

中文病历的实体识别方法、装置、设备及存储介质

本申请申明享有 2019 年 4 月 19 日递交的申请号为 201910316061.1、名称为“中文病
5 历的实体识别方法、装置、设备及存储介质”的中国专利申请的优先权，该中国专利申请的
整体内容以参考的方式结合在本申请中。

技术领域

本发明涉及自然语言处理领域，涉及一种中文病历的实体识别方法、装置、设备及存
10 储介质。

背景技术

目前对于命名实体识别在病例上的应用的需求很大，比如对病例的查询、搜索、整理
等。

15 发明人意识到，现有基于深度学习的中文命名实体识别的效果很难提升，而且之前都
是应用在其他语言上，比如英语。因为深度学习模型的限制和各个语言间语言特性的不同，
这使命名实体任务在中文上的应用受到了限制。又因为通用领域、其他领域与医疗领域之
间的差异，使其在医疗领域中病例的方向上的应用有所限制。

发明内容

20 本发明要解决的技术问题是为了克服现有技术中基于深度学习的中文命名实体识别准
确率不高的问题，提出了一种中文病历的实体识别方法、装置、设备及存储介质，通过对
中文病例中的文本内容抽取相应的特征转换成特征向量，然后将特征向量作为模型的输入，
以提高实体识别的准确率。

25 本发明是通过下述技术方案来解决上述技术问题：

一种中文病历的实体识别方法，包括以下步骤：

识别出所述中文病历中包含的个人信息,根据第一对应规则输出与所述个人信息对应的
30 第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量；

利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应
规则对应每个字输出用于表征每个字在词组中位置的第三特征向量；

识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与所述每个
字的偏旁对应的第三特征向量；

对所述中文病历做 n-gram 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量；

5 利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量；

根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量和所述第五特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集；

将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

10 本发明还公开了一种中文病历的实体识别装置，包括：

第一特征向量生成模块，用于识别出所述中文病历中包含的个人信息,根据第一对应规则输出与所述个人信息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量；

15 第二特征向量生成模块，用于利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的所述第二特征向量；

第三特征向量生成模块，用于识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与所述每个字的偏旁对应的第三特征向量；

20 第四特征向量生成模块，用于对所述中文病历做 n-gram 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量；

第五特征向量生成模块，用于利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量；

25 向量集生成模块，用于根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量和所述第五特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集；

实体识别模型，用于将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

30 本发明还公开了一种计算机设备，包括存储器和处理器，所述存储器上存储有计算机程序，所述计算机程序被所述处理器执行时实现如下步骤：

识别出所述中文病历中包含的个人信息,根据第一对应规则输出与所述个人信息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量；

利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的所述第二特征向量；

识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与所述每个字的偏旁对应的第三特征向量；

- 5 对所述中文病历做 n-gram 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量；

利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量；

- 10 根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量和所述第五特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集；

将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

- 15 本发明还公开了一种计算机可读存储介质，所述计算机可读存储介质内存储有计算机程序，所述计算机程序可被至少一个处理器所执行，以实现如下步骤：

识别出所述中文病历中包含的个人信息,根据第一对应规则输出与所述个人信息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量；

利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的所述第二特征向量；

- 20 识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与所述每个字的偏旁对应的第三特征向量；

对所述中文病历做 n-gram 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量；

- 25 利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量；

根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量和所述第五特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集；

- 30 将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

本发明通过先识别出中文病历中的实体信息转换成特征向量，然后将中文病历整体转换成的向量集作为模型的输入，以提高模型对实体抽取的准确率。

附图说明

图 1 示出了本发明中文病历的实体识别方法一实施例的流程图；

图 2 示出了本发明中文病历的实体识别装置一实施例的结构图；

5 图 3 示出了本发明计算机设备一实施例的硬件架构示意图。

具体实施方式

下面通过实施例的方式进一步说明本发明，但并不因此将本发明限制在所述的实施例范围之中。

10 首先，本发明提出一种中文病历的实体识别方法。

在一实施例中，如图 1 所示，所述的中文病历的实体识别方法包括如下步骤：

步骤 01：识别出所述中文病历中包含的个人信息，根据第一对应规则输出与所述个人信息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量。

15 识别个人信息具体可采用正则表达式匹配。所谓正则表达式是对字符串操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。

例如这里要识别个人信息，就要先创建一个用于匹配个人信息的正则表达式，具体的表达式根据使用的计算机程序设计语言而定，每种语言都定义有一套字符表达方法；然后运行创建的正则表达式与所述病人基本信息进行匹配，就可以识别出所述病人基本信息中
20 包含的个人信息了。

这里所述个人信息主要指病人类型和病人年龄，选择识别这些个人信息，是因为基本每份病例中都能体现出病人类型（男、女、青年、老人、青年、儿童、婴儿等）和病人年龄，而对应不同的病人类型和病人年龄，医生可能会采取相对应的治疗手段和检查检验的方式，因此基于病人类型和病人年龄的识别有利于对病历的分析，这里选择从病人基本信息中识别出病人类型和病人年龄这两个特征。
25

由于病人基本信息包括病人类型和病人年龄两种，因此这里第一对应规则也有两种，分别是病人类型与特征向量的对应规则和病人年龄与特征向量的对应规则。

具体地，病人类型与特征向量的对应规则包括两种：其一，特征向量的长度等于所述病人类型的种类数量；所述特征向量中每一维度对应所述病人类型中的一个种类；所述特征
30 向量通过所述病人类型对应维度的向量值的改变表征对应的所述病人类型。其二，特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述病人类型。

