



(12) 发明专利

(10) 授权公告号 CN 115455160 B

(45) 授权公告日 2024. 08. 06

(21) 申请号 202211071561.1

G06F 40/284 (2020.01)

(22) 申请日 2022.09.02

G06F 40/295 (2020.01)

(65) 同一申请的已公布的文献号

G06N 3/04 (2023.01)

申请公布号 CN 115455160 A

G06N 3/08 (2023.01)

(43) 申请公布日 2022.12.09

G06F 18/214 (2023.01)

(73) 专利权人 腾讯科技(深圳)有限公司

G06F 18/22 (2023.01)

地址 518057 广东省深圳市南山区高新区

G06N 20/00 (2019.01)

科技中一路腾讯大厦35层

(56) 对比文件

CN 112905768 A, 2021.06.04

(72) 发明人 杨韬

审查员 朱琦

(74) 专利代理机构 深圳市深佳知识产权代理事

务所(普通合伙) 44285

专利代理师 聂秀娜

(51) Int. Cl.

G06F 16/332 (2019.01)

G06F 16/35 (2019.01)

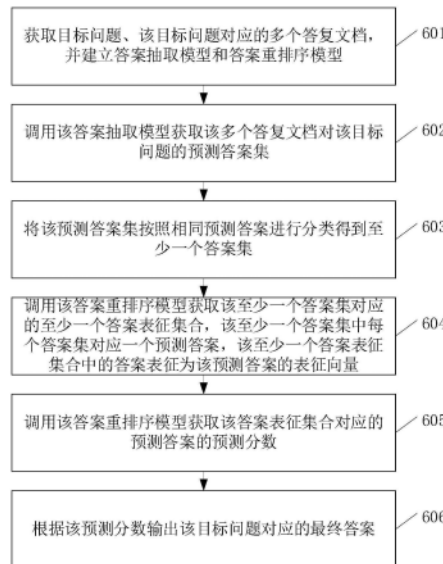
权利要求书4页 说明书17页 附图8页

(54) 发明名称

一种多文档阅读理解方法、装置、设备及存储介质

(57) 摘要

本申请实施例提供了一种多文档阅读理解方法、装置、设备及存储介质,用于获取抽取相同答案的答复文档之间的语义特征,从而增加答案重排序模型中的排序效果,提升多文档阅读理解的准确性。包括:获取目标问题、目标问题对应的多个答复文档并建立答案抽取模型和答案重排序模型;调用答案抽取模型获取多个答复文档对目标问题的预测答案集;将预测答案集按照相同预测答案进行分类得到至少一个答案集;调用该答案重排序模型获取至少一个答案集对应的至少一个答案表征集合;调用答案重排序模型获取答案表征集合对应的预测答案的预测分数;根据预测分数输出目标问题对应的最终答案。本申请可应用于人工智能领域。



1. 一种多文档阅读理解方法,其特征在于,包括:

获取目标问题、所述目标问题对应的多个答复文档,并建立答案抽取模型和答案重排序模型;

调用所述答案抽取模型获取所述多个答复文档对所述目标问题的预测答案集;

将所述预测答案集按照相同预测答案进行分类得到至少一个答案集;

调用所述答案重排序模型获取所述至少一个答案集对应的至少一个答案表征集合,包括:调用所述答案重排序模型获取所述目标问题与第一答复文档集中各个答复文档的第一输出向量集合,所述第一答复文档集包括所述至少一个答案集中第一答案集对应的答复文档;从所述第一输出向量集合中选择所述第一答复文档集对应的预测答案的字符串进行池化得到所述第一答案集中各个预测答案的中间答案表征集合;对所述中间答案表征集合进行自注意力处理得到所述第一答案集对应的预测答案的第一答案表征;依此类推,遍历得到所述至少一个答案集中各个答案集对应的预测答案的答案表征,并归类为所述至少一个答案表征集合,所述第一答案表征包含于所述至少一个答案表征集合,所述至少一个答案集中每个答案集对应一个预测答案,所述至少一个答案表征集合中的答案表征为所述预测答案的表征向量;

调用所述答案重排序模型获取所述答案表征集合对应的预测答案的预测分数;

根据所述预测分数输出所述目标问题对应的最终答案。

2. 根据权利要求1所述的方法,其特征在于,所述调用所述答案重排序模型获取所述目标问题与第一答复文档集中各个答复文档的第一输出向量集合包括:

将所述目标问题的每个字符与所述第一答复文档集中每个答复文档的每个字符,利用开始字符与间隔字符进行拼接得到第一词序列集合;

将所述第一词序列集合输入所述答案重排序模型得到所述第一输出向量集合。

3. 根据权利要求1所述的方法,其特征在于,所述对所述中间答案表征集合进行自注意力处理得到所述第一答案集对应的预测答案的第一答案表征包括:

对所述中间答案表征集合中的每个中间答案表征进行自注意力处理得到自注意预测分数集合;

对所述自注意预测分数集合中进行归一化处理得到归一化预测分数集合;

对所述归一化预测分数集合中的各个归一化预测分数进行加权求和得到所述第一答案表征。

4. 根据权利要求1至3中任一项所述的方法,其特征在于,所述方法还包括:

获取训练样本集、并建立初始答案抽取模型和初始答案重排序模型,所述训练样本集包括问题样本集、所述问题样本集对应的答复文档样本集;

利用所述训练样本集训练所述初始答案抽取模型和所述初始答案重排序模型得到所述答案抽取模型和所述答案重排序模型。

5. 根据权利要求4所述的方法,其特征在于,所述初始答案抽取模型和所述初始答案重排序模型共享相同的编码层,所述利用所述训练样本集训练所述初始答案抽取模型和所述初始答案重排序模型得到所述答案抽取模型和所述答案重排序模型包括:

调用所述编码层获取所述训练样本集对应的第二输出向量;

将所述第二输出向量输入所述初始答案抽取模型得到第一损失值和训练答案集;

将所述训练答案集输入所述初始答案重排序模型得到第二损失值；

利用所述第一损失值和所述第二损失值之和反向调整所述初始答案抽取模型和所述初始答案重排序模型的权重参数得到所述答案抽取模型和所述答案重排序模型。

6. 根据权利要求5所述的方法,其特征在于,所述将所述输出向量输入所述初始答案抽取模型得到第一损失值和训练答案集包括:

将所述输出向量输入所述初始答案抽取模型得到所述训练答案集;

获取所述输出向量的训练答案集中每个训练答案对应的开始词的预测分数和结束词的预测分数;

利用激活函数将所述开始词的预测分数转化为第一概率值,并利用所述激活函数将所述结束词的预测分数转化为第二概率值;

根据所述第一概率值和所述第二概率值得到所述第一损失值;

所述将所述训练答案集输入所述初始答案重排序模型得到第二损失值包括:

将所述训练答案集输入所述初始答案重排序模型得到训练答案集对应的训练答案表征集合;

获取所述训练答案表征集合中各个训练答案对应的训练预测分数集合;

利用激活函数将所述训练预测分数集合中的各个训练预测分数转化为第三概率值;

利用交叉熵和所述第三概率值得到所述第二损失值。

7. 一种多文档阅读理解装置,其特征在于,包括:

获取模块,用于获取目标问题、所述目标问题对应的多个答复文档、答案抽取模型和答案重排序模型;

处理模块,用于调用所述答案抽取模型获取所述多个答复文档对所述目标问题的预测答案集;将所述预测答案集按照相同预测答案进行分类得到至少一个答案集;调用所述答案重排序模型获取所述至少一个答案集对应的至少一个答案表征集合,包括:调用所述答案重排序模型获取所述目标问题与第一答复文档集中各个答复文档的第一输出向量集合,所述第一答复文档集包括所述至少一个答案集中第一答案集对应的答复文档;从所述第一输出向量集合中选择所述第一答复文档集对应的预测答案的字符串进行池化得到所述第一答案集中各个预测答案的中间答案表征集合;对所述中间答案表征集合进行自注意力处理得到所述第一答案集对应的预测答案的第一答案表征;依此类推,遍历得到所述至少一个答案集中各个答案集对应的预测答案的答案表征,并归类为所述至少一个答案表征集合,所述第一答案表征包含于所述至少一个答案表征集合,所述至少一个答案集中每个答案集对应一个预测答案,所述至少一个答案表征集合中的答案表征为所述预测答案的表征向量;调用所述答案重排序模型获取所述答案表征集合对应的预测答案的预测分数;

输出模块,用于根据所述预测分数输出所述目标问题对应的最终答案。

8. 根据权利要求7所述的装置,其特征在于,所述处理模块,具体用于:

将所述目标问题的每个字符与所述第一答复文档集中每个答复文档的每个字符,利用开始字符与间隔字符进行拼接得到第一词序列集合;

将所述第一词序列集合输入所述答案重排序模型得到所述第一输出向量集合。

9. 根据权利要求7所述的装置,其特征在于,所述处理模块,具体用于:

对所述中间答案表征集合中的每个中间答案表征进行自注意力处理得到自注意预测

分数集合；

对所述自注意预测分数集合中进行归一化处理得到归一化预测分数集合；

对所述归一化预测分数集合中的各个归一化预测分数进行加权求和得到所述第一答案表征。

10. 根据权利要求7至9中任一项所述的装置,其特征在于,还包括:训练模块;

所述获取模块,还用于获取训练样本集、并建立初始答案抽取模型和初始答案重排序模型,所述训练样本集包括问题样本集、所述问题样本集对应的答复文档样本集;

所述训练模块,用于利用所述训练样本集训练所述初始答案抽取模型和所述初始答案重排序模型得到所述答案抽取模型和所述答案重排序模型。

11. 根据权利要求10所述的装置,其特征在于,所述初始答案抽取模型和所述初始答案重排序模型共享相同的编码层,所述训练模块,具体用于:

