(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109726397 B
(45)授权公告日 2024.02.02

(21)申请号 201811614094.6
(22)申请日 2018.12.27
(65)同一申请的已公布的文献号
申请公布号 CN 109726397 A
(43)申请公布日 2019.05.07
(73)专利权人 网易(杭州)网络有限公司
地址 310052 浙江省杭州市滨江区长河街道网商路599号4幢7层
(72)发明人 吴庆洲
(74)专利代理机构 北京律智知识产权代理有限公司 11438
专利代理师 袁礼君 阚梓瑄
(51)Int.Cl.
G06F 40/295(2020.01)

(56)对比文件
CN 107797989 A,2018.03.13
CN 108304376 A,2018.07.20
CN 108536679 A,2018.09.14
US 2018357225 A1,2018.12.13
WO 2018023981 A1,2018.02.08
US 2018329886 A1,2018.11.15

审查员 朱为琦

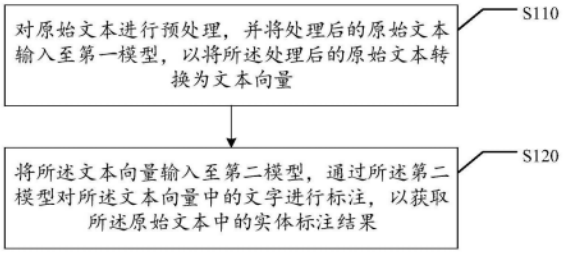
权利要求书3页 说明书10页 附图5页

(54)发明名称

中文命名实体的标注方法、装置、存储介质和电子设备

(57)摘要

本公开涉及一种中文命名实体的标注方法及装置、存储介质和电子设备。该中文命名实体的标注方法包括:对原始文本进行预处理,并将处理后的原始文本输入至第一模型,以将所述处理后的原始文本转换为文本向量;将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果。本公开通过将原始文本输入至第一模型后得到的文本向量,接着输入至第二模型,能准确地对原始文本进行实体的标注。



CN 109726397 B

1.一种中文命名实体的标注方法,其特征在于,包括:

对原始文本中的文字用空格进行分隔,并将空格分隔后的文字进行部首拆分,以获取处理后的原始文本;

将所述处理后的原始文本输入至第一模型,以将所述处理后的原始文本转换为文本向量;所述第一模型为根据训练数据对一语言模型进行训练获得的,所述训练数据包括第一文本样本、第二文本样本、第一文本样本对应的第一文本向量样本、第二文本样本对应的第二文本向量样本,所述第二文本向量样本为所述第二文本样本经预处理后的部首样本进行向量化处理后得到的;其中,所述语言模型为双向长短期记忆模型,在根据所述训练数据对所述语音模型进行训练的过程中,利用所述语言模型提取所述训练数据中每个文字的最后一个部首的输出隐状态,以将正向输出的隐状态序列和反向输出的隐状态序列进行拼接得到完整的隐状态序列;

将所述文本向量输入至第二模型,以获取所述文本向量中的文字之间的关联信息,根据所述关联信息对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果。

2.根据权利要求1所述的中文命名实体的标注方法,其特征在于,在所述将用空格分隔后的文字进行部首的拆分,以获取所述处理后的原始文本之前,所述方法还包括:

通过预设程序获取目标文本库中的文字,并对所述目标文本库中的文字进行部首拆分以获取与所述目标文本库中的文字对应的部首;

根据所述目标文本库中的文字和对应的所述部首,形成部首词典。

3.根据权利要求2所述的中文命名实体的标注方法,其特征在于,所述将用空格分隔后的文字进行部首的拆分,以获取所述处理后的原始文本,包括:

基于所述部首词典,对所述原始文本中的文字进行部首拆分,以获取所述处理后的原始文本。

4.根据权利要求1所述的中文命名实体的标注方法,其特征在于,在所述对原始文本进行预处理,并将处理后的原始文本输入至第一模型,以将所述处理后的原始文本转换为文本向量之前,所述方法还包括:

获取所述训练数据;

根据所述训练数据对一语言模型进行训练,以获取所述第一模型。

5.根据权利要求4所述的中文命名实体的标注方法,其特征在于,

所述获取所述训练数据,包括:

通过向量模型对所述第一文本样本进行向量化处理,以获取所述第一文本向量样本;

将所述第二文本样本中的文字用空格进行分隔,并将用空格分隔后的文字进行部首拆分,以获取部首样本;

通过预训练语言模型对所述部首样本进行向量化处理,以获取所述第二文本向量样本;

根据所述第一文本样本、所述第一文本向量样本、所述第二文本样本和所述第二文本向量样本,确定所述训练数据。

6.根据权利要求1所述的中文命名实体的标注方法,其特征在于,所述第二模型包括双向神经网络子模型和条件随机场子模型。

7.根据权利要求6所述的中文命名实体的标注方法,其特征在于,所述将所述文本向量输入至所述第二模型,以获取所述文本向量中的文字之间的关联信息,包括:

将所述文本向量输入至所述双向神经网络子模型,以通过所述双向神经网络子模型将所述文本向量的双向隐藏状态进行拼接,以获取所述文本向量中的文字之间的关联信息。

8.根据权利要求7所述的中文命名实体的标注方法,其特征在于,所述根据所述关联信息对所述文本向量中的文字进行标注,以获取所述原始文本的实体标注结果,包括:

基于所述文本向量中的文字之间的关联信息,将由所述双向神经网络模型输出的文本输入至所述条件随机场模型,以对所述文本向量中的文字进行标注,得到所述原始文本的实体标注结果。

9.根据权利要求1所述的中文命名实体的标注方法,其特征在于,在所述将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果之前,所述方法还包括:

获取第三文本向量样本;

根据预设标注规范对所述第三文本向量样本中的文字进行标注,以获取第一标注文本样本;

根据所述第三文本向量样本和所述第一标注文本样本,对一序列标注模型进行训练,以获取所述第二模型。

10.根据权利要求5所述的中文命名实体的标注方法,其特征在于,在所述将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果之前,所述方法还包括:

根据预设标注规范对所述第一文本向量样本和/或所述第二文本向量样本中的文字进行标注,以获取目标标注文本样本;

根据所述第一文本向量样本和/或所述第二文本向量样本及所述目标标注文本样本,对一序列标注模型进行训练,以获取所述第二模型。

11.根据权利要求1所述的中文命名实体的标注方法,其特征在于,所述方法还包括:

输出所述原始文本的实体标注结果,所述实体标注结果包括所述原始文本中的文字的标注信息、所述原始文本中的命名实体以及所述命名实体的数量。

12.一种中文命名实体的标注装置,其特征在于,所述装置包括:

文本向量转换模块,用于对原始文本中的文字用空格进行分隔,并将空格分隔后的文字进行部首拆分,以获取处理后的原始文本,并将所述处理后的原始文本输入至第一模型,以将所述处理后的原始文本转换为文本向量;所述第一模型为根据训练数据对一语言模型进行训练获得的,所述训练数据包括第一文本样本、第二文本样本、第一文本样本对应的第一文本向量样本、第二文本样本对应的第二文本向量样本,所述第二文本向量样本为所述第二文本样本经预处理后的部首样本进行向量化处理后得到的;其中,所述语言模型为双向长短期记忆模型,在根据所述训练数据对所述语言模型进行训练的过程中,利用所述语言模型提取所述训练数据中每个文字的最后一个部首的输出隐状态,以将正向输出的隐状态序列和反向输出的隐状态序列进行拼接得到完整的隐状态序列;

实体标注模块,用于将所述文本向量输入至第二模型,以获取所述文本向量中的文字之间的关联信息,根据所述关联信息对所述文本向量中的文字进行标注,以获取所述原始

文本中的实体标注结果。

13.一种存储介质,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根据权利要求1至11中任一项所述的中文命名实体的标注方法。

14.一种电子设备,其特征在于,包括:

处理器;以及

存储器,用于存储所述处理器的可执行指令;

其中,所述处理器配置为经由执行所述可执行指令来执行权利要求1至11中任一项所述的中文命名实体的标注方法。

中文命名实体的标注方法、装置、存储介质和电子设备

技术领域

[0001] 本公开涉及计算机技术领域，更具体地，涉及一种中文命名实体的标注方法、中文命名实体的标注装置、计算机存储介质和电子设备。

背景技术

[0002] 随着计算机科学领域和人工智能领域的蓬勃发展，命名实体识别成为自然语言处理领域中的一个重点研究问题。命名实体是目标文本中基本的信息元素，是正确理解目标文本的基础；命名实体识别是指从文本中识别出相关实体，并标注出其位置以及类型。汉语作为象形文字，与西方语言相比，缺少显示的标记，在语法、语义、语用方面也更加灵活，这就使中文的实体识别任务往往更具挑战性。

[0003] 相关技术中的中文命名识别的方式大致分为三类：基于词典和规则的方法、基于特征模板的方法和基于神经网络的方法。但很多情况下不可避免的需要将文本进行分词，难以避免由分词错误带来的问题，同时，由于无法捕捉到汉字的构成信息，大大降低了中文命名实体的识别准确性。

[0004] 需要说明的是，在上述背景技术部分发明的信息仅用于加强对本公开的背景的理解，因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

[0005] 本公开的目的在于提供一种中文命名实体的标注方法及装置、计算机存储介质和电子设备，进而至少在一定程度上克服由于分词错误和忽略了汉字的组成信息而导致的中文命名实体标注准确度低等问题。为实现以上技术效果，本公开采用如下技术方案。

[0006] 本公开的其他特性和优点将通过下面的详细描述变得显然，或部分地通过本公开的实践而习得。

[0007] 根据本公开的一个方面，提供一种中文命名实体的标注方法，所述方法包括：对原始文本进行预处理，并将处理后的原始文本输入至第一模型，以将所述处理后的原始文本转换为文本向量；将所述文本向量输入至第二模型，通过所述第二模型对所述文本向量中的文字进行标注，以获取所述原始文本中的实体标注结果。

[0008] 在本公开的一种示例性实施例中，所述对原始文本进行预处理，并将处理后的原始文本输入至第一模型，以将所述处理后的原始文本转换为文本向量，包括：将所述原始文本中的文字用空格进行分隔；将用空格分隔后的文字进行部首拆分，以获取所述处理后的原始文本；将所述处理后的原始文本输入至所述第一模型，以将所述处理后的原始文本转换为文本向量。

[0009] 在本公开的一种示例性实施例中，在所述将用空格分隔后的文字进行部首的拆分，以获取所述处理后的原始文本之前，所述方法还包括：通过预设程序获取目标文本库中的文字，并对所述目标文本库中的文字进行部首拆分以获取与所述目标文本库中的文字对应的部首；根据所述目标文本库中的文字和对应的所述部首，形成部首词典。

[0010] 在本公开的一种示例性实施例中,所述将用空格分隔后的文字进行部首的拆分,以获取所述处理后的原始文本,包括:基于所述部首词典,对所述原始文本中的文字进行部首拆分,以获取所述处理后的原始文本。

[0011] 在本公开的一种示例性实施例中,在所述对原始文本进行预处理,并将处理后的原始文本输入至第一模型,以将所述处理后的原始文本转换为文本向量之前,所述方法还包括:获取训练数据,所述训练数据包括文本样本和与所述文本样本对应的文本向量样本;根据所述训练数据对一语言模型进行训练,以获取所述第一模型。

[0012] 在本公开的一种示例性实施例中,所述文本样本包括第一文本样本和第二文本样本;所述文本向量样本包括第一文本向量样本和第二文本向量样本;所述获取训练数据,包括:

[0013] 通过向量模型对所述第一文本样本进行向量化处理,以获取所述第一文本向量样本;将所述第二文本样本中的文字用空格进行分隔,并将用空格分隔后的文字进行部首拆分,以获取部首样本;通过预训练语言模型对所述部首样本进行向量化处理,以获取所述第二文本向量样本;根据所述第一文本样本、所述第一文本向量样本、所述第二文本样本和所述第二文本向量样本,确定所述训练数据。

[0014] 在本公开的一种示例性实施例中,所述第二模型包括双向神经网络子模型和条件随机场子模型;所述将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果,包括:

[0015] 将所述文本向量输入至所述第二模型,以获取所述文本向量中的文字之间的关联信息;根据所述关联信息对所述文本向量中的文字进行标注,以获取所述原始文本的实体标注结果。

[0016] 在本公开的一种示例性实施例中,所述将所述文本向量输入至所述第二模型,以获取所述文本向量中的文字之间的关联信息,包括:将所述文本向量输入至所述双向神经网络子模型,以通过所述双向神经网络子模型将所述文本向量的双向隐藏状态进行拼接,以获取所述文本向量中的文字之间的关联信息。

[0017] 在本公开的一种示例性实施例中,所述根据所述关联信息对所述文本向量中的文字进行标注,以获取所述原始文本的实体标注结果,包括:基于所述文本向量中的文字之间的关联信息,将由所述双向神经网络模型输出的文本输入至所述条件随机场模型,以对所述文本向量中的文字进行标注,得到所述原始文本的实体标注结果。

[0018] 在本公开的一种示例性实施例中,在所述将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果之前,所述方法还包括:

[0019] 获取第三文本向量样本;根据预设标注规范对所述第三文本向量样本中的文字进行标注,以获取第一标注文本样本;根据所述第三文本向量样本和所述第一标注文本样本,对一序列标注模型进行训练,以获取所述第二模型。

[0020] 在本公开的一种示例性实施例中,在所述将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果之前,所述方法还包括:

