



(12) 发明专利申请

(10) 申请公布号 CN 113886424 A

(43) 申请公布日 2022. 01. 04

(21) 申请号 202111222792.3

(22) 申请日 2021.10.20

(71) 申请人 京东科技信息技术有限公司

地址 100176 北京市北京经济技术开发区
科创十一街18号院2号楼6层601

(72) 发明人 刘瑞雪 袁韶祖 祝天刚 陈蒙

(74) 专利代理机构 北京清亦华知识产权代理事
务所(普通合伙) 11201

代理人 罗岚

(51) Int. Cl.

G06F 16/242 (2019.01)

G06F 16/2455 (2019.01)

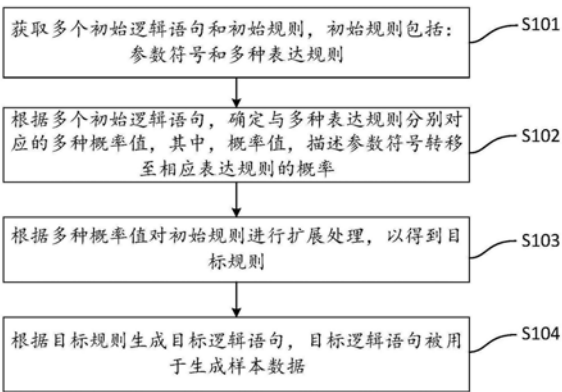
权利要求书2页 说明书11页 附图5页

(54) 发明名称

样本数据生成方法、装置、电子设备及存储
介质

(57) 摘要

本公开提出一种样本数据生成方法、装置、
电子设备及存储介质,该方法包括获取多个初始
逻辑语句和初始规则,初始规则包括:参数符号
和多种表达规则;根据多个初始逻辑语句,确定
与多种表达规则分别对应的多种概率值,其中,
概率值,描述参数符号转移至相应表达规则的
概率;根据多种概率值对初始规则进行扩展处理,
以得到目标规则;以及根据目标规则生成目标逻
辑语句,目标逻辑语句被用于生成样本数据。通
过本公开能够有效地生成多样化的结构化查询
语言逻辑语句,使得所生成的逻辑语句能够有
效地覆盖相应领域的真实分布情况,保障了逻辑
语句的真实性,并且降低对人工资源的依赖,从
而提升生成效率和生成准确性,辅助提升后续
样本数据的生成效果。



1. 一种样本数据生成方法,其特征在于,所述方法包括:
获取多个初始逻辑语句和初始规则,所述初始规则包括:参数符号和多种表达规则;
根据所述多个初始逻辑语句,确定与所述多种表达规则分别对应的多种概率值,其中,所述概率值,描述所述参数符号转移至相应所述表达规则的概率;
根据所述多种概率值对所述初始规则进行扩展处理,以得到目标规则;以及
根据所述目标规则生成目标逻辑语句,所述目标逻辑语句被用于生成样本数据。
2. 如权利要求1所述的方法,其特征在于,所述根据所述多个初始逻辑语句,确定与所述多种表达规则分别对应的多种概率值,包括:
确定与所述多种表达规则分别对应的多个第一次数,所述第一次数,指示在所述多个初始逻辑语句之中所述参数符号转移至相应所述表达规则的次数;
确定与所述多种表达规则分别对应的多个第二次数,所述第二次数,指示在所述多个初始逻辑语句之中所述参数符号转移至其它表达规则的次数;
将所述第一次数和对应所述第二次数的比值,作为所述相应表达规则的概率值。
3. 如权利要求1所述的方法,其特征在于,所述根据所述多种概率值对所述初始规则进行扩展处理,以得到目标规则,包括:
从所述多种概率值中,确定大于概率阈值的目标概率值;
确定与所述目标概率值对应的目标表达规则,和所述目标表达规则对应的目标参数符号;
将所述目标参数符号、所述目标表达规则作为所述目标规则。
4. 如权利要求3所述的方法,其特征在于,所述根据所述目标规则生成目标逻辑语句,包括:
根据所述目标参数符号和所述目标表达规则形成参考逻辑语句;
确定与所述目标参数符号对应的多种可选符号值;
采用所述多种可选符号值分别对所述参考逻辑语句进行扩展,以得到与所述多种可选符号值分别对应的多种目标逻辑语句。
5. 如权利要求4所述的方法,其特征在于,所述确定与所述目标参数符号对应的多种可选符号值,包括:
获取数据库表格,所述数据库表格包括:多种列字段,和与所述列字段分别对应的多种列值;
从所述数据库表格中,确定与所述目标参数符号匹配的目标列字段;
将所述目标列字段对应的多种列值作为所述多种可选符号值。
6. 如权利要求1所述的方法,其特征在于,在所述根据所述目标规则生成目标逻辑语句之后,还包括:
将所述目标逻辑语句输入至预训练的语句生成模型之中,以得到所述语句生成模型输出的与所述目标逻辑语句匹配的目标自然问句,其中,所述目标逻辑语句和所述目标自然问句被共同作为所述样本数据。
7. 一种样本数据生成装置,其特征在于,所述装置包括:
获取模块,用于获取多个初始逻辑语句和初始规则,所述初始规则包括:参数符号和多种表达规则;

确定模块,用于根据所述多个初始逻辑语句,确定与所述多种表达规则分别对应的多种概率值,其中,所述概率值,描述所述参数符号转移至相应所述表达规则的概率;

扩展模块,用于根据所述多种概率值对所述初始规则进行扩展处理,以得到目标规则;

生成模块,用于根据所述目标规则生成目标逻辑语句,所述目标逻辑语句被用于生成样本数据。

8.如权利要求7所述的装置,其特征在于,所述确定模块,包括:

第一确定子模块,用于确定与所述多种表达规则分别对应的多个第一次数,所述第一次数,指示在所述多个初始逻辑语句之中所述参数符号转移至相应所述表达规则的次数;

第二确定子模块,用于确定与所述多种表达规则分别对应的多个第二次数,所述第二次数,指示在所述多个初始逻辑语句之中所述参数符号转移至其它表达规则的次数;

第三确定子模块,用于将所述第一次数和对应所述第二次数的比值,作为所述相应表达规则的概率值。

9.如权利要求7所述的装置,其特征在于,所述扩展模块,具体用于:

从所述多种概率值中,确定大于概率阈值的目标概率值;

确定与所述目标概率值对应的目标表达规则,和所述目标表达规则对应的目标参数符号;

