



(12) 发明专利申请

(10) 申请公布号 CN 114238635 A

(43) 申请公布日 2022. 03. 25

(21) 申请号 202111523218.1

G06F 40/30 (2020.01)

(22) 申请日 2021.12.13

(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 刘瑞雪 袁韶祖 祝天刚 陈蒙

(74) 专利代理机构 北京路浩知识产权代理有限
公司 11002

代理人 吴刚

(51) Int. Cl.

G06F 16/35 (2019.01)

G06F 16/335 (2019.01)

G06F 40/169 (2020.01)

G06F 40/205 (2020.01)

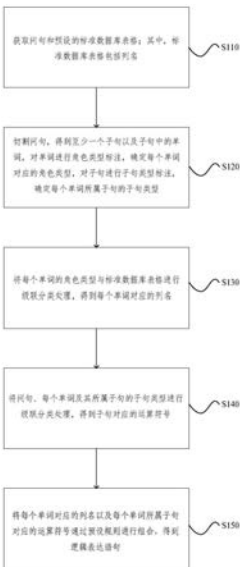
权利要求书2页 说明书11页 附图4页

(54) 发明名称

一种序列标注的方法、装置、电子设备及存
储介质

(57) 摘要

本公开提供一种序列标注的方法、装置、电
子设备及存储介质,该方法包括:获取问句和预
设的标准数据库表格;切割问句,得到至少一个
子句以及子句中的单词,对单词进行角色类型标
注,确定每个单词对应的角色类型,对子句进行
子句类型标注,确定每个单词所属子句的子句类
型;将每个单词的角色类型与标准数据库表格进
行级联分类处理,得到每个单词对应的列名;将
问句、每个单词及其所属子句的子句类型进行级
联分类处理,得到子句对应的运算符;将每个
单词对应的列名以及每个单词所属子句对应的
运算符通过预设规则进行组合,得到逻辑表达
语句。该方法较基于深度学习模型的方法设置更
加简洁,并且可以有效捕捉问句之间的本地信
息。



1. 一种序列标注的方法,其特征在于,包括:

获取问句和预设的标准数据库表格;其中,所述标准数据库表格包括列名;

切割所述问句,得到至少一个子句以及子句中的单词,对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型;

将每个所述单词的角色类型与标准数据库表格进行级联分类处理,得到每个所述单词对应的列名;

将所述问句、每个单词及其所属子句的子句类型进行级联分类处理,得到所述子句对应的运算符号;

将每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,得到逻辑表达语句。

2. 根据权利要求1所述的序列标注的方法,其特征在于,所述方法还包括:基于所述逻辑表达语句在所述标准数据库表格中进行查询,获取查询结果。

3. 根据权利要求1所述的序列标注的方法,其特征在于,所述对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,包括:

对所述单词和所述标准数据库表格中的列名进行拼接,得到整体序列;

将所述整体序列输入至预训练的语言模型中,生成所述整体序列对应的隐向量表示;

对所述问句和每个单词进行注意力机制运算,得到所述问句的注意力表示;

将所述隐向量表示和所述问句的注意力表示进行合并得到合并结果,将所述合并结果输入至角色类型标注模型中,得到每个所述单词对应的角色类型,其中,所述角色类型包括初始列名或初始列值。

4. 根据权利要求3所述的序列标注的方法,其特征在于,所述对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型,包括:

将所述隐向量表示和所述问句的注意力表示进行合并得到合并结果,将所述合并结果输入至子句类型标注模型中,得到每个所述单词所属子句的子句类型。

5. 根据权利要求1所述的序列标注的方法,其特征在于,所述将每个所述单词的角色类型与标准数据库表格进行级联分类处理,得到每个所述单词对应的列名,包括:

将所述标准数据库表格中列名的隐向量表示和每个单词对应的角色类型进行注意力机制的交互匹配;

根据匹配结果获取每个所述单词对应的列名。

6. 根据权利要求1所述的序列标注的方法,其特征在于,所述将所述问句、每个单词及其所属子句的子句类型进行级联分类处理,得到所述子句对应的运算符号,包括:

基于每个单词所属子句的子句类型,将所述每个子句中的单词和所述问句进行注意力机制的交互匹配;

根据匹配结果获取所述子句对应的运算符号。

7. 根据权利要求6所述的序列标注的方法,其特征在于,所述根据匹配结果获取所述子句对应的运算符号,包括:

所述子句类型包括条件类型和问题类型;

在所述子句类型为条件类型的情况下,所述子句对应的运算符号为聚合运算符;

在所述子句类型为问题类型的情况下,所述子句对应的运算符号为函数运算符。

8.一种序列标注的装置,其特征在于,包括:

获取模块,用于获取问句和预设的标准数据库表格;其中,所述标准数据库表格包括列名;

切割模块,用于切割所述问句,得到至少一个子句以及子句中的单词,对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型;

第一处理模块,用于将每个所述单词的角色类型与标准数据库表格进行级联分类处理,得到每个所述单词对应的列名;

第二处理模块,用于将所述问句、每个单词及其所属子句的子句类型进行级联分类处理,得到所述子句对应的运算符号;

组合模块,用于将每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,得到逻辑表达语句。

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7任一项所述的序列标注的方法步骤。

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7任一项所述的序列标注的方法步骤。

一种序列标注的方法、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,尤其涉及一种序列标注的方法、装置、电子设备及存储介质。

背景技术

[0002] 目前自然语言转化为逻辑表达语句的技术主要有两种方法,基于规则的方法和基于深度学习模型的方法。但两种方法均存在一些缺陷,基于规则的方法,需要设置大量的词典和同义词词典,人力成本较高,并且该方法只能涵盖字典中的词语,灵活性和鲁棒性较差;基于深度学习模型的方法,构建专用模块很复杂,无法捕获相互依赖关系,也无法捕捉语句之间的内部信息。

