(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110727782 A
(43)申请公布日 2020.01.24

(21)申请号 201911004179.7

(22)申请日 2019.10.22

(71)申请人 苏州思必驰信息科技有限公司
地址 215123 江苏省苏州市苏州工业园区新平街388号腾飞创新园14栋

(72)发明人 陈海龙 杜斌

(74)专利代理机构 北京商专永信知识产权代理事务所(普通合伙) 11400
代理人 黄谦 车江华

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/36(2019.01)
G06F 16/335(2019.01)
G06F 16/33(2019.01)

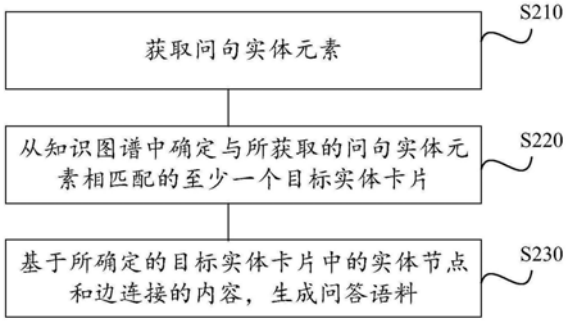
权利要求书2页 说明书6页 附图5页

(54)发明名称

问答语料生成方法及系统

(57)摘要

本发明公开一种问答语料生成方法及系统，该方法包括：获取问句实体元素；从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片，所述知识图谱包括多个实体卡片，所述实体卡片包括多个实体节点和关于不同实体节点之间的边连接；基于所确定的目标实体卡片中的实体节点和边连接的内容，生成问答语料。由此，利用知识图谱通过机器处理方式而生成问答语料，可以解决因人工录入方式而导致的人工成本巨大和问答语料无法标准化的问题。



CN 110727782 A

1.一种问答语料生成方法,包括:

获取问句实体元素;

从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片,所述知识图谱包括多个实体卡片,所述实体卡片包括多个实体节点和关于不同实体节点之间的边连接;

基于所确定的目标实体卡片中的实体节点和边连接的内容,生成问答语料。

2.如权利要求1所述的方法,其中,所述获取问句实体元素包括:

获取问句模板;

按照模板实体元素提取条件,从所述问句模板中提取所述问句实体元素。

3.如权利要求2所述的方法,其中,所述获取问句模板包括:

获取至少一个候选问句模板;

针对各个候选问句模板,确定该候选问句模板是否满足所述模板实体元素提取条件;以及

在所述候选问句模板中选择满足所述模板实体元素提取条件的问句模板。

4.如权利要求2或3所述的方法,还包括:

获取问答模板对,每一问答模板对包括具有针对同一问句实体元素的槽位的问句模板和答案模板;

其中,所述基于所确定的目标实体卡片中的实体节点和边连接的内容,生成问答语料包括:

基于所确定的目标实体卡片中的实体节点和边连接的内容对所述问句模板进行填槽,以生成针对所述问句实体元素的问句语句;

基于所确定的目标实体卡片中的实体节点和边连接的内容对所述答案模板进行填槽,以生成针对所述问句实体元素的答案语句;

基于所述问句语句和所述答案语句生成问答语料。

5.如权利要求4所述的方法,其中,针对同一问句实体元素存在多个问答模板对。

6.如权利要求1所述的方法,其中,所述问句实体元素包括实体类别。

7.如权利要求1所述的方法,其中,所述问答语料包括以下中的至少一种语料类型:文本语料、语音语料和视频语料。

8.一种问答语料生成系统,包括:

实体元素获取单元,被配置为获取问句实体元素;

目标卡片确定单元,被配置为从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片,所述知识图谱包括多个实体卡片,所述实体卡片包括多个实体节点和关于不同实体节点之间的边连接;

问答语料生成单元,被配置为基于所确定的目标实体卡片中的实体节点和边连接的内容,生成问答语料。

9.一种电子设备,其包括: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1-7中任一项所述方法的步骤。

