



(12) 发明专利

(10) 授权公告号 CN 112106056 B

(45) 授权公告日 2025. 06. 24

(21) 申请号 201980030899.2
(22) 申请日 2019.05.09
(65) 同一申请的已公布的文献号
申请公布号 CN 112106056 A
(43) 申请公布日 2020.12.18
(30) 优先权数据
62/668,963 2018.05.09 US
62/729,335 2018.09.10 US
(85) PCT国际申请进入国家阶段日
2020.11.06
(86) PCT国际申请的申请数据
PCT/US2019/031580 2019.05.09
(87) PCT国际申请的公布数据
W02019/217722 EN 2019.11.14
(73) 专利权人 甲骨文国际公司
地址 美国加利福尼亚

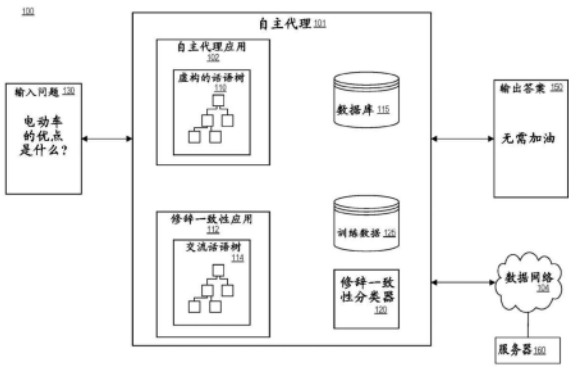
(72) 发明人 B·加里斯基
(74) 专利代理机构 中国贸促会专利商标事务所
有限公司 11038
专利代理师 刘玉洁
(51) Int.Cl.
G06F 40/289 (2020.01)
G06F 40/295 (2020.01)
G06F 40/35 (2020.01)
(56) 对比文件
CN 102165518 A,2011.08.24
CN 106383875 A,2017.02.08
Boris Galitsky.Matching parse
thickets for open domain question
answering.Data & Knowledge
Engineering.2016,第107卷第24-50页.
审查员 辛铁钢
权利要求书4页 说明书38页 附图20页

(54) 发明名称

构造虚构的话语树来提高回答聚敛性问题的能力

(57) 摘要

用于改进复杂的、多句子、聚敛性问题的问答式召回的系统和方法。更具体而言,自主代理访问初始答案,该初始答案部分地回答从用户设备接收到的问题。代理将问题和初始答案表示为话语树。代理从话语树中识别问题中未由答案解决的实体。代理从诸如文本语料库之类的附加资源中形成附加的话语树。附加的话语树在修辞上将未解决的实体与答案连接起来。代理将该话语树指定为虚构的话语树。虚构的话语树在与初始答案话语树组合时,用于生成相对于现有解决方案改进的答案。



1. 一种方法, 包括:

使用计算设备并从问题中构造包括问题实体的问题话语树, 其中问题话语树表示问题的基本话语单元之间的修辞相互关系;

使用计算设备从文本语料库访问初始答案;

使用计算设备从初始答案中构造包括答案实体的答案话语树, 其中答案话语树表示初始答案的基本话语单元之间的修辞相互关系;

使用计算设备确定指示答案实体与问题实体的相关性的分数低于阈值;

通过以下方式生成虚构的话语树:

根据文本语料库创建附加的话语树;

确定附加的话语树包括将问题实体与答案实体连接的修辞关系;

提取包括问题实体、答案实体和修辞关系的附加的话语树的子树, 从而生成虚构的话语树; 以及

输出由答案话语树和虚构的话语树的组合表示的答案。

2. 如权利要求1所述的方法, 其中访问初始答案包括:

确定文本的一部分的答案相关性分数; 以及

响应于确定答案相关性分数大于阈值, 选择文本的所述一部分作为初始答案。

3. 如权利要求1所述的方法, 其中虚构的话语树包括表示修辞关系的节点, 所述方法还包括通过将所述节点连接到答案实体来将虚构的话语树集成到答案话语树中。

4. 如权利要求1所述的方法, 其中创建附加的话语树包括:

为多个附加的话语树中的每个附加的话语树计算分数, 该分数指示包括到相应的附加的话语树中的一个或多个答案实体的映射的问题实体的数量; 以及

从所述多个附加的话语树中选择具有最高分数的附加的话语树。

5. 如权利要求1所述的方法, 其中创建附加的话语树包括:

通过对 (a) 问题话语树和 (b) 相应的附加的答案话语树中的一个或多个话语树应用经训练的分类模型, 为多个附加的话语树中的每个附加的话语树计算分数; 以及

从所述多个附加的话语树中选择具有最高分数的附加的话语树。

6. 如权利要求1所述的方法, 其中问题包括多个关键字, 并且其中访问初始答案包括:

通过执行多个电子文档的搜索, 基于包括关键字的搜索查询, 获得多个答案;

对于所述多个答案中的每个答案, 确定指示问题与相应答案之间的匹配程度的答案分数; 以及

从所述多个答案中选择具有最高分数的答案作为初始答案。

7. 如权利要求1所述的方法, 其中计算分数包括:

将经训练的分类模型应用于 (a) 问题话语树和 (b) 答案话语树中的一个或多个话语树; 以及

从分类模型接收分数。

8. 如权利要求1所述的方法, 其中构造话语树包括:

访问包括多个片段的句子, 其中至少一个片段包括动词和多个词语, 每个词语包括词语在片段内的角色, 其中每个片段是基本话语单元; 以及

生成表示所述多个片段之间的修辞相互关系的话语树, 其中话语树包括多个节点, 每

个非终端节点表示所述多个片段中的两个片段之间的修辞相互关系,话语树的节点中的每个终端节点与所述多个片段中的一个片段相关联。

9.如权利要求1所述的方法,还包括:

根据问题话语树确定包括问题根节点的问题交流话语树,其中交流话语树是包括交流动作的话语树,并且其中所述生成还包括:

根据虚构的话语树确定答案交流话语树,其中答案交流话语树包括答案根节点;

通过识别问题根节点和答案根节点是相同的来合并交流话语树;

通过将预测模型应用于合并后的交流话语树,计算问题交流话语树与答案交流话语树之间的互补程度;以及

响应于确定互补程度高于阈值,输出与虚构的话语树对应的最终答案。

10.如权利要求1所述的方法,其中话语树表示文本的多个片段之间的修辞相互关系,其中话语树包括多个节点,每个非终端节点表示所述多个片段中的两个片段之间的修辞相互关系,话语树的节点中的每个终端节点与所述多个片段中的一个片段相关联;并且其中构造交流话语树包括:

通过以下方式将具有动词的每个片段与动词签名匹配:

访问多个动词签名,其中每个动词签名包括该片段的动词和题元角色序列,其中题元角色描述动词和相关词语之间的关系;

对于所述多个动词签名中的每个动词签名,确定相应签名的与该片段中的词语的角色匹配的多个题元角色;

基于特定动词签名包括最大数量的匹配,从所述多个动词签名中选择所述特定动词签名;以及

将所述特定动词签名与该片段相关联。

11.一种计算机实现的方法,包括:

为问题构造包括多个问题实体的问题话语树;

从初始答案中构造包括多个答案实体的答案话语树;

在所述多个问题实体中的第一问题实体与所述多个答案实体中的答案实体之间建立映射,该映射建立所述答案实体与所述第一问题实体的相关性;

响应于确定所述多个问题实体中的第二问题实体未被所述多个答案实体中的任何答案实体解决,通过将附加答案对应的附加的话语树与答案话语树组合来生成虚构的话语树;

根据问题话语树确定问题交流话语树;

根据虚构的话语树确定答案交流话语树;

通过应用预测模型、问题交流话语树和答案交流话语树,计算问题交流话语树与答案交流话语树之间的互补程度;以及

响应于确定互补程度高于阈值,输出与虚构的话语树对应的最终答案。

12.一种系统,包括:

存储非暂态计算机可执行程序指令的计算机可读介质;以及

通信地耦合到计算机可读介质的用于执行非暂态计算机可执行程序指令的处理设备,其中执行非暂态计算机可执行程序指令将处理设备配置为执行操作,所述操作包括:

使用计算设备从问题中构造包括问题实体的问题话语树,其中问题话语树表示问题的基本话语单元之间的修辞相互关系;

使用计算设备并从文本语料库访问初始答案;

使用计算设备从初始答案中构造包括答案实体的答案话语树,其中答案话语树表示初始答案的基本话语单元之间的修辞相互关系;

使用计算设备确定指示答案实体与问题实体的相关性的分数低于阈值;

通过以下方式生成虚构的话语树:

根据文本语料库创建附加的话语树;

确定附加的话语树包括将问题实体与答案实体连接的修辞关系;

提取包括问题实体、答案实体和修辞关系的附加的话语树的子树,从而生成虚构的话语树;以及

输出由答案话语树和虚构的话语树的组合表示的答案。

13. 如权利要求12所述的系统,其中访问初始答案包括:

确定文本的一部分的答案相关性分数;以及

响应于确定答案相关性分数大于阈值,选择文本的所述一部分作为初始答案。

14. 如权利要求12所述的系统,其中虚构的话语树包括表示修辞关系的节点,所述操作还包括:通过将所述节点连接到答案实体来将虚构的话语树集成到答案话语树中。

15. 如权利要求12所述的系统,其中创建附加的话语树包括:

为多个附加的话语树中的每个附加的话语树计算分数,该分数指示包括到相应的附加的话语树中的一个或多个答案实体的映射的问题实体的数量;以及

从所述多个附加的话语树中选择具有最高分数的附加的话语树。

16. 如权利要求12所述的系统,其中创建附加的话语树包括:

通过对(a)问题话语树和(b)相应的附加的答案话语树中的一个或多个话语树应用经训练的分类模型,为多个附加的话语树中的每个附加的话语树计算分数;以及

从所述多个附加的话语树中选择具有最高分数的附加的话语树。

17. 如权利要求12所述的系统,其中问题包括多个关键字,并且其中访问初始答案包括:

通过执行多个电子文档的搜索,基于包括关键字的搜索查询,获得多个答案;

对于所述多个答案中的每个答案,确定指示问题与相应答案之间的匹配程度的答案分数;以及

从所述多个答案中选择具有最高分数的答案作为初始答案。

18. 如权利要求12所述的系统,其中计算分数包括:

将经训练的分类模型应用于(a)问题话语树和(b)答案话语树中的一个或多个话语树;以及

从分类模型接收分数。

19. 如权利要求12所述的系统,其中构造话语树包括:

访问包括多个片段的句子,其中至少一个片段包括动词和多个词语,每个词语包括词语在片段内的角色,其中每个片段是基本话语单元;以及

生成表示所述多个片段之间的修辞相互关系的话语树,其中话语树包括多个节点,每

个非终端节点表示所述多个片段中的两个片段之间的修辞相互关系,话语树的节点中的每个终端节点与所述多个片段中的一个片段相关联。

20. 如权利要求12所述的系统,所述操作还包括:

根据问题话语树确定包括问题根节点的问题交流话语树,其中交流话语树是包括交流动作的话语树,并且其中所述生成还包括:

根据虚构的话语树确定答案交流话语树,其中答案交流话语树包括答案根节点;

通过识别问题根节点和答案根节点是相同的来合并交流话语树;

通过将预测模型应用于合并后的交流话语树,计算问题交流话语树与答案交流话语树之间的互补程度;以及

响应于确定互补程度高于阈值,输出与虚构的话语树对应的最终答案。

21. 一种系统,包括:

存储非暂态计算机可执行程序指令的计算机可读介质;以及

通信地耦合到计算机可读介质的用于执行非暂态计算机可执行程序指令的处理设备,其中执行非暂态计算机可执行程序指令将处理设备配置为执行如权利要求11中所述的计算机实现的方法。

22. 一种存储指令的非暂态计算机可读介质,所述指令在由一个或多个处理器执行时,使得所述一个或多个处理器执行如权利要求1-10中任一项所述的方法。

23. 一种存储指令的非暂态计算机可读介质,所述指令在由一个或多个处理器执行时,使得所述一个或多个处理器执行如权利要求11中所述的计算机实现的方法。

24. 一种包括指令的计算机程序产品,所述指令在由计算机的一个或多个处理器执行时,使得所述计算机执行如权利要求1-10中任一项所述的方法。

25. 一种包括指令的计算机程序产品,所述指令在由计算机的一个或多个处理器执行时,使得所述计算机执行如权利要求11中所述的计算机实现的方法。

构造虚构的话语树来提高回答聚敛性问题的能力

[0001] 相关申请的交叉引用

[0002] 本申请要求于2018年9月10日提交的美国临时申请No.62/729,335和2018年5月9日提交的美国临时申请No.62/668,963的权益,这两个美国临时申请的全部内容通过引用并入本文。

背景技术

[0003] 语言学是语言的科学研究。语言学的一方面是将计算机科学应用于人类自然语言,诸如英语。由于处理器的速度和存储器的容量的极大提高,语言学的计算机应用正在增加。例如,启用计算机的对语言话语的分析促进了可以回答来自用户的问题的许多应用,诸如自动化代理。使用自主代理或“聊天机器人(chatbot)”来回答问题、促进讨论、管理对话以及提供社交推广(social promotion)越来越流行。

[0004] 自主代理可以通过基于利用诸如数据库之类的特定资源找到的信息生成答案或通过查询搜索引擎来为从用户设备接收的查询服务。但是有时,单个资源或搜索引擎结果无法完全解决复杂的用户查询或聚敛性(convergent)问题。聚敛性问题是需要具有高度准确性的答案的问题。

[0005] 因此,当在多个资源中提供答案时,现有系统无法形成对问题的完整且准确的答案。因此,需要新的解决方案。

发明内容

[0006] 本文描述的各方面使用虚构的话语树来改进针对复杂的、多句子、聚敛性问题的问答式召回(recall)。更具体而言,改进的自主代理访问针对从用户设备接收的问题的初始答案。初始答案部分而非完全解决问题。代理将问题和初始答案表示为话语树,并识别问题中未在答案中解决的实体。代理访问附加资源,诸如文本语料库,并确定将缺失实体与答案中的另一个实体进行修辞连接的答案。代理选择这个附加资源,从而形成虚构的话语树,该虚构的话语树在与答案的话语树组合时,可以用来生成相对于现有解决方案改进的答案。

[0007] 在一方面,一种方法包括使用计算设备并根据问题构造包括问题实体的问题话语树。问题话语树表示问题的基本话语单元之间的修辞相互关系(relationship)。该方法还包括使用计算设备并且从文本语料库中访问初始答案。该方法还包括使用计算设备从初始答案构造包括答案实体的答案话语树。答案话语树表示初始答案的基本话语单元之间的修辞相互关系。该方法还包括使用计算设备确定指示答案实体与问题实体的相关性的分数低于阈值。该方法还包括生成虚构的话语树。生成虚构的话语树包括从文本语料库中创建附加的话语树。生成虚构的话语树包括确定附加的话语树包括将问题实体与答案实体连接的修辞关系。生成虚构的话语树包括提取包括问题实体、答案实体和修辞关系的附加话语树的子树,从而生成虚构的话语树。生成虚构的话语树包括输出由答案话语树和虚构的话语树的组合表示的答案。

[0008] 在示例中,访问初始答案包括确定文本的一部分的答案相关性分数,以及响应于确定答案相关性分数大于阈值,选择文本的一部分作为初始答案。

[0009] 在示例中,虚构的话语树包括表示修辞关系的节点。该方法还包括通过将节点连接到答案实体来将虚构的话语树集成到答案话语树中。

[0010] 在示例中,创建附加话语树包括针对每个附加话语树计算指示若干个问题实体的分数,该若干个问题实体包括到相应的附加话语树中的一个或多个答案实体的映射。创建附加的话语树包括从附加的话语树中选择具有最高分数的附加话语树。

[0011] 在示例中,创建附加的话语树包括针对每个附加的话语树,通过对 (a) 问题话语树和 (b) 相应的附加答案话语树中的一个或多个应用经训练的分类模型来计算分数;以及从附加的话语树中选择具有最高分数的附加话语树。

[0012] 在示例中,问题包括关键字,并且访问初始答案包括通过执行电子文档的搜索基于包括关键字的搜索查询来获得答案。访问初始答案包括针对每个答案确定指示问题与相应答案之间的匹配程度的答案分数。访问初始答案包括从答案中选择具有最高分数的答案作为初始答案。

[0013] 在示例中,计算分数包括将经训练的分类模型应用于 (a) 问题话语树和 (b) 答案话语树中的一个或多个;以及从分类模型接收分数。

[0014] 在示例中,构造话语树包括访问包括片段的句子。至少一个片段包括动词和词语(word),每个词语包括词语在片段中的角色。每个片段都是基本的话语单元。构建话语树还包括生成表示片段之间的修辞相互关系的话语树。话语树包括节点,每个非终端节点表示两个片段之间的修辞相互关系,话语树的节点中的每个终端节点与片段中的一个片段相关联。

[0015] 在示例中,该方法还包括从问题话语树确定包括问题根节点的问题交流话语树。交流话语树是包括交流动作的话语树。该生成还包括从虚构的话语树确定答案交流话语树。答案交流话语树包括答案根节点。该生成包括通过识别问题根节点和答案根节点相同来合并交流话语树。该生成包括通过将预测模型应用于合并后的交流话语树来计算问题交流话语树与答案交流话语树之间的互补程度。该生成包括响应于确定互补程度高于阈值,输出与虚构的话语树对应的最终答案。

[0016] 在示例中,话语树表示文本片段之间的修辞相互关系。话语树包括节点。表示片段中的两个片段与话语树的节点中的每个终端节点之间的修辞相互关系的每个非终端节点都与片段中的一个片段相关联。构造交流话语树包括将具有动词的每个片段与动词签名进行匹配。该匹配包括访问动词签名。动词签名包括片段的动词和题元角色(thematic role)序列。题元角色描述了动词和相关词语之间的关系。匹配还包括对于动词签名中的每个动词签名,确定与片段中的词语的角色匹配的相应签名的题元角色。该匹配还包括基于特定动词签名包括最大数量的匹配来从动词签名中选择该特定动词签名。该匹配还包括将特定动词签名与片段相关联。

[0017] 在示例中,一种方法包括为问题构造包括问题实体的问题话语树。该方法还包括从初始答案构造包括答案实体的答案话语树。该方法还包括在各问题实体的第一问题实体和各答案实体中的答案实体之间建立映射,该映射建立答案实体与第一问题实体的相关性。该方法还包括响应于确定问题实体中的第二问题实体未被任何答案实体解决,通过将

与附加答案对应的附加话语树与答案话语树组合来生成虚构的话语树。该方法还包括从问题话语树确定问题交流话语树。该方法还包括从虚构的话语树确定答案交流话语树。该方法还包括通过应用预测模型、问题交流话语树和答案交流话语树来计算问题交流话语树和答案交流话语树之间的互补程度。该方法还包括响应于确定互补程度高于阈值,输出与虚构的话语树对应的最终答案。

[0018] 上述方法可以被移植为包括一个或多个处理设备和/或非暂态计算机可读介质的系统,在该系统上可以存储程序指令,该程序指令可以使一个或多个处理器执行关于上述方法描述的操作。

附图说明

[0019] 下面参考以下附图详细描述本发明的说明性方面。

[0020] 图1示出了根据一方面的示例性修辞分类环境。

[0021] 图2描绘了根据一方面的话语树的示例。

[0022] 图3描绘了根据一方面的话语树的进一步示例。

[0023] 图4描绘了根据一方面的说明性模式。

[0024] 图5描绘了根据一方面的分层二叉树的节点链接表示。

[0025] 图6描绘了根据一方面的图5中的表示的示例性缩进文本编码。

[0026] 图7描绘了根据一方面的关于财产税的示例请求的示例性DT。

[0027] 图8描绘了对图7中表示的问题的示例性响应。

[0028] 图9图示了根据一方面的用于官方答案的话语树。

[0029] 图10图示了根据一方面的用于原始答案的话语树。

[0030] 图11图示了根据一方面的用于第一代理的声明(claim)的交流话语树。

[0031] 图12图示了根据一方面的用于第二代理的声明的交流话语树。

[0032] 图13图示了根据一方面的用于第三代理的声明的交流话语树。

[0033] 图14图示了根据一方面的解析丛林(parsing thickets)。

[0034] 图15图示了根据一方面的用于构建交流话语树的示例性处理。

[0035] 图16图示了根据一方面的用于构建虚构的话语树的示例性处理。

[0036] 图17描绘了根据一方面的问题、答案的示例话语树和两个虚构的话语树。

[0037] 图18描绘了用于实现这些方面之一的分布式系统的简化图。

[0038] 图19是根据一方面的系统环境的部件的简化框图,通过该系统环境,可以将一方面系统的部件所提供的服务作为云服务供应。

[0039] 图20图示了其中可以实现本发明的各个方面的示例性计算机系统。

具体实施方式

[0040] 如以上所讨论的,用于自主代理的现有系统存在缺陷。例如,这样的系统无法回答复杂的、多句子或聚敛性查询。这些系统也可能依赖于构建起来既困难又昂贵的本体或领域中不同概念之间的相互关系。此外,一些现有的解决方案采用基于知识图的方法,这可能会限制表达性和覆盖范围。

[0041] 作为对照,本文描述的各方面能够通过采用与领域无关的话语分析来回答复杂的

用户查询。本文描述的各方面使用虚构的话语树来验证并在一些情况下完成问题与答案之间的修辞链接,从而改进针对复杂的、多句子、聚敛性问题的问答式召回。虚构的话语树是表示问题的初始答案和补充的附加答案的组合的话语树。以这种方式,虚构的话语树表示对用户查询的完整答案。

[0042] 为了使给定答案与给定问题相关,答案的实体应覆盖问题的实体。“实体”具有独立且独特的存在。示例包括对象、地点和人物。实体也可以是主体或主题,诸如“电动汽车”、“制动器”或“法国”。

[0043] 但是在一些情况下,问题中的一个或多个实体在初始答案中未得到解决。例如,用户可能询问他的汽车的“引擎”,但初始答案根本没有讨论“引擎”。在其它情况下,替代地,答案中会出现更具体的实体,但是它们与问题实体的连接并不立即是明显的。继续前面的示例,初始答案可能包含实体“火花塞”或“变速器”,而未将这些实体连接到“引擎”。在其它情况下,一些答案实体可能未在问题中明确提及,而是在问题中被假定。为了提供完整的答案,缺失的实体应被考虑,并在适当时被解释。

[0044] 为了填补这一间隙,本公开的某些方面使用虚构的话语树。例如,自主代理应用可以识别提供问题中存在的、答案中缺失的实体之间的链接的答案。该应用将虚构的答案话语树的相关片段组合到表示问题的完整答案的虚构的话语树中。因此,虚构的话语树提供了可以用于验证答案,并且在一些情况下扩充答案的修辞链接。然后,应用可以向用户设备呈现完整答案。

[0045] 通过创建利用从不同来源按需挖掘的文档中获得的树片段来扩充的虚构的话语树,所公开的解决方案消除了对本体的需要。例如,考虑汽车领域,本体可能表示制动器与车轮或变速器与引擎之间的相互关系。作为对照,各方面获得与给定作者的思想结构无关的答案的规范话语表示。

[0046] 在另一个示例中,所公开的解决方案可以使用交流话语树来验证文本的两个部分之间的修辞一致性。例如,与自主代理应用一起工作的修辞一致性应用可以验证问题与答案、问题与答案的一部分、或者初始答案与由虚构的话语树表示的答案之间的修辞一致性,例如,样式。

[0047] “交流话语树”或“CDT”包括补充有交流动作的话语树。交流动作是个人基于相互协商和论证而采取的合作动作,或者如本领域中已知的以其它方式的动作。通过合并识别交流动作的标签,可以在比仅仅是修辞关系和基本话语单元(EDU)的语法丰富的特征集上进行交流话语树的学习。利用这种特征集,可以使用诸如分类之类的附加技术来确定问题与答案或请求-响应对之间的修辞一致性水平,从而实现改进的自动化代理。在这样做时,计算系统实现能够智能地回答问题和其它消息的自主代理。

