(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111079418 B
(45)授权公告日 2023.12.05

(21)申请号 201911078307.2

(22)申请日 2019.11.06

(65)同一申请的已公布的文献号
申请公布号 CN 111079418 A

(43)申请公布日 2020.04.28

(73)专利权人 科大讯飞股份有限公司
地址 230088 安徽省合肥市高新开发区望江西路666号

(72)发明人 尹坤 刘权 陈志刚 王智国 胡国平

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 程琛

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/242(2020.01)
G06F 40/30(2020.01)
G06N 3/0442(2023.01)
G06N 3/045(2023.01)
G06N 3/088(2023.01)

(56)对比文件
CN 106528526 A,2017.03.22
CN 108460013 A,2018.08.28
CN 108628823 A,2018.10.09
CN 110297913 A,2019.10.01
KR 20190015797 A,2019.02.15
CN 110222201 A,2019.09.10
CN 104899304 A,2015.09.09
CN 109388795 A,2019.02.26
CN 108536754 A,2018.09.14
CN 108874997 A,2018.11.23
CN 109614614 A,2019.04.12
CN 109657239 A,2019.04.19
CN 109858041 A,2019.06.07
CN 109871538 A,2019.06.11
CN 109902307 A,2019.06.18
CN 109918680 A,2019.06.21
CN 110020438 A,2019.07.16
CN 110134969 A,2019.08.16
CN 110222330 A,2019.09.10
CN 110287961 A,2019.09.27
KR 20150050140 A,2015.05.08
KR 20190065665 A,2019.06.12
US 2017197152 A1,2017.07.13
WO 2015080558 A1,2015.06.04

审查员 白红昌

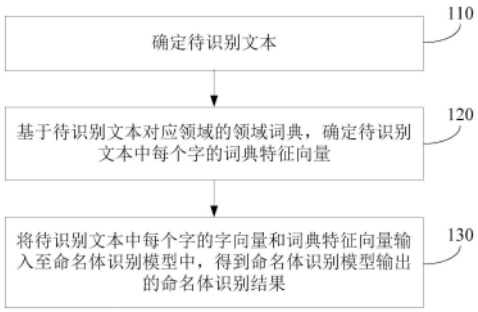
权利要求书2页 说明书11页 附图5页

(54)发明名称

命名体识别方法、装置、电子设备和存储介质

(57)摘要

本发明实施例提供一种命名体识别方法、装置、电子设备和存储介质,其中方法包括:确定待识别文本;基于待识别文本对应领域的领域词典,确定待识别文本中每个字的词典特征向量;将待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到命名体识别模型输出的命名体识别结果;其中,命名体识别模型是基于样本文本中每个样本字的字向量和词典特征向量,以及每个样本字的命名体标记训练得到的。本发明实施例提供的方法、装置、电子设备和存储介质,克服了由于词条冲突导致的识别准确性低的问题,提高了命名体识别的准确性。



CN 111079418 B

1.一种命名体识别方法,其特征在于,包括:

确定待识别文本;

基于所述待识别文本对应领域的领域词典,确定所述待识别文本中每个字的词典特征向量;

将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果;

其中,所述命名体识别模型是基于样本文本中每个样本字的字向量和词典特征向量,以及每个样本字的命名体标记训练得到的;

命名体识别模型基于注意力机制确定待识别文本中每个字的词典特征向量所对应的注意力特征向量,并基于每个字的注意力特征向量输出命名体识别结果,所述注意力机制用于基于任一字的语义,确定不同词典特征与所述任一字的语义之间的相关性,基于相关性调整词典特征向量中不同词典特征的权重,计算得到权重调整后的所述注意力特征向量。

2.根据权利要求1所述的命名体识别方法,其特征在于,所述命名体识别模型包括输入编码层、词典特征选择层和标签预测层;

对应地,所述将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果,具体包括:

将所述待识别文本中每个字的字向量和词典特征向量输入至所述输入编码层,得到所述输入编码层输出的每个字的隐层向量;

将每个字的隐层向量和词典特征向量输入至所述词典特征选择层,得到所述词典特征选择层输出的每个字的注意力特征向量;

将每个字的注意力特征向量输入至所述标签预测层,得到所述标签预测层输出的所述命名体识别结果。

3.根据权利要求2所述的命名体识别方法,其特征在于,所述将每个字的隐层向量和词典特征向量输入至所述词典特征选择层,得到所述词典特征选择层输出的每个字的注意力特征向量,具体包括:

基于任一字的隐层向量和词典特征向量,确定所述任一字相对于每一词典特征的权重;

基于所述任一字相对于每一词典特征的权重,对所述任一字的词典特征向量进行加权,得到所述任一字的注意力特征向量。

4.根据权利要求3所述的命名体识别方法,其特征在于,所述将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果,之前还包括:

基于损失函数对初始模型进行训练,得到所述命名体识别模型;

其中,所述损失函数包括识别结果损失函数和权重损失函数,其中所述识别结果损失函数对应于所述命名体识别结果,所述权重损失函数对应于字相对于每一词典特征的权重。

5.根据权利要求1所述的命名体识别方法,其特征在于,所述领域词典包括对应于不同命名体类型的词典;

对应地,所述基于所述待识别文本对应领域的领域词典,确定所述待识别文本中每个字的词典特征向量,具体包括:

基于对应于任一命名体类型的词典,确定所述待识别文本中每一字对应于所述任一命名体类型的词典特征;

基于任一字对应于每一命名体类型的词典特征,确定所述任一字的词典特征向量。

6.根据权利要求5所述的命名体识别方法,其特征在于,所述基于任一字对应于每一命名体类型的词典特征,确定所述任一字的词典特征向量,具体包括:

向量化任一字对应于每一命名体类型的词典特征,得到所述任一字的特征向量;

稀疏化处理所述任一字的特征向量,得到所述任一字的词典特征向量。

7.一种命名体识别装置,其特征在于,包括:

文本确定单元,用于确定待识别文本;

词典匹配单元,用于基于所述待识别文本对应领域的领域词典,确定所述待识别文本中每个字的词典特征向量;

命名体识别单元,用于将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果;

其中,所述命名体识别模型是基于样本文本中每个样本字的字向量和词典特征向量,以及每个样本字的命名体标记训练得到的;

命名体识别模型基于注意力机制确定待识别文本中每个字的词典特征向量所对应的注意力特征向量,并基于每个字的注意力特征向量输出命名体识别结果,所述注意力机制用于基于任一字的语义,确定不同词典特征与所述任一字的语义之间的相关性,基于相关性调整词典特征向量中不同词典特征的权重,计算得到权重调整后的所述注意力特征向量。

8.根据权利要求7所述的命名体识别装置,其特征在于,所述命名体识别模型包括输入编码层、词典特征选择层和标签预测层;

对应地,所述命名体识别单元包括:

编码子单元,用于将所述待识别文本中每个字的字向量和词典特征向量输入至所述输入编码层,得到所述输入编码层输出的每个字的隐层向量;

特征选择子单元,用于将每个字的隐层向量和词典特征向量输入至所述词典特征选择层,得到所述词典特征选择层输出的每个字的注意力特征向量;