调用所述编码层获取所述训练样本集对应的第二输出向量;

将所述第二输出向量输入所述初始答案抽取模型得到第一损失值和训练答案集;

将所述训练答案集输入所述初始答案重排序模型得到第二损失值;

利用所述第一损失值和所述第二损失值之和反向调整所述初始答案抽取模型和所述初始答案重排序模型的权重参数得到所述答案抽取模型和所述答案重排序模型。

12. 根据权利要求11所述的装置,其特征在于,所述训练模块,具体用于:

将所述输出向量输入所述初始答案抽取模型得到所述训练答案集;

获取所述输出向量的训练答案集中每个训练答案对应的开始词的预测分数和结束词的预测分数;

利用激活函数将所述开始词的预测分数转化为第一概率值,并利用所述激活函数将所述结束词的预测分数转化为第二概率值;

根据所述第一概率值和所述第二概率值得到所述第一损失值;

所述训练模块,具体还用于:

将所述训练答案集输入所述初始答案重排序模型得到训练答案集对应的训练答案表征集合;

获取所述训练答案表征集合中各个训练答案对应的训练预测分数集合;

利用激活函数将所述训练预测分数集合中的各个训练预测分数转化为第三概率值;

利用交叉熵和所述第三概率值得到所述第二损失值。

13. 一种计算机设备,其特征在于,包括:存储器、处理器以及总线系统;

其中,所述存储器用于存储程序;

所述处理器用于执行所述存储器中的程序,所述处理器用于根据程序代码中的指令执行权利要求1至6中任一项所述的方法;

所述总线系统用于连接所述存储器以及所述处理器,以使所述存储器以及所述处理器进行通信。

14. 一种计算机可读存储介质,包括指令,当其在计算机上运行时,使得计算机执行如权利要求1至6中任一项所述的方法。

15. 一种计算机程序产品,其特征在于,所述计算机程序产品包括计算机指令,所述计算机指令存储在计算机可读存储介质中;计算机设备的处理器从所述计算机可读存储介质

读取所述计算机指令,所述处理器执行所述计算机指令,使得所述计算机设备执行如权利要求1至6中任一项所述的方法。

## 一种多文档阅读理解方法、装置、设备及存储介质

### 技术领域

[0001] 本申请涉及人工智能领域,尤其涉及一种多文档阅读理解方法、装置、设备及存储介质。

### 背景技术

[0002] 机器阅读理解(machine reading comprehension,MRC)是自然语言处理(natural language processing,NLP)领域的一个重要任务,其旨在让机器根据给定的问题和文章,抽取相关的信息和知识,从而得到答案。相对于自然语言处理中的命名实体识别(named entity recognition,NER),关系抽取等基础任务,MRC是一个更复杂,更高层的任务,其对语义的理解要求更高,抽取的文本信息更多。其可以应用于多个领域,比如搜索引擎。

[0003] 搜索引擎是人们获取信息的一个非常重要途径。即大量用户通过搜索引擎搜索查询,这其中包括了大量的问答型查询操作。但传统的网页搜索引擎一般做的都是“匹配”的工作,即查询的问题与答复文档之间做相关性匹配,而不是更精准的理解这个问题。

[0004] 智能问答技术正好可以弥补传统搜索引擎的这个局限。用户向系统提交自然语言查询,系统直接返回满足用户需求的答案,减少人工参与成本,将用户获取信息和知识的过程变成一问一答的方式。对用户来说,帮助用户最快时间获取到答案,搜索体验比较好,对内容提供方来说,答案首条置顶展示,可以获得更多曝光和流量,有助于内容生态的建设。因此目前急需一种可以提升多文档阅读理解的准确性的方案。

### 发明内容

[0005] 本申请实施例提供了一种多文档阅读理解方法、装置、设备及存储介质,用于将多个答复文档提取的预测答案进行分类,并获取抽取出相同答案的答复文档之间的语义特征,从而增加答案重排序模型中的排序效果,提升多文档阅读理解的准确性。

[0006] 有鉴于此,本申请一方面提供一种多文档阅读理解方法,包括:获取目标问题、该目标问题对应的多个答复文档,并建立答案抽取模型和答案重排序模型;

[0007] 调用该答案抽取模型获取该多个答复文档对该目标问题的预测答案集;

[0008] 将该预测答案集按照相同预测答案进行分类得到至少一个答案集;

[0009] 调用该答案重排序模型获取该至少一个答案集对应的至少一个答案表征集合,该至少一个答案集中每个答案集对应一个预测答案,该至少一个答案表征集合中的答案表征为该预测答案的表征向量;

[0010] 调用该答案重排序模型获取该答案表征集合对应的预测答案的预测分数;

[0011] 根据该预测分数输出该目标问题对应的最终答案。

[0012] 本申请另一方面提供一种多文档阅读理解装置,包括:获取模块,用于获取目标问题、该目标问题对应的多个答复文档,并建立答案抽取模型和答案重排序模型;

[0013] 处理模块,用于调用该答案抽取模型获取该多个答复文档对该目标问题的预测答案集;将该预测答案集按照相同预测答案进行分类得到至少一个答案集;调用该答案重排

序模型获取该至少一个答案集对应的至少一个答案表征集合,该至少一个答案表征集合中的答案表征为该至少一个答案集中每个答案集对应的预测答案的表征向量;调用该答案重排序模型获取该答案表征集合对应的预测答案的预测分数;

[0014] 输出模块,用于根据该预测分数输出该目标问题对应的最终答案。

[0015] 在一种可能的设计中,在本申请实施例的另一方面的另一种实现方式中,该处理模块,具体用于调用该答案重排序模型获取该目标问题与第一答复文档集中各个答复文档的第一输出向量集合,该第一答复文档集包括该至少一个答案集中第一答案集对应的答复文档;

[0016] 从该第一输出向量集合中选择该第一答复文档集对应的预测答案的字符串进行池化得到该第一答案集中各个预测答案的中间答案表征集合;

[0017] 对该中间答案表征集合进行自注意力处理得到该第一答案集对应的预测答案的第一答案表征;

[0018] 依此类推,遍历得到该至少一个答案集中各个答案集对应的预测答案的答案表征,并归类为该至少一个答案表征集合,该第一答案表征包含于该至少一个答案表征集合。

[0019] 在一种可能的设计中,在本申请实施例的另一方面的另一种实现方式中,该处理模块,具体用于将该目标问题的每个字符与该第一答复文档集中每个答复文档的每个字符,利用开始字符与间隔字符进行拼接得到第一词序列集合;

[0020] 将该第一词序列集合输入该答案重排序模型得到该第一输出向量集合。

[0021] 在一种可能的设计中,在本申请实施例的另一方面的另一种实现方式中,该处理模块,具体用于对该中间答案表征集合中的每个中间答案表征进行自注意力处理得到自注意预测分数集合;

[0022] 对该自注意预测分数集合中进行归一化处理得到归一化预测分数集合;

[0023] 对该归一化预测分数集合中的各个归一化预测分数进行加权求和得到该第一答案表征。

[0024] 在一种可能的设计中,在本申请实施例的另一方面的另一种实现方式中,该获取模块,还用于获取训练样本集,并建立初始答案抽取模型和初始答案重排序模型,该训练样本集包括问题样本集、该问题样本集对应的答复文档样本集;

[0025] 该装置还包括训练模块,用于利用该训练样本集训练该初始答案抽取模型和该初始答案重排序模型得到该答案抽取模型和该答案重排序模型。

[0026] 在一种可能的设计中,在本申请实施例的另一方面的另一种实现方式中,该初始答案抽取模型和该初始答案重排序模型共享相同的编码层,该训练模块,具体用于调用该编码层获取该训练样本集对应的第二输出向量;

[0027] 将该第二输出向量输入该初始答案抽取模型得到第一损失值和训练答案集;

[0028] 将该训练答案集输入该初始答案重排序模型得到第二损失值;

[0029] 利用该第一损失值和该第二损失值之和反向调整该初始答案抽取模型和该初始答案重排序模型的权重参数得到该答案抽取模型和该答案重排序模型。

[0030] 在一种可能的设计中,在本申请实施例的另一方面的另一种实现方式中,该训练模块,具体用于将该输出向量输入该初始答案抽取模型得到该训练答案集;

[0031] 获取该输出向量的训练答案集中每个训练答案对应的开始词的预测分数和结束

词的预测分数；

[0032] 利用激活函数将该开始词的预测分数转化为第一概率值,并利用该激活函数将该结束词的预测分数转化为第二概率值；

[0033] 根据该第一概率值和该第二概率值得到该第一损失值；

[0034] 该训练模块,具体用于将该训练答案集输入该初始答案重排序模型得到训练答案集对应的训练答案表征集合；

[0035] 获取该训练答案表征集合中各个训练答案对应的训练预测分数集合；

[0036] 利用激活函数将该训练预测分数集合中的各个训练预测分数转化为第三概率值；

[0037] 利用交叉熵和该第三概率值得到该第二损失值。

[0038] 本申请另一方面提供一种计算机设备,包括:存储器、处理器以及总线系统；

[0039] 其中,存储器用于存储程序；

[0040] 处理器用于执行存储器中的程序,处理器用于根据程序代码中的指令执行上述各方面的方法；

[0041] 总线系统用于连接存储器以及处理器,以使存储器以及处理器进行通信。

[0042] 本申请的另一面提供了一种计算机可读存储介质,计算机可读存储介质中存储有指令,当其在计算机上运行时,使得计算机执行上述各方面的方法。

[0043] 本申请的另一个方面,提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述各方面所提供的方法。

