(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112650846 B
(45)授权公告日 2024.08.23

(21)申请号 202110040888.1
(22)申请日 2021.01.13
(65)同一申请的已公布的文献号
申请公布号 CN 112650846 A
(43)申请公布日 2021.04.13
(73)专利权人 北京智通云联科技有限公司
地址 100020 北京市朝阳区慧忠北里219号楼19幢六层601号
(72)发明人 侯志强 柳晶晶 刘锋 谭培波
(74)专利代理机构 北京八月瓜知识产权代理有限公司 11543
专利代理师 李斌
(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 40/205(2020.01)
G06F 40/242(2020.01)
G06F 40/289(2020.01)
G06F 40/30(2020.01)
(56)对比文件
CN 111428483 A,2020.07.17
审查员 周循

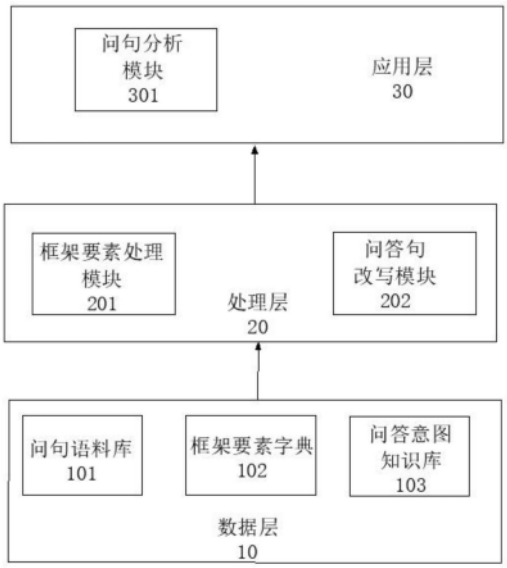
权利要求书1页 说明书6页 附图4页

(54)发明名称

一种基于问句框架的问答意图知识库构建系统及方法

(57)摘要

本发明提供了一种基于问句框架的问答意图知识库构建系统及方法。包括:数据层,包括问句语料库、框架要素字典和问答意图知识库;用于存储文件、读写文件和修改文件;处理层,包括框架要素处理模块和问答句改写模块,用于改写问句;应用层,包括问句分析模块,用于输出改写问句所形成的候选目标词串。该基于问句框架的问答意图知识库构建系统及方法改善了现有技术中识别框架要素困难且无法自动得到问句解析的答句形式的问题。



CN 112650846 B

1.一种基于问句框架的问答意图知识库构建系统,其特征在于,包括:

数据层,包括问句语料库、框架要素字典和问答意图知识库;用于存储文件、读写文件和修改文件;

所述框架要素字典的格式包括框架的名称和框架要素代号;所述框架要素字典包括问句解析,所述问句解析包括第一层和第二层,所述第一层用于序列解析,所述第二层用于蕴含关系和层次结构解析;

处理层,包括框架要素处理模块和问答句改写模块,用于改写句子;

应用层,包括问句分析模块,用于输出改写句子所形成的候选目标词串。

2.根据权利要求1所述的基于问句框架的问答意图知识库构建系统,其特征在于,所述问句语料库包括序号、问句来源和问句,用来记录所述问句的相关信息。

3.根据权利要求1所述的基于问句框架的问答意图知识库构建系统,其特征在于,所述问答意图知识库包括问句目标词串和问答意图解析,所述问答意图解析包括第一部分和第二部分,所述第一部分为框架的名称,所述第二部分为答句模板。

4.根据权利要求1所述的基于问句框架的问答意图知识库构建系统,其特征在于,所述框架要素处理模块用于从所述框架要素字典中查找出词串。

5.根据权利要求4所述的基于问句框架的问答意图知识库构建系统,其特征在于,所述改写模块用于对所述句子中的词串进行字符替换,完成对所述句子的改写,每一次改写后的句子将作为新的原始句子加入到改写句子集合中进行累加,直到所有的框架要素字符串都使用过,得到改写句子集合。

