(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 113051920 A
(43)申请公布日 2021.06.29

(21)申请号 202110285605.X

(22)申请日 2021.03.17

(71)申请人 的卢技术有限公司
地址 210038 江苏省南京市栖霞区恒泰路8号汇智科技园A1栋

(72)发明人 于兴文

(74)专利代理机构 南京经纬专利商标代理有限公司 32200
代理人 罗运红

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/242(2020.01)

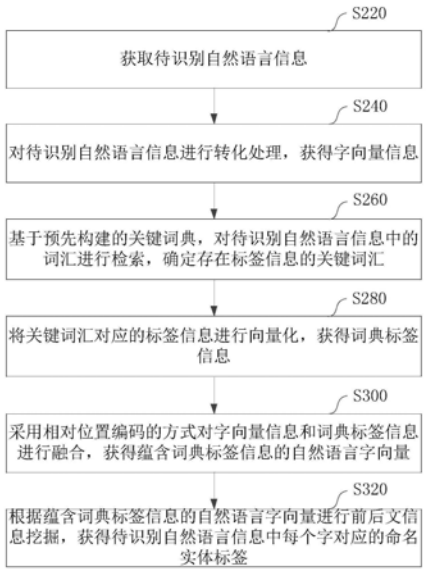
权利要求书2页 说明书6页 附图2页

(54)发明名称

命名实体识别方法、装置、计算机设备和存储介质

(57)摘要

本申请涉及一种命名实体识别方法、装置、计算机设备和存储介质。该方法包括:获取待识别自然语言信息;对所述待识别自然语言信息进行转化处理,获得字向量信息;基于预先构建的关键词典,对所述待识别自然语言信息中的词汇进行检索,确定存在标签信息的关键词汇;将所述关键词汇对应的标签信息进行向量化,获得词典标签信息;采用相对位置编码的方式对所述字向量信息和所述词典标签信息进行融合,获得蕴含词典标签信息的自然语言字向量;根据所述蕴含词典标签信息的自然语言字向量进行前后文信息挖掘,获得所述待识别自然语言信息中每个字对应的命名实体标签,采用本方法能够提高命名实体识别结果法人准确性。



CN 113051920 A

1.一种命名实体识别方法,其特征在于,所述方法包括:

获取待识别自然语言信息;

对所述待识别自然语言信息进行转化处理,获得字向量信息;

基于预先构建的关键词典,对所述待识别自然语言信息中的词汇进行检索,确定存在标签信息的关键词汇;

将所述关键词汇对应的标签信息进行向量化,获得词典标签信息;

采用相对位置编码的方式对所述字向量信息和所述词典标签信息进行融合,获得蕴含词典标签信息的自然语言字向量;

根据所述蕴含词典标签信息的自然语言字向量进行前后文信息挖掘,获得所述待识别自然语言信息中每个字对应的命名实体标签。

2.根据权利要求1所述的方法,其特征在于,所述对所述待识别自然语言信息进行转化处理,获得字向量信息的步骤,包括:

对所述待识别自然语言信息进行转化处理,获得数字标识信息;

将所述数字标识信息输入ALBERT模型进行编码,获得字向量信息。

3.根据权利要求1所述的方法,其特征在于,所述采用相对位置编码的方式对所述字向量信息和所述词典标签信息进行融合,获得蕴含词典标签信息的自然语言字向量的步骤,包括:

根据所述关键词汇,获取所述关键词汇在所述待识别自然语言信息的位置;

根据所述关键词汇在所述待识别自然语言信息的位置、所述字向量信息和所述词典标签信息,获得蕴含词典标签信息的自然语言字向量。

4.根据权利要求3所述的方法,其特征在于,所述根据所述关键词汇在所述待识别自然语言信息的位置、所述字向量信息和所述词典标签信息,获得蕴含词典标签信息的自然语言字向量,包括:

将所述关键词汇在所述待识别自然语言信息的位置、所述字向量信息和所述词典标签信息,输入到Transformer模型中进行信息融合,输出蕴含词典标签信息的自然语言字向量。

5.根据权利要求1所述的方法,其特征在于,所述根据所述蕴含词典标签信息的自然语言字向量进行前后文信息挖掘,获得所述待识别自然语言信息中每个字对应的命名实体标签的步骤,包括:

