



(12) 发明专利申请

(10) 申请公布号 CN 114610742 A

(43) 申请公布日 2022. 06. 10

(21) 申请号 202210157135.3

(22) 申请日 2022.02.21

(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 袁韶祖 陈蒙 祝天刚

(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 40/289 (2020.01)

G06F 40/30 (2020.01)

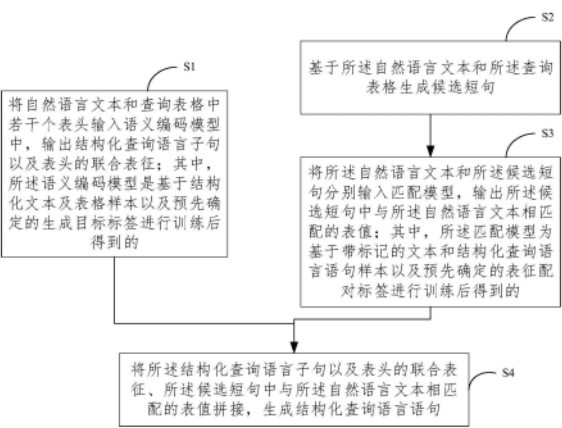
权利要求书2页 说明书11页 附图3页

(54) 发明名称

结构化查询语言语句生成方法和系统

(57) 摘要

本公开提供一种结构化查询语言语句生成方法和系统,所述生成方法包括:将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的;基于自然语言文本和查询表格生成候选短句,将自然语言文本和候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;将结构化查询语言子句以及表头的联合表征、表值拼接,生成结构化查询语言语句。本公开保证了模型运行效率,显著提升了匹配的准确性。



1. 一种结构化查询语言语句生成方法,其特征在于,包括:

将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的;

基于所述自然语言文本和所述查询表格生成候选短句;

将所述自然语言文本和所述候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;其中,所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的;

将所述结构化查询语言子句以及表头的联合表征、所述候选短句中与所述自然语言文本相匹配的表值拼接,生成结构化查询语言语句。

2. 根据权利要求1所述的结构化查询语言语句生成方法,其特征在于,所述基于所述自然语言文本和所述查询表格生成候选短句包括:

基于所述自然语言文本,在查询表格中选中候选单元,选择所述查询表格中与所述候选单元对应的表值;

依次将候选单元和所述候选短句中与所述自然语言文本相匹配的表值拼接,生成所述候选短句。

3. 根据权利要求1所述的结构化查询语言语句生成方法,其特征在于,所述将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征,包括:

将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出若干个子任务对应的结构化查询语言子句以及表头的联合表征;

将若干个子任务对应的结构化查询语言子句以及表头的联合表征进行拼接,形成结构化查询语言子句以及表头的联合表征。

4. 根据权利要求1所述的结构化查询语言语句生成方法,其特征在于,所述语义编码模型在训练过程中采用的损失函数为:

$$\text{Loss}_1 = - \sum_{i=1}^{n_1} y_i^1 \log \hat{y}_i^1$$

其中, y_i^1 表示预先确定的生成目标标签, $\hat{y}_i^1$ 表示语义编码模型训练时的输出结果, n_1 表示语义编码模型输出结果的分类数。

5. 根据权利要求1所述的结构化查询语言语句生成方法,其特征在于,所述匹配模型在训练过程中采用的损失函数为:

$$\text{Loss}_2 = - \sum_{i=1}^{n_2} y_i^2 \log \hat{y}_i^2$$

其中, y_i^2 表示预先确定的表征配对标签, $\hat{y}_i^2$ 表示匹配模型训练时的输出结果, n_2 表示匹配模型输出结果的分类数。

6. 一种结构化查询语言语句生成系统,其特征在于,包括:

语义编码模块,用于将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的;

候选短句生成模块,用于基于所述自然语言文本和所述查询表格生成候选短句;

表值匹配模块,用于将所述自然语言文本和所述候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;其中,所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的;

拼接模块,用于将所述结构化查询语言子句以及表头的联合表征、所述候选短句中与所述自然语言文本相匹配的表值拼接,生成结构化查询语言语句。

7.根据权利要求6所述的结构化查询语言语句生成系统,其特征在于,所述候选短句生成模块,具体用于:

基于所述自然语言文本,在查询表格中选中候选单元,选择所述查询表格中与所述候选单元对应的表值;

依次将候选单元和所述候选短句中与所述自然语言文本相匹配的表值拼接,生成所述候选短句。

8.根据权利要求6所述的结构化查询语言语句生成系统,其特征在于,所述语义编码模块具体用于:

将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出若干个子任务对应的结构化查询语言子句以及表头的联合表征;

