(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 114218948 A
(43)申请公布日 2022.03.22

(21)申请号 202111536935.8

(22)申请日 2021.12.15

(71)申请人 广州华多网络科技有限公司
地址 511442 广东省广州市番禺区南村镇万达广场B1栋24层

(72)发明人 王锋

(74)专利代理机构 广州利能知识产权代理事务所(普通合伙) 44673
代理人 王增鑫

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06F 40/216(2020.01)
G06F 40/242(2020.01)
G06F 16/35(2019.01)
G06F 16/9532(2019.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

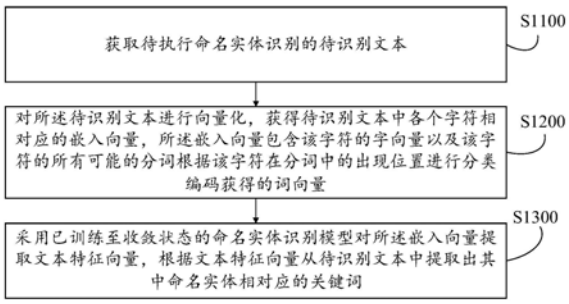
权利要求书2页 说明书14页 附图5页

(54)发明名称

关键词识别方法及其装置、设备、介质、产品

(57)摘要

本申请公开一种关键词识别方法及其装置、设备、介质、产品，所述方法包括：获取待执行命名实体识别的待识别文本；对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量；采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词。本申请在待识别文本的编码过程中根据各个字符在分词中出现的位置不同，对包含字符的分词进行分类编码，改进了编码所得的嵌入向量的语义表示能力，使命名实体识别模型能够提升命名实体识别的准确度。



CN 114218948 A

1.一种关键词识别方法,其特征在于,包括如下步骤:

获取待执行命名实体识别的待识别文本;

对所述待识别文本进行向量化,获得待识别文本中各个字符相对应的嵌入向量,所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量;

采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量,根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词。

2.根据权利要求1所述的关键词识别方法,其特征在于,对所述待识别文本进行向量化,获得待识别文本中各个字符相对应的嵌入向量,包括如下步骤:

将所述待识别文本与预设词典相匹配,获得包含待识别文本中每个字符相对应的所有分词;

针对待识别文本中的每个字符,根据该字符在包含该字符的分词中出现的位置不同,将该字符的所有分词划分为多个分词子集;

针对待识别文本中的每个字符,对该字符相对应的各个分词子集中分词相对应的词向量进行词压缩,获得编码了各个分词子集的词向量的各个分类向量;

拼接待识别文本中的每个字符及其所有分类向量,获得各个字符相对应的嵌入向量。

3.根据权利要求2所述的关键词识别方法,其特征在于,针对待识别文本中的每个字符,根据该字符在包含该字符的分词中出现的位置不同,将该字符的所有分词划分为多个分词子集,包括如下步骤:

针对待识别文本中的每个字符,将首字为该字符的所有分词构造为第一分词子集;

针对待识别文本中的每个字符,将尾字为该字符的所有分词构造为第二分词子集;

针对待识别文本中的每个字符,将中间位置包含该字符的所有分词构造为第三分词子集;

针对待识别文本中的每个字符,将仅包含该字符的分词构造为第四分词子集。

4.根据权利要求2所述的关键词识别方法,其特征在于,针对待识别文本中的每个字符,对该字符相对应的各个分词子集中分词相对应的词向量进行词压缩,获得编码了各个分词子集的词向量的各个分类向量的步骤中,以如下任意一种方式实施所述的压缩:

将分词子集中的所有分词的词向量求均值实现词压缩,获得该分词集合相应的分类向量;

将分词子集中的所有分词的词向量进行加权后求均值实现词压缩,获得该分词集合相应的分类向量。

5.根据权利要求1所述的关键词识别方法,其特征在于,采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量,根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词,包括如下步骤:

采用所述命名实体模型中的文本特征提取模型提取所述嵌入向量的深层语义信息,获得相应的文本特征向量;

采用所述命名实体模型中的条件随机场模型根据所述文本特征向量进行词性标注,根据词性标注结果从所述待识别文本中提取出多个表示命名实体的关键词。

6.根据权利要求1至5中任意一项所述的关键词识别方法,其特征在于,采用已训练至

收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词的步骤之后，包括如下步骤：

根据所述关键词构造搜索表达式，调用商品搜索引擎获取与该搜索表达式相匹配的商品列表，将该商品列表推送至提供所述待识别文本的搜索请求方。

7.根据权利要求1至5中任意一项所述的关键词识别方法，其特征在于，采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词的步骤之后，包括如下步骤：

利用所述关键词对携带所述待识别文本的商品对象进行标注，使所述关键词构成所述商品对象的画像标签。

8.一种关键词识别装置，其特征在于，包括：

文本获取模块，用于获取待执行命名实体识别的待识别文本；

文本编码模块，用于对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量；

实体识别模块，用于采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词。

9.一种计算机设备，包括中央处理器和存储器，其特征在于，所述中央处理器用于调用运行存储于所述存储器中的计算机程序以执行如权利要求1至6中任意一项所述的方法的步骤。

10.一种计算机可读存储介质，其特征在于，其以计算机可读指令的形式存储有依据权利要求1至6中任意一项所述的方法所实现的计算机程序，该计算机程序被计算机调用运行时，执行相应的方法所包括的步骤。

11.一种计算机程序产品，包括计算机程序/指令，其特征在于，该计算机程序/指令被处理器执行时实现权利要求1至6任意一项中所述方法的步骤。

关键词识别方法及其装置、设备、介质、产品

技术领域

[0001] 本申请涉及电商信息技术领域，尤其涉及一种关键词识别方法及其相应的装置、计算机设备、计算机可读存储介质，以及计算机程序产品。

背景技术

[0002] 命名实体识别在电商领域的搜索、推荐和用户画像分析等都起着重要作用，例如用户搜索时，搜索框对用户搜索的词汇联想，可以引导用户搜索想要搜索的商品词，提高搜索效率，而词汇联想中的联想词需要基于商品库中的商品，命名实体识别方法就可以识别出商品词。根据用户的搜索，识别出搜索关键词的tag，例如商品词、品牌等，可以用户TermWei ght、提高搜索精排的效果。同时用户的搜索商品、点击商品、加购、下单商品登行为，可以使用嵌套实体识别方法识别出用户这些行为的偏好，用于用户画像分析。

[0003] 命名实体识别(NER)可以被当作序列标注问题来解决，其中实体边界和类别标签被联合预测。不同于英文命名实体识别，中文没有明显的单词边界，字符序列下不同分词会有不同含义，导致命名实体识别任务更加困难。一种直观的方法是先分词再进行词级别的序列标注，但这种方法会带来分词的错误传递问题。另一种是直接基于字符级的序列标注，但这种方法忽略了词级别的信息。

[0004] 有很多工作试着通过使用词典来提高Chinese NER的性能。作为代表Lattice LSTM在多种公共Chinese NER数据集上获得新的benchmark。Lattice LSTM结构将词信息融合到字信息中，Lattice LSTM是在LSTM基础上改进而来，它允许节点能接收更远的信息，将模型从链式结构拓展到图式结构，Lattice LSTM将各种分词结果都引入到模型，词信息可以远程传入到节点中，最终路径通过模型探索。Lattice LSTM保留了所有可能的词典匹配结果，避免了启发性选择一个匹配结果给每个字符所引入的错误传播，同时引入了预训练模型，提高了性能。但是，Lattice LSTM的这一实现原理也存在不少问题，例如，其一、Lattice LSTM中，第j个字只能获取以它为结尾的词的信息和它之前的时刻状态的词的信息，但无法获取中间包含该字的词的信息。其二、距离最近的上下文对当前token的预测至关重要，但在Lattice LSTM中，词的表示融合了它之前的所有时刻的状态信息，包括字和词，这些信息会对词本身的含义造成干扰；其三、基于这种编码机制，导致Lattice LSTM训练速度比较慢，需要大量的数据样本才能促使其收敛。