10.一种存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现权利要求1-7中任一项所述方法的步骤。

问答语料生成方法及系统

技术领域

[0001] 本发明属于互联网技术领域，尤其涉及问答语料生成方法及系统。

背景技术

[0002] 随着互联网技术的不断发展，QA系统(question-and-answer，问答系统)在诸如智能客服、机器人领域等方面都取得了很大的进展。

[0003] QA系统的性能与QA系统所采用问答语料样本集有直接的关系，一般问答语料样本集越全，QA系统的性能就越强。目前针对问答语料样本集的确定过程，一般依赖于人工操作。

[0004] 但是，本申请的发明人在实践本申请的过程中发现目前相关技术至少存在以下问题：若问答对数量大，人工录入会耗费较大财力且容易导致较高的出错率，而难以管理。另外，如果针对一个问题存在多种问法，其对应的答案方式还可以有图片、音频和视频等多种方式，导致人工录入就更难以完成。进一步地，因为录入人员的差异化，可能会导致录入操作比较随意和口语化等，不能形成标准化的问答语料。

发明内容

[0005] 本发明实施例提供一种问答语料生成方法及系统，用于至少解决上述技术问题之一。

[0006] 第一方面，本发明实施例提供一种问答语料生成方法，包括：获取问句实体元素；从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片，所述知识图谱包括多个实体卡片，所述实体卡片包括多个实体节点和关于不同实体节点之间的边连接；基于所确定的目标实体卡片中的实体节点和边连接的内容，生成问答语料。

[0007] 第二方面，本发明实施例提供一种问答语料生成系统，包括：实体元素获取单元，被配置为获取问句实体元素；目标卡片确定单元，被配置为从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片，所述知识图谱包括多个实体卡片，所述实体卡片包括多个实体节点和关于不同实体节点之间的边连接；问答语料生成单元，被配置为基于所确定的目标实体卡片中的实体节点和边连接的内容，生成问答语料。

[0008] 第三方面，本发明实施例提供一种电子设备，其包括：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上述方法的步骤。

[0009] 第四方面，本发明实施例提供一种存储介质，其上存储有计算机程序，该程序被处理器执行时实现上述方法的步骤。

[0010] 本发明实施例的有益效果在于：通过将问句实体元素与知识图谱进行匹配而得到对应的目标实体卡片，并利用目标实体卡片中的内容信息来生成问答语料。由此，利用知识图谱通过机器处理方式而生成问答语料，解决了因人工录入方式而导致的人工成本巨大和

问答语料无法标准化的问题。

附图说明

[0011] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0012] 图1示出了根据本发明一实施例的知识图谱的示意图；

[0013] 图2示出了根据本发明第一实施例的问答语料生成方法的流程图；

[0014] 图3示出了根据本发明第二实施例的问答语料生成方法的流程图；

[0015] 图4示出了根据本发明第三实施例的问答语料生成方法的流程图；

[0016] 图5示出了根据本发明第四实施例的问答语料生成方法的流程图；

[0017] 图6示出了根据本发明一实施例的问答语料生成系统的结构框图。

具体实施方式

[0018] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0019] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。

[0020] 本发明可以在由计算机执行的计算机可执行指令的一般上下文中描述，例如程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、元件、数据结构等等。也可以在分布式计算环境中实践本发明，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

[0021] 在本发明中，“模块”、“系统”等等指应用于计算机的相关实体，如硬件、硬件和软件的组合、软件或执行中的软件等。详细地说，例如，元件可以、但不限于是运行于处理器的过程、处理器、对象、可执行元件、执行线程、程序和/或计算机。还有，运行于服务器上的应用程序或脚本程序、服务器都可以是元件。一个或多个元件可在执行的过程和/或线程中，并且元件可以在一台计算机上本地化和/或分布在两台或多台计算机之间，并可以由各种计算机可读介质运行。元件还可以根据具有一个或多个数据包的信号，例如，来自一个与本地系统、分布式系统中另一元件交互的，和/或在因特网的网络通过信号与其它系统交互的数据的信号通过本地和/或远程过程来进行通信。

