



(21) 申请号 202210158516.3

(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 16/22 (2019.01)

G06K 9/62 (2022.01)

G06F 40/30 (2020.01)

权利要求书2页 说明书12页 附图2页

(54) 发明名称

结构化查询语言语句处理方法和系统、设备
和介质

(57) 摘要

本公开提供一种结构化查询语言语句处理方法和系统、设备和介质,所述结构化查询语言语句处理方法包括:获取表格信息,将表格信息输入至结构化查询语言语句处理模型,输出结构化查询语言语句,其中,将第一训练样本分别输入第一老师模型和第一学生模型中进行蒸馏训练,得到第二学生模型;将第二训练样本分别输入第二老师模型和第二学生模型中进行蒸馏训练,得到训练好的第二学生模型;结构化查询语言语句处理模型是基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到,从而可以实现在保证结构化查询语言语句的准确率和精度的基础上,克服了线上部署时内存不足和推理时间过慢的缺陷。

获取表格信息,将所述表格信息输入至所述结构化查询语言语句处理模型,输出结构化查询语言语句

S1

1. 一种结构化查询语言语句处理方法,其特征在于,包括:

获取表格信息,将所述表格信息输入至所述结构化查询语言语句处理模型,输出结构化查询语言语句;

其中,将第一训练样本分别输入第一老师模型和第一学生模型中进行蒸馏训练,得到第二学生模型,所述第一老师模型为基于所述第一训练样本经过预先训练获得的;

将第二训练样本分别输入第二老师模型和所述第二学生模型中进行蒸馏训练,得到训练好的第二学生模型,所述第二老师模型为基于所述第二训练样本经过预先训练获得的;

所述结构化查询语言语句处理模型是基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到。

2. 根据权利要求1所述的结构化查询语言语句处理方法,其特征在于,所述第一训练样本包括未标记的中文训练语料样本以及预先确定的句子表征标签,所述第一学生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 表示预先确定的句子表征标签, $\hat{y}_i^1$ 表示第一学生模型训练时的输出结果, n_1 表示第一学生模型输出结果的分类数。

3. 根据权利要求1所述的结构化查询语言语句处理方法,其特征在于,所述第二训练样本包括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,所述第二学生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, $\hat{y}_i^2$ 表示第二学生模型训练时的输出结果, n_2 表示第二学生模型输出结果的分类数。

4. 根据权利要求1所述的结构化查询语言语句处理方法,其特征在于,所述第二训练样本的构造方法包括:

获取带标记的文本;

基于所述带标记的文本,构造包含数据集表中的表头、运算和表值的第一范式,将所述第一范式改写成正结构化查询语言语句样本;

将所述带标记的文本与正结构化查询语言语句样本相匹配的标签作为表征配对成功标签;

将所述数据集表中的其他表头和表值以及所述运算构造成第二范式,将所述第二范式改写成负结构化查询语言语句样本;

将所述带标记的文本与负结构化查询语言语句样本不匹配的标签作为表征配对失败标签。

5. 根据权利要求1所述的结构化查询语言语句处理方法,其特征在于,所述获取表格信息,将所述表格信息输入至所述结构化查询语言语句处理模型,输出结构化查询语言语句,

包括：

将所述表格信息输入至所述结构化查询语言语句处理模型，输出若干个子任务对应的结构化查询语言语句；

所述输出若干个子任务对应的结构化查询语言语句之后，包括：

将若干个子任务对应的结构化查询语言语句进行拼接，形成目标结构化查询语言语句。

6. 根据权利要求1所述的结构化查询语言语句处理方法，其特征在于，所述结构化查询语言语句处理模型在训练过程中采用的损失函数为：

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

其中， y_i 表示预先确定的语言标签， $\hat{y}_i$ 表示输出的结构化查询语言语句， n 表示输出的结构化查询语言语句的类别数。

7. 一种结构化查询语言语句处理系统，其特征在于，包括：

结构化查询语言语句处理模块，用于获取表格信息，将所述表格信息输入至所述结构化查询语言语句处理模型，输出结构化查询语言语句；

其中，所述系统还包括：第一蒸馏训练模块，用于将第一训练样本分别输入第一老师模型和第一学生模型中进行蒸馏训练，得到第二学生模型，所述第一老师模型为基于所述第一训练样本经过预先训练获得的；

第二蒸馏训练模块，用于将第二训练样本分别输入第二老师模型和所述第二学生模型中进行蒸馏训练，得到训练好的第二学生模型，所述第二老师模型为基于所述第二训练样本经过预先训练获得的；

