



(12) 发明专利

(10) 授权公告号 CN 102439596 B

(45) 授权公告日 2015. 07. 22

(21) 申请号 201080023190. 9

G06F 17/27(2006. 01)

(22) 申请日 2010. 05. 14

G06F 17/26(2006. 01)

(30) 优先权数据

12/470, 492 2009. 05. 22 US

(56) 对比文件

US 2003204400 A1, 2003. 10. 30, 说明书第 4、17-24、26、28、34、37、40 段。

(85) PCT国际申请进入国家阶段日

2011. 11. 21

US 2003204400 A1, 2003. 10. 30, 说明书第 4、17-24、26、28、34、37、40 段。

(86) PCT国际申请的申请数据

PCT/US2010/035033 2010. 05. 14

US 2004102957 A1, 2004. 05. 27, 说明书第 57 段。

(87) PCT国际申请的公布数据

W02010/135204 EN 2010. 11. 25

审查员 李劲娴

(73) 专利权人 微软公司

地址 美国华盛顿州

(72) 发明人 W·B·多兰 C·J·布罗克特

J·J·卡斯蒂略 L·H·范德温德

(74) 专利代理机构 上海专利商标事务所有限公司 31100

代理人 蔡悦

(51) Int. Cl.

G06F 17/30(2006. 01)

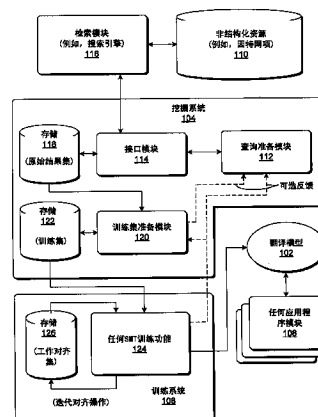
权利要求书2页 说明书12页 附图8页

(54) 发明名称

从非结构化资源挖掘短语对

(57) 摘要

挖掘系统应用查询来从非结构化资源检索结果项。非结构化资源可以对应于可通过网络访问的资源项的储存库。被检索的结果项可以对应于与资源项相关联的文本段(例如,句子段)。挖掘系统通过过滤结果项和建立相应的结果项对,来产生结构化训练集。训练系统可以使用该训练集来产生统计翻译模型。翻译模型可以用于单语种上下文中,以在单一语言的语义相关的短语之间翻译。翻译模型也可以用于双语上下文中,以在两种相应的语言表示的短语之间翻译。还描述了翻译模型的各种应用。



1. 一种用于使用电气数据处理功能来创建用于训练统计翻译模型的训练集的方法 (600), 包括:

构建 (606) 查询, 所述查询包括用查询语言表示的查询项;

向电气数据检索模块呈现 (608) 所述查询, 所述检索模块被配置成执行搜索操作以标识与所述查询的查询项匹配的文档;

从所述检索模块接收 (610) 结果集, 所述结果集提供由所述检索模块在与所述查询的查询项匹配的文档中所标识的结果项, 其中所述结果项也用所述查询语言来表示; 以及

对所述结果集执行 (612) 处理, 以产生结构化训练集, 所述训练集标识所述结果集内的结果项配对, 其中所述训练集中的一个单独的结果项配对包括:

从所述检索模块接收到的第一结果项, 所述第一结果项包括用所述查询语言表示的第一多个单词, 以及

从所述检索模块接收到的第二结果项, 所述第二结果项包括也用所述查询语言表示的第二多个单词,

所述训练集提供一个基础, 通过该基础, 电气训练系统能够学习所述统计翻译模型。

2. 如权利要求 1 所述的方法, 其特征在于, 所述检索模块是搜索引擎, 所述搜索引擎处理所述查询以标识所述结果项, 并且各个结果项包括相关联的各个文档的概述或摘要。

3. 如权利要求 2 所述的方法, 其特征在于, 所述结果集采用用于通过广域网接收文档的资源定位符来通过广域网接收。

4. 如权利要求 1 所述的方法, 其特征在于, 所述执行处理包括基于至少一个考虑因素来约束所述结果集中的所述结果项。

5. 如权利要求 4 所述的方法, 其特征在于, 所述约束包括基于与结果项相关联的评分在所述结果项中标识用于成对匹配的候选。

6. 如权利要求 5 所述的方法, 其特征在于, 所述约束包括基于所述候选和与所述结果集相关联的相应的词汇签名之间的一致性来标识所述用于成对匹配的候选。

7. 如权利要求 5 所述的方法, 其特征在于, 所述约束包括基于反映各个结果项与其它各结果项的相似度的相似度分数来标识所述用于成对匹配的候选。

8. 如权利要求 5 所述的方法, 其特征在于, 所述约束包括基于所述结果项和已标识的结果项的集群之间的关联来标识所述用于成对匹配的候选。

9. 如权利要求 1 所述的方法, 其特征在于, 所述执行处理包括, 对于每一个结果集, 标识所述结果集内的结果项配对。

10. 如权利要求 1 所述的方法, 其特征在于, 所述结果集内的所述结果项对应于单语种文本内容。

11. 如权利要求 1 所述的方法, 其特征在于, 所述结果集内的所述结果项对应于双语文本内容。

12. 如权利要求 1 所述的方法, 其特征在于, 所述结果项包括由所述检索模块从非结构化资源检索到的文本段, 所述文本段对应于所述非结构化资源内的相应的资源项的摘录。

13. 如权利要求 1 所述的方法, 其特征在于, 还包括基于所述训练集来生成所述统计翻译模型, 并应用所述统计翻译模型, 所述应用包括下列各项中的一项:

使用所述统计翻译模型来扩展搜索查询;

使用所述统计翻译模型来促进文档索引决策；
使用所述统计翻译模型来修改文本内容；或
使用所述统计翻译模型来扩展广告信息。

14. 一种用于使用电气数据处理功能来创建用于训练统计翻译模型的训练集的系统，包括：

用于构建查询的装置，所述查询包括用查询语言表示的查询项；

用于向电气数据检索模块呈现所述查询的装置，所述检索模块被配置成执行搜索操作以标识与所述查询的查询项匹配的文档；

用于从所述检索模块接收结果集的装置，所述结果集提供由所述检索模块在与所述查询的查询项匹配的文档中所标识的结果项，其中所述结果项也用所述查询语言来表示；以及

用于对所述结果集执行处理，以产生结构化训练集的装置，所述训练集标识所述结果集内的结果项配对，其中所述训练集中的一个单独的结果项配对包括：

从所述检索模块接收到的第一结果项，所述第一结果项包括用所述查询语言表示的第一多个单词，以及

从所述检索模块接收到的第二结果项，所述第二结果项包括也用所述查询语言表示的第二多个单词，

所述训练集提供一个基础，通过该基础，电气训练系统能够学习所述统计翻译模型。

从非结构化资源挖掘短语对

背景技术

[0001] 近年来,人们对统计机器翻译技术有相当大的兴趣。此技术通过首先建立训练集来操作。传统上,训练集提供并行的文本语料库,如第一语言的文本的正文以及第二语言的文本的对应的正文。训练模块使用统计技术来确定第一文本正文最有可能映射到第二文本正文的方式。此分析导致生成翻译模型。在解码阶段,翻译模型可以被用来将第一语言的文本的实例映射到第二语言的对应的文本实例。

[0002] 统计翻译模型的有效性常常取决于用于产生翻译模型的训练集的稳健性。然而,提供高质量训练集是具有挑战性的任务。部分地,这是因为训练模块通常要求大量的训练数据,但是,只有少量的预先建立的并行语料库类型的资源用于提供这样的信息。在传统的状况下,可以通过手动生成并行文本,例如,通过使用人类翻译者,来获取训练集。然而,手动生成这些文本是极其耗时的任务。

[0003] 存在许多以比较自动化的方式标识并行文本的技术。例如,考虑一个网站以多种不同的语言传达相同信息的情况,信息的每一个版本都与单独的网络地址(例如,单独的URL)相关联。在一种技术中,检索模块可以检查搜索索引以试图标识这些并行文档,例如,基于URL内的特征信息。然而,此技术可以提供对相对数量有限的并行文本的访问。此外,此方法还可以取决于在很多状况下可能不成立的假设。

[0004] 上面的示例已经被框入在两个不同的自然语言之间转换文本的模型的上下文中。还提出了单语种模型。这样的模型试图改述输入文本以与输入文本相同的语言产生输出文本。例如,在一种应用中,这种类型的模型可用于修改用户的搜索查询,例如,通过标识表达搜索查询的额外的方式。

[0005] 单语种模型具有上面所提及的相同缺点。实际上,查找相同语言内的预先存在的并行语料库是特别具有挑战性的。即,在双语上下文中,有预先存在的生成不同的语言的并行文本以适应不同的读者的本国语言的需要。以相同语言生成文本的并行版本的需要则有限得多。

[0006] 尽管如此,这样的单语种信息确实少量地存在。例如,常规的同义词库提供关于相同语言的带有类似的含义的单词的信息。在另一种状况下,一些书已经被不同的翻译者翻译成相同语言。不同的翻译可以充当并行的单语种语料库。然而,这种类型的并行信息太专业化,以至于难以有效地用于比较一般的上下文中。此外,如前所述,只有相对较小的量的这种类型的信息。

[0007] 已经作出了自动标识涉及相同主题的单语种文档的正文,然后挖掘这些文档以发现并行句子的存在的努力。然而,在某些状况下,这些方法依赖于可能会限制它们的有效性和通用性的上下文特定的假设。除这些困难之外,文本可以以多种多样的方式被改述;因此,标识单语种上下文中的并行性潜在地是比标识双语上下文中的相关的文本更加复杂的任务。

