(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109492215 A
(43)申请公布日 2019.03.19

(21)申请号 201811090023.0

(22)申请日 2018.09.18

(71)申请人 平安科技(深圳)有限公司
地址 518033 广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼

(72)发明人 黄萍 汪伟 肖京

(74)专利代理机构 广州华进联合专利商标代理有限公司 44224
代理人 王宁

(51)Int.Cl.
G06F 17/27(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

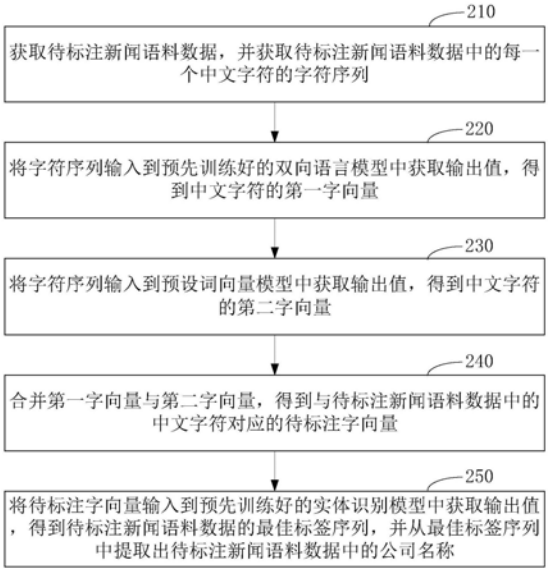
权利要求书2页 说明书13页 附图3页

(54)发明名称

新闻实体识别方法、装置、计算机设备和存储介质

(57)摘要

本申请涉及一种基于自然语言处理的新闻实体识别方法、装置、计算机设备和存储介质。该方法包括:获取待标注新闻语料数据,并获取待标注新闻语料数据中的每一个中文字符的字符序列;将字符序列输入到预先训练好的双向语言模型中获取输出值,得到中文字符的第一字向量;将字符序列输入到预设词向量模型中获取输出值,得到中文字符的第二字向量;合并第一字向量与第二字向量,得到中文字符对应的待标注字向量;将待标注字向量输入到预先训练好的实体识别模型,得到待标注新闻语料数据的最佳标签序列,并从最佳标签序列中提取出待标注新闻语料数据中的公司名称。采用本方法能够使得有歧义的公司名称的识别结果减少,提高公司名称识别结果的准确度。



CN 109492215 A

1.一种新闻实体识别方法,所述方法包括:

获取待标注新闻语料数据,并获取所述待标注新闻语料数据中的每一个中文字符的字符序列;

将所述字符序列输入到预先训练好的双向语言模型中获取输出值,得到所述中文字符的第一字向量;

将所述字符序列输入到预设词向量模型中获取输出值,得到所述中文字符的第二字向量;

合并所述第一字向量与所述第二字向量,得到与所述待标注新闻语料数据中的中文字符对应的待标注字向量;

将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值,得到所述待标注新闻语料数据的最佳标签序列,并从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称。

2.根据权利要求1所述的方法,其特征在于,所述实体识别模型包括双向循环神经网络以及条件随机场模型;

所述将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值,得到所述待标注新闻语料数据的最佳标签序列的步骤,包括:

将所述待标注字向量输入到双向循环神经网络中,获取所述待标注新闻语料数据中的中文字符对应的特征向量;

将所述特征向量输入至条件随机场模型中,获取所述待标注新闻语料数据的最佳标签序列。

3.根据权利要求1所述的方法,其特征在于,所述将所述字符序列输入到预先训练好的双向语言模型中获取输出值的步骤之前,还包括:

建立一个双向递归神经网络模型,其中所述双向神经网络模型包括前向递归神经网络隐含层以及后向递归神经网络隐含层;

获取无标注语料样本,将所述无标注语料样本中的中文字符转化为对应的字符序列;

将所述无标注语料样本的字符序列输入至所述双向的递归神经网络模型中,利用所述无标注语料样本的字符序列对所述前向递归神经网络隐含层以及所述后向递归神经网络隐含层进行无监督训练,得到所述双向语言模型。

4.根据权利要求3所述的方法,其特征在于,所述将所述字符序列输入到预先训练好的双向语言模型中获取输出值,得到所述中文字符的第一字向量的步骤,包括:

将所述待标注新闻语料数据的字符序列输入至所述前向递归神经网络隐含层中,获取前向隐状态序列;

将所述待标注新闻语料数据的字符序列输入至所述后向递归神经网络隐含层中,获取后向隐状态序列;

合并所述前向隐状态序列以及所述后向隐状态序列生成所述中文字符的第一字向量。

5.根据权利要求3所述的方法,其特征在于,所述得到所述双向语言模型的步骤之后,包括:

获取新闻语料训练样本,并获取所述新闻语料训练样本中的每一个中文字符的字符序列,其中,所述新闻语料训练样本中的中文字符具有对应的字符标签;

将所述新闻语料训练样本中的字符序列分别输入至所述双向语言模型以及所述预设词向量模型中获取输出值，得到第三字向量以及第四字向量；

合并所述第三字向量与所述第四字向量，得到与所述新闻语料训练样本中的中文字符对应的字向量样本；

根据所述新闻语料训练样本中的中文字符对应的字向量样本以及字符标签，对所述实体识别模型进行有监督训练。

6.根据权利要求1所述的方法，其特征在于，所述条件随机场模型中采用的标注规则为BIOES模式；

所述从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称的步骤，包括：

从最佳标签序列中提取连续标注的B标签、I标签以及E标签对应的中文字符，将提取得到的中文字符组合作为识别出的企业名称。

7.一种新闻实体识别装置，其特征在于，所述装置包括：

新闻语料获取模块，用于获取待标注新闻语料数据，并获取所述待标注新闻语料数据中的每一个中文字符的字符序列；

第一字向量获取模块，用于将所述字符序列输入到预先训练好的双向语言模型中获取输出值，得到所述中文字符的第一字向量；

第二字向量获取模块，用于将所述字符序列输入到预设词向量模型中获取输出值，得到所述中文字符的第二字向量；

待标注字向量获取模块，用于合并所述第一字向量与所述第二字向量，得到与所述待标注新闻语料数据中的中文字符对应的待标注字向量；

公司名称获取模块，用于将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到所述待标注新闻语料数据的最佳标签序列，并从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称。

8.根据权利要求7所述的装置，其特征在于，所述第一字向量获取模块用于将所述待标注新闻语料数据的字符序列输入至所述前向递归神经网络隐含层中，获取前向隐状态序列；将所述待标注新闻语料数据的字符序列输入至所述后向递归神经网络隐含层中，获取后向隐状态序列；合并所述前向隐状态序列以及所述后向隐状态序列生成所述中文字符的第一字向量。

9.一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求1至6中任一项所述新闻实体识别方法的步骤。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1至6中任一项所述新闻实体识别方法的步骤。