预测子单元,用于将每个字的注意力特征向量输入至所述标签预测层,得到所述标签预测层输出的所述命名体识别结果。

9.一种电子设备,包括存储器、处理器及存储在存储器上并可在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程序时实现如权利要求1至6中任一项所述的命名体识别方法的步骤。

10.一种非暂态计算机可读存储介质,其上存储有计算机程序,其特征在于,该计算机程序被处理器执行时实现如权利要求1至6中任一项所述的命名体识别方法的步骤。

命名体识别方法、装置、电子设备和存储介质

技术领域

[0001] 本发明涉及自然语言处理技术领域,尤其涉及一种命名体识别方法、装置、电子设备和存储介质。

背景技术

[0002] 命名体识别(Named Entity Recognition,NER)是自然语言处理过程中的重要步骤,被广泛应用到信息抽取、信息检索、信息推荐以及机器翻译等任务中。所谓命名体,是指自然语言中例如人名、地名、机构名称以及歌曲名称等具有特定意义的专有名词。

[0003] 现有技术中对于命名体的识别方法通常是基于领域词典的匹配方法,将待识别文本与对应领域的领域词典进行匹配,以确定待识别文本中的命名体。然而,领域词典中的词条可能存在冲突,导致命名体识别的准确性较低。

发明内容

[0004] 本发明实施例提供一种命名体识别方法、装置、电子设备和存储介质,用以解决领域词典中的词条冲突导致命名体识别准确性低的问题。

[0005] 第一方面,本发明实施例提供一种命名体识别方法,包括:

[0006] 确定待识别文本;

[0007] 基于所述待识别文本对应领域的领域词典,确定所述待识别文本中每个字的词典特征向量;

[0008] 将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果;

[0009] 其中,所述命名体识别模型是基于样本文本中每个样本字的字向量和词典特征向量,以及每个样本字的命名体标记训练得到的。

[0010] 优选地,所述命名体识别模型包括输入编码层、词典特征选择层和标签预测层;

[0011] 对应地,所述将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果,具体包括:

[0012] 将所述待识别文本中每个字的字向量和词典特征向量输入至所述输入编码层,得到所述输入编码层输出的每个字的隐层向量;

[0013] 将每个字的隐层向量和词典特征向量输入至所述词典特征选择层,得到所述词典特征选择层输出的每个字的注意力特征向量;

[0014] 将每个字的注意力特征向量输入至所述标签预测层,得到所述标签预测层输出的所述命名体识别结果。

[0015] 优选地,所述将每个字的隐层向量和词典特征向量输入至所述词典特征选择层,得到所述词典特征选择层输出的每个字的注意力特征向量,具体包括:

[0016] 基于任一字的隐层向量和词典特征向量,确定所述任一字相对于每一词典特征的权重;

[0017] 基于所述任一字相对于每一词典特征的权重,对所述任一字的词典特征向量进行加权,得到所述任一字的注意力特征向量。

[0018] 优选地,所述将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果,之前还包括:

[0019] 基于损失函数对初始模型进行训练,得到所述命名体识别模型;

[0020] 其中,所述损失函数包括识别结果损失函数和权重损失函数,其中所述识别结果损失函数对应于所述命名体识别结果,所述权重损失函数对应于字相对于每一词典特征的权重。

[0021] 优选地,所述领域词典包括对应于不同命名体类型的词典;

[0022] 对应地,所述基于所述待识别文本对应领域的领域词典,确定所述待识别文本中每个字的词典特征向量,具体包括:

[0023] 基于对应于任一命名体类型的词典,确定所述待识别文本中每一字对应于所述任一命名体类型的词典特征;

[0024] 基于任一字对应于每一命名体类型的词典特征,确定所述任一字的词典特征向量。

[0025] 优选地,所述基于任一字对应于每一命名体类型的词典特征,确定所述任一字的词典特征向量,具体包括:

[0026] 向量化任一字对应于每一命名体类型的词典特征,得到所述任一字的特征向量;

[0027] 稀疏化处理所述任一字的特征向量,得到所述任一字的词典特征向量。

[0028] 第二方面,本发明实施例提供一种命名体识别装置,包括:

[0029] 文本确定单元,用于确定待识别文本;

[0030] 词典匹配单元,用于基于所述待识别文本对应领域的领域词典,确定所述待识别文本中每个字的词典特征向量;

[0031] 命名体识别单元,用于将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果;

[0032] 其中,所述命名体识别模型是基于样本文本中每个样本字的字向量和词典特征向量,以及每个样本字的命名体标记训练得到的。

[0033] 第三方面,本发明实施例提供一种电子设备,包括处理器、通信接口、存储器和总线,其中,处理器,通信接口,存储器通过总线完成相互间的通信,处理器可以调用存储器中的逻辑指令,以执行如第一方面所提供的方法的步骤。

[0034] 第四方面,本发明实施例提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现如第一方面所提供的方法的步骤。

[0035] 本发明实施例提供的一种命名体识别方法、装置、电子设备和存储介质,通过命名体识别模型输出命名体识别结果,能够减弱噪音词条对命名体识别的影响,从而克服基于领域词典的命名体识别方法由于词条冲突导致的识别准确性低的问题,提高命名体识别的准确性。

附图说明

[0036] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现

有技术描述中所需要使用的附图作一简单地介绍,显而易见地,下面描述中的附图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的前提下,还可以根据这些附图获得其他的附图。

[0037] 图1为本发明实施例提供的命名体识别方法的流程示意图;

[0038] 图2为本发明实施例提供的命名体识别模型的运行流程示意图;

[0039] 图3为本发明实施例提供的注意力特征向量的计算方法的流程示意图;

[0040] 图4为本发明实施例提供的词典特征向量的确定方法的流程示意图;

[0041] 图5为本发明实施例提供的词典特征生成示意图;

[0042] 图6为本发明实施例提供的词典特征向量的稀疏化示意图;

[0043] 图7为本发明实施例提供的命名体识别模型的训练方法的流程示意图;

[0044] 图8为本发明实施例提供的命名体识别模型结构示意图;

[0045] 图9为本发明实施例提供的命名体识别装置的结构示意图;

[0046] 图10为本发明实施例提供的电子设备的结构示意图。

具体实施方式

[0047] 为使本发明实施例的目的、技术方案和优点更加清楚,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。

[0048] 语音助手是一款智能型的手机应用,通过智能对话与即时问答的智能交互,辅助用户进行相关操作。语义理解作为其中一个重要环节,起到让机器听懂、理解用户的查询的作用。各专有领域的语义理解一般包含两个任务:意图识别和NER识别。在工业界,虽然早期的专有领域的NER系统取得了巨大的应用,但是需要大量的人力物力去编写规则和发现特征。近年来,随着深度学习的热潮,将深度学习应用到NER任务中也取得了巨大的进步。但是,通用的NER方法均需要大量的有监督训练数据,而有监督训练数据极其昂贵,需要消耗大量的人力物力。

[0049] 专有领域的NER一个大的困难点在于训练数据稀少且难以获得。为了解决这个问题,如何利用外部知识开始得到广泛的研究。这些方法大概可以分为两类:

