(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116484870 B
(45)授权公告日 2024.01.05

(21)申请号 202211101583.8
(22)申请日 2022.09.09
(65)同一申请的已公布的文献号
申请公布号 CN 116484870 A
(43)申请公布日 2023.07.25
(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层
(72)发明人 杨静怡 孙明明 李平
(74)专利代理机构 北京钲霖知识产权代理有限公司 11722
专利代理师 李英艳 杨继成
(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/284(2020.01)
G06N 5/02(2023.01)

(56)对比文件
CN 112507040 A,2021.03.16
CN 112269884 A,2021.01.26
CN 103577398 A,2014.02.12
CN 107783960 A,2018.03.09
CN 112148871 A,2020.12.29
CN 111160030 A,2020.05.15
CN 111460787 A,2020.07.28
CN 112183059 A,2021.01.05
CN 112835927 A,2021.05.25
CN 114266258 A,2022.04.01
CN 114330293 A,2022.04.12
CN 114756691 A,2022.07.15
CN 114970543 A,2022.08.30
GB 201419051 D0,2014.12.10
JP 2021125182 A,2021.08.30
US 10387575 B1,2019.08.20
US 2010228693 A1,2010.09.09
US 2016267117 A1,2016.09.15
US 2019236469 A1,2019.08.01
US 2021232770 A1,2021.07.29
US 2011106843 A1,2011.05.05
US 2017147646 A1,2017.05.25
李英.越南语短语树到依存树的转换研究.计算机科学与探索.2016,第11卷(第04期),599-607.
吴泰中.基于转移神经网络的中文AMR解析.中文信息学报.2019,第33卷(第04期),1-11.

审查员 张子瑜

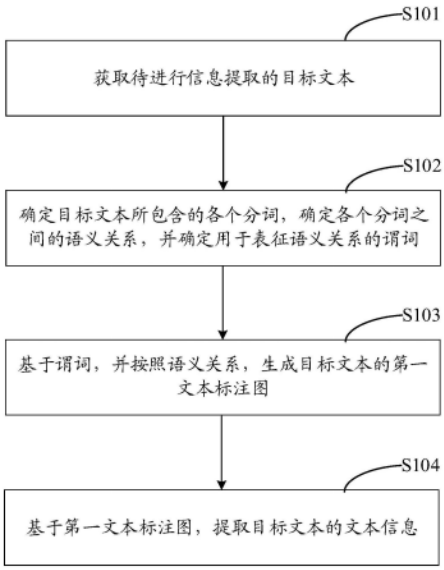
权利要求书4页 说明书12页 附图12页

(54)发明名称

提取文本信息的方法、装置、设备及介质

(57)摘要

本公开提供了一种提取文本信息的方法、装置、设备、介质及计算机产品,涉及计算机技术领域,尤其涉及人工智能领域的知识图谱、自然语言处理技术。具体实现方案为:获取待进行信息提取的目标文本;确定所述目标文本所包含的各个分词,确定各个所述分词之间的语义关系,并确定用于表征所述语义关系的谓词;基于所述谓词,并按照所述语义关系,生成所述目标文本的第一文本标注图;基于所述第一文本标注图,提取所述目标文本的文本信息。



CN 116484870 B

1.一种提取文本信息的方法,包括:

获取待进行信息提取的目标文本;

确定所述目标文本所包含的各个分词,确定各个所述分词之间的语义关系,并确定用于表征所述语义关系的谓词;

基于所述谓词,并按照所述语义关系,生成所述目标文本的第一文本标注图;

基于所述第一文本标注图,提取所述目标文本的文本信息;

其中,所述基于所述谓词,并按照所述语义关系,生成所述目标文本的第一文本标注图,包括:

基于所述谓词,确定多个层级,所述多个层级中包括最高层级,以及不同于所述最高层级的其他层级;

所述最高层级中包括单一的第一节点,所述第一节点用于标识所述各个分词;

基于所述谓词以及所述语义关系,确定所述其他层级中各层级包括第二节点以及第三节点;

所述第二节点用于标识所述各个分词中的一个或多个分词,所述第三节点用于标识单一的所述谓词,且所述第三节点所标识的谓词用于表征所述其他层级中同一层级内各个第二节点所标识的分词之间的语义关系;

基于所述第一节点、所述第二节点以及所述第三节点,生成所述目标文本对应所述多个层级的第一文本标注图。

2.根据权利要求1所述的方法,其中,所述基于所述谓词,确定多个层级,包括:

确定所述谓词的数量,并确定所述数量个所述谓词之间的主次关系;

按照所述主次关系,确定具有层级关系的所述数量个其他层级;

所述层级关系用于表征相邻两层级中的较高层级和较低层级;

其中,所述较高层级对应主要关系谓词,所述较低层级对应次要关系谓词;

将所述数量个其他层级中层级关系最高的层级,作为与所述最高层级相邻的较低层级,得到所述多个层级。

3.根据权利要求1或2所述的方法,其中,所述基于所述第一节点、所述第二节点以及所述第三节点,生成所述目标文本的第一文本标注图,包括:

针对所述多个层级中任意两个相邻层级,分别确定所述相邻层级中较高层级中存在的目标节点,所述目标节点包含有所述相邻层级中较低层级中全部所述第二节点所包含的各个分词;

针对所述多个层级中任意两个相邻层级,将所述相邻层级中较高层级中所包括的目标节点与较低层级中所包括的第二节点以及第三节点分别通过边连接,生成所述目标文本的第一文本标注图。

4.根据权利要求1所述的方法,其中,所述基于所述第一节点、所述第二节点以及所述第三节点,生成所述目标文本对应所述多个层级的第一文本标注图,包括:

生成所述目标文本的第二文本标注图;

其中,所述第二文本标注图中包括节点和边,所述目标文本中各个分词与所述第二文本标注图中各个节点之间一一对应,每一所述边连接有具有父子关系的两个节点,用于标识所述两个节点所对应的分词之间的依存关系;

将所述第二文本标注图转换为包括所述第一节点、所述第二节点以及所述第三节点的文本标注图,得到所述第一文本标注图。

5.根据权利要求4所述的方法,其中,所述将所述第二文本标注图转换为包括所述第一节点、所述第二节点以及所述第三节点的文本标注图,包括:

确定所述第二文本标注图中标识词性为指定谓词的指定边;

在所述指定边连接的两节点之间插入第一目标节点,所述第一目标节点用于标识所述指定谓词;

将所述第一目标节点转换为所述指定边连接的两节点的公共父节点,得到第一转换后的第二文本标注图;

在所述第一转换后的第二文本标注图中,确定与所述第一目标节点之间存在公共子节点的第二目标节点;

将所述第一目标节点转换为所述第二目标节点的子节点,并将所述第一目标节点转换为所述公共子节点的父节点,并将所述第二目标节点转换为所述公共子节点的祖父节点,得到第二转换后的第二文本标注图;

针对所述第二转换后的第二文本标注图,对每一非叶节点所包含的分词进行补充,以使每一所述非叶节点所包含的分词为所述非叶节点的各个子节点所包含分词的并集;

将补充分词后的第二文本标注图,作为所述第一文本标注图。

6.根据权利要求5所述的方法,其中,所述指定谓词包括以下之一或组合:

修饰关系谓词、并联关系谓词、同位语关系谓词以及丢失谓词。

7.一种提取文本信息的装置,包括:

获取模块,用于获取待进行信息提取的目标文本;

确定模块,用于确定所述目标文本所包含的各个分词,确定各个所述分词之间的语义关系,并确定用于表征所述语义关系的谓词;

生成模块,用于基于所述谓词,并按照所述语义关系,生成所述目标文本的第一文本标注图;

处理模块,用于基于所述第一文本标注图,提取所述目标文本的文本信息;

其中,所述生成模块采用如下方式基于所述谓词,并按照所述语义关系,生成所述目标文本的第一文本标注图:

基于所述谓词,确定多个层级,所述多个层级中包括最高层级,以及不同于所述最高层级的其他层级;

所述最高层级中包括单一的第一节点,所述第一节点用于标识所述各个分词;

基于所述谓词以及所述语义关系,确定所述其他层级中各层级包括第二节点以及第三节点;

所述第二节点用于标识所述各个分词中的一个或多个分词,所述第三节点用于标识单一的所述谓词,且所述第三节点所标识的谓词用于表征所述其他层级中同一层级内各个第二节点所标识的分词之间的语义关系;

基于所述第一节点、所述第二节点以及所述第三节点,生成所述目标文本对应所述多个层级的第一文本标注图。

8.根据权利要求7所述的装置,其中,所述生成模块采用如下方式基于所述谓词,确定

多个层级：

确定所述谓词的数量，并确定所述数量个所述谓词之间的主次关系；

按照所述主次关系，确定具有层级关系的所述数量个其他层级；

所述层级关系用于表征相邻两层级中的较高层级和较低层级；

其中，所述较高层级对应主要关系谓词，所述较低层级对应次要关系谓词；

将所述数量个其他层级中层级关系最高的层级，作为与所述最高层级相邻的较低层级，得到所述多个层级。

9.根据权利要求7或8所述的装置，其中，所述生成模块采用如下方式基于所述第一节点、所述第二节点以及所述第三节点，生成所述目标文本的第一文本标注图：

针对所述多个层级中任意两个相邻层级，分别确定所述相邻层级中较高层级中存在的目标节点，所述目标节点包含有所述相邻层级中较低层级中全部所述第二节点所包含的各个分词；

针对所述多个层级中任意两个相邻层级，将所述相邻层级中较高层级中所包括的目标节点与较低层级中所包括的第二节点以及第三节点分别通过边连接，生成所述目标文本的第一文本标注图。

10.根据权利要求7所述的装置，其中，所述生成模块采用如下方式基于所述第一节点、所述第二节点以及所述第三节点，生成所述目标文本对应所述多个层级的第一文本标注图：

生成所述目标文本的第二文本标注图；

其中，所述第二文本标注图中包括节点和边，所述目标文本中各个分词与所述第二文本标注图中各个节点之间一一对应，每一所述边连接有具有父子关系的两个节点，用于标识所述两个节点所对应的分词之间的依存关系；