新闻实体识别方法、装置、计算机设备和存储介质

技术领域

[0001] 本申请涉及自然语言处理技术领域，特别是涉及一种新闻实体识别方法、装置、计算机设备和存储介质。

背景技术

[0002] 目前，对新闻语料中公司名称的识别的主导技术是监督式学习，例如传统的隐马尔科夫模型(HMMS)、条件随机场算法(Conditional Random Field algorithm，CRF)等机器学习方法，以及近几年被提出来的循环神经网络(Recurrent Neural Network，RNN)等人工神经网络方法，这些方法在序列标注方法上已经取得了较好的表现。但是，机器学习方法过度的依赖人工提取特征和特定的语言资源，神经网络方法虽然在模型中能够实现特征的自动抽取，但有需要大量的有标注或注释的语料进行参数训练，而新闻语料中的公司名称已标注或注释的语料相对而言较少，难以训练复杂的模型；而新闻语料中的公司名称常常会以公司名称的简称出现，或者会出现一些新的公司名称，利用该模型对新闻语料文本的公司名称的识别效果并不理想。

发明内容

[0003] 基于此，有必要针对传统的新闻实体识别模型对公司名称的识别准确度低的技术问题，提供一种新闻实体识别方法、装置、计算机设备和存储介质。

[0004] 一种新闻实体识别方法，所述方法包括：

[0005] 获取待标注新闻语料数据，并获取所述待标注新闻语料数据中的每一个中文字符的字符序列；

[0006] 将所述字符序列输入到预先训练好的双向语言模型中获取输出值，得到所述中文字符的第一字向量；

[0007] 将所述字符序列输入到预设词向量模型中获取输出值，得到所述中文字符的第二字向量；

[0008] 合并所述第一字向量与所述第二字向量，得到与所述待标注新闻语料数据中的中文字符对应的待标注字向量；

[0009] 将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到所述待标注新闻语料数据的最佳标签序列，并从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称。

[0010] 在其中一个实施例中，所述实体识别模型包括双向循环神经网络以及条件随机场模型；

[0011] 所述将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到所述待标注新闻语料数据的最佳标签序列的步骤，包括：

[0012] 将所述待标注字向量输入到双向循环神经网络中，获取所述待标注新闻语料数据中的中文字符对应的特征向量；

[0013] 将所述特征向量输入至条件随机场模型中，获取所述待标注新闻语料数据的最佳标签序列。

[0014] 在其中一个实施例中，所述将所述字符序列输入到预先训练好的双向语言模型中获取输出值的步骤之前，还包括：

[0015] 建立一个双向递归神经网络模型，其中所述双向神经网络模型包括前向递归神经网络隐含层以及后向递归神经网络隐含层；

[0016] 获取无标注语料样本，将所述无标注语料样本中的中文字符转化为对应的字符序列；

[0017] 将所述无标注语料样本的字符序列输入至所述双向的递归神经网络模型中，利用所述无标注语料样本的字符序列对所述前向递归神经网络隐含层以及所述后向递归神经网络隐含层进行无监督训练，得到所述双向语言模型。

[0018] 在其中一个实施例中，所述将所述字符序列输入到预先训练好的双向语言模型中获取输出值，得到所述中文字符的第一字向量的步骤，包括：

[0019] 将所述待标注新闻语料数据的字符序列输入至所述前向递归神经网络隐含层中，获取前向隐状态序列；

[0020] 将所述待标注新闻语料数据的字符序列输入至所述后向递归神经网络隐含层中，获取后向隐状态序列；

[0021] 合并所述前向隐状态序列以及所述后向隐状态序列生成所述中文字符的第一字向量。

[0022] 在其中一个实施例中，所述得到所述双向语言模型的步骤之后，包括：

[0023] 获取新闻语料训练样本，并获取所述新闻语料训练样本中的每一个中文字符的字符序列，其中，所述新闻语料训练样本中的中文字符具有对应的字符标签；

[0024] 将所述新闻语料训练样本中的字符序列分别输入至所述双向语言模型以及所述预设词向量模型中获取输出值，得到第三字向量以及第四字向量；

[0025] 合并所述第三字向量与所述第四字向量，得到与所述新闻语料训练样本中的中文字符对应的字向量样本；

[0026] 根据所述新闻语料训练样本中的中文字符对应的字向量样本以及字符标签，对所述实体识别模型进行有监督训练。

[0027] 在其中一个实施例中，所述条件随机场模型中采用的标注规则为BIOES模式；

[0028] 所述从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称的步骤，包括：

[0029] 从最佳标签序列中提取连续标注的B标签、I标签以及E标签对应的中文字符，将提取得到的中文字符组合作为识别出的企业名称。

[0030] 一种新闻实体识别装置，所述装置包括：

[0031] 新闻语料获取模块，用于获取待标注新闻语料数据，并获取所述待标注新闻语料数据中的每一个中文字符的字符序列；

[0032] 第一字向量获取模块，用于将所述字符序列输入到预先训练好的双向语言模型中获取输出值，得到所述中文字符的第一字向量；

[0033] 第二字向量获取模块，用于将所述字符序列输入到预设词向量模型中获取输出

值，得到所述中文字符的第二字向量；

[0034] 待标注字向量获取模块，用于合并所述第一字向量与所述第二字向量，得到与所述待标注新闻语料数据中的中文字符对应的待标注字向量；

[0035] 公司名称获取模块，用于将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到所述待标注新闻语料数据的最佳标签序列，并从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称。

[0036] 在其中一个实施例中，所述第一字向量获取模块用于将所述待标注新闻语料数据的字符序列输入至所述前向递归神经网络隐含层中，获取前向隐状态序列；将所述待标注新闻语料数据的字符序列输入至所述后向递归神经网络隐含层中，获取后向隐状态序列；合并所述前向隐状态序列以及所述后向隐状态序列生成所述中文字符的第一字向量。

[0037] 一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现以下步骤：

[0038] 获取待标注新闻语料数据，并获取所述待标注新闻语料数据中的每一个中文字符的字符序列；

[0039] 将所述字符序列输入到预先训练好的双向语言模型中获取输出值，得到所述中文字符的第一字向量；

[0040] 将所述字符序列输入到预设词向量模型中获取输出值，得到所述中文字符的第二字向量；

[0041] 合并所述第一字向量与所述第二字向量，得到与所述待标注新闻语料数据中的中文字符对应的待标注字向量；

[0042] 将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到所述待标注新闻语料数据的最佳标签序列，并从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称。

[0043] 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现以下步骤：

[0044] 获取待标注新闻语料数据，并获取所述待标注新闻语料数据中的每一个中文字符的字符序列；

[0045] 将所述字符序列输入到预先训练好的双向语言模型中获取输出值，得到所述中文字符的第一字向量；

[0046] 将所述字符序列输入到预设词向量模型中获取输出值，得到所述中文字符的第二字向量；

[0047] 合并所述第一字向量与所述第二字向量，得到与所述待标注新闻语料数据中的中文字符对应的待标注字向量；

[0048] 将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到所述待标注新闻语料数据的最佳标签序列，并从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称。

[0049] 上述新闻实体识别方法、装置、计算机设备和存储介质，先通过双向语言模型获取待标注新闻语料数据中中文字符的第一字向量，然后将第一字向量与通过预设词向量模型获取的第二字向量进行级联合并，作为待标注新闻语料数据中中文字符的待识别字向量，