发明内容

[0003] 本公开提供一种序列标注的方法、装置、电子设备及存储介质,用以解决现有技术中自然语言转化为逻辑表达语句的问题,实现了设置简洁,并且可以有效捕捉问句之间的本地信息。

[0004] 第一方面,本公开提供一种序列标注的方法,包括:

[0005] 获取问句和预设的标准数据库表格;其中,所述标准数据库表格包括列名;

[0006] 切割所述问句,得到至少一个子句以及子句中的单词,对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型;

[0007] 将每个所述单词的角色类型与标准数据库表格进行级联分类处理,得到每个所述单词对应的列名;

[0008] 将所述问句、每个单词及其所属子句的子句类型进行级联分类处理,得到所述子句对应的运算符号;

[0009] 将每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,得到逻辑表达语句。

[0010] 根据本公开提供的序列标注的方法,所述方法还包括:基于所述逻辑表达语句在所述标准数据库表格中进行查询,获取查询结果。

[0011] 根据本公开提供的序列标注的方法,所述对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,包括:

[0012] 对所述单词和所述标准数据库表格中的列名进行拼接,得到整体序列;

[0013] 将所述整体序列输入至预训练的语言模型中,生成所述整体序列对应的隐向量表示;

[0014] 对所述问句和每个单词进行注意力机制运算,得到所述问句的注意力表示;

[0015] 将所述隐向量表示和所述问句的注意力表示进行合并得到合并结果,将所述合并结果输入至角色类型标注模型中,得到每个所述单词对应的角色类型,其中,所述角色类型

包括初始列名或初始列值。

[0016] 根据本公开提供的序列标注的方法,所述对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型,包括:

[0017] 将所述隐向量表示和所述问句的注意力表示进行合并得到合并结果,将所述合并结果输入至子句类型标注模型中,得到每个所述单词所属子句的子句类型。

[0018] 根据本公开提供的序列标注的方法,所述将每个所述单词的角色类型与标准数据库表格进行级联分类处理,得到每个所述单词对应的列名,包括:

[0019] 将所述标准数据库表格中列名的隐向量表示和每个单词对应的角色类型进行注意力机制的交互匹配;

[0020] 根据匹配结果获取每个所述单词对应的列名。

[0021] 根据本公开提供的序列标注的方法,所述将所述问句、每个单词及其所属子句的子句类型进行级联分类处理,得到所述子句对应的运算符号,包括:

[0022] 基于每个单词所属子句的子句类型,将所述每个子句中的单词和所述问句进行注意力机制的交互匹配;

[0023] 根据匹配结果获取所述子句对应的运算符号。

[0024] 根据本公开提供的序列标注的方法,所述根据匹配结果获取所述子句对应的运算符号,包括:

[0025] 所述子句类型包括条件类型和问题类型;

[0026] 在所述子句类型为条件类型的情况下,所述子句对应的运算符号为聚合运算符;

[0027] 在所述子句类型为问题类型的情况下,所述子句对应的运算符号为函数运算符;

[0028] 第二方面,本公开提供一种序列标注的装置,包括:

[0029] 获取模块,用于获取问句和预设的标准数据库表格;其中,所述标准数据库表格包括列名;

[0030] 切割模块,用于切割所述问句,得到至少一个子句以及子句中的单词,对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型;

[0031] 第一处理模块,用于将每个所述单词的角色类型与标准数据库表格进行级联分类处理,得到每个所述单词对应的列名;

[0032] 第二处理模块,用于将所述问句、每个单词及其所属子句的子句类型进行级联分类处理,得到所述子句对应的运算符号;

[0033] 组合模块,用于将每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,得到逻辑表达语句。

[0034] 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序列标注的方法步骤。

[0035] 第四方面,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如上述任一项所述的序列标注的方法步骤。

[0036] 本公开提供了一种序列标注的方法、装置、电子设备及存储介质,对获取的问句进行切割,得到至少一个子句以及子句中的单词,分别对单词进行角色类型标注和对子句进

行子句类型标注,得到每个单词对应的角色类型,以及每个单词所属子句的子句类型,对所获取的问句进行基于两种类型的分类判断,能够准确的与标准数据库中的数据进行匹配;之后将得到的每个单词的角色类型与标准数据库表格进行级联分类处理,得到每个单词对应的列名,将问句、每个单词及其所属子句的子句类型进行级联分类处理,得到子句对应的运算符,分别得到与标准数据库中对应的列名和与子句对应的运算符,能够对所获取的整个语句中的所有的信息进行完整的匹配;根据所得到的每个单词对应的列名以及每个单词所属子句对应的运算符通过预设规则进行组合,最终得到逻辑表达语句。本公开提出使用序列标注的方法解决自然语言转化为逻辑表达语句的问题,该方法较基于深度学习模型的方法设置更加简洁,并且可以有效捕捉问句之间的本地信息。

附图说明

[0037] 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。

[0038] 图1是本公开提供的序列标注方法的流程示意图;

[0039] 图2是本公开提供的确定每个单词对应的角色类型的流程示意图;

[0040] 图3是本公开提供的基于序列标注和级联方法的自然语言转化为逻辑表达语句技术框架;

[0041] 图4是本公开提供的序列标注的装置的结构示意图;

[0042] 图5是本公开提供的一种电子设备的实体结构示意图。

具体实施方式

[0043] 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。

[0044] 序列标注 (Sequence Tagging) 是自然语言处理 (NLP) 中最基础的任务,应用十分广泛,如分词、词性标注 (POS tagging)、命名实体识别 (Named Entity Recognition,NER)、关键词抽取、语义角色标注 (Semantic Role Labeling)、槽位抽取 (Slot Filling) 等实质上都属于序列标注的范畴。在本申请中,序列标注是指在自然语言处理过程中将所获取的问句与预设的标准数据库表格进行匹配的任务。其中,按照自然语言处理的具体任务的不同,可以对所获取的问句进行切割,切割的具体形式可以是一个词组,也可以是一个子句。