将所述第二文本标注图转换为包括所述第一节点、所述第二节点以及所述第三节点的文本标注图，得到所述第一文本标注图。

11.根据权利要求10所述的装置，其中，所述生成模块采用如下方式将所述第二文本标注图转换为包括所述第一节点、所述第二节点以及所述第三节点的文本标注图：

确定所述第二文本标注图中标识词性为指定谓词的指定边；

在所述指定边连接的两节点之间插入第一目标节点，所述第一目标节点用于标识所述指定谓词；

将所述第一目标节点转换为所述指定边连接的两节点的公共父节点，得到第一转换后的第二文本标注图；

在所述第一转换后的第二文本标注图中，确定与所述第一目标节点之间存在公共子节点的第二目标节点；

将所述第一目标节点转换为所述第二目标节点的子节点，并将所述第一目标节点转换为所述公共子节点的父节点，并将所述第二目标节点转换为所述公共子节点的祖父节点，得到第二转换后的第二文本标注图；

针对所述第二转换后的第二文本标注图，对每一非叶节点所包含的分词进行补充，以使每一所述非叶节点所包含的分词为所述非叶节点的各个子节点所包含分词的并集；

将补充分词后的第二文本标注图，作为所述第一文本标注图。

12.根据权利要求11所述的装置,其中,所述指定谓词包括以下之一或组合:

修饰关系谓词、并联关系谓词、同位语关系谓词以及丢失谓词。

13.一种电子设备,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-6中任一项所述的方法。

14.一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使所述计算机执行根据权利要求1-6中任一项所述的方法。

提取文本信息的方法、装置、设备及介质

技术领域

[0001] 本公开涉及计算机技术领域,尤其涉及人工智能领域的知识图谱、自然语言处理技术。

背景技术

[0002] 开放信息提取(Open Information Extraction,OIE)是知识计算的重要基础构件,其通过在开放的自由文本中提取事实,进而将所提取的事实应用于文本信息领域的诸多场景。

[0003] 相关技术中,能够将开放信息提取应用于文本标注图的构建。例如,在构建过程中,将文本拆分为多个分词,进而按照文本中各分词之间的依存关系,构建具有结构化信息的开放信息标注(Open Information Annotation,OIA)图。

发明内容

[0004] 本公开提供了一种提取文本信息的方法、装置、设备、介质及计算机产品。

[0005] 根据本公开的一方面,提供了一种提取文本信息的方法。

[0006] 获取待进行信息提取的目标文本;确定所述目标文本所包含的各个分词,确定各个所述分词之间的语义关系,并确定用于表征所述语义关系的谓词;基于所述谓词,并按照所述语义关系,生成所述目标文本的第一文本标注图;基于所述第一文本标注图,提取所述目标文本的文本信息。

[0007] 根据本公开的另一方面,提供了一种提取文本信息的装置,包括:

[0008] 获取模块,用于获取待进行信息提取的目标文本;确定模块,用于确定所述目标文本所包含的各个分词,确定各个所述分词之间的语义关系,并确定用于表征所述语义关系的谓词;生成模块,用于基于所述谓词,并按照所述语义关系,生成所述目标文本的第一文本标注图;处理模块,用于基于所述第一文本标注图,提取所述目标文本的文本信息。

[0009] 根据本公开的另一方面,提供了一种电子设备,包括:

[0010] 至少一个处理器;以及

[0011] 与所述至少一个处理器通信连接的存储器;其中,

[0012] 所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行上述涉及的方法。

[0013] 根据本公开的另一方面,提供了一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使所述计算机执行上述涉及的方法。

[0014] 根据本公开的另一方面,提供了一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时实现上述涉及的方法。

[0015] 应当理解,本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征,也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0016] 附图用于更好地理解本方案,不构成对本公开的限定。其中:

[0017] 图1是一种OIA图的结构示意图;

[0018] 图2是本公开示出的一种提取文本信息的方法流程图;

[0019] 图3是本公开示出的一种第一文本标注图的示意图;

[0020] 图4是本公开示出的一种基于谓词,并按照语义关系,生成目标文本的第一文本标注图的方法流程图;

[0021] 图5是本公开示出的一种基于谓词,确定多个层级的方法流程图;

[0022] 图6是本公开示出的一种通过方式一生成第一文本标注图的方法流程图;

[0023] 图7是本公开示出的一种通过方式二生成第一文本标注图的方法流程图;

[0024] 图8是本公开示出的一种通过规则引擎生成HOIA图的流程示意图;

[0025] 图9是本公开示出的一种将第二文本标注图转换为第一文本标注图的方法流程图;

[0026] 图10是本公开示出的一种第二文本标注图的示意图;

[0027] 图11是本公开示出的一种插入第一目标节点后的第二文本标注图的示意图;

[0028] 图12是本公开示出的一种第一转换后的第二文本标注图的示意图;

[0029] 图13是本公开示出的一种第二转换后的第二文本标注图的示意图;

[0030] 图14是根据本公开的提取文本信息的装置框图;

[0031] 图15示出了可以用来实施本公开的实施例的示例电子设备的示意性框图。

[0032] 实施方式

[0033] 以下结合附图对本公开的示范性实施例做出说明,其中包括本公开实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本公开的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0034] 本公开应用于提取文本信息的场景,例如可以是对文本信息进行开放域信息抽取、知识图谱构建、命名实体识别、共指消解、零指消解、开放域问答和/或信息检索的场景。其中,进行文本信息提取的一种方式是应用文本标注图进行提取。

[0035] 开放信息提取是知识计算的重要基础构件,其通过在开放的自由文本中提取事实,进而将所提取的事实应用于文本信息领域的诸多场景。

[0036] 相关技术中,能够将开放信息提取应用于文本标注图的构建。例如,在构建过程中,将文本拆分为多个分词,进而按照文本中各分词之间的依存关系,构建具有结构化信息的OIA图。其中,OIA图通常采用以下概念表达文本的语言信息。

[0037] 常量:表示实体,例如“太阳系”、“公司”等;或者,表示实体/事件/关系的状态,例如:“昂贵”、“难以”等。

[0038] 函数:f(arg1,·····)→{e},表示实体的查询或实体的代表。其中,arg1表示函数的参数,{e}是函数返回的一些未知实体集。例如,如下语言描述可以表示为函数:“X的总裁”,“当Y时”,其中X和Y表示函数的参数。函数在英语文本中较为常见,例如,what从句、where从句、of短语或者of从句等,均可以表示为函数。

[0039] 谓词:p(arg1,····,argn)→{0,1},表示实体、谓词之间的事实关系和逻辑联

系。其中arg1、argn表示谓词的参数,0、1表示谓词的真假状态,0表示谓词为假状态,1表示谓词为真状态。例如,如下语言描述可以表示为谓词:“X购买了Y”、“X说Y”、“Y,因为Z”。

[0040] 图1是一种OIA图的结构示意图。以图1为例,图1是文本“The Voice of Amierica presents differing points of views on a wide variety of issues”的OIA图,OIA图中包括节点(例如,0 | present |(4,)|事件)和边(例如,pred.arg.1)。其中,对于任一节点而言,由左至右对应的四项信息分别用于表征“节点标号”、“节点所标识的分词”、“节点所标识分词在完整文本中所处序列位置”以及“节点所标识分词的词性”。可见,对于OIA图而言,文本中各个分词与OIA图中各个节点之间一一对应,每一边连接有具有父子关系的两个节点,用于标识两个节点所对应的分词之间的依存关系。

[0041] 在此基础上,相关技术中能够通过基于文本构建的OIA图进行文本信息提取。然而,分析OIA图的结构可知,OIA图本身存在如下问题:

[0042] (1)缺少对谓词结构的清晰标注。当同一谓词对应有多个论元(argument)时,难以通过对OIA图的解析,得到文本内部结构的层次关系,该问题尤为突出的体现在用于表征语义角色(Semantic role)的内部结构。以文本“The Voice of Amierica presents differing points of views on a wide variety of issues”为例,文本中通常涉及有多个介词谓词(例如,该文本所包含的“on”及“of”等),而如图1所示,在基于该文本生成的OIA图中,多个谓词之间或以平行(或线性)方式关联,或未被从节点中提取以进行细化拆分。在此基础上,通过对OIA图的解读,无法清楚获知多个谓词之间的关联关系,因而也无法得到文本内部的层次关系。

[0043] (2)无法实现对复杂名词短语的清晰标注。例如,在标注文本的实际场景中,往往存在需要提取嵌套结构实体的情况,由于OIA图只能表征分词间的依存关系,而不能表示嵌套结构,因此无法清晰标注复杂名词短语的内部结构。

[0044] (3)无法解决复杂结构的歧义。由于OIA图不具有层次性,当文本中同时存在对同一名词结构进行修饰的多个修饰词时,多个修饰词会在OIA图中以平行或线性的方式出现在同一个层次,从而导致OIA图表达了一个存在歧义的文本信息。例如,针对文本“old man and woman with hats”,由于文本中的“old”和“hats”可以修饰“man”,也可以修饰“women”,或是同时修饰“man”和“women”,因此,该文本同时存在多种解读方式。若生成该文本的OIA图,则在后续分析OIA图时,无法确定OIA图所标识的文本释义具体为多种解读方式中的哪一种,因而导致文本的表达出现歧义。

[0045] 进一步的,基于OIA图存在的上述缺陷,相关技术中通过OIA图进行文本信息提取的方式,存在效率低及精度差的问题。

[0046] 鉴于此,本公开提出了一种提取文本信息的方法,该方法保留引用上述概念,并采用与相关技术不同的标注图生成方式,得到了一种能够对文本层次化信息进行有效标注的文本标注图,用以解决相关技术中存在的上述问题。具体的,可以确定文本中各分词间的语义关系,并以此确定用于表征分词间语义关系的谓词。进一步的,可以通过所得到的谓词及语义关系,生成具有层次化信息的文本标注图---层次化开放信息标注(Hierarxchical Open Information Annotation,HOIA)图。以下为便于理解,对生成HOIA图的方式进行示例性说明。其中,以下为便于描述,将目标文本的HOIA图称为第一文本标注图。