[0005] 因此，至少在命名实体识别相关的模型的编码过程中，仍有可改进的空间。

发明内容

[0006] 本申请的首要目的在于解决上述问题至少之一而提供一种关键词识别方法及其相应的装置、计算机设备、计算机可读存储介质、计算机程序产品。

[0007] 为满足本申请的各个目的，本申请采用如下技术方案：

[0008] 适应本申请的目的之一而提供的一种关键词识别方法，包括如下步骤：

[0009] 获取待执行命名实体识别的待识别文本；

[0010] 对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量；

[0011] 采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词。

[0012] 深化的实施例中，对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，包括如下步骤：

[0013] 将所述待识别文本与预设词典相匹配，获得包含待识别文本中每个字符相对应的所有分词；

[0014] 针对待识别文本中的每个字符，根据该字符在包含该字符的分词中出现的位置不同，将该字符的所有分词划分为多个分词子集；

[0015] 针对待识别文本中的每个字符，对该字符相对应的各个分词子集中分词相对应的词向量进行词压缩，获得编码了各个分词子集的词向量的各个分类向量；

[0016] 拼接待识别文本中的每个字符及其所有分类向量，获得各个字符相对应的嵌入向量。

[0017] 具体化的实施例中，针对待识别文本中的每个字符，根据该字符在包含该字符的分词中出现的位置不同，将该字符的所有分词划分为多个分词子集，包括如下步骤：

[0018] 针对待识别文本中的每个字符，将首字为该字符的所有分词构造为第一分词子集；

[0019] 针对待识别文本中的每个字符，将尾字为该字符的所有分词构造为第二分词子集；

[0020] 针对待识别文本中的每个字符，将中间位置包含该字符的所有分词构造为第三分词子集；

[0021] 针对待识别文本中的每个字符，将仅包含该字符的分词构造为第四分词子集。

[0022] 具体化的实施例中，针对待识别文本中的每个字符，对该字符相对应的各个分词子集中分词相对应的词向量进行词压缩，获得编码了各个分词子集的词向量的各个分类向量的步骤中，以如下任意一种方式实施所述的压缩：

[0023] 将分词子集中的所有分词的词向量求均值实现词压缩，获得该分词集合相应的分类向量；

[0024] 将分词子集中的所有分词的词向量进行加权后求均值实现词压缩，获得该分词集合相应的分类向量。

[0025] 深化的实施例中，采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词，包括如下步骤：

[0026] 采用所述命名实体模型中的文本特征提取模型提取所述嵌入向量的深层语义信息，获得相应的文本特征向量；

[0027] 采用所述命名实体模型中的条件随机场模型根据所述文本特征向量进行词性标注，根据词性标注结果从所述待识别文本中提取出多个表示命名实体的关键词。

[0028] 扩展的一种实施例中，采用已训练至收敛状态的命名实体识别模型对所述嵌入向

量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词的步骤之后，包括如下步骤：

[0029] 根据所述关键词构造搜索表达式，调用商品搜索引擎获取与该搜索表达式相匹配的商品列表，将该商品列表推送至提供所述待识别文本的搜索请求方。

[0030] 扩展的另一实施例中，采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词的步骤之后，包括如下步骤：

[0031] 利用所述关键词对携带所述待识别文本的商品对象进行标注，使所述关键词构成所述商品对象的画像标签。

[0032] 适应本申请的目的之一而提供的一种关键词识别装置，包括：文本获取模块、文本编码模块，以及实体识别模块，其中，所述文本获取模块，用于获取待执行命名实体识别的待识别文本；所述文本编码模块，用于对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量；所述实体识别模块，用于采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词。

[0033] 深化的实施例中，所述文本编码模块，包括：文本分词子模块，用于将所述待识别文本与预设词典相匹配，获得包含待识别文本中每个字符相对应的所有分词；分类构造子模块，用于针对待识别文本中的每个字符，根据该字符在包含该字符的分词中出现的位置不同，将该字符的所有分词划分为多个分词子集；分类压缩子模块，用于针对待识别文本中的每个字符，对该字符相对应的各个分词子集中分词相对应的词向量进行词压缩，获得编码了各个分词子集的词向量的各个分类向量；向量合成子模块，用于拼接待识别文本中的每个字符及其所有分类向量，获得各个字符相对应的嵌入向量。

[0034] 具体化的实施例中，所述分类构造子模块，包括：第一构造单元，用于针对待识别文本中的每个字符，将首字为该字符的所有分词构造为第一分词子集；第二构造单元，用于针对待识别文本中的每个字符，将尾字为该字符的所有分词构造为第二分词子集；第三构造单元，用于针对待识别文本中的每个字符，将中间位置包含该字符的所有分词构造为第三分词子集；第四构造单元，用于针对待识别文本中的每个字符，将仅包含该字符的分词构造为第四分词子集。

[0035] 具体化的实施例中，所述分类压缩子模块中采用如下任意一个单元实现：平均压缩单元，用于将分词子集中的所有分词的词向量求均值实现词压缩，获得该分词集合相应的分类向量；加权压缩单元，用于将分词子集中的所有分词的词向量进行加权后求均值实现词压缩，获得该分词集合相应的分类向量。

[0036] 深化的实施例中，所述实体识别模块，包括：表示学习子模块，用于采用所述命名实体模型中的文本特征提取模型提取所述嵌入向量的深层语义信息，获得相应的文本特征向量；实体提取子模块，用于采用所述命名实体模型中的条件随机场模型根据所述文本特征向量进行词性标注，根据词性标注结果从所述待识别文本中提取出多个表示命名实体的关键词。

[0037] 扩展的一种实施例中，本申请的关键词识别装置，还包括：搜索执行模块，用于根

据所述关键词构造搜索表达式，调用商品搜索引擎获取与该搜索表达式相匹配的商品列表，将该商品列表推送至提供所述待识别文本的搜索请求方。

[0038] 扩展的另一实施例中，本申请的关键词识别装置，还包括：标注执行模块，用于利用所述关键词对携带所述待识别文本的商品对象进行标注，使所述关键词构成所述商品对象的画像标签。

[0039] 适应本申请的目的之一而提供的一种计算机设备，包括中央处理器和存储器，所述中央处理器用于调用运行存储于所述存储器中的计算机程序以执行本申请所述的关键词识别方法的步骤。

[0040] 适应本申请的另一目的而提供的一种计算机可读存储介质，其以计算机可读指令的形式存储有依据所述的关键词识别方法所实现的计算机程序，该计算机程序被计算机调用运行时，执行该方法所包括的步骤。

[0041] 适应本申请的另一目的而提供的一种计算机程序产品，包括计算机程序/指令，该计算机程序/指令被处理器执行时实现本申请任意一种实施例中所述方法的步骤。

[0042] 相对于现有技术，本申请的优势如下：

[0043] 首先，本申请主要对需要进行命名实体识别的待识别文本的编码过程进行改进，对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，使所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量，通过分类加强各个分类的共性特征的表示，从而指导命名识别模型的表示学习过程，使模型表示学习所获得的深层语义信息更能准确表示各个字之间的语义相关性，从而由该模型基于更为精准的深层语义信息进行词性标注，在词性标注的基础上准确识别出各个命名实体，获得相对应的关键词。

[0044] 其次，本申请在编码过程中，考虑到中文没有明显的单词边界的特点，字符序列下不同分词会有不同含义，因此根据各个字符在分词的具体出现位置进行分类，由此使各个分类表现了字符所处位置所承载的含义，更符合中文的特点，因此，据其编码获得的嵌入向量，可使命名实体识别模型更易于对中文进行命名实体的识别，使基于中文的命名实体识别任务更加高效准确。

[0045] 其次，由于针对编码过程进行改进提升了命名实体识别的准确性，因此，可使相应的命名实体模型的训练过程所需的数据样本的总量降低，使模型更易被训练至收敛状态，提升模型训练效率，节省模型训练成本。

[0046] 此外，本申请基于准确性的提升而从待识别文本中获得关键词，所述的关键词可用于在线搜索、搜索关键词联想、数据画像、推荐搜索等场景中，能够获得语义更为精准的匹配效果。

附图说明

[0047] 本申请上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

[0048] 图1为本申请的关键词识别方法的典型实施例的流程示意图；

[0049] 图2为本申请实施例中待识别文本的编码过程的流程示意图；

[0050] 图3为本申请实施例中根据字符位置进行编码的过程的流程示意图；

[0051] 图4为本申请实施例中命名实体识别模型进行识别的过程的流程示意图；

[0052] 图5为本申请的关键词识别方法的扩展实施例之一的流程示意图；

[0053] 图6为本申请的关键词识别方法的另一扩展实施例的流程示意图；

[0054] 图7为本申请的关键词识别装置的原理框图；

[0055] 图8为本申请所采用的一种计算机设备的结构示意图。

具体实施方式

[0056] 下面详细描述本申请的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本申请，而不能解释为对本申请的限制。

