



(12) 发明专利申请

(10) 申请公布号 CN 114528312 A

(43) 申请公布日 2022. 05. 24

(21) 申请号 202210141701.1

(22) 申请日 2022.02.16

(71) 申请人 京东科技信息技术有限公司

地址 100176 北京市大兴区北京经济技术
开发区科创十一街18号院2号楼6层
601

(72) 发明人 祝天刚 刘瑞雪 袁韶祖 陈蒙

(74) 专利代理机构 北京路浩知识产权代理有限
公司 11002

专利代理师 王治东

(51) Int.Cl.

G06F 16/242 (2019.01)

G06F 40/18 (2020.01)

G06K 9/62 (2022.01)

权利要求书2页 说明书11页 附图4页

(54) 发明名称

一种结构化查询语言语句的生成方法和装
置

(57) 摘要

本公开提供一种结构化查询语言语句的生成方法和装置,其中方法包括:获取自然语言语句和所述自然语言语句对应的表格;将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句;其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。实现了能够准确地获取特定领域的自然语言语句对应的结构化查询语言语句。

获取自然语言语句和所述自然语言语句
对应的表格

将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句,其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的

1. 一种结构化查询语言语句的生成方法,其特征在于,包括:

获取自然语言语句和所述自然语言语句对应的表格;

将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句;

其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。

2. 根据权利要求1所述的结构化查询语言语句的生成方法,其特征在于,所述方法还包括:

将所述结构化查询语言语句中条件类型和问题类型分别对应的列名与表格中的列名相匹配,确定匹配结果;

基于匹配结果,在表格中进行查询,得出查询结果。

3. 根据权利要求1所述的结构化查询语言语句的生成方法,其特征在于,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的,包括:

对预训练语言模型进行微调,获取微调好的预训练语言模型;

将结构化查询语言样本语句输入至微调好的预训练语言模型中,生成对应的自然语言样本语句;其中,所述结构化查询语言样本语句是基于所述表格,通过全排列的方式获得的;

将所述自然语言样本语句和所述结构化查询语言样本语句形成第一文本对;

将所述第一文本对作为训练数据对转换模型进行训练,得到预训练的转换模型。

4. 根据权利要求3所述的结构化查询语言语句的生成方法,其特征在于,所述对预训练语言模型进行微调之前,包括:

获取通用领域中的自然语言语句和所述自然语言语句对应的查询语言语句;

将所述查询语言语句进行格式转换,得到结构化查询语言语句;

基于所述自然语言语句和所述结构化查询语言语句,形成第二文本对。

5. 根据权利要求3所述的结构化查询语言语句的生成方法,其特征在于,所述对预训练语言模型进行微调,获取微调好的预训练语言模型,包括:

将所述第二文本对输入至预训练语言模型中,对预训练语言模型进行微调,获取微调好的预训练语言模型。

6. 一种结构化查询语言语句的生成装置,其特征在于,包括:

获取模块,用于获取自然语言语句和所述自然语言语句对应的表格;

输出模块,用于将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句;

其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。

7. 根据权利要求6所述的结构化查询语言语句的生成装置,其特征在于,所述装置还包括:

匹配模块,用于将所述结构化查询语言语句中条件类型和问题类型分别对应的列名与表格中的列名相匹配,确定匹配结果;

查询模块,用于基于匹配结果,在表格中进行查询,得出查询结果。

8.根据权利要求6所述的结构化查询语言语句的生成装置,其特征在于,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的,包括:

对预训练语言模型进行微调,获取微调好的预训练语言模型;

将结构化查询语言样本语句输入至微调好的预训练语言模型中,生成对应的自然语言样本语句;其中,所述结构化查询语言样本语句是基于所述表格,通过全排列的方式获得的;

将所述自然语言样本语句和所述结构化查询语言样本语句形成第一文本对;

将所述第一文本对作为训练数据对转换模型进行训练,得到预训练的转换模型。

9.一种电子设备,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,其特征在于,所述处理器执行所述程序时实现如权利要求1至5任一项所述结构化查询语言语句的生成方法。

10.一种非暂态计算机可读存储介质,其上存储有计算机程序,其特征在于,该计算机程序被处理器执行时实现如权利要求1至5任一项所述结构化查询语言语句的生成方法。

一种结构化查询语言语句的生成方法和装置

技术领域

[0001] 本公开涉及自然语言处理技术领域,尤其涉及一种结构化查询语言语句的生成方法和装置。

背景技术

[0002] 自然语言语句转结构化查询语言语句是指将用户输入的、围绕用户自己提供的表格数据的问题语句转化为面向数据库的结构化查询语言语句。它可以充当数据库的智能接口,让不熟悉数据库的用户能够快速找到自己想要的数据库。

[0003] 现有技术中,大多数方法是先收集大量的自然语言语句到结构化查询语言语句的文本对作为训练数据,对模型进行离线训练;再基于用户上传的表格,对用户的问题语句解析成结构化查询语言语句。然而,这样存在以下问题:离线收集的自然语言语句到结构化查询语言语句的文本对训练数据是通用领域的数据库,其领域知识很难覆盖到用户上传的表格数据所涉及的领域。当用户上传特定领域的表格的时候,模型的解析效果会大打折扣。