所述结构化查询语言语句处理模型是基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到。

8. 根据权利要求7所述的结构化查询语言语句处理系统，其特征在于，所述结构化查询语言语句处理模块还用于：

将所述表格信息输入至所述结构化查询语言语句处理模型，输出若干个子任务对应的结构化查询语言语句；

所述系统还包括：拼接模块，用于在输出若干个子任务对应的结构化查询语言语句之后，将若干个子任务对应的结构化查询语言语句进行拼接，形成目标结构化查询语言语句。

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6任一项所述的结构化查询语言语句处理方法的步骤。

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6任一项所述的结构化查询语言语句处理方法的步骤。

结构化查询语言语句处理方法和系统、设备和介质

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处理方法和系统、电子设备和存储介质。

背景技术

[0002] 表格问答是根据表格信息把人类的自然语言转换成机器能理解的结构化查询语言语句,从数据库查询到客户想要的答案。目前生成结构化查询语言语句的做法可以采用使用规则生成的办法,这种方式首先设置字段及表的描述表项;填充与所述描述表项对应的描述信息,生成数据字典;根据数据字典及结构化查询语言语句生成规则,生成结构化查询语言语句,该方法准确率较低。另外一种模型生成,一般是通过预训练模型生成结构化查询语言语句。基于预训练模型的表格问答在准确率上有一定优势,然而基于预训练的做法是需要使用预训练模型,不仅模型大而且推理速度慢,在计算资源有限的线上部署会面临内存不足和推理时间过慢等问题。

发明内容

[0003] 本公开提供一种结构化查询语言语句处理方法和系统、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处理方法,包括:

[0005] 获取表格信息,将所述表格信息输入至所述结构化查询语言语句处理模型,输出结构化查询语言语句;

[0006] 其中,将第一训练样本分别输入第一老师模型和第一学生模型中进行蒸馏训练,得到第二学生模型,所述第一老师模型为基于所述第一训练样本经过预先训练获得的;

[0007] 将第二训练样本分别输入第二老师模型和所述第二学生模型中进行蒸馏训练,得到训练好的第二学生模型,所述第二老师模型为基于所述第二训练样本经过预先训练获得的;

[0008] 所述结构化查询语言语句处理模型是基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到。

[0009] 根据本公开提供的结构化查询语言语句处理方法,其中,所述第一训练样本包括未标记的中文训练语料样本以及预先确定的句子表征标签,所述第一学生模型在训练过程中采用的损失函数为:

$$[0010] \quad \text{Loss}_1 = - \sum_{i=1}^{n_1} y_i^1 \log \hat{y}_i^1$$

[0011] 其中, y_i^1 表示预先确定的句子表征标签, $\hat{y}_i^1$ 表示第一学生模型训练时的输出结果, n_1 表示第一学生模型输出结果的分类数。

[0012] 根据本公开提供的结构化查询语言语句处理方法,其中,所述第二训练样本包括

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,所述第一学生模型、第二学生模型在训练过程中采用的损失函数为:

$$[0013] \quad \text{Loss}_2 = - \sum_{i=1}^{n_2} y_i^2 \log \hat{y}_i^2$$

[0014] 其中, y_i^2 表示预先确定的表征配对标签, $\hat{y}_i^2$ 表示第二学生模型训练时的输出结果, n_2 表示第二学生模型输出结果的分类数。

[0015] 根据本公开提供的结构化查询语言语句处理方法,其中,所述第二训练样本的构造方法包括:

[0016] 获取带标记的文本;

[0017] 基于所述带标记的文本,构造包含数据集表中的表头、运算和表值的第一范式,将所述第一范式改写成正结构化查询语言语句样本;

[0018] 将所述带标记的文本与正结构化查询语言语句样本相匹配的标签作为表征配对成功标签;

[0019] 将所述数据集表中的其他表头和表值以及所述运算构造成第二范式,将所述第二范式改写成负结构化查询语言语句样本;

[0020] 将所述带标记的文本与负结构化查询语言语句样本不匹配的标签作为表征配对失败标签。

[0021] 根据本公开提供的结构化查询语言语句处理方法,其中,所述获取表格信息,将所述表格信息输入至所述结构化查询语言语句处理模型,输出结构化查询语言语句,包括:

[0022] 将所述表格信息输入至所述结构化查询语言语句处理模型,输出若干个子任务对应的结构化查询语言语句;

[0023] 所述将所述表格信息输入至所述结构化查询语言语句处理模型,输出若干个子任务对应的结构化查询语言语句之后,包括:

[0024] 将若干个子任务对应的结构化查询语言语句进行拼接,形成目标结构化查询语言语句。

[0025] 根据本公开提供的结构化查询语言语句处理方法,其中,所述结构化查询语言语句处理模型在训练过程中采用的损失函数为:

$$[0026] \quad \text{Loss} = - \sum_{i=1}^n y_i \log \hat{y}_i$$

[0027] 其中, y_i 表示预先确定的语言标签, $\hat{y}_i$ 表示输出的结构化查询语言语句, n 表示输出的结构化查询语言语句的类别数。

[0028] 本公开还提供了一种结构化查询语言语句处理系统,包括:

[0029] 结构化查询语言语句处理模块,用于获取表格信息,将所述表格信息输入至所述结构化查询语言语句处理模型,输出结构化查询语言语句;

[0030] 其中,所述系统还包括:第一蒸馏训练模块,用于将第一训练样本分别输入第一老师模型和第一学生模型中进行蒸馏训练,得到第二学生模型,所述第一老师模型为基于所述第一训练样本经过预先训练获得的;

[0031] 第二蒸馏训练模块,用于将第二训练样本分别输入第二老师模型和所述第二学生模型中进行蒸馏训练,得到训练好的第二学生模型,所述第二老师模型为基于所述第二训练样本经过预先训练获得的;

[0032] 所述结构化查询语言语句处理模型是基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到。

[0033] 根据本公开提供的结构化查询语言语句处理系统,其中,所述结构化查询语言语句处理模块还用于:

[0034] 将所述表格信息输入至所述结构化查询语言语句处理模型,输出若干个子任务对应的结构化查询语言语句;

[0035] 所述系统还包括:拼接模块,用于在输出若干个子任务对应的结构化查询语言语句之后,将若干个子任务对应的结构化查询语言语句进行拼接,形成目标结构化查询语言语句。

[0036] 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处理方法的步骤。

[0037] 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处理方法的步骤。

[0038] 本公开提供的结构化查询语言语句处理方法,通过将第一训练样本分别输入第一老师模型和第一学生模型中进行蒸馏训练,得到初步训练后的第二学生模型,再将第二训练样本分别输入第二老师模型和第二学生模型中进行蒸馏训练,得到训练好的第二学生模型,然后再基于训练好的第二学生模型训练得到结构化查询语言语句处理模型,由于第二学生模型为经过蒸馏训练得到的模型,与传统的基于表格问答的语句处理模型相比,极大地减少表格问答的模型计算量和所需的内存容量;另外通过蒸馏训练后的第二学生模型训练生成结构化查询语言语句处理模型,使训练后的模型尽可能保留表格问答数据集的一些语义信息,从而在保证结构化查询语言语句的准确率和精度的基础上,克服了线上部署时内存不足和推理时间过慢的缺陷。

附图说明

[0039] 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。

[0040] 图1是本公开实施例提供的结构化查询语言语句处理方法的流程示意图之一;

[0041] 图2是本公开实施例提供的结构化查询语言语句处理方法的流程示意图之二;

[0042] 图3是本公开实施例提供的结构化查询语言语句处理方法中的语义匹配训练数据方式的示意图;

[0043] 图4是本公开提供的结构化查询语言语句处理系统的示意图;

[0044] 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

[0045] 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。

[0046] 本公开实施例公开了一种结构化查询语言语句处理方法，参见图1，所述结构化查询语言语句处理方法包括：

[0047] S1：获取表格信息，将所述表格信息输入至所述结构化查询语言语句处理模型，输出结构化查询语言语句。

[0048] 其中，结构化查询语言语句指的是SQL语句。结构化查询语言 (Structured Query Language, SQL) 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。

[0049] 本公开实施例中，将第一训练样本分别输入第一老师模型和第一学生模型中进行蒸馏训练，得到第二学生模型；其中，所述第一老师模型为基于所述第一训练样本经过预先训练获得的；

[0050] 模型蒸馏训练是一种将模型小型化的技术，它通常包含一个teacher模型（老师模型）和student模型（学生模型）。蒸馏时，teacher模型的预测信息作为标签，来对student模型进行监督学习，其不同于传统离线蒸馏的方式，在对student模型训练之前，先输入部分的未标记的数据，利用teacher模型输出标签作为监督信息再输入到student模型中，通过一些样本来迭代上述过程来完成蒸馏过程，这样就可以使用更少标注量的数据集，达到提升模型精度的目的。

[0051] 在蒸馏之前，需要获取第一训练样本，也就是为了让第一老师模型（可以采用12层BERT）模型和第一学生（可以采用4层Transformer）模型获得丰富的语义表征，可先通过输入海量的未标记的数据（因为不是和业务相关的，因此也叫通用数据）。这里利用了第一老师模型输出作为监督信息来控制第一学生模型的学习。本方法主要通过几个知名中文数据集和网上爬取的大量数据，组合成通用训练语料，完成对12层BERT和4层Transformer做了通用性的训练学习。

[0052] 然后，将第二训练样本分别输入第二老师模型和所述第二学生模型中进行蒸馏训练，得到训练好的第二学生模型；其中，所述第二老师模型为基于所述第二训练样本经过预先训练获得的；

