(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 112035675 A
(43)申请公布日 2020.12.04

(21)申请号 202010897823.4

(22)申请日 2020.08.31

(71)申请人 康键信息技术(深圳)有限公司
地址 518052 广东省深圳市前海深港合作区前湾一路1号A栋201室(入驻深圳市前海商务秘书有限公司)

(72)发明人 胡俊飞

(74)专利代理机构 北京市京大律师事务所 11321
代理人 姚维

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 16/35(2019.01)
G06F 40/242(2020.01)
G06F 40/289(2020.01)
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06N 3/02(2006.01)

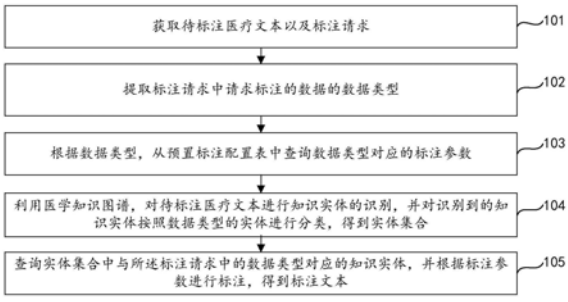
权利要求书2页 说明书13页 附图7页

(54)发明名称

医疗文本标注方法、装置、设备及存储介质

(57)摘要

本发明涉及人工智能技术领域，公开了一种医疗文本标注方法、装置、设备及存储介质。该方法通过设置标注配置表，在接收到标注请求后，根据请求中请求的数据类型调用标注配置表中配置的标注参数来对待标注医疗文本中的知识实体进行批量标注，并且该标注过程中，还使用了医疗知识图谱进行医疗的知识实体进行识别，知识图谱中存在相似或者相同的实体，基于比对即可帮助系统进行快速识别和快熟标注，这样的实现方式不仅提高了用户对医疗文本的实体命名的快速标注，还保证了标注实体的精确度，大大提高了标注效率和用户的使用体验。此外，本发明还涉及区块链技术，医疗文本和实体可存储于区块链中。



CN 112035675 A

1.一种医疗文本标注方法,其特征在于,所述医疗文本标注方法包括:

获取待标注医疗文本以及标注请求;

提取所述标注请求中请求标注的数据的数据类型;

根据所述数据类型,从预置标注配置表中查询所述数据类型对应的标注参数;

利用医学知识图谱,对所述待标注医疗文本进行知识实体的识别,并对识别到的知识实体按照数据类型的实体进行分类,得到实体集合;

查询所述实体集合中与所述标注请求中的数据类型对应的知识实体,并根据所述标注参数进行标注,得到标注文本。

2.根据权利要求1所述的医疗文本标注方法,其特征在于,所述方法还包括:

获取文本修改用户的修改请求;

提取所述修改请求中的配置项目和具体配置参数,其中所述配置项目信息包括标注的数据类型、标注标签号、标签名称和标签颜色,以及标签调用的快捷方式;

基于所述配置项目,在预置标注配置表模板中添加对应的表头名称,并根据所述具体配置参数设置所述表头名称下的显示数据,形成所述标注配置表。

3.根据权利要求2所述的医疗文本标注方法,其特征在于,在所述基于所述配置项目,在预置标注配置表模板中添加对应的表头名称,并根据所述具体配置参数设置所述表头名称下的显示数据,形成所述标注配置表之后,还包括:

获取历史医学实体,并调用自然语言处理技术对所述历史医学本体进行学习,得到实体识别模型;

利用所述所述标注配置表中的标注参数对所述实体识别模型进行标注训练,得到标注模型,其中所述标注模型用于对所述待标注医疗文本中的知识实体进行识别以及预标注。

4.根据权利要求3所述的医疗文本标注方法,其特征在于,所述利用医学知识图谱,对所述待标注医疗文本进行知识实体的识别,并对识别到的知识实体按照数据类型的实体进行分类,得到实体集合包括:

利用自然语言处理技术,提取所述待标注医疗文本中的句子,得到医学句子集合;

提取所述医学知识图谱中的实体名称,并基于所述实体名称逐一匹配所述医学句子集合中的每个句子的词语,确定句子中符合匹配条件的词语,得到词语集合;

根据语义识别算法,分析所述词语集合中每个词语的实体语义,并基于所述实体语义与所述数据类型之间的对应关系,对所述词语集合进行分类,得到多个知识实体组;

将所述多个知识实体组作为实体集合输出。

5.根据权利要求4所述的医疗文本标注方法,其特征在于,所述查询所述实体集合中与所述标注请求中的数据类型对应的知识实体,并根据所述标注参数进行标注,得到标注文本包括:

根据所述标注请求请求的数据类型,查询所述实体集合中与所述数据类型对应的知识实体组;

根据所述数据类型对应的标签颜色,将所述知识实体组中的每个词语标注为所述数据类型对应的标签颜色,并在所述词语上显示对应的标签名称和修改用户的名称,得到预打标数据;

将所述待标注医疗文本中与所述知识实体组对应的词语替换为所述预打标数据,生成

标注文本。

6.根据权利要求5所述的医疗文本标注方法，其特征在于，在所述查询所述实体集合中与所述标注请求中的数据类型对应的知识实体，并根据所述标注参数进行标注，得到标注文本之后，还包括：

调用实体命名纠错算法，识别被标注的知识实体的命名是否正确；

若不正确，则查询互联网词典，选择与所述知识实体相似的实体名称进行替换，得到新的标注文本；或者，通知修改用户进行人工修改。

7.根据权利要求6所述的医疗文本标注方法，其特征在于，在所述查询互联网词典，选择与所述知识实体相似的实体名称进行替换，得到新的标注文本之后，还包括：

获取替换后的实体名称作为训练集，对所述标注模型进行模型优化和训练，得到优化后的标注模型。

8.一种医疗文本标注装置，其特征在于，所述医疗文本标注装置包括：

接收模块，用于获取待标注医疗文本以及标注请求；

提取模块，用于提取所述标注请求中请求标注的数据的数据类型；

查询模块，用于根据所述数据类型，从预置标注配置表中查询所述数据类型对应的标注参数；

识别模块，用于利用医学知识图谱，对所述待标注医疗文本进行知识实体的识别，并对识别到的知识实体按照数据类型的实体进行分类，得到实体集合；

标注模块，用于查询所述实体集合中与所述标注请求中的数据类型对应的知识实体，并根据所述标注参数进行标注，得到标注文本。

9.一种医疗文本标注设备，其特征在于，所述医疗文本标注设备包括：存储器和至少一个处理器，所述存储器中存储有指令，所述存储器和所述至少一个处理器通过线路互连；

所述至少一个处理器调用所述存储器中的所述指令，以使得所述医疗文本标注设备执行如权利要求1-7中任一项所述的医疗文本标注方法。

10.一种计算机可读存储介质，所述计算机可读存储介质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如权利要求1-7中任一项所述的医疗文本标注方法。

医疗文本标注方法、装置、设备及存储介质

技术领域

[0001] 本申请涉及人工智能技术领域，具体涉及一种医疗文本标注方法、装置、设备及存储介质。

背景技术

[0002] 随着移动互联网和社交网络的推广，产生了大量的用户生成文本(User Generated Content，简称UGC)，由于文化背景和表述习惯的不同，人们往往会使用不同的词语和表述方式表达类似的内容。尤其是医疗领域中，由于其医药名称的命名并不是规范的语义命名方式，因此，其识别难度更加大。

[0003] 目前的实现方式，主要是采用自然语言处理技术来实现自动标注识别。而通过自然语言处理技术还是通过语义的方式来计算识别，从而进行标注，并且在标注后要纠正计算机的错误识别，这就使得完成一次的识别其需要的计算量过大，并且医药名称的语义还不能使用常规的语义解析来来识别，还需要进行大量的训练学习，这非常不利于开发，且成本也高，在使用时，其效率和准确率都比较低。

发明内容

[0004] 本发明的主要目的是解决现有的医疗文本标注效率和准确度较低的技术问题。

[0005] 本发明第一方面提供了一种医疗文本标注方法，所述医疗文本标注方法包括：

[0006] 获取待标注医疗文本以及标注请求；

[0007] 提取所述标注请求中请求标注的数据的数据类型；

[0008] 根据所述数据类型，从预置标注配置表中查询所述数据类型对应的标注参数；

[0009] 利用医学知识图谱，对所述待标注医疗文本进行知识实体的识别，并对识别到的知识实体按照数据类型的实体进行分类，得到实体集合；

[0010] 查询所述实体集合中与所述标注请求中的数据类型对应的知识实体，并根据所述标注参数进行标注，得到标注文本。

[0011] 可选地，在本发明第一方面的第一种实现方式中，所述方法还包括：

[0012] 获取文本修改用户的修改请求；

[0013] 提取所述修改请求中的配置项目和具体配置参数，其中所述配置项目信息包括标注的数据类型、标注标签号、标签名称和标签颜色，以及标签调用的快捷方式；

[0014] 基于所述配置项目，在预置标注配置表模板中添加对应的表头名称，并根据所述具体配置参数设置所述表头名称下的显示数据，形成所述标注配置表。

[0015] 可选地，在本发明第一方面的第二种实现方式中，在所述基于所述配置项目，在预置标注配置表模板中添加对应的表头名称，并根据所述具体配置参数设置所述表头名称下的显示数据，形成所述标注配置表之后，还包括：