[0045] 图1为本公开实施例提供的序列标注方法的流程示意图,如图1所示,该方法包括:

[0046] 步骤110,获取问句和预设的标准数据库表格;其中,标准数据库表格包括列名。

[0047] 该步骤中,问句是指有明确的语义、语法正确以及提出问题的句子,如:“价钱大于30万且出产在中国的汽车牌子一共有几个?”,该问句含有明确的语义信息(询问价钱大于30万和出产在中国的汽车牌子的数量),且合乎正常的语法表达,而“汽车牌子价钱大于30

万出产在中国一共有几个”，虽然有部分语义信息表明在询问符合要求的汽车牌子的数量和出产，但不合乎正常的语法表达，因此不能作为问句，并且，在该问句中是提出问题的，“一共有几个？”是个需要回答的问句，如果一个句子为“价钱大于30万且出产在中国的汽车牌子有3个”是不符合本申请获取问句的要求。

[0048] 预设的标准数据库表格是指数据库中存储了大量的表格。每一个表格包含多个列名，不同的列名用于存储不同的数据。可以对不同的列名进行分类处理，表格的分类结果是用于指示表格中存储的数据的类型的信息。比如：以“价钱大于30万且出产在中国的汽车牌子一共有几个？”为获取的问句，则它对应的标注数据库中存在的列名可以有“价格、品牌和产地”，“288900、659900、349400”等属于价格这一列名中的数据，“品牌1、品牌2、品牌3”等属于品牌这一列名中的数据，“中国、德国、美国”等属于产地这一列名中的数据。可以理解的是，预设的标准数据库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。

[0049] 步骤120，切割问句，得到至少一个子句以及子句中的单词，对单词进行角色类型标注，确定每个单词对应的角色类型，对子句进行子句类型标注，确定每个单词所属子句的子句类型。

[0050] 该步骤中，对获取的问句进行切割，即将问句进行拆解，使句子变成以单词为单位的样式或者以子句为单位的样式。对问句进行切割的方式可以为通过分词软件或者人工分词等方式，在本公开的实施例中不做具体限定。

[0051] 标注是指通过数据加工，借助标记工具或模型，对数据进行加工的一种行为。通常标注的类型包括：图像标注、语音标注、文本标注、视频标注等种类。

[0052] 具体地，以获取的问句“价钱大于30万且出产在中国的汽车牌子一共有几个？”为例对该步骤进一步解释说明。

[0053] 对问句“价钱大于30万且出产在中国的汽车牌子一共有几个？”进行拆解，可以拆解为“价钱、出产、牌子”等这样单个名词的样式，也可以拆解为“价钱大于30万、出产在中国”或“汽车牌子一共有几个？”这样子句的样式。在拆解成“价钱、出产、牌子”等这样单个名词样式的情况下，分别判断“价钱”、“出产”、“牌子”属于什么角色类型。在拆解成“价钱大于30万、出产在中国”或“汽车牌子一共有几个？”这样子句的样式下，分别判断“价钱大于30万”、“出产在中国”或“汽车牌子一共有几个？”属于什么子句类型。

[0054] 步骤130，将每个单词的角色类型与标准数据库表格进行级联分类处理，得到每个单词对应的列名。

[0055] 该步骤中，级联是指多个对象之间的映射关系，以级联的方式建立数据之间的关系。分类指按照种类、等级或性质分别归类。

[0056] 具体地，以获取的问句为“价钱大于30万且出产在中国的汽车牌子一共有几个？”说明，将“价钱”“出产”“牌子”与标准库中的“价格”“产地”“品牌”分别进行映射对应，得出“价钱”对应“价格”、“出产”对应“产地”以及“牌子”对应“品牌”这样的结果。

[0057] 步骤140，将问句、每个单词及其所属子句的子句类型进行级联分类处理，得到子句对应的运算符号。

[0058] 该步骤中，运算符号指表示子句或单词之间连接关系的一种特殊表征形式，基于判读出的子句类型的结果，得出在不同子句类型的子句中包含的运算符号。

[0059] 具体地,以获取的问句为“价钱大于30万且出产在中国的汽车牌子一共有几个?”说明,获取的单词可以为“价钱”、“出产”、“牌子”,获取的子句可以为“价钱大于30万”、“出产在中国”或“汽车牌子一共有几个?”,经过级联分类处理后,可以得出在“价钱大于30万”这个子句中对应的运算符号为“大于”;在“出产在中国”这个子句中对应的运算符号为“等于”,“汽车牌子一共有几个?”这个子句中对应的运算符号为“SUM”。

[0060] 步骤150,将每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,得到逻辑表达语句。

[0061] 该步骤中,规则是指是运行、运作规律所遵循的法则,可以为顺序、逻辑等其他规则,可以理解的是,预设规则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。

[0062] 具体地,问句“价钱大于30万且出产在中国的汽车牌子一共有几个?”通过预设规则进行组合,得到的逻辑表达语句为“子句类型1运算符号1(列名1)from table子句类型2(列名2)运算符号2 30万and(列名3)运算符号3中国”。

[0063] 本公开提供的一种序列标注的方法,对获取的问句进行切割,得到至少一个子句以及子句中的单词,分别对单词进行角色类型标注和对子句进行子句类型标注,得到每个单词对应的角色类型,以及每个单词所属子句的子句类型,对所获取的问句进行基于两种类型的分类判断,能够准确的与标准数据库中的数据进行匹配;之后将得到的每个单词的角色类型与标准数据库表格进行级联分类处理,得到每个单词对应的列名,将问句、每个单词及其所属子句的子句类型进行级联分类处理,得到子句对应的运算符号,分别得到与标准数据库中对应的列名和与子句对应的运算符号,能够对所获取的整个语句中的所有的信息进行完整的匹配;根据所得到的每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,最终得到逻辑表达语句。该方法较基于深度学习模型的方法设置更加简洁,并且可以有效捕捉问句之间的本地信息。