将所述蕴含词典标签信息的自然语言字向量输入条件随机场模型,进行前后文信息挖掘,输出所述待识别自然语言信息中每个字对应的命名实体标签。

6.根据权利要求1所述的方法,其特征在于,所述将所述关键词汇对应的标签信息进行向量化,获得词典标签信息的步骤,包括:

将所述关键词汇对应的标签信息进行数字化编码,获得数字化标签信息;

对所述数字化标签信息进行向量化,获得词典标签信息。

7.一种命名实体识别装置,其特征在于,所述装置包括:

信息获取模块,用于获取待识别自然语言信息;

信息转化模块,用于对所述待识别自然语言信息进行转化处理,获得字向量信息;

检索模块,用于基于预先构建的关键词典,对所述待识别自然语言信息中的词汇进行

检索，确定存在标签信息的关键词汇；

向量化模块，用于将所述关键词汇对应的标签信息进行向量化，获得词典标签信息；

融合模块，用于采用相对位置编码的方式对所述字向量信息和所述词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量；

信息挖掘模块，用于根据所述蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得所述待识别自然语言信息中每个字对应的命名实体标签。

8.一种计算机设备，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中任一项所述方法的步骤。

9.一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中任一项所述的方法的步骤。

命名实体识别方法、装置、计算机设备和存储介质

技术领域

[0001] 本申请涉及信息识别技术领域，特别是涉及一种命名实体识别方法、装置、计算机设备和存储介质。

背景技术

[0002] 命名实体识别(简称ner)任务是自然语言学习中的一项重要任务，其目的是从给定文本中抽取出所需的、关键的信息实体。ner任务是信息抽取、问答系统、知识图谱等智能服务的重要基础工具，为复杂分析任务提供资料与特征信息。

[0003] ner任务有多种解决途径，传统工程化的ner任务解决途径为对业务数据进行统计分析，总结归纳并维护一组与需求有关的专业核心词库，在对自然语言进行分析时，依据专业核心词库对文本序列进行词抽取。该方法可以一定程度上保证ner任务的识别质量，人工可以对其结果进行完全的干预。

[0004] 随着深度学习技术的发展，采用深度网络进行实体识别逐渐成为ner任务的一种更有效的解决途径。多种网络模型，如BiLSTM-CRF(序列标注算法)、BERT(是一种预训练模型，全称是Bidirectional Encoder Representation from Transformers)、ALBERT(基于BERT改进的一种预训练模型)，都可从语义的角度归纳自然语言的词或字，在结合上下文信息基础上更灵活、精准的抽取关键的实体。

[0005] 在当前工程化项目中，为使项目兼具可控性与灵活高效性，往往综合使用基于词典规则的命名实体识别方法与基于深度学习网络模型的命名实体识别方法。一方面维护业务词库，为命名实体识别的效果进行兜底，通过更改词库内容，适应不同业务需求下的命名实体识别需要。另一方面使用深度网络模型，根据自然语言的语义信息抽取关键实体。

[0006] 然而，虽然同时使用两种方法，但两种方法的信息并未进行结合，最终结果往往以其中一种方法的结果为主导，例如：当挖掘“谷雨在谷雨这一天认真工作”这句话中时间实体信息时，若时间词典中包含“谷雨”一词，采用词典的方法会将前后两个“谷雨”都归纳为“节气”的标签。采用深度学习的方法很大几率会将前者归类为“人名”后者归纳为“节气”，但由于训练集质量与数量的限制，其模型表现效果存在不稳定性。在实际项目中，往往采用两种方法中一种的结果作为最终的命名实体识别结果，这样的结果虽然较单独使用一种方法时更优，但无法考虑上下文语义关系，进行实体抽取时不灵活，因此，目前命名实体识别结果的准确性低。

发明内容

[0007] 基于此，有必要针对上述技术问题，提供一种能够提高命名实体识别结果准确性的命名实体识别方法、装置、计算机设备和存储介质。

[0008] 一种命名实体识别方法，所述方法包括：

[0009] 获取待识别自然语言信息；

[0010] 对所述待识别自然语言信息进行转化处理，获得字向量信息；

[0011] 基于预先构建的关键词典，对所述待识别自然语言信息中的词汇进行检索，确定存在标签信息的关键词汇；

[0012] 将所述关键词汇对应的标签信息进行向量化，获得词典标签信息；

[0013] 采用相对位置编码的方式对所述字向量信息和所述词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量；