[0044] 从以上技术方案可以看出,本申请实施例具有以下优点:将多个答复文档提取的预测答案进行分类,并获取抽取出相同答案的答复文档之间的语义特征,从而增加答案重排序模型中的排序效果,提升多文档阅读理解的准确性。

## 附图说明

[0045] 图1为问答系统的一个流程示意图；

[0046] 图2为本申请实施例中多文档阅读理解方法的应用系统架构的一个架构示意图；

[0047] 图3为本申请实施例中答案抽取模型和答案重排序模型的一个系统架构示意图；

[0048] 图4a为本申请实施例中答案抽取模型和答案重排序模型的一个训练架构示意图；

[0049] 图4b为本申请实施例中答案抽取模型和答案重排序模型的一个网络架构示意图；

[0050] 图5a为本申请实施例中答案抽取模型和答案重排序模型的另一个训练架构示意图；

[0051] 图5b为本申请实施例中答案抽取模型和答案重排序模型的另一个网络架构示意图；

[0052] 图6为本申请实施例中多文档阅读理解方法的一个实施例示意图；

[0053] 图7为本申请实施例中多文档阅读理解装置的一个实施例示意图；

[0054] 图8为本申请实施例中多文档阅读理解装置的一个实施例示意图；

[0055] 图9为本申请实施例中多文档阅读理解装置的一个实施例示意图；

[0056] 图10为本申请实施例中多文档阅读理解装置的一个实施例示意图。

## 具体实施方式

[0057] 本申请实施例提供了一种多文档阅读理解方法、装置、设备及存储介质,用于将多个答复文档提取的预测答案进行分类,并获取抽取相同答案的答复文档之间的语义特征,从而增加答案重排序模型中的排序效果,提升多文档阅读理解的准确性。

[0058] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等(如果存在)是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本申请的实施例例如能够以除了在这里图示或描述的那些以外的顺序实施。此外,术语“包括”和“对应于”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0059] 鉴于本申请会涉及到一些专业术语,因此,下面将对这部分专业术语先进行介绍。

[0060] 分词,将连续的字序列按照一定的规范重新组合成词序列的过程。通过让计算机模拟人对句子的理解,达到识别词的效果。

[0061] 实体词,实体是指能够独立存在的、作为一切属性的基础和万物本原的东西,即实体词是指能够表示实体的词语。名词与代词为实体词,例如“张三”、“妻子”为实体词。

[0062] 意图词,意图是清楚地意识到要争取实现的目标,意图词是指能够表示问题的词语,例如“谁”、“是谁”、“在哪”为意图词。

[0063] 机器学习(Machine Learning,ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0064] 深度学习(Deep Learning,DL)是机器学习的分支,是一种试图使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法。

[0065] 神经网络(Neural Network,NN),在机器学习和认知科学领域的一种模仿生物神经网络结构和功能的深度学习模型。

[0066] 信息抽取(information extraction),即从自然语言文本中,抽取特定的事件或事实信息,帮助用户将海量内容自动分类、提取和重构。该特定的事件或事实信息通常包括实体(entity)、关系(relation)、事件(event)。例如,从新闻中抽取时间、地点、关键人物,或者从技术文档中抽取产品名称、开发时间、性能指标等。由于信息抽取能从自然语言中抽取信息框架和用户感兴趣的事实信息,无论是在知识图谱、信息检索、问答系统还是在情感分析、文本挖掘中,信息抽取都具有广泛应用。信息抽取主要包括三个子任务:实体抽取与链指,关系抽取,以及事件抽取。其中,实体抽取与链指,即为命名实体识别。关系抽取,即为三元组(triple)抽取,主要用于抽取实体间的关系。事件抽取,相当于一种多元关系的抽取。

[0067] 关系抽取(Relation Extraction,RE),给定一个实体对和包含该实体对的文本,旨基于文本判断该实体对的语义关系。例如,给定实体对(M国,国家总统)和文本(“总统候

选人A在最近的大选中击败了总统候选人B,成为M国的下一任总统…)”,用户希望识别出实体“总统候选人A”和“M国”之间的关系为“国家总统”。在关系抽取中,通常会预定义好一组关系,如“国家总统”。

[0068] 关系分类(Relation Classification,RC),关系抽取的一种建模方式,即将关系抽取转换为分类问题,其中每个关系对应一个类别。

[0069] 问答系统(Question Answering,QA),给定一段文本和一个问题,问答系统能从该文本中识别出文本的答案位置。

[0070] 机器阅读理解(machine reading comprehension,MRC)是自然语言处理(natural language processing,NLP)领域的一个重要任务,其旨在让机器根据给定的问题和文章,抽取相关的信息和知识,从而得到答案。相对于自然语言处理中的命名实体识别(named entity recognition,NER),关系抽取等基础任务,MRC是一个更复杂,更高层的任务,其对语义的理解要求更高,抽取的文本信息更多。其可以应用于多个领域,比如搜索引擎。搜索引擎是人们获取信息的一个非常重要途径。即大量用户通过搜索引擎搜索查询,这其中包括了大量的问答型查询操作。但传统的网页搜索引擎一般做的都是“匹配”的工作,即查询的问题与答复文档之间做相关性匹配,而不是更精准的理解这个问题。智能问答技术正好可以弥补传统搜索引擎的这个局限。用户向系统提交自然语言查询,系统直接返回满足用户需求的答案,减少人工参与成本,将用户获取信息和知识的过程变成一问一答的方式。对用户来说,帮助用户最快时间获取到答案,搜索体验比较好,对内容提供方来说,答案首条置顶展示,可以获得更多曝光和流量,有助于内容生态的建设。目前传统的智能问答技术主要包括三个重要模块:文章检索模块、答案抽取模块和答案重排序模块,其具体流程可以如图1所示:

[0071] 获取输入的查询字符,通过问答意图过滤得到具体的查询问题;然后通过文章检索模块从段落索引库检测到相关的答复文档;然后将多文档输入至多文档答案抽取模块和答案重排序模块得到最终答案。其中,答案抽取模块的作用就是从检索出来的每个段落里抽取答案,而答案重排序模块就是对所有抽取出的答案在做一个统一的排序,最终选择分数最高的那一个作为这个查询问题(query)的最终答案。即答案抽取和答案重排序是采用pipeline方式,也就是先进行答案抽取,再进行答案重排序。答案重排序一般都是采用基于人工特征的方式,即会构造一些人工特征。比如抽取答案的答复文档的个数,答案在答复文档中的位置,答案抽取模块的分数等特征训练一个逻辑回归(Logistic Regression,LR)线性模型或者梯度提升决策树(Gradient Boosting Decision Tree,GBDT)树型模型。但是在这个过程中,抽取各个答案的答复文档之间相互独立,并未利用上其相互之间的语义特征。

[0072] 为了解决这一问题,本申请提供如下技术方案:获取目标问题、该目标问题对应的多个答复文档,并建立答案抽取模型和答案重排序模型;调用该答案抽取模型获取该多个答复文档对该目标问题的预测答案集;将该预测答案集按照相同预测答案进行分类得到至少一个答案集;调用该答案重排序模型获取该至少一个答案集对应的至少一个答案表征集合,该至少一个答案集中每个答案集对应一个预测答案,该至少一个答案表征集合中的答案表征为该预测答案的表征向量;调用该答案重排序模型获取该答案表征集合对应的预测答案的预测分数;根据该预测分数输出该目标问题对应的最终答案。这样将多个答复文档

提取的预测答案进行分类,并获取抽取出相同答案的答复文档之间的语义特征,从而增加答案重排序模型中的排序效果,提升多文档阅读理解的准确性。

[0073] 本申请提供的方法应用于如图2所示的系统架构,图2为本申请实施例中系统架构的一个架构示意图,如图2所示,系统架构包括服务器和终端设备,且客户端(即搜索引擎或者社交软件小程序)部署于终端设备上,其中,客户端可以通过浏览器的形式运行于终端设备上,也可以通过独立的应用程序(application,APP)的形式运行于终端设备上等,对于客户端的具体展现形式,此处不做限定。本申请涉及的服务器可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network,CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。终端设备可以是智能手机、平板电脑、笔记本电脑、掌上电脑、个人电脑、智能电视、智能手表、车载设备、可穿戴设备等,但并不局限于此。终端设备以及服务器可以通过有线或无线通信方式进行直接或间接地连接,本申请在此不做限制。服务器和终端设备的数量也不做限制。本申请提供的方案可以由终端设备独立完成,也可以由服务器独立完成,还可以由终端设备与服务器配合完成,对此,本申请并不做具体限定。其中,本申请中涉及的答复文档可以存储于数据库(Database)。其中,该数据库简而言之可视为电子化的文件柜——存储电子文件的处所,用户可以对文件中的数据进行新增、查询、更新、删除等操作。所谓“数据库”是以一定方式储存在一起、能与多个用户共享、具有尽可能小的冗余度、与应用程序彼此独立的数据集合。数据库管理系统(Database Management System,DBMS)是为管理数据库而设计的电脑软件系统,一般具有存储、截取、安全保障、备份等基础功能。数据库管理系统可以依据它所支持的数据库模型来作分类,例如关系式、可扩展标记语言(Extensible Markup Language,XML);或依据所支持的计算机类型来作分类,例如服务器群集、移动电话;或依据所用查询语言来作分类,例如结构化查询语言(Structured Query Language,SQL)、XQuery;或依据性能冲量重点来作分类,例如最大规模、最高运行速度;亦或其他分类方式。不论使用哪种分类方式,一些DBMS能够跨类别,例如,同时支持多种查询语言。

[0074] 可以理解的是,在本申请的具体实施方式中,涉及到答复文档等相关的数据,当本申请以上实施例运用到具体产品或技术中时,需要获得用户许可或者同意,且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准。