[0047] 图2是本公开示出的一种提取文本信息的方法流程图,如图2所示,包括以下步骤

S101至步骤S104。

[0048] 在步骤S101中,获取待进行信息提取的目标文本。

[0049] 在步骤S102中,确定目标文本所包含的各个分词,确定各个分词之间的语义关系,并确定用于表征语义关系的谓词。

[0050] 在步骤S103中,基于谓词,并按照语义关系,生成目标文本的第一文本标注图。

[0051] 在步骤S104中,基于第一文本标注图,提取目标文本的文本信息。

[0052] 本公开实施例中,目标文本的第一文本标注图是通过谓词,并按照各个分词之间的语义关系生成的。为便于理解,如下以图3为例,对第一文本标注图的结构进行示例性说明。

[0053] 图3是本公开示出的一种第一文本标注图的示意图。示例的,如图3所示,针对目标文本“Al-Zaman:American forces killed Shaikh Abdullah al-Ani,the preacher at the mosque in the town of Qaim,near the Syrian border.”,第一文本标注图包括包含最高层级在内的多个层级。其中,对于最高层级(示例的,最高层级为多个层级中的第一层级)包含用于标识各个分词的第一节点(示例的,第一节点为Al-Zaman:American forces killed Shaikh Abdullah al-Ani,the preacher at the mosque in the town of Qaim,near the Syrian border.|((0,28),)|实体)。

[0054] 在此基础上,对于同一层级内的多个节点分析可知,除最高层级外的其他层级分别包含多个第二节点和单一的第三节点,第二节点用于标识一个或多个分词,第三节点用于标识一个谓词,且第三节点所标识的谓词用于表征同一层级内各个第二节点所标识分词之间的语义关系。如图3所示,以第二层级为例,标号为“3”、“7”及“11”的节点即为第二层级的第二节点,标号为“0”的节点即为第二层级的第三节点。其中,第三节点所包含的分词“:”,即用于表征分词“Al-Zaman”、分词“.”以及分词“American forces killed Shaikh Abdullah al-Ani,the preacher at the mosque in the town of Qaim,near the Syrian border”之间的语义关系。可见,针对第一文本标注图,同一层级内不同节点之间的关联关系已被清晰标注。

[0055] 进一步的,对于不同层级间的关联关系分析可知,对于第一文本标注图而言,相邻层级间的节点通过特定方式相连接。具体的,针对多个层级中任意两个相邻层级,相邻层级的高层级中存在目标节点,目标节点包含有相邻层级的低层级中各个第二节点所包含的各个分词,且目标节点通过不同边分别与低层级中的各个节点相连接。以图3为例,在彼此相邻的第三层级与第四层级中,第三层级为高层级,第四层级为低层级。针对第四层级中标号为“2”、“5”和“6”的第二节点,第三层级中存在标号为“12”的目标节点,满足使目标节点标识有第四层级中各个第二节点所标识的分词。可见,针对第一文本标注图的相邻层级,可根据高层级中目标节点及低层级中第二节点分别标识的各个分词,得到相邻层级间的关联关系,以使最终生成的第一文本标注图具有清晰的层次化结构。

[0056] 此外,可以理解的是,对于第一文本标注图而言,除最高层级外的每一层级,分别基于目标文本中的一个谓词构建。换言之,第一文本标注图中,除最高层级外的其他层级的数量与谓词的数量相一致。

[0057] 综上可知,通过本公开实施例提供的方法生成的HOIA图,同一层级内的不同节点之间具有清晰的语义关联关系,进而在通过HOIA图进行信息提取时,可直接得到“实体(例

如某人物)+场景事件(例如某动作)+实体概念(例如人物类概念)+时间”的信息组合。可见,该方法可实现更加直接全面的信息提取。

[0058] 并且,随着层级的加深,文本分析依次递进,后续解读HOIA图时,可以清晰地确定出HIOA图中相邻层级间的层次化逻辑,进而改善通过标注图推导出存在多种歧义文本的问题。以文本“old man and woman with hats”为例,若确定“old”仅用于修饰“man”,“with hats”仅用于修饰“woman”,则针对该文本的HOIA图,第二层级配置有第二节点“old man”、第二节点“woman with hats”以及第三节点“and”,且第三层级可对第二节点“woman with hats”做进一步拆分。在此基础上,在通过该文本的就HOIA图进行信息提取时,不会就文本解读出多种含义,该方法可以减小提取到歧义信息的可能性。

[0059] 此外,可以理解的是,上述以英文的目标文本生成第一文本标注图的流程仅是本公开一示例性实施方式,本公开对目标文本所采用的语种并不限制。

[0060] 本公开实施例中,第一文本标注图包括多个层级,且各层级分别配置有一个或多个节点。示例的,在确定各个分词之间的语义关系,以及用于表征语义关系的谓词的情况下,可以根据谓词确定多个层级,并结合谓词以及语义关系,确定各层级所包含的节点。进一步的,可以通过确定出的多个层级以及各个节点,生成目标文本的第一标注图。

[0061] 图4是本公开示出的一种基于谓词,并按照语义关系,生成目标文本的第一文本标注图的方法流程图,如图4所示,包括以下步骤S401至步骤S403。

[0062] 在步骤S401中,基于谓词,确定多个层级。

[0063] 本公开实施例中,多个层级中包括最高层级,以及不同于最高层级的其他层级。其中,最高层级中包括单一的第一节点,第一节点用于标识目标文本中的各个分词。此外,其他层级中各层级分别包括多个节点,且所包括的节点通过如下步骤S402确定。

[0064] 在步骤S402中,基于谓词以及语义关系,确定其他层级中各层级包括第二节点以及第三节点。

[0065] 其中,第二节点用于标识各个分词中的一个或多个分词,第三节点用于标识单一的谓词。并且,同一层级中各第二节点所标识的各个分词之间的语义关系,通过该层级中第三节点所标识的谓词表征。示例的,如图3所示,第三层级中包含的第三节点用于标识谓词“killed”,第三层级中包含的第二节点分别用于标识分词“American forces”以及分词“Shaikh Abdullah al-Ani,the preacher at the mosque in the town of Qaim,near the Syrian border”,且谓词“killed”用于表征两分词之间的语义关系。

[0066] 在步骤S403中,基于第一节点、第二节点以及第三节点,生成目标文本对应多个层级的第一文本标注图。

[0067] 本公开实施例提供的方法,若仅存在一个谓词,则包括最高层级在内,第一文本标注图共有两个层级。若存在多个谓词,则除最高层级外的其他层级的数量与谓词数量相一致。

[0068] 进一步的,在构建第一文本标注图时,还需要确定多个其他层级之间的层级关系。本公开如下提供了一种确定多个其他层级之间的层级关系的可行方式。

[0069] 图5是本公开示出的一种基于谓词,确定多个层级的方法流程图,如图5所示,包括以下步骤S501至步骤S503。

[0070] 在步骤S501中,确定谓词的数量,并确定数量个谓词之间的主次关系。

[0071] 在步骤S502中,按照主次关系,确定具有层级关系的数量个其他层级。

[0072] 其中,层级关系用于表征相邻两层级中的较高层级和较低层级,并且对于相邻两层级而言,较高层级对应主要关系谓词,较低层级对应次要关系谓词。

[0073] 在步骤S503中,将数量个其他层级中层级关系最高的层级,作为与最高层级相邻的较低层级,得到多个层级。

[0074] 为便于理解,以下结合第一文本标注图,对按照主次关系确定具有层级关系的数量个其他层级进行解释说明。示例的,如图3所示,对于目标文本,首先需要通过谓词“:”对完整的目标文本做拆分,以得到标号为“3”、“7”及“11”的分词。在此基础上,对于标号为“11”的分词而言,可以通过分词中包含的谓词“killed”进行进一步拆分。在此基础上,与谓词“killed”相比,谓词“:”对应谓词主次关系中的主要关系,而谓词“killed”即对应谓词主次关系中的次要关系。

[0075] 本公开实施例中,可以针对包含多个谓词的目标文本,按照多个谓词之间的主次关系,确定多个层级之间的层级关系,通过该方法生成的第一文本标注图,各层级之间的逻辑关系较为清晰,便于进行文本信息提取。

[0076] 示例的,在确定多个层级以及各层级所包括的节点的情况下,可以通过如下两种方式生成目标文本的第一文本标注图。

[0077] 方式一:在确定第一节点、第二节点以及第三节点的情况下,按照特定方式将各节点通过边相连接,以生成目标文本的第一文本标注图。

[0078] 方式二:在确定第一节点、第二节点以及第三节点的情况下,参照第一节点、第二节点以及第三节点,将目标文本的OIA图(以下为便于描述,将目标文本的OIA图称为第二文本标注图)转换为第一文本标注图。

[0079] 为便于理解,本公开如下分别对以上述两种方式生成第一文本标注图的实施流程进行阐述。

[0080] 图6是本公开示出的一种通过方式一生成第一文本标注图的方法流程图,如图6所示,包括以下步骤。

[0081] 在步骤S601中,针对多个层级中任意两个相邻层级,分别确定相邻层级中较高层级中存在的目标节点,目标节点包含有相邻层级中较低层级中全部第二节点所包含的各个分词。

[0082] 在步骤S602中,针对多个层级中任意两个相邻层级,将相邻层级中较高层级中所包括的目标节点与较低层级中所包括的第二节点以及第三节点分别通过边连接,生成目标文本的第一文本标注图。

[0083] 本公开实施例提供的方法,可以在确定第一节点、第二节点以及第三节点的情况下,直接生成目标文本的第一文本标注图,该方法在保证标注效率的同时,通过第一文本标注图具有层次化结构的特性,满足对文本进行标注的实际需求。

[0084] 上述实施例中,确定第一节点、第二节点以及第三节点,并通过边将各节点连接以生成第一文本标注图的完整流程,可预训练的神经网络来完成。例如,基于人工预配置多个可作为金标准的HOIA图,并通过神经网络学习HOIA图的结构化信息,以使训练后的神经网络可以完成对文本的标注,生成文本的HOIA图。