[0014] 根据所述蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得所述待识别自然语言信息中每个字对应的命名实体标签。

[0015] 在其中一个实施例中，所述对所述待识别自然语言信息进行转化处理，获得字向量信息的步骤，包括：

[0016] 对所述待识别自然语言信息进行转化处理，获得数字标识信息；

[0017] 将所述数字标识信息输入ALBERT模型进行编码，获得字向量信息。

[0018] 在其中一个实施例中，所述采用相对位置编码的方式对所述字向量信息和所述词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量的步骤，包括：

[0019] 根据所述关键词汇，获取所述关键词汇在所述待识别自然语言信息的位置；

[0020] 根据所述关键词汇在所述待识别自然语言信息的位置、所述字向量信息和所述词典标签信息，获得蕴含词典标签信息的自然语言字向量。

[0021] 在其中一个实施例中，所述根据所述关键词汇在所述待识别自然语言信息的位置、所述字向量信息和所述词典标签信息，获得蕴含词典标签信息的自然语言字向量，包括：

[0022] 将所述关键词汇在所述待识别自然语言信息的位置、所述字向量信息和所述词典标签信息，输入到Transformer模型中进行信息融合，输出蕴含词典标签信息的自然语言字向量。

[0023] 在其中一个实施例中，所述根据所述蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得所述待识别自然语言信息中每个字对应的命名实体标签的步骤，包括：

[0024] 将所述蕴含词典标签信息的自然语言字向量输入条件随机场模型，进行前后文信息挖掘，输出所述待识别自然语言信息中每个字对应的命名实体标签。

[0025] 在其中一个实施例中，所述将所述关键词汇对应的标签信息进行向量化，获得词典标签信息的步骤，包括：

[0026] 将所述关键词汇对应的标签信息进行数字化编码，获得数字化标签信息；

[0027] 对所述数字化标签信息进行向量化，获得词典标签信息。

[0028] 一种命名实体识别装置，所述装置包括：

[0029] 信息获取模块，用于获取待识别自然语言信息；

[0030] 信息转化模块，用于对所述待识别自然语言信息进行转化处理，获得字向量信息；

[0031] 检索模块，用于基于预先构建的关键词典，对所述待识别自然语言信息中的词汇进行检索，确定存在标签信息的关键词汇；

[0032] 向量化模块，用于将所述关键词汇对应的标签信息进行向量化，获得词典标签信息；

[0033] 融合模块，用于采用相对位置编码的方式对所述字向量信息和所述词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量；

[0034] 信息挖掘模块，用于根据所述蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得所述待识别自然语言信息中每个字对应的命名实体标签。

[0035] 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所述方法的步骤。

[0036] 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所述的方法的步骤。

[0037] 上述命名实体识别方法、装置、计算机设备和存储介质，通过获取待识别自然语言信息；对所述待识别自然语言信息进行转化处理，获得字向量信息，基于预先构建的关键词典，对所述待识别自然语言信息中的词汇进行检索，确定存在标签信息的关键词汇，将所述关键词汇对应的标签信息进行向量化，获得词典标签信息，在保证运行效率的基础上融合了语义信息，采用相对位置编码的方式对所述字向量信息和所述词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量，该种融合方式使得信息的结合更加直接灵活，可对命名体识别的精度与灵活性产生极大的增益，并根据所述蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得所述待识别自然语言信息中每个字对应的命名实体标签，从而提高了命名实体识别结果法人准确性。

附图说明

[0038] 图1为一个实施例中命名实体识别方法的流程示意图；

[0039] 图2为一个实施例中命名实体识别装置的结构框图。

具体实施方式

[0040] 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。

[0041] 在一个实施例中，如图1所示，提供了一种命名实体识别方法，包括以下步骤：

[0042] 步骤S220，获取待识别自然语言信息。

[0043] 其中，待识别自然语言信息是需要进行命名实体识别的自然语言信息。

[0044] 步骤S240，对待识别自然语言信息进行转化处理，获得字向量信息。

[0045] 在一个实施例中，对待识别自然语言信息进行转化处理，获得字向量信息的步骤，包括：对待识别自然语言信息进行转化处理，获得数字标识信息；将数字标识信息输入ALBERT模型进行编码，获得字向量信息。