发明内容

[0008] 此处描述了从非结构化资源中精选出结构化训练集的挖掘系统。即,非结构化资源可以在重复的内容以及替换类型的内容方面隐而不见地丰富。重复的内容是指非结构化资源包括文本的相同实例的许多重复。替换类型的内容是指非结构化资源包括文本的在形式方面不同但是表达类似的语义内容的许多实例。挖掘系统展示并提取非结构化资源的这些特征,并通过该过程,将原始非结构化内容转换为用于训练翻译模型的结构化内容。在一种情况下,非结构化资源可以对应于可通过网络访问的资源项的储存库(例如,因特网可访问的资源项)。

[0009] 根据一个说明性实现,挖掘系统通过向检索模块提交查询来操作。检索模块使用查询来在非结构化资源内执行搜索,之后,它提供结果项。结果项可以对应于概括了在非结构化资源中所提供的相关联的资源项的文本段。挖掘系统通过过滤结果项并标识相应的结果项对来产生结构化训练集。训练系统可以使用该训练集来产生统计翻译模型。

[0010] 根据一个说明性方面,挖掘系统可以只基于查询的提交来标识结果项,无需预先标识针对相同主题的资源项的组。换言之,挖掘系统可以取将资源项的主题(例如,文档)作为整体的不可知的方法;挖掘系统在子文档片段级别展示非结构化资源内的结构。

[0011] 根据另一个说明性方面,训练集可包括对应于句子段的项。换言之,训练系统不依赖于句子级别的并行性的标识和利用(尽管训练系统也可以成功地处理包括完整的句子的训练集)。

[0012] 根据另一个说明性方面,翻译模型可以用于单语种上下文中,以在单一语言内将输入短语转换为输出短语,其中,输入短语和输出短语具有类似的语义内容,但是具有不同形式的表达。换言之,翻译模型可用于提供输入短语的释义版本。翻译模型也可以用于双语上下文中,以将第一语言的输入短语翻译为第二语言的输出短语。

[0013] 根据另一个说明性方面,描述了翻译模型的各种应用。

[0014] 上面的方法可以显现在各种类型的系统、组件,方法、计算机可读介质、数据结构、产品等等中。

[0015] 提供本发明内容是为了以精简的形式介绍一些概念;这些概念在以下详细描述中进一步描述。本概述并不旨在标识所要求保护主题的关键特征或必要特征,也不旨在用于限制所要求保护主题的范围。

[0016] 附图简述

[0017] 图 1 示出了用于创建和应用统计机器翻译模型的说明性系统。

[0018] 图 2 示出了图 1 的系统在网络相关的环境内的实现。

[0019] 图 3 示出了一个结果集内的一系列结果项的示例。图 1 的系统响应于向检索模块提交查询而返回结果集。

[0020] 图 4 示出了演示图 1 的系统如何在结果集内建立结果项的对的示例。

[0021] 图 5 示出了演示图 1 的系统如何基于对于不同的结果集执行的分析来创建训练集的示例。

[0022] 图 6 示出了呈现图 1 的系统的操作的概述的说明性过程。

[0023] 图 7 示出了用于在图 6 的过程内建立训练集的说明性过程。

[0024] 图 8 示出了用于应用使用图 1 的系统创建的翻译模型的说明性过程。

[0025] 图 9 示出了可以被用来实现前面的附图中所示出的特征的任何方面的说明性处

理功能。

[0026] 在整个说明书和附图中,使用相同的标号来指示相同的组件和特征。系列 100 的编号表示图 1 中的特征,系列 200 编号表示图 2 中的特征,系列 300 编号表示图 3 中的特征,以此类推。

具体实施方式

[0027] 本公开阐述了用于生成可用于建立统计翻译模型的训练集的功能。本公开还阐述了用于生成和应用统计翻译模型的功能。

[0028] 本发明是按如下方式来组织的。章节 A 描述了用于执行上面概括的功能的说明性系统。章节 B 描述了说明章节 A 的系统的操作的说明性方法。章节 C 描述了可以被用来实现章节 A 和 B 所描述的特征的任何方面的说明性处理功能。

[0029] 作为正文前的图文,一些附图是在一个或多个结构组件(各自称为功能、模块、特征、元件等等)的上下文中来描述概念的。附图所示出的各种组件可以以任何方式来实现,例如,通过软件、硬件(例如,离散逻辑组件等等)、固件等等,或这些实现的任何组合。在一种情况下,附图中所示出的将各种组件分离为不同的单元可以反映在实际实现中使用对应的不同的组件。可另选地,或者另外,附图中所示出的任何单个组件都可以通过多个实际组件来实现。可另选地,或者另外,对附图中的任何两个或更多单独的组件的描绘可以反映由单一实际组件所执行的不同的功能。要依次讨论的图 9 提供了关于附图所示出的功能的一个说明性实现的更多细节。

[0030] 其他附图以流程图形式描述了概念。以此形式,某些操作被描述为构成以某一顺序执行的不同的框。这样的实现是说明性的,非限制性的。可以将此处所描述的某些框分组在一起,并在单一操作中执行,可以将某些框分解为多个组件框,还可以以不同于此处所示出的顺序来执行某些框(包括以并行方式执行框)。流程图所示出的框可以通过软件、硬件(例如,离散逻辑组件等等)、固件、手动处理等等,或这些实现的任何组合来实现。

[0031] 至于术语,短语“被配置成”包含任何类型的功能可以被构建来执行已标识的操作的任何方式。功能可以被配置成使用,例如,软件、硬件(例如,离散逻辑组件等等)、固件等等,和/或其任何组合来执行操作。

[0032] 术语“逻辑”包含用于执行任务的任何功能。例如,流程图所示出的每一个操作都对应于用于执行该操作的逻辑组件。可以使用,例如,软件、硬件(例如,离散逻辑组件等等)、固件等等,和/或其任何组合来执行操作。

[0033] A. 说明性系统

[0034] 图 1 示出了用于生成和应用翻译模型 102 的说明性系统 100。翻译模型 102 对应于用于将输入短语映射到输出短语的统计机器翻译(SMT)模型,其中,这里“短语”是指任何一个或多个文本字符串。翻译模型 102 使用统计技术,而并非基于规则的方法来执行此操作。然而,在另一种实现中,翻译模型 102 可以通过结合基于规则的方法的一个或多个特点来补充其统计分析。

[0035] 在一种情况下,翻译模型 102 在单语种上下文中操作。这里,翻译模型 102 生成以与输入短语相同的语言表示的输出短语。换言之,输出短语可以被视为输入短语的释义的版本。在另一种情况下,翻译模型 102 在双语(或多语言)上下文中操作。这里,翻译模型

102 以与输入短语相比不同的语言来生成输出短语。在再一种情况下,翻译模型 102 在直译 (transliteration) 上下文中操作。这里,翻译模型以与输入短语相同的语言来生成输出短语,但是,输出短语以与输入短语相比不同的书写形式来表示。可以将翻译模型 102 应用于其他翻译方案。在所有这样的上下文中,单词“翻译”应该从广义上来理解,是指任何类型的文本信息从一种状态到另一种状态的转换。

[0036] 系统 100 包括三个主要组件:挖掘系统 104;训练系统 106;以及应用模块 108。作为概述,挖掘系统 104 产生用于训练翻译模型 102 的训练集。训练系统 106 根据训练集来应用迭代方法以导出翻译模型 102。应用模块 108 应用翻译模型 102,以在特定的与使用有关的方案中将输入短语映射到输出短语。

[0037] 在一种情况下,单一系统可以实现图 1 所示出的全部组件,通过单一实体或多个实体的任何组合来管理。在另一种情况下,任何两个或更多分开的系统可以实现图 1 所示出的任何两个或更多组件,也是通过单一实体或多个实体的任何组合来管理。不论是哪一种情况,图 1 所示出的组件可以位于单一站点或者分布在多个相应的站点中。下面的说明提供了关于图 1 所示出的组件的附加细节。

[0038] 以挖掘系统 104 开始,此组件通过从非结构化资源 110 检索结果项来操作。非结构化资源 110 表示资源项的任何局部化的或分布式源。资源项又可以对应于文本信息的任何单元。例如,非结构化资源 110 可以表示由诸如因特网之类的广域网所提供的资源项的分布式储存库。这里,资源项可以对应于任何类型的可通过网络访问的页面和 / 或相关联的文档。

[0039] 非结构化资源 110 被视为非结构化的,因为它没有按并行语料库的方式排列的先验。换言之,非结构化资源 110 不根据任何包罗万象的方案来将其资源项彼此相关联。尽管如此,非结构化资源 110 可以在重复的内容以及替换类型的内容方面隐而不见地丰富。重复的内容是指非结构化资源 110 包括文本的相同的实例的许多重复。替换类型的内容是指非结构化资源 110 包括文本的在形式方面不同但是表达类似的语义内容的许多实例。这意味着,存在非结构化资源 110 的可以被挖掘以便用于构建训练集的底层特点。

[0040] 挖掘系统 104 的一个用途是展示非结构化资源 110 的上文所描述的特征,并通过该过程,将原始非结构化内容转换为结构化内容,以便用于训练翻译模型 102。挖掘系统 104 部分地结合检索模块 116 使用查询准备模块 112 和接口模块 114 来实现此用途。查询准备模块 112 形成一组查询。每一个查询都可包括涉及目标主题的一个或多个查询项。接口模块 114 向检索模块 116 提交查询。检索模块 116 使用查询来在非结构化资源 110 内执行搜索。响应于此搜索,检索模块 116 返回不同的相应的查询的多个结果集。每一个结果集又都包括一个或多个结果项。结果项标识非结构化资源 110 内的相应的资源项。