用于新闻语料数据中公司名称的识别中，其中，第一字向量以及第二字向量有效表征中各个中文字符在待标注新闻语料中位置的语义以及语法含义，字符级别的知识能够为下游任务捕获到与语言相关的特征，在公司名称的识别过程中能够提供更多的语义特征，使得有歧义的公司名称的识别结果减少，提高公司名称识别结果的准确度。

附图说明

[0050] 图1为本发明一个实施例中新闻实体识别方法的应用场景图；

[0051] 图2为本发明一个实施例中新闻实体识别方法的流程示意图；

[0052] 图3为本发明另一个实施例中新闻实体识别方法的流程示意图；

[0053] 图4为本发明一个实施例中新闻实体识别装置的结构框图；

[0054] 图5为本发明另一个实施例中新闻实体识别装置的结构框图；

[0055] 图6为一个实施例中计算机设备的内部结构图。

具体实施方式

[0056] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0057] 本申请提供的新闻实体识别方法，可以应用于如图1所示的应用环境中。其中，终端102通过网络与服务器104通过网络进行通信。服务器104接收终端102发送的新闻语料数据，利用双向语言模型将新闻语料数据中的中文字符转换为第一字向量，利用预设词向量模型将中文字符转换为第二字向量，然后将第一字向量以及第二字向量进行级联获取待识别字向量，将待识别字向量输入至实体识别模型中以获取新闻语料数据中的公司名称。其中，终端102可以但不限于是各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务器104可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

[0058] 在一个实施例中，如图2所示，提供了一种新闻实体识别方法，以该方法应用于图1中的服务器为例进行说明，包括以下步骤：

[0059] 步骤S210：获取待标注新闻语料数据，并获取待标注新闻语料数据中的每一个中文字符的字符序列。

[0060] 本步骤中，字符序列可以独热码(One-Hot Encoder)，即在一个字符序列中只存在一个1其余全为0的序列，每个中文字符对应于一个字符序列。具体的，服务器获取终端输入的待标注新闻语料数据，并将该待标注新闻语料数据分解成单个中文字符的形式，并将每一个中文字符转为对应的字符序列。

[0061] 步骤S220：将字符序列输入到预先训练好的双向语言模型中获取输出值，得到中文字符的第一字向量。

[0062] 本步骤中，服务器将每个中文字符对应的字符序列输入至双向语言模型中，根据双向语言模型的输出中获取中文字符的第一字向量，其中，双向语言模型可以使用无任何标注的文本语料数据进行提前无监督训练，无需依赖有标注的文本语料数据，利用双向语言模型实现对中文字符的编码获得的第一字向量，表征着该中文字符在句子位置中的上下文语义以及语法。

[0063] 步骤S230:将字符序列输入到预设词向量模型中获取输出值,得到中文字符的第二字向量。

[0064] 本步骤中,预设词向量模型可以是利用word2vce训练的基于字的嵌入模型,服务器将每个中文字符对应的字符序列输入至预设词向量模型中,根据预设词向量模型的输出中获取中文字符的第二字向量。

[0065] 步骤S240:合并第一字向量与第二字向量,得到与待标注新闻语料数据中的中文字符对应的待标注字向量。

[0066] 本步骤中,服务器中获得每个中文符号对应第一字向量以及第二字向量后,将第一字向量以及第二字向量进行拼接,获得每个中文字符的待标注字向量。具体的,待标注新闻语料数据中的中文字符对应的第一字向量表示为h_1以及第二字向量表示为h_2,则将第一字向量h_1与第二字向量h_2进行级联拼接,获得待标注字向量$h_t=[h_1;h_2]$。

[0067] 步骤S250:将待标注字向量输入到预先训练好的实体识别模型中获取输出值,得到待标注新闻语料数据的最佳标签序列,并从最佳标签序列中提取出待标注新闻语料数据中的公司名称。

[0068] 本步骤中,服务器获得待标注新闻语料中每个中文字符对应的待标注字向量以后,将待标注字向量依次输入至实体识别模型中,以获取待标注新闻语料中中文字符的最佳标签序列,并根据最佳标签序列中每个中文字符对应的标注标签,获取待标注新闻语料中的公司名称。

[0069] 上述新闻实体识别方法中,先通过双向语言模型获取待标注新闻语料数据中中文字符的第一字向量,然后将第一字向量与通过预设词向量模型获取的第二字向量进行级联合并,作为待标注新闻语料数据中中文字符的待识别字向量,用于新闻语料数据中公司名称的识别中,其中,第一字向量以及第二字向量有效表征中各个中文字符在待标注新闻语料中位置的语义以及语法含义,字符级别的知识能够为下游任务捕获到与语言相关的特征,在公司名称的识别过程中能够提供更多的语义特征,使得有歧义的公司名称的识别结果减少,提高公司名称识别结果的准确度。

[0070] 在其中一个实施例中,实体识别模型包括双向循环神经网络以及条件随机场模型;将待标注字向量输入到预先训练好的实体识别模型中获取输出值,得到待标注新闻语料数据的最佳标签序列的步骤,包括:将待标注字向量输入到双向循环神经网络中,获取待标注新闻语料数据中的中文字符对应的特征向量;将特征向量输入至条件随机场模型中,获取待标注新闻语料数据的最佳标签序列。

[0071] 本实施例中,服务器将待标注新闻语料数据中每个中文字符对应的待标注字向量输入至双向循环神经网络中,通过双向循环神经网络获取该中文字符的特征向量,该特征向量中的数据可以理解为该中文字符分类到各个类型标签的分数值;服务器将中文字符的特征向量输入至条件随机场模型中,通过条件随机场模型选择使得待标注新闻语料数据具有最高预测分数的、各中文字符最佳标签,从而获取待标注新闻语料数据的最佳标签序列。

[0072] 具体的,双向循环神经网络包括前向循环神经网络层以及后向神经网络层;服务器将中文字符对应的待标注字向量h_t输入至双向循环神经网络的前向循环神经网络层,根据当前待标注字向量h_t的前一个待标注字向量h_{t-1}的隐状态向量$\overrightarrow{x_{t-1}}$计算当前待标注字向

量h_t的前向隐状态向量$\overrightarrow{x_t}$，并通过双向循环神经网络的后向循环神经网络层，根据当前待标注字向量h_t的后一个待标注字向量h_{t+1}的隐状态向量$\overleftarrow{x_{t+1}}$计算当前待标注字向量h_t的后向隐状态向量$\overleftarrow{x_t}$，然后将前向隐状态向量$\overrightarrow{x_t}$与后向隐状态向量$\overleftarrow{x_t}$进行拼接，获得待标注新闻语料数据中的中文字符对应的特征向量$x_t=\left[\overrightarrow{x_t};\overleftarrow{x_t}\right]$；获得中文字符对应的特征向量$x_t$后，将特征向量依次输入至条件随机场模型中，计算待标注新闻语料数据的最佳标签序列。