[0016] 获取历史医学实体，并调用自然语言处理技术对所述历史医学本体进行学习，得到实体识别模型；

[0017] 利用所述所述标注配置表中的标注参数对所述实体识别模型进行标注训练，得到标注模型，其中所述标注模型用于对所述待标注医疗文本中的知识实体进行识别以及预标注。

[0018] 可选地，在本发明第一方面的第三种实现方式中，所述利用医学知识图谱，对所述待标注医疗文本进行知识实体的识别，并对识别到的知识实体按照数据类型的实体进行分类，得到实体集合包括：

[0019] 利用自然语言处理技术，提取所述待标注医疗文本中的句子，得到医学句子集合；

[0020] 提取所述医学知识图谱中的实体名称，并基于所述实体名称逐一匹配所述医学句子集合中的每个句子的词语，确定句子中符合匹配条件的词语，得到词语集合；

[0021] 根据语义识别算法，分析所述词语集合中每个词语的实体语义，并基于所述实体语义与所述数据类型之间的对应关系，对所述词语集合进行分类，得到多个知识实体组；

[0022] 将所述多个知识实体组作为实体集合输出。

[0023] 可选地，在本发明第一方面的第四种实现方式中，所述查询所述实体集合中与所述标注请求中的数据类型对应的知识实体，并根据所述标注参数进行标注，得到标注文本包括：

[0024] 根据所述标注请求请求的数据类型，查询所述实体集合中与所述数据类型对应的知识实体组；

[0025] 根据所述数据类型对应的标签颜色，将所述知识实体组中的每个词语标注为所述数据类型对应的标签颜色，并在所述词语上显示对应的标签名称和修改用户的名称，得到预打标数据；

[0026] 将所述待标注医疗文本中与所述知识实体组对应的词语替换为所述预打标数据，生成标注文本。

[0027] 可选地，在本发明第一方面的第五种实现方式中，在所述查询所述实体集合中与所述标注请求中的数据类型对应的知识实体，并根据所述标注参数进行标注，得到标注文本之后，还包括：

[0028] 调用实体命名纠错算法，识别被标注的知识实体的命名是否正确；

[0029] 若不正确，则查询互联网词典，选择与所述知识实体相似的实体名称进行替换，得到新的标注文本；或者，通知修改用户进行人工修改。

[0030] 可选地，在本发明第一方面的第六种实现方式中，在所述查询互联网词典，选择与所述知识实体相似的实体名称进行替换，得到新的标注文本之后，还包括：

[0031] 获取替换后的实体名称作为训练集，对所述标注模型进行模型优化和训练，得到优化后的标注模型。

[0032] 本发明第二方面提供了一种医疗文本标注装置，所述医疗文本标注装置包括：

[0033] 接收模块，用于获取待标注医疗文本以及标注请求；

[0034] 提取模块，用于提取所述标注请求中请求标注的数据的数据类型；

[0035] 查询模块，用于根据所述数据类型，从预置标注配置表中查询所述数据类型对应的标注参数；

[0036] 识别模块，用于利用医学知识图谱，对所述待标注医疗文本进行知识实体的识别，并对识别到的知识实体按照数据类型的实体进行分类，得到实体集合；

[0037] 标注模块，用于查询所述实体集合中与所述标注请求中的数据类型对应的知识实体，并根据所述标注参数进行标注，得到标注文本。

[0038] 可选地，在本发明第二方面的第一种实现方式中，所述医疗文本标注装置还包括设置模块，其具体用于：

[0039] 获取文本修改用户的修改请求；

[0040] 提取所述修改请求中的配置项目和具体配置参数，其中所述配置项目信息包括标注的数据类型、标注标签号、标签名称和标签颜色，以及标签调用的快捷方式；

[0041] 基于所述配置项目，在预置标注配置表模板中添加对应的表头名称，并根据所述具体配置参数设置所述表头名称下的显示数据，形成所述标注配置表。

[0042] 可选地，在本发明第二方面的第二种实现方式中，所述医疗文本标注装置还包括训练模块，其具体用于：

[0043] 获取历史医学实体，并调用自然语言处理技术对所述历史医学本体进行学习，得到实体识别模型；

[0044] 利用所述所述标注配置表中的标注参数对所述实体识别模型进行标注训练，得到标注模型，其中所述标注模型用于对所述待标注医疗文本中的知识实体进行识别以及预标注。

[0045] 可选地，在本发明第二方面的第三种实现方式中，所述识别模块包括：

[0046] 分句单元，用于利用自然语言处理技术，提取所述待标注医疗文本中的句子，得到医学句子集合；

[0047] 匹配单元，用于提取所述医学知识图谱中的实体名称，并基于所述实体名称逐一匹配所述医学句子集合中的每个句子的词语，确定句子中符合匹配条件的词语，得到词语集合；

[0048] 识别单元，用于根据语义识别算法，分析所述词语集合中每个词语的实体语义，并基于所述实体语义与所述数据类型之间的对应关系，对所述词语集合进行分类，得到多个知识实体组；将所述多个知识实体组作为实体集合输出。

[0049] 可选地，在本发明第二方面的第四种实现方式中，所述标注模块包括：

[0050] 查询单元，用于根据所述标注请求请求的数据类型，查询所述实体集合中与所述数据类型对应的知识实体组；

[0051] 预处理单元，用于根据所述数据类型对应的标签颜色，将所述知识实体组中的每个词语标注为所述数据类型对应的标签颜色，并在所述词语上显示对应的标签名称和修改用户的名称，得到预打标数据；

[0052] 标注单元，用于将所述待标注医疗文本中与所述知识实体组对应的词语替换为所述预打标数据，生成标注文本。

[0053] 可选地，在本发明第二方面的第五种实现方式中，所述医疗文本标注装置还包括纠错模块，其具体用于：

[0054] 调用实体命名纠错算法，识别被标注的知识实体的命名是否正确；

[0055] 若不正确，则查询互联网词典，选择与所述知识实体相似的实体名称进行替换，得到新的标注文本；或者，通知修改用户进行人工修改。

[0056] 可选地，在本发明第二方面的第六种实现方式中，所述训练模块，其具体还用于：

[0057] 获取替换后的实体名称作为训练集，对所述标注模型进行模型优化和训练，得到优化后的标注模型。

[0058] 本发明第三方面提供了一种医疗文本标注设备，包括：存储器和至少一个处理器，所述存储器中存储有指令，所述存储器和所述至少一个处理器通过线路互连；

[0059] 所述至少一个处理器调用所述存储器中的所述指令，以使得所述医疗文本标注设备执行上述的医疗文本标注方法。

[0060] 本发明的第四方面提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机程序，当其在计算机上运行时，使得计算机执行上述的医疗文本标注方法。

[0061] 本发明提供的技术方案中，通过设置标注配置表，在接收到标注请求后，根据请求中请求的数据类型调用标注配置表中配置的标注参数来对待标注医疗文本中的知识实体进行批量标注，并且该标注过程中，还使用了医疗知识图谱进行医疗的知识实体进行识别，知识图谱中存在相似或者相同的实体，基于比对即可帮助系统进行快速识别和快熟标注，这样的实现方式不仅提高了用户对医疗文本的实体命名的快速标注，还保证了标注实体的精确度，大大提高了标注效率和用户的使用体验。

