(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 105528349 A
(43)申请公布日 2016.04.27

(21)申请号 201410513189.4

(22)申请日 2014.09.29

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
申请人 中国科学院自动化研究所

(72)发明人 赵军 刘康 何世柱 张轶博

(74)专利代理机构 北京龙双利达知识产权代理有限公司 11329
代理人 王君 肖鹏

(51)Int.Cl.
G06F 17/30(2006.01)

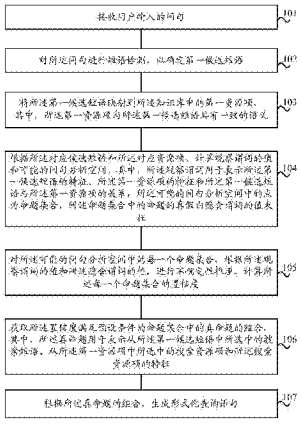
权利要求书5页 说明书33页 附图4页

(54)发明名称

知识库中问句解析的方法及设备

(57)摘要

本发明实施例提供一种知识库中问句解析的方法，包括：接收用户输入的问句；对问句进行短语检测确定候选短语；将候选短语映射到知识库中的资源项；进一步确定观察谓词的值和可能的问句分析空间。对可能的问句分析空间中的每一个命题集合，根据观察谓词和隐含谓词的值进行不确定性推理计算置信度，并获取置信度满足预设条件的命题集合中的真命题的组合；根据所述真命题的组合，生成形式化查询语句。本发明实施例利用观察谓词和隐含谓词，进行不确定性推理，能够将自然语言问句转化为形式化查询语句。并且，不确定性推理的方法能够应用于任何领域的知识库，具有领域扩展性，这样无需针对知识库人工地配置转换规则。



CN 105528349 A

1. 一种知识库中问句解析的方法，其特征在于，包括：

接收用户输入的问句；

对所述问句进行短语检测，以确定第一候选短语；

将所述第一候选短语映射到所述知识库中的第一资源项，其中，所述第一资源项与所述第一候选短语具有一致的语义；

根据所述第一候选短语和所述第一资源项，确定观察谓词的值和可能的问句分析空间，其中，所述观察谓词用于表示所述第一候选短语的特征、所述第一资源项的特征和所述第一候选短语与所述第一资源项的关系，所述可能的问句分析空间中的点为命题集合，所述命题集合中的命题的真假由隐含谓词的值表征；

对所述可能的问句分析空间中的每一个命题集合，根据所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度；

获取所述置信度满足预设条件的命题集合中的真命题的组合，其中，所述真命题用于表示从所述第一候选短语中所选中的搜索短语、从所述第一资源项中所选中的搜索资源项和所述搜索资源项的特征；

根据所述真命题的组合，生成形式化查询语句。

2. 根据权利要求 1 所述的方法，其特征在于，所述不确定性推理基于马尔科夫逻辑网络 MLN，所述 MLN 包括预定义的一阶公式以及所述一阶公式的权重。

3. 根据权利要求 2 所述的方法，其特征在于，在所述接收用户输入的问句之前，所述方法还包括：

从所述知识库中获取多个自然语言问句；

对所述多个自然语言问句进行短语检测，以确定所述多个自然语言问句的第二候选短语；

将所述第二候选短语映射到所述知识库中的第二资源项，其中，所述第二资源项与所述第二候选短语具有一致的语义；

根据所述第二候选短语和所述第二资源项，确定与所述多个自然语言问句对应的观察谓词的值；

获取人工标注的与所述多个自然语言问句对应的隐含谓词的值；

根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重。

4. 根据权利要求 3 所述的方法，其特征在于，所述一阶公式包括布尔公式和加权公式，所述布尔公式的权重为 $+\infty$，所述加权公式的权重为加权公式权重，所述人工标注的与所述多个自然语言问句对应的隐含谓词的值满足所述布尔公式，

根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重，包括：

根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述加权公式权重。

5. 根据权利要求 3 所述的方法，其特征在于，所述根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重，包括：

根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，采用差额注入松弛算法 MIRA，确定所述一阶公式的权重。

6. 根据权利要求 2 至 5 任一项所述的方法，其特征在于，所述 MLN 表示为 M，所述一阶公式表示为 Φ_i，所述一阶公式的权重表示为 w_i，所述命题集合表示为 y，

对所述问句分析空间中的每一个命题集合，根据所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度，包括：

根据 $p(\mathbf{y})=\frac{1}{Z}\exp\left(\sum_{(\phi_i,w_i)\in M} w_i \sum_{\mathbf{c}\in C^{n\phi_i}} f_{\mathbf{c}}^{\phi_i}(\mathbf{y})\right)$，计算所述每一个命题集合的置信度，

其中，Z 为归一化常数，$C^{n\phi_i}$ 为与一阶公式 Φ_i 对应的子公式的集合，c 为 $C^{n\phi_i}$ 的所述子公式的集合中的一个子公式，$f_{\mathbf{c}}^{\phi_i}$ 为二值函数，$f_{\mathbf{c}}^{\phi_i}(\mathbf{y})$ 表示在所述命题集合 y 下，所述一阶公式的真假。

7. 根据权利要求 1 至 6 任一项所述的方法，其特征在于，所述获取所述置信度满足预设条件的命题集合中的真命题的组合，包括：

确定所述置信度的值最大的命题集合，并获取所述置信度的值最大的命题集合中的真命题的组合。

8. 根据权利要求 1 至 7 任一项所述的方法，其特征在于，

所述第一候选短语的特征包括所述第一候选短语在所述问句中的位置、所述第一候选短语的主要词的词性、所述第一候选短语两两之间的依存路径上的标签，

所述第一资源项的特征包括所述第一资源项的类型、所述第一资源项两两之间的相关性值、所述第一资源项两两之间的参数匹配关系，

所述第一候选短语与所述第一资源项的关系包括所述第一候选短语与所述第一资源项的先验匹配得分，

所述根据所述第一候选短语和所述第一资源项，确定观察谓词的值，包括：

确定所述第一候选短语在所述问句中的位置；

采用 stanford 词性标注工具，确定所述第一候选短语的主要词的词性；

采用 stanford 依存句法分析工具，确定所述第一候选短语两两之间的依存路径上的标签；

从所述知识库中确定所述第一资源项的类型，其中，所述类型为实体或类别或关系；

从所述知识库中确定所述第一资源项两两之间的参数匹配关系；

将所述第一资源项两两之间的相似性系数，作为所述两个第一资源项两两之间的相关性值；

计算所述第一候选短语与所述第一资源项之间的先验匹配得分，所述先验匹配得分用于表示所述第一候选短语映射到所述第一资源项的概率。

9. 根据权利要求 1 至 8 任一项所述的方法，其特征在于，所述形式化查询语句为简单协议资源描述框架查询语句 SPARQL。

10. 根据权利要求 9 所述的方法，其特征在于，所述根据所述真命题的组合，生成形式

化查询语句，包括：

根据所述真命题的组合，利用 SPARQL 模板生成所述 SPARQL。

11. 根据权利要求 10 所述的方法，其特征在于，所述 SPARQL 模板包括 ASK WHERE 模板、SELECT COUNT(？ url)WHERE 模板和 SELECT？ url WHERE 模板，

所述根据所述真命题的组合，利用 SPARQL 模板生成所述 SPARQL，包括：

当所述问句为 Yes/No 问题时，根据所述真命题的组合，使用所述 ASK WHERE 模板生成所述 SPARQL；

当所述问句为 Normal 问题时，根据所述真命题的组合，使用所述 SELECT？ url WHERE 模板生成所述 SPARQL；

当所述问句为 Number 问题时，根据所述真命题的组合，使用所述 SELECT？ url WHERE 模板生成所述 SPARQL，或者，当使用所述 SELECT？ url WHERE 模板生成的 SPARQL 无法得到数值型答案时，使用所述 SELECT COUNT(？ url)WHERE 模板生成所述 SPARQL。

12. 根据权利要求 1 至 11 任一项所述的方法，其特征在于，所述对所述问句进行短语检测，以确定第一候选短语，包括：将所述问句中的词序列作为所述第一候选短语，其中，所述词序列满足：

所述词序列中所有连续的非停用词都以大写字母开头，或者，若所述词序列中所有连续的非停用词不都以大写字母开头，则所述词序列的长度小于四；

所述词序列的主要词的词性为 jj 或 nn 或 rb 或 vb，其中，jj 为形容词，nn 为名词，rb 为副词，vb 为动词；

所述词序列所包括的词不全为停用词。

13. 一种问答解析的设备，其特征在于，包括：

接收单元，用于接收用户输入的问句；

短语检测单元，用于对所述接收单元接收的所述问句进行短语检测，以确定第一候选短语；

映射单元，用于将所述短语检测单元确定的所述第一候选短语映射到知识库中的第一资源项，其中，所述第一资源项与所述第一候选短语具有一致的语义；

第一确定单元，用于根据所述第一候选短语和所述第一资源项，确定观察谓词的值和可能的问句分析空间，其中，所述观察谓词用于表示所述第一候选短语的特征、所述第一资源项的特征和所述第一候选短语与所述第一资源项的关系，所述可能的问句分析空间中的点为命题集合，所述命题集合中的命题的真假由隐含谓词的值表征；

第二确定单元，用于对所述可能的问句分析空间中的每一个命题集合，根据所述第一确定单元确定的所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度；

获取单元，用于获取所述第二确定单元确定的所述置信度满足预设条件的命题集合中的真命题的组合，其中，所述真命题用于表示从所述第一候选短语中所选中的搜索短语、从所述第一资源项中所选中的搜索资源项和所述搜索资源项的特征；

生成单元，用于根据所述真命题的组合，生成形式化查询语句。

14. 根据权利要求 13 所述的设备，其特征在于，所述不确定性推理基于马尔科夫逻辑网络 MLN，所述 MLN 包括预定义的一阶公式以及所述一阶公式的权重。

15. 根据权利要求 14 所述的设备，其特征在于，

所述获取单元，还用于从所述知识库中获取多个自然语言问句；

所述短语检测单元，还用于对所述获取单元接收的所述问句进行短语检测，以确定第一候选短语；

所述映射单元，还用于将所述第二候选短语映射到所述知识库中的第二资源项，其中，所述第二资源项与所述第二候选短语具有一致的语义；

所述第一确定单元，还用于根据所述第二候选短语和所述第二资源项，确定与所述多个自然语言问句对应的观察谓词的值；

所述获取单元，还用于获取人工标注的与所述多个自然语言问句对应的隐含谓词的值；

所述第二确定单元，还用于根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重。

16. 根据权利要求 15 所述的设备，其特征在于，所述一阶公式包括布尔公式和加权公式，所述布尔公式的权重为 + ∞，所述加权公式的权重为加权公式权重，所述人工标注的与所述多个自然语言问句对应的隐含谓词的值满足所述布尔公式，

所述第二确定单元，具体用于：根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述加权公式权重。

17. 根据权利要求 15 所述的设备，其特征在于，所述第二确定单元，具体用于：

根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，采用差额注入松弛算法 MIRA，确定所述一阶公式的权重。

18. 根据权利要求 14 至 17 任一项所述的设备，其特征在于，所述 MLN 表示为 M，所述一阶公式表示为 Φ_i，所述一阶公式的权重表示为 w_i，所述命题集合表示为 y，

所述第二确定单元，具体用于：

根据 $p(\mathbf{y})=\frac{1}{Z}\exp\left(\sum_{(\phi_i,w_i)\in M} w_i \sum_{\mathbf{c}\in C^{n\phi_i}} f_{\mathbf{c}}^{\phi_i}(\mathbf{y})\right)$，计算所述每一个命题集合的置信度，

其中，Z 为归一化常数，$C^{n\phi_i}$ 为与一阶公式 Φ_i 对应的子公式的集合，c 为 $C^{n\phi_i}$ 的所述子公式的集合中的一个子公式，$f_{\mathbf{c}}^{\phi_i}$ 为二值函数，$f_{\mathbf{c}}^{\phi_i}(\mathbf{y})$ 表示在所述命题集合 y 下，所述一阶公式的真假。

19. 根据权利要求 13 至 18 任一项所述的设备，其特征在于，所述获取单元，具体用于：

确定所述置信度的值最大的命题集合，并获取所述置信度的值最大的命题集合中的真命题的组合。

20. 根据权利要求 13 至 18 任一项所述的设备，其特征在于，

所述第一候选短语的特征包括所述第一候选短语在所述问句中的位置、所述第一候选短语的主要词的词性、所述第一候选短语两两之间的依存路径上的标签，

所述第一资源项的特征包括所述第一资源项的类型、所述第一资源项两两之间的相关性值、所述第一资源项两两之间的参数匹配关系，

所述第一候选短语与所述第一资源项的关系包括所述第一候选短语与所述第一资源项的先验匹配得分，

所述第一确定单元，具体用于：

确定所述第一候选短语在所述问句中的位置；

采用 stanford 词性标注工具，确定所述第一候选短语的主要词的词性；

采用 stanford 依存句法分析工具，确定所述第一候选短语两两之间的依存路径上的标签；

从所述知识库中确定所述第一资源项的类型，其中，所述类型为实体或类别或关系；

从所述知识库中确定所述第一资源项两两之间的参数匹配关系；

将所述第一资源项两两之间的相似性系数，作为所述两个第一资源项两两之间的相关性值；

计算所述第一候选短语与所述第一资源项之间的先验匹配得分，所述先验匹配得分用于表示所述第一候选短语映射到所述第一资源项的概率。

21. 根据权利要求 13 至 20 任一项所述的设备，其特征在于，所述形式化查询语句为简单协议资源描述框架查询语句 SPARQL。

22. 根据权利要求 21 所述的设备，其特征在于，所述生成单元，具体用于：

根据所述真命题的组合，利用 SPARQL 模板生成所述 SPARQL。

23. 根据权利要求 22 所述的设备，其特征在于，所述 SPARQL 模板包括 ASK WHERE 模板、SELECT COUNT(？ url)WHERE 模板和 SELECT？ url WHERE 模板，

所述生成单元，具体用于：

当所述问句为 Yes/No 问题时，根据所述真命题的组合，使用所述 ASK WHERE 模板生成所述 SPARQL；

当所述问句为 Normal 问题时，根据所述真命题的组合，使用所述 SELECT？ url WHERE 模板生成所述 SPARQL；

当所述问句为 Number 问题时，根据所述真命题的组合，使用所述 SELECT？ url WHERE 模板生成所述 SPARQL，或者，当使用所述 SELECT？ url WHERE 模板生成的 SPARQL 无法得到数值型答案时，使用所述 SELECT COUNT(？ url)WHERE 模板生成所述 SPARQL。

24. 根据权利要求 13 至 23 任一项所述的设备，其特征在于，所述短语检测单元，具体用于：

将所述问句中的词序列作为所述第一候选短语，其中，所述词序列满足：

所述词序列中所有连续的非停用词都以大写字母开头，或者，若所述词序列中所有连续的非停用词不都以大写字母开头，则所述词序列的长度小于四；

所述词序列的主要词的词性为 jj 或 nn 或 rb 或 vb，其中，jj 为形容词，nn 为名词，rb 为副词，vb 为动词；

所述词序列所包括的词不全为停用词。

知识库中问句解析的方法及设备

技术领域

[0001] 本发明实施例涉及通信领域，并且更具体地，涉及一种知识库中问句解析的方法及设备。

背景技术

[0002] 知识库(Knowledge Base，KB)是知识工程中结构化、易操作、易利用、全面有组织的知识集群，是针对某一个或某一些领域问题求解的需要，采用某一种或某几种知识表示方式在计算机存储器中存储、组织、管理和使用的互相联系的知识片集合。

[0003] 目前互联网上已经出现了大量的知识资源和知识社区，例如维基百科(Wikipedia)、百度百科、互动百科等。从这些知识资源中，已有研究已经挖掘出以实体、实体关系为核心的大规模知识库。除此之外，还存在一些领域知识库，如天气知识库、餐饮知识库等。