[0022] 最后，还需要说明的是，在本文中，术语“包括”、“包含”，不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0023] 如图1所示，根据本发明一实施例的知识图谱的示意图，在知识图谱中具有多个实

体卡片，例如针对实体“李白”和实体“苏东坡”的实体卡片，每一个实体卡片都是由边连接和实体节点组成的。如图1的示例中，各个方形框和圆框表示实体卡片所对应的实体节点(例如诗人、李白等)，各个实体节点之间的连线为边连接，其表示实体之间所存在的关联属性(例如职业、朝代、字号等)。

[0024] 应理解的是，知识图谱具有多个实体卡片，还可以包括未在图1中所示出的其他的实体卡片，各个实体卡片的实体的类型在此也不应限制，例如知识图谱还可以包括针对公司实体的实体卡片。并且，随着业务的需要和发展，可以对知识图谱补充和完善更多的实体卡片。

[0025] 如图2所示，本发明第一实施例的问答语料生成方法的流程，包括：

[0026] S210、获取问句实体元素。

[0027] 这里，问句实体元素表示用于描述实体的信息，例如实体类别(公司类实体或个人类实体等)或实体名称(例如XX公司)等，并且通过问句实体元素能够定向至对应的一个或多个实体。

[0028] 在本公开的一个示例中，直接基于用户输入操作而得到对应的问句实体元素。在本公开的另一示例中，获取问句模板，并按照模板实体元素提取条件从问句模板中提取问句实体元素。假设模板实体元素提取条件为“{}”，则针对“{公司}的/CEO/是谁”这个问句模板中的问句实体元素为{公司}。但是，如果问句模板中缺乏“}”，则会导致该问句模板无法满足模板实体元素提取条件(即无法提取出实体元素)，而被确定为无效问句模板。

[0029] S220、从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片。

[0030] 具体地，可以通过将问句实体元素与知识图谱进行文本匹配，从而将知识图谱中含有问句实体元素的实体卡片确定为目标实体卡片。参照如图1的示例，当所获取的问句实体元素为“诗人”时，针对“苏东坡”和“李白”的实体卡片都应被确定为目标实体卡片。

[0031] S230、基于所确定的目标实体卡片中的实体节点和边连接的内容，生成问答语料。

[0032] 示例性地，针对实体卡片“李白”的实体节点和边连接的内容为“朝代”为“唐朝”，则相应的问答语料可以是“李白是什么朝代的?——李白是唐朝的”，“李白是什么职业?-李白是诗人”，等等。由此，通过知识图谱自动生成问答语料。这里，针对问答语料中除了实体节点和边连接的内容之外的内容信息的获取方式，其一方面可以是通过预配置的模板信息填充来实现的，另一方面其还可以是通过语义模型来进行完善的，在此可不加限制。

[0033] 需说明的是，问答语料的语料类型可能是多样化的，例如文本语料、语音语料和/或视频语料。示例性地，当通过图谱解析到文本语料时，可以根据这个文本语料生成对应的语音语料或视频语料。

[0034] 如图3所示，根据本发明第二实施例的问答语料生成方法的流程，包括：

[0035] S310、获取至少一个候选问句模板。

[0036] 这里，可以是采用批量化的方式输入多个候选问句模板，从而提高语料生成效率。

[0037] S320、针对各个候选问句模板，确定该候选问句模板是否满足模板实体元素提取条件。

[0038] S330、在候选问句模板中选择满足模板实体元素提取条件的问句模板。

[0039] 这样，通过S320和S330的操作，可以实现筛除无效问句模板，确保所筛选出的问句

模板能够被用来提取问句实体元素。

[0040] S340、按照模板实体元素提取条件，从问句模板中提取问句实体元素。

[0041] S350、从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片。

[0042] S360、基于所确定的目标实体卡片中的实体节点和边连接的内容，生成问答语料。

[0043] 关于S340～S360的操作，可以结合如图2的示例的描述，在此便不赘述。在本实施例中，通过批量化的候选问句模板查询知识图谱，确定对应的问答语料，极大提高了语料生成的可靠性和操作效率。