[0073] 在其中一个实施例中，将字符序列输入到预先训练好的双向语言模型中获取输出值的步骤之前，还包括：建立一个双向递归神经网络模型，其中双向神经网络模型包括前向递归神经网络隐含层以及后向递归神经网络隐含层；获取无标注语料样本，将无标注语料样本中的中文字符转化为对应的字符序列；将无标注语料样本的字符序列输入至双向的递归神经网络模型中，利用无标注语料样本的字符序列对前向递归神经网络隐含层以及后向递归神经网络隐含层进行无监督训练，得到双向语言模型。

[0074] 本实施例中，无监督训练是指只将输入项的数据输入，而输出项的数据并不提供，也无须双向递归神经网络模型输出定义的输出项数据；双向语言模型包括前向递归神经网络隐含层以及后向递归神经网络隐含层；服务器在获取无标注语料样本后，将无标注语料样本中中文字符转化为对应的字符序列，并将这些字符序列作为输入项输入至双向递归神经网络模型中，双向递归神经网络模型的前向递归神经网络隐含层以及后向递归神经网络隐含层分别对输入的字符序列进行特征学习，以获取经过特征学习的双向语言模型。通过大量无标注语料样本训练的双向语言模型可以在给定前一个中文字符的情况下预测句子中下一个中文字符，在一定程度上表达了中文字符在句子内容中的语义和语法含义，为下游任务捕获更多与语言相关的特征，使得实体识别减少对专业领域中的标注数据的依赖，解决实体识别模型语料库不足的问题，使得实体识别模型更少依赖人工特征。

[0075] 在其中一个实施例中，将字符序列输入到预先训练好的双向语言模型中获取输出值，得到中文字符的第一字向量的步骤，包括：将待标注新闻语料数据的字符序列输入至前向递归神经网络隐含层中，获取前向隐状态序列；将待标注新闻语料数据的字符序列输入至后向递归神经网络隐含层中，获取后向隐状态序列；合并前向隐状态序列以及后向隐状态序列生成中文字符的第一字向量。

[0076] 本实施例中，服务器将待标注新闻语料数据中的中文字符的字符序列输入至双向语言模型中，通过前向递归神经网络隐含层，根据当前待标注字向量的前一个待标注字向量的隐状态向量计算当前待标注字向量的前向隐状态序列；并通过后向递归神经网络隐含层，根据当前待标注字向量的后一个待标注字向量的隐状态向量计算当前待标注字向量的后向隐状态序列，然后将前向隐状态序列与后向隐状态序列进行级联，获得第一字向量，其中，第一字向量包含着中文字符与前后中文字符的依赖关系，在后续对新闻文本语料进行实体识别时，能够提供更多与语言、语义相关的特征，有效识别公司名称任务减少对专业领域中的标注数据的依赖。

[0077] 在其中一个实施例中，得到双向语言模型的步骤之后，包括：获取新闻语料训练样本，并获取新闻语料训练样本中的每一个中文字符的字符序列，其中，新闻语料训练样本中的中文字符具有对应的字符标签；将新闻语料训练样本中的字符序列分别输入至双向语言

模型以及预设词向量模型中获取输出值，得到第三字向量以及第四字向量；合并第三字向量与第四字向量，得到与新闻语料训练样本中的中文字符对应的字向量样本；根据新闻语料训练样本中的中文字符对应的字向量样本以及字符标签，对实体识别模型进行有监督训练。

[0078] 本实施例中，新闻语料训练样本中，各个公司名称已经对应的标注上实体标签；在获得双向语言模型以后，可以利用新闻语料训练样本对实体识别模型进行有监督训练，其中，通过双向语言模型获得的第三字向量包含着中文字符与前后中文字符的依赖关系，能够提供更多与语言、语义相关的特征，在进行实体识别模型的训练时，有效减少实体识别模型对专业领域中的标注数据的依赖，在新闻语料中的公司名称已标注或注释的语料相对而言较少的情况下，有效提高实体识别模型对待标注字向量的最佳标签的标注的准确性，从而提高待标注新闻语料数据中的公司名称识别的准确性。

[0079] 在其中一个实施例中，条件随机场模型中采用的标注规则为BIOES模式；从最佳标签序列中提取出待标注新闻语料数据中的公司名称的步骤，包括：从最佳标签序列中提取连续标注的B标签、I标签以及E标签对应的中文字符，将提取得到的中文字符组合作为识别出的公司名称。

[0080] 本实施例中，BIOES模式下标签列表如下：B标签，即Begin，表示公司名称的开始字符；I标签，即Intermediate，表示公司名称的中间字符；E标签，即End，表示公司名称的结尾字符；S标签，即Single，表示单个字符；O标签，即Other，表示其他字符，用于标记无关字符；服务器在获取到待标注新闻语料数据中的最佳标签序列后，将连续标注的B标签、I标签以及E标签对应的中文字符确定为公司名称。

[0081] 在一个实施例中，如图3所示，提供了一种新闻实体识别方法，包括以下步骤：

[0082] 步骤S301：获取待标注新闻语料数据，并获取待标注新闻语料数据中的每一个中文字符的字符序列。

[0083] 本步骤中，服务器获取终端输入的待标注新闻语料数据，并将该待标注新闻语料数据分解成单个中文字符的形式，并将每一个中文字符转为对应的字符序列；例如，服务器获取终端输入待标注新闻语料数据为“发展集团转型成功”，并将“发展集团转型成功”分解为单个中文字符形式，每个中文字符转为对应的字符序列a_i，i＝1,2,…,n，在本实施例中n＝8，每个字符序列的维数可以根据具体情况设置。

[0084] 步骤S302：将待标注新闻语料数据的字符序列输入至双向语言模型的前向递归神经网络隐含层中，获取前向隐状态序列。

[0085] 具体的，服务器将待标注新闻语料数据每个中文字符对应的字符序列(a_1,a_2,a_3,a_4,a_5,a_6,a_7,a_8)输入至双向语言模型的前向循环神经网络层，根据当前字符序列a_i的前一个字符序列a_{i-1}的隐状态序列$\overrightarrow{h_{1,i-1}}$计算当前字符序列$a_i$的前向隐状态序列$\overrightarrow{h_{1,i}}$，获得待标注新闻语料数据中中文字符对应的前向隐状态序列$\left(\overrightarrow{h_{1,1}},\overrightarrow{h_{1,2}},\overrightarrow{h_{1,3}},\overrightarrow{h_{1,4}},\overrightarrow{h_{1,5}},\overrightarrow{h_{1,6}},\overrightarrow{h_{1,7}},\overrightarrow{h_{1,8}}\right)$。

[0086] 步骤S303：将待标注新闻语料数据的字符序列输入至双向语言模型的后向递归神经网络隐含层中，获取后向隐状态序列。

[0087] 具体的，服务器同时将待标注新闻语料数据每个中文字符对应的字符序列(a_1,a_2,a_3,a_4,a_5,a_6,a_7,a_8)输入至双向语言模型的前向循环神经网络层，根据当前字符序列a_i的后

一个字符序列a_{i+1}的隐状态序列$\overleftarrow{h_{1,i+1}}$计算当前字符序列$a_i$的前向隐状态序列$\overleftarrow{h_{1,i}}$，获得待标注新闻语料数据中中文字符对应的前向隐状态序列$\left(\overleftarrow{h_{1,1}},\overleftarrow{h_{1,2}},\overleftarrow{h_{1,3}},\overleftarrow{h_{1,4}},\overleftarrow{h_{1,5}},\overleftarrow{h_{1,6}},\overleftarrow{h_{1,7}},\overleftarrow{h_{1,8}}\right)$。