[0048] 某些定义

[0049] 如本文所使用的,“修辞结构理论”是提供了可以分析话语的连贯性的理论基础研究和学习领域。

[0050] 如本文所使用的,“话语树”或“DT”是指表示用于句子的一部分的句子的修辞关系的结构。

[0051] 如本文中所使用的,“修辞关系”、“修辞相互关系”或“连贯关系”或“话语关系”是指话语的两个段在逻辑上如何彼此连接。修辞关系的示例包括阐述(elaboration)、对比

(contrast)和归因(attribution)。

[0052] 如本文所使用的,“句子片段”或“片段”是可以从句子的其余部分分开的句子的一部分。片段是基本话语单元。例如,对于句子“Dutch accident investigators say that evidence points to pro-Russian rebels as being responsible for shooting down the plane”,两个片段是“Dutch accident investigators say that evidence points to pro-Russian rebels”和“as being responsible for shooting down the plane”。片段可以但不必包含动词。

[0053] 如本文所使用的,“签名”或“框架”是指片段中动词的属性。每个签名可以包括一个或多个题元角色。例如,对于片段“Dutch accident investigators say that evidence points to pro-Russian rebels”,动词是“say”并且动词“say”的这种特殊用法的签名可以为“施事者动词主题(agent verb topic)”,其中“investigators”是施事者,并且“evidence”是主题。

[0054] 如本文所使用的,“题元角色”是指签名中的用于描述一个或多个词语的角色的组成部分。继续前面的示例,“施事者”和“主题”是题元角色。

[0055] 如本文中所使用的,“核性(nuclearity)”是指哪个文本段、片段或区段(span)对于作者的目的而言更为关键。核心(nucleus)是更关键的区段,而卫星(satellite)是更不关键的区段。

[0056] 如本文所使用的,“连贯性(coherency)”是指将两个修辞关系链接在一起。

[0057] 如本文所使用的,“交流动词(communicative verb)”是指示交流的动词。例如,动词“deny(否认)”是交流动词。

[0058] 如本文所使用的,“交流动作(communicative action)”描述了由一个或多个施事者和施事者的主体执行的动作。

[0059] 图1示出了根据一方面的示例性修辞分类环境。图1描绘了自主代理101、输入问题130、输出答案150、数据网络104和服务器160。自主代理101可以包括以下一项或多项:自主代理应用102、数据库115、修辞一致性应用112、修辞一致性分类器120或训练数据125。服务器160可以是公共或私有互联网服务器,诸如用户问题和答案的公共数据库。服务器160提供的功能的示例包括搜索引擎和数据库。数据网络104可以是任何公共或私有网络、有线或无线网络、广域网、局域网或互联网。自主代理应用102和修辞一致性应用112可以在分布式系统1800内执行,如关于图18进一步讨论的。

[0060] 自主代理应用102可以访问输入问题130并生成输出答案150。输入问题130可以是单个问题或诸如聊天之类的问题流。如图所示,输入问题130是“What is an advantage of an electric car?(电动车的优点是什么?)”。自主代理应用102可以例如通过创建一个或多个虚构的话语树110来接收输入问题130并分析输入问题130。

[0061] 例如,自主代理应用102根据输入问题130创建问题话语树,并根据从诸如数据库之类的资源获得的初始答案来创建问题话语树。在一些情况下,自主代理应用102可以获得候选答案集合并选择最佳匹配作为初始答案。通过分析和比较问题话语树中的实体和答案话语树中的实体,自主代理应用102确定问题中未在答案中解决的一个或多个实体。

[0062] 继续该示例,自主代理应用102访问诸如文本语料库之类的附加资源,并从文本中确定一个或多个附加答案话语树。自主代理应用102确定附加答案话语树之一在缺失的问

题实体与问题或答案中的另一个实体之间建立了修辞链接,从而将特定的附加答案话语树指定为虚构的话语树。然后,自主代理应用102可以从由初始答案话语树和虚构的话语树表示的文本中形成完整答案。自主代理应用102提供答案作为输出答案150。例如,自主代理应用102输出文本“*There is no need for gasoline*(无需加油)”。

[0063] 在一些情况下,修辞一致性应用112可以与自主代理应用102协同工作,以促进输入问题130与输出答案150之间的改进的修辞一致性。例如,修辞一致性应用112可以通过使用一个或多个交流话语树114来验证一个或多个话语树或其部分之间的修辞一致性的程度。通过使用交流话语树,可以准确地建模问答与答案之间的修辞一致性和交流动作。例如,修辞一致性应用112可以验证输入问题和输出答案维持修辞一致性,从而不仅确保完整的响应性答案,而且还确保与该问题的风格一致的答案。

[0064] 例如,修辞一致性应用112可以创建表示输入问题130的问题交流话语树和用于每个候选答案的附加的交流话语树。修辞一致性应用112从候选答案中确定最合适的答案。可以使用不同的方法。在一方面,修辞一致性应用112可以为每个候选答案创建候选答案交流话语树,并将问题交流话语树与每个候选话语树进行比较。修辞一致性应用112然后可以识别问题交流话语树与候选答案交流话语树之间的最佳匹配。

[0065] 在另一个示例中,修辞一致性应用112为每个候选答案创建包括问题130和候选答案的问题-答案对。修辞一致性应用112将问题-答案对提供给预测模型,诸如修辞一致性分类器120。修辞一致性应用112使用经训练的修辞一致性分类器120确定问题-答案对是否高于阈值匹配程度,例如,指示答案是否解决了问题。如果否,那么修辞一致性应用112继续分析包括问题和不同答案的附加对,直到找到合适的答案为止。在一些情况下,修辞一致性应用112可以用训练数据125来训练修辞一致性分类器120。

[0066] 修辞结构理论与话语树

[0067] 语言学是语言的科学研究。例如,语言学可以包括句子的结构(句法),例如,主语-动词-宾语;句子的含义(语义),例如,狗咬人与人咬狗;以及说话者在会话中做什么,即,句子之外的语言分析或话语分析。

[0068] 话语的理论基础——修辞结构理论(RST)可以归功于Mann,William和Thompson,Sandra,“*Rhetorical structure theory:A Theory of Text organization*”,*Text-Interdisciplinary Journal for the Study of Discourse*,8(3):243-281,1988。类似于编程语言理论的语法和语义如何帮助实现现代软件编译器,RST帮助实现了话语的分析。更具体而言,RST在至少两个层次上放置结构块,诸如核性和修辞关系的第一层次,以及结构或模式的第二层次。话语解析器或其它计算机软件可以将文本解析成话语树。

[0069] 修辞结构理论对文本的逻辑组织进行建模,文本的逻辑组织是由作者采用的结构,它依赖于文本的各部分之间的关系。RST通过经由话语树形成文本的分层连接结构来模拟文本连贯性。修辞关系被分为协调(coordinate)和从属(subordinate)类别;这些关系跨两个或更多个文本区段保持,因此实现了连贯性。这些文本区段被称为基本话语单元(EDU)。句子中的子句和文本中的句子由作者逻辑连接。给定句子的含义与之前和之后的句子的含义相关。子句之间的这种逻辑关系被称为文本的连贯结构。RST是最流行的话语理论之一,它基于树状的话语结构——话语树(DT)。DT的叶子对应于EDU,即连续的原子文本区段。相邻的EDU通过连贯关系(例如,归因、顺序)相连,从而形成更高级别的话语单元。然后,

这些单元也受此关系链接的约束。通过关系而链接的EDU然后基于其相对重要性来区分:核心是关系的内核部分,而卫星是外围部分。如所讨论的,为了确定准确的请求-响应对,主题和修辞一致性两者都被分析。当说话者回答问题,诸如短语或句子时,说话者的答案应解决该问题的主题。在经由消息的种子文本隐式提出问题的情况下,不仅维持主题而且还与该种子的广义认知状态匹配的适当答案是期望的。

[0070] 修辞关系

[0071] 如所讨论的,本文描述的方面使用虚构的话语树。修辞关系可以用不同的方式描述。例如,Mann和Thompson描述了23种可能的关系。参见William C.Mann,William&Thompson,Sandra(1987) (“Mann and Thompson”)。Maite “Rhetorical structure theory: A Theory of Text organization,”Text-Interdisciplinary Journal for the Study of Discourse,8(3):243-281,1988。其它数量的关系也是可能的。

[0072]

关系名称	核心	卫星
对立	作者赞成的想法	作者不赞成的想法
背景	其理解被促进的文本	用于促进理解的文字
环境	表示在解释性上下文中	情况或时间的解释性上下文

[0073]

	发生的事件或想法的文本	
让步	作者确认的情况	明显不一致但作者也确认的情况
条件	其出现是由于条件情况的出现而导致的动作或情况	条件情况
阐述	基本信息	附加信息
使能	动作	旨在帮助读者执行动作的信息
评价	情况	关于情况的评价性评论
证据	声明	旨在增加读者对声明的信任的信息
解释	情况	对该情况的解释
证明	文本	支持作者表达该文本的权利的信息
动机	动作	旨在增加读者执行该动作的意愿的信息
非自愿原因	情况	导致该情况的另一种情况，但不是任何人的故意动作
非自愿结果	情况	由该情况引起的，而不是任何人的故意动作引起的另一种情况
否则（反条件）	由于条件情况没有出现而导致其出现的行动或情况	条件情况
目的	预期的情况	该情况背后的意图
重述	情况	该情况的重新表达
解决方案	支持需求的全部或部分满足的情况或方法	疑问、请求、问题或其它表达的需求
摘要	文本	该文本的简短摘要
自愿原因	情况	由于某人的故意动作，导致该情况的另一种情况
自愿结果	情况	由于某人的故意动作，由该情况导致的另一种情况

[0074] 一些经验研究假设大多数文本是使用核心-卫星关系构造的。参见Mann和Thompson 1988。但是其它关系并没有承载对核心的明确选择。此类关系的示例如下所示。

[0075]	关系名称	区段	其它区段
	对比	一个替代物	另一个替代物
	联合	(不受约束)	(不受约束)
	列表	项目	下一项目
	序列	项目	下一项目

[0076] 图2描绘了根据一方面的话语树的示例。图2包括话语树200。话语树包括文本区段201、文本区段202、文本区段203、关系210和关系228。图2中的数字对应于三个文本区段。图3对应于具有编号为1、2、3的三个文本区段的以下示例文本：

[0077] 1.Honolulu,Hawaii will be site of the 2017Conference on Hawaiian History(夏威夷火奴鲁鲁将成为2017年夏威夷历史大会的举办地)

[0078] 2.It is expected that 200historians from the U.S.and Asia will attend(预计来自美国和亚洲的200名历史学家将参加)

[0079] 3.The conference will be concerned with how the Polynesians sailed to Hawaii(会议将关注波利尼西亚人如何航行到夏威夷)

[0080] 例如,关系210或阐述描述了文本区段201和文本区段202之间的相互关系。关系228描绘了文本区段203和204之间的相互关系,即阐述。如所绘出的,文本区段202和203进一步阐述文本区段201。在上面的示例中,给定向读者通知会议的目标,文本区段1是核心。文本区段2和3提供关于会议的更多详细信息。在图2中,水平数字(例如,1-3、1、2、3)覆盖文本的(可能由进一步的区段组成的)区段;垂直线指示一个或多个核心;并且曲线表示修辞关系(阐述)并且箭头的方向从卫星指向核心。如果文本区段仅用作卫星而不用作核心,则删除该卫星仍将留下连贯的文本。如果从图2中删除核心,那么文本区段2和3将难以理解。

[0081] 图3描绘了根据一方面的话语树的另一个示例。图3包括组成部分301和302、文本区段305-307、关系310和关系328。关系310描绘了组成部分306和305以及307和305之间的相互关系310——使能。图3涉及以下文本区段：

[0082] 1.The new Tech Report abstracts are now in the journal area of the library near the abridged dictionary.(现在新的技术报告摘要在图书馆的靠近删减词典的期刊区域中)。

[0083] 2.Please sign your name by any means that you would be interested in seeing.(请以您有兴趣看到的任何方式签署您的名字)。

[0084] 3.Last day for sign-ups is 31May.(签署的最后一天是5月31日)。

[0085] 如可以看到的,关系328描绘了实体307和306之间的相互关系(该相互关系是使能)。图3图示了虽然核心可以被嵌套,但是仅存在一个最核心的文本区段。

[0086] 构造话语树

[0087] 话语树可以使用不同的方法生成。自底向上构造DT的方法的简单示例是：

[0088] (1) 通过以下方式将话语文本划分为单元：

[0089] (a) 单元大小可以取决于分析的目标而不同

[0090] (b) 典型地,单元是子句

[0091] (2) 检查每个单元及其邻居。它们之间保持着关系吗？

[0092] (3) 如果是,则标记该关系。

[0093] (4) 如果不是,则该单元可能位于更高级别关系的边界处。查看较大单元(区段)之间保持的关系。

[0094] (5) 继续,直到考虑了文本中的所有单元为止。

[0095] Mann和Thompson还描述了第二级别的构建块结构,称为模式应用。在RST中,修辞关系不直接映射到文本上;它们被装配到称为模式应用的结构上,并且这些结构又被装配到文本。模式应用是从被称为模式的更简单的结构得到的(如图4所示)。每个模式指示如何将文本的特定单元分解成其它更小的文本单元。修辞结构树或DT是模式应用的分层系统。模式应用链接多个连续的文本区段,并创建复杂的文本区段,该复杂的文本区段又可以由更高级别的模式应用链接。RST断言,每个连贯话语的结构都可以由单个修辞结构树来描述,该单个修辞结构树的顶部模式创建涵盖整个话语的区段。

[0096] 图4描绘了根据一方面的说明性模式。图4示出了联合模式是由核心组成的项目列表而没有卫星。图4描绘了模式401-406。模式401描绘了文本区段410和428之间的环境关系。模式402描绘了文本区段420和421之间的序列关系以及文本区段421和422之间的序列关系。模式403描绘了文本区段430和431之间的对比关系。模式404描绘了文本区段440和441之间的联合相互关系。模式405描绘了450和451之间的动机相互关系,以及452和451之间的使能相互关系。模式406描绘了文本区段460和462之间的联合相互关系。图4中示出了以下三个文本区段的联合模式的示例:

[0097] 1.Skies will be partly sunny in the New York metropolitan area today.
(今天纽约都会区的天空将部分晴朗。)

[0098] 2.It will be more humid,with temperatures in the middle 80's.(天气会更加潮湿,温度在80中部)

[0099] 3.Tonight will be mostly cloudy,with the low temperature between 65and 70.(今晚将大部分多云,低温在65至70之间。)

[0100] 虽然图2-图4描绘了话语树的一些图形表示,但其它表示也是可能的。

[0101] 图5描绘了根据一方面的分层二叉树的节点链接表示。如从图6中可以看到,DT的叶子对应于称为基本话语单元(EDU)的连续非重叠文本区段。相邻的EDU通过关系(例如,阐述、归因...)连接并形成更大的话语单元,这些更大的话语单元也通过关系连接。“RST中的话语分析涉及两个子任务:话语分割是识别EDU的任务,并且话语解析是将话语单元链接到带标记的树中的任务”。

[0102] 图5描绘了作为树上的叶子或终端节点的文本区段,每个按它们在全文本中出现的顺序编号,如图6所示。图5包括树500。树500包括例如节点501-507。节点指示相互关系。节点是非终端节点(诸如节点501),或者是终端节点(诸如节点502-507)。如可以看到的,节点503和504通过联合相互关系而相关。节点502、505、506和508是核心。虚线指示分支或文本区段是卫星。关系是灰色框中的节点。图6描绘了根据一方面的图5中的表示的示例性缩进文本编码。图6包括更适合计算机编程的文本600。“N”是核心,并且“S”是卫星。

[0103] 话语解析器的示例

[0104] 自动话语分割可以用不同的方法来执行。例如,给定一个句子,分割模型通过预测是否应在句子中的每个特定符号(token)之前插入边界来识别复合基本话语单元的边界。例如,一个框架顺序且独立地考虑句子中的每个符号。在这个框架中,分割模型逐个符号地

扫描句子,并使用二进制分类器(诸如支持向量机或逻辑回归)来预测在正被检查的符号之前插入边界是否是合适的。在另一个示例中,任务是顺序标记问题。一旦将文本分割为基本话语单元,就可以执行句子级别的话语解析,以构造话语树。可以使用机器学习技术。

[0105] 在本发明的一方面,使用了两个修辞结构理论(RST)话语解析器:依赖于组成语法的CoreNLPProcessor和使用依赖性语法的FastNLPProcessor。在Mihai Surdeanu,Thomas Hicks和Marco A.Valenzuela-Escarcega,“Two Practical Rhetorical Structure Theory Parsers”,Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies: Software Demonstrations(NAAACL HLT)2015中对它们进行了描述。

[0106] 此外,以上两个话语解析器,即,CoreNLPProcessor和FastNLPProcessor使用自然语言处理(NLP)进行句法解析。例如,Stanford CoreNLP给出了词语的基本形式、它们的词性、它们是否是公司、人员的名称等;对日期、时间和数值量进行规范化;根据短语和句法依赖关系标记句子的结构;指示哪些名词短语指代相同的实体。实际上,RST仍然是在许多话语的情况下可能起作用但在一些情况下可能不起作用的理论。存在许多变量,包括但不限于,连贯文本中有哪些EDU,即,使用了什么话语分割器、使用了哪些关系清单以及为EDU选择了哪些关系、用于训练和测试的文档的语料库,以及甚至使用了什么解析器。因此,例如,在上面引用的Surdeanu等人的“Two Practical Rhetorical Structure Theory Parsers”中,必须使用专用的度量对特定的语料库进行测试,以确定哪个解析器给出更好的性能。因此,与给出可预测结果的计算机语言解析器不同,话语解析器(和分段器)可能取决于训练和/或测试文本语料库而给出不可预测的结果。因此,话语树是可预测领域(例如,编译器)和不可预测领域(例如,像化学一样需要进行实验来确定哪些组合将给你期望结果)的混合。

[0107] 为了客观地确定话语分析的质量,正在使用一系列度量,例如来自Daniel Marcu,“The Theory and Practice of Discourse Parsing and Summarization”,MIT Press,2000年11月,ISBN:9780262123722的精确度(Precision)/查全率(Recall)/F1度量。精确度或正预测值是检索到的实例当中相关实例所占的比例,而查全率(也称为敏感度)是在相关实例的总数内已被检索的相关实例所占的比例。因此,精确度和查全率两者都基于对相关性的理解和量度。假设用于识别照片中的狗的计算机程序在包含12条狗和一些猫的图片中识别出8条狗。在识别出的八条狗中,五条实际上是狗(真阳性),而其余的是猫(假阳性)。该程序的精确度为 $5/8$,而它的查全率为 $5/12$ 。当搜索引擎返回30页,其中只有20页是相关的,而未能返回40个附加的相关页时,其精确度为 $20/30 = 2/3$,而其查全率为 $20/60 = 1/3$ 。因此,在这种情况下,精确度是“搜索结果的有用程度”,并且查全率是“结果的完整程度”。F1分数(也称为F-分数或F-量度)是测试的准确性的量度。它考虑了测试的精确度和查全率二者来计算分数: $F1 = 2 \times ((\text{精确度} \times \text{查全率}) / (\text{精确度} + \text{查全率}))$ 并且是精确度和查全率的调和平均值。F1分数为1时达到其最佳值(完美的精确度和查全率),为0时是最差值。

[0108] 自主代理或聊天机器人

[0109] 人类A和人类B之间的会话是一种话语的形式。例如,存在诸如**FaceBook®** Messenger、**WhatsApp®**、**Slack®**、SMS等之类的应用,除了较传统的电子邮件和语音

会话之外,A和B之间的会话典型地还可以经由消息进行。聊天机器人(也可以称为智能机器人或虚拟助手等)是例如取代人类B并在各种程度上模仿两个人类之间的会话的“智能”机器。示例最终目标是,人类A不能分辨出B是人类还是机器(由Alan Turing在1950年开发的图灵测试)。话语分析、包括机器学习在内的人工智能、以及自然语言处理已在通过图灵测试的长期目标方面取得了很大的进步。当然,随着计算机越来越具有搜索和处理大量数据储存库以及对数据执行复杂分析以包括预测分析的能力,长期目标是使聊天机器人像人一样并且与计算机组合。

[0110] 例如,用户可以通过会话交互与智能机器人平台交互。也称为会话用户界面(UI)的这种交互是最终用户和聊天机器人之间的对话,就像两个人类之间的对话一样。它可能很简单,就像最终用户对聊天机器人说“Hello”,然后聊天机器人以“Hi”响应并询问用户它能怎样提供帮助,或者可以是银行聊天机器人中的交易交互,诸如将资金从一个账户转移到另一个账户,或者在HR聊天机器人中的信息交互,诸如检查剩余假期,或者在零售聊天机器人中的询问FAQ,诸如如何处理退货。自然语言处理(NLP)和机器学习(ML)算法与其它方法的结合可以用于对最终用户意图进行分类。高级别的意图是最终用户想要完成的目标(例如,获得账户余额、进行购买)。意图本质上是客户输入到后端应执行的工作单元的映射。因此,基于用户在聊天机器人中说出的短语,将它们映射到特定且离散的用例或工作单元,例如,检查余额、转移资金和跟踪支出都是聊天机器人应支持的“用例”,并能够从最终用户以自然语言键入的自由文本条目中推断出应触发哪个工作单元。

[0111] 使AI聊天机器人像人类一样做出响应的基本原则是人类大脑可以制定和理解请求,然后对人类请求给出比机器好得多的良好响应。因此,如果模仿人类B,那么聊天机器人的请求/响应应该有显著改善。因此,问题的最初部分是人类大脑如何制定和理解请求?为了模仿,使用模型。RST和DT允许以正式且可重复的方式完成这一点。

[0112] 在高级别上,通常有两种类型的请求:(1)执行某个动作的请求;(2)对信息的请求,例如,问题。第一种类型具有其中创建工作单元的响应。第二种类型具有例如对该问题的良好答案的响应。例如,在一些方面,答案可以采取以下形式:AI从其广泛的(一个或多个)知识库中构造答案,或者通过搜索互联网或内联网或其它公共/私有可用数据源来匹配最佳现有答案。

[0113] 分析请求和响应对

[0114] 图7描绘了根据一方面的关于财产税的示例请求的示例性DT。节点标签是关系并且箭头线指向卫星。核心是实线。图7描绘了以下文本。

[0115] 请求:“My husbands’ grandmother gave him his grandfather’s truck. She signed the title over but due to my husband having unpaid fines on his license, he was not able to get the truck put in his name. I wanted to put in my name and paid the property tax and got insurance for the truck. By the time it came to sending off the title and getting the tag, I didn’t have the money to do so. Now, due to circumstances, I am not going to be able to afford the truck. I went to the insurance place and was refused a refund. I am just wondering that since I am not going to have a tag on this truck, is it possible to get the property tax refunded?”

[0116] 响应:“The property tax is assessed on property that you own. Just because you chose to not register it does not mean that you don't own it, so the tax is not refundable. Even if you have not titled the vehicle yet, you still own it within the boundaries of the tax district, so the tax is payable. Note that all states give you a limited amount of time to transfer title and pay the use tax. If you apply late, there will be penalties on top of the normal taxes and fees. You don't need to register it at the same time, but you absolutely need to title it within the period of time stipulated in state law.”

[0117] 分析上面的文本得出:“My husbands' grandmother gave him his grandfather's truck”通过“She signed the title over but due to my husband”进行阐述,后者又通过“having unpaid fines on his license, he was not able to get the truck put in his name”进行阐述,这又通过“I wanted to put in my name”、“and paid the property tax”和“and got insurance for the truck”进行阐述。

[0118] “My husbands' grand mother gave him his grandfather's truck. She signed the title over but due to my husband having unpaid fines on his license, he was not able to get the truck put in his name. I wanted to put in my name and paid the property tax and got insurance for the truck.”通过以下进行阐述:

[0119] “I didn't have the money”,这又通过“to do so”进行阐述,后者与“By the time”对比,“By the time”通过“it came to sending off the title”“and getting the tag”进行阐述,

[0120] “My husbands' grandmother gave him his grandfather's truck. She signed the title over but due to my husband having unpaid fines on his license, he was not able to get the truck put in his name. I wanted to put in my name and paid the property tax and got insurance for the truck. By the time it came to sending off the title and getting the tag, I didn't have the money to do so”与以下形成对比:

[0121] “Now, due to circumstances”,其通过“I am not going to be able to afford the truck”进行阐述,后者通过以下阐述:

[0122] “I went to the insurance place”

[0123] “and was refused a refund”。

[0124] “My husbands' grandmother gave him his grandfather's truck. She signed the title over but due to my husband having unpaid fines on his license, he was not able to get the truck put in his name. I wanted to put in my name and paid the property tax and got insurance for the truck. By the time it came to sending off the title and getting the tag, I didn't have the money to do so. Now, due to circumstances, I am not going to be able to afford the truck. I went to the insurance place and was refused a refund”通过以下阐述:

[0125] “I am just wondering that since I am not going to have a tag on this

truck, is it possible to get the property tax refunded?”