[0075] 可以理解的是,本申请实施例中,在进行多文档阅读理解之前,本申请实施例中还可以提供答案抽取模型与答案重排序模型的训练方法。一种示例性方案中,本申请实施例提供的答案抽取模型与答案重排序模型的训练方法,能够由计算机设备执行。下面介绍一下本申请实施例提供的答案抽取模型与答案重排序模型的训练方法的实施环境,图3是本申请实施例提供的一种答案抽取模型与答案重排序模型的训练方法的实施环境示意图。参见图3,该实施环境包括终端设备101和服务器102。终端设备101和服务器102能够通过有线或无线通信方式进行直接或间接地连接,本申请在此不做限制。

[0076] 在一些实施例中,终端设备101是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表、智能语音交互设备、智能家电以及车载终端等,但并不局限于此。终端设备101安装和运行有支持内容推荐的客户端,其中,客户端可以通过浏览器的形式运行于

终端设备101上,也可以通过独立的应用程序(application,APP)的形式运行于终端设备上,对于客户端的具体展现形式,此处不做限定。在一些实施例中,服务器102是独立的物理服务器,也能够是多个物理服务器构成的服务器集群或者分布式系统,还能够是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network,CDN)以及大数据和人工智能平台等基础云计算服务的云服务器。服务器102用于为支持虚拟场景的应用程序提供后台服务。在一些实施例中,服务器102承担主要计算工作,终端设备101承担次要计算工作,例如,该终端设备101向该服务器102提供样本数据,该服务器102提供该多目标推荐模型的训练过程;或者,服务器102和终端设备101二者之间采用分布式计算架构进行协同计算。

[0077] 可以理解的是,上述终端设备101的数量可以更多或者更少。比如上述终端设备101可以仅为一个,或者上述终端设备101可为几十个或者上百个,或者更多数量。即本申请实施例对于终端设备101的数量和设备类型不做限定。

[0078] 下面介绍一下本申请实施例提供的答案抽取模型与答案重排序模型的训练方法的训练架构,如图4a所示的训练架构中,获取该答案抽取模型的训练样本集;其中,该训练样本集中的真实的抽取答案作为监督数据对该答案抽取模型进行监督训练。在获取到训练样本集后,其整体的训练流程可以如下:该训练样本集输入该答案抽取模型输出预估答案和其他输出结果,然后该预估答案与该真实的抽取答案计算第一损失,并根据该第一损失反向调整该答案抽取模型的参数从而实现该答案抽取模型的训练,直到达到训练结束的条件;该预估答案按照相同的预估答案分为一个集合的规则分类得到多个预估答案集合,并将多个预估答案集合中的每一个预估答案集通过答案重排序模型的表示层和编码层得到该预估答案集合中答案的一个最终答案表征,并将该最终答案表征输入该答案重排序模型得到各个预估答案集对应的预估答案的预估分数,然后将该预估分数与该真实答案计算第二损失;最后根据该第二损失反向调整该答案重排序模型的参数从而实现该答案重排序模型的训练,直到达到训练结束的条件。可以理解的是,如图4a所示的训练架构,其答案抽取模型与该答案重排序模型的网络架构可以如图4b所示。

[0079] 可以理解的是,为了提升该答案抽取模型和该答案重排序模型的训练效果,下面介绍一下本申请实施例提供的答案抽取模型与答案重排序模型的训练方法的训练架构,如图5a所示的训练架构中,获取该答案抽取模型的训练样本集;其中,该训练样本集中的真实的抽取答案作为监督数据对该答案抽取模型进行监督训练。在获取到训练样本集后,其整体的训练流程可以如下:该训练样本集输入该答案抽取模型输出预估答案和其他输出结果,然后该预估答案与该真实的抽取答案计算第一损失;该预估答案按照相同的预估答案分为一个集合的规则分类得到多个预估答案集合,并将多个预估答案集合中的每一个预估答案集通过答案重排序模型的表示层和编码层得到该预估答案集合中答案的一个最终答案表征,并将该最终答案表征输入该答案重排序模型得到各个预估答案集对应的预估答案的预估分数,然后将该预估分数与该真实答案计算第二损失;最后取该第二损失和该第一损失值之和作为该答案抽取模型和该答案重排序模型的整体损失值,并根据该整体损失值反向调整该答案抽取模型和该答案重排序模型的参数从而实现该答案抽取模型和该答案重排序模型的训练,直到达到训练结束的条件。可以理解的是,如图5a所示的训练架构,其答案抽取模型与该答案重排序模型的网络架构可以如图5b所示。在此过程中,该答案抽取

模型通过公式 $\text{logits}_i = v\text{emb}_i$ 以及公式 $P_i = \text{softmax}(\text{logits}_i)$ 计算得到每个字符(即 token)的预测分数和概率值,然后根据各个字符的预测分数和概率值选择预测答案,并确定该预测答案的开始位置(即 start token)的概率和该预测答案的结束位置(即 end token)的概率,即该预测答案的开始字符和该预测答案的结束字符;然后再根据公式 $\text{Loss1} = -\log P_{\text{start}} - \log P_{\text{end}}$ 得到该答案抽取模型的第一损失;同时,该答案重排序模型通过公式 $\text{Loss2} = -\log(\text{Prob}_{\text{label}})$ 计算得到,然后该答案抽取模型与该答案重排序模型的整体损失为 $\text{Loss1} + \text{Loss2}$ 。

[0080] 本实施例中答案抽取指的是给定一个问题Q和一个或多个文本片段P(P1, P2, P3, ... Pn),找出问题答案A。机器阅读理解是给定一段文本Paragraph和问题Question,得到答案Answer。通常假定Answer就包含在原文中,因此机器阅读理解任务的目标就是得到一个span(start, end), start表示Answer的开始字符在Paragraph中的位置, end表示Answer的结束字符在Paragraph中的位置。根据问题和对应的短文进行自然语言理解,依据文本预测问题答案。抽取式阅读理解任务的答案为出现在原文中的连续文字,答案必须是文本中的一个范围。机器阅读理解有几种模式,即不同类型的问题,答案类型也不同,一般情况下,有三种问题:简单的问题,可以用简单的事实回答,答案通常是实体,答案简短;稍微复杂的叙述问题,答案略长;复杂的问题,通常是关于观点或意见。答案重排序是指将答案抽取中获取到的答案做一个统一的排序,最终选择分数最高的答案作为给定问题Q的Answer。

[0081] 本实施例中,该答案抽取模型和该答案重排序模型可以是GBDT、线性模型、深度模型,具体此处不做限定。

[0082] 需要说明的是,本申请实施例提供的一种多文档阅读理解方法的实施环境可以与该答案抽取模型和该答案重排序模型的训练方法的实施环境相同,也可以不同,本申请实施例对此不进行限制。

[0083] 结合上述介绍,下面将对本申请中多文档阅读理解方法进行介绍,请参阅图6,以终端设备为执行主体,以图5a所示的训练架构训练得到的答案抽取模型和答案重排序模型进行说明,本申请实施例中多文档阅读理解方法的一个实施例包括:

[0084] 601、获取目标问题、该目标问题对应的多个答复文档,并建立答案抽取模型和答案重排序模型。

[0085] 该终端设备通过其输入设备获取用户输入的目标问题,然后通过文章检索获取到该目标问题对应的多个答案文档;同时该终端设备还可以部署有答案抽取模型和答案重排序模型或者该终端设备与部署有该答案抽取模型和答案重排序模型的服务器进行相连。本实施例中,该答案抽取模型与该答案重排序模型可以采用上述图4a或图4b所示的训练方法进行训练得到。

[0086] 本实施例中,该文章检索也可以称为文本召回。其具体可以如下:即在用户通过终端的输入界面输入该目标问题,输入完成后,并向服务器发送该目标问题。服务器接收到该目标问题后,根据目标问题对待召回文本(即该答复文档)进行召回处理,得到对应的至少一个答复文档,其中,该答复文档可以是知识库中的文本,该知识库可以是用户预先在服务器中存储的数据库,也可以是未被存储至知识库中文本,例如网页上最新报道的新闻、公众号上的文本。在该过程中,该服务器在接收到该目标问题之后,可以对该目标问题进行分词

处理,得到该目标问题中的关键词,从而确定目标问题中的关键词的权重。服务器在获得目标问题中的关键词后,可以根据目标问题中的关键词以及答复文档,确定关键词与答复文档的相关度。在服务器获得关键词的权重、以及关键词与答复文档的相关度后,可以对关键词的权重、以及关键词与答复文档的相关度进行加权求和,得到目标问题与答复文档的相关度分数。在服务器得到目标问题与答复文档的相关度分数后,可以基于目标问题与答复文档的相关度分数,对答复文档进行降序排序,得到前N个答复文档,将前N个答复文档确定为答复文档,以便后续对答复文档进行相应的处理。

[0087] 602、调用该答案抽取模型获取该多个答复文档对该目标问题的预测答案集。

[0088] 该终端设备将该目标问题与多个答复文档输入该答案抽取模型得到该多个答复文档对该目标问题的预测答案集。即每一个答复文档都将抽取一个答案。一个示例性方案中,若该目标问题为“历史上谁发动了陈桥兵变”,其通过文章检索可以得到6个答复文档,每个答复文档对应的一个答案,该预测答案集的结果可以如下:其中4个答复文档的答案为“赵匡胤”,其中一个答复文档的答案为“赵光义”,其中一个答复文档的答案为“柴宗训”。