将若干个子任务对应的结构化查询语言子句以及表头的联合表征进行拼接,形成结构化查询语言子句以及表头的联合表征。

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任一项所述的结构化查询语言语句生成方法的步骤。

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任一项所述的结构化查询语言语句生成方法的步骤。

结构化查询语言语句生成方法和系统

技术领域

[0001] 本公开涉及自然语言处理业务技术领域,尤其涉及一种结构化查询语言语句生成方法和系统、电子设备和存储介质。

背景技术

[0002] 把人类的自然语言转换成成机器能理解的结构化查询语言语句,从数据库查询答案返回,这样可以充当语言-数据库的智能桥梁,让不熟悉数据库的用户只通过自然语言就能够快速地查询到自己想要的数据库,该技术在人机交互中有巨大的价值。目前,基于分解任务结构化查询语言生成方式中的表值生成非常复杂,这一部分需要两个子任务协同完成:先通过序列标注从自然语言中选出可能的候选表值,因为可能的表值有多个;再通过匹配将选出的多个表值和对应的条件子句对应上。此外,还需要一个后处理操作。由于标注任务和匹配任务都是相对复杂的操作,这既增加了模型层面的复杂度,同时后处理也非常繁琐,方法不简洁,非常容易出错。

发明内容

[0003] 本公开提供一种结构化查询语言语句生成方法和系统、电子设备和存储介质,用以解决现有技术中存在的技术缺陷。

[0004] 本公开提供一种结构化查询语言语句生成方法,包括:

[0005] 将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的;

[0006] 基于所述自然语言文本和所述查询表格生成候选短句;

[0007] 将所述自然语言文本和所述候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;其中,所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的;

[0008] 将所述结构化查询语言子句以及表头的联合表征、所述候选短句中与所述自然语言文本相匹配的表值拼接,生成结构化查询语言语句。

[0009] 根据本公开提供的结构化查询语言语句生成方法,其中,所述基于所述自然语言文本和所述查询表格生成候选短句包括:

[0010] 基于所述自然语言文本,在查询表格中选中候选单元,选择所述查询表格中与所述候选单元对应的表值;

[0011] 依次将候选单元和所述候选短句中与所述自然语言文本相匹配的表值拼接,形成所述候选短句。

[0012] 根据本公开提供的结构化查询语言语句生成方法,其中,所述将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征,包括:

[0013] 将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出若干个子任务对应的结构化查询语言子句以及表头的联合表征;

[0014] 将若干个子任务对应的结构化查询语言子句以及表头的联合表征进行拼接,形成结构化查询语言子句以及表头的联合表征。

[0015] 根据本公开提供的结构化查询语言语句生成方法,其中,所述语义编码模型在训练过程中采用的损失函数为:

$$[0016] \quad \text{Loss}_1 = - \sum_{i=1}^{n_1} y_i^1 \log \hat{y}_i^1$$

[0017] 其中, y_i^1 表示预先确定的生成目标标签, $\hat{y}_i^1$ 表示语义编码模型训练时的输出结果, n_1 表示语义编码模型输出结果的分类数。

[0018] 根据本公开提供的结构化查询语言语句生成方法,其中,所述匹配模型在训练过程中采用的损失函数为:

$$[0019] \quad \text{Loss}_2 = - \sum_{i=1}^{n_2} y_i^2 \log \hat{y}_i^2$$

[0020] 其中, y_i^2 表示预先确定的表征配对标签, $\hat{y}_i^2$ 表示匹配模型训练时的输出结果, n_2 表示匹配模型输出结果的分类数。

[0021] 本公开还提供了一种结构化查询语言语句生成系统,包括:

[0022] 语义编码模块,用于将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的;

[0023] 候选短句生成模块,用于基于所述自然语言文本和所述查询表格生成候选短句;

[0024] 表值匹配模块,用于将所述自然语言文本和所述候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;其中,所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的;

[0025] 拼接模块,用于将所述结构化查询语言子句以及表头的联合表征、所述候选短句中与所述自然语言文本相匹配的表值拼接,生成结构化查询语言语句。

[0026] 根据本公开提供的结构化查询语言语句生成系统,所述候选短句生成模块,具体用于:

[0027] 基于所述自然语言文本,在查询表格中选中候选单元,选择所述查询表格中与所述候选单元对应的表值;

[0028] 依次将候选单元和所述候选短句中与所述自然语言文本相匹配的表值拼接,生成所述候选短句。

[0029] 根据本公开提供的结构化查询语言语句生成系统,其中,所述语义编码模块具体用于:

[0030] 将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出若干个子任务对应的结构化查询语言子句以及表头的联合表征;

[0031] 将若干个子任务对应的结构化查询语言子句以及表头的联合表征进行拼接,形成

结构化查询语言子句以及表头的联合表征。

[0032] 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种所述结构化查询语言语句生成方法的步骤。

[0033] 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如上述任一种所述结构化查询语言语句生成方法的步骤。

[0034] 本公开通过语义编码模型生成结构化查询语言子句以及表头的联合表征,通过匹配模型生成自然语言文本和候选短句相匹配的表值,然后将结构化查询语言子句以及表头的联合表征以及表值相拼接形成结构化查询语言语句,采用了高效的两段式生成方法,减少了子任务数量并且不需要做后处理,简化了结构化查询语言语句生成流程,保证了模型运行效率,经过调优改良后的两段模型显著提升了匹配的准确性。

附图说明

[0035] 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。

[0036] 图1是本公开实施例提供的结构化查询语言语句生成方法的流程示意图;

