



(21) 申请号 202010266675.6

G06F 40/30 (2020.01)

(22) 申请日 2020.04.07

G06F 16/33 (2019.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 111798941 A

(43) 申请公布日 2020.10.20

(30) 优先权数据

16/375,403 2019.04.04 US

(73) 专利权人 IQVIA 有限公司

地址 美国新泽西州

(72) 发明人 N·杜舍夫 K·摩根

J·P·阿尔伯纳 L·格拉斯

S·萨克拉尼

(74) 专利代理机构 北京林达刘知识产权代理事

务所(普通合伙) 11277

专利代理师 刘新宇

(51) Int.Cl.

G16H 10/20 (2018.01)

G16H 10/60 (2018.01)

G16H 70/00 (2018.01)

G06F 40/295 (2020.01)

(56) 对比文件

CN 105069077 A, 2015.11.18

CN 105912625 A, 2016.08.31

US 10140421 B1, 2018.11.27

US 2019147993 A1, 2019.05.16

US 2009083218 A1, 2009.03.26

CN 106874643 A, 2017.06.20

CN 106934220 A, 2017.07.07

CN 105989080 A, 2016.10.05

伍大勇. 搜索引擎中命名实体查询处理相关技术研究.《中国博士学位论文全文数据库》.2014, (第1期), 第65-71段.

Kundet等. Clinical named entity recognition: Challenges and opportunities.《2016 IEEE International Conference on Big Data (Big Data)》.2016, 第1937-1944页.

审查员 赵伟华

权利要求书4页 说明书16页 附图5页

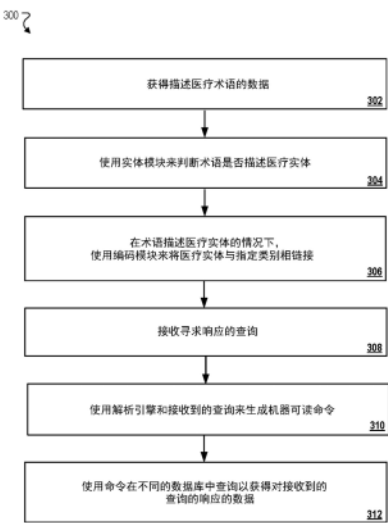
(54) 发明名称

用于生成临床查询的预测系统

(57) 摘要

本发明涉及一种用于生成临床查询的预测系统。包括编码在计算机存储介质上的计算机程序的方法、系统和装置用于生成预测系统,该预测系统获得并处理描述不同医学概念的术语的数据,以从用户查询生成命令。系统的实体模块判断术语是否描述与影响个人的健康护理状况相关联的医疗实体。在术语描述医疗实体的情况下,编码模块基于编码方案来将医疗实体与指定类别链接。系统接收用户查询。系统的解析引擎使用接收到的查询,以通过对照描述医疗实体的术语解析查询并基于用于将医疗实体链接至指定类别的编码方案来生成机器可读命令。系统使用该命令在不同的数据库中查询以获得用于生

成对所接收到的查询的响应的数据。



1. 一种计算机实现的方法,其使用系统来进行,所述方法包括:

获得包括多个术语的数据;

判断所述多个术语中的术语是否描述医疗实体;

利用所述系统的预测模型,确定所述多个术语中的各术语与所述医疗实体之间的相应置信度得分,其中所述置信度得分是基于所述术语与所述医疗实体之间的相似度来确定的;

利用用于与所述预测模型交互的编码模块,基于针对来自医疗信息的类别中的类别的编码方案以及与所述术语相关联的置信度得分,将所述医疗实体与所述类别相链接,其中所述编码模块通过采用一个或多个机器学习模型来进行所述链接,以针对包括级别层级结构的层级编码方案中的特定深度级别将所述医疗实体与所述类别相关联;

响应于接收到查询,至少通过基于所述编码方案和所述置信度得分对照与链接至所述类别的医疗实体有关的信息对所述查询进行解析,来生成机器可读命令;

在使用所述机器可读命令在一个或多个数据库中进行查询之后,生成对所述查询的回复;以及

提供所述回复作为输出以供在用户装置处显示,

其中,所述级别层级结构包括至少一个父类别,各所述至少一个父类别包括至少一个子类别,所述编码模块还被配置为:

针对所述至少一个子类别中的各子类别计算所述医疗实体的置信度得分,

响应于针对所述至少一个子类别至少之一的置信度得分低于预定义阈值,将所述医疗实体与所述至少一个子类别中的子类别相链接,以及

响应于针对所述至少一个子类别的置信度得分都高于所述预定义阈值,将所述医疗实体与所述父类别相链接。

2. 根据权利要求1所述的方法,其中,判断术语是否描述医疗实体包括:

判断所述置信度得分是否超过阈值置信度得分。

3. 根据权利要求1所述的方法,其中,将所述医疗实体与所述类别相链接包括:

获得针对所述类别的类别代码的列表;

确定所述术语和类别代码的列表中的相应类别代码之间的匹配;以及

基于用于描述所述医疗实体的术语和所述相应类别代码之间的匹配,将所述医疗实体与所述类别相链接。

4. 根据权利要求1所述的方法,其中,将所述医疗实体与所述类别相链接包括:

基于针对所述类别的编码方案来利用相应类别代码对所述医疗实体进行编码,

其中,针对所述类别的编码方案是包括所述级别层级结构的所述层级编码方案。

5. 根据权利要求4所述的方法,其中,利用相应类别代码对所述医疗实体进行编码包括:

对包括要编码的医疗实体的内容进行量化;

确定所述级别层级结构中的用于映射所述内容的级别的深度;以及

将所述内容中所包括的医疗实体与针对所述级别层级结构中的特定深度级别的相应类别代码相关联。

6. 根据权利要求3所述的方法,其中,所述医疗实体是疾病,以及确定所述匹配包括:

针对级别层级结构中的各级别生成相应匹配得分;以及
判断所述相应匹配得分是否超过阈值匹配得分。

7. 根据权利要求1所述的方法,其中,判断术语是否描述医疗实体包括:

对照实体特定数据集中的信息来进行所述术语的查找;以及

基于所述术语和所述实体特定数据集中的第一条目之间的匹配,判断为所述术语描述所述医疗实体。

8. 根据权利要求7所述的方法,其中,所述医疗实体与影响个人的健康护理状况相关联,以及其中所述医疗实体包括以下项中至少之一:

一个或多个医学疾病;

用于治疗所述一个或多个医学疾病的医疗药物;

与所述一个或多个医学疾病相关联的医疗过程;以及

用于描述与个人的健康护理状况相对应的多个医学发现的数据。

9. 根据权利要求7所述的方法,其中,所述实体特定数据集是基于包括以下项中至少之一的数据而生成的:

i) 用于描述多个疾病的预定义的一组信息;

ii) 用于描述多个药物的预定义的一组信息;

iii) 用于描述多个医疗过程的预定义的一组信息;以及

iv) 多个健康护理患者的电子医疗数据。

10. 根据权利要求1所述的方法,其中,获得包括多个术语的数据包括:

获得多个非结构化数据;以及

构造所述非结构化数据,使得能够对照所述系统的多个数据库中的信息来处理所述查询。

11. 一种用于查询的系统,包括:

一个或多个处理装置;以及

一个或多个非暂时性机器可读存储装置,其存储指令,所述指令能够由所述一个或多个处理装置执行以使得进行操作,所述操作包括:

获得包括多个术语的数据;

判断所述多个术语中的术语是否描述医疗实体;

利用所述系统的预测模型,确定所述多个术语中的各术语与所述医疗实体之间的相应置信度得分,其中所述置信度得分是基于所述术语与所述医疗实体之间的相似度来确定的;

利用用于与所述预测模型交互的编码模块,基于针对来自医疗信息的类别中的类别的编码方案以及与所述术语相关联的置信度得分,将所述医疗实体与所述类别相链接,其中所述编码模块通过采用一个或多个机器学习模型来进行所述链接,以针对包括级别层级结构的层级编码方案中的特定深度级别将所述医疗实体与所述类别相关联;

响应于接收到查询,至少通过基于所述编码方案和所述置信度得分对照与链接至所述类别的医疗实体有关的信息对所述查询进行解析,来生成机器可读命令;

在使用所述机器可读命令在一个或多个数据库中进行查询之后,生成对所述查询的回复;以及

提供所述回复作为输出以供在用户装置处显示，

其中，所述级别层级结构包括至少一个父类别，各所述至少一个父类别包括至少一个子类别，并且将所述医疗实体与所述类别相链接包括：

针对所述至少一个子类别中的各子类别计算所述医疗实体的置信度得分，

响应于针对所述至少一个子类别至少之一的置信度得分低于预定义阈值，将所述医疗实体与所述至少一个子类别中的子类别相链接，以及

响应于针对所述至少一个子类别的置信度得分都高于所述预定义阈值，将所述医疗实体与所述父类别相链接。

12. 根据权利要求11所述的系统，其中，判断术语是否描述医疗实体包括：

判断所述置信度得分是否超过阈值置信度得分。

13. 根据权利要求11所述的系统，其中，将所述医疗实体与所述类别相链接包括：

获得针对所述类别的类别代码的列表；

确定所述术语和类别代码的列表中的相应类别代码之间的匹配；以及

基于用于描述所述医疗实体的术语和所述相应类别代码之间的匹配，将所述医疗实体与所述类别相链接。

14. 根据权利要求11所述的系统，其中，将所述医疗实体与所述类别相链接包括：

基于针对所述类别的编码方案来利用相应类别代码对所述医疗实体进行编码，

其中，针对所述类别的编码方案是包括所述级别层级结构的所述层级编码方案。

15. 根据权利要求14所述的系统，其中，利用相应类别代码对所述医疗实体进行编码包括：

对包括要编码的医疗实体的内容进行量化；

确定所述级别层级结构中的用于映射所述内容的级别的深度；以及

将所述内容中所包括的医疗实体与针对所述级别层级结构中的特定深度级别的相应类别代码相关联。

16. 根据权利要求13所述的系统，其中，所述医疗实体是疾病，以及确定所述匹配包括：

针对级别层级结构中的各级别生成相应匹配得分；以及

判断所述相应匹配得分是否超过阈值匹配得分。

17. 根据权利要求11所述的系统，其中，判断术语是否描述医疗实体包括：

对照实体特定数据集中的信息来进行所述术语的查找；以及

基于所述术语和所述实体特定数据集中的第一条目之间的匹配，判断为所述术语描述所述医疗实体。

18. 根据权利要求17所述的系统，其中，所述医疗实体与影响个人的健康护理状况相关联，以及其中所述医疗实体包括以下项中至少之一：

一个或多个医学疾病；

用于治疗所述一个或多个医学疾病的医疗药物；

与所述一个或多个医学疾病相关联的医疗过程；以及

用于描述与个人的健康护理状况相对应的多个医学发现的数据。

19. 根据权利要求11所述的系统，其中，获得包括多个术语的数据包括：

获得多个非结构化数据；以及

构造所述非结构化数据,使得能够对照所述系统的多个数据库中的信息来处理所述查询。

20. 一个或多个非暂时性机器可读存储装置,其存储指令,所述指令能够由一个或多个处理装置执行以使得进行操作,所述操作包括:

获得包括多个术语的数据;

判断所述多个术语中的术语是否描述医疗实体;

利用系统的预测模型,确定所述多个术语中的各术语与所述医疗实体之间的相应置信度得分,其中所述置信度得分是基于所述术语与所述医疗实体之间的相似度来确定的;

利用用于与所述预测模型交互的编码模块,基于针对来自医疗信息的类别中的类别的编码方案以及与所述术语相关联的置信度得分,将所述医疗实体与所述类别相链接,其中所述编码模块通过采用一个或多个机器学习模型来进行所述链接,以针对包括级别层级结构的层级编码方案中的特定深度级别将所述医疗实体与所述类别相关联;

响应于接收到查询,至少通过基于所述编码方案和所述置信度得分对照与链接至所述类别的医疗实体有关的信息对所述查询进行解析,来生成机器可读命令;

在使用所述机器可读命令在一个或多个数据库中进行查询之后,生成对所述查询的回复;以及

提供所述回复作为输出以供在用户装置处显示,

其中,所述级别层级结构包括至少一个父类别,各所述至少一个父类别包括至少一个子类别,并且将所述医疗实体与所述类别相链接包括:

针对所述至少一个子类别中的各子类别计算所述医疗实体的置信度得分,

响应于针对所述至少一个子类别至少之一的置信度得分低于预定义阈值,将所述医疗实体与所述至少一个子类别中的子类别相链接,以及

响应于针对所述至少一个子类别的置信度得分都高于所述预定义阈值,将所述医疗实体与所述父类别相链接。

用于生成临床查询的预测系统

技术领域

[0001] 本说明书涉及预测计算系统。

背景技术

[0002] 作为健康护理过程的一部分,医师或其他医疗护理提供方可以进行临床试验、项目和其它活动,以评价药物或其它医疗治疗选项的受试者安全性和有效性。使用健康相关试验项目可以有助于标识用于改善整体患者健康并降低健康系统成本的新颖治疗选项。临床试验或项目可以是前瞻性地将人类参与者/受试者或人类受试者群体分配至一个或多个健康相关干预以评价对健康结果的影响的一项调查研究或多项调查研究。

发明内容

[0003] 作为健康护理过程的一部分,医师或其他健康护理提供方可以进行试验、项目和其它活动,以评价特定药物或其它医疗治疗选项的有效性。进行健康相关临床试验可以有助于标识用于改善整体患者健康并降低健康系统成本的新颖治疗选项。临床试验和其它受控项目通常由不同地理位置中的医疗设施处的一名或多名调研方进行,这些调研方与研究对象交互以评价药物治疗选项的有效性。在一些情况下,患者的医师可以与临床试验相关联,并且医师可以基于患者的诊断状况将患者转介为参与试验的候选。调研方、地理位置或这两者可以形成用于执行项目的实体。

[0004] 基于以上的上下文,本文描述了一种计算系统,该计算系统使用特定的计算规则或指令(例如,唯一算法)以基于接收到的用户输入来预测或生成命令。为了生成命令,系统被配置为使用一个或多个学习算法(例如,深度学习算法)来训练预测模型。预测模型用于处理使用系统的实体模块中的自然语言处理器(NLP)进行识别和提取的术语。预测模型可被训练为从语义上理解相关术语(例如,医学术语和临床术语)及其与其它医学术语的关系。术语可以从诸如教科书和在线资源等的信息源中提取,或者从诸如多个健康护理患者的电子医疗数据等的非结构化数据集中提取。

[0005] 编码模块使用一个或多个神经网络模型来对所提取的术语进行编码并将其与诸如疾病实体、药物实体、医疗过程(medical procedure)实体或各种其它类型的实体等的特定医疗实体相链接。系统利用预测模型关于编码后的医学术语的学习推断,以基于从用户接收到的查询来生成命令。例如,解析引擎可以将所提取的术语自动转化为机器可读命令,该机器可读命令对照医疗数据库进行处理以获得对用户查询的准确响应。因此,预测系统的至少一个目标是以计算高效的方式准确地解释包括与患者或一组患者有关的健康相关信息的用户查询。

[0006] 例如,用户查询代表预测系统的用户输入,该用户输入指定患者属性的列表。查询/用户输入可以是人类可读格式。所描述的技术使得系统能够以计算机可读格式快速且高效地生成相应的命令。然后,使用计算机可读命令来查询不同的电子健康记录(EHR),以基于列表中指定的属性标识满足给定条件的患者(例如,试验受试者)。例如,该命令可以用

于查询历史医疗记录,以在没有(例如,来自人类操作者的)人工干预的情况下得到洞察和信息。所得到的洞察可以包括准确的对临床试验参与的合格患者人群的估计和对不良事件的倾向的估计。

[0007] 本说明书中所描述的主题的一个方面可以体现在一种计算机实现的方法中,该方法包括:获得包括多个术语的第一组数据;判断所述多个术语中的术语是否描述医疗实体;响应于判断为所述术语描述所述医疗实体,基于针对类别的编码方案来将所述医疗实体与所述类别相链接;响应于接收到查询,通过对照所述第一组数据中的用于描述所述医疗实体的术语对所述查询进行解析并且基于所述编码方案,来生成机器可读命令;使用所述机器可读命令在多个数据库中进行查询;在使用所述机器可读命令在所述多个数据库中进行查询的情况下,响应于接收到的查询来获得第二组数据;以及提供所述第二组数据作为输出以供在用户装置处显示。

[0008] 这些和其它实现可以各自可选地包括以下特征中的一个或多个。例如,在一些实现中,判断术语是否描述医疗实体包括:基于所述第一组数据中所描述的术语和所述医疗实体之间的相似度的推断来生成置信度得分;以及判断所述置信度得分是否超过阈值置信度得分。

[0009] 在一些实现中,将所述医疗实体与类别相链接包括:获得针对该类别的类别代码的列表;确定术语和类别代码的列表中的相应类别代码之间的匹配;以及基于用于描述所述医疗实体的术语和所述相应类别代码之间的匹配,将所述医疗实体与指定类别相链接。

[0010] 在一些实现中,将所述医疗实体与类别相链接包括:基于针对指定类别的编码方案利用相应类别代码对医疗实体进行编码;以及针对指定类别的编码方案是包括级别层级结构的层级编码方案。

[0011] 在一些实现中,利用相应类别代码对医疗实体进行编码包括:对包括要编码的医疗实体的内容进行量化;确定级别层级结构中的用于映射所述内容的级别的深度;以及将所述内容中所包括的医疗实体与针对级别层级结构中的特定深度级别的相应类别代码相关联。

[0012] 在一些实现中,所述医疗实体是疾病,并且确定所述匹配包括:针对级别层级结构中的各级别生成相应匹配得分;以及判断相应匹配得分是否超过阈值匹配得分。

[0013] 在一些实现中,判断术语是否描述医疗实体包括:对照实体特定数据集中的信息对术语进行查找;以及基于该术语和实体特定数据集中的第一条目之间的匹配,判断为该术语描述所述医疗实体。

[0014] 在一些实现中,所述医疗实体与影响个人的健康护理状况相关联,以及所述医疗实体包括以下项至少之一:一个或多个医学疾病;用于治疗所述一个或多个医学疾病的医疗药物;与所述一个或多个医学疾病相关联的医疗过程;以及用于描述与个人的健康护理状况相对应的多个医学发现的数据。

[0015] 在一些实现中,实体特定数据集是基于包括以下各项中至少之一的数据而生成的:i) 用于描述多个疾病的一组预定义信息;(ii) 用于描述多个药物的一组预定义信息;(iii) 用于描述多个医疗过程的一组预定义信息;以及iv) 多个医学患者的电子医疗数据。

[0016] 在一些实现中,获得用于描述与多个医学概念相关的术语的数据包括:获得多个非结构化数据;以及构造非结构化数据,使得能够对照多个数据库中的信息来处理所述查

询。

[0017] 该方面和其它方面的其它实现包括相应的系统、装置和计算机程序,其被配置为进行在计算机存储装置上编码的方法的动作。一个或多个计算机或硬件电路的计算系统可以通过安装在系统上的软件、固件、硬件或它们的组合来配置,这些软件、固件、硬件或它们的组合在操作时使系统执行动作。一个或多个计算机程序可以通过具有如下指令来配置,这些指令在由数据处理装置执行时使装置执行动作。

[0018] 可以实施本说明书中所描述的主题以实现以下优点中的一个或多个。所描述的技术提供了可扩展计算系统,该可扩展计算系统是用于分析和解析结构化和非结构化数据集的全自动端到端预测解决方案。使用分析和解析功能,系统的预测模型被配置为使得可以使用基于预测模型所学到的数据推断而生成的机器可读命令来查询数据集中的信息。

[0019] 预测系统被配置为快速且高效地分析用于描述各种疾病和适应症、药物/治疗选项、以及医疗过程的多个数据集。例如,通过使用如下的三个步骤生成命令来证明系统的效率:(1)提取用于描述患者属性的实体;(2)将状况映射到标准化科学实体名称;(3)以及解释不同健康护理状况之间的关系,包括状况是否被否定。

[0020] 因此,系统提供了专门用于对医学语言术语进行识别和编码的解决方案,并且需要很少(或不需要)人工数据管护来实现命令或机器可读查询的期望准确度级别,其中这些命令或机器可读查询被生成且处理以获得对用户输入的响应。预测系统使用学习算法(例如,深度学习算法)来确定相关信息类别之间的关系,并使用信息类别之间的关系来直接查询医疗和调查数据库。

[0021] 本说明书中所描述的主题的一个或多个实现的细节在以下的附图部分和具体实施方式部分中阐述。根据具体实施方式部分、附图部分和权利要求书部分可以明显看出本主题的其它潜在特征、方面和优点。

附图说明

[0022] 图1示出用于生成机器可读命令的示例性计算系统的框图。

[0023] 图2示出图1的计算系统的示例性数据处理模块的框图。

[0024] 图3是用于使用图1的计算系统的预测模型来生成机器可读命令的示例性处理。

[0025] 图4示出用于处理非结构化医疗数据以生成机器可理解查询的示例性数据流。

[0026] 图5示出可以结合本文献中所描述的方法使用的计算系统的框图。

[0027] 各个附图中的相同附图标记和标示表示相同元件。

具体实施方式

[0028] 当前系统以非结构化数据格式存储医疗信息,并且信息可以分散在多个系统中,这使得信息难以被查询并链接到其它数据源。这种医疗信息的许多源是基于文本的源,并且信息可以由不同元素形成。然而,对于这些不同元素,不存在广泛遵循的国际或现有分类。

