(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 112199954 A
(43)申请公布日 2021.01.08

(21)申请号 202011080585.4

(22)申请日 2020.10.10

(71)申请人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼

(72)发明人 方春华

(74)专利代理机构 深圳市世联合知识产权代理有限公司 44385
代理人 汪琳琳

(51)Int.Cl.
*G06F 40/295*(2020.01)
*G06F 40/30*(2020.01)
*G06F 40/242*(2020.01)
*G06F 16/33*(2019.01)
*G10L 15/26*(2006.01)
*G10L 15/30*(2013.01)
*G10L 25/66*(2013.01)
*G10L 15/18*(2013.01)

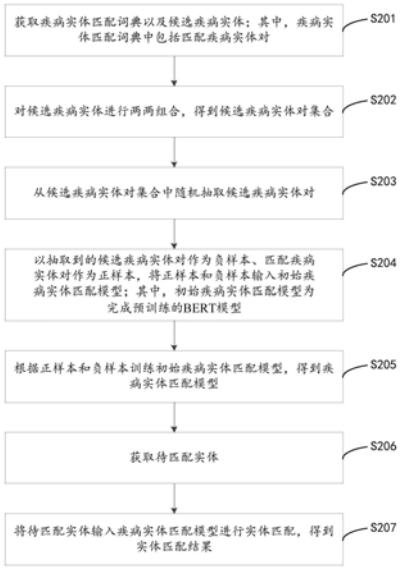
权利要求书2页 说明书12页 附图3页

(54)发明名称

基于语音语义的疾病实体匹配方法、装置及计算机设备

(57)摘要

本申请实施例属于人工智能领域，可应用于医疗科技领域，涉及一种基于语音语义的疾病实体匹配方法、装置、计算机设备及存储介质，所述方法包括：获取包含匹配疾病实体对的疾病实体匹配词典以及候选疾病实体；对候选疾病实体进行两两组合，得到候选疾病实体对集合，并从中随机抽取候选疾病实体对；以抽取到的候选疾病实体对作为负样本、匹配疾病实体对作为正样本，将正样本和负样本输入初始疾病实体匹配模型进行模型训练，得到疾病实体匹配模型；获取待匹配实体并输入疾病实体匹配模型，得到实体匹配结果。此外，本申请还涉及区块链技术，疾病实体匹配词典可存储于区块链中。本申请提高了疾病实体匹配效率。




CN 112199954 A

1.一种基于语音语义的疾病实体匹配方法,其特征在于,包括下述步骤:

获取疾病实体匹配词典以及候选疾病实体;其中,所述疾病实体匹配词典中包括匹配疾病实体对;

对所述候选疾病实体进行两两组合,得到候选疾病实体对集合;

从所述候选疾病实体对集合中随机抽取候选疾病实体对;

以抽取到的候选疾病实体对作为负样本、所述匹配疾病实体对作为正样本,将所述正样本和所述负样本输入初始疾病实体匹配模型;其中,所述初始疾病实体匹配模型为完成预训练的BERT模型;

根据所述正样本和所述负样本训练所述初始疾病实体匹配模型,得到疾病实体匹配模型;

获取待匹配实体;

将所述待匹配实体输入所述疾病实体匹配模型进行实体匹配,得到实体匹配结果。

2.根据权利要求1所述的基于语音语义的疾病实体匹配方法,其特征在于,在所述获取疾病实体匹配词典以及候选疾病实体的步骤之前还包括:

获取疾病语料信息;

通过语义信息识别所述疾病语料信息中的匹配疾病实体对;

基于识别到的匹配疾病实体对构建疾病实体匹配词典。

3.根据权利要求1所述的基于语音语义的疾病实体匹配方法,其特征在于,所述从所述候选疾病实体对集合中随机抽取候选疾病实体对的步骤包括:

获取所述候选疾病实体对集合在所述疾病实体匹配词典中的补集;

从所述补集中随机抽取预设数量的候选疾病实体对;

计算抽取到的候选疾病实体对的实体相似度;

筛选实体相似度小于相似度阈值的候选疾病实体对。

4.根据权利要求1所述的基于语音语义的疾病实体匹配方法,其特征在于,所述根据所述正样本和所述负样本训练所述初始疾病实体匹配模型,得到疾病实体匹配模型的步骤包括:

将所述正样本和所述负样本各自进行拼接,并添加样本标签,得到待处理样本;

将所述待处理样本输入所述初始疾病实体匹配模型的网络层,得到所述待处理样本的表征向量;

对所述表征向量进行计算,输出匹配预测概率;

根据所述匹配预测概率和所述样本标签计算模型损失;

根据所述模型损失调整所述初始疾病实体匹配模型的模型参数,直至模型收敛,得到疾病实体匹配模型。

5.根据权利要求1所述的基于语音语义的疾病实体匹配方法,其特征在于,在所述以抽取到的候选疾病实体对作为负样本、所述匹配疾病实体对作为正样本,将所述正样本和所述负样本输入初始疾病实体匹配模型的步骤之前还包括:

获取医学语料数据集;

将所述医学语料数据集输入BERT模型以进行预训练,得到初始疾病实体匹配模型。

6.根据权利要求1所述的基于语音语义的疾病实体匹配方法,其特征在于,所述将所述

待匹配实体输入所述疾病实体匹配模型进行实体匹配，得到实体匹配结果的步骤包括：

获取疾病实体词典；

将所述待匹配实体与所述疾病实体词典中的各疾病实体进行组合，得到第一待匹配实体对；

将所述第一待匹配实体对输入所述疾病实体匹配模型，得到匹配疾病实体对；

根据所述匹配疾病实体对，在所述疾病实体词典中确定与所述待匹配实体相匹配的疾病实体，并将确定的疾病实体作为实体匹配结果。

7.根据权利要求1所述的基于语音语义的疾病实体匹配方法，其特征在于，所述将所述待匹配实体输入所述疾病实体匹配模型进行实体匹配，得到实体匹配结果的步骤包括：

对所述待匹配实体进行两两组合，得到第二待匹配实体对；

将所述第二待匹配实体对输入所述疾病实体匹配模型，得到所述第二待匹配实体对中的匹配疾病实体对，并将得到的匹配疾病实体对作为实体匹配结果。

8.一种基于语音语义的疾病实体匹配装置，其特征在于，包括：

第一获取模块，用于获取疾病实体匹配词典以及候选疾病实体；其中，所述疾病实体匹配词典中包括匹配疾病实体对；

实体组合模块，用于对所述候选疾病实体进行两两组合，得到候选疾病实体对集合；

实体对抽取模块，用于从所述候选疾病实体对集合中随机抽取候选疾病实体对；

样本输入模块，用于以抽取到的候选疾病实体对作为负样本、所述匹配疾病实体对作为正样本，将所述正样本和所述负样本输入初始疾病实体匹配模型；其中，所述初始疾病实体匹配模型为完成预训练的BERT模型；