[0088] 步骤S304：合并前向隐状态序列以及后向隐状态序列生成中文字符的第一字向量。

[0089] 具体的，前向隐状态序列$\overleftarrow{h_{1,i}}$与后向隐状态序列$\overleftarrow{h_{1,i}}$进行拼接，获得待标注新闻语料数据中的中文字符对应的第一字向量$h_{1,i}=\left[\overleftarrow{h_{1,i}};\overleftarrow{h_{1,i}}\right]$，从而获得待标注新闻语料数据中中文字符对应的第一字向量$(h_{1,1},h_{1,2},h_{1,3},h_{1,4},h_{1,5},h_{1,6},h_{1,7},h_{1,8})$。

[0090] 步骤S305：将字符序列输入到预设词向量模型中获取输出值，得到中文字符的第二字向量。

[0091] 本步骤中，服务器将每个中文字符对应的字符序列a_i输入至预设word2vce词向量模型中，根据word2vce词向量模型的输出中获取中文字符的第二字向量$h_{2,i}$，从而获得待标注新闻语料数据中中文字符对应的第二字向量$(h_{2,1},h_{2,2},h_{2,3},h_{2,4},h_{2,5},h_{2,6},h_{2,7},h_{2,8})$。

[0092] 步骤S306：合并第一字向量与第二字向量，得到与待标注新闻语料数据中的中文字符对应的待标注字向量。

[0093] 本步骤中，服务器将待标注新闻语料数据中的中文字符对应的第一字向量$h_{1,i}$以及第二字向量$h_{2,i}$进行级联拼接，获得待标注字向量$h_{t,i}=[h_{1,i};h_{2,i}]$。

[0094] 步骤S307：将待标注字向量输入到双向循环神经网络中，获取待标注新闻语料数据中的中文字符对应的特征向量。

[0095] 步骤S308：将特征向量输入至条件随机场模型中，获取待标注新闻语料数据的最佳标签序列，其中，条件随机场模型中采用的标注规则为BIOES模式。

[0096] 具体的，服务器将待标注新闻语料数据中每个中文字符对应的待标注字向量$h_{t,i}$输入至双向循环神经网络中，通过双向循环神经网络获取该中文字符的特征向量，并将中文字符的特征向量输入至条件随机场模型中，通过条件随机场模型选择使得待标注新闻语料数据具有最高预测分数的、各中文字符最佳标签，从而获取待标注新闻语料数据的最佳标签序列，本实施例中，对于待标注新闻语料数据“发展集团转型成功”，最终获得的最佳标签序列为{B,I,I,E,0,0,0,0}。

[0097] 步骤S309：从最佳标签序列中提取连续标注的B标签、I标签以及E标签对应的中文字符，将提取得到的中文字符组合作为识别出的企业名称。

[0098] 具体的，从待标注新闻语料数据“发展集团转型成功”最终提取得到企业名称“发展集团”。

[0099] 上述新闻实体识别方法、装置、计算机设备和存储介质，先通过双向语言模型获取待标注新闻语料数据中中文字符的第一字向量，然后将第一字向量与通过预设词向量模型获取的第二字向量进行级联，作为待标注新闻语料数据中中文字符的待识别字向量，用于新闻语料数据中公司名称的识别中，其中，通过双向语言模型获取的第一字向量有效表征中各个中文字符在待标注新闻语料中位置的语义以及语法含义，能够为下游任务捕获很多与语言相关的特征，在公司名称的识别过程中能够提供更多的语义特征，使得有歧义的公司名称的识别结果减少，提高公司名称识别结果的准确度。

[0100] 应该理解的是，虽然图2以及图3的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图2以及图3中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0101] 在其中一个实施例中，如图4所示，提供了一种新闻实体识别装置，包括：新闻语料获取模块410、第一字向量获取模块420、第二字向量获取模块430、待标注字向量获取模块440和公司名称获取模块450，其中：

[0102] 新闻语料获取模块410，用于获取待标注新闻语料数据，并获取待标注新闻语料数据中的每一个中文字符的字符序列；

[0103] 第一字向量获取模块420，用于将字符序列输入到预先训练好的双向语言模型中获取输出值，得到中文字符的第一字向量；

[0104] 第二字向量获取模块430，用于将字符序列输入到预设词向量模型中获取输出值，得到中文字符的第二字向量；

[0105] 待标注字向量获取模块440，用于合并第一字向量与第二字向量，得到与待标注新闻语料数据中的中文字符对应的待标注字向量；

[0106] 公司名称获取模块450，用于将待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到待标注新闻语料数据的最佳标签序列，并从最佳标签序列中提取出待标注新闻语料数据中的公司名称。

[0107] 在其中一个实施例中，第一字向量获取模块用于将待标注新闻语料数据的字符序列输入至前向递归神经网络隐含层中，获取前向隐状态序列；将待标注新闻语料数据的字符序列输入至后向递归神经网络隐含层中，获取后向隐状态序列；合并前向隐状态序列以及后向隐状态序列生成中文字符的第一字向量。

[0108] 在其中一个实施例中，实体识别模型包括双向循环神经网络以及条件随机场模型；公司名称获取模块450用于将待标注字向量输入到双向循环神经网络中，获取待标注新闻语料数据中的中文字符对应的特征向量；将特征向量输入至条件随机场模型中，获取待标注新闻语料数据的最佳标签序列。

[0109] 在其中一个实施例中，如图5所示，提供了一种新闻实体识别装置，新闻实体识别装置还包括双向语言模型构建模块460，双向语言模型构建模块460用于建立一个双向递归神经网络模型，其中双向神经网络模型包括前向递归神经网络隐含层以及后向递归神经网络隐含层；获取无标注语料样本，将无标注语料样本中的中文字符转化为对应的字符序列；将无标注语料样本的字符序列输入至双向的递归神经网络模型中，利用无标注语料样本的字符序列对前向递归神经网络隐含层以及后向递归神经网络隐含层进行无监督训练，得到双向语言模型。

[0110] 在其中一个实施例中，双向语言模型构建模块460还用于获取新闻语料训练样本，并获取新闻语料训练样本中的每一个中文字符的字符序列，其中，新闻语料训练样本中的中文字符具有对应的字符标签；将新闻语料训练样本中的字符序列分别输入至双向语言模

型以及预设词向量模型中获取输出值，得到第三字向量以及第四字向量；合并第三字向量与第四字向量，得到与新闻语料训练样本中的中文字符对应的字向量样本；根据新闻语料训练样本中的中文字符对应的字向量样本以及字符标签，对实体识别模型进行有监督训练。

[0111] 在其中一个实施例中，条件随机场模型中采用的标注规则为BIOES模式；公司名称获取模块450用于从最佳标签序列中提取连续标注的B标签、I标签以及E标签对应的中文字符，将提取得到的中文字符组合作为识别出的企业名称。