[0029] 在从这些医疗源查询信息时,通常需要领域专业知识来识别相关术语、对术语进行分类、并将这些术语转换为可用于查询数据的合适格式。对这种医疗信息进行建模和结构化的能力使得用户能够得到某些洞察,这些洞察可以改善一组患者的健康状况。例如,至

少一种用途可以涉及查询患者数据库以准确地计算具有满足参与新临床试验的协议和标准的健康属性的患者池。

[0030] 在该上下文中,描述了用于生成如下预测模型的技术,该预测模型提取并从语义上理解相关临床术语以及术语之间的关系。使用计算机系统,以例如使用自然语言处理(NLP)和深度学习算法来从结构化和非结构化的数据集中提取临床术语。这些技术使用数据处理模块来对医疗实体(例如,适应症、药物、过程(procedure)等)进行识别、提取和分类,并参考用于描述医疗实体的术语来确定这些实体之间的关系。

[0031] 使用从这些处理获得的信息来生成机器可读命令,以查询不同数据库,从而得到与一组患者有关的各种洞察和信息,而无需人工的用户干预。例如,该命令可以用于查询历史医疗记录的数据库,以获得对参与临床试验的合格患者人群的估计或者获得对与临床试验相关的不良事件的发生倾向的估计。

[0032] 图1示出用于检测未诊断状况的示例性计算系统100的框图。系统100可以是被配置为处理输入数据以训练预测模型的预测计算系统。如本文所述,训练预测模型以进行与处理并从语义上理解相关临床和医疗术语有关的各种功能。

[0033] 系统100包括实体模块120、编码模块140、预测模型160和解析引擎170。在示例性训练阶段155期间,使用实体模块120和编码模块140来处理输入数据110,以计算用于生成经训练的预测模型160的学习推断。因此,训练阶段155与示例性预测模型相关联,并在系统100处进行以训练和生成预测模型160。

[0034] 如本文所述,经训练的预测模型160被配置为从各种输入文本中提取并在语义上理解相关临床术语及其关系。预测模型用于处理使用系统的实体模块中的自然语言处理器(NLP)进行识别和提取的术语。系统100利用预测模型关于编码后的医疗术语的学习推断,以基于从用户接收到的查询来生成命令。

[0035] 一般来说,系统100获得或接收输入数据110以在系统处进行处理。输入数据110可以包括结构化数据和非结构化数据。结构化数据可以包括各种类型的出版物和健康相关文本,诸如医学教科书、与健康护理相关的在线出版物、医学期刊、电子出版物、医学论文、基于web的文章、医学网站、或固有地格式化以供计算机系统数据进行数据提取和处理的信息的各种源等。非结构化数据可以包括与医疗活动、患者医疗记录、或健康护理事务相关的不同数据集,并且在下文中参考至少图2进行详细描述。

[0036] 实体模块120被配置为接收或获得输入数据110、并处理该数据以训练系统100的一个或多个预测模型。在一些实现中,实体模块120是执行高级通用编程语言以从输入数据110中识别和提取数据元素的命名实体识别(NER)模块。例如,实体模块120可以是使用编码指令来标识用于描述文献中的不同医学概念的医疗实体的术语和字(数据元素)的python模块。文献可以是电子文献(诸如医学教科书的数字版本),并且术语或字可以描述诸如疾病名称、药物治疗、药物化合物或医疗过程等的医疗实体。

[0037] 如下文参考图2更详细地描述,实体模块120包括各种数据搜索/查找和机器学习功能。这些功能可以用于对输入数据110中所包括的医疗文档或结构化数据源中的每个字或术语进行字典查找。在一些实现中,查找功能和机器学习功能是实体模块120的独立计算功能。如下文更详细地描述,各计算功能可以用于判断输入数据110中的术语或数据元素是否描述特定的医疗实体。

[0038] 例如,对照经管护的数据源(诸如实体特定医学术语的字典等)来搜索输入数据110的医疗文档中的每个字或术语。在一些实现中,进行字典查找包括从医疗文档中提取一个或多个n元,该一个或多个n元被识别为与经管护的数据源中的一个或多个术语相匹配。所提取的n元可以包括连续的数据项或元素(例如,字母或字),这些数据项或元素与经管护的数据源中的术语相匹配。例如,医疗文档可以具有形成用于描述诸如癌症等的特定医疗状况的字的字母。所提取的n元可以包括连续的字母或字,这些字母或字与经管护的数据源中的也用于描述与癌症相关的不同医学概念的术语相匹配。

[0039] 编码模块140被配置为接收或获得与医疗实体相关联的数据,其中数据包括用于描述医疗实体的相应术语。编码模块140对与医疗实体相关联的数据进行编码,并且编码数据用于训练系统100的一个或多个预测模型。在一些实现中,编码模块140是多用途编码模块,其执行高级通用编程语言以对与医疗实体相关联的数据进行编码。在一些情况下,与实体模块120非常相似,编码模块140也可以是使用编码指令来进行各种功能的python模块。

[0040] 编码模块140可以使用指令,以响应于实体模块120判断为术语描述医疗实体而将医疗实体与指定类别代码145相链接、关联或以其它方式对医疗实体进行编码。例如,编码模块140将医疗实体编码为指定实体组织,其中i) 疾病可以基于ICD-10代码进行编码,ii) 药物可以基于GPI代码进行编码,以及iii) 过程可以基于PRC_CD代码进行编码。ICD-10代码、GPI代码和PRC_CD代码各自与针对健康相关疾病、药物和医疗过程的相应编码方案相关联。

[0041] 如下文更详细地描述,医疗实体基于针对指定类别的编码方案与指定类别代码145相链接或利用指定类别代码145进行编码。在一些实现中,每个标识或提取的医疗实体都被编码为特定的命名(例如,官方疾病名称和相应类别代码)。在一些情况下,一个或多个医疗实体被编码为多于一个命名。例如,药物可被至少编码为特定GPI代码、特定ATC代码或这两者。

[0042] 编码模块140可以包括分立的计算元件,这些计算元件各自用于对疾病、药物或过程进行相应的编码操作。在一些实现中,至少参考健康相关疾病的医疗实体编码,针对类别代码145的编码方案是基于级别(或深度级别)层级结构的层级编码方案。以这种方式,利用相应类别代码145对健康相关疾病的医疗实体进行编码包括确定级别层级结构中的级别的深度、并确定用于映射包括与医疗实体相关联的数据的特定信息内容的适当级别。

[0043] 可以使用一个或多个神经网络150来确定级别层级结构中的级别的深度并确定用于映射与医疗实体相关联的数据的相应级别。编码模块140可以使用神经网络150的输出(例如,计算推断或置信度得分)来将信息内容中所包括的医疗实体与级别层级结构中的特定深度级别的相应类别代码145相关联。例如,可以使用各自代表各个神经网络150的推断输出的一个或多个得分来将医疗实体与级别层级结构中的特定深度级别以及该特定深度级别处的相应一个或多个类别代码145相关联。

[0044] 解析引擎170是被配置为将用户输入125转换为机器可读命令的语义查询解析器。例如,用户输入125可以代表人类可理解查询,并且命令生成器175被配置为基于对用户输入125进行的一个或多个查询解析操作来生成机器可读命令。解析引擎170使用命令生成器175,以响应于基于一个或多个查询解析操作的结果将用户输入125转换为机器命令而生成机器可读命令。

[0045] 所生成的命令可以代表机器可理解查询,这些机器可理解查询被配置用于对不同的医疗或信息数据库190进行处理。解析引擎170进行查询解析操作,以将所识别出的实体关系转化为命令。除了识别实体之外,解析引擎170还可以进行查询解析操作,以基于用户输入125中的特定关键字、用户输入125的文本中的字的特定位置、或形成用户输入125的术语的语义属性来确定术语之间的关系的类型。

[0046] 命令与计算机可理解且可执行的逻辑操作相关联。命令和逻辑操作可以使计算机查询数据库190的特定数据集,这可以包括对关系数据库的表执行过滤或合并以获得用于生成响应195的数据元素。在一些实现中,所生成的命令可以具有机器可读格式,诸如基于结构化查询语言(SQL)的查询格式等。在一些实现中,命令可被格式化为SQL查询,并且用于获得或管理示例性关系数据库190中所存储的数据,或者用于关系数据流管理系统中的流数据处理。

[0047] 系统100的示例性用例包括接收实体模块120,该实体模块120接收用于指定针对新临床试验的协议的数据。实体模块120用于基于标识和提取描述实体的一个或多个术语来识别协议中所包括的所有相关医疗实体。编码模块140使用针对各实体提取的一个或多个术语来对每个识别出的医疗实体进行分类。例如,解析引擎模块170可以将实体分类为用于获得可参与新临床试验的受试者的纳入或排除标准。解析引擎170的命令生成器175写入或生成集成了实体的命令(例如,查询)。机器可读命令也可以解释每个医疗实体及其相关术语是如何相关的。

[0048] 对照数据库(例如,医疗事务数据库)中的给定数据集来运行或处理命令。系统100响应于对照数据库处理命令而生成结果。该结果可以是具有满足新临床试验的纳入和排除标准的医疗属性的患者或潜在试验受试者的数量。如本文献中所使用的,受试者或试验受试者可以是参与临床试验的候选者、临床试验的参与者、或可能被标识为参与临床试验的(健康管理提供方的)现有患者。

[0049] 系统100可被配置为提高或增强至少使用预测模型160生成的输出判断的准确性。例如,系统100包括反馈回路180,该反馈回路180使得特定输出判断能够作为输入反馈至系统100。反馈回路180可以确保在至少由实体模块120和编码模块140处理的一组输入数据110中完全捕获离散参数。通过反馈输出的数据参数(诸如所生成的命令或计算出的干扰判断等),反馈回路180可以用于提高或增强预测模型160的准确性。

[0050] 例如,使用反馈回路180,系统100可以通过检测输出中的更细粒度的一组参数之间可能存在的新关系和共性来迭代地增强其预测能力。在一些情况下,系统100响应于再评价如下输出数据而获得准确性的这种迭代增强,其中该输出数据可以包括与用于描述不同医疗实体的术语或数据元素之间的关系有关的特定判断。系统100可以将反馈回路180连同实体模块120和编码模块140一起使用,从而以相对于传统系统而言降低的计算成本和更好的准确性来联合并迭代地处理各种类型的输入数据110(包括反馈数据)。

[0051] 在一些实现中,系统100的示例性训练阶段155可以是基于反馈回路180,其中嵌入式向量(输出)作为系统的输入反馈至系统100,然后分析系统的输入以迭代地增强预测模型160所生成的输出和判断的准确性。在其它实现中,除了反馈回路180之外,系统100还被配置为扩展可使用由模块执行的计算规则(例如,算法)以及该系统的计算元件所识别的不同医疗实体的当前列表。