[0004] 知识库的建设经历了由人工和群体智能添加到面向整个互联网利用机器学习和信息抽取技术自动获取的过程。早期的知识库是由专家人工构建。例如WordNet、CYC、CCD、HowNet、中国大百科全书等。但是随着信息技术的发展，传统人工构建的知识库逐渐暴露出规模小、知识少、更新慢的缺点；同时由专家构建的确定性知识框架也无法满足互联网有噪环境下大规模计算的需求。这也是CYC项目最终失败的原因之一。随着Web 2.0的飞速崛起，出现了大量基于群体智慧的网络知识库，包括Wikipedia、互动百科、百度百科等。以这些网络资源为基础，大量的自动半自动知识库构建方法被用来构建大型可用的知识库，比如YAGO，DBpedia，Freebase等。

[0005] 基于这些知识库，可以构建起知识库问答系统(Knowledge-base-based Question Answering)。与基于检索技术的问答系统相比，基于知识库的问答系统由于知识库规模的限制，对问题的覆盖率可能会较低，但其具备一定的推理能力。另外，在限定领域内会达到较高的准确率。因此，一些基于知识库的问答系统应运而生，有些成为独立的应用，有些作为已有产品的增强功能，比如苹果的siri、谷歌的知识图谱等。

[0006] 问答系统(Question Answering)是指不需要用户把问题分解成关键词，而直接以自然语言的形式提问，经过问答系统对用户的问题的处理，再从知识库或者互联网快速搜索出和用户的问题对应的答案，然后把答案直接返回给用户，而不是相关的网页。因此问答系统大大降低了用户的使用难度，它比传统的关键字检索和语义搜索技术等搜索引擎更加方便和高效。

[0007] 关联数据问答(Question Answering over Linked Data，QALD)评测比赛推动了问答系统的发展。其目标是针对大规模结构化的关联数据，将自然语言问句转换为结构化的简单协议资源描述框架查询语句(Simple Protocol and RDF(Resource Description Framework，资源描述框架)Query Language，SPARQL)，从而建立友好的自然语言查询接口。将自然语言问句转换为结构化的SPARQL，需要依赖于针对于知识库的转换规则。但是目前的问答系统中，转换规则都是人工配置，这样导致不仅耗费大量人力，而且领域扩展性也很

差。

发明内容

[0008] 本发明实施例提供一种基于知识库的问句解析的方法，不需要人工配置转换规则，并且是领域无关的。

[0009] 第一方面，提供了一种知识库中问句解析的方法，包括：

[0010] 接收用户输入的问句；

[0011] 对所述问句进行短语检测，以确定第一候选短语；

[0012] 将所述第一候选短语映射到所述知识库中的第一资源项，其中，所述第一资源项与所述第一候选短语具有一致的语义；

[0013] 根据所述第一候选短语和所述第一资源项，确定观察谓词的值和可能的问句分析空间，其中，所述观察谓词用于表示所述第一候选短语的特征、所述第一资源项的特征和所述第一候选短语与所述第一资源项的关系，所述可能的问句分析空间中的点为命题集合，所述命题集合中的命题的真假由隐含谓词的值表征；

[0014] 对所述可能的问句分析空间中的每一个命题集合，根据所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度；

[0015] 获取所述置信度满足预设条件的命题集合中的真命题的组合，其中，所述真命题用于表示从所述第一候选短语中所选中的搜索短语、从所述第一资源项中所选中的搜索资源项和所述搜索资源项的特征；

[0016] 根据所述真命题的组合，生成形式化查询语句。

[0017] 结合第一方面，在第一方面的第一种可能的实现方式中，所述不确定性推理基于马尔科夫逻辑网络 MLN，所述 MLN 包括预定义的一阶公式以及所述一阶公式的权重。

[0018] 结合第一方面或者第一方面的第一种可能的实现方式，在第一方面的第二种可能的实现方式中，在所述接收用户输入的问句之前，所述方法还包括：

[0019] 从所述知识库中获取多个自然语言问句；

[0020] 对所述多个自然语言问句进行短语检测，以确定所述多个自然语言问句的第二候选短语；

[0021] 将所述第二候选短语映射到所述知识库中的第二资源项，其中，所述第二资源项与所述第二候选短语具有一致的语义；

[0022] 根据所述第二候选短语和所述第二资源项，确定与所述多个自然语言问句对应的观察谓词的值；

[0023] 获取人工标注的与所述多个自然语言问句对应的隐含谓词的值；

[0024] 根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重。

[0025] 结合第一方面的第二种可能的实现方式，在第一方面的第三种可能的实现方式中，所述一阶公式包括布尔公式和加权公式，所述布尔公式的权重为 $+\infty$，所述加权公式的权重为加权公式权重，所述人工标注的与所述多个自然语言问句对应的隐含谓词的值满足所述布尔公式，

[0026] 根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对

应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重，包括：

[0027] 根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述加权公式权重。

[0028] 结合第一方面的第二种可能的实现方式，在第一方面的第四种可能的实现方式中，所述根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重，包括：

[0029] 根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，采用差额注入松弛算法 MIRA，确定所述一阶公式的权重。

[0030] 结合上述第一方面的任一种可能的实现方式，在第一方面的第五种可能的实现方式中，所述 MLN 表示为 M，所述一阶公式表示为 Φ_i，所述一阶公式的权重表示为 w_i，所述命题集合表示为 y，

[0031] 对所述问句分析空间中的每一个命题集合，根据所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度，包括：

[0032] 根据 $p(\mathbf{y})=\frac{1}{Z}\exp\left(\sum_{(\phi_i,w_i)\in M} w_i \sum_{\mathbf{c}\in C^{n\phi_i}} f_{\mathbf{c}}^{\phi_i}(\mathbf{y})\right)$，计算所述每一个命题集合的置信度，

[0033] 其中，Z 为归一化常数，$C^{n\phi_i}$ 为与一阶公式 Φ_i 对应的子公式的集合，c 为 $C^{n\phi_i}$ 的所述子公式的集合中的一个子公式，$f_{\mathbf{c}}^{\phi_i}$ 为二值函数，$f_{\mathbf{c}}^{\phi_i}(\mathbf{y})$ 表示在所述命题集合 y 下，所述一阶公式的真假。

[0034] 结合第一方面或者上述第一方面的任一种可能的实现方式，在第一方面的第六种可能的实现方式中，所述获取所述置信度满足预设条件的命题集合中的真命题的组合，包括：

[0035] 确定所述置信度的值最大的命题集合，并获取所述置信度的值最大的命题集合中的真命题的组合。

[0036] 结合第一方面或者上述第一方面的任一种可能的实现方式，在第一方面的第七种可能的实现方式中，

[0037] 所述第一候选短语的特征包括所述第一候选短语在所述问句中的位置、所述第一候选短语的主要词的词性、所述第一候选短语两两之间的依存路径上的标签，

[0038] 所述第一资源项的特征包括所述第一资源项的类型、所述第一资源项两两之间的相关性值、所述第一资源项两两之间的参数匹配关系，

[0039] 所述第一候选短语与所述第一资源项的关系包括所述第一候选短语与所述第一资源项的先验匹配得分，

[0040] 所述根据所述第一候选短语和所述第一资源项，确定观察谓词的值，包括：

[0041] 确定所述第一候选短语在所述问句中的位置；

[0042] 采用 stanford 词性标注工具，确定所述第一候选短语的主要词的词性；

[0043] 采用 stanford 依存句法分析工具，确定所述第一候选短语两两之间的依存路径上的标签；

[0044] 从所述知识库中确定所述第一资源项的类型，其中，所述类型为实体或类别或关系；

[0045] 从所述知识库中确定所述第一资源项两两之间的参数匹配关系；

[0046] 将所述第一资源项两两之间的相似性系数，作为所述两个第一资源项两两之间的相关性值；

[0047] 计算所述第一候选短语与所述第一资源项之间的先验匹配得分，所述先验匹配得分用于表示所述第一候选短语映射到所述第一资源项的概率。

[0048] 结合第一方面或者上述第一方面的任一种可能的实现方式，在第一方面的第八种可能的实现方式中，所述形式化查询语句为简单协议资源描述框架查询语句 SPARQL。

[0049] 结合第一方面的第八种可能的实现方式，在第一方面的第九种可能的实现方式中，所述根据所述真命题的组合，生成形式化查询语句，包括：

[0050] 根据所述真命题的组合，利用 SPARQL 模板生成所述 SPARQL。

[0051] 结合第一方面的第九种可能的实现方式，在第一方面的第十种可能的实现方式中，所述 SPARQL 模板包括 ASK WHERE 模板、SELECT COUNT(？url)WHERE 模板和 SELECT？url WHERE 模板，

[0052] 所述根据所述真命题的组合，利用 SPARQL 模板生成所述 SPARQL，包括：

[0053] 当所述问句为 Yes/No 问题时，根据所述真命题的组合，使用所述 ASK WHERE 模板生成所述 SPARQL；

[0054] 当所述问句为 Normal 问题时，根据所述真命题的组合，使用所述 SELECT？url WHERE 模板生成所述 SPARQL；

[0055] 当所述问句为 Number 问题时，根据所述真命题的组合，使用所述 SELECT？url WHERE 模板生成所述 SPARQL，或者，当使用所述 SELECT？url WHERE 模板生成的 SPARQL 无法得到数值型答案时，使用所述 SELECT COUNT(？url)WHERE 模板生成所述 SPARQL。

[0056] 结合第一方面或者上述第一方面的任一种可能的实现方式，在第一方面的第十一种可能的实现方式中，所述对所述问句进行短语检测，以确定第一候选短语，包括：将所述问句中的词序列作为所述第一候选短语，其中，所述词序列满足：

[0057] 所述词序列中所有连续的非停用词都以大写字母开头，或者，若所述词序列中所有连续的非停用词不都以大写字母开头，则所述词序列的长度小于四；

[0058] 所述词序列的主要词的词性为 jj 或 nn 或 rb 或 vb，其中，jj 为形容词，nn 为名词，rb 为副词，vb 为动词；

[0059] 所述词序列所包括的词不全为停用词。

[0060] 第二方面，提供了一种问答解析的设备，包括：

[0061] 接收单元，用于接收用户输入的问句；

[0062] 短语检测单元，用于对所述接收单元接收的所述问句进行短语检测，以确定第一候选短语；

[0063] 映射单元，用于将所述短语检测单元确定的所述第一候选短语映射到知识库中的第一资源项，其中，所述第一资源项与所述第一候选短语具有一致的语义；

[0064] 第一确定单元，用于根据所述第一候选短语和所述第一资源项，确定观察谓词的值和可能的问句分析空间，其中，所述观察谓词用于表示所述第一候选短语的特征、所述第一资源项的特征和所述第一候选短语与所述第一资源项的关系，所述可能的问句分析空间中的点为命题集合，所述命题集合中的命题的真假由隐含谓词的值表征；

[0065] 第二确定单元，用于对所述可能的问句分析空间中的每一个命题集合，根据所述第一确定单元确定的所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度；

[0066] 获取单元，用于获取所述第二确定单元确定的所述置信度满足预设条件的命题集合中的真命题的组合，其中，所述真命题用于表示从所述第一候选短语中所选中的搜索短语、从所述第一资源项中所选中的搜索资源项和所述搜索资源项的特征；

[0067] 生成单元，用于根据所述真命题的组合，生成形式化查询语句。

[0068] 结合第二方面，在第二方面的第一种可能的实现方式中，所述不确定性推理基于马尔科夫逻辑网络 MLN，所述 MLN 包括预定义的一阶公式以及所述一阶公式的权重。

[0069] 结合第二方面或者第二方面的第一种可能的实现方式，在第二方面的第二种可能的实现方式中，

[0070] 所述获取单元，还用于从所述知识库中获取多个自然语言问句；

[0071] 所述短语检测单元，还用于对所述获取单元接收的所述问句进行短语检测，以确定第一候选短语；

[0072] 所述映射单元，还用于将所述第二候选短语映射到所述知识库中的第二资源项，其中，所述第二资源项与所述第二候选短语具有一致的语义；

[0073] 所述第一确定单元，还用于根据所述第二候选短语和所述第二资源项，确定与所述多个自然语言问句对应的观察谓词的值；

[0074] 所述获取单元，还用于获取人工标注的与所述多个自然语言问句对应的隐含谓词的值；

[0075] 所述第二确定单元，还用于根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重。

[0076] 结合第二方面的第二种可能的实现方式，在第二方面的第三种可能的实现方式中，所述一阶公式包括布尔公式和加权公式，所述布尔公式的权重为 + ∞，所述加权公式的权重为加权公式权重，所述人工标注的与所述多个自然语言问句对应的隐含谓词的值满足所述布尔公式，

[0077] 所述第二确定单元，具体用于：根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述加权公式权重。

[0078] 结合第二方面的第二种可能的实现方式，在第二方面的第四种可能的实现方式中，所述第二确定单元，具体用于：

[0079] 根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，采用差额注入松弛算法 MIRA，确定所述一阶公式的权重。

[0080] 结合上述第二方面的任一种可能的实现方式，在第二方面的第五种可能的实现方式中，所述 MLN 表示为 M，所述一阶公式表示为 Φ_i，所述一阶公式的权重表示为 w_i，所述命题集合表示为 y，

[0081] 所述第二确定单元，具体用于：

[0082] 根据所述观察谓词的值和所述隐含谓词构建可能的世界，所述可能的世界表示为 y；

[0083] 根据 $p(\mathbf{y})=\frac{1}{Z}\exp\left(\sum_{(\phi_i,w_i)\in M} w_i \sum_{\mathbf{c}\in C^{n\phi_i}} f_{\mathbf{c}}^{\phi_i}(\mathbf{y})\right)$，计算所述每一个命题集合的置信度，

[0084] 其中，Z 为归一化常数，$C^{n\phi_i}$ 为与一阶公式 Φ_i 对应的子公式的集合，c 为 $C^{n\phi_i}$ 的所述子公式的集合中的一个子公式，$f_{\mathbf{c}}^{\phi_i}$ 为二值函数，$f_{\mathbf{c}}^{\phi_i}(\mathbf{y})$ 表示在所述命题集合 y 下，所述一阶公式的真假。

[0085] 结合第二方面或者上述第二方面的任一种可能的实现方式，在第二方面的第六种可能的实现方式中，所述获取单元，具体用于：

[0086] 确定所述置信度的值最大的命题集合，并获取所述置信度的值最大的命题集合中的真命题的组合。

[0087] 结合第二方面或者上述第二方面的任一种可能的实现方式，在第二方面的第七种可能的实现方式中，

[0088] 所述第一候选短语的特征包括所述第一候选短语在所述问句中的位置、所述第一候选短语的主要词的词性、所述第一候选短语两两之间的依存路径上的标签，

[0089] 所述第一资源项的特征包括所述第一资源项的类型、所述第一资源项两两之间的相关性值、所述第一资源项两两之间的参数匹配关系，

[0090] 所述第一候选短语与所述第一资源项的关系包括所述第一候选短语与所述第一资源项的先验匹配得分，

[0091] 所述第一确定单元，具体用于：

[0092] 确定所述第一候选短语在所述问句中的位置；

[0093] 采用 stanford 词性标注工具，确定所述第一候选短语的主要词的词性；

[0094] 采用 stanford 依存句法分析工具，确定所述第一候选短语两两之间的依存路径上的标签；

[0095] 从所述知识库中确定所述第一资源项的类型，其中，所述类型为实体或类别或关系；

[0096] 从所述知识库中确定所述第一资源项两两之间的参数匹配关系；

[0097] 将所述第一资源项两两之间的相似性系数，作为所述两个第一资源项两两之间的相关性值；

[0098] 计算所述第一候选短语与所述第一资源项之间的先验匹配得分，所述先验匹配得分用于表示所述第一候选短语映射到所述第一资源项的概率。

[0099] 结合第二方面或者上述第二方面的任一种可能的实现方式，在第二方面的第八种

可能的实现方式中，所述形式化查询语句为简单协议资源描述框架查询语句 SPARQL。

[0100] 结合第二方面的第八种可能的实现方式，在第二方面的第九种可能的实现方式中，所述生成单元，具体用于：

[0101] 根据所述真命题的组合，利用 SPARQL 模板生成所述 SPARQL。

[0102] 结合第二方面的第九种可能的实现方式，在第二方面的第十种可能的实现方式中，所述 SPARQL 模板包括 ASK WHERE 模板、SELECT COUNT(？ url)WHERE 模板和 SELECT？url WHERE 模板，