[0050] 基于语言模型:使用词级别的语言模型增强NER模型,或者使用字级别的语言模型增强NER模型。由于专有领域的语料在通用场景出现次数较少,所以基于通用语料训练的语言模型解决能力较为有限。

[0051] 基于领域词典模型:例如巧妙地应用词典特征的Lattice LSTM,或者基于领域词典的无监督NER模型。然而上述方法均忽略了领域词典之间的冲突,没有考虑噪音词条对于模型的影响。

[0052] 对此,本发明实施例提供了一种命名体识别方法,该方法可以用于电视点播领域,以根据用户的语音数据确定用户需要观看的演员或影片名称,进而为用户播放相应的视频,还可以应用于智能家居控制、语音助手的信息检索、信息推荐等领域,本发明实施例对此不作具体限定。

[0053] 图1为本发明实施例提供的命名体识别方法的流程示意图,如图1所示,该方法包

括:

[0054] 步骤110,确定待识别文本。

[0055] 此处,待识别文本即需要进行命名体识别的文本,待识别文本可以是用户直接输入的文本,也可以是通过对用户输入的语音数据进行语音识别后得到的文本,还可以应用OCR(Optical Character Recognition,光学字符识别)等文字识别技术对用户输入的图像进行文字识别得到的,本发明实施例对此不作具体限定。

[0056] 步骤120,基于待识别文本对应领域的领域词典,确定待识别文本中每个字的词典特征向量。

[0057] 具体地,领域词典对应于待识别文本的领域,例如在电视点播场景下,待识别文本用于进行视频点播,领域词典对应于视频领域,领域词典中包含的词条可以包括演员名称的词条,影片名称的词条以及影片类型的词条等。领域词典是基于对应领域的词条预先构建的,例如对应于视频领域,可以通过爬虫从视频网站中获取视频领域的词条,从而构建领域词典。

[0058] 在确定待识别文本后,可以将待识别文本与对应领域的领域词典进行匹配,进而得到待识别文本中每个字的词典特征向量。此处,针对任一字,该字的词典特征向量是该字的词典特征对应的向量,该字的词典特征可以用于表征该字是否命中领域词典中的词条,且在该字命中领域词典中的词条时,词典特征还可以用于表征该字在命中的词条中的位置,或者该字命中的词条的类型等,本发明实施例对此不作具体限定。

[0059] 步骤130,将待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到命名体识别模型输出的命名体识别结果;

[0060] 其中,命名体识别模型是基于样本文本中每个样本字的字向量和词典特征向量,以及每个样本字的命名体标记训练得到的。

[0061] 具体地,基于领域词典进行匹配时,任一字可能未命中任何词条,也可能命中一个词条,还有可能命中多个词条。在命中多个词条时,该字实际对应的可能只是多个词条中的一个,也就是多个词条中只有一个是正确命中的词条。例如待识别文本为“我想看电影无间道”,在领域词典中对应命中词条“无间道”和词条“道”,因此,对于“道”字而言,命中有两个词条,两个词条中,“无间道”为正确命中的词条,“道”为噪音词条。这一问题通常是由于领域词典内部的词条冲突导致的。

[0062] 针对这一问题,本发明实施例中采用命名体识别模型,通过命名体识别模型确定并输出命名体识别结果。

[0063] 在命中多个词条时,该字存在多个词典特征,对应地该字的词典特征向量为包含多个词典特征的向量。命名体识别模型能够基于该字的语义,确定不同词典特征与该字语义之间的相关性,进而从多个词典特征中选取出该字语义更为接近的词典特征,提高命名体识别精度。

[0064] 由此得到的命名体识别结果用于指示待识别文本中的命名体,还可以用于指示待识别文本中命名体的类型等,本发明实施例对此不作具体限定。

[0065] 在执行步骤130之前,还可以预先训练得到命名体识别模型,具体可以通过如下方式训练得到命名体识别模型:首先,收集大量样本文本,并基于样本文本对应领域的领域词典确定样本文本中每个样本字的词典特征向量。此外,通过人工标注的方式,确定样本文本

中每个样本字的命名体标记,此处命名体标记可以表征该样本字是否为命名体,还可以表征该样本字在命名体中的位置,或者该样本字对应命名体的类型等。随即基于样本文本中每个样本字的字向量、词典特征向量以及命名体标记对初始模型进行训练,从而得到命名体识别模型。

[0066] 本发明实施例提供的方法,通过命名体识别模型输出命名体识别结果,能够减弱噪音词条对命名体识别的影响,从而克服基于领域词典的命名体识别方法由于词条冲突导致的识别准确性低的问题,提高命名体识别的准确性。

[0067] 基于上述实施例,该方法中,命名体识别模型基于注意力机制确定待识别文本中每个字的词典特征向量所对应的注意力特征向量,并基于每个字的注意力特征向量输出命名体识别结果。

[0068] 具体地,在命中多个词条时,该字存在多个词典特征,对应地该字的词典特征向量为包含多个词典特征的向量。命名体识别模型中,注意力机制能够基于该字的语义,确定不同词典特征与该字语义之间的相关性,进而调整词典特征向量中不同词典特征的权重,并计算得到权重调整后的注意力特征向量,注意力特征向量在指示该字对应命中多个词条时,相对于不同词典特征的权重存在偏重,从而更好地突出与该字语义更为接近的词典特征,提高命名体识别精度。

[0069] 基于上述任一实施例,该方法中,命名体识别模型包括输入编码层、词典特征选择层和标签预测层。图2为本发明实施例提供的命名体识别模型的运行流程示意图,如图2所示,步骤130具体包括:

[0070] 步骤131,将待识别文本中每个字的字向量和词典特征向量输入至输入编码层,得到输入编码层输出的每个字的隐层向量。

[0071] 具体地,输入编码层用于基于输入的每个字的字向量和词典特征向量,分析对应字在待识别文本中的语义,并输出每个字的隐层向量。此处,隐层向量是对针对待识别文本中每个字的字向量和词典特征向量进行编码得到的。输入编码层可以是用于编码的双向长短时记忆网络BI-LSTM。

[0072] 步骤132,将每个字的隐层向量和词典特征向量输入至词典特征选择层,得到词典特征选择层输出的每个字的注意力特征向量。

[0073] 具体地,词典特征选择层用于基于输入的任一字的隐层向量和词典特征向量,通过注意力机制计算词典特征向量中不同词典特征与隐层向量之间的相关性,进而调整词典特征向量中不同词典特征的权重,输出调整后的注意力特征向量。

[0074] 步骤133,将每个字的注意力特征向量输入至标签预测层,得到标签预测层输出的命名体识别结果。

[0075] 具体地,标签预测层用于基于输入的每一字的注意力特征向量,分析每一字对应于不同词典特征的概率,进而得到每一实际词对应的词典特征,输出命名体识别结果。此处,标签预测层可以是双向长短时记忆网络BI-LSTM+条件随机场CRF的结构。

[0076] 本发明实施例提供的方法,结合隐层向量和词典特征向量进行注意力特征向量的计算,有助于减弱噪音词条对命名体识别的影响,提高命名体识别的准确性。

[0077] 基于上述任一实施例,图3为本发明实施例提供的注意力特征向量的计算方法的流程示意图,如图3所示,该方法中,步骤132具体包括:

[0078] 步骤1321,基于任一字的隐层向量和词典特征向量,确定该字相对于每一词典特征的权重。

[0079] 具体地,词典特征向量中包含有多个词典特征对应的向量,基于隐层向量以及词典特征向量中任一词典特征对应的向量,基于注意力机制可以得到隐层向量与该词典特征对应的向量之间的相关性。在得到隐层向量与词典特征向量中每一词典向量对应的向量之间的相关性后,即可确定每一词典向量的权重。此处,相关性越高,则对应权重越大,相关性越低,则对应权重越小。

[0080] 步骤1322,基于该字相对于每一词典特征的权重,对该字的词典特征向量进行加权,得到该字的注意力特征向量。

[0081] 具体地,在得到该字相对于每一词典特征的权重后,即可对该字的词典特征向量中包含的每一词典特征对应的向量进行加权,并将每一词典特征对应的向量加权之和作为该字的注意力特征向量。

[0082] 基于上述任一实施例,该方法中,步骤1321具体包括:

[0083] 基于如下公式,确定任一字相对于每一词典特征的权重:

[0084] $$\mathrm{weigth_i} = \delta\left(\mathrm{w} * \mathrm{h_i^T} * \mathrm{f_i'}\right);$$

[0085] 式中,$\mathrm{weigth_i}$为待识别文本中的第i个字相对于每一词典特征的权重,i为正整数;$\mathrm{h_i}$和$\mathrm{f'_i}$分别为第i个字的隐层向量和词典特征向量;w用于将$\mathrm{h_i}$映射至与$\mathrm{f'_i}$相同的维度,δ()为狄拉克函数。

[0086] 由上式可知,针对任一字,通过将该字的隐层向量和词典特征向量相乘,即可得到该字的语义与每一词典特征的相关性,在此基础上通过狄拉克函数对该字的语音与每一词典特征的相关性进行处理,即可得到该字相对于每一词典特征的权重。此处,该字相对于每一词典特征的权重之和为1。

[0087] 步骤1322具体包括:

[0088] 基于如下公式,计算第i个字的注意力特征向量:

[0089] $$\mathrm{f_i} = \sum_{\mathrm{i}}^{\mathrm{M}} \mathrm{weigth_i} * \mathrm{f_i'};$$

[0090] 式中,$\mathrm{f_i}$为待识别文本中的第i个字的注意力特征向量,M为预先设定的词典特征的总数量。

[0091] 由上式可知,针对任一字,该字的注意力特征向量即该字相对于每一词典特征的权重与该字的词典特征向量中对应词典特征的向量相乘后叠加得到的,即注意力特征向量为根据权重对每一词典特征的向量进行加权求和的结果。

[0092] 基于上述任一实施例,该方法中,步骤130之前还包括:基于损失函数对初始模型进行训练,得到命名体识别模型;其中,损失函数包括识别结果损失函数和权重损失函数,其中识别结果损失函数对应于命名体识别结果,权重损失函数对应于字相对于每一词典特征的权重。

[0093] 具体地,在命名体识别模型的训练过程中,由于样本字的词典特征向量中的词典特征与样本字的命名体标记之间存在相关性,为了使得词典特征的权重向命名体标记的方向学习,可以利用这种相关性,去监督模型中词典特征权重的学习,由此定义权重损失函

数。本发明实施例中,可以将识别结果损失函数与权重损失函数两者的和作为损失函数进行模型训练。

[0094] 进一步地,定义权重损失函数如下:

[0095] $$loss_{weight} = -\frac{1}{M}\sum_{i=1}^{M} y_i \log(weight_i)$$

[0096] 上式中,M为词典特征向量所表征的词典特征的个数。y_i表示第i个样本字的词典特征向量所表征的每一词典特征与该字的命名体标记是否相同,若相同则y_i中对应词典特征的位置设置为1,不相同则y_i中对应词典特征的位置设置为0。例如,样本文本“我想看刘德华演的无间道”中,样本字“德”的词典特征向量所表征的词典特征为“I_artist”,“德”的命名体标记为“I_artist”,两者相同,所以样本字“德”的词典特征I_artist对应的y取值为1,反之样本字“德”的另一词典特征“E_artist”与命名体标记“I_artist”不相同,故样本字“德”的词典特征E_artist对应的y取值0;$weight_i$则为第i个字相对于每一词典特征的权重。所以在模型训练阶段,为了让损失函数$loss_{weight}$往小的方向趋近,当y_i中当前词典对应的y取值为1时,即当前词典特征和命名体标记相同时,其对应的权重$weight_i$应往大的方向逼近。

[0097] 假设识别结果损失函数为:

[0098] $$loss_{slot} = \sum_{i=1}^{N} \log(p(y_i|s_i))$$

[0099] 式中,N为样本文本的字数,$p(y_i|s_i)$用于表示第i个样本字基于模型得到的预测结果y_i与命名体标记s_i相同的概率。由此得到对初始模型进行训练的损失函数为:

[0100] $$loss = loss_{slot} + \alpha\sum_{i=0}^{N} loss_{weight}$$

[0101] 其中,α为预设参数,用于调节$loss_{weight}$的权重,作为优选,α为0.25。

[0102] 目前基于领域词典的命名体识别方法中,仅考虑是否命中词条,而忽略了领域词典的类型,例如音乐垂直领域中,“刘德华”和“忘情水”都是专有领域词,但是“刘德华”是歌手名称,“忘情水”是歌曲名称,两者的类型并不相同,不同类型的词条带给模型的信息显然也不相同。基于上述任一实施例,该方法中,领域词典包括对应于不同命名体类型的词典。

[0103] 此处,领域词典内,不同命名体类型的词条存储在对应命名体类型的词典中,例如,在电视点播领域,可以按照命名体类型划分出演员词典、电影名称词典、电影类型词典等,其中演员词典中包含词条刘德华、梁朝伟、曾志伟等,电影名称词典中包含词条无间道等,电影类型词典包含词条犯罪、警匪等。

[0104] 图4为本发明实施例提供的词典特征向量的确定方法的流程示意图,如图4所示,步骤120具体包括:

[0105] 步骤121,基于对应于任一命名体类型的词典,确定待识别文本中每一字对应于该命名体类型的词典特征。

[0106] 此处,任一字对应于任一命名体类型的词典特征,用于表征该字是否命中该命令

体类型的词条,且在该字命中该命名体类型词条时,还可以用于表征该字在命中的词条中的位置。

[0107] 步骤122,基于任一字对应于每一命名体类型的词典特征,确定该字的词典特征向量。

[0108] 具体得,在得到任一字相对于每一命名体类型的词典特征后,即可对该字相对于每一命名体类型的词典特征进行向量化和拼接,得到该字的词典特征向量。

[0109] 本发明实施例提供的方法,词典特征向量中包含有命中词条的命名体类型,为命名体识别提供了更为丰富的信息,有助于提高命名体识别的准确性。

[0110] 基于上述任一实施例,该方法中,步骤122具体包括:向量化任一字对应于每一命名体类型的词典特征,得到该字的特征向量;稀疏化处理该字的特征向量,得到该字的词典特征向量。