附图说明

[0062] 图1为本发明实施例中医疗文本标注方法的第一个实施例示意图；

[0063] 图2为本发明实施例中医疗文本标注方法的第二个实施例示意图；

[0064] 图3为本发明实施例中步骤203的细化流程示意图；

[0065] 图4为本发明实施例中医疗文本标注方法的第三个实施例示意图；

[0066] 图5为本发明实施例中标注配置表的示意图；

[0067] 图6为本发明实施例中标注文本的一种示意图；

[0068] 图7为本发明实施例中标注文本的另一种示意图；

[0069] 图8为本发明实施例中医疗文本标注方法的第四个实施例示意图；

[0070] 图9为本发明实施例中医疗文本标注装置的一个实施例示意图；

[0071] 图10为本发明实施例中医疗文本标注装置的另一个实施例示意图；

[0072] 图11为本发明实施例中医疗文本标注设备的一个实施例示意图。

具体实施方式

[0073] 针对用户通过现有的医学文本标注方法对医学文本进行标注修改时精准度和处理效率较低的问题，本申请通过设置一种半自动的标注方法来提高对医学文本的修改识别，首先是根据不同的用户设置不同的标注方式，以及对不同用户的标注内容进行分配，然后基于医学知识图谱对医学文本进行实体的识别，从而生成预标注信息，并在医学文本中显示出来，用户可以根据标注的情况进行选择修改，从而提高了标注修改的效率，也保证了修改的精准度，同时也实现了多人同时标注修改，大大提高了处理效率。

[0074] 本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等(如果存在)是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外，术语“包括”或“具有”及其任何变形，意图在于覆盖不

排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0075] 为便于理解，下面对本发明实施例的具体流程进行描述，请参阅图1，本发明实施例中医疗文本标注方法的第一个实施例包括：

[0076] 101、获取待标注医疗文本以及标注请求；

[0077] 在该步骤中，这里的待标注医疗文本可以是医疗书籍、网络论坛知识和病历单，甚至还可以是诊断报告。

[0078] 在实际应用中，当医疗终端接收到的用户在终端上触发的标注操作时，医疗终端根据标注操作生成标注请求，该标注请求中包括有待标注的数据类型和待标注的医疗文本的名称。

[0079] 进一步的，其数据类型是根据用户在文本中选择的词语确定类型信息，并将该类型信息串接到触发指令中，同时还需将对应的文本的名称也串接到触发指令中，从而生成标注请求。

[0080] 102、提取标注请求中请求标注的数据的数据类型；

[0081] 在本实施例中，在接收到标注请求后，服务器或者是医疗系统的后台对请求进行解析，其解析主要是根据请求的生成格式进行解析，例如生成格式为触发指令+数据类型+文本名称，基于该格式将触发指令、数据类型和文本名称单独分割出来，然后根据触发指令启动文本的标注流程，然后加载文本名称对应的医疗文本显示在显示界面上。

[0082] 在实际应用中，该步骤还包括：将所述数据类型的信息显示在所述医疗文本中，或者是根据所述数据类型确定标注快捷键，并显示在所述医疗文本的边缘位置。

[0083] 103、根据数据类型，从预置标注配置表中查询数据类型对应的标注参数；

[0084] 在该步骤中，根据数据类型查询标注配置表中是否存在对应类型的配置记录，若存在，则读取对应的标注参数，例如：类型为seq，标签ID为101，名称为drug，快捷键为d，颜色为gree。

[0085] 在实际应用中，若标注配置表中没有对应类型的配置记录，则可以启动参数配置程序，基于配置的程序调度对应的配置界面，根据用户的当前评注需求进行设置类型名称、标签名称和ID、快捷键，以及标注的显示色彩等等参数，具体的，若标注中存在特殊字体或者词组命名时，还可以设置特殊字体或者词组命名的识别规则。

[0086] 104、利用医学知识图谱，对待标注医疗文本进行知识实体的识别，并对识别到的知识实体按照数据类型的实体进行分类，得到实体集合；

[0087] 该步骤中，所述医学知识图谱为预先通过模型算法训练学习得到的实体树状图，在该树状图上有多个节点，每个节点对应有实体名称以及实体对应的一类实体语音解析。

[0088] 具体的，其训练过程为：从医学数据库中获取到已经被医学专家通过人工标注的方式标注分类好的实体数据，基于该实体数据进行模型的训练，可选的，在根据图谱构建数据构建医疗知识图谱时，可以是通过将图谱构建数据输入到预置的图谱树中，构建医疗知识图谱，这里的图谱树指的是包含多个父节点和子节点的树型结构图，从而形成医学知识图谱。

[0089] 基于该医疗知识图谱对待标注医疗文本进行识别，提取中其中的各种类型的实

体，然后利用分类模型对实体进行分类，也可以是使用聚类算法进行分类，而使用聚类算法则是通过计算每个实体的语义或者是本体，基于语义或本体进行聚类，从而得到多个实体类别，从而得到实体集合。

[0090] 进一步的，还包括识别实体集合中每个实体类别的数据类型，并建立实体类别与数据类型之间的对应关系，并形成一个表格。

[0091] 105、查询实体集合中与所述标注请求中的数据类型对应的知识实体，并根据标注参数进行标注，得到标注文本。

[0092] 在该步骤中，这里的实体集合指的是通过医疗知识图谱识别并分类后的，包含有至少两组实体类别的集合，而在数据类型与实体集合中的数据类型进行匹配，根据匹配的结果提取出实体类别，然后将上述步骤中查询到的标注参数对实体类别中的所有实体进行标注，形成标注数据，将该标注数据映射到待标注医疗文本中，从而得到标注文本。

[0093] 通过对上述方法的实施例，基于设置标注参数，根据具体用户的标注请求来选择对应的数据类型进行标注，具体是据请求中请求的数据类型调用标注配置表中配置的标注参数来对待标注医疗文本中的知识实体进行批量标注，由于不同的请求选择的标注参数是不相同的，所以在多人同时请求时，其也可以进行区分，从而实现了可以进行多人合作，分配标注任务，只需选择文本跨度并对其进行标注即可，支持快捷键，因此您可以快速标注文本跨度，在标注过程中支持无效标签、存疑标签对数据的一个判断，标注操作实时更新，翻页自动保存，模型预打标与人工标注相结合，利用模型跑出的结果辅助人工标注，不同的实体标签配置生成不同颜色，人工核对标注结果是否正确，只需对模型预判错误的进行修改，提高标注的一致性及效率。

[0094] 请参阅图2，本发明实施例中医疗文本标注方法的第二个实施例包括：

[0095] 201、获取待标注医疗文本以及标注请求；

[0096] 202、提取标注请求中请求标注的数据的数据类型；

[0097] 203、根据数据类型，从预置标注配置表中查询数据类型对应的标注参数；

[0098] 在实际应用中，步骤201-203的具体实现与上述步骤101-103的具体实现相同，根据数据类型查询标注配置表中是否存在对应类型的配置记录，若存在，则读取对应的标注参数，例如：类型为seq，标签ID为101，名称为drug，快捷键为d，颜色为gree。

[0099] 在实际应用中，若标注配置表中没有对应类型的配置记录，则可以启动参数配置程序，基于配置的程序调度对应的配置界面，根据用户的当前评注需求进行设置类型名称、标签名称和ID、快捷键，以及标注的显示色彩等等参数，具体的，若标注中存在特殊字体或者词组命名时，还可以设置特殊字体或者词组命名的识别规则。

[0100] 204、利用自然语言处理技术，提取待标注医疗文本中的句子，得到医学句子集合；

[0101] 205、提取医学知识图谱中的实体名称，并基于实体名称逐一匹配所述医学句子集合中的每个句子的词语，确定句子中符合匹配条件的词语，得到词语集合；

[0102] 206、根据语义识别算法，分析词语集合中每个词语的实体语义，并基于实体语义与所述数据类型之间的对应关系，对词语集合进行分类，得到多个知识实体组；

[0103] 207、将多个知识实体组作为实体集合输出；

[0104] 具体的，上述步骤204-207实质上是实现了对医疗文本中每个句子的实体进行精准识别，在实际应用中，在对待标注医疗文本中的知识实体进行识别时，具体可以通过以下

方式来实现，如图3所示：

[0105] 301、输入待识别的中文医疗文本，进行预处理；

[0106] 具体的，首先根据标注词典对待识别的中文医疗文本数据中的每个句子S进行分词和标注，S＝(w1,w2,…wi,…wn)，wi表示对S进行分词之后，该句子中的第i个词语；

[0107] 然后对每个句子进行分字处理，S＝(c1,c2…ci…cm)，其中ci表示对句子S进行分字处理后的第i个字符；

[0108] 302、对于每一个句子S，对组成它的字、词、部首三个粒度的特征分别进行提取；

[0109] 具体的，首先词语特征的提取及向量表示；

[0110] 对于分词和标注之后的每个句子中的每个词语wi，将该词语的第一个字符用1表示，最后一个字符用3表示，出现在中间位置的字符统一编码为2；如果一个词语的长度小于2，则其所对应的向量在终止位置之后统一用0补齐；对于只由单个汉字独立构成的词，统一用全；