[0126] “I am just wondering”归因于:

[0127] “that”与“is it possible to get the property tax refunded?”是相同的单元,后者的条件为“since I am not going to have a tag on this truck”。

[0128] 如可以看到的,该主题的主要主体是“Property tax on a car(关于汽车的财产税)”。问题包括矛盾:一方面,所有财产都应交税,另一方面,所有权有些不完整。一个良好的响应必须既解决问题的主题又澄清不一致性。为了做到这一点,响应者就无论财产的注册状态如何,拥有的无论任何财产都要纳税的必要性做出更强烈的声明。这个示例是我们的Yahoo!Answers评估领域中正面训练集的成员。主题的主要主体是“Property tax on a car”。问题包括矛盾:一方面,所有财产都应交税,另一方面,所有权有些不完整。良好的答案/响应必须既解决问题的主题又澄清不一致性。读者可以观察到,由于问题包括对比的修辞关系,因此答案必须将它与类似的关系匹配才能令人信服。否则,即使对于那些不是领域专家的人来说,这个答案看起来也不完整。

[0129] 图8描绘了根据本发明的某些方面的对图7中表示的问题的示例性响应。中心核心是“the property tax is assessed on property”,其通过“that you own”进行阐述。“The property tax is assessed on property that you own”也是核心,其通过“Just because you chose to not register it does not mean that you don't own it, so the tax is not refundable. Even if you have not titled the vehicle yet, you still own it within the boundaries of the tax district, so the tax is payable. Note that all states give you a limited amount of time to transfer title and pay the use tax.”来阐述。

[0130] 核心:“The property tax is assessed on property that you own. Just because you chose to not register it does not mean that you don't own it, so the tax is not refundable. Even if you have not titled the vehicle yet, you still own it within the boundaries of the tax district, so the tax is payable. Note that all states give you a limited amount of time to transfer title and pay the use tax.”通过以下阐述:“there will be penalties on top of the normal taxes and fees”,其具有条件:“If you apply late”,这又通过“but you absolutely need to title it within the period of time stipulated in state law”和“You don't need to register it at the same time”的对比来阐述。

[0131] 通过比较图7的DT和图8的DT,可以确定响应(图8)与请求(图7)的匹配程度。在本发明的一些方面,以上框架至少部分地用于确定用于请求/响应的DT以及DT之间的修辞一致性。

[0132] 图9图示了根据一方面的用于官方答案的话语树。如图9所绘出的,官方答案或使命陈述指出:“The Investigative Committee of the Russian Federation is the main federal investigating authority which operates as Russia's Anti-corruption agency and has statutory responsibility for inspecting the police forces, combating police corruption and police misconduct, is responsible for conducting investigations into local authorities and federal governmental

bodies”。

[0133] 图10图示了根据一方面的用于原始答案的话语树。如图10所绘出的,另一个也许更为诚实的答案指出:“Investigative Committee of the Russian Federation is supposed to fight corruption.However,top-rank officers of the Investigative Committee of the Russian Federation are charged with creation of a criminal community.Not only that,but their involvement in large bribes,money laundering,obstruction of justice,abuse of power,extortion,and racketeering has been reported.Due to the activities of these officers,dozens of high-profile cases including the ones against criminal lords had been ultimately ruined.”

[0134] 答案的选择取决于上下文。修辞结构允许区分“官方”、“政治上正确”、基于模板的答案和“实际”、“原始”、“来自现场的报告”或“争议性”答案。(参见图9和图10)。有时,问题本身可以给出关于预期的答案的类别的提示。如果问题被表述为事实或定义性问题,而没有第二含义,那么第一类别的答案是合适的。否则,如果问题具有“告诉我实际上是什么”的含义,那么第二类别是合适的。一般而言,在从问题中提取修辞结构之后,选择具有相似、匹配或互补的修辞结构的合适答案更容易。

[0135] 官方答案基于阐述和联合,这在文本可能包含的争议方面是中立的(参见图9)。同时,原始答案包括对比关系。这种关系在期望代理做的和发现该代理已经做的短语之间被提取。

[0136] 请求-响应对的分类

[0137] 修辞一致性应用112可以确定给定的答案或响应(诸如从答案数据库105或公共数据库获得的答案)是否响应给定的问题或请求。更具体而言,修辞一致性应用112通过确定请求和响应之间的(i)相关性或(ii)修辞一致性中的一个或两个来分析请求和响应对正确还是不正确。修辞一致性可以在不考虑相关性的情况下被分析,而相关性可以被正交处理。

[0138] 修辞一致性应用112可以使用不同的方法来确定问题-答案对之间的相似性。例如,修辞一致性应用112可以确定单独的问题和单独的答案之间的相似性程度。替代地,修辞一致性应用112可以确定包括问题和答案的第一对与包括问题和答案的第二对之间的相似性的量度。

[0139] 例如,修辞一致性应用112使用经训练以预测匹配或不匹配答案的修辞一致性分类器120。修辞一致性应用112可以一次处理两个对,例如,<q1,a1>和<q2,a2>。修辞一致性应用112将q1与q2以及a1与a1进行比较,从而产生组合的相似性分数。这样的比较允许通过评估与具有已知标签的另一个问题/答案对的距离来确定未知问题/答案对是否包含正确答案。特别地,可以对未标记的对<q2,a2>进行处理,使得不是基于q2和a2共享的词语或结构来“猜测”正确性,而是可以基于这样的词语或结构将q2和a2两者与已标记的对<q2,a2>的它们的对应组成部分q1和a2进行比较。由于该方法针对答案的领域无关的分类,因此只能利用问题与答案之间的结构内聚性,而不能利用答案的“含义”。

[0140] 在一方面,修辞一致性应用112使用训练数据125来训练修辞一致性分类器120。以这种方式,修辞一致性分类器120被训练以确定问题和答案对之间的相似性。这是分类问题。训练数据125可以包括肯定训练集和否定训练集。训练数据125包括在肯定数据集中匹

配的请求-响应对和在否定数据集中任意的或较低相关性或适当性的请求-响应对。对于肯定数据集,选择具有指示答案或响应是否适合该问题的不同接受标准的各种领域。

[0141] 每个训练数据集包括训练对集合。每个训练集包括表示问题的问题交流话语树和表示答案的答案交流话语树以及问题与答案之间的预期互补程度。通过使用迭代处理,修辞一致性应用112向修辞一致性分类器120提供训练对,并从模型接收互补程度。修辞一致性应用112通过确定针对特定训练对的所确定的互补程度与预期的互补程度之间的差来计算损失函数。基于损失函数,修辞一致性应用112调整分类模型的内部参数以最小化损失函数。

[0142] 接受标准可能因应用而异。例如,对于社区问题回答、自动化的问题回答、自动化的和手动的客户支持系统、社交网络交流以及诸如消费者之类的个人关于其对产品的体验的书写内容(诸如评论和投诉),接受标准可能低。在以FAQ、专业社交网络(诸如“stackoverflow”)形式的科学文本、专业新闻、卫生和法律文件中,RR接受标准可能高。

[0143] 交流话语树(CDT)

[0144] 修辞一致性应用112可以创建、分析和比较交流话语树。交流话语树被设计为将修辞信息与言语行为结构组合。CDT包含弧线,该弧线标记有用于交流动作的表达式。通过组合交流动作,CDT实现对RST关系和交流动作的建模。CDT是解析丛林的简化。参见Galitsky, B, Ilvovsky, D.和Kuznetsov S O. Rhetoric Map of an Answer to Compound Queries Knowledge Trail Inc. ACL 2015, 681-686. (“Galitsky 2015”)。解析丛林是句子的解析树的组合,在一个图中这些句子的词语和部分之间具有话语级别相互关系。通过合并识别言语动作的标签,交流话语树的学习可以在比仅基本话语单元(EDU)的修辞关系和语法更丰富的特征集上发生。

[0145] 在示例中,分析了三方之间关于商业客机(即马来西亚航空航班17)坠机原因的争议。建立了所传达的论点的RST表示。在该示例中,三个有冲突的代理,即,荷兰调查员、俄罗斯联邦调查委员会和自称的顿涅茨克人民共和国,就此事交换了他们的意见。该示例图示了争议性冲突,其中各方都尽其所能责怪其对方。为了听起来更有说服力,各方不仅提出自己的声明,而且以拒绝对方声明的方式作出回应。为了实现这一目标,各方都试图匹配对方的声明的风格和话语。

[0146] 图11图示了根据一方面的用于第一代理的声明的交流话语树。图11描绘了交流话语树100,该树表示以下文本:“Dutch accident investigators say that evidence points to pro-Russian rebels as being responsible for shooting down plane. The report indicates where the missile was fired from and identifies who was in control of the territory and pins the downing of MH17 on the pro-Russian rebels”。

[0147] 如从图11可以看到的,CDT的非终端节点是修辞关系,并且终端节点是作为这些关系的主体的基本话语单元(短语、句子片段)。CDT的某些弧线标记有用于交流动作的表达式,包括行动者代理和这些动作的主体(正在传达的内容)。例如,用于阐述关系的核心节点(在左侧)被标记有say (Dutch, evidence),并且卫星被标记有responsible (rebels, shooting down)。这些标签并不旨在表达EDU的主体是evidence和shooting down,而是为了使该CDT与其它CDT匹配,以便找到它们之间的相似性。在这种情况下,仅通过修辞关系将

这些交流动作链接起来而不提供交流话语的信息将太有限,而不能表示所传达的内容和传达的方式的结构。RR对要具有相同或协调的修辞关系的要求太弱,因此需要匹配的节点的顶部上的弧线的CDT标签的一致性。

[0148] 该图的直边是句法关系,并且弯曲的弧线是话语关系,诸如回指(anaphora)、相同实体、子实体、修辞关系和交流动作。该图包含比仅仅用于各个句子的解析树的组合丰富得多的信息。除了CDT之外,解析丛林还可以在词语、关系、短语和句子的级别进行泛化(generalization)。言语动作是表达各个言语行为及其主体所涉及的施事者的逻辑谓词。如由诸如VerbNet之类的框架所提出的,逻辑谓词的论元根据各自的语义角色形成。参见Karin Kipper、Anna Korhonen、Neville Ryant、Martha Palmer, A Large-scale Classification of English Verbs, Language Resources and Evaluation Journal, 42 (1), pp.21-40, Springer Netherland, 2008和/或Karin Kipper Schuler, Anna Korhonen, Susan W. Brown, VerbNet overview, extensions, mappings and apps, Tutorial, NAACL-HLT 2009, Boulder, Colorado。

[0149] 图12图示了根据一方面的用于第二代理的声明的交流话语树。图12描绘了交流话语树1200,该树表示以下文本:“The Investigative Committee of the Russian Federation believes that the plane was hit by a missile, which was not produced in Russia. The committee cites an investigation that established the type of the missile”。

[0150] 图13图示了根据一方面的用于第三代理的声明的交流话语树。图13描绘了交流话语树1300,该树表示以下文本:“Rebels, the self-proclaimed Donetsk People’s Republic, deny that they controlled the territory from which the missile was allegedly fired. It became possible only after three months after the tragedy to say if rebels controlled one or another town”。

[0151] 如从交流话语树1100-1300可以看到的,响应不是任意的。响应该及与原始文本相同的实体。例如,交流话语树1200和1300与交流话语树1100相关。响应支持关于这些实体以及关于这些实体的动作的估计和观点的不一致。

[0152] 更具体而言,所涉及代理的答复需要反映第一种子消息的交流话语。作为简单的观察,由于第一代理使用归因来传达其声明,因此其它代理必须遵循这套声明,或者提供其自己的归因或者攻击支持者的归因的有效性,或者两者兼而有之。为了捕获各种各样的特征以用于需要如何在连续消息中保留种子消息的交流结构,可以学习相应CDT的对。

[0153] 为了验证请求-响应的一致性,仅话语关系或言语行为(交流动作)通常是不够的。如从图11-图13所绘出的示例可以看到的,代理之间的交互的话语结构和交互的种类是有用的。但是,交互的领域(例如,军事冲突或政治)或这些交互的主体(即,实体),不需要被分析。

[0154] 表示修辞关系和交流动作

[0155] 为了计算抽象结构之间的相似性,频繁使用两种方法:(1)在数值空间中表示这些结构,并将相似性表达为数字,这是统计学习方法,或者(2)使用结构表示,而无需数值空间,诸如树和图形,并且将相似性表达为最大公共子结构。将相似性表达为最大公共子结构被称为泛化。

[0156] 学习交流动作有助于表达和理解论元。计算动词词典帮助支持获取动作的实体,并提供基于规则的形式来表达其含义。动词表达所描述事件的语义以及该事件中的参与者之间的关系信息,并投射对该信息进行编码的语法结构。动词,尤其是交流动作动词,可能变化很大并且可能展现丰富的语义行为。作为响应,动词分类通过将动词组织到共享核心语义属性的组中来帮助学习系统应对这种复杂性。

[0157] VerbNet是一种这样的词典,它识别每个类中动词的语义角色和句法模式特性,并使句法模式与可以为该类的所有成员推断出的基础语义关系之间的连接明确。参见Karin Kipper,Anna Korhonen,Neville Ryant和Martha Palmer,Language Resources and Evaluation,Vol.42,No.1(2008年3月),21。类的每个句法框架或动词签名具有对应的语义表示,其详细描述了事件的过程中事件参与者之间的语义关系。

[0158] 例如,动词amuse(娱乐)是类似动词的聚类的一部分,这些动词具有类似的论元(语义角色)的结构,诸如amaze、anger、arouse、disturb和irritate(惊奇、生气、激发、打扰和烦躁)。这些交流动作的论元的角色如下:Experience(体验者,通常为有生命实体)、Stimulus(刺激)和Result(结果)。每个动词可以具有通过用于该动词在句子或框架中如何出现的句法特征区分的含义类别。例如,用于amuse的框架如下,使用以下关键名词短语(NP)、名词(N)、交流动作(V)、动词短语(VP)、副词(ADV):

[0159] NP V NP。示例:“The teacher amused the children”。语法:Stimulus V Experiencer。子句:amuse(Stimulus,E,Emotion,Experiencer),cause(Stimulus,E),emotional_state(result(E),Emotion,Experiencer)。

[0160] NP V ADV-Middle。示例:“Small children amuse quickly”。语法:Experiencer V ADV。子句:amuse(Experiencer,Prop):-,property(Experiencer,Prop),adv(Prop)。

[0161] NP V NP-PRO-ARB。示例“The teacher amused”。语法:Stimulus V.amuse(Stimulus,E,Emotion,Experiencer):.cause(Stimulus,E),emotional_state(result(E),Emotion,Experiencer)。

[0162] NP.cause V NP。示例“The teacher’s dolls amused the children”。语法:Stimulus<+genitive>('s)V Experiencer.amuse(Stimulus,E,Emotion,Experiencer):.cause(Stimulus,E),emotional_state(during(E),Emotion,Experiencer)。

[0163] NP V NP ADJ。示例“This performance bored me totally”。语法:Stimulus V Experiencer Result.amuse(Stimulus,E,Emotion,Experiencer).cause(Stimulus,E),emotional_state(result(E),Emotion,Experiencer),Pred(result(E),Experiencer)。

[0164] 交流动作可以被表征为多个聚类,例如:

[0165] 具有谓语补语的动词(appoint,characterize,dub,declare,conjecture,masquerade,orphan,captain,consider,classify),感知动词(see,sight,peer)。

[0166] 心理状态动词(amuse,admire,marvel,appeal),欲望动词(want,long)。

[0167] 判断动词(judgment),评估动词(assess,estimate),搜索动词(hunt,search,stalk,investigate,rummage,ferret),社交互动动词(correspond,marry,meet,battle),交流动词(transfer(message),inquire,interrogate,tell,manner(speaking),talk,chat,say,complain,advise,confess,lecture,overstate,promise)。避免动词(avoid),度量动词(register,cost,fit,price,bill),体貌动词(begin,complete,continue,stop,

establish,sustain)。

[0168] 本文所描述的方面比统计学习模型更具优势。与统计解决方案相比,使用分类系统的方面可以提供动词或类动词结构,该动词或类动词结构被确定为导致目标特征(诸如修辞一致性)。例如,统计机器学习模型将相似性表达为数字,这可能使得难以解释。

[0169] 表示请求-响应对

[0170] 表示请求-响应对基于对来促进基于分类的操作。在示例中,请求-响应对可以被表示为解析丛林。解析丛林是两个或更多个句子的解析树的表示,其中在一个图中句子的部分和词语之间具有话语级别关系。参见Galitsky (2015)。问题和答案之间的主题相似性可以被表达为解析丛林的公共子图。公共图节点的数量越多,相似度越高。

[0171] 图14图示了根据一方面的解析丛林。图14描绘了解析丛林1400,其包括解析树1401和用于对应的响应的解析树1402。

[0172] 解析树1401表示问题:“I just had a baby and it looks more like the husband I had my baby with.However it does not look like me at all and I am scared that he was cheating on me with another lady and I had her kid.This child is the best thing that has ever happened to me and I cannot imagine giving my baby to the real mom”。

[0173] 响应1402表示响应“Marital therapists advise on dealing with a child being born from an affair as follows.One option is for the husband to avoid contact but just have the basic legal and financial commitments.Another option is to have the wife fully involved and have the baby fully integrated into the family just like a child from a previous marriage”。

[0174] 图14表示用于表示关于文本的段落的语言学信息的贪心方法。该图的直边是句法关系,并且弯曲的弧线是话语关系,诸如回指、相同实体、子实体、修辞关系和交流动作。实心弧线用于相同实体/子实体/回指关系,并且虚线弧线用于修辞关系和交流动作。直边中的椭圆形标签表示句法关系。词元(Lemmas)写在节点的框中,并且词元形式写在节点的右侧。

[0175] 解析丛林1400包括比仅仅针对各个句子的解析树的组合丰富得多的信息。沿着句法关系的边以及话语关系的弧线在该图中导航允许将给定的解析丛林变换成语义上等效的形式,以与其它解析丛林匹配,从而执行文本相似性评估任务。为了形成段落的完整形式表示,表达尽可能多的链接。每个话语弧线都产生一对可能是潜在匹配的丛林短语。

[0176] 种子(请求)和响应之间的主题相似性被表达为解析丛林的公共子图。它们被可视化为连接的云。公共图节点的数量越多,相似度越高。对于修辞一致性,公共子图不必像它在给定文本中那样大。但是,种子与响应的修辞关系和交流动作是相关的,并且需要对应关系。

[0177] 交流动作的泛化

[0178] 两个交流动作 A_1 和 A_2 之间的相似性被定义为拥有 A_1 和 A_2 之间公共的特征的抽象动词。将两个动词的相似性定义为抽象的类动词结构支持归纳学习任务,诸如修辞一致性评估。在示例中,可以将以下两个常见动词(agree和disagree)之间的相似性泛化如下: $agree \sim disagree = verb(Interlocutor, Proposed_action, Speaker)$,其中Interlocutor(对话

者)是向Speaker(发言者)提议Proposed_action(提议的动作)、并且Speaker向其传达Speaker的响应的人。Proposed_action是如果Speaker接受或拒绝请求或建议时Speaker将执行的动作,并且Speaker是已向其提议特定动作并对所作出的请求或建议进行响应的人。

[0179] 在另一个示例中,动词agree和explain之间的相似性被表示如下: $\text{agree} \hat{=} \text{explain} = \text{verb}(\text{Interlocutor}, *, \text{Speaker})$ 。交流动作的主体在交流动作的上下文中被泛化,并且不相对于其它“物理”动作被泛化。因此,各方面将交流动作的单独的发生与对应的主体一起泛化。

[0180] 此外,可以将表示对话的交流动作序列与类似对话的其它此类序列进行比较。以这种方式,单独的交流动作的含义以及对话的动态话语结构(与经由修辞关系反映的其静态结构相比)被表示。泛化是发生在每个级别的复合结构表示。交流动作的词元与词元一起泛化,并且其语义角色与相应的语义角色一起泛化。

[0181] 文本作者使用交流动作来指示对话的结构或冲突。参见Searle, J. R. 1969, *Speech acts: an essay in the philosophy of language*. London: Cambridge University Press. 主体在这些动作的上下文中被泛化,并且不相对于其它“物理”动作泛化。因此,交流动作的单独的发生与它们的主体以及它们的对一起被泛化为话语“步骤”。

[0182] 也可以从匹配动词框架(诸如VerbNet)的角度来考虑交流动作的泛化。交流链接反映了与文本中多于一个代理的参与(或提及)相关联的话语结构。这些链接形成连接用于交流动作的词语(隐式指示人的交流意图的多个词语或者动词)的序列。

[0183] 交流动作包括行动者、正在对其采取行动的一个或多个施事者,以及描述该动作的特征的短语。交流动作可以被描述为以下形式的函数: $\text{verb}(\text{agent}, \text{subject}, \text{cause})$ (动词(施事者, 主体, 原因)), 其中动词表征所涉及的施事者之间的某种类型的交互(例如, 解释、确认、提醒、不同意、拒绝等), 主体是指传送的信息或描述的对象, 原因是指用于主体的动机或解释。

[0184] 场景(标记为有向图)是解析丛林 $G = (V, A)$ 的子图, 其中 $V = \{\text{action}_1, \text{action}_2 \dots \text{action}_n\}$ 是与交流动作对应的顶点的有限集合, 并且 A 是带标签的弧线的有限集合(有序的顶点对), 其分类如下:

[0185] 每个弧线 $\text{action}_i, \text{action}_j \in A_{\text{sequence}}$ 对应于引用相同主体(例如, $s_j = s_i$)或不同的主体的两个动作 $v_i, \text{ag}_i, s_i, c_i$ 和 $v_j, \text{ag}_j, s_j, c_j$ 的时间优先级。每个弧线 $\text{action}_i, \text{action}_j \in A_{\text{cause}}$ 对应于动作 action_i 和 action_j 之间的攻击相互关系, 其指示 action_i 的原因与 action_j 的主体或原因冲突。

[0186] 与施事者之间的交互场景相关联的解析丛林的子图具有一些明显的特征。例如, (1) 所有顶点都按时间排序, 使得所有顶点(初始和终止顶点除外)都存在一个传入弧和一个传出弧, (2) 对于 A_{sequence} 弧, 最多一个传入弧且仅一个传出弧是可允许的, 以及 (3) 对于 A_{cause} 弧, 给定顶点可以有多个传出弧, 以及许多传入弧。涉及的顶点可以与不同的施事者或相同的施事者(即, 当他自相矛盾时)相关联。为了计算解析丛林与其交流动作之间的相似性, 归纳子图、具有相似弧标签的相同配置的子图以及顶点的严格对应关系被分析。

[0187] 通过分析解析丛林的交流动作的弧线, 存在以下相似性: (1) 其主体来自T1的一个交流动作相对于其主体来自T2的另一个交流动作(不使用交流动作弧), 和 (2) 其主体来自T1的一对交流动作与来自T2的另一对交流动作进行比较(使用了交流动作弧)。

[0188] 泛化两个不同的交流动作基于它们的属性。参见 (Galitsky等人2013)。如在关于图14所讨论的示例中可以看到,可以将来自T1的一个交流动作cheating (husband,wife, another lady) 与来自T2,的第二交流动作avoid (husband,contact (husband,another lady)) 进行比较。泛化导致communicative_action (husband,*),其引入对A的约束,其形式为:如果在Q中作为CA的主体提到给定的施事者(=husband),则他(她)也应成为A中的(可能的另一个)CA的主体。可以始终泛化两个交流动作,而对于其主体不是这样:如果其泛化结果为空,则具有这些主体的交流动作的泛化结果也为空。