模型训练模块，用于根据所述正样本和所述负样本训练所述初始疾病实体匹配模型，得到疾病实体匹配模型；

第二获取模块，用于获取待匹配实体；

实体匹配模块，用于将所述待匹配实体输入所述疾病实体匹配模型进行实体匹配，得到实体匹配结果。

9.一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述处理器执行所述计算机可读指令时实现如权利要求1至7中任一项所述的基于语音语义的疾病实体匹配方法的步骤。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现如权利要求1至7中任一项所述的基于语音语义的疾病实体匹配方法的步骤。

# 基于语音语义的疾病实体匹配方法、装置及计算机设备

## 技术领域

[0001] 本申请涉及人工智能技术领域，尤其涉及一种基于语音语义的疾病实体匹配方法、装置及计算机设备。

## 背景技术

[0002] 病历是在医疗活动中记录的个体健康信息，病历中记录了疾病实体，即病人所患疾病的名称。病历中记载的疾病实体可能存在多种表达，例如，强迫性障碍和强迫症属于同一种疾病，因此经常需要判断两个疾病实体是否匹配。

[0003] 传统的疾病实体匹配，有的由人工进行判断，在疾病实体较多时，人工判断需要大量时间，效率低下。有的是借助计算机进行疾病实体匹配，例如对疾病实体进行属性匹配、上下文匹配等。然而，这些匹配技术都需要预先获取大规模的疾病语料，且对语料质量要求较高，因此语料的收集和预处理所需时间较长，导致疾病实体匹配的效率依然较低。

## 发明内容

[0004] 本申请实施例的目的在于提出一种基于语音语义的疾病实体匹配方法、装置、计算机设备及存储介质，以解决疾病实体匹配效率较低的问题。

[0005] 为了解决上述技术问题，本申请实施例提供一种基于语音语义的疾病实体匹配方法，采用了如下所述的技术方案：

[0006] 获取疾病实体匹配词典以及候选疾病实体；其中，所述疾病实体匹配词典中包括匹配疾病实体对；

[0007] 对所述候选疾病实体进行两两组合，得到候选疾病实体对集合；

[0008] 从所述候选疾病实体对集合中随机抽取候选疾病实体对；

[0009] 以抽取到的候选疾病实体对作为负样本、所述匹配疾病实体对作为正样本，将所述正样本和所述负样本输入初始疾病实体匹配模型；其中，所述初始疾病实体匹配模型为完成预训练的BERT模型；

[0010] 根据所述正样本和所述负样本训练所述初始疾病实体匹配模型，得到疾病实体匹配模型；

[0011] 获取待匹配实体；

[0012] 将所述待匹配实体输入所述疾病实体匹配模型进行实体匹配，得到实体匹配结果。

[0013] 进一步的，在所述获取疾病实体匹配词典以及候选疾病实体的步骤之前还包括：

[0014] 获取疾病语料信息；

[0015] 通过语义信息识别所述疾病语料信息中的匹配疾病实体对；

[0016] 基于识别到的匹配疾病实体对构建疾病实体匹配词典。

[0017] 进一步的，所述从所述候选疾病实体对集合中随机抽取候选疾病实体对的步骤包括：

[0018] 获取所述候选疾病实体对集合在所述疾病实体匹配词典中的补集；

[0019] 从所述补集中随机抽取预设数量的候选疾病实体对；

[0020] 计算抽取到的候选疾病实体对的实体相似度；

[0021] 筛选实体相似度小于相似度阈值的候选疾病实体对。

[0022] 进一步的，所述根据所述正样本和所述负样本训练所述初始疾病实体匹配模型，得到疾病实体匹配模型的步骤包括：

[0023] 将所述正样本和所述负样本各自进行拼接，并添加样本标签，得到待处理样本；

[0024] 将所述待处理样本输入所述初始疾病实体匹配模型的网络层，得到所述待处理样本的表征向量；

[0025] 对所述表征向量进行计算，输出匹配预测概率；

[0026] 根据所述匹配预测概率和所述样本标签计算模型损失；

[0027] 根据所述模型损失调整所述初始疾病实体匹配模型的模型参数，直至模型收敛，得到疾病实体匹配模型。

[0028] 进一步的，在所述以抽取到的候选疾病实体对作为负样本、所述匹配疾病实体对作为正样本，将所述正样本和所述负样本输入初始疾病实体匹配模型的步骤之前还包括：

[0029] 获取医学语料数据集；

[0030] 将所述医学语料数据集输入BERT模型以进行预训练，得到初始疾病实体匹配模型。

[0031] 进一步的，所述将所述待匹配实体输入所述疾病实体匹配模型进行实体匹配，得到实体匹配结果的步骤包括：

[0032] 获取疾病实体词典；

[0033] 将所述待匹配实体与所述疾病实体词典中的各疾病实体进行组合，得到第一待匹配实体对；

[0034] 将所述第一待匹配实体对输入所述疾病实体匹配模型，得到匹配疾病实体对；

[0035] 根据所述匹配疾病实体对，在所述疾病实体词典中确定与所述待匹配实体相匹配的疾病实体，并将确定的疾病实体作为实体匹配结果。

[0036] 进一步的，所述将所述待匹配实体输入所述疾病实体匹配模型进行实体匹配，得到实体匹配结果的步骤包括：

[0037] 对所述待匹配实体进行两两组合，得到第二待匹配实体对；

[0038] 将所述第二待匹配实体对输入所述疾病实体匹配模型，得到所述第二待匹配实体对中的匹配疾病实体对，并将得到的匹配疾病实体对作为实体匹配结果。

[0039] 为了解决上述技术问题，本申请实施例还提供一种基于语音语义的疾病实体匹配装置，采用了如下所述的技术方案：

[0040] 第一获取模块，用于获取疾病实体匹配词典以及候选疾病实体；其中，所述疾病实体匹配词典中包括匹配疾病实体对；

[0041] 实体组合模块，用于对所述候选疾病实体进行两两组合，得到候选疾病实体对集合；

[0042] 实体对抽取模块，用于从所述候选疾病实体对集合中随机抽取候选疾病实体对；

[0043] 样本输入模块，用于以抽取到的候选疾病实体对作为负样本、所述匹配疾病实体

对作为正样本，将所述正样本和所述负样本输入初始疾病实体匹配模型；其中，所述初始疾病实体匹配模型为完成预训练的BERT模型；

[0044] 模型训练模块，用于根据所述正样本和所述负样本训练所述初始疾病实体匹配模型，得到疾病实体匹配模型；

[0045] 第二获取模块，用于获取待匹配实体；

[0046] 实体匹配模块，用于将所述待匹配实体输入所述疾病实体匹配模型进行实体匹配，得到实体匹配结果。

[0047] 为了解决上述技术问题，本申请实施例还提供一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机程序，所述处理器执行所述计算机程序时实现上述所述的基于语音语义的疾病实体匹配方法的步骤。