[0111] 为0的20维向量来表示，得到词语的向量

[0112] 进一步的，字特征的提取和向量表示；

[0113] 利用现有的Word2Vec模型中的Skip-Gram算法对文本中的每一个字符进行训练，将每个字符用100维的数值向量进行表示，得到字向量；

[0114] 然后，部首特征的提取和向量表示；

[0115] 将文本中每个单字的部首拆分出来，设其部首所对应的会意字为P，则通过检索字向量字典可以得到该会意字所对应的100维字向量，将该100维向量看作是该字的部首向量，记为$V_i^p = (\theta_1, \theta_2 \cdots \theta_l),\ l=100$；

[0116] 303、将提取出的三种特征进行特征融合，得到用于实体识别和分类的特征的联合向量表示。

[0117] 在实际应用中，该步骤的实现可以是，首先字向量和部首向量的融合；

[0118] 使用逐点相加法对二者进行特征融合，将两个向量的对应分量逐个进行相加，用相加之后的分量作为融合之后新的特征向量的分量，记将字向量和部首向量融合之后的特征向量为x＝(x1,x2,…xl)，该过程可以用如下的公式表示：

[0119] x＝(x1,x2,…xl)＝(μ1+θ1,μ2+θ2,…μt+xl)；

[0120] 然后再对词语特征的融合；

[0121] 对于字+部首的特征向量与词语特征向量进行融合，由于二者的维度不相同，这里使用维度拼接的方法进行特征融合，记融合之后的最终特征向量为Y＝(y1…yd)，其中d表示融合之后特征的维度，则维度拼接的过程由如下的公式表示：

[0122] $$Y_i = \left(X, V_i^w\right) = (\eta_1, \eta_2 \cdots \eta_k, x_1, x_2, \cdots x_l)$$

[0123] 其中，d＝k+l，通过前面的条件可知，这里的d＝120，即最终得到的融合之后的特征向量为120维。

[0124] 通过对上述方法的实施例，只需导入数据到系统平台，平台根据预先设置的标注参数对用户指定的数据类型进行快速标注，并且该种标注方式可以实现多人同时标注，不同的人操作，其标注参数的设置还会不同，从而大大提高了标注的效率。

[0125] 进一步的，在标注过程中，使用了医疗知识图谱进行医疗的知识实体进行识别，从

而帮助系统进行快速识别和快熟标注，这样的实现方式不仅提高了用户对医疗文本的实体命名的快速标注，还保证了标注实体的精确度，大大提高了标注效率和用户的使用体验。

[0126] 208、查询实体集合中与标注请求中的数据类型对应的知识实体，并根据标注参数进行标注，得到标注文本。

[0127] 在本实施例中，根据不同的请求选择的标注颜色是不相同的，所以在多人同时请求时，其也可以进行区分，从而实现了可以进行多人合作，分配标注任务，只需选择文本跨度并对其进行标注即可，支持快捷键，因此您可以快速标注文本跨度，在标注过程中支持无效标签、存疑标签对数据的一个判断，标注操作实时更新，翻页自动保存，模型预打标与人工标注相结合，利用模型跑出的结果辅助人工标注，不同的实体标签配置生成不同颜色，人工核对标注结果是否正确，只需对模型预判错误的进行修改，提高标注的一致性及效率。

[0128] 请参阅图4，本发明实施例中医疗文本标注方法的第三个实施例包括：

[0129] 401、获取文本修改用户的修改请求；

[0130] 402、提取修改请求中的配置项目和具体配置参数；

[0131] 该步骤中，所述配置项目信息包括标注的数据类型、标注标签号、标签名称和标签颜色，以及标签调用的快捷方式。

[0132] 403、基于配置项目，在预置标注配置表模板中添加对应的表头名称，并根据具体配置参数设置所述表头名称下的显示数据，形成标注配置表；

[0133] 在实际应用中，其配置主要是以标签的形式设置，而这里的标签的可配置项参数设计包括但不限制于以下几种：数据类型、标签名称和标签颜色；

[0134] 具体的，还可以是设置标注数据类型、标签ID、关联批次、标签名称、标签快捷键、标签颜色；根据不同的数据标注类型设置相关参数，比如：标签标注，通过页面操作设置一个label，类型为seq，标签ID为101，名称为drug，快捷键为d，颜色为gree，这样就成功配置好一个label参数，在界面可增删改查可视化操作。

[0135] 404、获取待标注医疗文本以及标注请求；

[0136] 405、提取标注请求中请求标注的数据的数据类型；

[0137] 406、根据数据类型，从预置标注配置表中查询数据类型对应的标注参数；

[0138] 在实际应用中，上述步骤404-406的具体实现与上述实施例步骤101-103的具体实现相同，这里不再赘述。

[0139] 407、查询实体集合中与标注请求中的数据类型对应的知识实体，并根据标注参数进行标注，得到标注文本；

[0140] 该步骤中，在查询知识实体时，具体是通过根据所述标注请求请求的数据类型，查询所述实体集合中与所述数据类型对应的知识实体组；

[0141] 根据所述数据类型对应的标签颜色，将所述知识实体组中的每个词语标注为所述数据类型对应的标签颜色，并在所述词语上显示对应的标签名称和修改用户的名称，得到预打标数据；

[0142] 将所述待标注医疗文本中与所述知识实体组对应的词语替换为所述预打标数据，生成标注文本。

[0143] 在实际应用中，其标注参数一般通过页面操作设置一个label，类型为seq，标签ID为101，名称为drug，快捷键为d，颜色为gree，这样就成功配置好一个label参数。

[0144] 如图5-7所示，根据上述设置的标注参数进行标注时，通过在医学文本实体标注界面可导入标注数据，同时可根据提供的标注人员信息进行标注任务分配，根据导入批次以及初始化搜索条件查询到标注数据，根据分页和每页展示数量可自行设置，标注人员根据已分配的数据在界面进行标注工作，初次加载数据为模型预打标数据，根据预打标数据结果和已经配置的医学实体标签展示在界面上，如：实体标签为药物和疾病搭配不同的颜色区分展示，利用模型跑出的结果辅助人工标注，人工核对标注结果是否正确，只需对模型预判错误的进行修改，在需要修改的数据文本上用鼠标滑动产生跨度区间配合键盘快捷键操作，同时支持对无效数据、存疑数据标注判断处理，提高标注效率。

[0145] 408、查询实体集合中与标注请求中的数据类型对应的知识实体，并根据标注参数进行标注，得到标注文本。

[0146] 为了进一步提高标注的效率，在本实施例中，还可以利用标注模型来实现，而标注模型具体是基于上述提供的预先设置好的标注配置表中的具体设置参数来训练得到，如图8所示，基于标注模型进行快速标注的实现步骤如下：

[0147] 501、获取当前标注的人员设置不同的标注参数，形成标注配置表；

[0148] 具体的，获取文本修改用户的修改请求，提取修改请求中的配置项目和具体配置参数，其中，所述配置项目信息包括标注的数据类型、标注标签号、标签名称和标签颜色，以及标签调用的快捷方式，基于配置项目，在预置标注配置表模板中添加对应的表头名称，并根据具体配置参数设置所述表头名称下的显示数据，形成标注配置表。

[0149] 在实际应用中，其配置主要是以标签的形式设置，而这里的标签的可配置项参数设计包括但不限制于以下几种：数据类型、标签名称和标签颜色；

[0150] 具体的，还可以是设置标注数据类型、标签ID、关联批次、标签名称、标签快捷键、标签颜色；根据不同的数据标注类型设置相关参数，比如：标签标注，通过页面操作设置一个label，类型为seq，标签ID为101，名称为drug，快捷键为d，颜色为gree，这样就成功配置好一个label参数，在界面可增删改查可视化操作。

[0151] 502、利用自然语言对标注配置表中的标注参数进行学习，得到标注模型；

[0152] 在该步骤中，所述标注模型用于对所述待标注医疗文本中的知识实体进行识别以及预标注。具体训练过程包括：首先获取历史医学实体，并调用自然语言处理技术对所述历史医学本体进行学习，得到实体识别模型；

[0153] 然后利用所述所述标注配置表中的标注参数对所述实体识别模型进行标注训练，得到标注模型，其中所述标注模型用于对所述待标注医疗文本中的知识实体进行识别以及预标注。

[0154] 在实际应用中，具体是通过将由所述标注文本包括的多个词语组成的第一词语序列输入到预先训练的神经网络模型中，得到所述神经网络模型输出的所述第一词语序列的识别结果，基于识别结果来调整模型的参数，其中标注文本为预先通过人工标注分类的文本；而这里的识别结果为对第一词语序列的医疗实体的识别结果；进一步的，将该识别结果与标注的结果进行比对，基于比对的结果判断是否识别准确，若比对结果为低于预设概率值，则调整模型的参数重新训练，反之则继续训练下一个文本，直至全部文本训练完成后进入测试上线。