[0189] RST关系的泛化

[0190] 可以泛化话语树之间的一些关系,诸如表示相同类型的关系(表示关系,诸如对照;主题关系,诸如条件;以及多核心关系,诸如列表)的弧线可以被泛化。核心或由核心表示的情况用“N”指示。卫星或由卫星呈现的情况用“S”指示。“W”指示作者。“R”指示读者(听众)。情况是命题(propositions)、完成的动作或进行中的动作,以及交流动作和状态(包括信念、愿望、批准、解释、和解及其它)。具有上述参数的两个RST关系的泛化被表达为: $\text{rst1}(N1,S1,W1,R1) \wedge \text{rst2}(N2,S2,W2,R2) = (\text{rst1} \wedge \text{rst2})(N1 \wedge N2, S1 \wedge S2, W1 \wedge W2, R1 \wedge R2)$ 。

[0191] $N1, S1, W1, R1$ 中的文本作为短语被泛化。例如, $\text{rst1} \wedge \text{rst2}$ 可以被泛化如下:(1) 如果 $\text{relation_type}(\text{rst1}) \neq \text{relation_type}(\text{rst2})$,则泛化为空。(2) 否则,修辞关系的签名被泛化为句子: $\text{sentence}(N1, S1, W1, R1) \wedge \text{sentence}(N2, S2, W2, R2)$ 。参见Iruskieta,Mikel, Iria da Cunha和Maite Taboada.A qualitative comparison method for rhetorical structures:identifying different discourse structures in multilingual corpora.Lang Resources&Evaluation.2015年6月,第49卷,第2期。

[0192] 例如, $\text{rst-background} \wedge \text{rst-enablement}$ 的含义 = (S增加R理解N中的元素的能力) $\wedge$ (R理解S增加R执行N中的动作的能力) = increase-VB the-DT ability-NN of-IN R-NNto-IN。

[0193] 由于关系 $\text{rst-background} \wedge \text{rst-enablement}$ 不同,因此RST关系部分为空。然后,泛化作为相应RST关系的语言定义的表达式。例如,对于每个词语或诸如施事者之类的词语的占位符,如果该词语在每个输入短语中相同,则保留该词语(及其POS),如果该词语在这些短语之间不同,则移除该词语。可以将结果表达式解释为正式获得的两个不同RST关系的定义之间的公共含义。

[0194] 图14中绘出的问题和答案之间的两个弧线显示了基于RST关系“RST-对比(RST-contrast)”的泛化实例。例如,“I just had a baby”是与“it does not look like me”的RST对比,并与“husband to avoid contact”相关,后者是与“have the basic legal and financial commitments”的RST-对比。如可以看到的,答案不必与问题的动词短语相似,但问题和答案的修辞结构是相似的。并非答案中的所有短语都必须与问题中的短语匹配。例如,不匹配的短语与答案中与问题中的短语相关的短语具有某些修辞关系。

[0195] 建立交流话语树

[0196] 图15图示了根据一方面的用于构建交流话语树的示例性处理。修辞一致性应用112可以实现处理1500。如所讨论的,交流话语树实现改进的搜索引擎结果。

[0197] 在框1501处,处理1500涉及访问包括片段的句子。至少一个片段包括动词和词语,并且每个词语包括词语在片段内的角色,并且每个片段是基本话语单元。例如,修辞一致性

应用112访问诸如关于图13所描述的“Rebels, the self-proclaimed Donetsk People’s Republic, deny that they controlled the territory from which the missile was allegedly fired”之类的句子。

[0198] 继续该示例, 修辞一致性应用112确定该句子包括若干个片段。例如, 第一片段是“rebels..deny”。第二片段是“that they controlled the territory”。第三片段是“from which the missile was allegedly fired”。每个片段包含动词, 例如, 第一片段的“deny”和第二片段的“controlled”。但是, 片段不是必需包含动词。

[0199] 在框1502处, 处理1500涉及生成表示句子片段之间的修辞关系的话语树。话语树包括节点, 每个非终端节点表示句子片段中的两个句子片段之间的修辞关系, 并且话语树的节点中的每个终端节点与句子片段中的一个句子片段相关联。

[0200] 继续该示例, 修辞一致性应用112生成如图13所示的话语树。例如, 第三片段“from which the missile was allegedly fired”阐述“that they controlled the territory”。第二和第三片段一起与所发生的事情的归因相关, 即, 攻击不可能是叛乱分子, 因为他们不控制领土。

[0201] 在框1503处, 处理1500涉及访问多个动词签名。例如, 修辞一致性应用112访问例如VerbNet中的动词列表。每个动词与片段中的动词匹配或相关。例如, 对于第一片段, 动词为“deny”。相应地, 修辞一致性应用112访问与该动词deny相关的动词签名的列表。

[0202] 如所讨论的, 每个动词签名包括片段的动词以及一个或多个题元角色。例如, 签名包括名词短语 (NP)、名词 (N)、交流动作 (V)、动词短语 (VP) 或副词 (ADV) 中的一个或多个。题元角色描述动词与相关词语之间的相互关系。例如, “the teacher amused the children”具有与“small children amuse quickly”不同的签名。对于第一片段, 动词“deny”, 修辞一致性应用112访问与“deny”匹配的动词的动词签名或框架的列表。该列表是“NP V NP to be NP”、“NP V that S”和“NP V NP”。

[0203] 每个动词签名包含题元角色。题元角色是指动词在句子片段中的角色。修辞一致性应用112确定每个动词签名中的题元角色。题元角色的示例包括行动者、施事者、资产、属性、受益人、原因、位置目的地源、目的地、源、位置、体验者、范围、仪器、材料和产品、材料、产品、患者、谓词、接收者、刺激、词干、时间或主题。

[0204] 在框1504处, 处理1500涉及针对动词签名中的每个动词签名确定该相应签名的与片段内的词语的角色匹配的题元角色的数量。对于第一片段, 修辞一致性应用112确定动词“deny”仅具有三个角色: “施事者”、“动词”和“词干”。

[0205] 在框1505处, 处理1500涉及基于特定动词签名具有最大数量的匹配从动词签名中选择特定动词签名。例如, 再次参考图13, 第一片段“the rebels deny...that they control the territory”中的deny与动词签名deny“NP V NP”匹配, 并且“control”与control (rebel, territory) 进行匹配。动词签名被嵌套, 从而产生嵌套签名“deny (rebel, control (rebel, territory))”。

[0206] 虚构的话语树

[0207] 本文描述的某些方面使用虚构的话语树来改进针对复杂的、多句子、聚敛性问题的问答式 (Q/A) 召回。通过用从本体获得的树片段扩充答案的话语树, 各方面获得独立于给定作者的思想结构的答案的规范话语表示。

[0208] 如所讨论的,话语树(DT)可以表示文本是如何(特别地利用文本段落)组织的。话语级别分析可以用于其中学习语言结构可能必不可少的许多自然语言处理任务中。DT概述了实体之间的相互关系。存在在文本内引入实体和相关联的属性的不同方式。并非所有这些实体之间存在的修辞关系都出现在给定段落的DT中。例如,在另一个文本主体中,两个实体之间的修辞关系可能在给定段落的DT之外。

[0209] 因此,为了使问题和答案彼此完全相关,期望答案的更完整的话语树。这样的话语树将包括所涉及的实体之间的所有修辞关系。为了实现这一点,利用被识别为答案中缺失的某些修辞关系来扩充至少部分解决问题的答案的初始或最佳匹配话语树。

[0210] 这些缺失的修辞关系可以从文本语料库或从搜索诸如互联网之类的外部数据库获得。因此,不是依赖于可能具有在候选答案中缺失的实体的定义的本体,而是各方面在线挖掘这些实体之间的修辞关系。该过程避免了构建本体的需求,并且可以在常规搜索引擎之上实现。

[0211] 更具体而言,答案A与问题Q相关的基线要求是A的实体(E_n)覆盖Q的实体: $E-Q \subseteq E-A$ 。自然,一些答案实体E-A在Q中没有明确提及,但是需要这些答案实体以提供完整的答案A。因此,可以分析Q通过A的逻辑流程。但是,由于在 E_n 之间建立关系可能具有挑战性,因此可以对Q和A的近似逻辑流程进行建模,该近似逻辑流程可用与领域无关的术语 $EnDT-Q \sim EnDT-A$ 表示,并且以后被验证和/或扩充。

[0212] 在用于回答问题的初始答案中可能存在不同的缺陷。例如,一种情况是一些实体E在Q中未明确提及,而是被假设。另一种情况是,当用于回答Q的A中的一些实体没有出现在A中,而是在A中出现更具体的实体或更一般的实体。为了确定一些更具体的实体确实解决来自Q的问题,可以使用一些外部或附加源(在本文称为虚构的 $EnDT-A$)来建立这些相互关系。这种源包含关于 E_n 之间的内部相互关系的信息,其在Q和/或A中被省略,但被假定为是对等体已知的。因此,对于计算机实现的系统,在话语级别期望以下知识:

[0213] $EnDT-Q \sim EnDT-A + \text{虚构的} EnDT-A$ 。

[0214] 为了讨论的目的,引入了以下示例。第一个示例是问题“What is an advantage of electric car?(电动车的优点是什么)”并且对应的答案是“No need to for gas(无需加油)”。为了建立特定答案适合特定问题,使用广义实体advantage和常规名词实体car。更具体而言,链接了A{need,gas}中的显式实体。示出了可能的虚构的 $EnDT-A$ 的片段:[...No need...-Elaborate-Advantage]...[gas-Enablement-engine]...[engine-Enablement-car]。仅需要存在这些修辞链接的证据。

[0215] 在第二个示例中,使用虚构的话语树来改进搜索。用户向自主代理询问关于“a faulty brake switch can affect the automatic transmission Munro(故障的制动开关可能影响自动变速器Munro)”的信息。现有的搜索引擎可能会识别在给定搜索结果中找不到的某些关键字。但是,可以通过查找其中这些未识别的关键字与查询中出现的关键字在修辞上连接的文档来指示这些关键字与搜索结果如何相关。这一特征自然会一方面提高答案的相关性,并为用户提供关于答案中其关键字如何被解决的可解释性。

[0216] 虚构的话语树可以使搜索引擎能够解释搜索结果中缺失的关键字。在默认搜索中,munro缺失。但是,通过尝试将munro与问题中的实体进行修辞连接,自主代理应用102得知Munro是自动变速器的发明者。图16描绘了可以由自主代理应用102使用的示例处理。

[0217] 图16图示了根据一方面的用于构建虚构的话语树的示例性处理1600。自主代理应用102可以实现处理1600。处理1600解释了话语树(DT)如何促进问题和答案之间的改进匹配。例如,为了验证答案A(或初始答案)适合于给定问题Q,自主代理应用102验证DT-A和DT-Q是否一致,然后可选地用其它DT的片段来扩充DT-A,以确保Q中的所有实体在扩充的DT-A中都得到解决。为了讨论的目的,关于图17讨论了处理1600。

[0218] 图17描绘了根据一方面的问题、答案的示例话语树和两个虚构的话语树。图17描绘了问题话语树1710、答案话语树1720、附加答案话语树1730、附加答案话语树1740、链接1750-1754和节点1170-1174。

[0219] 问题话语树1710由以下问题形成:“[When driving the cruise control][the engine will turn off][when I want to accelerate,][although the check engine light was off.][I have turned on the ignition][and listen for the engine pump running][to see][if it is building up vacuum.][Could there be a problem with the brake sensor under the dash?][Looks like there could be a little play in the plug.]”。为了讨论的目的,方括号[]指示基本话语单元。

[0220] 返回图16,在方框1601处,处理1600涉及从问题构造问题话语树(DT-Q)。问题可以包括片段(或基本话语单元)。词语中的每个词语指示片段内各个词语的角色。

[0221] 自主代理应用102生成问题话语树。问题话语树表示片段之间的修辞关系,并且包含节点。每个非终端节点表示两个片段之间的修辞相互关系。话语树的节点中的每个终端节点都与片段之一相关联。

[0222] 问题话语树可以包括一个或多个问题实体。自主代理应用102可以识别这些实体。如图17所示,问题包含诸如“CRUISE CONTROL(巡航控制)”、“CHECK ENGINE LIGHT(检查引擎指示灯)”、“ENGINE PUMP(引擎泵)”、“VACUUM(真空吸尘器)”、“BRAKE SENSOR(制动传感器)”之类的实体。如后面所讨论的,答案解决了这些实体中的一些实体,而一些实体则没有被解决。

[0223] 在方框1602处,处理1600从文本语料库访问初始答案。可以从现有文档(诸如文件或数据库)中获得初始答案。也可以经由搜索本地或外部系统来获得初始答案。例如,自主代理应用102可以通过提交搜索查询来获得附加答案集合。搜索查询可以从问题中得出,例如,通过问题中识别的一个或多个关键字制定。自主代理应用102可以确定多个答案中的每个答案的相关性分数。答案分数或排名指示问题与相应答案之间的匹配程度。自主代理应用102选择分数高于阈值的答案作为初始答案。

[0224] 继续该示例,自主代理应用102访问以下初始答案“[A faulty brake switch can affect the cruise control.][If it is,][there should be a code][stored in the engine control module.][Since it is not an emissions fault,][the check engine light will not illuminate.][First of all,watch the tachometer][to see][if engine speed increases 200rpm][when this happens.][If it does,][the torque converter is unlocking transmission.]”

[0225] 在方框1603处,处理1600涉及从初始答案中构造包括答案实体的答案话语树。自主代理应用102形成用于初始答案的话语树。在方框1603处,自主代理应用102执行与关于方框1601所描述的类似的步骤。

[0226] 继续该示例,自主代理应用102生成答案话语树1720,其表示作为问题的初始答案的文本。答案话语树1720包括问题实体中的一些,具体而言,“CRUISE CONTROL(巡航控制)”、“CHECK ENGINE LIGHT(检查引擎指示灯)”和“TORQUE CONVERTER(转矩变换器)”。

[0227] 在方框1604处,处理1600涉及使用计算设备确定指示答案实体与问题实体的相关性的分数低于阈值。问题Q中的实体可以(例如,被答案中的一个或多个实体)解决或未被解决,即,答案A中没有实体解决特定实体。问题中未解决的实体由 E_0 指示。

[0228] 可以使用不同的方法来确定特定答案实体与特定问题实体如何相关性。例如,在一方面,自主代理应用102可以确定问题实体的确切文本匹配是否出现在答案实体中。在其它情况下,自主代理应用102可以确定指示问题实体的关键字的比例出现在答案实体中的分数。

[0229] 实体未被解决可能存在不同的原因。如图所示, E_0-Q 包括{ENGINE PUMP、BRAKE SENSOR和VACUUM}。例如,任一答案A与问题Q并不完全相关,因为答案省略了实体 E_0 中的一些。可替代地,相反,答案使用不同的实体。在答案A中可能忽略了 E_0-Q 。为了验证后一种可能性,使用背景知识来找到链接到 E_0-Q 和E-A两者的实体 E_{img} 。

[0230] 例如,可能不清楚 $E-A=TORQUE CONVERTER$ 如何连接到Q。为了验证该连接,获得来自维基百科(Wikipedia)的关于转矩变换器的文本片段,并构建 $DT-A_{img1}$ (1730)。各方面可以确定转矩变换器经由阐述的修辞关系与引擎连接。

[0231] 因此, $E-A=Torque Converter$ 确实与问题相关,如垂直的蓝色弧线所示。无需构建链接实体的离线本体以及学习实体之间的关系就可以做出这样的确定。替代地,使用话语级别上下文来确认A包括相关实体。

[0232] 继续该示例,自主代理应用102确定问题话语树1710中的其它实体,具体而言是那些引用“ENGINE PUMP”、“VACUUM”、“BRAKE SENSOR”和“TORQUE CONVERTER”的实体在答案话语树1720中未被解决。因此,分别通过链接1750和1751图示了E-Q和E-A之间的这些对应关系。

[0233] 在方框1605处,处理1600涉及根据文本语料库创建附加的话语树。附加的话语树是从文本创建的。在一些情况下,自主代理应用102根据评分机制从一个或多个文本中选择适当的文本。

[0234] 例如,自主代理应用102可以访问附加答案集合。自主代理应用102可以生成附加的虚构的话语树集合。每个附加的话语树与相应的附加答案对应。自主代理应用102针对每个附加的话语树计算分数,该分数指示包括到相应的附加话语树中的一个或多个答案实体的映射的问题实体的数量。自主代理应用102从附加的话语树集合中选择具有最高分数的附加话语树。一旦用虚构的 $DT-A_{img}$ 扩充了DT-A,就将搜索相关性测量为未解决的 E_0-Q 的倒数。

[0235] 例如,自主代理应用102获得候选者集合 A_s 。然后,对于 A_s 中的每个候选者 A_c ,自主代理应用102执行以下操作:

[0236] (a) 构建 $DT-A_c$;

[0237] (b) 建立映射 $E-Q \rightarrow E-A_c$;

[0238] (c) 识别 E_0-Q ;

[0239] (d) 从 E_0-Q 和 E_0-A_c (不在 E_0-Q 中的实体)形成查询;

[0240] (e) 从B获得针对查询d)的搜索结果,并构建虚构的DTs-A_c;以及

[0241] (f) 计算剩下的分数 $|E_0|$ 。

[0242] 自主代理应用102然后可以选择具有最佳分数的A。

[0243] 在一些情况下,可以使用机器学习方法将<EDT-Q, EDT-A>对分类为正确或不正确。示例训练集包括好(肯定)Q/A对和坏(否定)Q/A对。因此,选择DT内核学习方法(SVM TK, Joty和Moschitti 2014, Galitsky 2017),该方法将SVM学习应用于Q/A对的DT的所有子DT的集合。树内核系列方法对解析中的错误(语法和修辞)不是很敏感,因为错误的子树大部分是随机的,并且将在训练集的不同元素之间不太可能是常见的。

[0244] DT可以由每个子树类型的整数计数的向量V表示(不考虑其祖先):

[0245] $V(T) = (\text{类型1的子树的数量}, \dots)$ 。

[0246] 给定两个树段DT₁和DT₂,树内核函数 $K(EDT_1, EDT_2) \leq V(EDT_1)$ 和

[0247]
$$V(EDT_2) = \sum_i V(EDT_1)[i], V(EDT_2)[i] = \sum_{n1} \sum_{n2} \sum_i I_i(n_1) * I_i(n_2)$$

[0248] 其中 $n_1 \in N_1, n_2 \in N_2$,其中 N_1 和 N_2 分别是DT₁和DT₂中的所有节点的集合。

[0249] $I_1(n)$ 是指示符函数。

[0250] $I_1(n) = \{1, \text{类型i的子树出现并且在节点处有根}; \text{否则为} 0\}$

[0251] 继续该示例,尚不清楚问题中的E-Q“ENGINE PUMP”在初始答案中如何被解决。因此,自主代理应用102可以确定可以解决ENGINE PUMP实体的附加资源。

[0252] 在方框1606处,处理1600涉及确定附加的话语树包括将问题实体与答案实体连接的修辞关系。继续该示例,如图17所示,节点1770识别实体“VACUUM(真空吸尘器)”,并且节点1171识别实体“ENGINE(引擎)”。因此,均表示修辞关系“阐述”的节点1172和1173与节点1170和1171相关。以这种方式,节点1172和1173提供了问题中的“ENGINE”和在问题中但先前未解决的“VACUUM”之间的缺失链接。

[0253] 自主代理应用102将附加的话语树连接到E-A。如图所示,DT-A_{img2}经由阐述将VACUUM和ENGINE连接起来。因此,组合的DT-A包括真实的DT-A加上DT-A_{img1}和DT-A_{img2}。通过以领域独立的方式采用背景知识,真实的和虚构的DT对于证明答案是相关的都是必要的。

[0254] 在方框1607处,处理1600涉及提取包括问题实体、答案实体和修辞关系的附加话语树的子树,从而生成虚构的话语树。在一些情况下,自主代理应用102可以提取将问题实体与答案实体相关的子树或附加话语树的一部分。子树包括至少一个节点。在一些情况下,自主代理应用102可以通过将节点连接到答案实体或另一个实体将来自虚构话语树的子树集成到答案话语树中。

[0255] 在方框1608处,处理1600涉及输出由答案话语树和虚构的话语树的组合表示的答案。自主代理应用102可以将方框1607处识别的子树与答案话语树组合。自主代理应用102然后可以输出与组合的树对应的答案。

[0256] 实验结果

[0257] 对于事实和非事实问题的传统Q/A数据集以及SemEval和神经Q/A评估均不适合,因为问题被排序并且没那么复杂以观察话语级别分析的潜在贡献。为了评估,形成两个聚敛性Q/A集:

[0258] 1. 具有广泛主题的问题-答案对的Yahoo!Answer(Webscope 2017)集。从140k的集

合中选择3300个用户问题,其包括三到五个句子。大多数问题的答案都相当详细,因此没有按照句子长度对答案应用过滤。

[0259] 2.包括9300个汽车问题描述的Q/A对以及关于如何纠正它们的建议的汽车维修对话。

[0260] 对于这些集合中的每一个,我们从实际的Q/A对形成肯定对并从 $Q/A_{\text{similar-entities}}$ 形成否定对: $E-A_{\text{similar-entities}}$ 与E-A具有强的重叠性,但是 $A_{\text{similar-entities}}$ 并不是真正的正确、全面和确切的答案。因此,将Q/A简化为经由将Q/A对关联为正确对的类的精确度和查全率而进行测量的分类任务。

[0261]	来源	Yahoo! Answers			汽车修理 (Car Repair)		
	搜索方法	P	R	F1	P	R	F1
	基线 TF * IDF	41.8	42.9	42.3	42.5	37.4	39.8
	E-Q $\cap$ E-A	53.0	57.8	55.3	54.6	49.3	51.8
	EDT-Q $\cap$ EDT-A	66.3	64.1	65.1	66.8	60.3	63.4
	EDT-Q $\cap$ EDT-A + EDT- A_{imgi}	76.3	78.1	77.2	72.1	72.0	72.0
	用于评估 Q/A 准确性的 SVM TK	83.5	82.1	82.8	80.8	78.5	79.6

[0262] 对Q/A准确性的评估

[0263] 表1中的前两行示出了Q/A的基线性能,并表明在复杂的领域中,从关键字到匹配的实体的变换带来12%以上的性能提升。最下面的三行示出了应用话语分析时的Q/A准确性。确保DT-A和DT-Q之间基于规则的对应关系相对于基线提供了12%的增加,并且使用虚构的DT进一步增加10%。最后,从基于规则进行到机器学习的Q/A对应关系(SVM TK)给出了大约7%的性能增益。

[0264] 用修辞一致性分类器补充虚构的话语树

[0265] 通过使用交流话语树,修辞一致性应用112可以确定两个句子之间的互补性。例如,修辞一致性应用112可以确定问题的话语树和初始答案的话语树之间、问题的话语树和附加或候选答案的话语树之间、或答案的话语树和附加答案的话语树之间的互补程度。以这种方式,自主代理应用102确保通过在修辞一致性或风格方面完整的答案来解决复杂的问题。

[0266] 在示例中,修辞一致性应用112从问题构造问题交流话语树,并从初始答案构造答案交流话语树。修辞一致性应用112为问题语句确定问题交流话语树。问题话语树可以包括根节点。例如,再次参考图13和图15,示例问题语句是“are rebels responsible for the downing of the flight”。修辞分类应用102可以使用关于图15描述的处理1500。示例问题具有“阐述”的根节点。

[0267] 修辞一致性应用112为答案语句确定第二交流话语树。答案交流话语树可以包括根节点。继续上面的示例,修辞一致性应用创建了交流话语树,如图13所示,该树也具有标记为“阐述”的根节点。

[0268] 修辞一致性应用112通过识别问题根节点和答案根节点相同来关联交流话语树。修辞一致性应用112确定问题交流话语树和答案交流话语树具有相同的根节点。所产生的相关联的交流话语树如图17所示,并且可以标记为“请求-响应对”。

[0269] 修辞一致性应用112通过将预测模型应用于合并的话语树来计算问题交流话语树和答案交流话语树之间的互补程度。可以使用不同的机器学习技术。在一方面,修辞一致性应用112训练并使用修辞一致性分类器120。例如,修辞一致性应用112可以定义请求-响应对的肯定和否定类别。肯定类别包括修辞上正确的请求-响应对,否定类别包括相关的但修辞上对外的请求-响应对。对于每个请求-响应对,修辞一致性应用112可以通过解析每个句子并获得句子片段的动词签名来构建CDT。修辞一致性应用112将相关联的交流话语树对提供给修辞一致性分类器120,修辞一致性分类器120又输出互补程度。