[0041] 在一种情况下,挖掘系统 104 和检索模块 116 通过由相同实体或不同的相应的实体管理的相同系统来实现。在另一种情况下,挖掘系统 104 和检索模块 116 通过也是由相同实体或不同的相应的实体管理的两个相应的系统来实现。例如,在一种实现中,检索模块 116 表示诸如,但不仅限于,由美国华盛顿州雷德蒙市的微软公司所提供的 Live Search 引擎之类的搜索引擎。用户可以通过诸如由搜索引擎所提供的接口(例如,API 等等)之类的任何机制来访问搜索引擎。搜索引擎可以使用任何搜索策略和分级策略,响应于所提交的查询,标识和形成结果集。

[0042] 在一种情况下,结果集中的结果项对应于相应的文本段。在响应于对查询的提交而形成文本段时,不同的搜索引擎可以使用不同的策略。在很多情况下,文本段提供资源项的通过所提交的查询传递资源项的相关性的代表性的部分(例如,摘录)。为解释起见,文本段可以被视为它们的相关联的完整的资源项的摘要。更具体而言,在一种情况下,文本段可以对应于从底层的完整的资源项获取的一个或多个句子。在一种情况下,接口模块 114 和检索模块 116 可以形成包括句子段的资源项。在另一种情况下,接口模块 114 和检索模块 116 可以形成包括完整的句子(或诸如完整的段落等等之类的较大的文本单元)的资源项。接口模块 114 将结果集存储在存储 118 中。

[0043] 训练集准备模块 120(简称为“准备模块”)处理结果集中的原始数据以产生训练集。此操作包括两个组件操作,即,可以分开或一起执行的过滤和匹配。至于过滤操作,准备模块 120 基于一个或多个约束考虑因素来过滤原始结果项组。此处理的目标是标识作为用于成对匹配的适当候选的结果项的子集,从而从结果集中消除“噪声”。过滤操作产生经过过滤的结果集。至于匹配操作,准备模块 120 对经过过滤的结果集执行成对匹配。成对匹配标识结果集内的结果项对。准备模块 120 将由上面的操作所产生的训练集存储在存储 122 中。将在此说明的稍后的结合点提供关于准备模块 120 的操作的附加细节。

[0044] 训练系统 106 使用存储 122 中的训练集来训练翻译模型 102。为此,训练系统 106 可包括诸如短语类型的 SMT 功能之类的任何类型的统计机器翻译(SMT)功能 124。SMT 功能 124 通过使用统计技术来操作,以标识训练集中的模式。SMT 功能 124 使用这些模式来标识训练集内的短语的关联。

[0045] 更具体而言,SMT 功能 124 以迭代的方式执行其训练操作。在每一个阶段,SMT 功能 124 执行统计分析,该统计分析允许它达成关于训练集中的短语的成对地对齐的暂定的假设。SMT 功能 124 使用这些暂定的假设来重复其统计分析,允许它达成更新的暂定的假设。SMT 功能 124 重复此迭代操作,直到结束条件被视为被满足。存储 126 可以在由 SMT 功能 124 执行的处理过程中维持临时对齐信息(例如,以翻译表等等的形式)的工作集。在其处理结束时,SMT 功能 124 产生定义翻译模型 102 的统计参数。将在此说明的稍后的结合点提供关于 SMT 功能 124 的附加细节。

[0046] 应用模块 108 使用翻译模型 102 来将输入短语转换为语义相关的输出短语。如上文所指出的,输入短语和输出短语可以以相同语言或不同的相应的语言来表示。应用模块 108 可以在各种应用方案的上下文中执行此转换。将在此说明的稍后的结合点提供关于应用模块 108 和应用方案的附加细节。

[0047] 图 2 示出了图 1 的系统 100 的一个代表性实现。在此情况下,计算功能 202 可用于实现挖掘系统 104 和训练系统 106。计算功能 202 可以表示在单一站点中维护的或分布在多个站点上的任何处理功能,如由单一实体或多个实体的组合维护的。在一个代表性的情况下,计算功能 202 对应于任何类型的计算机设备,如个人台式计算设备,服务器类型的计算设备,等等。

[0048] 在一种情况下,非结构化资源 110 可以通过由网络环境 204 所提供的资源项的分布式储存库来实现。网络环境 204 可以对应于任何类型的局域网或广域网。例如,但不仅限于,网络环境 204 可以对应于因特网。这样的环境提供对潜在大量的资源项(例如,对应于可通过网络访问的页面和链接的内容项的)的访问。检索模块 116 可以以常规方式,例

如,使用网络爬行功能等等,维持网络环境 204 中的可用资源项的索引。

[0049] 图 3 示出了可以由检索模块 116 响应于查询 304 的提交而返回的假设结果集 302 的一部分的示例。此示例充当用于说明图 1 的挖掘系统 104 的一些概念性基础的媒介。

[0050] 查询 304“shingles zoster(带状疱疹)”涉及一种已知的疾病。查询被选择为带有足够的聚焦准确定位针对的主题,以排除大量的无关信息。在此示例中,“shingles”(带状疱疹)是指疾病的通用名称,而“zoster”(带状疱疹)(例如,如在 herpes zoster 中)是指疾病的比较正式的名称。因此,这种查询项的组合可以减少涉及单词“shingles”(带状疱疹)的无关的和非故意的含义的结果项的检索。

[0051] 结果集 302 包括被标记为 R1-RN 的一系列结果项;图 3 示出了这些结果项的小样本。每一个结果项都包括从对应的资源项中提取的文本段。在此情况下,文本段包括句子段。但是,接口模块 114 和检索模块 116 也可以被配置成提供包括完整的句子(或完整的段落等等)的资源项。

[0052] 疾病“shingles(带状疱疹)”具有突出的特征。例如,shingles 是由导致水痘的相同病毒(herpes zoster)的重新激活所引起的疾病。在被重新唤醒之后,病毒沿着身体的神经传播,导致外表上是微红的疼痛的皮疹,并有成小簇的水疱。当免疫系统受损害时,常常会出现该疾病,因此,会由身体的外伤,其他疾病,压力等等而触发。该疾病常常折磨老年人等等。

[0053] 可以预计不同的结果项包括聚焦于疾病的突出的特征的内容。结果,可以预计结果项重复某些警告短语。例如,如由实例 306 所指示的,多个结果项提及发生了疼痛的皮疹,如不同地表达的。如由实例 308 所指示的,多个结果项提及该疾病与免疫系统变弱相关联,如不同地表达的。如由实例 310 所指示的,多个结果项提及该疾病导致病毒沿着身体中的神经移动,如不同地表达的,等等。这些示例只是说明性的。其他结果项可能基本上与目标主题不相干。例如,结果项 312 在建筑材料的上下文中使用术语“shingles”,因此,与主题没有密切关系。但是,甚至这种无关的结果项 312 也可包括与其他结果项共享的短语。

[0054] 可以从结果集 302 中表现的模式收集各种洞察。这些洞察中的某些严密地涉及目标主题,即,疾病“shingles”(带状疱疹)。例如,挖掘系统 104 可以使用结果集 302 来推断“shingles”和“herpes zoster”是同义词。其他洞察涉及一般的医学领域。例如,挖掘系统 104 可以推断短语“painful rash”(疼痛的皮疹)可以在含义上代替短语“rash that is painful”(疼痛的皮疹)。此外,当讨论免疫系统(以及潜在地,其他主题)时,挖掘系统 104 还可以推断短语“impaired”可以在含义上替换为“weakened”或“compromised”。其他洞察可能具有全局性的或领域独立的范围。例如,挖掘系统 104 可以推断短语“moves along”可以在含义上代替“travels over”或“moves over”,而短语“elderly”可以替换为“old people”,或“old folks”,或“senior citizens”等等。这些等效性表现在结果集 302 内的医学上下文中,但是,它们也可以适用于其他上下文。例如,可以描述一个人的去上班的路为“travelling over”一条道路或者“moving along”一条道路。

[0055] 图 3 还示出了训练系统 106 可以用来标识短语之间有意义的相似度的一个机制。例如,结果项重复诸如“rash”、“elderly”、“nerves”、“immune system”等等之类的相同单词中的许多。这些频繁出现的单词可以充当研究文本段以了解语义相关短语的存在的锚点。例如,通过聚焦于与通常出现的短语“immune system”相关联的锚点,训练系统 106

可以得出结论：“impaired”、“weakened”，以及“compromised”可以对应于语义可互换的单词。训练系统 106 可以以逐段方式趋近此研究。即，它可以导出关于短语的对齐的暂定的假设。基于那些假设，它可以重复其研究以导出新的暂定的假设。在任何结合点，暂定的假设可以允许训练系统 106 导出对于结果项的相关性的额外的洞察；可另选地，假设可以表示后退一步，使进一步的分析模糊（在这样的情况下，可以修改假设）。通过此过程，训练系统 106 试图得出关于结果集内的短语的相关性的稳定的假设组。

[0056] 更一般而言，此示例还示出了挖掘系统 104 可以只基于查询的提交来标识结果项，无需预先标识针对相同主题的资源项的组（例如，底层文档）。换言之，挖掘系统 104 可以取将资源项的主题作为整体的不可知的方法。在图 3 的示例中，大部分资源项可能事实上确实涉及相同主题（疾病 shingles）。然而，(1) 这种相似度是只根据查询暴露的，而非对文档的元级别的分析，以及 (2) 没有要求资源项涉及相同主题。