[0057] 本技术领域技术人员可以理解，除非特意声明，这里使用的单数形式“一”、“一个”、“所述”和“该”也可包括复数形式。应该进一步理解的是，本申请的说明书中使用的措辞“包括”是指存在所述特征、整数、步骤、操作、元件和/或组件，但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、组件和/或它们的组。应该理解，当我们称元件被“连接”或“耦接”到另一元件时，它可以直接连接或耦接到其他元件，或者也可以存在中间元件。此外，这里使用的“连接”或“耦接”可以包括无线连接或无线耦接。这里使用的措辞“和/或”包括一个或更多个相关联的列出项的全部或任一单元和全部组合。

[0058] 本技术领域技术人员可以理解，除非另外定义，这里使用的所有术语(包括技术术语和科学术语)，具有与本申请所属领域中的普通技术人员的一般理解相同的意义。还应该理解的是，诸如通用字典中定义的那些术语，应该被理解为具有与现有技术的上下文中的意义一致的意义，并且除非像这里一样被特定定义，否则不会用理想化或过于正式的含义来解释。

[0059] 本技术领域技术人员可以理解，这里所使用的“客户端”、“终端”、“终端设备”既包括无线信号接收器的设备，其仅具备无发射能力的无线信号接收器的设备，又包括接收和发射硬件的设备，其具有能够在双向通信链路上，进行双向通信的接收和发射硬件的设备。这种设备可以包括：蜂窝或其他诸如个人计算机、平板电脑之类的通信设备，其具有单线路显示器或多线路显示器或没有多线路显示器的蜂窝或其他通信设备；PCS(Personal Communications Service，个人通信系统)，其可以组合语音、数据处理、传真和/或数据通信能力；PDA(Personal Digital Assistant，个人数字助理)，其可以包括射频接收器、寻呼机、互联网/内联网访问、网络浏览器、记事本、日历和/或GPS(Global Positioning System，全球定位系统)接收器；常规膝上型和/或掌上型计算机或其他设备，其具有和/或包括射频接收器的常规膝上型和/或掌上型计算机或其他设备。这里所使用的“客户端”、“终端”、“终端设备”可以是便携式、可运输、安装在交通工具(航空、海运和/或陆地)中的，或者适合于和/或配置为在本地运行，和/或以分布形式，运行在地球和/或空间的任何其他位置运行。这里所使用的“客户端”、“终端”、“终端设备”还可以是通信终端、上网终端、音乐/视频播放终端，例如可以是PDA、MID(Mobile Internet Device，移动互联网设备)和/或具有音乐/视频播放功能的移动电话，也可以是智能电视、机顶盒等设备。

[0060] 本申请所称的“服务器”、“客户端”、“服务节点”等名称所指向的硬件，本质上是具备个人计算机等效能力的电子设备，为具有中央处理器(包括运算器和控制器)、存储器、输

入设备以及输出设备等冯诺依曼原理所揭示的必要构件的硬件装置，计算机程序存储于其存储器中，中央处理器将存储在外存中的程序调入内存中运行，执行程序中的指令，与输入输出设备交互，借此完成特定的功能。

[0061] 需要指出的是，本申请所称的“服务器”这一概念，同理也可扩展到适用于服务器机群的情况。依据本领域技术人员所理解的网络部署原理，所述各服务器应是逻辑上的划分，在物理空间上，这些服务器既可以是互相独立但可通过接口调用的，也可以是集成到一台物理计算机或一套计算机机群的。本领域技术人员应当理解这一变通，而不应以此约束本申请的网络部署方式的实施方式。

[0062] 本申请的一个或数个技术特征，除非明文指定，既可部署于服务器实施而由客户端远程调用获取服务器提供的在线服务接口来实施访问，也可直接部署并运行于客户端来实施访问。

[0063] 本申请中所引用或可能引用到的神经网络模型，除非明文指定，既可部署于远程服务器且在客户端实施远程调用，也可部署于设备能力胜任的客户端直接调用，某些实施例中，当其运行于客户端时，其相应的智能可通过迁移学习来获得，以便降低对客户端硬件运行资源的要求，避免过度占用客户端硬件运行资源。

[0064] 本申请所涉及的各种数据，除非明文指定，既可远程存储于服务器，也可存储于本地终端设备，只要其适于被本申请的技术方案所调用即可。

[0065] 本领域技术人员对此应当知晓：本申请的各种方法，虽然基于相同的概念而进行描述而使其彼此间呈现共通性，但是，除非特别说明，否则这些方法都是可以独立执行的。同理，对于本申请所揭示的各个实施例而言，均基于同一发明构思而提出，因此，对于相同表述的概念，以及尽管概念表述不同但仅是为了方便而适当变换的概念，应被等同理解。

[0066] 本申请即将揭示的各个实施例，除非明文指出彼此之间的相互排斥关系，否则，各个实施例所涉的相关技术特征可以交叉结合而灵活构造出新的实施例，只要这种结合不背离本申请的创造精神且可满足现有技术中的需求或解决现有技术中的某方面的不足即可。对此变通，本领域技术人员应当知晓。

[0067] 本申请的一种关键词识别方法，可被编程为计算机程序产品，部署于客户端或服务器中运行而实现，例如在本申请的电商平台应用场景中，一般部署在服务器中实施，藉此可以通过访问该计算机程序产品运行后开放的接口，通过图形用户界面与该计算机程序产品的进程进行人机交互而执行该方法。

[0068] 请参阅图1，本申请的关键词识别方法在其典型实施例中，包括如下步骤：

[0069] 步骤S1100、获取待执行命名实体识别的待识别文本：

[0070] 所述的待识别文本可以从用户请求中获取，也可以从数据库例如商品数据库中商品对象的商品标题或本地存储空间中例如系统粘贴板中调用，视本申请所应用的下游任务的不同而灵活获取输入文本作为所述的待识别文本，用于执行命名实体识别，从待识别文本中提取出一个或多个命名实体相对应的关键词。

[0071] 所述的下游任务，视本申请的技术方案的具体应用场景而定，例如在线搜索、搜索关键词联想、数据画像、推荐搜索等。

[0072] 示例之一，在线搜索场景中，用户在客户端设备提交其原始搜索表达式后，该原始搜索表达式被作为所述的待识别文本进行命名实体识别后提取出多个关键词，再根据所述

关键词进行逻辑结合之后构成优化后的搜索表达式再执行在线搜索。

[0073] 示例之二，搜索关键词联想场景中，用户在输入其搜索关键词的过程中，其已输入的文本被作为所述的待识别文本进行命名实体识别，然后后台根据命名实体识别获得的关键词进行词语联想，将相关联想词显然给用户选定。

[0074] 示例之三，数据画像场景中，以电商领域为例，可以直接从客户端设备或者从商品数据库中获取商品对象的商品标题或商品详情等文本信息，作为所述的待识别文本，进行命名实体识别，获得相应的关键词，对相应的商品对象进行标注，将这些关键词作为商品对象的数据画像标签。

[0075] 示例之四，推荐搜索场景中，仍以电商领域为例，可以获取客户端设备的输入作为待识别文本，对其进行命名实体识别，根据识别所获得的关键词在商品数据库中搜索获取其相对应的商品对象，然后构造出商品对象推荐列表推送给用户端。

[0076] 诸如此类，表明可以根据其应用场景的不同而获取相应输入的文本信息来作为所述的待识别文本，以便启动本申请的命名识别过程。实际上，命名实体识别是自然语言处理技术中的一个项基础功能，可以广泛地适用于多种场合，对此，本领域技术人员均已知晓，因此，不应以应用领域为限约束本申请的创造精神所应涵盖的保护范围。

[0077] 较佳的情况下，鉴于本申请所实现的编码原理所体现出的优势，所述待识别文本为中文文本，或者其他能够以单字独立表示字义而以多字组合所成的词汇表示独立含义的语种。