[0270] 修辞一致性应用112确定互补程度高于阈值,然后将问题和答案语句识别为是互补的。修辞一致性应用112可以使用互补程度阈值来确定问题-答案对是否足够互补。例如,如果分类分数大于阈值,那么修辞一致性应用112可以使用答案。替代地,修辞一致性应用112可以丢弃答案并访问答案数据库105或用于另一个候选答案的公共数据库,并根据需要重复。

[0271] 在另一方面,修辞一致性应用112将从林内核学习应用于表示。丛林内核学习可以代替上述基于分类的学习发生。修辞一致性应用112为请求-响应对的解析树构建解析丛林对。修辞一致性应用112应用话语解析来获得用于请求-响应对的话语树对。修辞一致性应用112使话语树请求-响应和解析树请求-响应的基本话语单元对齐。修辞一致性应用112合并话语树请求-响应和解析树请求-响应的基本话语单元。

[0272] 相关工作

[0273] 在话语中的任何点处,一些实体被认为比其它实体更重要(在DT的核部分中出现),因此预期表现出不同的性质。在中心理论(Grosz等人,1995;Poesio等人,2004)中,实体重要性决定了如何用言语实现它们,包括它们之间的代词化关系。在其它话语理论中,可以经由话题性(Prince 1978)和认知可及性(Gundel等人,1993)来定义实体重要性。

[0274] Barzilay和Lapata (2008)自动将文本抽象为实体转换序列集合,并记录关于话语实体的分布、句法和参考信息。作者将一致性评估制定为学习任务,并表明他们基于实体的表示非常适合基于排名的生成和文本分类任务。

[0275] (Nguyen和Joty 2017)提出了基于卷积神经网络的局部相干模型,该卷积神经网络对文本的网格表示中的实体转变的分布式表示进行操作。局部相干模型可以对足够长的实体转换进行建模,并且可以合并特定于实体的特征而不会失去泛化能力。Kuyten等人2015年开发了一种搜索引擎,该搜索引擎利用文档中的话语结构来克服与信息检索中的词袋(bag-of-words)文档表示相关联的限制。该系统没有解决Q和A之间的修辞协调问题,但是给定Q,该搜索引擎可以检索相关的A和来自A的描述与查询的一些修辞关系的单独的语句。

[0276] 在该研究领域中回答问题是比诸如其中仅需要涉及一个或两个实体及其参数的斯坦福QA数据库(Rajpurkar等人,2016)的事实QA明显复杂得多的任务。为了回答“如何解决问题”的问题,需要维持连接问题中的实体的逻辑流程。由于不可避免地会省略Q中的一些实体,因此这些实体将可能需要从关于这些省略的实体和Q中呈现的实体的一些背景知

识文本中进行恢复。此外,逻辑流程需要补充Q的逻辑流程。

[0277] 特定于领域的本体,诸如与汽车的机械问题相关的本体,很难构建且构建成本很高。在这项工作中,提出了一种经由与领域无关的话语级别分析的替代方法。更具体而言,例如通过在诸如维基百科(Wikipedia)之类的背景知识语料库中找到文本片段来解决DT-A的未解决部分。因此,不需要必须维护所涉及实体之间的关系的本体。

[0278] 所提出的Q/A系统的虚构的DT特征提供了回答复杂聚敛性问题的准确性的显著提高。然而,与关注相关性的基线相比,用于答案样式匹配的DT将Q/A准确性提高了10%以上,依靠虚构的DT则还可以提高10%。

[0279] 本文描述的各方面分析了DT-A和DT-Q之间的互补关系,从而大大地减少了学习特征空间,从而使得从有限大小的可用数据集(诸如汽车维修列表)进行学习是合理的。

[0280] 图18描绘了用于实现这些方面之一的分布式系统1800的简化图。在所示方面中,分布式系统1800包括一个或多个客户端计算设备1802、1804、1806和1808,其被配置为通过一个或多个网络1810执行和操作客户端应用,诸如web浏览器、专有客户端(例如,Oracle Forms)等。服务器1812可以经由网络1810与远程客户端计算设备1802、1804、1806和1808通信地耦合。

[0281] 在各方面中,服务器811可以适于运行由系统的一个或多个部件提供的一个或多个服务或软件应用。服务或软件应用可以包括非虚拟和虚拟环境。虚拟环境可以包括用于虚拟事件、展览、模拟器、教室、购物交易场所和企业的环境,无论是二维还是三维(3D)表示形式、基于页面的逻辑环境还是其它形式。在一些方面中,可以将这些服务作为基于web的服务或云服务或在软件即服务(SaaS)模型下提供给客户端计算设备1802、1804、1806和/或1808的用户。操作客户端计算设备1802、1804、1806和/或1808的用户又可以利用一个或多个客户端应用与服务器1812进行交互,以利用由这些部件提供的服务。

[0282] 在图中绘出的配置中,分布式系统1800的软件部件1818、1820和1822被示出为在服务器1812上实现。在其它方面中,分布式系统1800的一个或多个部件和/或由这些部件提供的服务也可以由客户端计算设备1802、1804、1806和/或1808中的一个或多个来实现。然后,操作客户端计算设备的用户可以利用一个或多个客户端应用来使用由这些部件提供的服务。这些部件可以用硬件、固件、软件或其组合来实现。应该认识到的是,各种不同的系统配置是可能的,其可能与分布式系统1800不同。因此,图中所示的方面是用于实现方面系统的分布式系统的一个示例,并不旨在进行限制。

[0283] 客户端计算设备1802、1804、1806和/或1808可以是便携式手持设备(例如,**iPhone®**、蜂窝电话、**iPad®**、计算平板电脑、个人数字助理(PDA))或可穿戴设备(例如,Google **Glass®**头戴式显示器),它们运行诸如Microsoft Windows **Mobile®**和/或各种移动操作系统(诸如iOS、Windows Phone、Android、BlackBerry 10、Palm OS等)的软件,并且启用互联网、电子邮件、短消息服务(SMS)、**Blackberry®**或其它通信协议。客户端计算设备可以是通用个人计算机,作为示例,包括运行各种版本的Microsoft **Windows®**、Apple **Macintosh®**和/或Linux操作系统的个人计算机和/或膝上型计算机。客户端计算设备可以是运行任何各种可商业获得的**UNIX®**或类UNIX操作系统(包括但不限于各种GNU/Linux操作系统,诸如例如Google Chrome OS)的工作站计算机。替代地

或附加地,客户端计算设备1802、1804、1806和1808可以是能够通过(一个或多个)网络1810通信的任何其它电子设备,诸如瘦客户端计算机、启用互联网的游戏系统(例如,具有或不具有**Kinect®**姿势输入设备的微软Xbox游戏控制台)和/或个人消息传送设备。

[0284] 虽然示例性分布式系统1800被示出为具有四个客户端计算设备,但是可以支持任何数量的客户端计算设备。其它设备(诸如具有传感器的设备等)可以与服务器1812交互。

[0285] 分布式系统1800中的(一个或多个)网络1810可以是本领域技术人员熟悉的、可以利用各种可商业获得的协议中的任何协议支持数据通信的任何类型的网络,其中协议包括但不限于TCP/IP(传输控制协议/网际协议)、SNA(系统网络体系架构)、IPX(互联网报文交换)、AppleTalk,等等。仅仅作为示例,(一个或多个)网络1810可以是局域网(LAN),诸如基于以太网、令牌环等的LAN。(一个或多个)网络1810可以是广域网和互联网。它可以包括虚拟网络,包括但不限于虚拟专用网(VPN)、内联网、外联网、公共交换电话网(PSTN)、红外网络、无线网络(例如,依据电子电气学会(IEEE)802.18协议套件、**Bluetooth®**和/或任何其它无线协议当中任意一种操作的网络);和/或这些和/或其它网络的任意组合。

[0286] 服务器1812可以由一个或多个通用计算机、专用服务器计算机(作为示例,包括PC(个人计算机)服务器、**UNIX®**服务器、中型服务器、大型计算机、机架安装的服务器等)、服务器场、服务器集群或任何其它适当的布置和/或组合组成。服务器1812可以包括运行虚拟操作系统或涉及虚拟化的其他计算架构的一个或多个虚拟机。可以虚拟化一个或多个逻辑存储设备的灵活的池,以维护服务器的虚拟存储设备。服务器1812可以使用软件定义的网络来控制虚拟网络。在各种方面中,服务器1812可以适于运行在前述公开中所描述的一个或多个服务或软件应用。例如,服务器1812可以与用于执行以上根据本公开的方面描述的处理的服务器对应。

[0287] 服务器1812可以运行包括以上讨论的操作系统当中任意一种的操作系统,以及任何可商业获得的服务器操作系统。服务器1812还可以运行各种附加的服务器应用和/或中间层应用中的任何服务器应用和/或中间层应用,包括HTTP(超文本传输协议)服务器、FTP(文件传输协议)服务器、CGI(公共网关接口)服务器、**JAVA®**服务器、数据库服务器,等等。示例性数据库服务器包括但不限于从Oracle、Microsoft、Sybase、IBM(国际商业机器)等可商业获得的那些数据库服务器。

[0288] 在一些实现中,服务器1812可以包括一个或多个应用,以分析和整合从客户端计算设备802、804、806和808的用户接收到的数据馈送和/或事件更新。作为示例,数据馈送和/或事件更新可以包括,但不限于,**Twitter®**馈送、**Facebook®**更新或者从一个或多个第三方信息源接收到的实时更新和连续数据流,其可以包括与传感器数据应用、金融报价机、网络性能测量工具(例如,网络监视和流量管理应用)、点击流分析工具、汽车交通监视等相关的实时事件。服务器1812还可以包括一个或多个应用,以经由客户端计算设备1802、1804、1806和1808的一个或多个显示设备显示数据馈送和/或实时事件。

[0289] 分布式系统1800还可以包括一个或多个数据库1814和1816。数据库1814和1816可以驻留在各种位置中。作为示例,数据库1814和1816中的一个或多个可以驻留在服务器1812本地的(和/或驻留在服务器1812中的)非瞬态存储介质上。替代地,数据库1814和1816可以远离服务器1812,并且经由基于网络的连接或专用的连接与服务器1812通信。在一组

方面中,数据库1814和1816可以驻留在存储区域网络(SAN)中。类似地,用于执行服务器1812所具有的功能的任何必要的文件都可以适当地本地存储在服务器1812上和/或远程存储。在一组方面中,数据库1814和1816可以包括适于响应于SQL格式化的命令而存储、更新和检索数据的关系数据库,诸如由Oracle提供的数据库。

[0290] 图19是根据本公开的方面的系统环境1900的一个或多个部件的简化框图,通过该系统环境1900,由方面系统的一个或多个部件提供的服务可以作为云服务提供。在所示方面中,系统环境1900包括可以由用户使用以与提供云服务的云基础设施系统1902交互的一个或多个客户端计算设备1904、1906和1908。客户端计算设备可以被配置为操作客户端应用,诸如web浏览器、专有客户端应用(例如,Oracle Forms)或某种其它应用,这些客户端应用可以由客户端计算设备的用户用来与云基础设施系统1902交互以使用由云基础设施系统1902提供的服务。

[0291] 应该认识到的是,图中描绘的云基础设施系统1902可以具有除了所描绘的部件之外的其它部件。另外,图中所示的方面仅是可以结合本发明的方面的云基础设施系统的一个示例。在一些其它方面中,云基础设施系统1902可以具有比图中所示更多或更少的部件、可以组合两个或更多个部件、或者可以具有不同的部件配置或布置。

[0292] 客户端计算设备1904、1906和1908可以是与上面针对1002、1004、1006和1008所描述的设备类似的设备。

[0293] 虽然示例性系统环境1900被示出具有三个客户端计算设备,但是任何数量的客户端计算设备可以被支持。诸如具有传感器的设备等之类的其它设备可以与云基础设施系统1902交互。

[0294] (一个或多个)网络1910可以促进客户端计算设备1904、1906和1908与云基础设施系统1902之间的数据通信和交换。每个网络可以是本领域技术人员所熟悉的可以使用各种商业上可获得的协议(包括上面针对(一个或多个)网络1810所描述的那些协议)中的任何一种支持数据通信的任何类型的网络。

[0295] 云基础设施系统1002可以包括一个或多个计算机和/或服务器,其可以包括上面针对服务器1812所描述的那些计算机和/或服务器。

[0296] 在某些方面中,由云基础设施系统提供的服务可以包括按需对云基础设施系统的用户可用的许多服务,诸如在线数据存储和备份解决方案、基于Web的电子邮件服务、托管的办公室(office)套件和文档协作服务、数据库处理、受管理的技术支持服务等。由云基础设施系统提供的服务可以动态扩展以满足云基础设施系统的用户的需要。由云基础设施系统提供的服务的具体实例化在本文中被称为“服务实例”。一般而言,从云服务提供商的系统经由通信网络(诸如互联网)对用户可用的任何服务被称为“云服务”。通常,在公共云环境中,构成云服务提供商的系统的服务器和系统与用户自己的本地服务器和系统不同。例如,云服务提供商的系统可以托管应用,并且用户可以经由诸如互联网之类的通信网络按需订购和使用应用。

[0297] 在一些示例中,计算机网络云基础设施中的服务可以包括对存储装置、被托管的数据库、被托管的Web服务器、软件应用或由云供应商向用户提供的其它服务的受保护的计算机网络访问,或者如本领域中另外已知的那样。例如,服务可以包括通过互联网对云上的远程存储装置进行受密码保护的访问。作为另一个示例,服务可以包括基于Web服务的被托

管的关系数据库和脚本语言中间件引擎,以供联网的开发人员私有使用。作为另一个示例,服务可以包括对在云供应商的网站上托管的电子邮件软件应用的访问。

[0298] 在某些方面中,云基础设施系统1902可以包括以自助服务、基于订阅、弹性可扩展、可靠、高度可用和安全的方式交付给客户的应用、中间件和数据库服务产品的套件。这种云基础设施系统的示例是由本受让人提供的Oracle公共云。

[0299] 大量数据(有时称为大数据)可以由基础设施系统在许多级别和不同规模上托管和/或操纵。这样的数据可能包含数据集,这些数据集庞大而复杂以致于很难使用典型的数据库管理工具或传统的数据处理应用进行处理。例如,使用个人计算机或其基于机架的对应物可能难以存储、检索和处理兆字节级的数据。使用最新的关系数据库管理系统以及桌面统计数据和可视化程序包,可能难以处理这种大小的数据。它们可能需要运行数千台服务器计算机的超出常用软件工具的结构的大规模并行处理软件,才能在可容忍的经过时间内捕获、整理、管理和处理数据。

[0300] 分析人员和研究人员可以存储和操纵非常大的数据集,以可视化大量数据、检测趋势和/或以其他方式与数据进行交互。并行链接的数十个、数百个或数千个处理器可以对此类数据进行操作,以呈现该数据或模拟对该数据或其表示的事物的外力。这些数据集可能涉及结构化数据(例如,在数据库中组织的数据或根据结构化模型以其他方式进行组织的数据)和/或非结构化数据(例如,电子邮件、图像、数据blob(二进制大对象)、网页、复杂事件处理)。通过利用一个方面的能力来相对快速地将更多(或更少)计算资源集中在一个目标上,可以根据企业、政府机构、研究组织、私人、志趣相投的个人或组织或其他实体的需求,更好地利用云基础架构系统来对大型数据集执行任务。

[0301] 在各种方面中,云基础设施系统1002可以适于自动供应、管理和跟踪客户对由云基础设施系统1902供给的服务的订阅。云基础设施系统1002可以经由不同的部署模型来提供云服务。例如,可以依据公共云模型提供服务,在公共云模型中云基础设施系统1002被销售云服务的组织拥有(例如,被Oracle拥有),并且服务对一般公众或不同行业的企业可用。作为另一个示例,可以依据私有云模型来提供服务,在私有云模型中云基础设施系统1002仅针对单个组织操作,并且可以为该组织内的一个或多个实体提供服务。还可以依据社区云模型来提供云服务,在社区云模型中云基础设施系统1002和由云基础设施系统1002提供的服务由相关社区中的若干组织共享。云服务还可以依据混合云模型被提供,该混合云模型是两个或更多个不同模型的组合。

[0302] 在一些方面中,由云基础设施系统1002提供的服务可以包括在软件即服务(SaaS)类别、平台即服务(PaaS)类别、基础设施即服务(IaaS)类别或包括混合服务的其它服务类别下提供的一个或多个服务。客户经由订阅订单可以订购由云基础设施系统1902提供的一个或多个服务。云基础设施系统1002然后执行处理以提供客户的订阅订单中的服务。

[0303] 在一些方面中,由云基础设施系统1002提供的服务可以包括但不限于应用服务、平台服务和基础设施服务。在一些示例中,应用服务可以由云基础设施系统经由SaaS平台提供。SaaS平台可以被配置为提供落入SaaS类别的云服务。例如,SaaS平台可以提供在集成开发和部署平台上构建和交付按需应用套件的能力。SaaS平台可以管理和控制用于提供SaaS服务的底层软件和基础设施。通过利用由SaaS平台提供的服务,客户可以利用在云基础设施系统上执行的应用。客户可以获取应用服务,而无需客户购买单独的许可和支持。可

以提供各种不同的SaaS服务。示例包括但不限于为大型组织提供销售业绩管理、企业集成和业务灵活性的解决方案的服务。

[0304] 在一些方面中,平台服务可以由云基础设施系统经由PaaS平台提供。PaaS平台可以被配置为提供落入PaaS类别的云服务。平台服务的示例可以包括但不限于使组织(诸如Oracle)能够在共享的公共体系架构上整合现有应用以及充分利用平台提供的共享服务来构建新应用的能力的服务。PaaS平台可以管理和控制用于提供PaaS服务的底层软件和基础设施。客户可以获取由云基础架构系统提供的PaaS服务,而无需客户购买单独的许可和支持。平台服务的示例包括但不限于Oracle Java云服务(JCS)、Oracle数据库云服务(DBCS)等。

[0305] 通过利用由PaaS平台提供的服务,客户可以采用由云基础设施系统支持的编程语言和工具,并且还控制所部署的服务。在一些方面中,由云基础设施系统提供的平台服务可以包括数据库云服务、中间件云服务(例如,Oracle融合中间件服务)和Java云服务。在一个方面中,数据库云服务可以支持共享服务部署模型,该模型使得组织能够汇集数据库资源并且以数据库云的形式向客户供应数据库即服务。在云基础设施系统中,中间件云服务可以为客户提供开发和部署各种业务应用的平台,并且Java云服务可以为客户提供部署Java应用的平台。

[0306] 各种不同的基础设施服务可以由云基础设施系统中的IaaS平台提供。基础设施服务促进对底层计算资源(诸如存储装置、网络和其它基础计算资源)的管理和控制,以供客户利用由SaaS平台和PaaS平台提供的服务。

[0307] 在某些方面中,云基础设施系统1002还可以包括基础设施资源1930,用于向云基础设施系统的客户提供用于提供各种服务的资源。在一个方面中,基础设施资源1930可以包括预先集成和优化的硬件(诸如服务器、存储装置和联网资源)的组合,以执行由PaaS平台和SaaS平台提供的服务。

[0308] 在一些方面中,云基础设施系统1002中的资源可以由多个用户共享并且根据需要动态重新分配。此外,可以将资源分配给在不同时区的用户。例如,云基础设施系统1002可以使在第一时间区中的第一组用户能够在指定的小时数内利用云基础设施系统的资源,并且然后使相同资源能够被重新分配给位于不同时间区的另一组用户,从而使资源的利用率最大化。

[0309] 在某些方面中,可以提供由云基础设施系统1902的不同部件或模块以及由云基础设施系统1902提供的服务共享的多个内部共享服务1932。这些内部共享服务可以包括但不限于:安全和身份服务、集成服务、企业储存库服务、企业管理器服务、病毒扫描和白名单服务、高可用性、备份和恢复服务、启用云支持的服务、电子邮件服务、通知服务、文件传输服务等。

[0310] 在某些方面中,云基础设施系统1902可以提供云基础设施系统中的云服务(例如,SaaS、PaaS和IaaS服务)的综合管理。在一个方面中,云管理功能可以包括用于供应、管理和跟踪由云基础设施系统1902接收到的客户订阅的能力等。

[0311] 在一个方面中,如图中所绘出的,云管理功能可以由一个或多个模块提供,诸如订单管理模块1920、订单编排模块1922、订单供应模块1924、订单管理和监视模块1926,以及身份管理模块1928。这些模块可以包括一个或多个计算机和/或服务器或者使用一个或多

个计算机和/或服务器来提供,这些计算机和/或服务器可以是通用计算机、专用服务器计算机、服务器场、服务器集群或任何其它适当的布置和/或组合。

[0312] 在示例性操作1934中,使用客户端计算设备(诸如客户端计算设备1904、1906或1908)的客户可以通过请求由云基础设施系统1902提供的一个或多个服务并且针对由云基础设施系统1902供应的一个或多个服务下订阅订单来与云基础设施系统1902交互。在某些方面中,客户可以访问云用户界面(UI)(云UI 1919、云UI 1914和/或云UI 1916)并经由这些UI下订阅订单。云基础设施系统1902响应于客户下订单而接收到的订单信息可以包括识别客户以及客户想要订阅的由云基础设施系统1902供应的一个或多个服务的信息。

[0313] 在客户下订单之后,经由云UI 1010、1014和/或1014接收订单信息。

[0314] 在操作1936处,订单存储在订单数据库1918中。订单数据库1918可以是由云基础设施系统1902操作和与其它系统元件一起操作的若干数据库之一。

[0315] 在操作1938处,订单信息被转发到订单管理模块1920。在一些情况下,订单管理模块1920可以被配置为执行与订单相关的计费 and 记账功能,诸如验证订单、以及在验证后预订订单。

[0316] 在操作1940处,将关于订单的信息传送到订单编排模块1922。订单编排模块1922可以利用订单信息为客户下的订单编排服务和资源的供应。在一些情况下,订单编排模块1922可以使用订单供应模块1924的服务来编排资源的供应以支持所订阅的服务。

[0317] 在某些方面中,订单编排模块1922使得能够管理与每个订单相关联的业务过程并应用业务逻辑来确定订单是否应该进行到供应。在操作1942处,在接收到新订阅的订单时,订单编排模块1922向订单供应模块1924发送分配资源并配置履行订阅订单所需的那些资源的请求。订单供应模块1924使得能够为客户订购的服务分配资源。订单供应模块1924提供在由云基础设施系统1902提供的云服务 and 用于供应于提供所请求的服务的资源的物理实现层之间的抽象层。因此,订单编排模块1922可以与实现细节(诸如服务和资源是否实际上即时供应或预先供应并仅在请求后才分配/指派)隔离。

[0318] 在操作1944处,一旦供应了服务和资源,就可以通过云基础设施系统1902的订单供应模块1924向客户端计算设备1904、1906和/或1908上的客户发送所提供的服务的通知。

[0319] 在操作1946处,订单管理和监视模块1926可以管理和跟踪客户的订阅订单。在一些情况下,订单管理和监视模块1926可以被配置为收集订阅订单中的服务的使用统计信息,诸如,所使用的存储量、传输的数据量、用户的数量,以及系统运行时间量和系统停机时间量。

[0320] 在某些方面中,云基础设施系统1902可以包括身份管理模块1928。身份管理模块1928可以被配置为提供身份服务,诸如云基础设施系统1902中的访问管理和授权服务。在一些方面中,身份管理模块1928可以控制关于希望利用由云基础设施系统1902提供的服务的客户的信息。这样的信息可以包括认证这些客户的身份的信息以及描述这些客户被授权相对于各种系统资源(例如,文件、目录、应用、通信端口、存储器段等)执行哪些动作的信息。身份管理模块1928还可以包括对关于每个客户的描述性信息以及关于如何和由谁来访问和修改这些描述性信息的管理。

[0321] 图20示出了其中可以实现本发明的各种方面的示例性计算机系统2000。系统2000可以用于实现上述任何计算机系统。如图所示,计算机系统2000包括经由总线子系统2002