将所述目标参数符号、所述目标表达规则作为所述目标规则。

10.如权利要求9所述的装置,其特征在于,所述生成模块,具体用于:

根据所述目标参数符号和所述目标表达规则形成参考逻辑语句;

确定与所述目标参数符号对应的多种可选符号值;

采用所述多种可选符号值分别对所述参考逻辑语句进行扩展,以得到与所述多种可选符号值分别对应的多种目标逻辑语句。

11.如权利要求10所述的装置,其特征在于,所述生成模块,具体用于:

获取数据库表格,所述数据库表格包括:多种列字段,和与所述列字段分别对应的多种列值;

从所述数据库表格中,确定与所述目标参数符号匹配的目标列字段;

将所述目标列字段对应的多种列值作为所述多种可选符号值。

12.如权利要求7所述的装置,其特征在于,还包括:

输入模块,用于在所述根据所述目标规则生成目标逻辑语句之后,将所述目标逻辑语句输入至预训练的语句生成模型之中,以得到所述语句生成模型输出的与所述目标逻辑语句匹配的目标自然问句,其中,所述目标逻辑语句和所述目标自然问句被共同作为所述样本数据。

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权利要求1-6中任一项所述的方法。

样本数据生成方法、装置、电子设备及存储介质

技术领域

[0001] 本公开涉及自然语言处理技术领域,尤其涉及一种样本数据生成方法、装置、电子设备及存储介质。

背景技术

[0002] 在自然语言处理中,将自然语言转为结构化查询语言SQL语句(Natural Language to Structured Query Language,NL2SQL)是一项表格问答的核心技术,即,将自然语言问题转为数据库上可以执行的SQL查询语句。NL2SQL技术的输入为数据库表格和问句,输出为可执行性的SQL语句。近年来,随着深度学习技术的发展,通常聚焦于采用人工智能模型以解决NL2SQL技术中涉及的问题,包括:基于编码器-解码器(Encoder-Decoder,ED)结构的生成式模型。

[0003] 由于人工智能模型的训练依赖大量的高质量训练用样本数据,相关技术中通常使用基于规则的方法生成海量的样本数据。而基于规则的方法,是通过设置一系列固定规则,来为已知表格生成对应的SQL表达式,而后,根据规则语句模板生成与逻辑语句对应的自然问句,逻辑语句和基于逻辑语句生成的自然问句,均可以被作为训练模型用的样本数据。

[0004] 这种方式下,由于固定规则无法捕获逻辑语句的基本分布,另外在不同的领域中逻辑语句的分布是存在差异的,从而导致所生成的逻辑语句较为单一,不够多样化,较为依赖人工设计,从而导致影响后续样本数据的生成效率和生成效果。

发明内容

[0005] 本公开旨在至少在一定程度上解决相关技术中的技术问题之一。

[0006] 为此,本公开的目的在于提出一种样本数据生成方法、装置、电子设备及存储介质,能够有效地生成多样化的结构化查询语言逻辑语句,使得所生成的逻辑语句能够有效地覆盖相应领域的真实分布情况,保障了逻辑语句的真实性,并且降低对人工资源的依赖,从而提升生成效率和生成准确性,辅助提升后续样本数据的生成效果。

[0007] 为达到上述目的,本公开第一方面实施例提出的样本数据生成方法,包括:获取多个初始逻辑语句和初始规则,初始规则包括:参数符号和多种表达规则;根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,其中,概率值,描述参数符号转移至相应表达规则的概率;根据多种概率值对初始规则进行扩展处理,以得到目标规则;以及根据目标规则生成目标逻辑语句,目标逻辑语句被用于生成样本数据。

[0008] 本公开第一方面实施例提出的方法,通过获取多个初始逻辑语句和初始规则,初始规则包括:参数符号和多种表达规则,根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,其中,概率值,描述参数符号转移至相应表达规则的概率,根据多种概率值对初始规则进行扩展处理,以得到目标规则,以及根据目标规则生成目标逻辑语句,目标逻辑语句被用于生成样本数据,能够有效地生成多样化的结构化查询语言逻辑语句,使

得所生成的逻辑语句能够有效地覆盖相应领域的真实分布情况,保障了逻辑语句的真实性,并且降低对人工资源的依赖,从而提升生成效率和生成准确性,辅助提升后续样本数据的生成效果。

[0009] 为达到上述目的,本公开第二方面实施例提出的样本数据生成装置,包括:获取模块,用于获取多个初始逻辑语句和初始规则,初始规则包括:参数符号和多种表达规则;确定模块,用于根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,其中,概率值,描述参数符号转移至相应表达规则的概率;扩展模块,用于根据多种概率值对初始规则进行扩展处理,以得到目标规则;生成模块,用于根据目标规则生成目标逻辑语句,目标逻辑语句被用于生成样本数据。

[0010] 本公开第二方面实施例提出的样本数据生成装置,通过获取多个初始逻辑语句和初始规则,初始规则包括:参数符号和多种表达规则,根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,其中,概率值,描述参数符号转移至相应表达规则的概率,根据多种概率值对初始规则进行扩展处理,以得到目标规则,以及根据目标规则生成目标逻辑语句,目标逻辑语句被用于生成样本数据,能够有效地生成多样化的结构化查询语言逻辑语句,使得所生成的逻辑语句能够有效地覆盖相应领域的真实分布情况,保障了逻辑语句的真实性,并且降低对人工资源的依赖,从而提升生成效率和生成准确性,辅助提升后续样本数据的生成效果。

[0011] 根据本公开第三方面,提供了一种电子设备,包括:至少一个处理器;以及与至少一个处理器通信连接的存储器;其中,存储器存储有可被至少一个处理器执行的指令,指令被至少一个处理器执行,以使至少一个处理器能够执行本公开第一方面实施例的样本数据生成方法。

[0012] 根据本公开第四方面,提出了一种存储有计算机指令的非瞬时计算机可读存储介质,所述计算机指令用于使所述计算机执行本公开第一方面实施例的样本数据生成方法。

[0013] 根据本公开第五方面,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本公开第一方面实施例的样本数据生成方法。

[0014] 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

[0015] 本公开上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解,其中:

[0016] 图1是本公开一实施例提出的样本数据生成方法的流程示意图;

[0017] 图2是本公开实施例中数据库表格示意图;

[0018] 图3是本公开实施例中上下文无关文法CFG示意图;