下面以病人类型为五种“男、女、老人、婴儿、儿童”为例，说明两种病人类型与特

征向量的对应规则：

对应第一种对应规则，特征向量的长度为 5，假设初始的特征向量为[0, 0, 0, 0, 0]，特征向量中每个维度对应一种病人类型，假设该特征向量中的每个维度从前往后依次对应的病人类型为“男、女、老人、婴儿、儿童”。基于上述对应规则，如果病人基本信息中识别出病人类型为“男”，那么初始的特征向量中对应病人类型“男”的第一维度的特征值由 0 变为 1，即用特征向量[1, 0, 0, 0, 0]表示病人类型“男”；如果识别出病人类型为“老人”，那么用特征向量[0, 0, 1, 0, 0]来表示。

对应第二种对应规则，特征向量的长度为 1，即初始的特征向量为[0]，用不同的数字对应五种病人类型，这里假设用数字 1, 2, 3, 4, 5 依次对应病人类型“男、女、老人、婴儿、儿童”。基于上述对应原则，如果病人基本信息中识别出病人类型为“男”，那么初始的特征向量的特征值由 0 变为 1，即用特征向量[1]表示病人类型“男”；如果识别出病人类型为“老人”，那么用特征向量[3]来表示。

病人年龄与特征向量的对应规则具体为：特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述病人年龄，所述向量值等于所述病人年龄。

下面以病人年龄为“78 岁”为例，具体说明病人年龄与特征向量的对应规则。根据该对应规则，特征向量的长度为 1，即初始的特征向量为[0]，通过识别到病人年龄为 78，那么将初始的特征向量的向量值由 0 改为 78，即用特征向量[78]表示病人年龄为 78 岁。

所述中文病历中每个字对应相同的所述第一特征向量是指，假设识别出某一中文病历中病人为男性，年龄为 78 岁，该中文病历中每个字的特征向量中均包含表征性别为男的第一特征向量[1, 0, 0, 0, 0]或[1]以及表征年龄为 78 岁的第一特征向量[78]。

步骤 02：利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的第二个特征向量。

由于是针对中文病历，因此分词工具也对应采用中文分词工具，这里所述分词工具都是现有的，常见的有 jieba、SnowNLP、THULAC、NLPIR 等，不再详述。

通过分词把句子中的单个字、词区分出来，也可以把标点分出来，以便后续实体的识别。

利用分词工具对病历进行分词，以对病历内容“直肠腹膜返折上方未及肿块，结合术前肠镜及病理术中诊断为直肠肛管癌，决定行 Miles 术”分词为例，经过分词得到“直肠腹膜/返折/上方/未及/肿块/, /结合/术前/肠镜/及/病理/术中诊断/为/直肠肛管癌/, /决定/行/Miles 术/”。

所述第二对应规则具体为：特征向量的长度为 4；所述特征向量的前三个维度用于表征包含两个字以上词组，其中第一个维度的向量值的改变用于表征位于所述词组中首位的字，

第二个维度的向量值的改变用于表征位于所述词组中中间的字，第三个维度的向量值的改变用于表征位于所述词组中末尾的字；所述特征向量的第四个维度用于表征单字词组，所述第四个维度的向量值的改变用于表征单字词组中的字。

以四字词组“直肠腹膜”为例，每个字对应有一个初始的特征向量，根据第一对应规则，特征向量的长度为4，因此这里每个字对应的初始的特征向量均为 $[0, 0, 0, 0]$ 。由于词组“直肠腹膜”为四字词组，因此仅用到特征向量的前三个维度。“直”位于该词组的首位，则相应改变初始的特征向量中第一个维度的向量值（由0改为1），即“直”的特征向量为 $[1, 0, 0, 0]$ ；“肠”和“腹”均位于该词组的中间位置，因此这两个字的特征向量相同，都相应改变初始的特征向量中第二个维度的向量值（由0改为1），即“肠”和“腹”的特征向量均为 $[0, 1, 0, 0]$ ；“膜”位于该词组的末位，则相应改变初始的特征向量中第三个维度的向量值（由0改为1），即“膜”的特征向量为 $[0, 0, 1, 0]$ 。

再以单字词组“及”为例，该字对应的初始的特征向量也为 $[0, 0, 0, 0]$ ，由于是单字词组，仅用到特征向量的第四个维度，即改变初始的特征向量中第四个维度的向量值（由0改为1），得到“及”的特征向量为 $[0, 0, 0, 1]$ 。

步骤 03：识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与
所述每个字的偏旁对应的第三特征向量。

首先，通过将所述中文病历中的每个字与预设的偏旁字典进行匹配，输出匹配到的偏旁，从而识别出每个字的偏旁，其中所述偏旁字典包含所有中文字与对应偏旁的关联关系。具体地，预设有一个包含所有中文字与对应偏旁的关联关系的偏旁字典，将所述中文病历中的每个字与预设的偏旁字典进行匹配，输出匹配到的偏旁。

所述第三对应规则具体包括两种：其一，特征向量的长度等于预设的实体偏旁的数量；所述特征向量中每一维度对应一个所述实体偏旁；所述特征向量通过所述实体偏旁对应维度的向量值的改变表征对应的包含所述实体偏旁的字；其二，特征向量的长度为1；所述特征向量通过不同的向量值对应表征包含不同所述实体偏旁的字。

实体偏旁根据具体需要进行预设，比如预设最有效的两种实体偏旁分别是病字框（“疒”）和月字旁（“月”），也可以添加其他偏旁作为实体偏旁，比如（竹字头“竹”、骨字旁“骨”）等。