[0103] 所述生成单元，具体用于：

[0104] 当所述问句为 Yes/No 问题时，根据所述真命题的组合，使用所述 ASK WHERE 模板生成所述 SPARQL；

[0105] 当所述问句为 Normal 问题时，根据所述真命题的组合，使用所述 SELECT？url WHERE 模板生成所述 SPARQL；

[0106] 当所述问句为 Number 问题时，根据所述真命题的组合，使用所述 SELECT？url WHERE 模板生成所述 SPARQL，或者，当使用所述 SELECT？url WHERE 模板生成的 SPARQL 无法得到数值型答案时，使用所述 SELECT COUNT(？url)WHERE 模板生成所述 SPARQL。

[0107] 结合第二方面或者上述第二方面的任一种可能的实现方式，在第二方面的第十一种可能的实现方式中，所述短语检测单元，具体用于：

[0108] 将所述问句中的词序列作为所述第一候选短语，其中，所述词序列满足：

[0109] 所述词序列中所有连续的非停用词都以大写字母开头，或者，若所述词序列中所有连续的非停用词不都以大写字母开头，则所述词序列的长度小于四；

[0110] 所述词序列的主要词的词性为 jj 或 nn 或 rb 或 vb，其中，jj 为形容词，nn 为名词，rb 为副词，vb 为动词；

[0111] 所述词序列所包括的词不全为停用词。

[0112] 本发明实施例基于预定义的不确定性推理网络，能够用于将用户输入的自然语言问句转换为结构化的 SPARQL。本发明实施例中，该预定义的不确定性推理网络能够应用于任何领域的知识库，具有领域扩展性，这样无需针对知识库人工地配置转换规则。

附图说明

[0113] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0114] 图 1 是本发明一个实施例的知识库中问句解析的方法的流程图。

[0115] 图 2 是本发明一个实施例的依存分析树的一例。

[0116] 图 3 是本发明另一个实施例的知识库中问句解析的方法的示意图。

[0117] 图 4 是本发明一个实施例的资源项查询图的另一例。

[0118] 图 5 是本发明一个实施例的确定加权公式权重的方法的流程图。

[0119] 图 6 是本发明一个实施例的问句解析的设备的框图。

[0120] 图 7 是本发明另一个实施例的问句解析的设备的框图。

具体实施方式

[0121] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0122] 在知识库问答系统中，需将自然语言问句(natural language question)转换为形式化查询语句。例如，形式化查询问句为结构化查询语句(Structure Query Language，SQL)或SPARQL。一般地，SPARQL以主体-属性-对象(subject-property-object，SPO)三元组形式(triple format)表示。

[0123] 例如：自然语言问句“Which software has been developed by organization founded in California,USA？”所对应的SPARQL为：

[0124] ？url_answer rdf:type dbo:Software

[0125] ？url_answer db:developer ？x1

[0126] ？x1 rdf:type dbo:Company

[0127] ？x1 dbo:foundationPlace dbr:California。

[0128] 将自然语言问句转换为形式化查询语句，需要依赖于针对于知识库的转换规则。也就是说，不同的知识库所对应的转换规则也是不同的。但是目前的问答系统中，需要人工地对每个知识库的转换规则进行人工配置。对于某一个知识库，人工地收集一些问题，并确定问题的答案，根据这些问题人工地总结出一些规律作为转换规则。也就是说，人工配置的转换规则没有领域扩展性，针对某一个知识库所配置的转换规则不能用于另外一个知识库。并且，由于自然语言问句中存在大量的歧义，也会导致人工配置的转换规则缺乏鲁棒性。

[0129] 自然语言处理(Natural Language Processing，NLP)是计算科学、人工智能和语言学科中用于描述机器语言与自然语言之间的关系的工具。NLP涉及人机交互。NLP的任务(tasks)可以包括：自动监督(Automatic summarization)、互参分辨率(Coreference resolution)、引语分析(Discourse analysis)、机器翻译(Machine translation)、形态分割(Morphological segmentation)、命名实体识别(Named entity recognition，NER)、自然语言生成(Natural language generation)、自然语言理解(Natural language understanding)、光学字符识别(Optical character recognition，OCR)、词性标注(Part-of-speech tagging)、句法分析(Parsing)、问答系统(Question answering)、关系提取(Relationship extraction)、断句(Sentence breaking)、情绪分析(Sentiment analysis)、语音识别(Speech recognition)、语音分割(Speech segmentation)、话题分割与识别(Topic segmentation and recognition)、词分割(Word segmentation)、词义消歧(Word sense disambiguation)、信息检索(Information retrieval，IR)、信息抽取(Information extraction，IE)、语音处理(Speech processing)等。

[0130] 具体地，斯坦福(Stanford)自然语言处理(Natural Language Processing，NLP)工具是针对上述NLP的不同任务所设计的。本发明实施例中采用了Stanford NLP工具。例如，其中的词性标注工具可以用于确定一个问句中的每一个单词(word)的词性

(Part-of-speech)。

[0131] 不确定性推理泛指除精确推理以外的其他各种推理问题。包括不完备、不精确知识的推理，模糊知识的推理，非单调性推理等。

[0132] 不确定性推理过程实际上是一种从不确定的初始证据出发，通过运用不确定性知识，最终推出具有一定不确定性但却又是合理或基本合理的结构的思维过程。

[0133] 不确定性推理的类型有数值方法和非数值方法，其中数值方法包括基于概率的方法。具体地，基于概率的方法是基于概率论的有关理论发展起来的方法，如可信度方法，主观贝叶斯 (Bayes) 方法、证据理论等。

[0134] 其中，马尔科夫逻辑网络是不确定性推理网络中较为常用的一种。

[0135] 马尔科夫逻辑网络 (Markov Logic Network，MLN) 是一种结合一阶逻辑 (First-Order Logic，FOL) 和马尔科夫网络 (Markov Network) 的统计关系学习 (Statistical Relational Learning) 框架。马尔科夫逻辑网络与传统的一阶逻辑的不同之处在于：传统的一阶逻辑要求所有的规则之间不允许有冲突，如果某一个命题不能同时满足所有规则，则其为假；而在马尔科夫逻辑网络中，每个规则都有一个权重，一个命题会按照一个概率为真。

[0136] 其中，一阶逻辑 (First-Order Logic，FOL) 也可以称为谓词逻辑或一阶谓词逻辑，由若干一阶谓词规则组成。一阶谓词规则由四种类型的符号组成，即常量、变量、函数和谓词。其中，常量指定义域里一个简单的对象；变量可以指定义域里若干对象；函数表示一组对象到一个对象的映射；谓词指定义域中若干对象之间的关系、或者对象的属性。变量和常量可以有类型。一个类型的变量仅能从定义类型的对象集中取值。一个项可以是任意地表示一个对象的表达式。原子是作用于一组项的谓词。一个常项是指没有变量的项。一个闭原子 (ground atom) 或闭谓词 (ground predicate) 是指所有参数均为常项的原子或谓词。一般地，规则是从原子开始，用连接词 (如蕴含关系、等价关系等) 和量词 (如全称量词和存在量词) 递归地建立起来。在一阶逻辑中，通常把规则表示成从句的形式。一个可能世界 (a possible world) 是指给所有可能出现的闭原子都赋予了真值。一阶逻辑可看作是在一个可能世界的集合上建立一系列硬规则，即如果一个世界违反了其中的某一条规则，那么这个世界的存在概率即为零。

[0137] MLN 的基本思想是让那些硬规则有所松弛，即当一个世界违反了其中的一条规则，那么这个世界存在的可能性将降低，但并非不可能。一个世界违反的规则越少，那么这个世界存在的可能性就越大。为此，给每个规则都加上了一个特定的权重，它反映了对满足该规则的可能世界的约束力。若一个规则的权重越大，则对于满足和不满足该规则的两个世界而言，它们之间的差异将越大。

[0138] 这样，通过设计不同的一阶逻辑公式 (高阶规则模板)，马尔科夫逻辑网络能够很好的结合语言特征和知识库限制。该概率框架中的逻辑公式能够对软规则限制进行建模。马尔科夫逻辑 (Markov Logic) 中一组加权的公式集合就称为一个马尔科夫逻辑网络。

[0139] 具体地，在 MLN 中，可以包括一阶公式和惩罚 (penalty)。闭原子可以以某种惩罚违法对应的一阶公式。

[0140] 其中，一阶公式中包括一阶谓词、逻辑联结词 (logical connectors) 和变量。

[0141] 图 1 是本发明一个实施例的知识库中问句解析的方法的流程图。图 1 所示的方法

包括：

[0142] 101，接收用户输入的问句。

[0143] 102，对所述问句进行短语检测，以确定第一候选短语。

[0144] 103，将所述第一候选短语映射到所述知识库中的第一资源项，其中，所述第一资源项与所述第一候选短语具有一致的语义。

[0145] 104，根据所述对应候选短语和所述对应资源项，计算观察谓词的值和可能的问句分析空间，其中，所述观察谓词用于表示所述第一候选短语的特征、所述第一资源项的特征和所述第一候选短语与所述第一资源项的关系，所述可能的问句分析空间中的点为命题集合，所述命题集合中的命题的真假由隐含谓词的值表征。

[0146] 105，对所述可能的问句分析空间中的每一个命题集合，根据所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度。

[0147] 106，获取所述置信度满足预设条件的命题集合中的真命题的组合，其中，所述真命题用于表示从所述第一候选短语中所选中的搜索短语、从所述第一资源项中所选中的搜索资源项和所述搜索资源项的特征。

[0148] 107，根据所述真命题的组合，生成形式化查询语句。

[0149] 本发明实施例利用观察谓词和隐含谓词，进行不确定性推理，能够将自然语言问句转化为形式化查询语句。并且，本发明实施例中，不确定性推理的方法能够应用于任何领域的知识库，具有领域扩展性，这样无需针对知识库人工地配置转换规则。

[0150] 可理解，本发明实施例中，在101中用户输入的问句为自然语言问句(natural language question)。

[0151] 例如，该自然语言问句为“Give me all actors who were born in Berlin.”。

[0152] 进一步地，在102中，可通过短语检测(phrase detection)，识别出问句中的词(token)序列。可选地，可将所述问句中的词序列作为所述第一候选短语。其中，词序列又称为多词序列或词语序列或词项或n元词序列或n-gram(s)，是指n个连续的单词组成的序列。

[0153] 可理解，102中可确定多个第一候选短语。

[0154] 可选地，102中，可将满足如下的限定的词序列作为第一候选短语：

[0155] (1)、所述词序列中所有连续的非停用词都以大写字母开头；或者，若所述词序列中所有连续的非停用词不都以大写字母开头，则所述词序列的长度小于四。

[0156] (2)、所述词序列的主要词(head word)的词性为jj或nn或rb或vb，其中，jj为形容词，nn为名词，rb为副词，vb为动词。

[0157] (3)、所述词序列所包括的词不全为停用词。

[0158] 同时，所有连续的大写字母开头的非停用词必须在同一个词序列中。

[0159] 可理解，本发明实施例中，head word也可以称为重要词或主导词等，并且可以从词性标注集合中获取词性的表示符号。

[0160] 举例来说，“United States Court of Appeals for the District of Columbia Circuit”中所有连续的非停用词都以大写字母开头，为一个候选短语。可以理解，所有连续的非停用词都以大写字母开头的词序列一般为专有名词。

[0161] 其中，词序列的长度是指词序列所包括的词的个数。例如，词序列“born in”的长

度为 2。

[0162] 其中，可以采用 stanford 的词性标注工具来确定每一个词的词性。

[0163] 举例来说，英文的停用词 (stop words) 有“a”、“an”、“the”、“that”等。中文的停用词有“一个”、“一些”、“不但”等。

[0164] 例如，在问句“Give me all actors who were born in Berlin”中，所确定的第一候选短语包括 :actors、who、born in、in、Berlin。

[0165] 具体地，可以表示为表一的形式，其中表一的第一列为所述第一候选短语的短语标识。

[0166] 表一

[0167]

11	actors
12	who
13	born in
14	in
15	Berlin

[0168] 本发明实施例中，103 可以理解为是将每个第一候选短语映射到知识库中的第一资源项。本发明实施例中，103 也可以称为短语映射 (phrase mapping)。具体地，一个第一候选短语可能映射到多个第一资源项。第一资源项的类型可以为实体 (Entity) 或类别 (Class) 或关系 (Relation)。

[0169] 举例来说，假设该知识库为 DBpedia。103 具体为 :

[0170] 将第一候选短语映射到实体 (Entity)，考虑到 DBpedia 中的实体来自于 Wikipedia 中的实体页面，首先收集 Wikipedia 中的锚文本 (anchor text)、重定向页面和消歧页面，并利用 Wikipedia 中的锚文本、重定向页面和消歧页面构建第一候选短语与实体之间的对应辞典，当第一候选短语匹配到实体的提及 (mention) 短语的时候，那么该实体即为与该第一候选短语语义一致的第一资源项。

[0171] 将第一候选短语映射到类别 (Class)，考虑到有词汇变种的情况，特别是同义词，例如，短语 film、movie 和 show 都可以映射到类别 dbo:Film。首先利用 word2vec 工具把第一候选短语中所有的词转换为向量形式，知识库中类别的向量形式为其标签 (对应 rdfs:label 关系) 的向量形式 ;然后计算第一候选短语与每个类别在向量上的余弦相似度 ;最后将余弦相似度值最大的 N 个类别作为与该第一候选短语语义一致的第一资源项。

[0172] 其中，word2vec 工具是一种将词 (word) 转换成向量 (vector) 的工具。例如，可以是由谷歌 (google) 开发并提供的一段开放代码，具体可以参见 :http://code.google.com/p/word2vec/。

[0173] 将第一候选短语映射到关系 (Relation)，使用 PATTY 和 ReVerb 所定义的关系模板作为资源。首先计算 DBpedia 中的关系与 PATTY 和 ReVerb 所定义的关系模板 (relation patterns) 在实例上的对齐，也就是统计 DBpedia 中满足关系模板的关系的实例对。然后，

如果第一候选短语能够匹配关系模板，那么，将满足关系模板的关系作为与该第一候选短语语义一致的第一资源项。

[0174] 其中，PATTY 和 ReVerb 所定义的关系模板可以参见 Nakashole 等人于 2012 在 EMNLP 发表的“Patty:a taxonomy of relational patterns with semantic types”，以及 Fader 等人于 2011 在 EMNLP 发表的“Identifying relations for open information extraction”。

[0175] 这样，通过 103，可以将第一候选短语映射到第一资源项，具体地，每一个第一候选短语映射到至少一个第一资源项。并且，具有映射关系的第一候选短语和第一资源项具有一致的语义。

[0176] 其中，若一个第一候选短语映射到多个第一资源项，说明该一个第一候选短语具有歧义。

[0177] 例如，在问句“Give me all actors who were born in Berlin”中，在 103 中，可确定第一候选短语 actors、who、born in、in、Berlin 映射为第一资源项如表二所示。其中，表二的第一列为第一候选短语，第二列为第一资源项，第三列为第一资源项的标识。并且，第一候选短语“in”映射到五个第一资源项。

[0178] 表二

[0179]

actors	dbo:Actor	21
who	dbo:Person	22
born in	dbo:birthPlace	23
in	dbo:headquarter	24
in	dbo:league	25
in	dbo:location	26
in	dbo:ground	27
in	dbo:locationCity	28
Berlin	dbr:Berlin	29

[0180] 本发明实施例中，104 可以理解为是特征抽取 (feature extraction) 的过程。

[0181] 具体地，本发明实施例定义隐含谓词 (hidden predicates)。隐含谓词可以包括如下的形式：

[0182] hasphrase(p)，表示候选短语 p 被选中。

[0183] hasResource(p,r)，表示资源项 r 被选中，且候选短语 p 映射到资源项 r。

[0184] hasRelation(p,r,rr)，表示资源项 p 和资源项 r 之间的参数匹配关系 rr 被选中。

[0185] 可理解，其中，p 可以为候选短语的短语标识，p 和 r 可以为资源项的标识。其中，

参数匹配关系 rr 为可以为以下一种 :1_1、1_2、2_1 和 2_2。

[0186] 具体地，本发明实施例中，参数匹配关系 rr 可以为以下一种 :1_1、1_2、2_1 和 2_2。那么，资源项 p 和资源项 r 之间的参数匹配关系为 m1_m2 表示资源项 p 的第 m1 个参数与资源项 r 的第 m2 个参数对齐。其中，m1 为 1 或 2，m2 为 1 或 2。

[0187] 如表三所示，为上述参数匹配关系的具体举例。其中，表三的第三列给出了一个问句，以解释第二列中的参数匹配关系。

[0188] 表三

[0189]

1_1	dbo:height 1_1 dbr:Michael Jordan	How tall is Michael Jordan?
1_2	dbo:River 1_2 dbo:crosses	Which river does the Brooklyn Bridge cross?
2_1	dbo:creator 2_1 dbr:Walt Disney	Which television shows were created by Walt Disney?
2_2	dbo:birthplace 2_2 dbo:capital	Which actors were born in the capital of American?