[0112] 关于新闻实体识别装置的具体限定可以参见上文中对于新闻实体识别方法的限定，在此不再赘述。上述新闻实体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。

[0113] 在一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图6所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储待识别新闻语料数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种新闻实体识别方法。

[0114] 本领域技术人员可以理解，图6中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

[0115] 在一个实施例中，提供了一种计算机设备，包括存储器和处理器，该存储器存储有计算机程序，该处理器执行计算机程序时实现以下步骤：

[0116] 获取待标注新闻语料数据，并获取所述待标注新闻语料数据中的每一个中文字符的字符序列；

[0117] 将所述字符序列输入到预先训练好的双向语言模型中获取输出值，得到所述中文字符的第一字向量；

[0118] 将所述字符序列输入到预设词向量模型中获取输出值，得到所述中文字符的第二字向量；

[0119] 合并所述第一字向量与所述第二字向量，得到与所述待标注新闻语料数据中的中文字符对应的待标注字向量；

[0120] 将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到所述待标注新闻语料数据的最佳标签序列，并从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称。

[0121] 在其中一个实施例中，所述实体识别模型包括双向循环神经网络以及条件随机场模型；处理器执行计算机程序实现所述将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到所述待标注新闻语料数据的最佳标签序列的步骤时，具体实现以下步骤：将所述待标注字向量输入到双向循环神经网络中，获取所述待标注新闻语料数据

中的中文字符对应的特征向量;将所述特征向量输入至条件随机场模型中,获取所述待标注新闻语料数据的最佳标签序列。

[0122] 在其中一个实施例中,处理器执行计算机程序时还实现以下步骤:建立一个双向递归神经网络模型,其中所述双向神经网络模型包括前向递归神经网络隐含层以及后向递归神经网络隐含层;获取无标注语料样本,将所述无标注语料样本中的中文字符转化为对应的字符序列;将所述无标注语料样本的字符序列输入至所述双向的递归神经网络模型中,利用所述无标注语料样本的字符序列对所述前向递归神经网络隐含层以及所述后向递归神经网络隐含层进行无监督训练,得到所述双向语言模型。

[0123] 在其中一个实施例中,处理器执行计算机程序实现所述将所述字符序列输入到预先训练好的双向语言模型中获取输出值,得到所述中文字符的第一字向量的步骤时,具体实现以下步骤:将所述待标注新闻语料数据的字符序列输入至所述前向递归神经网络隐含层中,获取前向隐状态序列;将所述待标注新闻语料数据的字符序列输入至所述后向递归神经网络隐含层中,获取后向隐状态序列;合并所述前向隐状态序列以及所述后向隐状态序列生成所述中文字符的第一字向量。

[0124] 在其中一个实施例中,处理器执行计算机程序时还实现以下步骤:获取新闻语料训练样本,并获取所述新闻语料训练样本中的每一个中文字符的字符序列,其中,所述新闻语料训练样本中的中文字符具有对应的字符标签;将所述新闻语料训练样本中的字符序列分别输入至所述双向语言模型以及所述预设词向量模型中获取输出值,得到第三字向量以及第四字向量;合并所述第三字向量与所述第四字向量,得到与所述新闻语料训练样本中的中文字符对应的字向量样本;根据所述新闻语料训练样本中的中文字符对应的字向量样本以及字符标签,对所述实体识别模型进行有监督训练。

[0125] 在其中一个实施例中,所述条件随机场模型中采用的标注规则为BIOES模式;处理器执行计算机程序实现所述从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称的步骤时,具体实现以下步骤:从最佳标签序列中提取连续标注的B标签、I标签以及E标签对应的中文字符,将提取得到的中文字符组合作为识别出的企业名称。

[0126] 在其中一个实施例中,提供了一种计算机可读存储介质,其上存储有计算机程序,计算机程序被处理器执行时实现以下步骤:

[0127] 获取待标注新闻语料数据,并获取所述待标注新闻语料数据中的每一个中文字符的字符序列;

[0128] 将所述字符序列输入到预先训练好的双向语言模型中获取输出值,得到所述中文字符的第一字向量;

[0129] 将所述字符序列输入到预设词向量模型中获取输出值,得到所述中文字符的第二字向量;

[0130] 合并所述第一字向量与所述第二字向量,得到与所述待标注新闻语料数据中的中文字符对应的待标注字向量;

[0131] 将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值,得到所述待标注新闻语料数据的最佳标签序列,并从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称。

[0132] 在其中一个实施例中,所述实体识别模型包括双向循环神经网络以及条件随机场

模型;计算机程序被处理器执行实现所述将所述待标注字向量输入到预先训练好的实体识别模型中获取输出值,得到所述待标注新闻语料数据的最佳标签序列的步骤时,具体实现以下步骤:将所述待标注字向量输入到双向循环神经网络中,获取所述待标注新闻语料数据中的中文字符对应的特征向量;将所述特征向量输入至条件随机场模型中,获取所述待标注新闻语料数据的最佳标签序列。

[0133] 在其中一个实施例中,计算机程序被处理器执行时还实现以下步骤:建立一个双向递归神经网络模型,其中所述双向神经网络模型包括前向递归神经网络隐含层以及后向递归神经网络隐含层;获取无标注语料样本,将所述无标注语料样本中的中文字符转化为对应的字符序列;将所述无标注语料样本的字符序列输入至所述双向的递归神经网络模型中,利用所述无标注语料样本的字符序列对所述前向递归神经网络隐含层以及所述后向递归神经网络隐含层进行无监督训练,得到所述双向语言模型。

[0134] 在其中一个实施例中,计算机程序被处理器执行实现所述将所述字符序列输入到预先训练好的双向语言模型中获取输出值,得到所述中文字符的第一字向量的步骤时,具体实现以下步骤:将所述待标注新闻语料数据的字符序列输入至所述前向递归神经网络隐含层中,获取前向隐状态序列;将所述待标注新闻语料数据的字符序列输入至所述后向递归神经网络隐含层中,获取后向隐状态序列;合并所述前向隐状态序列以及所述后向隐状态序列生成所述中文字符的第一字向量。

[0135] 在其中一个实施例中,计算机程序被处理器执行时还实现以下步骤:获取新闻语料训练样本,并获取所述新闻语料训练样本中的每一个中文字符的字符序列,其中,所述新闻语料训练样本中的中文字符具有对应的字符标签;将所述新闻语料训练样本中的字符序列分别输入至所述双向语言模型以及所述预设词向量模型中获取输出值,得到第三字向量以及第四字向量;合并所述第三字向量与所述第四字向量,得到与所述新闻语料训练样本中的中文字符对应的字向量样本;根据所述新闻语料训练样本中的中文字符对应的字向量样本以及字符标签,对所述实体识别模型进行有监督训练。

[0136] 在其中一个实施例汇总,所述条件随机场模型中采用的标注规则为BIOES模式;计算机程序被处理器执行实现所述从所述最佳标签序列中提取出待标注新闻语料数据中的公司名称的步骤时,具体实现以下步骤:从最佳标签序列中提取连续标注的B标签、I标签以及E标签对应的中文字符,将提取得到的中文字符组合作为识别出的企业名称。