[0021] 根据预设标注规范对所述第一文本向量样本和/或所述第二文本向量样本中的文

字进行标注,以获取目标标注文本样本;根据所述第一文本向量样本和/或所述第二文本向量样本及所述目标标注文本样本,对一序列标注模型进行训练,以获取所述第二模型。

[0022] 在本公开的一种示例性实施例中,所述第四文本向量样本为所述第一文本向量样本和/或所述第二文本向量样本。

[0023] 在本公开的一种示例性实施例中,所述方法还包括:输出所述原始文本的实体标注结果,所述实体标注结果包括所述原始文本中的文字的标注信息、所述原始文本中的命名实体以及所述命名实体的数量。

[0024] 根据本公开的一个方面,提供一种中文命名实体的标注装置,所述中文命名实体的标注装置包括:文本向量转换模块,用于对原始文本进行预处理,并将处理后的原始文本输入至第一模型,以将所述处理后的原始文本转换为文本向量;实体标注模块,用于将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果。

[0025] 根据本公开的一个方面,提供一种计算机存储介质,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上述任意一项所述的中文命名实体的标注方法。

[0026] 根据本公开的一个方面,提供一种电子设备,包括:处理器;以及存储器,用于存储所述处理器的可执行指令;其中,所述处理器配置为经由执行所述可执行指令来执行上述任意一项所述的中文命名实体的标注方法。

[0027] 本公开的示例性实施方式中的中文命名实体的标注方法,将原始文本输入至第一模型后得到文本向量,接着将文本向量输入至第二模型,以实现对原始文本的实体标注。一方面,通过第一模型对处理后的原始文本进行向量化处理,无需将文本进行分词处理,避免了由于分词错误而导致的实体划分错误的问题;同时,通过第一模型处理后得到的文本向量,能更好地表示原始文本中的汉字,提高了命名实体标注的准确性;另一方面,以第一模型处理后得到的文本向量为基础,通过第二模型对原始文本进行实体标注,两个模型的结合使实体标注过程更具可靠性。

[0028] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本公开。

附图说明

[0029] 通过参考附图阅读下文的详细描述,本公开示例性实施方式的上述以及其他目的、特征和优点将变得易于理解。在附图中,以示例性而非限制性的方式示出了本公开的若干实施方式,其中:

[0030] 图1示意性地示出了根据本公开示例性实施方式的中文命名实体的标注方法的流程图;

[0031] 图2示意性地示出了根据本公开示例性实施方式的对原始文本进行预处理的示意图;

[0032] 图3示意性地示出了根据本公开示例性实施方式的第一模型对处理后的原始文本进行文本向量化处理的示意图;

[0033] 图4示意性地示出了根据本公开示例性实施方式的获取训练数据的流程图;

[0034] 图5示意性地示出了根据本公开示例性实施方式的通过第二模型对文本向量中的

文字进行标注,以获取原始文本中的实体标注结果的流程图;

[0035] 图6示意性地示出了根据本公开示例性实施方式的BiLSTM-CRF序列标注模型的示意图;

[0036] 图7示意性地示出了根据本公开示例性实施方式的基于第二模型对文本向量中的文字进行标注的示意图;

[0037] 图8示意性地示出了根据本公开示例性实施方式的获取第二模型的流程图;

[0038] 图9A-9B示意性地示出了根据本公开示例性实施方式的原始文本的部分标注结果的示意图;

[0039] 图10示意性地示出了根据本公开示例性实施方式的中文命名实体的标注装置的结构示意图;

[0040] 图11示意性地示出了根据本公开示例性实施方式的存储介质的示意图;以及

[0041] 图12示意性地示出了根据本公开示例性实施方式的电子设备的框图。

[0042] 在附图中,相同或对应的标号表示相同或对应的部分。

具体实施方式

[0043] 现在将参考附图更全面地描述示例性实施方式。然而,示例性实施方式能够以多种形式实施,且不应被理解为限于在此阐述的范例;相反,提供这些实施例使得本公开将更加全面和完整,并将示例性实施方式的构思全面地传达给本领域的技术人员。图中相同的附图标记表示相同或类似的结构,因而将省略它们的详细描述。

[0044] 此外,所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施例中。在下面的描述中,提供许多具体细节从而给出对本公开的实施例的充分理解。然而,本领域技术人员将意识到,可以实践本公开的技术方案而没有所述特定细节中的一个或更多,或者可以采用其它的方法、组元、装置、步骤等。在其它情况下,不详细示出或描述公知结构、方法、装置、实现或者操作以避免模糊本公开的各方面。

[0045] 附图中所示的方框图仅仅是功能实体,不一定必须与物理上独立的实体相对应。即,可以采用软件形式来实现这些功能实体,或在一个或多个软件硬化的模块中实现这些功能实体或功能实体的一部分,或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0046] 在本领域的相关技术中,中文的命名实体的识别主要有两种方式:基于词的中文命名实体识别是将文本进行分词后,基于LSTM(Long Short-Term Memory,长短期记忆网络)-CRF(Conditional Random Field,条件随机场)模型完成实体的标注;基于字的中文命名实体识别,无需分词,基于LSTM-CRF模型完成实体的标注。

[0047] 相应地,相关技术中的中文命名实体标注方法存在如下缺陷:分词的错误可能导致实体边界划分错误,进而影响到实体的标注结果;无法完全获取到汉字的构字信息,而忽略了汉字的构字信息在一定程度上会降低对中文命名实体标注的准确性。

[0048] 命名实体识别是机器翻译、问答系统、信息提取和面向语义网的元数据标注等领域的重要基础工作,由于汉字以及汉语言缺少显示的标记,在语法、语义、语用等方面也更加灵活,使得中文的命名实体标注更具挑战性。基于此,在本公开示例性实施方式中,首先提供了一种中文命名实体的标注方法。

[0049] 图1示出了本公开示例性实施方式的中文命名实体的标注方法的流程图,参考图1所示,该中文命名实体的标注方法可以包括以下步骤:

[0050] 步骤S110:对原始文本进行预处理,并将处理后的原始文本输入至第一模型,以将所述处理后的原始文本转换为文本向量;

[0051] 步骤S120:将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果。

[0052] 根据本公开示例性实施方式中的中文命名实体的标注方法,一方面,通过第一模型对处理后的原始文本进行向量化处理,无需将文本进行分词处理,避免了由于分词错误而导致的实体划分错误的问题;同时,通过第一模型处理后得到的文本向量,能够更好地表示原始文本中的汉字,提高了命名实体标注的准确性;另一方面,以第一模型处理后得到的文本向量为基础,通过第二模型对原始文本进行实体标注,两个模型的结合使实体标注的过程更具可靠性。

[0053] 下面将对本公开的示例性实施方式中的中文命名实体的标注方法进行进一步的说明。

[0054] 在步骤S110中,对原始文本进行预处理,并将处理后的原始文本输入至第一模型,以将所述处理后的原始文本转换为文本向量。

[0055] 在本公开的示例性实施方式中,在将原始文本输入至第一模型之前,需要对原始文本进行预处理。该预处理是将原始文本进行处理以获得可以作为第一模型的输入的文本格式的过程,具体可以包括:首先将原始文本中的文字用空格进行分隔;然后将用空格分隔后的文字进行部首拆分,将原始文本按部首拆分可以获取汉字的隐藏信息,例如通常部首中有“鸟”的汉字,往往与家禽类相关联。举例而言,图2示出了对原始文本进行预处理的示意图,如图2所示,首先将原始文本“阴阳师的”用空格进行分隔,然后将分隔后的原始文本进行部首的拆分,以得到处理后的原始文本。