发明内容

[0004] 本公开提供一种结构化查询语言语句的生成方法和装置,用以解决现有技术中不能针对特定领域将自然语言语句转为查询语言语句的缺陷,实现了能够准确地获取特定领域的自然语言语句对应的查询语言语句。

[0005] 第一方面,本公开提供一种结构化查询语言语句的生成方法,包括:

[0006] 获取自然语言语句和所述自然语言语句对应的表格;

[0007] 将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句;

[0008] 其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。

[0009] 根据本公开提供的结构化查询语言语句的生成方法,所述方法还包括:

[0010] 将所述结构化查询语言语句中条件类型和问题类型分别对应的列名与表格中的列名相匹配,确定匹配结果;

[0011] 基于匹配结果,在表格中进行查询,得出查询结果。

[0012] 根据本公开提供的结构化查询语言语句的生成方法,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的,包括:

[0013] 对预训练语言模型进行微调,获取微调好的预训练语言模型;

[0014] 将结构化查询语言样本语句输入至微调好的预训练语言模型中,生成对应的自然语言样本语句;其中,所述结构化查询语言样本语句是基于所述表格,通过全排列的方式获得的;

[0015] 将所述自然语言样本语句和所述结构化查询语言样本语句形成第一文本对;

[0016] 将所述第一文本对作为训练数据对转换模型进行训练,得到预训练的转换模型。

[0017] 根据本公开提供的结构化查询语言语句的生成方法,所述对预训练语言模型进行微调之前,包括:

[0018] 获取通用领域中的自然语言语句和所述自然语言语句对应的查询语言语句;

[0019] 将所述查询语言语句进行格式转换,得到结构化查询语言语句;

[0020] 基于所述自然语言语句和所述结构化查询语言语句,形成第二文本对。

[0021] 根据本公开提供的结构化查询语言语句的生成方法,所述对预训练语言模型进行微调,获取微调好的预训练语言模型,包括:

[0022] 将所述第二文本对输入至预训练语言模型中,对预训练语言模型进行微调,获取微调好的预训练语言模型。

[0023] 第二方面,本公开提供一种结构化查询语言语句的生成装置,包括:

[0024] 获取模块,用于获取自然语言语句和所述自然语言语句对应的表格;

[0025] 输出模块,用于将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句;

[0026] 其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。

[0027] 根据本公开提供的结构化查询语言语句的生成方法,所述装置还包括:

[0028] 匹配模块,用于将所述结构化查询语言语句中条件类型和问题类型分别对应的列名与表格中的列名相匹配,确定匹配结果;

[0029] 查询模块,用于基于匹配结果,在表格中进行查询,得出查询结果。

[0030] 根据本公开提供的结构化查询语言语句的生成装置,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的,包括:

[0031] 对预训练语言模型进行微调,获取微调好的预训练语言模型;

[0032] 将结构化查询语言样本语句输入至微调好的预训练语言模型中,生成对应的自然语言样本语句;其中,所述结构化查询语言样本语句是基于所述表格,通过全排列的方式获得的;

[0033] 将所述自然语言样本语句和所述结构化查询语言样本语句形成第一文本对;

[0034] 将所述第一文本对作为训练数据对转换模型进行训练,得到预训练的转换模型。

[0035] 第三方面,本公开还提供一种电子设备,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,所述处理器执行所述程序时实现如上述任一项所述的结构化查询语言语句的生成方法。

[0036] 第四方面,本公开还一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现如上述任一项所述的结构化查询语言语句的生成方法。

[0037] 本公开提供的结构化查询语言语句的生成方法和装置,将提出的自然语言语句以及对应的表格输入至一个预训练的转换模型中,输出自然语言语句对应的结构化查询语言语句。由于转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到,从而使转换模型具有了将特定领域的自然语言语句转换为查询语言语句的能力,这样便可以在应用过程中,准确地获取特定领域的自然语言语句对应的结构化查询语言语句。

附图说明

[0038] 为了更清楚地说明本公开或现有技术中的技术方案,下面将对实施例或现有技术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公开的一些实施例,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。

[0039] 图1是本公开实施例提供的结构化查询语言语句的生成方法的流程示意图;

[0040] 图2是本公开提供的获取预训练的转换模型的流程图;

[0041] 图3是本公开提供的预训练语言模型进行微调之前的流程示意图;

[0042] 图4是本公开提供的结构化查询语言语句的生成装置的结构示意图;

[0043] 图5是本公开提供的电子设备的结构示意图。

具体实施方式

[0044] 为使本公开实施例的目的、技术方案和优点更加清楚,下面将结合本公开实施例中的附图,对本公开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公开实施例一部分实施例,而不是全部的实施例。基于本公开实施例中的实施例,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,都属于本公开实施例保护的范围。