[0089] 本实施例中,在该终端设备在将该目标问题与多个答复文档输入该答案抽取模型之前,还需要进行数据预处理。即首先对目标问题和答复文档进行tokenization的处理,即将句子转为字符级的序列;之后将目标问题序列和答复文档序列相连接并以“【SEP】”分隔,并在序列的开头增加“【CLS】”,形成【CLS】目标问题【SEP】答复文档【SEP】,并在连接的序列后做padding处理。经过预处理后,输入问题+答复文档序列的长度大于答案抽取模型中的BERT网络规定的最大序列长度时,将答复文档以一定的步长分割为若干段分别与问题连接,答复文档分割段与答复文档分割段之间有一定长度的重叠部分,为了保证分割后的答复文档尽可能不削减完整答复文档的语义。假设该目标问题为“历史上谁发动了陈桥兵变”,该答复文档1为“宋太祖赵匡胤在发动陈桥兵变之后建立了北宋”,该答复文档2为“陈桥兵变为北宋开国皇帝发动政变的事件名称”。则其进行数据预处理之后得到的词序列可以如下所示,即目标问题与答复文档1的词序列为“【CLS】【历】【史】【上】【谁】【发】【动】【了】【陈】【桥】【兵】【变】【SEP】【宋】【太】【祖】【赵】【匡】【胤】【在】【发】【动】【陈】【桥】【兵】【变】【之】【后】【建】【立】【了】【北】【宋】【SEP】”;而该目标问题与答复文档2的词序列为“【CLS】【历】【史】【上】【谁】【发】【动】【了】【陈】【桥】【兵】【变】【SEP】【陈】【桥】【兵】【变】【为】【北】【宋】【开】【国】【皇】【帝】【发】【动】【政】【变】【的】【事】【件】【名】【称】【SEP】”。该终端设备在对该目标问题和该答复文档处理得到词序列之后,对该词序列输入该答案抽取模型的编码层得到该目标问题和该答复文档中各个字符对应的输出向量;并对该输出向量进行混合预测答案得到该预测答案集。可以理解的是,该预测答案的开始词为该预测答案在该词序列中的开始字符对应的token(也称为starttoken),该预测答案的结束词为该预测答案在该词序列中的结束字符对应的token(也称为endtoken)。一种示例性方案中,根据上面的答复文档1得到的预测答案为【赵】【匡】【胤】,则该预测答案的开始词为【赵】,其结束词为【胤】。

[0090] 本实施例中,该答案抽取模型可以采用自注意力(attention)机制的想法,分别训练用于预测答案的开始位置和答案的结束位置,从而抽取得到该预测答案。可以理解的是,该答案抽取模型也可采用其他方式进行预测答案的抽取,具体此处不做限定。

[0091] 603、将该预测答案集按照相同预测答案进行分类得到至少一个答案集。

[0092] 该终端设备在获取到该目标问题对应的预测答案集之后,将该预测答案集按照相

同预测答案进行分类得到至少一个答案集。如步骤602中所示的示例性方案中,若该预测答案集中包括6个预测答案,其中有4个预测答案均为“赵匡胤”,则将这四个预测答案归于一个答案集,而另外的一个“赵光义”归于一个答案集,另外的一个“柴宗训”归于一个答案集。

[0093] 604、调用该答案重排序模型获取该至少一个答案集对应的至少一个答案表征集合,该至少一个答案集中每个答案集对应一个预测答案,该至少一个答案表征集合中的答案表征为该预测答案的表征向量。

[0094] 该终端设备调用该答案重排序模型获取该目标问题与第一答复文档集中各个答复文档的第一输出向量集合,该第一答复文档集包括该至少一个答案集中第一答案集对应的答复文档;从该第一输出向量集合中选择该第一答复文档集对应的预测答案的字符串进行池化得到该第一答案集中各个预测答案的中间答案表征集合;对该中间答案表征集合进行自注意力处理得到该第一答案集对应的预测答案的第一答案表征;依此类推,遍历得到该至少一个答案集中各个答案集对应的预测答案的答案表征,并归类为该至少一个答案表征集合,该第一答案表征包含于该至少一个答案表征集合。

[0095] 具体来说,该终端设备可以该终端设备在获取得到该第一输出向量集合时可以采用如上述步骤603中的数据预处理方式,即将该目标问题的每个字符与该第一答复文档集中每个答复文档的每个字符,利用开始字符与间隔字符进行拼接得到第一词序列集合;将该第一词序列集合输入该答案重排序模型得到该第一输出向量集合;然后该终端设备从该第一输出向量集合中选择该第一答复文档集对应的预测答案的字符串进行池化得到该第一答案集中各个预测答案的中间答案表征集合;然后再对该中间答案表征集合中的每个中间答案表征进行自注意力处理得到自注意预测分数集合;对该自注意预测分数集合中进行归一化处理得到归一化预测分数集合;对该归一化预测分数集合中的各个归一化预测分数进行加权求和得到该第一答案表征。

[0096] 一种示例性方案中,假设第一答案集对应的答案为“赵匡胤”,其对应的答复文档为答复文档1和答复文档2,且该目标问题为“历史上谁发动了陈桥兵变”,该答复文档1为“宋太祖赵匡胤在发动陈桥兵变之后建立了北宋”,该答复文档2为“陈桥兵变为北宋开国皇帝发动政变的事件名称”。则其进行数据预处理之后得到的词序列可以如下所示,即目标问题与答复文档1的词序列为“【CLS】【历】【史】【上】【谁】【发】【动】【了】【陈】【桥】【兵】【变】【SEP】【宋】【太】【祖】【赵】【匡】【胤】【在】【发】【动】【陈】【桥】【兵】【变】【之】【后】【建】【立】【了】【北】【宋】【SEP】”;而该目标问题与答复文档2的词序列为“【CLS】【历】【史】【上】【谁】【发】【动】【了】【陈】【桥】【兵】【变】【SEP】【陈】【桥】【兵】【变】【为】【北】【宋】【开】【国】【皇】【帝】【发】【动】【政】【变】【的】【事】【件】【名】【称】【SEP】”。若该第二答案集对应的答案为“赵光义”,该答复文档对应的答复文档3,且该目标问题为“历史上谁发动了陈桥兵变”,该答复文档3为“赵光义和其兄在发动陈桥兵变之后建立了北宋”,则该目标问题与答复文档3的词序列为“【CLS】【历】【史】【上】【谁】【发】【动】【了】【陈】【桥】【兵】【变】【SEP】【赵】【光】【义】【和】【其】【兄】【在】【发】【动】【陈】【桥】【兵】【变】【之】【后】【建】【立】【了】【北】【宋】【SEP】”。若该第三答案集对应的答案为“柴宗训”,该答复文档对应的答复文档4,且该目标问题为“历史上谁发动了陈桥兵变”,该答复文档4为“柴宗训在陈桥兵变之后沦为亡国之君”,则该目标问题与答复文档3的词序列为“【CLS】【历】【史】【上】【谁】【发】【动】【了】【陈】【桥】【兵】【变】【SEP】”

【柴】【宗】【训】【在】【陈】【桥】【兵】【变】【之】【后】【沦】【为】【亡】【国】【之】【君】【SEP】”。然后再将上述词序列输入该答案重排序模型的编码层得到该各个词序列的输出向量集合。本实施例中为了方便该终端设备从该输出向量集合中获取到该预测答案,可以在进行数据预处理时,对该答案抽取模型抽取出的预测答案进行标注处理。本实施例中以该第一答案集为“赵匡胤”进行说明,则该终端设备再将第一答案集中的各个词序列中的“赵匡胤”截取出来,并进行池化处理得到第一答案集对应的答案的中间答案表征。

[0097] 本实施例中,该池化处理可以采用公式1:

[0098] 该公式1为: $V_i = \text{avg\_pooling}(V_t, V_{t+1}, \dots, V_{t+m})$ ;其中,该 $V_t, V_{t+1}, \dots, V_{t+m}$ 用于指示某一个答复文档中预测答案对应的输出向量序列,比如答复文档1中的“赵匡胤”对应的输出向量 $(V_t, V_{t+1}, V_{t+2})$ ;该 $V_i$ 用于指示第i个答复文档的中间答案表征。在第一答案集中该终端设备将获取到2个中间答案表征,即答案文档1中的“赵匡胤”对应的中间答案表征 $V_1$ ,答案文档2中的“赵匡胤”对应的中间答案表征 $V_2$ 。

[0099] 然后该终端设备在得到该中间答案表征之后,利用该自注意力处理机制得到该第一答案集对应的预测答案的答案表征。本实施例中,该终端设备采用自注意力处理机制可以通过公式2至公式4:

[0100] 该公式2为: $S_i = V^T \tanh(WV_i + B)$ ;

[0101] 该公式3为: $a_k = \frac{S_i}{S_1 + S_2 + \dots + S_i}$ ;

[0102] 该公式4为: $V_{\text{answer}} = a_1 V_1 + a_2 V_2 + \dots + a_k V_i$ ;

[0103] 其中,该公式2中的 $S_i$ 用于指示每一个中间答案表征对应的自注意分数,该 $V^T$ 用于指示对参数矩阵 $V$ 求秩,该 $W$ 为一个参数矩阵;该 $B$ 为参数向量。该公式3中的 $a_k$ 用于指示对该中间答案表征对应的自注意分数进行归一化处理得到归一化的自注意分数;该公式4中的 $V_{\text{answer}}$ 用于指示该答案集对应的预测答案的答案表征。一种示例性方案中,答复文档1中的“赵匡胤”对应的输出向量 $(V_t, V_{t+1}, V_{t+2})$ ;该 $V_i$ 用于指示第i个答复文档的中间答案表征。在第一答案集中该终端设备将获取到2个中间答案表征,即答案文档1中的“赵匡胤”对应的中间答案表征 $V_1$ ,答案文档2中的“赵匡胤”对应的中间答案表征 $V_2$ 。而通过该自注意力处理机制处理后得到该答案文档1中的“赵匡胤”对应的自注意力分数为80,该答案文档2中的“赵匡胤”对应的自注意力分数为80;则该答案文档1中的“赵匡胤”对应的归一化自注意力分数为0.5,该答案文档2中的“赵匡胤”对应的归一化自注意力分数为0.5;其该第一答案集中对应的“赵匡胤”的答案表征为“ $0.5V_1 + 0.5V_2$ ”。