[0111] 具体地,在得到任一字相对于每一命名体类型的词典特征后,向量化词典特征,得到用0或1表示的向量化的词典特征,随即对向量化的相对于每一命名体类型的词典特征进行拼接,得到该字的特征向量。

[0112] 图5为本发明实施例提供的词典特征生成示意图,如图5所示,待识别文本为“我要看刘德华饰演的无间道”,基于领域词典,确定待识别文本命中演员词典中的词条“刘德华”和“刘德”,以及电影名称词典中的词条“无间道”和“道”。由此得到“刘”字对应的词典特征为B_artist,表示“刘”为一个演员名的开始字符;“德”字对应的词典特征为I_artist和E_artist,分别表示“德”为一个演员名的中间字符,以及“德”为一个演员名的结尾字符;“华”对应的词典特征为E_artist,表示“华”为一个演员名的结尾字符。“无”字对应的词典特征为B_name,表示“无”为一个电影名称的开始字符;“间”字对应的词典特征为I_name,表示“间”为一个电影名称的中间字符;“道”字对应的词典特征为E_name和S_name,分别表示“道”为一个电影名称的结尾字符,以及“道”为一个电影名称的单个字符。

[0113] 假设词典特征的数量为M,针对任一字,初始化一个长度为M的全0向量,然后在该字对应的词典特征处设置1。针对“德”字,对其相对于每一命名体类型的词典特征进行向量化,得到该字的特征向量如下表所示:

[0114]

词典特征列表	B_artist	I_artist	E_artist	S_artist	……	……	S_name
词典特征索引位置	0	1	2	3	……	……	M-1
“德”的特征向量	0	1	1	0	0	0	0

[0115] 由此得到的特征向量中仅包含有0和1,其信息量过强,因此需要对特征向量进行稀疏化处理,并将稀疏化处理后的特征向量作为词典特征向量应用于命名体识别。

[0116] 进一步地,特征向量的稀疏化处理可以通过如下步骤实现:

[0117] 初始化一个形状为[M,$label_{dim}$]的矩阵,其中M为词典特征的数量,$label_{dim}$为词典特征向量的维度,本发明实施例将该矩阵记为L。

[0118] 图6为本发明实施例提供的词典特征向量的稀疏化示意图,如图6所示,将特征向量与矩阵L相乘,即可得到稀疏化的词典特征向量。图6中,“德”字的词典特征向量为[M,$label_{dim}$]的矩阵,其中,词典特征I_artist和E_artist在矩阵中对应的列为填充有斜线的圆圈,表示这两列有值,其余空白的圆圈均为0。

[0119] 在此基础上,还可以将图6中的矩阵平展为一维列向量,以便于与该字对应的字向

量进行拼接,本发明实施例对此不作具体限定。

[0120] 基于上述任一实施例,图7为本发明实施例提供的命名体识别模型的训练方法的流程示意图,如图7所示,该方法包括:

[0121] 首先,建立领域词典。

[0122] 其次,预先收集大量待识别的样本文本,作为命名体识别模型的训练样本数据。

[0123] 然后,对训练样本数据中每个待识别的样本文本进行领域词典特征匹配,得到样本文本的词典特征。通常,词典特征匹配是通过字符串匹配的方法查询领域词典中所有可能存在的词条,假设用样本文本为“我想看刘德华演的无间道”,则通过领域词典的字符串匹配后,可命中2个词条:演员词典中的刘德华、电影词典中的无间道。

[0124] 接着,基于样本文本,以及样本文本的词典特征,确定样本文本中每个样本字的字向量,以及每个样本字的词典特征向量,同时标注每个字的命名体标记。

[0125] 最后,根据样本文本中每个样本字的字向量、词典特征向量以及命名体标记,构建并训练命名体识别模型。

[0126] 基于上述任一实施例,图8为本发明实施例提供的命名体识别模型结构示意图,如图8所示,命名体识别模型包括输入编码层、词典特征选择层、标签预测层。其中,输入编码层输入每个字的字向量和词典特征向量,即特征向量的稀疏向量;输入编码层中包含为一个BI-LSTM网络,并每个字的字向量和词典特征向量进行编码并输出每个字的隐层向量;词典特征选择层,用于基于任一字的隐层向量和词典特征向量,通过注意力机制计算词典特征向量中不同词典特征与隐层向量之间的相关性,调整词典特征向量中不同词典特征的权重,从而选择重要的词典特征,降低噪音词典特征的影响;标签预测层输出每个字对应于每种词典特征的概率,选择概率值最大的词典特征作为当前字的命名体识别结果。

[0127] 基于上述任一实施例,图9为本发明实施例提供的命名体识别装置的结构示意图,如图9所示,该装置包括文本确定单元910、词典匹配单元920和命名体识别单元930;

[0128] 其中,文本确定单元910用于确定待识别文本;

[0129] 词典匹配单元920用于基于所述待识别文本对应领域的领域词典,确定所述待识别文本中每个字的词典特征向量;

[0130] 命名体识别单元930用于将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果;

[0131] 其中,所述命名体识别模型是基于样本文本中每个样本字的字向量和词典特征向量,以及每个样本字的命名体标记训练得到的。

[0132] 本发明实施例提供的装置,通过命名体识别模型输出命名体识别结果,能够减弱噪音词条对命名体识别的影响,从而克服基于领域词典的命名体识别方法由于词条冲突导致的识别准确性低的问题,提高命名体识别的准确性。

[0133] 基于上述任一实施例,该装置中,所述命名体识别模型包括输入编码层、词典特征选择层和标签预测层;

[0134] 对应地,所述命名体识别单元930包括:

[0135] 编码子单元,用于将所述待识别文本中每个字的字向量和词典特征向量输入至所述输入编码层,得到所述输入编码层输出的每个字的隐层向量;

[0136] 特征选择子单元,用于将每个字的隐层向量和词典特征向量输入至所述词典特征

选择层,得到所述词典特征选择层输出的每个字的注意力特征向量;

[0137] 预测子单元,用于将每个字的注意力特征向量输入至所述标签预测层,得到所述标签预测层输出的所述命名体识别结果。

[0138] 基于上述任一实施例,该装置中,特征选择子单元包括:

[0139] 权重调节模块,用于基于任一字的隐层向量和词典特征向量,确定所述任一字相对于每一词典特征的权重;

[0140] 加权模块,用于基于所述任一字相对于每一词典特征的权重,对所述任一字的词典特征向量进行加权,得到所述任一字的注意力特征向量。

[0141] 基于上述任一实施例,该装置还包括:

[0142] 模型训练单元,用于基于损失函数对初始模型进行训练,得到所述命名体识别模型;

[0143] 其中,所述损失函数包括识别结果损失函数和权重损失函数,其中所述识别结果损失函数对应于所述命名体识别结果,所述权重损失函数对应于字相对于每一词典特征的权重。

[0144] 基于上述任一实施例,该装置中,所述领域词典包括对应于不同命名体类型的词典;

[0145] 对应地,所述词典匹配单元920包括:

[0146] 词典特征确定子单元,用于基于对应于任一命名体类型的词典,确定所述待识别文本中每一字对应于所述任一命名体类型的词典特征;