[0037] 图2是本公开实施例提供的结构化查询语言语句生成方法中的语义编码模型的示意图;

[0038] 图3是本公开实施例提供的结构化查询语言语句生成方法中的匹配模型的示意图;

[0039] 图4是本公开提供的结构化查询语言语句生成系统的示意图;

[0040] 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

[0041] 为使本公开的目的、技术方案和优点更加清楚,下面将结合本公开中的附图,对本公开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一部分实施例,而不是全部的实施例。基于本公开中的实施例,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,都属于本公开保护的范围。

[0042] 本公开实施例公开了一种结构化查询语言语句生成方法,参见图1,所述结构化查询语言语句生成方法包括:

[0043] S1、将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的。

[0044] 本实施例中,结构化查询语言语句(Structured Query Language,SQL)为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。

[0045] SQL(Structured Query Language,结构化查询语言)语句本身是没法输入到语义

编码模型里学习,需要将SQL语句改成文本语句,就有了结构化查询语言子句以及表头的联合表征。其中,语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的。语义编码模型生成的内容不包含value(表值),通过语义编码模型可以得到除了value以外的所有内容,也就是生成了除了value以外的结构化查询语言子句以及表头的联合表征。查询表格中若干个表头可以是查询表格中第一行的所有表头。

[0046] 模型输入为自然语言文本和查询表格中的多个表头,输出为结构化查询语言子句以及表头的联合表征。如图2所示,可以接着通过不同的子任务获取到Agg(max,min等聚合操作),sel_column(选择子句的列),op(=,>等),where_column(条件子句的列),conn(AND等连接符)的具体值。这部分SQL语句生成方法可以采取各个击破的原理,把一个SQL语句生成问题分解成多个子任务。

[0047] S2、基于所述自然语言文本和所述查询表格生成候选短句;

[0048] S3、将所述自然语言文本和所述候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;其中,所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的;

[0049] 在自然语言文本和候选短句中进行一一匹配,如果自然语言文本和候选短句相匹配,则被选中,代表这个value是当前候选单元的值,反之,如果自然语言文本和候选短句不匹配,在不被选中,继续匹配下一个,直到找到匹配的value。所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的,利用匹配模型可以生成自然语言文本和候选短句相匹配的表值,优势是将原本的序列标注和匹配两个任务简化为了一个匹配任务,同时因为候选值是取自查询表格中的value,是实际存在的,因而不需要对value修正和配置额外的词表。匹配模型的输出是输出是二分类,即当前的value是否为符合SQL语句的value,这个过程需要把表中一列做为候选单元,然后将一列的value都过一遍,最终是得到一个符合匹配条件的value,也就是通过匹配模型可以直接得到value。

[0050] S4、将所述结构化查询语言子句以及表头的联合表征、所述候选短句中与所述自然语言文本相匹配的表值拼接,生成结构化查询语言语句。

[0051] 将S1得到的结构化查询语言子句以及表头的联合表征和S3得到的所述候选短句中与所述自然语言文本相匹配的表值拼接在一起,可以生成完整的结构化查询语言语句。S3得到的value要和语义编码模型的输出合并成最终的SQL语句,也就是语义编码模型生成的8个子任务加上匹配模型生成的value共同构成SQL语句。

[0052] 本公开实施例通过语义编码模型生成结构化查询语言子句以及表头的联合表征,通过匹配模型生成自然语言文本和候选短句相匹配的表值,然后将结构化查询语言子句以及表头的联合表征以及表值相拼接形成结构化查询语言语句,采用了高效的两段式生成方法,减少了子任务数量并且不需要做后处理,简化了结构化查询语言语句生成流程,保证了模型运行效率,经过调优改良后的两段模型显著提升了匹配的准确性。

[0053] 根据本公开实施例提供的结构化查询语言语句生成方法,其中,所述基于所述自然语言文本和所述查询表格生成候选短句包括:

[0054] 基于所述自然语言文本,在查询表格中选中候选单元,选择所述查询表格中与所述候选单元对应的表值;

[0055] 依次将候选单元和所述候选短句中与所述自然语言文本相匹配的表值拼接,形成所述候选短句。

[0056] 如图3所示,图3中的示例是通过选中的where(条件子句)列(日期)将表中对应的value(昨天、今天、明天)作为候选全部取出来,然后依次将候选单元和value值拼接成为几个候选短句。

[0057] 根据本公开实施例提供的结构化查询语言语句生成方法,其中,所述将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征,包括:

[0058] 将自然语言文本和查询表格中若干个表头输入语义编码模型中,输出若干个子任务对应的结构化查询语言子句以及表头的联合表征;

[0059] 将若干个子任务对应的结构化查询语言子句以及表头的联合表征进行拼接,形成结构化查询语言子句以及表头的联合表征。

[0060] 多个子任务均是语义编码模型生成的,也就是SQL语句的各个子句是分开生成的,语义编码模型的输入都是自然语言文本和查询表格中若干个表头,最终输出若干个子任务对应的结构化查询语言子句以及表头的联合表征。然后将若干个子任务对应的结构化查询语言子句以及表头的联合表征进行拼接,就形成了结构化查询语言子句以及表头的联合表征。