[0056] 其中,在将用空格分隔后的文字进行部首的拆分,以获取处理后的原始文本之前,可以先获取部首词典,该部首词典是对原始文本进行处理的依据,基于部首词典可以确定原始文本中的文字的组成信息,进而确定原始文本中的文字的部首拆分情况。具体的,可以通过预设程序获取目标文本库中的文字;然后对目标文本库中的文字进行部首拆分,以获取与目标文本库中的文字对应的部首;最后根据获取的目标文本库中的文字和该些文字对应的部首,形成部首词典。其中预设程序可以是网络爬虫程序,网络爬虫程序可以按照预设规则自动爬取目标文本库中的文字以及文字对应的部首信息,当然,预设程序还可以是其他具有获取文字以及文字对应的部首信息功能的脚本程序;目标文本库可以是万维网中的网络百科全书(如维基百科、百度百科等),也可以是语料库(如《人民日报》标注语料库)中的语料,本公开对此不做具体限定。在形成部首词典后,可以基于部首词典对原始文本中的文字进行部首拆分,以得到与原始文本中的文字对应的部首。

[0057] 进一步的,可以将处理后的原始文本输入至第一模型,以将处理后的原始文本转换为文本向量。图3示出了第一模型对处理后的原始文本进行文本向量化处理的示意图,如图3所示,将用空格分隔且已进行部首拆分处理后的原始文本输入至已训练好的第一模型,基于第一模型中已训练得到的文本向量样本的动态嵌入,对处理后的原始文本进行向量化处理,以获得与原始文本对应的字向量文本。

[0058] 当然,在对原始文本进行预处理,并将处理后的原始文本输入至第一模型,以将处理后的原始文本转换为文本向量之前,可以根据训练数据对语言模型进行训练,以获得第一模型。具体的,首先获取训练数据,该训练数据包括文本样本和与文本样本对应的文本向量样本。在本公开的实施方式中,文本样本可以包括第一文本样本和第二文本样本;文本向量样本可以包括第一文本向量样本和第二文本向量样本,图4示出了获取训练数据的流程图,如图4所示,该过程可以包括如下步骤:

[0059] 步骤S410:通过向量模型对所述第一文本样本进行向量化处理,以获取所述第一文本向量样本。

[0060] 在本公开的示例性实施方式中,向量模型是指可以用来训练词向量的模型,例如可以是Word2Vec模型(Word to Vector,词向量模型),第一文本样本作为Word2Vec模型的训练语料,可以是通过收集维基百科和百度百科获得的文本,也可以是语料库(如《人民日报》标注语料库)中的语料。通过Word2Vec模型训练第一文本样本,得到与第一文本样本对应的第一文本化向量样本。需要说明的是,向量模型的种类以及第一文本样本还可以根据实际训练情况进行选择,本公开对此不做具体限定。

[0061] 步骤S420:将所述第二文本样本中的文字用空格进行分隔,并将用空格分隔后的文字进行部首拆分,以获取部首样本。

[0062] 在本公开的示例性实施方式中,第二文本样本可以为与上述的第一文本样本相同的样本,也可以是区别于第一文本样本的样本,例如可以是来自于不同语料库中的语料或者来自同一语料库中的不同语料部分等,本公开对此不做特殊限定。具体的预处理过程的示意图可以继续参照图2所示,当然,在将用空格分隔后的文字进行部首拆分时,同样需要基于预先形成的部首词典,本公开对此不再赘述。

[0063] 步骤S430:通过预训练语言模型对所述部首样本进行向量化处理,以获取所述第二文本向量样本。

[0064] 在本公开的示例性实施方式中,由于在步骤S410中,仅通过向量模型对第一文本样本的向量化处理得到的第一文本向量样本,并未考虑到汉字的构字信息,因此还可以基于预训练语言模型对第二文本样本进行向量化处理得到第二向量文本样本,以作为Word2Vec模型训练得到的第一文本向量样本的补充,以提高后续通过第一模型进行文本样本向量化处理的准确性。其中,该预训练语言模型可以是预先训练好的BiLSTM模型,当然,也可以根据实际需要选择相应的预训练语言模型。

[0065] 步骤S440:根据所述第一文本样本、所述第一文本向量样本、所述第二文本样本和所述第二文本向量样本,确定所述训练数据。

[0066] 在本公开的示例性实施方式中,根据上述获取的第一文本样本、与第一文本样本对应的第一文本向量样本、第二文本样本和与第二文本样本对应的第二文本向量样本,确定训练数据。

[0067] 进一步的,在获取到训练数据后,基于训练数据对一语言模型进行训练,以获取第一模型。其中语言模型具体可以是BiLSTM模型(Bi-directional Long Term Memory network,双向长短时记忆模型)。具体而言,在对BiLSTM模型进行训练时,当将训练数据中的文本样本(包括第一文本样本和第二文本样本)输入至BiLSTM模型后,BiLSTM模型将会提取每个文字的部首特征,由于BiLSTM模型是双向的循环神经网络,首先将会提取到每个字

的最后一个部首的输出隐状态;然后将正向LSTM输出的隐状态序列和反向的LSTM序列在各个位置的隐状态进行拼接,以得到完整的隐状态序列,由此得到的隐状态是由前后向的LSTM的输出链接组成的,包含文本样本中的每一句的开头字、结尾字的传播信息。在此过程中,基于与第一文本样本对应的第一向量文本样本、与第二文本样本对应的第二向量文本样本,对该模型的参数进行调整,直至形成第一文本向量样本与第二文本向量样本的动态嵌入,进而能够得到更准确的文本向量化表示。由于该语言模型的训练是基于获取的训练数据(包括第一文本样本、第一文本向量样本、第二文本样本和第二文本向量样本)训练得到的,当再次输入一处理后的文本时,将会准确输出与该处理后的文本对应的文本向量化表示,提高了对原始文本的向量化处理的准确性,进而能更准确的表示文本中的汉字。

[0068] 在步骤S120中,将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果。

[0069] 在本公开的示例性实施方式中,第二模型是指用于对文本向量进行预测标注的模型,该模型可以包括双向神经网络子模型和条件随机场子模型。具体的,图5示出了通过第二模型对文本向量中的文字进行标注,以获取原始文本中的实体标注结果的流程图,如图5所示,该过程可以包括步骤S510和步骤S520:在步骤S510中,将文本向量输入至第二模型,以获取文本向量中的文字之间的关联信息,也就是说,将文本向量输入至双向神经网络子模型,通过双向神经网络子模型将文本向量的双向隐状态进行拼接,以获取文本向量中的文字之间的关联信息;在步骤S520中,根据关联信息对文本向量中的文字进行标注,以获取原始文本的实体标注结果,具体而言,基于步骤S510获得的文本向量中的文字之间的关联信息,将由双向神经网络模型输出的文本输入至条件随机场模型,以对文本向量中的文字进行标注,得到原始文本中的实体标注结果。