[0057] 前进到图 4，此图示出了准备模块 120（图 1 的）可用于在结果集 (R_A) 内建立初始结果项配对 (R_{A1} - R_{AN}) 的方式。这里，准备模块 120 可以在结果集中的每一个结果项和每一个其他结果项之间建立链接（排除结果项的自我相同的配对）。例如，第一对将结果项 R_{A1} 与结果项 R_{A2} 连接。第二对将结果项 R_{A1} 与结果项 R_{A3} 连接，依次类推。在实践中，准备模块 120 可以基于一个或多个过滤考虑因素来约束结果项之间的关联。章节 B 将提供关于准备模块 120 可以约束结果项的成对匹配的方式的附加信息。

[0058] 为重复，以上面的方式配对的结果项可以对应于它们的相应的资源项的任何部分，包括句子段。这意味着，挖掘系统 104 可以建立训练集，而无需明确的标识并行句子的任务。换言之，训练系统 106 不取决于对句子级别的并行性的利用。然而，训练系统 106 也可以成功地处理其中结果项包括完整的句子（较大的文本单元）的训练集。

[0059] 图 5 示出了来自不同的结果集的成对地映射可以被组合以形成存储 122 中的训练集的方式。即，查询 Q_A 导致结果集 R_A ，而 R_A 又导致成对匹配的结果集 TS_A 。查询 Q_B 导致结果集 R_B ，而 R_B 又导致成对匹配的结果集 TS_B ，依次类推。准备模块 120 组合并级联这些不同的成对匹配的结果集以创建训练集。作为整体，训练集在结果项之间建立初始临时对齐的集合，供进一步的调查。训练系统 106 以迭代的方式对训练集进行操作，以标识揭示确实相关的文本段的对齐的子集。最终，训练系统 106 试图标识在对齐内表现出的语义相关的短语。

[0060] 作为本节中的最后一点，注意，在图 1 中，在系统 100 的不同的组件之间绘制了虚线。这用图形方式表示了由任何组件取得的结论可用于修改其他组件的操作。例如，SMT 功能 124 可以取得对准备模块 120 执行其初始过滤以及结果集的配对的方式有影响的某些结论。准备模块 120 可以接收此反馈，并作为响应，修改其过滤或匹配行为。在另一种情况下，SMT 功能 124 或准备模块 120 可以达成关于某些查询组成策略的有效性的结论，例如，对查询组成策略提取在重复的内容以及替换类型的内容方面丰富的结果集的能力影响。查询准备模块 112 可以接收此反馈，并作为响应，修改其行为。更具体而言，在一种情况下，SMT 功能 124 或准备模块 120 可以发现包括在另一轮查询内可能有用的关键术语或关键短语，导致额外的结果集供分析。在系统 100 内可以存在供反馈的其他机会。

[0061] B. 说明性过程

[0062] 图 6-8 示出了说明图 1 的系统 100 的操作的一种方式的过程 (600、700、800)。由

于在章节 A 已经介绍了系统 100 的操作的底层的原理,在此部分以概述的方式说明某些操作。

[0063] 从图 6 开始,此图示出了表示挖掘系统 104 和训练系统 106 的操作的概述的过程 600。更具体而言,操作的第一阶段描述了由挖掘系统 104 执行的挖掘操作 602,而操作的第二阶段描述了由训练系统 106 执行的训练操作 604。

[0064] 在框 606 中,挖掘系统 104 通过构建一组查询来启动过程 600。挖掘系统 104 可以使用不同的策略来执行此任务。在一种情况下,挖掘系统 104 可以提取一组由用户先前向搜索引擎提交的实际查询,例如,从查询日志等等中获得的。在另一种情况下,挖掘系统 104 可以基于任何引用源或引用源的组合来构建“人工的”查询。例如,挖掘系统 104 可以从诸如 Wikipedia 等等之类的百科全书引用源的分类索引或从同义词库等等提取查询项。只作为一个示例,挖掘系统 104 可以使用引用源来生成包括不同的疾病名称的查询集合。挖掘系统 104 可以利用一个或多个其他项来补充疾病名称,以帮助聚焦返回的结果集。例如,挖掘系统 104 可以将每一个常见的疾病名称与其正式的医学等效名称结合,如在“shingles AND zoster”中那样。或者,挖掘系统 104 可以将每一个疾病名称与在某种程度上与疾病名称正交的另一查询项结合,如“shingles AND prevention”(带状疱疹 AND 预防)等等。

[0065] 更加广泛地考虑,框 606 中的查询选择可以受不同的包罗万象的目标来控制。在一种情况下,挖掘系统 104 可以试图准备聚焦于特定领域的查询。这种策略在对在某种程度上偏重于该特定领域的短语进行表面处理时有效。在另一种情况下,挖掘系统 104 可以试图准备细查较宽的范围的领域的查询。这种策略在对本质上比较领域独立的短语进行表面处理(surcacing)时有效。在任何情况下,挖掘系统 104 试图获取在重复的内容以及替换类型的内容方面丰富的结果项,如上文所讨论的。此外,查询本身仍是从非结构化资源中提取并行性的主要媒介,而并非任何类型的对资源项之间的相似的主题的先验的分析。

[0066] 最后,挖掘系统 104 可以接收揭示其对查询的选择的有效性的反馈。基于此反馈,挖掘系统 104 可以修改控制它如何构建查询的规则。另外,反馈可以标识可以被用来形成查询的特定关键字或关键短语。

[0067] 在框 608 中,挖掘系统 104 向检索模块 116 提交查询。检索模块 116 又使用查询来在非结构化资源 110 内执行搜索操作。

[0068] 在框 610 中,挖掘系统 104 从检索模块 116 接收返回的结果集。结果集包括相应的结果项组。每一个结果项都可以对应于从非结构化资源 110 内的对应的资源项中提取的文本段。

[0069] 在框 612 中,挖掘系统 104 对结果集执行初始处理以产生训练集。如上文所描述的,此操作可包括两个组件。在过滤组件中,挖掘系统 104 约束结果集以去除或边缘化对标识语义相关的短语不太可能有用的信息。在匹配组件中,挖掘系统 104 标识结果项对,例如,一个集一个集地。图 4 用图形方式示出了说明性结果集的上下文中的此操作。图 7 提供关于在框 612 中执行的操作的附加细节。

[0070] 在框 614 中,训练系统 106 使用统计技术来对训练集进行操作,以导出翻译模型 102。可以使用任何统计机器翻译方法来执行此操作,诸如任何类型的面向短语的方法。一般而言,翻译模型 102 可以被表示成 $P(y|x)$,后者定义了输出短语 y 表示给定输入短语 x 的概率。通过使用贝叶斯规则,这可以被表达为 $P(y|x) = P(x|y)P(y)/P(x)$ 。训练系统 106 操

作以基于对训练集的研究来暴露由此表达式所定义的概率,带有趋向于最大化 $P(x|y)P(y)$ 的从输入短语 x 学习映射的目标。如上文所指出的,调查本质上是迭代的。在操作的每一个阶段,训练系统 106 可以取得关于训练集内的短语(以及文本段作为整体)的对齐的暂定的结论。在面向短语的 SMT 方法中,可以使用翻译表等等来表达暂定的结论。

[0071] 在框 616 中,训练系统 616 确定是否已经达到结束条件,指示已经实现了令人满意的对齐结果。可以使用任何度量来作出此确定,诸如已知的双语评估替代技术(BLEU)分数。

[0072] 在框 618 中,如果还没有实现满意的结果,则训练系统 106 修改用于训练的其假设中的任何一个。这具有修改关于结果项内的短语如何彼此相关联(以及文本段作为整体如何彼此相关联)的主要的起作用的假设的效果。

[0073] 当满足了结束条件时,训练系统 106 将具有经标识的训练集内的语义相关的短语之间的映射。定义这些映射的参数建立了翻译模型 102。构成对这样的翻译模型 102 的使用的基础的假设是,文本的新遇到的实例将类似于训练集内发现的模式。

[0074] 图 6 的过程可以以不同的方式变化。例如,在替换实现中,框 614 中的训练操作可以使用统计分析和基于规则的分析的组合来导出翻译模型 102。在另一修改方案中,框 614 中的训练操作可以将训练任务分解为多个子任务,事实上,创建多个翻译模型。然后,训练操作可以将多个翻译模型合并为单一翻译模型 102。在另一修改方案中,可以使用诸如从同义词库等等获得的信息之类的引用源,来初始化或“预准备”框 614 中的训练操作。其他修改方案也是可以的。

[0075] 图 7 示出了提供关于由挖掘系统 104 在图 6 的框 612 中执行的过滤和匹配过程的附加细节的过程 700。

[0076] 在框 702 中,挖掘系统 104 基于一个或多个考虑因素来过滤原始结果集。此操作具有标识被视为用于成对匹配的最适当的候选的结果项的子集的效果。此操作有助于降低训练集的复杂性和训练集中的噪声量(例如,通过消除或边缘化被评估为具有低相关性的结果项)。

[0077] 在一种情况下,挖掘系统 104 可以基于与结果项相关联的评分来将结果项标识为用于成对匹配的适当的候选。否定地说,挖掘系统 104 可以去除具有低于规定的相关性阈值的评分的结果项。

[0078] 可另选地,或另外地,挖掘系统 104 可以生成相应的结果集的词汇签名,这些签名表达在结果集内发现的典型的文本特点(例如,基于在结果集中出现的单词的共同性)。然后,挖掘系统 104 可以将每一个结果项与和其结果集相关联的词汇签名进行比较。挖掘系统 104 可以基于此比较来将结果项标识为用于成对匹配的适当的候选。从反面来说,挖掘系统 104 可以去除与它们的词汇签名相差规定的量的结果项。不太正式地陈述,挖掘系统 104 可以去除在它们的相应的结果集内“突出的”结果项。