[0085] 相应的,除上述通过方式一生成第一文本标注图外,还可以通过如下步骤实现以

上述方式二生成第一文本标注图。

[0086] 图7是本公开示出的一种通过方式二生成第一文本标注图的方法流程图,如图7所示,包括以下步骤。

[0087] 在步骤S701中,生成目标文本的第二文本标注图。

[0088] 本公开实施例中,目标文本的第二文本标注图表征基于目标文本生成的OIA图,其结构特性与前述涉及的OIA图相一致。例如,第二文本标注图中包括节点和边,目标文本中各个分词与第二文本标注图中各个节点之间一一对应,每一边连接有具有父子关系的两个节点,用于标识两个节点中子节点的词性,第二文本标注图中具有父子关系的两个节点所对应的分词之间具有依存关系。

[0089] 在步骤S702中,将第二文本标注图转换为包括第一节点、第二节点以及第三节点的文本标注图,得到第一文本标注图。

[0090] 本公开实施例提供的方法,可以实现将OIA图转换为HOIA图,该方法在提供另一种生成HOIA图的可行实施方式的同时,使HIOA图的构建适配于已生成OIA图的场景,进而实现对已生成OIA图的回收利用。

[0091] 示例的,可通过预配置的OIA语法分析器和规则引擎,生成目标文本的第一文本标注图。

[0092] 图8是本公开示出的一种通过规则引擎生成HOIA图的流程示意图,如图8所示,可以通过OIA语法分析器,将目标文本转换为OIA图,或直接获取待进行信息提取的目标文本的OIA图,进而将所得到的OIA图输入规则引擎,由规则引擎完成对OIA图的特定转换步骤,以得到目标文本的HOIA图。其中,通过OIA语法分析器得到OIA图的方式与相关技术中以OIA图进行文本标注的方式并无本质区别,本公开在此不做赘述,如下主要阐述通过规则引擎将OIA图转换为HOIA图的具体流程。为便于描述,将在指定边连接的两节点之间插入的节点称为第一目标节点,将与第一目标节点之间存在公共子节点的节点称为第二目标节点。

[0093] 图9是本公开示出的一种将第二文本标注图转换为第一文本标注图的方法流程图,如图9所示,包括以下步骤S901至步骤S907。

[0094] 在步骤S901中,确定第二文本标注图中标识词性为指定谓词的指定边。

[0095] 在步骤S902中,在指定边连接的两节点之间插入第一目标节点,第一目标节点用于标识指定谓词。

[0096] 在步骤S903中,将第一目标节点转换为指定边连接的两节点的公共父节点,得到第一转换后的第二文本标注图。

[0097] 在步骤S904中,在第一转换后的第二文本标注图中,确定与第一目标节点之间存在公共子节点的第二目标节点。

[0098] 在步骤S905中,将第一目标节点转换为第二目标节点的子节点,并将第一目标节点转换为公共子节点的父节点,并将第二目标节点转换为公共子节点的祖父节点,得到第二转换后的第二文本标注图。

[0099] 在步骤S906中,针对第二转换后的第二文本标注图,对每一非叶节点所包含的分词进行补充,以使每一非叶节点所包含的分词为非叶节点的各个子节点所包含分词的并集。

[0100] 其中,可以理解的是,非叶节点是指文本标注图中除最低层级外的其他各个层级

所包含的各个节点。

[0101] 在步骤S907中,将补充分词后的第二文本标注图,作为第一文本标注图。

[0102] 在此基础上,为便于理解上述步骤S901至步骤S907,以下结合图10至图13,对第二文本标注图的转换流程进行示例性说明。其中,图10是本公开示出的一种第二文本标注图的示意图,图11是本公开示出的一种插入第一目标节点后的第二文本标注图的示意图,图12是本公开示出的一种第一转换后的第二文本标注图的示意图,图13是本公开示出的一种第二转换后的第二文本标注图的示意图。

[0103] 示例的,如图10所示,针对第二文本标注图,标识词性为指定谓词的指定边例如可以是标注有“同位语关系谓词”的边。在此基础上,标号为“2”及标号为“5”的两个节点即为第三节点和第四节点。在标号为“2”的节点与标号为“5”的节点之间插入第一目标节点,即可得到插入第一目标节点后的第二文本标注图。如图11所示,针对插入第一目标节点后的第二文本标注图,存在标识有“as:pred.arg.1”的边,其中“as:”表示所连接的两个节点具有反向的父子关系。在此基础上,将标识“as:pred.arg.1”转换为用于标识正向父子关系的标识“pred.arg.1”,即可得到第一转换后的第二文本标注图。如图12所示,针对第一转换后的第二文本标注图,标号为“1”的节点即为与第一目标节点之间存在公共子节点的第二目标子节点,二者间的公共子节点即为标号为“2”的节点。在此基础上,将标号为“9”的第一目标节点转换为标号为“2”的节点与标号为“1”的第二目标子节点之间的中间节点,即可得到第二转换后的第二文本标注图。如图13所示,针对第二转换后的第二文本标注图,各节点或各层级之间的层级关系并不明确,因而还需要对各个节点所标识的分词进行补充。示例的,可以按照“每一非叶节点所包含的分词为非叶节点的各个子节点所包含分词的并集”的方式进行分词补充。如图13所示,以标号为“0”的节点为例,对标号为“0”的节点进行分词补充,补充后标号为“0”节点应标识有目标文本的全部分词。在对各个非叶节点进行分词补充后,即可得到如图3所示的第一文本标注图。

[0104] 本公开实施例中涉及的指定谓词,例如可以包括修饰关系谓词(modification)、并联关系谓词(parataxis)、同位语关系谓词(appositive)以及丢失谓词(missing)之一或组合。并且,需要说明的是,上述仅是示例性列举了指定谓词的可选项,而并不说明指定谓词仅限于此。

[0105] 此外,考虑到直接通过神经网络学习HOIA图的方式,训练过程存在大量信息沉余,且考虑到上述将OIA图转换为HOIA图的各个中间结果同样可以作为神经网络的学习目标。一实施方式中,可以将上述涉及的第二转换后的第二文本标注图作为训练神经网络的金标准,以使神经网络基于输入文本输出文本的第二转换后的OIA图。进一步的,可通过在神经网络的输出衔接补充分词的相关流程,用以得到输入文本所对应的HOIA图,该方法同样可直接基于目标文本生成HOIA图,且相较于神经网络直接学习HOIA图的方式,该方法得到的第一文本标注图更加贴合HOIA的真实结构,具有更高的标注精度。

[0106] 基于相同的构思,本公开实施例还提供一种提取文本信息的装置。

[0107] 可以理解的是,本公开实施例提供的提取文本信息的装置为了实现上述功能,其包含了执行各个功能相应的硬件结构和/或软件模块。结合本公开实施例中所公开的各示例的模块及算法步骤,本公开实施例能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行,取决于技术方案的特定应用

和设计约束条件。本领域技术人员可以对每个特定的应用来使用不同的方法来实现所描述的功能,但是这种实现不应认为超出本公开实施例的技术方案的范围。

[0108] 图14是根据本公开的提取文本信息的装置框图。参照图14,该装置1400包括获取模块1401、确定模块1402、生成模块1403和处理模块1404。

[0109] 获取模块1401,用于获取待进行信息提取的目标文本。确定模块1402,用于确定目标文本所包含的各个分词,确定各个分词之间的语义关系,并确定用于表征语义关系的谓词。生成模块1403,用于基于谓词,并按照语义关系,生成目标文本的第一文本标注图。处理模块1404,用于基于第一文本标注图,提取目标文本的文本信息。

[0110] 一种实施方式中,生成模块1403采用如下方式基于谓词,并按照语义关系,生成目标文本的第一文本标注图:基于谓词,确定多个层级,多个层级中包括最高层级,以及不同于最高层级的其他层级。最高层级中包括单一的第一节点,第一节点用于标识各个分词。基于谓词以及语义关系,确定其他层级中各层级包括第二节点以及第三节点。第二节点用于标识各个分词中的一个或多个分词,第三节点用于标识单一的谓词,且第三节点所标识的谓词用于表征其他层级中同一层级内各个第二节点所标识的分词之间的语义关系。基于第一节点、第二节点以及第三节点,生成目标文本对应多个层级的第一文本标注图。

[0111] 一种实施方式中,生成模块1403采用如下方式基于谓词,确定多个层级:确定谓词的数量,并确定数量个谓词之间的主次关系。

[0112] 按照主次关系,确定具有层级关系的数量个其他层级。层级关系用于表征相邻两层级中的较高层级和较低层级。其中,较高层级对应主要关系谓词,较低层级对应次要关系谓词。将数量个其他层级中层级关系最高的层级,作为与最高层级相邻的较低层级,得到多个层级。

[0113] 一种实施方式中,生成模块1403采用如下方式基于第一节点、第二节点以及第三节点,生成目标文本的第一文本标注图:针对多个层级中任意两个相邻层级,分别确定相邻层级中较高层级中存在的目标节点,目标节点包含有相邻层级中较低层级中全部第二节点所包含的各个分词。针对多个层级中任意两个相邻层级,将相邻层级中较高层级中所包括的目标节点与较低层级中所包括的第二节点以及第三节点分别通过边连接,生成目标文本的第一文本标注图。

[0114] 一种实施方式中,生成模块1403采用如下方式基于第一节点、第二节点以及第三节点,生成目标文本对应多个层级的第一文本标注图:生成目标文本的第二文本标注图。其中,第二文本标注图中包括节点和边,目标文本中各个分词与第二文本标注图中各个节点之间一一对应,每一边连接有具有父子关系的两个节点,用于标识两个节点所对应的分词之间的依存关系。将第二文本标注图转换为包括第一节点、第二节点以及第三节点的文本标注图,得到第一文本标注图。

[0115] 一种实施方式中,生成模块1403采用如下方式将第二文本标注图转换为包括第一节点、第二节点以及第三节点的文本标注图:确定第二文本标注图中标识词性为指定谓词的指定边。在指定边连接的两节点之间插入第一目标节点,第一目标节点用于标识指定谓词。将第一目标节点转换为指定边连接的两节点的公共父节点,得到第一转换后的第二文本标注图。在第一转换后的第二文本标注图中,确定与第一目标节点之间存在公共子节点的第二目标节点。将第一目标节点转换为第二目标节点的子节点,并将第一目标节点转换