6.根据权利要求5所述的基于问句框架的问答意图知识库构建系统,其特征在于,所述问句分析模块用于建立读入问句列表,对所述改写句子所形成的框架目标词串按照词串长度进行逆向排序输出。

7.一种基于问句框架的问答意图知识库构建方法,其特征在于,所述方法具体包括:

S101,根据框架要素字典和问句文件构建句子框架要素字典;

S102,对所述句子框架要素字典进行循环;

S103,对现有句子目标词串集合进行循环,形成新的候选候选目标词串集合,将句子保留在新的候选目标词串集合中;

S104,用句子框架要素字典替换目标词串中的对应的词,更新所述候选目标词串集合;

S105,按候选目标词串长度排序,输出候选目标词串。

8.根据权利要求7所述的基于问句框架的问答意图知识库构建方法,其特征在于,构建句子框架要素字典具体包括:查找所述问句文件中的每个句子,当所述框架要素字典中的词在所述句子中出现时,将所述句子收集在所述框架要素字典中形成句子框架要素字典。

一种基于问句框架的问答意图知识库构建系统及方法

技术领域

[0001] 本发明涉及构建问答意图知识库技术领域,尤其是涉及一种基于问句框架的问答意图知识库构建系统及方法。

背景技术

[0002] 句子意图(句子框架)就是句子的在现实物质世界的所指即语义,语义有很多种,一般采用框架语义学(FrameNet)的方法,根据所处的场景决定框架的名称和框架要素,根据句子中的谓语或者动词定义框架的目标词。这种以谓语或者动词这个句子的一部分来定义整个句子的目标词并决定框架要素,在实际中出现以下问题:

[0003] (1)实体歧义无法消除,无法识别框架要素

[0004] 比如“毛坝3井深是多少”,这里的“毛坝3“和“毛坝3井”都是2个真实存在的但完全不同类型的实体,那么问句中的实体到底是“毛坝3井深”还是“毛坝3井深”呢?这个歧义问题在词层面无法解决,只有在更高的句子层面通过知识库才能把提问者的真实意图和要素校正出来。

[0005] (2)无动词句子无法识别框架

[0006] 英文是一种以动词为主的语言,因此以动词为主来识别框架和框架要素是成功的,但是很明确,但是在问句意图(框架)识别的时候就无法定义出目标词,这样就无法确定问句的框架和框架要素,导致无法对问句以及答句进行解析。

[0007] 比如“毛坝3井深”,这个句子在问答场景下的语义是非常明确的,就是问“毛坝3”这个集气站所包含的各井的“井深”,但是问句中没有动词只有名词,NLP无法识别这种句子的框架和框架要素。

[0008] (3)无虚词序列也无法识别框架

[0009] 借鉴槽位方法,在句子中把实体去掉,以留下的虚词序列作为目标词进行框架识别,这样的结果由于只应用了一半的信息,因此,也无法识别句子的框架和框架要素。

[0010] 对于有虚词的句子如“毛坝3井深是多少”可以通过去掉实体词“毛坝3”、“井深”并保留槽位位置得到一个虚词序列目标词“,,是多少”,通过这个目标词可以识别句子的框架,但是对于没有虚词的句子比如“毛坝3井深”,这种虚词序列方法也无法识别句子的框架和框架要素。

[0011] (4)对问句的解析不能自动得到答句的形式

[0012] 由于问句和答句是成对出现的,问句不同答句也不同,以保持问句和答句的用词、语气、语义的一致性。但是单独对问句进行而不考虑答句的话,就不能得到符合场景和语义的流畅的答句形式。

发明内容

[0013] 本发明的目的在于提供一种基于问句框架的问答意图知识库构建方法,该基于问句框架的问答意图知识库构建方法能够解决现有技术中识别框架要素困难且无法自动得

到问句解析的答句形式的问题。

[0014] 为了实现上述目的,本发明提供如下技术方案:

[0015] 一种基于问句框架的问答意图知识库构建系统,包括:数据层,包括问句语料库、框架要素字典和问答意图知识库;用于存储文件、读写文件和修改文件;

[0016] 处理层,包括框架要素处理模块和问答句改写模块,用于改写句子;

[0017] 应用层,包括问句分析模块,用于输出改写句子所形成的候选目标词串。

[0018] 在上述技术方案的基础上,本发明还可以做如下改进:

[0019] 进一步地,所述问句语料库包括序号、问句来源和问句,用来记录所述问句的相关信息。

[0020] 进一步地,所述框架要素字典的格式包括框架的名称和框架要素代号;所述框架要素字典包括问句解析,所述问句解析包括第一层和第二层,所述第一层用于序列解析,所述第二层用于蕴含关系和层次结构解析。

[0021] 进一步地,所述问答意图知识库包括问句目标词串和问答意图解析,所述问答意图解析包括第一部分和第二部分,所述第一部分为框架的名称,所述第二部分为答句模板。

[0022] 进一步地,所述框架要素处理模块用于从所述框架要素字典中查找出词串。

[0023] 进一步地,所述改写模块用于对所述句子中的词串进行字符替换,完成对所述句子的改写,每一次改写后的句子将作为新的原始句子加入到改写句子集合中进行累加,直到所有的框架要素字符串都使用过,得到改写句子集合。

[0024] 进一步地,所述问句分析模块用于建立读入问句列表,对所述改写句子所形成的框架目标词串按照词串长度进行逆向排序输出。

[0025] 一种基于问句框架的问答意图知识库构建方法,所述方法具体包括:

[0026] S101,根据框架要素字典和问句文件构建句子框架要素字典;

[0027] S102,对所述句子框架要素字典进行循环;

[0028] S103,对现有句子目标词串集合进行循环,形成新的候选候选目标词串集合,将句子保留在新的候选目标词串集合中;

[0029] S104,用句子框架要素字典替换目标词串中的对应的词,更新所述候选目标词串集合;

[0030] S105,按候选目标词串长度排序,输出候选目标词串。

[0031] 进一步地,构建句子框架要素字典具体包括:查找所述问句文件中的每个句子,当所述框架要素字典中的词在所述句子中出现时,将所述句子收集在所述框架要素字典中形成句子框架要素字典。

[0032] 本发明具有如下优点:

[0033] 本发明中的基于问句框架的问答意图知识库构建系统及方法,最大限度的采用了问句中所有字以及语序的信息,最大限度地保留了问句目标词串的语义,不仅能有效地消除词级的对象歧义,也能通过问句的目标词串实现对问句的框架名称的识别、框架要素的解析和答句的生成,解决了问答系统中问答意图的解析;解决了现有技术中识别框架要素困难且无法自动得到问句解析的答句形式的问题。

附图说明

[0034] 为了更清楚地说明本发明具体实施方式或现有技术中的技术方案,下面将对具体实施方式或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图是本发明的一些实施方式,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。

[0035] 图1为本发明实施例中问答意图知识库构建系统的示意图;

[0036] 图2为本发明实施例中问答意图知识库构建方法的流程示意图;

[0037] 图3为本发明实施例中问答框架名称编码规则以及框架要素的定义示意图;

[0038] 图4为本发明实施例中问句语料库的格式的示意图;

[0039] 图5为本发明实施例中框架要素字典格式的示意图;

[0040] 图6为本发明实施例中问答图知识库格式的示意图。

[0041] 附图标记说明:

[0042] 数据层10,问句语料库101,框架要素字典102,问答意图知识库103,处理层20,框架要素处理模块201,问答句改写模块202,应用层30,问句分析模块301。

具体实施方式

[0043] 下面将结合实施例对本发明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做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。