[0078] 步骤S1200、对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量：

[0079] 本步骤旨在对所述待识别文本进行编码，实现向量化，而获得相应的嵌入向量。具体而言，对于所述待识别文本，以中文为例，针对其中的每个字符先分别进行编码。

[0080] 在针对每个字符进行编码时，先根据预设的词典获得每个字符相对应的分词集，每个分词集中包含其相应的字符在待识别文本中的所有可能的分词。例如，以“中山西路”的“山”字为例，其根据词典确定出的分词集中可能包含{“中山”；“中山西”；“中山西路”；“山西”；“山西路”；“山”}等分词。通过该示例可以看出，单字“山”在其分词集的各个分词中的出现位置，存在多种位置情况，包括出现于首字位置、尾字位置、中间字位置、独字位置，其中，为理解的便利，首字位置可标记为B，尾字位置标记为E、中间字位置标记为M、独字位置标记为S。根据这几种不同出现位置，便可根据出现的不同构造出不同的分类，相应获得：

[0081] B＝{“山西”；“山西路”}

[0082] E＝{“中山”}

[0083] M＝{“中山西”；“中山西路”}

[0084] S＝{“山”}

[0085] 由此可见，根据每个字符在其可能的分词中所出现的位置的不同，可以划分出多个相应的分词子集。然后针对不同的分词子集获取词向量进行综合，综合的方式包括对各个词向量求均值或求加权值，再将综合获得的向量表示结合到该字符的字向量中，构成该字符相应的向量表示，由各个字符的向量表示构造成待识别文本相对应的嵌入向量。这一过程中，根据所述各个字符在其可能的分词所出现的位置不同获得多个分词子集，然后针

对每个分词子集获得其中各个分词的词向量的综合结果，这一综合结果独立表示了相应位置的分词的共性特征，最后通过这些共性特征叠加到字向量中，构成单字相对应的向量表示，使每个单字的向量表示又综合了各个出现位置相对应的共性特征，据此，通过这些共性特征的表示，突出了不同出现位置相对应的含义的表征，使最终获得的待识别文本的嵌入向量综合了全方面的语义信息，后续据此进行命名实体识别，自然可以获得更丰富的语义。

[0086] 步骤S1300、采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词：

[0087] 本申请可采用现有技术中各种已知的命名实体识别模型用于实施命名实体的识别，此类模型通常是基于Transformer底层网络架构实现的，包括但不限于集成了Lattice LSTM基础模型和条件随机场模型的网络架构、由Bert独立承担或Bert模型结合条件随机场模型所搭建的网络架构等，因这些模型用于实施命名实体识别的技术已为本领域技术人员所熟知，故恕不赘述。当然，所述的命名实体模型在投入本申请使用之前应被事先训练至收敛状态，使其习得相应的能力，该能力使得该模型能根据本申请前一步骤所揭示的原理对待识别文本进行编码获得的嵌入向量进行表示学习获得深层语义信息相对应的文本特征向量，并在该文本特征向量的基础上实现词性标注，根据词性标注结果获得各个命名实体相对应的关键词。

[0088] 通过所述的命名实体识别模型提取出待识别文本的过程中，由于经本申请编码获得的嵌入向量根据字符在分词中的出现位置表现了相对应的语义特征，其语义信息更为丰富，对应的高层语义向量所包含的语义也更为丰富，对于中文这一类单字及多字结合均可表义的语言而言，能够实现良好的表示学习效果，因此，能够更准确地对此类语种的文本实现命名实体识别。

[0089] 通过本实施例的介绍，可以看出，本申请具有丰富的积极效果，包括但不限于如下各个方面：

[0090] 首先，本申请主要对需要进行命名实体识别的待识别文本的编码过程进行改进，对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，使所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量，通过分类加强各个分类的共性特征的表示，从而指导命名识别模型的表示学习过程，使模型表示学习所获得的深层语义信息更能准确表示各个字之间的语义相关性，从而由该模型基于更为精准的深层语义信息进行词性标注，在词性标注的基础上准确识别出各个命名实体，获得相对应的关键词。

[0091] 其次，本申请在编码过程中，考虑到中文没有明显的单词边界的特点，字符序列下不同分词会有不同含义，因此根据各个字符在分词的具体出现位置进行分类，由此使各个分类表现了字符所处位置所承载的含义，更符合中文的特点，因此，据其编码获得的嵌入向量，可使命名实体识别模型更易于对中文进行命名实体的识别，使基于中文的命名实体识别任务更加高效准确。

[0092] 其次，由于针对编码过程进行改进提升了命名实体识别的准确性，因此，可使相应的命名实体模型的训练过程所需的数据样本的总量降低，使模型更易被训练至收敛状态，提升模型训练效率，节省模型训练成本。

[0093] 此外，本申请基于准确性的提升而从待识别文本中获得关键词，所述的关键词可

用于在线搜索、搜索关键词联想、数据画像、推荐搜索等场景中，能够获得语义更为精准的匹配效果。

[0094] 请参阅图2，深化的实施例中，所述步骤S1200、对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，包括如下步骤：

[0095] 步骤S1210、将所述待识别文本与预设词典相匹配，获得包含待识别文本中每个字符相对应的所有分词：

[0096] 命名实体识别模型对应设置有一个词典，该词典中包含各种指代命名实体的词汇。为了实现待识别文本的词嵌入，先根据该词典，基于待识别文本的每一字符，从该词典中匹配出包含该字符的所有可能的分词，针对每个字符构造出其分词集，该分词集中包含出现该字符的所有可能的分词。

[0097] 步骤S1220、针对待识别文本中的每个字符，根据该字符在包含该字符的分词中出现的位置不同，将该字符的所有分词划分为多个分词子集：

[0098] 接下来针对待识别文本中的每个字符，对其相对应的分词集进行划分出多个分词子集，划分的依据参阅前一实施例所述，具体可按照如下公式所表示：

[0099] $$B(c_i)=\{w_{i,k},\forall w_{i,k}\in L,i<k\leq n\}$$

[0100] $$E(c_i)=\{w_{j,i},\forall w_{j,i}\in L,1\leq j<i\}$$

[0101] $$M(c_i)=\{w_{j,k},\forall w_{j,k}\in L,1\leq j<i<k\leq n\}$$

[0102] $$S(c_i)=\{c_i,\exists c_i\in L\}$$

[0103] 其中c_i表示具体字符，L表示命名实体识别模型相对应的预设词典；n为待识别文本的总长度，i、j、k均表现该具体字符在待识别文本中的出现位置的序号，w表示由其两个下标所限定的待识别文本中位置所对应的分词，B、M、E、S分别如前所述，表示首字位置、尾字位置、中间字位置、独字位置。

[0104] 请参阅图3，一个实施例中，根据此处的公式示例，可以将本步骤实现为包括如下的具体步骤：

[0105] 步骤S1221、针对待识别文本中的每个字符，将首字为该字符的所有分词构造为第一分词子集：

[0106] 针对每个字符，将首字为该字符的所有分词提取出来，构造为第一分词子集，例如前文示例中的：B＝{“山西”；“山西路”}。

[0107] 步骤S1222、针对待识别文本中的每个字符，将尾字为该字符的所有分词构造为第二分词子集：

[0108] 针对每个字符，将尾字为该字符的所有分词提取出来，构造为第二分词子集，例如前文示例中的：E＝{“中山”}

[0109] 步骤S1223、针对待识别文本中的每个字符，将中间位置包含该字符的所有分词构造为第三分词子集：

[0110] 针对每个字符，将中间位置包含该字符的所有分词提取出来，构造为第三分词子集，例如前文示例中的：M＝{“中山西”；“中山西路”}

[0111] 步骤S1224、针对待识别文本中的每个字符，将仅包含该字符的分词构造为第四分

词子集：

[0112] 针对每个字符，将仅包含该字符的分词也即其独字对应的分词提取出来，构造为第四分词子集，例如前文示例中的：S＝{“山”}。

[0113] 到此，便完成了根据公式对每个字符相对应的分词的分类，使每个字符均对应有B、M、E、S四个集合(允许空集)，获得相对应的多个分词子集，后续可根据各个分词子集获得其中的分词对应的词向量。

[0114] 步骤S1230、针对待识别文本中的每个字符，对该字符相对应的各个分词子集中分词相对应的词向量进行词压缩，获得编码了各个分词子集的词向量的各个分类向量：

[0115] 针对待识别文本中的每个字符，可针对其各个分词子集进行词压缩，以获得每个分词子集相对应的分类向量。具体而言，获取所述各个分词子集中各个分词相应的词向量，基于同一分词子集的各个词向量，对该分词子集进行词向量压缩，具体的压缩方式可以采用如下任意一种：