[0070] 其中,第二模型可以是BiLSTM-CRF模型,图6示出了BiLSTM-CRF序列标注模型的示意图,由图6可知,将文本向量至输入至BiLSTM-CRF模型,将输出与该文本向量中的文字对应的标注结果。图7示出了基于第二模型对文本向量中的文字进行标注的示意图,如图7所示,将文本向量输入至第二模型后,通过BiLSTM模型将文本向量的双向隐状态进行拼接获取文本向量中的文字之间的关联信息后,输入至CRF模型对文本向量中的文字进行标注,例如“阴”的标注结果为“B-NG”,其中“B”在实体标注规范中代表实体的开始,NG(Name of Game,游戏名词)可以为在对模型进行训练时的预先设定好的标注标签。

[0071] 当然,在将文本向量输入至第二模型,通过第二模型对文本向量中的文字进行标注,以获取原始文本中的实体标注结果之前,需要根据文本向量样本和与文本向量样本对应的标注文本样本对一序列标注模型进行训练,以获取第二模型。具体而言,图8示出了获取第二模型的流程图,参照图8所示,该流程包括如下步骤:

[0072] 步骤S810:获取第三文本向量样本。

[0073] 在本公开的示例性实施方式中,作为第二模型的训练数据,该第三文本向量样本可以是上述的通过向量模型处理得到的第一文本向量样本,可以是通过预训练语言模型处理得到的第二文本向量样本,或者可以是第一样本向量样本和第二向量样本的集合;当然,第三文本向量样本还可以是区别于上述的第一文本向量样本和第二文本向量样本的文本向量,例如可以是通过对预设文本库中的文本进行向量化处理后得到的文本向量样本,本公开对此不做具体限定。

[0074] 步骤S820:根据预设标注规范对所述第三文本向量样本中的文字进行标注,以获取第一标注文本样本。

[0075] 在本公开的示例性实施方式中,预设标注规范可以是BIOES标注规范,也可以是BIO标注规范。其中,在BIOES标注规范中,B是实体的开始,I是实体的中间,E是实体的末尾,O是非实体,S是单独成实体;在BIO标注规范中,B是实体的开始,I是实体的中间或结尾,O是非实体,当然,预设标注规范也可以是其它的标注规范,本公开对此不做具体限定。根据预设的标注规范对第三文本向量中的文字进行标注,可以获得第一标注文本;当然,在根据预设标注规范对第三文本向量进行标注时,还可以根据实际标注需求,设置相应的标注标签,例如上述的NG游戏标签,等等,本公开对此不做具体限定。

[0076] 步骤S830:根据所述第三文本向量样本和所述第一标注文本样本,对一序列标注模型进行训练,以获取所述第二模型。

[0077] 在本公开的示例性实施方式中,根据第三文本向量样本和第一标注文本对一序列模型进行训练,以获取第二模型,通过对模型不断优化,使得当输入一文本向量后即可输出对该文本向量的标注结果。其中,当第三文本向量样本为第一文本向量样本和/或第二文本向量样本时,可以对第一文本向量样本和/或第二文本向量样本中的文字进行标注,以获得目标标注文本样本,并根据第一文本向量样本和/或第二文本向量样本及该目标标注文本样本,对一序列标注模型进行训练,以获取第二模型。

[0078] 另外,在本公开的示例性实施方式中,还可以将原始文本的实体标注结果输出,该实体标注结果包括原始文本中的文字的标注信息、原始文本中的命名实体以及命名实体的数量,图9A-9B示出了原始文本的部分标注结果的示意图,如图9A所示为原始文本的示例图,如图9B所示,在将原始文本中的文字的标注信息输出的同时,还将原始文本中的实体以及相应的实体数量输出,例如输出“阴阳师”以及“2”,说明原始文本中包括两个实体“阴阳师”。需要说明的是,图9A-9B仅是原始文本以及输出的原始文本的实体标注结果的部分示例,本公开包括但不限于上述示例的形式。

[0079] 此外,在本公开的示例性实施方式中,还提供了一种中文命名实体的标注装置,参考图10所示,该中文命名实体的标注装置1000可以包括文本向量转换模块1010以及实体标注模块1020。具体地,

[0080] 文本向量转换模块1010,用于对原始文本进行预处理,并将处理后的原始文本输入至第一模型,以将所述处理后的原始文本转换为文本向量;

[0081] 实体标注模块1020,用于将所述文本向量输入至第二模型,通过所述第二模型对所述文本向量中的文字进行标注,以获取所述原始文本中的实体标注结果。

[0082] 上述装置中各模块/单元的具体细节在方法部分的实施方式中已经详细说明,因此不再赘述。

[0083] 此外,在本公开示例性实施方式中,还提供了一种能够实现上述方法的计算机存储介质。其上存储有能够实现本说明书上述方法的程序产品。在一些可能的实施例中,本公开的各个方面还可以实现为一种程序产品的形式,其包括程序代码,当所述程序产品在终端设备上运行时,所述程序代码用于使所述终端设备执行本说明书上述“示例性方法”部分中描述的根据本公开各种示例性实施例的步骤。

[0084] 参考图11所示,描述了根据本公开的示例性实施方式的用于实现上述方法的程序

产品1100,其可以采用便携式紧凑盘只读存储器(CD-ROM)并包括程序代码,并可以在终端设备,例如个人电脑上运行。然而,本公开的程序产品不限于此,在本文件中,可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0085] 所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括:具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0086] 计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了可读程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读介质,该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0087] 可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于无线、有线、光缆、RF等等,或者上述的任意合适的组合。

[0088] 可以以一种或多种程序设计语言的任意组合来编写用于执行本公开操作的程序代码,所述程序设计语言包括面向对象的程序设计语言—诸如Java、C++等,还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中,远程计算设备可以通过任意种类的网络,包括局域网(LAN)或广域网(WAN),连接到用户计算设备,或者,可以连接到外部计算设备(例如利用因特网服务提供商来通过因特网连接)。

[0089] 此外,在本公开的示例性实施例中,还提供了一种能够实现上述方法的电子设备。所属技术领域的技术人员能够理解,本公开的各个方面可以实现为系统、方法或程序产品。因此,本公开的各个方面可以具体实现为以下形式,即:完全的硬件实施例、完全的软件实施例(包括固件、微代码等),或硬件和软件方面结合的实施例,这里可以统称为“电路”、“模块”或“系统”。

[0090] 下面参照图12来描述根据本公开的这种实施例的电子设备1200。图12显示的电子设备1200仅仅是一个示例,不应对本公开实施例的功能和使用范围带来任何限制。