[0053] 对于第二学生模型和第二老师模型的训练过程，同第一老师模型和第一学生模型的训练过程一样，采用的是模型蒸馏训练，都属于蒸馏的预学习过程。其中，第二老师模型是基于所述第二训练样本经过预先训练获得的。

[0054] 最后，基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到结构化查询语言语句处理模型。

[0055] 需要解释的是，本公开是通过表格样本以及预先确定的语言标签对训练好的第二学生模型进行有监督训练得到最终的结构化查询语言语句处理模型。

[0056] 本公开实施例提供的结构化查询语言语句处理方法，通过将第一训练样本分别输入第一老师模型和第一学生模型中进行蒸馏训练，得到初步训练后的第二学生模型，再将

第二训练样本分别输入第二老师模型和第二学生模型中进行蒸馏训练,得到训练好的第二学生模型,然后再基于训练好的第二学生模型训练得到结构化查询语言语句处理模型,由于第二学生模型为经过蒸馏训练得到的模型,与传统的基于表格问答的语句处理模型相比,极大地减少表格问答的模型计算量和所需的内存容量;另外通过蒸馏训练后的第二学生模型训练生成结构化查询语言语句处理模型,使训练后的模型尽可能保留表格问答数据集的一些语义信息,从而在保证结构化查询语言语句的准确率和精度的基础上,克服了线上部署时内存不足和推理时间过慢的缺陷。

[0057] 根据本公开实施例提供的结构化查询语言语句处理方法,其中,所述第一训练样本包括未标记的中文训练语料样本以及预先确定的句子表征标签,所述第一学生模型在训练过程中采用的损失函数为:

$$[0058] \quad \text{Loss}_1 = - \sum_{i=1}^{n_1} y_i^1 \log \hat{y}_i^1$$

[0059] 其中, y_i^1 表示预先确定的句子表征标签, $\hat{y}_i^1$ 表示第一学生模型训练时的输出结果, n_1 表示第一学生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,第一类是句子表征配对,另一类是句子表征不配对,此时, n_1 为2。

[0060] 第一训练样本采用的是大规模的中文训练语料,以实现预训练,让模型尽可能接触到各种各样的句子获取到丰富的中文知识,通过执行多次迭代操作对所述第一学生模型进行蒸馏训练,可以根据所述第一学生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,迭代是为了监督让第一老师模型和第一学生模型的输出一致。

[0061] 根据本公开实施例提供的结构化查询语言语句处理方法,其中,所述第二训练样本包括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,所述第二学生模型在训练过程中采用的损失函数为:

$$[0062] \quad \text{Loss}_2 = - \sum_{i=1}^{n_2} y_i^2 \log \hat{y}_i^2$$

[0063] 其中, y_i^2 表示预先确定的表征配对标签, $\hat{y}_i^2$ 表示第二学生模型训练时的输出结果, n_2 表示第二学生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,第一类是句子表征配对成功,另一类是句子表征配对不成功,此时, n_2 为2。

[0064] 需要说明的是,SQL语句本身是没法输入到预训练模型里学习,需要将SQL语句生成文本语句,就有了(文本语句,SQL语句)的表征,这是正样本,SQL语句中的成分替换下改的句子就是负样本,就可以利用第二学生模型学习了,是为了让模型能够学习到特定领域的内容,例如,SQL语句的知识和文本语句的知识。通过执行多次迭代操作对所述第二学生模型进行蒸馏训练,可以根据所述第二学生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处理方法,其中,所述第二训练样本的构造方法包括:

[0066] 获取带标记的文本；

[0067] 基于所述带标记的文本，构造包含数据集表中的表头、运算和表值的第一范式，将所述第一范式改写成正结构化查询语言语句样本；

[0068] 将所述带标记的文本与正结构化查询语言语句样本相匹配的标签作为表征配对成功标签；

[0069] 将所述数据集表中的其他表头和表值以及所述运算构造成第二范式，将所述第二范式改写成负结构化查询语言语句样本；

[0070] 将所述带标记的文本与负结构化查询语言语句样本不匹配的标签作为表征配对失败标签。

[0071] 数据集表是数据集以表格的形式展示的列表，匹配需要有正样本，和负样本，正样本就是SQL语句本身，也就是说，把SQL语句改写成自然语言和原自然语言来配对训练。负样本就用表中的其他值来替换，此时是不匹配的，将所述带标记的文本与负结构化查询语言语句样本不匹配的标签作为表征配对失败标签。这样构造的优势是不但能够让模型关注到表头和表值的信息，同时使用运算符将其写成一个短句，也有利于获取较长的语义信息。

[0072] 根据本公开实施例提供的结构化查询语言语句处理方法，其中，所述获取表格信息，将所述表格信息输入至所述结构化查询语言语句处理模型，输出结构化查询语言语句，包括：