[0155] 在本实施例中，利用训练的标注模型对待标注医疗文本进行标注时，具体是实现

步骤如下：

[0156] 首先，同理根据请求中的数据类型确定对应的标注参数；

[0157] 然后将待标注医疗文本输入到标注模型中，按照数据类型进行实体的识别，得到实体组，然后读取查询到的参数至标注模型中，标注模型对实体组中的每个实体进行标注，得到标注数据，将该标注数据作为替换文本替换待标注医疗文本中对应的实体，从而得到标注文本。

[0158] 503、获取标注人员信息进行标注任务分配，并确定每个任务中标注的数据类型；

[0159] 504、根据数据类型从标注配置表中查询对应的标注参数；

[0160] 具体的，所述标注参数为数据类型、标签ID、关联批次、标签名称、标签快捷键和标签颜色。

[0161] 505、将待标注的医疗文本和标注参数输入到标注模型中进行识别标注，得到预标注数据；

[0162] 506、根据预标注数据标注医疗文本，得到标注文本；

[0163] 507、调用实体命名纠错算法，识别被标注的知识实体的命名是否正确；

[0164] 508、若不正确，则查询互联网词典，选择与知识实体相似的实体名称进行替换，得到新的标注文本。

[0165] 本实施例中，在利用上述的模型或者技术算法进行标注后，可能会存在一定的错差，即是本身其文本中的实体名称会存在问题，通过标注之后，用户可以进行适应的修改调整。

[0166] 在实际应用中，该修改调整可以是人工修改，也可以是机器自动修改，而对于机器自动修改，这需要调用纠错模型，该纠错模型也是需要进行训练得到，通过采集医学中常有重名或者是同音不同命的医学名词形成训练集进行训练。

[0167] 进一步的，为了提高标注模型的精准度，在标注结果被修改后，记录其修改的实体名称和修改前的实体名称，在收集数量达到一定数量级时，将这些记录中的数据作为新的训练集，输入至标注模型中优化训练，具体的：获取替换后的实体名称作为训练集，对所述标注模型进行模型优化和训练，得到优化后的标注模型。

[0168] 在实际应用中，文本数据的实际标注工作结果有利于深度学习序列标注模型算法的学习，通过已有的深度学习模型算法对大规模的无监督数据进行预达标，选择置信度较高的数据与人工标注相结合，标注平台系统自动化记录医学实体标注的文本结果，将二次标注的结果进行质检，最终标注数据进行模型优化和训练，提高深度学习模型的准确度，有利于完成自动化实体抽取任务。

[0169] 在本实施例中，在接收到标注请求时，还包括，检测标签请求中的标注用户的数量，若为多个，则将标注用户进行优先级的排序，依次进行标注的处理。

[0170] 通过上述方案的实施，实现了可以根据所需导入数据到系统平台，支持多种语言标注，支持RESTful风格调用，可以进行多人合作，分配标注任务，只需选择文本跨度并对其进行标注即可，支持快捷键，因此您可以快速标注文本跨度，在标注过程中支持无效标签、存疑标签对数据的一个判断，标注操作实时更新，翻页自动保存，模型预打标与人工标注相结合，利用模型跑出的结果辅助人工标注，不同的实体标签配置生成不同颜色，人工核对标注结果是否正确，只需对模型预判错误的进行修改，提高标注的一致性及效率。

[0171] 上面对本发明实施例中医疗文本标注方法进行了描述，下面对本发明实施例中医疗文本标注装置进行描述，请参阅图9，本发明实施例中医疗文本标注装置的第一个实施例包括：

[0172] 接收模块901，用于获取待标注医疗文本以及标注请求；

[0173] 提取模块902，用于提取所述标注请求中请求标注的数据的数据类型；

[0174] 查询模块903，用于根据所述数据类型，从预置标注配置表中查询所述数据类型对应的标注参数；

[0175] 识别模块904，用于利用医学知识图谱，对所述待标注医疗文本进行知识实体的识别，并对识别到的知识实体按照数据类型的实体进行分类，得到实体集合；

[0176] 标注模块905，用于查询所述实体集合中与所述标注请求中的数据类型对应的知识实体，并根据所述标注参数进行标注，得到标注文本。

[0177] 在本实施例中，所述医疗文本标注装置运行上述医疗文本标注方法，该方法通过设置标注配置表，在接收到标注请求后，根据请求中请求的数据类型调用标注配置表中配置的标注参数来对待标注医疗文本中的知识实体进行批量标注，并且该标注过程中，还使用了医疗知识图谱进行医疗的知识实体进行识别，知识图谱中存在相似或者相同的实体，基于比对即可帮助系统进行快速识别和快熟标注，这样的实现方式不仅提高了用户对医疗文本的实体命名的快速标注，还保证了标注实体的精确度，大大提高了标注效率和用户的使用体验。

[0178] 请参阅图10，本发明实施例中医疗文本标注装置的第二个实施例，该医疗文本标注装置具体包括：

[0179] 接收模块901，用于获取待标注医疗文本以及标注请求；

[0180] 提取模块902，用于提取所述标注请求中请求标注的数据的数据类型；

[0181] 查询模块903，用于根据所述数据类型，从预置标注配置表中查询所述数据类型对应的标注参数；

[0182] 识别模块904，用于利用医学知识图谱，对所述待标注医疗文本进行知识实体的识别，并对识别到的知识实体按照数据类型的实体进行分类，得到实体集合；

[0183] 标注模块905，用于查询所述实体集合中与所述标注请求中的数据类型对应的知识实体，并根据所述标注参数进行标注，得到标注文本。

[0184] 其中，所述医疗文本标注装置还包括设置模块906，其具体用于：

[0185] 获取文本修改用户的修改请求；

[0186] 提取所述修改请求中的配置项目和具体配置参数，其中所述配置项目信息包括标注的数据类型、标注标签号、标签名称和标签颜色，以及标签调用的快捷方式；

[0187] 基于所述配置项目，在预置标注配置表模板中添加对应的表头名称，并根据所述具体配置参数设置所述表头名称下的显示数据，形成所述标注配置表。

[0188] 其中，所述医疗文本标注装置还包括训练模块907，其具体用于：

[0189] 获取历史医学实体，并调用自然语言处理技术对所述历史医学本体进行学习，得到实体识别模型；

[0190] 利用所述所述标注配置表中的标注参数对所述实体识别模型进行标注训练，得到标注模型，其中所述标注模型用于对所述待标注医疗文本中的知识实体进行识别以及预标

注。

[0191] 可选地，所述识别模块904包括：

[0192] 分句单元9041，用于利用自然语言处理技术，提取所述待标注医疗文本中的句子，得到医学句子集合；

[0193] 匹配单元9042，用于提取所述医学知识图谱中的实体名称，并基于所述实体名称逐一匹配所述医学句子集合中的每个句子的词语，确定句子中符合匹配条件的词语，得到词语集合；

[0194] 识别单元9043，用于根据语义识别算法，分析所述词语集合中每个词语的实体语义，并基于所述实体语义与所述数据类型之间的对应关系，对所述词语集合进行分类，得到多个知识实体组；将所述多个知识实体组作为实体集合输出。

[0195] 其中，所述标注模块905包括：

[0196] 查询单元9051，用于根据所述标注请求请求的数据类型，查询所述实体集合中与所述数据类型对应的知识实体组；

[0197] 预处理单元9052，用于根据所述数据类型对应的标签颜色，将所述知识实体组中的每个词语标注为所述数据类型对应的标签颜色，并在所述词语上显示对应的标签名称和修改用户的名称，得到预打标数据；

[0198] 标注单元9053，用于将所述待标注医疗文本中与所述知识实体组对应的词语替换为所述预打标数据，生成标注文本。

[0199] 其中，所述医疗文本标注装置还包括纠错模块908，其具体用于：

[0200] 调用实体命名纠错算法，识别被标注的知识实体的命名是否正确；

[0201] 若不正确，则查询互联网词典，选择与所述知识实体相似的实体名称进行替换，得到新的标注文本；或者，通知修改用户进行人工修改。

[0202] 可选地，所述训练模块906具体还用于：

[0203] 获取替换后的实体名称作为训练集，对所述标注模型进行模型优化和训练，得到优化后的标注模型。

[0204] 综上，基于不同的请求选择的标注颜色是不相同的，所以在多人同时请求时，其也可以进行区分，从而实现了可以进行多人合作，分配标注任务，只需选择文本跨度并对其进行标注即可，支持快捷键，因此您可以快速标注文本跨度，在标注过程中支持无效标签、存疑标签对数据的一个判断，标注操作实时更新，翻页自动保存，模型预打标与人工标注相结合，利用模型跑出的结果辅助人工标注，不同的实体标签配置生成不同颜色，人工核对标注结果是否正确，只需对模型预判错误的进行修改，提高标注的一致性及效率。