[0052] 图2示出与上文参考图1所描述的系统100的模块相对应的示例性数据处理模块200的框图。如图2的实现所示,系统100的数据处理模块200可以包括附加计算元件。这些附加计算元件可以代表上述的相应模块120、140和170的子系统。例如,附加元件可以进行计算功能,以在系统100的示例性训练阶段期间处理输入数据,从而生成经训练的预测模型160。

[0053] 如上文所示,在训练阶段155期间处理的输入数据110包括结构化数据和非结构化数据。例如,非结构化数据202可以包括用于描述一大组患者或受试者的健康和医疗属性的信息。非结构化数据202还可以包括用于描述一组临床试验的赞助和执行细节的信息。执行细节可以指定每项试验的诸如医疗设施、地理位置和调研方等的信息、以及每项试验在一组临床试验中的纳入和排除标准。

[0054] 在一些情况下,非结构化数据202包括健康护理事务信息的多个相应数据集以及个人患者或预期受试者的人口统计信息的多个相应数据集。健康护理事务信息可以从用于描述医师和患者之间的互动的数据、由受试者的物理或电子医疗记录(EMR)得到的数据、由处方记录/医疗理赔得到的数据(“Rx/Dx数据”)、与患者或受试者所使用的处方或治疗选项相关的数据、或与健康护理事务和健康护理活动相关的其它数据源获得。人口统计信息可以包括患者标识号、受试者的年龄、受试者的性别和/或首选代名词、受试者的地理区域或地址位置、以及与个人有关的其它标识数据。

[0055] 实体模块120使用编码指令(例如,基于python编程语言的指令)来标识文献中的医学概念并进行命名实体识别。如上文所述,实体模块120包括各种数据搜索/查找和机器学习功能。特别地,实体模块120包括数据查找引擎210和机器学习引擎220。在一些实现中,查找引擎210和机器学习引擎220是实体模块120的独立计算元件。引擎212、220各自被配置为识别输入数据110中的特定数据元素,并提取这些数据元素或将其标记为描述特定医疗实体或与特定医疗实体相关联。

[0056] 查找引擎210被配置为对输入数据110中的每个术语或字进行字典查找。查找引擎210对照经管护的一个或多个实体特定字典来进行查找任务。实体特定字典可以根据多个数据源进行管护。数据源可以在系统100的内部(内部源)或系统100的外部(外部源),或者两者兼有。内部源包括从电子医疗记录(EMR)、疾病字典、药物字典、过程字典或每种的组合获得的数据。外部源可以包括政府机构和卫生组织(包括国内和国际组织,诸如世界卫生组织等)所管理的网站或信息数据库。

[0057] 从EMR获得的数据包括用于描述多个患者的健康护理和医疗状况的信息,并且可以是结构化数据或非结构化数据。疾病字典可与ICD-10疾病数据库相对应,其中ICD-10是对定义国际疾病和相关健康问题统计分类(ICD)的编码和分类方案的文献的修订版。过程字典可以与用于将过程名称映射到相应过程代码的过程代码映射数据库相对应。

[0058] 药物字典可以与用于将药物名称映射到相应GPI代码的GPI代码映射数据库相对应。通用产品标识符(GPI)是至少根据药物的主要治疗用途来标识药物的14字符层级分类系统。药物字典可以与用于将药物映射到解剖治疗化学(ATC)分类系统的相应ATC代码的ATC代码映射数据库相对应。ATC代码用于例如根据药物作用的器官或系统及其治疗、药理和化学性质对药物的活性成分进行分类。

[0059] 机器学习(ML)引擎220被配置为根据输入数据110来分析文本的给定部分、并标识

或识别文本中的每个字或术语是否与医疗实体相对应。ML引擎可以包括一个或多个经训练的神经网络。ML引擎220使用从基于特定算法训练的神经网络得到的特定计算规则(例如, ML算法)来进行分析和术语识别功能。神经网络的架构可以是双向LSTM-CRF(长短期记忆条件随机场)。

[0060] 神经网络响应于使用诸如“金质”和“银质”数据集等的一个或多个数据集训练神经网络而得到或学习特定计算规则。在一些实现中,金质数据集是从上文描述的一个或多个外部数据源获得的。银质数据集是使用上文还描述的经管护的实体特定字典而生成的。在系统100的实现阶段期间,可以使用迁移学习进一步提高与ML引擎220相关联的示例性数据模型的性能。例如,可以关于不同任务对深度学习模型进行预训练,其中大量的训练数据可用或者以无需人工生成的训练数据的无监督方式进行训练。预训练的模型可以基于迁移学习针对特定任务进行微调(即,改写),以改进模型的性能并减少训练数据量。以这种方式,针对第一任务进行训练的模型可以更高效地再用于基于处理与第一任务相关的数据时学到的推断来学习第二相关任务。

[0061] 现在参考编码模块140,如上所述,医疗实体基于针对指定类别的编码方案来与指定类别相链接或利用指定类别进行编码。编码模块140可以包括分立的计算元件,这些计算元件各自用于对疾病、药物或过程进行相应的编码操作。例如,如图2所示,编码模块140包括疾病编码器230、药物编码器235和过程编码器240。

[0062] 疾病编码器230是用以进行用于对与健康相关疾病相关联的医疗实体进行编码的特定编码操作的分立计算元件。疾病编码器230被配置为将给定疾病名称(例如,医疗实体)映射到ICD-10代码中。ICD-10代码提供了疾病名称和状况的标准化命名。用于ICD-10疾病编码的类别代码145是基于层级组织的编码方案。例如,在级别层级结构中,最高级别是级别1,第二级别是级别2,第三级别是级别3,第四级别是级别4,并且第五级别是级别5。疾病编码器230被配置为将针对疾病名称的医疗实体映射或编码直至级别5。

[0063] 疾病编码器230包括各种数据搜索/查找和机器学习功能。特别地,实体模块120包括数据查找引擎232和ML引擎234。在一些实现中,查找引擎232和ML引擎234是疾病编码器230的独立计算元件。查找引擎232对与经管护的实体特定疾病字典中的疾病名称相关的医疗实体进行字典查找。查找引擎232被配置为针对与特定疾病名称相匹配的给定医疗实体检索一个或多个相应ICD-10代码。在一些情况下,实体特定疾病字典根据从一个或多个电子医疗记录(EMR)得到的数据进行管护。

[0064] ML引擎234包括一个或多个神经网络150。如上所示,ML引擎234可以使用神经网络150来确定用于对与医疗实体相关联的数据进行映射或编码的相应级别。ML引擎234参考级别层级结构中的级别(级别1~级别5)的深度来确定相应级别。在一些实现中,ML引擎234包括用以对ICD-10命名的各级别(例如,级别1~5)进行编码操作的一个神经网络模型150。神经网络模型的计算推断输出(例如,相似度得分)可以用于将医疗实体与级别层级结构中的特定深度级别以及该深度级别处的相应类别(或子类别)代码145相关联。例如,ML引擎234可以针对给定用户查询生成疾病类别树中的各节点(或级别)的得分。类别树可以包括父级别以及子级别相对应的至少一个子类别。

[0065] 对于给定用户查询,ML引擎234使用得分来决定是否应包括特定父类别。例如,对于给定的具有子类别“A1”、“A2”和“A3”的疾病类别“A”以及用户查询“Q”,ML引擎234首先针

对“Q”生成三个相应的相似度得分：得分S1、得分S2和得分S3。这些得分各自可以与“A”子类别相对应。系统100判断任意得分是否低于预定义阈值。如果至少一个得分低于预定义阈值，则系统预测类别“A”的子类别(A1、A2等)作为输出，而如果所有得分S1、S2等高于预定义阈值，则系统预测类别“A”作为输出。算法可被训练成不仅针对查询疾病类别预测高相似度得分，而且还针对所有查询疾病子类别级别预测高相似度得分。因此，使用这些方法，系统100可以标识级别以及该级别的正确代码。

[0066] 在其它实现中，ML引擎234包括五个神经网络模型，其中这五个神经网络模型与神经网络150相对应。在该实现中，ML引擎234使用相应的神经网络模型来对ICD-10命名的特定级别(例如，级别1~5)进行编码操作。可以使用各个神经网络模型的计算推断输出来将医疗实体与级别层级结构中的特定深度级别以及针对该深度级别的相应类别代码145相关联。

[0067] 例如，各神经网络模型可以计算置信度得分(例如，输出)，该置信度得分用于确定官方疾病名称的最佳匹配以及ICD-10命名中针对给定疾病名称的特定级别的最佳匹配。作为示例，对于医疗实体“肺癌”，来自特定神经网络模型的置信度得分可以表示：最匹配的级别1ICD-10官方疾病名称是“肿瘤”，或者最匹配的级别2ICD-10官方疾病名称是“呼吸和胸内器官的恶性肿瘤”。例如，如果“肿瘤”的置信度得分超过另一种疾病名称的置信度得分(或阈值得分)，则疾病编码器230判断为最匹配的级别1ICD-10官方疾病名称是“肿瘤”。在一些实现中，神经网络模型具有相同的架构。例如，该架构可以是基于具有产生单个置信度/相似度得分的单个输出的孪生递归神经网络(RNN)。在一些实现中，关于来自一组EMR的一个或多个现有映射对各神经网络模型进行训练。

[0068] 药物编码器235是用以进行用于对与药物相关联的医疗实体进行编码的特定编码操作的分立计算元件。药物编码器235可被配置为针对与通用或药品名称相匹配的给定医疗实体检索相匹配的GPI-10代码。药物编码器235还可被配置为针对与通用或药品名称相匹配的给定医疗实体检索相匹配的ATC代码。在一些实现中，基于对照经管护的实体特定药物字典的数据条目进行精确和模糊搜索来进行映射或编码。药物字典可以包括GPI-10代码、药品名称、通用药品名称和ATC代码的列表，包括与不同药物的活性成分和化学性质相关的数据。在一些情况下，药物字典是使用示例性内部GPI-10通用名称映射字典或ATC代码映射字典来管护的。

[0069] 过程编码器240是用以进行用于对与医疗过程相关联的医疗实体进行编码的特定编码操作的分立计算元件。过程编码器240被配置为针对与特定过程名称相匹配的给定医疗实体(例如，医疗过程)检索相匹配的PRC_CD代码。在一些实现中，基于对照经管护的实体特定过程字典的数据条目的精确和模糊搜索来进行映射或编码。过程字典可以包括过程代码(PRC_CD)和过程名称的列表。在一些情况下，过程字典是使用示例性内部PRC_CD过程名称映射字典来管护的。

[0070] 编码模块140还可以包括一个或多个附加编码器245。例如，附加编码器245可以用于对描述与患者体征和症状、患者的生命、患者的实验室测试结果(例如，BMI或胆固醇数字)相关联的医学发现的数据进行编码。在一些实现中，附加编码器245可以用于对描述可能与医学发现相关联的时间实体的数据进行编码。例如，时间实体可以指示医学发现是在最近一个月内、是在最近一年内、还是在两年前获得或确定的。在其它实现中，附加编码器

