(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 113892150 A
(43)申请公布日 2022.01.04

(21)申请号 202080038894.7
(22)申请日 2020.06.05
(30)优先权数据
16/454,311 2019.06.27 US
(85)PCT国际申请进入国家阶段日
2021.11.25
(86)PCT国际申请的申请数据
PCT/IB2020/055332 2020.06.05
(87)PCT国际申请的公布数据
WO2020/261002 EN 2020.12.30
(71)申请人 国际商业机器公司
地址 美国纽约阿芒克
(72)发明人 J·卡森 C·姆瓦拉布
T·H·罗杰斯 C·艾伦
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 马明月
(51)Int.Cl.
G16H 50/70(2006.01)
G06N 3/04(2006.01)

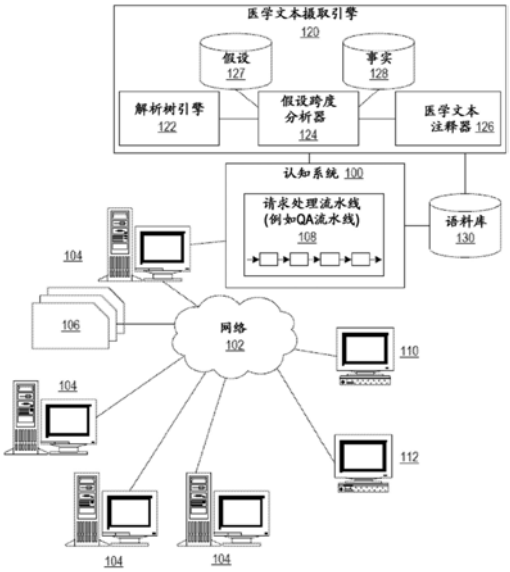
权利要求书4页 说明书26页 附图11页

(54)发明名称

计算跨度的深度学习方法

(57)摘要

公开了一种使用自然语言处理器的方法、系统和计算机程序产品。包括导入突出显示和未突出显示的训练文本，每个训练文本都包括训练节点，对训练文本进行单热编码，使用训练文本训练投影模型，使用投影模型处理突出显示的训练文本，以及使用突出显示的经处理的训练文本训练分类器模型。还包括导入包括新节点的新文本、对新文本进行单热编码、使用投影模型处理新文本以及使用分类器模型确定新节点之一是否在抢手的类中。



CN 113892150 A

1.一种使用自然语言处理器的方法,该方法包括:

导入包括第一组多个训练节点的突出显示的训练文本;

导入包括第二组多个训练节点的未突出显示的训练文本;

单热编码突出显示和未突出显示的训练文本;

使用突出显示和未突出显示的训练文本训练投影模型;

使用投影模型处理突出显示的训练文本;

使用突出显示的经处理的训练文本训练分类器模型;

导入包含多个新节点的新文本;

单热编码新文本;

使用投影模型处理新文本;以及

使用分类器模型确定多个新节点之一是否在抢手的类中。

2.根据权利要求1的所述的方法,其中抢手的类是假设文本跨度的成员。

3.根据权利要求1的所述的方法,还包括:

输出突出显示的新文本,该文本指示多个新节点中的每一个都在抢手的类中。

4.根据权利要求1的所述的方法,还包括:

使用突出显示和未突出显示的训练文本训练单热编码器。

5.根据权利要求1的所述的方法,还包括:

使用分类器模型处理突出显示的经处理的训练向量,以确定每个节点是否在抢手的类中;

将每个节点是否在抢手的类中的确定与每个节点的突出显示进行比较;以及

调整分类器模型以增加与突出显示相同的确定的次数。

6.根据权利要求1的所述的方法,还包括:

进行特征选择;以及

在训练投影模型之前,根据特征选择从突出显示和未突出显示的训练文本中删除节点。

7.一种训练自然语言处理器的方法,该方法包括:

导入包括第一组多个训练节点的突出显示的训练文本;

导入包括第二组多个训练节点的未突出显示的训练文本;

将突出显示的训练文本转换为突出显示的训练转换表;

将未突出显示的训练文本转换为未突出显示的训练转换表;

使用突出显示和未突出显示的训练转换表训练单热编码器;

对突出显示的训练转换表进行单热编码以生成突出显示的训练向量;

对未突出显示的训练转换表进行单热编码以生成未突出显示的训练向量;

使用突出显示和未突出显示的训练向量训练投影模型;

使用投影模型处理突出显示的训练向量,生成突出显示的经处理的训练向量;和

使用突出显示的经处理的训练向量训练分类器模型,其中分类器模型确定节点是否在抢手的类中。

8.根据权利要求7所述的方法,其中抢手的类是假设文本跨度的成员或事实文本跨度的成员。

9.根据权利要求7所述的方法,还包括:

将突出显示的训练文本转换为突出显示的解析树;以及

将未突出显示的训练文本转换为未突出显示的解析树。

10.根据权利要求7所述的方法,还包括:

使用分类器模型处理突出显示的经处理的训练向量,以确定每个节点是否在抢手的类中;

将每个节点是否在抢手的类中的确定与每个节点的突出显示进行比较;以及

调整分类器模型以增加与突出显示相同的确定的次数。

11.根据权利要求7所述的方法,还包括:

进行特征选择;以及

在训练投影模型之前,根据特征选择从突出显示和未突出显示的训练向量中删除一列。

12.一种在跨度中查找节点的系统,该系统包括:

多个突出显示的解析树,代表标记的自然语言文本;

多个未突出显示的解析树,代表未标记的自然语言文本;

新解析树,代表新自然语言文本;

自然语言处理(NLP)学习机,配置为处理多个突出显示的解析树、多个未突出显示的解析树和新解析树,其中NLP学习机包括计算处理器;以及

存储器,耦合到计算处理器,其中存储器包括指令,当由计算处理器执行时,该指令具体地配置计算处理器并使计算处理器:

导入包括第一组多个训练节点的突出显示的训练文本;

导入包括第二组多个训练节点的未突出显示的训练文本;

单热编码突出显示和未突出显示的训练文本;

使用突出显示和未突出显示的训练文本训练投影模型;

使用投影模型处理突出显示的训练文本;

使用突出显示的经处理的训练文本训练分类器模型;

导入包含多个新节点的新文本;

单热编码新文本;

使用投影模型处理新文本;以及

使用分类器模型确定多个新节点之一是否在抢手的类中。

13.根据权利要求12所述的系统,其中抢手的类是假设文本跨度的成员。

14.根据权利要求12所述的系统,其中,存储器还包括指令,当由计算处理器执行时,该指令具体地配置计算处理器并且使计算处理器:

输出突出显示的新文本,该文本指示多个新节点中的每一个都在抢手的类中。

15.根据权利要求12所述的系统,其中,存储器还包括指令,当由计算处理器执行时,该指令具体地配置计算处理器并且使计算处理器:

使用突出显示和未突出显示的训练文本训练单热编码器。

16.根据权利要求12所述的系统,其中,存储器还包括指令,当由计算处理器执行时,该指令具体地配置计算处理器并且使计算处理器:

使用分类器模型处理突出显示的经处理的训练向量,以确定每个节点是否在抢手的类中;

将每个节点是否在抢手的类中的确定与每个节点的突出显示进行比较;以及

调整分类器模型以增加与突出显示相同的确定的次数。

17.根据权利要求12所述的系统,其中,存储器还包括指令,当由计算处理器执行时,该指令具体地配置计算处理器并且使计算处理器:

进行特征选择;以及

在训练投影模型之前,根据特征选择从突出显示和未突出显示的训练文本中删除节点。

18.一种在跨度中查找节点的系统,该系统包括:

多个突出显示的解析树,代表标记的自然语言文本;

多个未突出显示的解析树,代表未标记的自然语言文本;

新解析树,代表新自然语言文本;

自然语言处理(NLP)学习机,配置为处理多个突出显示的解析树、多个未突出显示的解析树和新解析树,其中NLP学习机包括计算处理器;以及

存储器,耦合到计算处理器,其中存储器包括指令,当由计算处理器执行时,该指令具体地配置计算处理器并使计算处理器:

将突出显示的训练文本转换为突出显示的训练转换表;

将未突出显示的训练文本转换为未突出显示的训练转换表;

使用突出显示和未突出显示的训练转换表训练单热编码器;

对突出显示的训练转换表进行单热编码以生成突出显示的训练向量;

对未突出显示的训练转换表进行单热编码以生成未突出显示的训练向量;

使用突出显示和未突出显示的训练向量训练投影模型;

使用投影模型处理突出显示的训练向量,生成突出显示的经处理的训练向量;和

使用突出显示的经处理的训练向量训练分类器模型,其中分类器模型确定节点是否在抢手的类中。

19.根据权利要求18所述的系统,其中抢手的类是假设文本跨度的成员或事实文本跨度的成员。

20.根据权利要求18所述的系统,其中,存储器还包括指令,当由计算处理器执行时,该指令具体地配置计算处理器并且使计算处理器:

将突出显示的训练文本转换为突出显示的解析树;以及

将未突出显示的训练文本转换为未突出显示的解析树。

21.根据权利要求18所述的系统,其中,存储器还包括指令,当由计算处理器执行时,该指令具体地配置计算处理器并且使计算处理器:

使用分类器模型处理突出显示的经处理的训练向量,以确定每个节点是否在抢手的类中;

将每个节点是否在抢手的类中的确定与每个节点的突出显示进行比较;以及

调整分类器模型以增加与突出显示相同的确定的次数。

22.根据权利要求18所述的系统,其中,存储器还包括指令,当由计算处理器执行时,该

指令具体地配置计算处理器并且使计算处理器：

进行特征选择;以及

在训练投影模型之前,根据特征选择从突出显示和未突出显示的训练向量中删除一列。

23.一种计算机程序产品,包括计算机可读存储介质,其中存储有计算机可读程序以查找跨度中的节点,其中计算机可读程序在计算设备上执行时,具体地配置计算设备并使计算设备：

导入包括第一组多个训练节点的突出显示的训练文本；

导入包括第二组多个训练节点的未突出显示的训练文本；

单热编码突出显示和未突出显示的训练文本；

使用突出显示和未突出显示的训练文本训练投影模型；

使用投影模型处理突出显示的训练文本；

使用突出显示的经处理的训练文本训练分类器模型；

导入包含多个新节点的新文本；

单热编码新文本；

使用投影模型处理新文本;以及

使用分类器模型确定多个新节点之一是否在抢手的类中。

24.根据权利要求23所述的计算机程序产品,其中抢手的类是假设文本跨度的成员。

25.根据权利要求23所述的计算机程序产品,其中计算机可读程序进一步在计算设备上执行时,具体地配置计算设备并使计算设备：

进行特征选择;以及

在训练投影模型之前,根据特征选择从突出显示和未突出显示的训练文本中删除节点。

计算跨度的深度学习方法

背景技术

[0001] 本申请总体上涉及一种改进的数据处理装置和方法，更具体地，提供一种机制来对文本中的假设陈述，例如医学文本、司法陈述和其他文本文件语料库，进行机器学习。

[0002] 决策支持系统存在于许多不同的行业中，在这些行业中，人类专家在检索和分析信息方面需要帮助。在整个应用中使用的一个示例是医疗保健行业中使用的诊断系统。诊断系统可以分为使用结构化知识的系统、使用非结构化知识的系统和使用临床决策公式、规则、树或算法的系统。最早的诊断系统使用结构化知识或经典的、手动构建的知识库。随着开发的进展，更复杂的概率推理能力被添加，然后使用非结构化知识的系统开始出现。最近，已经为许多医学疾病开发了临床决策规则，并且已经开发了计算机系统来帮助从业者和患者应用这些规则。

发明内容

[0003] 根据本公开的一些实施例，公开了一种用于使用自然语言处理器的方法、系统和计算机程序产品。包括导入突出显示(highlighted)和未突出显示的训练文本，每个训练文本都包括训练节点，对训练文本进行单热(one-hot)编码，使用训练文本训练投影模型，使用投影模型处理突出显示的训练文本，以及使用经处理的突出显示的训练文本训练分类器模型。还包括导入包括新节点的新文本、对新文本进行单热编码、使用投影模型处理新文本以及使用分类器模型确定新节点之一是否在抢手的类中。

[0004] 根据本公开的一些实施例，公开了一种用于使用自然语言处理器的方法和系统。包括导入突出显示的训练文本，每个训练文本都包括训练节点，将训练文本转换为训练转换表，使用训练转换表训练单热编码器，并对训练转换表进行单热编码以生成训练向量。还包括使用训练向量训练投影模型，使用投影模型处理突出显示训练向量以生成经处理的突出显示训练向量，使用经处理的突出显示训练向量训练分类器模型，其中分类器模型确定节点是否在抢手的类中。

附图说明

[0005] 现在将参考附图仅通过示例的方式描述本发明的实施例，其中：

图1描绘了计算机网络中认知医疗保健系统的说明性实施例的示意图。

图2是其中实现说明性实施例的方面的示例数据处理系统的框图。

图3是描绘根据说明性实施例的医疗保健认知系统的元素的交互的示例图。

图4是根据说明性实施例的示例注释的解析树数据结构表示的示例，由医学专业人员撰写，可以是患者电子病历(EMR)的一部分。

图5是根据说明性实施例的句子的另一解析树数据结构的示例，其中执行对应于忽略触发器的节点的消歧。

图6A是根据说明性实施例的句子的另一解析树数据结构的示例，描述了句子的节点和连接边。

图6B是沿两次遍历的解析树的两个表的示例。

图6C是表格形式的两次遍历及其对应的单热编码向量的示例比较表。

图7是寻找解析树数据结构的跨度的示例方法的流程图。

图8是使用解析树训练自然语言处理(NLP)系统对自然语言文本进行操作的方法的流程图。

图9是NLP学习机对自然语言文本进行操作的流程图。

具体实施方式

[0006] 当对文本部分(例如医学文本、司法文本等)执行自然语言处理时,能够区分针对实际事实陈述的文本部分和包括假设陈述的文本部分通常很重要。例如,在执行医学文本和自然语言处理以促进患者治疗的情况下,能够将对于更准确的治疗建议很重要的实际事件与可能代表可能导致患者错误诊断和治疗的可能性的文本的假设部分区分开来通常是至关重要的。大多数情况下,医疗记录既包含描述实际发生的事实,也包含表明与患者讨论过但实际上并未发生的计划(又名假设)。例如,患者的电子病历(EMR)可能有实验室报告,表明进行了特定的实验室测试,并且从实验室测试中获得了特定的结果。这将是关于患者发生的实际事实事件的示例。此外,医生可能在患者的EMR中有自己的注释,表明医生与患者讨论的潜在程序或事件(例如,"We recommended that the patient have a mammogram performed(我们建议患者进行乳房X光检查)")。这种潜在的程序或事件实际上并未发生,但代表了患者的潜在计划。虽然与患者讨论了项目,但它们实际上是假设性的,因为当时不知道是否将会发生该程序或事件。

[0007] 对于认知系统,例如可从纽约阿蒙克的国际商业机器公司获得的IBM**Watson®**认知系统,实际情况通常是此类医学文本中最重要的部分,因为治疗建议是基于实际事件和事实。然而,计划的动作或未实施的动作、可能的事件等(即,假设)也可以记录在医学文本中,即使它们尚未代表实际事件或事实。为了提高这种认知系统的准确性,能够将表示假设的文本内容部分与表示实际事实和事件的文本内容部分区分开来将是有益的。因此,认知系统生成的治疗建议将基于代表实际事实和事件的部分。

[0008] 说明性实施例提供了用于摄取电子文本、文档或文本内容的其他部分并分析文本内容以区分针对假设的文本部分与针对实际发生的实际事实或事件的文本部分的机制。出于以下描述的目的,将提供针对医学文本和认知医学治疗建议系统的上下文中的实现的说明性实施例。然而,此类示例实施例不应被视为限制性上下文。特别地,应当理解,在不脱离本发明的精神和范围的情况下,可以针对除医学文本之外的各种领域的任何类型的文本来实施各种其他实施例。因此,例如,下文描述的机制可以关于司法文本或任何其他类型的文本来实施,这些文本可能包括假设部分和事实部分,并且文本的假设部分和事实部分之间的区别随后用于执行分析、认知或其他处理文本以生成结果。

[0009] 在其中说明性实施例的机制将文本的事实部分与文本的假设部分区分开的医疗建议系统实施例的上下文中,说明性实施例的机制可以摄取各种类型的医学文本并将说明性实施例的机制应用于这些医学文本。这些医学文本可以包括例如患者电子病历(EMR),其中医疗服务提供者(例如,医生、护士、医院、医学实验室、药房和医疗保险公司)可以贡献内容以包含在EMR中。因此,来自每个来源的医学文本可能包含事实(例如,实际发生的事件、

事件或结果)和假设(例如,计划或其他实际上并未发生的可能性)。

[0010] 在某些情况下,单个陈述或医学文本可能包含事实和假设,例如在示例陈述中,“Given her node positivity and lack of comorbidities,we recommend neoadjuvant therapy(鉴于她的淋巴结阳性且没有合并症,我们建议新辅助治疗)”在这种情况下,在为患者提出治疗建议时,希望知道患者具有淋巴结阳性并且没有合并症的事实。然而,对于治疗建议系统来说,知道患者实际上并未接受新辅助治疗而不是将这部分陈述也解释为事实也是至关重要的.而不是系统认为患者实际上已经接受了新辅助治疗,系统应该能够确定声明的这部分是指对未来计划的建议(即假设)而不是事件发生的事实。因此,系统可以忽略语句的这部分或简单地将这部分与语句的其余部分区别对待。