以预设的实体偏旁包括病字框（“疒”）和月字旁（“月”）两种为例，具体说明两种对应规则。

对应第一种对应规则，特征向量的长度为2，对应初始的特征向量为 $[0, 0]$ ，其中第一个维度的向量值的改变用于表征病字框（“疒”），第二个维度的向量值的改变用于表征月字旁（“月”）。那么，针对“直肠腹膜”中的后三个字都是月字旁（“月”），月字旁（“月”）

属于预设的实体偏旁，因此相应的第二特征向量均是相同的[0, 1]，针对“直肠腹膜”中的第一个字“直”既不是月字旁（“月”）也不是病字框（“疒”），不属于预设的实体偏旁，那么该字的第二特征向量即为初始特征向量[0, 0]；而针对“病理”中的“病”字为病字框（“疒”），那么相应的第二特征向量为[1, 0]。

5 对应第二种对应规则，由于特征向量的长度固定为 1，对应初始的特征向量为[0]，用不同的向量值 1, 2 分别对应表征病字框（“疒”）和月字旁（“月”）。那么，针对“直肠腹膜”中的后三个字都是月字旁（“月”），相应的第二特征向量均是相同的[2]，针对“直肠腹膜”中的第一个字“直”既不是月字旁（“月”）也不是病字框（“疒”），那么该字的第二特征向量即为初始特征向量[0]；而针对“病理”中的“病”字为病字框（“疒”），那么相应的第二特征向量为[1]。

若预设的实体偏旁还包括竹字头（“竹”）和骨字旁（“骨”），对应第一种对应规则时，特征向量的长度即为 4，初始的特征向量为[0, 0, 0, 0]；对应第二种对应规则是，初始的特征向量为[0]，用不同的向量值 1, 2, 3, 4 分别对应表征病字框（“疒”）、月字旁（“月”）、竹字头（“竹”）和骨字旁（“骨”），还可以预设更多的实体偏旁，依次类推，不再赘述。

15 步骤 04：对所述中文病历做 n-gram 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量。

遍历之前需要对中文病历预处理，通常需要将标点符合去除。

20 n 取值等于或小于所述中文病历长度的非零自然数。优选地，n 一般取原始医学字典中最长的词组的长度。对所有小于 n 的自然数都输入病例一次。以 n=5 为例，需要对中文病历先后做 5-gram 遍历、4-gram 遍历、3-gram 遍历、2-gram 遍历、1-gram 遍历。

所谓 n-gram 遍历是自然语言处理的常用方法，实际上就是分词，n 为分词后每个词组的长度，从第一个字开始取前 5 个字组成词组，从第二个字开始取前 5 个字组成词组，依次类推，即每个词组的首字为中文病历中的第 i 个字，每个词组的末字为中文病历中的第 25 i+n-1 个字， $1 \leq i \leq (\text{中文病历长度} - n + 1)$ 。

以中文病历内容“直肠腹膜返折上方未及肿块”为例，经 3-gram 遍历后得到的结果为“直肠腹”、“肠腹膜”、“腹膜返”、“膜返折”、“返折上”、“折上方”、“上方未”、“方未及”、“未及肿”、“及肿块”。

30 所述原始医学词典可以选用现有的任意一种医疗术语词典。在原始医学词典中，每个词组都对应一种实体类别，根据该词组分出的若干个词组在前缀词典和后缀词典的实体类别沿用该词组在原始医学词典中对应的实体类别。

所述前缀词典的构建具体为：识别出所述原始医学词典中多于两个字的词组，依次将

识别出的词组的前 i 个字存入前缀词典， i 为小于该词组长度且大于该词组长度的一半的自然数，其中该词组长度的一半取整数。

以原始医学词典中的词组“左侧粗隆间骨折”为例，该词组的长度为 7，该词组长度的
一半为 3.5，取整数为 3，因此 i 的取值为 4、5、6， $i=6$ 对应“左侧粗隆间骨”， $i=5$ 对
5 应“左侧粗隆间”， $i=4$ 对应“左侧粗隆”，因此根据词组“左侧粗隆间骨折”构建的前
缀词典包含“左侧粗隆间骨”、“左侧粗隆间”、“左侧粗隆”。

所述后缀词典的构建具体为：识别出所述原始医学词典中多于两个字的词组；将识别
出的所述词组的后 i 个字存入后缀词典， i 为小于该词组长度且大于等于该词组长度的一
半的自然数，其中该词组长度的一半取整数。

10 以原始医学词典中的词组“左侧粗隆间骨折”为例，该词组的长度为 7，该词组长度的
一半为 3.5，取整数为 3，因此 i 的取值为 3、4、5、6， $i=6$ 对应“侧粗隆间骨折”， $i=5$
对应“粗隆间骨折”， $i=4$ 对应“隆间骨折”， $i=3$ 对应“间骨折”，因此根据词组“左侧
粗隆间骨折”构建的后缀词典包含“侧粗隆间骨折”、“粗隆间骨折”、“隆间骨折”、“间骨
折”。

15 匹配时，三个词典（原始医学词典、前缀词典和后缀词典）可以同时进行匹配，也可
以设定匹配顺序。无论采用哪种匹配方法，只要词组匹配到其中一个词典后，就停止与其
他两个词典的匹配。这里的匹配要求为完全匹配。匹配结果包括匹配和不匹配两种。当匹
配结果为匹配时，所述匹配结果包含有匹配的词典名词和匹配的医学名词及该医学名词对
应的实体类别