[0044] 如图4所示，根据本发明第三实施例的问答语料生成方法的流程，包括：

[0045] S410、获取问答模板对。

[0046] 这里，每一问答模板对包括具有针对同一问句实体元素的槽位的问句模板和答案模板，例如问句实体元素为“公司”，其对应的槽位是该“公司”的名称。

[0047] S420、按照模板实体元素提取条件，从问句模板中提取问句实体元素。

[0048] 这样，基于一个问答模板对能够得出对应的问句实体元素。

[0049] S430、从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片。

[0050] S440、基于所确定的目标实体卡片中的实体节点和边连接的内容对问句模板进行填槽，以生成针对问句实体元素的问句语句。

[0051] S450、基于所确定的目标实体卡片中的实体节点和边连接的内容对答案模板进行填槽，以生成针对问句实体元素的答案语句。

[0052] S460、基于问句语句和答案语句生成问答语料。

[0053] 示例性地，问题模板可以是“(我?想知道)?那?个?${#company}的?CEO是谁呢?”，在第一实体卡片中的company的槽位对应为“腾讯”。相应地，答案模板可以是“${company.name}的CEO是${company.ceo.name}”，在第一实体卡片中的{company.ceo.name}的槽位对应为“马化腾”。

[0054] 在一些实施方式中，针对同一问句实体元素存在多个问答模板对，例如每个问答模板对分别对应于不同的问答样式或问答标的。例如，针对问答实体元素“company”可能存在一个问答模板对是针对“CEO”的问答标的，而在另一个问答模板对中是针对“注册资本”的问答标的。另外，问答样式可以实现针对同一问答标的下问答语料的样式的多样化，例如第一样式为“李白是唐朝的吗?——是的，李白是唐朝的”，第二样式为“李白生活在哪个朝代——李白生活在唐朝”。由此，实现了多样化的问答语料。

[0055] 图5示出了根据本发明第四实施例的问答语料生成方法的流程示意图。在该流程的示例中，将KG(Knowledge Graph,知识图谱)与QA系统(即，问答系统)进行关联，基于KG系统构建QA系统的语料，对同一类事物，定义统一的问题答案对模板，批量生成问答对，供QA系统使用。QA系统可发送成对的问题模板和答案模板，也可只发送问题模板，使用知识卡片自带的默认答案模板来形成问答对语料。

[0056] 如图5所示，QA系统批量发送同一类事物的多个问题模板。然后，判断问题模板是否有效，以确定是否进行KG转QA的操作。例如，判断问题模板是否满足模板实体元素提取条件，并将无效的问题模板丢弃，将有效的问题模板提交给KG系统。

[0057] KG系统经过自然语言处理(NLP)操作,可以得到问题对应的事物概念(即,问句实体元素或实体类别)、对应的目标属性(即边连接所指示的在实体节点之间的属性描述,例如CEO),同时查询此问句实体元素下所有实体卡片。在事物概念中的目标属性包含了默认的答案模板(例如xx的CEO是xxx)。事物概念中的目标属性与问题模板结合,生成问题文本。每个实体卡片的目标属性与答案模板结合,生成答案文本。进而,由问题文本与答案文本组成了这个实体卡片的问答对。针对多个实体卡片,一一对应的生成多个问答对,全部可以作为QA系统的问答语料,完善了QA系统的性能。

[0058] 通过本发明实施例,在KG系统与QA系统之间构建了信息流,为QA系统提供大批量、高质量、可定制、标准化的问答对。由此,通过自动化操作丰富了QA系统的语料数据集,提高了问答性能效果和效率,还节省了人工制作语料的成本。

