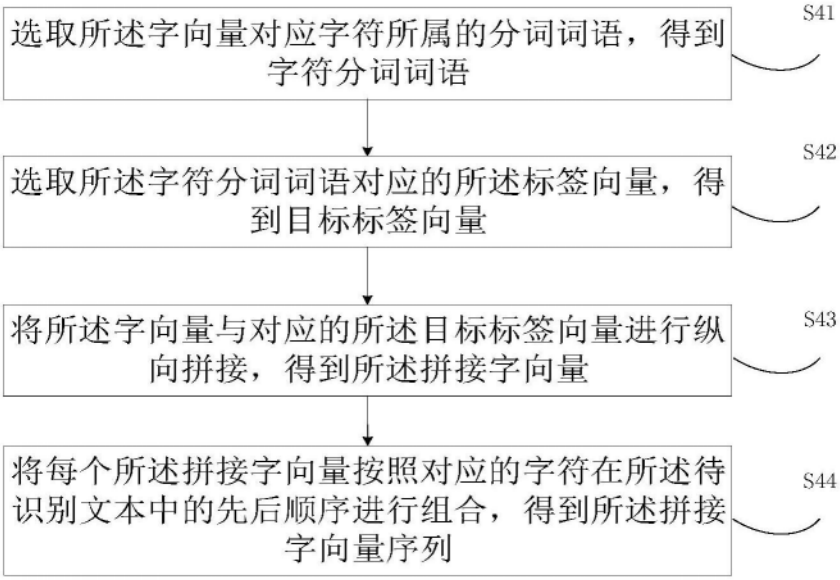
选取所述字向量对应字符所属的分词词语，得到字符分词词语
S41
选取所述字符分词词语对应的所述标签向量，得到目标标签向量
S42
将所述字向量与对应的所述目标标签向量进行纵向拼接，得到所述拼接字向量
S43
将每个所述拼接字向量按照对应的字符在所述待识别文本中的先后顺序进行组合，得到所述拼接字向量序列
S44

图2

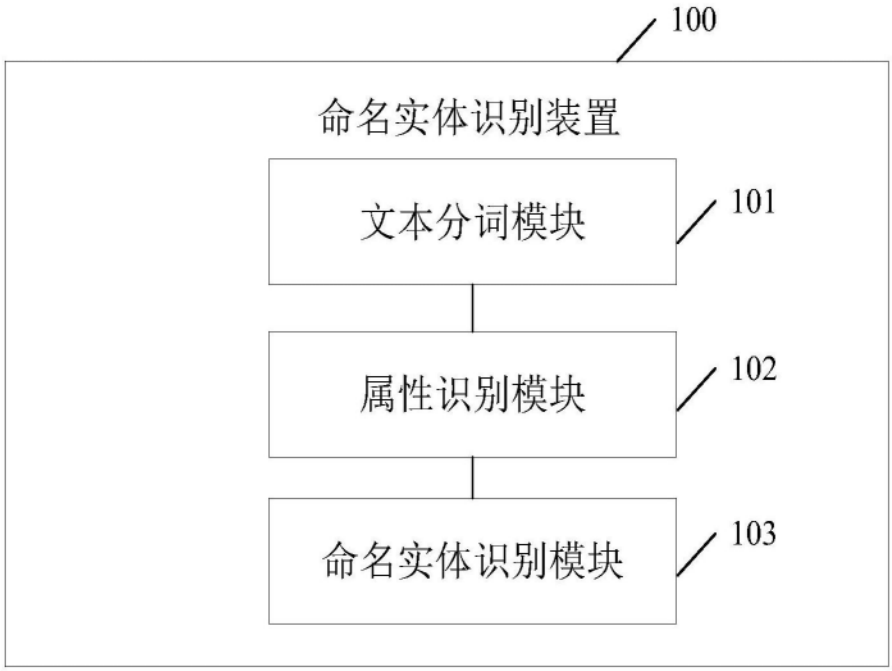
100
命名实体识别装置
文本分词模块
101
属性识别模块
102
命名实体识别模块
103

图3

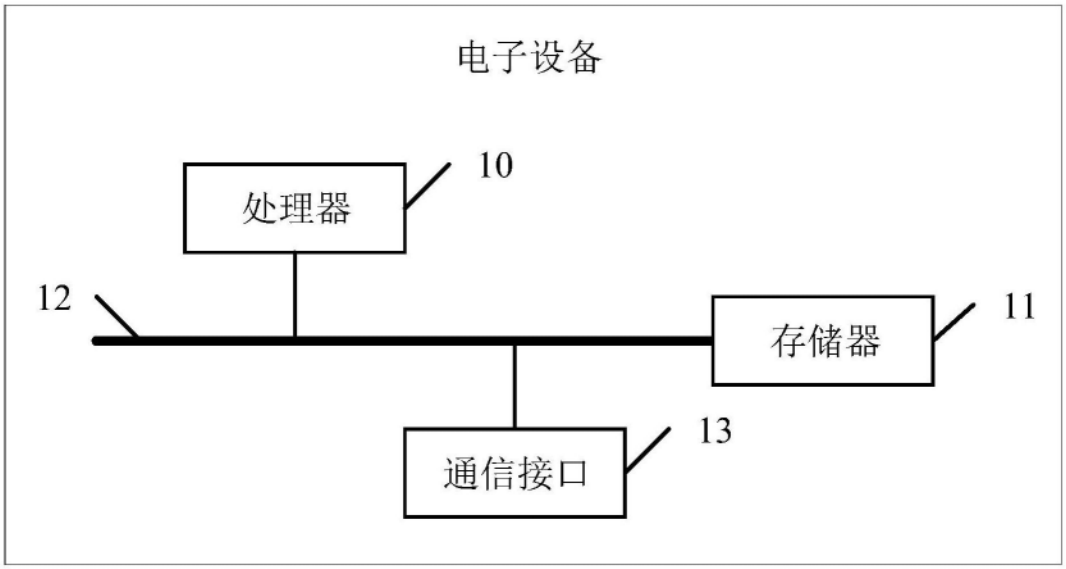

图4