(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101796508 B
(45)授权公告日 2013.03.06

(21)申请号 200880105563.X

(22)申请日 2008.08.29

(30)优先权数据
60/969,426 2007.08.31 US
60/969,483 2007.08.31 US
12/200,962 2008.08.29 US

(85)PCT申请进入国家阶段日
2010.03.01

(86)PCT申请的申请数据
PCT/US2008/074935 2008.08.29

(87)PCT申请的公布数据
WO2009/029903 EN 2009.03.05

(73)专利权人 微软公司
地址 美国华盛顿州

(72)发明人 M·范登伯格 R·克鲁奇 F·萨尔维蒂 G·L·蒂奥内 D·安

(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 陈斌

(51)Int.Cl.
G06F 17/20(2006.01)
G06F 17/27(2006.01)

(56)对比文件
US 6185592 B1,2001.02.06, 第5栏第8行-第8栏第64行, 第11栏权利要求1-第12栏权利要求7、说明书附图2.
US 2001/0014852 A1,2001.08.16, 全文.
CN 1898670 A,2007.01.17, 全文.

审查员 刘曼

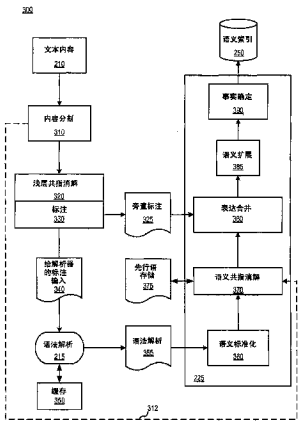
权利要求书 2 页 说明书 10 页 附图 5 页

(54)发明名称

歧义敏感自然语言处理系统中的共指消解

(57)摘要

本文中描述了歧义敏感自然语言处理系统中的共指消解的技术。将指代消解功能集成到自然语言处理系统中的技术可以处理要在信息搜索和检索系统内索引的文档。歧义感知特征以及歧义消解功能可以和共指消解协同操作。可以通过文本内容中的内嵌标记或通过外部实体映射来支持共指实体的标注以及歧义解释。文档内表达的信息可以在形式上按事实、或文本中实体之间的关系来组织。扩展可以支持将多个别名或歧义应用到进行索引的实体从而将该实体的所有可能的指代或解释捕获到索引中。可选的存储的描述可以支持通过原始描述或共指描述来检索事实。



CN 101796508 B

1. 一种用于集成共指消解机制的方法，所述方法包括：

检索文本的一部分；

识别所述文本的部分中的共指；

从所述文本的部分中提取事实，所述事实表示实体之间的关系；及

基于识别的共指扩展所述事实以包括共指含义。

2. 如权利要求 1 所述的方法，其特征在于，识别共指结合来自语法解析的信息。

3. 如权利要求 1 所述的方法，其特征在于，识别共指结合来自语义映射的信息。

4. 如权利要求 1 所述的方法，其特征在于，识别共指包括识别歧义共指。

5. 如权利要求 1 所述的方法，其特征在于，还包括识别所述文本的部分中的歧义。

6. 如权利要求 5 所述的方法，其特征在于，还包括基于识别的歧义扩展所述事实以包括有歧义的含义。

7. 如权利要求 1 所述的方法，其特征在于，还包括将经扩展的事实存储到可操作以支持信息检索的索引中。

8. 如权利要求 7 所述的方法，其特征在于，还包括响应于搜索查询从所述索引中检索经扩展的事实。

9. 如权利要求 1 所述的方法，其特征在于，还包括在所述文本的部分中标注识别的共指。

10. 如权利要求 2 所述的方法，其特征在于，还包括缓存来自所述语法解析的信息。

11. 一种用于集成共指消解机制的系统，所述系统包括：

用于检索文本的一部分的装置；

用于识别所述文本的部分中的共指的装置；

用于从所述文本的部分中提取事实的装置，所述事实表示实体之间的关系；及

用于基于识别的共指扩展所述事实以包括共指含义的装置。

12. 如权利要求 11 所述的系统，其特征在于，识别共指结合来自语法解析的信息。

13. 如权利要求 11 所述的系统，其特征在于，识别共指结合来自语义映射的信息。

14. 如权利要求 11 所述的系统，其特征在于，用于识别共指的装置包括用于识别歧义共指的装置。

15. 如权利要求 11 所述的系统，其特征在于，还包括用于识别所述文本的部分中的歧义的装置。

16. 如权利要求 15 所述的系统，其特征在于，还包括用于基于识别的歧义扩展所述事实以包括有歧义的含义的装置。

17. 如权利要求 11 所述的系统，其特征在于，还包括用于将经扩展的事实存储到可操作以支持信息检索的索引中的装置。

18. 如权利要求 17 所述的系统，其特征在于，还包括用于响应于搜索查询从所述索引中检索经扩展的事实的装置。

19. 如权利要求 11 所述的系统，其特征在于，还包括用于在所述文本的部分中标注识别的共指的装置。

20. 一种用于集成共指消解机制的方法，所述方法包括：

检索文本的一部分；

识别所述文本的部分中的共指；

识别所述文本的部分中的歧义；

从所述文本的部分中提取事实，所述事实表示实体之间的关系；

基于识别的共指扩展所述事实以包括共指含义；

将经扩展的事实存储到可操作以支持信息检索的索引中；及

响应于搜索查询从所述索引中检索经扩展的事实。

歧义敏感自然语言处理系统中的共指消解

[0001] 背景

[0002] 在自然语言中，用不同的描述来指代实体并不少见。例如，代词常用于代替名词。此外，各种其他描述或不同形式的指代可用于指代实体。作为示例考虑下述文本片段：

[0003] “Pablo Picasso was born in Malaga(巴勃罗·毕加索出生在马拉加)”

[0004] “The Spanish painter became famous for his varied styles(这位西班牙画家以其多变的风格著称)”

[0005] “Among his paintings is the large-scale Guernica(在他的画作中有大幅的格尔尼卡)”

[0006] “He painted this disturbing masterpiece during the Spanish Civil War(他在西班牙内战期间画出了这幅令人不安的杰作)”

[0007] “Picasso died in 1973(毕加索逝世于1973年)”

[0008] 出现了多种语言变体。例如，使用了两个不同的名字：“Pablo Picasso”和“Picasso”。限定性描述“the Spanish painter”和两个代词“his”和“he”都用于指毕加索。两个不同的表达用于指代画作：该作品的名称“Guernica”和说明性描述“this disturbing masterpiece”。

[0009] 两个语言表达如果有相同的指代对象则可称为共指。换句话说，如果两者指代相同的实体则可称为共指。第二短语可以是回指第一短语的回指语。因此，第一短语是第二短语的先行语。确定回指语的指代对象可能需要知道先行语的指代对象。在文档内查找共指表达、回指语及其先行语的一般任务可称为共指消解。共指消解是确定两个表达指代相同的指代对象而不必确定该指代对象为何的处理。指代消解则是确定指代对象为何的处理。

[0010] 对于共指表达的集群，无论其回指关系如何，这些表达都可称为彼此的别名。根据上述示例，表达“Pablo Picasso”、“the Spanish painter”、“his”、“he”和“Picasso”形成了指代毕加索的别名集群。

[0011] 自然语言表达通常显示出歧义。当表达可以解释为一种以上含义时会出现歧义。例如，句子“The duck is ready to eat(鸭子可以吃了/鸭子要吃了)”可以解释为表明鸭子已烹饪好，或鸭子饿了需要喂食。