[0061] 根据本公开实施例提供的结构化查询语言语句生成方法,其中,所述语义编码模型在训练过程中采用的损失函数为:

$$[0062] \quad \text{Loss}_1 = - \sum_{i=1}^{n_1} y_i^1 \log \hat{y}_i^1$$

[0063] 其中, y_i^1 表示预先确定的生成目标标签, $\hat{y}_i^1$ 表示语义编码模型训练时的输出结果, n_1 表示语义编码模型输出结果的分类数。在语义编码模型在训练过程中,有多个任务,每个任务对应的 n_1 的数字也可能会不同,如果有max和min两个字段就叫2分类,如果有5个叫5分类。

[0064] 通过执行多次迭代操作对所述语义编码模型进行训练,也就是深度学习的过程,可以根据所述语义编码模型在训练过程中采用的损失函数确定当前的迭代操作是否达到预设的收敛条件,若是达到收敛条件,可以停止迭代。

[0065] 根据本公开实施例提供的结构化查询语言语句生成方法,其中,所述匹配模型在训练过程中采用的损失函数为:

$$[0066] \quad \text{Loss}_2 = - \sum_{i=1}^{n_2} y_i^2 \log \hat{y}_i^2$$

[0067] 其中, y_i^2 表示预先确定的表征配对标签, $\hat{y}_i^2$ 表示匹配模型训练时的输出结果, n_2 表示匹配模型输出结果的分类数。该训练过程中,训练时的输出结果具有两类,第一类是输出自然语言文本和候选短句相匹配,另一类是输出自然语言文本和候选短句不匹配,此时, n_2 为2。

[0068] 通过执行多次迭代操作对所述匹配模型进行训练,也就是深度学习的过程,可以

根据所述匹配模型在训练过程中采用的损失函数确定当前的迭代操作是否达到预设的收敛条件,若是达到收敛条件,可以停止迭代。

[0069] 根据本公开实施例提供的结构化查询语言语句生成方法,其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的,包括:

[0070] 所述语义编码模型是基于RoBERTa模型、结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的。

[0071] 语义编码模型可以选用RoBERTa这种已经通过通用中文预料训练好的模型来进一步优化,相当于RoBERTa模型已经有了一个初始化的权重,可以理解为在大量通用语料上训练完了之后直接用于语义编码模型,可以直接在RoBERTa模型的基础上进一步通过样本进行训练,其中,RoBERTa主要是在BERT (Bidirectional Encoder Representations from Transformers,词向量)基础上做了几点调整:1) 训练时间更长,batch size更大,训练数据更多;2) 移除了next predict loss;3) 训练序列更长;4) 动态调整Masking机制。RoBERTa相较于BERT进行了多项改进和调整,它的预训练使用了更多的数据。在实验中验证该模型比BERT效果提升1%。

[0072] 为了进一步地理解本实施例的方法,在一个具体实例中,本实施例的结构化查询语言语句生成方法包括:

[0073] 首先,利用第一段模型(也就是语义编码模型),生成除value(可能是数值型和字符型)以外的结构化查询子句:

[0074] 第一段模型为生成模型,第一段模型是利用结构化文本(text-SQL)及表格样本(对应的一个表格)以及预先确定的生成目标标签训练的到的,也就是SQL语句各个部分。第一段模型的输入是一致的,输出是包含8个子任务。其中6个为基础任务,s_sum:select了多少个;w_sum:where有多少列;order:是否存在order by;agg(聚合操作):当前列的agg(max,min,count)操作是什么;conn_op(连接符)当前列和value的运算符(>,<=)是什么;conns_op:当前SQL语句的连接符(And,or)是什么。

[0075] 除了以上6个基本任务之外,还增加了两个辅助分类任务,即当前列是否被select(选择子句)选中 and 当前where column(条件子句列)是否被选中。这两个子任务会让模型有更多的学习标签,能够充分利用好训练数据,对于其他子任务有促进作用。增加的两个子任务准确率提升约1%。

[0076] 和其他模型不同的是,提了进一步提高模型的效果,我们将预训练模型BERT替换为语义效果更好的RoBERTa,RoBERTa相较于BERT进行了多项改进和调整,它的预训练使用了更多的数据。在实验中验证该模型比BERT效果提升1%。

[0077] 其次,利用第二段模型(匹配模型),生成输出所述候选短句中与所述自然语言文本相匹配的表值:

[0078] 第二段模型我们通过选中的where列将表中对应的value全部取出来,然后依次将这些列和value拼成的短句与自然语言中进行一一匹配,如果选中就是该where列的value值。这样就将原本的序列标注和匹配两个任务简化为了一个匹配任务,因为使用了表中的value值,因此得到的SQL语句的value是正确的,不需要后处理并配置额外的词表。

[0079] 为了保证模型的运算速度,这里舍弃了使用预训练的BERT作为匹配模型,而是使用了轻量级的匹配模型CDSSM。该模型仅由几层CNN组成,推理速度极快,因此几乎不会降低