[0064] 本公开实施例提供的方法,还包括:基于逻辑表达语句在标准数据库表格中进行查询,获取查询结果。

[0065] 该步骤中,查询是指将获取的逻辑表达语句与标准数据库表格进行匹配,获取匹配结果。

[0066] 具体地,以问句“价钱大于30万且出产在中国的汽车牌子一共有几个?”为例进行说明,所得到的逻辑表达语句为“子句类型1运算符号1(列名1)from table子句类型2(列名2)运算符号2 30万and(列名3)运算符号3中国”。

[0067] 参照表1所示,表示本公开提供的标准数据库表格。在逻辑表达语句中只涉及到3个列名,对应的标准数据库表格,可以看出涉及的3个列名分别是“品牌、价格、产地”,即逻辑表达语句中的列名1为品牌,列名2为价格,列名3为产地;根据上述实施例可以得出运算符号1为“SUM”,运算符号2为“大于”,运算符号3为“等于”。

[0068] 经过与标准数据库表格的匹配,得出结果为,“子句类型1SUM(品牌)from table子句类型2(价格)>30万and(产地)=中国”。

[0069] 表1

[0070]

车型	品牌	价格	级别	能源	产地	排量
a	A	288900	一	汽油	中国	1
b	B	65990	二	汽油	中国	2

c	C	349400	三	汽油	中国	3
---	---	--------	---	----	----	---

[0071] 图2为本公开实施例提供的确定每个单词对应的角色类型的流程示意图,如图2所示,包括:

[0072] 步骤210,对单词和标准数据库表格中的列名进行拼接,得到整体序列。

[0073] 该步骤中,拼接指将问句切割后的所有单词和这些单词与标准数据库中对应的列名,连接在一起。拼接的方式可以通过拼接软件完成,也可通过人工方式完成。

[0074] 步骤220,将整体序列输入至预训练的语言模型中,生成整体序列对应的隐向量表示。

[0075] 该步骤中,预训练的语言模型可以为ELMO、GPT、BERT等模型,也可以是LSTM等循环神经网络模型。本公开实施例中使用BERT模型。BERT模型的全称是:Bidirectional Encoder Representations from Transformer。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NLP任务中作微调,最终应用于该NLP任务。

[0076] 具体地,将整体序列输入到BERT模型中,但为了获取到更好的特征表示,首先在每句开头嵌入新的特征符号[CLX],其次,在列名的拼接过程中,为了更好学到每个列名的独立特征表示,不仅嵌入了分隔符[sep]还根据每个列的不同数值类型(例如,数值类,字符类,是否类等)使用不同的类别表示符号[t]将其进行分割。举例说明:

[0077] “[CLX]A的价钱和排量是多少[sep][t₁]价格[sep][t₂]品牌[sep][t₂]产地[sep][t₁]排量[sep][t₁]”,随后将上述例子输入至BERT模型中进行编码,获取到整个序列的隐向量表示。

[0078] $Q_h; C_h = \text{BERT}([Q; C])$ (1)

[0079] 在公式(1)中,Q表示问句,C表示问句中切割的单词, Q_h 表示问句的隐向量表示, C_h 表示单词的隐向量表示。

[0080] 步骤230,对问句和每个单词进行注意力机制运算,得到问句的注意力表示。

[0081] 该步骤中,注意力机制就是在众多信息中把注意力集中放在重要的点上,选出关键信息,而忽略其他不重要的信息。提出的注意力机制(问句和列名交互时使用)包括当前提出的多种注意力交互方式,比如基于自注意力机制的self-attention或者机器翻译中常用的Dot-product Attention等。

[0082] 具体地,通过Attention注意力机制的运算方法,获取基于列名的问句注意力表示 Q_{att} (column_attentive_query),具体公式为:

[0083] $Q_{att} = \text{Attention}(Q_h, C_h)$ (2)

[0084] 步骤240,将隐向量表示和问句的注意力表示进行合并得到合并结果,将合并结果输入至角色类型标注模型中,得到每个单词对应的角色类型,其中,所述角色类型包括初始列名或初始列值。

[0085] 该步骤中,在角色类型标注模型中进行标注的方法可以是算法CRF(条件随机场),也可以是其他的算法或其他模型,在此并不做具体的限制。

[0086] 条件随机场(conditional random field algorithm,CRF)算法是一类用于序列预测问题的统计建模方法。条件随机场算法为这些任务提供了优于隐马尔可夫模型和最大熵模型的多个优点,包括减少了这些模型中的强条件独立假设等。条件随机场算法还避免

了基于有向图模型的最大熵马尔可夫模型和其他无向马尔可夫模型的基本限制(在这两种图模型中,解码时会更加偏向于具有很少后继状态的隐藏状态)。条件随机场中的参数估计通常通过前向后向算法来实现,而条件随机场中的概率预测则常常通过维特比算法实现。

[0087] 对应地,角色类型包括初始列名或初始列值,结合问句“价钱大于30万且出产在中国的汽车牌子一共有几个?”进行说明,切割问句后的单词为“价钱”、“出产”、“牌子”,将这些命名为初始列名;“30万”、“中国”为初始列值。

[0088] 具体地,使用序列标注的算法CRF(条件随机场)进行特定任务的序列标注,得到每个单词对应的角色类型,用 L_r 表示,具体公式为:

$$[0089] \quad L_r = \text{CRF}(W_h[Q_{att}:Q_h]) \quad (3)$$

[0090] 本公开实施例提供的方法,对子句进行子句类型标注,确定每个单词所属子句的子句类型,包括:

[0091] 将隐向量表示和问句的注意力表示进行合并得到合并结果,将合并结果输入至子句类型标注模型中,得到每个单词所属子句的子句类型。

[0092] 该步骤中,使用序列标注的算法CRF(条件随机场)进行特定任务的序列标注,得到每个单词所属子句的子句类型,用 L_p 表示,具体公式为:

$$[0093] \quad L_p = \text{CRF}(W_h[Q_{att}:Q_h]) \quad (4)$$

[0094] 基于上述任一实施例,该方法中,步骤130具体包括:将标准数据库表格中列名的隐向量表示和每个单词对应的角色类型进行注意力机制的交互匹配;根据匹配结果获取每个单词对应的列名。

[0095] 对应地,进行注意力机制的交互匹配可以为标准数据库表格中列名的隐向量表示与每个单词对应的初始列名的角色类型,也可以为标准数据库表格中列名的隐向量表示与每个单词对应的初始列值的角色类型,具体公式为:

$$[0096] \quad L_{r_att} = \text{Attention}(L_r, C_h) \quad (5)$$

[0097] 其中, L_{r_att} 表示注意力机制的交互, L_r 表示角色类型。

[0098] 基于注意力机制的交互,得出列名,用 L_c 表示,具体公式为:

$$[0099] \quad L_c = \text{Softmax}(W_h * (L_{r_att} + L_r)) \quad (6)$$

[0100] 基于上述任一实施例,该方法中,步骤140具体包括:

[0101] 基于每个单词所属子句的子句类型,将每个子句中的单词和问句进行注意力机制的交互匹配;根据匹配结果获取子句对应的运算符。

[0102] 对应地, L_{p_att} 表示注意力机制的交互,公式为

$$[0103] \quad L_{p_att} = \text{Attention}(L_p, Q_h) \quad (7)$$

[0104] 运算符用 L_a 表示,公式为:

$$[0105] \quad L_a = \text{Softmax}(W_h * (L_{p_att} + Q_h)) \quad (8)$$

[0106] 本公开实施例提供的方法,根据匹配结果获取子句对应的运算符,包括:

[0107] 子句类型包括条件类型和问题类型;

[0108] 在子句类型为条件类型的情况下,子句对应的运算符为聚合运算符;

[0109] 在子句类型为问题类型的情况下,子句对应的运算符为函数运算符;

[0110] 具体地,在获取的问句为“价钱大于30万且出产在中国的汽车牌子一共有几个?”的情况下,“价钱大于30万”、“出产在中国”为条件类型(where),“汽车牌子一共有几个”为

问题类型(select)；

[0111] 对应地,聚合运算符可以为“>,<,! =, =”等其他;函数运算符可以为“SUM,AVG,COUNT,NONE”等其他。“价钱大于30万”、“出产在中国”为条件类型,聚合运算符为“>”和“=”,“汽车牌子一共有几个”为问题类型,对应的函数运算符为SUM。

[0112] 进一步地,对本公开的实施例做进一步补充说明:具体参照图3所示,表示基于序列标注和级联方法的自然语言转化为逻辑表达语句技术框架,其中涉及到标准数据库表格,参照表1所示。

[0113] (1) 获取的问句为“A的价钱和排量是多少?”对该问句进行切割,形成“A”、“价钱”、“排量”单词形式或者“A的”、“价钱和排量是多少”子句形式。

[0114] (2) 将问句中的单词与标准数据库表格中的列名进行拼接,如图3中从下往上的第一行所示,表示拼接完成的整体序列。

[0115] (3) 将整体序列输入至BERT模型中,能够获取到整个序列的隐向量表示,也就是说获取的隐向量表示包括问句中单词的隐向量表示和标准数据库表格中列名的隐向量表示。

[0116] (4) 完成task1,即使用序列标注CRF算法,基于(3)结果获取问句中每个单词所属的角色类型。{column(列名)/value(列值)/operator(条件运算符)/aggregator(每个select表头的运算符)等}。

[0117] (5) 完成task2,即使用序列标注CRF算法,基于(3)结果获取问句中每个单词所属子句的子句类型{select(问题类型)/where(条件类型)}。

[0118] (6) 完成task3,使用级联分类的方式对所得到的每个单词的角色类型和每个单词所属子句的子句类型进行处理,从而获取到每个单词的初始列名/列值对应到标准数据库表格中的列名和每个单词所属子句对应的运算符号。

[0119] (7):将每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,将识别的各类信息转化为逻辑表达语句。

[0120] 其中,将整体序列输入至BERT模型中,是进行编码工作,编码模式包括BIO编码模式和BILOU编码模式,在本公开的实施例中以BIO编码模式进行说明。

[0121] 对于BIO编码模式,其包括三种标签,分别为标签B、标签I和标签O,其中,标签B表示段的开始(Begin),标签I表示段的内部(Intermediate),包括结尾,标签O表示不属于任何段的词(即Other)。

[0122] 图3中涉及的其他具体单词的解释参照表2名词解释。

[0123] 表2

名词	select	where	operator	aggregator	cond_conn_op
解释	select表头是哪些	where条件包括[column,operator,value] eg.(价格>30万)	条件运算符 Eg.><!=	每个select表头的运算符 Eg.max,avg,min,count	where 连接符 Eg. and,or, none

[0125] 基于上述任一实施例,图4为本公开实施例提供的序列标注的装置的结构示意图,如图4所示,该装置包括获取模块410,切割模块420,第一处理模块430,第二处理模块440以及组合模块450。其中,

[0126] 获取模块410,用于获取问句和预设的标准数据库表格;其中,标准数据库表格包括列名;

[0127] 切割模块420,用于切割问句,得到至少一个子句以及子句中的单词,对单词进行角色类型标注,确定每个单词对应的角色类型,对子句进行子句类型标注,确定每个单词所属子句的子句类型;

[0128] 第一处理模块430,用于将每个单词的角色类型与标准数据库表格进行级联分类处理,得到每个单词对应的列名;

[0129] 第二处理模块440,用于将问句、每个单词及其所属子句的子句类型进行级联分类处理,得到子句对应的运算符号;