[0044] 如图1所示,本发明提供了一种基于问句框架的问答意图知识库构建系统,包括:

[0045] 数据层10,包括问句语料库101、框架要素字典102和问答意图知识库103;用于存储文件、读写文件和修改文件;

[0046] 处理层20,包括框架要素处理模块201和问答句改写模块202,用于改写句子;

[0047] 应用层30,包括问句分析模块301,用于输出改写句子所形成的候选目标词串。

[0048] 句子意图(句子框架)就是句子的在现实物质世界的所指即语义,语义有很多种,一般采用框架语义学(FrameNet)的方法,根据所处的场景决定框架的名称和框架要素,根据句子中的谓语或者动词定义框架的目标词。

[0049] 比如社交场景下定义一个名为“言谈交际”框架,一般具有“发送者”、“接收者”、“信息”、“媒介”等框架要素,而句子中包含的动词如“说”、“讲”、“谈话”、“命令”、“告诉”、“讨论”、“提醒”、“问”、“承诺”、“警告”、“威胁”等都是这个框架的目标词。

[0050] 对于具体的句子,通过识别目标词确定句子所属的框架。如“张三告诉李四机场在哪里”,“告诉”这个目标词说明该句属于“言谈交际”框架,“张三”、“李四”、“机场在哪里”等都是该框架的框架要素。

[0051] 将问句中的实体词用框架要素进行替换,构成一个虚实结合的完整词串,这样最大限度地利用了句子中所有字的信息,构造出最完整的句子框架目标词串。

[0052] 具体操作是,首先根据问句确定整个问答(包括问句和答句)框架以及问答框架要素代码(如框架编号F111,框架要素代码T,时间,O对象,P参数)。

[0053] 其次,根据句子的框架和框架要素代码,构建框架要素字典102;然后用框架要素

代码替换原问句和答句中对应的实体词,对问句和答句进行改写(如“2020年1月3号T1001井产气量是多少”改写为“TOP是多少F111#TOP是Q”)。

[0054] 将改写后的问句定义为在所选框架下的目标词用于识别框架,将改写后的答句作为答句模板和问句一起构成对应该句的问答框架;将所有问句的问答框架放在一起,构成整个问答知识库。

[0055] 进一步地,如图4所示,所述问句语料库101包括序号、问句来源和问句,用来记录所述问句的相关信息。

[0056] 用来记录问句相关信息,这些信息可以扩展如增加地域和提问人等信息,为未来进行更精准的问答做准备。

[0057] 进一步地,如图5所示,所述框架要素字典102的格式包括框架的名称和框架要素代号;所述框架要素字典102包括问句解析,考虑到中文自然语言中时间的粒度以及自然语言的蕴含关系如“产多少油”,所述问句解析包括第一层和第二层,所述第一层用于序列解析,所述第二层用于蕴含关系和层次结构解析。其中问答句框架名称定义和框架要素定义如图2所示,其中问句定义TOPVM5个元素,答句定义为TOPVMQ等6个元素。

[0058] 进一步地,如图6问答图知识库格式所示,所述问答意图知识库103包括问句目标词串和问答意图解析,所述问答意图解析包括第一部分和第二部分,其中的问答意图解析包括用“@@@”隔开的第一部分和第二部分,所述第一部分为框架的名称,所述第二部分为答句模板。

[0059] 进一步地,所述框架要素处理模块201用于从所述框架要素字典102中查找出词串。

[0060] 进一步地,所述改写模块用于对所述句子中的词串进行字符替换,完成对所述句子的改写,每一次改写后的句子将作为新的原始句子加入到改写句子集合中进行累加,直到所有的框架要素字符串都使用过,得到改写句子集合。

[0061] 原问句经改写后的目标词串集合,既包含了原来句子虚词部分构建的句式,也包含了实词部分构建的框架或者槽位,因此这种方法利用了句子中信息,是对句子框架目标词最完整的构建方式。