20 第四特征向量用于区分实体类别，特征向量的长度等实体类别的数量，例如共分为六
类实体，分别代表疾病和诊断、症状和体征、身体部位、检查和检验、手术、药物，那么
对应特征向量的长度为 6，则初始的特征向量即为 $[0, 0, 0, 0, 0, 0]$ ，每个维度的向量值对应
一个实体类别。

当匹配结果为不匹配时，对应每个字输出初始的特征向量。

25 当匹配结果为匹配时，则根据相匹配到的词组的实体类别，将该实体类别对应维度的
向量值改变，根据与不同词典中的词组相匹配，改变规则有所不同，因此需要根据所述匹
配结果中包含的词典名词调用适用的对应规则，具体如下：

当与原始医学词典相匹配时，适用的对应规则具体为：特征向量的长度等于所述实体
类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始
30 向量值改为第一向量值、第二向量值或第三向量值对应表征单个字在词组中的首位、中间
位或末位。

下面以词组“直肠腹膜”为例做具体说明，其中第一向量值、第二向量值和第三向量

值分别取 1, 2, 3。该词组出现在原始医学词典, 且与身体部位这一实体类别相关联, 因此这四个字均要改变第三维度的向量值。再根据各个字在词组中的具体位置, “直” 位于该词组的首位, 将第三维度的向量值由 0 改为 1, 即“直”的特征向量为 $[0, 0, 1, 0, 0, 0]$; “肠”和“腹”均位于该词组的中间位置, 因此这两个字的特征向量相同, 均将第三维度的向量值由 0 改为 2, 即“肠”和“腹”的特征向量均为 $[0, 0, 2, 0, 0, 0]$; “膜”位于该词组的末位, 将第三维度的向量值由 0 改为 3, 即“膜”的特征向量为 $[0, 0, 3, 0, 0, 0]$ 。

当与前缀词典相匹配时, 适用的对应规则具体为: 特征向量的长度等于所述实体类别的数量; 所述特征向量中每一维度对应一个所述实体类别; 所述特征向量通过将初始向量值改为第一向量值或第二向量值对应表征单个字在词组中的首位或非首位。

下面以词组“血细胞”为例做具体说明, 其中第一向量值和第二向量值分别取 1, 2。该词组出现在前缀词典中, 且与检查和检验这一实体类别相关联, 因此这三个字均要改变第四维度的向量值。再根据各个字在词组中的具体位置, “血”位于该词组的首位, 将第四维度的向量值由 0 改为 1, 即“血”的特征向量为 $[0, 0, 0, 1, 0, 0]$; “细”和“胞”均位于该词组的非首位, 均将第四维度的向量值由 0 改为 2, 即“细”和“胞”的特征向量均为 $[0, 0, 0, 2, 0, 0]$ 。

当与后缀词典相匹配时, 适用的对应规则具体为: 特征向量的长度等于所述实体类别的数量; 所述特征向量中每一维度对应一个所述实体类别; 所述特征向量通过将初始向量值改为第二向量值或第三向量值对应表征单个字在词组中的非末位或末位。

下面以词组“彩超”为例做具体说明, 其中第二向量值和第三向量值分别取 2, 3。该词组出现在后缀词典中, 且与检查和检验这一实体类别相关联, 因此这三个字要改变第四维度的向量值。再根据各个字在词组中的具体位置, “彩”位于该词组的非末位, 将第四维度的向量值由 0 改为 2, 即“彩”的特征向量为 $[0, 0, 0, 2, 0, 0]$; “超”位于该词组的末位, 将第四维度的向量值由 0 改为 3, 即“超”的特征向量为 $[0, 0, 0, 3, 0, 0]$ 。

需要注意的是, 经过 n-gram 遍历, 会将一段语句按不同的字数划分多次, 因此每个字会得到 n 个特征向量, 但这个特征向量只会有两种可能, 要么不匹配输出的初始的特征向量, 要么匹配输出的相应的特征向量 (各次遍历下匹配输出的相应的特征向量是相同的)。只要有一次匹配, 则最终对应该字输出的是该字相应的特征向量, 除非每次都不匹配, 则最终对应该字输出的是初始的特征向量。

步骤 05: 利用中文拼音转换工具将所述中文病历中每个字转换为拼音, 根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量。

所述中文拼音转换工具为现有技术, 可以用 python 包作为转换工具使用。转化后的拼音可以不表示声调, 也可以用 1、2、3、4 表示声调, 以“匹”为例, 转换成拼音可以是“pi”,

也可以是“pi1”。

所述第五对应规则具体为：特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述拼音。

以“匹”的拼音“pi”为例，根据对应规则，特征向量的长度定义为 1，特征向量的初始值为[0]，每个拼音都预设有的数字编号，假设“pi”对应的数字编号为 20，那么用该 20 替换特征向量中的初始向量值 0，即“pi”对应的特征向量为[20]。

步骤 06：根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量、所述第五特征向量和所述第六特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集。

假设某个字的初始向量为[0]，第一特征向量、第二特征向量、第三特征向量、第四特征向量和第五特征向量分别对应[1a]、[2b]、[3c]、[4d]、[5e]，那么该字最终的特征向量为[0, 1a, 2b, 3c, 4d, 5e]。

如果还有别的特征向量，还可以往后拼接，不限数量；此外，拼接顺序不限于上述所示。

以中文病历中包含如下内容“性别：男……年龄：78……直肠腹膜……”为例，对各特征向量的拼接做进一步说明。

根据步骤 01，该中文病历中每个字对应的第一特征向量有两个，分别是表征性别为男性的[1]（采用第二种病人类型与特征向量的对应规则）和表征年龄为 78 岁的[78]。