[0019] 图4是本公开实施例中初始逻辑语句示意图;

[0020] 图5是本公开另一实施例提出的样本数据生成方法的流程示意图;

[0021] 图6是根据本公开一实施例提出的样本数据生成装置的结构示意图;

[0022] 图7是根据本公开另一实施例提出的样本数据生成装置的结构示意图;

[0023] 图8示出了适于用来实现本公开实施方式的示例性电子设备的框图。

具体实施方式

[0024] 下面详细描述本公开的实施例,所述实施例的示例在附图中示出,其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的,仅用于解释本公开,而不能理解为对本公开的限制。相反,本公开的实施例包括落入所附加权利要求书的精神和内涵范围内的所有变化、修改和等同物。

[0025] 图1是本公开一实施例提出的样本数据生成方法的流程示意图。

[0026] 其中,需要说明的是,本实施例的样本数据生成方法的执行主体为样本数据生成装置,该装置可以由软件和/或硬件的方式实现,该装置可以配置在电子设备中,电子设备可以包括但不限于终端、服务器端等。

[0027] 如图1所示,该样本数据生成方法,包括:

[0028] S101:获取多个初始逻辑语句和初始规则,初始规则包括:参数符号和多种表达规则。

[0029] 其中,当前待处理的逻辑语句,可以被称为初始逻辑语句,该逻辑语句,可以是结构化查询语言SQL语句(Structured Query Language,SQL),或者,该逻辑语句,也可以是其它类型的可以被数据库读取并解析的逻辑语句,对此不做限制。

[0030] 本公开实施例中上述的初始逻辑语句,可以被用于辅助生成多样化的目标逻辑语句,也即是说,目标逻辑语句可以是基于初始逻辑语句扩展生成得到的,并且,基于本公开实施例提供的样本数据生成方法,能够使得生成的多种目标逻辑语句有效地覆盖相应领域的真实分布情况。

[0031] 本公开实施例中,以逻辑语句是结构化查询语言SQL语句进行示例,对此不做限制。

[0032] 其中,初始逻辑语句可以例如,Select SUM(品牌)from Table where价格>300000and产地=德国。

[0033] 该初始逻辑语句,能够被用于从数据库表格中获取到相应的数据。

[0034] 数据库表格,可以如图2所示,图2是本公开实施例中数据库表格示意图,其中,该数据库表格包括:多种列字段,和与列字段分别对应的多种列值,多种列字段例如,车型、品牌、价格、级别、能源、产地、排量,对应各种列字段,分别对应多种列值。

[0035] 上述在获取多个初始逻辑语句后,还可以获取初始规则,初始规则包括:参数符号和多种表达规则,该初始规则,可以被用于辅助生成多种目标逻辑语句。

[0036] 其中,该初始规则,指的是语法表达规则,例如,上下文无关文法(Context-Free Grammar,CFG)、无限制文法、上下文相关文法、上下文无关文法和正规文法等,对此不做限制。

[0037] 本公开实施例中,以初始规则是上下文无关文法CFG进行示例,该上下文无关文法CFG包括:参数符号和多种表达规则。

[0038] 举例而言,一并参见图3和图4,图3是本公开实施例中上下文无关文法CFG示意图,图4是本公开实施例中初始逻辑语句示意图,在图4中,以树状图表述该初始逻辑语句进行示例。

[0039] 上下文无关文法CFG由 $G=(M,T,R,S)$ 描述:

[0040] M:非终结符号(Non-terminal symbols):

[0041] 例如,“Select,Where,Conds,Cond,A,C,op”。

[0042] T:终结符号(Terminal symbols):

[0043] 例如,SQL具体的列/值/预算符构成了SQL的实际内容,例如图4方框中的“品牌,价格,200000”等。

[0044] R:一系列生产规则(production rules):

[0045] 例如,Select包括:SELECT A或者SELECT AA,或者Conds包括:Cond或者Cond AND Cond。

[0046] S:起始符号,此处为树最顶端的“SQL”。

[0047] 则上述的M、T、S即可以被称为参数符号,而R涉及到的表达规则,即可以被称为多种表达规则,对此不做限制。

[0048] 上述图3和图4中所示,可以是与初始逻辑语句【Select SUM(品牌)from Table where价格>300000and产地=德国】对应的上下文无关文法CFG,则本公开实施例中在获取多个初始逻辑语句之后,可以获取与每个初始逻辑语句分别对应的初始规则,从而保障基于少量初始逻辑语句,完成整体的逻辑语句的扩展处理逻辑。

[0049] S102:根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,其中,概率值,描述参数符号转移至相应表达规则的概率。

[0050] 上述在获取多个初始逻辑语句和初始规则之后,可以根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值。

[0051] 也即是说,上述的多个初始逻辑语句,可以是少量的逻辑语句,而后,可以参考少量初始逻辑语句,结合初始规则执行相应的扩展处理逻辑。

[0052] 在扩展处理时,可以根据多个初始逻辑语句,确定与上下文无关文法CFG中的各种表达规则分别对应的概率值,该概率值,描述上下文无关文法CFG中的参数符号转移至相应表达规则的概率,该概率值,可以被用于辅助后续扩展得到多种目标逻辑语句。

[0053] 在根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值时,可以将多个初始逻辑语句输入至预训练的概率确定模型中,以得到概率确定模型输出的与多种表达规则分别对应的多种概率值,该概率值,在含义方面可以用于表述相应表达规则,在多个初始逻辑语句当中的分布概率情况,在数值方面,表达初始规则中的参数符号转移至相应表达规则的概率。

[0054] 可选地,一些实施例中,在根据所述多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,可以是确定与多种表达规则分别对应的多个第一次数,第一次数,指示在多个初始逻辑语句之中参数符号转移至相应表达规则的次数,确定与多种表达规则分别对应的多个第二次数,第二次数,指示在多个初始逻辑语句之中参数符号转移至其它表达规则的次数,以及将第一次数和对应第二次数的比值,作为相应表达规则的概率值,从而使得概率值能够与初始逻辑语句所处相应领域相匹配,使得后续基于概率值得到的目标规则能够描述出相应领域的真实分布情况,并且,概率值的确定方式较为简捷,保障了样本数据生成方法的执行效率。

[0055] 上述过程,可以被称为根据少量初始逻辑语句SQL训练初始规则的过程,在所述多个初始逻辑语句之中所述参数符号转移至相应所述表达规则的次数,可以被称为第一次