[0045] 自然语言语句转查询语言语句是语义解析、语义理解中的重要任务。单轮会话场景下,给定自然语言语句和数据库表列名集合作为输入,自然语言语句转查询语言语句的目的是解析自然语言语句的语义,建立语句、数据库模式和查询语言三者间的映射,生成语句对应的查询语言语句。

[0046] 本公开实施例提出的一种结构化查询语言语句的生成方法,是针对特定领域的相关数据来训练一个自然语言语句转查询语言语句的模型。在本公开实施例中具体是采用基于预训练语言模型T5PEGASUS结合文本增强技术,设计并开发了自动批量生成特定领域相关的自然语言语句到结构化查询语言语句的文本对的方法,用来生成大量训练数据,以训练一个高质量的自然语言语句转结构化查询语言语句的模型,用来很好的解析用户根据自己特定领域的表格数据提出的问题。

[0047] 图1为本公开实施例提供的结构化查询语言语句的生成方法的流程示意图,该方法包括:

[0048] 步骤110,获取自然语言语句和所述自然语言语句对应的表格。

[0049] 该步骤中,自然语言指的是人类用于书面交流或口头交流的日常语言。自然语言的示例包括中文、英文、德文、西班牙文、法文等等。在以下描述中,将以中文或英文作为自然语言的示例。然而应当理解,这仅仅出于示例的目的,而无意于限制本公开的范围。本公开的实施例可以适用于各种不同的自然语言。

[0050] 自然语言语句指的是有明确的语义、语法正确以及提出问题的句子,如:“A产品在5月份的销售额是多少?”,该语句含有明确的语义信息(询问A产品且在5月份的销售额的值),且合乎正常的语法表达,并且,在该语句中是提出问题的,“是多少?”是个需要回答的语句,如果一个句子为“A产品在5月份的销售额是1万”是不符合本申请获取语句的要求。

[0051] 表格,既是一种可视化交流模式,又是一种组织整理数据的手段,并且,表格表示

了用户的特定领域的一些实体。表头由一些列名组成,每个列名对应于一列。在表上列名必须唯一,不允许重名,不同的列名用于存储不同的数据。比如:以“A产品在5月份的销售额是1万?”为获取的语句,则它对应的表格中存在的列名必须有“产品类型、月份和销售额”,也可以包含其他的列名,如“产品产地、数量”等。“288900、659900、349400”等属于销售额这一列名中的数据,“产品A、产品B、产品C”等属于产品类型这一列名中的数据,“3、4、5”等属于月份这一列名中的数据。可以理解的是,表格可以由本领域技术人员依据实际需求或者应用场景自主设置。

[0052] 对应地,获取的自然语言语句是基于表格提出的,针对用户提供的表格,想要获知表格中某类产品的数据,即获取的语句与表格是相互对应的。

[0053] 步骤120,将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句。

[0054] 其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。

[0055] 该步骤中,结构化查询语言语句指的是SQL语句,具体的SQL (Structured Query Language) 为结构化查询语言,是一种特殊目的的编程语言,是一种数据库查询和程序设计语言,用于存取数据以及查询、更新和管理关系数据库系统。

[0056] 结构化查询语言是高级的非过程化编程语言,允许用户在高层数据结构上工作。它不要求用户指定对数据的存放方法,也不需要用户了解具体的数据存放方式,所以具有完全不同底层结构的不同数据库系统,可以使用相同的结构化查询语言语句作为数据输入与管理的接口。结构化查询语言语句可以嵌套,这使它具有极大的灵活性和强大的功能。

[0057] 将自然语言语句转换为结构化查询语言语句的目的是为了可以计算机能够识别,从而在数据库中方便查询,并能将查询响应的结构化数据反馈给用户。

[0058] 预训练的转换模型是指已经用训练数据集训练好了的模型,且基于自监督方法训练的语言模型,其模型的训练数据与用户表格数据领域相关,该模型能后用来很好的解析用户根据自己领域的表格数据提出的问题。

[0059] 特定领域是指用户上传的表格数据所涉及的领域,根据自己想要查询的数据所建立的一个专属领域,特定领域的体系结构是为共享一个共同特性的一类工作负载而定制的。

[0060] 再者,特定领域是一种为解决特定问题而专门设置的一个区域操作。这点与通用领域不同,而且针对某个特定的领域,所提出的问题能够很自然地方便地表述出来,也常常比通用领域更快地解决问题,得出答案。

[0061] 比如说,用户想要查询关于“品牌A的汽车的相关数据”,特定领域就是指的是针对汽车的品牌进行分类的相关的参数或数据,而通用领域不仅仅涉及到汽车,还会有自行车、电动车领域相关的数据;再比如,用户想要查询涉及人工智能领域相关的算法都有哪些?