根据步骤 02，直肠腹膜为一个四字词组，这四个字对应的第二特征向量分别为[1, 0, 0, 0]、[0, 1, 0, 0]、[0, 1, 0, 0]、[0, 0, 1, 0]。

根据步骤 03，直肠腹膜中后面三个字均为月字旁（“月”），因此这三个字对应的第三特征向量均为[2]。

根据步骤 04，直肠腹膜出现在原始医学词典，且与身体部位这一实体类别相关联，因此这四个字对应的第四特征向量分别为[0, 0, 1, 0, 0, 0]、[0, 0, 2, 0, 0, 0]、[0, 0, 2, 0, 0, 0]、[0, 0, 3, 0, 0, 0]。

根据步骤 05，中文病历中的包含的内容转换为拼音，再根据每个拼音预设的对应数字编号得到第五特征向量，这里假设中文病历中包含的内容“性别：男……年龄：78……直肠腹膜……”中每个字对应的数字编号分别为 7、8、9、10、11、12、13、14、15、16、17，则对应每个字的第五特征向量分别为[7]、[8]、[9]、[10]、[11]、[12]、[13]、[14]、[15]、[16]、[17]。

假设拼接顺序为第一特征向量至第五特征向量依次拼接，且每个字的初始向量均为[0]，那么针对该中文病历中的内容“性别：男……直肠腹膜”这几个字得到的向量集应为

[0, 1, 78, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7]
 [0, 1, 78, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 8] [0, 1, 78, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 9]
 [0, 1, 78, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 14] [0, 1, 78, 0, 1, 0, 0, 0, 2, 0, 0, 0, 2, 0, 0, 0, 15]
 [0, 1, 78, 0, 1, 0, 0, 0, 2, 0, 0, 0, 2, 0, 0, 0, 16] [0, 1, 78, 0, 0, 1, 0, 2, 0, 0, 0, 3, 0, 0, 0, 17]。

5 步骤 07：将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

这里所述模型是指深度神经网络模型，例如双向 LSTM+CRF，也可以是传统的机器学习模型。模型训练时，给模型定义输入向量和对应的输出值，模型经过训练后，一旦识别到输入的向量集中包含某段特定的向量值，模型就可以识别出特定的实体特征。例如输入向量集

10 [0, 1, 78, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7] [0, 1, 78, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 8]
 [0, 1, 78, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 9] [0, 1, 78, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 14]
 [0, 1, 78, 0, 1, 0, 0, 0, 2, 0, 0, 0, 2, 0, 0, 0, 15] [0, 1, 78, 0, 1, 0, 0, 0, 2, 0, 0, 0, 2, 0, 0, 0, 16]
 [0, 1, 78, 0, 0, 1, 0, 2, 0, 0, 0, 3, 0, 0, 0, 17]，根据不同维度上的向量值，可以识别出这向量集表征的是一 78 岁男性的中文病历，并且最后四个特征向量表征的是一个身体部位的四字词组。

15 本实施例通过先识别出中文病历中的实体转换成特征向量，最终生成向量集作为模型的输入，可以有效提高模型对实体提取的准确率。

其次，本发明提出了一种中文病历的实体识别装置，所述装置 20 可以被分割为一个或者多个模块。

20 例如，图 2 示出了所述中文病历的实体识别装置 20 一实施例的结构图，该实施例中，所述装置 20 可以被分割为第一特征向量生成模块 201、第二特征向量生成模块 202、第三特征向量生成模块 203、第四特征向量生成模块 204、第五特征向量生成模块 205、向量集生成模块 206 和实体识别模型 207。以下描述将具体介绍所述模块 201-207 的具体功能。

25 所述第一特征向量生成模块 201，用于识别出所述中文病历中包含的个人信息，根据第一对应规则输出与所述个人信息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量。

所述第二特征向量生成模块 202 用于利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的

30 第二特征向量。

所述第三特征向量生成模块 203 用于识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与每个字的偏旁对应的第三特征向量。

所述第四特征向量生成模块 204 用于对所述中文病历做 n-gram 遍历，将遍历后得到各

个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量。

所述第五特征向量生成模块 205 用于利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量。

5 所述向量集生成模块 206 用于根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量和所述第五特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集。

所述实体识别模型 207 用于将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

10

再次，本发明还提出一种计算机设备。

参阅图 3 所示，是本发明计算机设备一实施例的硬件架构示意图。本实施例中，所述计算机设备 2 是一种能够按照事先设定或者存储的指令，自动进行数值计算和/或信息处理的设备。例如，可以是智能手机、平板电脑、笔记本电脑、台式计算机、机架式服务器、刀片式服务器、塔式服务器或机柜式服务器（包括独立的服务器，或者多个服务器所组成的服务器集群）等。如图所示，所述计算机设备 2 至少包括，但不限于，可通过系统总线相互通信连接存储器 21、处理器 22 以及网络接口 23。其中：

所述存储器 21 至少包括一种类型的计算机可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器 21 可以是所述计算机设备 2 的内部存储单元，例如该计算机设备 2 的硬盘或内存。在另一些实施例中，所述存储器 21 也可以是所述计算机设备 2 的外部存储设备，例如该计算机设备 2 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。当然，所述存储器 21 还可以既包括所述计算机设备 2 的内部存储单元也包括其外部存储设备。本实施例中，所述存储器 21 通常用于存储安装于所述计算机设备 2 的操作系统和各类应用软件，例如用于实现所述中文病历的实体识别方法的计算机程序等。此外，所述存储器 21 还可以用于暂时地存储已经输出或者将要输出的各类数据。