[0190] 其中，“dbo:height 1_1dbr:Michael Jordan”表示资源项 dbo:height 与资源项 dbr:Michael Jordan 之间的参数匹配关系为 1_1。即，资源项 dbo:height 的第 1 个参数与资源项 dbr:Michael Jordan 的第 1 个参数对齐。

[0191] 可理解，隐含谓词的值为 1 表示相应的候选短语、资源项、资源项和资源项之间的参数匹配关系被选中。隐含谓词的值为 0 表示相应的候选短语、资源项、资源项和资源项之间的参数匹配关系没有被选中。换句话说，隐含谓词的值为 1 表示相应的命题为真，隐含谓词的值为 0 表示相应的命题为假。

[0192] 例如，结合表一，hasphrase(11) = 1，表示“候选短语 actors 被选中”这个命题为真。hasphrase(11) = 1，表示“候选短语 actors 被选中”这个命题为假。

[0193] 这样，对于 102 和 103 所确定第一候选短语和第一资源项，能够基于隐含谓词构建可能的问句分析空间 (possible question parse space)。具体地，可能的问句分析空间中的一个点表示一个命题集合。一个命题集合包括一组命题，并且这一组命题是由一组隐含谓词的值还表示的。可理解，一个命题集合中的一组命题的真假由对应的隐含谓词的值来表征。

[0194] 具体地，本发明实施例还定义观察谓词 (observed predicates) 用于表示所述第一候选短语的特征、所述第一资源项的特征和所述第一候选短语与所述第一资源项的关系。

[0195] 其中，所述第一候选短语的特征包括所述第一候选短语在所述问句中的位置、所述第一候选短语的主要词的词性、所述第一候选短语两两之间的依存路径上的标签等。

[0196] 其中，所述第一资源项的特征包括所述第一资源项的类型、所述第一资源项两两之间的相关性值、所述第一资源项两两之间的参数匹配关系等。

[0197] 其中，所述第一候选短语与所述第一资源项的关系包括所述第一候选短语与所述

第一资源项的先验匹配得分。

[0198] 那么，可理解，104 中确定观察谓词的值包括：确定所述第一候选短语在所述问句中的位置；采用 stanford 的词性标注工具，确定所述第一候选短语的主要词的词性；采用 stanford 依存句法分析工具，确定所述第一候选短语两两之间的依存路径上的标签；从所述知识库中确定所述第一资源项的类型，其中，所述类型为实体或类别或关系；从所述知识库中确定所述第一资源项两两之间的参数匹配关系，其中，所述参数匹配关系为以下一种：1_1、1_2、2_1 和 2_2。将所述第一资源项两两之间的相似性系数，作为所述两个第一资源项两两之间的相关性值；计算所述第一候选短语与所述第一资源项之间的先验匹配得分，所述先验匹配得分用于表示所述第一候选短语映射到所述第一资源项的概率。

[0199] 具体地，从所述知识库中确定所述第一资源项两两之间的参数匹配关系，包括：从所述知识库中确定第一资源项 r1 和第一资源项 r2 之间的参数匹配关系 m1_m2，用于表示所述第一资源项 r1 的第 m1 个参数与所述第一资源项 r2 的第 m2 个参数对齐。其中，所述第一资源项包括所述第一资源项 r1 和所述第一资源项 r2，m1 为 1 或 2，m2 为 1 或 2。

[0200] 具体地，观察谓词可以包括如下的形式：

[0201] phraseIndex(p,i,j)，表示候选短语 p 在问句中的起始位置 i 和结束位置 j。

[0202] phrasePosTag(p,pt)，表示候选短语 p 的主要词 (head word) 的词性 pt。

[0203] 具体地，可以采用 stanford 词性标注工具确定主要词的词性。

[0204] phraseDepTag(p,q,dt)，表示候选短语 p 和候选短语 q 之间的依存路径上的标签 dt。

[0205] 具体地，可以采用 stanford 依存分析 (stanford dependency parser) 工具建立问句的依存分析树 (dependency parse trees)，根据所述依存分析树进行特征提取，从而确定两个候选短语之间的依存路径上的标签。

[0206] 例如，问句“Give me all actors who were born in Berlin.”的依存分析树如图 2 所示。

[0207] phraseDepOne(p,q)，表示当候选短语 p 和候选短语 q 之间的依存路径上的标签只有一个时，该谓词为真，否则为假。

[0208] 可理解，观察谓词中的谓词 phraseDepOne(p,q) 只包括结果为真的谓词。

[0209] hasMeanWord(p,q)，表示当候选短语 p 和候选短语 q 之间的依存路径上的词全部为停用词或者词性为 dt、in、wdt、to、cc、ex、pos 或 wp 时，hasMeanWord(p,q) 为假，否则为真。

[0210] 其中，dt 为限定词，in 为介词 in，wdt 为以 w 开头的疑问词，to 为介词 to，cc 为连接词，ex 为存在词 there，pos 为所有格结尾词，wp 为疑问代词。其中，以 w 开头的疑问词如 what、which 等，连接词如 and、but、or 等。具体地，可以从词性标注集合获取上述词性的表示符号。

[0211] 可理解，观察谓词中的谓词 hasMeanWord(p,q) 只包括结果为真的谓词。

[0212] resourceType(r,rt)，表示资源项 r 的类型为 rt。其中 rt 为 E 或 C 或 R。E 表示实体 (Entity)，C 表示类别 (Class)，R 表示关系 (Relation)。

[0213] priorMatchScore(p,r,s)，表示候选短语 p 与资源项 r 之间的先验匹配得分 s。

[0214] 举例来说，假设知识库为 DBpedia。

[0215] 具体地，若资源项 r 的类型为 E，首先收集 Wikipedia 中的锚文本、重定向页面和消歧页面，候选短语 p 匹配到资源项 r 的提及短语，可将对应的频率作为先验匹配得分。其中，对应的频率是指候选短语 p 链接到资源项 r 的次数除以候选短语 p 链出的总次数。

[0216] 具体地，若资源项 r 的类型为 C，候选短语 p 与资源项 r 的先验匹配得分可以为 $\gamma \cdot s_1+(1-\gamma) \cdot s_2$。其中，$\gamma$ 为 0 至 1 之间的任意值，例如 $\gamma = 0.6$。s_1 为资源项 r 的标签与候选短语 p 之间的 Levenshtein 距离，s_2 为候选短语 p 的向量与资源项 r 的向量之间的余弦相似性度量值。其中，Levenshtein 距离可以参见 Navarro 于 2001 年在 ACM Comput. Surv. 发表的“A guided tour to approximate string matching”。其中，s_2 的计算可以参见 Mikolov 等人于 2010 年在 INTERSPEECH 发表的“Recurrent neural network based language model”。

[0217] 具体地，若资源项 r 的类型为 R，候选短语 p 与资源项 r 的先验匹配得分可以为 $\alpha \cdot s_1+\beta \cdot s_2+(1-\alpha-\beta) \cdot s_3$。其中，$\alpha$ 和 β 为 0 至 1 之间的任意值，且 $\alpha+\beta < 1$，例如 $\alpha = 0.3$，$\beta = 0.3$。s_1 为资源项 r 的标签与候选短语 p 之间的 Levenshtein 距离，s_2 为候选短语 p 的向量与资源项 r 的向量之间的余弦相似性度量值，s_3 为资源项 r 与关系模板的匹配集合的 Jaccard 系数。其中，关系模板为如前所述的 PATTY 和 ReVerb 所定义的关系模板。s_3 的计算可以参见 Yahya 等人于 2012 年在 EMNLP 发表的“Natural language questions for the web of data”。

[0218] hasRelatedness(p,q,s)，表示资源项 p 和资源项 q 之间的相关性值 s。该相关性值 s 的取值区间为 0 至 1。具体地，该相关性值 s 可以为资源项 p 和资源项 q 的相似性系数。可选地，该相似性系数也可以称为 Jaccard 相似性系数或 Jaccard 系数或相似度评价系数。

[0219] 例如，参见 Yahya 等人于 2012 年在 EMNLP 发表的“Natural language questions for the web of data”，资源项 p 和资源项 q 的相似性系数可以等于资源项 p 和资源项 q 的入度集合的 Jaccard 系数。

[0220] isTypeCompatible(p,q,rr)，表示资源项 p 和资源项 q 之间的参数匹配关系 rr。

[0221] 具体地，本发明实施例中，参数匹配关系 rr 可以为以下一种：1_1、1_2、2_1 和 2_2。具体地，参数匹配关系可如前所述，为避免重复，这里不再赘述。

[0222] hasQueryResult(p,q,o,rr1,rr2)，表示资源项 p、资源项 q 和资源项 o 之间的参数匹配关系。具体地，资源项 p 和资源项 q 之间具有参数匹配关系 rr1，资源项 q 和资源项 o 之间具有参数匹配关系 rr2。

[0223] 可理解，上述所描述的观察谓词中，phraseIndex(p,i,j)、phrasePosTag(p,pt)、phraseDepTag(p,q,dt)、phraseDepOne(p,q) 和 hasMeanWord(p,q) 用于表示所述候选短语的特征。resourceType(r,rt)、hasRelatedness(p,q,s)、isTypeCompatible(p,q,rr) 和 hasQueryResult(p,q,o,rr1,rr2) 用于表示所述资源项的特征。priorMatchScore(p,r,s) 用于表示所述候选短语与所述资源项之间的关系。

[0224] 其中，p 和 q 可以为候选短语的短语标识，p、q、r 和 o 可以为资源项的标识。

[0225] 这样，基于 102 和 103 所确定的第一候选短语和第一资源项，能够确定相应的观察谓词的值。

[0226] 例如，对问句“Give me all actors who were born in Berlin”，在表一和表二的

基础上，可以在 104 计算观察谓词的值。具体地，其中观察谓词的值为 1 的表达式包括：

[0227] phraseIndex(11,3,3)

[0228] phraseIndex(12,4,4)

[0229] phraseIndex(13,6,7)

[0230] phraseIndex(14,7,7)

[0231] phraseIndex(15,8,8)

[0232] phrasePosTag(11,nn)

[0233] phrasePosTag(12,wp)

[0234] phrasePosTag(13,vb)

[0235] phrasePosTag(14,in)

[0236] phrasePosTag(15,nn)

[0237] phraseDepTag(11,13,rcmod)

[0238] phraseDepTag(12,13,nsubjpass)

[0239] phraseDepTag(12,14,nsubjpass)

[0240] phraseDepTag(13,15,pobj)

[0241] phraseDepTag(14,15,pobj)

[0242] phraseDepOne(11,13)

[0243] phraseDepOne(12,13)

[0244] phraseDepOne(12,14)

[0245] phraseDepOne(13,15)

[0246] phraseDepOne(14,15)

[0247] hasMeanWord(12,14)

[0248] resourceType(21,E)

[0249] resourceType(22,E)

[0250] resourceType(23,R)

[0251] resourceType(24,R)

[0252] resourceType(25,R)

[0253] resourceType(26,R)

[0254] resourceType(27,R)

[0255] resourceType(28,R)

[0256] resourceType(29,E)

[0257] priorMatchScore(11,21,1.000000)

[0258] priorMatchScore(12,22,1.000000)

[0259] priorMatchScore(13,23,1.000000)

[0260] priorMatchScore(14,24,1.000000)

[0261] priorMatchScore(14,25,1.000000)

[0262] priorMatchScore(14,26,1.000000)

[0263] priorMatchScore(14,27,1.000000)

[0264] priorMatchScore(14,28,1.000000)

[0265] priorMatchScore(15,29,1.000000)

[0266] hasRelatedness(21,23,1.000000)

[0267] hasRelatedness(22,23,1.000000)

[0268] hasRelatedness(22,24,0.440524)

[0269] hasRelatedness(22,25,0.425840)

[0270] hasRelatedness(22,26,0.226393)

[0271] hasRelatedness(22,27,0.263207)

[0272] hasRelatedness(23,29,0.854583)

[0273] hasRelatedness(24,29,0.816012)

[0274] hasRelatedness(26,29,0.532818)

[0275] hasRelatedness(27,29,0.569732)

[0276] hasRelatedness(28,29,0.713400)

[0277] isTypeCompatible(21,23,1_1)

[0278] isTypeCompatible(22,23,1_1)

[0279] isTypeCompatible(22,23,1_2)

[0280] isTypeCompatible(22,24,1_2)

[0281] isTypeCompatible(22,25,1_1)

[0282] isTypeCompatible(22,26,1_1)

[0283] isTypeCompatible(22,26,1_2)

[0284] isTypeCompatible(22,27,1_2)

[0285] isTypeCompatible(23,29,2_1)

[0286] isTypeCompatible(24,29,2_1)

[0287] isTypeCompatible(26,29,2_1)

[0288] isTypeCompatible(27,29,2_1)

[0289] isTypeCompatible(28,29,2_1)

[0290] hasQueryResult(21,23,29,1_1,2_1)

[0291] hasQueryResult(22,23,29,1_1,2_1)

[0292] hasQueryResult(22,26,29,1_1,2_1)

[0293] 可理解,观察谓词的值为 1,即表示对应的命题为真。

[0294] 例如,其中,phraseIndex(11,3,3) 的值为 1,表示“第一候选短语 actors 在问句中的起始位置 i 和结束位置 j 均为 3”这一命题为真。其中,11 为候选短语“actors”的短语标识,如表一所示。

[0295] 其中,phrasePosTag(13,vb) 的值为 1,表示“第一候选短语 born in 的主要词为 born,其词性 vb”这一命题为真。其中,13 为候选短语“born in”的短语标识,如表一所示。

[0296] 其中,phraseDepTag(13,15,pobj) 的值为 1,表示“第一候选短语 born in 和第一候选短语 Berlin 依存路径上的标签为 pobj”这一命题为真。其中,13 为候选短语“born in”的短语标识,15 为候选短语“Berlin”的短语标识,如表一所示。