[0062] 特定领域就是指用户想要查询的人工智能相关的内容,而通用领域可以为计算机编程相关的算法,而计算机编程相关的算法是包含人工智能领域的,因此,人工智能领域针对用户提出的问题是属于特定领域,而计算机编程领域是通用领域。

[0063] 具体地,在获取的语句为“A产品在5月份的销售额是多少?”的情况下,该语句所对应的表格如表1所示。

[0064] 表1

[0065]	产品类型	月份	产地	销售额
	A	5	中国	288900
	B	6	中国	659900
	C	7	中国	349400

[0066] 将获取的语句“A产品在5月份的销售额是多少?”和表1,输入至预训练的转换模型中,该语句和表1的对应关系是:A产品对应表1中的产品类型中的“A”,5月份对应表1中的“5”,销售额是多少对应表1中销售额的数值。

[0067] 对应输出的结构化查询语言语句,具体表现为:

[0068] [SELECT]

[0069] 销售额

[0070] FROM

[0071] 表格1

[0072] [WHERE-COL]

[0073] 月份

[0074] [WHERE-VAL]

[0075] 5月

[0076] [WHERE-COL]

[0077] 产品类型

[0078] [WHERE-VAL]

[0079] A

[0080] 本公开提供的结构化查询语言语句的生成方法,将提出的自然语言语句以及对应的表格输入至一个预训练的转换模型中,输出自然语言语句对应的结构化查询语言语句。由于转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到,从而使转换模型具有了将特定领域的自然语言语句转换为结构化查询语言语句的能力,这样便可以在应用过程中,准确地获取特定领域的自然语言语句对应的结构化查询语言语句。

[0081] 本公开实施例提供的方法,还包括:

[0082] 将所述结构化查询语言语句中条件类型和问题类型分别对应的列名与表格中的列名相匹配,确定匹配结果。

[0083] 该步骤中,通常结构化查询语言语句是数据查询语句,里面会包含select查询子句,而条件语句则是可选的。除了select语句和where语句,结构化查询语言语句中还可能包括left join外链这样的子语句,或者group分组这样的子语句,或者order排序这样的子语句。由于这样的语句,不涉及选择问题,因此不作为筛选组件中的可选内容。

[0084] 对于复杂的结构化查询语言语句来说,可以对类型可以进行划分,如集合运算类型,条件操作类型,行计算嵌套,问题类型等。针对不同的类型对每个类型再分解成“select”、“where”、“having”、“groupby”、“orderby”等形式的子句,如where子句“where年份=2017”,不同关键字子句却有着相似的结构,比如所有的子句都包含列名,对于每个子句层次进行归纳划分,得到列名。

[0085] 在本公开实施例中选中“select”表示条件类型，“where”表示问题类型。可以理解的是，选用的类型可以由本领域技术人员依据实际需求或者应用场景自主设置，并不做具体的限定。

[0086] 对应的，在“where”表示条件类型，“select”表示问题类型的情况下，select标识是指根据用户实际需求设定用于从解析文本中确定select子句的标识信息。where标识是指根据用户实际需求设定用于从解析文本中确定where子句的标识信息。

[0087] 以获取的语句为“A产品在5月份的销售额是多少？”进行解释说明，并参照表1所示。

[0088] 在该语句中“A产品”和“在5月份”子句表示条件类型，用where表示；“销售额是多少”子句表示问题类型，用select表示。

[0089] 在表1中“产品类型、月份、产地、销售额”表示列名“A、5、中国、288900、B、6、中国、659900、C、7、中国、349400”表示列值。

[0090] 将where中的“A产品和在5月份”子句与表1列名中的“产品类型、月份、销售额”进行匹配；将select中的子句“销售额是多少”与表1列名中的“销售额”进行匹配；匹配完成后，根据表1得出匹配结果为与表格中的第二行相对应，如下表2所示。

[0091] 表2

[0092]

A	5	中国	288900
---	---	----	--------

[0093] 基于匹配结果，在表格中进行查询，得出查询结果。

[0094] 该步骤中，查询的任务主要就是配合数据检索系统的表格结构等元数据解析出相应的子句对应的操作体，即得出结果。

[0095] 具体地，所查询的结果为“288900”，即A产品在5月份的销售额是288900。

[0096] 基于上述任一实施例，该方法中，预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的，参照图2所示，表示获取预训练的转换模型的流程图，包括：

[0097] 步骤210，对预训练语言模型进行微调，获取微调好的预训练语言模型。

[0098] 该步骤中，预训练语言模型可以是BERT (Bidirectional Encoder Representations from Transformers，来自Transformer的双向编码器表示) 语言模型，也可以是T5 PEGSUS预训练语言模型。

[0099] 其中，BERT模型的目标是利用大规模无标注语料训练、获得文本的包含丰富语义信息的表示，然后将文本的语义表示在特定自然语言处理任务中作微调，最终应用于该自然语言处理任务。PEGSUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization)，指专门为摘要定制的预训练模型，T5 PEGSUS预训练语言模型是开源的用于文本生成任务的中文预训练语言模型，在本公开实施例中使用T5 PEGSUS预训练语言模型。

[0100] 微调指的是在预训练语言模型上针对下游任务进行有监督学习，其中，下游任务例如可以包括情感分析任务、阅读理解任务、意图识别任务、命名实体识别任务、语义相似度匹配任务等。