[0104] 依此类推,该终端设备对于其他答案也进行同样的处理,从而得到该目标问题对应的多个答案表征集合。

[0105] 605、调用该答案重排序模型获取该答案表征集合对应的预测答案的预测分数。

[0106] 该终端设备将该答案表征集合中的各个答案表征输入该答案重排序模型得到各类预测答案的预测分数。

[0107] 本实施例中,该终端设备根据该答案表征计算得到预测分数的可采用如下公式5:

[0108] 公式5为: $\text{logits} = WV_{\text{answer}}$ ;其中,该logits用于指示该答案表征对应的预测答案的预测分数。

[0109] 可以理解的是,该终端设备在训练该答案重排序模型的过程中,还可以根据该预

测答案获取相应的损失值,具体可以采用如下公式6和公式7:

[0110] 公式6为: $\text{Prob}=\text{softmax}(\text{logits})$ ;

[0111] 公式7为: $\text{Loss}=-\log(\text{Prob}_{\text{label}})$ ;

[0112] 其中,该公式6用于指示对预测分数通过激活函数得到概率值,然后公式7用于指示对该概率值进行交叉熵处理得到该答案重排序模型的损失值,而该 $\text{Prob}_{\text{label}}$ 就是标注为0或者1的一个概率值。

[0113] 606、根据该预测分数输出该目标问题对应的最终答案。

[0114] 该终端设备选择该预测分数最高的答案输出为该目标问题对应的最终答案。

[0115] 本实施例中,将多个答复文档提取的预测答案进行分类,并获取抽取相同答案的答复文档之间的语义特征,从而增加答案重排序模型中的排序效果,提升多文档阅读理解的准确性。

[0116] 下面对本申请中的多文档阅读理解装置进行详细描述,请参阅图7,图7为本申请实施例中多文档阅读理解装置的一个实施例示意图,多文档阅读理解装置20包括:

[0117] 获取模块201,用于获取目标问题、该目标问题对应的多个答复文档,并建立答案抽取模型和答案重排序模型;

[0118] 处理模块202,用于调用该答案抽取模型获取该多个答复文档对该目标问题的预测答案集;将该预测答案集按照相同预测答案进行分类得到至少一个答案集;调用该答案重排序模型获取该至少一个答案集对应的至少一个答案表征集合,该至少一个答案集中每个答案集对应一个预测答案,该至少一个答案表征集合中的答案表征为该预测答案的表征向量该;调用该答案重排序模型获取该答案表征集合对应的预测答案的预测分数;

[0119] 输出模块203,用于根据该预测分数输出该目标问题对应的最终答案。

[0120] 本申请实施例中,提供了一种多文档阅读理解装置。采用上述装置,将多个答复文档提取的预测答案进行分类,并获取抽取相同答案的答复文档之间的语义特征,从而增加答案重排序模型中的排序效果,提升多文档阅读理解的准确性。

[0121] 可选地,在上述图7所对应的实施例的基础上,本申请实施例提供的多文档阅读理解装置20的另一实施例中,

[0122] 该处理模块202,具体用于调用该答案重排序模型获取该目标问题与第一答复文档集中各个答复文档的第一输出向量集合,该第一答复文档集包括该至少一个答案集中第一答案集对应的答复文档;

[0123] 从该第一输出向量集合中选择该第一答复文档集对应的预测答案的字符串进行池化得到该第一答案集中各个预测答案的中间答案表征集合;

[0124] 对该中间答案表征集合进行自注意力处理得到该第一答案集对应的预测答案的第一答案表征;

[0125] 依此类推,遍历得到该至少一个答案集中各个答案集对应的预测答案的答案表征,并归类为该至少一个答案表征集合,该第一答案表征包含于该至少一个答案表征集合。

[0126] 本申请实施例中,提供了一种多文档阅读理解装置。采用上述装置,将答案集中的相同答案分为一个集合,并将同一个集合中的答复文档的各个词序列进行池化处理之后,再将池化后的表征向量进行自注意力处理得到该相同答案的最终答案表征,这样将各个不同答复文档的上下文都进行了处理,从而使得该最终答案表征更具有丰富的语义特征,进

而提高多文档阅读理解的准确性。

[0127] 可选地,在上述图7所对应的实施例的基础上,本申请实施例提供的多文档阅读理解装置20的另一实施例中,该处理模块202,具体用于将该目标问题的每个字符与该第一答复文档集合中每个答复文档的每个字符,利用开始字符与间隔字符进行拼接得到第一词序列集合;

[0128] 将该第一词序列集合输入该答案重排序模型得到该第一输出向量集合。

[0129] 本申请实施例中,提供了一种多文档阅读理解装置。采用上述装置,采用对各个字符进行输出向量表示,可以提升对答案的表征,从而提高多文档阅读理解的准确性。

[0130] 可选地,在上述图7所对应的实施例的基础上,本申请实施例提供的多文档阅读理解装置20的另一实施例中,该处理模块202,具体用于对该中间答案表征集合中的每个中间答案表征进行自注意力处理得到自注意预测分数集合;

[0131] 对该自注意预测分数集合中进行归一化处理得到归一化预测分数集合;

[0132] 对该归一化预测分数集合中的各个归一化预测分数进行加权求和得到该第一答案表征。

[0133] 本申请实施例中,提供了一种多文档阅读理解装置。采用上述装置,对于各个答案的预测分数进行加权求和之后,将各个不同答复文档的上下文都进行了处理,从而使得该最终答案表征更具有丰富的语义特征,进而提高多文档阅读理解的准确性。

[0134] 可选地,在上述图7所对应的实施例的基础上,本申请实施例提供的多文档阅读理解装置20的另一实施例中,

[0135] 该获取模块,还用于获取训练样本集,并建立初始答案抽取模型和初始答案重排序模型,该训练样本集包括问题样本集、该问题样本集对应的答复文档样本集;

[0136] 如图8所示,该装置还包括训练模块204,用于利用该训练样本集训练该初始答案抽取模型和该初始答案重排序模型得到该答案抽取模型和该答案重排序模型。

[0137] 本申请实施例中,提供了一种多文档阅读理解装置。采用上述装置,根据训练样本集进行模型训练,同时在训练过程中将多个答复文档提取的预测答案按类进行分类,获取相同答案的答复文档之间的语义特征,从而增加重排序网络中的排序效果,提升多文档阅读理解的准确性。

[0138] 可选地,在上述图8所对应的实施例的基础上,本申请实施例提供的多文档阅读理解装置20的另一实施例中,该初始答案抽取模型和该初始答案重排序模型共享相同的编码层,该训练模块204,具体用于调用该编码层获取该训练样本集对应的第二输出向量;

[0139] 将该第二输出向量输入该初始答案抽取模型得到第一损失值和训练答案集;

[0140] 将该训练答案集输入该初始答案重排序模型得到第二损失值;

[0141] 利用该第一损失值和该第二损失值之和反向调整该初始答案抽取模型和该初始答案重排序模型的权重参数得到该答案抽取模型和该答案重排序模型。

[0142] 本申请实施例中,提供了一种多文档阅读理解装置。采用上述装置,答案抽取模型与该答案重排序模型共享编码层,使得两个模型可以联合学习,同时提高答案抽取模型与答案重排序模型的效果。

[0143] 可选地,在上述图8所对应的实施例的基础上,本申请实施例提供的多文档阅读理解装置20的另一实施例中,该训练模块204,具体用于将该输出向量输入该初始答案抽取模

型得到该训练答案集；

[0144] 获取该输出向量的训练答案集中每个训练答案对应的开始词的预测分数和结束词的预测分数；

[0145] 利用激活函数将该开始词的预测分数转化为第一概率值,并利用该激活函数将该结束词的预测分数转化为第二概率值；

[0146] 根据该第一概率值和该第二概率值得到该第一损失值；

[0147] 该训练模块204,具体用于将该训练答案集输入该初始答案重排序模型得到训练答案集对应的训练答案表征集合；

[0148] 获取该训练答案表征集合中各个训练答案对应的训练预测分数集合；

[0149] 利用激活函数将该训练预测分数集合中的各个训练预测分数转化为第三概率值；

[0150] 利用交叉熵和该第三概率值得到该第二损失值。

[0151] 本申请实施例中,提供了一种多文档阅读理解装置。采用上述装置,答案抽取模型与该答案重排序模型共享编码层,同时提高答案抽取模型与答案重排序模型的效果。

[0152] 本申请提供的多文档阅读理解装置可用于服务器,请参阅图9,图9是本申请实施例提供的一种服务器结构示意图,该服务器300可因配置或性能不同而产生比较大的差异,可以包括一个或一个以上中央处理器(central processing units,CPU)322(例如,一个或一个以上处理器)和存储器332,一个或一个以上存储应用程序342或数据344的存储介质330(例如一个或一个以上海量存储设备)。其中,存储器332和存储介质330可以是短暂存储或持久存储。存储在存储介质330的程序可以包括一个或一个以上模块(图示没标出),每个模块可以包括对服务器中的一系列指令操作。更进一步地,中央处理器322可以设置为与存储介质330通信,在服务器300上执行存储介质330中的一系列指令操作。

[0153] 服务器300还可以包括一个或一个以上电源326,一个或一个以上有线或无线网络接口350,一个或一个以上输入输出接口358,和/或,一个或一个以上操作系统341,例如Windows Server™,Mac OS X™,Unix™,Linux™,FreeBSD™等等。