[0073] 将所述表格信息输入至所述结构化查询语言语句处理模型，输出若干个子任务对应的结构化查询语言语句。

[0074] 多个子任务均是结构化查询语言语句处理模型生成的，也就是SQL语句的各个子句是分开生成的，结构化查询语言语句处理模型的输入都是表格信息，最终输出若干个子任务对应的结构化查询语言语句；

[0075] 所述将所述表格信息输入至所述结构化查询语言语句处理模型，输出若干个子任务对应的结构化查询语言语句之后，包括：

[0076] 将若干个子任务对应的结构化查询语言语句进行拼接，形成目标结构化查询语言语句。

[0077] 也就是将多个若干个子任务对应的结构化查询语言语句进行连接起来，就形成了SQL语句。

[0078] 根据本公开实施例提供的结构化查询语言语句处理方法，其中，所述第一老师模型、第二老师模型分别包括BERT (Bidirectional Encoder Representations from Transformers, 词向量) 层，所述第一学生模型、第二学生模型分别包括transformer (转换器) 层；所述BERT层的层数大于transformer层的层数。

[0079] BERT的网络架构使用的是多层Transformer结构，其最大的特点是抛弃了传统的RNN和CNN，通过Attention机制将任意位置的两个单词的距离转换成1，有效的解决了NLP中棘手的长期依赖问题。Transformer是一个encoder-decoder的结构，由若干个编码器和解码器堆叠形成。编码器由Multi-Head Attention和一个全连接组成，用于将输入语料转化成特征向量。解码器的输入为编码器的输出以及已经预测的结果，由Masked Multi-Head Attention, Multi-Head Attention以及一个全连接组成，用于输出最后结果的条件概率。

[0080] 根据本公开实施例提供的结构化查询语言语句处理方法，其中，对单个样本，这里

以SQL语句的运算符为例,所述结构化查询语言语句处理模型在训练过程中采用的损失函数为:

$$[0081] \quad \text{Loss} = - \sum_{i=1}^n y_i \log \hat{y}_i$$

[0082] 其中, y_i 表示预先确定的语言标签, $\hat{y}_i$ 表示输出的结构化查询语言语句, n 表示输出的结构化查询语言语句的类别数,如果有7个类别,那么 n 可以是7。在结构化查询语言语句处理模型在训练过程中,有多个任务,每个任务对应的 n 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。

[0083] 训练采用的样本包括表格样本以及预先确定的语言标签,通过执行多次迭代操作对所述结构化查询语言语句处理模型进行训练,可以根据所述结构化查询语言语句处理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,迭代是通过深度学习,提升结构化查询语言语句处理模型的精准度。

[0084] 为了进一步地理解本实施例的方法,如图2所示,在一个具体实例中,本实施例的结构化查询语言语句处理方法包括:

[0085] S21、通用性蒸馏:

[0086] 蒸馏之前,为了让第一老师模型(12层BERT模型)和第一学生模型(4层Transformer模型)获得丰富的语义表征,先通过输入海量的未标记的数据(因为不是和业务相关的,因此也叫通用数据)。这里利用了第一老师模型的输出作为监督信息来控制第一学生模型的学习。这里本公开实施例主要通过JDDC等几个知名中文数据集和网上爬取的大量数据,组合成通用训练语料,完成对12层BERT模型和4层Transformer模型做了通用性的训练学习。

[0087] S22、匹配任务融合SQL语句语义的蒸馏:

[0088] 由于原始的蒸馏方法使用一个匹配任务做蒸馏,这是因为其最终做的任务就是匹配。对于表格问答和匹配任务并无直接联系,因而如何将一个表格问答数据构造成一个匹配任务进行蒸馏学习就成了该任务学习的关键。为了让模型能预先学习到query(查询)和SQL语句信息,蒸馏学习必须要同时考虑到query和SQL语句的信息。同时SQL语句中既包含了表格中的表头信息,又包含表值的信息,只使用query或者只是用值的信息都不利于模型全面学习表格问答中的语义信息。

[0089] 为此,本公开实施例设计了一种基于SQL语句语义的训练方式。首先将SQL语句的where子句构造成一个范式[表头,运算,表值],然后将范式改写写成一个短句。匹配训练需要正负样本,这里使用表格中其他表头和表值来替换范式中的表头和表值构造负样本。举个例子,如图3中的where子句范式[时间,=,10月1日],改写成正确的短句就是时间是10月1日。把时间或者10月1日使用表中其他信息替换就是负样本。这样构造的优势是不但能够让模型关注到表头和表值的信息,同时使用运算符将其写成一个短句,也有利于获取较长的语义信息。

[0090] 得到匹配语料后,将成对的数据分别输入到第二老师模型(BERT模型)和第二学生模型(4层Transformer模型)中,这里要求BERT模型和Transformer模型每一层的输入都要