[0101] 不做微调会出现以下问题：从头开始训练，需要大量的数据，计算时间和计算资源；存在模型不收敛，参数不够优化，准确率低，模型泛化能力低，容易过拟合等风险。而进

行微调能够有效避免上述可能存在的问题。

[0102] 步骤220,将结构化查询语言样本语句输入至微调好的预训练语言模型中,生成对应的自然语言样本语句;其中,结构化查询语言样本语句是基于表格,通过全排列的方式获得的。

[0103] 该步骤中,全排列是指从 n 个不同元素中任取 m ($m \leq n$) 个元素,按照一定的顺序排列起来,叫做从 n 个不同元素中取出 m 个元素的一个排列。当 $m=n$ 时所有的排列情况叫全排列。

[0104] 具体地,对这一步骤进行举例说明,并且参照表1所示:

[0105] 结构化查询语言样本语句为:

[0106] “select

[0107] 销售额

[0108] from

[0109] 表1

[0110] Where

[0111] C产品 7月份”

[0112] 则该结构化查询语言样本语句对应的自然语言样本语句为“C产品在7月份的销售额是多少?”。

[0113] 步骤230,将自然语言样本语句和结构化查询语言样本语句形成第一文本对。

[0114] 该步骤中,第一文本对指的是,在特定领域的数据集中,将获取的自然语言样本语句与结构化查询语言样本语句一一对应起来,即每一自然语言语句对应一个结构化查询语言语句,形成类似[自然语言语句,结构化查询语言语句]这样的样式。

[0115] 步骤240,将第一文本对作为训练数据对转换模型进行训练,得到预训练的转换模型。

[0116] 具体地,将步骤230中获取的每一对的[自然语言语句,结构化查询语言语句],作为训练数据对转换模型进行训练,需要说明的是在步骤230获取了大量的第一文本对,基于这些第一文本对模型进行训练。

[0117] 基于上述任一实施例,参照图3所示,表示预训练语言模型进行微调之前的流程示意图,该方法中,包括:

[0118] 步骤310,获取通用领域中的自然语言语句和自然语言语句对应的查询语言语句。

[0119] 该步骤中,通用领域,顾名思义,在某个领域内,全部都可用,包含某类实体的共同属性,与特定领域是相对的。

[0120] 比如说,用户想要查询“关于品牌A的汽车的相关数据”,通用领域指的是包含所有类型的车以及相关属性,汽车,自行车、电动车等。而与通用领域相对的特定领域就是只针对汽车的品牌进行分类的相关的参数或数据。

[0121] 步骤320,将查询语言语句进行格式转换,得到结构化查询语言语句。

[0122] 具体地,格式转换指的是将原有的表达形式转换成另一种表达形式,进行格式转换的目的是对查询语言语句压缩格式化排版,把查询语言语句进行整齐的显示。进行格式转换的方式可以是安装查询语言语句转结构化查询语言语句的相关应用,也可以在线进行转化。可以理解的是,这仅仅出于示例的目的,而无意于限制本公开的范围。

[0123] 以举例方式进行具体说明。

[0124] 获取的查询语言语句为“select销售额from表1Where C产品7月份”

[0125] 对应的,结构化查询语言样本语句为:

[0126] “select

[0127] 销售额

[0128] from

[0129] 表1

[0130] Where

[0131] C产品 7月份”

[0132] 步骤330,基于自然语言语句和结构化查询语言语句,形成第二文本对。

[0133] 该步骤中,第二文本对指的是,在通用领域的数据集中,将获取的自然语言语句和结构化查询语言语句一一对应起来,即每一自然语言语句对应一个结构化查询语言语句,形成类似[自然语言语句,结构化查询语言语句]这样的样式。

[0134] 基于上述任一实施例,该方法中,步骤210具体包括:

[0135] 将第二文本对输入至预训练语言模型中,对预训练语言模型进行微调,获取微调好的预训练语言模型。

[0136] 该步骤中,基于步骤330中获取的第二文本对[自然语言语句,结构化查询语言语句],将第二文本对输入至T5 PEGSUS预训练语言模型中,T5 PEGSUS预训练语言模型基于获取的第二文本对进行微调。

[0137] 进一步地,对本公开的实施例做进一步补充说明:

[0138] 具体方法如下:

[0139] (1)对于通用领域的自然语言语句到查询语言语句的文本对数据,我们将查询语言语句根据属性名(表格列名)、属性值(表格内的列值)结合查询语言语句关键字如SELECT(问题类型)、WHERE(条件类型)和聚合操作符(如MAX,MIN,COUNT,SUM,AVG)进行格式转换,转换之后的文本对为[自然语言语句,结构化查询语言语句],如[XXX产品5月的销售额是多少?,[SELECT]销售额[WHERE-COL]月份[WHERE-VAL]5月[WHERE-COL]产品名[WHERE-VAL]XXX]。其中“[SELECT]销售额”是查询语言语句中SELECT关键字及其对应的表格列名,“[WHERE-COL]产品名和[WHERE-VAL]XXX”分别是查询语言语句中WHERE关键字及其对应的表格列名和列值,COL代表列名,VAL代表列值。该数据来自于通用领域的自然语言语句到结构化查询语言语句的文本对数据。