[0154] 上述实施例中由服务器所执行的步骤可以基于该图9所示的服务器结构。

[0155] 本申请提供的多文档阅读理解装置可用于终端设备,请参阅图10,为了便于说明,仅示出了与本申请实施例相关的部分,具体技术细节未揭示的,请参照本申请实施例方法部分。在本申请实施例中,以终端设备为智能手机为例进行说明:

[0156] 图10示出的是与本申请实施例提供的终端设备相关的智能手机的部分结构的框图。参考图10,智能手机包括:射频(radio frequency,RF)电路410、存储器420、输入单元430、显示单元440、传感器450、音频电路460、无线保真(wireless fidelity,WiFi)模块470、处理器480、以及电源490等部件。本领域技术人员可以理解,图10中示出的智能手机结构并不构成对智能手机的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件布置。

[0157] 下面结合图10对智能手机的各个构成部件进行具体的介绍:

[0158] RF电路410可用于收发信息或通话过程中,信号的接收和发送,特别地,将基站的下行信息接收后,给处理器480处理;另外,将设计上行的数据发送给基站。通常,RF电路410包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放大器(low noise amplifier,LNA)、双工器等。此外,RF电路410还可以通过无线通信与网络和其他设备通信。

上述无线通信可以使用任一通信标准或协议,包括但不限于全球移动通讯系统(global system of mobile communication,GSM)、通用分组无线服务(general packet radio service,GPRS)、码分多址(code division multiple access,CDMA)、宽带码分多址(wideband code division multiple access,WCDMA)、长期演进(long term evolution,LTE)、电子邮件、短消息服务(short messaging service,SMS)等。

[0159] 存储器420可用于存储软件程序以及模块,处理器480通过运行存储在存储器420的软件程序以及模块,从而执行智能手机的各种功能应用以及数据处理。存储器420可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等;存储数据区可存储根据智能手机的使用所创建的数据(比如音频数据、电话本等)等。此外,存储器420可以包括高速随机存取存储器,还可以包括非易失性存储器,例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

[0160] 输入单元430可用于接收输入的数字或字符信息,以及产生与智能手机的用户设置以及功能控制有关的键信号输入。具体地,输入单元430可包括触控面板431以及其他输入设备432。触控面板431,也称为触摸屏,可收集用户在其上或附近的触摸操作(比如用户使用手指、触笔等任何适合的物体或附件在触控面板431上或在触控面板431附近的操作),并根据预先设定的程式驱动相应的连接装置。可选的,触控面板431可包括触摸检测装置和触摸控制器两个部分。其中,触摸检测装置检测用户的触摸方位,并检测触摸操作带来的信号,将信号传送给触摸控制器;触摸控制器从触摸检测装置上接收触摸信息,并将它转换成触点坐标,再送给处理器480,并能接收处理器480发来的命令并加以执行。此外,可以采用电阻式、电容式、红外线以及表面声波等多种类型实现触控面板431。除了触控面板431,输入单元430还可以包括其他输入设备432。具体地,其他输入设备432可以包括但不限于物理键盘、功能键(比如音量控制按键、开关按键等)、轨迹球、鼠标、操作杆等中的一种或多种。

[0161] 显示单元440可用于显示由用户输入的信息或提供给用户的信息以及智能手机的各种菜单。显示单元440可包括显示面板441,可选的,可以采用液晶显示器(liquid crystal display,LCD)、有机发光二极管(organic light-emitting diode,OLED)等形式来配置显示面板441。进一步的,触控面板431可覆盖显示面板441,当触控面板431检测到在其上或附近的触摸操作后,传送给处理器480以确定触摸事件的类型,随后处理器480根据触摸事件的类型在显示面板441上提供相应的视觉输出。虽然在图10中,触控面板431与显示面板441是作为两个独立的部件来实现智能手机的输入和输入功能,但是在某些实施例中,可以将触控面板431与显示面板441集成而实现智能手机的输入和输出功能。

[0162] 智能手机还可包括至少一种传感器450,比如光传感器、运动传感器以及其他传感器。具体地,光传感器可包括环境光传感器及接近传感器,其中,环境光传感器可根据环境光线的明暗来调节显示面板441的亮度,接近传感器可在智能手机移动到耳边时,关闭显示面板441和/或背光。作为运动传感器的一种,加速计传感器可检测各个方向上(一般为三轴)加速度的大小,静止时可检测出重力的大小及方向,可用于识别智能手机姿态的应用(比如横竖屏切换、相关游戏、磁力计姿态校准)、振动识别相关功能(比如计步器、敲击)等;至于智能手机还可配置的陀螺仪、气压计、湿度计、温度计、红外线传感器等其他传感器,在此不再赘述。

[0163] 音频电路460、扬声器461,传声器462可提供用户与智能手机之间的音频接口。音频电路460可将接收到的音频数据转换后的电信号,传输到扬声器461,由扬声器461转换为声音信号输出;另一方面,传声器462将收集的声音信号转换为电信号,由音频电路460接收后转换为音频数据,再将音频数据输出处理器480处理后,经RF电路410以发送给比如另一智能手机,或者将音频数据输出至存储器420以便进一步处理。

[0164] WiFi属于短距离无线传输技术,智能手机通过WiFi模块470可以帮助用户收发电子邮件、浏览网页和访问流式媒体等,它为用户提供了无线的宽带互联网访问。虽然图10示出了WiFi模块470,但是可以理解的是,其并不属于智能手机的必须构成,完全可以根据需要在不改变发明的本质的范围内而省略。

[0165] 处理器480是智能手机的控制中心,利用各种接口和线路连接整个智能手机的各个部分,通过运行或执行存储在存储器420内的软件程序和/或模块,以及调用存储在存储器420内的数据,执行智能手机的各种功能和处理数据,从而对智能手机进行整体监测。可选的,处理器480可包括一个或多个处理单元;可选的,处理器480可集成应用处理器和调制解调处理器,其中,应用处理器主要处理操作系统、用户界面和应用程序等,调制解调处理器主要处理无线通信。可以理解的是,上述调制解调处理器也可以不集成到处理器480中。

[0166] 智能手机还包括给各个部件供电的电源490(比如电池),可选的,电源可以通过电源管理系统与处理器480逻辑相连,从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。

[0167] 尽管未示出,智能手机还可以包括摄像头、蓝牙模块等,在此不再赘述。

[0168] 上述实施例中由终端设备所执行的步骤可以基于该图10所示的终端设备结构。

[0169] 本申请实施例中还提供一种计算机可读存储介质,该计算机可读存储介质中存储有计算机程序,当其在计算机上运行时,使得计算机执行如前述各个实施例描述的方法。

[0170] 本申请实施例中还提供一种包括程序的计算机程序产品,当其在计算机上运行时,使得计算机执行前述各个实施例描述的方法。

[0171] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统,装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0172] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统,装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0173] 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单元来实现本实施例方案的目的。

[0174] 另外,在本申请各个实施例中的各功能单元可以集成在一个处理单元中,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单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。

[0175] 所述集成的单元如果以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,本申请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全部或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)执行本申请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(read-only memory, ROM)、随机存取存储器(random access memory, RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0176] 以上所述,以上实施例仅用以说明本申请的技术方案,而非对其限制;尽管参照前述实施例对本申请进行了详细的说明,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申请各实施例技术方案的精神和范围。