模型效率。由于模型相对简单,模型需要细致的优化才能取得较好效果。具体优化如下:

[0080] 1) CDSSM默认模型使用卷积核长度为3,即 $\text{filter}=3$,经验证效果极差;经调参后发现卷积核越大效果越差,最终实验得出 $\text{filter}=1,2$ (卷积核同时使用1和2)效果最好。这是因为该匹配任务匹配的value为单词或短语,长度较短,因此设置小的卷积核有利于更好的理解语义。

[0081] 2) CDSSM训练时需要预先将样本中遇到的词初始化成词向量,若只将query初始化,可能不能覆盖Table中的单词,会出现OOV(out of vocabulary)问题;于是将table中出现的词也加入到词表中,降低了OOV的出现频率。

[0082] 3) 此外,在实验验证中,发现增大训练集正负样本的比率,同时增高epoch轮数,设置训练集样本正负比例=1:20,模型迭代数为40epoch时模型能够取得最佳效果。

[0083] 表1

[0084]	模型	训练集样本(正负)	value ACC	迭代 数	卷积 核	total ACC
	CDSSM	1: 10	30%	20	3	28%
	CDSSM	1: 1	41.9%	20	3	39.3%
	CDSSM	1: 1	59%	20	1, 2	55%
	CDSSM	1: 10 (并加入 table 中的词表)	69.4%	20	1, 2	63.8%
	CDSSM	1: 15 (并加入 table 中的词表)	83%	40	1, 2	76%
	CDSSM	1: 20 (并加入 table 中的词表)	87%	40	1, 2	78.3%

[0085] 上表1为本模块的调参实验,实验数据为在真实业务场景中经过人工标注的1w条高质量数据,经过调优value单项生成准确率由30%提升至87%,最终SQL语句生成结果准确率为78.3%。

[0086] 最终,将第一段模型生成的除value以外的结构化查询子句和第二段模型生成的输出所述候选短句中与所述自然语言文本相匹配的所述候选短句中与所述自然语言文本相匹配的表值拼接,得到完整的结构化查询语言语句。

[0087] 本公开实施例公开了一种结构化查询语言语句生成系统,参见图4,所述结构化查询语言语句生成系统包括:

[0088] 语义编码模块10,用于将自然语言文本和表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的;

[0089] 候选短句生成模块20,用于基于所述自然语言文本和所述查询表格生成候选短句;

[0090] 表值匹配模块30,用于将所述自然语言文本和所述候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;其中,所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的;

[0091] 拼接模块40,用于将所述结构化查询语言子句以及表头的联合表征、所述候选短句中与所述自然语言文本相匹配的表值拼接,生成结构化查询语言语句。

[0092] 将语义编码模块10得到的结构化查询语言子句以及表头的联合表征和表值匹配模块30得到的所述候选短句中与所述自然语言文本相匹配的表值拼接在一起,可以生成完整的结构化查询语言语句。表值匹配模块30得到的value要和语义编码模型的输出合并成最终的SQL语句,也就是语义编码模型生成的8个子任务加上匹配模型生成的value共同构成SQL语句。

[0093] 根据本公开实施例提供的结构化查询语言语句生成系统,所述候选短句生成模块20,具体用于:

[0094] 基于所述自然语言文本,在查询表格中选中候选单元,选择所述查询表格中与所述候选单元对应的表值;

[0095] 依次将候选单元和所述候选短句中与所述自然语言文本相匹配的表值拼接,形成所述候选短句。

[0096] 可以通过选中的where(条件子句)列(日期)将表中对应的value(昨天、今天、明天)作为候选全部取出来,然后依次将候选单元和value值拼接成为几个候选短句。

[0097] 根据本公开实施例提供的结构化查询语言语句生成系统,其中,所述语义编码模块10具体用于:

[0098] 将自然语言文本和表格中若干个表头输入语义编码模型中,输出若干个子任务对应的结构化查询语言子句以及表头的联合表征;

[0099] 将若干个子任务对应的结构化查询语言子句以及表头的联合表征进行拼接,形成结构化查询语言子句以及表头的联合表征。

[0100] 多个子任务均是语义编码模型生成的,也就是SQL的各个子句是分开生成的,语义编码模型的输入都是自然语言文本和查询表格中第一行的所有表头,最终输出若干个子任务对应的结构化查询语言子句以及表头的联合表征。然后将若干个子任务对应的结构化查询语言子句以及表头的联合表征进行拼接,就形成了结构化查询语言子句以及表头的联合表征。

[0101] 根据本公开实施例提供的结构化查询语言语句生成系统,其中,所述语义编码模型在训练过程中采用的损失函数为:

$$[0102] \quad \text{Loss}_1 = - \sum_{i=1}^{n_1} y_i^1 \log \hat{y}_i^1$$

[0103] 其中, y_i^1 表示预先确定的生成目标标签, $\hat{y}_i^1$ 表示语义编码模型训练时的输出结果, n_1 表示语义编码模型输出结果的分类数。