[0011] 为了将描述实际事实的医学文本部分与涉及假设的文本部分区分开来,说明性实施例提供了实现不假设句子结构的可概括方法的机制。说明性实施例使用两组字典数据结构。第一个是旨在识别与医疗建议认知系统在执行医疗建议分析时可能忽略的内容的假设部分相对应的术语和短语的字典数据结构集。第二个是旨在区分与内容的事实部分相关联的术语和短语,这些术语和短语应该用作执行此类医疗建议分析的基础的第二字典数据结构集。此外,还使用了解析树,其中包括应用词典的文本内容的增强表示。注释的跨度(例如,假设或事实注释)是通过查看以匹配字典条目为根的子树来确定的。例如,如果解析树的节点与假设字典数据结构中的假设术语或短语匹配,则以匹配的假设术语或短语为根的子树可以被注释为假设的。由说明性实施例的机制实现的方法很容易针对以前未见过的情况进行调整,例如通过不同的或更新的假设术语/短语的词典。

[0012] 说明性实施例可以在执行医学文本的自然语言处理的医疗建议系统的后端部分中操作。在后端系统中,可以使用包括实现本公开的一个或多个说明性实施例的一个或多个模型的若干自然语言处理模型来分析医学文本。这种分析的结果是一组带注释的医学文本,可以由医疗建议认知系统在机器学习和对特定患者EMR的实际应用方面使用以提供特定患者医疗建议。

[0013] 在开始更详细地讨论说明性实施例的各个方面之前,首先应当理解,在整个描述中,术语“机制”将用于指代执行各种操作、功能之类的本公开的元件。“机制”,如本文所使用的术语,可以是装置、过程或计算机程序产品形式的说明性实施例的功能或方面的实现。在程序的情况下,程序由一个或多个设备、装置、计算机、数据处理系统等实现。在计算机程序产品的情况下,由包含在计算机程序产品中或计算机程序产品上的计算机代码或指令表示的逻辑由一个或多个硬件设备执行,以实现与特定“机制”相关的功能或操作。因此,本文描述的机制可以被实现为专用硬件、在通用硬件上执行的软件、存储在介质上使得指令易于由专用或通用硬件执行的软件指令、用于执行功能的过程或方法,或以上任何一项的组合。

[0014] 本公开和权利要求可以使用关于说明性实施例的特定特征和元素的术语“一个”、“至少一个”和“一个或多个”。应当理解,这些术语和短语旨在说明在特定说明性实施例中存在至少一个特定特征或元素,但也可以存在多个。即,这些术语/短语不旨在将描述或权利要求限制为存在单个特征/元素或要求存在多个这样的特征/元素。相反,这些术语/短语仅需要至少单个特征/元素,并且多个这样的特征/元素可能在说明书和权利要求的范围内。

[0015] 此外，应当理解的是，如果这里使用术语“引擎”来描述本公开的实施例和特征，则不应限制用于完成和/或执行归因于引擎和/或由引擎执行的动作、步骤、过程等的实现。引擎可以是但不限于执行指定功能的软件、硬件和/或固件或其任何组合，包括但不限于与加载的或在机器可读存储器中存储的适当软件相结合并由处理器执行的。此外，除非另有说明，否则与特定引擎相关联的任何名称都是为了方便参考，并不旨在限制特定实现。另外，归因于引擎的任何功能可以由多个引擎同等地执行，并入和/或与相同或不同类型的另一引擎的功能组合，或者分布在各种配置的一个或多个引擎上。

[0016] 此外，应当理解，以下公开针对说明性实施例的各种元素使用多个各种示例来进一步说明说明性实施例的示例实现方式并帮助理解说明性实施例的机制。这些示例旨在是非限制性的并且并非穷尽用于实现说明性实施例的机制的各种可能性。鉴于本公开，对于本领域的普通技术人员来说显而易见的是，对于这些各种元件，除了这里提供的示例之外，或者替代这里提供的示例，还有许多其他替代实施方式可以被利用，而不脱离本发明的精神和范围。

[0017] 本发明可以是任何可能的技术细节集成水平的系统、方法和/或计算机程序产品。计算机程序产品可以包括其上具有用于使处理器执行本发明的方面的计算机可读程序指令的计算机可读存储介质(或介质)。

[0018] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是－－但不限于－－电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如，通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0019] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0020] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、集成电路配置数据或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等，以及过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网

络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读程序指令的状态信息来个性化定制电子电路,例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA),该电子电路可以执行计算机可读程序指令,从而实现本发明的各个方面。

[0021] 这里参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0022] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0023] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上,使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0024] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分,所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0025] 如上所述,本公开提供了用于分析医学文本以及区分文本的假设部分和文本的事实部分的机制,以及注释这样的文本部分使得医疗建议认知系统在执行机器学习和/或医疗建议操作时可以将这样的文本部分包括或排除在的进一步考虑之外。说明性实施例的机制通过通知系统可以准确地依赖作为实际事实指示哪些部分,而不能依赖潜在事实(即假设)的哪些部份来提高医疗建议认知系统的准确性。通过这种方式,医疗建议认知系统能够将最终的医疗建议建立在事实事件的基础上,而不会受到医疗文本的假设部分的错误影响。

[0026] 说明性实施例的机制利用混合方法,该方法涉及使用专门构造的字典数据结构集以及解析树数据结构集。专门构造的字典数据结构集包括一组假设字典数据结构,这些结构指定指示内容的假设部分的术语或短语,其中假设字典数据结构集中的这些术语或短语在本文中被称为“忽略触发器”。专门构造的字典数据结构集还包括一组事实字典数据结构,这些结构指定指示内容的事实部分的术语或短语,其中事实字典数据结构集中的这些

术语或短语在本文中被称为“确认触发器”。忽略触发器和确认触发器与从解析树中获得的部分文本内容(例如,文档、段落、句子、短语等)的系统视图相结合,从而实现了更通用的方法。

[0027] 忽略和确认触发器与解析树的组合允许将解析树的部分识别为对应于内容的假设部分,这里也称为“假设跨度”,以及与内容的事实部分相关联的解析树的其他部分,这里也称为“事实跨度”。在与这些内容部分相关联的元数据中,这些不同的跨度可以分别被注释为假设的或事实的。注释的跨度或内容的部分然后可以由医疗建议认知系统处理以便忽略对应于假设跨度的内容部分(例如,可以将零权重应用于内容的这些部分或者可以提供逻辑以提供假设跨度中的信息的其他评估作为医疗专业人员的计划)。

在一些说明性实施例中,包含在假设跨度内的注释可以被移除以生成修剪的解析树,被提供给医疗建议认知系统以用于执行治疗建议认知操作。在其他说明性实施例中,不是给予假设跨度零权重或从解析树中修剪这些跨度,而是与事实跨度内的注释相比,可以给予假设跨度内的注释相对较低的权重,以便仍然允许提供来自假设跨度的一些影响,但通过相对较低的权重来减轻它们的影响。

[0029] 说明性实施例可以用在许多不同类型的数据处理环境中。为了提供用于说明性实施例的特定元素和功能的描述的上下文,下文提供图1-3作为其中可以实现说明性实施例的方面的示例环境。应当理解,图1-3仅是示例并且不旨在断言或暗示关于可以实现本发明的方面或实施例的环境的任何限制。在不脱离本发明的精神和范围的情况下,可以对所描绘的环境进行许多修改。

[0030] 图1-3旨在描述用于医疗保健应用的示例认知系统(在此也称为“医疗保健认知系统”),实现请求处理流水线(例如问答(QA)流水线(也称为QA流水线或问答流水线))、请求处理方法和请求处理计算机程序产品,利用其实现说明性实施例的机制。这些请求可以作为结构化或非结构化请求消息、自然语言问题或用于请求由医疗保健认知系统执行的操作的任何其他合适的格式来提供。如下文更详细描述的,在本发明的实施例的示例认知系统中实现的特定医疗保健应用是用于提供医疗建议的医疗保健应用,因此,医疗保健认知系统这里也可以被称为医疗建议认知系统。

[0031] 应当理解,虽然在下文的示例中被示为具有单个请求处理流水线,但医疗保健认知系统实际上可以具有多个请求处理流水线。每个请求处理流水线可以被单独训练和/或配置为处理与不同域相关联的请求,或者被配置为根据期望的实现对输入请求(或使用QA流水线的实现中的问题)执行相同或不同的分析。例如,在一些情况下,可以训练第一请求处理流水线对针对第一医学疾病域(例如,各种类型的血液疾病)的输入请求进行操作,而可以训练另一请求处理流水线针对另一医学疾病域(例如,各种类型的癌症)的输入请求进行回答。在其他情况下,例如,请求处理流水线可以配置为提供不同类型的认知功能或支持不同类型的医疗保健应用,例如一个请求处理流水线用于患者诊断,另一个请求处理流水线用于医疗建议,另一个请求处理流水线被配置用于患者监控等。

[0032] 此外,每个请求处理流水线可以具有它们摄取和操作的自己的关联语料库(例如,在上述示例中,一个用于血液疾病领域文档的语料库和另一个用于癌症诊断领域相关文档的语料库)。在某些情况下,请求处理流水线可能都在相同的输入问题域上运行,但可能具有不同的配置(例如,不同的注释者或不同的经训练的注释者,从而生成不同的分析和潜在

的答案)。医疗保健认知系统可以提供额外的逻辑来将输入问题路由到适当的请求处理流水线,例如基于输入请求的确定域,组合和评估由多个请求处理流水线执行的处理产生的最终结果,以及其他有助于利用多个请求处理流水线的控制和交互逻辑。

[0033] 如上所述,可以利用说明性实施例的机制的一种类型的请求处理流水线是问答(QA)流水线。下文对本发明示例实施例的描述将利用QA流水线作为请求处理流水线的示例,该请求处理流水线可以被扩充以包括根据一个或多个说明性实施例的机制。应当理解,虽然将在实现对输入问题进行操作的一个或多个QA流水线的认知系统的上下文中描述本发明的实施例,但是说明性实施例不限于此。相反,说明性实施例的机制可以对不作为"问题"提出但被格式化为请求认知系统使用相关联的语料库和特定配置对指定的输入数据集执行认知操作的请求进行操作用于配置认知系统的信息。例如,不是询问"What diagnosis applies to patient P?(什么诊断适用于患者P?)"的自然语言问题,而是认知系统可以接收"generate diagnosis for patient P(为患者P生成诊断)"等的请求。应当理解,QA系统流水线的机制可以以与输入自然语言问题类似的方式对请求进行操作,并稍作修改。在某些情况下,如果特定实现需要,可以将请求转换为自然语言问题以供QA系统流水线处理。

[0034] 如下文将更详细地讨论的,说明性实施例可以被集成到、增强和扩展医疗保健认知系统的这些QA流水线(或请求处理流水线)机制的功能,针对注释摄取的医学文本和对这些摄取的医学文本进行操作。因此,可以执行区分医学文本的假设部分和医学文本的事实部分的基于医疗保健的操作。特别地,在一些说明性实施例中,医学文本可以包括患者EMR并且基于医疗保健的操作可以包括基于患者的EMR提供医疗建议。通过这种方式,医疗保健认知系统提供了一个针对医疗建议的决策支持系统。

[0035] 鉴于上文,在描述说明性实施例的机制如何被集成到并增强这样的认知系统和请求处理流水线或QA流水线机制之前,重要的是首先了解认知系统以及实施QA流水线的认知系统中的问答创建是如何实施的。应当理解,图1-3中描述的机制仅是示例并且不旨在陈述或暗示关于实施说明性实施例的认知系统机制的类型的任何限制。在不脱离本发明的精神和范围的情况下,可以在本发明的各种实施例中实现对图1-3中所示的示例认知系统的许多修改。

[0036] 作为概述,认知系统是专门的计算机系统或一组计算机系统,配置有硬件和/或软件逻辑(与软件在其上执行的硬件逻辑相结合)以模拟人类认知功能。这些认知系统将类似人类的特征应用于传达和操纵思想,结合数字计算的固有优势,可以大规模解决高精度和弹性的问题。认知系统执行一个或多个计算机实现的认知操作,这些操作近似于人类的思维过程,并使人与机器以更自然的方式交互,从而扩展和放大人类的专业知识和认知。认知系统包括人工智能逻辑,例如基于自然语言处理(NLP)的逻辑,以及机器学习逻辑,可以作为专用硬件、在硬件上执行的软件或在硬件上执行的专用硬件和软件的任何组合。认知系统的逻辑实现认知操作,示例包括但不限于问答、识别语料库中不同部分内容内的相关概念、智能搜索算法,例如互联网网页例如,搜索医疗诊断和治疗建议以及其他类型的建议生成(例如,特定用户感兴趣的项目、潜在的新联系人建议等)。

[0037] IBM**Watson®**是一种这样的认知系统的示例,可以处理人类可读语言并以比人类快得多的速度和更大规模地以类似人类的高精度识别文本段落之间的推理。一般而言,此类认知系统能够执行但不限于以下一项或多项功能:

驾驭人类语言和理解的复杂性；
摄取和处理大量结构化和非结构化数据；
产生和评估假设；
权衡和评估仅基于相关证据的反应；
提供针对具体情况的建议、见解和指导；
通过机器学习过程在每次迭代和交互中提高知识和学习；
在影响点做出决策(例如，上下文指导)；
与任务成比例；
扩展和放大人类的专业知识和认知；
从自然语言中识别出共鸣的、类人的属性和特征；
从自然语言中推断出各种特定于语言或不可知的属性；
从数据点(图像、文本、语音)(例如，记忆和回忆)中以高度相关性进行回忆；
预测和感知基于经验的模拟人类认知的情境意识；和
根据自然语言和具体证据回答问题；

[0038] 一方面，认知系统提供用于使用问答流水线或系统(QA系统)和/或处理可以或不可以作为自然语言问题提出的请求来回答向这些认知系统提出的问题的机制。QA流水线或系统是在数据处理硬件上执行的人工智能应用，它回答与以自然语言呈现的给定主题域相关的问题。QA流水线从各种来源接收输入，包括通过网络的输入、电子文档或其他数据的语料库、来自内容创建者的数据、来自一个或多个内容用户的信息以及来自其他可能的输入源的其他此类输入。数据存储设备存储数据语料库。内容创建者在文档中创建内容，以用作QA流水线的数据语料库的一部分。该文档可以包括用于QA系统的任何文件、文本、文章或数据源。例如，QA流水线访问有关领域或主题领域(例如，金融领域、医学领域、法律领域等)的知识体系，其中知识体系(知识库)可以以各种配置进行组织(例如，域特定信息的结构化存储库，例如本体，或与域相关的非结构化数据，或关于域的自然语言文档的集合)。

[0039] 内容用户向实现QA流水线的认知系统输入问题。然后，QA流水线通过评估文档、文档部分、语料库中的数据部分等，使用数据语料库中的内容回答输入问题。当流程评估文档的给定部分的语义内容时，该流程可以使用各种约定从QA流水线中查询此类文档(例如，将查询作为格式良好的问题发送到QA流水线，然后由QA流水线解释并提供包含一个或多个问题答案的响应)。语义内容是基于能指(signifier)之间的关系的内容，例如单词、短语、符号和符号，以及它们代表什么、它们的外延或内涵。换句话说，语义内容是解释表达的内容，例如通过使用自然语言处理。

[0040] 如下文将更详细描述的，QA流水线接收输入问题，解析问题以提取问题的主要特征，使用提取的特征来制定查询，然后将这些查询应用于数据语料库.基于对数据语料库的查询应用，QA流水线生成一组假设或输入问题的候选答案，方法是查看数据语料库中可能包含的部分数据对输入问题的有用回答。然后，QA流水线使用各种推理算法对输入问题的语言以及在应用查询期间发现的数据语料库的每个部分中使用的语言进行深入分析。可能应用了数百甚至数千种推理算法，每种算法执行不同的分析(例如，比较、自然语言分析、词法分析等)并生成分数。例如，一些推理算法可能会查看输入问题语言内的术语和同义词与数据语料库中找到的部分的匹配情况。其他推理算法可能会查看语言中的时间或空间特

征,而其他推理算法可能会评估数据语料库部分的来源并评估其准确性。

[0041] 从各种推理算法获得的分数表明潜在响应在何种程度上被输入问题基于该推理算法的特定关注领域推断出来。然后根据统计模型对每个结果分数进行加权。统计模型捕获推理算法在QA流水线的训练期间在特定域的两个相似段落之间建立推理时的执行情况。统计模型用于总结QA流水线对于问题推断出潜在响应(即候选答案)的证据的置信度。对每个候选答案重复此过程,直到QA流水线将表面上的候选答案识别为明显强于其他答案,从而为输入问题生成最终答案或排名答案集。

[0042] 如上所述,QA流水线机制通过访问来自数据或信息语料库(也称为内容语料库)的信息、对其进行分析、然后基于对该数据的分析来生成答案结果来操作。从数据语料库访问信息通常包括:用于回答有关结构化记录集合中的内容的问题的数据库查询,以及提供文档链接集合以响应针对非结构化数据集合(文本、标记语言等)的查询的搜索。传统的问答系统能够根据数据语料库和输入问题生成答案,验证数据语料库问题集合的答案,使用数据语料库纠正数字文本中的错误,并从潜在答案池中选择问题的答案(即候选答案)。

[0043] 内容创建者,例如文章作者、电子文档创建者、网页作者、文档数据库创建者等在编写他们的内容之前确定这些内容中描述的产品、解决方案和服务的用例。因此,内容创建者知道该内容打算在该内容解决的特定主题中回答什么问题。在数据语料库的每个文档中对问题进行分类(例如根据角色、信息类型、任务等,与问题相关联)允许QA流水线更快速有效地识别包含与具体查询相关内容的文档。内容还可以回答内容创建者没有考虑的可能对内容用户有用的其他问题。问题和答案可以由内容创建者验证以包含在给定文档的内容中。这些功能有助于提高QA流水线的准确性、系统性能、机器学习和置信度。内容创建者、自动化工具等注释或以其他方式生成元数据以提供可由QA流水线使用以识别内容的这些问题和答案属性的信息。