[0297] 上述其他的观察谓词的值为 1 的表达式的含义可以参照上述解释,为避免重复,这里不再赘述。

[0298] 可理解，还包括观察谓词的值为 0 的表达式，为节省篇幅，这里不再罗列。

[0299] 可选地，本发明实施例中，也可以用谓词 resource 表示资源项的标识。

[0300] 例如，结合表二可知，以下谓词的值为 1 ：

[0301] resource(21,dbo:Actor)

[0302] resource(22,dbo:Person)

[0303] resource(23,dbo:birthPlace)

[0304] resource(24,dbo:headquarter)

[0305] resource(25,dbo:league)

[0306] resource(26,dbo:location)

[0307] resource(27,dbo:ground)

[0308] resource(28,dbo:locationCity)

[0309] resource(29,dbr:Berlin)

[0310] 可理解，本发明实施例中，102 和 103 中所确定的第一候选短语和第一资源项是有歧义的。本发明实施例通过不确定性推理来消除所述第一候选短语和所述第一资源项的歧义。

[0311] 不确定性推理是根据不确定性信息作出推理和决策。不确定性推理网络可以处理不完整的和带有噪音的数据集，用概率测度的权重来描述数据间的相关性，旨在解决数据的不一致性和不确定性。

[0312] 本发明实施例中，105 中的不确定性推理所使用的模型可以为如下的任意一种 ：贝叶斯网络 (Bayesian Network)、似然关系模型 (Probabilistic relational models)、贝叶斯逻辑程序模型 (Bayesian logic programs)、关系马尔科夫网 (Relational Markov Network)、马尔科夫逻辑网 (Markov Logic Network)、概率软化逻辑 (Probabilistic Soft Logic)。本发明对此不作限定。

[0313] 可选地，本发明实施例中，105 中的不确定性推理是基于马尔科夫逻辑网络 (Markov Logic Network，MLN) 的，其中，所述 MLN 包括预定义的一阶公式以及所述一阶公式的权重。也就是说，不确定性推理所使用的模型为 MLN。

[0314] 可选地，本发明实施例中，一阶公式可以包括布尔公式 (Boolean formulas) 和加权公式 (weighted formulas)。其中，布尔公式的权重为 + ∞，布尔公式可以理解为一阶逻辑中的一阶逻辑公式，表示硬规则 (hard constraints)，也可以称为硬公式 (hard formulas，hf)，是所有的闭原子必须满足的限制条件。加权公式的权重为加权公式权重，是软规则 (soft constraints)，也可以称为软公式 (soft formulas，sf)，闭原子可以以某种惩罚违法。

[0315] 其中，一阶公式是由一阶谓词、逻辑联结词和变量所组成的。其中，一阶谓词可以包括前述的观察谓词和隐含谓词。

[0316] 应注意，本发明实施例中，MLN 也可以包括二阶公式、一阶公式、所述二阶公式的权重、以及所述一阶公式的权重。或者，MLN 也可以包括更高阶的公式及权重，本发明对此不作限定。

[0317] 具体地，布尔公式如表四所示。其中，符号“_”表示逻辑变量中的任意常量。|•| 表示公式中为真的闭原子的个数。

[0318] 表四

[0319]

hf1	hasPhase(p)⇒hasResource(p,_)
hf2	hasResource(p,_)⇒hasPhase(p)
hf3	|hasResource(p,_)|≤1
hf4	!hasPhase(p)⇒!hasResource(p,r)
hf5	hasResource(_,r)⇒hasRelation(r,_,_)∨hasRelation(_,r,_)
hf6	|hasRelation(r1,r2,_)|≤1
hf7	hasRelation(r1,r2,_)⇒hasResource(_,r1)∧hasResource(_,r2)
hf8	phraseIndex(p1,s1,e1)∧phraseIndex(p2,s2,e2)∧overlap(s1,e1,s2,e2)∧hasPhase(p1)⇒!hasPhase(p2)
hf9	resourceType(r,E)⇒!hasRelation(r,_,2_1)∧!hasRelation(r,_,2_2)
hf10	resourceType(r,E)⇒!hasRelation(_,r,2_1)∧!hasRelation(r,_,2_2)
hf11	resourceType(r,C)⇒!hasRelation(r,_,2_1)∧!hasRelation(r,_,2_2)
hf12	resourceType(r,C)⇒!hasRelation(_,r,2_1)∧!hasRelation(r,_,2_2)
hf13	!isTypeCompatible(r1,r2,rr)⇒!hasRelation(r1,r2,rr)

[0320] 具体地，表四中的含义如下：

[0321] hf1：表示如果一个短语 p 被选中，那么该短语 p 至少映射到一个资源项。

[0322] hf2：表示如果一个短语 p 到资源项的映射被选中，那么该短语 p 必须被选中。

[0323] hf3：表示一个短语 p 只能映射到一个资源项。

[0324] hf4：表示如果一个短语 p 没有被选中，那么任何一个短语 p 到资源项的映射关系都不被选中。

[0325] hf5：表示如果一个短语到资源项 r 的映射被选中，那么，该资源项 r 至少与其他的一个资源项有关系。

[0326] hf6：表示两个资源项 r1 和 r2 只能有一个参数匹配关系。

[0327] hf7：表示如果两个资源项 r1 和 r2 存在参数匹配关系，那么，至少有一个短语到资源项 r1 的映射被选中且至少有一个短语到资源项 r2 的映射被选中。

[0328] hf8：表示任意两个被选中的短语没有重叠。这里的重叠可以用在问句中的位置表征。

[0329] hf9、hf10、hf11、hf12：表示如果一个资源项 r 的类型为实体或类别，那么，该资源项 r 不能有第二个参数与其他资源项对齐。

[0330] hf13：表示两个资源项 r1 和 r2 之间的参数匹配关系必须一致。

[0331] 可理解，表四中，逻辑联结词“∧”表示与 (and)，逻辑联结词“∨”表示或 (or)，逻辑联结词“！”表示非 (not)。

[0332] 具体地，加权公式如表五所示。其中，符号“+”表示逻辑变量的每个常数都应该单独设置权重。

[0333] 表五

[0334]

sf1	priorMatchScore(p,r,s)⇒hasPhase(p)
sf2	priorMatchScore(p,r,s)⇒hasResource(p,r)
sf3	phrasePosTag(p,pt+)∧resourceType(r,rt+)⇒hasResource(p,r)
sf4	phraseDepTag(p1,p2,dp+)∧hasResource(p1,r1)∧hasResource(p2,r2)⇒hasRelation(r1,r2,rr+)
sf5	phraseDepTag(p1,p2,dp+)∧hasResource(p1,r1)∧hasResource(p2,r2)∧!hasMeanWord(p1,p2)⇒hasRelation(r1,r2,rr+)
sf6	phraseDepTag(p1,p2,dp+)∧hasResource(p1,r1)∧hasResource(p2,r2)∧phraseDepOne(p1,p2)⇒hasRelation(r1,r2,rr+)
sf7	hasRelatedness(r1,r2,s)∧hasResource(_,r1)∧hasResource(_,r2)⇒hasRelation(r1,r2,_)
sf8	hasQueryResult(r1,r2,r3,rr1,rr2)⇒hasRelation(r1,r2,rr1) ∧hasRelation(r2,r3,rr2_)

[0335] 具体地，表五中的含义如下：

[0336] sf1、sf2：表示短语 p 映射到资源项 r 的先验匹配得分 s 越大，短语 p 和资源项 r 被选中的概率越大。

[0337] sf3：表示短语 p 的主要词的词性与该短语 p 映射到的资源项 r 的类型有某种关联。

[0338] sf4、sf5、sf6：表示两个短语 p1 和 p2 之间的依存路径上的标签与两个资源项 r1 和 r2 之间的参数匹配关系有某种关联，其中，短语 p1 映射到资源项 r1，短语 p2 映射到资源项 r2。

[0339] sf7：表示两个资源项 r1 和 r2 之间的相关性值越大，这两个资源项 r1 和 r2 之间有参数匹配关系的可能性越大。

[0340] sf8：表示如果一个资源项三元组存在查询结果，那么，这三个资源项之间应该具有相应的参数匹配关系。

[0341] 应注意，本发明实施例中，加权公式权重可以是人工设置的。例如，可以是由知识库的管理者或专家预设置的经验值。

[0342] 本发明实施例中，加权公式权重也可以是通过学习的方法，经过训练所得到的。

[0343] 可理解，对于不同的知识库，加权公式权重一般不同。本发明实施例中，表四所示的布尔公式可以理解为所有的知识库满足的通用的规则。表五所示的加权公式可以理解为针对不同的知识库，加权公式权重不同的特定的规则。

[0344] 本发明实施例中，也可以将布尔公式和加权公式统称为“元规则”。即，“元规则”是适用于不同的领域的知识库的规则。

[0345] 本发明实施例中，105 也可以称为推理（Inference）或联合推理（joint Inference）或联合消歧（joint disambiguation）。具体地，可以使用 thebeast 工具进行联合推理。可选地，对所述问句分析空间中的每一个命题集合，可以根据所述观察谓词的值和所述隐含谓词的值，，采用切平面方法（cutting plane method 或 cutting plane approach），计算所述每一个命题集合的置信度。具体地，thebeast 工具可以参见：https://code.google.com/p/thebeast/。

[0346] 可理解，置信度也可以称为可信度。并且，可以采用无向图模型的极大似然估计，计算所述每一个命题集合的置信度。

[0347] 可选地，所述 MLN 表示为 M，所述一阶公式表示为 Φ_i，所述一阶公式的权重表示为 w_i，所述命题集合表示为 y，那么，105 可以为：

[0348] 根据 $p(\mathbf{y})=\frac{1}{Z}\exp\left(\sum_{(\phi_i,w_i)\in M} w_i \sum_{\mathbf{c}\in C^{n\phi_i}} f_{\mathbf{c}}^{\phi_i}(\mathbf{y})\right)$，计算所述每一个命题集合的置信度。

[0349] 其中，Z 为归一化常数，$C^{n\phi_i}$ 为与一阶公式 Φ_i 对应的子公式的集合，c 为 $C^{n\phi_i}$ 的所述子公式的集合中的一个子公式，$f_{\mathbf{c}}^{\phi_i}$ 为二值函数，$f_{\mathbf{c}}^{\phi_i}(\mathbf{y})$ 表示在所述命题集合 y 下，所述一阶公式的真假。

[0350] 其中，二值函数（binary feature function）$f_{\mathbf{c}}^{\phi_i}$ 的值为 1 或 0。具体地，在所述命题集合 y 下，当子公式 c 为真时，$f_{\mathbf{c}}^{\phi_i}$ 为 1。否则为 0。

[0351] 可选地，在 105 中可以设置一个最大循环次数。例如，该最大循环次数为 100。

[0352] 这样，在 105 中计算每一个命题集合的置信度之后，可以得到与可能的问句分析空间对应的置信度集合，并且所述置信度集合中的每一个置信度对应一个命题集合。

[0353] 进一步地，在 106 中，可以从可能的问句分析空间的多个命题集合中选择一个或者几个命题集合，并且所选择的一个或者几个命题集合的置信度满足预设条件。

[0354] 可选地，在 106 中，可以确定置信度的值最大的一个命题集合，并获取所述置信度的值最大的一个命题集合中的真命题的组合。

[0355] 或者，可选地，在 106 中，可以确定置信度的值最大的多个命题集合，并获取所述置信度的值最大的多个命题集合中的真命题的组合。本发明对此不作限定。

[0356] 由于命题集合中的命题的真假由隐含谓词的值来表征，那么，可理解，在 106 中获取真命题的组合，即获取隐含谓词的值为 1 的组合。并且，所述真命题用于表示从所述第一候选短语中所选中的搜索短语、从所述第一资源项中所选中的搜索资源项和所述搜索资源项的特征。

[0357] 例如，对问句“Give me all actors who were born in Berlin.”，所确定的隐含谓词的值为 1 的表达式如下：

[0358] hasphrase(11)

[0359] hasphrase(13)

[0360] hasphrase(15)

[0361] hasResource(11,21)

[0362] hasResource(13,23)

[0363] hasResource(15,29)

[0364] hasRelation(21,23,1_1)

[0365] hasRelation(23,29,2_1)

[0366] 进一步地，可以在107生成形式化查询语句。可选地，形式化查询语句可以为SQL。或者，本发明实施例中，形式化查询语句可以为SPARQL，相应地，107也可以称为SPARQL生成(SPARQL Generation)的过程。

[0367] 可选地，107可以为：根据所述真命题的组合，利用SPARQL模板生成所述SPARQL。

[0368] 具体地，可利用所述真命题的组合，构建SPARQL的三元组，进一步，利用SPARQL模板生成SPARQL。

[0369] 具体地，自然语言问句可以分为三类：Yes/No，Number和Normal。相应地，SPARQL模板也包括ASK WHERE模板、SELECT COUNT(？url)WHERE模板和SELECT？url WHERE模板。

[0370] 那么，当所述问句为Yes/No问题时，根据所述真命题的组合，使用ASK WHERE模板生成所述SPARQL。

[0371] 当所述问句为Normal问题时，根据所述真命题的组合，使用SELECT？url WHERE模板生成所述SPARQL。

[0372] 当所述问句为Number问题时，根据所述真命题的组合，使用SELECT？url WHERE模板生成所述SPARQL，或者，当使用SELECT？url WHERE模板生成的SPARQL无法得到数值型答案时，使用SELECT COUNT(？url)WHERE模板生成所述SPARQL。

[0373] 例如，问句“Give me all actors who were born in Berlin.”为Normal问题，所生成的SPARQL为：

[0374] SELECT？url WHERE{

[0375] ？x rdf:type dbo:Actor.

[0376] ？x dbo:birthplace dbr:Berlin.