[0046] 其中，首先将输入的待识别自然语言信息转化为对应的数字标识信息，即使用数字标号代替文本，方便后续的计算，具体地：通过建立字符与数字的对应字典，通过查字典将字符序列(即待识别自然语言信息)转化为数字序列(即数字标识信息)，从含义上来说，字符序列与数字序列是同等含义。将数字标识信息输入ALBERT(A little bert)模型，输出字向量信息，ALBERT(A little bert)模型是一种语言模型，是bert模型的改进模型之一，可根据自然语言序列的前后文对序列内的字信息进行编码，输出字向量信息；将一个数字序列作为输入，就会输出一个512维向量。

[0047] 步骤S260，基于预先构建的关键词典，对待识别自然语言信息中的词汇进行检索，

确定存在标签信息的关键词汇。

[0048] 其中，预先构建的关键词典是业务中人工总结的实体词典，词典中应包含词与其对应标签的标签信息，如在挖掘节气实体时，“谷雨”就可以被归纳为节气实体词典。检索待识别自然语言信息中是否出现过词典内词汇，若存在，则确定为是存在标签信息的关键词汇，则记录该关键词汇、该关键词汇在待识别自然语言信息中出现的位置、该关键词汇的标签信息。

[0049] 步骤S280，将关键词汇对应的标签信息进行向量化，获得词典标签信息。

[0050] 在一个实施例中，将关键词汇对应的标签信息进行向量化，获得词典标签信息的步骤，包括：将关键词汇对应的标签信息进行数字化编码，获得数字化标签信息；对数字化标签信息进行向量化，获得词典标签信息。

[0051] 其中，标签信息转化为一个数，然后获取这个数对应的向量。例如：有3个标签A、B、C，随机初始化一个[3,512]的向量，第一行向量代表A，第二行向量代表B，第三行向量代表C。现在输入一个c标签，将其映射到3，然后映射到第三个向量。

[0052] 步骤S300，采用相对位置编码的方式对字向量信息和词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量。

[0053] 在一个实施例中，采用相对位置编码的方式对字向量信息和词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量的步骤，包括：根据关键词汇，获取关键词汇在待识别自然语言信息的位置；根据关键词汇在待识别自然语言信息的位置、字向量信息和词典标签信息，获得蕴含词典标签信息的自然语言字向量。

[0054] 其中，对于待识别自然语言中的字信息，其位置编码为每个字对应的索引位置，如“谷雨在谷雨这一天努力工作”，共12个字，其每个字的编码从前到后分别为0至12。对于词典标签信息，其位置编码对应于该关键词汇在待识别自然语言中出现的位置，如上例中“人名”标签对应于第一个“谷雨”，其位置索引为“[0,1]”。将字向量信息和词典标签信息与关键词汇在待识别自然语言信息的位置共同输入到Transformer模型中，进行信息融合，最终输出蕴含词典标签信息的自然语言字向量。

[0055] 步骤S320，根据蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得待识别自然语言信息中每个字对应的命名实体标签。

[0056] 在一个实施例中，根据蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得待识别自然语言信息中每个字对应的命名实体标签的步骤，包括：将蕴含词典标签信息的自然语言字向量输入条件随机场模型，进行前后文信息挖掘，输出待识别自然语言信息中每个字对应的命名实体标签。

[0057] 其中，为进一步挖掘字向量的前后文信息，蕴含词典标签信息的自然语言字向量还需输入条件随机场(CRF)模型，最终输出与待识别自然语言信息中的每个字对应的实体标签。例如目前一共有3个实体标签，一个字向量512维，将其输入CRF模型中，CRF模型会将其转化为一个3维的向量，输出如[0.1,0.2,0.7]，这三维分别代表着三个标签对应权重，权重最大数字代表的标签即为该字对应的结果标签。

[0058] 上述命名实体识别方法，通过获取待识别自然语言信息；对待识别自然语言信息进行转化处理，获得字向量信息，基于预先构建的关键词典，对待识别自然语言信息中的词汇进行检索，确定存在标签信息的关键词汇，将关键词汇对应的标签信息进行向量化，获得