[0079] 可另选地,或另外地,挖掘系统 104 可以生成相似度分数,这些分数标识每一个结果项相对于结果集内的每一个其他结果项的相似度如何。挖掘系统 104 可以依赖于任何相似度度量来作出此确定,诸如,但不仅限于,余弦相似度度量。挖掘系统 104 可以基于这些相似度分数来将结果项标识为用于成对匹配的适当的候选。从反面来说,挖掘系统 104 可以标识对于匹配来说不是好的候选的结果项的对,因为它们彼此相差超过规定量(如相似

度分数所揭示的)。

[0080] 可另选地,或另外地,挖掘系统 104 可以对结果集内的结果项执行聚类分析,以确定类似的结果项的组,例如,使用 k-最近邻聚类技术或任何其他聚类技术。然后,挖掘系统 104 可以将每一个集群内的结果项标识为用于成对匹配的适当的候选,但不是跨不同的集群的候选。

[0081] 挖掘系统 104 可以执行其他操作以过滤或“清理”从非结构化资源 110 收集到的结果项。框 702 导致生成经过滤的结果集。

[0082] 在框 704 中,挖掘系统 104 标识经过滤的结果集内的对。如已经讨论的,图 4 示出了如何在说明性结果集的上下文内执行此操作。

[0083] 在框 706 中,挖掘系统 104 可以组合框 704 的结果(与单个结果集相关联)以提供训练集。如已经讨论的,图 5 示出了如何执行此操作。

[0084] 虽然框 704 被示为与框 702 分开,以有助于说明,但是,框 702 和 704 可以作为集成的操作来执行。此外,框 702 和 704 的过滤和匹配操作可以分布到操作的多个阶段。例如,挖掘系统 104 可以对结果项遵循框 706 执行进一步的过滤。此外,训练系统 106 可以在其迭代过程的过程中对结果项执行进一步的过滤(如图 6 的框 614-618 所示)。

[0085] 作为另一种变体,框 704 是在单个结果集内建立结果项对的上下文中描述的。然而,在另一种模式下,挖掘系统 104 可以跨不同的结果集建立候选对。

[0086] 图 8 示出了描述翻译模型 102 的说明性应用的过程 800。

[0087] 在框 802 中,应用模块 108 接收输入短语。

[0088] 在框 804 中,应用模块 108 使用翻译模型 102 来将输入短语转换为输出短语。

[0089] 在框 806 中,应用模块 108 基于输出短语来生成输出结果。不同的应用模块可以提供不同的相应的输出结果以实现不同的相应的优点。

[0090] 在一种情况下,应用模块 108 可以使用翻译模型 102 来执行查询修改操作。这里,应用模块 108 可以将输入短语当作搜索查询。应用模块 108 可以使用输出短语来替换或补充搜索查询。例如,如果输入短语是“shingles”,则应用模块 108 可以使用输出短语“zoster”来生成补充查询“shingles AND zoster”。然后,应用模块 108 可以向搜索引擎呈现扩展的查询。

[0091] 在另一种情况下,应用模块 108 可以使用翻译模型 102 来编制索引分类决策。这里,应用模块 108 可以从要被分类的文档中提取任何文本内容,并将该文本内容当作输入短语。应用模块 108 可以使用输出短语来收集关于文档的主题的额外的洞察,该洞察又可用于提供文档的适当的分类。

[0092] 在另一种情况下,应用模块 108 可以使用翻译模型 102 来执行任何类型的文本修改操作。这里,应用模块 108 可以将输入短语当作文本修改的候选。应用模块 108 可以使用输出短语来建议可以修改输入短语的方式。例如,假设输入短语对应于相当冗长的文本“rash that is painful”。应用模块 108 可以建议可以将此输入短语替换为比较简洁的“painful rash”。在提出此建议时,应用模块 108 可以纠正原始短语中的任何语法上的和/或拼写错误(假设输出短语不包含语法上的和/或拼写错误)。在一种情况下,应用模块 108 可以给用户提供关于他或她如何修改输入短语的多种选择,与允许用户估计不同的修改的适当性的一些类型的信息耦合。例如,应用模块 108 通过指出这种表达您的思想的方

式被 80% 的作者使用（只引用一个代表性的示例）来注解特定修改。可另选地，应用模块 108 可以基于一个或多个考虑因素来自动作出修改。

[0093] 在另一种文本修改情况下，应用模块 108 可以使用翻译模型 102 来执行文本截断操作。例如，应用模块 108 可以接收原文本，以便在诸如移动电话设备等等之类的小屏幕查看设备上呈现。应用模块 108 可以使用翻译模型 102 来将被视为输入短语的文本转换为文本的缩写的版本。在另一种情况下，应用模块 108 可以使用此方法来缩短原始短语，以使得它与诸如类似于 Twitter 的通信机制之类的对其消息施加大小约束的任何消息传输机制兼容。

[0094] 在另一种文本修改情况下，应用模块 108 可以使用翻译模型 102 来概括一个或短语。例如，应用模块 108 可以使用此方法来缩短原始摘要的长度。在另一种情况下，应用模块 108 可以使用此方法来基于文本的较长的片断来建议标题。可另选地，应用模块 108 可以使用翻译模型 102 来扩展文档或短语。

[0095] 在另一种情况下，应用模块 108 可以使用翻译模型 102 来执行广告信息的扩展。这里，例如，广告商可能已经选择了与广告内容（例如，网页或其他可通过网络访问的内容）相关联的初始触发关键字。如果最终用户输入了这些触发关键字，或者如果用户以别的方式正在消费与这些触发关键字相关联的内容，则广告机制可以将用户定向到与触发关键字相关联的广告内容。这里，应用模块 108 可以将初始触发关键字集合视为要使用翻译模型 102 扩展的输入短语。可另选地，或另外，应用模块 108 可以将广告内容本身当作输入短语。然后，应用模块 108 可以使用翻译模型 102 来建议与广告内容关联的文本。广告商可以基于建议的文本来提供一个或多个触发关键字。

[0096] 上文所描述的应用是代表性的，且非穷尽性的。其他应用也是可以的。

[0097] 在上面的讨论中，作出了输出短语以与输入短语相同的语言来表示的假设。在此情况下，输出短语可以被视为输入短语的释义。在另一种情况下，挖掘系统 104 和训练系统 106 可用于产生将第一语言的短语转换为另一种语言（或多种其他语言）的对应的短语的翻译模型 102。

[0098] 为在双语或多语言上下文中操作，挖掘系统 104 可以如上文参考双语或多语言信息所描述的相同基本操作。在一种情况下，挖掘系统 104 可以通过在网络环境内提交并行查询来建立双语结果集。即，挖掘系统 104 可以提交一个组以第一语言表示的查询和另一组以第二语言表示的查询。例如，挖掘系统 104 可以提交短语“rash zoster”来生成英语结果集，提交短语“zoster erupción de piel”来生成英语结果集的西班牙语对应结果集。然后，挖掘系统 104 可以建立将英语结果项链接到西班牙语结果项的对。这种匹配操作的目标是提供允许训练系统 106 标识用英语和西班牙语表示的语义相关的短语之间的链接的训练集。

[0099] 在另一种情况下，挖掘系统 104 可以提交组合了英语和西班牙语关键术语的查询，诸如在查询“shingles rash erupción de piel”的情况下。在此方法中，可以预计检索模块 116 提供组合了以英语表示的结果项和以西班牙语表示的结果项的结果集。然后，挖掘系统 104 可以在此混合结果集中的不同的结果项之间建立链接，无需鉴别结果项是以英语表示的还是以西班牙语表示的。训练系统 106 可以基于混合训练集中的底层模式来生成单一翻译模型 102。在使用中，可以在单语种模式下应用翻译模型 102，在这种模式下，翻译

模型被约束为以与输入短语相同语言生成输出短语。或者,翻译模型 102 可以在双语模式下操作,在这种模式下,翻译模型被约束为以与输入短语相比不同的语言的生成输出短语。或者,翻译模型 102 可以在无约束的模式下操作,在无约束的模式下,翻译模型同时提供两种语言的结果。

[0100] 代表性的处理功能

[0101] 图 9 阐述了可以被用来实现上文所描述的功能的任何方面的说明性电气数据处理功能 900。参考图 1 和 2,例如,图 9 所示出的处理功能 900 的类型可用于实现系统 100 或计算功能 202 的任何方面,等等。在一种情况下,处理功能 900 可以对应于包括一个或多个处理设备的任何类型的计算设备。

[0102] 处理功能 900 可包括诸如 RAM 902 和 ROM 904 之类的易失性和非易失性存储器,以及一个或多个处理设备 906。处理功能 900 还可任选地包括诸如硬盘模块、光盘模块等等之类的各种介质设备 908。当处理设备 900 执行由存储器(例如, RAM 906、ROM 902 或在别处)维护的指令时,处理功能 904 可以执行上文所标识的各种操作。更一般而言,指令及其他信息可以存储在任何计算机可读介质 910 上,包括,但不仅限于,静态存储器存储设备、磁存储设备、光存储设备等等。术语“计算机可读介质”还包含多个存储设备。术语“计算机可读介质”还包含,例如,通过有线、电缆、无线传输等等,从第一位置传输到第二位置的信号。

[0103] 处理功能 900 还包括用于接收来自用户的各种输入(通过输入模块 912),以及用于向用户提供各种输出(通过输出模块)的输入/输出模块 914。一个特定的输出机制可包括呈现模块 916 和相关联的图形用户界面(GUI)918。处理功能 900 还可以包括用于通过一个或多个通信管道 920 与其他设备交换数据的一个或多个网络接口 922。一个或多个通信总线 924 可通信地将上文所描述的组件耦合在一起。

[0104] 尽管用结构特征和/或方法动作专用的语言描述了本主题,但可以理解,所附权利要求书中定义的主题不必限于上述具体特征或动作。相反,上述具体特征和动作是作为实现权利要求的示例形式公开的。