[0012] 共指消解和歧义消解是可用于机械地支持人类使用者通常表达的语言的自然语言处理操作的两个示例。信息处理系统，诸如支持信息搜索的文本索引和查询，可以得益于自然语言处理系统的更多应用。

[0013] 基于这些以及其他考虑在此作出本公开。

[0014] 概要

[0015] 本文描述歧义敏感自然语言处理系统中的共指消解的技术。特别是，描述了将共指消解功能集成到用于处理要索引到信息搜索和检索系统中的文档的系统的技术。该集成可以在自然语言文档内用支持共指消解的信息和有歧义的含义增强索引

[0016] 根据本发明的一个方面，共指消解系统所提供的信息可以集成到自然语言处理系统中并提高自然语言处理系统的性能。这样的系统的一个示例是文档索引和检索系统。

[0017] 根据本发明的另一个方面，歧义感知特征以及歧义消解功能可以协同自然语言处理系统中的共指消解进行操作。共指实体的标注以及有歧义的解释可以由文本表达中的内嵌标记支持或替代地由外部实体映射支持。

[0018] 根据本发明的又一个方面，可以从要索引的文本中提取事实。文本中表达的信息可以在形式上按事实来组织。用在此意义上时，事实可以是文本中包含的任何信息，且不必是真实的。事实可以表示为实体之间的关系。事实可以作为存储在索引中的实体之间的关系存储在语义索引中。在基于事实的检索系统中，如果文档包含与通过对查询的解析确定的事实相匹配的事实，则可以检索到该文档。

[0019] 根据本发明的再一个方面，扩展处理可以支持将多重别名或歧义应用到进行索引的实体。这样的扩展可以支持对捕获到语义索引中的给定实体提供附加的可能指代或解释。可选的存储的描述可以支持通过原始描述或共指描述来检索事实。

[0020] 应理解，上述主题也可以实现为计算机控制的装置、计算机处理、计算机系统，或实现为制品，如计算机可读介质。这些和各种其他特征将通过阅读下面的详细说明和参照相关附图变得明显。

[0021] 提供本概要以便用简化的形式引入下面在详细说明中进一步描述的选择的概念。本概要不意图要求保护的主题的识别关键特征或核心特征，也不意图将本概要用于限制要求保护的主题的范围。此外，要求保护的主题不限于解决在本公开的任何部分提及的任何或所有缺点的实施方式。

[0022] 附图简述

[0023] 图 1 是示出根据本发明实施例的各方面的信息搜索系统的网络架构图；

[0024] 图 2 是示出根据本发明实施例的各方面的自然语言索引和查询系统的各个组件的功能框图；

[0025] 图 3 是示出根据本发明实施例的各方面的自然语言处理系统中的共指消解和歧义消解的功能框图；

[0026] 图 4 是示出根据本发明实施例的各方面对共指消解进行歧义敏感索引的处理的各方面的逻辑流程图；及

[0027] 图 5 是示出能够实现本发明实施例的各方面的计算系统的示意性计算机硬件和软件架构的计算机架构图。

[0028] 详细说明

[0029] 下面的详细说明涉及歧义敏感自然语言处理系统中的共指消解的技术。通过使用本发明的技术和概念，共指消解功能可以被集成到对要索引的以供在信息搜索和检索系统中使用的文档进行处理的自然语言处理系统中。该集成可以用支持对要索引的自然语言文档进行共指消解的信息来增强索引。

[0030] 虽然在与计算机系统上的操作系统和应用程序的执行结合执行的程序模块的一般上下文中描述本文所述的主题，本领域技术人员应理解，其他实施方式可以结合其他类型的程序模块实现。一般地，程序模块包括例程、程序、组件、数据结构，及执行特定任务或实现特定抽象数据类型的其他类型的结构。此外，本领域技术人员应理解，本文所述主题可以用其他计算机系统配置实现，包括手持设备、多处理器系统、基于微处理器的或可编程的消费者电子设备、小型机、大型机等等。

[0031] 在下面的详细说明中，参考构成说明书的部分的附图，附图示意性地示出具体实施例或示例。现参考附图，其中类似的标号在多个附图中表示类似的元素，描述用于歧义敏感自然语言处理系统中的共指消解的计算系统和方法的各方面。

[0032] 现参考图 1，提供关于用于实现本发明的示意性操作环境的细节。特别地，网络架构图 100 示出根据本发明实施例的各方面的信息搜索系统。客户机计算机 110A-110D 可以通过网络 140 连接到服务器 120 以获取与自然语言引擎 130 关联的信息。虽然示出了四个客户机计算机 110A-110D，应理解可以使用任意数量的客户机计算机 110A-110D。客户机计算机 110A-110D 可以在地理上分布于网络 140 上、设在同一位置，或其任意组合。虽然示出了单个服务器 120，应理解服务器 120 的功能可以分布在任意数量的多个服务器 120 中。这样的多个服务器 120 可以设在同一位置、地理上分布于网络 140 上，或其任意组合。

[0033] 根据一个或多个实施例，自然语言引擎 130 可以支持搜索引擎功能。在搜索引擎情景中，用户查询可以从客户机计算机 110A-110D 通过网络 140 提交到服务器 120 上。用户查询可以为自然语言格式。在服务器处，自然语言引擎 130 可以处理自然语言查询以基于从自然语言查询提取的语法和语义支持搜索。这样的搜索的结果可以从服务器 120 通过网络 140 发送回客户机计算机 110A-110D。

[0034] 一个或多个搜索索引可以存储在服务器 120 处或关联于服务器 120。搜索索引中的信息可以由一组源信息或语料库填充。例如，在 Web 搜索实现中，可以从网络 140 上的各种 Web 服务器（未示出）上的各种 Web 站点收集和索引内容。这样的收集和索引可以由服务器 120 上或另一个计算机上（未示出）执行的软件执行。收集可以由 Web 爬行器或蜘蛛应用程序执行。自然语言引擎 130 可以应用于所收集的信息，以便基于自然语言引擎 130 提取的语法和语义对从语料库中收集的自然语言内容进行索引。索引和搜索将参考图 2 更详细地说明。

[0035] 客户机计算机 110A-110D 可以作为服务器 120 的终端客户机、超文本浏览器客户机、图形显示客户机，或其他联网客户机。例如，客户机计算机 110-110D 处的 Web 浏览器应用程序可以支持与服务器 120 处的 Web 服务器应用程序连接。这样的浏览器可以使用控件、插件，或小程序来支持连接到服务器 120。客户机计算机 110A-110D 也可以使用其他定制程序、应用，或模块与服务器 120 连接。客户机计算机 110A-110D 可以是桌面计算机、膝上型计算机、手持设备、移动终端、移动电话、电视机顶盒、网亭、服务器、终端、瘦客户机，或任意其他计算机化的设备。

[0036] 网络 140 可以是能够支持客户机计算机 110A-110D 和服务器 120 之间的通信的任何通信网络。网络 140 可以是有线网络、无线网络、光网络、无线电网络、分组交换网络、电路交换网络，或其任意组合。网络 140 可以使用任何拓扑结构，且网络 140 的链接可以支持任何联网技术、协议或带宽，如以太网、DSL、电缆调制解调器、ATM、SONET、MPLS、PSTN、POTS 调制解调器、PONS、HFC、卫星、ISDN、WiFi、WiMax、移动蜂窝，或其任意组合，或任何其他数据互连或联网机制。网络 140 可以是内联网、互联网、因特网、万维网、LAN、WAN、MAN，或用于互连计算机系统的任意其他网络。