与多个外围子系统通信的处理单元2004。这些外围子系统可以包括处理加速单元2006、I/O子系统2008、存储子系统2018和通信子系统2024。存储子系统2018包括计算机可读存储介质2022和系统存储器2010。

[0322] 总线子系统2002提供用于让计算机系统2000的各种部件和子系统按意图彼此通信的机制。虽然总线子系统2002被示意性地示出为单条总线,但是总线子系统的替代方面可以利用多条总线。总线子系统2002可以是若干种类型的总线结构中的任何一种,包括使用任何各种总线体系架构的存储器总线或存储器控制器、外围总线、以及局部总线。例如,这种体系架构可以包括工业标准体系架构 (ISA) 总线、微通道体系架构 (MCA) 总线、增强型 ISA (EISA) 总线、视频电子标准协会 (VESA) 局部总线和外围部件互连 (PCI) 总线,其可以被实现为按IEEE P2086.1标准制造的Mezzanine总线。

[0323] 可以被实现为一个或多个集成电路 (例如,常规微处理器或微控制器) 的处理单元2004控制计算机系统2000的操作。一个或多个处理器可以被包括在处理单元2004中。这些处理器可以包括单核或多核处理器。在某些方面中,处理单元2004可以被实现为一个或多个独立的处理单元2032和/或2034,其中在每个处理单元中包括单核或多核处理器。在其它方面中,处理单元2004也可以被实现为通过将两个双核处理器集成到单个芯片中而形成的四核处理单元。

[0324] 在各种方面中,处理单元2004可以响应于程序代码执行各种程序并且可以维护多个并发执行的程序或进程。在任何给定的时间,要被执行的程序代码中的一些或全部代码可以驻留在 (一个或多个) 处理单元2004中和/或存储子系统2018中。通过适当的编程, (一个或多个) 处理单元2004可以提供上述各种功能。计算机系统2000可以附加地包括处理加速单元2006,其可以包括数字信号处理器 (DSP)、专用处理器,等等。

[0325] I/O子系统2008可以包括用户接口输入设备和用户接口输出设备。用户接口输入设备可以包括键盘、诸如鼠标或轨迹球之类的定点设备、结合到显示器中的触摸板或触摸屏、滚动轮、点击轮、拨盘、按钮、开关、键盘、具有语音命令识别系统的音频输入设备、麦克风以及其它类型的输入设备。用户接口输入设备可以包括,例如,运动感测和/或姿势识别设备,诸如的Microsoft **Kinect®**运动传感器,其使得用户能够使用姿势和口头命令通过自然用户接口来控制诸如Microsoft **Xbox®** 360游戏控制器之类的输入设备并与该输入设备交互。用户接口输入设备也可以包括眼睛姿势识别设备,诸如从用户检测眼睛活动 (例如,当拍摄照片和/或做出菜单选择时的“眨眼”) 并且将眼睛姿势转换为到输入设备 (例如, Google **Glass®**) 中的输入的Google **Glass®**眨眼检测器。此外,用户接口输入设备可以包括使用户能够通过语音命令与语音识别系统 (例如, **Siri®**导航器) 交互的语音识别感测设备。

[0326] 用户接口输入设备也可以包括但不限于三维 (3D) 鼠标、操纵杆或指向棒、游戏面板和绘图板,以及音频/视觉设备,诸如扬声器、数码相机、数码摄像机、便携式媒体播放器、网络摄像头、图像扫描仪、指纹扫描仪、条形码阅读器3D扫描仪、3D打印机、激光测距仪和视线跟踪设备。此外,用户接口输入设备可以包括,例如,医学成像输入设备,诸如计算机断层扫描、磁共振成像、正电子发射断层摄影术、医疗超声设备。用户接口输入设备也可以包括,例如,诸如MIDI键盘、数字乐器等之类的音频输入设备。

[0327] 用户接口输出设备可以包括显示子系统、指示灯,或者诸如音频输出设备之类的非视觉显示器,等等。显示子系统可以是阴极射线管(CRT)、诸如使用液晶显示器(LCD)或等离子显示器的平板设备、投影设备、触摸屏,等等。一般而言,术语“输出设备”的使用意在包括用于从计算机系统2000向用户或其它计算机输出信息的所有可能类型的设备和机制。例如,用户接口输出设备可以包括,但不限于,可视地传达文本、图形和音频/视频信息的各种显示设备,诸如监视器、打印机、扬声器、耳机、汽车导航系统、绘图仪、语音输出设备以及调制解调器。

[0328] 计算机系统2000可以包括包含软件元件、被示为当前位于系统存储器2010中的存储子系统2018。系统存储器2010可以存储可加载并且可在处理单元2004上执行的程序指令,以及在这些程序的执行期间所产生的数据。

[0329] 取决于计算机系统2000的配置和类型,系统存储器2010可以是易失性的(诸如随机存取存储器(RAM))和/或非易失性的(诸如只读存储器(ROM)、闪存存储器,等等)。RAM通常包含可被处理单元2004立即访问和/或目前正被处理单元2004操作和执行的数据和/或程序模块。在一些实现中,系统存储器2010可以包括多种不同类型的存储器,例如静态随机存取存储器(SRAM)或动态随机存取存储器(DRAM)。在一些实现中,包含有助于诸如在启动期间在计算机系统2000内的元件之间传送信息的基本例程的基本输入/输出系统(BIOS),通常可以被存储在ROM中。作为示例,但不是限制,系统存储器2010也示出了应用程序2012、程序数据2014以及操作系统2016,应用程序2012可以包括客户端应用、web浏览器、中间层应用、关系数据库管理系统(RDBMS)等。作为示例,操作系统2016可以包括各种版本的Microsoft **Windows®**、Apple **Macintosh®** 和/或Linux操作系统、各种可商业获得的 **UNIX®**或类UNIX操作系统(包括但不限于各种GNU/Linux操作系统、Google **Chrome®**操作系统等)和/或诸如iOS、**Windows®** Phone、**Android®** OS、**BlackBerry®** 100S和**Palm®** OS操作系统之类的移动操作系统。

[0330] 存储子系统2018也可以提供用于存储提供一些方面的功能的基本编程和数据结构的有形计算机可读存储介质。当被处理器执行时提供上述功能的软件(程序、代码模块、指令)可以被存储在存储子系统2018中。这些软件模块或指令可以被处理单元2004执行。存储子系统2018也可以提供用于存储根据本发明被使用的数据的储存库。

[0331] 存储子系统2018也可以包括可被进一步连接到计算机可读存储介质2023的计算机可读存储介质读取器2020。与系统存储器2010一起,并且可选地与系统存储器2010相结合,计算机可读存储介质2022可以全面地表示用于临时和/或更持久地包含、存储、发送和检索计算机可读信息的远程、本地、固定和/或可移除存储设备加存储介质。

[0332] 包含代码或代码的部分的计算机可读存储介质2022也可以包括本领域已知或使用的任何适当的介质,包括存储介质和通信介质,诸如但不限于,以用于信息的存储和/或传输的任何方法或技术实现的易失性和非易失性、可移除和不可移除介质。这可以包括有形的非暂时性计算机可读存储介质,诸如RAM、ROM、电可擦除可编程ROM(EEPROM)、闪存存储器或其它存储器技术、CD-ROM、数字多功能盘(DVD)或其它光学存储装置、磁带盒、磁带、磁盘存储装置或其它磁存储设备,或者其它有形的计算机可读介质。当被指定时,这还可以包括非有形的暂时性计算机可读介质,诸如数据信号、数据传输,或者可以被用来发送期望信

息并且可以被计算系统2000访问的任何其它介质。

[0333] 作为示例,计算机可读存储介质2022可以包括从不可移除的非易失性磁介质读取或对不可移除的非易失性磁介质写入的硬盘驱动器、从可移除的非易失性磁介质读取或对可移除的非易失性磁介质写入的磁盘驱动器、以及从可移除的非易失性光盘(诸如CD ROM、DVD和**Blu-Ray®**盘或其它光学介质)读取或对可移除的非易失性光盘写入的光盘驱动器。计算机可读存储介质2022可以包括,但不限于,**Zip®**驱动器、闪存卡、通用串行总线(USB)闪存驱动器、安全数字(SD)卡、DVD盘、数字视频带,等等。计算机可读存储介质2022也可以包括基于非易失性存储器的固态驱动器(SSD)(诸如基于闪存存储器的SSD、企业闪存驱动器、固态ROM等)、基于易失性存储器的SSD(诸如固态RAM、动态RAM、静态RAM)、基于DRAM的SSD,磁阻RAM(MRAM) SSD,以及使用基于DRAM和闪存存储器的SSD的组的混合SSD。盘驱动器及其关联的计算机可读介质可以为计算机系统2000提供计算机可读指令、数据结构、程序模块及其它数据的非易失性存储。

[0334] 通信子系统2024提供到其它计算机系统和网络的接口。通信子系统2024用作用于从其它系统接收数据和从计算机系统2000向其它系统发送数据的接口。例如,通信子系统2024可以使计算机系统2000能够经由互联网连接到一个或多个设备。在一些方面中,通信子系统2024可以包括用于(例如,使用蜂窝电话技术,诸如3G、4G或EDGE(用于全球演进的增强型数据速率)之类的先进数据网络技术,WiFi(IEEE 802.10系列标准),或其它移动通信技术,或其任意组合)访问无线语音和/或数据网络的射频(RF)收发器部件、全球定位系统(GPS)接收器部件和/或其它部件。在一些方面中,作为无线接口的附加或者替代,通信子系统2024可以提供有线网络连接(例如,以太网)。

[0335] 在一些方面中,通信子系统2024也可以代表可以使用计算机系统2000的一个或多个用户接收结构化和/或非结构化数据馈送2026、事件流2028、事件更新2030等形式的输入通信。

[0336] 作为示例,通信子系统2024可被配置为实时地从社交媒体网络和/或其它通信服务的用户接收非结构化的数据馈送2026,诸如**Twitter®**馈送、**Facebook®**更新、诸如丰富站点摘要(RSS)馈送之类的web馈送和/或来自一个或多个第三方信息源的实时更新。

[0337] 此外,通信子系统2024也可被配置为接收连续数据流形式的数据,这可以包括本质上可以是连续的或无界的没有明确终止的实时事件的事件流2028和/或事件更新2030。产生连续数据的应用的示例可以包括,例如,传感器数据应用、金融报价机、网络性能测量工具(例如,网络监视和流量管理应用)、点击流分析工具、汽车交通监视,等等。

[0338] 通信子系统2024也可被配置为向一个或多个数据库输出结构化和/或非结构化数据馈送2026、事件流2028、事件更新2030,等等,这一个或多个数据库可以与耦合到计算机系统2000的一个或多个流式数据源计算机通信。

[0339] 计算机系统2000可以是各种类型之一,包括手持便携式设备(例如,**iPhone®**蜂窝电话、**iPad®**计算平板电脑、PDA)、可穿戴设备(例如,**Google Glass®**头戴式显示器)、PC、工作站、大型机、信息亭、服务器机架、或任何其它数据处理系统。

[0340] 由于计算机和网络的不断变化的本质,在图中绘出的计算机系统2000的描述仅仅旨在作为具体的示例。具有比图中绘出的系统更多或更少部件的许多其它配置是可能的。

例如,定制的硬件也可以被使用和/或特定的元素可以用硬件、固件、软件(包括小应用程序)或其组合来实现。另外,也可以采用到诸如网络输入/输出设备之类的其它计算设备的连接。基于本文提供的公开内容和教导,本领域普通技术人员将认识到实现各种方面的其它方式和/或方法。

[0341] 在前述的说明书中,本发明的各方面参考其具体方面进行了描述,但本领域技术人员将认识到的是,本发明不限于此。上述发明的各个特征和方面可以被单独或联合使用。此外,在不脱离本说明书的更广泛精神和范围的情况下,方面可以在除本文所述的环境和应用之外的任何数目的环境 and 应用中被使用。相应地,本说明书和附图应当被认为是说明性而不是限制性的。