245可以用于将数据编码为标识患者所属的年龄组(例如,成人、新生儿、大于18岁),或者编码为标识基因标记(例如,癌症或由于特定基因突变引起的其它异常)。

[0071] 如上所述,编码模块120和编码模块140进行交互以执行用于定义系统100的训练阶段155的各种数据处理和信号处理功能。执行这些功能能够完成训练阶段155,使得系统100可以生成经训练的预测模型160。

[0072] 在训练阶段155期间,查找引擎210对照经管护的实体特定字典在输入数据110中进行数据元素的搜索。例如,查找引擎210被配置为提取一个或多个单个字(例如,一元)、提取两个连续字(例如,二元)或提取三个连续字(例如,三元)。在一些实现中,查找引擎210按顺序提取一元、二元和三元,使得首先提取一元,其次提取二元,再次提取三元。在其它实现中,查找引擎210不以特定顺序提取一元、二元和三元(统称为“n元”)。

[0073] 所提取的n元以精确匹配模式或模糊模式对照经管护的实体特定字典进行匹配。在例如使用经训练的预测模型160的系统100的实现阶段期间,所提取的n元以模糊模式对照经管护的实体特定字典进行匹配。在一些实现中,模糊模式匹配是基于示例性最小编辑距离算法。例如,如果查找引擎210检测到输入术语/字或短语与特定字典中的实体之间的匹配,则利用相应的匹配医疗实体来标记特定匹配字或短语。

[0074] 解析引擎170是被配置为将用户输入125转换成机器可读命令的语义查询解析器。如以下更详细地所述,基于接收到的查询,解析引擎170通常被配置为:i)进行命名实体识别以从查询中提取患者属性;ii)将一个或多个患者属性编码成标准格式;iii)如果提到多个患者属性,则确定应如何组合这些属性中的一个或多个;iv)判断是否应否定任何患者属性标准;以及v)基于前述步骤i)-iv)中的一个或多个的结果,构建在医疗数据库190中进行搜索的计算机可理解查询。解析引擎170包括命名实体识别(NER)和编码引擎250、实体否定引擎255、以及连接词引擎260。对于给定的查询输入,NER和编码引擎250被配置为标识与一个或多个医学概念、患者诊断、患者处方历史和先前提供给患者的医疗过程有关的所有患者属性。实体否定引擎255可以包括用于标识是包括还是排除特定患者属性的否定标识符的软件指令。

[0075] 连接词引擎260可以包括用于标识查询输入中的连接词模式的软件指令。例如,在查询多个患者属性时,连接词引擎260可以标识示例用户输入中的一个或多个连接词模式。作为示例,连接词引擎260被配置为区分以下两个查询:i)“发热且(AND)疼痛的患者”与ii)“发热或(OR)疼痛的患者”。连接词引擎260被配置为基于各查询的所标识的连接词模式来区分第一查询和第二查询。例如,在第一查询i)中,“发烧”和“疼痛”属性这两者必须同时存在,而在第二查询ii)中,存在任意属性就足够了。

[0076] 图3是用于基于所述的技术来生成机器可读命令的示例处理300。处理300可以使用上述的系统100的一个或多个计算元件以及本文中所述的其它组件和指令来实现。

[0077] 现在参考处理300,系统100获得描述与多个不同的医学概念有关的术语的数据(302)。在一些实现中,所获得的数据是输入数据110中所包括的结构化数据或非结构化数据。例如,非结构化数据202包括从以下数据所获得的健康护理事务信息:描述医师和患者之间的交互的数据、从受试者的物理或电子医疗记录(EMR)得到的数据、从处方记录/医疗理赔得到的数据(“Rx/Dx数据”)、或者与患者或受试者所使用的处方或治疗选项相关的数据。

[0078] 系统100使用实体模块来判断术语是否描述医疗实体(304)。医疗实体可以与影响个人的健康护理状况相关联。例如,在系统100的训练阶段155期间,实体模块120在诸如患者EMR等的医疗文档或非结构化数据源中进行各字或术语的字典查找。医疗文档可以具有用于形成描述诸如癌症(例如,医疗实体)等的特定医疗状况的字的字母。在实体模块120检测到描述医疗实体或与医疗实体相关联的字与经管护的数据源中的也用于描述与癌症相关的医学概念的术语相匹配时,提取这些术语或字。

[0079] 在系统100的实现阶段期间,判断术语是否描述医疗实体可以包括:预测模型160计算用于确定例如用户输入125中的术语与描述诸如癌症等的医疗实体相关的置信度的推断。预测模型160可以生成表示术语描述医疗实体的置信度的置信度得分。例如,查询可以是“示出正在接受化疗的患者的列表吗?”。预测模型160可以基于超过阈值置信度得分的置信度得分来判断为术语“化疗”描述了医疗实体即癌症。

[0080] 系统100使用编码模块来将医疗实体与指定类别链接(306)。例如,至少在训练阶段155期间,当在输入数据110中识别的术语描述了医疗实体时,编码模块140用于将医疗实体与指定类别相链接。基于针对指定类别的编码方案来将医疗实体与指定类别链接或者利用指定类别对医疗实体进行编码。例如,编码模块140获得指定类别的类别代码145的列表,并且确定用于描述医疗实体的术语和类别代码列表中的相应一个或多个类别代码之间的匹配。然后,编码模块140基于所确定的用于描述医疗实体的术语和相应类别代码之间的匹配来将医疗实体与指定类别链接。

[0081] 在一些实现中,针对指定类别的编码方案是包括级别层级结构的层级编码方案。在该实现中,利用相应的类别代码对医疗实体进行编码包括:i) 量化包括要编码的医疗实体的信息内容;ii) 确定用于映射信息内容的级别层级结构中的级别的深度;以及iii) 针对级别层级结构中的特定深度级别,将信息内容中所包括的医疗实体与相应的类别代码相关联。对信息内容进行量化可以包括用于标识给定查询的正确类别级别的任务。例如,在疾病编码的情况下,术语“感染”可能不包含足以将医疗实体映射到特定类型的感染(诸如将医疗实体映射到“肺结核”等)的信息。因此,在这种情况下,如以上参考计算给定类别树的相似度得分所述,系统100可以将“感染”映射到包含所有“感染”的疾病的最高类别。

[0082] 预测计算系统100接收寻求响应的查询(308)。例如,该查询可以是作为适合参与评价新癌症治疗药物的疗效的临床试验的候选的受试者列表的请求。系统100的解析引擎170使用所接收到的查询来生成机器可读命令(310)。在一些实现中,该命令是响应于解析引擎170对照描述医疗实体的术语来解析查询并且基于用于将医疗实体链接到指定类别的编码方案而生成的。

[0083] 解析引擎170可以进行查询解析操作,以基于用户输入125中的特定关键字、用户输入125的文本中的字的特定位置或形成用户输入125的术语的语义属性来确定术语之间的关系类型。在一些实现中,解析引擎170使用一个或多个语义解析功能来确定用户输入125中的术语的语义属性。语义解析功能可以包括:例如通过应用从可被预先训练以检测输入短语或查询的句子语法的机器学习解决方案得到的特定计算规则,来从用户输入125中提取句子语法。

[0084] 系统100使用命令在一个或多个数据库中查询,以获得用于生成对所接收到的查询的响应的数据(312)。例如,经训练的预测模型160可以与实体模块120和编码模块140交

互,以将一个或多个患者属性标识为条件(例如,健康相关或医疗状况)。解析引擎170的语义解析功能可用于确定:(1)各条件是否被否定;以及(2)在指定多于一个的条件时,在不同条件之间存在什么关系。例如,可以使用诸如逻辑AND(与)运算、逻辑OR(或)运算或逻辑NOT(非)运算等的一个或多个逻辑操作数来确定或定义在不同条件之间存在的关系。

[0085] 图4示出用于处理可以包括在输入数据110中的结构化和非结构化医疗数据的示例数据流400。与上述的处理300类似,数据流400可以使用上述的系统100的一个或多个计算元件以及本文中所述的其它组件和指令来实现。

[0086] 现在参考数据流400,系统100使用实体模块120来进行用于标识文本文件中所包括的医学概念或医疗实体的示例任务(410)。例如,文本文件可以是诸如患者的EMR等的非结构化数据202中所包括的医疗期刊或文本。如图所示,所标识的文本可以是在较大一组术语、字或短语内标识的一个或多个术语。该较大一组术语可以标识在过去两年内患有某种类型的医疗诊断的特定患者。在一个实现中,该较大一组术语也可以表明患者“无<肾虚>病史”。在一些情况下,所标识的特定患者的文本是例如在系统100的训练阶段155期间从患者的EMR获得的非结构化数据202。在图4的实现中,实体模块120所标识的一个或多个术语包括“MRSA相关感染”和“肾虚”。

[0087] 使用上述技术,实体模块120与编码模块140交互,以检索由所标识的术语描述的一个或多个医疗实体的官方名称(415)。例如,实体模块120判断所标识的术语描述或涉及医疗实体420、即耐甲氧西林金黄色葡萄球菌(MRSA),其是导致身体的不同部位感染的细菌。实体模块120还判断所标识的术语描述或涉及医疗实体425即“肾虚”,其可能对应于包括急性肾衰竭或慢性肾病的其它医学概念或实体。

[0088] 例如,实体模块120确定为一个或多个术语描述医疗实体420、425包括:i)使用查找引擎210或ML引擎220来对照经管护的实体特定数据集中的信息进行较大一组术语的查找;ii)确定一个或多个术语与经管护的实体特定数据集中的信息之间的匹配;以及iii)基于所确定的一个或多个术语与经管护的实体特定数据集中的信息之间的匹配来确定一个或多个术语描述医疗实体420、425。

[0089] 编码模块140使用编码器230、235、240、245至少之一来检索由所标识的术语描述的一个或多个医疗实体420、425的官方名称。例如,编码模块140可以使用疾病编码器230的查找引擎232或ML引擎234而在经管护的实体特定疾病字典中进行与疾病名称有关的医疗实体的字典查找。查找引擎232和ML引擎234各自被配置为检索:i)与特定疾病名称匹配的针对医疗实体420的相应ICD-10代码430;以及ii)与特定疾病名称匹配的针对医疗实体425的相应ICD-10代码435。

[0090] 检索与一个或多个术语相关联的官方疾病名称可以包括:在一个或多个术语描述医疗实体420、425时,使用编码模块140将医疗实体420、425与指定类别相链接。基于针对指定类别的编码方案来将医疗实体420、425与指定类别链接。

[0091] 通常,医疗实体420、425可以至少包括以下项至少之一:a)与特定医学概念有关的医学疾病;b)用于治疗与特定医学概念有关的医学疾病的医疗药物;(c)同与特定医学概念有关的医学疾病相关联的医疗过程;或者d)用于描述与患者的健康护理或医疗状况相对应的多个医学发现的数据。