[0044] 对这样的内容进行操作,QA流水线使用多个密集分析机制生成输入问题的答案,这些分析机制评估内容以识别最可能的答案(即,对于输入问题的候选答案)。最可能的答案作为候选答案的排名列表输出,这些候选答案根据其相对分数或在评估候选答案期间计算的置信度进行排名,作为具有最高排名分数或置信度的单个最终答案,或与输入问题的最佳匹配,或排名列表和最终答案的组合。

[0045] 图1描绘了在计算机网络102中实现请求处理流水线108的认知系统100的一个说明性实施例的示意图,请求处理流水线108在一些实施例中可以是问答(QA)流水线。出于本公开的目的,假设请求处理流水线108被实现为对输入问题形式的结构化和/或非结构化请求进行操作的QA流水线。认知系统100在一个或多个计算设备104(包括一个或多个处理器和一个或多个存储器,并且可能包括本领域公知的任何其他计算设备元件,包括总线、存储设备、通信接口等)上实现连接到计算机网络102。网络102包括经由一个或多个有线和/或无线数据通信链路彼此通信并与其他设备或组件通信的多个计算设备104,其中每个通信链路包括电线、路由器、交换机、发射器、接收器中的一个或多个,或类似。认知系统100和网络102通过他们各自的计算设备110-112为一个或多个认知系统用户实现问题处理和答案生成(QA)功能。认知系统100的其他实施例可以与除了在此描述的那些之外的组件、系统、子系统和/或设备一起使用。

[0046] 认知系统100被配置为实现从各种源接收输入的QA流水线108。例如,认知系统100

从网络102、电子文档语料库106、认知系统用户(未示出)和/或其他数据和其他可能的输入源接收输入。在一个实施例中,认知系统100的一些或全部输入通过网络102路由。网络102上的各种计算设备104包括内容创建者和QA系统用户的接入点。一些计算设备104包括用于存储数据语料库106的数据库的设备(其在图1中被示为单独的实体仅用于说明目的)。数据语料库106的部分还可以提供在一个或多个其他网络附加存储设备上、一个或多个数据库中或图1中未明确显示的其他计算设备中。在各种实施例中,网络102包括本地网络连接和远程连接,使得认知系统100可以在任何规模的环境中运行,包括本地和全球(例如,互联网)。

[0047] 在一个实施例中,内容创建者在数据语料库106的文档中创建内容以用作认知系统100的数据语料库的一部分。文档包括在认知系统100中使用的任何文件、文本、文章或数据源。QA系统用户经由网络连接或到网络102的因特网连接访问认知系统100,并向认知系统输入问题100,这些问题由数据语料库106中的内容回答。在一个实施例中,问题是使用自然语言形成的。认知系统100经由QA流水线108解析和解释问题,并且向认知系统用户(例如,经由认知系统用户设备110)提供包含对问题的一个或多个答案的响应。在一些实施例中,认知系统100在候选答案的排名列表中向用户提供响应,而在其他说明性实施例中,认知系统100提供单个最终答案或最终答案与其他候选答案的排名列表的组合。

[0048] 认知系统100实现QA流水线108,包括用于处理输入问题和数据语料库106的多个阶段。QA流水线108基于对输入问题和数据语料库106的处理来生成输入问题的答案。下文将关于图3更详细地描述QA流水线108。

[0049] 在一些说明性实施例中,认知系统100可以是可从纽约阿芒克的国际商业机器公司获得的IBM**Watson®**认知系统,被以下描述的说明性实施例的机制增强。如前所述,IBM **Watson®** 认知系统的QA流水线接收输入问题,然后对其进行解析以提取问题的主要特征,然后使用这些特征制定应用于数据语料库的查询。基于对数据语料库的查询应用,通过查看数据语料库中可能包含有价值的数据的部分数据,生成一组假设或输入问题的候选答案。对输入问题的回答。IBM**Watson®**认知系统的QA流水线然后使用各种推理算法对输入问题的语言以及在查询应用过程中发现的数据语料库的每个部分中使用的语言执行深入分析。

[0050] 然后根据统计模型对从各种推理算法获得的分数进行加权,该统计模型总结了IBM**Watson®**认知系统的QA流水线对于由问题推断潜在响应(即候选答案)的证据所具有的置信水平。对于每个候选答案重复该过程以生成候选答案的排序列表,然后将其呈现给提交输入问题的用户,或者从中选择最终答案并呈现给用户。例如,可以从IBM Corporation网站、IBM Redbooks等获得关于IBM**Watson®**认知系统的QA流水线的更多信息。例如,有关IBM**Watson®**认知系统的QA流水线的信息可以在Yuan等人的“Watson and Healthcare”,IBM developerWorks,2011和“The Era of Cognitive Systems:An Inside Look at IBM Watson and How It Works”,Rob High,IBM Redbooks,2012。

[0051] 如上所述,虽然从客户端设备到认知系统100的输入可以以自然语言问题的形式提出,但是说明性实施例不限于此。相反,输入问题实际上可以被格式化或结构化为可以使用结构化和/或非结构化输入分析来解析和分析的任何合适类型的请求,包括但不限于认

知的自然语言解析和分析机制。IBMWatson®等系统，以确定执行认知分析的基础并提供认知分析的结果。在基于医疗保健的认知系统的情况下，该分析可能涉及处理患者EMR、来自一个或多个语料库的医疗指导文档等，以提供面向医疗保健的认知系统结果。

[0052] 在本公开的上下文中，认知系统100可以提供用于协助基于医疗保健的操作的认知功能。例如，根据特定的实施，基于医疗保健的操作可以包括患者诊断、医疗建议系统、医疗实践管理系统、个人患者护理计划生成和监控、用于各种目的的患者EMR评估，例如用于识别适合进行医学试验或特定类型的医学治疗等的患者等。因此，认知系统100可以是在医疗或医疗保健类型域中操作的医疗保健认知系统100，并且可以通过请求处理流水线108输入作为结构化或非结构化请求、自然语言输入问题、或类似。

[0053] 在一个说明性实施例中，认知系统100是医疗建议系统，分析信息语料库中的与医学指南和其他医学文档相关的患者的EMR以生成关于如何治疗患者的医学疾病或病症的医学治疗建议。在其他说明性实施例中，该领域可以是司法领域，认知系统100提供关于法律案例和法律文本的假设和事实陈述的认知分析。例如，认知系统100可以基于区分受害者、证人或被告的记录、陈述等中的假设来提供建议。例如，可以使用说明性实施例的机制进行分析声明“受害者的手机在车里，我们相信受害者将她的手机放在车里”，以区分受害者的手机在车里的事实与受害者他/她自己实际上将手机放在车里的假设。然后可以基于区分事实部分与假设部分来执行建议或其他认知或算法操作。

[0054] 如图1所示，并且再次参考医疗建议认知系统实现，根据说明性实施例的机制，认知系统100被进一步增强以包括在专用硬件中实现的逻辑、在硬件上执行的软件、或在硬件上执行的专用硬件和软件的任何组合，用于实现医学文本摄取引擎120，这可以例如使用服务器104来实现。

医学文本摄取引擎120本身实现解析树引擎122、假设跨度分析器124和医学文本注释器126。此外，假设跨度分析器124具有关联的假设字典数据结构127和事实字典数据结构128，假设跨度分析器124利用它们来识别解析树内的假设和事实跨度，如下文所述。

[0055] 医学文本摄取引擎120可以对存在于语料库130中的任何医学文本内容进行操作，并且对这个医学文本进行操作以注释医学文本作为摄取操作的一部分。摄取操作生成医学文本的内存表示以供认知系统100在执行其认知操作时使用，例如利用流水线108的基于医疗保健的认知操作。这些医学文本可能包括医学指南文件、医学立场文件、健康保险指南或任何其他可能存在事实和/或假设陈述的医学信息。在一些说明性实施例中，语料库130中的医学文本可以包括具有存储在其中的一个或多个患者的患者EMR的患者登记册。这些患者EMR可以包括从患者的各种不同医疗信息来源获得的信息，包括医生生成的EMR、机构生成的EMR(例如来自医疗实践、医院、紧急护理机构等)、药房生成的记录、医疗实验室记录等。该信息可以一起编译成患者的EMR或患者的EMR集。或者，该信息可以单独存储在与患者标识符相关联的单独数据结构中。

[0056] 如上所述，医学文本可以包括内容的事实部分和假设部分。医学文本摄取引擎120操作以从语料库130检索这样的医学文本，例如响应于接收到的请求或作为在接收特定请求之前发生的一般摄取操作的一部分。例如，认知系统100可以接收为指定患者生成医疗建议的请求。作为响应，认知系统100可以请求医学文本摄取引擎120从语料库130摄取指定患者的EMR。或者，可以摄取语料库130的患者登记册中的多个患者的多个EMR作为一部分医学

文本摄取引擎120的初始化或周期性过程。在任一情况下，医学文本摄取引擎120对患者EMR的医学文本或视情况而定的其他医学文本进行操作，以区分医学文本中内容的假设部分(假设陈述或短语)和事实部分的内容。通过向与医学文本相关联的元数据添加注释来相应地注释医学文本。注释的医学文本可以作为医学文本的内存表示提供给认知系统100，认知系统100可以在其上执行其认知操作。

[0057] 为了生成带注释的医学文本，医学文本摄取引擎120从语料库130接收或检索医学文本。医学文本然后由解析树引擎122使用逻辑解析技术解析以生成解析树。不管解析树引擎122使用的特定解析技术如何，由解析树引擎122基于对医学文本的分析生成的结果解析树数据结构提供医学文本中部分文本内容的结构表示(例如医学文本中的句子)。解析树提供了部分文本内容(例如，句子)的分层可视化，从而能够推断标记(即对应于解析树节点的单词或短语)之间的关系。

[0058] 假设跨度分析器124实现了一种混合技术，用于在分析树数据结构中搜索匹配在假设字典数据结构127(忽略触发器)和事实字典数据结构128(确认触发器)中指定的忽略触发器或确认触发器的标记。假设字典数据结构127指定指示假设语句或语句的假设部分的那些术语和短语。事实字典数据结构128指定那些指示事实陈述或陈述的一部分的术语和短语。同样，假设是对实际未发生的事物的指示，例如动作、事件、状态或条件的指定，或实际上并未实际发生的其他潜在事件。另一方面，事实是实际发生的事情(即事件、动作、状态或条件的指定，或实际发生的其他类型的事件)。在医学文本的上下文中，假设通常与未来计划或与患者治疗相关的潜在条件/结果相关联，这些条件/结果可能会或可能不会发生。另一方面，事实与患者当前或过去的状况、对患者执行的当前或过去的程序、以及实际发生的其他患者状况或状态信息和事件信息相关联。

[0059] 例如，假设字典数据结构127可以包括将术语“discussed(讨论)”识别为忽略触发器的条目。也就是说，在本示例的上下文中，已确定术语“discussed(讨论)”在医学文本(例如患者的EMR)中使用时表示潜在的未来事件，因为它通常指的是医生与患者讨论可能的治疗方法或患者实际尚未发生的可能情况或状态(例如，“I discussed performing a nipple-sparing mastectomy with the patient(我与患者讨论过进行保留乳头的乳房切除)”)。因此，术语“discussed(讨论)”的实例是忽略与术语“discussed(讨论)”相关联的医学文本部分的触发器。应当理解，一大组忽略触发术语和短语可以被标识为假设的指示，例如“recommended(推荐)”、“advised(建议)”和“planned(计划)”等，并且可以包括在假设词典数据中结构127。

[0060] 类似地，事实词典数据结构128可以包括将术语“revealed(揭示)”识别为确认触发器的条目。也就是说，在本示例的上下文中，已确定术语“revealed(揭示)”在医学文本(例如患者的EMR)中使用时表示已发生的患者的实际事件、状态或状况(例如，“Results of the biopsy revealed that the tumor was malignant(活检结果揭示肿瘤是恶性的)”)。因此，术语“revealed(揭示)”的实例是用于确认医学文本的部分与事实陈述或陈述的事实部分相关联的触发器。应当理解，一大组确认触发术语和短语可以被识别为指示事实陈述或陈述的一部分，例如“results(结果)”、“resulted(导致)”、“the patent has(患者有)”，并且可以包含在事实字典数据结构128中。

[0061] 假设跨度分析器124使用假设字典数据结构127和事实字典数据结构128来搜索由

解析树引擎122生成的解析树数据结构，以识别与节点相关联的令牌的解析树数据结构内的实例匹配忽略触发器或确认触发器。在解析树数据结构中搜索两组触发器，然后基于解析树和匹配节点识别相应的文本跨度。跨度被标识为匹配特定触发器的节点的子树。因此，假设跨度是对应于匹配忽略触发器的节点的解析树数据结构的子树部分。事实跨度是与匹配确认触发器的节点对应的解析树数据结构的子树部分。可以在假设跨度内发现事实跨度的情况可能是这样的，在这种情况下，事实跨度从假设跨度中移除并且被认为与确认触发器相关联并且因此指向文本的事实部分。下面将更详细地描述由假设跨度分析器124执行的操作。

[0062] 假设跨度分析器124识别由解析树引擎122生成的解析树数据结构内的假设和事实跨度，并将该信息提供给医学文本注释器126。医学文本注释器126处理假设跨度并基于解析的医学文本的子树创建注释(元数据)，该子树指示医学文本的哪些部分与假设陈述或陈述的假设部分相关联，以及陈述的哪些部分相关联。医学文本与事实陈述或陈述的事实部分相关联。医学文本注释器126基于在假设跨度中找到的元组以及与它们在解析树模式中使用的比较来执行触发术语的名词-动词消歧。换言之，假设跨度分析器124的输出被医学文本注释器126使用以找到处理假设跨度内的注释的方法(例如，忽略与假设跨度关联的所有注释，将与假设跨度关联的注释转换为其他注释等)。除了由对医学文本操作的其他注释器生成的其他注释之外，还可以提供这些注释，并且可以将这些注释存储在与医学文本相关联的元数据中。该元数据可以存储为单独但相关联的数据结构，或者可以存储为包含医学文本内容的数据结构的一部分(例如，作为患者EMR数据结构的一部分)。应当理解，一旦对患者的EMR数据结构的一部分执行此操作，则无需再次执行该操作，因为元数据具体标识了EMR数据结构的哪些部分是假设的，哪些不是。然而，在新内容已被添加到患者EMR、执行对词典127-128的修改等的情况下，说明性实施例的机制可再次对患者EMR进行操作。

[0063] 可以将得到的带注释的医学文本数据结构提供给认知系统100，以用于对医学文本执行认知操作。在一些说明性实施例中，这些认知操作利用假设/事实注释来确定作为认知操作的一部分对医学文本的每个部分加权多少。例如，在一些说明性实施例中，与医学文本的元数据中的假设注释相关联的医学文本部分可以通过将零权重因子与医学文本的这些部分相关联而基本上被忽略，而与事实注释相关联的医学文本部分被赋予预定义的权重，该权重可以根据特定实现由医学文本的其他方面的其他权重进行修改。在一些说明性实施例中，元数据本身可以包括医学文本的修剪解析树表示，其中修剪解析树对应于原始解析树，但是对应于假设的文本跨度的子树已从解析树中删除或修剪，从而导致认知系统在执行其认知操作时忽略医学文本的那些部分。

[0064] 在一个说明性实施例中，认知系统100执行的认知操作是医疗建议认知操作，其忽略与假设注释相关联的医学文本部分并且治疗建议仅基于医学文本中与事实注释相关的部分或与假设注释没有明确关联的部分(例如，医学文本的其他部分与假设注释或事实注释无关，因此是不确定的)。

[0065] 应当理解，虽然在所描绘的实施例中示出了假设和事实字典数据结构127-128，但是说明性实施例不要求存在两种类型的数据结构以执行它们的操作。相反，在一些说明性实施例中，可以仅利用假设字典数据结构127，使得与忽略触发器不匹配的解析树的任何部分或者作为与在假设字典数据结构127中阐述的与忽略触发器匹配的节点相关联的子树的

一部分被认为与内容的事实部分相关联。因此，在该实施例中，仅执行对忽略触发器的搜索，解析树中的任何其他内容都被认为是事实。

[0066] 因此，说明性实施例提供用于区分文本陈述的假设部分和文本陈述的事实部分的机制。基于这种区别，适当的注释被应用于文本陈述的部分，然后可以用来修改基于文本执行的认知操作。特别地，与文本陈述的事实部分相比，文本陈述的假设部分可能被给予相对较少的权重或考虑，并且在对文本执行认知操作时在某些情况下可能完全被忽略。

[0067] 如上所述，本公开可以提供对认知系统操作方式的特定改进。这种认知系统在一个或多个数据处理系统或计算设备上实现。图2是其中实现了说明性实施例的方面的示例数据处理系统的框图。数据处理系统200是诸如图1中的服务器104或客户端110之类的计算机的示例，实现本公开的说明性实施例过程的计算机可用代码或指令位于其中。在一个说明性实施例中，图2表示服务器计算设备，例如服务器104，实现认知系统和QA系统流水线(例如图1所示的认知系统100和QA系统流水线108)，该流水线被增强以包括下文描述的说明性实施例的附加机制。