[0037] 应理解，除了所示网络环境，自然语言引擎 130 可以本地地运行。例如，服务器 120 和客户机计算机 110A-110D 可以组合到单个计算设备上。这样的组合系统可以支持本地或远程存储的搜索索引。

[0038] 现参考图 2，功能框图示出根据一个示例实施例的自然语言引擎 130 的各个组件。如上所述，自然语言引擎 130 可以支持信息搜索。为了支持这样的搜索，执行内容获取处理 200。与内容获取 200 相关的操作从作为文本内容 210 提供的文档中提取信息。该信息可被存储在可用于搜索的语义索引 250 中。与用户搜索 205 相关的操作可以支持处理用户输入的搜索查询。用户查询可以采取自然语言问题 260 的形式。自然语言引擎 130 可以分析用户输入以将查询转换为要与语义索引 250 中表示的信息相比较的表示。语义索引 250 中的信息的内容和结构可以支持快速匹配和检索与查询的含义或自然语言问题 260 相关的文档或文档部分。

[0039] 文本内容 210 可以包括非常宽泛意义的文档。这样的文档的示例可以包括网页、文本文档、扫描文档、数据库、信息列表、其他因特网内容，或任意其他信息源。该文本内容 210 可以提供被搜索的信息语料库。处理文本内容 210 可以在语法解析 215 和语义映射 225 两个阶段中进行。初步语言处理步骤可以在语法解析 215 之前或开始时进行。例如，可以在句子边界处分割文本内容 210。专有名词可以识别为特定人、地点、对象或事件的名字或名称。此外，可以确定有含义的词尾的语法属性。例如，在英语中，以“s”结尾的名词可能是复数名词，而以“s”结尾的动词可能是第三人称单数动词。

[0040] 解析 215 可以由语法分析系统执行，如 Xerox 语言环境 (XLE)，该环境在此仅作为一般示例提供，而不限制本发明的可能的实施方式。解析器 215 可以将句子转换为明确单词之间的语法关系的表示。解析器 215 可以应用与所使用的特定语言关联的语法 220。例如，解析器 215 可以应用英语的语法 220。语法 220 可以形式化为例如词汇功能语法 (LFG) 或其他适合的解析机制，如基于中心语驱动短语结构语法 (HPSG)、组合范畴语法 (CCG)、概率上下文无关语法 (PCFG) 或任何其他语法形式的解析机制。语法 220 可以指明构建给定语言中有含义的句子的可能方式。解析器 215 可以将语法 220 的规则应用于文本内容 210 中的字符串。

[0041] 可以为各种语言提供语法 220。例如，已为英语、法语、德语、汉语和日语创建了 LFG 语法。可以通过手工获取来发展语法 220，其中语法规则由语言学家或词典编者定义。或者，机器学习获取可以涉及自动观察和分析来自大语料库的大量文本示例以自动地确定语法规则。手工定义和机器学习的组合也可以用于获取语法 220 的规则。

[0042] 解析器 215 可以应用语法 220 到文本内容 210 以确定语法结构。在基于 LFG 的解析情况下，语法结构由成分结构 (c- 结构) 和功能结构 (f- 结构) 组成。c- 结构可以表示成分短语和单词的分层结构。f- 结构可以编码 c- 结构的各种成分之间的角色和关系。f- 结构还可以表示从单词的形式派生的信息。例如，可以在 f- 结构中指明名词的复数或动词的时态。

[0043] 在解析处理 215 之后的语义映射处理 225 期间，可以从语法结构中提取信息并将其与有关单词在句子中的含义的信息组合。可以提供句子的语义映射或语义表示以作为内容语义 240。语义映射 225 可以用各个单词的概念属性增强解析器 215 提供的语法关系。结果可以被转换为来自文本内容 210 中的句子的含义的表示。语义映射 225 可以确定单词在句子中充当的角色。例如，执行动作的主体、用于执行该动作的事物，或受该动作影响的事物。为搜索索引的目的，单词及其角色可以被存储在语义索引 250 中。因此，从语义索引 250 中检索不仅取决于单独的单词，还取决于单词在文本内容 210 内出现的句子中的含义。

语义映射 225 可以支持词语的消歧、确定先行语关系，及通过同义词、上义词、或下义词扩展词语。

[0044] 语义映射 225 可以应用知识资源 230 作为用于从句子中提取语义的规则和方法。知识资源可以通过手工定义和机器学习两者获取，如对于语法 220 的获取所述。语义映射 225 处理可以提供语义可扩展标记语言（语义 XML 或 semxml）表示的内容语义 240。也可以使用任何适合的表示语言，如以 PROLOG、LISP、JSON、YAML 或其他语言写出的表达。内容语义 240 可以指明单词在文本内容 210 的句子中充当的角色。可以将内容语义 240 提供给索引处理 245。

[0045] 索引可以支持表示大语料库的信息从而可以在索引内快速识别单词和短语的位置。传统的搜索引擎可以使用关键词作为搜索项从而索引将用户指明的关键词映射到这些关键词出现的文章或文档。除了单词自身，语义索引 250 还可以表示单词的语义。可以在内容获取 200 和用户搜索 205 两者期间都向单词指派语义关系。针对语义索引 250 进行的查询不仅可以基于单词，还可以基于特定角色的单词。角色是单词在语义索引 250 内存储的句子或短语中充当的角色。语义索引 250 可以视为反向索引，反向索引是可快速搜索的数据库，其条目为语义单词（即，具有给定角色的单词）及指向这些单词所出现的文档或网页的指针。语义索引 250 可以支持混合索引。这样的混合索引可以结合关键词索引和语义索引两者的特征和功能。

[0046] 可以用自然语言问题 260 的形式支持查询的用户输入。可以通过与内容获取 200 中使用的自然语言管线类似或等同的自然语言管线解析查询。即，自然语言问题 260 可以由解析器 265 处理以提取语法结构。在语法解析 265 之后，可以处理自然语言问题 260 以进行语义映射 270。语义映射 270 可以提供在检索处理 280 中针对如上所述的语义索引 250 使用的问题语义 275。检索处理 280 可以支持混合索引查询，其中可以单独地或组合地提供关键词索引检索和语义索引检索两者。

[0047] 响应于用户查询，来自语义索引 250 的检索 280 的结果以及问题语义 275 可以通知评级处理 285。评级可以充分利用关键词和语义信息两者。在评级 285 期间，可以按各种量度对检索 280 获得的结果排序以尝试使最合乎需要的结果更加接近要作为结果呈现 290 提供给用户的检索到的信息的顶部。

[0048] 现参考图 3，功能框图示出根据本发明实施例的各方面的自然语言处理系统 300 内的共指消解和歧义消解。作为示例应用，自然语言处理系统 300 可以支持用于文档索引和检索的信息搜索引擎。这样的自然语言支持的搜索引擎可以基于语言学分析扩展存储在其索引内的信息。该系统还可以支持用语言学方式分析查询以发现用户查询中的意图。本文所述的共指消解和歧义消解特征可以相关于语法解析 215、语义映射 225 和语义索引 245 进行操作，如参考图 2 所述。共指消解可以对文本内容 210 直接执行，或使用来自解析 215 或语义映射 225 操作的信息。