[0137] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,的计算机程序可存储于一非易失性计算机可读取存储介质中,该计算机程序在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0138] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例

中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0139] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

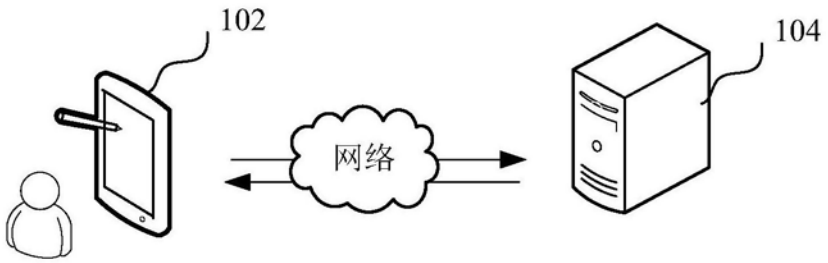

102
104
网络

图1

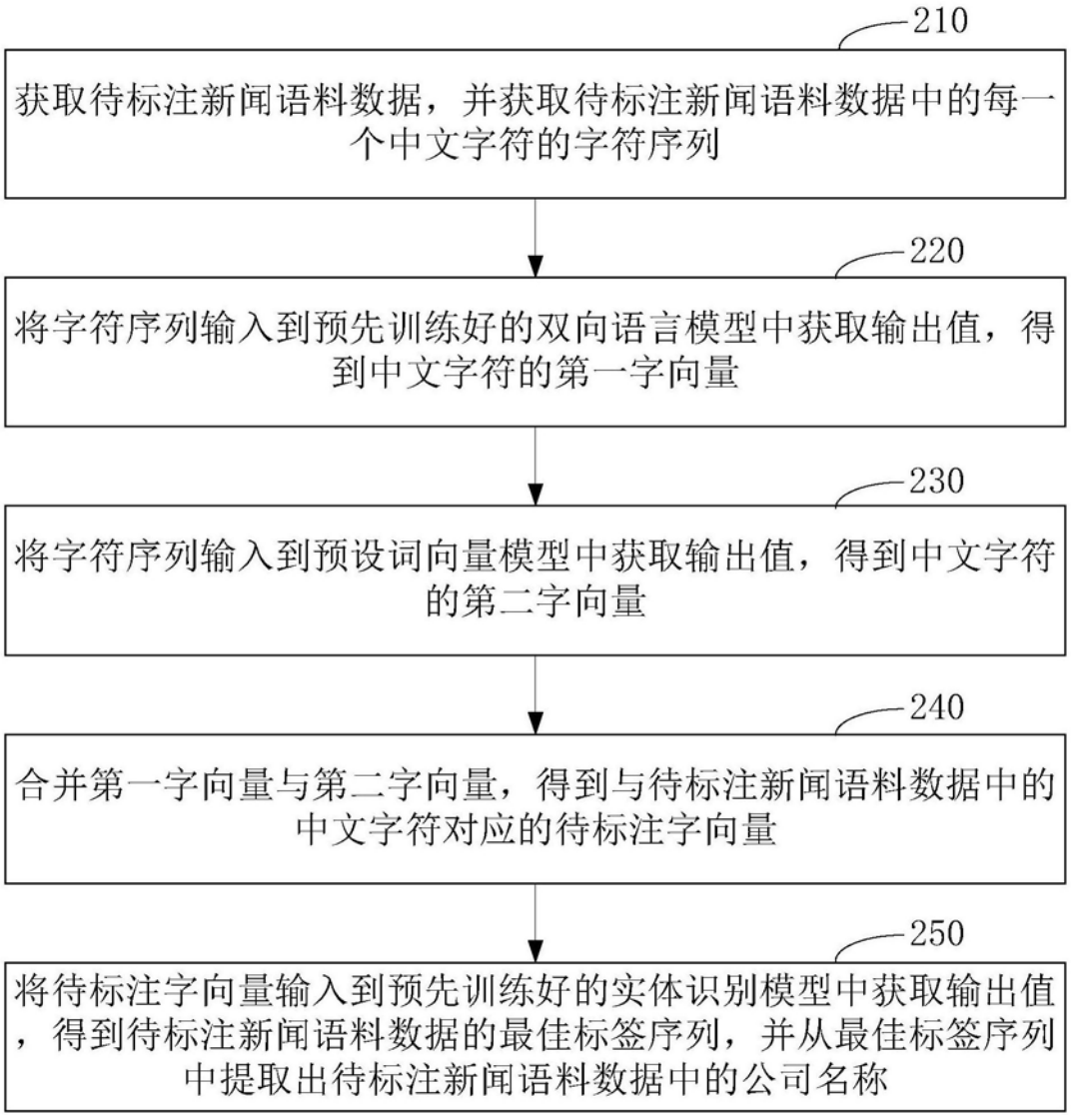

210
获取待标注新闻语料数据，并获取待标注新闻语料数据中的每一个中文字符的字符序列
220
将字符序列输入到预先训练好的双向语言模型中获取输出值，得到中文字符的第一字向量
230
将字符序列输入到预设词向量模型中获取输出值，得到中文字符的第二字向量
240
合并第一字向量与第二字向量，得到与待标注新闻语料数据中的中文字符对应的待标注字向量
250
将待标注字向量输入到预先训练好的实体识别模型中获取输出值，得到待标注新闻语料数据的最佳标签序列，并从最佳标签序列中提取出待标注新闻语料数据中的公司名称