[0130] 组合模块450,用于将每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,得到逻辑表达语句。本公开提供的一种序列标注的装置,首先对获取的问句进行切割,得到至少一个子句以及子句中的单词,分别对单词进行角色类型标注和对子句进行子句类型标注,得到每个单词对应的角色类型,以及每个单词所属子句的子句类型,对所获取的问句进行基于两种类型的分类判断,能够准确的与标准数据库中的数据进行匹配;之后将得到的每个单词的角色类型与标准数据库表格进行级联分类处理,得到每个单词对应的列名,将问句、每个单词及其所属子句的子句类型进行级联分类处理,得到子句对应的运算符号,分别得到与标准数据库中对应的列名和与子句对应的运算符号,能够对所获取的整个语句中的所有信息进行完整的匹配;根据所得到的每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,最终得到逻辑表达语句。本文提出使用序列标注的装置解决现有技术中自然语言转化为逻辑表达语句的问题,并且可以有效捕捉问句之间的本地信息。

[0131] 基于上述任一实施例,所述装置还包括查询模块,用于基于所述逻辑表达语句在所述标准数据库表格中进行查询,获取查询结果。

[0132] 基于上述任一实施例,所述切割模块420中的对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,包括拼接子模块、生成子模块、运算符模块以及第一合并子模块。其中,拼接子模块用于对所述单词和所述标准数据库表格中的列名进行拼接,得到整体序列;生成子模块用于将所述整体序列输入至预训练的语言模型中,生成所述整体序列对应的隐向量表示;运算符模块用于对所述问句和每个单词进行注意力机制运算,得到所述问句的注意力表示;第一合并子模块用于将所述隐向量表示和所述问句的注意力表示进行合并得到合并结果,将所述合并结果输入至角色类型标注模型中,得到每个所述单词对应的角色类型,其中所述角色类型包括初始列名或初始列值。

[0133] 基于上述任一实施例,所述切割模块420中的对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型,包括:第二合并子模块,用于将所述隐向量表示和所述问句的注意力表示进行合并得到合并结果,将所述合并结果输入至子句类型标注模型中,得到每个所述单词所属子句的子句类型。

[0134] 基于上述任一实施例,所述第一处理模块430包括:

[0135] 将所述标准数据库表格中列名的隐向量表示和每个单词对应的角色类型进行注意力机制的交互匹配;

[0136] 根据匹配结果获取每个所述单词对应的列名。

[0137] 基于上述任一实施例,所述第二处理模块440包括:

[0138] 基于每个单词所属子句的子句类型,将所述每个子句中的单词和所述问句进行注意力机制的交互匹配;

[0139] 根据匹配结果获取所述子句对应的运算符号。

[0140] 基于上述任一实施例,所述第二处理模块440中的根据匹配结果获取所述子句对应的运算符号,包括:

[0141] 所述子句类型包括条件类型和问题类型;

[0142] 在所述子句类型为条件类型的情况下,所述子句对应的运算符号为聚合运算符;

[0143] 在所述子句类型为问题类型的情况下,所述子句对应的运算符号为函数运算符;

[0144] 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,以执行序列标注的方法,该方法包括:获取问句和预设的标准数据库表格;其中,所述标准数据库表格包括列名;切割所述问句,得到至少一个子句以及子句中的单词,对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型;将每个所述单词的角色类型与标准数据库表格进行级联分类处理,得到每个所述单词对应的列名;将所述问句、每个单词及其所属子句的子句类型进行级联分类处理,得到所述子句对应的运算符号;将每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,得到逻辑表达语句。

[0145] 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。

[0146] 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序列标注的方法,该方法包括:获取问句和预设的标准数据库表格;其中,所述标准数据库表格包括列名;切割所述问句,得到至少一个子句以及子句中的单词,对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型;将每个所述单词的角色类型与标准数据库表格进行级联分类处理,得到每个所述单词对应的列名;将所述问句、每个单词及其所属子句的子句类型进行级联分类处理,得到所述子句对应的运算符号;将每个单词对应的列名以及每个单词所属子句对应的运算符号通过预设规则进行组合,得到逻辑表达语句。

[0147] 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序列标注的方法,该方法包括:获取问句和预设的标准数据库表格;其中,所述标准数据库表格包括列名;切割所述问句,得到至少一个子句以及子句中的单词,对所述单词进行角色类型标注,确定每个所述单词对应的角色类型,对所述子句进行子句类型标注,确定每个所述单词所属子句的子句类型;将每个所述单词的角色类型与标准数据库表格进行级联分类处理,得到每个所述单词对应的列名;将所述问句、每个单词及其所属子句的子句类型进行级联分类处理,得到所述子句对应的运算符;将每个单词对应的列名以及每个单词所属子句对应的运算符通过预设规则进行组合,得到逻辑表达语句。