[0049] 如图所示，共指消解 320、370 可以直接对分割的文档执行，还可以作为语义映射 225 的一部分执行。共指消解 320、370 的这两次出现可以合并或其信息输出可以合并。应理解，共指消解还可以出现在语法解析 215 和语义映射 225 之间。共指消解还可以出现在自然语言处理管线中的任何其他阶段。在自然语言处理系统内的各种位置可以有一个、两个或多个共指消解组件或阶段。可以分析文本内容 210 以得到要存储到语义索引 250 中的

信息。搜索可以涉及查询语义索引 250 以得到期望的信息。

[0050] 可以对组成文本内容 210 的文档执行内容分割 310。可以分割文档以实现更加高效且可能更准确的共指消解 320。共指消解 320 可以在整个文档中考虑可能的指代关系。对于长文档，大量时间可能用于比较远距离表达。当考虑处理速度时，在共指消解 320 之前进行文档的内容分割 310 可以显著减少用于处理的时间。内容分割 310 可以有效地减少进行探索以尝试共指消解 320 的内容文本 210 的量。

[0051] 内容分割 310 可以向语义共指消解 370 提供信息以指示新文件片段何时开始。可以作为分割信号 312 或通过将标记插入到内容文档片段中提供这样的信息。也可以使用包含元信息的外部文件或其他机制。

[0052] 文档的结构可以用于识别指代关系不太可能跨过的片段边界。文档结构可以通过明确的标记如段落边界、章，或通过章节标题来推断。文档结构也可以通过语言学处理发现。超过指定长度的片段可以进一步分割为子片段。期望的子片段长度可以例如按句子的数量或单词的数量表示。

[0053] 在没有可靠的文档结构时，可以应用启发式准则或统计准则。可以指明这样的准则以倾向于将共指放在一起同时将片段大小限制为预定的最大值。也可以应用分割文本内容 210 文档的各种其他方法。内容分割 310 还可以指定整个文档作为一个片段。

[0054] 共指消解 320、370 可以用于识别内容文本 210 中的共指和别名。例如，在索引句子“He painted Guernica（他画出了格尔尼卡）”时，关键是确定“he”指代毕加索。在使用基于事实的检索时尤其如此。消解毕加索的代词别名可以支持索引毕加索画出了格尔尼卡这一事实，而不是索引用处不大的事实：某位男性“he”画出了格尔尼卡。没有这样的识别和索引代词指代对象的能力，则难以使用基于事实的检索方法响应于查询“Picasso painted（毕加索画出）”检索到该文档。当可返回其他方式不能返回的与查询相关的文档时，系统的检索能力得到提高。

[0055] 可以将标注 330 应用于文本内容 210 以支持跟踪实体和可能的共指关系。也可以在文本内容 210 内标注或标记消解判断中的置信值。可以通过将明确的标注记号添加到文本中来记录消解判断。例如，给定文本“John visited Mary.Hemet her in 2003”。可以这样应用标注 330：“[E1 ;0.9John]visited[E2 ;0.8Mary].[E1 ;0.9He]met[E2 ;0.8her]in 2003”。其中单词“John”和“He”可以相关以作为具有置信值 0.9 的实体一 E1。类似地，单词“Mary”和“her”可以相关以作为具有置信值 0.8 的实体二 E2。置信值可以指示共指消解 320 判断的置信度的度量。标注可以直接编码共指判断，或标注可以充当连接所标注文本中的相关词语与旁置标注 325 中的附加信息的标识符。

[0056] 共指消解 320 判断可以作为构建语义映射 225 的处理的一部分使用。共指消解 320 系统使用的指代表达可以通过文本内容 210 中的内嵌标注集成到语义映射 225 的输入表示中。也可以独立地在外部旁置实体映射 325 中提供指代。

[0057] 在大量文档集合的文本内容 210 中，如万维网，相同的句子可能在不同的上下文中出现多次。这些不同的上下文可以为共指消解 320 提供不同的候选。由于语法解析 215 的计算成本高，因此在缓存中保存句子的解析结果是有用的。这样的缓存机制 350 可以在将来碰到某句子时支持快速检索解析信息。

[0058] 如果共指消解 320 应用于出现在不同上下文中的单个句子，其可以识别相同指代

表达的不同共指关系，因为共指可以取决于上下文。因此可以插入不同的实体标识符以嵌入文本中。例如，出现在两个不同文档中的文本“He is smart”可以用两个不同的标识符标注：“[E21He]is smart”和“[E78He]is smart”。其中第一文档中的单词“He”和第二文档中的单词“He”指代不同的人。

[0059] 可以有不同的信息源用于浅层共指消解 320。例如，除了共指消解 320 期间执行的表达检测，可以有系统专用于查找文本内容 210 中的专有名词。这些不同的源可以识别冲突的消解信息。例如，冲突消解可能出现在跨边界处。例如，两个系统可能已识别了下面的冲突指代表达：

[0060] “[John]told[George Washington][Irving]was a great writer”

[0061] “[John]told[George][Washington Irving]was a great writer”

[0062] 考虑下面的跨边界冲突：第一个字符串中的 [George Washington] 与第二个字符串中的 [George] 冲突。第一个字符串中的 [George Washington] 还与第二个字符串中的 [Washington Irving] 冲突。基于置信度信息或上下文因素，可以迭代地应用不同的策略以消解该冲突或保留该冲突。在“丢弃”策略中，两个或多个冲突的边界可以通过丢弃置信度最低的边界来消解。在“合并”策略中，当两个或多个边界在相容的上下文中同等地似是而非时，边界可以相应地移动。例如，“[Mr.John]Smith”和“Mr.[John Smith]”可以合并成“[Mr.John Smith]”。在“保留”策略中，在边界的配置及其置信值既不支持合并也不支持丢弃时，可以通过保持多个边界作为歧义输出来保留多个边界。例如，“[Alexander theGreat]”和“[Alexander][the Great]”可以作为可选的歧义消解提供。

[0063] 解析组件 215 可以是支持直接解析歧义输入的歧义感知解析器，其中语法解析 355 可以保留歧义。或者，可能需要单独地解析歧义输入消解，且可以将多个输出结构单独地传递给语义组件 225。语义处理 225，如下文中进一步详述，可以对语法解析器 215 的每个输出应用多次。这可以对不同的语法输入得到不同的语义输出。或者，语义映射 225 可以组合各种输入并一致地处理这些输入。

[0064] 语义映射 225 可以具有语义标准化 360。句子的多个有歧义的语法解析 355 输出可以共享含义同时具有不同的形式。例如，这可以出现在被动语言的标准化中。考虑“John gave Mary a present”，单词“John”是主语，“Mary”是间接宾语。考虑“a present was given to Mary by John”，主语是“Mary”而“John”是宾语。标准化 360 可以提供这样的输出，其中这两个示例相同地表示为“John”是语义主语而“Mary”是语义间接宾语。或者，“John”可以识别为动作者而“Mary”识别为接受者。类似地，可以对“Rome's destruction of Carthage”和“Rome destroyed Carthage”提供等同的表示。

[0065] 语义标准化还可以增加关于所解析句子中的不同单词的信息。例如，可以在辞典中识别单词并将其与其同义词、上义词、可能的别名及其他词汇信息关联。

[0066] 基于语义的共指消解 370 可以基于语法和语义信息消解表达。例如，“Johnsaw Bill.He greeted him”可以将“he”消解为“John”并将“him”消解为“Bill”。可以做出该消解因为“he”和“John”都是主语，而“him”和“Bill”都是宾语。