[0059] 如图6所示,根据本发明一实施例的问答语料生成系统600,包括:实体元素获取单元610,被配置为获取问句实体元素;目标卡片确定单元620,被配置为从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片,所述知识图谱包括多个实体卡片,所述实体卡片包括多个实体节点和关于不同实体节点之间的边连接;问答语料生成单元630,被配置为基于所确定的目标实体卡片中的实体节点和边连接的内容,生成问答语料。

[0060] 上述本发明实施例的系统可用于执行本发明中相应的方法实施例,并相应的达到上述本发明方法实施例所达到的技术效果,这里不再赘述。

[0061] 本发明实施例中可以通过硬件处理器(hardware processor)来实现相关功能模块。

[0062] 另一方面,本发明实施例提供一种存储介质,其上存储有计算机程序,该程序被处理器执行如上的问答语料生成方法的步骤。

[0063] 上述产品可执行本申请实施例所提供的方法,具备执行方法相应的功能模块和有益效果。未在本实施例中详尽描述的技术细节,可参见本申请实施例所提供的方法。

[0064] 本申请实施例的客户端以多种形式存在,包括但不限于:

[0065] (1)移动通信设备:这类设备的特点是具备移动通信功能,并且以提供话音、数据通信为主要目标。这类终端包括:智能手机(例如iPhone)、多媒体手机、功能性手机,以及低端手机等。

[0066] (2)超移动个人计算机设备:这类设备属于个人计算机的范畴,有计算和处理功能,一般也具备移动上网特性。这类终端包括:PDA、MID和UMPC设备等,例如iPad。

[0067] (3)便携式娱乐设备:这类设备可以显示和播放多媒体内容。该类设备包括:音频、视频播放器(例如iPod),掌上游戏机,电子书,以及智能玩具和便携式车载导航设备。

[0068] (4)其他具有数据交互功能的电子装置。

[0069] 以上所描述的系统实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0070] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到各实施方式可借助软件加通用硬件平台的方式来实现,当然也可以通过硬件。基于这样的理解,上述技术

方案本质上或者说对相关技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0071] 最后应说明的是：以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

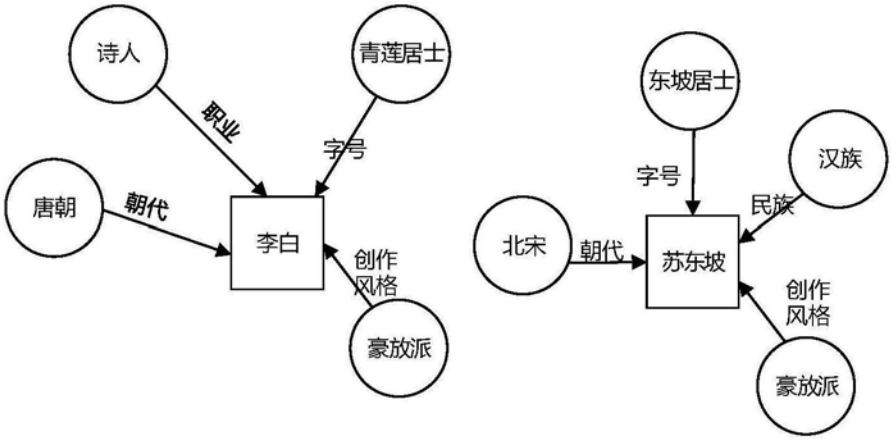

诗人
青莲居士
职业
字号
唐朝
朝代
李白
创作
风格
豪放派
东坡居士
字号
汉族
民族
北宋
朝代
苏东坡
创作
风格
豪放派

图1


S210
获取问句实体元素
S220
从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片
S230
基于所确定的目标实体卡片中的实体节点和边连接的内容，生成问答语料