[0116] 可选的一种压缩方式中，将分词子集中的所有分词的词向量求均值实现词压缩，获得该分词集合相应的分类向量：

[0117] 此种方式主要是基于求均值原理实现的，请参考公式：

[0118] $$v(S)=\frac{1}{|S|}\sum_{w\epsilon S}e^{w}(w)$$

[0119] 该公式中，S表示具体分词子集，w表示分词子集中的分词，e^{w}表示分词w的词向量。

[0120] 根据该公式，可以理解，通过将分词子集中的各个分词的词向量进行加和求均值，可以获得该分词子集相应的分类向量，即v^{S}，从而实现对该具体分词子集中分词的词向量的综合表示，通过该分类向量实现对相应的字符在该分词子集中各分词所出现的位置的共性进行表征。

[0121] 可选的另一种压缩方式中，将分词子集中的所有分词的词向量进行加权后求均值实现词压缩，获得该分词集合相应的分类向量：

[0122] 此种方式主要是基于求加权均值原理实现的，请参考公式：

[0123] $$v(S)=\frac{1}{|S|}\sum_{w\epsilon S}z(w)e^{w}(w)$$

[0124] 比较此一公式与前一公式不难理解，两者不同之处仅在于此一压缩方式中，为各个词向量匹配了相应的权重z(w)。较佳的实施例中，此处所匹配的权重是静态权重，使用每个分词在一个静态数据集上出现的频率作为权重。采用这种方式匹配权重，能够使得命名实体识别模型在训练阶段能加快训练速率。在匹配权重的过程中，对于同一分词集合的权重的归一化，应考虑整个分词集合的情况。一般来说，所述静态数据集来自于训练数据和验证数据。其中，如果包含子串w的子串a被匹配，w的频率无需增加。

[0125] 至此，采用加权求均值的方式，也便相应为每个字符获得其各个分词子集相对应的分类向量，每个字符的各个分类向量可分别表示为：$v^{S}(B)$、$v^{S}(E)$、$v^{S}(M)$、$v^{S}(S)$。

[0126] 步骤S1240、拼接待识别文本中的每个字符及其所有分类向量，获得各个字符相对应的嵌入向量：

[0127] 最后，令公式为：

[0128] $e^s(B,M,E,S)=,v^s(B);v^s(E);v^s(M);v^s(S)-$

[0129] 继而：

[0130] $x^c\leftarrow,x^c;e^s(B,M,E,S)-$

[0131] 也即，针对所述待识别文本中的每个字符，将其字向量及其分词子集相对应的所有分类向量简单拼接，即可获得各个字符相对应的嵌入向量，各个字符相对应的嵌入向量的组合也就构成了待识别文本的嵌入向量，从而获得了待识别文本的向量化表示，完成对待识别文本的编码过程。

[0132] 本实施例，适应命名实体模型进行表示学习所需，详细揭示了对待识别文本进行编码的具体过程，可以看出，通过分类向量实现对各个分词子集的词向量的压缩，抽取了各分词子集有关于字符位置信息的共性特征的语义，有助于提升表示学习过程中的语义表示效果，从而为命名实体识别奠定可靠的基础。在实践中，具体的编码过程可以做灵活的变动，例如，允许每个字符相对应的个别分词子集为空集，而不必要求每个分词子集都包含至少一个分词。

[0133] 请参阅图4，深化的实施例中，所述步骤S1300、采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词，包括如下步骤：

[0134] 步骤S1310、采用所述命名实体模型中的文本特征提取模型提取所述嵌入向量的深层语义信息，获得相应的文本特征向量:

[0135] 所述的文本特征提取模型，优选采用Lattice LSTM实施，该模型基于预先编码获得的所述待识别文本的嵌入向量，参考上下文对其进行表示学习，获得相应的文本特征向量。

[0136] 步骤S1320、采用所述命名实体模型中的条件随机场模型根据所述文本特征向量进行词性标注，根据词性标注结果从所述待识别文本中提取出多个表示命名实体的关键词：

[0137] 继续，将所述文本特征向量输入条件随机场模型(CRF)进行词性标注，在条件随机场模型的作用下，用Lattice LSTM输出的概率矩阵和CRF的状态转移矩阵结合进行预测，完成词性标注，根据词性标注结果即可提取出待识别文本中的多个命名实体相对应的关键词。

[0138] 同理，LSTM也可替换为Bert之类的基于Transformer内核的模型，尽管这些模型本身也可独立担当词性标注的任务，但结合条件随机场能显著提升命名实体提取的准确率，因此予以推荐。

[0139] 本实施例中，通过给出更为具体的模型结构，更进一步地表明本申请的技术方案的实用性，实践中，采用这样的模型，根据本申请的编码获得的待识别文本的嵌入向量进行命名实体的识别，更为准确，并且，采用这一嵌入向量，在模型的训练过程中，能够提升模型收敛速度，从而节省训练成本。

[0140] 请参阅图5，扩展的一种实施例中，所述步骤S1300、采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词的步骤之后，包括如下步骤：

[0141] 步骤S1400、根据所述关键词构造搜索表达式，调用商品搜索引擎获取与该搜索表

达式相匹配的商品列表，将该商品列表推送至提供所述待识别文本的搜索请求方：

[0142] 本实施例中，适应电商领域在线搜索应用场景的需要，所述的待识别文本为用户输入的原始搜索表达式，用户在其终端界面完成所述待识别文本的输入，确认提交后，被服务器获得，服务器根据本申请的技术方案，通过前述各个实施例的相关步骤，对所述待识别文本进行编码获得其相应的嵌入向量，然后采用命名实体识别模型根据该嵌入向量预测出命名实体，获得相应的关键词。在此基础上，根据预设规则将各个关键词进行组合，例如将各个关键词设置为逻辑与运算，构成新的搜索表达式。根据修正后的搜索表达式，调用商品搜索引擎，对商品数据库执行搜索，获得匹配该搜索表达式的商品列表，然后将该商品列表推送至作为请求方的所述用户的客户设备中显示。

[0143] 本实施例将本申请其他实施例的技术方案进一步应用于电商领域的在线搜索应用场景中，展示出本申请的技术优势。不难理解，由于本申请在先介绍的有益效果的存在，其经命名识别获得的关键词更为精准，因此，根据预设逻辑结合规则确定出来的新的搜索表达式，以其在商品数据库中进行搜索获得的商品列表，更能匹配用户在待识别文本中表达的意图，实现商品的精准搜索。

[0144] 请参阅图6，扩展的另一实施例中，所述步骤S1300、采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词的步骤之后，包括如下步骤：

[0145] 步骤S1500、利用所述关键词对携带所述待识别文本的商品对象进行标注，使所述关键词构成所述商品对象的画像标签：

[0146] 本实施例中，适应电商领域对商品数据库中的商品对象进行数据画像的需要，以所述商品对象的商品标题和/或商品详情文本等作为确定数据画像标签的素材，作为所述的待识别文本，然后，通过前述各个实施例的相关步骤，对所述待识别文本进行编码获得其相应的嵌入向量，采用命名实体识别模型根据该嵌入向量预测出命名实体，获得相应的关键词。这些关键词便可作为对所述商品对象进行数据画像所需的标签，直接用这些标签对所述商品对象进行标注，便可完成商品对象的数据画像。

[0147] 本实施例将本申请其他实施例的技术方案进一步应用于电商领域的在线搜索应用场景中，展示出本申请的技术优势。不难理解，由于本申请在先介绍的有益效果的存在，其经命名识别获得的关键词更为精准，因此，将这些关键词作为商品对象的数据画像标签，可以快速高效且精准地实现对商品对象的数据画像标注。

[0148] 请参阅图7，适应本申请的目的之一而提供的一种关键词识别装置，是对本申请的关键词识别方法的功能化体现，该装置包括：文本获取模块1100、文本编码模块1200，以及实体识别模块1300，其中，所述文本获取模块1100，用于获取待执行命名实体识别的待识别文本；所述文本编码模块1200，用于对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量；所述实体识别模块1300，用于采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词。

[0149] 深化的实施例中，所述文本编码模块1200，包括：文本分词子模块，用于将所述待识别文本与预设词典相匹配，获得包含待识别文本中每个字符相对应的所有分词；分类构

造子模块，用于针对待识别文本中的每个字符，根据该字符在包含该字符的分词中出现的位置不同，将该字符的所有分词划分为多个分词子集；分类压缩子模块，用于针对待识别文本中的每个字符，对该字符相对应的各个分词子集中分词相对应的词向量进行词压缩，获得编码了各个分词子集的词向量的各个分类向量；向量合成子模块，用于拼接待识别文本中的每个字符及其所有分类向量，获得各个字符相对应的嵌入向量。