30 所述处理器 22 在一些实施例中可以是中央处理器（Central Processing Unit, CPU）、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器 22 通常用于控制所述计算机设备 2 的总体操作，例如执行与所述计算机设备 2 进行数据交互或者通信相关的控制和

处理等。本实施例中，所述处理器 22 用于运行所述存储器 21 中存储的程序代码或者处理数据，例如运行用于实现所述中文病历的实体识别方法的计算机程序等。

所述网络接口 23 可包括无线网络接口或有线网络接口，该网络接口 23 通常用于在所述计算机设备 2 与其他计算机设备之间建立通信连接。例如，所述网络接口 23 用于通过网络将所述计算机设备 2 与外部终端相连，在所述计算机设备 2 与外部终端之间的建立数据传输通道和通信连接等。所述网络可以是企业内部网（Intranet）、互联网（Internet）、全球移动通讯系统（Global System of Mobile communication, GSM）、宽带码分多址（Wideband Code Division Multiple Access, WCDMA）、4G 网络、5G 网络、蓝牙（Bluetooth）、Wi-Fi 等无线或有线网络。

需要指出的是，图 3 仅示出了具有组件 21-23 的计算机设备 2，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

在本实施例中，存储于存储器 21 中的用于实现所述中文病历的实体识别方法的计算机程序可以被一个或多个处理器（本实施例为处理器 22）所执行，以完成以下步骤的操作：

步骤 01：识别出所述中文病历中包含的个人信息，根据第一对应规则输出与所述个人信息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量；

步骤 02：利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的所述第二特征向量；

步骤 03：识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与每个字的偏旁对应的第三特征向量；

步骤 04：对所述中文病历做 n-gram 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量；

步骤 05：利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与每个字的拼音对应的第五特征向量；

步骤 06：根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量、所述第五特征向量和所述第六特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集；

步骤 07：将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

此外，本发明一种计算机可读存储介质，所述计算机可读存储介质为非易失性可读存储介质，其内存储有计算机程序，所述计算机程序可被至少一个处理器所执行，以实现上述中文病历的实体识别方法或装置的操作。

其中，计算机可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘等。在一些实施例中，计算机可读存储介质可以是计算机设备的内部存储单元，例如
5 该计算机设备的硬盘或内存。在另一些实施例中，计算机可读存储介质也可以是计算机设备的外部存储设备，例如该计算机设备上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。当然，计算机可读存储介质还可以既包括计算机设备的内部存储单元也包括其外部存储设备。本实施例中，计算机可读存储介质通常用于存储安装于计算机设备的操作系统和各类应用软件，例如前
10 述用于实现所述中文病历的实体识别方法的计算机程序等。此外，计算机可读存储介质还可以用于暂时地存储已经输出或者将要输出的各类数据。

虽然以上描述了本发明的具体实施方式，但是本领域的技术人员应当理解，这仅是举例说明，本发明的保护范围是由所附权利要求书限定的。本领域的技术人员在不背离本发明的原理和实质的前提下，可以对这些实施方式做出多种变更或修改，但这些变更和修改均落入本发明的保护范围。

权利要求书

1. 一种中文病历的实体识别方法，包括以下步骤：

识别出所述中文病历中包含的个人信息,根据第一对应规则输出与所述个人信息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量；

利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的所述第二特征向量；

识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与所述每个字的偏旁对应的第三特征向量；

对所述中文病历做 n-gram 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量；

利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量；

根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量和所述第五特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集；

将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

2. 根据权利要求 1 所述的中文病历的实体识别方法，所述识别出所述中文病历中包含的个人信息包括以下步骤：创建用于匹配个人信息的正则表达式；用所述正则表达式与所述病人基本信息进行匹配，以识别出所述病人基本信息中包含的个人信息；所述第一对应规则包括病人类型与特征向量的对应规则和病人年龄与特征向量的对应规则；

所述病人类型与特征向量的对应规则包括：特征向量的长度等于所述病人类型的种类数量；所述特征向量中每一维度对应所述病人类型中的一个种类；所述特征向量通过所述病人类型对应维度的向量值的改变表征对应的所述病人类型；

或者，特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述病人类型；

所述病人年龄与特征向量的对应规则包括：特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述病人年龄，所述向量值等于所述病人年龄。

3. 根据权利要求 1 所述的中文病历的实体识别方法，所述第二对应规则包括：特征向量的长度为 4；所述特征向量的前三个维度用于表征包含两个字以上词组，

其中第一个维度的向量值的改变用于表征位于所述词组中首位的字，第二个维度的向量值的改变用于表征位于所述词组中中间的字，第三个维度的向量值的改变用于表征位于所述词组中末尾的字；所述特征向量的第四个维度用于表征单字词组，所述第四个维度的向量值的改变用于表征单字词组中的字。

5 4. 根据权利要求 1 所述的中文病历的实体识别方法，所述识别出所述中文病历中每个字的偏旁具体包括以下步骤：将所述中文病历中的每个字与预设的偏旁字典进行匹配，输出匹配到的偏旁，所述偏旁字典包含所有中文字与对应偏旁的关联关系；

10 所述第三对应规则包括：特征向量的长度等于预设的实体偏旁的数量；所述特征向量中每一维度对应一个所述实体偏旁；所述特征向量通过所述实体偏旁对应维度的向量值的改变表征对应的包含所述实体偏旁的字；