数,在所述多个初始逻辑语句之中所述参数符号转移至其它表达规则的次数,可以被称为第二次数,而后,可以将每个表达规则对应的第一次数和对应所述第二次数的比值,作为所述相应表达规则的概率值。

[0056] 举例而言,为了计算初始逻辑语句SQL所处领域场景下的概率,可以首先根据在该领域场景下已有的少量初始逻辑语句SQL以及上述得到的初始规则,计算出每个表达规则的概率值,而后,可以触发后续通过替换终结符号的方式,构建出更多的目标逻辑语句。

[0057] 其中,每个非终结符转移到其对应某个表达规则的概率值计算如下:

$$p(N \rightarrow a) = \frac{C(N \rightarrow a)}{\sum C(N \rightarrow b)}$$

;

[0059] 其中,C指的是某条表达规则出现的次数,C(N→a)指的是当前M(非终结符)转移到a表达规则的次数,C(N→b)则指的是当前M转移到所有可能表达规则的次数。

[0060] 当然,也可以采用其它任意可能的方式来实现确定与多种表达规则分别对应的多种概率值,例如上述模型的方式、或者工程学方式等等,对此不做限制。

[0061] S103:根据多种概率值对初始规则进行扩展处理,以得到目标规则。

[0062] 举例而言,根据少量初始逻辑语句SQL训练初始规则以得到概率上下文无关文法(Probabilistic Context Free Grammar,PCFG),训练得到的PCFG表达规则,即可以被称为目标规则。

[0063] 目标规则PCFG表达式如下:

[0064] $G = (M, T, R, S)$ 。

[0065] 其中,概率上下文无关文法PCFG是与上下文无关文法CFG相似的语法表达规则,其中,概率上下文无关文法PCFG中的P指是一系列表达规则(R)的转移概率,例如,Select表达规则是转移到(SELECT A)还是转移到(SELECT AA)。

[0066] S104:根据目标规则生成目标逻辑语句,目标逻辑语句被用于生成样本数据。

[0067] 上述在根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,并根据多种概率值对初始规则进行扩展处理,以得到目标规则之后,可以根据目标规则生成目标逻辑语句,目标逻辑语句被用于生成样本数据。

[0068] 举例而言,在根据目标规则生成目标逻辑语句时,可以确定数据库表格,而后,根据数据库表格包含的数据内容对初始逻辑语句中的T:终结符号(Terminal symbols)进行多种可能的替换,以生成多种替换后的逻辑语句并作为目标逻辑语句。

[0069] 举例而言,CFG:Select A1,A2 WHERE Cond(A op Value);

[0070] 替换:A1→“价格”,A2→“产地”A→“品牌”op→“等于”Value→“宝马”,生成的目标逻辑语句是SQL:Select价格,产地Where品牌=宝马。

[0071] 本实施例中,通过获取多个初始逻辑语句和初始规则,初始规则包括:参数符号和多种表达规则,并根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,其中,概率值,描述参数符号转移至相应表达规则的概率情况,根据多种概率值对初始规则进行扩展处理,以得到目标规则,以及根据目标规则生成目标逻辑语句,目标逻辑语句被用于生成样本数据,能够有效地生成多样化的结构化查询语言逻辑语句,使得所生成的逻辑语句能够有效地覆盖相应领域的真实分布情况,保障了逻辑语句的真实性,并且降低对人工

资源的依赖,从而提升生成效率和生成准确性,辅助提升后续样本数据的生成效果。

[0072] 图5是本公开另一实施例提出的样本数据生成方法的流程示意图。

[0073] 如图5所示,该样本数据生成方法,包括:

[0074] S501:获取多个初始逻辑语句和初始规则,初始规则包括:参数符号和多种表达规则。

[0075] S502:根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,其中,概率值,描述参数符号转移至相应表达规则的概率。

[0076] S501-S502的描述说明可以具体参见上述实施例,在此不再赘述。

[0077] S503:从多种概率值中,确定大于概率阈值的目标概率值。

[0078] 上述在根据所述多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值之后,可以从多种概率值中,确定大于概率阈值的目标概率值,也即是说,本公开实施例中支持从目标规则中包括的多种表达规则中确定出转移概率较大的目标表达规则,从而辅助基于该目标表达规则进行逻辑语句的扩展,不仅仅使得所生成的逻辑语句有效地覆盖相应领域的真实分布情况,还实现生成具有更为显著转移可能的逻辑语句,全面地保障逻辑语句的生成质量。

[0079] S504:确定与目标概率值对应的目标表达规则,和目标表达规则对应的目标参数符号。

[0080] S505:将目标参数符号、目标表达规则作为目标规则。

[0081] S506:根据目标参数符号和目标表达规则形成参考逻辑语句。

[0082] 举例而言,在根据在该领域场景下已有的少量初始逻辑语句SQL以及上述得到的初始规则,计算出每个表达规则的概率值后,可以确定大于概率阈值的目标概率值,即确定出现概率较大(即具有更为显著的转移可能)的逻辑语句SQL是:Select A1,A2 WHERE Cond (A op Value),或者Select A WHERE Cond (A op Value)AND Cond (A op Value)等等,对此不做限制。

[0083] 其中,Select A1,A2 WHERE Cond (A op Value),或者Select A WHERE Cond (A op Value)AND Cond (A op Value)中涉及的表达规则,即可以被称为目标表达规则,涉及的参数符号,即可以被称为目标参数符号,该示例中,目标参数符号转移至目标表达规则的概率值,可以是大于概率阈值的目标概率值。

[0084] 从而依据目标参数符号和目标表达规则形成的逻辑语句,可以被称为参考逻辑语句,即上述的,Select A1,A2 WHERE Cond (A op Value),或者Select A WHERE Cond (A op Value)AND Cond (A op Value)。

[0085] 举例而言,如果SQL语句中包含排序相关的语法,则目标规则也可以同样地被扩充。

[0086] 例如:

[0087] 自然语句:根据价格从高到齐列出产地是德国的汽车品牌。

[0088] 参考逻辑语句:Select品牌Where产地=德国Order by价格DESC

[0089] 目标规则:

[0090] SQL:=SELECT|SELECT WHERE|SELECT WHERE ORDER。

[0091] S507:确定与目标参数符号对应的多种可选符号值。

[0092] 其中,目标参数符号,可以具体是上述示例中的非终结符号M,从而可以依据如图2中所示的数据库表格,确定与非终结符号M对应的多种可选符号值(与非终结符号M对应的实际领域场景中的取值,可以被称为可选符号值)。