[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。

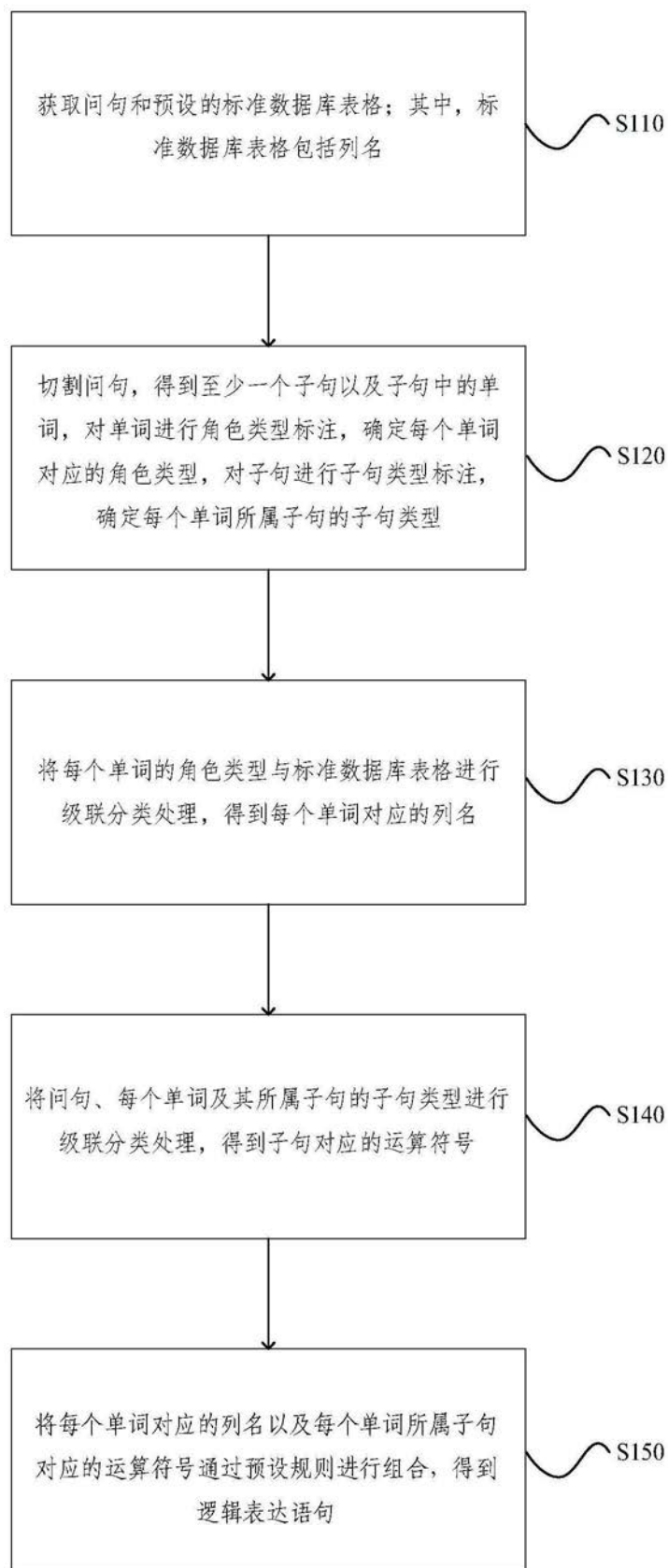


图1

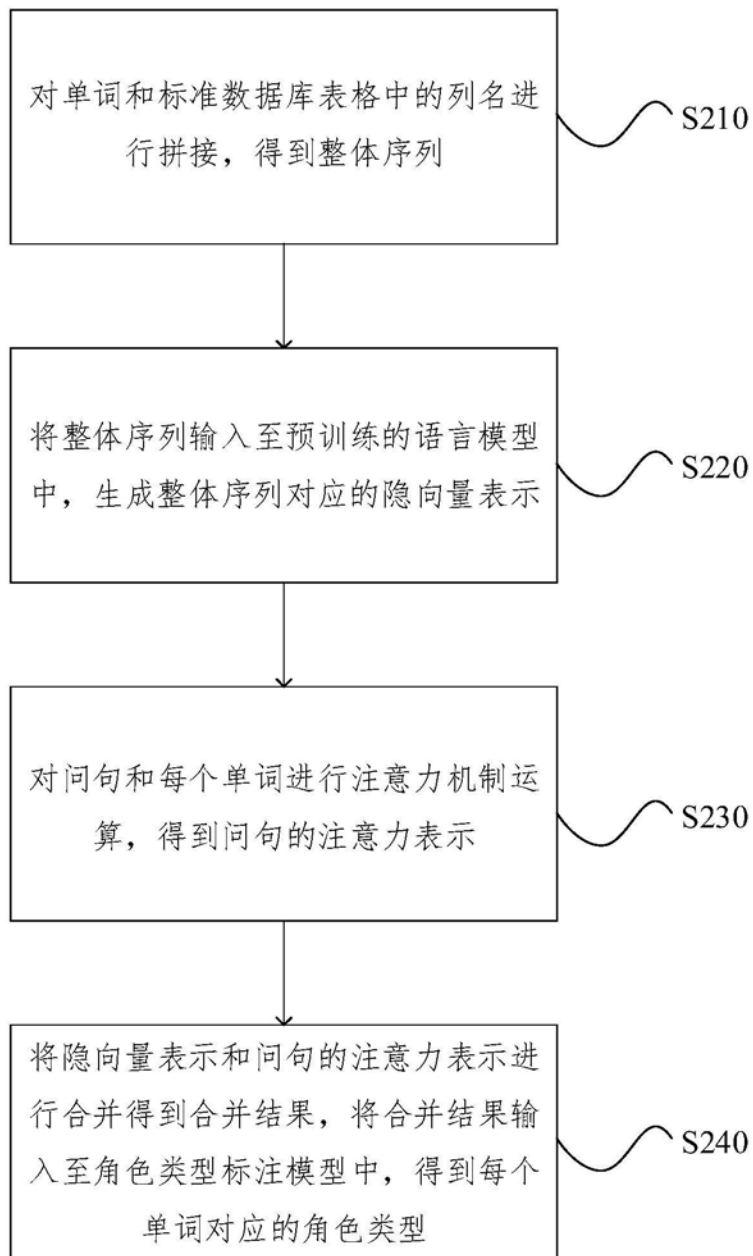