 或者，特征向量的长度为 1；所述特征向量通过不同的向量值对应表征包含不同所述实体偏旁的字。

15 5. 根据权利要求 1 所述的中文病历的实体识别方法，所述前缀词典的构建包括以下步骤：识别出所述原始医学词典中多于两个字的词组；将识别出的所述词组的前 i 个字存入前缀词典，i 为小于该词组长度且大于该词组长度的一半的自然数，其中该词组长度的一半取整数；

20 所述后缀词典的构建包括以下步骤：识别出所述原始医学词典中多于两个字的词组；将识别出的所述词组的后 i 个字存入后缀词典，i 为小于该词组长度且大于等于该词组长度的一半的自然数，其中该词组长度的一半取整数。

 6. 根据权利要求 1 所述的中文病历的实体识别方法，所述第四对应规则包括与原始医学词典相匹配时适用的对应规则、与前缀词典相匹配时适用的对应规则以及与所述后缀词典相匹配时适用的对应规则；

25 所述与原始医学词典相匹配时适用的对应规则包括：特征向量的长度等于所述实体类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始向量值改为第一向量值、第二向量值或第三向量值对应表征单个字在词组中的首位、中间位或末位；

30 所述与前缀词典相匹配时适用的对应规则包括：特征向量的长度等于所述实体类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始向量值改为第一向量值或第二向量值对应表征单个字在词组中的首位或非首位；

 所述与后缀词典相匹配时适用的对应规则包括：特征向量的长度等于所述实体

类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始向量值改为第二向量值或第三向量值对应表征单个字在词组中的非末位或未位。

7. 根据权利要求 1 所述的中文病历的实体识别方法，第五对应规则包括：特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述拼音。

8. 一种中文病历的实体识别装置，包括：

第一特征向量生成模块，用于识别出所述中文病历中包含的个人信息，根据第一对应规则输出与所述个人信息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量；

第二特征向量生成模块，用于利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的所述第二特征向量；

第三特征向量生成模块，用于识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与所述每个字的偏旁对应的第三特征向量；

第四特征向量生成模块，用于对所述中文病历做 n-gram 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量；

第五特征向量生成模块，用于利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量；

向量集生成模块，用于根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量和所述第五特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集；

实体识别模型，用于将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

9. 一种计算机设备，包括存储器和处理器，所述存储器上存储有计算机程序，所述计算机程序被所述处理器执行时实现如下步骤：

识别出所述中文病历中包含的个人信息，根据第一对应规则输出与所述个人信息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量；

利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的所述第二特征向量；

识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与所

述每个字的偏旁对应的第三特征向量；

对所述中文病历做 n-gram 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量；

5 利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量；

根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量和所述第五特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集；

10 将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

10、根据权利要求 9 所述的计算机设备，所述识别出所述中文病历中包含的个人信息包括以下步骤：创建用于匹配个人信息的正则表达式；用所述正则表达式与所述病人基本信息进行匹配，以识别出所述病人基本信息中包含的个人信息；

所述第一对应规则包括病人类型与特征向量的对应规则和病人年龄与特征向量的对应规则；

所述病人类型与特征向量的对应规则包括：特征向量的长度等于所述病人类型的种类数量；所述特征向量中每一维度对应所述病人类型中的一个种类；所述特征向量通过所述病人类型对应维度的向量值的改变表征对应的所述病人类型；

或者，特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述病人类型；

所述病人年龄与特征向量的对应规则包括：特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述病人年龄，所述向量值等于所述病人年龄。

11、根据权利要求 9 所述的计算机设备，所述第二对应规则包括：特征向量的长度为 4；所述特征向量的前三个维度用于表征包含两个字以上词组，其中第一个维度的向量值的改变用于表征位于所述词组中首位的字，第二个维度的向量值的改变用于表征位于所述词组中中间的字，第三个维度的向量值的改变用于表征位于所述词组中末尾的字；所述特征向量的第四个维度用于表征单字词组，所述第四个维度的向量值的改变用于表征单字词组中的字；

所述第五对应规则包括：特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述拼音。

12、根据权利要求 9 所述的计算机设备，所述识别出所述中文病历中每个字的偏旁具体包括以下步骤：

将所述中文病历中的每个字与预设的偏旁字典进行匹配，输出匹配到的偏旁，所述偏旁字典包含所有中文字与对应偏旁的关联关系；

所述第三对应规则包括：特征向量的长度等于预设的实体偏旁的数量；所述特征向量中每一维度对应一个所述实体偏旁；所述特征向量通过所述实体偏旁对应维度的向量值的改变表征对应的包含所述实体偏旁的字；

或者，特征向量的长度为 1；所述特征向量通过不同的向量值对应表征包含不同所述实体偏旁的字。

13、根据权利要求 9 所述的计算机设备，所述前缀词典的构建包括以下步骤：识别出所述原始医学词典中多于两个字的词组；将识别出的所述词组的前 i 个字存入前缀词典，i 为小于该词组长度且大于该词组长度的一半的自然数，其中该词组长度的一半取整数；

所述后缀词典的构建包括以下步骤：识别出所述原始医学词典中多于两个字的词组；将识别出的所述词组的后 i 个字存入后缀词典，i 为小于该词组长度且大于等于该词组长度的一半的自然数，其中该词组长度的一半取整数。

14、根据权利要求 9 所述的计算机设备，所述第四对应规则包括与原始医学词典相匹配时适用的对应规则、与前缀词典相匹配时适用的对应规则以及与后缀词典相匹配时适用的对应规则；