[0377] }

[0378] 可选地，107可包括：根据所述真命题的组合，生成查询资源图，其中，所述查询资源图包括顶点和边，所述顶点包括所述搜索短语、所述搜索资源项，并且，每个顶点中所述搜索短语映射到该顶点中所述搜索资源项。所述边表示相连的两个顶点中两个所述搜索资源项之间的参数匹配关系；进一步根据所述查询资源图生成所述SPARQL。

[0379] 具体地，可以将所述查询资源图中相互连接的三个所述搜索资源项，作为所述SPARQL的三元组，其中，位于所述相互连接的三个所述搜索资源项的中间的搜索资源项的类型为关系。

[0380] 这样，本发明实施例中，可以将自然语言问句转化为SPARQL。并且所采用的预定义的一阶公式是领域无关的，即预定义的布尔公式和加权公式可以运用于所有的知识库，具有可扩展性。也就是说，采用本发明实施例所提供的方法，无需人工地设置转换规则。

[0381] 举例来说，如图 3 所示，是本发明的问句解析的一例。

[0382] 301，接收用户输入的问句。假设该问句为自然语言问句：“Which software has been developed by organization founded in California, USA？”

[0383] 302，对 301 输入的问句进行短语检测 (phrase detection)，确定第一候选短语。

[0384] 具体地，302 可以参见前述实施例中的 102，为避免重复，这里不再赘述。

[0385] 例如，所确定的第一候选短语包括：software，developed，developed by，organizations，founded in，founded，California，USA。

[0386] 303，对 302 中所确定的第一候选短语进行短语映射 (phrase mapping)，将第一候选短语映射到第一资源项。

[0387] 具体地，303 可以参见前述实施例中的 103，为避免重复，这里不再赘述。

[0388] 例如，将第一候选短语 software 映射到：dbo:Software，dbr:Software 等。这里不再一一罗列。

[0389] 304，通过特征提取 (feature extraction)，确定观察谓词的值，并构建可能的问句分析空间。

[0390] 具体地，304 可以参见前述实施例中的 104，为避免重复，这里不再赘述。

[0391] 应注意，这里不再一一罗列。

[0392] 305，通过联合推理 (Inference)，计算所述每一个命题集合的置信度，并获取所述置信度满足预设条件的命题集合中的真命题的组合。

[0393] 具体地，305 可以参见前述实施例中的 105 和 106，为避免重复，这里不再赘述。

[0394] 其中，所述真命题的组合即其中隐含谓词的值为 1 的组合。

[0395] 例如，所确定的隐含谓词的值为 1 的表达式为

[0396] hasPhrase(software)，

[0397] hasPhrase(developed by)，

[0398] hasPhrase(organizations)，

[0399] hasPhrase(founded in)，

[0400] hasPhrase(California)；

[0401] hasResource(software,dbo:Software)，

[0402] hasResource(developed by,dbo:developer)，

[0403] hasResource(California,dbr:California)，

[0404] hasResource(organizations,dbo:Company)，

[0405] hasResource(founded in,dbo:foundationPlace)；

[0406] hasRelation(dbo:Software,dbo:developer,1_1)，

[0407] hasRelation(dbo:developer,dbo:Company,2_1)，

[0408] hasRelation(dbo:Company,dbo:foundationPlace,1_1)，

[0409] hasRelation(dbo:foundationPlace,dbr:California,2_1)。

[0410] 306，生成资源项查询图。

[0411] 具体地，该资源项查询图也可以称为语义项查询图 (Semantic Items Query Graph)。

[0412] 具体地，资源项查询图中的顶点可包括：搜索资源项、搜索资源项的类型、映射到

所述搜索资源项的搜索短语在问句中的位置。

[0413] 具体地，资源项查询图中的边包括：所述边相连的两个顶点中的两个搜索资源项之间的参数匹配关系。

[0414] 应注意，资源项查询图中搜索资源项之间的关系是二元关系。

[0415] 可选地，资源项查询图中的顶点可包括：搜索短语，搜索资源项、搜索资源项的类型、映射到所述搜索资源项的搜索短语、以及所述搜索短语在问句中的位置。如图 4 是资源项查询图的另一例。包括顶点 311 至 315。

[0416] 其中，顶点 311 包括：搜索资源项 dbo:Software、搜索资源项的类型 Class、搜索短语 Software 和搜索短语在问句中的位置 11。其中，搜索短语 Software 映射到搜索资源项 dbo:Software。

[0417] 其中，顶点 312 包括：搜索资源项 dbo:developer、搜索资源项的类型 Relation、搜索短语 developed by 和搜索短语在问句中的位置 45。其中，搜索短语 Software 映射到搜索资源项 dbo:Software。

[0418] 其中，顶点 313 包括：搜索资源项 dbo:Company、搜索资源项的类型 Class、搜索短语 organizations 和搜索短语在问句中的位置 66。其中，搜索短语 organizations 映射到搜索资源项 dbo:Company。

[0419] 其中，顶点 314 包括：搜索资源项 dbo:foundationPlace、搜索资源项的类型 Relation、搜索短语 founded in 和搜索短语在问句中的位置 78。其中，搜索短语 founded in 映射到搜索资源项 dbo:foundationPlace。

[0420] 其中，顶点 315 包括：搜索资源项 dbr:California、搜索资源项的类型 Entity、搜索短语 California 和搜索短语在问句中的位置 99。其中，搜索短语 California 映射到搜索资源项 dbr:California。

[0421] 其中，顶点 311 与顶点 312 之间的边 1_1 表示搜索资源项 dbo:Software 与搜索资源项 dbo:developer 之间的参数匹配关系为 1_1。

[0422] 其中，顶点 312 与顶点 313 之间的边 2_1 表示搜索资源项 dbo:developer 与搜索资源项 dbo:Company 之间的参数匹配关系为 2_1。

[0423] 其中，顶点 313 与顶点 314 之间的边 1_1 表示搜索资源项 dbo:Company 与搜索资源项 dbo:foundationPlace 之间的参数匹配关系为 1_1。

[0424] 其中，顶点 315 与顶点 314 之间的边 1_2 表示搜索资源项 dbr:California 与搜索资源项 dbo:foundationPlace 之间的参数匹配关系为 1_2。

[0425] 307，SPARQL 生成 (SPARQL genaration)。

[0426] 具体地，将资源项查询图中的二元关系转换为三元关系。

[0427] 也就是，将资源项查询图中相互连接的三个搜索资源项，具有三元关系，并且，位于所述相互连接的三个搜索资源项的中间的搜索资源项的类型为关系。

[0428] 例如，301 中的自然语言问句为 Normal 问题，使用 SELECT ? url WHERE 模板，所生成的 SPARQL 为：

[0429] SELECT ? url WHERE{

[0430] ? url_answer rdf:type dbo:Software

[0431] ? url_answer dbo:developer ? x1

[0432] ? x1 rdf:type dbo:Company

[0433] ? x1 dbo:foundationPlace dbr:California

[0434] }

[0435] 这样，本发明实施例中，可以将自然语言问句转化为SPARQL。并且所采用的预定义的一阶公式是领域无关的，即预定义的布尔公式和加权公式可以运用于所有的知识库，具有可扩展性。也就是说，采用本发明实施例所提供的方法，无需人工地设置转换规则。

[0436] 并且，可理解，本发明实施例中，预定义的布尔公式和加权公式是语言无关的，即具有语言扩展性。例如，既可以用于英文知识库，也可以用于中文知识库。

[0437] 如前所述，本发明实施例中，105中不确定性推理可以基于MLN。其中，所述MLN包括预定义的一阶公式以及所述一阶公式的权重。

[0438] 可选地，所述一阶公式可以包括布尔公式和加权公式。布尔公式的权重为+∞，加权公式的权重为加权公式权重。其中，加权公式权重可以是通过学习的方法，经过训练所得到的。那么，可理解，在101之前，如图5所示，还可包括：

[0439] 401，从所述知识库中获取多个自然语言问句。

[0440] 402，对所述多个自然语言问句进行短语检测，以确定所述多个自然语言问句的第二候选短语。

[0441] 403，将所述第二候选短语映射到所述知识库中的第二资源项，其中，所述第二资源项与所述第二候选短语具有一致的语义。

[0442] 404，根据所述第二候选短语和所述第二资源项，确定与所述多个自然语言问句对应的观察谓词的值。

[0443] 405，获取人工标注的与所述多个自然语言问句对应的隐含谓词的值。

[0444] 406，根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重。

[0445] 这样，本发明实施例中，基于预定义的一阶公式，能够通过学习的方法，确定针对知识库的一阶公式的权重，并可以作为针对知识库的转换规则。这样，无需人工设置转换规则，并且预定义的马尔科夫逻辑网络MLN的一阶公式具有可扩展性，能够应用与任何的知识库。

[0446] 具体地，问答系统知识库包括问题库，问题库中包括多个自然语言问句。那么，401中可以是从问答系统知识库中的问题库获取多个自然语言问句。本发明实施例对多个自然语言问句的数目不作限定。例如，多个自然语言问句可以为1千个自然语言问句。

[0447] 例如，可以从关联数据问答系统(Question Answering over Linked Data，QALD)的问题库Q1的训练集(training set)获取110个自然语言问题。

[0448] 本发明实施例中，402的过程可以参见前述实施例的102的过程，403的过程可以参见前述实施例的103的过程，404的过程可以参见前述实施例的104的过程。为避免重复，这里不再赘述。这样，针对401中的多个自然语言问句，能够确定与所述多个自然语言文件分别对应的观察谓词的值。

[0449] 可理解，在405之前，需人工地标注所述多个自然语言问句中每个自然语言问句对应的隐含谓词的值，也就是说，405中所获取的与所述多个自然语言问句对应的隐含谓词的值是人工标注的(hand-labeled)。

[0450] 可选地，一阶公式包括布尔公式和加权公式。布尔公式的权重为 + ∞，加权公式的权重为加权公式权重。那么，405 中人工标注的隐含谓词的值满足所述布尔公式。相应地，406 中，通过训练确定所述一阶公式的权重，即通过训练确定所述加权公式权重。其中，无向图可以包括马尔科夫网络 (Markov Network，MN)。

[0451] 可选地，406 中，可以根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，采用差额注入松弛算法 (Margin Infused Relaxed Algorithm，MIRA)，确定所述一阶公式的权重。

[0452] 具体地，在 406 中，可以使用 thebeast 工具学习加权公式权重。在参数学习的过程中，可以先将加权公式权重初始化为 0，再使用 MIRA 更新所述加权公式权重。可选地，在训练的过程中，还可以设置训练的最大循环次数，例如训练的最大循环次数为 10。

[0453] 举例来说，表五中的 sf3 的加权公式权重可如表六所示。从表六可以看出，候选短语的主要词的词性为 nn 时，该候选短语映射到类型为 E 的资源项的可能性比较大。

[0454] 表六

[0455]

候选短语的主要词的词性	候选短语映射到的资源项的类型	加权公式权重
nn	E	2.11
nn	C	0.243

[0456]

nn	R	0.335
vb	R	0.517
wp	C	0.143
wr	C	0.025

[0457] 这样，通过图 5 所示的实施例，可以确定任何一个知识库的加权公式权重，从而可以得到针对任何一个知识库的转换规则。

[0458] 可理解，本发明实施例中，确定一阶公式的权重的方法是一种数据驱动的方式，可以适用于不同的知识库。在大大减少人力的情况下，可以提高知识库的问答解析的效率。

[0459] 应理解，本发明实施例中，也可以根据所构建的无向图，进行结构学习，进而学习到二阶公式甚至更高阶的公式，进一步根据所学习到的二阶公式或更高阶的公式构建新的无向图，并学习二阶公式或更高阶的公式所对应的权重。本发明对此不作限定。

[0460] 图 6 是本发明一个实施例的问句解析的设备的框图。图 6 所示的设备 500 包括：接收单元 501、短语检测单元 502、映射单元 503、第一确定单元 504、第二确定单元 505、获取单元 506、和生成单元 507。

[0461] 接收单元 501，用于接收用户输入的问句。

[0462] 短语检测单元 502，用于对所述接收单元 501 接收的所述问句进行短语检测，以确定第一候选短语。

[0463] 映射单元503，用于将所述短语检测单元502确定的所述第一候选短语映射到知识库中的第一资源项，其中，所述第一资源项与所述第一候选短语具有一致的语义。

[0464] 第一确定单元504，用于根据所述第一候选短语和所述第一资源项，确定观察谓词的值和可能的问句分析空间，其中，所述观察谓词用于表示所述第一候选短语的特征、所述第一资源项的特征和所述第一候选短语与所述第一资源项的关系，所述可能的问句分析空间中的点为命题集合，所述命题集合中的命题的真假由隐含谓词的值表征。

[0465] 第二确定单元505，用于对所述可能的问句分析空间中的每一个命题集合，根据第一确定单元504确定所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度。

[0466] 获取单元506，用于获取所述置信度满足预设条件的命题集合中的真命题的组合，其中，所述真命题用于表示从所述第一候选短语中所选中的搜索短语、从所述第一资源项中所选中的搜索资源项和所述搜索资源项的特征。

[0467] 生成单元507，用于根据所述获取单元506获取的所述真命题的组合，生成形式化查询语句。

[0468] 本发明实施例利用观察谓词和隐含谓词，进行不确定性推理，能够将自然语言问句转化为形式化查询语句。并且，本发明实施例中，不确定性推理的方法能够应用于任何领域的知识库，具有领域扩展性，这样无需针对知识库人工地配置转换规则。

[0469] 可选地，作为一个实施例，所述不确定性推理基于马尔科夫逻辑网络MLN，所述MLN包括预定义的一阶公式以及所述一阶公式的权重。

[0470] 可选地，作为另一个实施例，

[0471] 所述获取单元506，还用于从所述知识库中获取多个自然语言问句；

[0472] 所述短语检测单元502，还用于对所述获取单元506接收的所述问句进行短语检测，以确定第一候选短语；

[0473] 所述映射单元503，还用于将所述第二候选短语映射到所述知识库中的第二资源项，其中，所述第二资源项与所述第二候选短语具有一致的语义；

[0474] 所述第一确定单元504，还用于根据所述第二候选短语和所述第二资源项，确定与所述多个自然语言问句对应的观察谓词的值；

[0475] 所述获取单元506，还用于获取人工标注的与所述多个自然语言问句对应的隐含谓词的值；

[0476] 所述第二确定单元505，还用于根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重。

[0477] 可选地，作为另一个实施例，所述一阶公式包括布尔公式和加权公式，所述布尔公式的权重为+∞，所述加权公式的权重为加权公式权重，所述人工标注的与所述多个自然语言问句对应的隐含谓词的值满足所述布尔公式，所述第二确定单元505，具体用于：根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述加权公式权重。

[0478] 可选地，作为另一个实施例，所述第二确定单元505，具体用于：根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述

一阶公式，构建无向图，采用差额注入松弛算法 MIRA，确定所述一阶公式的权重。

[0479] 可选地，作为另一个实施例，所述 MLN 表示为 M，所述一阶公式表示为 Φ_i，所述一阶公式的权重表示为 w_i，所述命题集合表示为 y，第二确定单元 505，具体用于：

[0480] 根据 $p(\mathbf{y})=\frac{1}{Z}\exp\left(\sum_{(\phi_i,w_i)\in M} w_i \sum_{\mathbf{c}\in C^{n\phi_i}} f_{\mathbf{c}}^{\phi_i}(\mathbf{y})\right)$，计算所述每一个命题集合的置信度，其中，Z 为归一化常数，$C^{n\phi_i}$ 为与一阶公式 Φ_i 对应的子公式的集合，c 为 $C^{n\phi_i}$ 的所述子公式的集合中的一个子公式，$f_{\mathbf{c}}^{\phi_i}$ 为二值函数，$f_{\mathbf{c}}^{\phi_i}(\mathbf{y})$ 表示在所述命题集合 y 下，所述一阶公式的真假。

[0481] 可选地，作为另一个实施例，获取单元 506，具体用于：确定所述置信度的值最大的命题集合，并获取所述置信度的值最大的命题集合中的真命题的组合。

[0482] 可选地，作为另一个实施例，

[0483] 所述第一候选短语的特征包括所述第一候选短语在所述问句中的位置、所述第一候选短语的主要词的词性、所述第一候选短语两两之间的依存路径上的标签，

[0484] 所述第一资源项的特征包括所述第一资源项的类型、所述第一资源项两两之间的相关性值、所述第一资源项两两之间的参数匹配关系，

[0485] 所述第一候选短语与所述第一资源项的关系包括所述第一候选短语与所述第一资源项的先验匹配得分，

[0486] 所述第一确定单元 504，具体用于：

[0487] 确定所述第一候选短语在所述问句中的位置；

[0488] 采用 stanford 词性标注工具，确定所述第一候选短语的主要词的词性；

[0489] 采用 stanford 依存句法分析工具，确定所述第一候选短语两两之间的依存路径上的标签；

[0490] 从所述知识库中确定所述第一资源项的类型，其中，所述类型为实体或类别或关系；

[0491] 从所述知识库中确定所述第一资源项两两之间的参数匹配关系；

[0492] 将所述第一资源项两两之间的相似性系数，作为所述两个第一资源项两两之间的相关性值；

[0493] 计算所述第一候选短语与所述第一资源项之间的先验匹配得分，所述先验匹配得分用于表示所述第一候选短语映射到所述第一资源项的概率。

[0494] 可选地，作为另一个实施例，所述形式化查询语句为简单协议资源描述框架查询语句 SPARQL。

[0495] 可选地，作为另一个实施例，所述生成单元 507，具体用于：

[0496] 根据所述真命题的组合，利用 SPARQL 模板生成所述 SPARQL。

[0497] 可选地，作为另一个实施例，所述 SPARQL 模板包括 ASK WHERE 模板、SELECT COUNT(? url)WHERE 模板和 SELECT ? url WHERE 模板，