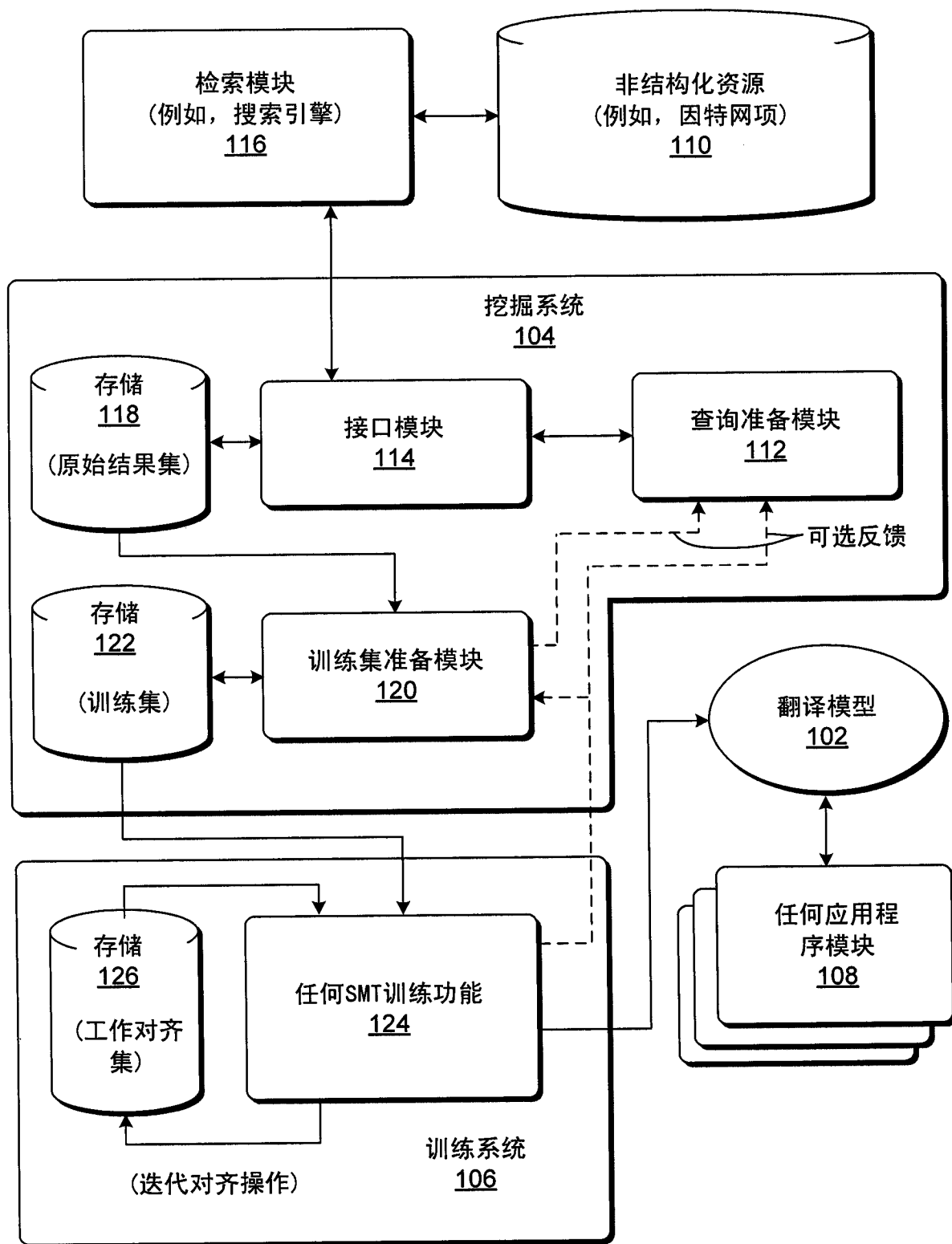


图 1

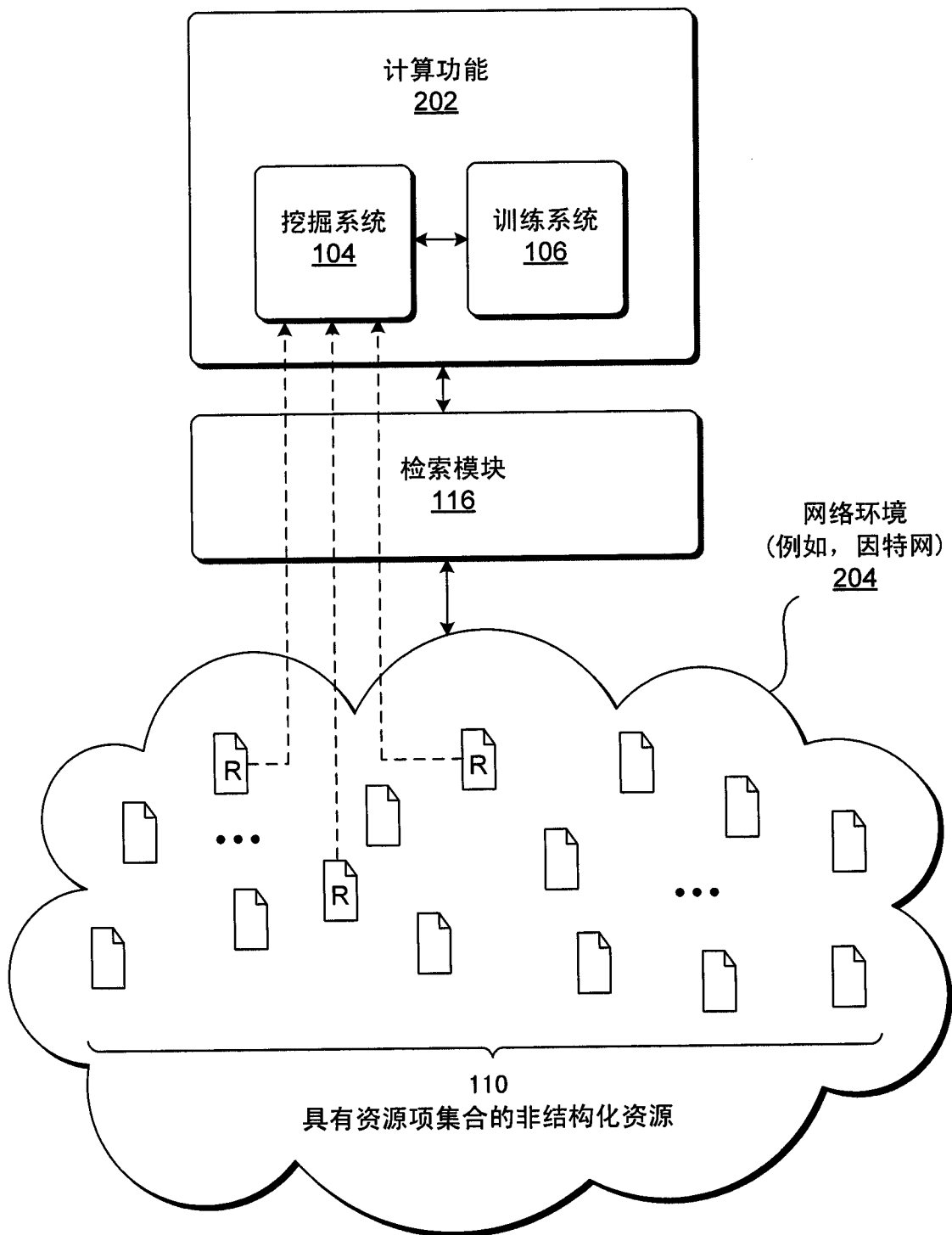


图 2

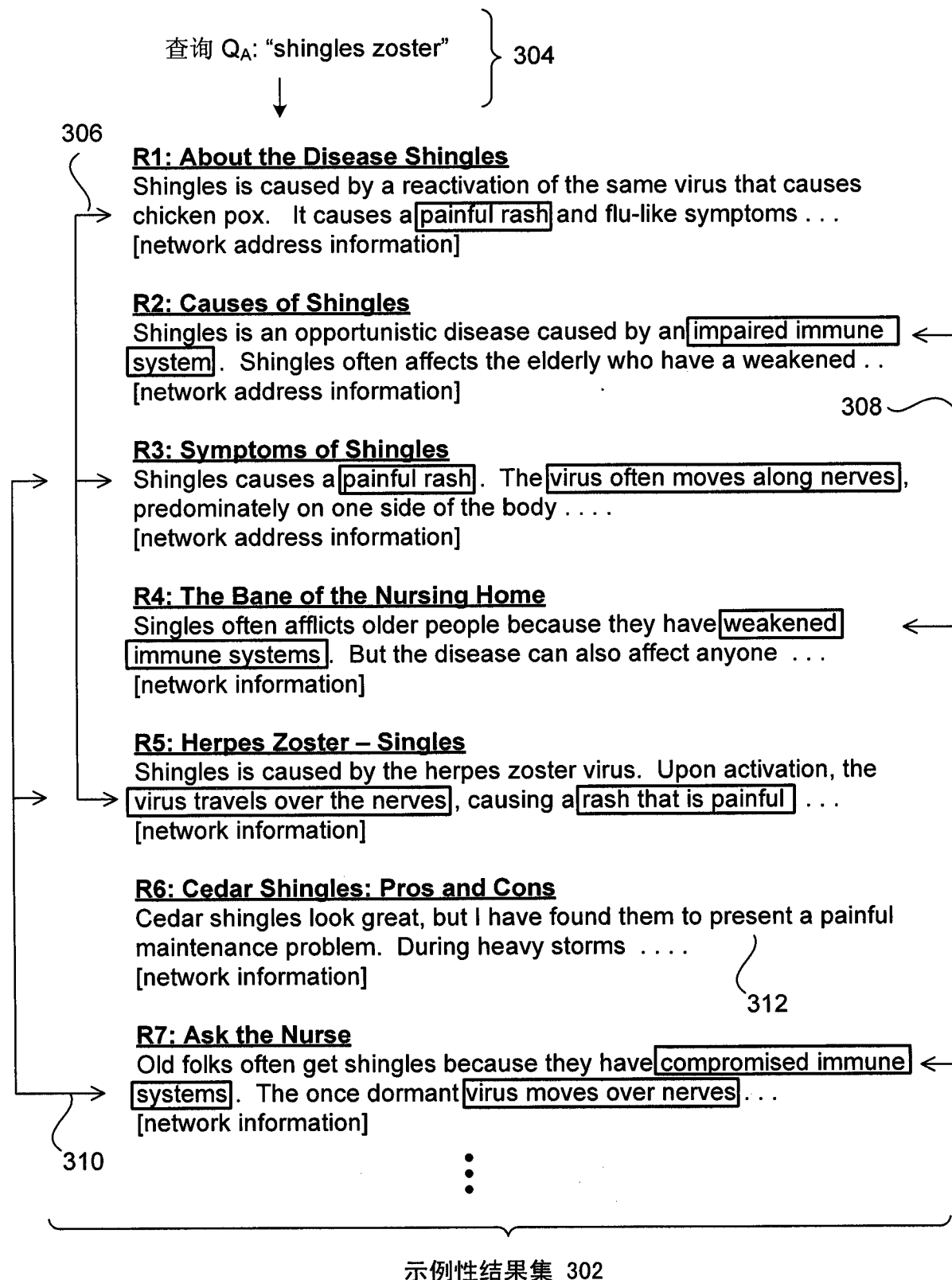


图 3