图2

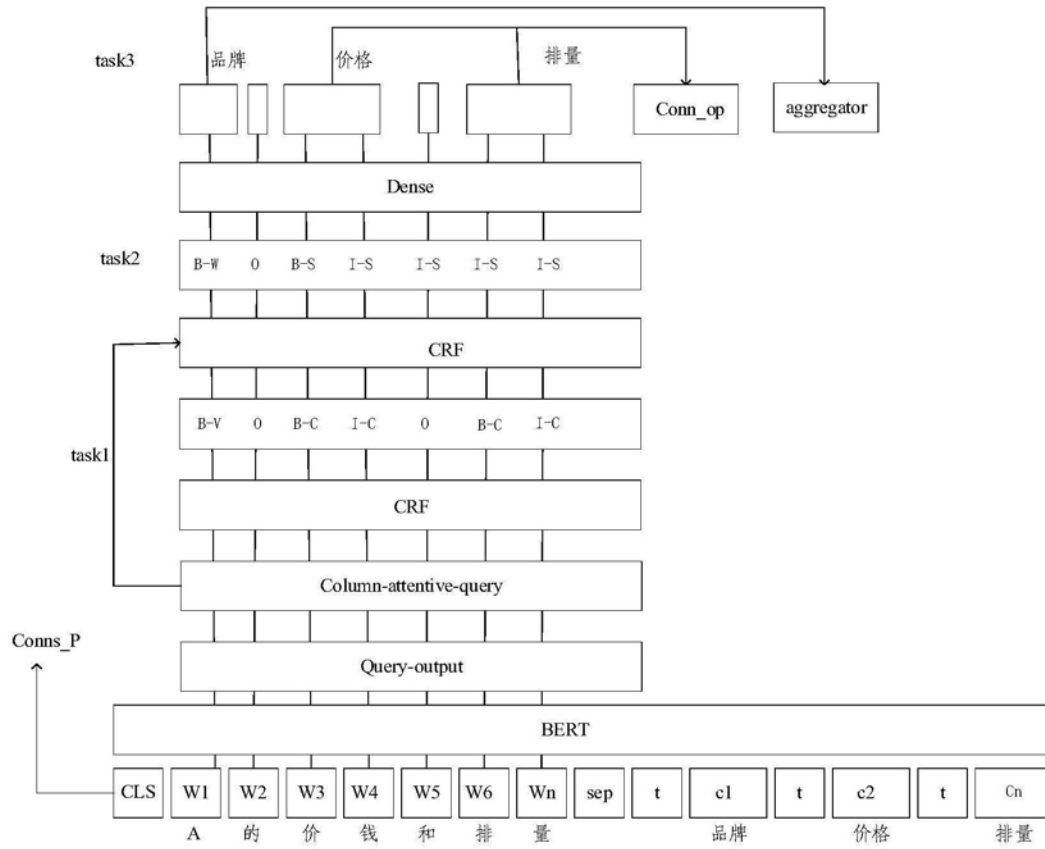


图3

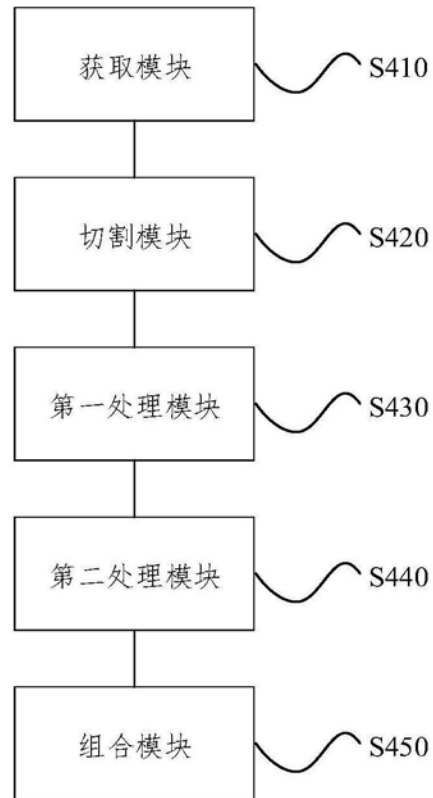


图4

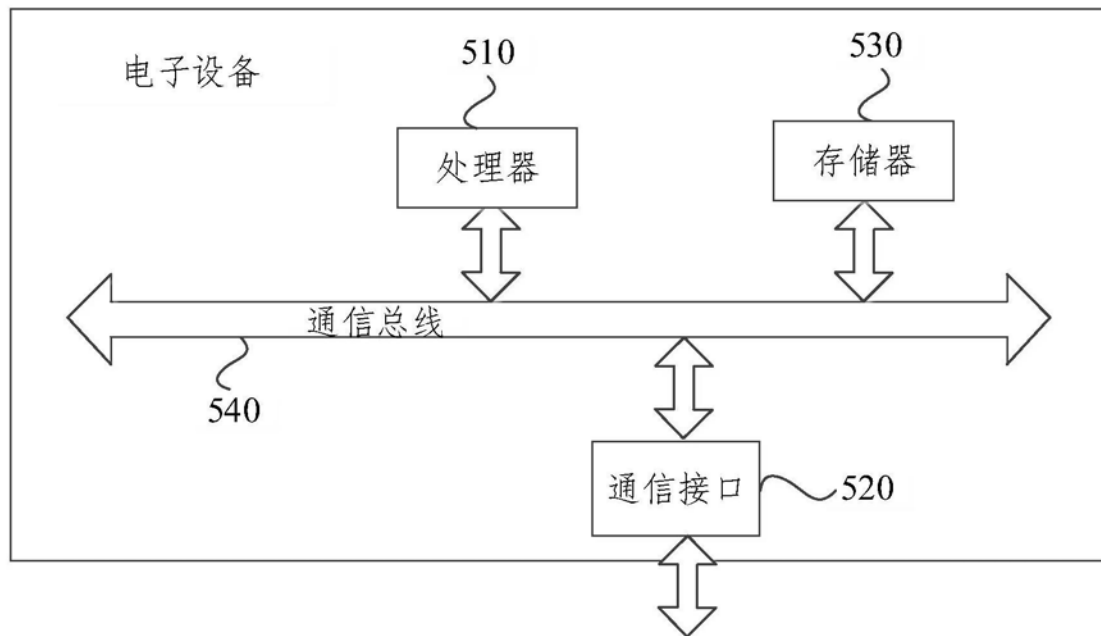


图5