[0048] 为了解决上述技术问题，本申请实施例还提供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，所述计算机程序被处理器执行时实现上述所述的基于语音语义的疾病实体匹配方法的步骤。

[0049] 与现有技术相比，本申请实施例主要有以下有益效果：获取疾病实体匹配词典以及候选疾病实体后，对候选疾病实体进行两两组合以构建负样本，以疾病实体匹配词典作为正样本；将正样本和负样本输入初始疾病实体匹配模型以进行充分训练，初始疾病实体匹配模型可以是完成预训练的BERT模型，具有丰富的语义信息，当训练样本规模较小时也可以获得精准的匹配效果，缩短了训练所需时间，提高了疾病实体匹配模型的训练效率；训练完成后，疾病实体匹配模型即可对输入的待匹配实体进行实体匹配，提高了疾病实体匹配的效率。

## 附图说明

[0050] 为了更清楚地说明本申请中的方案，下面将对本申请实施例描述中所需要使用的附图作一个简单介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0051] 图1是本申请可以应用于其中的示例性系统架构图；

[0052] 图2是根据本申请的基于语音语义的疾病实体匹配方法的一个实施例的流程图；

[0053] 图3是根据本申请的基于语音语义的疾病实体匹配装置的一个实施例的结构示意图；

[0054] 图4是根据本申请的计算机设备的一个实施例的结构示意图。

## 具体实施方式

[0055] 除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同；本文中在申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请；本申请的说明书和权利要求书及上述附图说明中的术语“包括”和“具有”以及它们的任何变形，意图在于覆盖不排他的包含。本申请的说明书和权利要求书或上述附图中的术语“第一”、“第二”等是用于区别不同对象，而不是用于描述特定顺序。

[0056] 在本文中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包

含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

[0057] 为了使本技术领域的人员更好地理解本申请方案，下面将结合附图，对本申请实施例中的技术方案进行清楚、完整地描述。

[0058] 如图1所示，系统架构100可以包括终端设备101、102、103，网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0059] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、102、103上可以安装有各种通讯客户端应用，例如网页浏览器应用、购物类应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等。

[0060] 终端设备101、102、103可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、MP3播放器(Moving Picture ExpertsGroup Audio Layer III，动态影像专家压缩标准音频层面3)、MP4(Moving PictureExperts Group Audio Layer IV，动态影像专家压缩标准音频层面4)播放器、膝上型便携计算机和台式计算机等等。

[0061] 服务器105可以是提供各种服务的服务器，例如对终端设备101、102、103上显示的页面提供支持的后台服务器。

[0062] 需要说明的是，本申请实施例所提供的基于语音语义的疾病实体匹配方法一般由服务器执行，相应地，基于语音语义的疾病实体匹配装置一般设置于服务器中。本申请可应用于医疗科技领域。

[0063] 应该理解，图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

[0064] 继续参考图2，示出了根据本申请的基于语音语义的疾病实体匹配方法的一个实施例的流程图。所述的基于语音语义的疾病实体匹配方法，包括以下步骤：

[0065] 步骤S201，获取疾病实体匹配词典以及候选疾病实体；其中，疾病实体匹配词典中包括匹配疾病实体对。

[0066] 在本实施例中，基于语音语义的疾病实体匹配方法运行于其上的电子设备(例如图1所示的服务器)可以通过有线连接方式或者无线连接方式与终端设备进行通信。需要指出的是，上述无线连接方式可以包括但不限于3G/4G连接、WiFi连接、蓝牙连接、WiMAX连接、Zigbee连接、UWB(ultra wideband)连接、以及其他现在已知或将来开发的无线连接方式。

[0067] 其中，疾病实体匹配词典用于记录匹配疾病实体对；匹配疾病实体对可以是匹配的疾病实体的组合。候选疾病实体可以是单独的疾病实体，用于构建训练样本。

[0068] 具体地，服务器接收到模型训练指令后，从数据库中获取疾病实体匹配词典以及候选疾病实体，或者从终端接收疾病实体匹配词典以及候选疾病实体。本申请对疾病实体匹配词典的规模要求不高，小规模的疾病实体匹配词典即可满足训练需求，节约了构建疾病实体匹配词典的人力成本以及时间成本。

[0069] 需要强调的是，为进一步保证上述疾病实体匹配词典的私密和安全性，上述疾病实体匹配词典还可以存储于一区块链的节点中。

[0070] 本申请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。

[0071] 步骤S202,对候选疾病实体进行两两组合,得到候选疾病实体对集合。

[0072] 具体地,服务器将候选疾病实体进行两两组合,得到多组候选疾病实体对,全部候选疾病实体对构成候选疾病实体对集合。举例说明,当候选疾病实体有100个时,两两组合后得到$C_{100}^{2}=4950$组候选疾病实体对,4950组候选疾病实体对组成了候选疾病实体对集合。

[0073] 步骤S203,从候选疾病实体对集合中随机抽取候选疾病实体对。

[0074] 具体地,服务器可以不必将整个候选疾病实体对集合用于训练。当候选疾病实体较多时,候选疾病实体对集合规模也会较大。为了提高处理速度,服务器可以随机从候选疾病实体对集合中抽取预设数量的候选疾病实体对。

[0075] 步骤S204,以抽取到的候选疾病实体对作为负样本、匹配疾病实体对作为正样本,将正样本和负样本输入初始疾病实体匹配模型;其中,初始疾病实体匹配模型为完成预训练的BERT模型。

[0076] 具体地,服务器输入初始疾病实体匹配模型的样本既包含正样本,又包含负样本,以充分训练初始疾病实体匹配模型;其中,抽取到的候选疾病实体将作为负样本,疾病实体匹配词典中的匹配疾病实体对作为正样本。

[0077] 服务器将正样本和负样本输入初始疾病实体匹配模型,初始疾病实体匹配模型可以是完成了预训练的BERT(Bidirectional Encoder Representation from Transformers)模型。

[0078] 在一个实施例中,上述步骤S205之前还可以包括:获取医学语料数据集;将医学语料数据集输入BERT模型以进行预训练,得到初始疾病实体匹配模型。

[0079] 其中,医学语料数据集可以是医学语料信息组成的数据集。

[0080] 具体地,服务器获取医学语料数据集,医学语料数据集中的医学语料信息可以来自各个医学疾病领域。服务器根据医学语料数据集对BERT模型进行预训练,得到初始疾病实体匹配模型。BERT模型学习了丰富的语义信息,使得初始疾病实体匹配模型在样本规模有限的情况下,也可以有效地进行训练,且在训练完毕后,面对不同领域的疾病实体时能够达到较高的匹配准确率。