[0062] 进一步地,所述问句分析模块301用于建立读入问句列表,对改写问句所形成的框架目标词串按照词串长度进行逆向排序输出。供人们选择和校验合适的框架目标词串用。对于一句简单问句“401-1井产气量是多少”,由于字典的数量很大,一般在100万量级,将输出163个可能的框架目标词串,其中只有“OP是多少”这一句是正确的问句框架目标词串。问句越复杂,包含的框架要素越多,输出的框架目标词串的数量就越大。

[0063] 一种基于问句框架的问答意图知识库构建方法,所述方法具体包括:

[0064] S101,构建句子框架要素字典102;

[0065] 本步骤中,根据框架要素字典102和问句文件构建句子框架要素字典102;打开框架要素字典102,将整个字典读入内存,以加快处理速度。打开输入问句文件,实现对问句文件中的每一句进行处理;按句读入问句文件中的句子。构建句子框架要素字典102dic每个问句所对应的框架要素是不一样的,因此,需要对每句单独进行查找。

[0066] 只要框架要素字典102中的词在句子中出现,都要收集在句子框架要素字典102中。这个句子框架要素字典102的数量很大,而且可能存在复杂的包含关系,比如“1”可能是

月份数据也可能是实体的井号,这时需要全部收入句子框架要素字典102中。

[0067] S102,对句子框架要素字典102进行循环;

[0068] 本步骤中,对所述句子框架要素字典102进行循环;按照句子框架要素字典102dic进行循环;构成对整个要素字典的循环。对字典中出现每一种情况都要进行穷举,检查每一种句子改写构造候选目标词串的可能性。

[0069] S103,对现有句子目标词串集合进行循环;

[0070] 本步骤中,对现有句子目标词串集合进行循环,形成新的候选候选目标词串集合,将句子保留在新的候选目标词串集合中;对现有句子目标词串集合进行循环;构成目标词串集合对单个要素字典的更新。因为句子每替换一个要素字典中的词,候选目标词串集合的数量都将增大,因此这里的现有句子目标词串集合是个可变的不断增加的集合。

[0071] 这里是对所有的目标词串进行一次更新。将句子保留在新的候选目标词串集合中;将使原始句一直保留在目标候选词串集合中,实现候选词串集合的扩张。采用集合新增的方式,实现句子的自动去重。

[0072] S104,更新候选目标词串集合;

[0073] 本步骤中,用句子框架要素字典102替换目标词串中的对应的词,更新所述候选目标词串集合;用句子框架要素字典102dic替换目标词串中的对应的词;新词替换后的句子也要增加到目标词串集合中。需要注意的是,如果句子里面有几个相同的要素词,每次也只能替换其中的一个要素词,剩下的要等后来的要素词进行替换。

[0074] 与步骤S102一起构成对下一个要素词的处理,关键点在于目标词串集合是一个更新了的新集合,累加有从原始句子到中间每个字典处理后的所有不重复的候选目标词串;

[0075] S105,输出候选目标词串;

[0076] 本步骤中,按候选目标词串长度排序,输出候选目标词串。

[0077] 进一步地,构建句子框架要素字典102具体包括:查找所述问句文件中的每个句子,当所述框架要素字典102中的词在所述句子中出现时,将所述句子收集在所述框架要素字典102中形成句子框架要素字典102。

[0078] 按候选目标词串长度排序;按候选目标词串长度由小到大排序,这是因为一般而言,字数越多的长词语义越明确,因此替代后,整个目标词串越短的越有可能是所需要的正确的目标词串。这样的排序可以减少人查找正确的目标词串的时间。输出所有的候选目标词串文件;对于每一问句构造一个目标词串序列,然后记录在一个文件里面进行输出,供人工检验用。

[0079] 该基于问句框架的问答意图知识库103构建方法使用过程如下:

[0080] 使用时,操作人员根据框架要素字典102和问句文件构建句子框架要素字典102;对所述句子框架要素字典102进行循环;对现有句子目标词串集合进行循环,形成新的候选候选目标词串集合,将句子保留在新的候选目标词串集合中;用句子框架要素字典102替换目标词串中的对应的词,更新所述候选目标词串集合;按候选目标词串长度排序,输出候选目标词串。

[0081] 在本发明的描述中,需要理解的是,术语"中心"、"纵向"、"横向"、"长度"、"宽度"、"厚度"、"上"、"下"、"前"、"后"、"左"、"右"、"竖直"、"水平"、"顶"、"底"、"内"、"外"、"顺时针"、"逆时针"等指示的方位或位置关系为基于附图所示的方位或位置关系,仅是为了

便于描述本发明和简化描述,而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作,因此不能理解为对本发明的限制。

[0082] 此外,术语"第一"、"第二"仅用于描述目的,而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此,限定有"第一"、"第二"的特征可以明示或者隐含地包括或者更多个所述特征。在本发明的描述中,"多个"的含义是两个或两个以上,除非另有明确具体的限定。此外,术语"安装"、"相连"、"连接"应做广义理解,例如,可以是固定连接,也可以是可拆卸连接,或一体地连接;可以是机械连接,也可以是电连接;可以是直接相连,也可以通过中间媒介间接相连,可以是两个元件内部的连通。对于本领域的普通技术人员而言,可以具体情况理解上述术语在本发明中的具体含义。

[0083] 最后应说明的是:以上各实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述各实施例对本发明进行了详细的说明,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,或者对其中部分或者全部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发明各实施例技术方案的范围。