图2

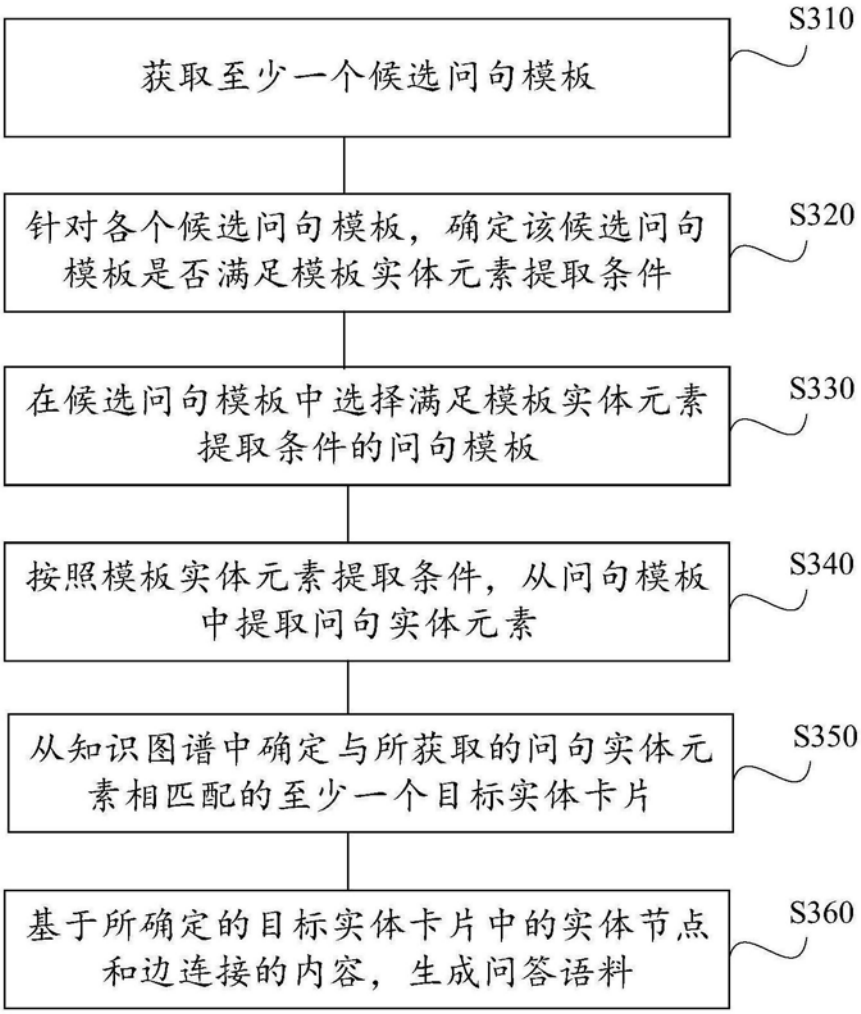

S310
获取至少一个候选问句模板
S320
针对各个候选问句模板，确定该候选问句模板是否满足模板实体元素提取条件
S330
在候选问句模板中选择满足模板实体元素提取条件的问句模板
S340
按照模板实体元素提取条件，从问句模板中提取问句实体元素
S350
从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片
S360
基于所确定的目标实体卡片中的实体节点和边连接的内容，生成问答语料