[0498] 所述生成单元 507，具体用于：

[0499] 当所述问句为 Yes/No 问题时，根据所述真命题的组合，使用所述 ASK WHERE 模板

生成所述 SPARQL ；

[0500] 当所述问句为 Normal 问题时，根据所述真命题的组合，使用所述 SELECT ? url WHERE 模板生成所述 SPARQL ；

[0501] 当所述问句为 Number 问题时，根据所述真命题的组合，使用所述 SELECT ? url WHERE 模板生成所述 SPARQL，或者，当使用所述 SELECT ? url WHERE 模板生成的 SPARQL 无法得到数值型答案时，使用所述 SELECT COUNT(? url)WHERE 模板生成所述 SPARQL。

[0502] 可选地，作为另一个实施例，所述短语检测单元 502，具体用于 ：

[0503] 将所述问句中的词序列作为所述第一候选短语，其中，所述词序列满足 ：

[0504] 所述词序列中所有连续的非停用词都以大写字母开头，或者，若所述词序列中所有连续的非停用词不都以大写字母开头，则所述词序列的长度小于四 ；

[0505] 所述词序列的主要词的词性为 jj 或 nn 或 rb 或 vb，其中，jj 为形容词，nn 为名词，rb 为副词，vb 为动词 ；

[0506] 所述词序列所包括的词不全为停用词。

[0507] 可选地，作为另一个实施例，设备 500 可以是知识库的服务器。

[0508] 设备 500 能够实现图 1 至图 5 的实施例中由设备实现的各个过程，为避免重复，这里不再赘述。

[0509] 图 7 是本发明另一个实施例的问句解析的设备的框图。图 7 所示的设备 600 包括 ：处理器 601、接收电路 602、发送电路 603 和存储器 604。

[0510] 接收电路 602，用于接收用户输入的问句。

[0511] 处理器 601，用于对所述接收电路 602 接收的所述问句进行短语检测，以确定第一候选短语。

[0512] 处理器 601，还用于将所述第一候选短语映射到知识库中的第一资源项，其中，所述第一资源项与所述第一候选短语具有一致的语义。

[0513] 处理器 601，还用于根据所述第一候选短语和所述第一资源项，确定观察谓词的值和可能的问句分析空间，其中，所述观察谓词用于表示所述第一候选短语的特征、所述第一资源项的特征和所述第一候选短语与所述第一资源项的关系，所述可能的问句分析空间中的点为命题集合，所述命题集合中的命题的真假由隐含谓词的值表征。

[0514] 处理器 601，还用于对所述可能的问句分析空间中的每一个命题集合，根据第一确定单元 504 确定所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度。

[0515] 接收电路 602，还用于获取所述置信度满足预设条件的命题集合中的真命题的组合，其中，所述真命题用于表示从所述第一候选短语中所选中的搜索短语、从所述第一资源项中所选中的搜索资源项和所述搜索资源项的特征。

[0516] 处理器 601，还用于根据所述真命题的组合，生成形式化查询语句。

[0517] 本发明实施例利用观察谓词和隐含谓词，进行不确定性推理，能够将自然语言问句转化为形式化查询语句。并且，本发明实施例中，不确定性推理的方法能够应用于任何领域的知识库，具有领域扩展性，这样无需针对知识库人工地配置转换规则。

[0518] 设备 600 中的各个组件通过总线系统 605 耦合在一起，其中总线系统 605 除包括数据总线之外，还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见，在图

7 中将各种总线都标为总线系统 605。

[0519] 上述本发明实施例揭示的方法可以应用于处理器 601 中，或者由处理器 601 实现。处理器 601 可能是一种集成电路芯片，具有信号的处理能力。在实现过程中，上述方法的各步骤可以通过处理器 1001 中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器 1001 可以是通用处理器、数字信号处理器 (Digital Signal Processor，DSP)、专用集成电路 (Application Specific Integrated Circuit，ASIC)、现成可编程门阵列 (Field Programmable Gate Array，FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。可以实现或者执行本发明实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本发明实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器 604，处理器 601 读取存储器 604 中的信息，结合其硬件完成上述方法的步骤。

[0520] 可以理解，本发明实施例中的存储器 604 可以是易失性存储器或非易失性存储器，或可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器 (Read-Only Memory，ROM)、可编程只读存储器 (Programmable ROM，PROM)、可擦除可编程只读存储器 (Erasable PROM，EPROM)、电可擦除可编程只读存储器 (Electrically EPROM，EEPROM) 或闪存。易失性存储器可以是随机存取存储器 (Random Access Memory，RAM)，其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的 RAM 可用，例如静态随机存取存储器 (Static RAM，SRAM)、动态随机存取存储器 (Dynamic RAM，DRAM)、同步动态随机存取存储器 (Synchronous DRAM，SDRAM)、双倍数据速率同步动态随机存取存储器 (Double Data Rate SDRAM，DDR SDRAM)、增强型同步动态随机存取存储器 (Enhanced SDRAM，ESDRAM)、同步连接动态随机存取存储器 (Synchlink DRAM，SLDRAM) 和直接内存总线随机存取存储器 (Direct Rambus RAM，DR RAM)。本文描述的系统和方法的存储器 604 旨在包括但不限于这些和任意其它适合类型的存储器。

[0521] 可以理解的是，本文描述的这些实施例可以用硬件、软件、固件、中间件、微码或其组合来实现。对于硬件实现，处理单元可以实现在一个或多个专用集成电路 (Application Specific Integrated Circuits，ASIC)、数字信号处理器 (Digital Signal Processing，DSP)、数字信号处理设备 (DSP Device，DSPD)、可编程逻辑设备 (Programmable Logic Device，PLD)、现场可编程门阵列 (Field-Programmable Gate Array，FPGA)、通用处理器、控制器、微控制器、微处理器、用于执行本申请所述功能的其它电子单元或其组合中。

[0522] 当在软件、固件、中间件或微码、程序代码或代码段中实现实施例时，它们可存储在例如存储部件的机器可读介质中。代码段可表示过程、函数、子程序、程序、例程、子例程、模块、软件分组、类、或指令、数据结构或程序语句的任意组合。代码段可通过传送和 / 或接收信息、数据、自变量、参数或存储器内容来稿合至另一代码段或硬件电路。可使用包括存储器共享、消息传递、令牌传递、网络传输等任意适合方式来传递、转发或发送信息、自变量、参数、数据等。

[0523] 对于软件实现，可通过执行本文所述功能的模块 (例如过程、函数等) 来实现本文所述的技术。软件代码可存储在存储器单元中并通过处理器执行。存储器单元可以在处理

器中或在处理器外部实现，在后一种情况下存储器单元可经由本领域已知的各种手段以通信方式耦合至处理器。

[0524] 可选地，作为一个实施例，所述不确定性推理基于马尔科夫逻辑网络 MLN，所述 MLN 包括预定义的一阶公式以及所述一阶公式的权重。

[0525] 本发明实施例中，存储器 604 可用于存储资源项、以及资源项的类型等。存储器 604 还可用于存储所述一阶公式。存储器 604 还可用于存储 SPARQL 模板。

[0526] 可选地，作为另一个实施例，

[0527] 所述接收电路 602，还用于从所述知识库中获取多个自然语言问句；

[0528] 所述处理器 601，还用于对所述问句进行短语检测，以确定第一候选短语；

[0529] 所述处理器 601，还用于将所述第二候选短语映射到所述知识库中的第二资源项，其中，所述第二资源项与所述第二候选短语具有一致的语义；

[0530] 所述处理器 601，还用于根据所述第二候选短语和所述第二资源项，确定与所述多个自然语言问句对应的观察谓词的值；

[0531] 所述接收电路 602，还用于获取人工标注的与所述多个自然语言问句对应的隐含谓词的值；

[0532] 所述处理器 601，还用于根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重。

[0533] 可选地，作为另一个实施例，所述一阶公式包括布尔公式和加权公式，所述布尔公式的权重为 + ∞，所述加权公式的权重为加权公式权重，所述人工标注的与所述多个自然语言问句对应的隐含谓词的值满足所述布尔公式，

[0534] 所述处理器 601，具体用于：根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述加权公式权重。

[0535] 可选地，作为另一个实施例，所述处理器 601，具体用于：

[0536] 根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，采用差额注入松弛算法 MIRA，确定所述一阶公式的权重。

[0537] 可选地，作为另一个实施例，所述 MLN 表示为 M，所述一阶公式表示为 Φ_i，所述一阶公式的权重表示为 w_i，所述命题集合表示为 y，处理器 601，具体用于：

[0538] 根据 $p(\mathbf{y})=\frac{1}{Z}\exp\left(\sum_{(\phi_i,w_i)\in M} w_i \sum_{\mathbf{c}\in C^{n\phi_i}} f_{\mathbf{c}}^{\phi_i}(\mathbf{y})\right)$，计算所述每一个命题集合的置信度，其中，Z 为归一化常数，$C^{n\phi_i}$ 为与一阶公式 Φ_i 对应的子公式的集合，c 为 $C^{n\phi_i}$ 的所述子公式的集合中的一个子公式，$f_{\mathbf{c}}^{\phi_i}$ 为二值函数，$f_{\mathbf{c}}^{\phi_i}(\mathbf{y})$ 表示在所述命题集合 y 下，所述一阶公式的真假。

[0539] 可选地，作为另一个实施例，接收电路 602，具体用于：确定所述置信度的值最大的命题集合，并获取所述置信度的值最大的命题集合中的真命题的组合。

[0540] 可选地，作为另一个实施例，

[0541] 所述第一候选短语的特征包括所述第一候选短语在所述问句中的位置、所述第一候选短语的主要词的词性、所述第一候选短语两两之间的依存路径上的标签，

[0542] 所述第一资源项的特征包括所述第一资源项的类型、所述第一资源项两两之间的相关性值、所述第一资源项两两之间的参数匹配关系，

[0543] 所述第一候选短语与所述第一资源项的关系包括所述第一候选短语与所述第一资源项的先验匹配得分，

[0544] 所述处理器 601，具体用于：

[0545] 确定所述第一候选短语在所述问句中的位置；

[0546] 采用 stanford 词性标注工具，确定所述第一候选短语的主要词的词性；

[0547] 采用 stanford 依存句法分析工具，确定所述第一候选短语两两之间的依存路径上的标签；

[0548] 从所述知识库中确定所述第一资源项的类型，其中，所述类型为实体或类别或关系；

[0549] 从所述知识库中确定所述第一资源项两两之间的参数匹配关系；

[0550] 将所述第一资源项两两之间的相似性系数，作为所述两个第一资源项两两之间的相关性值；

[0551] 计算所述第一候选短语与所述第一资源项之间的先验匹配得分，所述先验匹配得分用于表示所述第一候选短语映射到所述第一资源项的概率。

[0552] 可选地，作为另一个实施例，所述形式化查询语句为简单协议资源描述框架查询语句 SPARQL。

[0553] 可选地，作为另一个实施例，所述处理器 601，具体用于：

[0554] 根据所述真命题的组合，利用 SPARQL 模板生成所述 SPARQL。

[0555] 可选地，作为另一个实施例，所述 SPARQL 模板包括 ASK WHERE 模板、SELECT COUNT(？ url)WHERE 模板和 SELECT？ url WHERE 模板，

[0556] 所述处理器 601，具体用于：

[0557] 当所述问句为 Yes/No 问题时，根据所述真命题的组合，使用所述 ASK WHERE 模板生成所述 SPARQL；

[0558] 当所述问句为 Normal 问题时，根据所述真命题的组合，使用所述 SELECT？ url WHERE 模板生成所述 SPARQL；

[0559] 当所述问句为 Number 问题时，根据所述真命题的组合，使用所述 SELECT？ url WHERE 模板生成所述 SPARQL，或者，当使用所述 SELECT？ url WHERE 模板生成的 SPARQL 无法得到数值型答案时，使用所述 SELECT COUNT(？ url)WHERE 模板生成所述 SPARQL。

[0560] 可选地，作为另一个实施例，所述处理器 601，具体用于：

[0561] 将所述问句中的词序列作为所述第一候选短语，其中，所述词序列满足：

[0562] 所述词序列中所有连续的非停用词都以大写字母开头，或者，若所述词序列中所有连续的非停用词不都以大写字母开头，则所述词序列的长度小于四；

[0563] 所述词序列的主要词的词性为 jj 或 nn 或 rb 或 vb，其中，jj 为形容词，nn 为名词，rb 为副词，vb 为动词；

[0564] 所述词序列所包括的词不全为停用词。

[0565] 可选地，作为另一个实施例，设备 600 可以是知识库的服务器。

[0566] 设备 600 能够实现图 1 至图 5 的实施例中由设备实现的各个过程，为避免重复，这里不再赘述。

[0567] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

[0568] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0569] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0570] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0571] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

[0572] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括 :U 盘、移动硬盘、只读存储器 (Read-Only Memory，ROM)、随机存取存储器 (Random Access Memory，RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0573] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

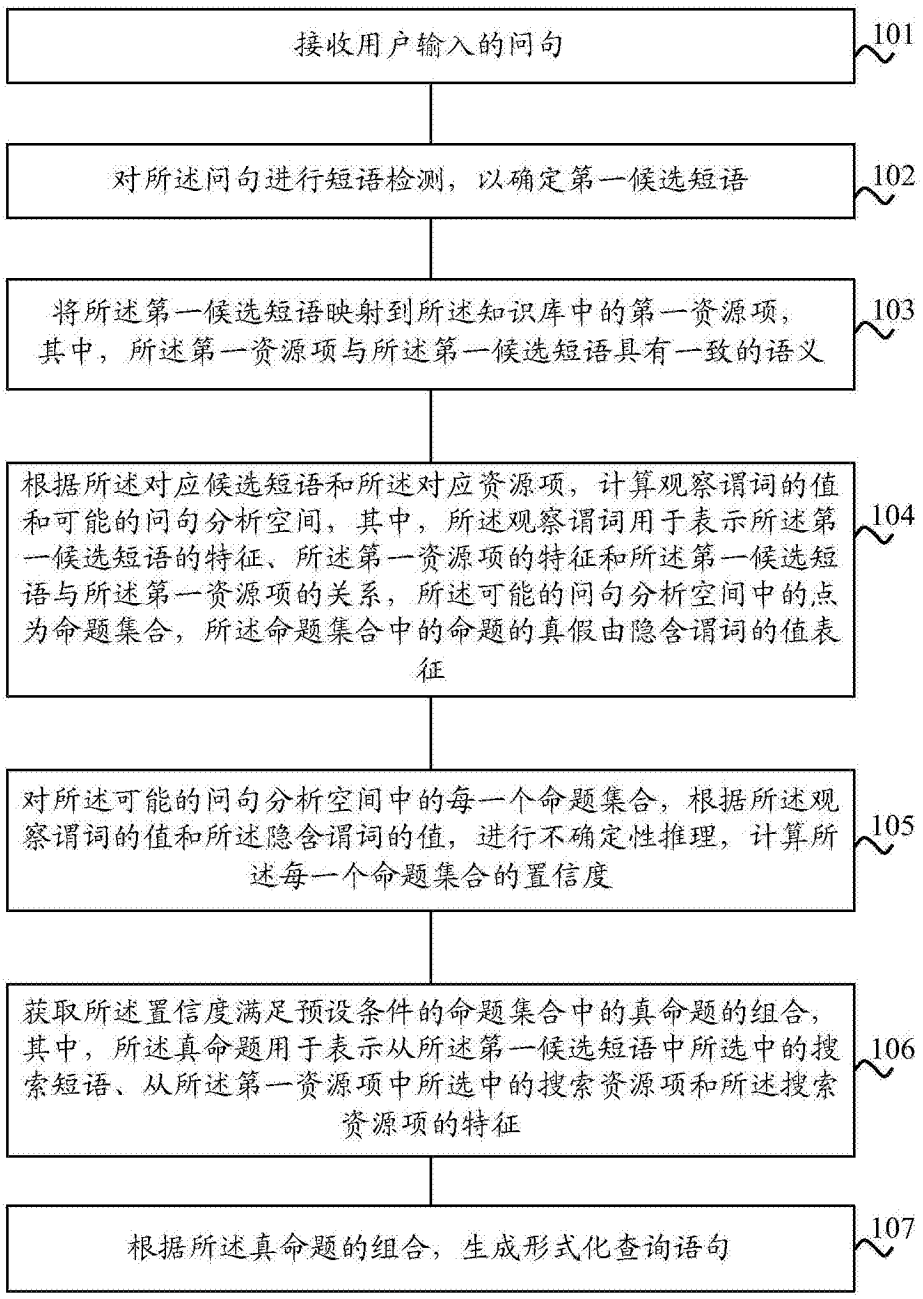

接收用户输入的问句
101
对所述问句进行短语检测，以确定第一候选短语
102
将所述第一候选短语映射到所述知识库中的第一资源项，其中，所述第一资源项与所述第一候选短语具有一致的语义
103
根据所述对应候选短语和所述对应资源项，计算观察谓词的值和可能的问句分析空间，其中，所述观察谓词用于表示所述第一候选短语的特征、所述第一资源项的特征和所述第一候选短语与所述第一资源项的关系，所述可能的问句分析空间中的点为命题集合，所述命题集合中的命题的真假由隐含谓词的值表征
104
对所述可能的问句分析空间中的每一个命题集合，根据所述观察谓词的值和所述隐含谓词的值，进行不确定性推理，计算所述每一个命题集合的置信度
105
获取所述置信度满足预设条件的命题集合中的真命题的组合，其中，所述真命题用于表示从所述第一候选短语中所选中的搜索短语、从所述第一资源项中所选中的搜索资源项和所述搜索资源项的特征
106
根据所述真命题的组合，生成形式化查询语句
107