[0150] 具体化的实施例中，所述分类构造子模块，包括：第一构造单元，用于针对待识别文本中的每个字符，将首字为该字符的所有分词构造为第一分词子集；第二构造单元，用于针对待识别文本中的每个字符，将尾字为该字符的所有分词构造为第二分词子集；第三构造单元，用于针对待识别文本中的每个字符，将中间位置包含该字符的所有分词构造为第三分词子集；第四构造单元，用于针对待识别文本中的每个字符，将仅包含该字符的分词构造为第四分词子集。

[0151] 具体化的实施例中，所述分类压缩子模块中采用如下任意一个单元实现：平均压缩单元，用于将分词子集中的所有分词的词向量求均值实现词压缩，获得该分词集合相应的分类向量；加权压缩单元，用于将分词子集中的所有分词的词向量进行加权后求均值实现词压缩，获得该分词集合相应的分类向量。

[0152] 深化的实施例中，所述实体识别模块1300，包括：表示学习子模块，用于采用所述命名实体模型中的文本特征提取模型提取所述嵌入向量的深层语义信息，获得相应的文本特征向量；实体提取子模块，用于采用所述命名实体模型中的条件随机场模型根据所述文本特征向量进行词性标注，根据词性标注结果从所述待识别文本中提取出多个表示命名实体的关键词。

[0153] 扩展的一种实施例中，本申请的关键词识别装置，还包括：搜索执行模块，用于根据所述关键词构造搜索表达式，调用商品搜索引擎获取与该搜索表达式相匹配的商品列表，将该商品列表推送至提供所述待识别文本的搜索请求方。

[0154] 扩展的另一实施例中，本申请的关键词识别装置，还包括：标注执行模块，用于利用所述关键词对携带所述待识别文本的商品对象进行标注，使所述关键词构成所述商品对象的画像标签。

[0155] 为解决上述技术问题，本申请实施例还提供计算机设备。如图8所示，计算机设备的内部结构示意图。该计算机设备包括通过系统总线连接的处理器、计算机可读存储介质、存储器和网络接口。其中，该计算机设备的计算机可读存储介质存储有操作系统、数据库和计算机可读指令，数据库中可存储有控件信息序列，该计算机可读指令被处理器执行时，可使得处理器实现一种关键词识别方法。该计算机设备的处理器用于提供计算和控制能力，支撑整个计算机设备的运行。该计算机设备的存储器中可存储有计算机可读指令，该计算机可读指令被处理器执行时，可使得处理器执行本申请的关键词识别方法。该计算机设备的网络接口用于与终端连接通信。本领域技术人员可以理解，图8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。

[0156] 本实施方式中处理器用于执行图7中的各个模块及其子模块的具体功能，存储器存储有执行上述模块或子模块所需的程序代码和各类数据。网络接口用于向用户终端或服

务器之间的数据传输。本实施方式中的存储器存储有本申请的关键词识别装置中执行所有模块/子模块所需的程序代码及数据，服务器能够调用服务器的程序代码及数据执行所有子模块的功能。

[0157] 本申请还提供一种存储有计算机可读指令的存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行本申请任一实施例的关键词识别方法的步骤。

[0158] 本申请还提供一种计算机程序产品，包括计算机程序/指令，该计算机程序/指令被一个或多个处理器执行时实现本申请任一实施例所述方法的步骤。

[0159] 本领域普通技术人员可以理解实现本申请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，该计算机程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，前述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory，ROM)等计算机可读存储介质，或随机存储记忆体(Random Access Memory，RAM)等。

[0160] 综上所述，本申请在待识别文本的编码过程中根据各个字符在分词中出现的位置不同，对包含字符的分词进行分类编码，改进了编码所得的嵌入向量的语义表示能力，使命名实体识别模型能够据此而提升命名实体识别的准确度，特别适合处理中文文本，据此获得的关键词可以提升多种下游任务的执行效果。

[0161] 本技术领域技术人员可以理解，本申请中已经讨论过的各种操作、方法、流程中的步骤、措施、方案可以被交替、更改、组合或删除。进一步地，具有本申请中已经讨论过的各种操作、方法、流程中的其他步骤、措施、方案也可以被交替、更改、重排、分解、组合或删除。进一步地，现有技术中的具有与本申请中公开的各种操作、方法、流程中的步骤、措施、方案也可以被交替、更改、重排、分解、组合或删除。

[0162] 以上所述仅是本申请的部分实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本申请的保护范围。

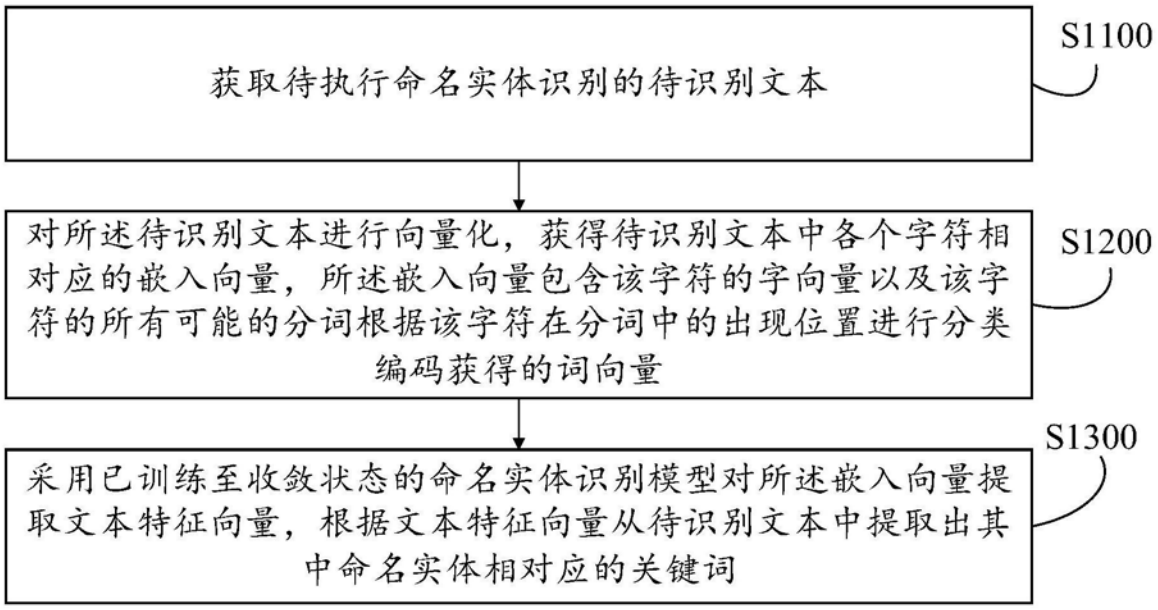

获取待执行命名实体识别的待识别文本
S1100
对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量
S1200
采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词
S1300