图3

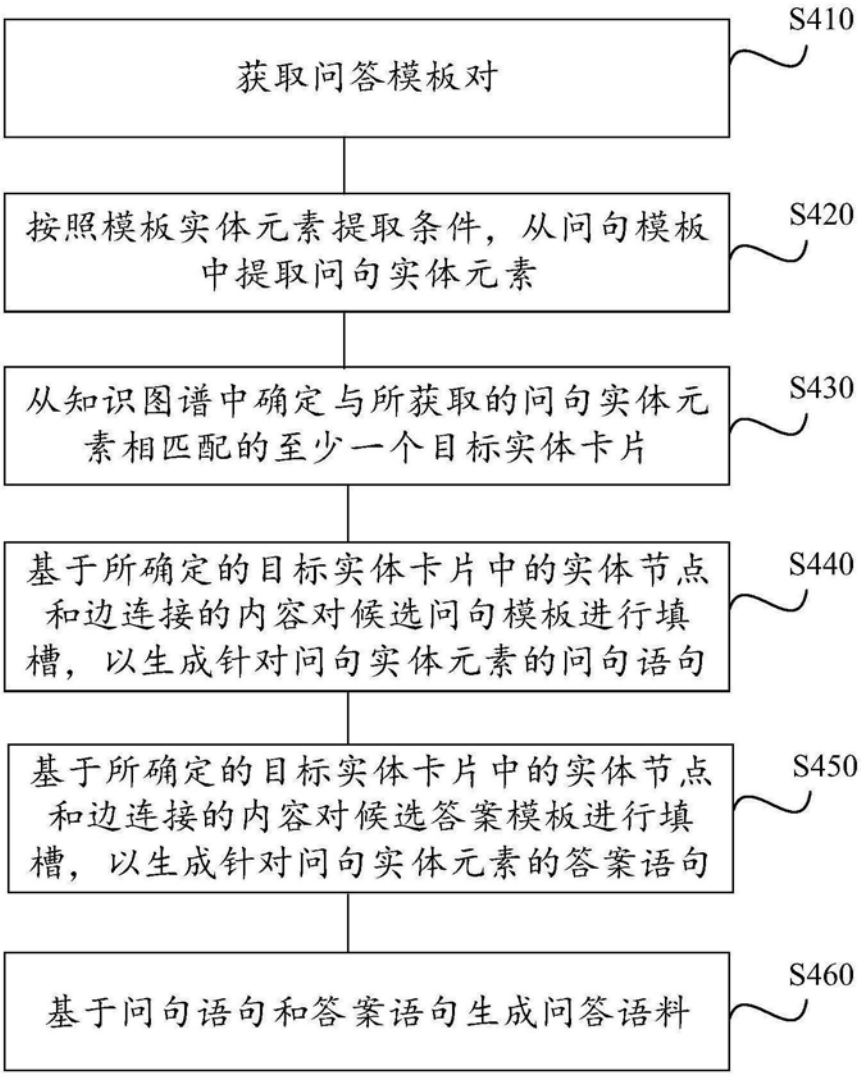

S410
获取问答模板对
S420
按照模板实体元素提取条件，从问句模板中提取问句实体元素
S430
从知识图谱中确定与所获取的问句实体元素相匹配的至少一个目标实体卡片
S440
基于所确定的目标实体卡片中的实体节点和边连接的内容对候选问句模板进行填槽，以生成针对问句实体元素的问句语句
S450
基于所确定的目标实体卡片中的实体节点和边连接的内容对候选答案模板进行填槽，以生成针对问句实体元素的答案语句
S460
基于问句语句和答案语句生成问答语料