[0081] BERT模型中使用了Masked language model即遮蔽语言模型,用以克服从左到右的预训练与以及无法利用下文信息的单向局限性,遮蔽语言模型能够表征融合上下文信息。

[0082] 遮蔽语言模型随机将一定比例的token(自然语言处理中的单位,例如可以是单词)替换成mask(掩膜),然后将mask对应位置的最后一层隐藏层的输出送入softmax(逻辑回归)层,用来预测被mask掉的token所对应的原始字符串。

[0083] BERT模型将大量在下游自然语言处理任务中做的操作转移到预训练词向量中,通过BERT获得词向量后,在词向量基础上加入分类器。例如对于句子对或实体对分类任务,在预训练的基础上,根据下游任务进行微调,BERT模型获取最后一层的表征,加上softmax层

预测概率。最后一层的表征可以学习到语义级别的信息，并且利用了前面各层的信息。

[0084] 本实施例中，通过医学语料数据集对BERT模型进行训练，使得BERT模型学习到丰富的语义信息，保证了疾病实体匹配的准确性。

[0085] 步骤S205，根据正样本和负样本训练初始疾病实体匹配模型，得到疾病实体匹配模型。

[0086] 具体地，服务器将正样本和负样本输入初始疾病实体匹配模型，初始疾病实体匹配模型根据输入的样本分别输出匹配预测结果，匹配预测结果可以是一个二分类的结果。

[0087] 初始疾病实体匹配模型根据匹配预测结果与样本标签计算模型损失，其中，正样本的样本标签取一个值，负样本的样本标签取另一个值。服务器以减小模型损失为目标对初始疾病实体匹配模型进行参数调整，然后根据正样本和负样本继续对初始疾病实体匹配模型进行训练，直至模型收敛，得到疾病实体匹配模型。

[0088] 在一个实施例中，可以根据Focal Loss损失函数计算模型损失。

[0089] 步骤S206，获取待匹配实体。

[0090] 其中，待匹配实体为输入的疾病实体，用于疾病实体匹配。

[0091] 具体地，得到疾病实体匹配模型后即可进行疾病实体匹配。用户可以通过终端输入待匹配实体，由终端将待匹配实体发送至服务器。

[0092] 步骤S207，将待匹配实体输入疾病实体匹配模型进行实体匹配，得到实体匹配结果。

[0093] 具体地，服务器将待匹配实体输入疾病实体匹配模型，疾病实体匹配模型既可以对单独的待匹配实体进行实体匹配，输出与之匹配的疾病实体作为匹配结果；也可以对多个待匹配实体进行处理，输出多个待匹配实体中的匹配疾病实体对作为实体匹配结果。

[0094] 本实施例中，获取疾病实体匹配词典以及候选疾病实体后，对候选疾病实体进行两两组合以构建负样本，以疾病实体匹配词典作为正样本；将正样本和负样本输入初始疾病实体匹配模型以进行充分训练，初始疾病实体匹配模型可以是完成预训练的BERT模型，具有丰富的语义信息，当训练样本规模较小时也可以获得精准的匹配效果，缩短了训练所需时间，提高了疾病实体匹配模型的训练效率；训练完成后，疾病实体匹配模型即可对输入的待匹配实体进行实体匹配，提高了疾病实体匹配的效率。

[0095] 进一步的，上述步骤S201之前还可以包括：获取疾病语料信息；通过语义信息识别疾病语料信息中的匹配疾病实体对；基于识别到的匹配疾病实体对构建疾病实体匹配词典。

[0096] 其中，疾病语料信息可以是疾病相关的语料信息。

[0097] 具体地，服务器获取疾病语料信息，疾病语料信息可以通过爬虫获取。爬虫可以爬取疾病相关的词条页面，得到疾病语料信息。服务器根据语义知识库对疾病语料信息进行语义标注，根据语义标注结果得到疾病语料信息中的匹配疾病实体对。举例说明，疾病相关的词条页面中记录了“Y1又名Y2”，服务器通过语义信息得到Y1和Y2可以作为匹配疾病实体对。根据识别到的匹配疾病实体对，服务器可以构建出疾病实体匹配词典。

[0098] 疾病语料信息也可以人工选取并输入服务器，匹配疾病实体对可以由人工对疾病语料信息进行标注。

[0099] 本实施例中，基于疾病语料信息构建的疾病实体匹配词典用于训练初始疾病实体

匹配模型，保证了模型训练的顺利实现。

[0100] 进一步的，上述步骤S203可以包括：获取候选疾病实体对集合在疾病实体匹配词典中的补集；从补集中随机抽取预设数量的候选疾病实体对；计算抽取到的候选疾病实体对的实体相似度；筛选实体相似度小于相似度阈值的候选疾病实体对。

[0101] 具体地，服务器先求候选疾病实体对集合在疾病实体匹配词典中的补集，从而删除已经存在于疾病实体匹配词典中的候选疾病实体对，再从补集中抽取预设数量的候选疾病实体对。

[0102] 服务器计算实体相似度，实体相似度是候选疾病实体对中两个候选疾病实体间的相似度。实体相似度的计算有多种方法，例如通过Jaccard系数、N-Gram(又称N元模型)、Levenshtein距离(也称文本编辑距离)、余弦相似度等方法计算实体相似度。服务器可以单独采用上述的一种方法，也可以综合采用上述方法中的多种。

[0103] 其中，采用Jaccard系数时，将候选疾病实体以字符为单位进行划分，计算公式如下：

[0104] $$Jaccard(A,B)=\frac{len(A\cap B)}{len(A\cup B)} \quad (1)$$

[0105] 其中，A和B表示候选疾病实体，Jaccard(A,B)表示实体相似度，len(A∩B)表示A与B中相同字符的个数，len(A∪B)表示组成A与B所需的非重复字符的个数。

[0106] 在通过N-Gram计算实体相似度时，将候选疾病实体按长度N切分得到词组，其中，上一个词组的尾为下一个词组的头，例如，将“糖尿病”解析为{“$糖”,“糖尿”,“尿病”,“病$”}，其中$为填充字符，N值一般取2或者3。再以如下公式计算实体相似度：

[0107] $$Jaccard(M,N)=\frac{len(A\cap B)}{len(A\cup B)} \quad (2)$$

[0108] 其中，M和N表示候选疾病实体，Jaccard(M,N)是M与N之间的实体相似度；len(M∩N)表示M与N中相同词组的个数，len(M∪N)表示组成M与N所需的非重复词组的个数。