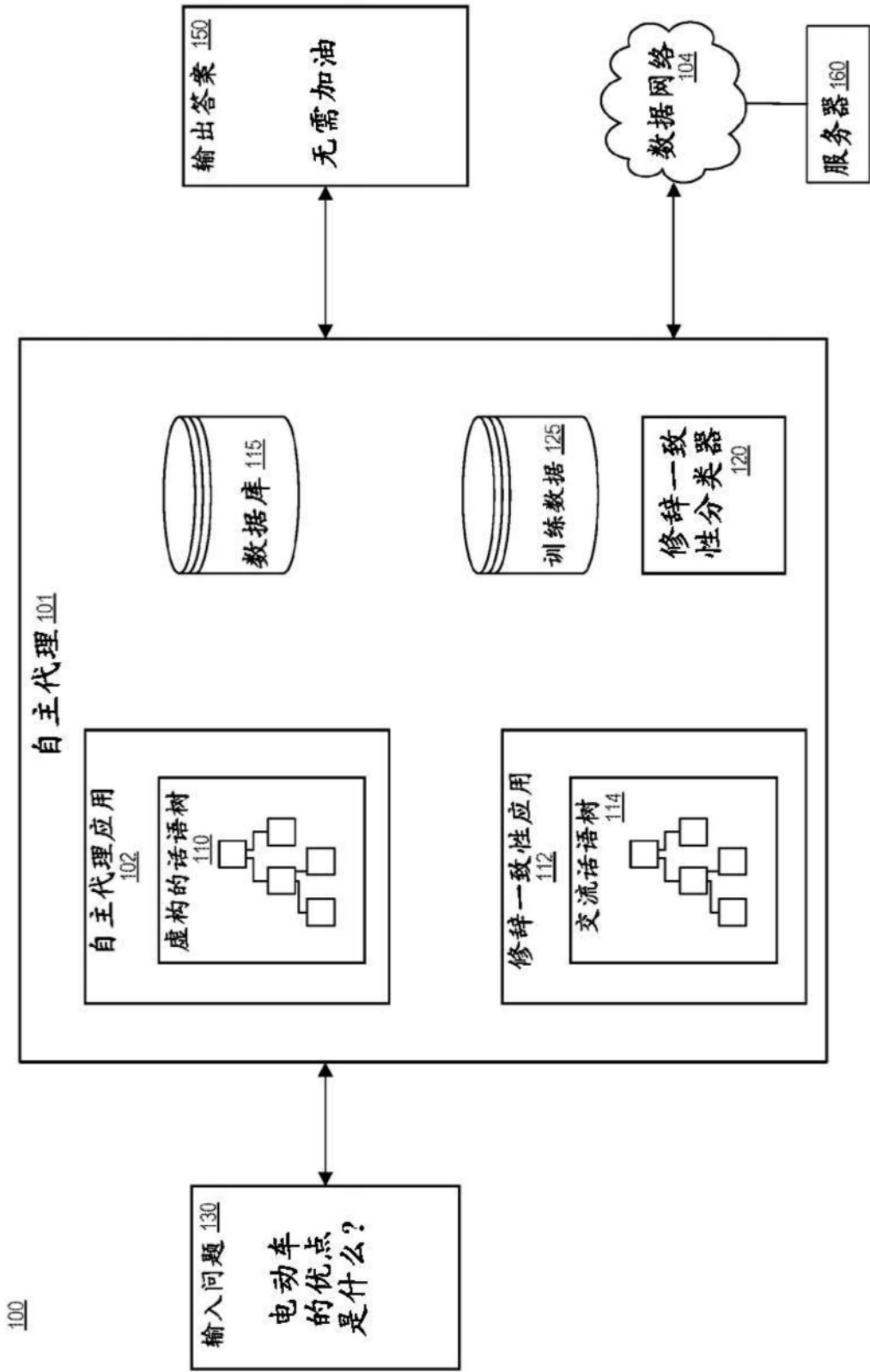


图1

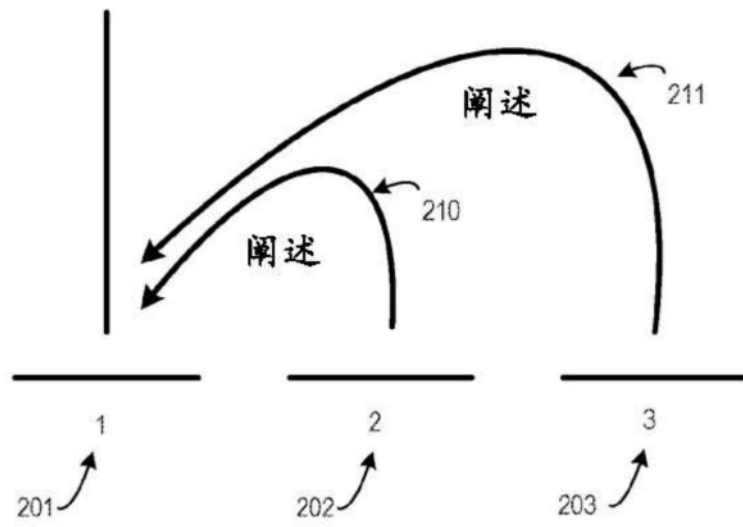
200

图2

300

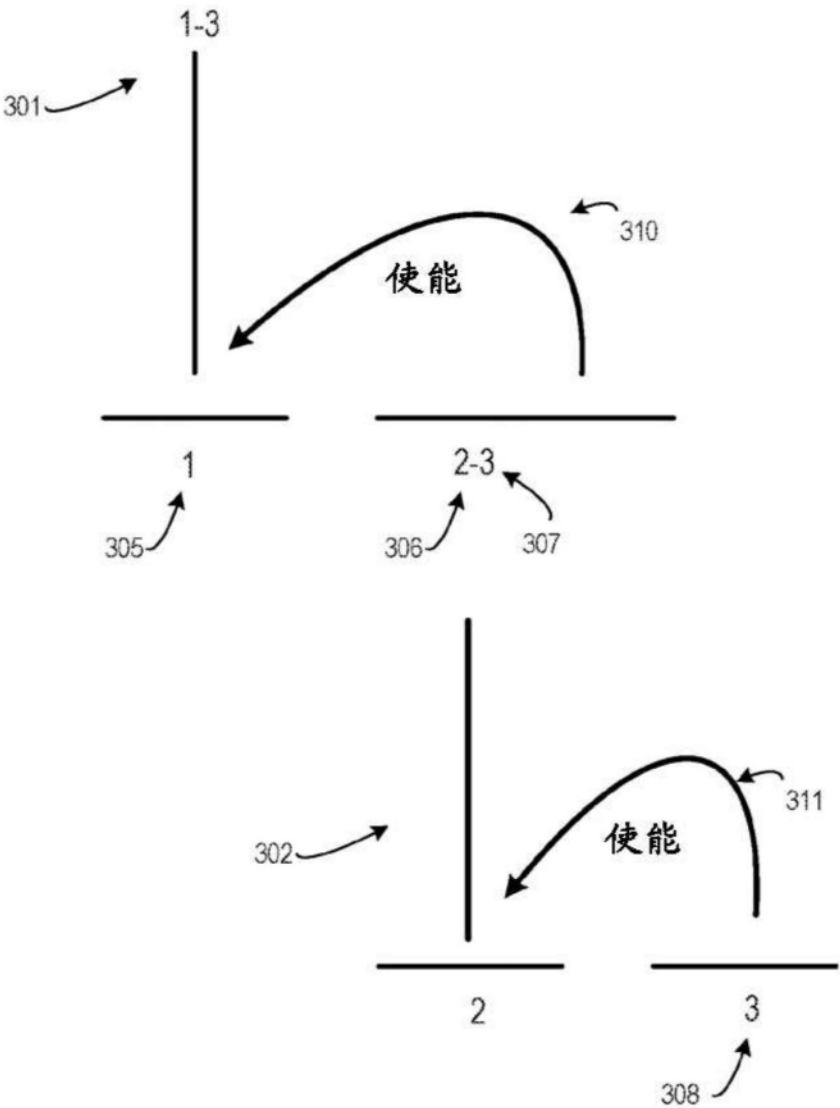


图3

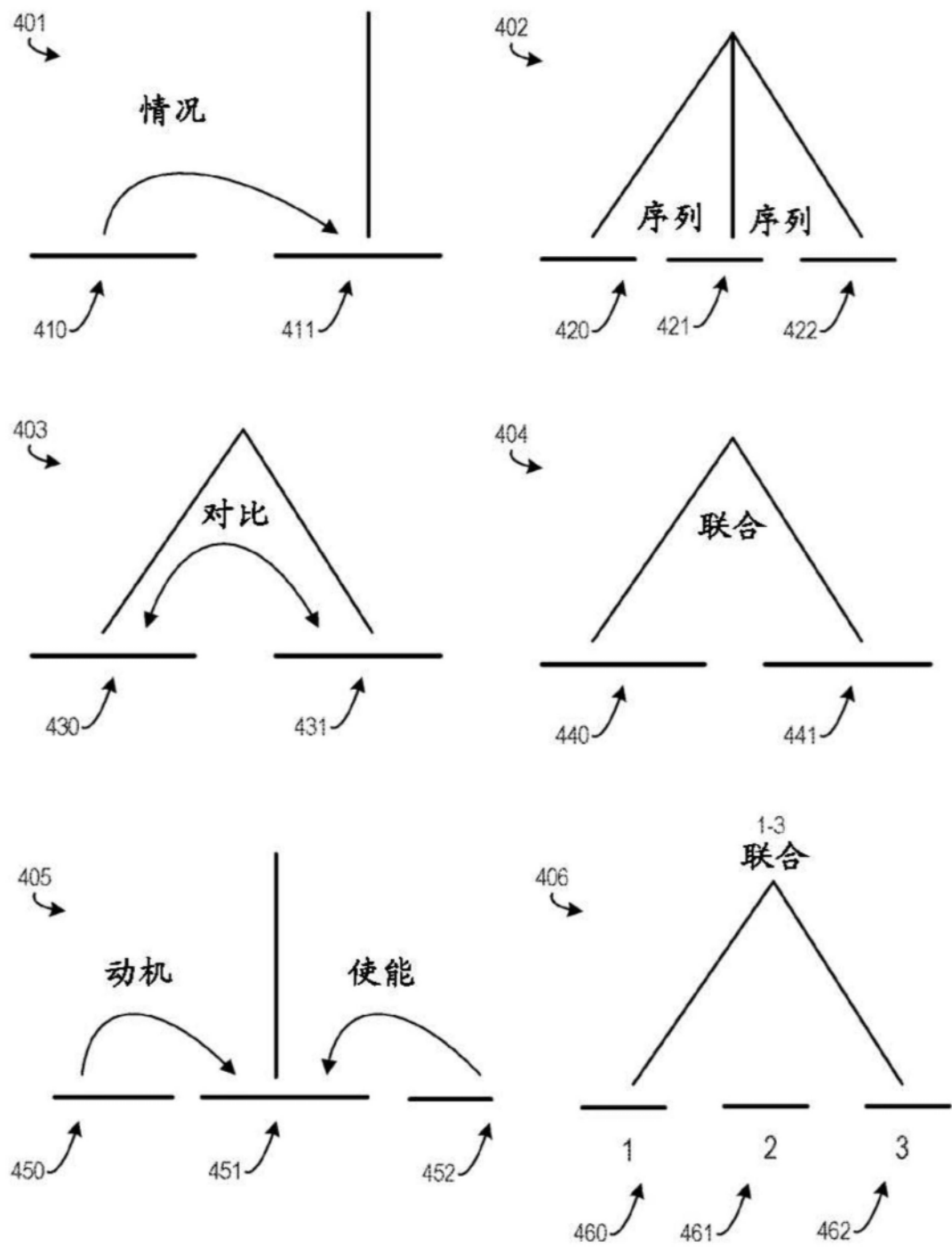


图4

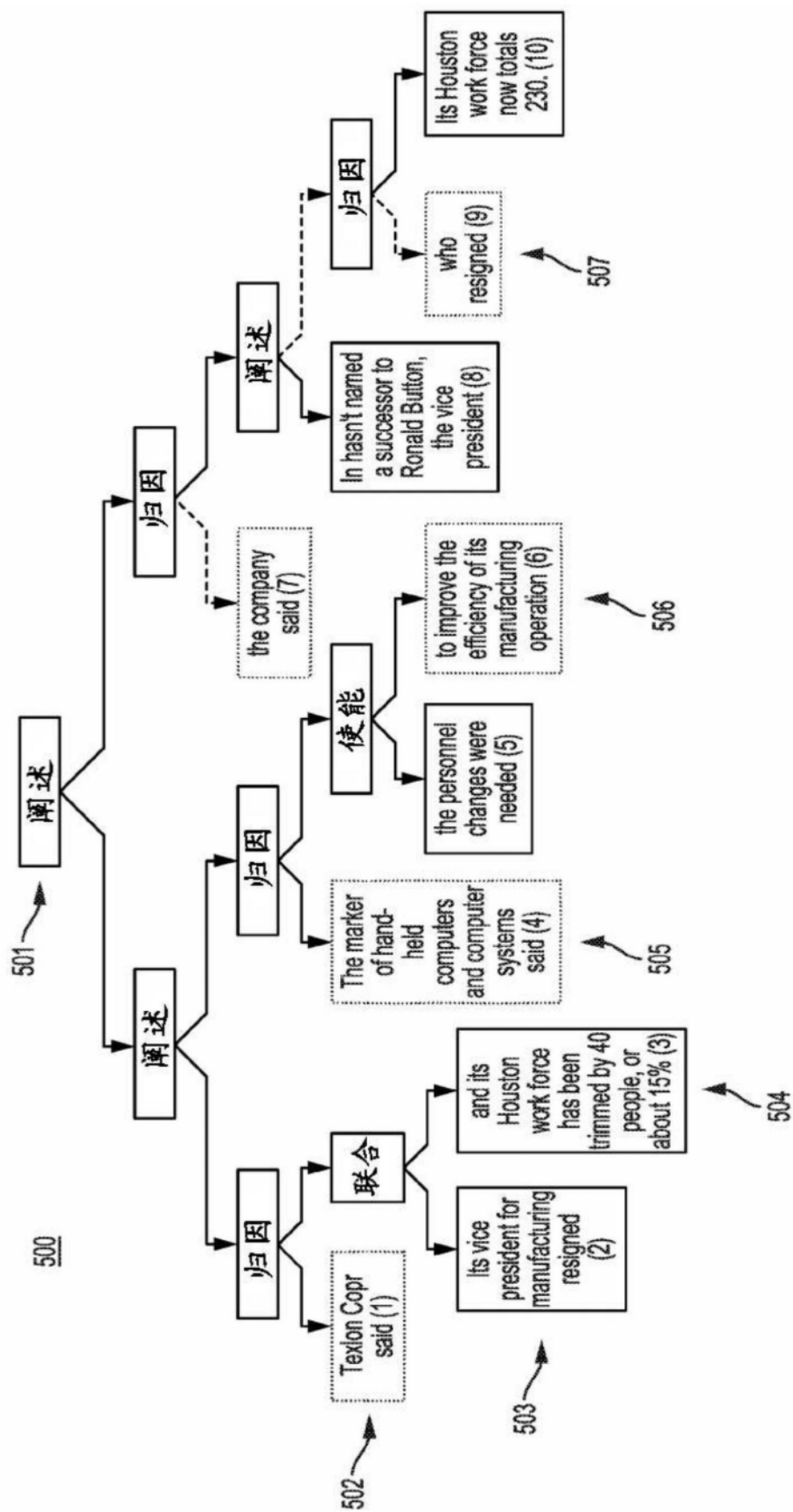


图5

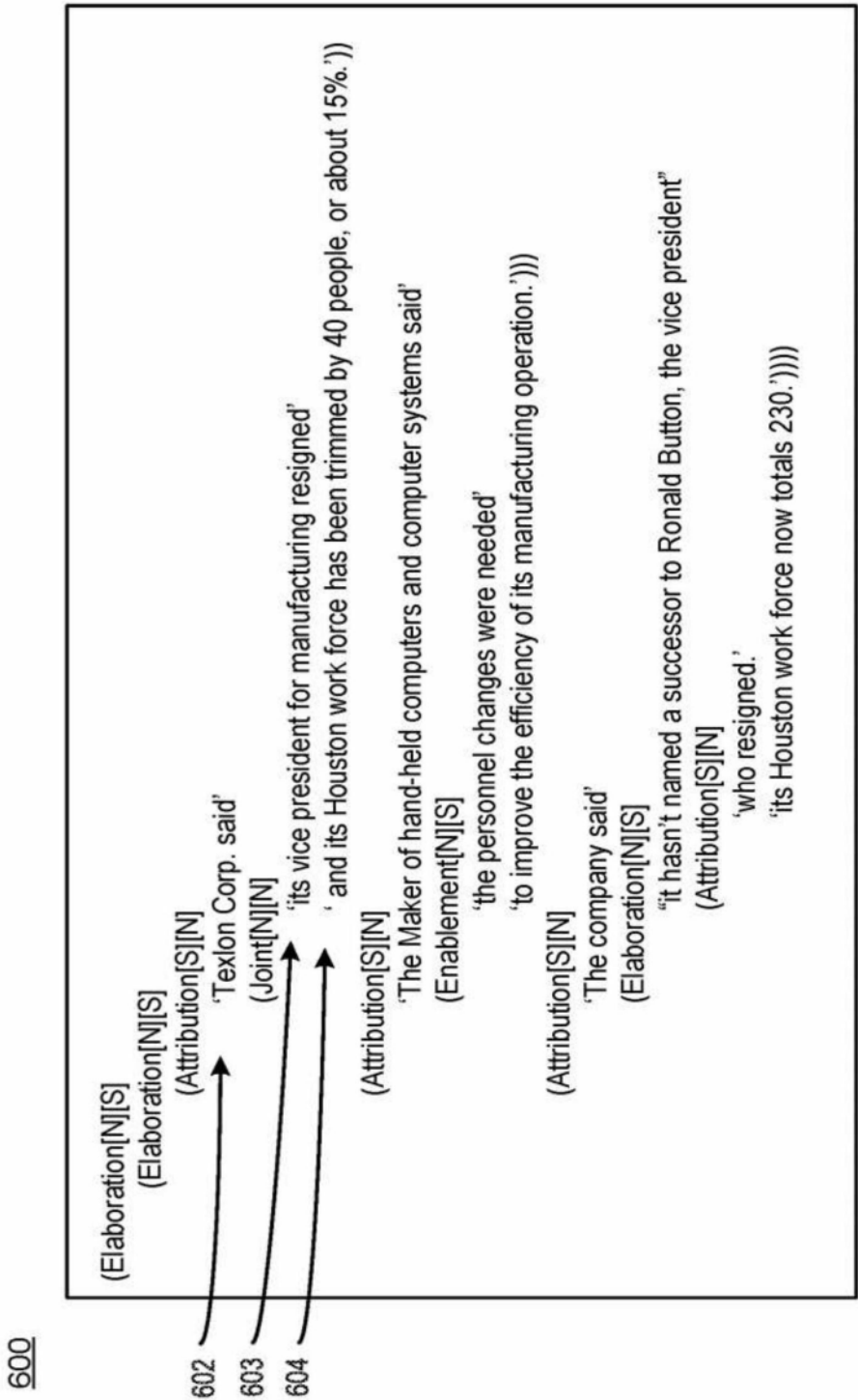


图6

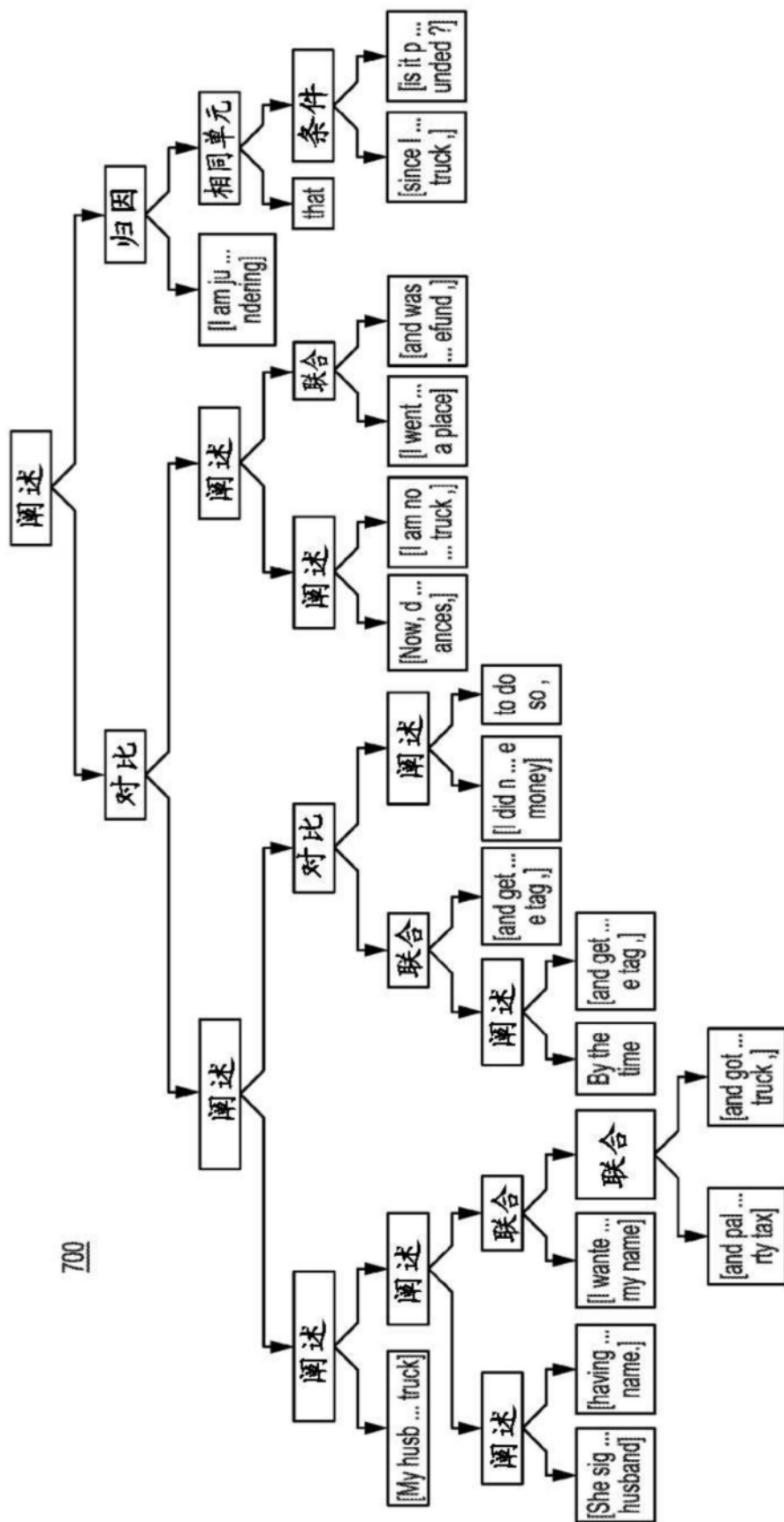


图7

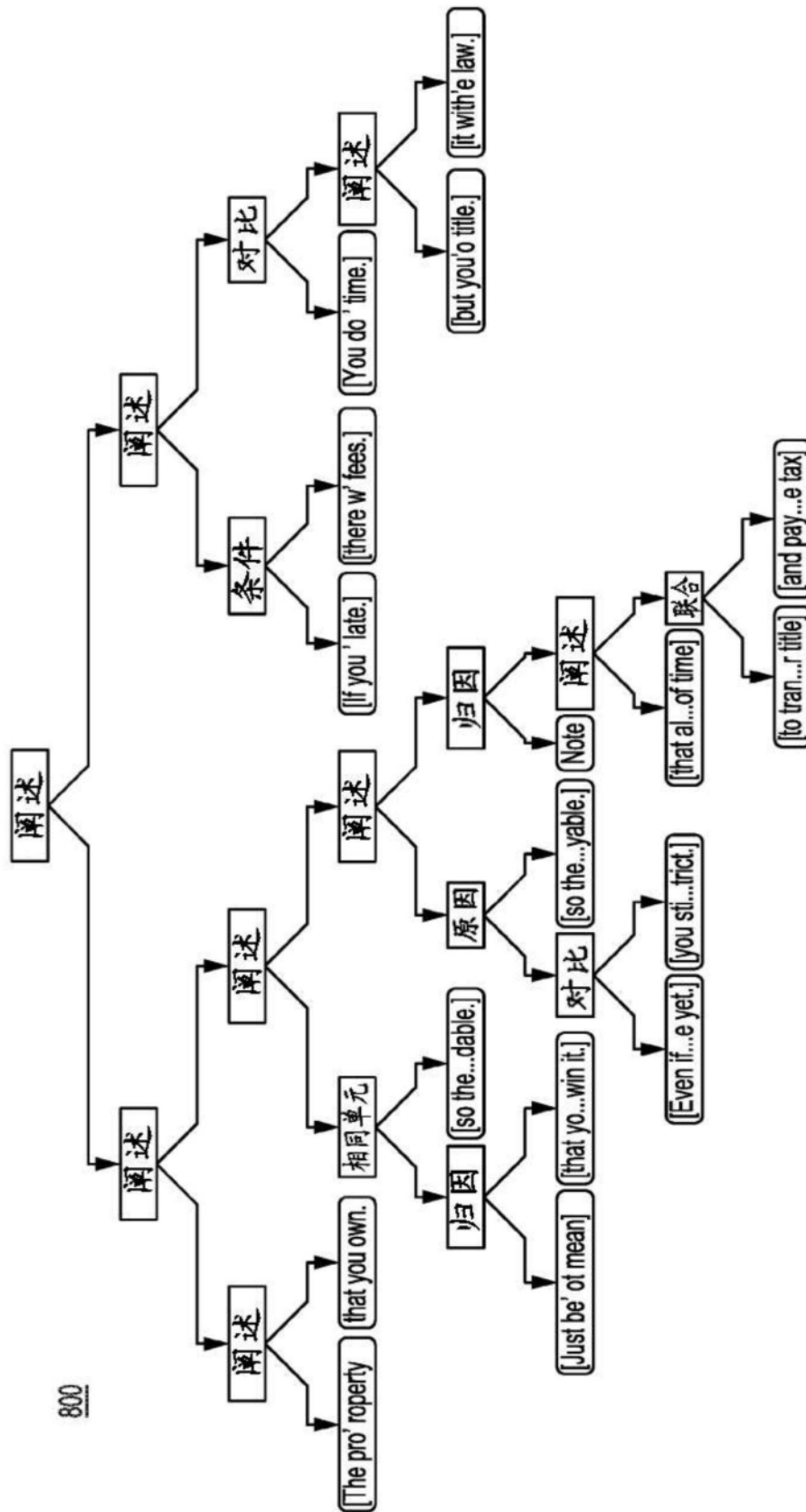


图8

900

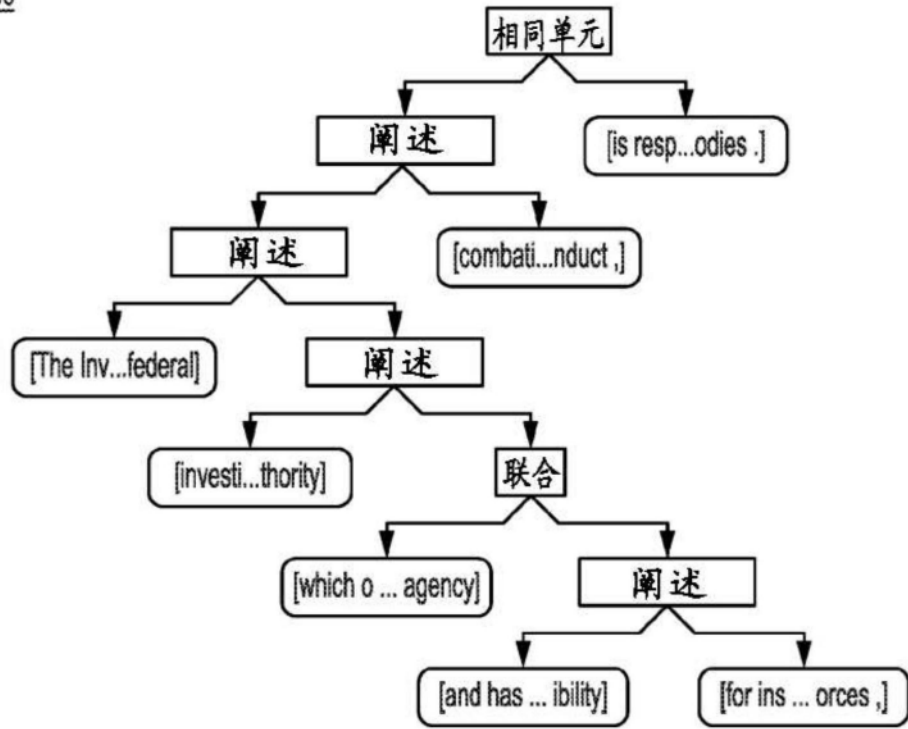


图9

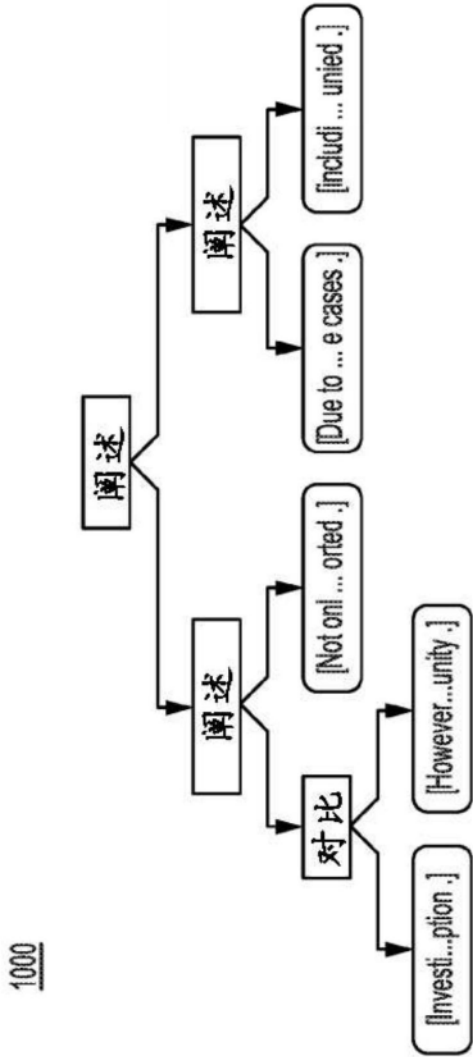