[0068] 在所描绘的示例中，数据处理系统200采用包括北桥和存储器控制器中枢(NB/MCH)202以及南桥和输入/输出(I/O)控制器中枢(SB/ICH)204的中枢架构。处理单元206、主存储器208和图形处理器210连接到NB/MCH 202。图形处理器210通过加速图形端口(AGP)连接到NB/MCH 202。

[0069] 在所描绘的示例中，局域网(LAN)适配器212连接到SB/ICH 204。音频适配器216、键盘和鼠标适配器220、调制解调器222、只读存储器(ROM)224、硬盘驱动器(HDD)226、CD-ROM驱动器230、通用串行总线(USB)端口和其他通信端口232，以及PCI/PCIe设备234通过总线238和总线240连接到SB/ICH 204。PCI/PCIe设备可以例如包括，用于笔记本电脑的以太网适配器、附加卡和PC卡。PCI使用卡总线控制器，而PCIe不使用。ROM 224可以是例如闪存基本输入/输出系统(BIOS)。HDD 226和CD-ROM驱动器230通过总线240连接到SB/ICH 204。HDD 226和CD-ROM驱动器230可以使用例如集成驱动电子设备(IDE)或串行高级技术附件(SATA)接口。超级I/O(SIO)设备236连接到SB/ICH 204。

[0070] 操作系统在处理单元206上运行。操作系统协调并提供对图2中的数据处理系统200内的各种组件的控制。作为客户端，操作系统是可商购的操作系统，例如Microsoft **Windows®**。面向对象的编程系统，例如Java™编程系统，可以与操作系统一起运行，并提供从Java™程序或在数据处理系统200上执行的应用对操作系统的调用。

[0071] 作为服务器，数据处理系统200可以是例如**IBM®** eServer™System**p®**计算机系统，运行Advanced Interactive Executive**(AIX®)**操作系统或**LINUX®**操作系统。数据处理系统200可以是在处理单元206中包括多个处理器的对称多处理器(SMP)系统。或者，可以采用单处理器系统。

[0072] 用于操作系统、面向对象的编程系统和应用或程序的指令位于诸如HDD 226之类的存储设备上，并且被加载到主存储器208中以供处理单元206执行。本发明说明性实施例的过程由处理单元206使用计算机可用程序代码来执行，该程序代码位于诸如主存储器208、ROM 224之类的存储器中，或者例如位于一个或多个外围设备226和230中。

[0073] 诸如图2中所示的总线238或总线240之类的总线系统由一个或多个总线组成。当

然，可以使用任何类型的通信结构或架构来实现总线系统，这些通信结构或架构提供在附接到该结构或架构的不同组件或设备之间的数据传输。诸如调制解调器222或网络适配器212之类的通信单元包括一个或多个用于发送和接收数据的设备。存储器可以是例如主存储器208、ROM 224或诸如在NB/MCH 202中找到的高速缓存。

[0074] 本领域的普通技术人员将理解，图1和图2中描绘的硬件可以根据实施方式而变化。除了或代替图1和图2所示的硬件，还可以使用其他内部硬件或外围设备，例如闪存、等效的非易失性存储器或光盘驱动器等。此外，在不脱离本发明的精神和范围的情况下，说明性实施例的过程可以应用于除前面提到的SMP系统之外的多处理器数据处理系统。

[0075] 此外，数据处理系统200可以采用多种不同数据处理系统中的任一种的形式，包括客户端计算设备、服务器计算设备、平板计算机、膝上型计算机、电话或其他通信设备、个人数字助理(PDA)等。在一些说明性示例中，数据处理系统200可以是配置有闪存以提供用于存储例如操作系统文件和/或用户生成的数据的非易失性存储器的便携式计算设备。本质上，数据处理系统200可以是任何已知的或以后开发的数据处理系统，而没有架构限制。

[0076] 图3是图示根据一个说明性实施例的医疗保健认知系统的元素的交互的示例图。图3的示例图描绘了被配置为为患者提供医疗建议的医疗保健认知系统300的实现。然而，应当理解，这仅是示例实施方式，并且可以在医疗保健认知系统300的其他实施方式中实施其他医疗保健操作。

[0077] 此外，应当理解，虽然图3将患者302和用户306描绘为人物，但是可以使用计算设备、医疗设备和/或类似物来执行与这些实体的交互以及这些实体之间的交互，使得实体302和306实际上可以是计算设备(例如，客户端计算设备)。例如，患者302和用户306之间的交互304、314、316和330可以口头进行(例如，医生会见患者)并且可以涉及使用一种或多种医疗仪器、监测设备或诸如此类，以收集可以作为患者属性318输入到保健认知系统300的信息。用户306和医疗保健认知系统300之间的交互将通过用户计算设备(未示出)电子化，例如图1中的客户端计算设备110或112，通过一个或多个数据通信和潜在的一个或多个数据网络与医疗保健认知系统300通信链接。

[0078] 如图3所示，根据一个说明性实施例，患者302向用户306(例如医疗保健从业者、技术员等)呈现医学疾病或病症的症状304。用户306可以通过问题314和响应316交换与患者302交互，其中用户收集关于患者302、症状304和患者302的医学疾病或状况的更多信息。应当理解，问题/响应实际上也可以代表用户306使用各种医疗设备从患者302收集信息(例如，血压监测器、温度计、与患者相关联的可穿戴健康和活动监测设备，例如FitBit®、可穿戴心脏监测器或可以监测患者302的一个或多个医疗特征的任何其他医疗设备)。在某些情况下，此类医疗设备可以是医院或医疗中心通常使用的医疗设备，以监测出现在医院病床上以进行观察或医疗的患者的生命体征和医疗状况。

[0079] 作为响应，用户302向医疗保健认知系统300提交请求308，例如经由客户端计算设备上的用户界面，该用户界面被配置为允许用户以医疗保健认知系统300可以解析和处理的格式向医疗保健认知系统300提交请求。请求308可以包括或伴随有识别患者属性318的信息。这些患者属性318可以包括例如患者302的标识符，从该标识符可以检索患者的患者EMR 322、关于患者的人口统计信息、症状304以及从对问题的响应316获得的其他相关信息314或从用于监测或收集关于患者302状况的数据的医疗设备获得的信息。可能与医疗保健

认知系统300对患者的认知评估相关的关于患者302的任何信息可以包括在请求308和/或患者属性318中。

[0080] 医疗保健认知系统300提供了一种认知系统，该认知系统被具体配置为执行特定实施方式的面向医疗保健的认知操作。在所描绘的示例中，该面向医疗保健的认知操作旨在向用户306提供治疗建议328，以帮助用户306基于他们报告的症状304和通过问题314和响应316处理和/或医疗设备监控/数据收集收集的关于患者302的其他信息来治疗患者302。医疗保健认知系统300利用从医学语料库和其他源数据326、治疗指导数据324和与患者302相关联的患者EMR 322收集的信息对请求308和患者属性318进行操作以生成一个或多个治疗建议328。治疗建议328可以以与从患者属性318和数据源322-326获得的相关联的支持证据的排序顺序呈现，指示关于为什么提供治疗建议328以及为什么以它排序的方式排序的推理。

[0081] 例如，基于请求308和患者属性318，医疗保健认知系统300可以对请求进行操作，例如通过使用这里描述的QA流水线类型处理来解析请求308和患者属性318以确定正在请求什么以及根据患者识别生成请求的标准属性318，并且可以执行用于生成查询的各种操作，这些查询被发送到数据源322-326以检索数据、生成候选治疗建议(或输入问题的答案)，并基于在数据源中找到的支持证据对这些候选治疗建议进行评分322-326。在所描绘的示例中，患者EMR 322是从各种来源(例如，医院、实验室、医生办公室、健康保险公司、药房等)收集患者数据的患者信息库。患者EMR 322以信息可由医疗保健认知系统300检索和处理的方式(结构化、非结构化或结构化和非结构化格式的混合)存储关于个体患者(例如患者302)的各种信息。患者信息可能包括关于患者的各种人口统计信息、关于患者的个人联系信息、就业信息、健康保险信息、实验室报告、医生就诊报告、医院图表、关于先前诊断、症状、治疗、处方信息等的历史信息.基于患者302的标识符，来自该患者储存库的患者对应EMR 322可由医疗保健认知系统300检索并搜索/处理以生成治疗建议328。

[0082] 治疗指导数据324提供医学知识的知识库，用于基于患者的属性318和患者的EMR 322中呈现的历史信息来识别患者的潜在治疗。治疗指导数据324可以从医疗机构(例如，美国医学协会)发布的官方治疗指南和政策中获得，可以从广泛接受的医师医学和参考文本(例如，医师的案头参考)、保险公司指南或类似中获得。治疗指导数据324可以以可由医疗保健认知系统300摄取的任何合适的形式提供，包括结构化和非结构化格式。

[0083] 在一些情况下，这样的治疗指导数据324可以以规则的形式提供，这些规则指示需要存在和/或不需要存在的标准，以便相应的治疗适用于特定患者治疗特定症状或医学疾病/病症。例如，治疗指导数据324可以包括治疗建议规则，指示对于地西他滨的治疗，使用这种治疗的严格标准是患者302小于或等于60岁，患有急性髓系白血病(AML)，并且没有心脏病的证据。因此，对于年龄为59岁、患有AML并且在其患者属性318或患者EMR中没有任何证据表明心脏病证据的患者302，存在以下治疗规则条件：

年龄<＝60岁＝59(MET)；

患者患有AML＝AML(MET)；和

心脏病＝假(MET)

[0084] 由于关于该患者302的特定信息满足治疗规则的所有标准，所以地西他滨的治疗是该患者302考虑的候选治疗。然而，如果患者已经69岁，则不会满足第一个标准，并且地西

他滨治疗将不是该患者302考虑的候选治疗。医疗保健认知系统300可以基于摄取的治疗指导数据324评估各种潜在的治疗建议，以通过基于从患者EMR 322以及医学语料库和其他源数据326获得的证据数据对这些候选治疗进行评分来识别候选治疗的子集以供医疗保健认知系统300进一步考虑。

[0085] 例如，可以采用数据挖掘过程来挖掘源322和326中的数据以识别支持和/或反驳候选治疗对由患者的患者属性318和EMR 322所表征的特定患者302的适用性的证据数据。例如，对于处理规则的每个标准，数据挖掘的结果提供了一组证据，支持在标准为“满足(MET)”的情况下和在标准为“不满足(NOT MET)”的情况下进行处理。医疗保健认知系统300根据各种认知逻辑算法处理证据以生成每个候选治疗建议的置信度分数，指示相应候选治疗建议对患者302有效的置信度。然后可以根据它们的置信度对候选治疗建议进行排名，并作为治疗建议的排名列表328呈现给用户306。在某些情况下，仅返回最高排名或最终答案作为治疗建议328。治疗建议328可以以可由医疗保健认知系统300评估的潜在证据可访问的方式呈现给用户306，例如通过从上到下的界面，以便用户306可以识别医疗保健认知系统300提供治疗建议328的原因。

[0086] 根据这里的说明性实施例，医疗保健认知系统300被扩充以包括医学文本摄取引擎340，例如，可以是图1中的医学文本摄取引擎120。医学文本摄取引擎340对数据322-326的一个或多个语料库进行操作以摄取该一个或多个语料库322-326以生成医疗保健认知系统300可使用的医学文本的内存表示以执行其认知操作。摄取操作包括分析医学文本以识别医学文本的各种特征，例如医学文本中使用的各种术语和短语的词性、指示医学文本中概念实例的本体相关性以及医学文本的其他注释以生成元数据注释，医疗保健认知系统300可以使用该注释来执行其认知操作。语料库322-326的其他适当处理，如关于认知系统摄取机制一般已知的，也可以作为摄取操作的一部分来实现。

[0087] 根据说明性实施例，医学文本摄取引擎340被扩充以包括用于执行分析以区分一个或多个语料库322-326的医学文本中的文本的假设部分和文本的事实部分的逻辑。在一个说明性实施例中，医学文本摄取引擎340分析患者EMR 322以区分和注释文本的假设部分和文本的事实部分。得到的带注释的医学文本然后可以被医疗保健认知系统300用来执行认知操作，例如医疗建议、给予文本的假设部分和事实部分适当的权重(例如，假设部分的权重为零，文本的事实部分的权重大于零)。

[0088] 例如，医学文本摄取引擎340可以从患者EMR语料库322中检索患者EMR 323，该语料库可以是患者登记册等。患者EMR 323的文本内容然后可由解析树引擎342分析以生成表示文本内容的解析树数据结构。解析树数据结构包括表示文本中标记的节点，其中标记是术语或短语，以及连接表示节点之间关系的节点的边。此外，一些节点可能表示文本部分之间的逻辑关系(例如，AND、OR、ANDNOT等)。节点可以具有相关联的属性，包括词性属性，当确定节点是对应于忽略触发器还是确认触发器时，这些属性可以用于辅助分析，如下文所讨论的。

[0089] 虽然图3描绘了患者302和用户306之间的交互，用户306可以是医疗保健从业者，例如医生、护士、医生助理、实验室技术员或任何其他医疗保健工作者，说明性实施例不需要这样。相反，患者302可以直接与医疗保健认知系统300交互而不必经历与用户306的交互，并且用户306可以与医疗保健认知系统300交互而不必与患者302交互。例如，在第一种

情况下,患者302可以直接基于由患者302提供给医疗保健认知系统300的症状304从医疗保健认知系统300请求308治疗建议328。此外,医疗保健认知系统300实际上可以具有用于自动向患者302提出问题314并从患者302接收响应316以协助数据收集以生成治疗建议328的逻辑。在后一种情况下,用户306可以通过连同患者属性318发送请求308并响应于来自医疗保健认知系统300的治疗建议,仅基于先前收集并呈现在患者EMR 322中的信息进行操作。因此,图3中的描述仅是示例,并且在不脱离本公开的精神和范围的情况下可以进行许多修改时,不应被解释为需要所描述的特定交互。

[0090] 因此,说明性实施例提供了用于分析诸如医学文本之类的文档的自然语言内容的机制,以识别引用假设事件、状态、条件等的文本部分并将这些假设与参考实际情况的文本部分区分开来。为文本的各个部分提供相应的注释以基于这种分析的结果将它们识别为假设的或事实的,然后将这些注释提供给认知系统以在执行其认知操作时使用。

[0091] 如上所述,在一些说明性实施例中,这些认知操作可以包括执行机器学习的机器学习模型,例如用于确定适当医疗建议的机器学习。例如,作为由机器学习模型执行的机器学习操作的一部分,可以从语料库的患者登记册检索多个患者的患者EMR,并用于绘制患者属性和相应的规定治疗之间的相关性。例如,可以在患者EMR中识别各种医学疾病、患者属性(例如,年龄、性别、身高、体重、特定实验室结果等)以及医务人员开出的相应治疗方法,并用于生成机器学习医疗建议模型。这种机器学习可以将这些医学疾病、患者属性和处方治疗相关联,识别语料库或语料库中的其他确证证据,包括其他医学文本,例如指南、立场文件等,并产生对治疗建议相关性的置信度.

[0092] 例如,图4是由医学专业人员撰写的示例笔记的示例解析树数据结构表示,可以是患者的EMR的一部分。在所描绘的示例中,解析树用于声明,“We discussed the fact that the chemotherapy would most likely put her into menopause and not allow her to have more children(我们讨论了这样一个事实,即化疗很可能会使她进入更年期,并且不允许她生育更多孩子)”。

[0093] 将解析树数据结构提供给假设跨度分析器344,其分析解析树数据结构的每个节点以识别匹配由假设字典数据结构347指定的忽略触发器并确认由事实字典数据结构348指定的触发器的节点。例如,假设跨度分析器344可以接收医学文本的每个句子的解析树数据结构,或者根据特定实现,接收来自医学文本摄取引擎340检索的医学文本的任何大小的文本部分的解析树。对于解析树数据结构中的每个节点,确定该节点的标记是否对应于假设字典数据结构347中指定的忽略触发器。如果是,则将该节点的词性属性与忽略触发器的词性属性进行比较,以确定该词性中是否存在匹配,该匹配是动词词性。如果节点的词性属性是动词,并且节点的父节点的词性是动词,则选择该节点的子树作为忽略子树,该节点的父节点是忽略子树。

[0094] 对父节点的词性标签进行检查以确定该句子是被动句还是主动句,例如包含“was recommended(被建议)”的句子表示被动句。如果触发器是“recommended(建议)”并且“recommended(建议)”被解析树识别为动词,并且其父节点是“was”,则假设子树从“was”而不是“recommended”开始。例如,这是为了捕获诸如“were discussed(被讨论)”之类的短语,其中“discussed(讨论)”是所识别的节点,而“were”是所识别节点的父节点。如果节点和父节点不是动词,则选择节点的子树,该节点是忽略子树的根。

[0095] 动词作为该过程的目标的原因是一些术语或短语可以用作多个词性(例如,名词和动词)。然而,在一些实施方式中,假设的触发术语或短语更常被用作动词,因此,作为动词的触发术语的识别可能指示文本的假设跨度。应当理解,其他实现可以对词性进行更复杂的分析并且可以不依赖于节点令牌和忽略触发器的词性是否是动词。

[0096] 对于忽略子树的每个节点,确定该节点是否对应于确认触发器。如果忽略子树的节点与确认触发器匹配,则选择该节点的子树并且从忽略子树中移除该确认子树。删除了任何确认子树的结果忽略子树返回用于带有忽略注释或假设注释的注释,而确认子树返回用于确认或事实注释。不对应于忽略子树的解析树数据结构的树或子树也可以用确认注释或事实注释进行注释,或者可以不以其他方式不注释确认/忽略注释,这取决于具体实现。