[0109] 当采用Levenshtein距离时，Levenshtein距离越小，实体相似度越高。

[0110] 得到实体相似度之后，服务器获取预设的相似度阈值，将实体相似度与相似度阈值相比较，删去实体相似度大于或等于相似度阈值的候选疾病实体对，保留实体相似度小于相似度阈值的候选疾病实体对，以去除具有较高相似度的候选疾病实体对。

[0111] 候选疾病实体对将作为负样本，已经存在于疾病实体匹配词典中的候选疾病实体对以及实体相似度较高的候选疾病实体对将对模型训练产生负面影响，需要进行去除。

[0112] 本实施例中，通过对候选疾病实体对集合求补集，以及计算实体对相似度，从而去除相似度较高的候选疾病实体对，保证了根据候选疾病实体对构建的负样本的准确性。

[0113] 进一步的，上述步骤S205可以包括：将正样本和负样本各自进行拼接，并添加样本标签，得到待处理样本；将待处理样本输入初始疾病实体匹配模型的网络层，得到待处理样本的表征向量；对表征向量进行计算，输出匹配预测概率；根据匹配预测概率和样本标签计算模型损失；根据模型损失调整初始疾病实体匹配模型的模型参数，直至模型收敛，得到疾病实体匹配模型。

[0114] 具体地，正样本和负样本同时输入初始疾病实体匹配模型。初始疾病实体匹配模型对正样本和负样本的处理方式相同，在两个候选疾病实体间添加【SEP】字符，然后拼接在

一起;再在拼接后的字符串首尾分别加【CLS】、【SEP】字符;服务器还可以添加样本标签,其中,正样本的样本标签一致,负样本的样本标签一致,得到待处理样本。

[0115] 待处理样本被输入初始疾病实体匹配模型的网络层,输出待处理样本的表征向量sequence_output,在一个实施例中,表征向量的维度可以是1*768。服务器将表征向量进行矩阵运算,再乘上偏置矩阵[1,2],再加上softmax(逻辑回归)层,得到匹配预测概率,匹配预测概率是1*2的向量,分别表示两个实体匹配和不匹配的概率。服务器根据匹配预测概率和样本标签计算交叉熵得到模型损失,以减小模型损失为目标调整初始疾病实体匹配模型的模型参数,然后重新进行训练,直至模型收敛,得到疾病实体匹配模型。当模型收敛时,模型损失小于预设的损失阈值。

[0116] 本实施例中,对样本进行处理输出匹配预测概率,并根据样本标签计算模型损失,根据模型损失对模型进行微调直至模型收敛,得到的疾病实体匹配模型可以准确地进行疾病实体的匹配判断。

[0117] 进一步的,在一个实施例中,上述步骤S207可以包括:获取疾病实体词典;将待匹配实体与疾病实体词典中的各疾病实体进行组合,得到第一待匹配实体对;将第一待匹配实体对输入疾病实体匹配模型,得到匹配疾病实体对;根据匹配疾病实体对,在疾病实体词典中确定与待匹配实体相匹配的疾病实体,并将确定的疾病实体作为实体匹配结果。

[0118] 其中,疾病实体词典可以是记录疾病实体的词典。

[0119] 具体地,可以使用疾病实体匹配模型进行单个待匹配疾病实体的匹配。用户可以通过终端输入待匹配实体。服务器获取待匹配实体,并读取存储的疾病实体词典。疾病实体词典中记录了大量的疾病实体,服务器将待匹配实体与疾病实体词典中的各疾病实体逐一组合,得到多组第一待匹配实体对。服务器将第一待匹配实体对输入疾病实体匹配模型,以判断第一待匹配实体对中的待匹配实体与疾病实体是否匹配,若可以匹配,将被标记为匹配疾病实体对。服务器将匹配疾病实体对中来自疾病实体词典的疾病实体作为实体匹配结果,并将实体匹配结果输出至终端,以展示与待匹配实体相匹配的疾病实体,使得用户无需再从互联网中搜索、查找与待匹配实体相关的疾病实体,方便高效。

[0120] 服务器还可以查询待匹配实体是否存在于疾病实体词典,若不存在,则将待匹配实体补充到疾病实体词典中,以扩充疾病实体词典,提高对待匹配实体的匹配能力。

[0121] 本实施例中,只需输入待匹配实体,疾病实体匹配模型将待匹配实体与疾病实体词典中的疾病实体一一进行匹配判断,可以快速地对待匹配实体实现实体匹配。

[0122] 进一步的,在另一个实施例中,上述步骤S207还可以包括:对待匹配实体进行两两组合,得到第二待匹配实体对;将第二待匹配实体对输入疾病实体匹配模型,得到第二待匹配实体对中的匹配疾病实体对,并将得到的匹配疾病实体对作为实体匹配结果。

[0123] 具体地,疾病实体匹配模型还可以同时对多个待匹配实体进行处理,输出多个待匹配实体中的匹配疾病实体对。

[0124] 在应用时,用户可以同时输入多个待匹配实体,服务器先对多个待匹配实体进行两两组合得到第二待匹配实体对,然后将第二待匹配实体对输入疾病实体匹配模型,即可快速识别出多个待匹配实体中存在的匹配疾病实体对,并将得到的匹配疾病实体对作为实体匹配结果输出至终端进行展示。

[0125] 本实施例中,从多个待匹配实体中筛选匹配疾病实体对时,将待匹配实体两两组

合输入疾病实体匹配模型，即可快速对所有实体组合进行判断，提高了匹配效率。

[0126] 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，是可以通过计算机可读指令来指令相关的硬件来完成，该计算机可读指令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)等非易失性存储介质，或随机存储记忆体(Random Access Memory，RAM)等。

[0127] 应该理解的是，虽然附图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，其可以以其他的顺序执行。而且，附图的流程图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，其执行顺序也不必然是依次进行，而是可以与其他步骤或者其他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。

[0128] 进一步参考图3，作为对上述图2所示方法的实现，本申请提供了一种基于语音语义的疾病实体匹配装置的一个实施例，该装置实施例与图2所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0129] 如图3所示，本实施例所述的基于语音语义的疾病实体匹配装置300包括：第一获取模块301、实体组合模块302、实体对抽取模块303、样本输入模块304、模型训练模块305、第二获取模块306以及实体匹配模块307，其中：

[0130] 第一获取模块301，用于获取疾病实体匹配词典以及候选疾病实体；其中，疾病实体匹配词典中包括匹配疾病实体对。

[0131] 实体组合模块302，用于对候选疾病实体进行两两组合，得到候选疾病实体对集合。

[0132] 实体对抽取模块303，用于从候选疾病实体对集合中随机抽取候选疾病实体对。

[0133] 样本输入模块304，用于以抽取到的候选疾病实体对作为负样本、匹配疾病实体对作为正样本，将正样本和负样本输入初始疾病实体匹配模型；其中，初始疾病实体匹配模型为完成预训练的BERT模型。