[0093] 可选地,一些实施例中,在确定与所述目标参数符号对应的多种可选符号值时,可以是获取数据库表格,数据库表格包括:多种列字段,和与列字段分别对应的多种列值,从数据库表格中,确定与目标参数符号匹配的目标列字段,将目标列字段对应的多种列值作为多种可选符号值,从而实现快速地融合实际领域场景中的内容,避免了消耗大量的人力工作,并且降低了对专业知识的依赖性,辅助后续快速地构建目标逻辑语句。

[0094] 数据库表格的列字段和与列字段对应的多种列值的示意,可以如上述图2所示,可以将如图2中与非终结符号M匹配的目标列字段是:车型,品牌,价格,能源,产地等,而后,可以从图2中读取与车型,品牌,价格,能源,产地等分别对应的多种列值,作为多种可选符号值。

[0095] S508:采用多种可选符号值分别对参考逻辑语句进行扩展,以得到与多种可选符号值分别对应的多种目标逻辑语句。

[0096] 上述在确定与目标参数符号对应的多种可选符号值车型,品牌,价格,能源,产地等,进行随机替换即可获得大量目标逻辑语句SQL:

[0097] CFG:Select A1,A2 WHERE Cond(A op Value);

[0098] 替换:A1->“价格”,A2->“产地”A->”品牌”op->“等于”Value->“宝马”;

[0099] 目标逻辑语句SQL:Select价格,产地Where品牌=宝马。

[0100] S509:将目标逻辑语句输入至预训练的语句生成模型之中,以得到语句生成模型输出的与目标逻辑语句匹配的目标自然问句,其中,目标逻辑语句和目标自然问句被共同作为样本数据。

[0101] 上述在对初始逻辑语句进行扩展得到多种目标逻辑语句之后,可根据目标逻辑语句生成匹配的自然问句,该匹配的自然问句即可以被称为目标自然问句。

[0102] 可选地,可以结合文本生成相关的技术来生成与目标逻辑语句匹配的目标自然问句,或者,也可以采用序列至序列到生成式模型(sequence-to-sequence)来生成与目标逻辑语句匹配的目标自然问句,对此不做限制。

[0103] 上述的语句生成模型可以是预先训练得到的。

[0104] 在训练得到语句生成模型时,可以使用自监督的方式进行训练,对此不做限制。

[0105] 举例而言,该语句生成模型可以是生成式的预训练(Generative Pre-Training, GPT)模型,或者也可以使用采用序列到序列模型构建的降噪自编码器模型等文本生成任务相关的预训练模型,可以首先使用少量的SQL&自然语言(Natural Language,NL)问句(即自然问句)的语料对该模型进行微调(fine-tuning),使其能够执行根据给定的SQL生成对应的自然问句的任务。

[0106] 实验表明,使用少量的数据对该模型进行微调(fine-tuning)后,该模型可以生成顺畅的自然问句,而后,可以根据上述生成的目标逻辑语句结合该生成式的预训练模型生成表述多样,语言流畅且贴近特定场景的目标自然问句。

[0107] 上述生成的目标逻辑语句和目标自然问句,可以被共同作为样本数据,以进行各类NL2SQL任务模型的预训练。

[0108] 本实施例中,能够有效地生成多样化的结构化查询语言逻辑语句,使得所生成的逻辑语句能够有效地覆盖相应领域的真实分布情况,保障了逻辑语句的真实性,并且降低对人工资源的依赖,从而提升生成效率和生成准确性,辅助提升后续样本数据的生成效果。支持从目标规则中包括的多种表达规则中确定出转移概率较大的目标表达规则,从而辅助基于该目标表达规则进行逻辑语句的扩展,不仅仅使得所生成的逻辑语句有效地覆盖相应领域的真实分布情况,还实现生成具有更为显著转移可能的逻辑语句,全面地保障逻辑语句的生成质量。并且,使用预训练语言模型,能够有效地节省数据资源和人工资源来构建语句模版,还能够有效地保证生成的自然问句的表达多样性和流畅性。

[0109] 图6是根据本公开一实施例提出的样本数据生成装置的结构示意图。

[0110] 如图6所示,该样本数据生成装置60,包括:

[0111] 获取模块601,用于获取多个初始逻辑语句和初始规则,初始规则包括:参数符号和多种表达规则;

[0112] 确定模块602,用于根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,其中,概率值,描述参数符号转移至相应表达规则的概率;

[0113] 扩展模块603,用于根据多种概率值对初始规则进行扩展处理,以得到目标规则;

[0114] 生成模块604,用于根据目标规则生成目标逻辑语句,目标逻辑语句被用于生成样本数据。

[0115] 在本公开的一些实施例中,如图7所示,图7是根据本公开另一实施例提出的样本数据生成装置的结构示意图。确定模块602,包括:

[0116] 第一确定子模块6021,用于确定与多种表达规则分别对应的多个第一次数,第一次数,指示在多个初始逻辑语句之中参数符号转移至相应表达规则的次数;

[0117] 第二确定子模块6022,用于确定与多种表达规则分别对应的多个第二次数,第二次数,指示在多个初始逻辑语句之中参数符号转移至其它表达规则的次数;

[0118] 第三确定子模块6023,用于将第一次数和对应第二次数的比值,作为相应表达规则的概率值。

[0119] 在本公开的一些实施例中,如图7所示,扩展模块603,具体用于:

[0120] 从所述多种概率值中,确定大于概率阈值的目标概率值;

[0121] 确定与目标概率值对应的目标表达规则,和目标表达规则对应的目标参数符号;

[0122] 将目标参数符号、目标表达规则作为目标规则。

[0123] 在本公开的一些实施例中,如图7所示,生成模块604,具体用于:

[0124] 根据目标参数符号和目标表达规则形成参考逻辑语句;

[0125] 确定与目标参数符号对应的多种可选符号值;

[0126] 采用多种可选符号值分别对参考逻辑语句进行扩展,以得到与多种可选符号值分别对应的多种目标逻辑语句。

[0127] 在本公开的一些实施例中,如图7所示,生成模块604,具体用于:

[0128] 获取数据库表格,数据库表格包括:多种列字段,和与列字段分别对应的多种列值;

[0129] 从数据库表格中,确定与目标参数符号匹配的目标列字段;