[0097] 如果确定与忽略触发器匹配的节点的令牌的词性是名词而不是动词,则可以对该节点的令牌对应的其他自然语言资源进行附加分析以生成关于节点的令牌是否可能表示假设的置信度得分。例如,可以分析来自字典数据结构的定义信息,该定义信息指示令牌的各种用途的词性和各种用途的时态信息、n-gram等,以生成标记表示假设的文本跨度,因此匹配忽略触发器。执行此分析是因为根据在文本中使用标记的方式,相同的标记可能代表忽略触发器和确认触发器。例如,请考虑以下句子中的“considering(考虑)”一词:

(1)“The patient has been strongly considering a prophylactic mastectomy on the right breast for ultimate risk reduction(患者一直强烈考虑对右乳房进行预防性乳房切除术,以最终降低风险)。”

(2)“The patient has been advised considering the prophylactic mastectomy on the right breast for ultimate risk reduction(已建议患者考虑对右乳房进行预防性乳房切除术,以最终降低风险)。”

[0098] 在上面的句子(1)中,术语“considering(考虑)”是一个忽略触发因素,因为它描述了患者接受预防性乳房切除术的假设未来可能性。在上面的句子(2)中,术语“considering(考虑)”是确认触发,因为该术语指的是发生的实际事件(即医疗专业人员建议患者进行预防性乳房切除术)。在句子(2)中,根据与令牌相关的词性和时态信息以及词典中的词性和时态信息进行名词-动词消歧,以确定令牌“considering(考虑)”的实例是忽略触发器还是确认触发器。

[0099] 用于消除这两个句子歧义的n-gram将是不同的:<名词><副词>考虑(considering)<名词过程>和<名词><动词>considering(考虑)<名词过程>。由于第一个句子与训练集中的元组匹配,因此句子(1)将被识别为假设,而句子(2)则不会。

[0100] 返回到图4,对应于图4中所示的解析树400的句子图示了具有忽略触发器和对应的不包括嵌入式确认子树的忽略子树的句子的简单示例。如图4所示,具有“discussed(讨论)”的令牌的节点402与假设字典数据结构347中的相应忽略触发器相匹配。将该节点402作为包括节点402的子节点的忽略子树的根节点,假设跨度分析器344在忽略子树中搜索作为“discussed(讨论)”节点402的兄弟节点或子节点的任何确认触发器匹配,但在这个例子中没有。结果,以“discussed(讨论)”节点402为根的整个树400被选择为忽略子树并且被医学文本注释器346标记为具有忽略或假设注释的注释。

[0101] 带注释的忽略子树400然后可以由医疗保健认知系统300处理以执行具有给予忽略子树400的适当权重的认知操作。在一些说明性实施例中,该加权涉及在执行相应认知操

作时忽略忽略子树400。在一些说明性实施例中，该认知操作是由关于医疗建议使用的医疗保健认知系统的机器学习模型执行的机器学习操作。在一些说明性实施例中，该认知操作是向用户请求(例如图3中的用户请求308)提供医疗建议的操作。在其他说明性实施例中，可能受归因于假设文本跨度的有效性、信任或置信度影响的其他认知操作可以基于由说明性实施例的机制生成的假设(或忽略)注释和事实(或确认)注释进行操作。

[0102] 图5是根据一个说明性实施例的其中执行对应于忽略触发器的节点的消歧的句子的另一解析树数据结构的示例。如图5所示，解析树500对应于以下语句：“Undergoing a nipple-sparing mastectomy results in an insensate nipple with an up to 15% risk of partial nipple necrosis(进行保留乳头的乳房切除术会导致无知觉的乳头，乳头部分坏死的风险高达15%)。”当人们查看这句话的解析树500时，可以看到术语“results in(导致)”是一个短语，它捕获了所有可能是假设跨度的标记，并且具有足够的泛化性，不会导致其他任何错误的注释情况。

[0103] 将解析树500中每个节点的每个标记与假设字典数据结构347中的忽略触发器进行比较，节点502被正确识别为匹配忽略触发器，但这个例子中该标记与“名词(noun)”词性相关联。因此，基于字典信息、时态信息、n-gram、本体论信息等执行对应于节点502的令牌的消歧。消歧尝试将节点502的标记的特征匹配到句子的其他部分(即节点502的子树的其他部分以消歧标记的语言使用)。例如，可以将标记的术语的定义与句子的其他部分进行比较以确定它是否与句子的其他词性相匹配。

[0104] 例如，取节点502的子树，对应的句子是“A nipple-sparing mastectomy results in an insensate nipple(保留乳头的乳房切除术导致无知觉的乳头)”。元组或n-gram的相应数据集以及包含医学调整本体的相应元组说明了上述句子的词性模式如下：

<名词><动词><名词>(这是一个直句解析元组)

医疗调整的本体元组是：

<名词-过程><动词><名词-身体-部分>(这是为域调整的句子解析元组)

[0105] 元组是从训练集获得的。上述元组中的<名词-过程>匹配句子中的“nipple-sparing mastectomy(保留乳头的乳房切除术)”，<名词-身体-部分>匹配“insensate nipple(无知觉的乳头)”，从元组的数据集中，预计触发器是一个动词不是名词(正如XSG所标识的那样)。由于句子与元组匹配，因此推断触发器确实必须是动词而不是名词，并且可以将其识别为假设语句。

[0106] 可以在该示例中用于消除节点502的标记的歧义的术语“results(结果)”的字典定义如下：

(1)由于行为、环境、场所等而产生、出现或进行；成为结果。

(2)以特定的方式或事物终止或结束。

[0107] 通过分析该信息，可以确定节点502的标记“results(结果)”被用作句子中的动词，因此，很可能是引用假设的文本跨度的忽略触发器。因此，节点502的子树将被识别为忽略子树并且可以进一步进行分析如上所述关于确认触发器的分析。也就是说，一旦识别出词性，就为术语解析定义。根据句型匹配的集合，定义可以帮助确认该“触发器”确实是正确的。对于此示例，句型之一包括“名词-结果或计算”。“result(结果)”的定义包括术语“outcome(结果)”。主题专家指出的一组这些模式将有助于确认可以是不同词性的术语的

使用。

[0108] 作为另一个例子,考虑句子“A mastectomy performed had good results(进行的乳房切除术有很好的结果)”。这句话对应的元组或n-gram如下:

<名词><动词><形容词><名词>

医疗调整的本体元组是:

<名词-过程><动词-动作-过去时><名词-结果/计算>

[0109] 对该元组、字典定义、本体信息等的分析结果表明术语“result(结果)”为名词使用,使其不是忽略触发器匹配,因为它不是动词。如果不使用该元组,则术语“result(结果)”在这句话中实际上是可以被视为忽略触发器。在这个特定的句子中,匹配的元组是名词:mastectomy(乳房切除术)、动词:performed(执行)、动词-动作-过去时:had和名词-结果/计算:good results(很好的结果)。从训练集数据可知,该元组与事实而非假设相关联。因此,发现该句子与元组匹配,说明性实施例的机制将术语“results(结果)”识别为确认触发器而不是忽略触发器。

[0110] 为了识别句子是否与特定元组匹配,在一些说明性实施例中,说明性实施例的机制可以相对于句子对元组进行评分。对于每个元组模式,存在匹配元组模式的最大分数,使得当在假设跨度或包含假设跨度的自然语言内容内找到所有词性时,最大分数与元组相关联。模式中的每个匹配项都被赋予一个权重,名词和动词的权重最高,主语的权重次之。元组模式的得分是元组模式每个匹配部分的加权值的总和,当得分高于阈值时,可以确定元组模式已经充分匹配,项应该是被视为触发器(例如,假设或确认触发器取决于元组是用于确认还是假设触发器识别)。

[0111] 例如,在上述元组模式中,各个词性的权重可以如下:<名词>(2)<动词>(6)<形容词>(1)<名词过程>(3),<动词-动作过去时>(2)<名词-结果/计算>(4),最高得分为18。用于确认触发器的阈值对正确的词性具有很高的权重,因此,示例阈值分数可以确定为10,这样如果文本的一部分与元组模式的部分匹配以生成10或更大的加权分数,则将其视为触发器。如果这些元组模式中的一个以上与其对应的阈值匹配,则可以根据匹配的数量执行触发器的确认。

[0112] 应当理解,可以针对医学文本内的文本的每个部分(例如,每个句子)执行用于识别忽略子树和确认子树的上述过程,使得分析整个医学文本以识别忽略(假设)子树和确认(事实)子树。忽略子树代表假设跨度,而确认子树代表事实跨度。这些假设跨度和事实跨度可以提供给医学文本注释器(例如图3中的医学文本注释器346),其在医学文本的元数据325中生成相应的忽略(假设)注释和确认(事实)注释(例如,EMR 323)指向医学文本中相应的假设跨度和事实跨度。医学文本(例如,EMR 323)和元数据325被返回到医疗保健认知系统300以用于执行认知操作。

[0113] 在一些说明性实施例中,对应于说明性实施例的机制所标识的忽略子树的文本的假设跨度的识别和这种文本的假设跨度的注释可以在执行机器学习操作以学习医学疾病、患者属性和治疗的相关性时用于忽略假设跨度。因此,当机器学习操作遇到被注释为假设文本跨度的部分文本时,该部分文本将被忽略并且不作为机器学习操作的一部分进行处理。在一些说明性实施例中,可以确定,虽然本质上是假设的,但是假设的文本跨度仍然可以提供对医学疾病、患者属性和治疗的相关性的有效性的一些洞察,可以不忽略文本的这

些部分，取而代之的是与被确定为与事实内容相关联的文本的其他部分相比，在评估期间可能赋予它们相对较小的权重。因此，例如，当识别对相关性的证据支持时，与被识别为事实性质的文本的其他部分相比，假设的文本跨度将提供相对较少量的支持/反对相关性的证据支持。

[0114] 类似地，认知操作可以包括针对特定识别出的患者的治疗建议的实际运行时间确定，例如上面图3的上下文中所描述的。在这种情况下，当生成要返回给用户306的治疗建议328时，医疗保健认知系统300可以执行对假设文本跨度的类似考虑。也就是说，当基于语料库322-326中的其他证据信息评估患者EMR以确定适当的治疗时，可以忽略文本的假设跨度或赋予相对较小的权重，这取决于特定的实施方式。

[0115] 因此，在具有处理器和至少一个存储器的数据处理系统中提供了机制，其中，至少一个存储器具有由处理器执行并配置处理器以执行与上述说明性实施例中的一者或一者以上相对应的操作的指令。在一个说明性实施例中，这些操作包括：(1)数据处理系统接收自然语言内容；(2)由数据处理系统对自然语言内容进行分析，生成解析树，其中，解析树是自然语言内容的分层表示，包括与自然语言内容中的术语或短语对应的节点和链接节点的边；(3)由数据处理系统对解析树数据结构进行处理，以识别自然语言内容中的一个或多个假设触发器实例，其中假设触发是表示假设陈述的术语或短语；和(4)数据处理系统根据自然语言内容进行认知运算，其中认知操作是在自然语言内容的与一个或多个所识别的假设触发器实例相对应的部分被赋予比自然语言内容的其他部分相对较低的权重的情况下执行的。

[0116] 在一些说明性实施例中，这些操作进一步包括由数据处理系统移除对应于一个或多个假设触发器实例的解析树数据结构的一个或多个子树数据结构，从而生成假设修剪后的解析树数据结构，其中认知操作是基于假设修剪后的解析树数据结构进行的。在其他说明性实施例中，执行认知操作包括由数据处理系统基于对自然语言内容中的一个或多个假设触发器实例的识别来训练自然语言处理(NLP)系统的模型，以及由NLP系统根据训练好的模型对自然语言内容进行自然语言处理。

[0117] 在更进一步的说明性实施例中，处理解析树数据结构还包括，对于在解析树数据结构中找到的假设触发器的每个实例：使用字典数据结构分析假设触发器以确定假设触发器的词性属性；并且利用所确定的词性属性确定假设触发器是否对应于假设语句的度量。此外，利用所确定的词性属性确定假设触发器是否对应于假设语句的度量可以包括：生成对应于假设触发器的子树数据结构的元组表示；从字典数据结构中检索存在于假设触发器中的术语的一个或多个字典定义；以及基于子树数据结构的元组表示与一个或多个字典定义的相关性确定假设触发器的词性属性。响应于词性属性指示假设触发器是名词，确定与假设触发器对应的子树数据结构不指向假设语句。

[0118] 在其他说明性实施例中，NLP系统是医疗建议系统，并且认知操作包括基于患者电子病历的内容生成治疗建议。此外，数据处理系统可以是医疗建议系统的后端数据处理系统。

[0119] 在一些说明性实施例中，处理解析树数据结构进一步包括处理解析树数据结构以识别事实触发器的实例，其中事实触发器是指示事实陈述的术语或短语。此外，该操作可以包括确定假设子树中是否存在事实子树，以及在进一步处理修改后的假设子树之前，从假

设子树中移除事实子树以生成修改后的假设子树。

[0120] 图6A是根据说明性实施例的用于描述句子的节点604和连接边606的句子的另一解析树数据结构600的示例。图6B是沿着两次遍历的解析树600的表602的示例。图6C是向量表617中的两次遍历和它们对应的单热编码向量的比较表614的示例(尽管表617的仅一部分是可见的)。图6A-6C基于这句话,“She is aware that if a nipple sparing mastectomy was performed,there would be no sensation,stimulation,or arousal(她知道如果进行了保留乳头的乳房切除术,将不会有任何感觉、刺激或唤醒)”。在这句话中,有两个假设的触发器,一个是“if(如果)”,另一个是“would be(将)”。现在将一起讨论图6A-6C。

[0121] 在说明性实施例中,解析树600包括许多节点604,例如:if节点604A、was节点604B、would节点604C、nipple节点604D、performed节点604E、a节点604F、sparing节点604G和mastectomy节点604H。虽然所有节点604都由边606连接,但一些节点604直接连接到与其相邻的其他节点604(见图6A)。

[0122] 每个表602包括行608和列610,其中行608包含节点604的属性。每个表602代表一个遍历,它是从触发节点(例如if节点604A)到目标节点的路径。在说明性实施例中,表602A表示从if节点604A到performed节点604E的遍历,并且表602B表示从if节点604A到mastectomy节点604H的遍历。在表602A和602B的说明性实施例中,行608B和608F包括“Part of Speech(词性)”,行608C和608G包括“Distance(距离)”,行608D和608H包括“Highlighting(突出显示)”。在其他实施例中,表602可以包括节点604的其他属性,例如“slot name(槽名)”(与两个节点604之间的边606相关联的标签,例如“vadj”、“obj”、“top”、“subj”等)、“horizontal(水平)”(放射边缘606的水平方向,例如左右)和“vertical(垂直)”(放射边缘606的垂直方向,例如上下)。

[0123] 在说明性实施例中,行608C和608G指的是距目标所在的触发器的距离。例如,在表602B中,触发器是if节点604A并且距离“4”可以计算到mastectomy节点604H。然而,在一些实施例中,该距离被分成具有预定或可配置的一个或多个宽度的区间。例如,如果一个bin有两个位置宽,则节点604H的值将为“2”,如果一个bin为四个位置宽,则节点604H的值将为“1”。这样的特征可以用于识别不同的句子可能具有有效相似的节点,但是这些节点可能与触发器的距离不同。在一些实施例中,每个非距离属性包括其各自的距离以区分它出现在遍历的哪个阶段。例如,如果这种做法被用于制作树表602A,那么行608B将是:“Part of Speech(词性)”subconj_0;verb(动词)_1;verb(动词)_2”。

[0124] 树表602A表示从触发器节点604A到目标节点604E的遍历,因此列610B-610D按该顺序排列。类似地,树表602B表示从触发器节点604A到目标节点604H的遍历,因此列610F-610J按该顺序排列。此外,树表602可以转换成考虑表614。更具体地,考虑表614可以通过将每次遍历的值布置成单行来创建。更具体地,树表602A的行608A-608D在行608I中彼此相邻,并且树表602B的行608E-608H在行608J中彼此相邻。

[0125] 在说明性实施例中,向量表617包括向量618A和618B。向量618A是考虑表614中的行608I的单热编码的结果,并且向量618B是考虑表614中的行608J的单热编码的结果。向量表617被构造为好像产生行的遍历608I和608J是训练数据集中存在的唯一遍历。从而,列610表示考虑表614中存在的每个唯一值(即,属性类型和属性值的每个组合都被表示,尽管

并非所有的列610在图6C中都是可见的)。这样做时,来自考虑表614的属性值可以被转换为“1”和“0”(例如,“mastectomy”向量是[1 1 1 0 1 1 1 1 1 0…],而“performed”向量是是[1 1 0 1 0 0 1 1 0 1…])。在一些实施例中,没有如图6C中所示的向量表617,而是存在单独向量的集合和什么属性类型/值组合在哪个位置的单独映射。

[0126] 在说明性实施例中,行608E和608J指的是对应的节点604是被突出显示还是被标记。为简单起见,图6A中的“highlighted(突出显示)”节点604已用星号(*)标记。解析树600中某些节点604的突出显示已由第三方(例如由人类或其他机器/算法)执行以标记解析树600中的假设跨度616A。如下文将解释的,这是为机器学习而完成的和训练目的。虽然解析树600还包括另一个假设跨度616B,但仅跨度616A已被突出显示,因为突出显示跨度616A和616B将不允许将分析树600用于训练。这是因为每个节点604由于其与单个触发器的语法关系而在跨度616之内或之外。当一系列文本包含两个触发器时,具有一个触发器的节点将不会与另一个触发器跨接,反之亦然。相反,如果使用跨度616B进行训练是有益的,则可以制作另一个解析树(未显示),仅突出显示跨度616B。