[0091] 如图12所示,电子设备1200以通用计算设备的形式表现。电子设备1200的组件可以包括但不限于:上述至少一个处理单元1210、上述至少一个存储单元1220、连接不同系统组件(包括存储单元1220和处理单元1210)的总线1230、显示单元1240。

[0092] 其中,所述存储单元存储有程序代码,所述程序代码可以被所述处理单元1210执行,使得所述处理单元1210执行本说明书上述“示例性方法”部分中描述的根据本公开各种示例性实施例的步骤。

[0093] 存储单元1220可以包括易失性存储单元形式的可读介质,例如随机存取存储单元

(RAM)1221和/或高速缓存存储单元1222,还可以进一步包括只读存储单元(ROM)1223。

[0094] 存储单元1220还可以包括具有一组(至少一个)程序模块1225的程序/实用工具1224,这样的程序模块1225包括但不限于:操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0095] 总线1230可以为表示几类总线结构中的一种或多种,包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

[0096] 电子设备1200也可以与一个或多个外部设备1300(例如键盘、指向设备、蓝牙设备等)通信,还可与一个或者多个使得用户能与该电子设备1200交互的设备通信,和/或与使得该电子设备1200能与一个或多个其它计算设备进行通信的任何设备(例如路由器、调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口1250进行。并且,电子设备1200还可以通过网络适配器1260与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器1260通过总线1230与电子设备1200的其它模块通信。应当明白,尽管图中未示出,可以结合电子设备1200使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0097] 通过以上的实施例的描述,本领域的技术人员易于理解,这里描述的示例实施例可以通过软件实现,也可以通过软件结合必要的硬件的方式来实现。因此,根据本公开实施例的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM,U盘,移动硬盘等)中或网络上,包括若干指令以使得一台计算设备(可以是个人计算机、服务器、终端装置、或者网络设备等)执行根据本公开实施例的方法。

[0098] 此外,上述附图仅是根据本公开示例性实施例的方法所包括的处理的示意性说明,而不是限制目的。易于理解,上述附图所示的处理并不表明或限制这些处理的时间顺序。另外,也易于理解,这些处理可以是例如在多个模块中同步或异步执行的。

[0099] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本公开的其他实施例。本公开旨在涵盖本公开的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的,本公开的真正范围和精神由权利要求指出。

[0100] 应当理解的是,本公开并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本公开的范围仅由所附的权利要求来限。

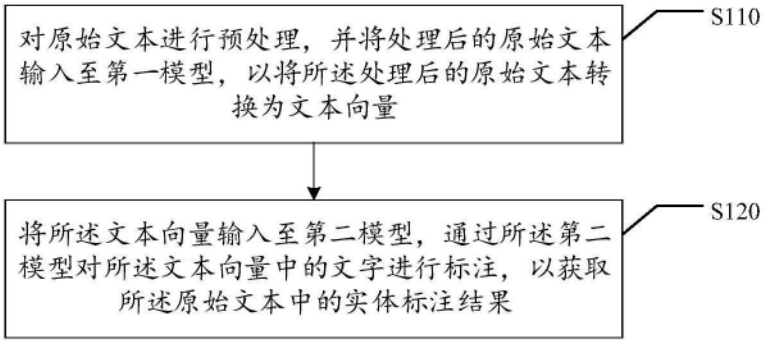

对原始文本进行预处理，并将处理后的原始文本输入至第一模型，以将所述处理后的原始文本转换为文本向量
S110
将所述文本向量输入至第二模型，通过所述第二模型对所述文本向量中的文字进行标注，以获取所述原始文本中的实体标注结果
S120

图1

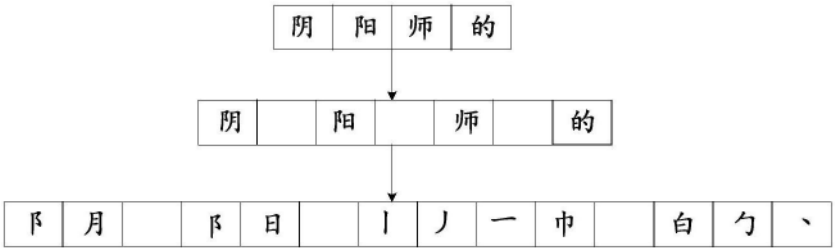

阴 阳 师 的
阴 阳 师 的
阝 月 阝 日 丨 丿 一 巾 白 勹 丶

图2

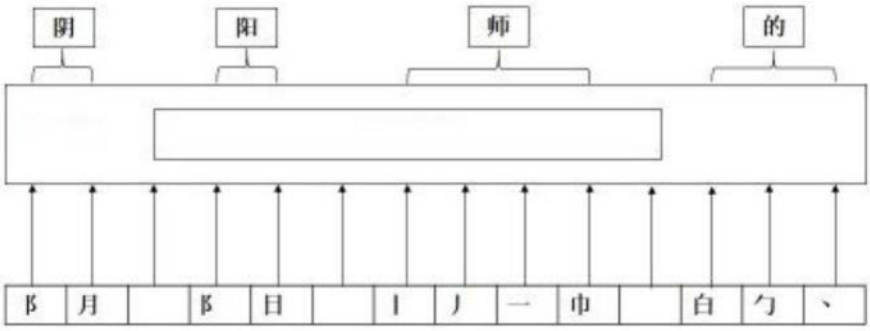

阴 阳 师 的
阝 月 阝 日 丨 丿 一 巾 白 勹 丶

图3

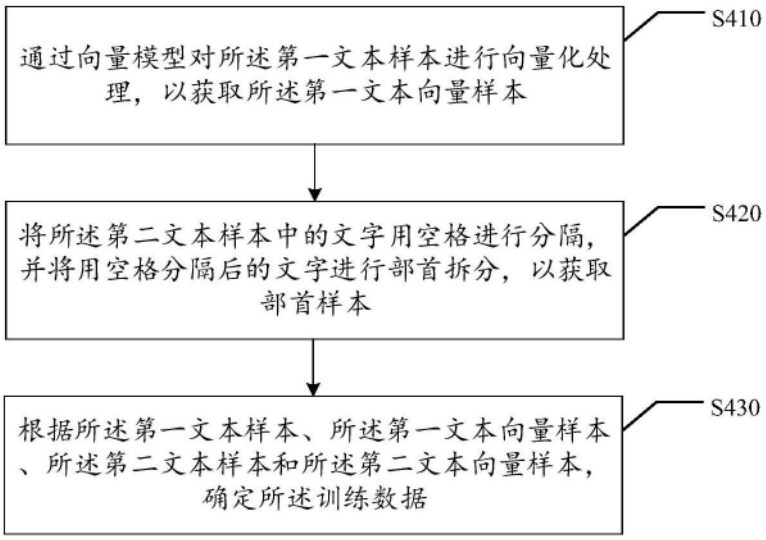

通过向量模型对所述第一文本样本进行向量化处理，以获取所述第一文本向量样本
S410
将所述第二文本样本中的文字用空格进行分隔，并将用空格分隔后的文字进行部首拆分，以获取部首样本
S420
根据所述第一文本样本、所述第一文本向量样本、所述第二文本样本和所述第二文本向量样本，确定所述训练数据
S430