图4

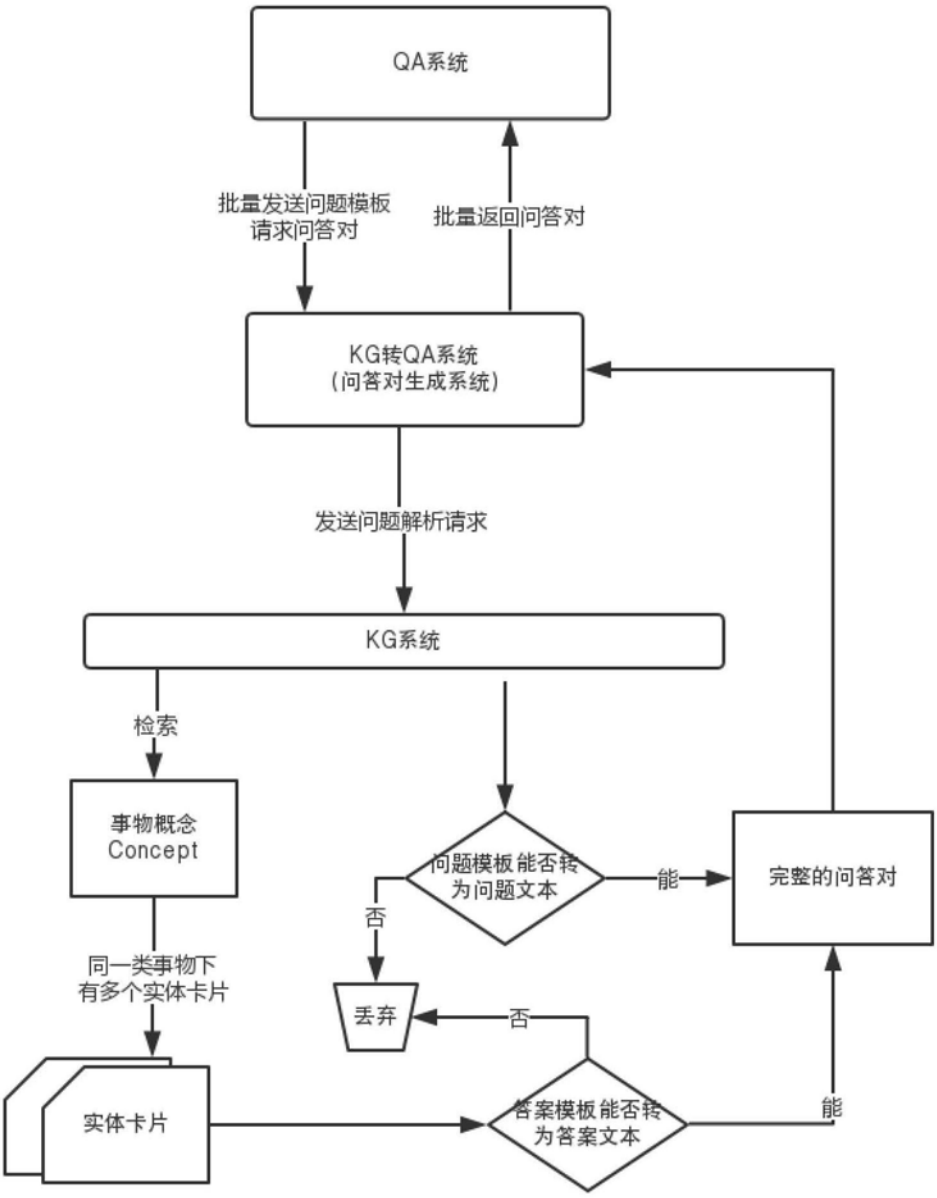

QA系统
批量发送问题模板
请求问答对
批量返回问答对
KG转QA系统
(问答对生成系统)
发送问题解析请求
KG系统
检索
事物概念
Concept
同一类事物下
有多个实体卡片
实体卡片
问题模板能否转
为问题文本
能
否
完整的问答对
丢弃
否
答案模板能否转
为答案文本
能

图5

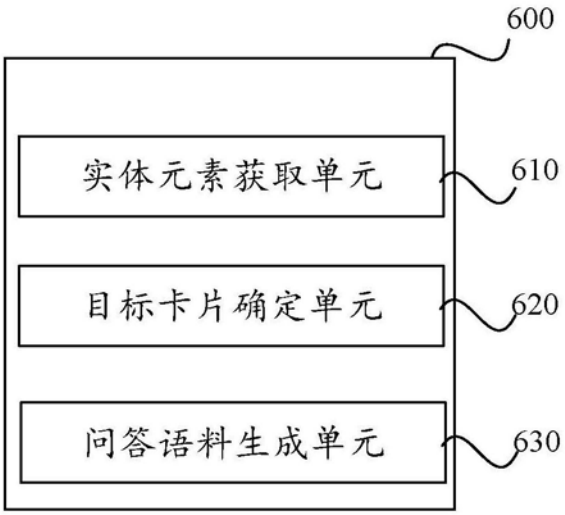


图6