[0127] 解析树600、表602、考虑表614以及向量618A和618B的特征允许在解析树600上执行结构化查询语言(SQL)操作。因此,执行机器学习模型的推理算法可以从训练集中编码的遍历中识别属性类型/值组合的出现。然后,可以根据NLP系统审查的组合选择最佳标签。

[0128] 虽然图6B中只有两个树表602,图6C中只有两个对应的考虑表614和向量618,但对解析树的完整分析可以包含更多的表和向量。可以为解析树600中的任何触发器和任何目标之间的每个可能的遍历制作考虑表602。因此,也可以为任何触发器和解析树600中的任何目标之间的每个可能的遍历制作向量618。

[0129] 图7是例如分别使用图4、5和6A的解析树数据结构400、500和600之一来寻找解析树的跨度的方法700的流程图。方法700开始于多边形702。在多边形704,解析树中的触发器(例如,假设触发器)被识别为起点,并且树表以触发器节点作为第二列(沿着行标签),该列由触发器节点的属性填充。在多边形706,从先前节点(在第一次迭代中,这将是触发器节点)发出的每条边被遍历到下一个相邻节点(它们现在是“目标节点”)。在多边形708,每个目标节点将其属性作为列输入到与前一节点相邻的其自己的树表中。

[0130] 在多边形710,每个树表被转换成考虑表中的单行(例如,如先前关于图6C所描述的)。在一些实施例中,可以在多边形710将多个树表转换成单个考虑表,例如,通过将每个树表转换为单行,并且在考虑表中堆叠行。树表可以来自多个并行节点(即与触发器距离相同的节点),例如图4中的“that”、“chemotherapy(化疗)”和“and”节点或图6A中的“was”和“would”节点。或者,树表可以来自多个串行节点(即沿着触发器单次遍历的节点),例如图4中的“and”、“allow”和“her”节点或图6A中的“was”和“performed”节点。

[0131] 方法700在多边形714继续,其中确定连接到先前节点的更多先前未访问节点(例如,未列表节点)的可用性。如果有更多尚未制表的节点要制表,则遍历相应的边,这些未制表的节点成为新的目标节点。然后重复多边形706-714,直到所有新的目标节点(即,连接到触发器的未列表节点)都已被列表并且没有更多节点要列表为止。例如,可以通过广度优先搜索或深度优先搜索来确定从触发器遍历解析树的顺序。

[0132] 一旦节点列表已经完成,方法700在多边形718继续,解析树分析解析树中是否存在更多触发器。如果存在一个或多个剩余触发器(例如,另一个假设触发器),则方法700返

回到多边形704以产生新的跨度。如果不是，则方法700在多边形720结束。

[0133] 方法700的特征允许系统地分析句子或短语的解析树以找到给定触发器的节点跨度，例如，通过医学文本摄取引擎120中的一个或多个模块(如图1所示)。可以以多种不同方式使用此功能。例如，寻找参与者的临床试验可以使用NLP处理器通过评估他们的病史来选择候选人，并能够将其中的事实信息与假设的受试者分开。再举一个例子，如果NLP处理器更有能力将事实信息与问题中的假设主题分开，那么NLP处理器回答自然语言问题的速度和准确性可以提高。再例如，NLP处理器可以分析句子并标记其中的触发器和跨度，以用作其他学习机器(例如NLP处理器)的训练材料。再例如，可以分析关联个人的社交网络以找到满足特定标准的属性，例如，用户的哪些关联是近亲(例如，在用户的两代内共享共同祖先)。在这样的实施例中，分析方法可以包括检查节点(例如，人)是否已经被分析以避免重复分析相同节点的循环。此外，在这样的实施例中，分析方法可以包括搜索深度限制以在从用户的直接连接跳转一定次数后停止搜索继续。这可能是因为连接链越长，另一个用户是近亲的可能性就会显着降低。

[0134] 图8是使用解析树来训练自然语言处理(NLP)系统以例如对自然语言文本进行操作的方法800的流程图。方法800可以使用例如服务器104和/或医学文本摄取引擎120(如图1所示，尽管具有不同于上面讨论的假设跨度分析器124的跨度分析器)来实现。在所示实施例中，方法800在多边形802开始，并且在多边形804，NLP训练机导入训练文本，其中一些被突出显示而一些未被突出显示，并且将它们转换成训练解析树。在多边形806，例如通过对训练分析树执行方法700，从训练分析树制作训练考虑表。在一些实施例中，当处理未突出显示的训练解析树时，根据解析树中的每个可能的遍历来制作树表，使得每个节点-目标对具有一个树表。这是因为之前未在未突出显示的分析树中识别触发器节点，尽管突出显示的分析树可能被视为相同，虽然它们标记了触发器节点。在这样的实施例中，训练考虑表还表示每个节点-目标对。

[0135] 在所示实施例中，在多边形808，训练考虑表用于训练单热编码器。一旦该训练完成，在多边形810，所有训练考虑表都被单热编码。在多边形812，通过分析训练单热编码向量，例如使用主成分分析(PCA)、广义Hebbian算法、和/或各种深度学习(即神经网络)技术，例如自动编码和/或嵌入。一旦该训练完成，在多边形814，使用投影模型处理训练的单热编码向量。更具体地说，只有源自突出显示解析树的处理训练向量被投影到多边形814。在多边形816，使用训练投影向量(即源自突出显示解析树的训练向量)训练分类器模型。具体地，例如，来自多边形810的“true(真)”数学向量可以用作正训练示例，来自多边形810的“false(假)”数学向量可以用作负训练示例。在多边形818，还可以使用分类器模型对突出显示的解析树进行分类，以便NLP学习机可以将每个节点的分类与每个节点的突出显示(或缺少)进行比较，以查看它们匹配的频率。然后可以调整分类器模型以增加分类与突出显示的匹配数量。从而，通过机器学习改进分类器模型。该训练过程在NLP学习机离线的情况下执行，并且方法800在多边形820结束。

[0136] 方法800的特征允许训练NLP学习机以在自然语言文本中寻找跨度(例如，假设跨度)。这可以使用相对较少的标记跨度和相对大量的未标记跨度来完成，与仅使用标记跨度相比，这提高了NLP学习机的有效性。然而，自然地，未标记跨度不需要标记跨度所需的所有时间和精力，但公开的NLP学习机(包括投影模型)可以使用未标记跨度来放大从标记跨度

学习的效果。此外，由于没有从初始考虑制表(即，在多边形806)剪裁节点，与已知跨度成员节点(例如，兄弟节点)不直接相关的节点仍被分析。但是由于在过程中稍后(即在多边形816)检查目标节点的突出显示和标记跨度“true(真)”或“false(假)”，NLP学习机仍然可以学习裁剪不属于跨度的节点正在创建。此外，分析是在不使用必须设计然后维护的大量复杂的黑名单和白名单行的情况下完成的。此外，在不完整的黑名单和/或白名单被提供用于裁剪的情况下，根据方法800训练的NLP学习机可以使用其根据上下文裁剪的知识来回退而不是依赖于列表。

[0137] 在一些实施例中，方法800还包括多边形822。这由从多边形810延伸到多边形822以及从多边形822延伸到多边形812的虚箭头表示。在这样的实施例中，方法800将改为通过多边形822进行从多边形810直接移动到多边形812。在多边形822，基于标记跨度的分析，例如基于卡方分析和/或信噪比分析，进行特征选择。特征选择可以指示训练编码向量的哪些列更强烈地预测正在解决的特定问题(例如，在新的未标记文本中找到假设的或否定的跨度)。但是，假设跨度的特征选择可能与否定跨度的特征选择不同。因此，可以根据正在寻找的答案进一步缩小特征选择范围(例如，寻找假设的跨度或寻找否定的跨度)。选择一个特征可能允许删除一些列(即一些文本)，因为它们可能无法预测问题的解决方案。例如，可以在训练多边形812的投影模型之前去除训练编码向量的某些列以便找到假设跨度。

[0138] 在该替代实施例中，在多边形812训练投影模型之前设置特征选择。这使得投影模型将更加适应所寻求的答案是什么。这可以提高获得结果的准确性。但是，如果所需的答案发生变化，则可能需要重新训练投影模型以更好地适应新的广受欢迎的答案。如果用于学习的数据集相对较小，那么再训练可能不是一项大工程。但是如果学习数据集很大，那么不包括多边形822的更通用的学习方法可能是有益的，因为它不需要重新训练(尽管它的预测可能不那么准确)。

[0139] 图9是对开始于多边形902的自然语言文本进行操作的NLP学习机的方法900的流程图。在多边形904，NLP学习机导入一系列新的自然语言文本(例如，句子)，其转换为新的解析树。在多边形906(其可以类似于图8中的多边形806)，新的考虑表是从新解析树中的每个可能的遍历中产生的，使得每个新的节点-目标对都有一个表。在多边形908(可以类似于图8中的多边形810)，新的考虑表被单热编码，并且在多边形910(可以类似于图8中的多边形814)，新的编码向量使用投影模型。在多边形912，来自新文本的节点被分类为是否属于广受欢迎的跨度类型，并且NLP学习机可以提供输出，例如，指示哪些节点是该跨度成员的突出显示的跨度。然后方法900在多边形914结束。因此，使用方法900，NLP学习机可以在线操作来自真实世界源(未示出)的新文本，例如，从假设段落中确定事实段落。

[0140] 本发明的各种实施例的描述是出于说明的目的而呈现的，但并非旨在穷举或限于所公开的实施例。在不脱离所描述实施例的范围和精神的情况下，许多修改和变化对于本领域普通技术人员来说将是显而易见的。选择此处使用的术语以最好地解释实施例的原理、实际应用或对市场中发现的技术的技术改进，或者使本领域普通技术人员能够理解此处公开的实施例。

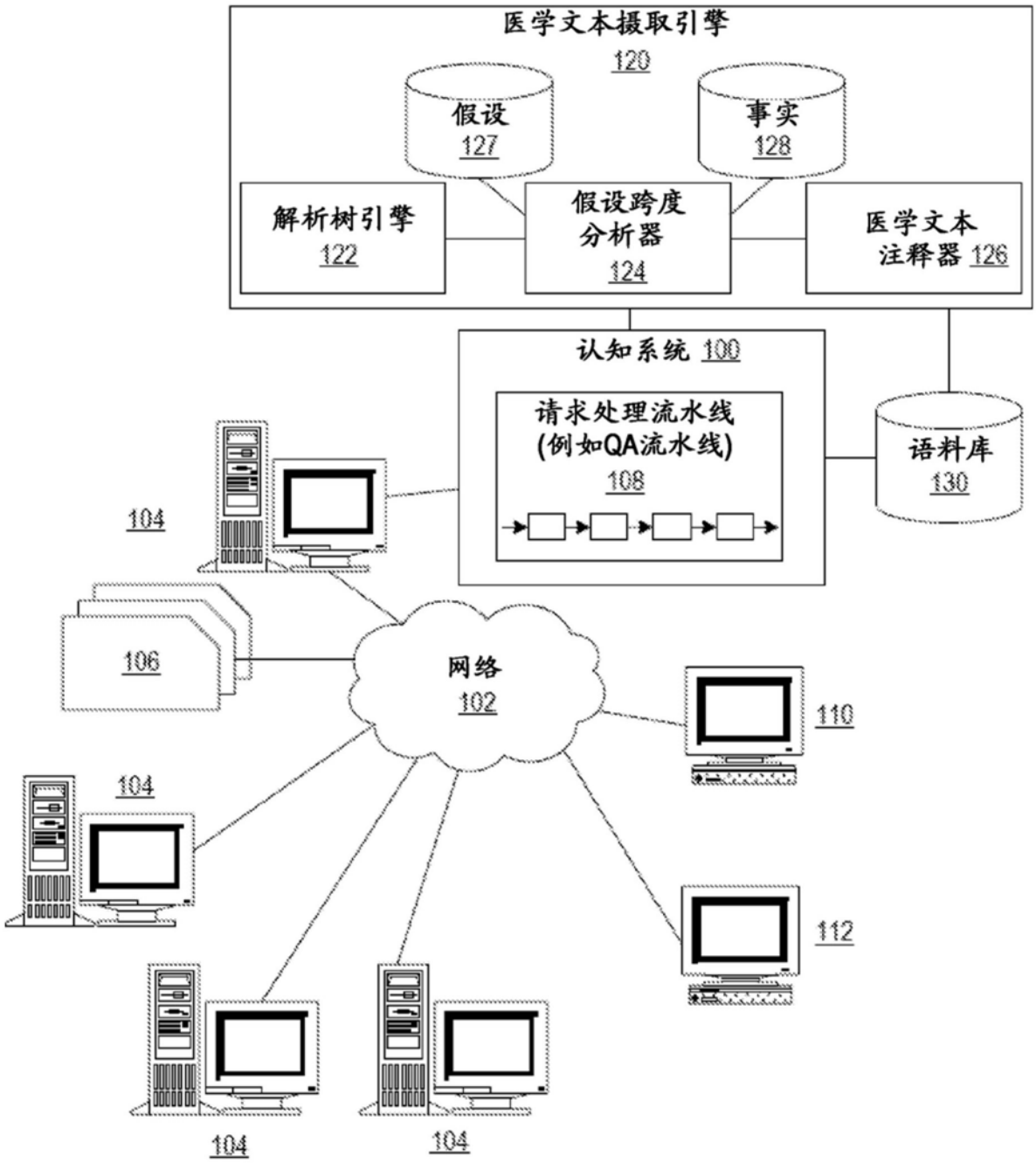

医学文本摄取引擎
120
假设
127
事实
128
解析树引擎
122
假设跨度
分析器
124
医学文本
注释器 126
认知系统 100
请求处理流水线
(例如QA流水线)
108
语料库
130
104
106
网络
102
110
104
112
104
104

图1

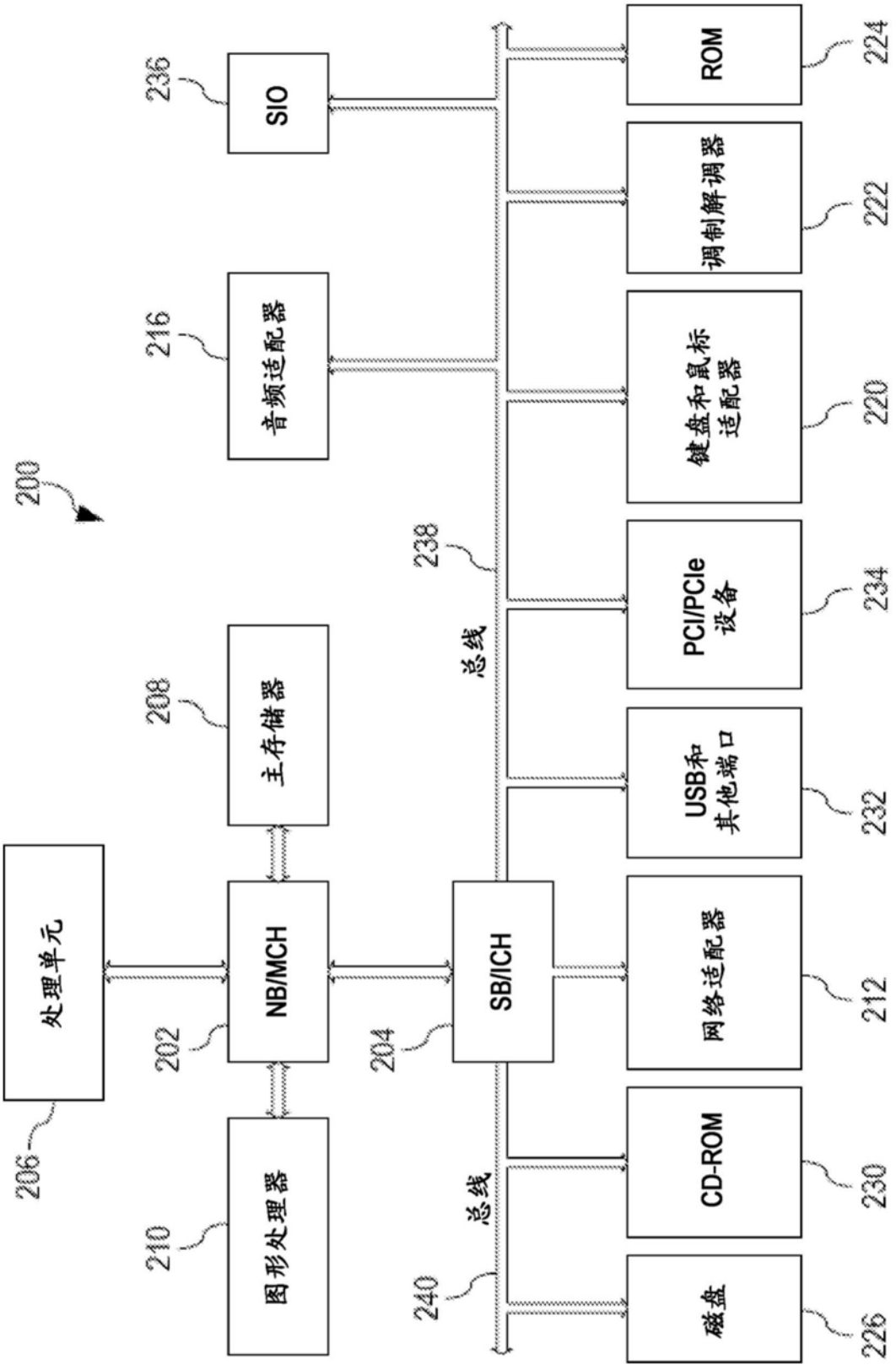

200
处理单元
206
NB/MCH
202
SB/ICH
204
主存储器
208
图形处理器
210
音频适配器
216
SIO
236
总线
238
总线
240
ROM
224
调制解调器
222
键盘和鼠标适配器
220
PCI/PCIe设备
234
USB和其他端口
232
网络适配器
212
CD-ROM
230
磁盘
226