[0134] 模型训练模块305，用于根据正样本和负样本训练初始疾病实体匹配模型，得到疾病实体匹配模型。

[0135] 第二获取模块306，用于获取待匹配实体。

[0136] 实体匹配模块307，用于将待匹配实体输入疾病实体匹配模型进行实体匹配，得到实体匹配结果。

[0137] 本实施例中，获取疾病实体匹配词典以及候选疾病实体后，对候选疾病实体进行两两组合以构建负样本，以疾病实体匹配词典作为正样本；将正样本和负样本输入初始疾病实体匹配模型以进行充分训练，初始疾病实体匹配模型可以是完成预训练的BERT模型，具有丰富的语义信息，当训练样本规模较小时也可以获得精准的匹配效果，缩短了训练所需时间，提高了疾病实体匹配模型的训练效率；训练完成后，疾病实体匹配模型即可对输入的待匹配实体进行实体匹配，提高了疾病实体匹配的效率。

[0138] 在本实施例的一些可选的实现方式中，上述基于语音语义的疾病实体匹配装置300还包括：信息获取模块、实体对识别模块以及词典构建模块，其中：

[0139] 信息获取模块，用于获取疾病语料信息。

[0140] 实体对识别模块，用于通过语义信息识别疾病语料信息中的匹配疾病实体对。

[0141] 词典构建模块，用于基于识别到的匹配疾病实体对构建疾病实体匹配词典。

[0142] 本实施例中，基于疾病语料信息构建的疾病实体匹配词典用于训练初始疾病实体匹配模型，保证了模型训练的顺利实现。

[0143] 在本实施例的一些可选的实现方式中，上述实体对抽取模块303包括：补集获取子模块、实体对抽取子模块、相似计算子模块以及实体对筛选子模块，其中：

[0144] 补集获取子模块，用于获取候选疾病实体对集合在疾病实体匹配词典中的补集。

[0145] 实体对抽取子模块，用于从补集中随机抽取预设数量的候选疾病实体对。

[0146] 相似计算子模块，用于计算抽取到的候选疾病实体对的实体相似度。

[0147] 实体对筛选子模块，用于筛选实体相似度小于相似度阈值的候选疾病实体对。

[0148] 本实施例中，通过对候选疾病实体对集合求补集，以及计算实体对相似度，从而去除相似度较高的候选疾病实体对，保证了根据候选疾病实体对构建的负样本的准确性。

[0149] 在本实施例的一些可选的实现方式中，上述模型训练模块305包括：样本拼接子模块、样本输入子模块、向量计算子模块、损失计算子模块以及参数调整子模块，其中：

[0150] 样本拼接子模块，用于将正样本和负样本各自进行拼接，并添加样本标签，得到待处理样本。

[0151] 样本输入子模块，用于将待处理样本输入初始疾病实体匹配模型的网络层，得到待处理样本的表征向量。

[0152] 向量计算子模块，用于对表征向量进行计算，输出匹配预测概率。

[0153] 损失计算子模块，用于根据匹配预测概率和样本标签计算模型损失。

[0154] 参数调整子模块，用于根据模型损失调整初始疾病实体匹配模型的模型参数，直至模型收敛，得到疾病实体匹配模型。

[0155] 本实施例中，对样本进行处理输出匹配预测概率，并根据样本标签计算模型损失，根据模型损失对模型进行微调直至模型收敛，得到的疾病实体匹配模型可以准确地进行疾病实体的匹配判断。

[0156] 在本实施例的一些可选的实现方式中，上述基于语音语义的疾病实体匹配装置300还包括：数据集获取模块以及数据集输入模块，其中：

[0157] 数据集获取模块，用于获取医学语料数据集。

[0158] 数据集输入模块，用于将医学语料数据集输入BERT模型以进行预训练，得到初始疾病实体匹配模型。

[0159] 本实施例中，通过医学语料数据集对BERT模型进行训练，使得BERT模型学习到丰富的语义信息，保证了疾病实体匹配的准确性。

[0160] 在本实施例的一些可选的实现方式中，上述实体匹配模块307包括：词典获取子模块、第一组合子模块、第一输入子模块以及实体确定子模块，其中：

[0161] 词典获取子模块，用于获取疾病实体匹配词典。

[0162] 第一组合子模块，用于将待匹配实体与疾病实体匹配词典中的各疾病实体进行组合，得到第一待匹配实体对。

[0163] 第一输入子模块，用于将第一待匹配实体对输入疾病实体匹配模型，得到匹配疾

病实体对。

[0164] 实体确定子模块，用于根据匹配疾病实体对，在疾病实体匹配词典中确定与待匹配实体相匹配的疾病实体，并将确定的疾病实体作为实体匹配结果。

[0165] 本实施例中，只需输入待匹配实体，疾病实体匹配模型将待匹配实体与疾病实体词典中的疾病实体一一进行匹配判断，可以快速地对待匹配实体实现实体匹配。

[0166] 在本实施例的另一些可选的实现方式中，上述实体匹配模块307基于语音语义的疾病实体匹配包括：第二组合子模块以及第二输入子模块，其中：

[0167] 第二组合子模块，用于对待匹配实体进行两两组合，得到第二待匹配实体对。

[0168] 第二输入子模块，用于将第二待匹配实体对输入疾病实体匹配模型，得到第二待匹配实体对中的匹配疾病实体对，并将得到的匹配疾病实体对作为实体匹配结果。

[0169] 本实施例中，从多个待匹配实体中筛选匹配疾病实体对时，将待匹配实体两两组合输入疾病实体匹配模型，即可快速对所有实体组合进行判断，提高了匹配效率。

[0170] 为解决上述技术问题，本申请实施例还提供计算机设备。具体请参阅图4，图4为本实施例计算机设备基本结构框图。

[0171] 所述计算机设备4包括通过系统总线相互通信连接存储器41、处理器42、网络接口43。需要指出的是，图中仅示出了具有组件41-43的计算机设备4，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。其中，本技术领域技术人员可以理解，这里的计算机设备是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路(Application Specific Integrated Circuit，ASIC)、可编程门阵列(Field－Programmable Gate Array，FPGA)、数字处理器(Digital Signal Processor，DSP)、嵌入式设备等。

[0172] 所述计算机设备可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述计算机设备可以与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互。