[0140] (2)使用T5 PEGSUS预训练语言模型,在上述格式的通用领域的自然语言语句到结构化查询语言语句的文本对数据上进行微调(fine tuning)。T5 PEGSUS预训练语言模型是开源的用于文本生成任务的中文预训练语言模型。

[0141] (3)根据用户上传的特定领域数据表格,通过全排列的形式,枚举出大量的如(1)中提及的“结构化查询语言语句”,其格式为“[SELECT]销售额[WHERE-COL]月份[WHERE-VAL]5月[WHERE-COL]产品名[WHERE-VAL]XXX”。

[0142] 其中,枚举指的是列出某些有穷序列集的所有成员的程序,或者是一种特定类型对象的计数。这两种类型经常(但不总是)重叠。是一个被命名的整型常数的集合,例如表示星期的SUNDAY、MONDAY、TUESDAY、WEDNESDAY、THURSDAY、FRIDAY、SATURDAY就是一个枚举。

[0143] (4) 使用(2)中微调好的T5 PEGSUS预训练语言模型,输入(3)中枚举出的结构化查询语言语句,生成对应的自然语言语句,最终形成了[自然语言语句,结构化查询语言语句]文本对,基于形成的[自然语言语句,结构化查询语言语句]文本对作为自然语言语句转结构化查询语言语句模型的训练语料。

[0144] (5) 使用(4)中生成的自然语言语句到结构化查询语言语句的文本对作为训练数据,训练一个基于用户特定领域表格数据相关的自然语言语句转结构化查询语言语句模型。

[0145] 基于上述任一实施例,图4为本公开实施例提供的结构化查询语言语句的生成装置的结构示意图,如图4所示,该装置包括获取模块410和输出模块420。

[0146] 获取模块410,用于获取自然语言语句和自然语言语句对应的表格。

[0147] 输出模块420,用于将自然语言语句和对应的表格输入至预训练的转换模型中,输出自然语言语句对应的结构化查询语言语句。

[0148] 其中,预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。

[0149] 本公开提供的结构化查询语言语句的生成装置,将提出的自然语言语句以及对应的表格输入至一个预训练的转换模型中,输出自然语言语句对应的结构化查询语言语句。由于转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到,从而使转换模型具有了将特定领域的自然语言语句转换为查询语言语句的能力,这样便可以在应用过程中,准确地获取特定领域的自然语言语句对应的结构化查询语言语句。

[0150] 基于上述任一实施例,所述装置还包括匹配模块和查询模块。其中,匹配模块,用于将结构化查询语言语句中条件类型和问题类型分别对应的列名与表格中的列名相匹配,确定匹配结果。查询模块,用于基于匹配结果,在表格中进行查询,得出查询结果。

[0151] 基于上述任一实施例,预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的,包括:

[0152] 获取子单元,用于对预训练语言模型进行微调,获取微调好的预训练语言模型。

[0153] 生成子单元,用于将结构化查询语言样本语句输入至微调好的预训练语言模型中,生成对应的自然语言样本语句;其中,结构化查询语言样本语句是基于表格,通过全排列的方式获得的。

[0154] 形成子单元,用于将自然语言样本语句和结构化查询语言样本语句形成第一文本对。

[0155] 训练子单元,用于将第一文本对作为训练数据对转换模型进行训练,得到预训练的转换模型。

[0156] 基于上述任一实施例,在进行获取子单元对预训练语言模型进行微调之前,包括:

[0157] 获取通用领域中的自然语言语句和自然语言语句对应的查询语言语句。

[0158] 将查询语言语句进行格式转换,得到结构化查询语言语句。

[0159] 基于自然语言语句和结构化查询语言语句,形成第二文本对。

[0160] 基于上述任一实施例,该装置中的获取子单元具体包括:

[0161] 将第二文本对输入至预训练语言模型中,对预训练语言模型进行微调,获取微调

好的预训练语言模型。

[0162] 图5示例了一种电子设备的实体结构示意图,如图5所示,该电子设备可以包括:处理器(processor)510、通信接口(Communications Interface)520、存储器(memory)530和通信总线540,其中,处理器510,通信接口520,存储器530通过通信总线540完成相互间的通信。处理器510可以调用存储器530中的逻辑指令,以执行结构化查询语言语句的生成方法,该方法包括:获取自然语言语句和所述自然语言语句对应的表格;将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句;其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。

[0163] 此外,上述的存储器530中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读取存储介质中。基于这样的理解,本公开实施例的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,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)执行本公开各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0164] 另一方面,本公开还提供一种计算机程序产品,所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序,所述计算机程序包括程序指令,当所述程序指令被计算机执行时,计算机能够执行上述各方法所提供的结构化查询语言语句的生成方法,该方法包括:获取自然语言语句和所述自然语言语句对应的表格;将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句;其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。