[0104] 通过执行多次迭代操作对所述语义编码模型进行训练,也就是深度学习的过程,

可以根据所述语义编码模型在训练过程中采用的损失函数确定当前的迭代操作是否达到预设的收敛条件,若是达到收敛条件,可以停止迭代。

[0105] 根据本公开实施例提供的结构化查询语言语句生成系统,其中,所述匹配模型在训练过程中采用的损失函数为:

$$[0106] \quad \text{Loss}_2 = - \sum_{i=1}^{n_2} y_i^2 \log \hat{y}_i^2$$

[0107] 其中, y_i^2 表示预先确定的表征配对标签, $\hat{y}_i^2$ 表示匹配模型训练时的输出结果, n_2 表示匹配模型输出结果的分类数。

[0108] 通过执行多次迭代操作对所述匹配模型进行训练,也就是深度学习的过程,可以根据所述匹配模型在训练过程中采用的损失函数确定当前的迭代操作是否达到预设的收敛条件,若是达到收敛条件,可以停止迭代。

[0109] 根据本公开实施例提供的结构化查询语言语句生成系统,其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的,包括:

[0110] 所述语义编码模型是基于RoBERTa模型、结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的。

[0111] 语义编码模型可以选用RoBERTa这种已经通过通用中文预料训练好的模型来进一步优化,相当于RoBERTa模型已经有了一个初始化的权重,可以理解为在大量通用语料上训练完了之后直接用于语义编码模型,可以直接在RoBERTa模型的基础上进一步通过样本进行训练,其中,RoBERTa主要是在BERT(Bidirectional Encoder Representations from Transformers,词向量)基础上做了几点调整:1) 训练时间更长,batch size更大,训练数据更多;2) 移除了next predict loss;3) 训练序列更长;4) 动态调整Masking机制。RoBERTa相较于BERT进行了多项改进和调整,它的预训练使用了更多的数据。在实验中验证该模型比BERT效果提升1%。

[0112] 图5示例了一种电子设备的实体结构示意图,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,以执行一种结构化查询语言语句生成方法,该方法包括:

[0113] S1、将自然语言文本和表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的;

[0114] S2、基于所述自然语言文本和所述查询表格生成候选短句;

[0115] S3、将所述自然语言文本和所述候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;其中,所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的;

[0116] S4、将所述结构化查询语言子句以及表头的联合表征、所述候选短句中与所述自然语言文本相匹配的表值拼接,生成结构化查询语言语句。

[0117] 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时,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。

[0118] 另一方面,本公开实施例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,计算机能够执行一种结构化查询语言语句生成方法,该方法包括:

[0119] S1、将自然语言文本和表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的;

[0120] S2、基于所述自然语言文本和所述查询表格生成候选短句;

[0121] S3、将所述自然语言文本和所述候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;其中,所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的;

[0122] S4、将所述结构化查询语言子句以及表头的联合表征、所述候选短句中与所述自然语言文本相匹配的表值拼接,生成结构化查询语言语句。

[0123] 又一方面,本公开实施例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一种结构化查询语言语句生成方法,该方法包括:

[0124] S1、将自然语言文本和表格中若干个表头输入语义编码模型中,输出结构化查询语言子句以及表头的联合表征;其中,所述语义编码模型是基于结构化文本及表格样本以及预先确定的生成目标标签进行训练后得到的;

[0125] S2、基于所述自然语言文本和所述查询表格生成候选短句;

[0126] S3、将所述自然语言文本和所述候选短句分别输入匹配模型,输出所述候选短句中与所述自然语言文本相匹配的表值;其中,所述匹配模型为基于带标记的文本和结构化查询语言语句样本以及预先确定的表征配对标签进行训练后得到的;