词典标签信息，在保证运行效率的基础上融合了语义信息，采用相对位置编码的方式对字向量信息和词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量，该种融合方式使得信息的结合更加直接灵活，可对命名体识别的精度与灵活性产生极大的增益，并根据蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得待识别自然语言信息中每个字对应的命名实体标签，从而提高了命名实体识别结果法人准确性。

[0059] 应该理解的是，虽然图1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1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。

[0060] 在一个实施例中，如图2所示，提供了一种命名实体识别装置，包括：信息获取模块310、信息转化模块320、检索模块330、向量化模块340、融合模块350和信息挖掘模块360。

[0061] 信息获取模块310，用于获取待识别自然语言信息。

[0062] 信息转化模块320，用于对待识别自然语言信息进行转化处理，获得字向量信息。

[0063] 检索模块330，用于基于预先构建的关键词典，对待识别自然语言信息中的词汇进行检索，确定存在标签信息的关键词汇。

[0064] 向量化模块340，用于将关键词汇对应的标签信息进行向量化，获得词典标签信息。

[0065] 融合模块350，用于采用相对位置编码的方式对字向量信息和词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量。

[0066] 信息挖掘模块360，用于根据蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得待识别自然语言信息中每个字对应的命名实体标签。

[0067] 在一个实施例中，信息转化模块320还用于：对待识别自然语言信息进行转化处理，获得数字标识信息；将数字标识信息输入ALBERT模型进行编码，获得字向量信息。

[0068] 在一个实施例中，融合模块350还用于：根据关键词汇，获取关键词汇在待识别自然语言信息的位置；根据关键词汇在待识别自然语言信息的位置、字向量信息和词典标签信息，获得蕴含词典标签信息的自然语言字向量。

[0069] 在一个实施例中，融合模块350还用于：将关键词汇在待识别自然语言信息的位置、字向量信息和词典标签信息，输入到Transformer模型中进行信息融合，输出蕴含词典标签信息的自然语言字向量。

[0070] 在一个实施例中，信息挖掘模块360还用于：将蕴含词典标签信息的自然语言字向量输入条件随机场模型，进行前后文信息挖掘，输出待识别自然语言信息中每个字对应的命名实体标签。

[0071] 在一个实施例中，向量化模块340还用于：将关键词汇对应的标签信息进行数字化编码，获得数字化标签信息；对数字化标签信息进行向量化，获得词典标签信息。

[0072] 关于命名实体识别装置的具体限定可以参见上文中对于命名实体识别方法的限定，在此不再赘述。上述命名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。

[0073] 在一个实施例中，提供一种计算机设备，包括存储器和处理器，存储器存储有计算机程序，处理器执行计算机程序时实现上述的命名实体识别方法的步骤。

[0074] 在一个实施例中，提供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上述的命名实体识别方法的步骤。