一致,因此模型每层在输出时都做了约束,这里使用了CrossEntropy (交叉熵) 损失确保第二老师模型和第二学生模型的输出一致。训练完成,就得到融合SQL语句语义的蒸馏模型了。

[0091] S22、表格问答模型训练:

[0092] 得到融合SQL语句语义的蒸馏模型后,使用蒸馏后模型作为预训练模型训练表格问答的多个子任务,这里可以将模型被分为如下的7个子任务,当模型输出了这些子任务,将子任务连接起来就是最后的SQL语句。例如:

[0093] s_sum:select了多少个;

[0094] w_sum:where有多少列;

[0095] order:是否存在order by;

[0096] agg:当前列的agg (max,min,count) 操作是什么;

[0097] conn_op当前列和value的运算符(>,<=) 是什么;

[0098] conns_op:当前sql语句的连接符(And,or) 是什么。

[0099] w_value:where中值是多少

[0100] 详解:一条SQL语句基本结构由select agg sel_column from table where where_column1 op1 value1 conn where_colnum2 op2 value2等多个部分组成,所以模型直接将自然语言自动转为如此复杂的SQL语句并正确执行是非常困难。这里本专利把一个SQL语句生成问题分解成多个子任务,即上面以展示出的模块,每个模块都代表一个子任务。模型得到每个子任务的结果,最后再将各个部分拼接在一起就是输出的SQL语句。

[0101] 为了验证蒸馏后模型的效果,本公开实施例还进行了定量分析。通过表1中的性能分析可以得出,相比较原始的BERT模型,本专利蒸馏后的模型比原来快了接近一倍,每百条数据的处理速度只有6.46s;而大小只有45M,大约只有原来的1/9。这样一来不仅极大提升了回答的效率,并且对计算机内存空间的要求上也降低了,可以节省更多的计算资源。

[0102] 表1

模型	模型大小	运行时间(每百条)
BERT	45M	6.46s
蒸馏后的4层Transformer	393M	11.59s

[0104] 表2中展示了准确率分析,和通用的4层Transformer模型相比,本公开实施例提出的融合SQL语句语义的蒸馏训练方式使模型的准确率提升了5%。

[0105] 表2

模型	Agg 和 sel 列准确率	条件子句 准确率	条件列 准确率	条件值 准确率	连接符 准确率	总准 准确率
通用 4 层 Transformer 模型	0.7	0.89	0.94	0.91	0.97	0.65
融合 sql 语义蒸馏 4 层 Transformer 模 型	0.74	0.91	0.99	0.92	0.98	0.70

[0107] 本公开实施例公开了一种结构化查询语言语句处理系统,参见图4,所述结构化查

询语言语句处理系统包括：

[0108] 结构化查询语言语句处理模块401,用于获取表格信息,将所述表格信息输入至所述结构化查询语言语句处理模型,输出结构化查询语言语句;

[0109] 其中,所述系统还包括:第一蒸馏训练模块402,用于将第一训练样本分别输入第一老师模型和第一学生模型中进行蒸馏训练,得到第二学生模型,所述第一老师模型为基于所述第一训练样本经过预先训练获得的;

[0110] 第二蒸馏训练模块403,用于将第二训练样本分别输入第二老师模型和所述第二学生模型中进行蒸馏训练,得到训练好的第二学生模型,所述第二老师模型为基于所述第二训练样本经过预先训练获得的;

[0111] 其中,所述结构化查询语言语句处理模型是基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到。

[0112] 根据本公开实施例提供的结构化查询语言语句处理系统,其中,所述第一训练样本包括未标记的中文训练语料样本以及预先确定的句子表征标签,所述第一学生模型在训练过程中采用的损失函数为:

$$[0113] \quad \text{Loss}_1 = - \sum_{i=1}^{n_1} y_i^1 \log \hat{y}_i^1$$

[0114] 其中, y_i^1 表示预先确定的句子表征标签, $\hat{y}_i^1$ 表示第一学生模型训练时的输出结果, n_1 表示第一学生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,第一类是句子表征配对,另一类是句子表征不配对,此时, n_1 为2。

[0115] 第一训练样本采用的是大规模的中文训练语料,以实现预训练,让模型尽可能接触到各种各样的句子获取到丰富的中文知识,通过执行多次迭代操作对所述第一学生模型进行蒸馏训练,可以根据所述第一学生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,迭代是为了监督让teacher和student的输出一致。

[0116] 根据本公开实施例提供的结构化查询语言语句处理系统,其中,所述第二训练样本包括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,所述第二学生模型在训练过程中采用的损失函数为:

$$[0117] \quad \text{Loss}_2 = - \sum_{i=1}^{n_2} y_i^2 \log \hat{y}_i^2$$

[0118] 其中, y_i^2 表示预先确定的表征配对标签, $\hat{y}_i^2$ 表示第二学生模型训练时的输出结果, n_2 表示第二学生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,第一类是句子表征配对成功,另一类是句子表征配对不成功,此时, n_2 为2。