图4

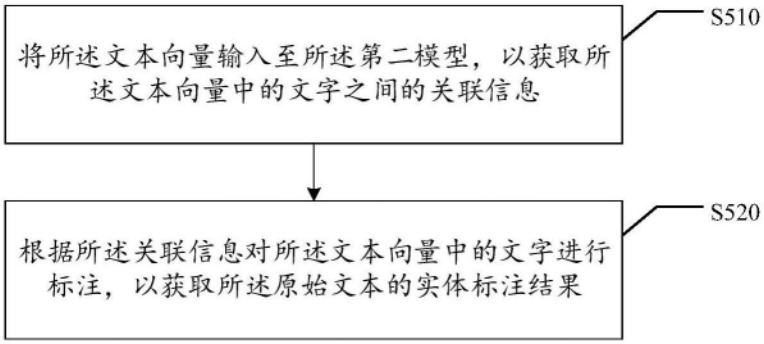

将所述文本向量输入至所述第二模型，以获取所述文本向量中的文字之间的关联信息
S510
根据所述关联信息对所述文本向量中的文字进行标注，以获取所述原始文本的实体标注结果
S520

图5

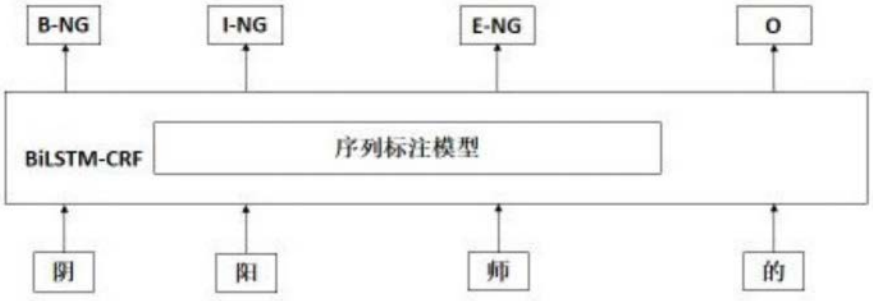

B-NG
I-NG
E-NG
O
BiLSTM-CRF
序列标注模型
阴
阳
师
的

图6

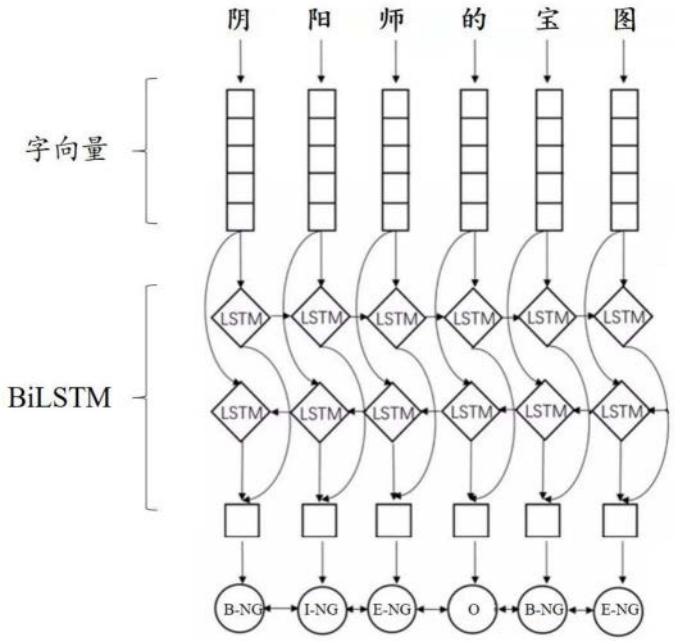

阴
阳
师
的
宝
图
字向量
BiLSTM
LSTM
B-NG
I-NG
E-NG
O
B-NG
E-NG

图7

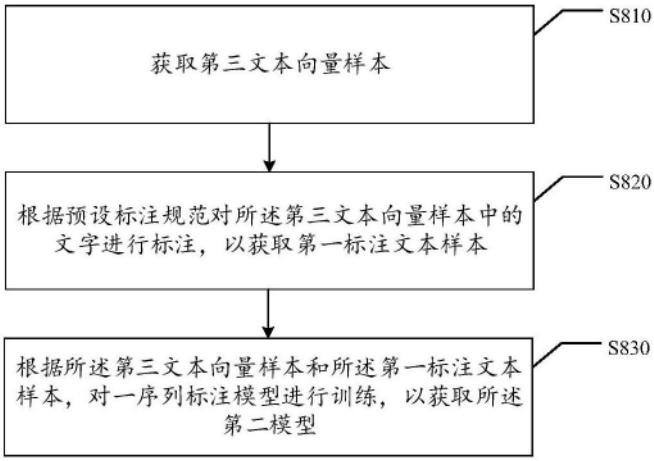

S810
获取第三文本向量样本
S820
根据预设标注规范对所述第三文本向量样本中的文字进行标注，以获取第一标注文本样本
S830
根据所述第三文本向量样本和所述第一标注文本样本，对一序列标注模型进行训练，以获取所述第二模型

图8

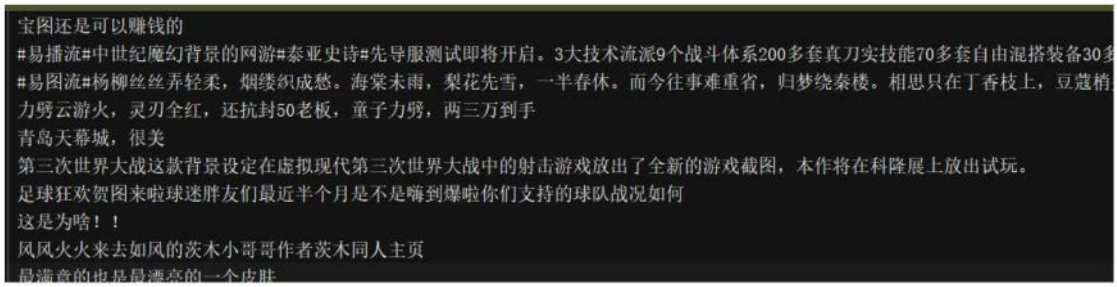

宝图还是可以赚钱的
#易播流#中世纪魔幻背景的网游#泰亚史诗#先导服测试即将开启。3大技术流派9个战斗体系200多套真刀实技能70多套自由混搭装备30多
#易图流#杨柳丝丝弄轻柔，烟缕织成愁。海棠未雨，梨花先雪，一半春休。而今往事难重省，归梦绕秦楼。相思只在丁香枝上，豆蔻梢
力劈云游火，灵刃全红，还抗封50老板，童子力劈，两三万到手
青岛天幕城，很美
第三次世界大战这款背景设定在虚拟现代第三次世界大战中的射击游戏放出了全新的游戏截图，本作将在科隆展上放出试玩。
足球狂欢贺图来啦球迷胖友们最近半个月是不是嗨到爆啦你们支持的球队战况如何
这是为啥！！
风风火火来去如风的茨木小哥哥作者茨木同人主页
最满意的也是最漂亮的一个皮肤