所述与原始医学词典相匹配时适用的对应规则包括：特征向量的长度等于所述实体类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始向量值改为第一向量值、第二向量值或第三向量值对应表征单个字在词组中的首位、中间位或未位；

所述与前缀词典相匹配时适用的对应规则包括：特征向量的长度等于所述实体类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始向量值改为第一向量值或第二向量值对应表征单个字在词组中的首位或非首位；

所述与后缀词典相匹配时适用的对应规则包括：特征向量的长度等于所述实体类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始向量值改为第二向量值或第三向量值对应表征单个字在词组中的非末位或未位。

15. 一种非易失性计算机可读存储介质，所述计算机可读存储介质内存储有计算机程序，所述计算机程序可被至少一个处理器所执行，以实现如下步骤：

识别出所述中文病历中包含的个人信息,根据第一对应规则输出与所述个人信

息对应的第一特征向量，所述中文病历中每个字对应相同的所述第一特征向量；

利用分词工具对所述中文病历进行分词，以分词后得到的词组为单位，根据第二对应规则对应每个字输出用于表征每个字在词组中位置的所述第二特征向量；

识别出所述中文病历中每个字的偏旁，根据第三对应规则对应每个字输出与所
5 述每个字的偏旁对应的第三特征向量；

对所述中文病历做 **n-gram** 遍历，将遍历后得到各个词组分别与预设的原始医学词典、前缀词典和后缀词典进行匹配，根据匹配结果与第四对应规则对应每个字输出相应的第四特征向量；

利用中文拼音转换工具将所述中文病历中每个字转换为拼音，根据第五对应规则对应每个字输出与所述每个字的拼音对应的第五特征向量；
10

根据拼接规则将所述第一特征向量、所述第二特征向量、所述第三特征向量、所述第四特征向量和所述第五特征向量对应拼接在每个字的初始向量之后，以得到用于表征所述中文病历的向量集；

将用于表征所述中文病历的向量集输入训练好的模型以抽取其中的实体。

16、根据权利要求 15 所述的计算机可读存储介质，所述识别出所述中文病历中包含的个人信息包括以下步骤：创建用于匹配个人信息的正则表达式；用所述正则表达式与所述病人基本信息进行匹配，以识别出所述病人基本信息中包含的个人信息；所述第一对应规则包括病人类型与特征向量的对应规则和病人年龄与特征向量的对应规则；
15

所述病人类型与特征向量的对应规则包括：特征向量的长度等于所述病人类型的种类数量；所述特征向量中每一维度对应所述病人类型中的一个种类；所述特征向量通过所述病人类型对应维度的向量值的改变表征对应的所述病人类型；
20

或者，特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述病人类型；

所述病人年龄与特征向量的对应规则包括：特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述病人年龄，所述向量值等于所述病人年龄。
25

17、根据权利要求 15 所述的计算机可读存储介质，所述第二对应规则包括：特征向量的长度为 4；所述特征向量的前三个维度用于表征包含两个字以上词组，其中第一个维度的向量值的改变用于表征位于所述词组中首位的字，第二个维度的向量值的改变用于表征位于所述词组中中间的字，第三个维度的向量值的改变用于表征位于所述词组中末尾的字；所述特征向量的第四个维度用于表征单字词组，所述第四个维度的向量值的改变用于表征单字词组中的字；
30

所述第五对应规则包括：特征向量的长度为 1；所述特征向量通过不同的向量值对应表征不同所述拼音。

18、根据权利要求 15 所述的计算机可读存储介质，所述识别出所述中文病历中每个字的偏旁具体包括以下步骤：将所述中文病历中的每个字与预设的偏旁字典进行匹配，输出匹配到的偏旁，所述偏旁字典包含所有中文字与对应偏旁的关联关系；

所述第三对应规则包括：特征向量的长度等于预设的实体偏旁的数量；所述特征向量中每一维度对应一个所述实体偏旁；所述特征向量通过所述实体偏旁对应维度的向量值的改变表征对应的包含所述实体偏旁的字；

或者，特征向量的长度为 1；所述特征向量通过不同的向量值对应表征包含不同所述实体偏旁的字。

19、根据权利要求 15 所述的计算机可读存储介质，所述前缀词典的构建包括以下步骤：识别出所述原始医学词典中多于两个字的词组；将识别出的所述词组的前 i 个字存入前缀词典，i 为小于该词组长度且大于该词组长度的一半的自然数，其中该词组长度的一半取整数；

所述后缀词典的构建包括以下步骤：识别出所述原始医学词典中多于两个字的词组；将识别出的所述词组的后 i 个字存入后缀词典，i 为小于该词组长度且大于等于该词组长度的一半的自然数，其中该词组长度的一半取整数。

20、根据权利要求 15 所述的计算机可读存储介质，所述第四对应规则包括与原始医学词典相匹配时适用的对应规则、与前缀词典相匹配时适用的对应规则以及与所述词典相匹配时适用的对应规则；

所述与原始医学词典相匹配时适用的对应规则包括：特征向量的长度等于所述实体类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始向量值改为第一向量值、第二向量值或第三向量值对应表征单个字在词组中的首位、中间位或未位；

所述与前缀词典相匹配时适用的对应规则包括：特征向量的长度等于所述实体类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始向量值改为第一向量值或第二向量值对应表征单个字在词组中的首位或非首位；

所述与后缀词典相匹配时适用的对应规则包括：特征向量的长度等于所述实体类别的数量；所述特征向量中每一维度对应一个所述实体类别；所述特征向量通过将初始向量值改为第二向量值或第三向量值对应表征单个字在词组中的非末位或未位。