图 1

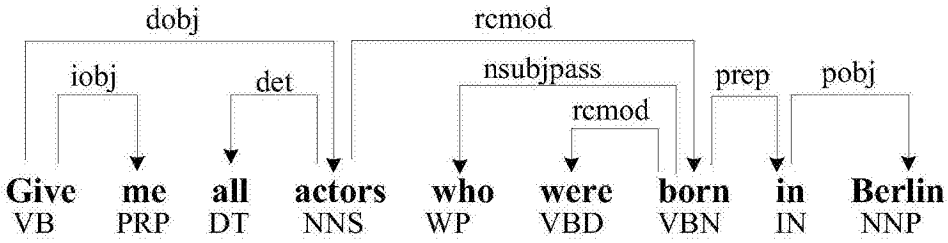

dobj
rcmod
iobj
det
nsubjpass
prep
pobj
rcmod
Give VB
me PRP
all DT
actors NNS
who WP
were VBD
born VBN
in IN
Berlin NNP

图 2

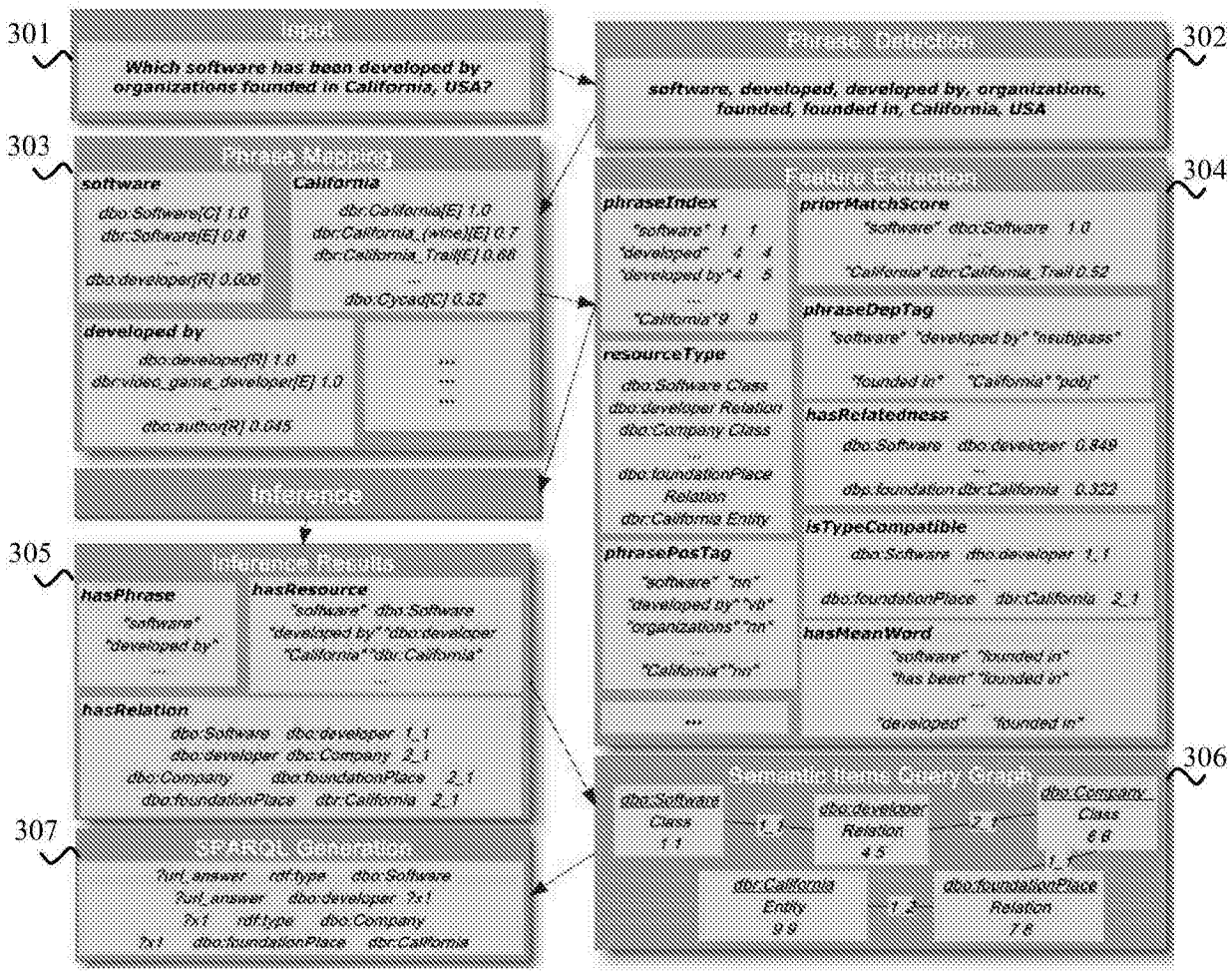

301
Which software has been developed by organizations founded in California, USA?
302
software, developed, developed by, organizations, founded, founded in, California, USA
303
304
305
306
307
Inference
Inference Results
SPARQL Generation
Semantic Items Query Graph

图 3

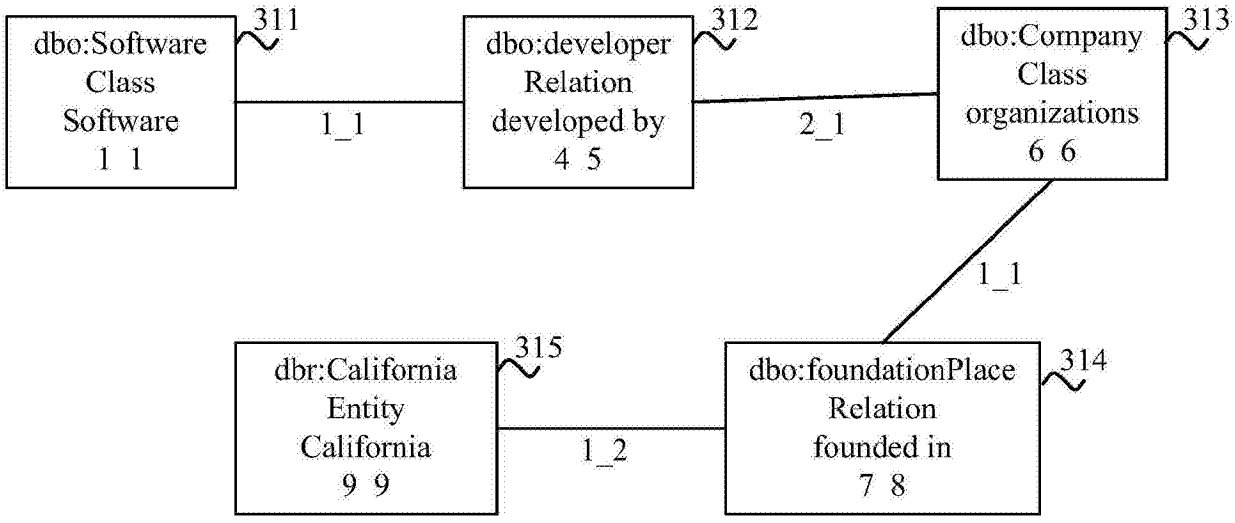

dbo:Software Class Software 1 1
311
1_1
dbo:developer Relation developed by 4 5
312
2_1
dbo:Company Class organizations 6 6
313
1_1
dbr:California Entity California 9 9
315
1_2
dbo:foundationPlace Relation founded in 7 8
314

图 4

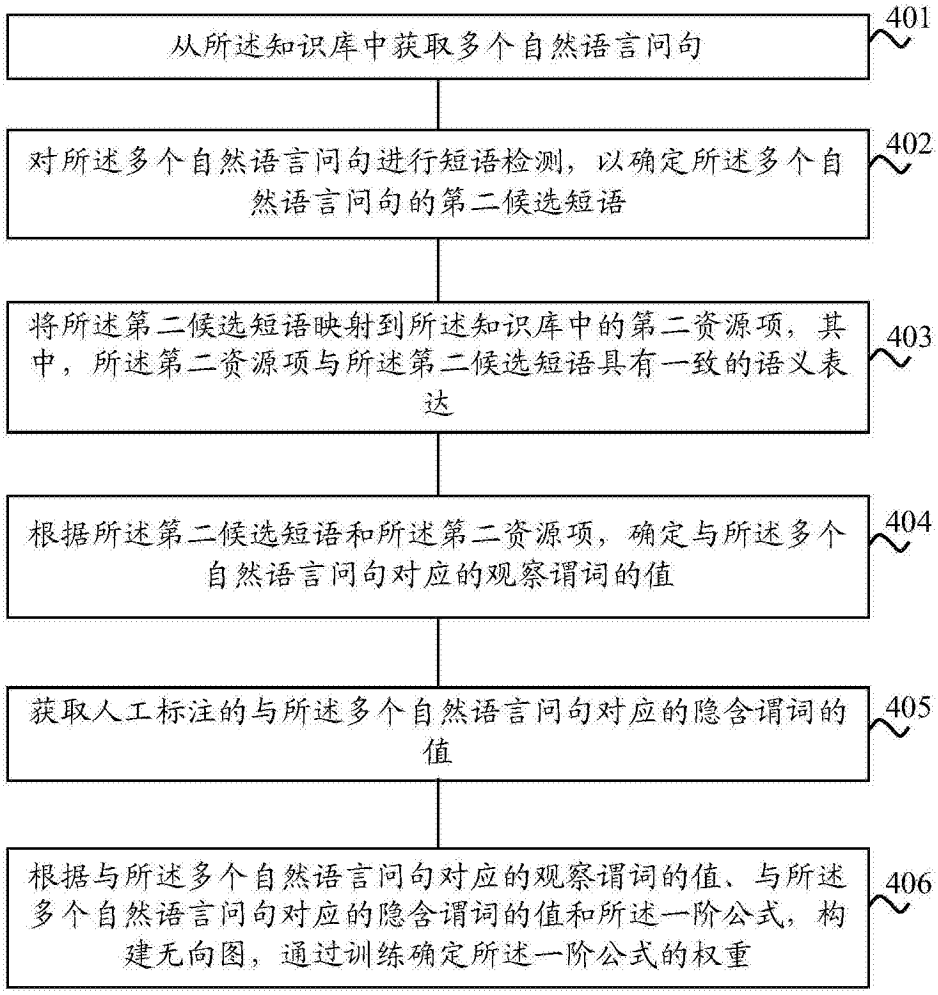

从所述知识库中获取多个自然语言问句
401
对所述多个自然语言问句进行短语检测，以确定所述多个自然语言问句的第二候选短语
402
将所述第二候选短语映射到所述知识库中的第二资源项，其中，所述第二资源项与所述第二候选短语具有一致的语义表达
403
根据所述第二候选短语和所述第二资源项，确定与所述多个自然语言问句对应的观察谓词的值
404
获取人工标注的与所述多个自然语言问句对应的隐含谓词的值
405
根据与所述多个自然语言问句对应的观察谓词的值、与所述多个自然语言问句对应的隐含谓词的值和所述一阶公式，构建无向图，通过训练确定所述一阶公式的权重
406

图 5

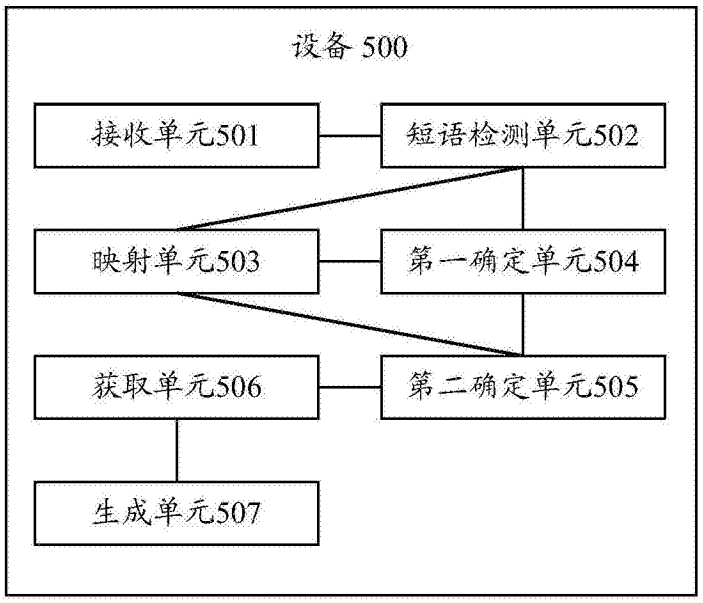

设备 500
接收单元501
短语检测单元502
映射单元503
第一确定单元504
获取单元506
第二确定单元505
生成单元507

图 6

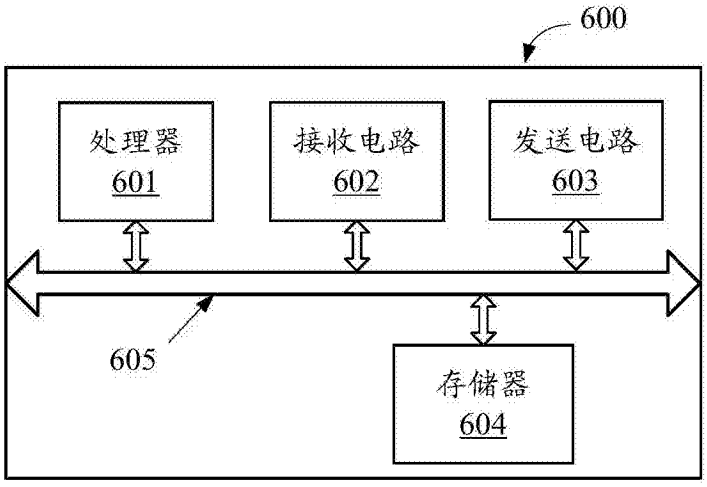

600
处理器
601
接收电路
602
发送电路
603
605
存储器
604

图 7