[0130] 将目标列字段对应的多种列值作为多种可选符号值。

[0131] 在本公开的一些实施例中,如图7所示,还包括:

[0132] 输入模块605,用于在根据目标规则生成目标逻辑语句之后,将目标逻辑语句输入至预训练的语句生成模型之中,以得到语句生成模型输出的与目标逻辑语句匹配的目标自然问句,其中,目标逻辑语句和目标自然问句被共同作为样本数据。

[0133] 与上述图1至图5实施例提供的样本数据生成方法相对应,本公开还提供一种样本数据生成装置,由于本公开实施例提供的样本数据生成装置与上述图1至图5实施例提供的样本数据生成方法相对应,因此在样本数据生成方法的实施方式也适用于本公开实施例提供的样本数据生成装置,在本公开实施例中不再详细描述。

[0134] 本实施例中,通过获取多个初始逻辑语句和初始规则,初始规则包括:参数符号和多种表达规则,并根据多个初始逻辑语句,确定与多种表达规则分别对应的多种概率值,其中,概率值,描述参数符号转移至相应表达规则的概率情况,根据多种概率值对初始规则进行扩展处理,以得到目标规则,以及根据目标规则生成目标逻辑语句,目标逻辑语句被用于生成样本数据,能够有效地生成多样化的结构化查询语言逻辑语句,使得所生成的逻辑语句能够有效地覆盖相应领域的真实分布情况,保障了逻辑语句的真实性,并且降低对人工资源的依赖,从而提升生成效率和生成准确性,辅助提升后续样本数据的生成效果。

[0135] 为了实现上述实施例,本公开还提出一种电子设备,包括:存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,处理器执行程序时,实现如本公开前述实施例提出的样本数据生成方法。

[0136] 为了实现上述实施例,本公开还提出一种非临时性计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现如本公开前述实施例提出的样本数据生成方法。

[0137] 为了实现上述实施例,本公开还提出一种计算机程序产品,当计算机程序产品中的指令处理器执行时,执行如本公开前述实施例提出的样本数据生成方法。

[0138] 图8示出了适于用来实现本公开实施方式的示例性电子设备的框图。图8显示的电子设备12仅仅是一个示例,不应对本公开实施例的功能和使用范围带来任何限制。

[0139] 如图8所示,电子设备12以通用计算设备的形式表现。电子设备12的组件可以包括但不限于:一个或者多个处理器或者处理单元16,系统存储器28,连接不同系统组件(包括系统存储器28和处理单元16)的总线18。总线18表示几类总线结构中的一种或多种,包括存储器总线或者存储器控制器,外围总线,图形加速端口,处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说,这些体系结构包括但不限于工业标准体系结构(Industry Standard Architecture;以下简称:ISA)总线,微通道体系结构(Micro Channel Architecture;以下简称:MAC)总线,增强型ISA总线、视频电子标准协会(Video Electronics Standards Association;以下简称:VESA)局域总线以及外围组件互连(Peripheral Component Interconnection;以下简称:PCI)总线。

[0140] 电子设备12典型地包括多种计算机系统可读介质。这些介质可以是任何能够被电子设备12访问的可用介质,包括易失性和非易失性介质,可移动的和不可移动的介质。

[0141] 存储器28可以包括易失性存储器形式的计算机系统可读介质,例如随机存取存储器(Random Access Memory;以下简称:RAM) 30和/或高速缓存存储器32。电子设备12可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例,存

储系统34可以用于读写不可移动的、非易失性磁介质(图8未显示,通常称为“硬盘驱动器”)。

[0142] 尽管图8中未示出,可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器,以及对可移动非易失性光盘(例如:光盘只读存储器(Compact Disc Read Only Memory;以下简称:CD-ROM)、数字多功能只读光盘(Digital Video Disc Read Only Memory;以下简称:DVD-ROM)或者其它光介质)读写的光盘驱动器。在这些情况下,每个驱动器可以通过一个或者多个数据介质接口与总线18相连。存储器28可以包括至少一个程序产品,该程序产品具有一组(例如至少一个)程序模块,这些程序模块被配置以执行本公开各实施例的功能。

[0143] 具有一组(至少一个)程序模块42的程序/实用工具40,可以存储在例如存储器28中,这样的程序模块42包括但不限于操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块42通常执行本公开所描述的实施例中的功能和/或方法。

[0144] 电子设备12也可以与一个或多个外部设备14(例如键盘、指向设备、显示器24等)通信,还可与一个或者多个使得用户能与该电子设备12交互的设备通信,和/或与使得该电子设备12能与一个或多个其它计算设备进行通信的任何设备(例如网卡,调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口22进行。并且,电子设备12还可以通过网络适配器20与一个或者多个网络(例如局域网(Local Area Network;以下简称:LAN),广域网(Wide Area Network;以下简称:WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器20通过总线18与电子设备12的其它模块通信。应当明白,尽管图中未示出,可以结合电子设备12使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0145] 处理单元16通过运行存储在系统存储器28中的程序,从而执行各种功能应用以及数据处理,例如实现前述实施例中提及的样本数据生成方法。

[0146] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本公开的其它实施方案。本公开旨在涵盖本公开的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的,本公开的真正范围和精神由下面的权利要求指出。

[0147] 应当理解的是,本公开并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本公开的范围仅由所附的权利要求来限制。

[0148] 需要说明的是,在本公开的描述中,术语“第一”、“第二”等仅用于描述目的,而不能理解为指示或暗示相对重要性。此外,在本公开的描述中,除非另有说明,“多个”的含义是两个或两个以上。

[0149] 流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为,表示包括一个或更多个用于实现特定逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分,并且本公开的优选实施方式的范围包括另外的实现,其中可以不按所示出或讨论的顺序,包括根据所涉及的功能按基本同时的方式或按相反的顺序,来执行功能,这应被本公开的实施例所属技术领域的技术人员所理解。

[0150] 应当理解,本公开的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中,多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如,如果用硬件来实现,和在另一实施方式中一样,可用本领域公知的下列技术中的任一项或他们的组合来实现:具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路,具有合适的组合逻辑门电路的专用集成电路,可编程门阵列(PGA),现场可编程门阵列(FPGA)等。

[0151] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,该程序在执行时,包括方法实施例的步骤之一或其组合。

[0152] 此外,在本公开各个实施例中的各功能单元可以集成在一个处理模块中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现,也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用,也可以存储在一个计算机可读取存储介质中。