[0127] S4、将所述结构化查询语言子句以及表头的联合表征、所述候选短句中与所述自然语言文本相匹配的表值拼接,生成结构化查询语言语句。

[0128] 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。

[0129] 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。

[0130] 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。

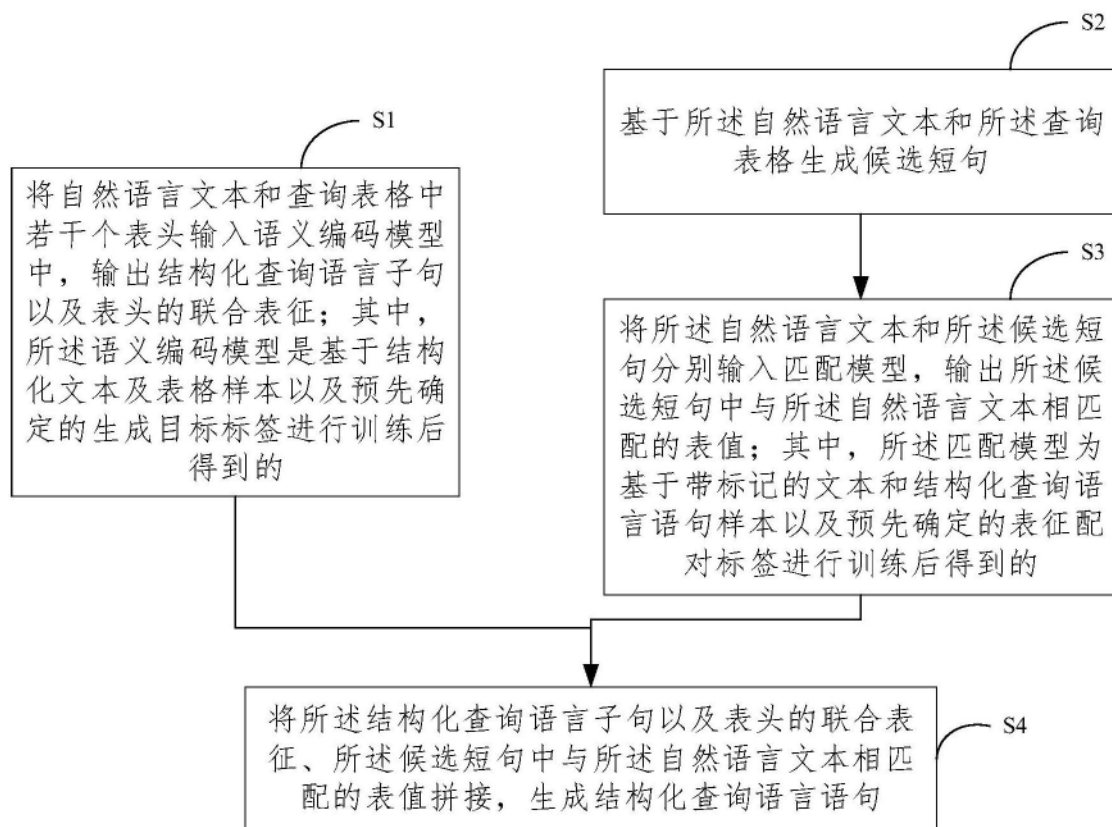


图1

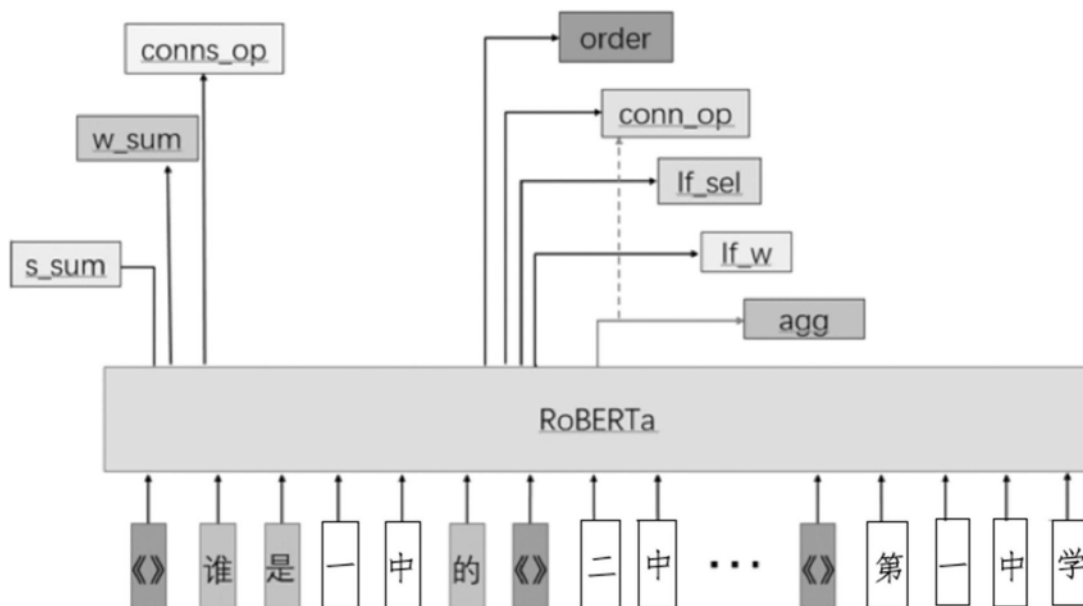


图2

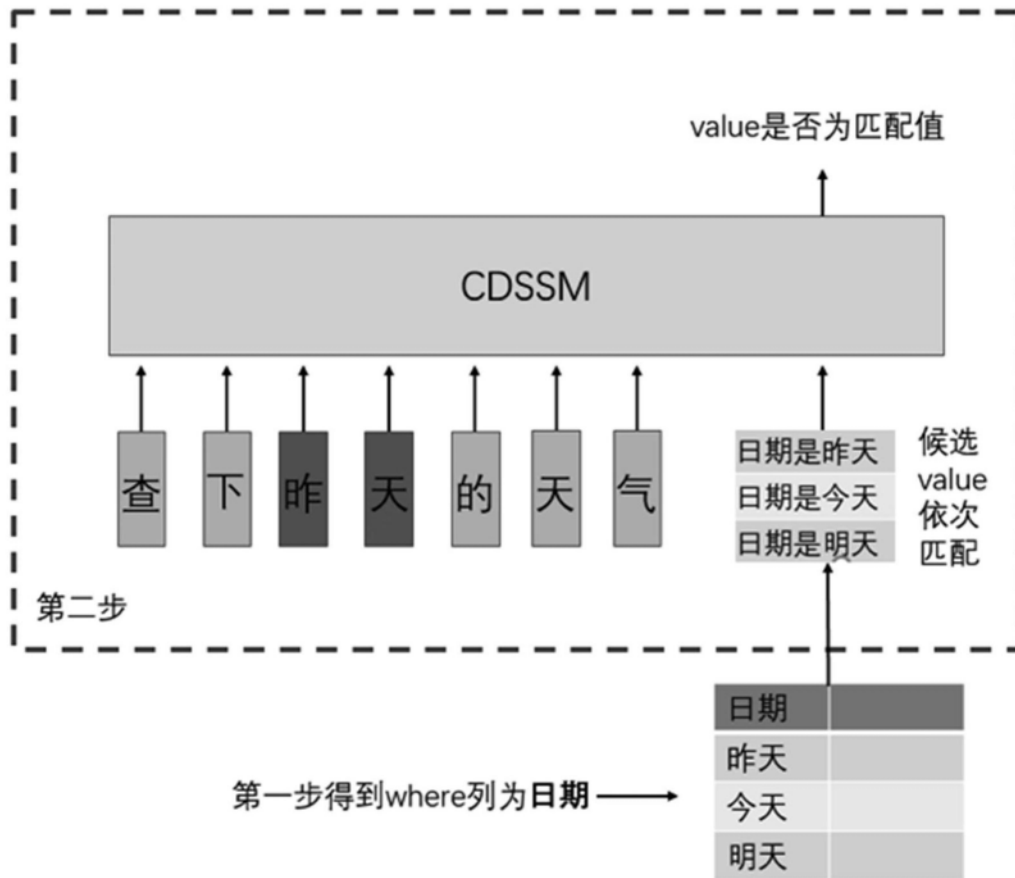


图3

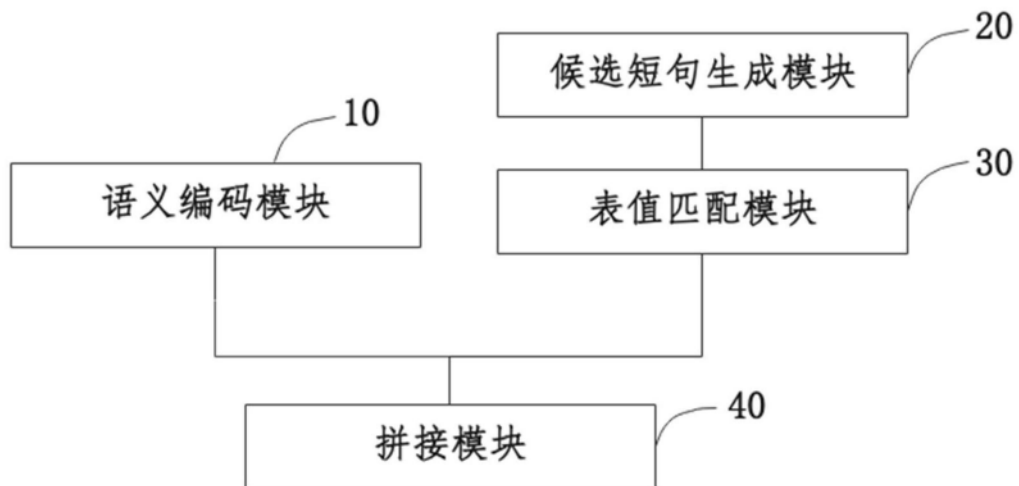


图4

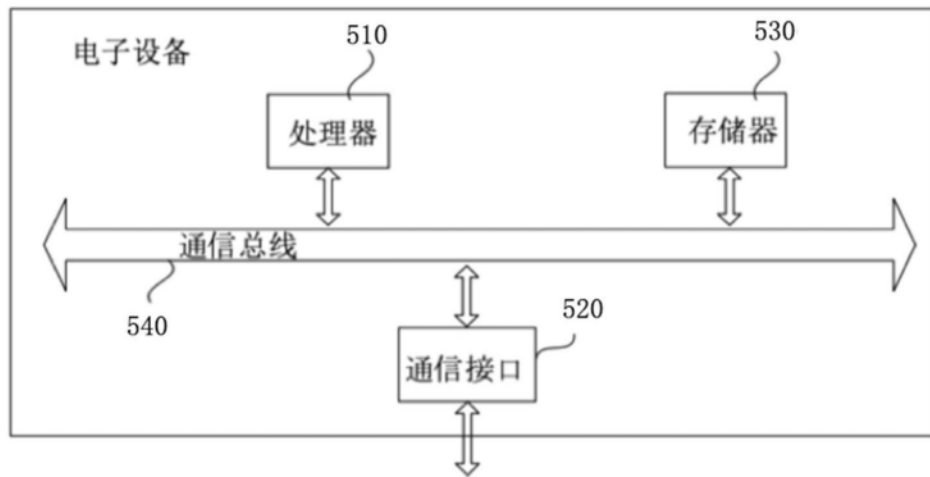


图5