[0173] 所述存储器41至少包括一种类型的可读存储介质，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41可以是所述计算机设备4的内部存储单元，例如该计算机设备4的硬盘或内存。在另一些实施例中，所述存储器41也可以是所述计算机设备4的外部存储设备，例如该计算机设备4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41还可以既包括所述计算机设备4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41通常用于存储安装于所述计算机设备4的操作系统和各类应用软件，例如基于语音语义的疾病实体匹配方法的计算机可读指令等。此外，所述存储器41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。

[0174] 所述处理器42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42通常用于控制所述计算机设备4的总体操作。本实施例中，所述处理器42用于运行所述存储器41中存储的计算机可读指令或者处理数据，例如运行所述基于语音语义的疾病实体匹配方法的计算机可读指

令。

[0175] 所述网络接口43可包括无线网络接口或有线网络接口，该网络接口43通常用于在所述计算机设备4与其他电子设备之间建立通信连接。

[0176] 本实施例中提供的计算机设备可以执行上述基于语音语义的疾病实体匹配方法的步骤。此处基于语音语义的疾病实体匹配方法的步骤可以是上述各个实施例的基于语音语义的疾病实体匹配方法中的步骤。

[0177] 本实施例中，获取疾病实体匹配词典以及候选疾病实体后，对候选疾病实体进行两两组合以构建负样本，以疾病实体匹配词典作为正样本；将正样本和负样本输入初始疾病实体匹配模型以进行充分训练，初始疾病实体匹配模型可以是完成预训练的BERT模型，具有丰富的语义信息，当训练样本规模较小时也可以获得精准的匹配效果，缩短了训练所需时间，提高了疾病实体匹配模型的训练效率；训练完成后，疾病实体匹配模型即可对输入的待匹配实体进行实体匹配，提高了疾病实体匹配的效率。

[0178] 本申请还提供了另一种实施方式，即提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，所述计算机可读指令可被至少一个处理器执行，以使所述至少一个处理器执行如上述的基于语音语义的疾病实体匹配方法的步骤。

[0179] 本实施例中，获取疾病实体匹配词典以及候选疾病实体后，对候选疾病实体进行两两组合以构建负样本，以疾病实体匹配词典作为正样本；将正样本和负样本输入初始疾病实体匹配模型以进行充分训练，初始疾病实体匹配模型可以是完成预训练的BERT模型，具有丰富的语义信息，当训练样本规模较小时也可以获得精准的匹配效果，缩短了训练所需时间，提高了疾病实体匹配模型的训练效率；训练完成后，疾病实体匹配模型即可对输入的待匹配实体进行实体匹配，提高了疾病实体匹配的效率。

[0180] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的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(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，空调器，或者网络设备等)执行本申请各个实施例所述的方法。

[0181] 显然，以上所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例，附图中给出了本申请的较佳实施例，但并不限制本申请的专利范围。本申请可以以许多不同的形式来实现，相反地，提供这些实施例的目的是使对本申请的公开内容的理解更加透彻全面。尽管参照前述实施例对本申请进行了详细的说明，对于本领域的技术人员来而言，其依然可以对前述各具体实施方式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效替换。凡是利用本申请说明书及附图内容所做的等效结构，直接或间接运用在其他相关的技术领域，均同理在本申请专利保护范围之内。