[0067] 可以通过检查词语所出现的文档片段来执行浅层共指消解 320。相反，语义共指消解 370 或深层共指消解可以一次处理一个句子。句子的可能的先行语可以放置到先行语存储 375 中以使在后句子的语义共指消解 370 可以访问先前引入的元素。先行语可以和关于

其在句子中的语法功能和角色的信息、其在文本中的距离、关于其与其他先行语的关系的信息，及各种其他信息一起存储。

[0068] 表达合并380可以组合来自浅层共指消解320、旁置标注325的表达以及来自语义共指消解370的信息。可以使用字符串对齐或标注330识别要组合的项的信息。也可以使用组合相同文本的两个标注的其他机制。

[0069] 语法解析215可以是可任选地检测的指代表达的自然集成点。解析器可以支持句子中的推断结构，如成分或语法关系，如主语和宾语。支持歧义的语法解析器215可以识别句子的多种可选的结构表示。在一个示例中，通过仅保留那些每个指代表达的左边界与短语中相容部分的开始重合的表示，可以使用来自共指消解320的信息过滤语法解析器215的输出。例如，共指消解可以确定如在“[E0John]told[E1George][E2Washington Irving]was a great writer”中的共指对象。语法解析器215可以单独地提供四种解析可能：

[0070] 1.[John]and[George]and[Washington Irving]

[0071] 2.[John]and[George]and[Washington]and[Irving]

[0072] 3.[John]and[George Washington]and[Irving]

[0073] 4.[John]and[George Washington Irving]

[0074] 可以过滤出编号为3和4的解析器可能性，因为其按指代消解320的规定与实体E2“Washington Irving”的左边界不相容。

[0075] 扩展处理385可以将附加信息添加到表示中。例如，对于“John sold a carfrom Bill(约翰卖车给比尔)”，扩展385可以附加地输出“Bill bought a car fromJohn(比尔从约翰那儿买车)”的表示。类似地，对于“John killed Bill(约翰杀死了比尔)”，扩展385可以附加地输出“Bill died(比尔死了)”的表示。

[0076] 传统的搜索引擎可以响应于用户查询基于匹配的关键词或项检索文档。在这些传统系统中，可以根据诸如查询中的多少个项出现在文档内、这些项出现的频繁程度，或这些项一起出现的紧密程度等因素对文档评级。

[0077] 考虑示例查询“Picasso painted”及包含“Picasso was born in Malaga. Hepainted Guernica”的第一示例文档与包含“Picasso ′ s friend Matisse paintedprolifically”的第二示例文档。在所有其他条件相同的情况下，传统的系统可能使第二文档评级高于第一文档，因为单词“Picasso”和“painted”在第二文档中更紧密地在一起。相反，能够消解第一文档中的单词“He”指代毕加索的系统可以基于该知识正确地使第一文档得到更高评级。假设查询“Picassopainted”反映找到毕加索画出了什么的用户意图，则第一文档显然是更加相关的结果。

[0078] 自然语言处理系统300可以具有不同的架构。在一个实施例中，可以提供管线，其中来自语言处理的一个阶段的信息作为后续阶段的输入传递。应理解，可以使用可操作以从自然语言文本内容210中提取要进行索引的事实的任何其他架构实现这些方法。

[0079] 现参考图4，提供有关用于歧义敏感自然语言处理系统中的共指消解的本发明实施例的附加细节。特别地，图4是流程图，其示出用根据本发明实施例的各方面的共指消解进行歧义敏感索引的处理400的各方面。

[0080] 应理解，本文所述的逻辑操作可实现为(1)计算机执行的步骤序列或在计算系统上运行的程序模块和/或(2)互连的机器逻辑电路或计算系统内的电路模块。具体实现是

取决于计算系统的性能和其他需求作出的选择。相应地，本文所述的逻辑操作可以不同地指状态操作、结构装置、步骤，或模块。这些操作、结构装置、步骤和模块可以实现为软件、固件、专用数字逻辑，及其任意组合。还应理解可以执行比附图所示和本文所述的操作更多或更少的操作。这些操作还可以顺序执行、并行执行，或以本文所述顺序不同的顺序执行。

[0081] 例程 400 开始于操作 410，其中可以检索文本内容 210 的部分以供解析和索引。在操作 420，可以分割文本内容 210 以界定消解处理可在其上进行搜索和解析的文本区域。分割可以基于文本中的结构，如句子、段落、页、章或节。分割也可以基于单词数量、句子数量，或者空间或复杂性的其他量度。

[0082] 在操作 430，可以在文本内容 210 内消解共指。使用操作 430 中确定的边界，可以识别和匹配共指。可以确定别名群集。表面结构可以用于提供“浅层”消解。可以标注共指消解期间引起的歧义。这样的标注 340 可以作为文本内容 210 内的标记提供或通过使用外部实体映射提供。类似的标注也可以用于以实体编号标记指代和指代对象。还可以提供标注以指示所确定的共指消解的置信度水平。

[0083] 在操作 440，语法解析可以将句子转换为明确单词间语法关系的表示。解析器 215 可以应用关联于具体语言的语法 220 来提供语法解析 355 信息。

[0084] 在操作 450，可以从文本内容 210 中提取语义表示。文本内容 210 中的文档内表达的信息可以在形式上按文本内实体之间的关系的表示来组织。这些关系可以指一般意义的事实。

[0085] 在操作 455，来自语法解析 215 的语法解析 355 信息输出可以用于支持深层共指消解 370。也可以充分利用操作 450 期间产生的语义表示。

[0086] 在操作 460，来自浅层共指消解操作 430 的表达可以和来自深层共指消解操作 455 的信息集成。支持歧义的语法解析器 215 可以识别句子的多种可选的结构表示。来自共指消解的信息可以用于过滤语法解析器 215 的输出。

[0087] 在操作 470，可以将文本内容 210 的语义扩展为包括所选择的隐含表达。在操作 475，可以从表达内容文本中的实体间关系、事件和事务状态的语义表示中提取事实。在操作 480，可以将事实和实体存储到语义索引 250 中。

[0088] 例程 400 可以在操作 480 之后终止。然而，应理解，可以重复地或连续地应用例程 400 以检索要要应用到语义索引 250 的文本内容 210 部分。

[0089] 现参考图 5，示意性计算机架构 500 可以执行本文所述用于歧义敏感自然语言处理系统中的共指消解的软件组件。图 5 所示的计算机架构示出常规的桌面计算机、膝上型计算机或服务器计算机并可用于执行本文所述的软件组件的任何方面。然而应理解，所述的软件组件也可以在其他示例计算环境中执行，如移动设备、电视机、机顶盒、网亭、车辆信息系统、移动电话、嵌入系统，或其他环境。客户机计算机 110A-110D 或服务器计算机 120 中的任何一个或多个可以实现为根据各实施例的计算机系统 500。

[0090] 图 5 示出的计算机架构可以包括中央处理单元 10(CPU)、包括随机存取存储器 14(RAM) 和只读存储器 16(ROM) 的系统存储器 13，及可将系统存储器 13 耦合到 CPU 10 的系统总线 11。基本输入 / 输出系统可以存储在 ROM16 中，并包含有助于诸如在启动期间在计算机 500 内的元件之间传输信息的基本例程。计算机 500 还可以包括大容量存储设备 15，用于存储操作系统 18、软件、数据和各种程序模块，如与自然语言引擎 130 关联的程序