为公共子节点的父节点,并将第二目标节点转换为公共子节点的祖父节点,得到第二转换后的第二文本标注图。针对第二转换后的第二文本标注图,对每一非叶节点所包含的分词进行补充,以使每一非叶节点所包含的分词为非叶节点的各个子节点所包含分词的并集。将补充分词后的第二文本标注图,作为第一文本标注图。

[0116] 一种实施方式中,指定谓词包括以下之一或组合:修饰关系谓词、并联关系谓词、同位语关系谓词以及丢失谓词。

[0117] 关于上述实施例中的装置,其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述,此处将不做详细阐述说明。

[0118] 根据本公开的实施例,本公开还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0119] 图15示出了可以用来实施本公开的实施例的示例电子设备1500的示意性框图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0120] 如图15所示,设备1500包括计算单元1501,其可以根据存储在只读存储器(ROM)1502中的计算机程序或者从存储单元1508加载到随机访问存储器(RAM)1503中的计算机程序,来执行各种适当的动作和处理。在RAM 1503中,还可存储设备1500操作所需的各种程序和数据。计算单元1501、ROM 1502以及RAM 1503通过总线1504彼此相连。输入/输出(I/O)接口1505也连接至总线1504。

[0121] 设备1500中的多个部件连接至I/O接口1505,包括:输入单元1506,例如键盘、鼠标等;输出单元1507,例如各种类型的显示器、扬声器等;存储单元1508,例如磁盘、光盘等;以及通信单元1509,例如网卡、调制解调器、无线通信收发机等。通信单元1509允许设备1500通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0122] 计算单元1501可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元1501的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元1501执行上文所描述的各个方法和处理,例如提取文本信息的方法。例如,在一些实施例中,提取文本信息的方法可被实现为计算机软件程序,其被有形地包含于机器可读介质,例如存储单元1508。在一些实施例中,计算机程序的部分或者全部可以经由ROM 1502和/或通信单元1509而被载入和/或安装到设备1500上。当计算机程序加载到RAM 1503并由计算单元1501执行时,可以执行上文描述的提取文本信息的方法的一个或多个步骤。备选地,在其他实施例中,计算单元1501可以通过其他任何适当的方式(例如,借助于固件)而被配置为执行提取文本信息的方法。

[0123] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算

机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0124] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器或控制器,使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行,作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0125] 在本公开的上下文中,机器可读介质可以是有形的介质,其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备,或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0126] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0127] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0128] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,也可以为分布式系统的服务器,或者是结合了区块链的服务器。

[0129] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本公开公开的技术方案所期望的结果,本文在此不进行限制。

[0130] 上述具体实施方式,并不构成对本公开保护范围的限制。本领域技术人员应该明白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本公开

的精神和原则之内所作的修改、等同替换和改进等,均应包含在本公开保护范围之内。

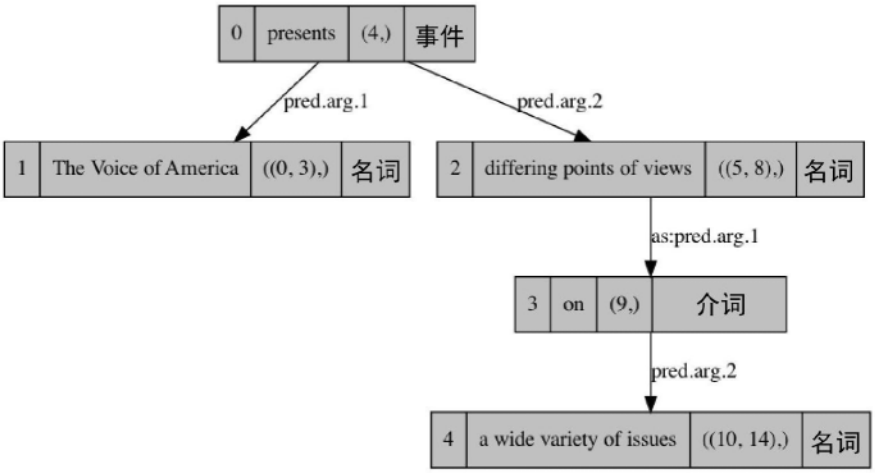

0
presents
(4,)
事件
pred.arg.1
pred.arg.2
1
The Voice of America
((0, 3),)
名词
2
differing points of views
((5, 8),)
名词
as:pred.arg.1
3
on
(9,)
介词
pred.arg.2
4
a wide variety of issues
((10, 14),)
名词

图1

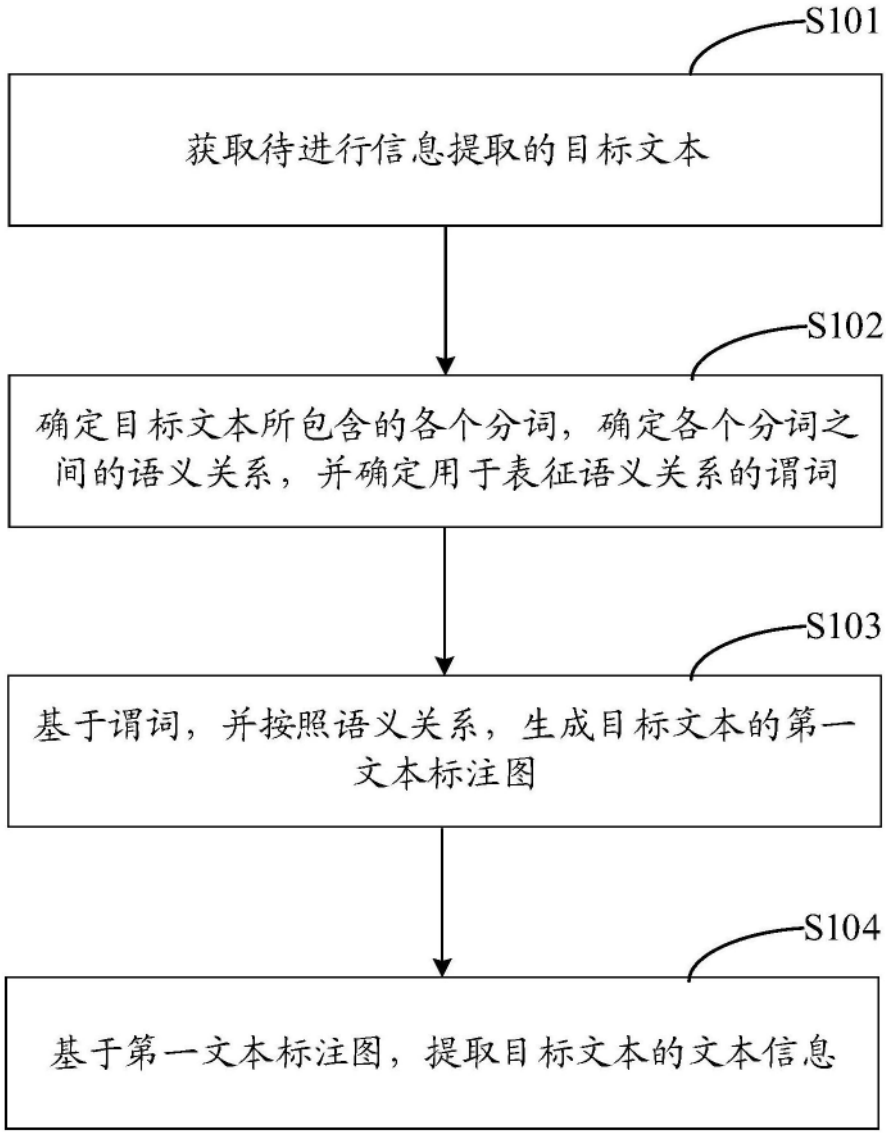

S101
获取待进行信息提取的目标文本
S102
确定目标文本所包含的各个分词，确定各个分词之间的语义关系，并确定用于表征语义关系的谓词
S103
基于谓词，并按照语义关系，生成目标文本的第一文本标注图
S104
基于第一文本标注图，提取目标文本的文本信息

图2

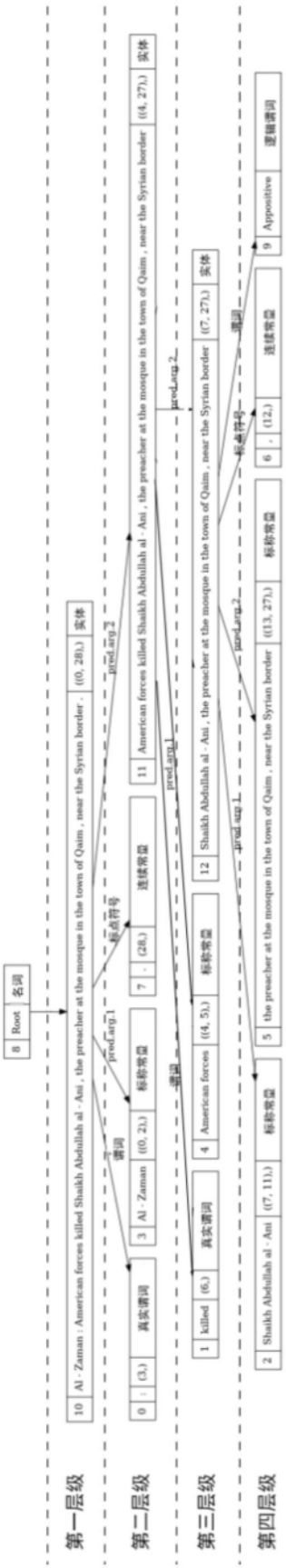

第一层级
第二层级
第三层级
第四层级
8 Root 名词
10 Al · Zaman : American forces killed Shaikh Abdullah al · Ani , the preacher at the mosque in the town of Qaim , near the Syrian border . ((0, 28),) 实体
0 : (3,) 真实谓词
3 Al · Zaman ((0, 2),) 标称常量
7 . (28,) 连续常量
11 American forces killed Shaikh Abdullah al · Ani , the preacher at the mosque in the town of Qaim , near the Syrian border ((4, 27),) 实体
1 killed (6,) 真实谓词
4 American forces ((4, 5),) 标称常量
12 Shaikh Abdullah al · Ani , the preacher at the mosque in the town of Qaim , near the Syrian border ((7, 27),) 实体
2 Shaikh Abdullah al · Ani ((7, 11),) 标称常量
5 the preacher at the mosque in the town of Qaim , near the Syrian border ((13, 27),) 标称常量
6 . (12,) 连续常量
9 Appositive 逻辑谓词
谓词
pred_arg 1
pred_arg 2
标点符号