[0119] 根据本公开实施例提供的结构化查询语言语句处理系统,还包括第二训练样本构造模块,用于:

[0120] 获取带标记的文本;

[0121] 基于所述带标记的文本,构造包含数据集表中的表头、运算和表值的第一范式,将所述第一范式改写成正结构化查询语言语句样本;

[0122] 将所述带标记的文本与正结构化查询语言语句样本相匹配的标签作为表征配对成功标签；

[0123] 将所述数据集表中的其他表头和表值以及所述运算构造成第二范式，将所述第二范式改写成负结构化查询语言语句样本；

[0124] 将所述带标记的文本与负结构化查询语言语句样本不匹配的标签作为表征配对失败标签。

[0125] 根据本公开实施例提供的结构化查询语言语句处理系统，其中，所述结构化查询语言语句处理模块401，具体用于：将所述表格信息输入至所述结构化查询语言语句处理模型，输出若干个子任务对应的结构化查询语言语句。

[0126] 多个子任务均是结构化查询语言语句处理模型生成的，也就是SQL的各个子句是分开生成的，结构化查询语言语句处理模型的输入都是表格信息，最终输出若干个子任务对应的结构化查询语言语句。

[0127] 所述系统还包括：拼接模块，用于在输出若干个子任务对应的结构化查询语言语句之后，将若干个子任务对应的结构化查询语言语句进行拼接，形成目标结构化查询语言语句。

[0128] 根据本公开实施例提供的结构化查询语言语句处理系统，其中，对单个样本，这里以SQL的运算符为例，所述结构化查询语言语句处理模型在训练过程中采用的损失函数为：

$$[0129] \quad \text{Loss} = - \sum_{i=1}^n y_i \log \hat{y}_i$$

[0130] 其中， y_i 表示预先确定的语言标签， $\hat{y}_i$ 表示输出的结构化查询语言语句， n 表示输出的结构化查询语言语句的类别数，如果有7个类别，那么 n 可以是7。在结构化查询语言语句处理模型在训练过程中，有多个任务，每个任务对应的 n 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。

[0131] 训练采用的样本包括表格样本以及预先确定的语言标签，通过执行多次迭代操作对所述结构化查询语言语句处理模型进行训练，可以根据所述结构化查询语言语句处理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，迭代是通过深度学习，提升结构化查询语言语句处理模型的精准度。

[0132] 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处理方法，该方法包括：

[0133] 获取表格信息，将所述表格信息输入至所述结构化查询语言语句处理模型，输出结构化查询语言语句；

[0134] 其中，将第一训练样本分别输入第一老师模型和所述第一学生模型中进行蒸馏训练，得到第二学生模型，所述第一老师模型为基于所述第一训练样本经过预先训练获得的；

[0135] 将第二训练样本分别输入第二老师模型和所述第二学生模型中进行蒸馏训练，得

到训练好的第二学生模型,所述第二老师模型为基于所述第二训练样本经过预先训练获得的;

[0136] 所述结构化查询语言语句处理模型是基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到。

[0137] 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。

[0138] 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处理方法,该方法包括:

[0139] 获取表格信息,将所述表格信息输入至所述结构化查询语言语句处理模型,输出结构化查询语言语句;

[0140] 其中,将第一训练样本分别输入第一老师模型和所述第一学生模型中进行蒸馏训练,得到第二学生模型,所述第一老师模型为基于所述第一训练样本经过预先训练获得的;

[0141] 将第二训练样本分别输入第二老师模型和所述第二学生模型中进行蒸馏训练,得到训练好的第二学生模型,所述第二老师模型为基于所述第二训练样本经过预先训练获得的;

[0142] 所述结构化查询语言语句处理模型是基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到。

[0143] 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处理方法,该方法包括:

[0144] 获取表格信息,将所述表格信息输入至所述结构化查询语言语句处理模型,输出结构化查询语言语句;

[0145] 其中,将第一训练样本分别输入第一老师模型和所述第一学生模型中进行蒸馏训练,得到第二学生模型,所述第一老师模型为基于所述第一训练样本经过预先训练获得的;

[0146] 将第二训练样本分别输入第二老师模型和所述第二学生模型中进行蒸馏训练,得到训练好的第二学生模型,所述第二老师模型为基于所述第二训练样本经过预先训练获得的;