模块。自然语言引擎 130 可以执行本文所述的软件组件的部分。关联于自然语言引擎 130 的语义索引 250 可以存储在大容量存储设备 15 内。

[0091] 大容量存储设备 15 可以通过连接到总线 11 的大容量存储控制器（未示出）连接到 CPU 10。大容量存储设备 15 及其相关的计算机可读介质可以为计算机 500 提供非易失性存储。虽然对本文包含的计算机可读介质的描述指大容量存储设备，如硬盘或 CD-ROM 驱动器，本领域技术人员应理解计算机可读介质可以是可由计算机 500 存取的任何可用的计算机存储介质。

[0092] 作为示例而非限制，计算机可读介质可以包括以任何方法或技术实现的用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的易失性和非易失性、可移动和不可移动介质。例如，计算机可读介质包括但不限于 RAM、ROM、EPROM、EEPROM、闪存或其他固态存储器技术、CD-ROM、数字多功能盘 (DVD)、HD-DVD、BLU-RAY，或其他光学存储、盒式磁带、磁带、磁盘存储或其他磁存储设备，或可用于存储所需信息并可由计算机 500 存取的任何其他介质。

[0093] 根据各种实施例，计算机 500 可以使用通过网络，如网络 140 到远程计算机的逻辑连接以在联网环境中操作。计算机 500 可以通过连接到总线 11 的网络接口单元 19 连接到网络 140。应理解，也可以使用网络接口单元 19 来连接到其他类型的网络和远程计算机系统。计算机 500 还可以包括用于接收和处理来自多种其他设备的输入的输入 / 输出控制器 12，其他设备可包括键盘、鼠标，或电子触笔（未示出）。类似地，输入 / 输出控制器 12 可以提供输出到视频显示、打印机，或其他类型的输出设备（亦未示出）。

[0094] 如上文简述，多个程序模块和数据文件可以存储在计算机 500 的大容量存储设备 15 和 RAM 14 中，包括操作系统 18，该操作系统适合用于控制联网的桌面计算机、膝上型计算机、服务器计算机，或其他计算环境的操作。大容量存储设备 15、ROM 16 和 RAM 14 还可以存储一个或多个程序模块。特别地，大容量存储设备 15、ROM 16 和 RAM 14 可以存储由 CPU 10 执行的自然语言引擎 130。自然语言引擎 130 可以包括用于执行参考图 2-4 详述的处理的部分的软件组件。大容量存储设备 15、ROM 16 和 RAM 14 还可以存储其他类型的程序模块。大容量存储设备 15、ROM 16 和 RAM 14 还可以存储关联于自然语言引擎 130 的语义索引 250。

[0095] 基于上文所述，应理解本文中提供了歧义敏感自然语言处理系统中的共指消解的技术。虽然用特定于计算机结构特征、方法步骤和计算机可读介质的语言描述本文提供的主题，应理解本申请的权利要求限定的发明不必限于本文所述的这些具体特征、步骤或介质。相反，具体特征、步骤和介质是作为实现权利要求的示例形式公开的。

[0096] 上文所述主题是仅通过示例提供且不应认为是限制性的。可以对本文所述主题做出各种修改和改变而不遵循示出和描述的示例实施例和应用，且不偏离在本申请的权利要求中阐述的本发明的实质和范围。