[0147] 词典向量确定子单元,用于基于任一字对应于每一命名体类型的词典特征,确定所述任一字的词典特征向量。

[0148] 基于上述任一实施例,该装置中,所述词典向量确定子单元具体用于:

[0149] 向量化任一字对应于每一命名体类型的词典特征,得到所述任一字的特征向量;

[0150] 稀疏化处理所述任一字的特征向量,得到所述任一字的词典特征向量。

[0151] 图10为本发明实施例提供的电子设备的结构示意图,如图10所示,该电子设备可以包括:处理器(processor)1010、通信接口(Communications Interface)1020、存储器(memory)1030和通信总线1040,其中,处理器1010,通信接口1020,存储器1030通过通信总线1040完成相互间的通信。处理器1010可以调用存储器1030中的逻辑指令,以执行如下方法:确定待识别文本;基于所述待识别文本对应领域的领域词典,确定所述待识别文本中每个字的词典特征向量;将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果;其中,所述命名体识别模型是基于样本文本中每个样本字的字向量和词典特征向量,以及每个样本字的命名体标记训练得到的。

[0152] 此外,上述的存储器1030中的逻辑指令可以通过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。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案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中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,

Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0153] 本发明实施例还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现以执行上述各实施例提供的方法,例如包括:确定待识别文本;基于所述待识别文本对应领域的领域词典,确定所述待识别文本中每个字的词典特征向量;将所述待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中,得到所述命名体识别模型输出的命名体识别结果;其中,所述命名体识别模型是基于样本文本中每个样本字的字向量和词典特征向量,以及每个样本字的命名体标记训练得到的。

[0154] 以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元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。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0155] 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各实施方式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。基于这样的理解,上述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,该计算机软件产品可以存储在计算机可读存储介质中,如ROM/RAM、磁碟、光盘等,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0156] 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,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例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同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

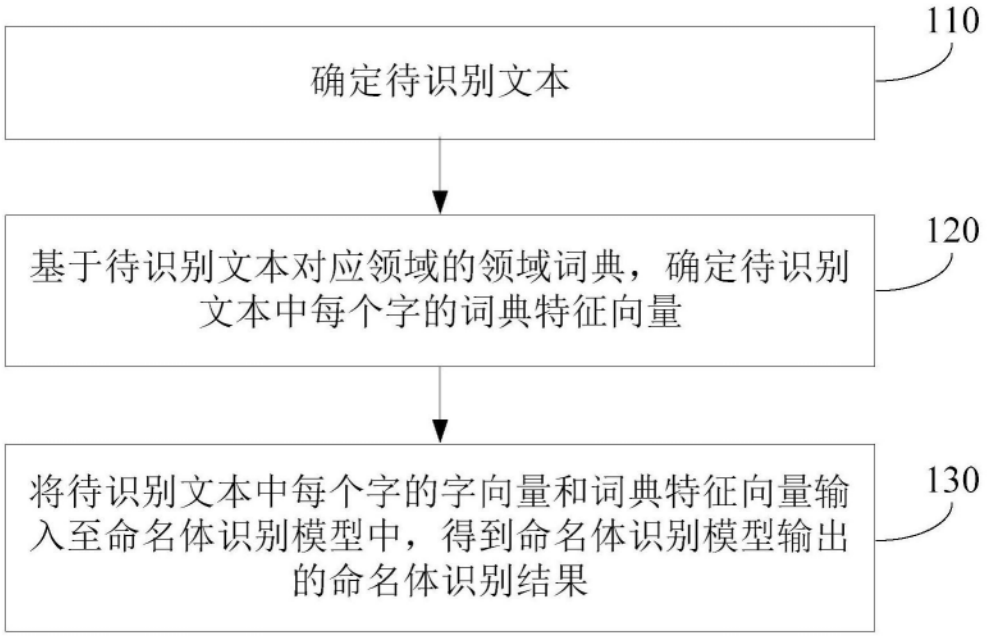

110
确定待识别文本
120
基于待识别文本对应领域的领域词典，确定待识别文本中每个字的词典特征向量
130
将待识别文本中每个字的字向量和词典特征向量输入至命名体识别模型中，得到命名体识别模型输出的命名体识别结果

图1

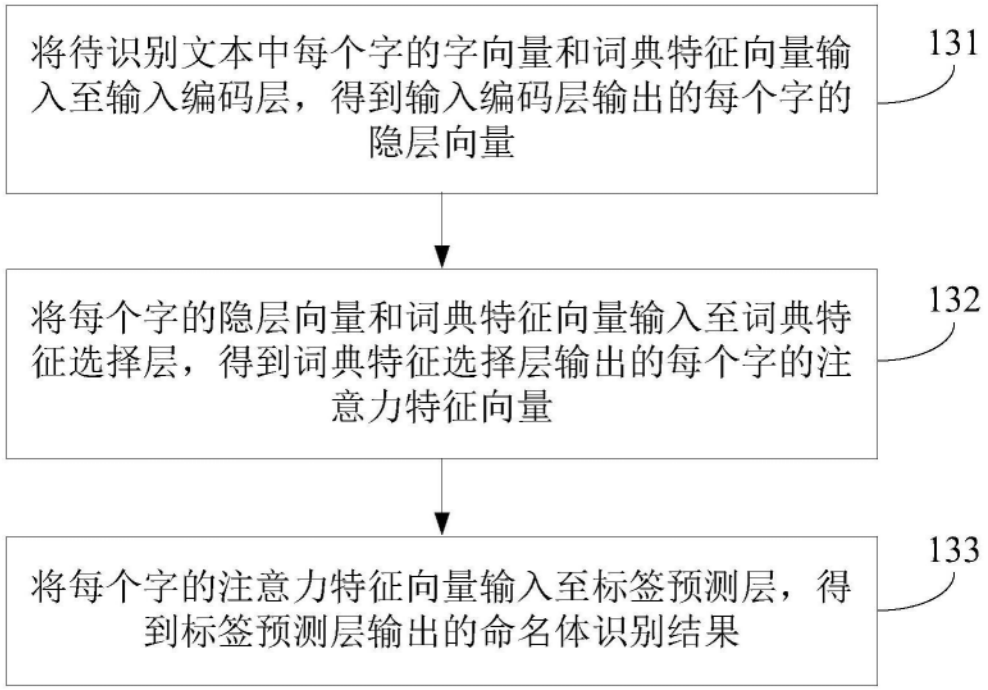

131
将待识别文本中每个字的字向量和词典特征向量输入至输入编码层，得到输入编码层输出的每个字的隐层向量
132
将每个字的隐层向量和词典特征向量输入至词典特征选择层，得到词典特征选择层输出的每个字的注意力特征向量
133
将每个字的注意力特征向量输入至标签预测层，得到标签预测层输出的命名体识别结果