[0147] 所述结构化查询语言语句处理模型是基于训练好的第二学生模型、表格样本以及预先确定的语言标签进行训练后得到。

[0148] 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。

[0149] 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。

[0150] 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。

获取表格信息，将所述表格信息输入至所述结构化查询语言语句处理模型，输出结构化查询语言语句

S1

图1

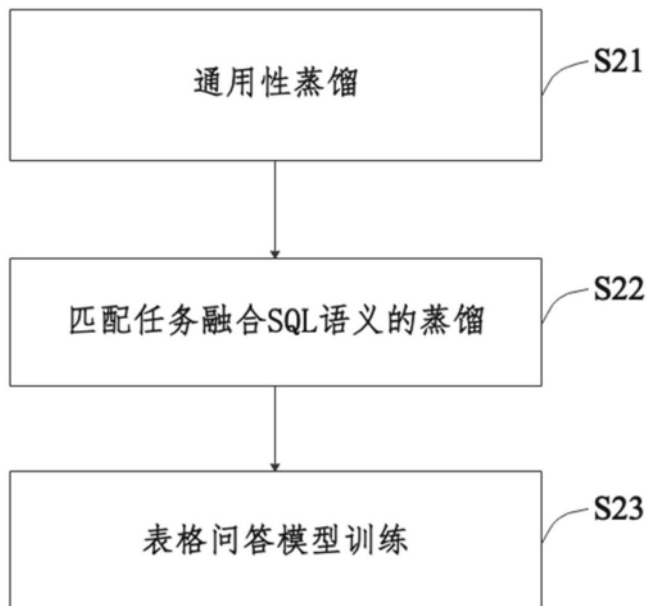


图2

Question:帮我查询一下10月1日的天气，谢谢

Sql: select 天气 from 天气数据 where 时间=10月1日

时间是10月5日 ×

时间不是是10月1日 ×

地点是10月1日 ×

时间是10月1日 ✓

图3

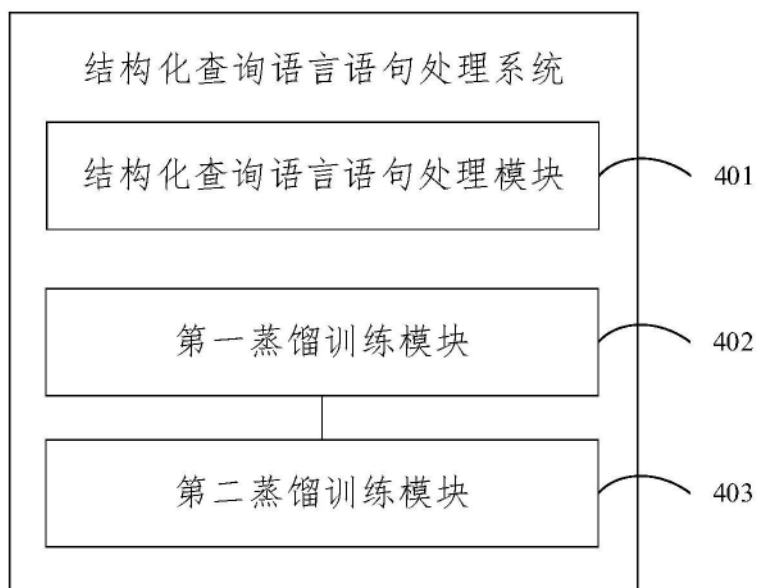


图4

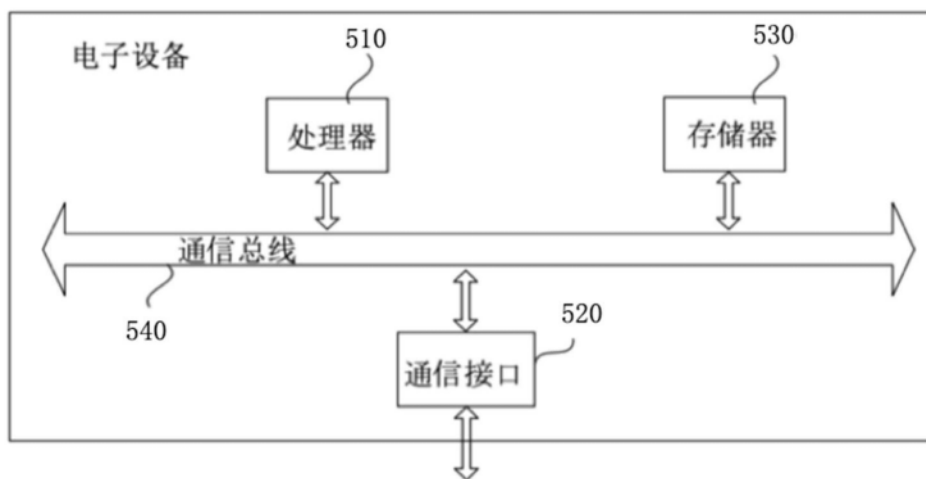


图5