图10

1100

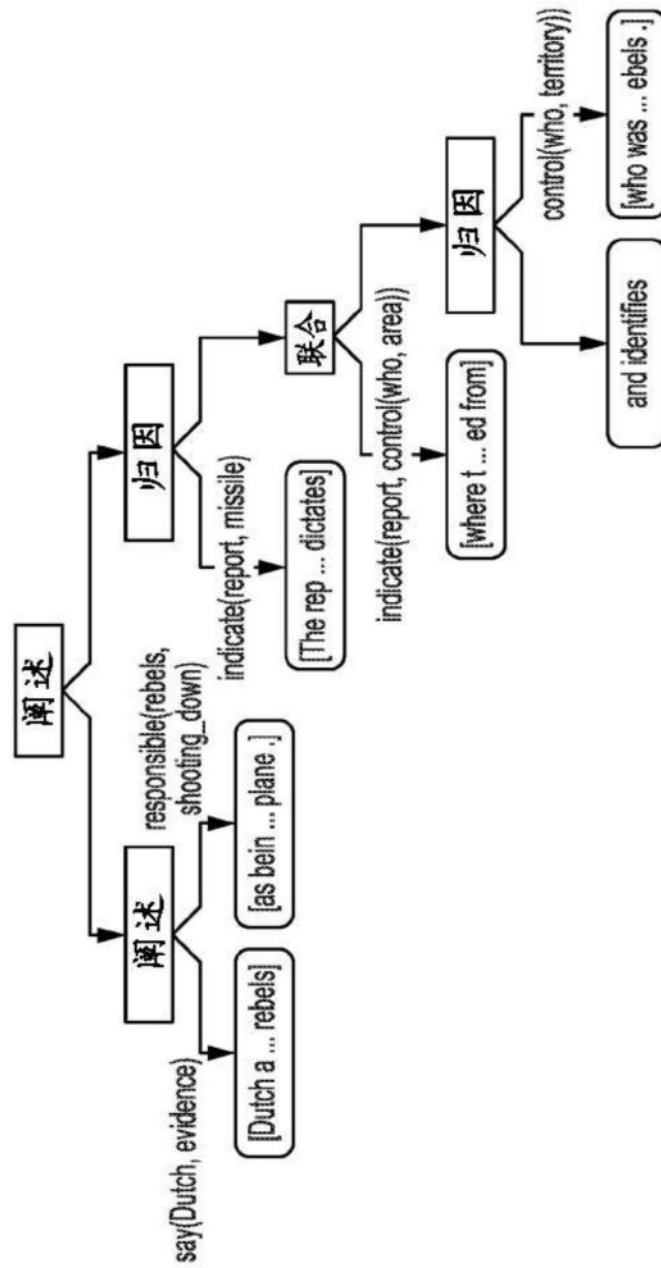


图11

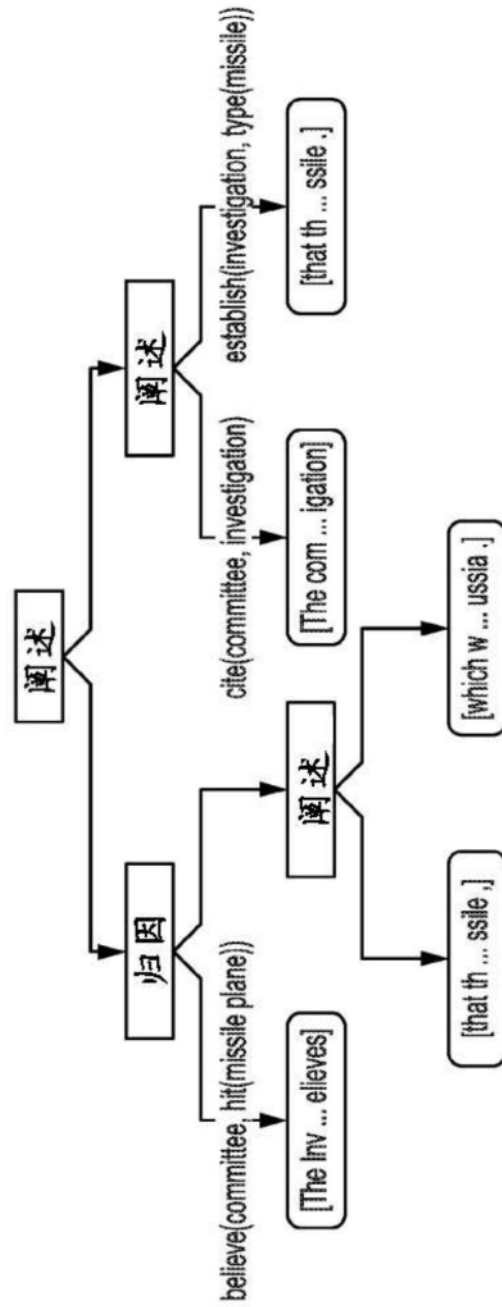


图12

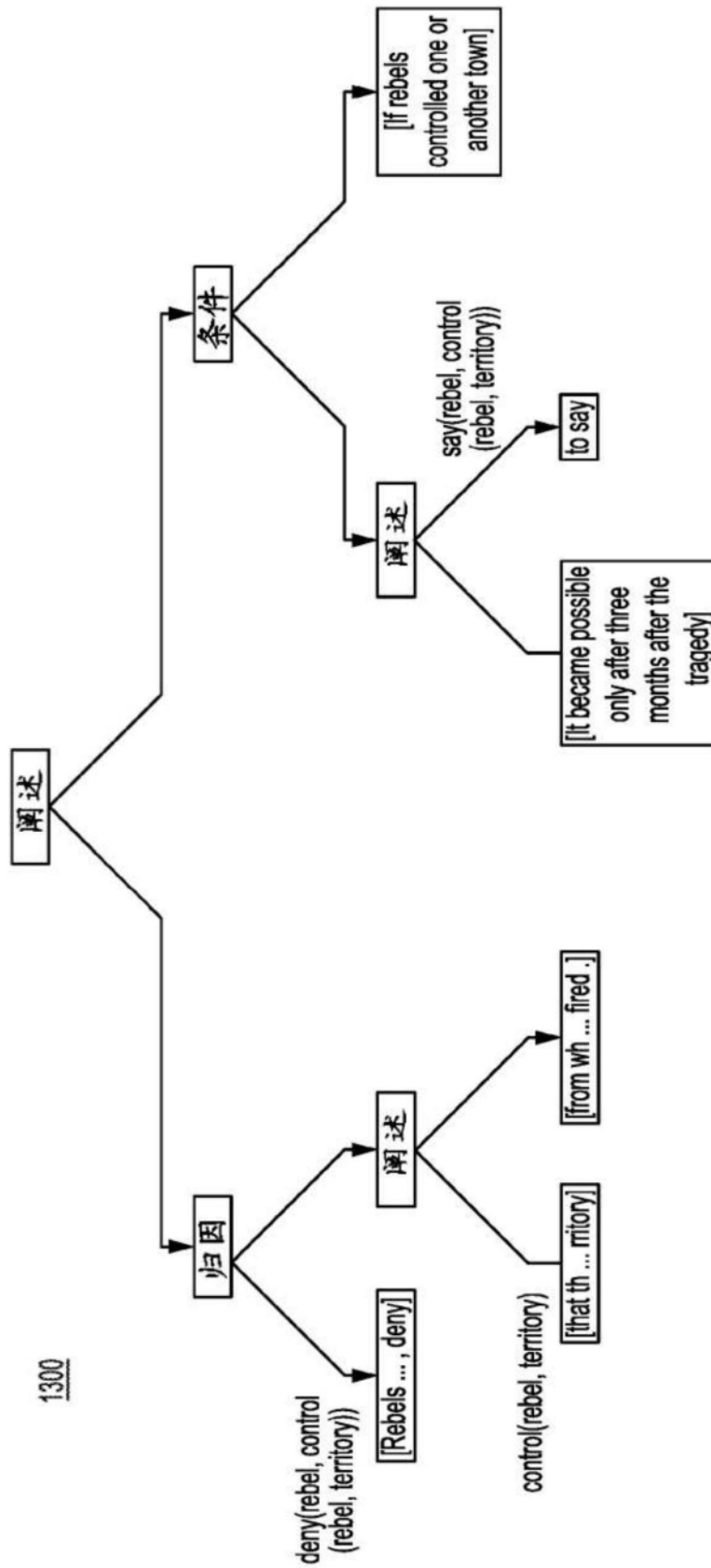


图13

1500

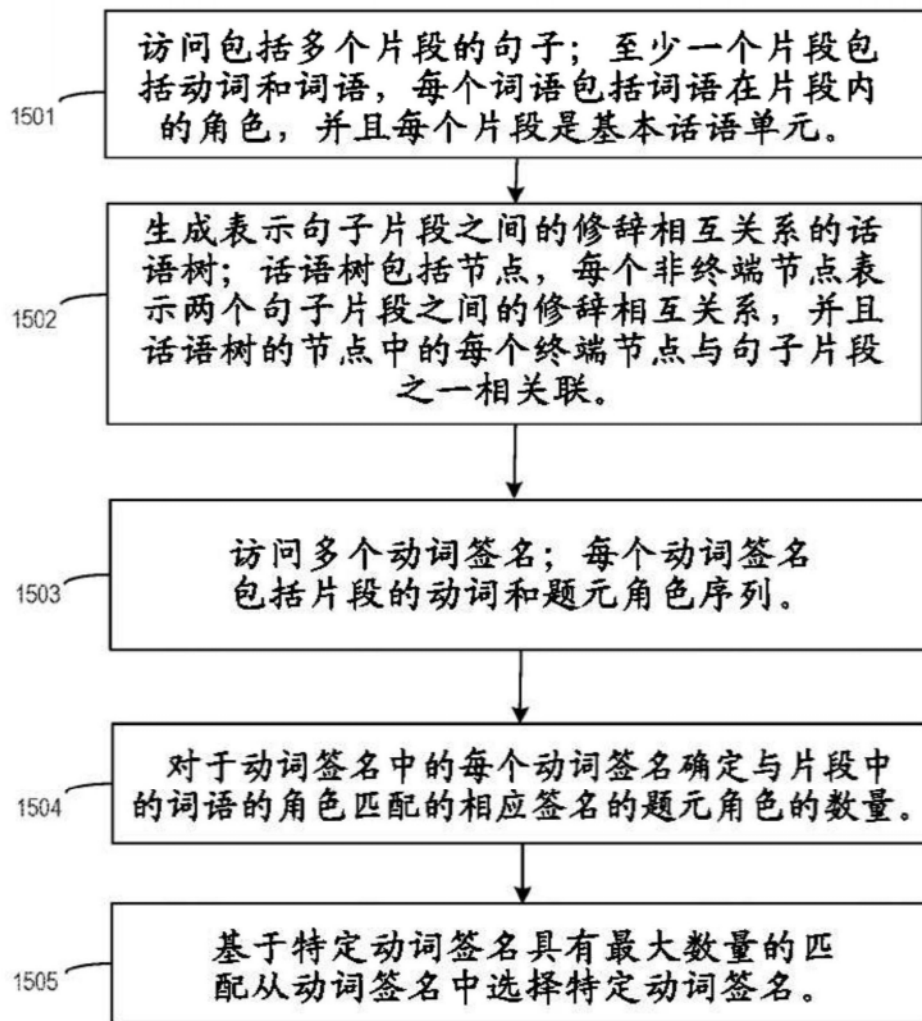


图15

1600

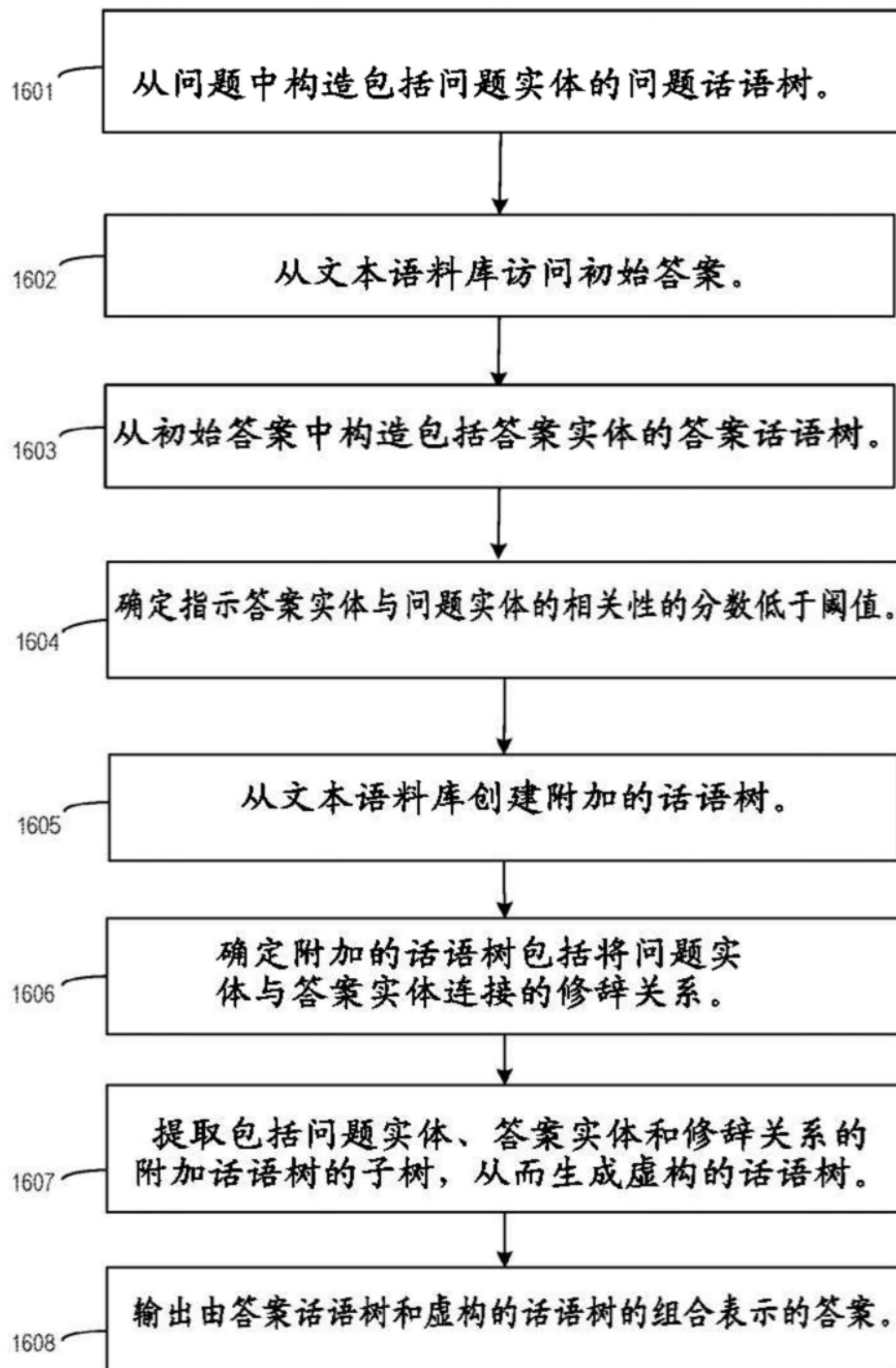


图16

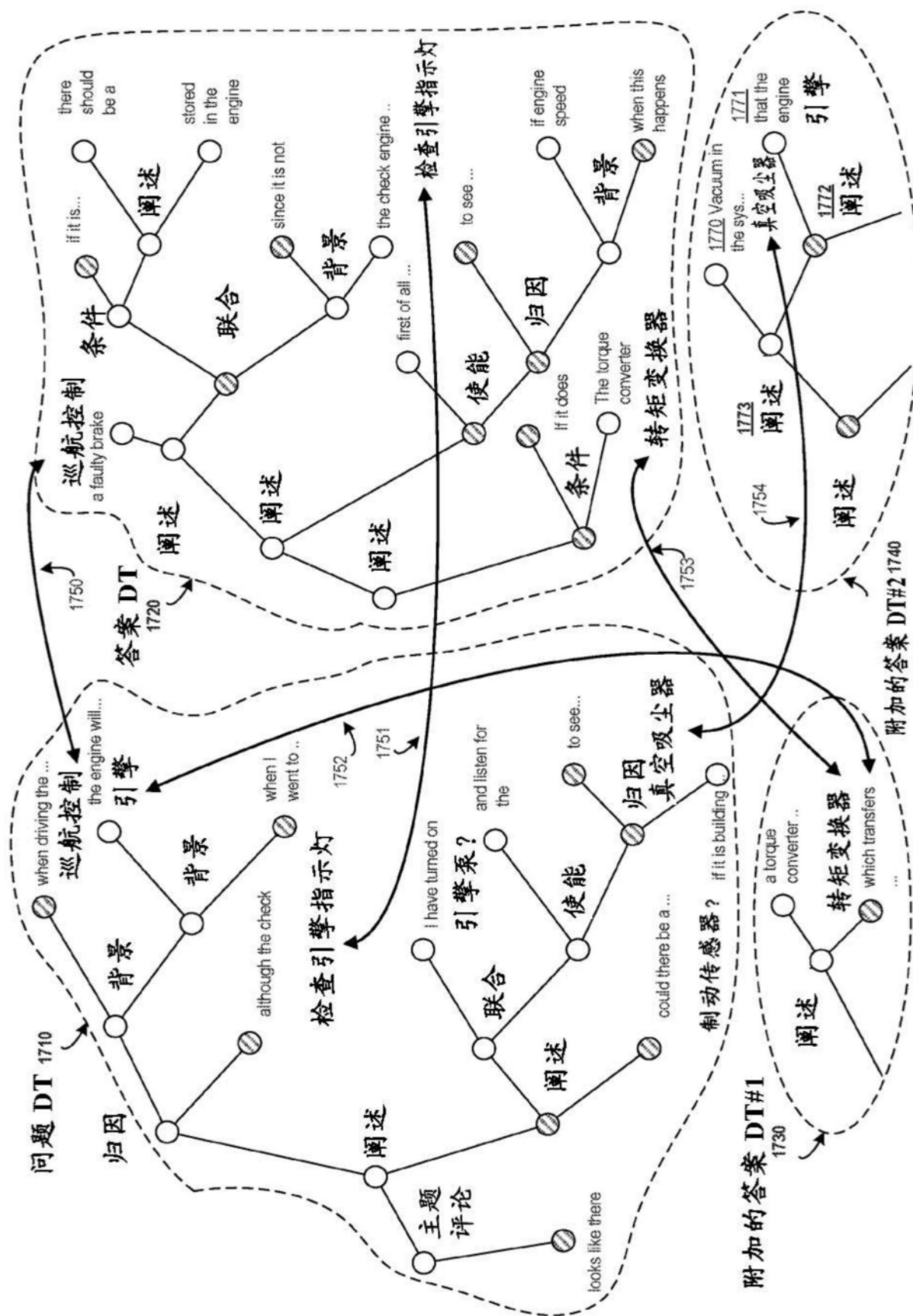


图17

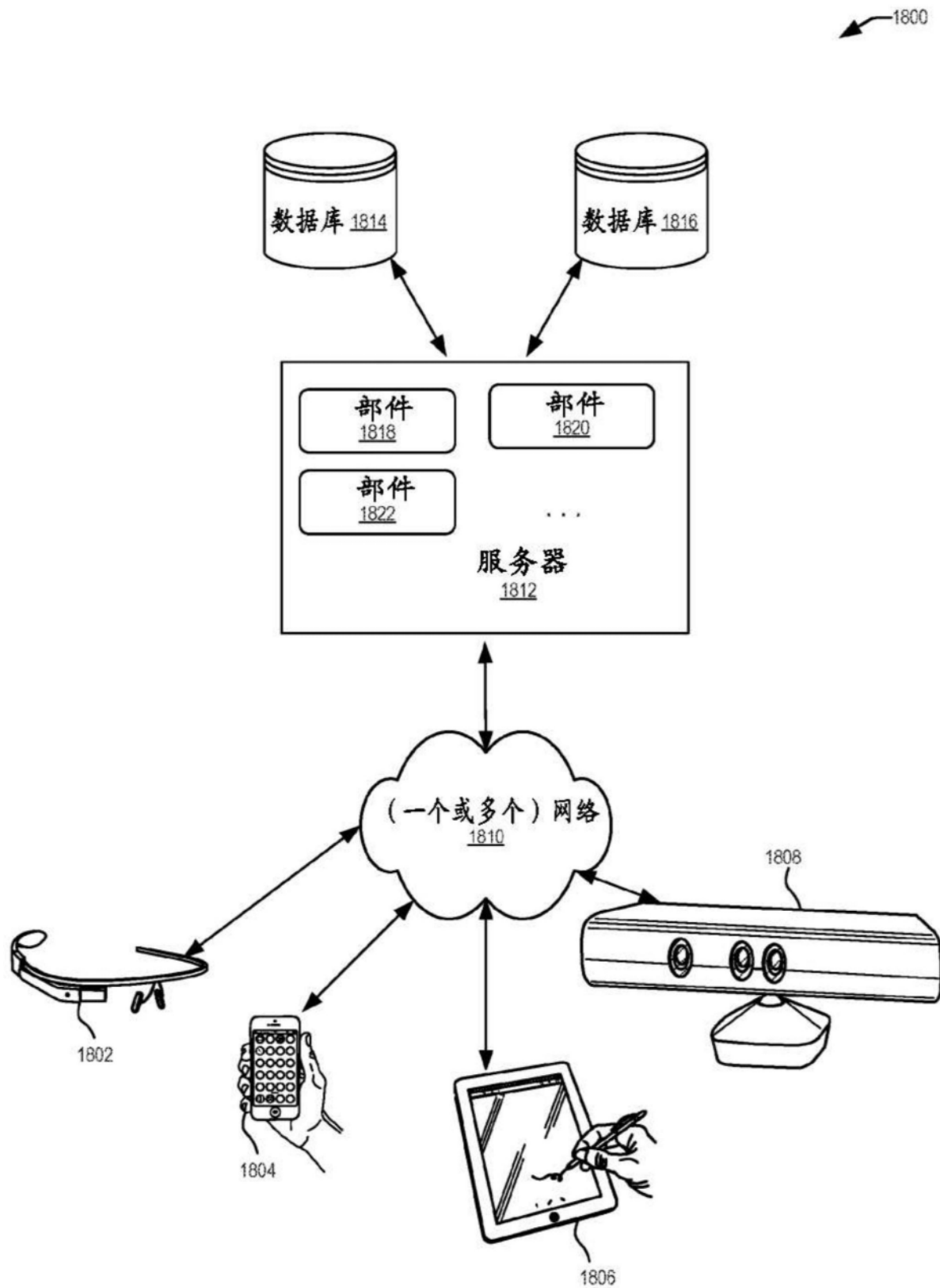


图18

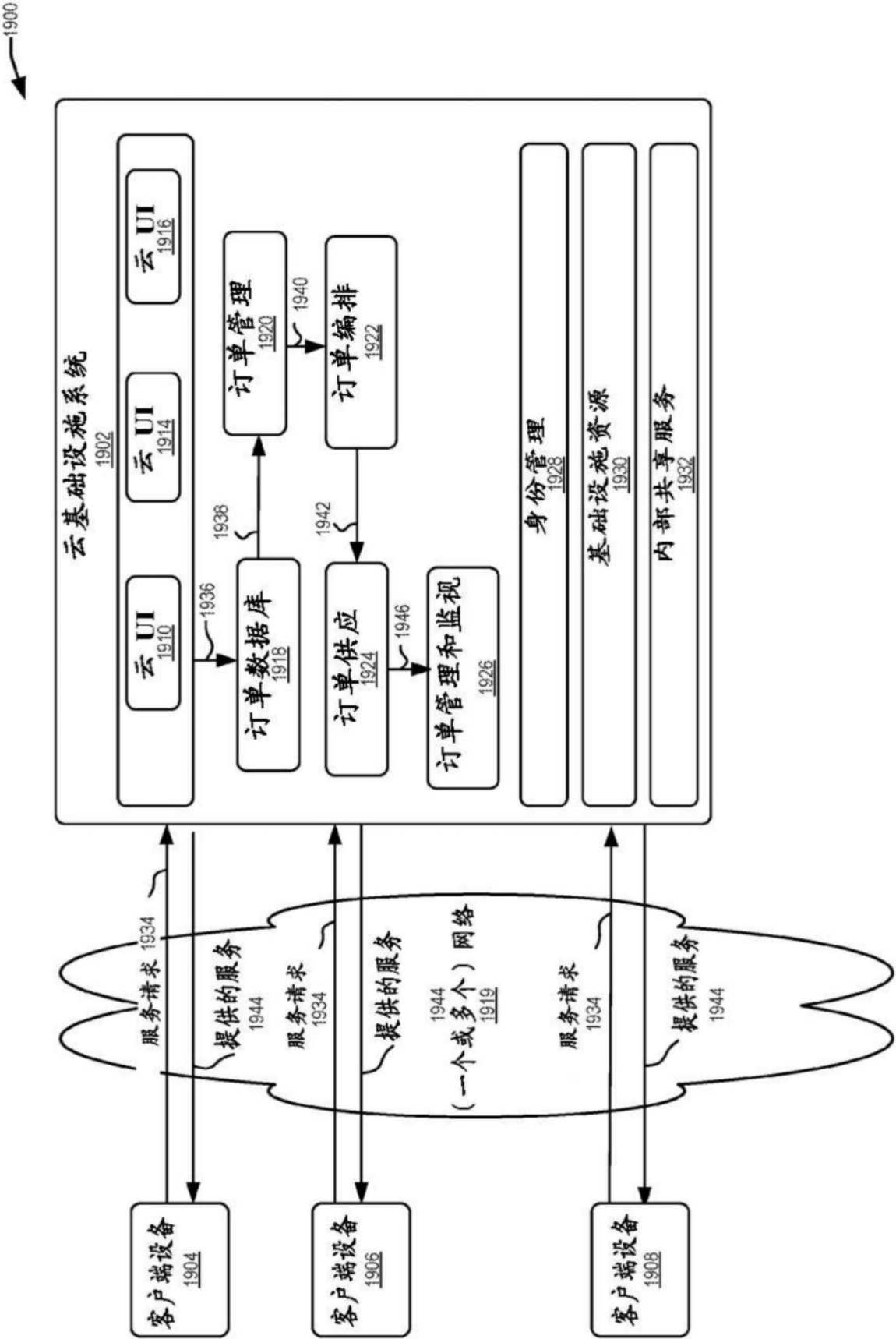


图19

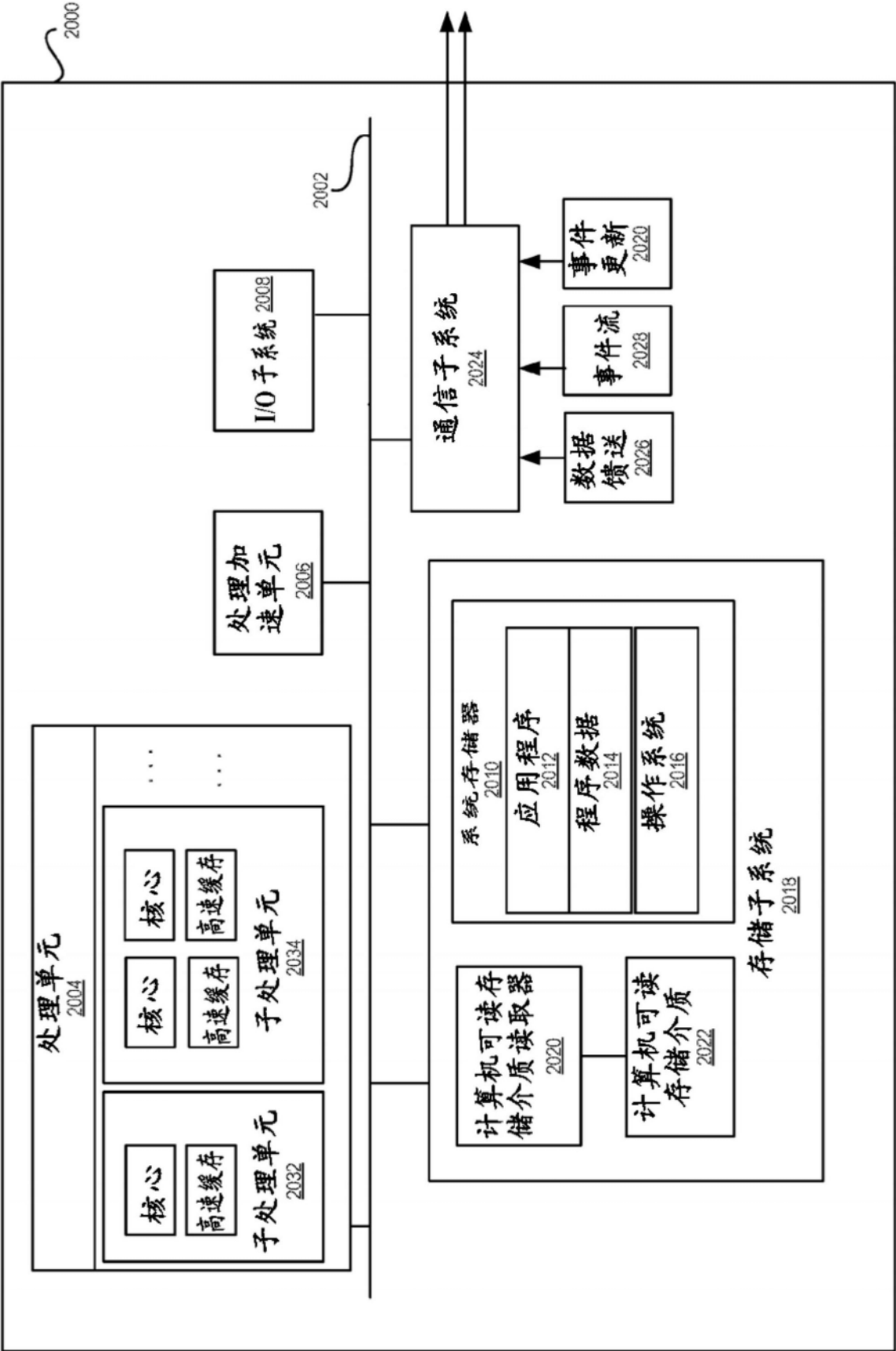


图20