[0205] 上面图9和图10从模块化功能实体的角度对本发明实施例中的医疗文本标注装置进行详细描述，下面从硬件处理的角度对本发明实施例中医疗文本标注设备进行详细描述，而医疗文本标注装置可以插件的形式设置与所述医疗文本标注设备种实现对话术的识别。

[0206] 图11是本发明实施例提供的一种医疗文本标注设备的结构示意图，该医疗文本标注设备600可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上处理器(central processing units,CPU)610(例如，一个或一个以上处理器)和存储器620，一个或一个以上存储应用程序633或数据632的存储介质630(例如一个或一个以上海量存储设

备)。其中,存储器620和存储介质630可以是短暂存储或持久存储。存储在存储介质630的程序可以包括一个或一个以上模块(图示没标出),每个模块可以包括对医疗文本标注设备600中的一系列指令操作。更进一步地,处理器610可以设置为与存储介质630通信,在医疗文本标注设备600上执行存储介质630中的一系列指令操作,以实现上述医疗文本标注方法的步骤。

[0207] 医疗文本标注设备600还可以包括一个或一个以上电源640,一个或一个以上有线或无线网络接口650,一个或一个以上输入输出接口660,和/或,一个或一个以上操作系统631,例如Windows Serve,Mac OS X,Unix,Linux,FreeBSD等等。本领域技术人员可以理解,图11示出的医疗文本标注设备结构并不构成对本申请提供的医疗文本标注设备的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件布置。

[0208] 本发明所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain),本质上是一个去中心化的数据库,是一串使用密码学方法相关联产生的数据块,每一个数据块中包含了一批次网络交易的信息,用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

[0209] 本发明还提供一种计算机可读存储介质,该计算机可读存储介质可以为非易失性计算机可读存储介质,该计算机可读存储介质也可以为易失性计算机可读存储介质,所述计算机可读存储介质中存储有指令,当所述指令在计算机上运行时,使得计算机执行上述各实施例提供的医疗文本标注方法的步骤。

[0210] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统,装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0211] 所述集成的单元如果以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全部或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(read-only memory,ROM)、随机存取存储器(random access memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0212] 以上所述,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。

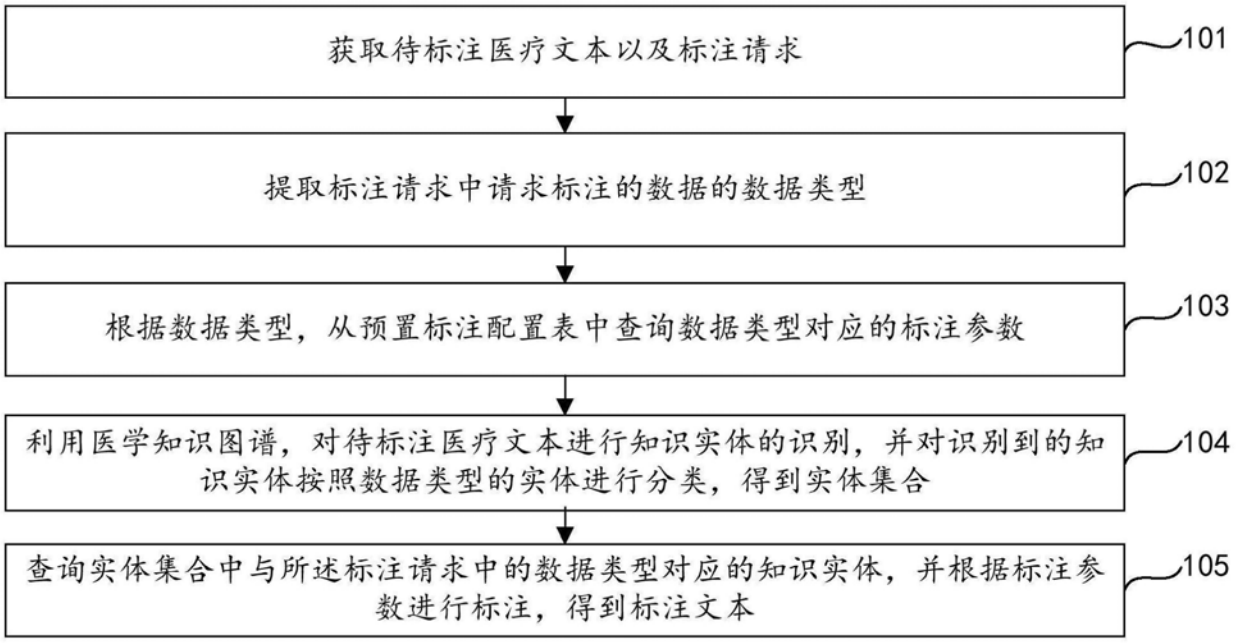

获取待标注医疗文本以及标注请求
101
提取标注请求中请求标注的数据的数据类型
102
根据数据类型，从预置标注配置表中查询数据类型对应的标注参数
103
利用医学知识图谱，对待标注医疗文本进行知识实体的识别，并对识别到的知识实体按照数据类型的实体进行分类，得到实体集合
104
查询实体集合中与所述标注请求中的数据类型对应的知识实体，并根据标注参数进行标注，得到标注文本
105