[0165] 又一方面,本公开还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现以执行上述各提供的结构化查询语言语句的生成方法,该方法包括:获取自然语言语句和所述自然语言语句对应的表格;将所述自然语言语句和所述对应的表格输入至预训练的转换模型中,输出所述自然语言语句对应的结构化查询语言语句;其中,所述预训练的转换模型是基于特定领域的自然语言样本语句和结构化查询语言样本语句作为训练数据得到的。

[0166] 以上所描述的装置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。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0167] 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必需的通用领域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现有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。

[0168] 最后应说明的是:以上实施例仅用以说明本公开的技术方案,而非对其限制;尽管参照前述实施例对本公开进行了详细的说明,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公开各实施例技术方案的精神和范围。

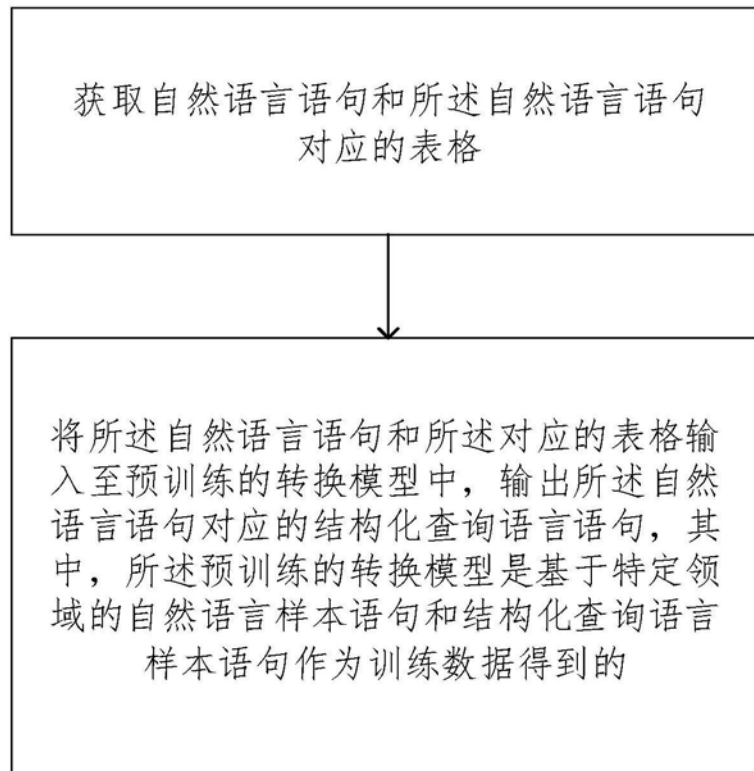


图1

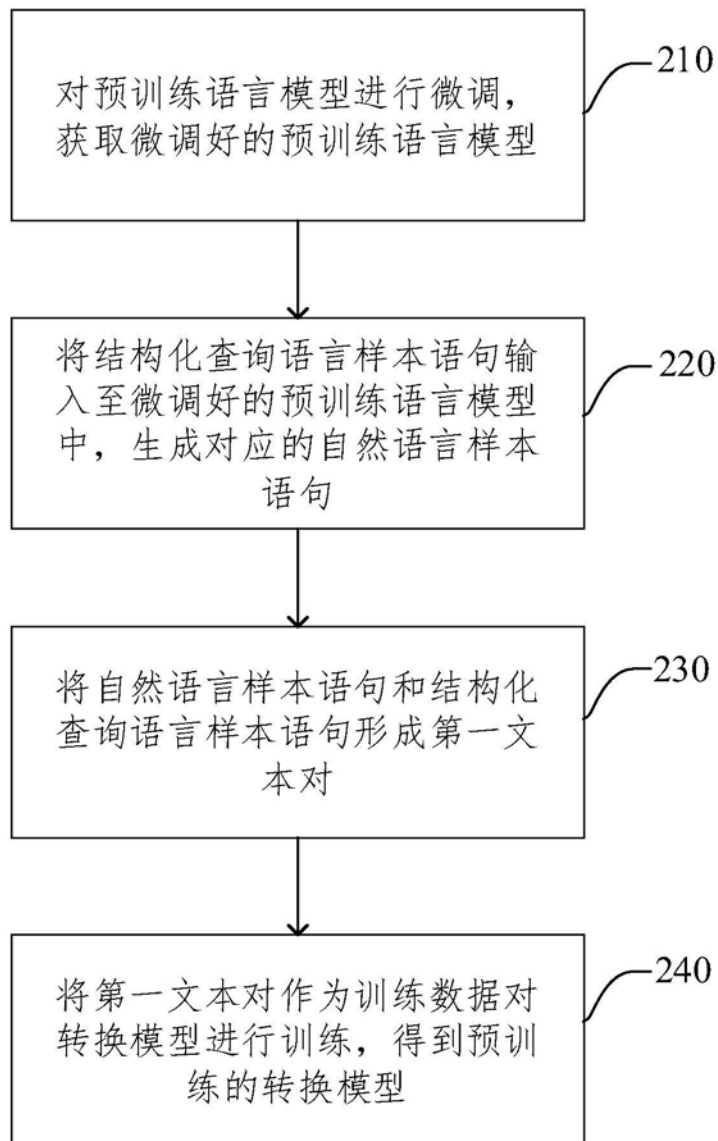


图2

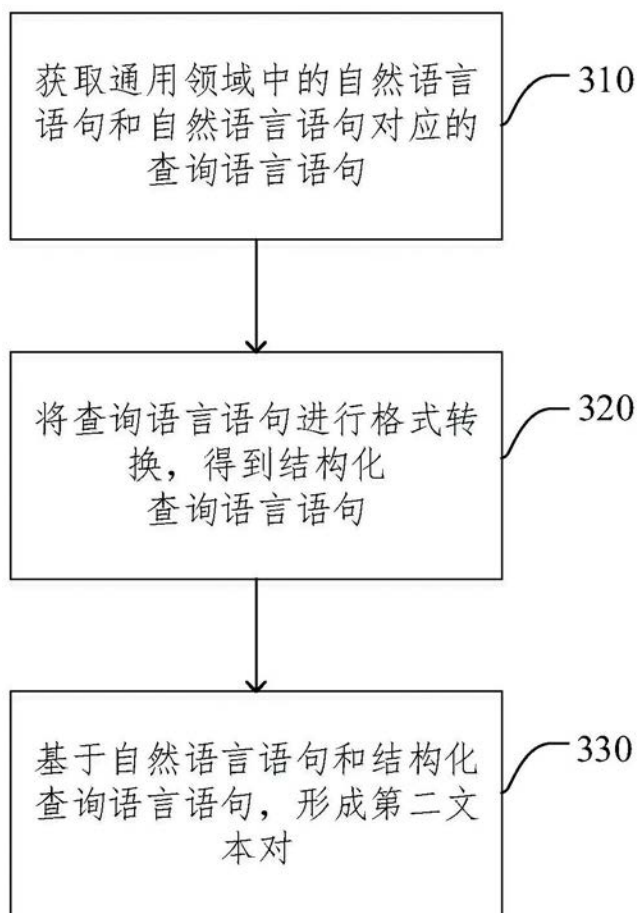


图3

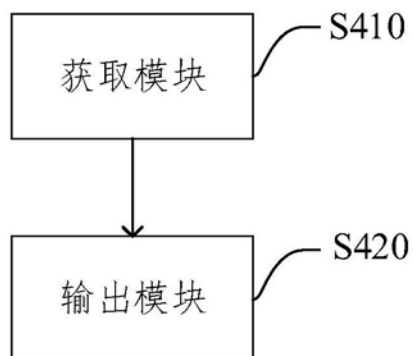


图4

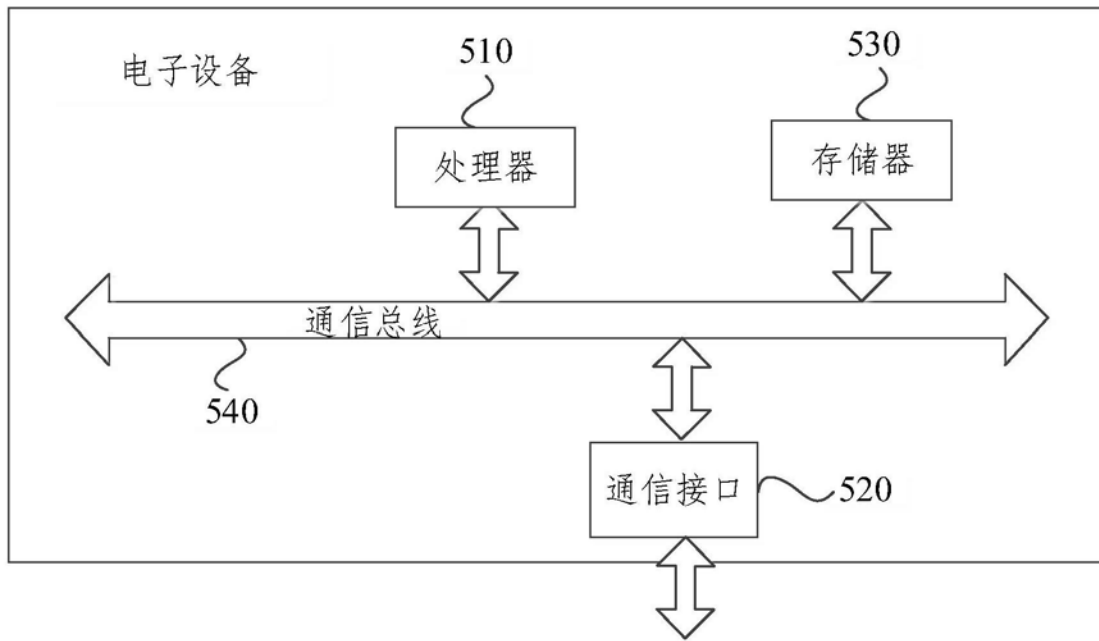


图5