[0092] 在一些实现中,基于包括以下项中的一个或多个的数据来生成经管护的实体特定

数据集：i) 描述多个类型的疾病的预定义的一组信息，例如上述的经管护的实体特定疾病字典；ii) 描述多个药物的预定义的一组信息，例如上述的经管护的实体特定药物字典；iii) 描述多个医疗过程的预定义的一组信息，例如上述的经管护的实体特定过程字典；或者iv) 多个健康护理患者的电子医疗数据，例如上述的患者EMR数据。

[0093] 使用上述的技术，解析引擎170与预测模型160交互以生成计算机可理解查询(440)。例如，解析引擎170使用接收到的查询125来生成机器可读命令450。查询125可以是用于表明“向我示出在过去2个月内诊断为<MRSA相关感染>且无<肾虚>病史的患者”的用户查询。在一些实现中，该命令是通过对照描述医疗实体的术语解析查询并且基于用于将医疗实体链接到指定类别的编码方案而生成的。

[0094] 对于示例查询输入125，解析引擎170使用NER和编码引擎250来标识与患者的诊断和医疗病史相关联的所有患者属性。实体否定引擎255的否定标识符用于标识是包括还是排除某些患者属性。例如，实体否定模块255例如通过包括子命令“诊断NOT IN N00-N99”，可以使得所生成的命令被构造成将有肾虚病史的患者从数据库搜索中排除，其中“N00-N99”对应于急性肾衰竭或慢性肾病的示例ICD-10类别代码。

[0095] 连接词引擎260用于标识查询输入中的连接词模式。例如，如上所述，在查询多个患者属性时，连接词引擎260标识示例用户输入/查询125中的连接词模式。因而，连接词引擎260被配置为标识查询125中的“and”连接词。以这种方式，解析引擎170可以生成从患者数据库获得具有如下患者属性的患者列表的机器可读命令450，这些患者属性表示疾病类别代码B95.62 (MSRA相关感染) 且 (AND) 诊断日期范围为NOW (当前日期) 至过去2个月的诊断。机器可读命令450还可以包括用以使得数据库搜索返回已诊断出NOT IN疾病类别代码“N00-N99”的病状的患者否定子命令。

[0096] 图5是可用于将本文所述的系统和方法实现为客户端或者一个或多个服务器的计算装置500、550的框图。计算装置500和550意图表示各种形式的数字计算机，诸如膝上型电脑、台式电脑、工作站、个人数字助理、服务器、刀片式服务器、大型机和其它适当的计算机等。这里所示的组件、这些组件的连接和关系、以及这些组件的功能仅是示例性的，而并不是限制本文中所述和/或要求保护的实现。

[0097] 计算装置500包括处理器502、存储器504、存储装置506、连接至存储器504和高速扩展端口510的高速接口508、以及连接至低速总线514和存储装置506的低速接口512。各个组件502、504、506、508、510和512使用各种总线互连，并且可以安装在通用主板上或以其它适当的方式安装。处理器502可以处理供在计算装置500内执行用的指令，这些指令包括存储器504中或存储装置506上所存储的用以将GUI的图形信息显示在外部输入/输出装置(诸如连接至高速接口508的显示器516等)上的指令。在其它实现中，可以适当地使用多个处理器和/或多个总线以及多个存储器和存储器类型。此外，可以连接多个计算装置500，其中各装置提供所需操作的一部分，例如作为服务器阵列、一组刀片式服务器或多处理器系统。

[0098] 存储器504存储计算装置500内的信息。在一个实现中，存储器504是计算机可读介质。在一个实现中，存储器504是一个或多个易失性存储器单元。在另一实现中，存储器504是一个或多个非易失性存储器单元。

[0099] 存储装置506能够为计算装置500提供大容量存储。在一个实现中，存储装置506是计算机可读介质。在各种不同的实现中，存储装置506可以是软盘装置、硬盘装置、光盘装置

或磁带装置、闪速存储器或其它类似的固态存储器装置、或者包括存储区域网络或其它结构中的装置的装置阵列。在一个实现中, 计算机程序产品被有形地体现在信息载体中。计算机程序产品包含执行时进行诸如上述方法等的一个或多个方法的指令。信息载体是诸如存储器504、存储装置506或处理器502上的存储器等的计算机或机器可读介质。

[0100] 高速控制器508管理计算装置500的带宽密集型操作, 而低速控制器512管理较低带宽密集型操作。这种职责分配仅是示例性的。在一个实现中, 高速控制器508例如经由图形处理器或加速器连接至存储器504、显示器516, 并且连接至可以接受各种扩展卡(未示出)的高速扩展端口510。在该实现中, 低速控制器512连接至存储装置506和低速扩展端口514。低速扩展端口(其可以包括各种通信端口, 例如, USB、蓝牙(Bluetooth)、以太网(Ethernet)、无线以太网)可以连接至诸如键盘、指点装置、扫描器等的一个或多个输入/输出装置, 或者例如经由网络适配器连接至诸如交换机或路由器等的联网装置。

[0101] 如图所示, 计算装置500可以以多个不同的形式实现。例如, 计算装置500可被实现为标准服务器520, 或者可以在一组这样的服务器中多次实现。计算装置500也可被实现为机架服务器系统524的一部分。另外, 计算装置500可以在诸如膝上型计算机522等的个人计算机中实现。可选地, 来自计算装置500的组件可以与诸如装置550等的移动装置(未示出)中的其它组件组合。各个这样的装置可以包含计算装置500、550中的一个或多个, 并且整个系统可以由彼此通信的多个计算装置500、550组成。

[0102] 计算装置550包括处理器552、存储器564、诸如显示器554等的输入/输出装置、通信接口566和收发器568以及其它组件。该装置550还可以配备有诸如微型驱动器或其它装置等的存储装置, 以提供附加的存储。各个组件550、552、564、554、566和568使用各种总线互连, 并且这些组件中的数个组件可以安装在通用主板上或者以其它适当的方式安装。

[0103] 处理器552可以处理供在计算装置550内执行的指令, 这些指令包括存储器564中所存储的指令。处理器还可以包括单独的模拟和数字处理器。处理器例如可以提供装置550的其它组件的协调, 诸如用户界面的控制、装置550所运行的应用、以及由装置550进行的无线通信等。

[0104] 处理器552可以经由连接至显示器554的显示接口556和控制接口558与用户进行通信。显示器554例如可以是TFT LCD显示器或OLED显示器、或者其它适当的显示技术。显示接口556可以包括用于驱动显示器554以向用户呈现图形和其它信息的适当电路。控制接口558可以接收来自用户的命令并对这些命令进行转换以提交至处理器552。另外, 可以提供与处理器552进行通信的外部接口562, 以使得装置550能够与其它装置进行近区域通信。外部接口562例如可以提供例如经由对接过程的有线通信、或者例如经由蓝牙或其它这样的技术的无线通信。

[0105] 存储器564存储计算装置550内的信息。在一个实现中, 存储器564是计算机可读介质。在一个实现中, 存储器564是一个或多个易失性存储器单元。在另一实现中, 存储器564是一个或多个非易失性存储器单元。扩展存储器574还可以经由扩展接口572设置和连接到装置550, 该扩展接口572例如可以包括SIMM卡接口。这样的扩展存储器574可以为装置550提供额外的存储空间, 或者也可以存储装置550的应用或其它信息。具体地, 扩展存储器574可以包括用以执行或补充上述处理的指令, 并且还可以包括安全信息。因而, 例如, 扩展存储器574可被设置为装置550的安全模块, 并且可以编程有许可装置550的安全使用的指令。

另外,可以经由SIMM卡提供安全应用程序以及附加信息,诸如以不可入侵的方式将标识信息放置在SIMM卡上等。

[0106] 如以下所论述的,存储器例如可以包括闪速存储器和/或MRAM存储器。在一个实现中,计算机程序产品被有形地体现在信息载体中。计算机程序产品包含执行时进行诸如上述方法等的一个或多个方法的指令。信息载体是计算机或机器可读介质,诸如存储器564、扩展存储器574或处理器552上的存储器等。

[0107] 装置550可以经由通信接口566进行无线通信,该通信接口566可以在需要的情况下包括数字信号处理电路。通信接口566可以提供各种模式或协议下的通信,诸如GSM语音呼叫、SMS、EMS或MMS消息、CDMA、TDMA、PDC、WCDMA、CDMA2000或GPRS等。这样的通信例如可以经由射频收发器568发生。另外,可以发生诸如使用蓝牙、WiFi或其它这样的收发器(未示出)等的短距离通信。另外,GPS接收器模块570可以向装置550提供可由在装置550上运行的应用适当使用的附加无线数据。

[0108] 装置550还可以使用音频编解码器560来以可听见的方式进行通信,该音频编解码器560可以从用户接收口头信息并将该口头信息转换成可用的数字信息。音频编解码器560同样可以诸如经由(例如在装置550的手机中的)扬声器等为用户产生可听声音。这样的声音可以包括来自语音电话呼叫的声音,可以包括所记录的声音(例如语音信息、音乐文件等),并且还可以包括由装置550上运行的应用所产生的声音。

[0109] 如该图所示,计算装置550可以以多个不同的形式实现。例如,计算装置550可被实现为蜂窝电话580。计算装置550也可被实现为智能电话582、个人数字助理或其它类似移动装置的一部分。

[0110] 这里所述的系统和技术的各种实现可以在数字电子电路、集成电路、特别设计的ASIC、计算机硬件、固件、软件和/或它们的组合中实现。这些各种实现可以包括在包括至少一个可编程处理器的可编程系统上可执行和/或可解释的一个或多个计算机程序中的实现,该至少一个可编程处理器可以是专用或通用的,被连接以相对于存储系统、至少一个输入装置和至少一个输出装置进行数据和指令的接收和发送。

[0111] 这些计算机程序(也称为程序、软件、软件应用程序或代码)包括可编程处理器的机器指令,并且可以用高级程序和/或面向对象的编程语言、以及/或者用汇编/机器语言来实现。如本文所使用的,术语“机器可读介质”、“计算机可读介质”是指包括用于接收机器指令作为机器可读信号的机器可读介质的、任何计算机程序产品、装置和/或装置,例如磁盘、光盘、存储器、用于向可编程处理器提供机器指令和/或数据的可编程逻辑器件(PLD)。术语“机器可读信号”是指用于向可编程处理器提供机器指令和/或数据的任何信号。

[0112] 为了提供与用户的交互,这里所述的系统和技术可以在计算机上实现,该计算机具有用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或LCD(液晶显示器)监视器)、以及用户可以向计算机提供输入所利用的键盘和指点装置(例如,鼠标或追踪球)。其它种类的装置也可用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的感官反馈,例如视觉反馈、听觉反馈或触觉反馈;并且来自用户的输入可以以包括声学、语音或触觉输入的任何形式接收。

[0113] 这里所述的系统和技术可以在计算系统中实现,该计算系统包括例如作为数据服务器的后端组件,或者包括诸如应用服务器等的中间件组件,或者包括前端组件(诸如具有