图2

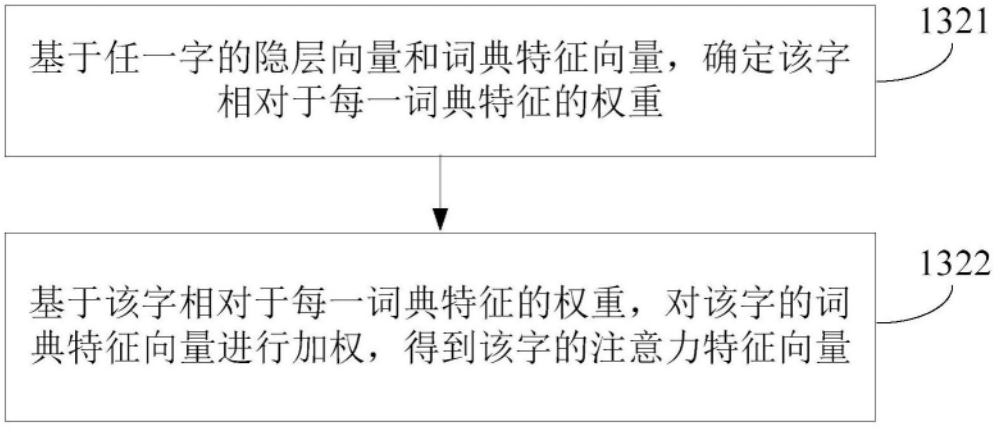

基于任一字的隐层向量和词典特征向量，确定该字相对于每一词典特征的权重
1321
基于该字相对于每一词典特征的权重，对该字的词典特征向量进行加权，得到该字的注意力特征向量
1322

图3

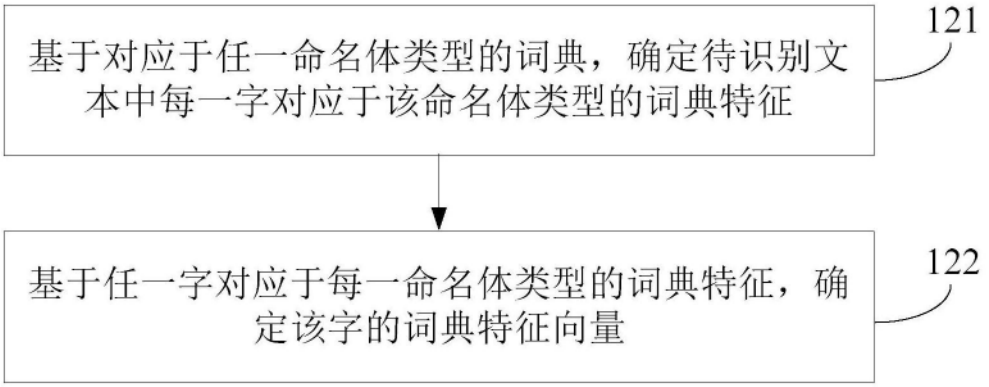

基于对应于任一命名体类型的词典，确定待识别文本中每一字对应于该命名体类型的词典特征
121
基于任一字对应于每一命名体类型的词典特征，确定该字的词典特征向量
122

图4

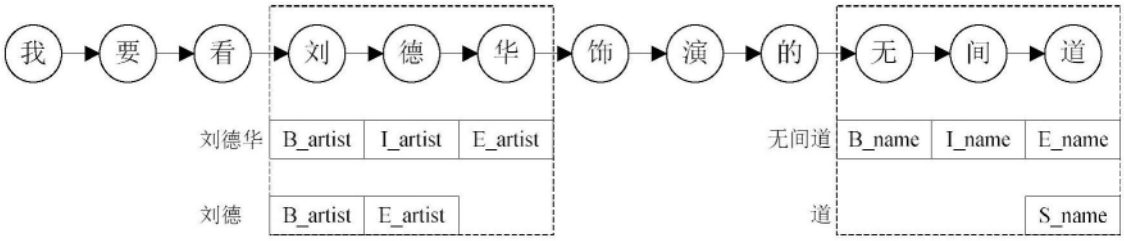

我
要
看
刘
德
华
饰
演
的
无
间
道
刘德华
B_artist
I_artist
E_artist
刘德
B_artist
E_artist
无间道
B_name
I_name
E_name
道
S_name