图1

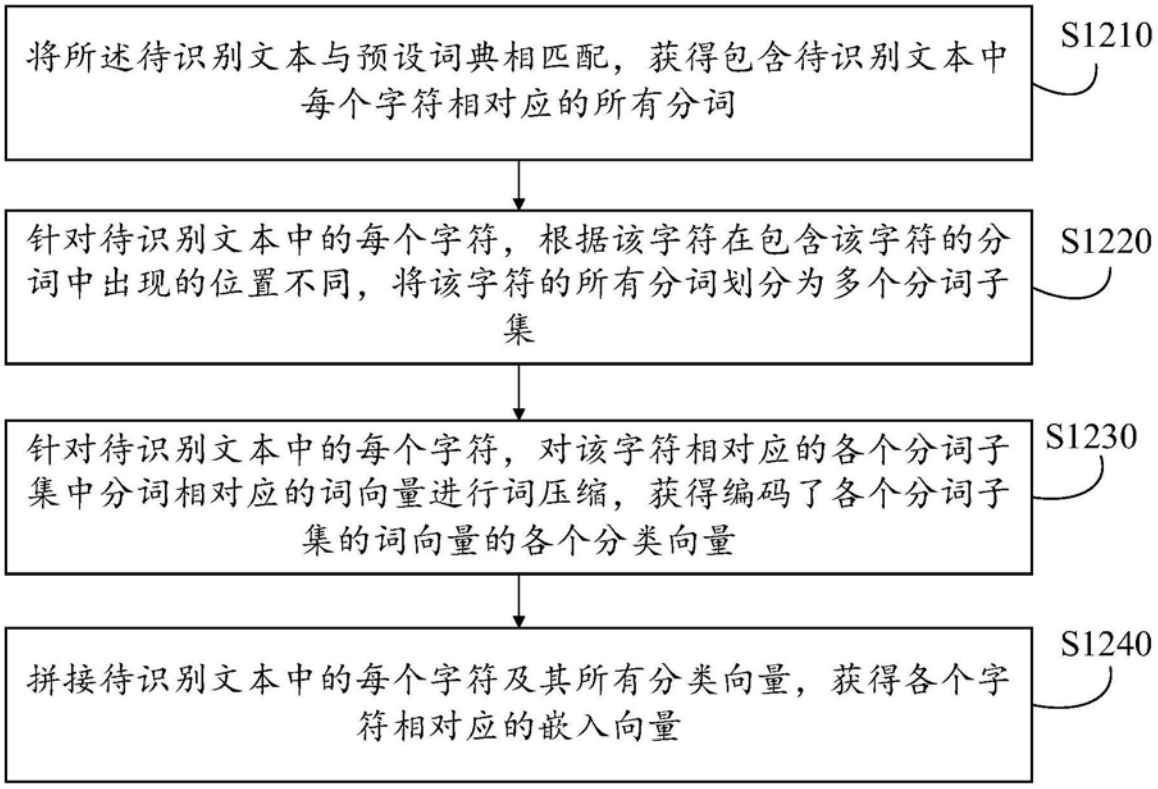

将所述待识别文本与预设词典相匹配，获得包含待识别文本中每个字符相对应的所有分词
S1210
针对待识别文本中的每个字符，根据该字符在包含该字符的分词中出现的位置不同，将该字符的所有分词划分为多个分词子集
S1220
针对待识别文本中的每个字符，对该字符相对应的各个分词子集中分词相对应的词向量进行词压缩，获得编码了各个分词子集的词向量的各个分类向量
S1230
拼接待识别文本中的每个字符及其所有分类向量，获得各个字符相对应的嵌入向量
S1240

图2

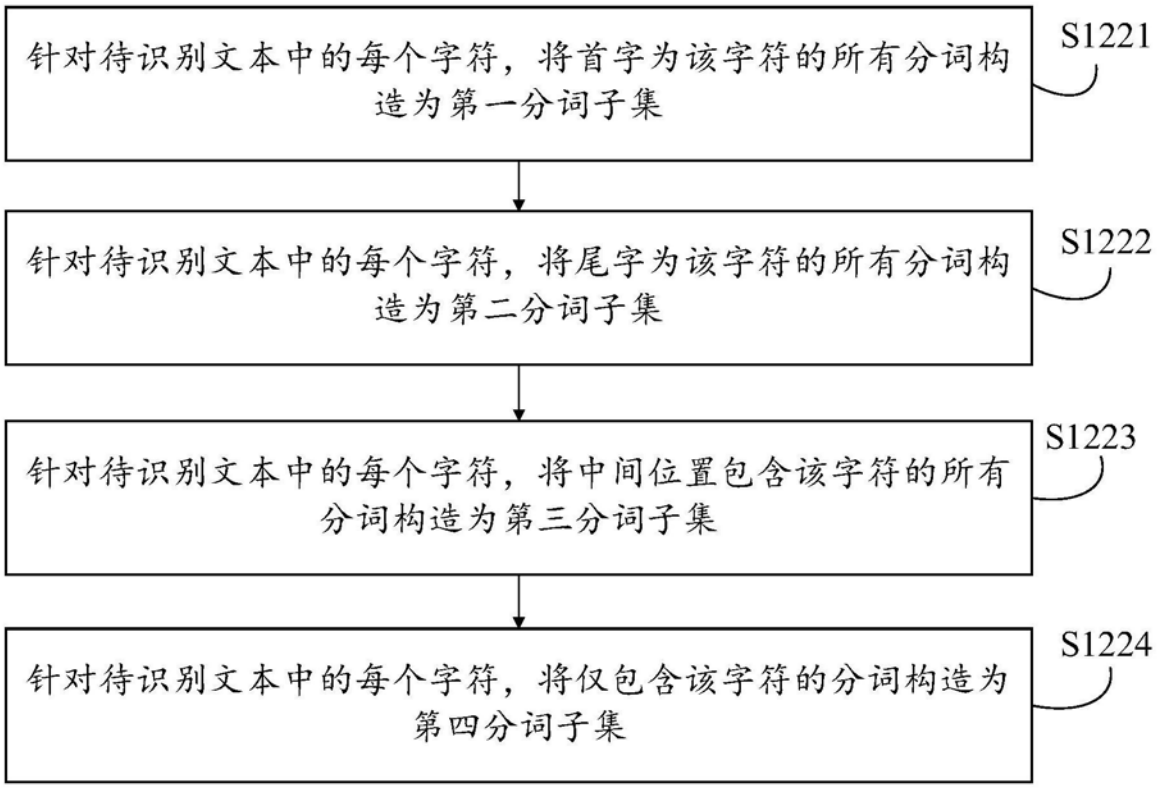

针对待识别文本中的每个字符，将首字为该字符的所有分词构造为第一分词子集
S1221
针对待识别文本中的每个字符，将尾字为该字符的所有分词构造为第二分词子集
S1222
针对待识别文本中的每个字符，将中间位置包含该字符的所有分词构造为第三分词子集
S1223
针对待识别文本中的每个字符，将仅包含该字符的分词构造为第四分词子集
S1224

图3

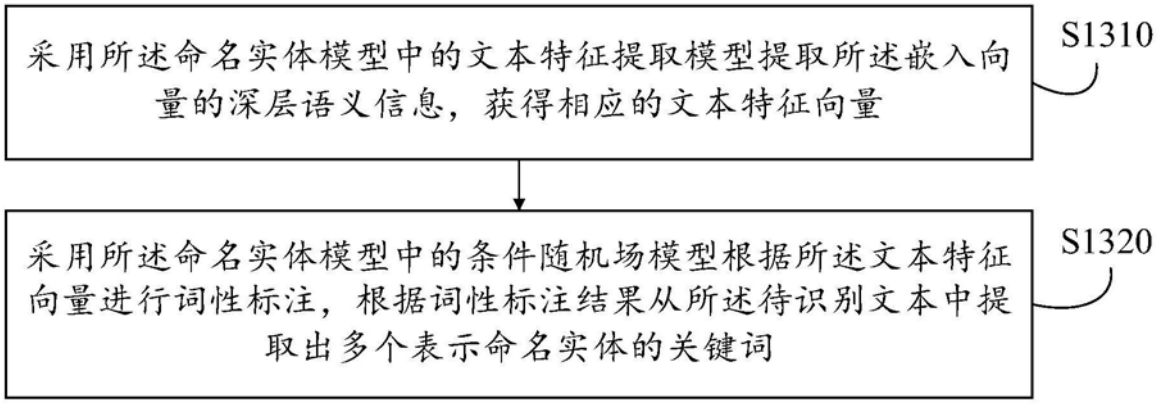

采用所述命名实体模型中的文本特征提取模型提取所述嵌入向量的深层语义信息，获得相应的文本特征向量
S1310
采用所述命名实体模型中的条件随机场模型根据所述文本特征向量进行词性标注，根据词性标注结果从所述待识别文本中提取出多个表示命名实体的关键词
S1320