图形用户界面的客户端计算机、或者用户可以利用这里所述的系统和技术的实现进行交互所经由的Web浏览器等),或者这样的后端组件、中间件组件或前端组件的任何组合。系统的组件可以通过诸如通信网络等的任何形式或介质的数字数据通信互连。通信网络的示例包括局域网(“LAN”)、广域网(“WAN”)和因特网。

[0114] 计算系统可以包括客户端和服务端。客户端和服务端通常彼此远离,并且通常经由通信网络进行交互。客户端和服务端的关系借助于在各个计算机上运行并且彼此具有客户端-服务端关系的计算机程序而产生。

[0115] 另外,特定数据可以在被存储或使用之前以一个或多个方式处置,使得移除个人可识别的信息。例如,在一些实施例中,可以对用户的身份进行处理,使得不能为用户确定个人可识别的信息,或者可以对获得位置信息的用户的地理位置(诸如城市、ZIP代码或州级等)进行概括,使得无法确定用户的特定位置。因此,用户可以控制收集与用户有关的哪些信息、如何使用该信息、以及向用户提供哪些信息。

[0116] 已经说明了数个实施例。然而,应当理解,可以在未背离本发明的精神和范围的情况下进行各种修改。因此,其它实施例在所附权利要求书的范围内。尽管本说明书包含许多特定实现细节,但这些细节不应被解释为对可能要求保护的范围的限制,而是应被解释为对可以特定于特定实施例的特征的描述。在本说明书中在单独实施例的上下文中所述的特定特征也可以在单个实施例中组合地实现。

[0117] 相反,在单个实施例的上下文中所描述的各种特征还可以在多个实施例中单独实现或者以任何合适的子组合实现。此外,尽管以上可以将特征描述为以某些组合起作用并且甚至最初如此要求保护这些特征,但在一些情况下可以从组合中消除所要求保护的组合中的一个或多个特征,并且所要求保护的组合可以针对子组合或子组合的变形。

[0118] 同样,尽管在附图中以特定顺序描绘了操作,但这不应该被理解为要求以所示的特定顺序或者以序列顺序来进行这些操作、或者进行所有所示操作以实现期望结果。在某些情形中,多任务和并行处理可以是有利的。此外,上述实施例中的各种系统组件的分离不应被理解为在所有实施例中都要求这种分离,并且应当理解,所描述的程序组件和系统通常可以一起集成在单个软件产品中或者封装到多个软件产品中。

[0119] 已经描述了主题的特定实施例。其它实施例在所附权利要求书的范围内。例如,权利要求书中所列举的动作可以以不同的顺序进行,并且仍然实现期望结果。作为一个示例,附图中所描绘的处理并非必须需要所示的特定顺序或序列顺序来实现期望的结果。