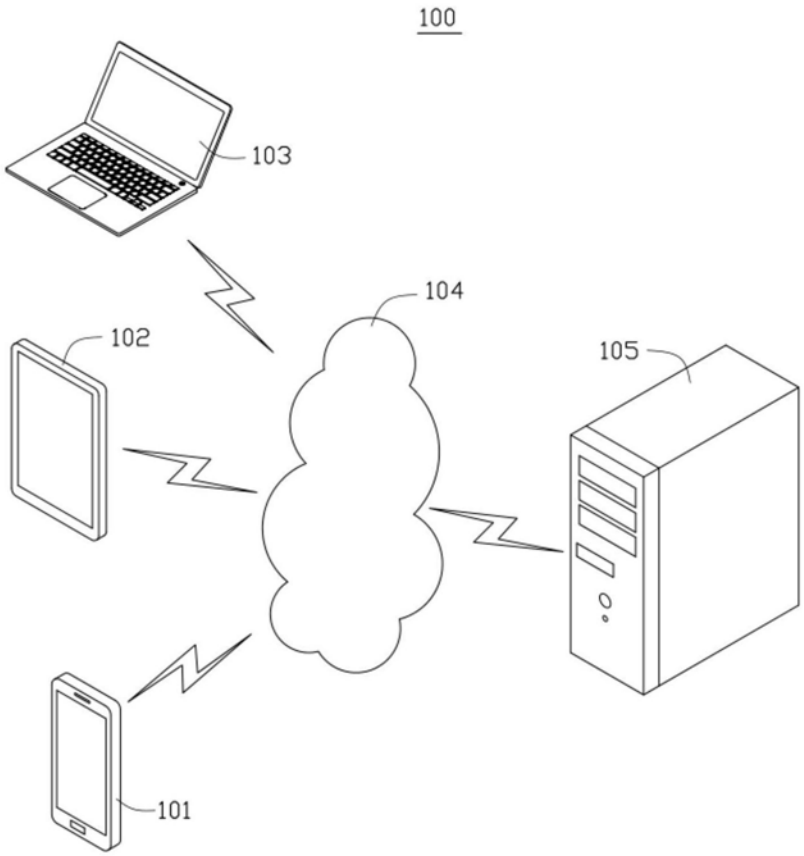



图1

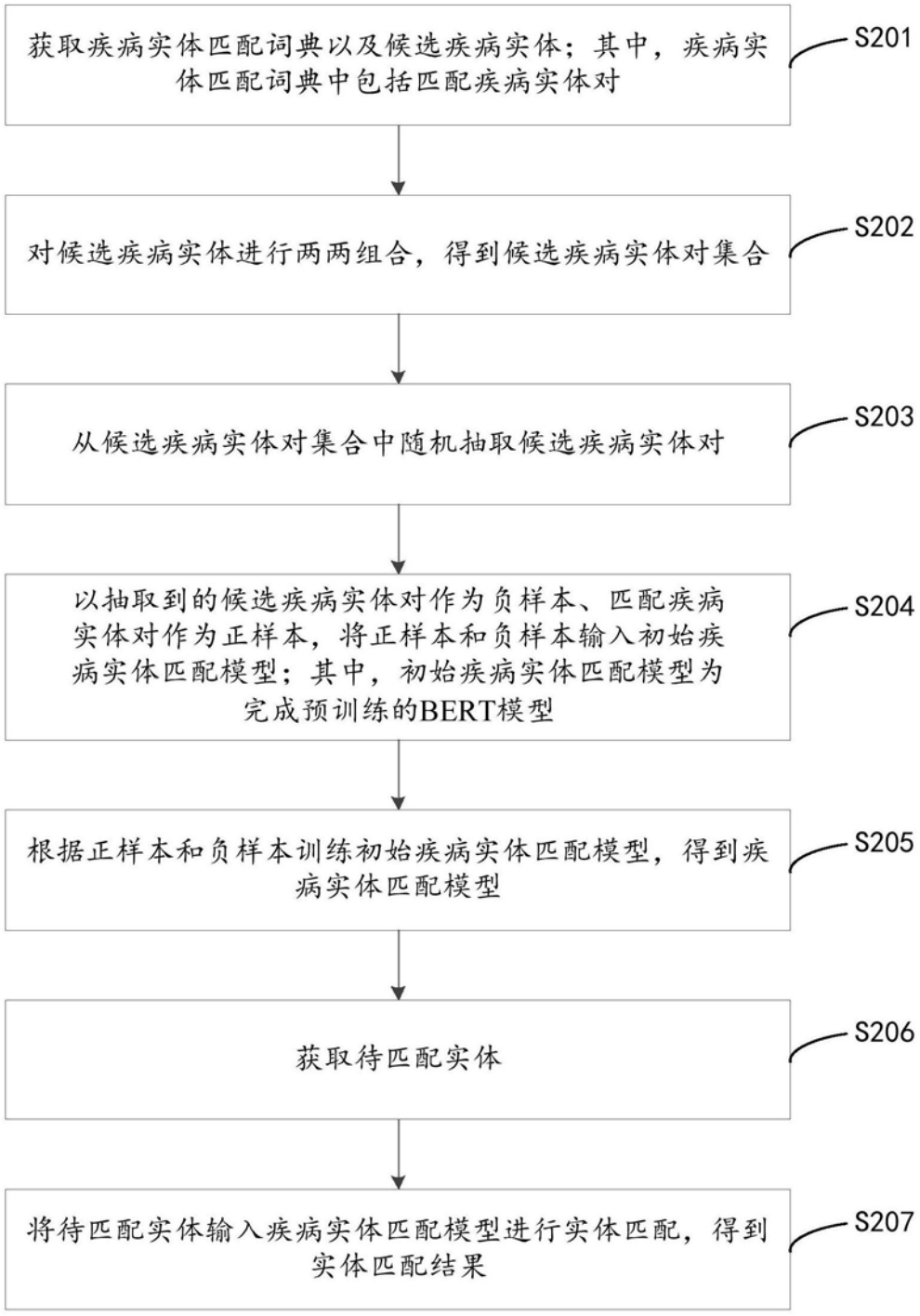

获取疾病实体匹配词典以及候选疾病实体；其中，疾病实体匹配词典中包括匹配疾病实体对
S201
对候选疾病实体进行两两组合，得到候选疾病实体对集合
S202
从候选疾病实体对集合中随机抽取候选疾病实体对
S203
以抽取到的候选疾病实体对作为负样本、匹配疾病实体对作为正样本，将正样本和负样本输入初始疾病实体匹配模型；其中，初始疾病实体匹配模型为完成预训练的BERT模型
S204
根据正样本和负样本训练初始疾病实体匹配模型，得到疾病实体匹配模型
S205
获取待匹配实体
S206
将待匹配实体输入疾病实体匹配模型进行实体匹配，得到实体匹配结果
S207

图2

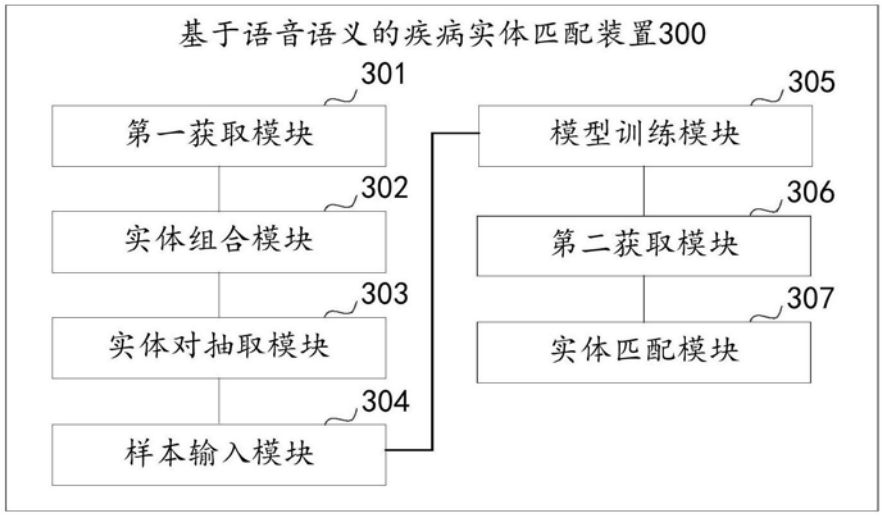

基于语音语义的疾病实体匹配装置300
301
第一获取模块
302
实体组合模块
303
实体对抽取模块
304
样本输入模块
305
模型训练模块
306
第二获取模块
307
实体匹配模块

图3

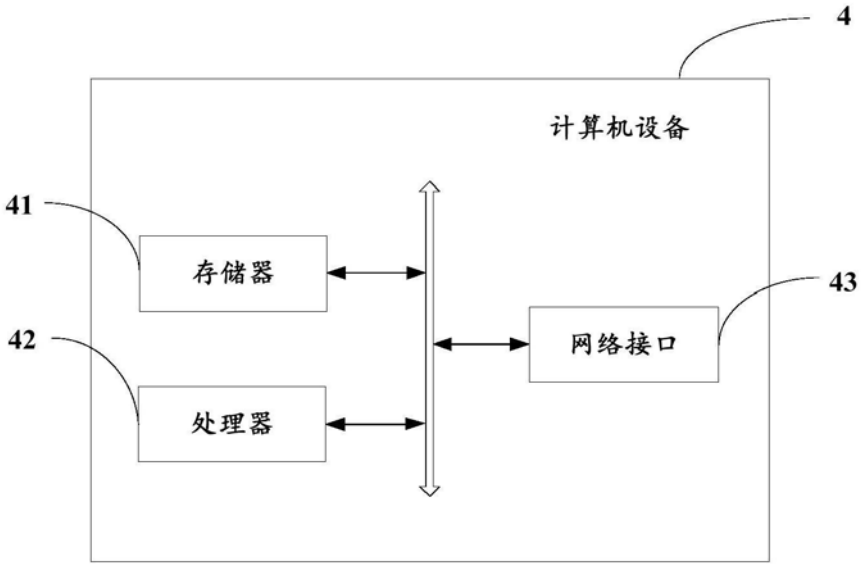

4
计算机设备
41
存储器
42
处理器
43
网络接口

图4