图2

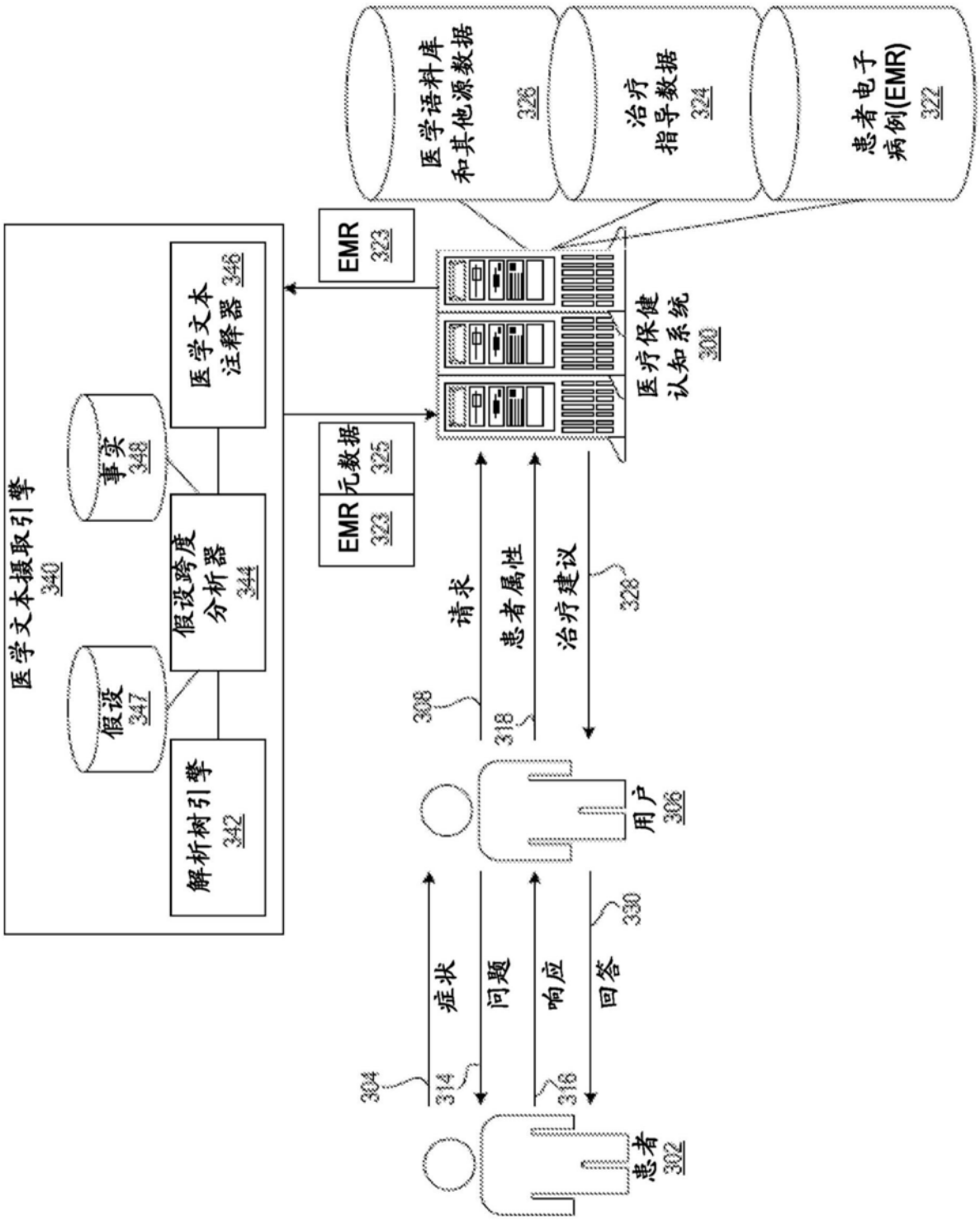

医学语料库
和其他源数据
326
治疗
指导数据
324
患者电子
病例(EMR)
322
EMR
323
医学文本
注释器
346
事实
348
假设跨度
分析器
344
假设
347
解析树引擎
342
医学文本摄取引擎
340
EMR元数据
325
EMR
323
医疗保健
认知系统
300
请求
患者属性
治疗建议
328
308
318
用户
306
症状
问题
响应
回答
330
304
314
316
患者
302

图3

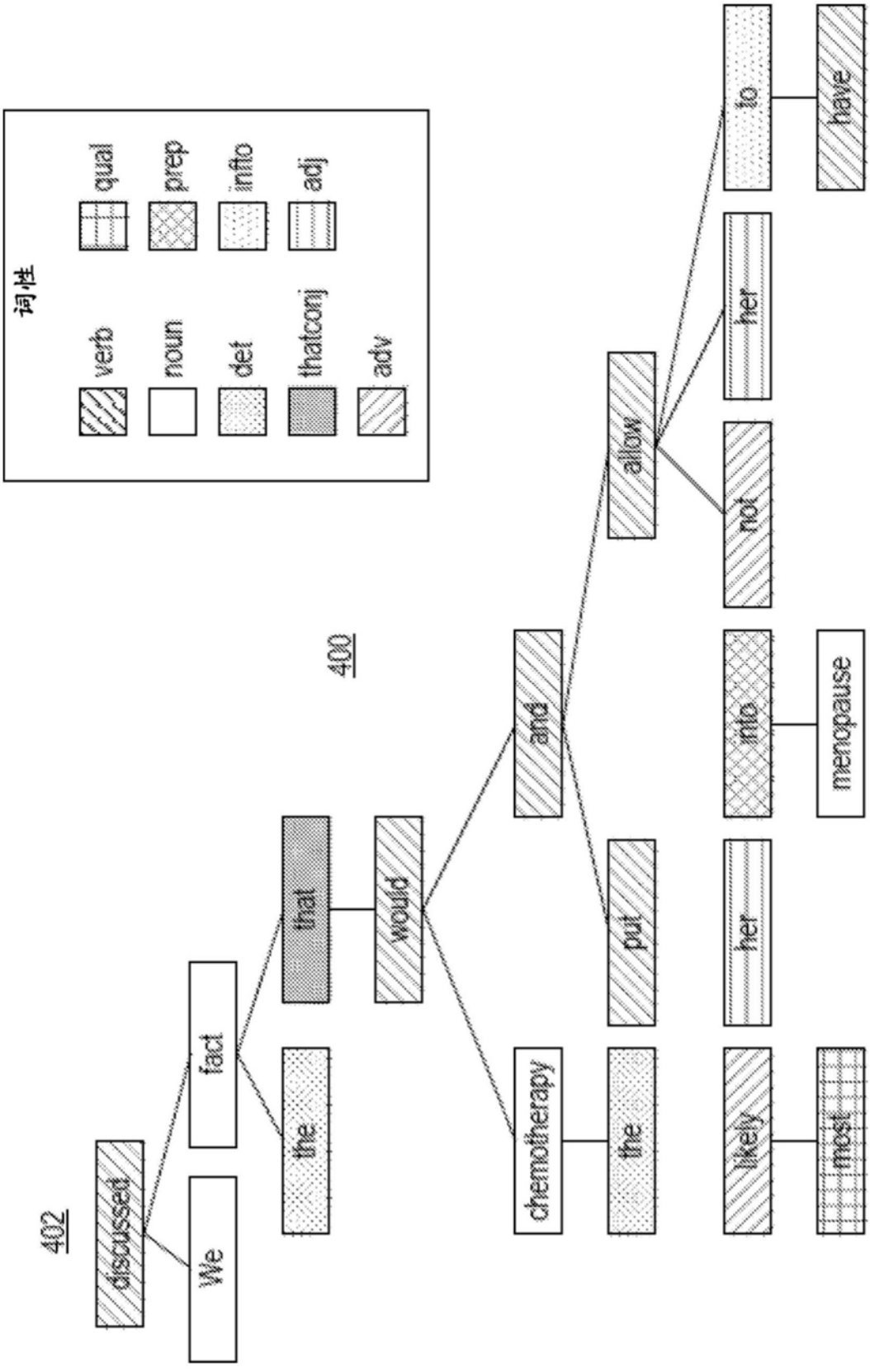

词性
qual
prep
infto
adj
verb
noun
det
thatconj
adv
400
402
We
discussed
fact
the
that
would
and
allow
to
have
her
not
into
menopause
put
her
chemotherapy
the
likely
most

图4

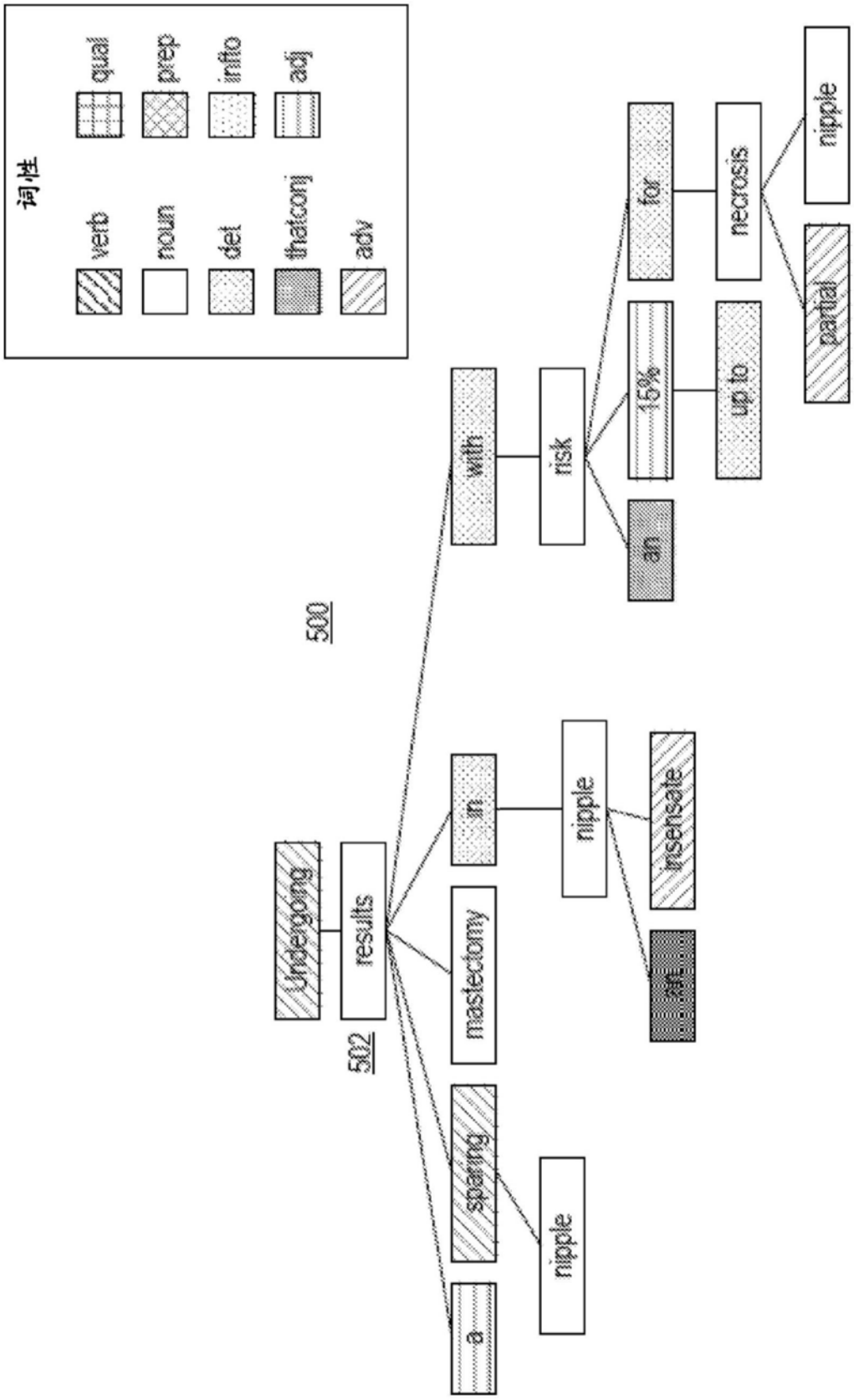

词性
qual
prep
infto
adj
verb
noun
det
thatconj
adv
500
502
Undergoing
results
with
risk
for
16%
up to
an
necrosis
nipple
partial
in
nipple
mastectomy
insensate
pain
sparing
nipple
e

图5

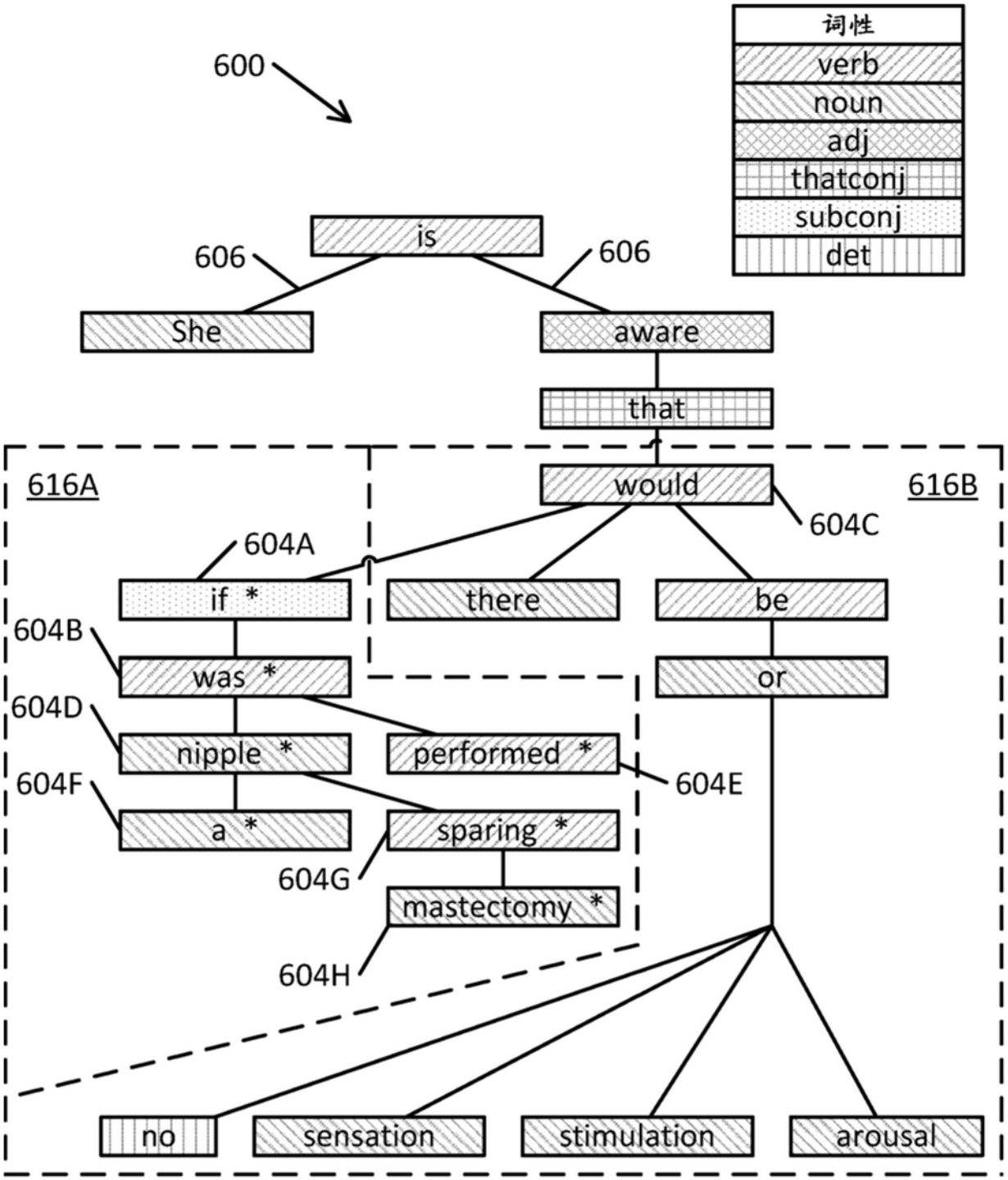

词性
verb
noun
adj
thatconj
subconj
det
600
is
606
606
She
aware
that
616A
616B
would
604C
604A
if *
there
be
604B
was *
or
604D
nipple *
performed *
604E
604F
a *
sparing *
604G
mastectomy *
604H
no
sensation
stimulation
arousal