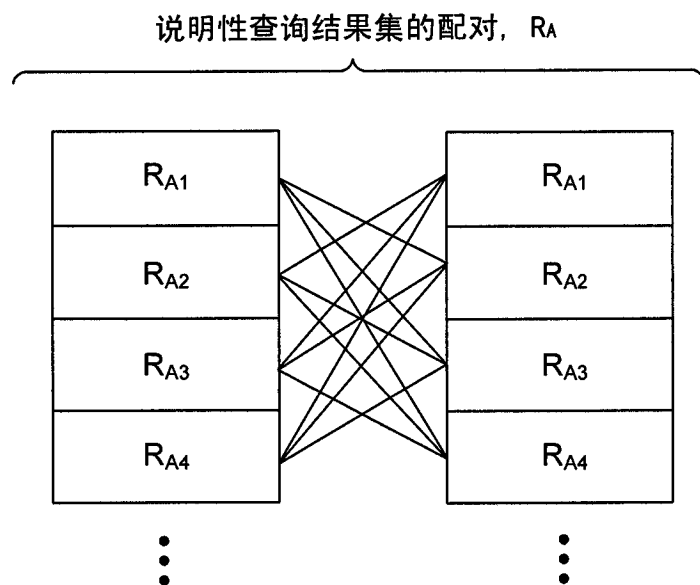


图 4

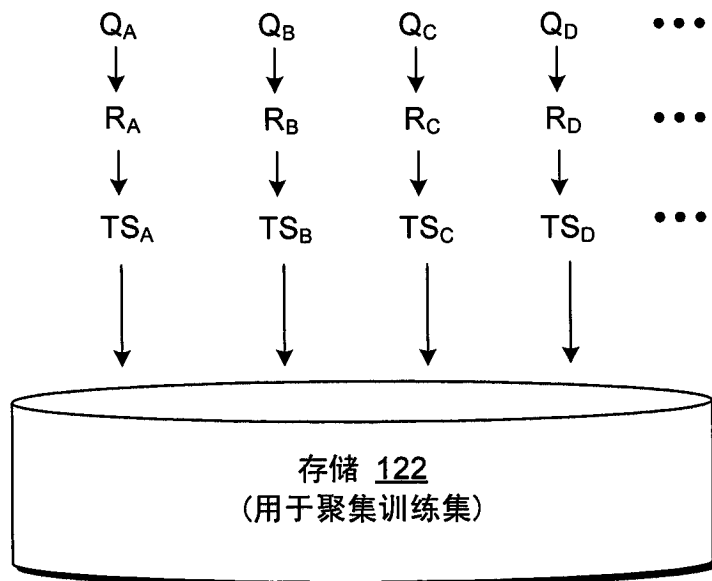


图 5

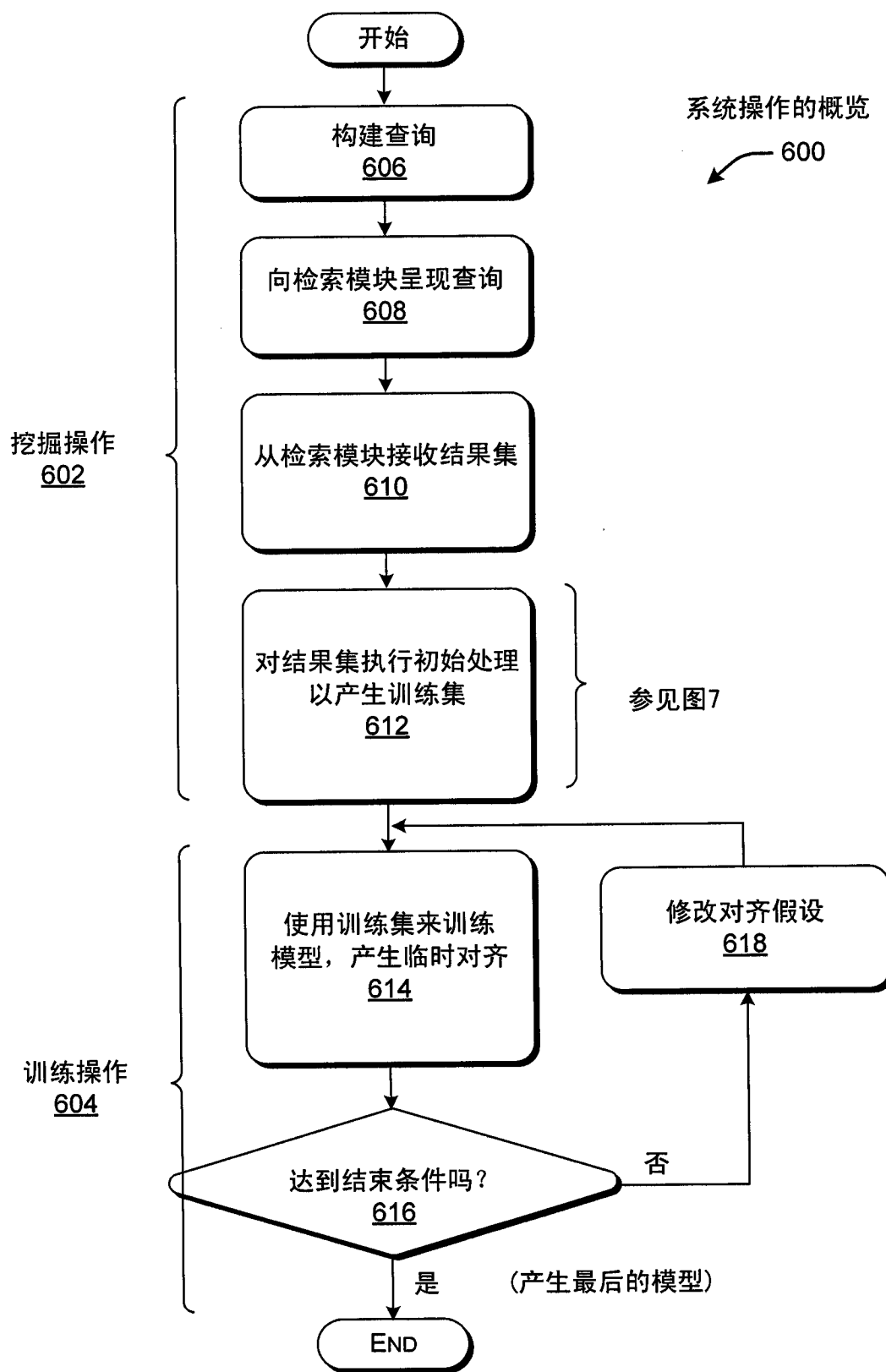


图 6