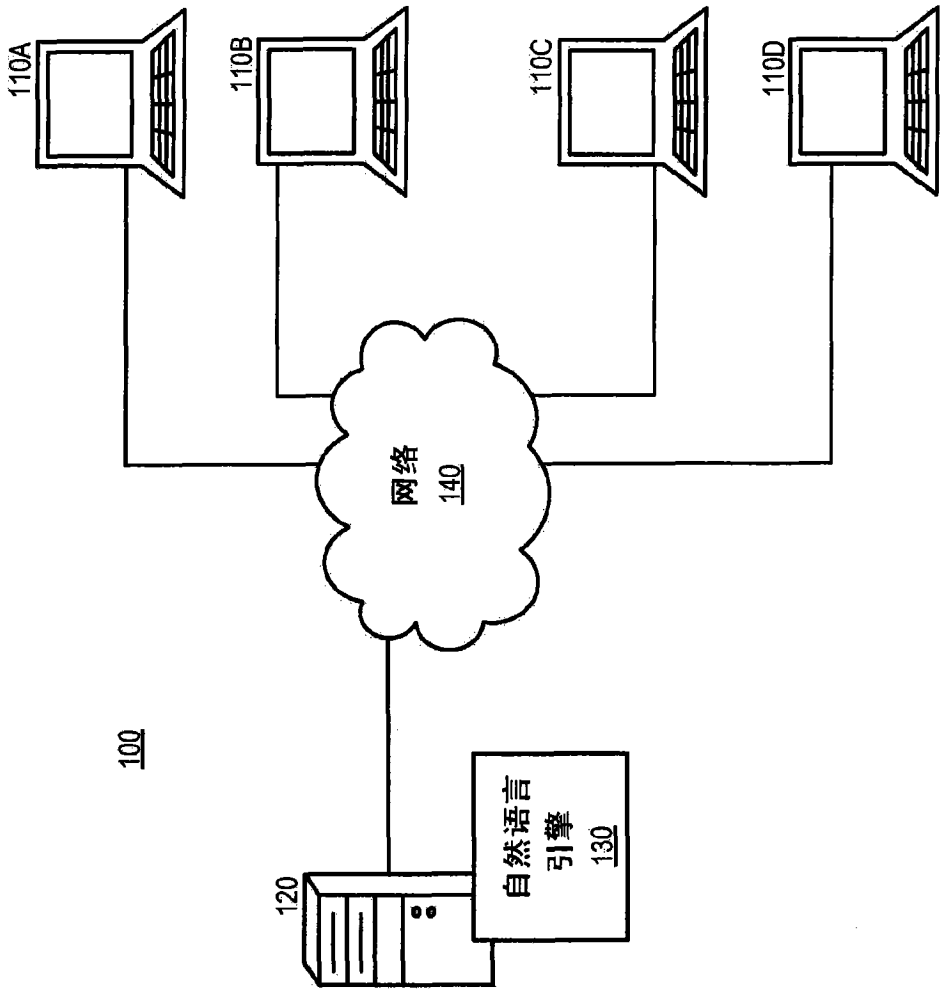


图 1

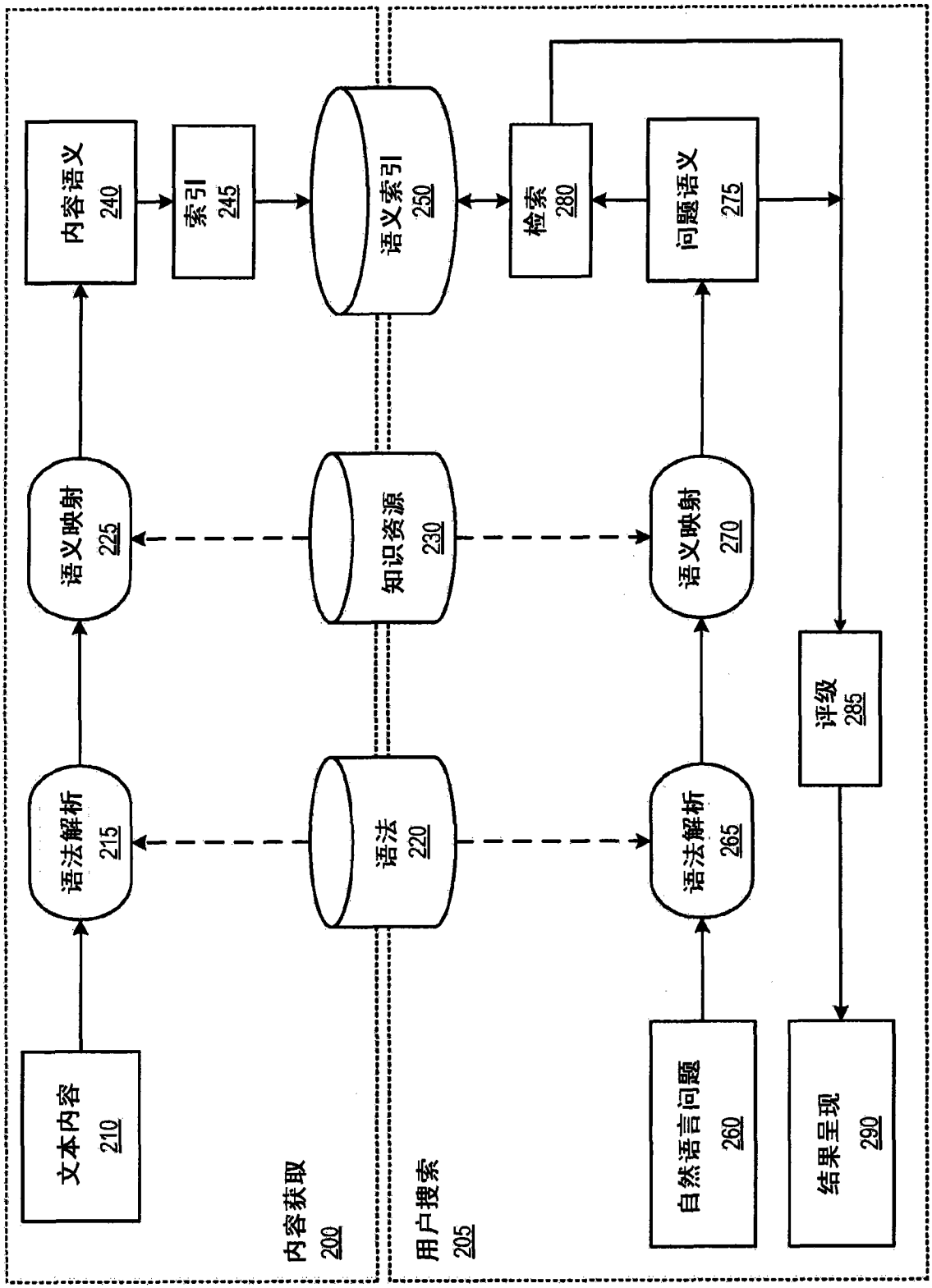

内容语义 240
索引 245
语义索引 250
检索 280
问题语义 275
语义映射 225
知识资源 230
语义映射 270
评级 285
语法解析 215
语法 220
语法解析 265
文本内容 210
自然语言问题 260
结果呈现 290
内容获取 200
用户搜索 205

图 2

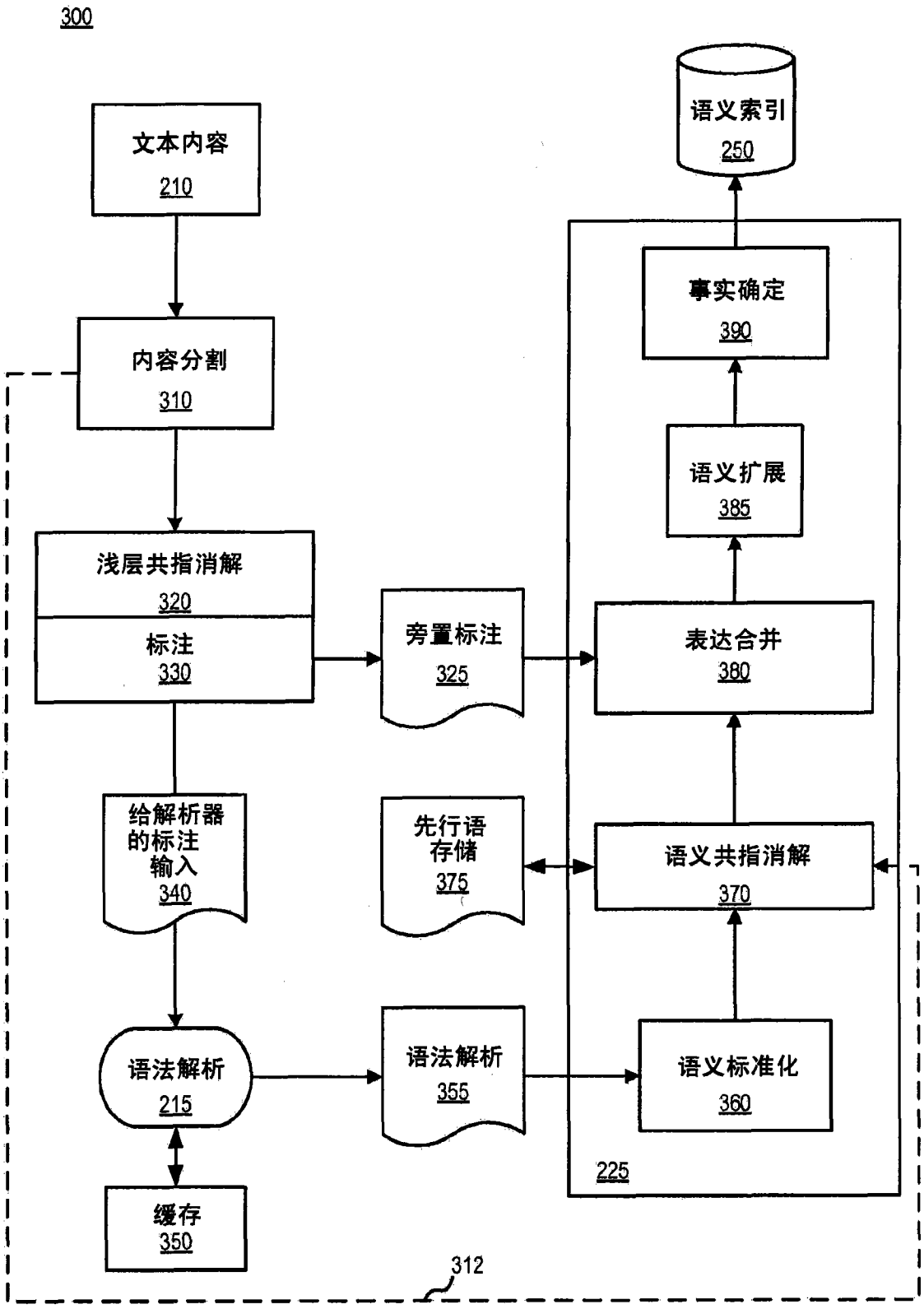

300
文本内容
210
内容分割
310
浅层共指消解
320
标注
330
给解析器的标注输入
340
语法解析
215
缓存
350
旁置标注
325
先行语存储
375
语法解析
355
312
语义索引
250
事实确定
390
语义扩展
385
表达合并
380
语义共指消解
370
语义标准化
360
225