图2

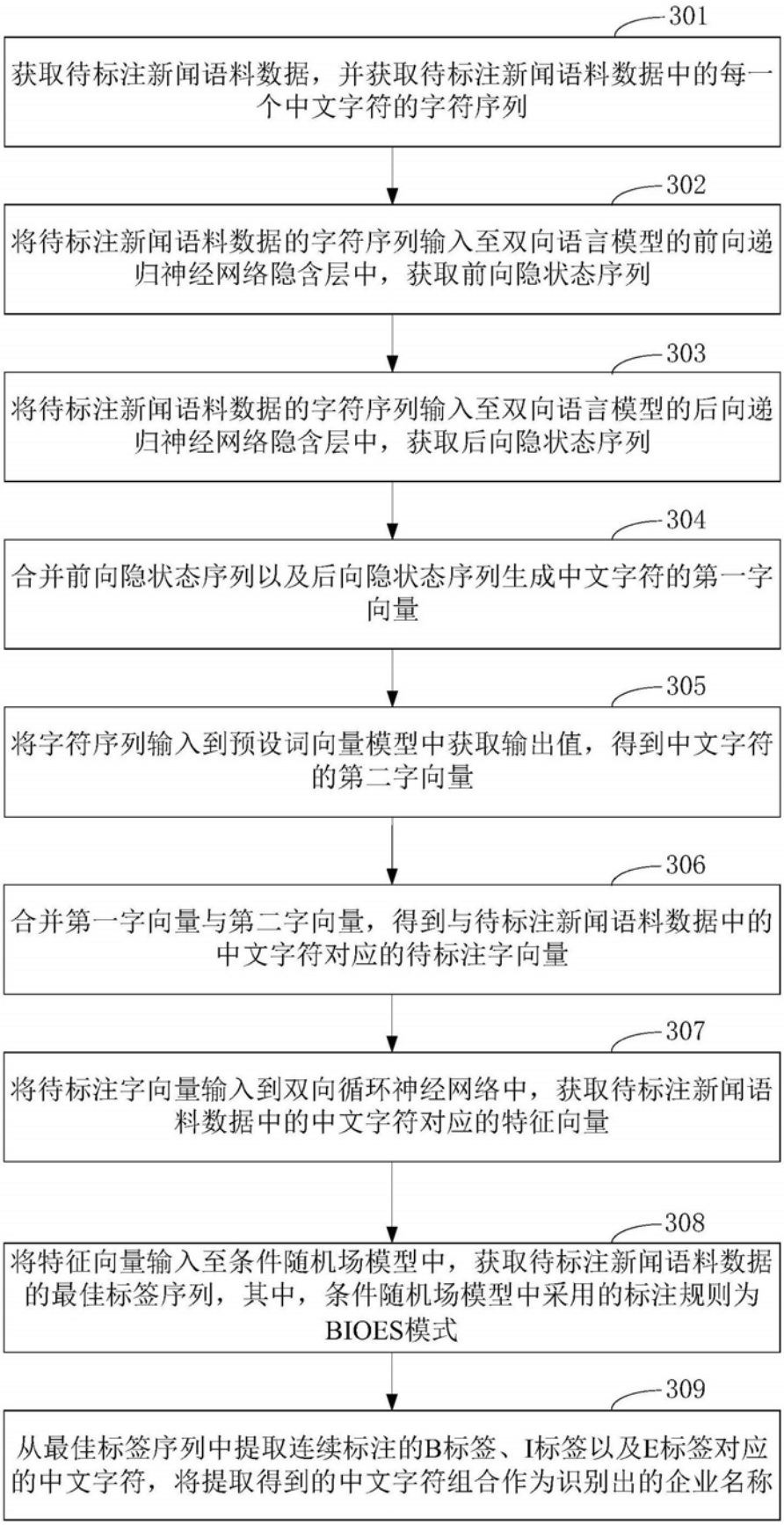

301
获取待标注新闻语料数据，并获取待标注新闻语料数据中的每一个中文字符的字符序列
302
将待标注新闻语料数据的字符序列输入至双向语言模型的前向递归神经网络隐含层中，获取前向隐状态序列
303
将待标注新闻语料数据的字符序列输入至双向语言模型的后向递归神经网络隐含层中，获取后向隐状态序列
304
合并前向隐状态序列以及后向隐状态序列生成中文字符的第一字向量
305
将字符序列输入到预设词向量模型中获取输出值，得到中文字符的第二字向量
306
合并第一字向量与第二字向量，得到与待标注新闻语料数据中的中文字符对应的待标注字向量
307
将待标注字向量输入到双向循环神经网络中，获取待标注新闻语料数据中的中文字符对应的特征向量
308
将特征向量输入至条件随机场模型中，获取待标注新闻语料数据的最佳标签序列，其中，条件随机场模型中采用的标注规则为BIOES模式
309
从最佳标签序列中提取连续标注的B标签、I标签以及E标签对应的中文字符，将提取得到的中文字符组合作为识别出的企业名称

图3

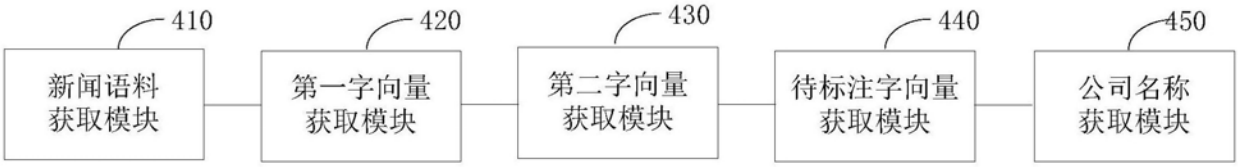

410
新闻语料获取模块
420
第一字向量获取模块
430
第二字向量获取模块
440
待标注字向量获取模块
450
公司名称获取模块

图4

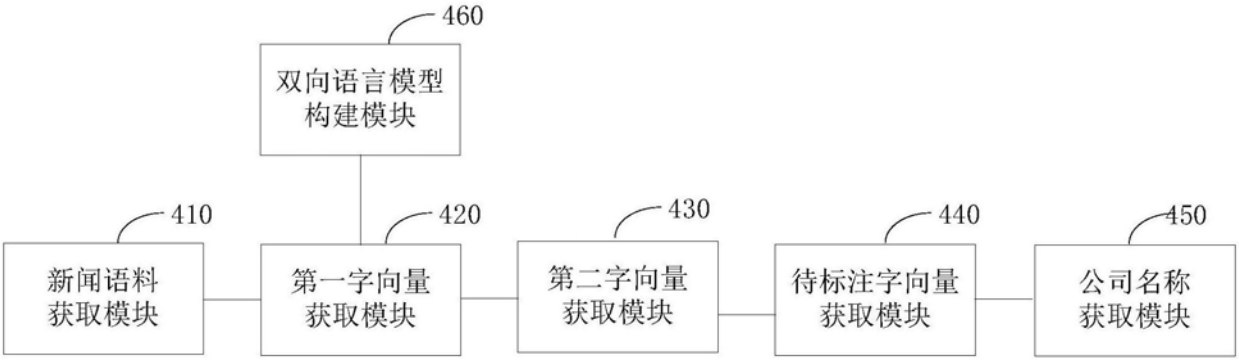

460
双向语言模型构建模块
410
新闻语料获取模块
420
第一字向量获取模块
430
第二字向量获取模块
440
待标注字向量获取模块
450
公司名称获取模块

图5

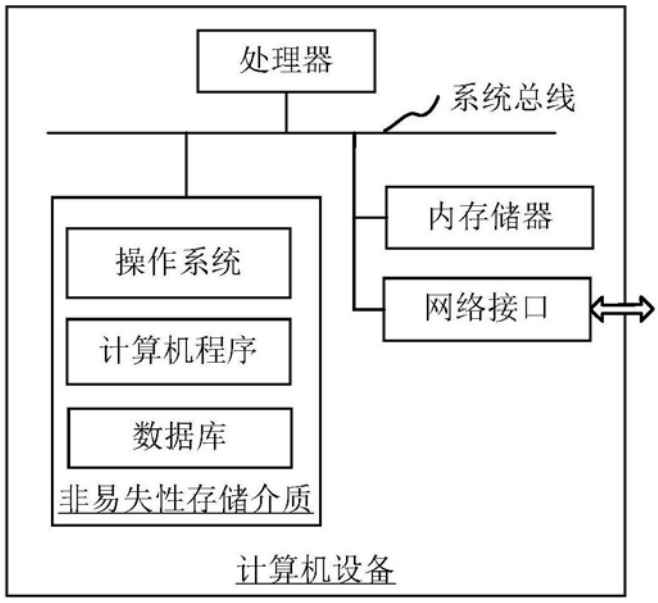

处理器
系统总线
内存储器
网络接口
操作系统
计算机程序
数据库
非易失性存储介质
计算机设备

图6