[0153] 上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0154] 在本说明书的描述中,参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本公开的至少一个实施例或示例中。在本说明书中,对上述术语的示意性表述不一定指的是相同的实施例或示例。而且,描述的具体特征、结构、材料或者特点可以在任何一个或多个实施例或示例中以合适的方式结合。

[0155] 尽管上面已经示出和描述了本公开的实施例,可以理解的是,上述实施例是示例性的,不能理解为对本公开的限制,本领域的普通技术人员在本公开的范围内可以对上述实施例进行变化、修改、替换和变型。

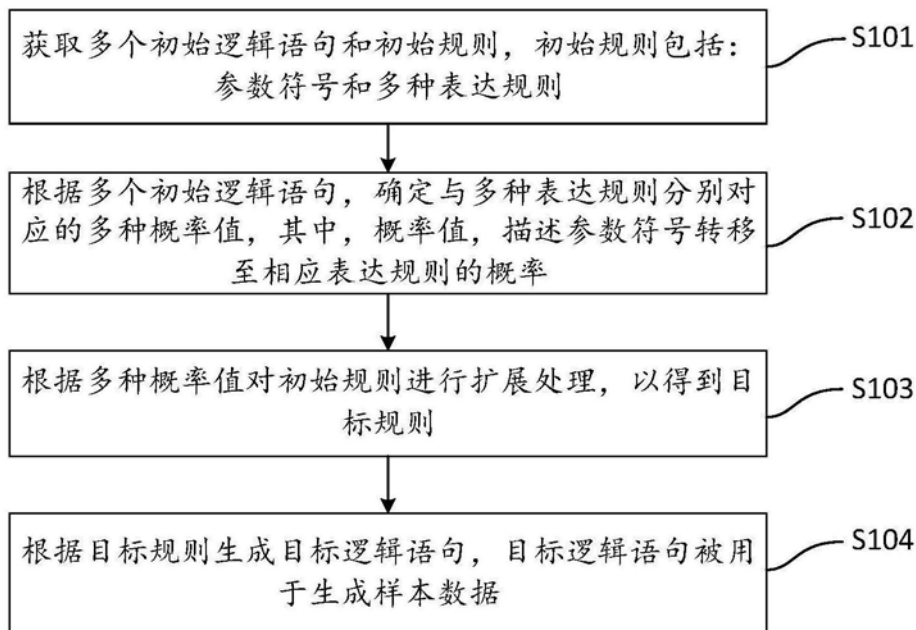


图1

车型	品牌	价格	级别	能源	产地	排量	列名
帕萨特	大众	288900	中型车	汽油	德国	2	
5系	宝马	659900	中大型车	汽油	德国	3	
A4L	奥迪	349400	中型车	汽油	德国	1.4	列值

图2

SQL	::=	Select Select Where ...
Select	::=	SELECT A SELECT A A ...
Where	::=	WHERE Conds
Conds	::=	Cond Cond AND Conds ...
Cond	::=	A op value A op SQL
A	::=	C min C max C avg C ...
C	::=	table.column
op	::=	= != > >= < like in ...