[0075] 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。

[0076] 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。

[0077] 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。

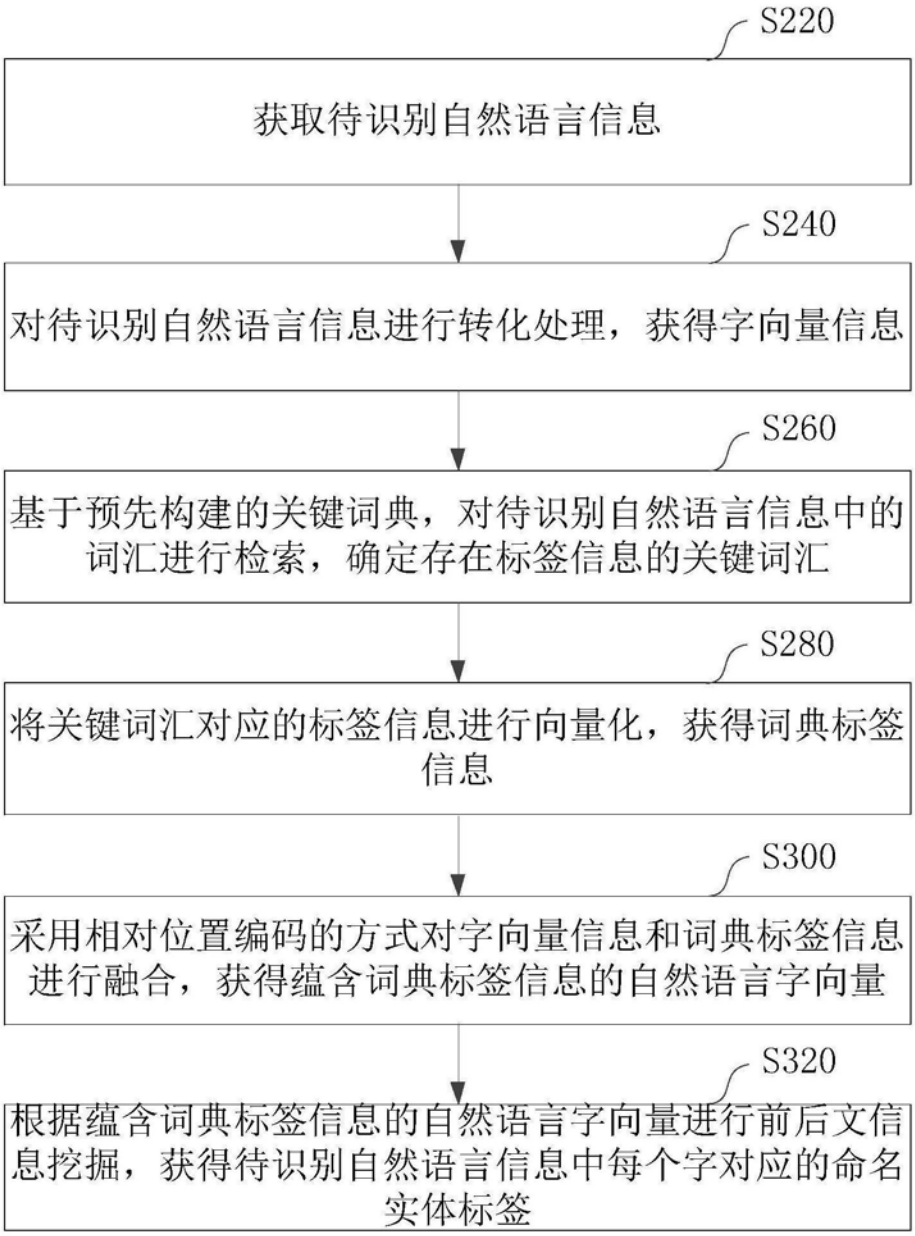

S220
获取待识别自然语言信息
S240
对待识别自然语言信息进行转化处理，获得字向量信息
S260
基于预先构建的关键词典，对待识别自然语言信息中的词汇进行检索，确定存在标签信息的关键词汇
S280
将关键词汇对应的标签信息进行向量化，获得词典标签信息
S300
采用相对位置编码的方式对字向量信息和词典标签信息进行融合，获得蕴含词典标签信息的自然语言字向量
S320
根据蕴含词典标签信息的自然语言字向量进行前后文信息挖掘，获得待识别自然语言信息中每个字对应的命名实体标签

图1

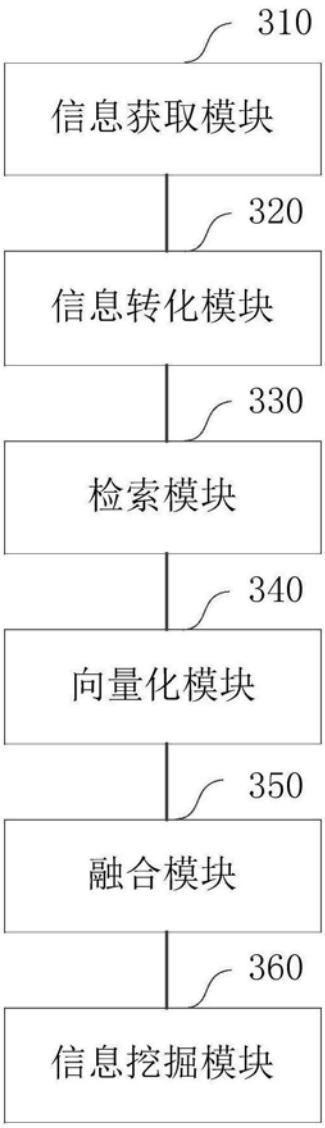


图2