图1

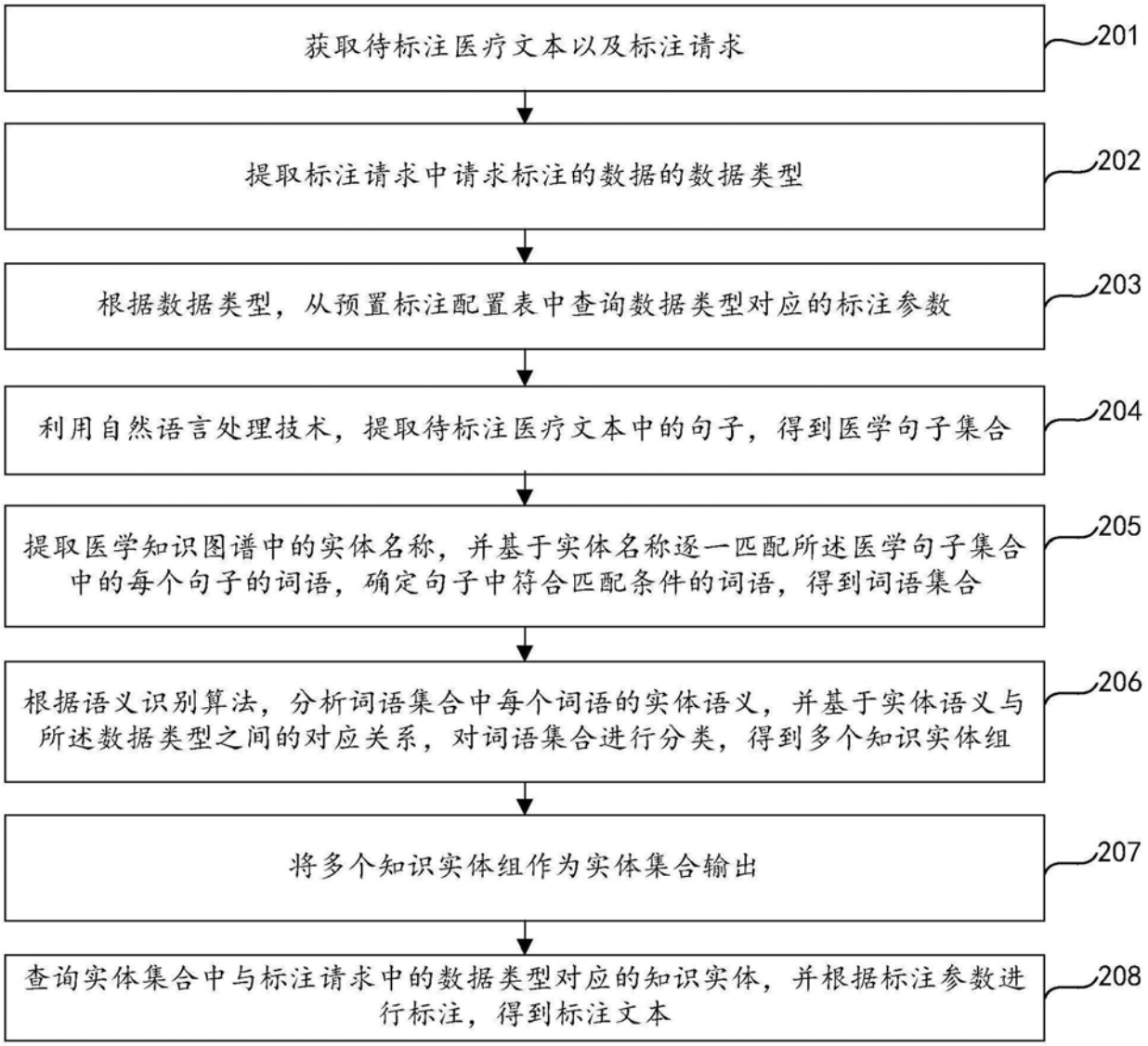

获取待标注医疗文本以及标注请求
201
提取标注请求中请求标注的数据的数据类型
202
根据数据类型，从预置标注配置表中查询数据类型对应的标注参数
203
利用自然语言处理技术，提取待标注医疗文本中的句子，得到医学句子集合
204
提取医学知识图谱中的实体名称，并基于实体名称逐一匹配所述医学句子集合中的每个句子的词语，确定句子中符合匹配条件的词语，得到词语集合
205
根据语义识别算法，分析词语集合中每个词语的实体语义，并基于实体语义与所述数据类型之间的对应关系，对词语集合进行分类，得到多个知识实体组
206
将多个知识实体组作为实体集合输出
207
查询实体集合中与标注请求中的数据类型对应的知识实体，并根据标注参数进行标注，得到标注文本
208

图2

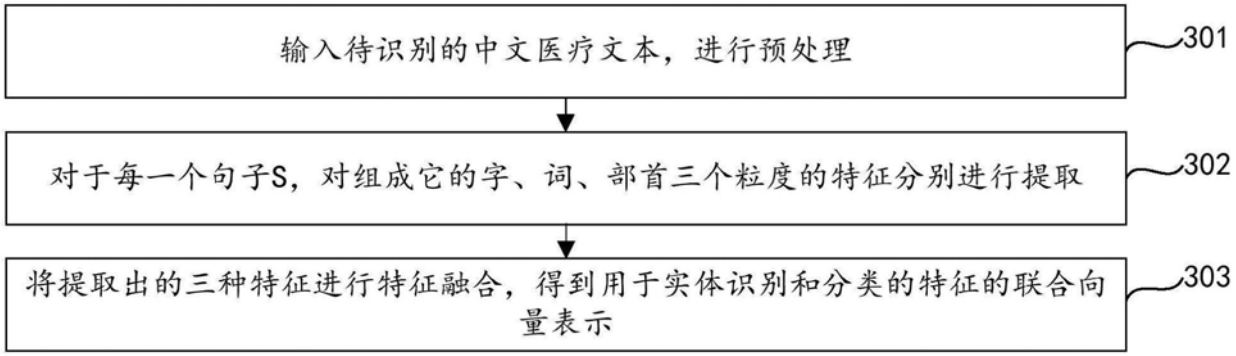

输入待识别的中文医疗文本，进行预处理
301
对于每一个句子S，对组成它的字、词、部首三个粒度的特征分别进行提取
302
将提取出的三种特征进行特征融合，得到用于实体识别和分类的特征的联合向量表示
303

图3

图4

标签ID	标签名称	标签快捷键code	标签颜色	修改人	修改时间	操作
Drug	d	68	green	刘某某	2019-10-12 10:30:32	修改 删除
药物	v	86	blue	系统管理员	2019-10-16 10:40:35	修改 删除

图5

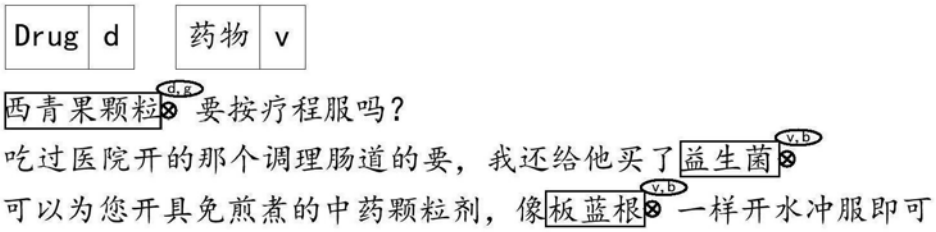

Drug d
药物 v
西青果颗粒 d,g 要按疗程服吗？
吃过医院开的那个调理肠道的要，我还给他买了益生菌 v,b
可以为您开具免煎煮的中药颗粒剂，像板蓝根 v,b 一样开水冲服即可

图6


Drug d
药物 v
目前的情况可以口服西替利嗪 d,g 乌蛇止痒丸 v,b 外用用氟芬那酸丁酯软膏 v,b 止痒，一天两次，均不含激素
要吃蒙脱石散 v,b 吗？肚脐贴那个牌子的呢？
目前可以使用维生素b1 v,b 甲钴胺 v,b 营养神经

图7

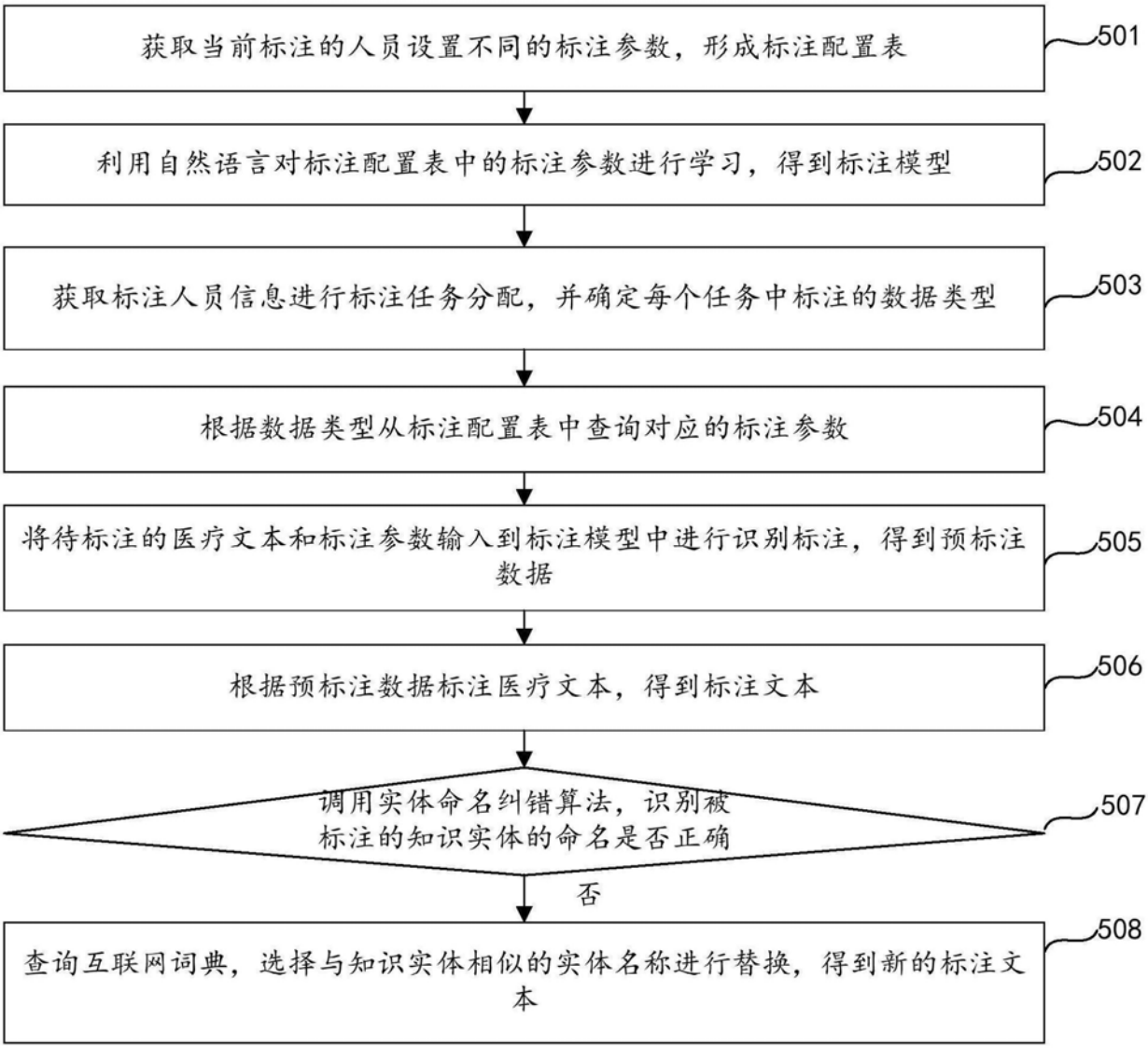

获取当前标注的人员设置不同的标注参数，形成标注配置表
501
利用自然语言对标注配置表中的标注参数进行学习，得到标注模型
502
获取标注人员信息进行标注任务分配，并确定每个任务中标注的数据类型
503
根据数据类型从标注配置表中查询对应的标注参数
504
将待标注的医疗文本和标注参数输入到标注模型中进行识别标注，得到预标注数据
505
根据预标注数据标注医疗文本，得到标注文本
506
调用实体命名纠错算法，识别被标注的知识实体的命名是否正确
507
否
查询互联网词典，选择与知识实体相似的实体名称进行替换，得到新的标注文本
508