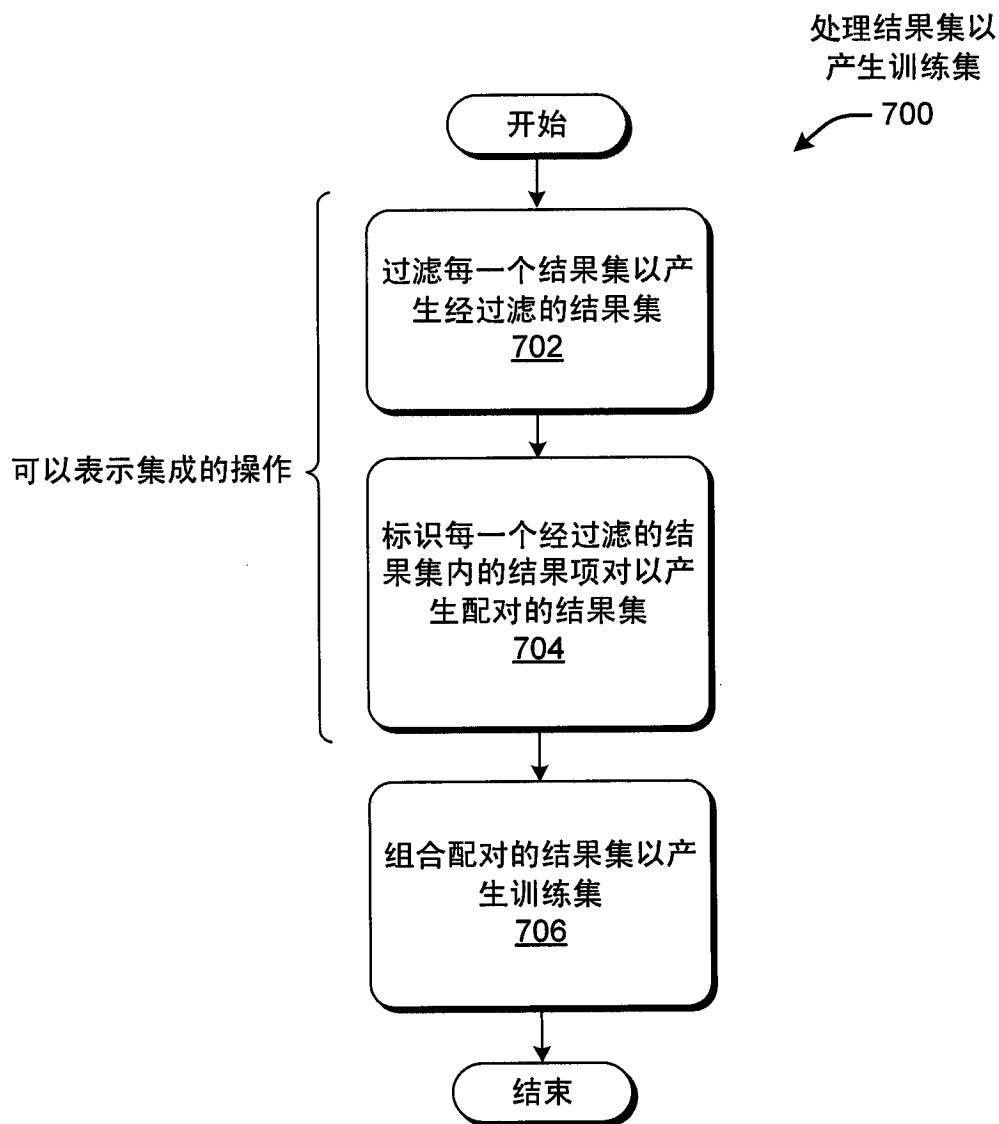


图 7

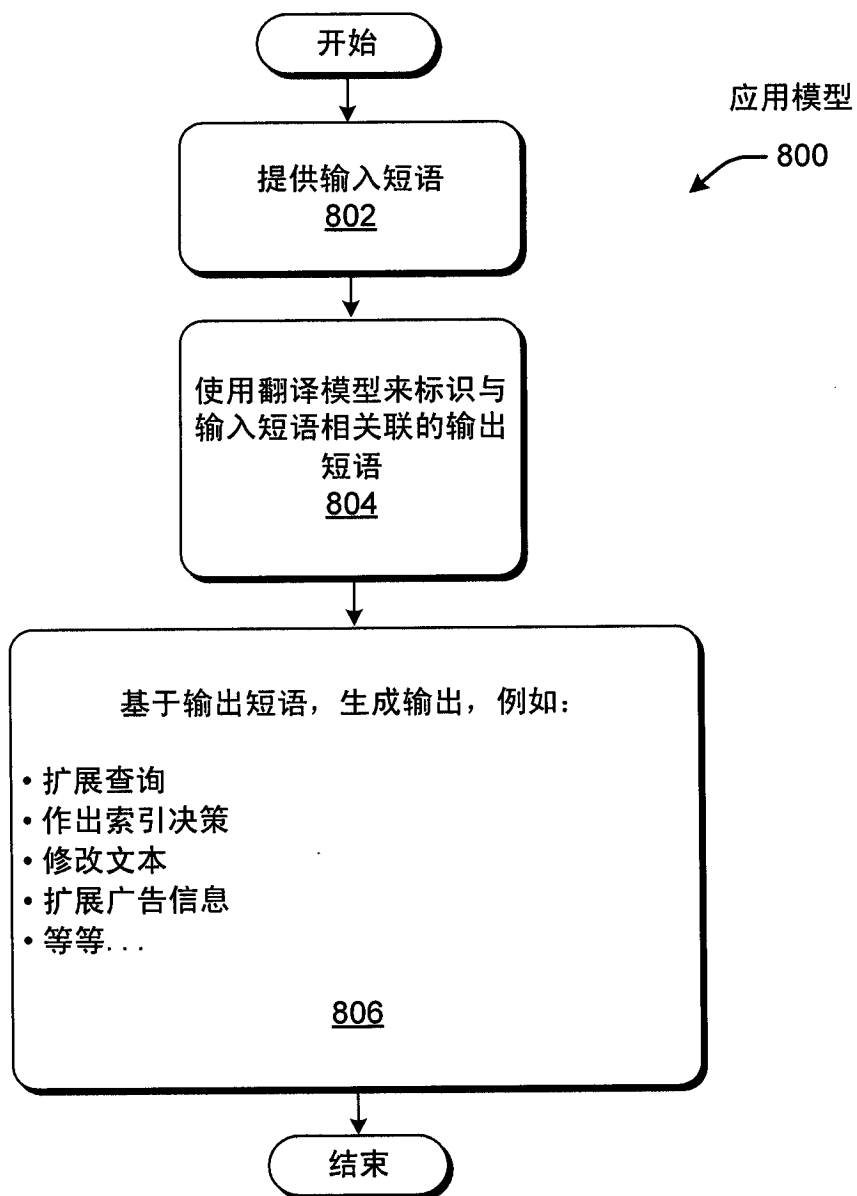


图 8

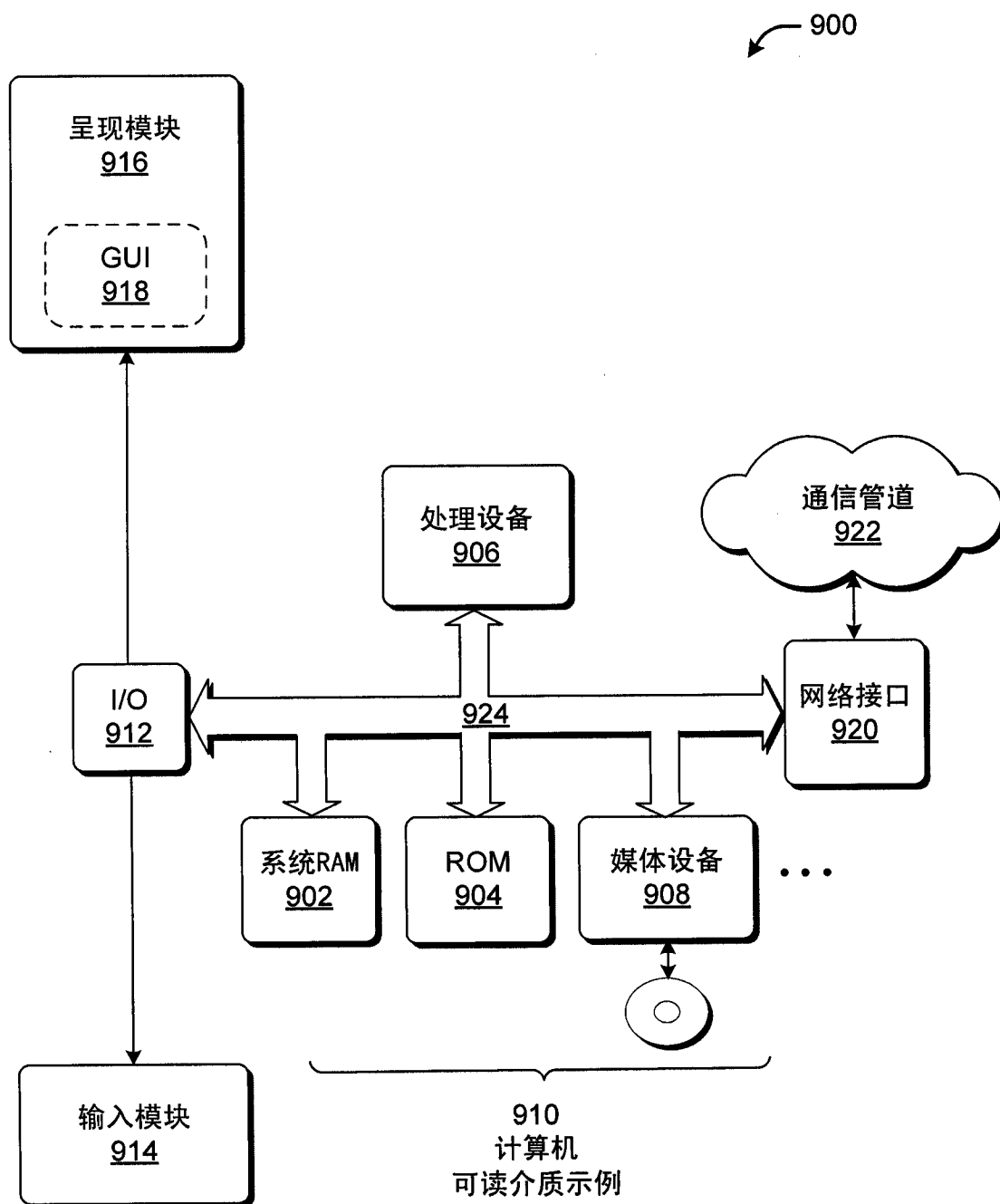


图 9