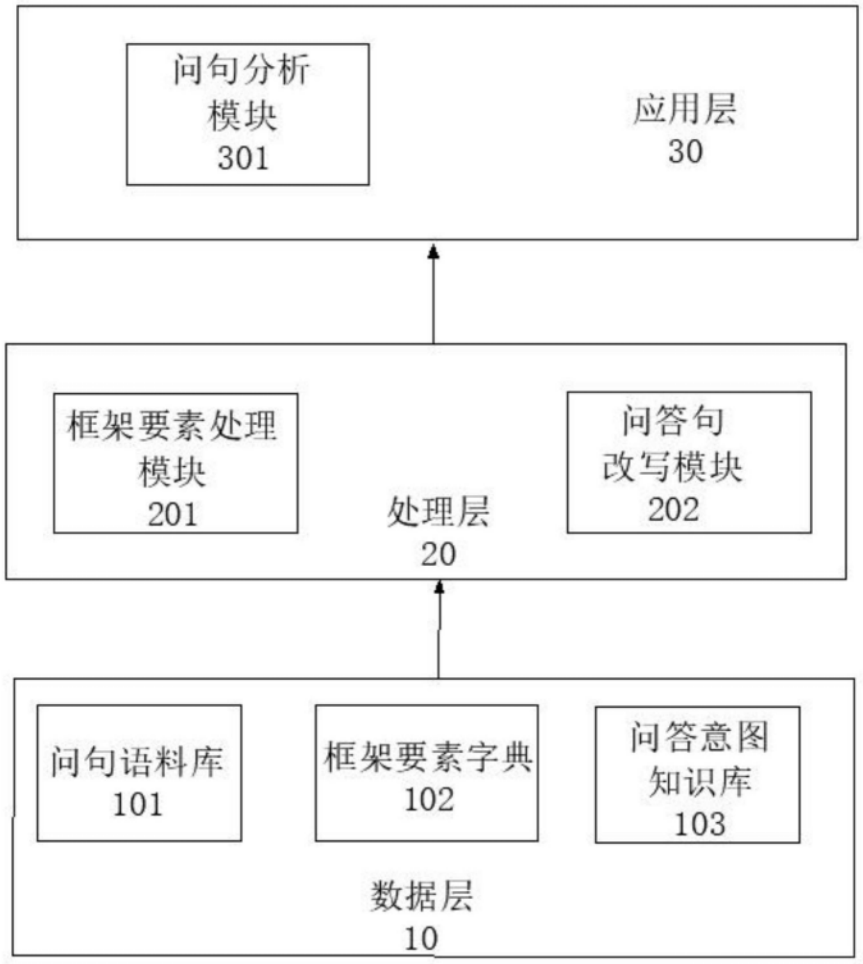

问句分析
模块
301
应用层
30
框架要素处理
模块
201
处理层
20
问答句
改写模块
202
问句语料库
101
框架要素字典
102
问答意图
知识库
103
数据层
10

图1

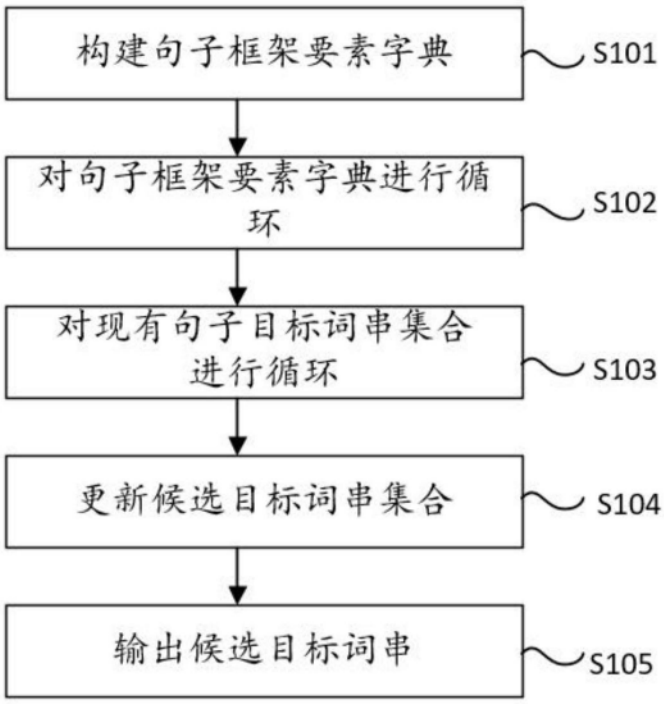

构建句子框架要素字典
S101
对句子框架要素字典进行循环
S102
对现有句子目标词串集合进行循环
S103
更新候选目标词串集合
S104
输出候选目标词串
S105