图3

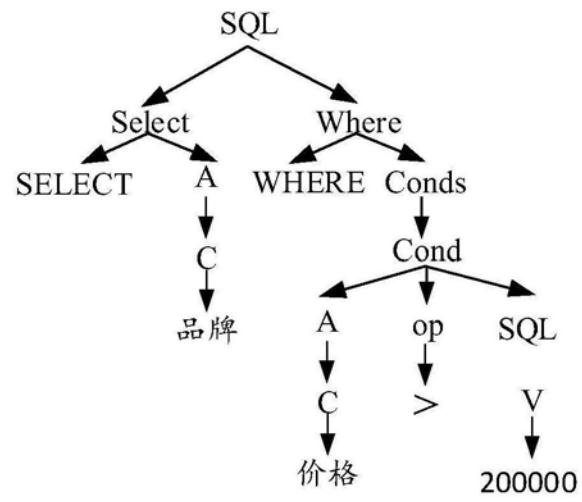


图4

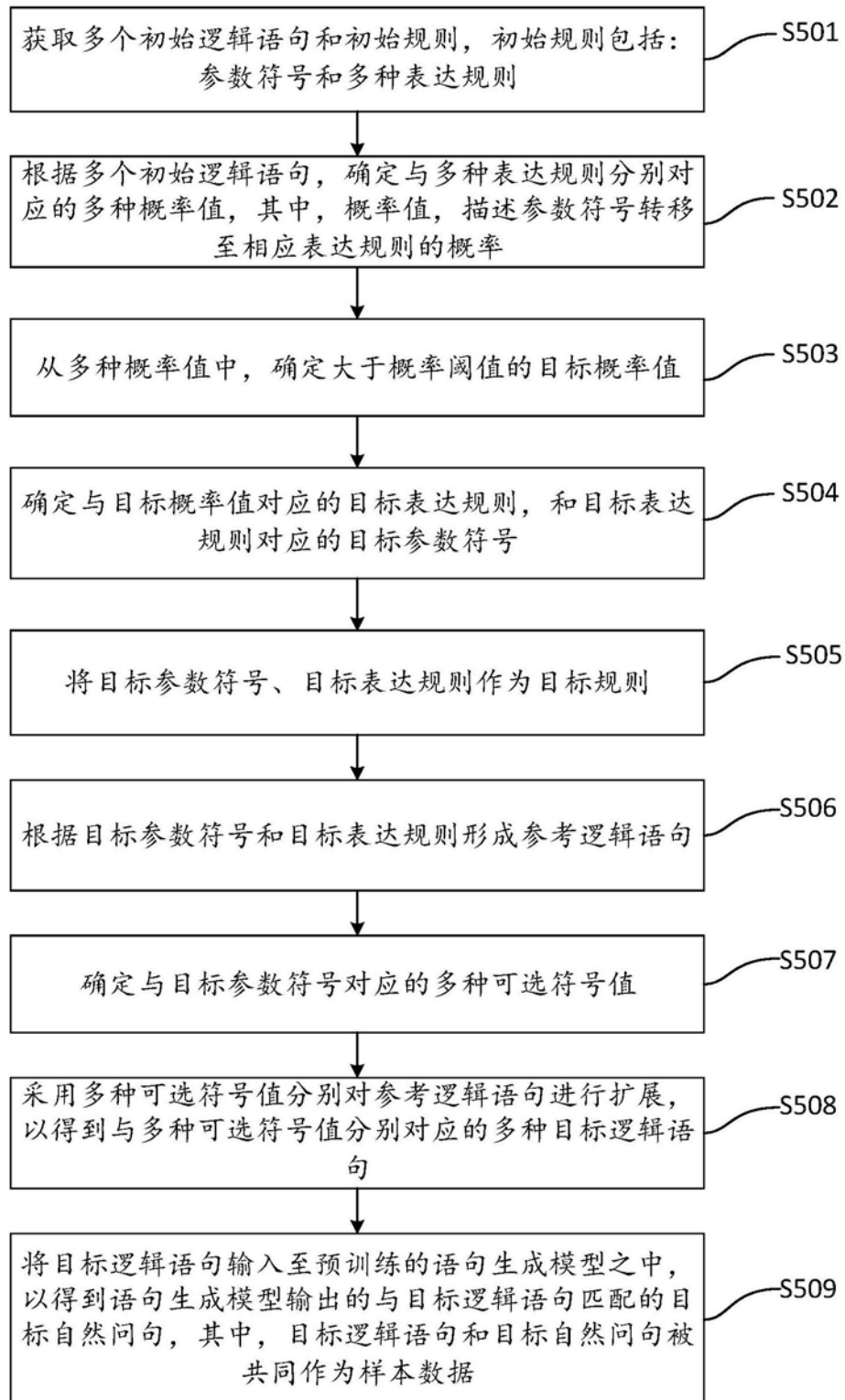


图5



图6

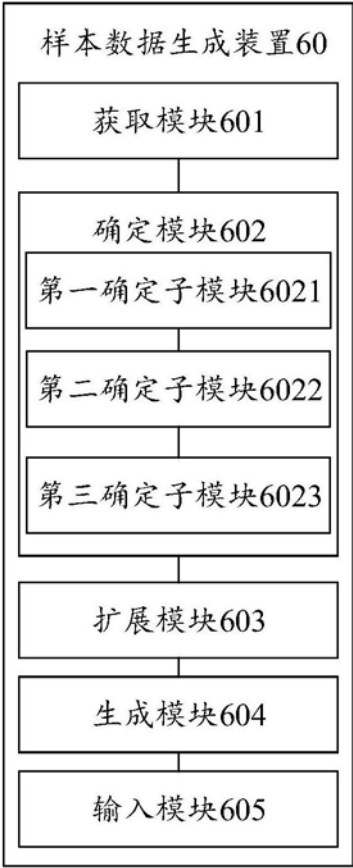


图7

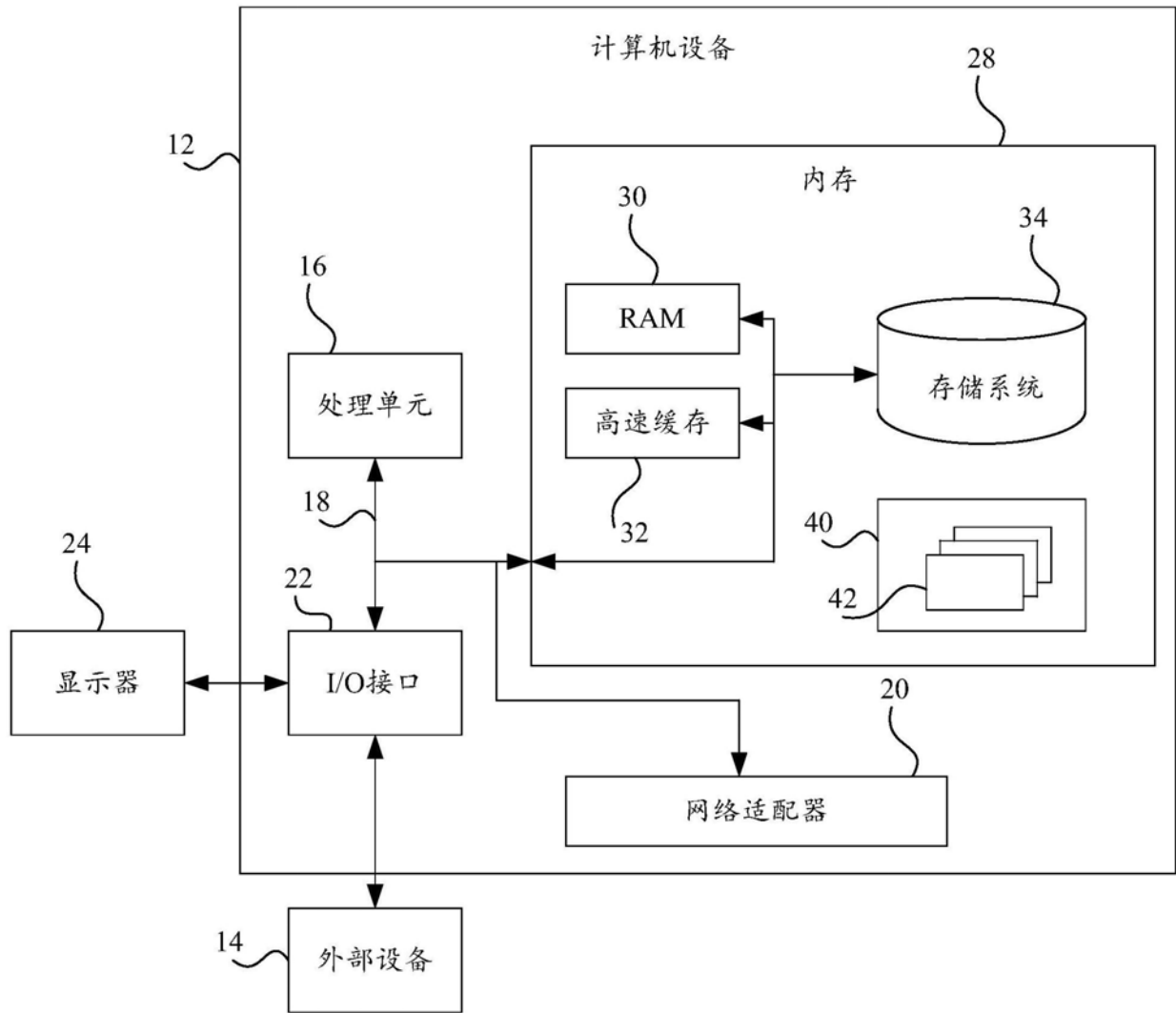


图8