图5

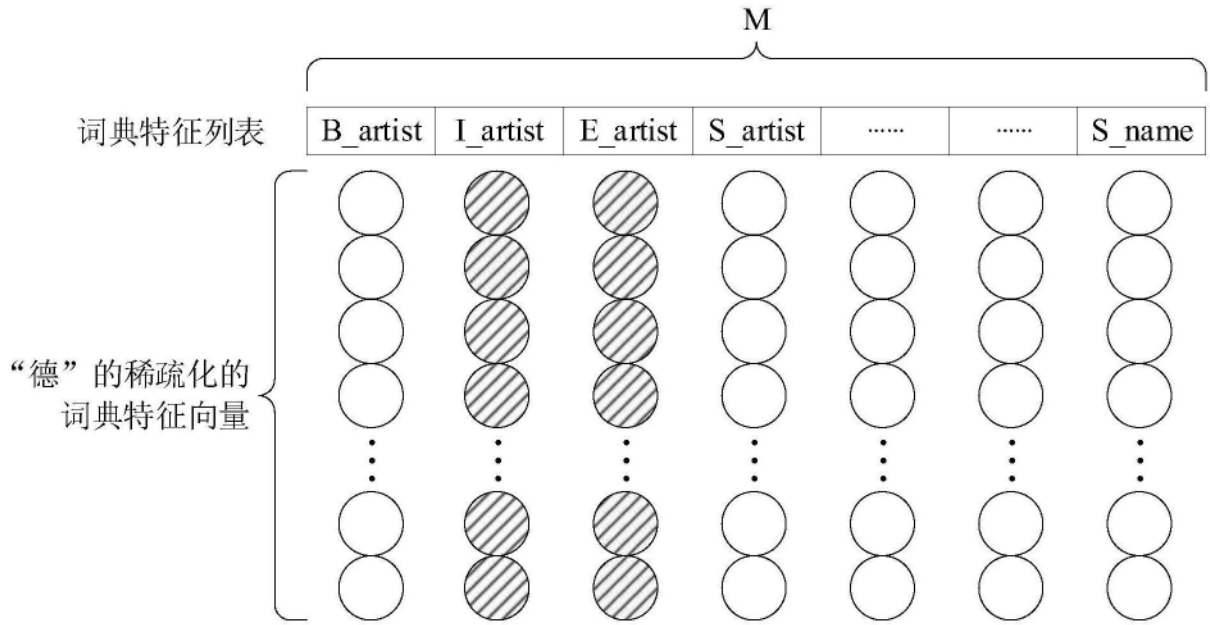

M
词典特征列表
B_artist
I_artist
E_artist
S_artist
.....
.....
S_name
“德”的稀疏化的
词典特征向量

图6

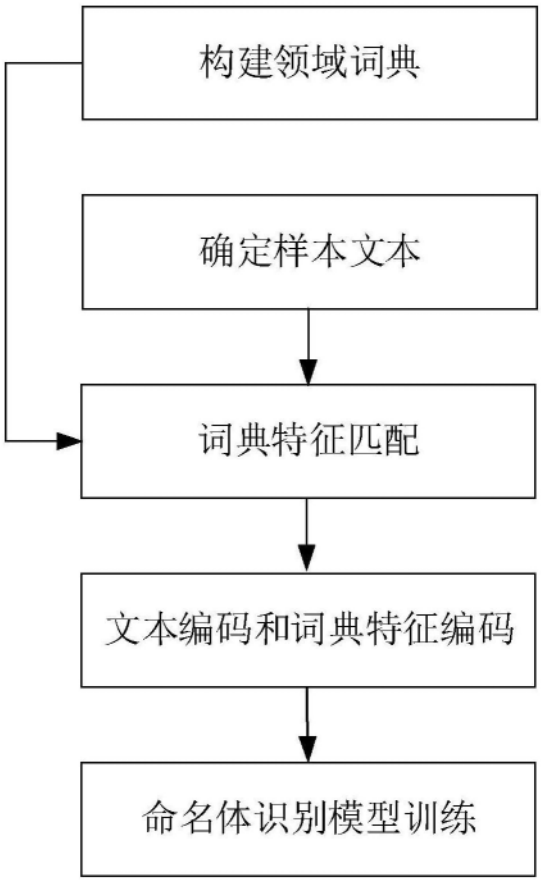

构建领域词典
确定样本文本
词典特征匹配
文本编码和词典特征编码
命名体识别模型训练

图7

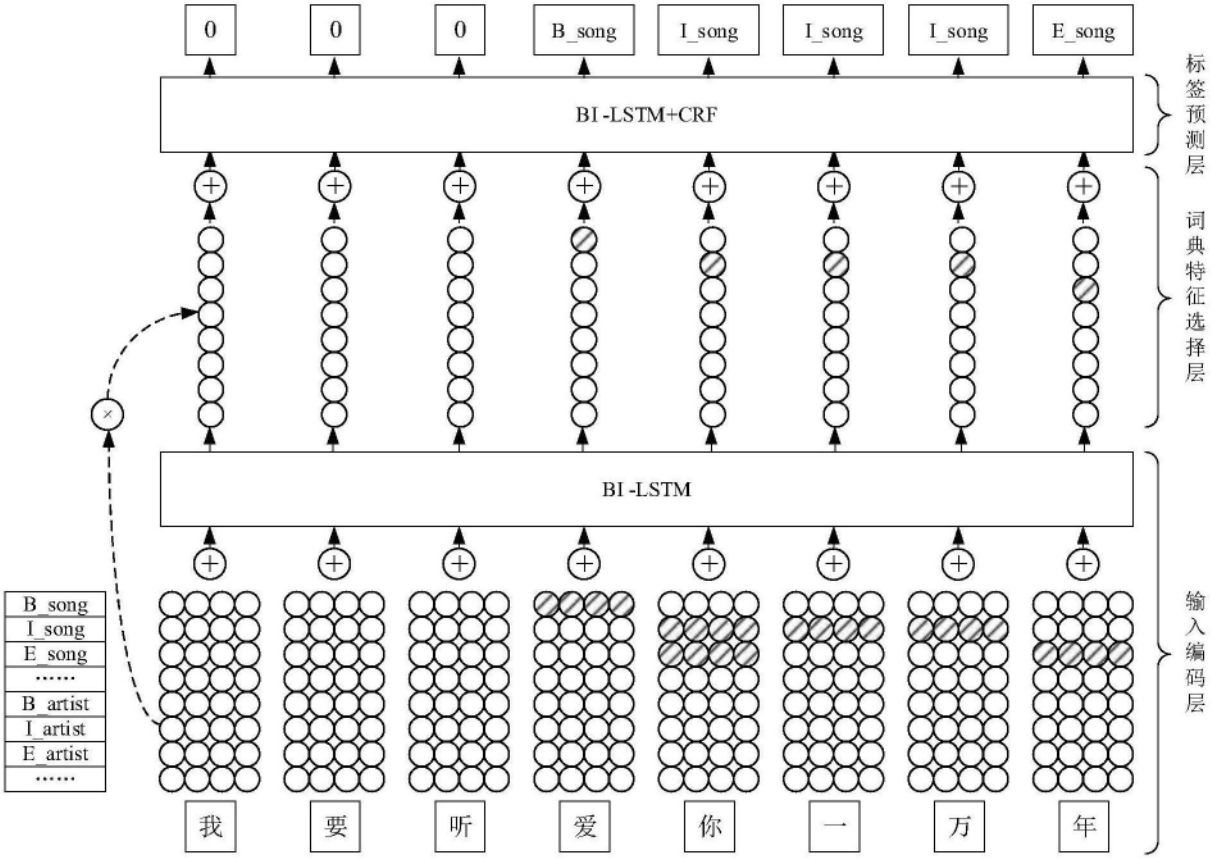

0
0
0
B_song
I_song
I_song
I_song
E_song
BI-LSTM+CRF
标签预测层
词典特征选择层
BI-LSTM
输入编码层
B_song
I_song
E_song
.....
B_artist
I_artist
E_artist
.....
我
要
听
爱
你
一
万
年

图8

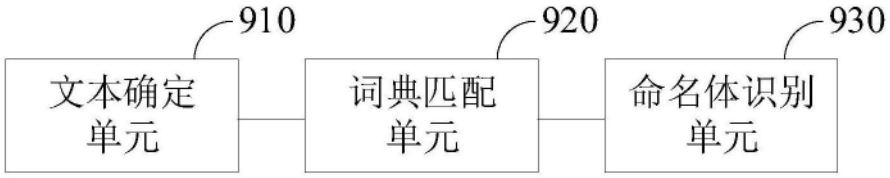

910
920
930
文本确定单元
词典匹配单元
命名体识别单元

图9

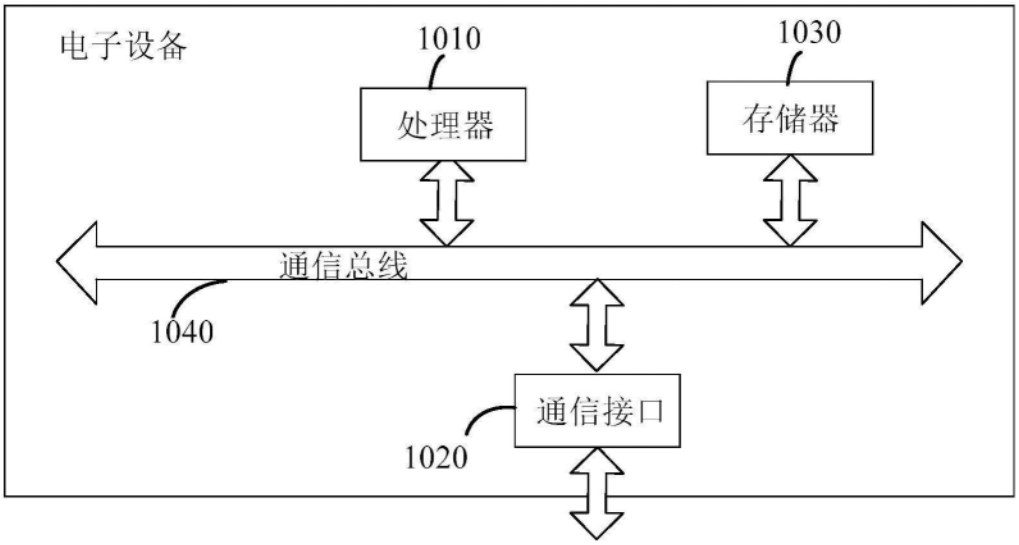


图10