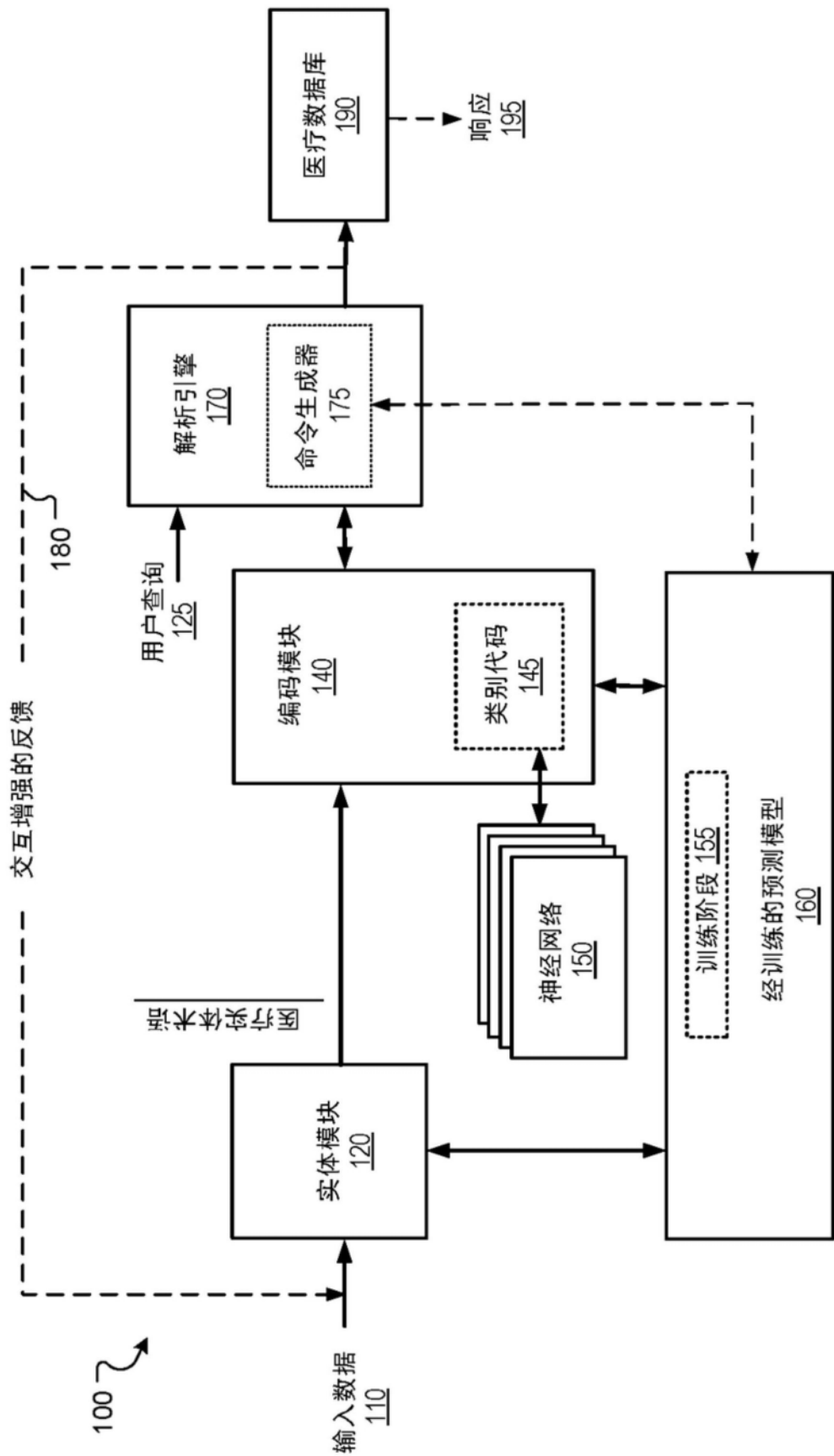


图1

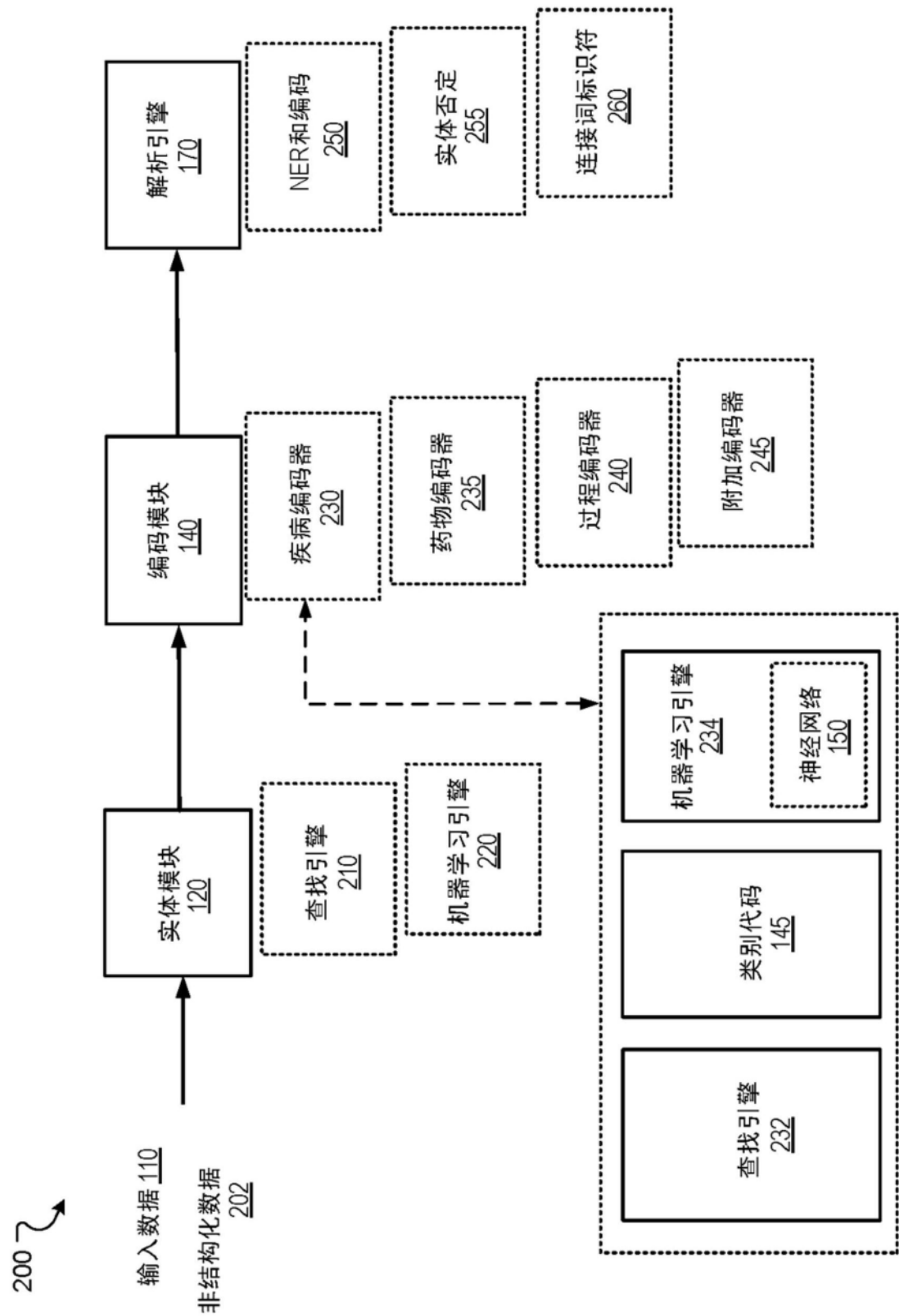


图2

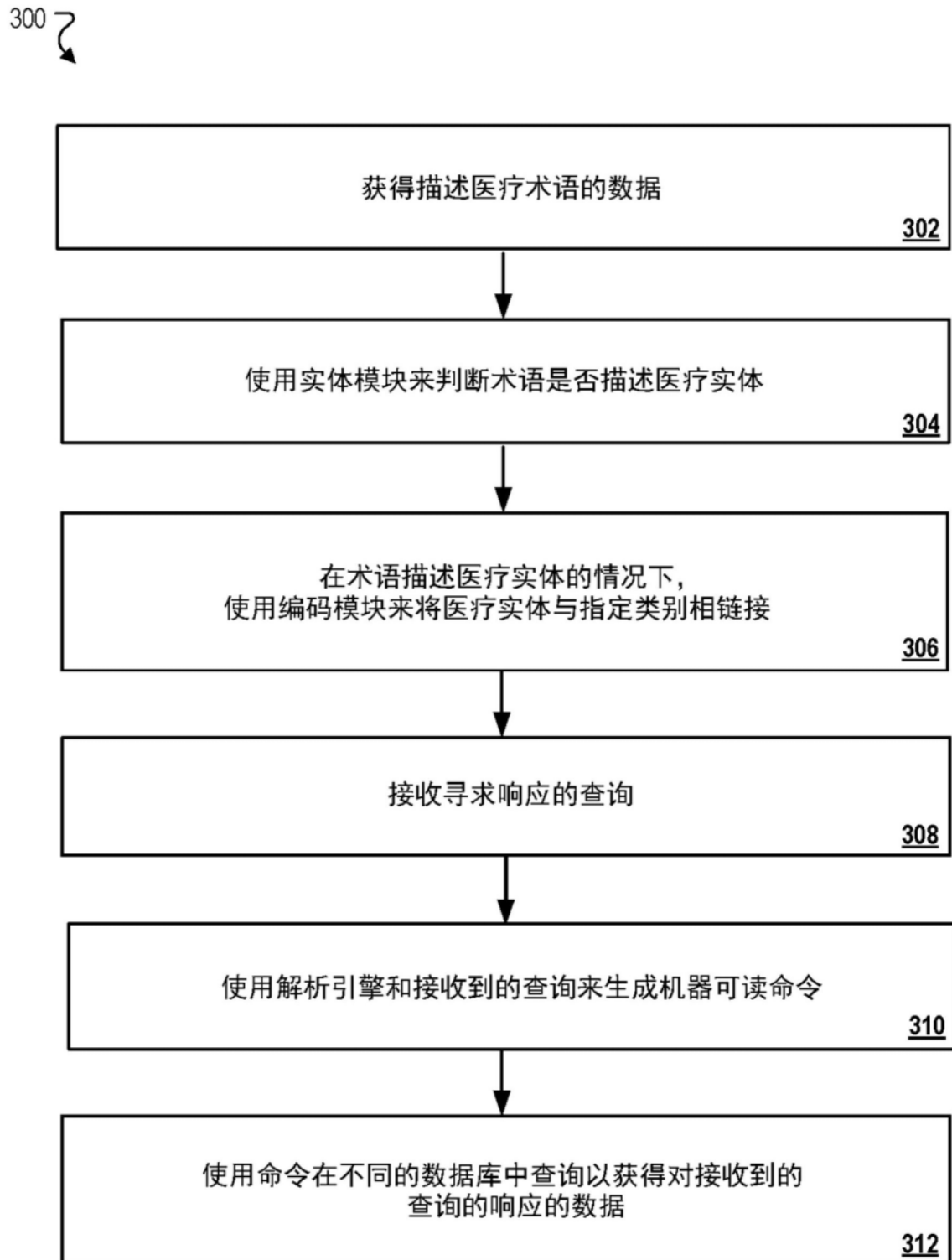


图3

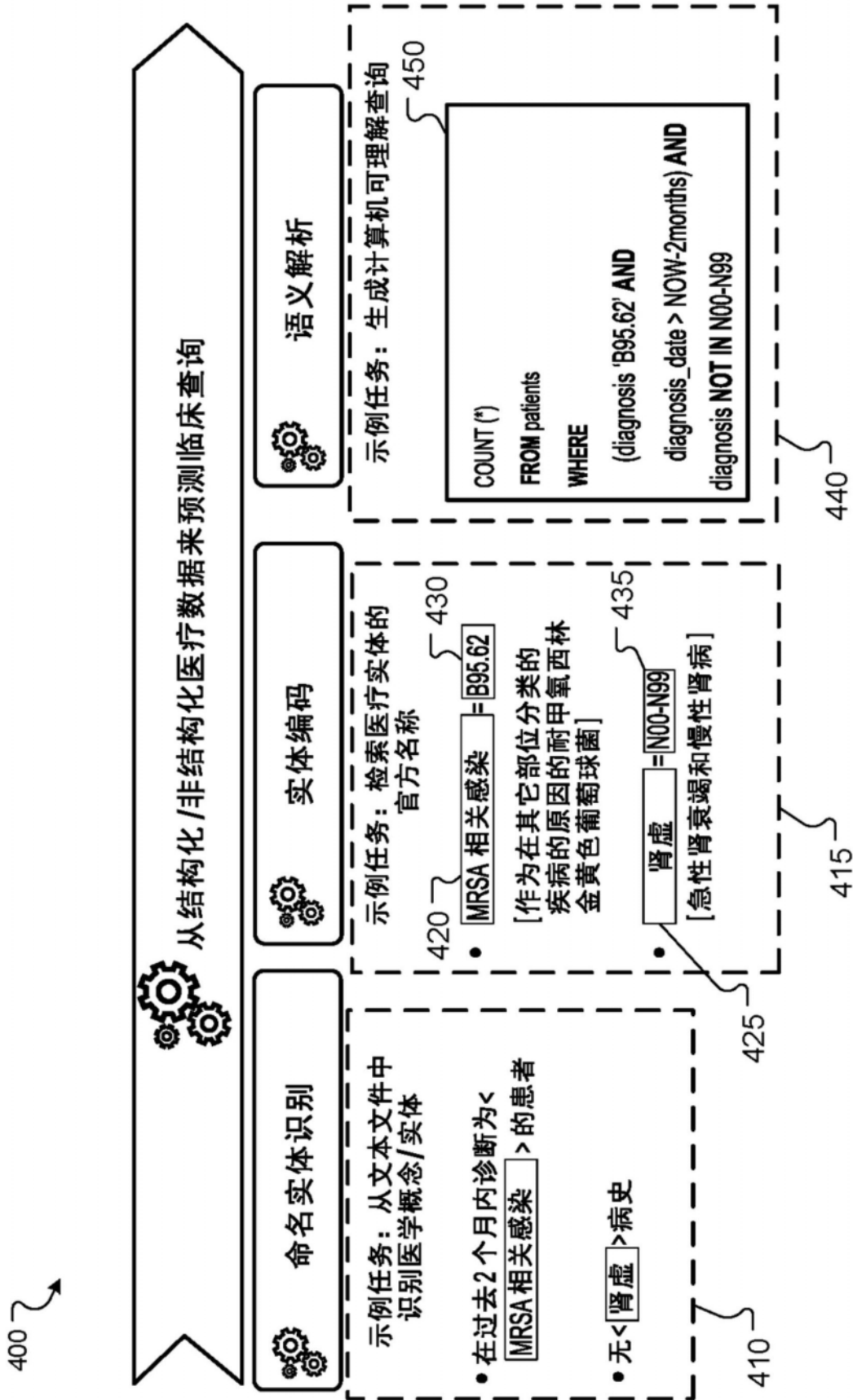


图4

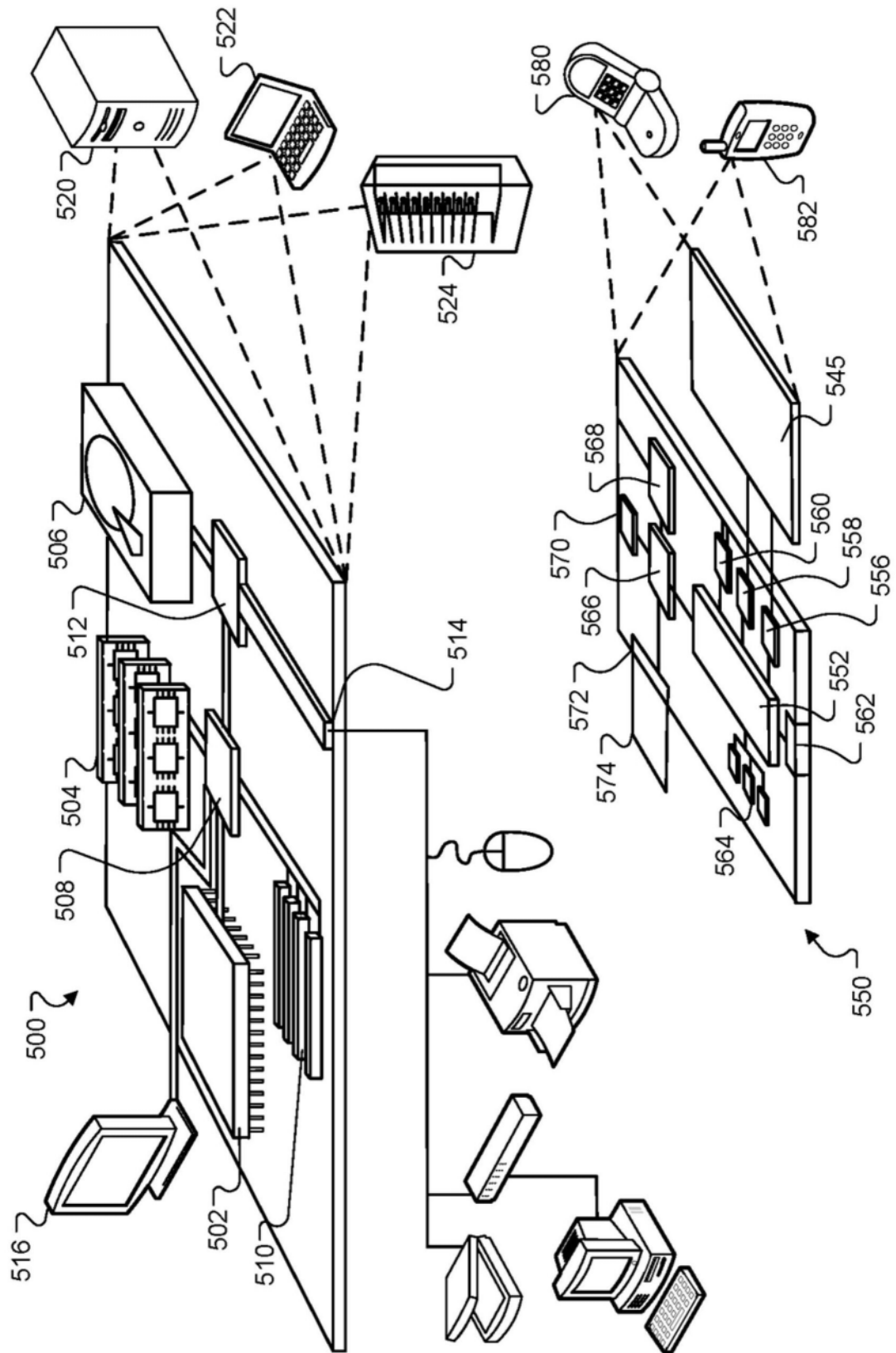


图5