图3

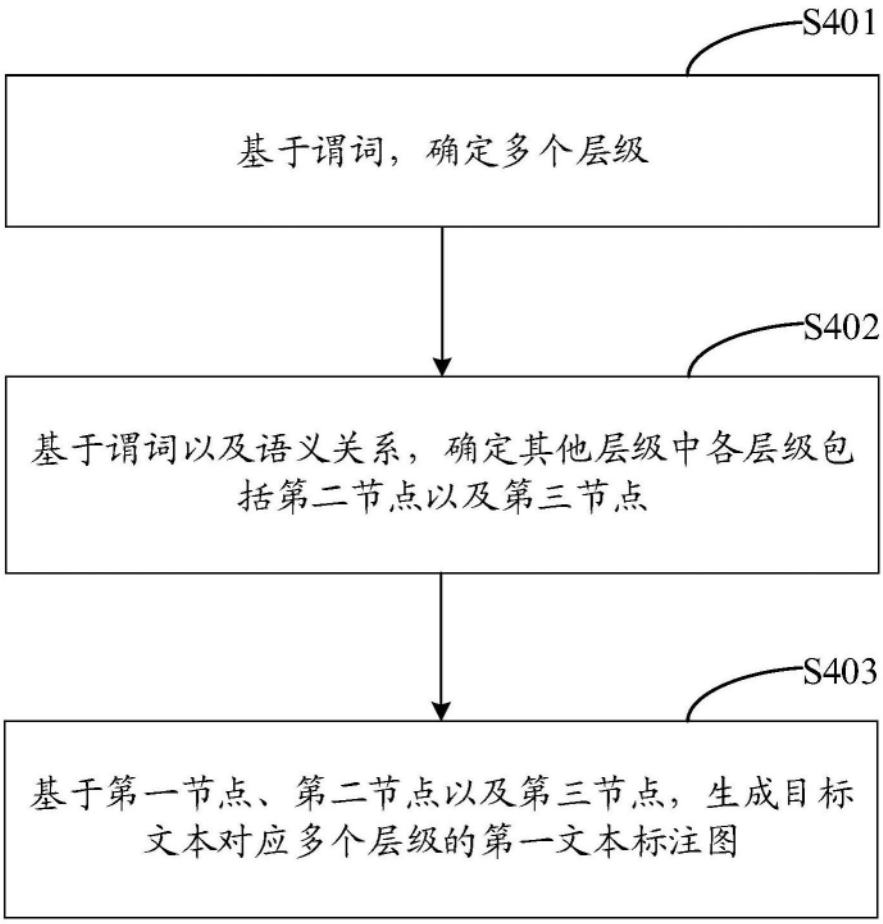

S401
基于谓词，确定多个层级
S402
基于谓词以及语义关系，确定其他层级中各层级包括第二节点以及第三节点
S403
基于第一节点、第二节点以及第三节点，生成目标文本对应多个层级的第一文本标注图

图4

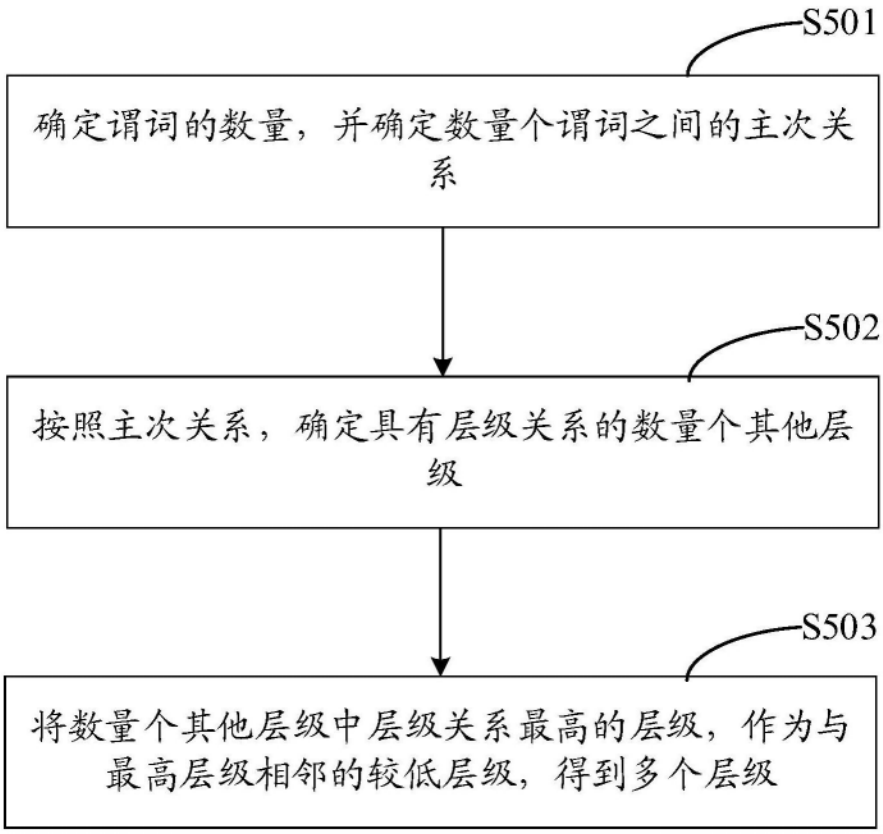

S501
确定谓词的数量，并确定数量个谓词之间的主次关系
S502
按照主次关系，确定具有层级关系的数量个其他层级
S503
将数量个其他层级中层级关系最高的层级，作为与最高层级相邻的较低层级，得到多个层级

图5

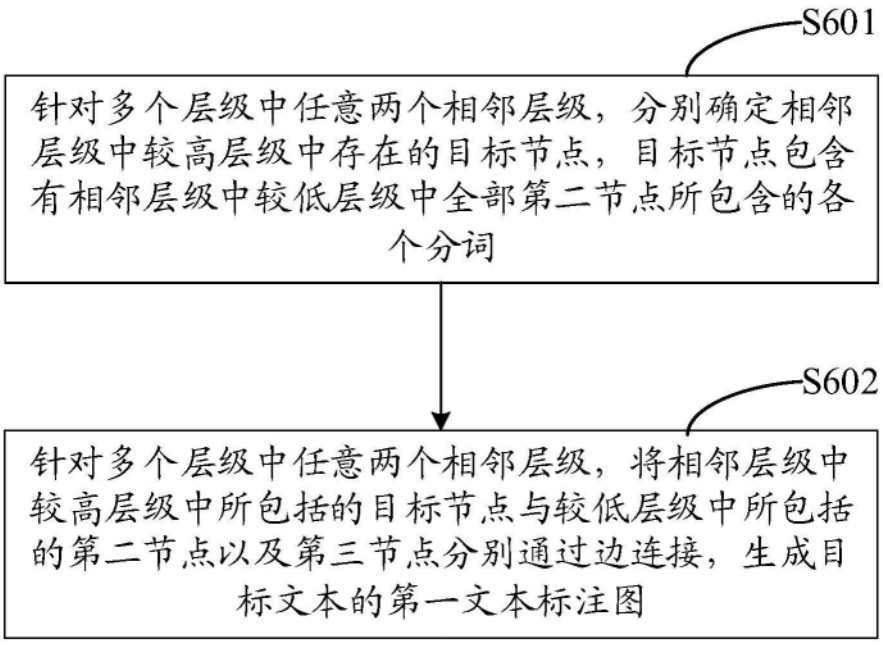

S601
针对多个层级中任意两个相邻层级，分别确定相邻层级中较高层级中存在的目标节点，目标节点包含有相邻层级中较低层级中全部第二节点所包含的各个分词
S602
针对多个层级中任意两个相邻层级，将相邻层级中较高层级中所包括的目标节点与较低层级中所包括的第二节点以及第三节点分别通过边连接，生成目标文本的第一文本标注图