图6A

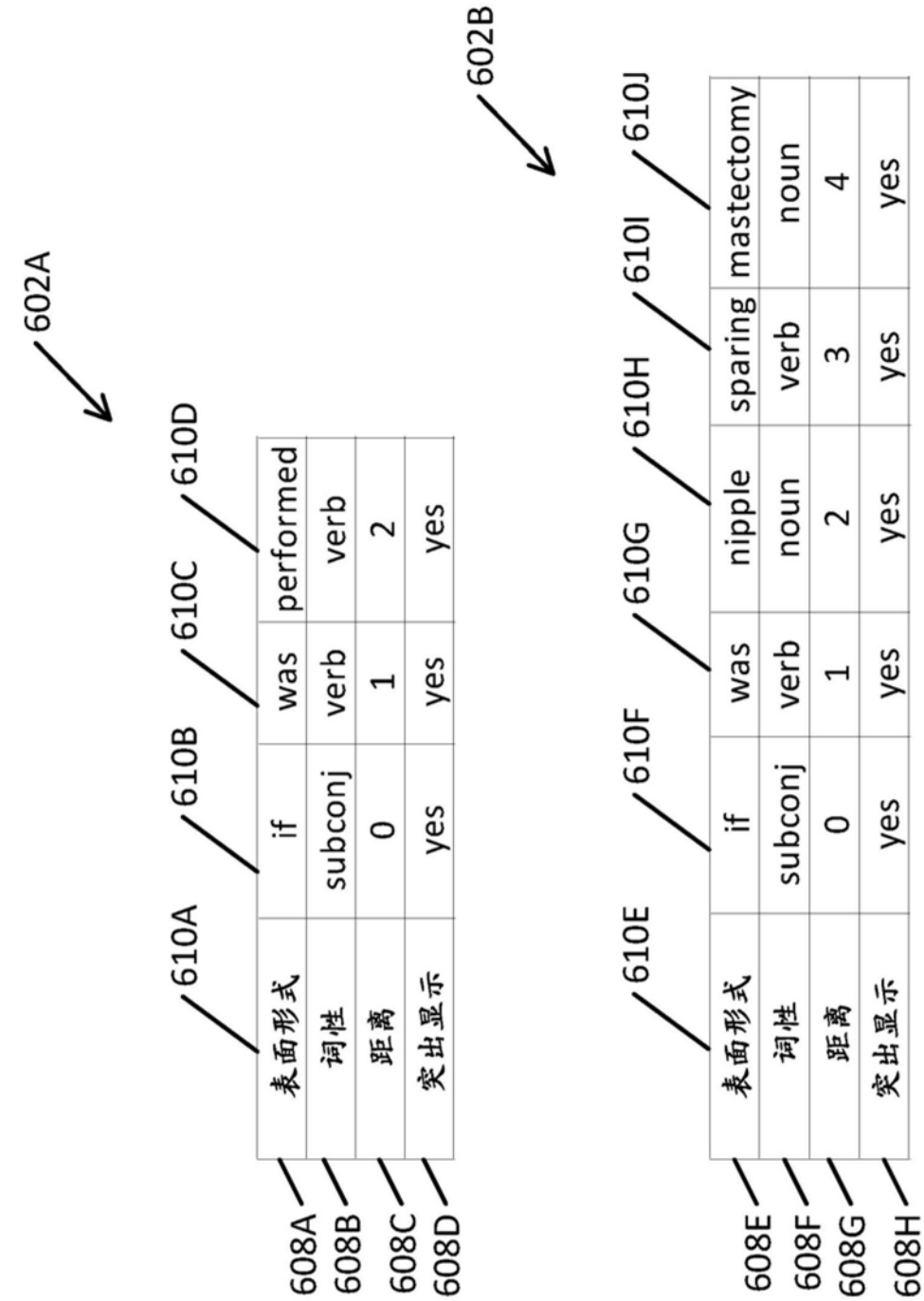

602A
610A
610B
610C
610D
表面形式
词性
距离
突出显示
if
subconj
0
yes
was
verb
1
yes
performed
verb
2
yes
608A
608B
608C
608D
602B
610E
610F
610G
610H
610I
610J
表面形式
词性
距离
突出显示
if
subconj
0
yes
was
verb
1
yes
nipple
noun
2
yes
sparing
verb
3
yes
mastectomy
noun
4
yes
608E
608F
608G
608H

图6B

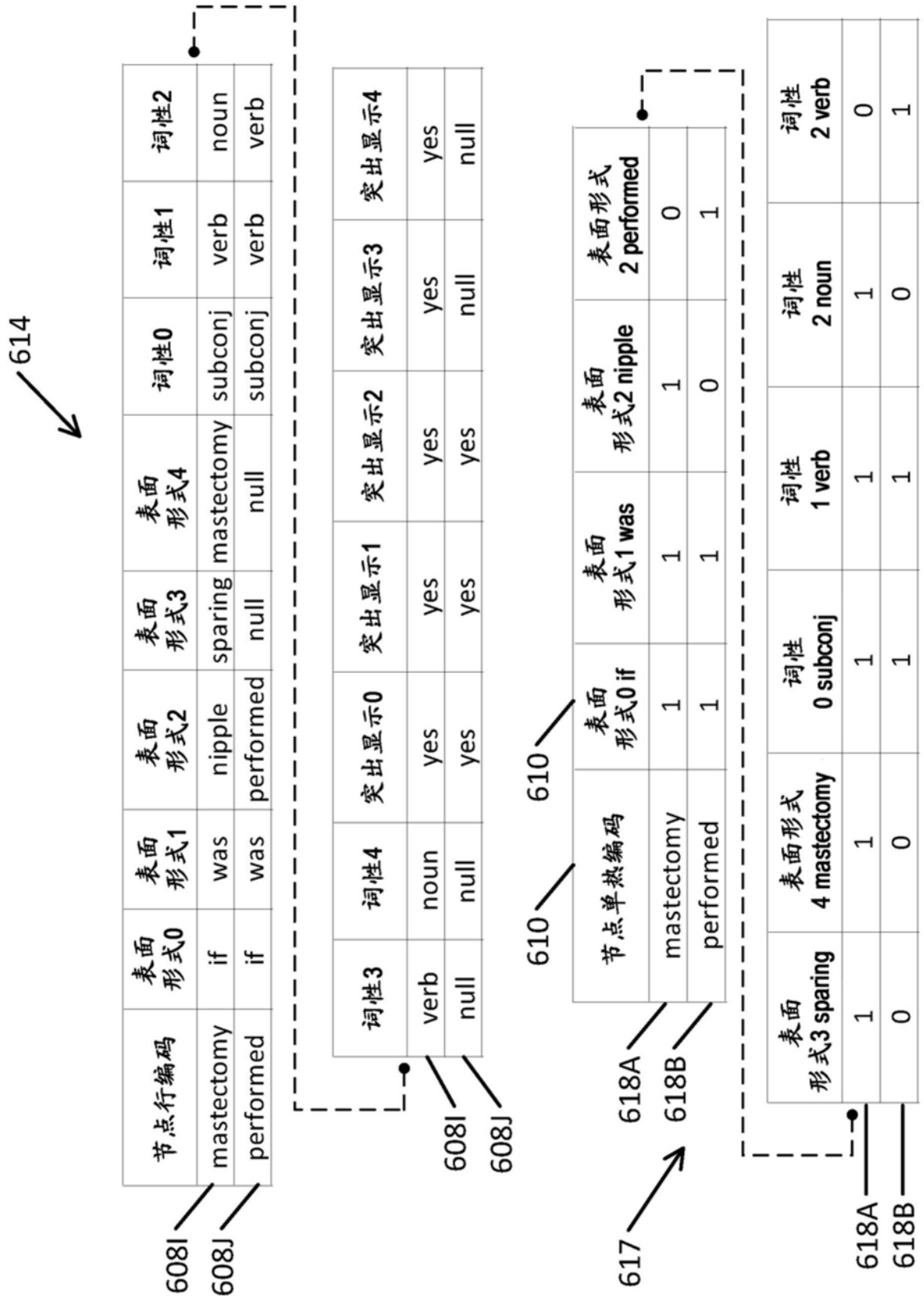

614
节点行编码
表面形式0
表面形式1
表面形式2
表面形式3
表面形式4
词性0
词性1
词性2
mastectomy
if
was
nipple
sparing
mastectomy
subconj
verb
noun
performed
if
was
performed
null
null
subconj
verb
verb
608I
608J
词性3
词性4
突出显示0
突出显示1
突出显示2
突出显示3
突出显示4
verb
noun
yes
yes
yes
yes
yes
null
null
yes
yes
yes
null
null
608I
608J
610
610
节点单热编码
表面形式0 if
表面形式1 was
表面形式2 nipple
表面形式2 performed
mastectomy
1
1
1
0
performed
1
1
0
1
618A
618B
617
表面形式3 sparing
表面形式4 mastectomy
词性0 subconj
词性1 verb
词性2 noun
词性2 verb
1
1
1
1
0
0
0
0
1
1
618A
618B

图6C

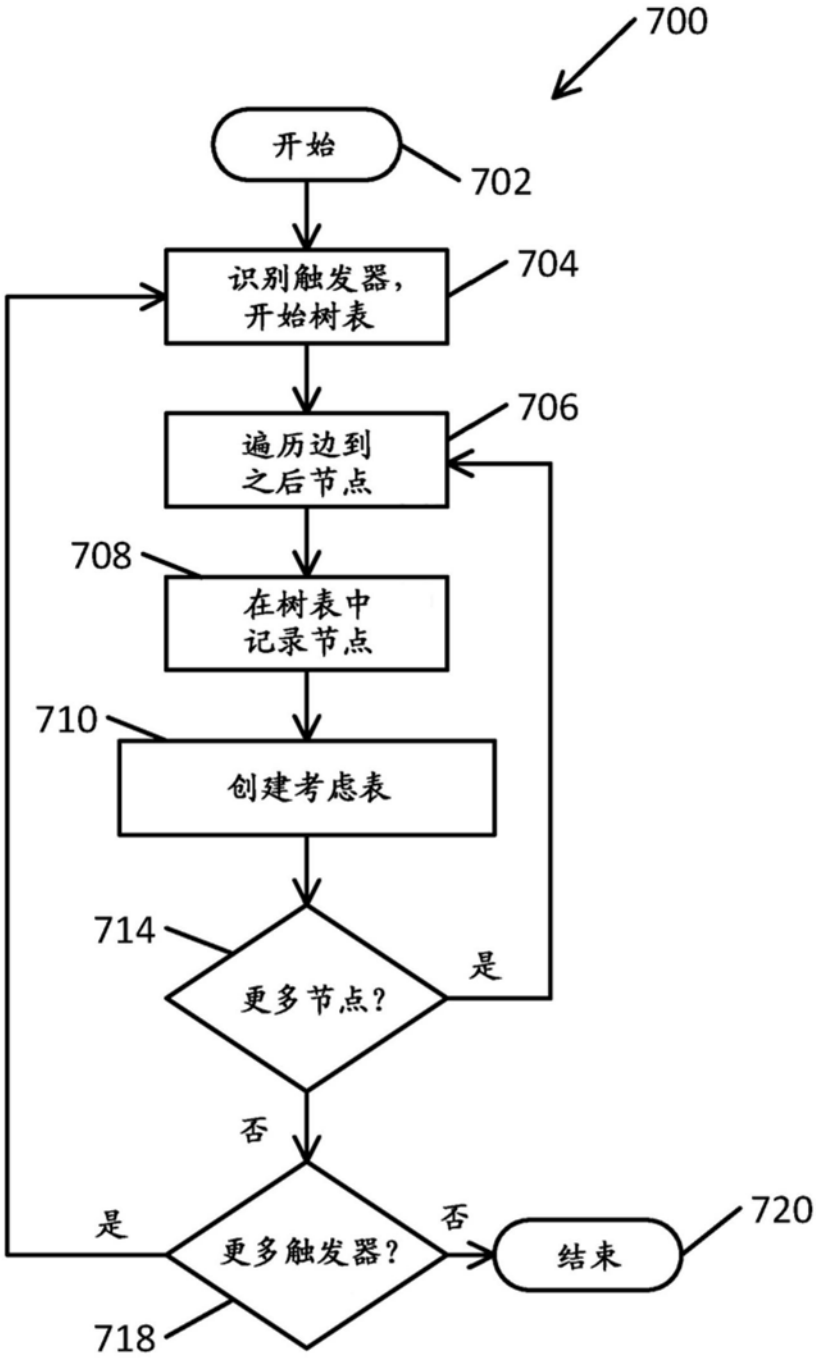

700
开始
702
识别触发器，
开始树表
704
遍历边到
之后节点
706
708
在树表中
记录节点
710
创建考虑表
714
更多节点？
是
否
是
更多触发器？
否
结束
720
718

图7

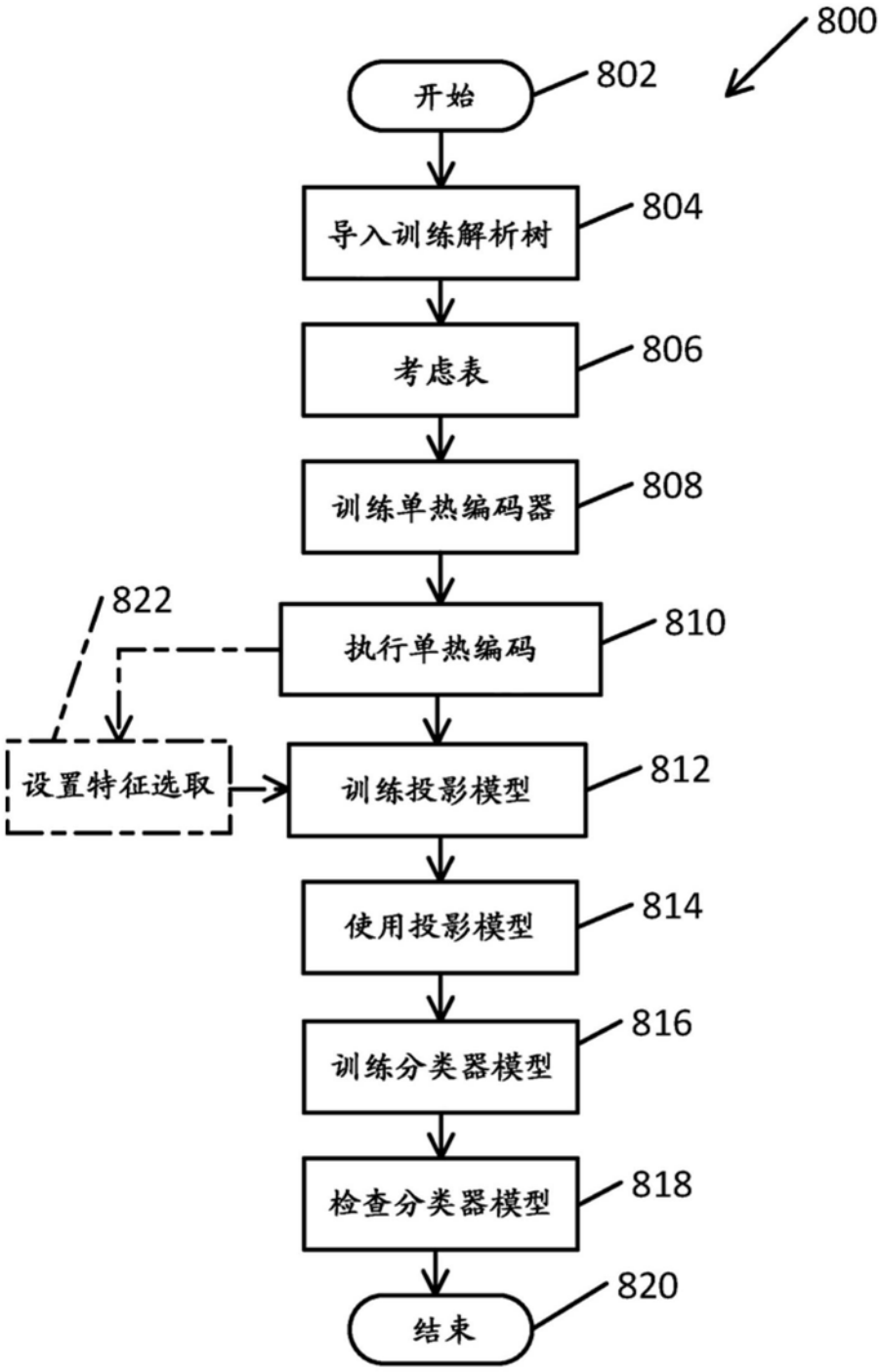

800
开始
802
导入训练解析树
804
考虑表
806
训练单热编码器
808
822
执行单热编码
810
设置特征选取
训练投影模型
812
使用投影模型
814
训练分类器模型
816
检查分类器模型
818
结束
820

图8

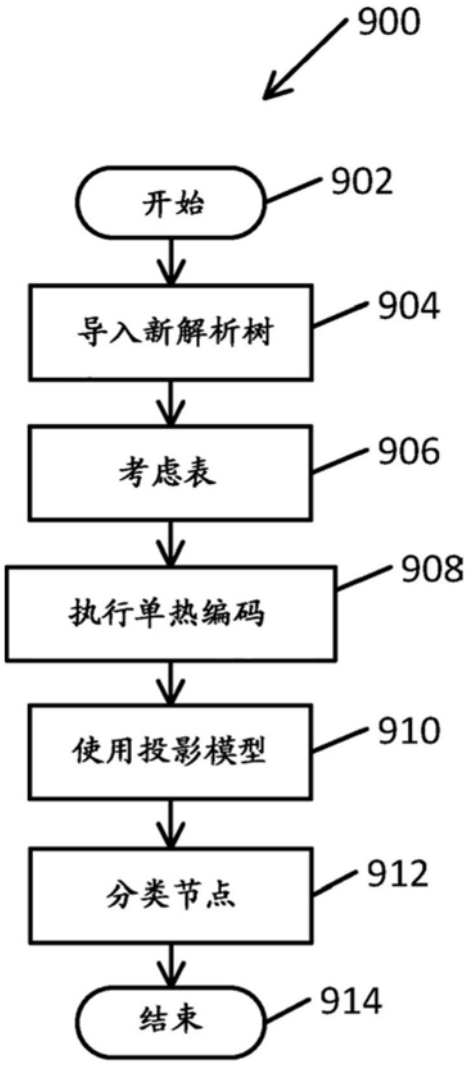

900
开始
902
导入新解析树
904
考虑表
906
执行单热编码
908
使用投影模型
910
分类节点
912
结束
914

图9