图2

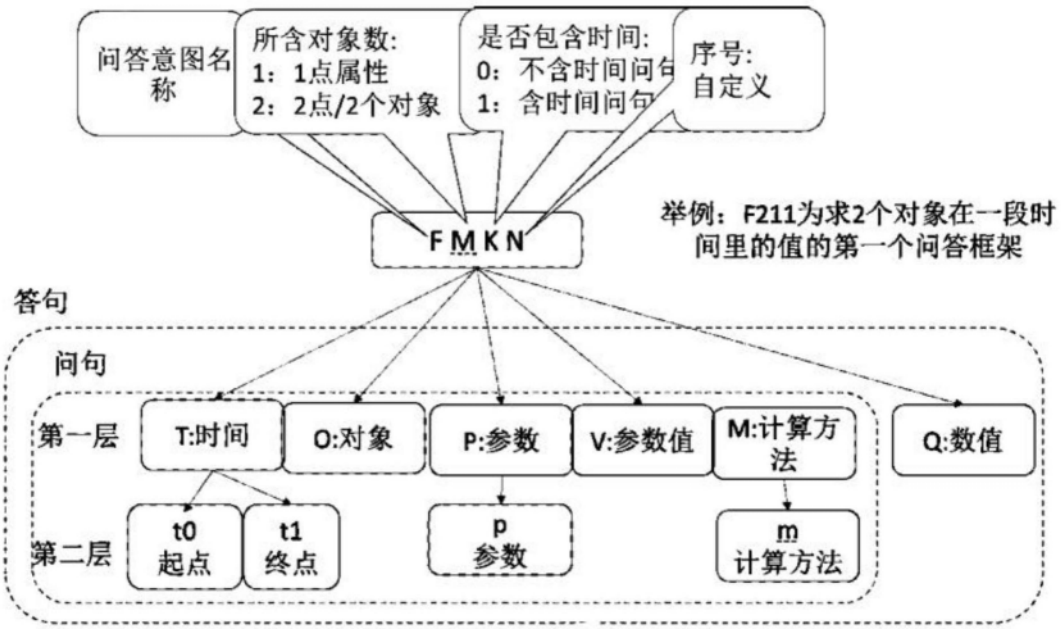

问答意图名称
所含对象数:
1：1点属性
2：2点/2个对象
是否包含时间:
0：不含时间问句
1：含时间问句
序号:
自定义
F M K N
举例：F211为求2个对象在一段时间里的值的第一个问答框架
答句
问句
第一层
T:时间
O:对象
P:参数
V:参数值
M:计算方法
Q:数值
第二层
t0
起点
t1
终点
p
参数
m
计算方法

图3

序号	来源	问句
1	中原油田分公司_问答记录汇总表0515xy(1).xls	105-1H的补心高是多少?
2		普光105-1H的补心高是多少?
71		1#水处理站的污水处理量
72		普光污水处理量
73		1#水处理站站的污水处理量
74	普光问题0513.docx	普光201集气站有多少口井
75		普光201有多少口井
76		402-1井2019年11月5日的产液量是多少
77		402-1井2019年11月5日的产气量是多少
78		402-2井月产气量是多少
79		202-2井昨天的产气量是多少
80		402-2井2019年11月17日的日产气量是多少
81		普光气田的气量和昨天相比增加还是减少
82		普光气田的2016年5月的月产气量同比2016年4月变化多少
83		回注井是什么井?
84		回注井
85		普光气田的产量是多少?
86	问答-时间问句集.txt	毛坝502-1井前三月的月产气量是多少
87		毛坝502-1井前六月的月产气量是多少
88		毛坝502-1井前一月的月产气量是多少
89		TK001井产液量这周比上周减少多少?

图4

名称	框架要素代号
康村组地层参与设计地层及岩性描述	O
康村组	O
阿克库勒组地层参与设计地层及岩性描述	O
阿克库勒组	O
地层参与钻井地质监督日报--地质分层	O
一体化孔板流量计设备参与设备月度运行数据	O
一保累计时间	P
一保	P
一区生产管理区块参与生产管理区块表	O
一区（十一区）生产管理区块参与生产管理区块表	O
一区（十一区）	O
2020年	T
注水井	V
多少	Q

图5

问句的目标词串	问答意图解析
井O的P是啥？	F101@@@井O的P是Q。
从O号到T，OP多少	F111@@@从O号到T，OPQ。
从TO的P是多少？	F111@@@从TO的P是Q。
对于T来说OP的数是多少	F111@@@对于T来说OP的数是Q。
开发单元OT的P是多少	F111@@@开发单元OT的P是Q。
总共有多少O在O	F201@@@总共有QO在O。
总共有多少O在V	F102@@@总共有QO在V。
总共有多少V在O	F102@@@总共有QV在O。
我想知道O的P	F101@@@O的P是Q。
有哪些O位于O？	F201@@@有QO位于O。
有哪些O在O？	F201@@@有QO在O？ 。
有哪些O在O	F201@@@有QO在O。
有多少O位于O？	F201@@@有QO位于O。
有多少O在O	F201@@@有QO在O。
O有多少O	F201@@@O有QO。
O的P是多少？	F101@@@O的P是Q。
O的P是多少	F101@@@O的P是Q。
O的PT增加还是减少	F111@@@O的PT是Q。
OP	F101@@@OP是Q。

图6