图8

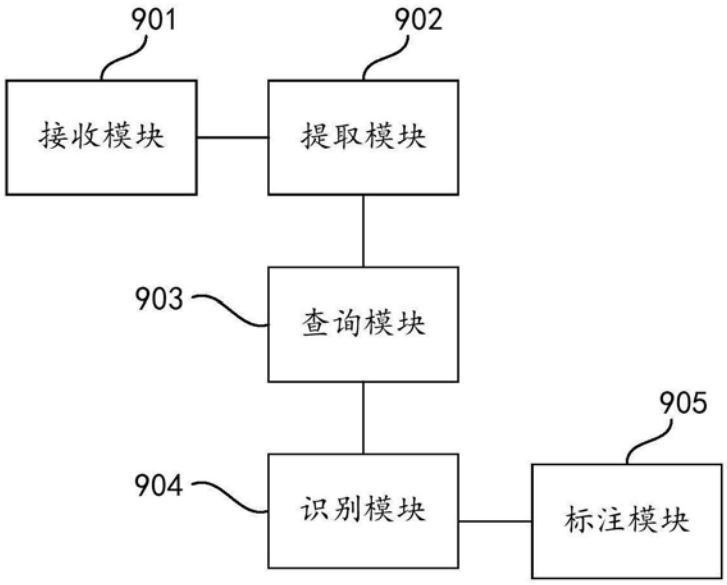

901
902
接收模块
提取模块
903
查询模块
905
904
识别模块
标注模块

图9

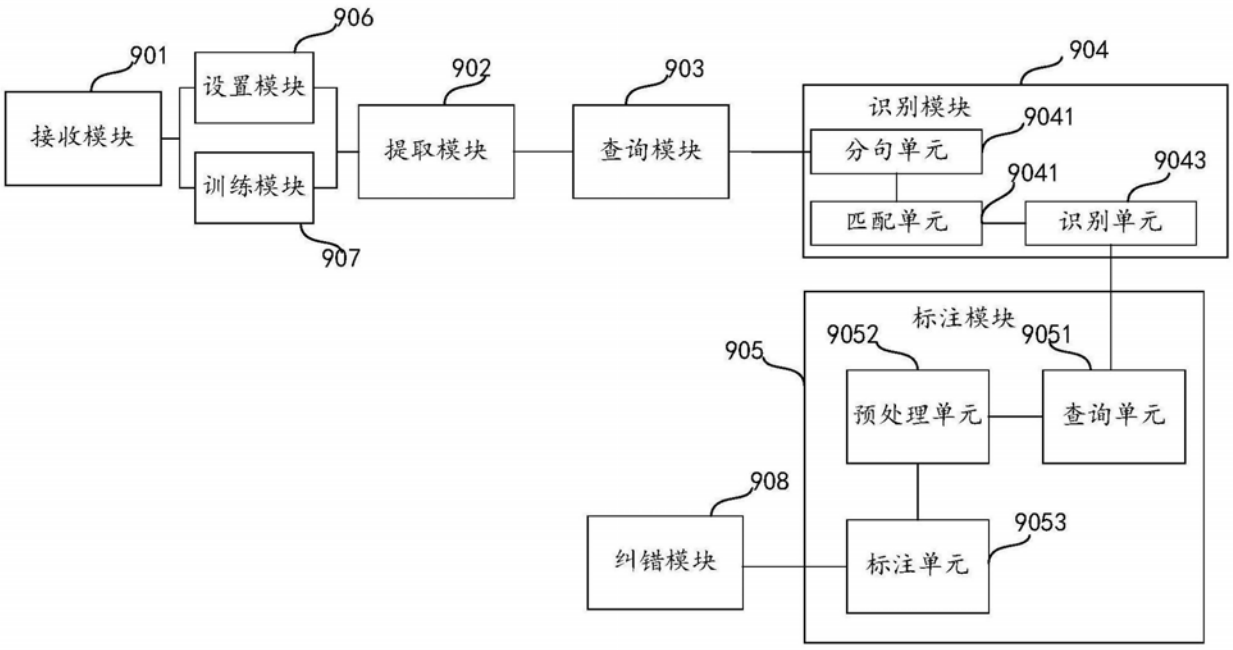

906
901
设置模块
902
903
904
识别模块
9041
分句单元
接收模块
训练模块
提取模块
查询模块
9041
9043
匹配单元
识别单元
907
标注模块
9052
9051
905
预处理单元
查询单元
908
9053
纠错模块
标注单元

图10

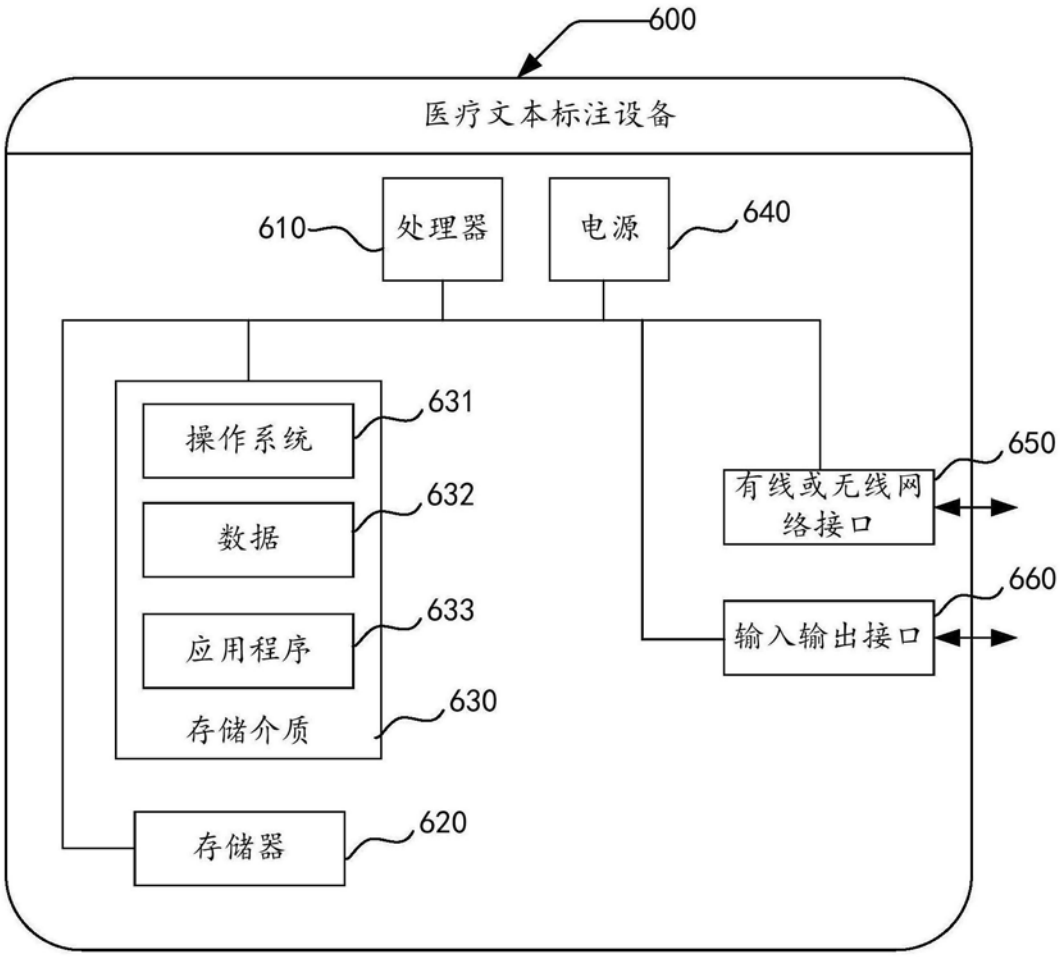

600
医疗文本标注设备
610
处理器
电源
640
631
操作系统
632
数据
633
应用程序
630
存储介质
620
存储器
650
有线或无线网络接口
660
输入输出接口

图11