图9A

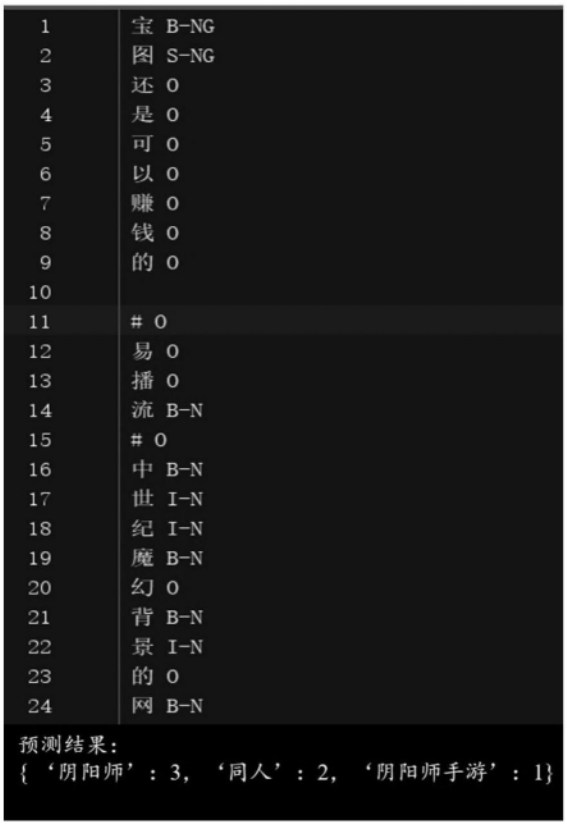

1 宝 B-NG
2 图 S-NG
3 还 O
4 是 O
5 可 O
6 以 O
7 赚 O
8 钱 O
9 的 O
10
11 # O
12 易 O
13 播 O
14 流 B-N
15 # O
16 中 B-N
17 世 I-N
18 纪 I-N
19 魔 B-N
20 幻 O
21 背 B-N
22 景 I-N
23 的 O
24 网 B-N
预测结果:
{ '阴阳师' : 3, '同人' : 2, '阴阳师手游' : 1}

图9B

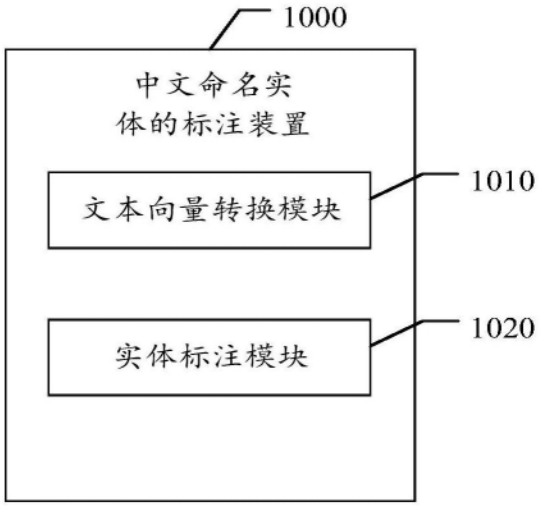

1000
中文命名实体的标注装置
文本向量转换模块
1010
实体标注模块
1020

图10

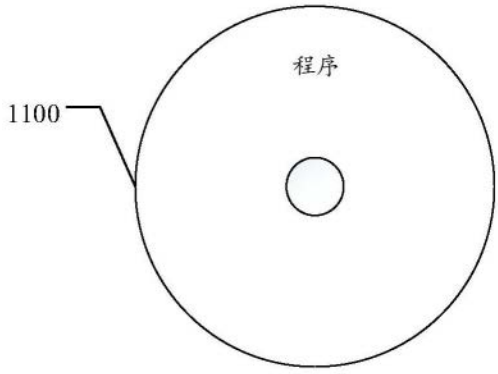

程序
1100

图11

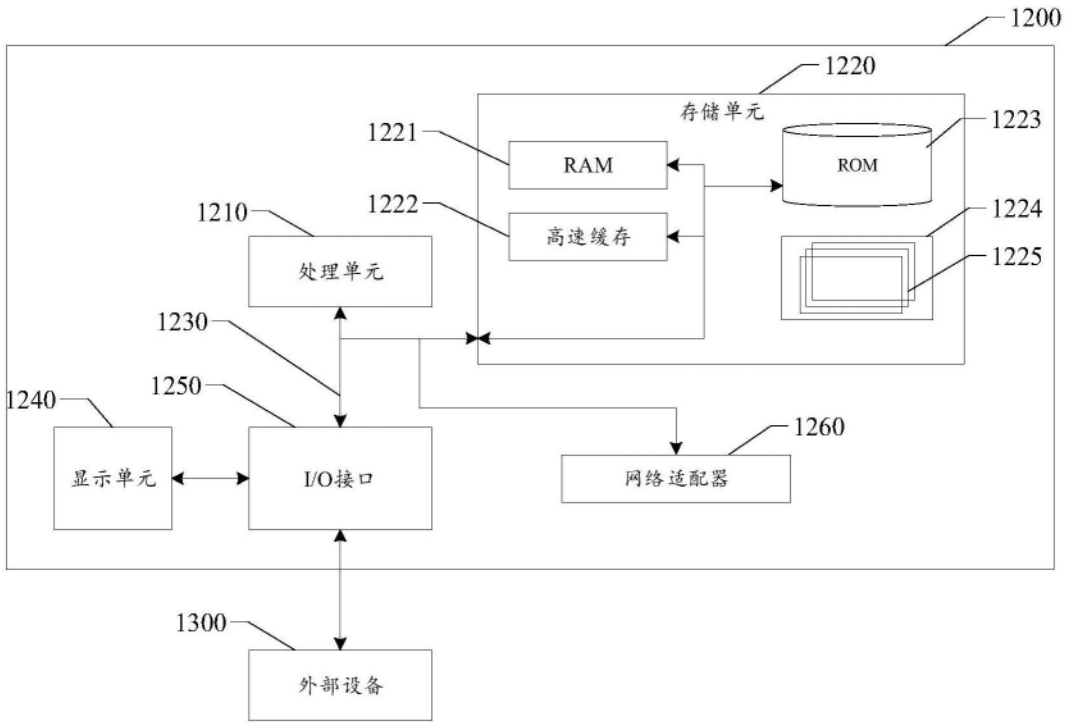

1200
1220
存储单元
1221
RAM
1223
ROM
1222
高速缓存
1210
处理单元
1224
1225
1230
1250
1240
显示单元
I/O接口
1260
网络适配器
1300
外部设备

图12