图6

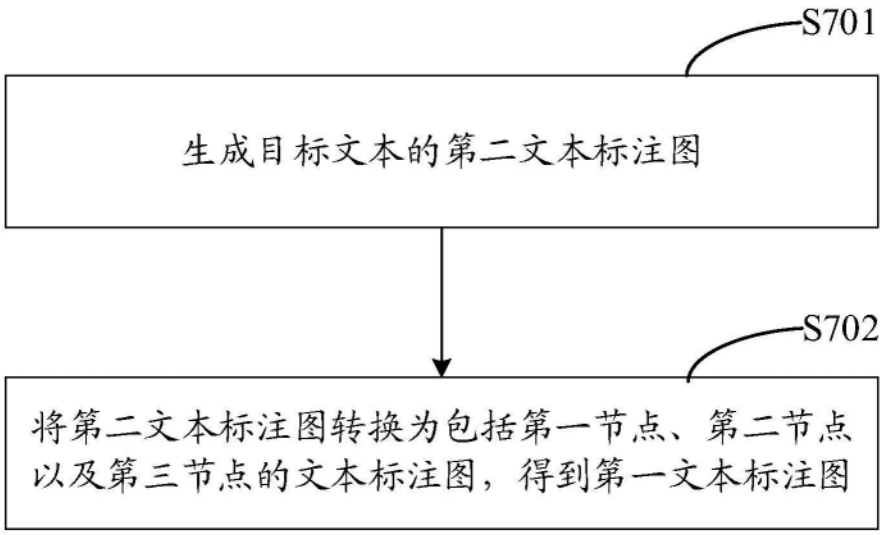

S701
生成目标文本的第二文本标注图
S702
将第二文本标注图转换为包括第一节点、第二节点以及第三节点的文本标注图，得到第一文本标注图

图7

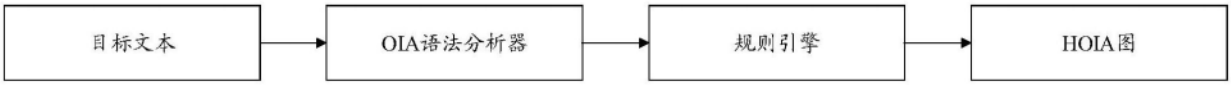

目标文本
OIA语法分析器
规则引擎
HOIA图

图8

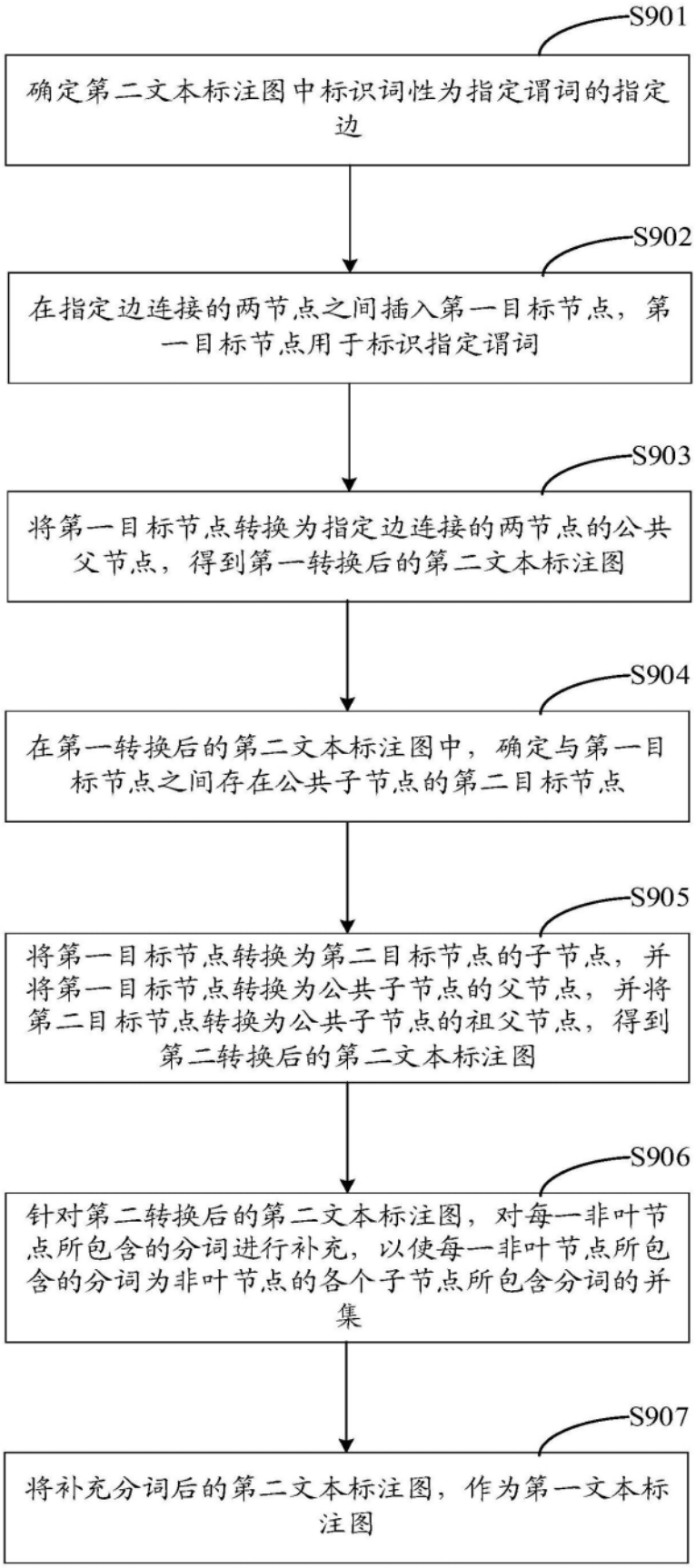

S901
确定第二文本标注图中标识词性为指定谓词的指定边
S902
在指定边连接的两节点之间插入第一目标节点，第一目标节点用于标识指定谓词
S903
将第一目标节点转换为指定边连接的两节点的公共父节点，得到第一转换后的第二文本标注图
S904
在第一转换后的第二文本标注图中，确定与第一目标节点之间存在公共子节点的第二目标节点
S905
将第一目标节点转换为第二目标节点的子节点，并将第一目标节点转换为公共子节点的父节点，并将第二目标节点转换为公共子节点的祖父节点，得到第二转换后的第二文本标注图
S906
针对第二转换后的第二文本标注图，对每一非叶节点所包含的分词进行补充，以使每一非叶节点所包含的分词为非叶节点的各个子节点所包含分词的并集
S907
将补充分词后的第二文本标注图，作为第一文本标注图

图9

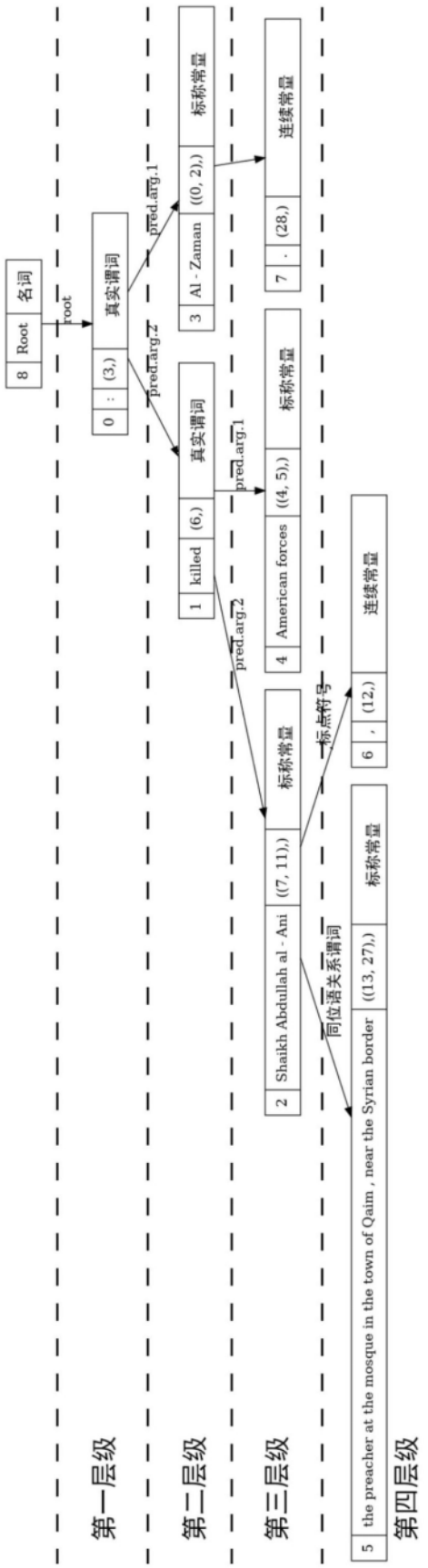

第一层级
第二层级
第三层级
第四层级
8 Root 名词
root
0 : (3,)
真实谓词
pred.arg.1
pred.arg.2
3 Al - Zaman ((0, 2),)
标称常量
7 . (28,)
连续常量
1 killed (6,)
真实谓词
pred.arg.1
pred.arg.2
4 American forces ((4, 5),)
标称常量
2 Shaikh Abdullah al - Ani ((7, 11),)
标称常量
标点符号
6 , (12,)
连续常量
同位语关系谓词
5 the preacher at the mosque in the town of Qaim , near the Syrian border ((13, 27),)
标称常量

图10

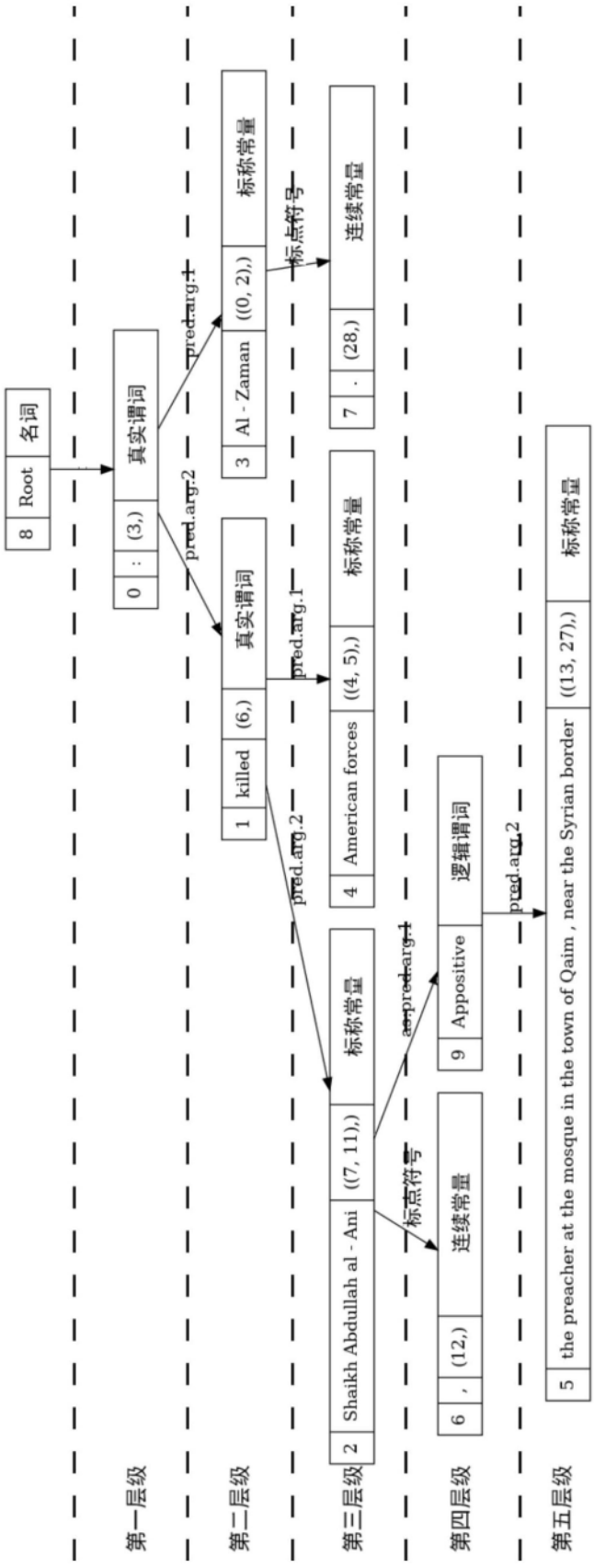

8 Root 名词
0 : (3,) 真实谓词
1 killed (6,) 真实谓词
2 Shaikh Abdullah al - Ani ((7, 11),) 标称常量
3 Al - Zaman ((0, 2),) 标称常量
4 American forces ((4, 5),) 标称常量
5 the preacher at the mosque in the town of Qaim , near the Syrian border ((13, 27),) 标称常量
6 , (12,) 连续常量
7 . (28,) 连续常量
9 Appositive 逻辑谓词
pred.arg.1
pred.arg.2
标点符号
第一层级
第二层级
第三层级
第四层级
第五层级