图 3

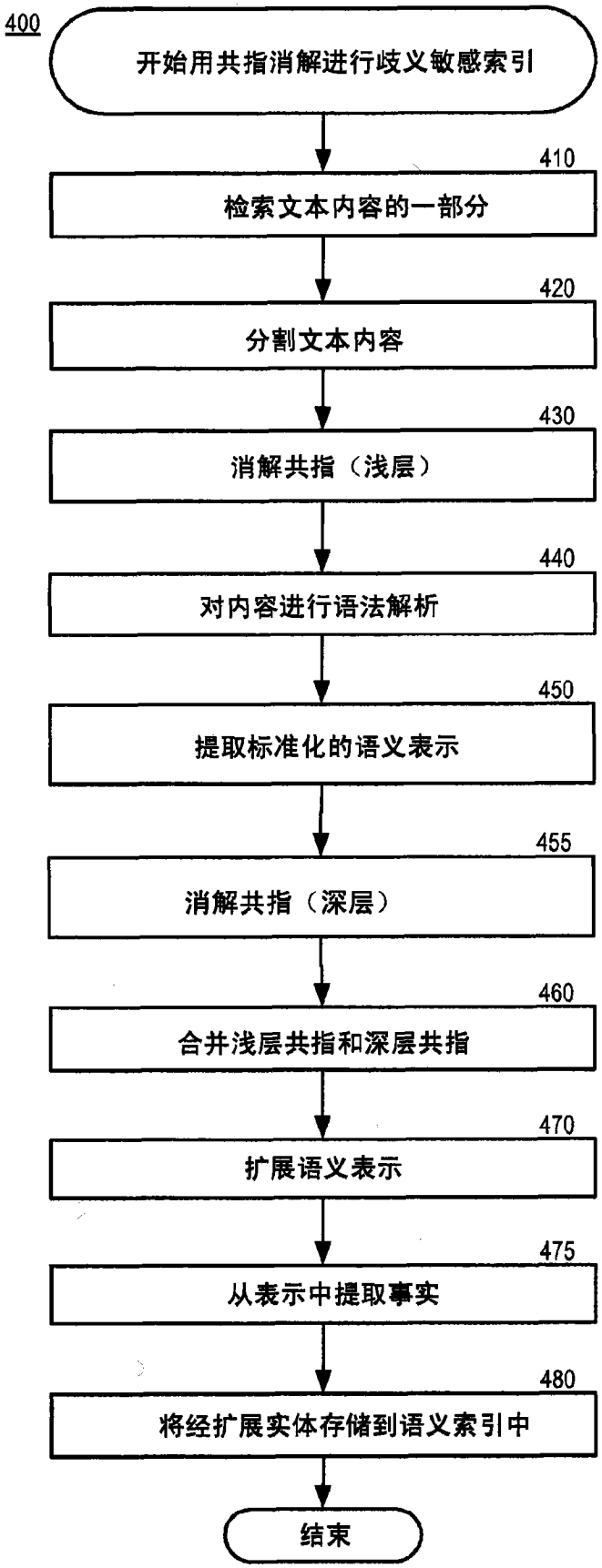

400
开始用共指消解进行歧义敏感索引
410
检索文本内容的一部分
420
分割文本内容
430
消解共指（浅层）
440
对内容进行语法解析
450
提取标准化的语义表示
455
消解共指（深层）
460
合并浅层共指和深层共指
470
扩展语义表示
475
从表示中提取事实
480
将经扩展实体存储到语义索引中
结束

图 4

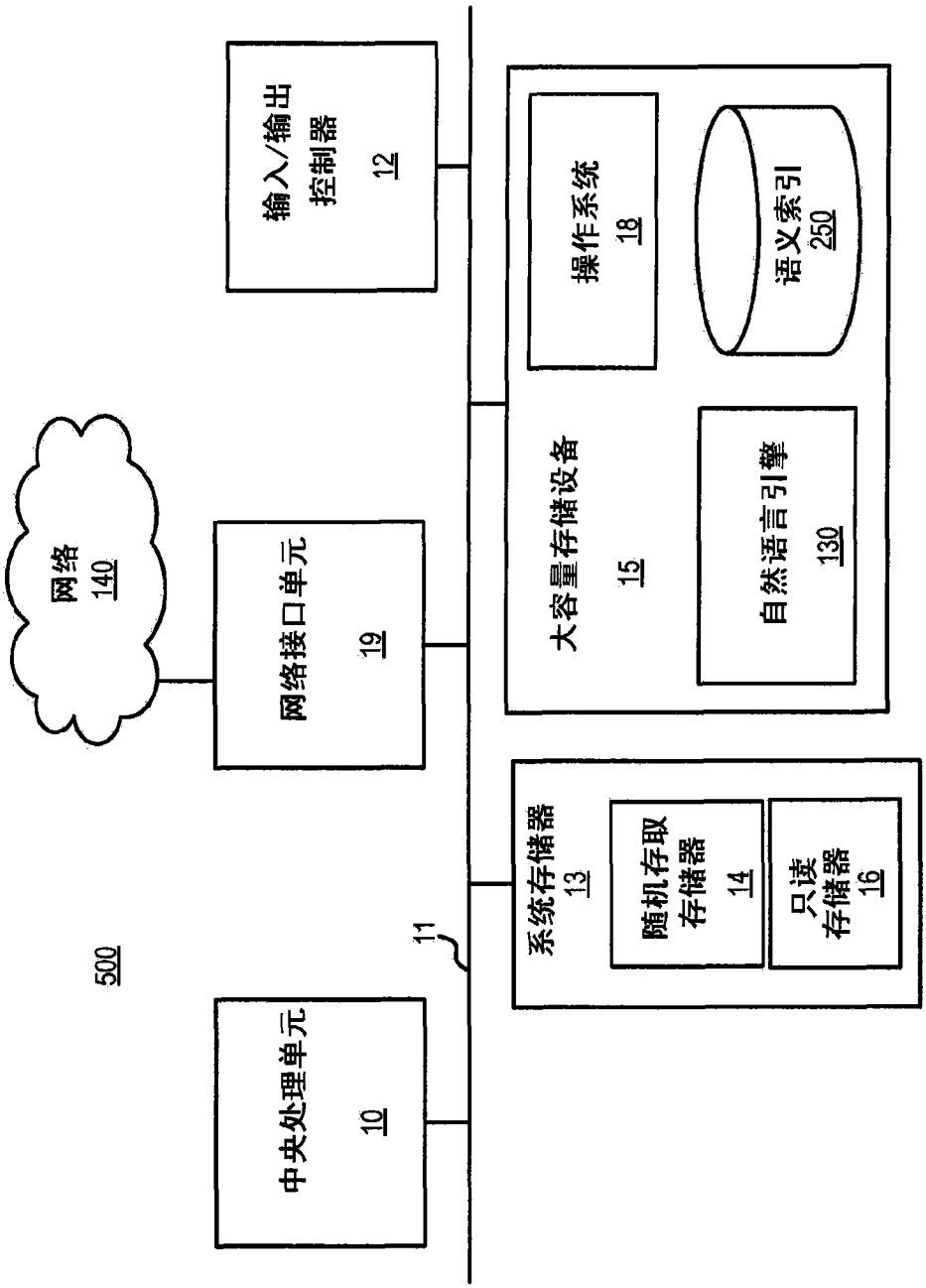

输入/输出控制器 12
操作系统 18
语义索引 250
大容量存储设备 15
自然语言引擎 130
网络 140
网络接口单元 19
系统存储器 13
随机存取存储器 14
只读存储器 16
11
500
中央处理单元 10

图 5