图4

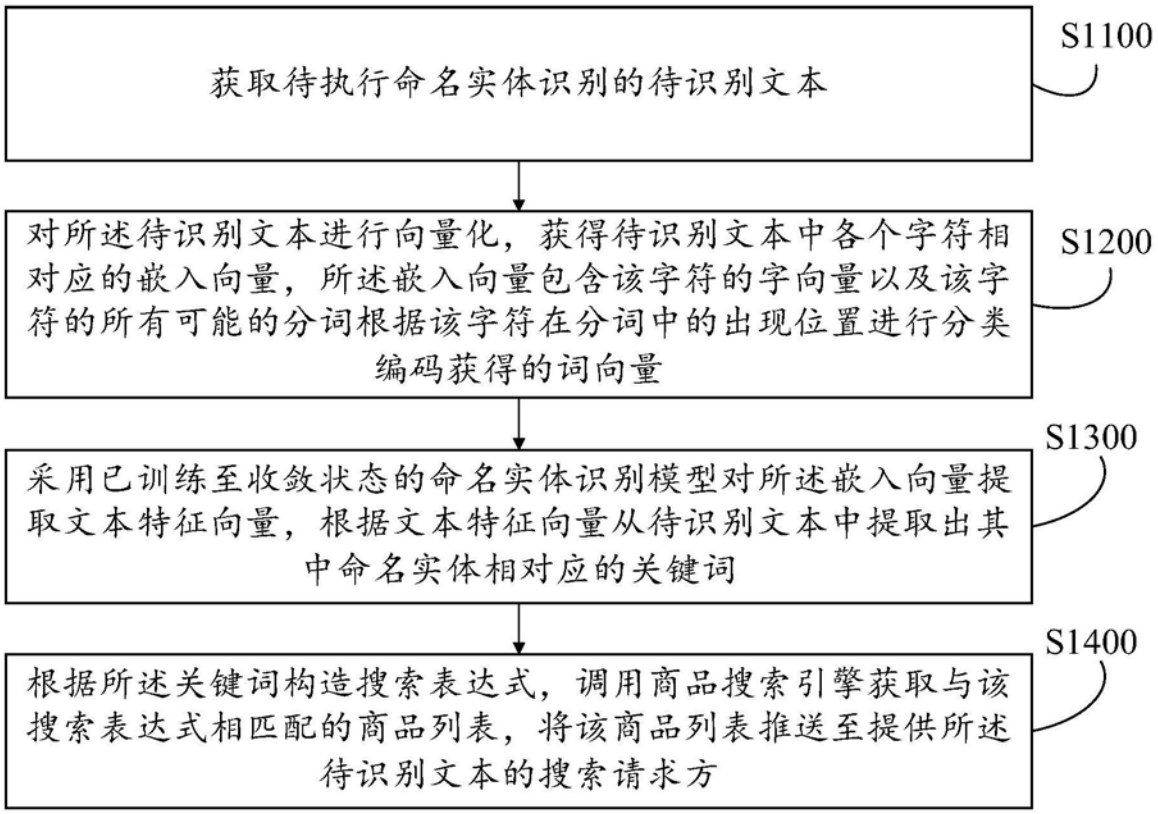

获取待执行命名实体识别的待识别文本
S1100
对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量
S1200
采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词
S1300
根据所述关键词构造搜索表达式，调用商品搜索引擎获取与该搜索表达式相匹配的商品列表，将该商品列表推送至提供所述待识别文本的搜索请求方
S1400

图5

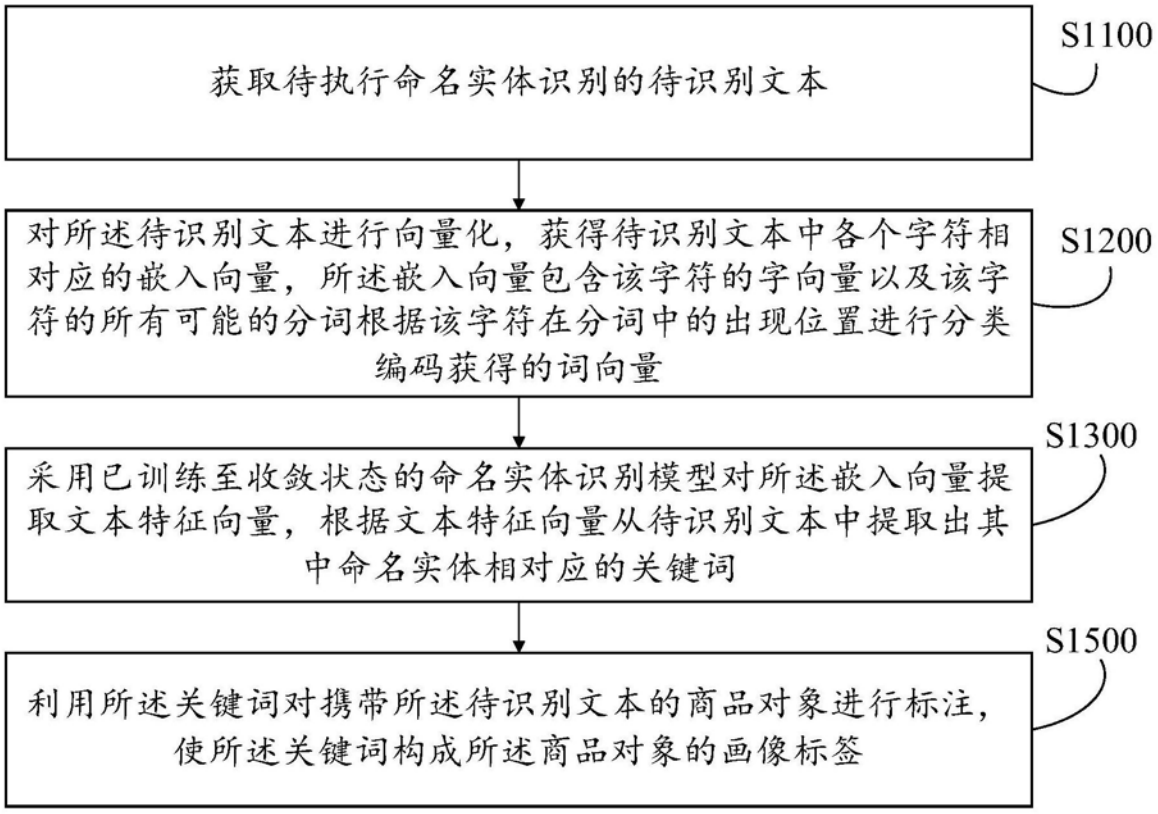

获取待执行命名实体识别的待识别文本
S1100
对所述待识别文本进行向量化，获得待识别文本中各个字符相对应的嵌入向量，所述嵌入向量包含该字符的字向量以及该字符的所有可能的分词根据该字符在分词中的出现位置进行分类编码获得的词向量
S1200
采用已训练至收敛状态的命名实体识别模型对所述嵌入向量提取文本特征向量，根据文本特征向量从待识别文本中提取出其中命名实体相对应的关键词
S1300
利用所述关键词对携带所述待识别文本的商品对象进行标注，使所述关键词构成所述商品对象的画像标签
S1500

图6

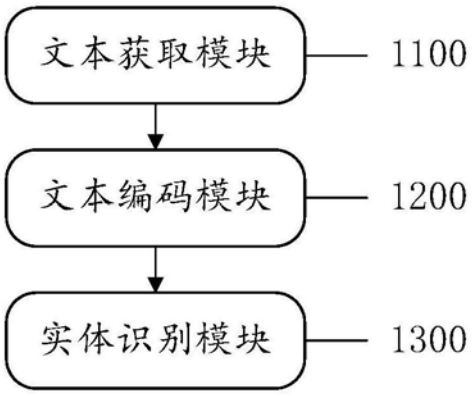

文本获取模块
1100
文本编码模块
1200
实体识别模块
1300

图7

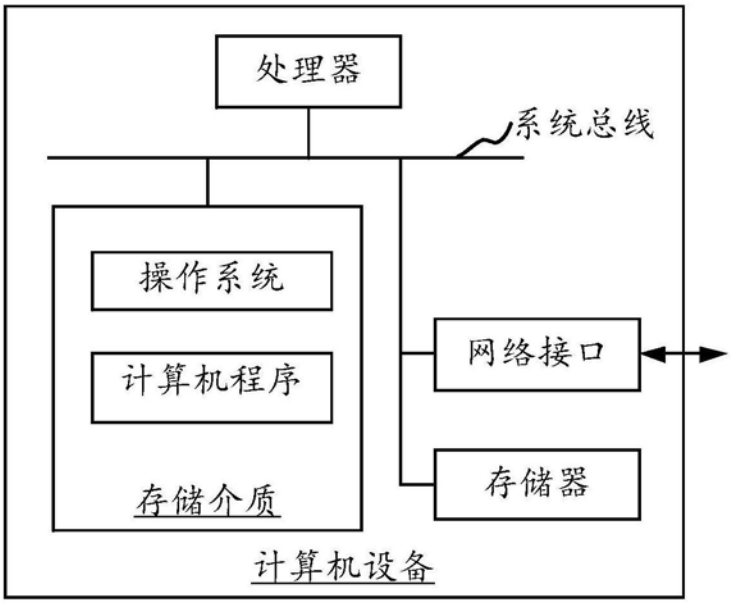


图8