图11

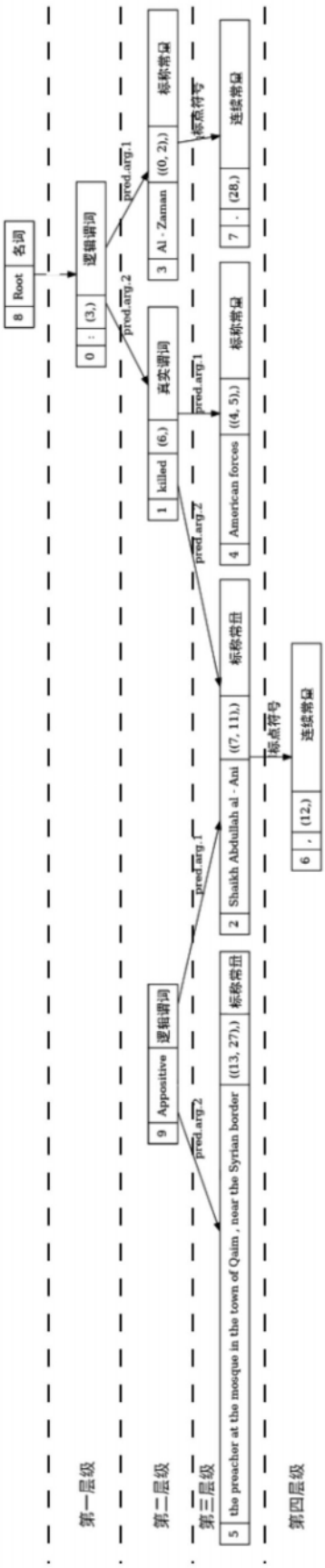

第一层级
第二层级
第三层级
第四层级
8 Root 名词
0 : (3,) 逻辑谓词
3 Al - Zaman ((0, 2),) 标称带量
7 . (28,) 连续带量
1 killed (6,) 真实谓词
4 American forces ((4, 5),) 标称带量
9 Appositive 逻辑谓词
2 Shaikh Abdullah al - Ani ((7, 11),) 标称带量
5 the preacher at the mosque in the town of Qaim , near the Syrian border ((13, 27),) 标称带量
6 , (12,) 连续带量
pred.arg.1
pred.arg.2
标点符号

图12

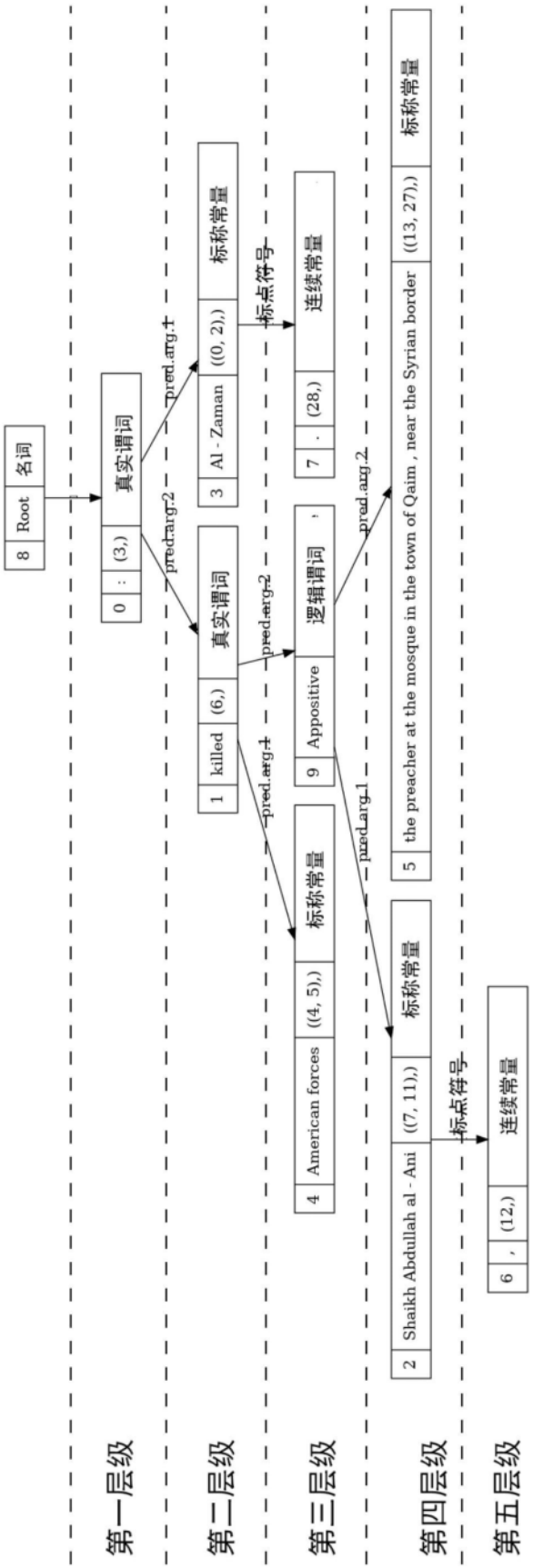

8 Root 名词
0 : (3,) 真实谓词
1 killed (6,) 真实谓词
3 Al - Zaman ((0, 2),) 标称常量
7 . (28,) 连续常量
4 American forces ((4, 5),) 标称常量
9 Appositive 逻辑谓词 ,
2 Shaikh Abdullah al - Ani ((7, 11),) 标称常量
5 the preacher at the mosque in the town of Qaim , near the Syrian border ((13, 27),) 标称常量
6 , (12,) 连续常量
pred.arg.1
pred.arg.2
标点符号
第一层级
第二层级
第三层级
第四层级
第五层级

图13

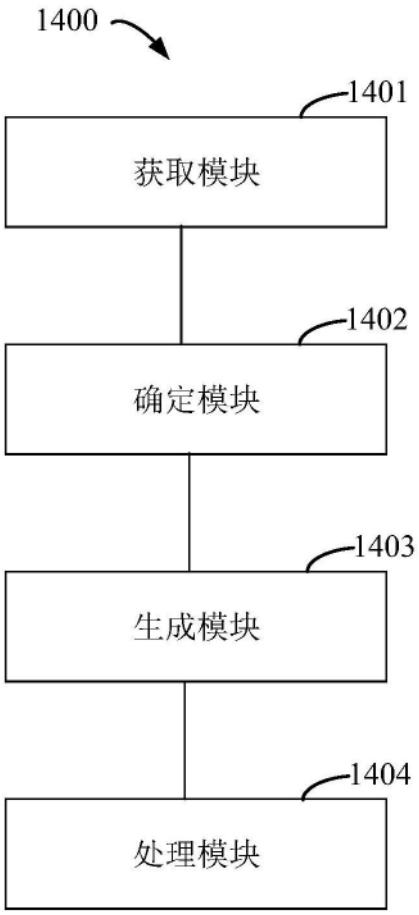

1400
1401
获取模块
1402
确定模块
1403
生成模块
1404
处理模块

图14

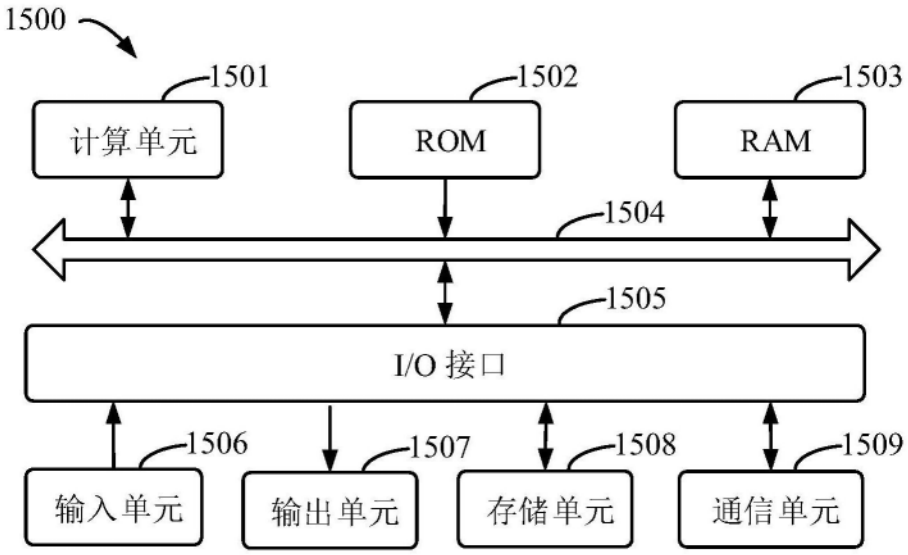

1500
1501
计算单元
1502
ROM
1503
RAM
1504
1505
I/O 接口
1506
输入单元
1507
输出单元
1508
存储单元
1509
通信单元

图15