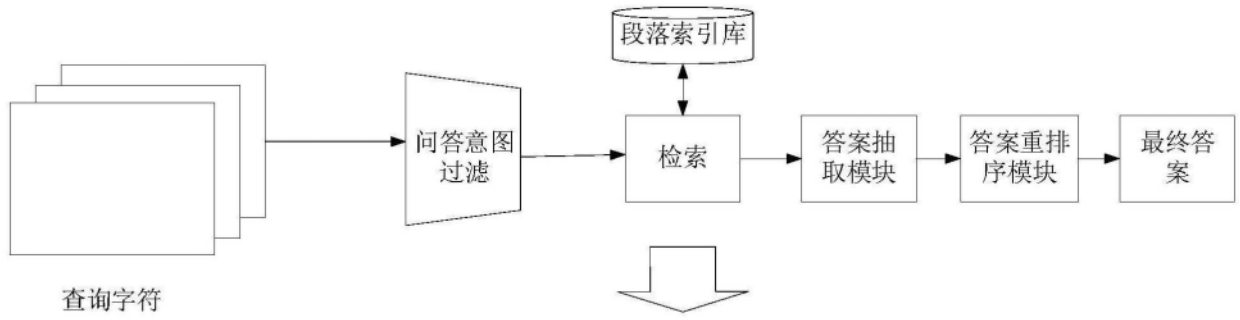


图1

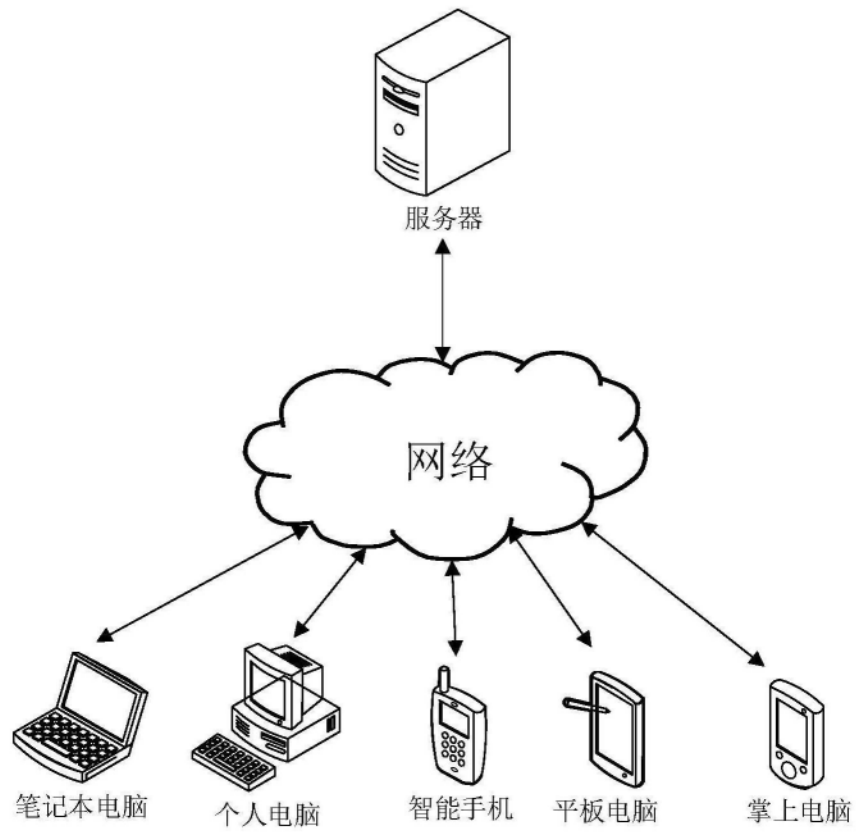


图2

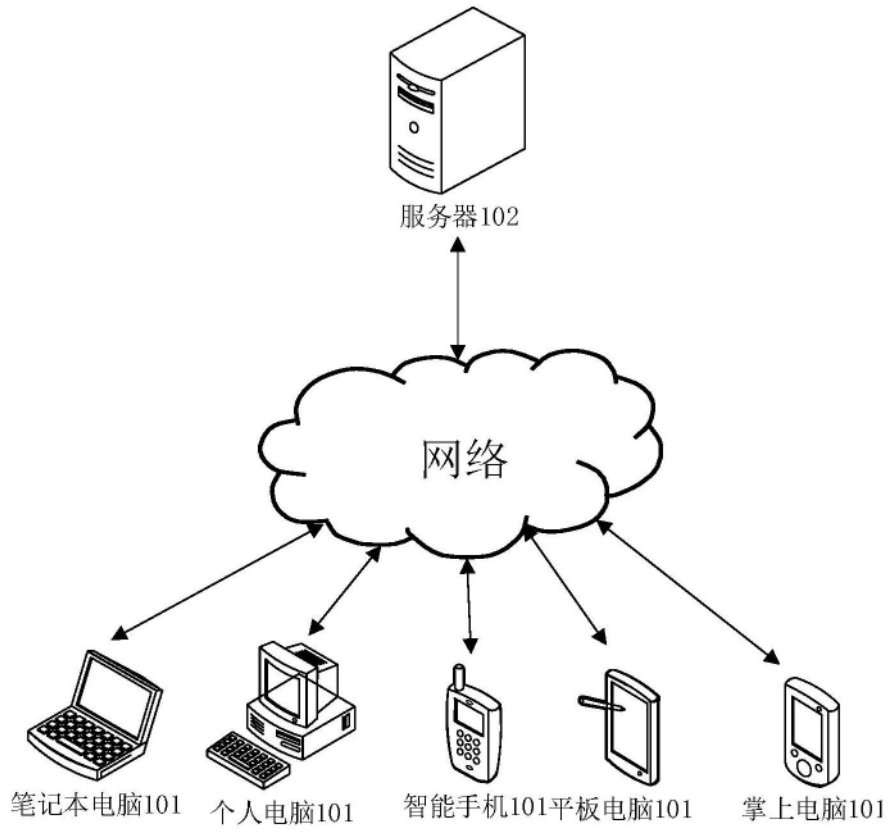


图3

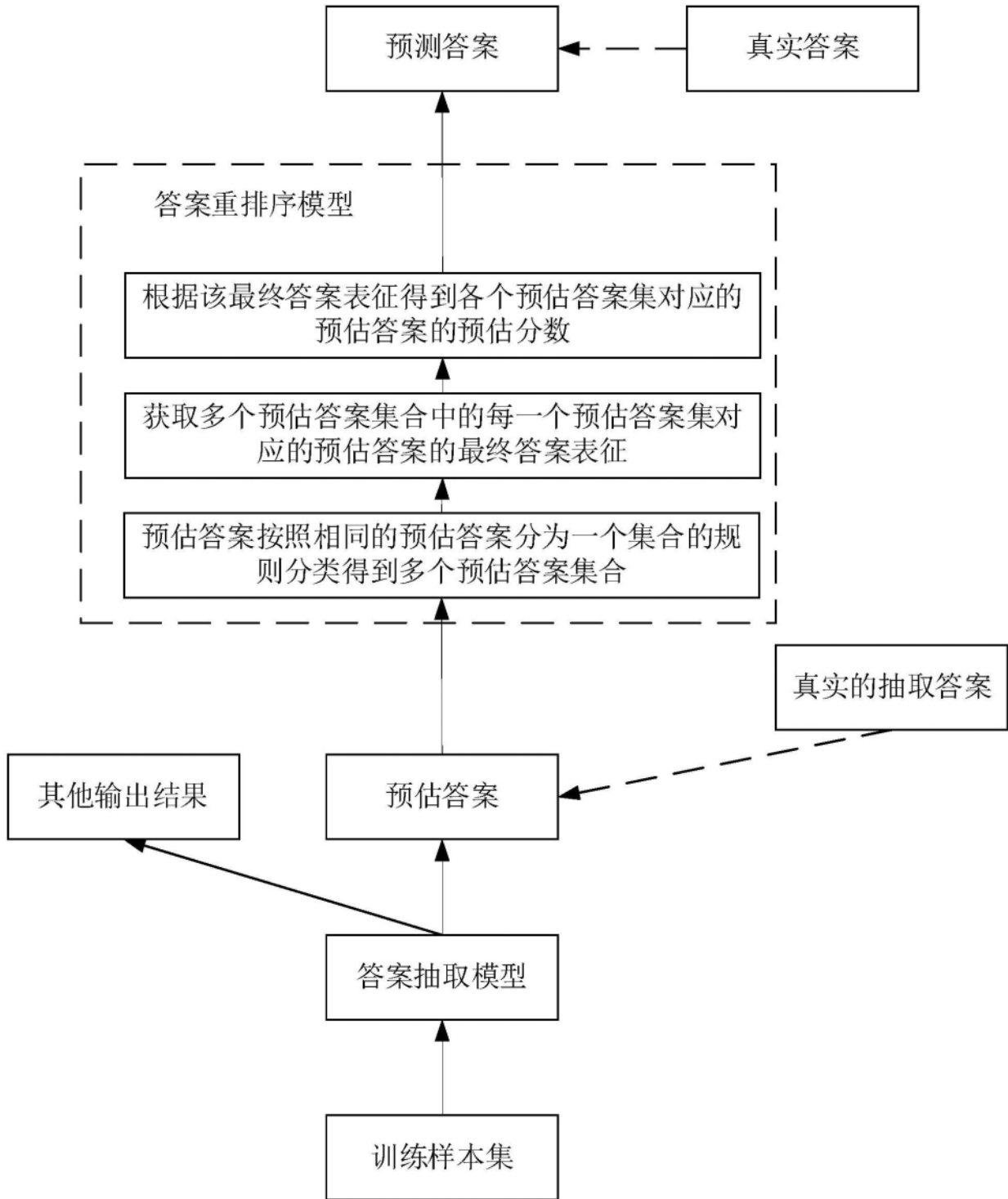


图4a

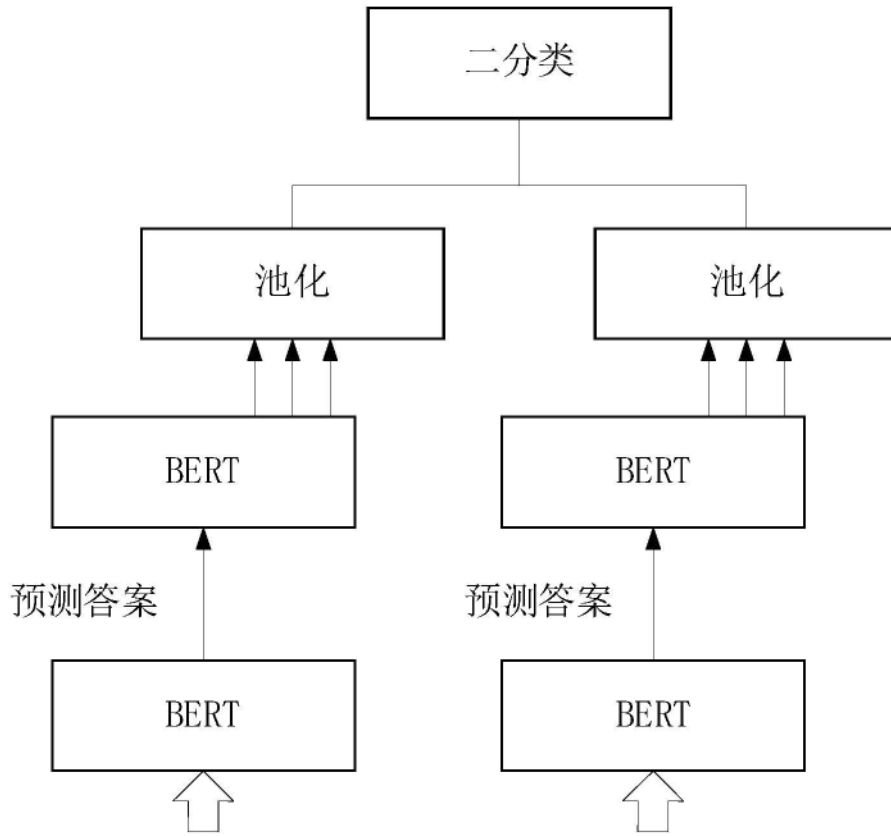


图4b

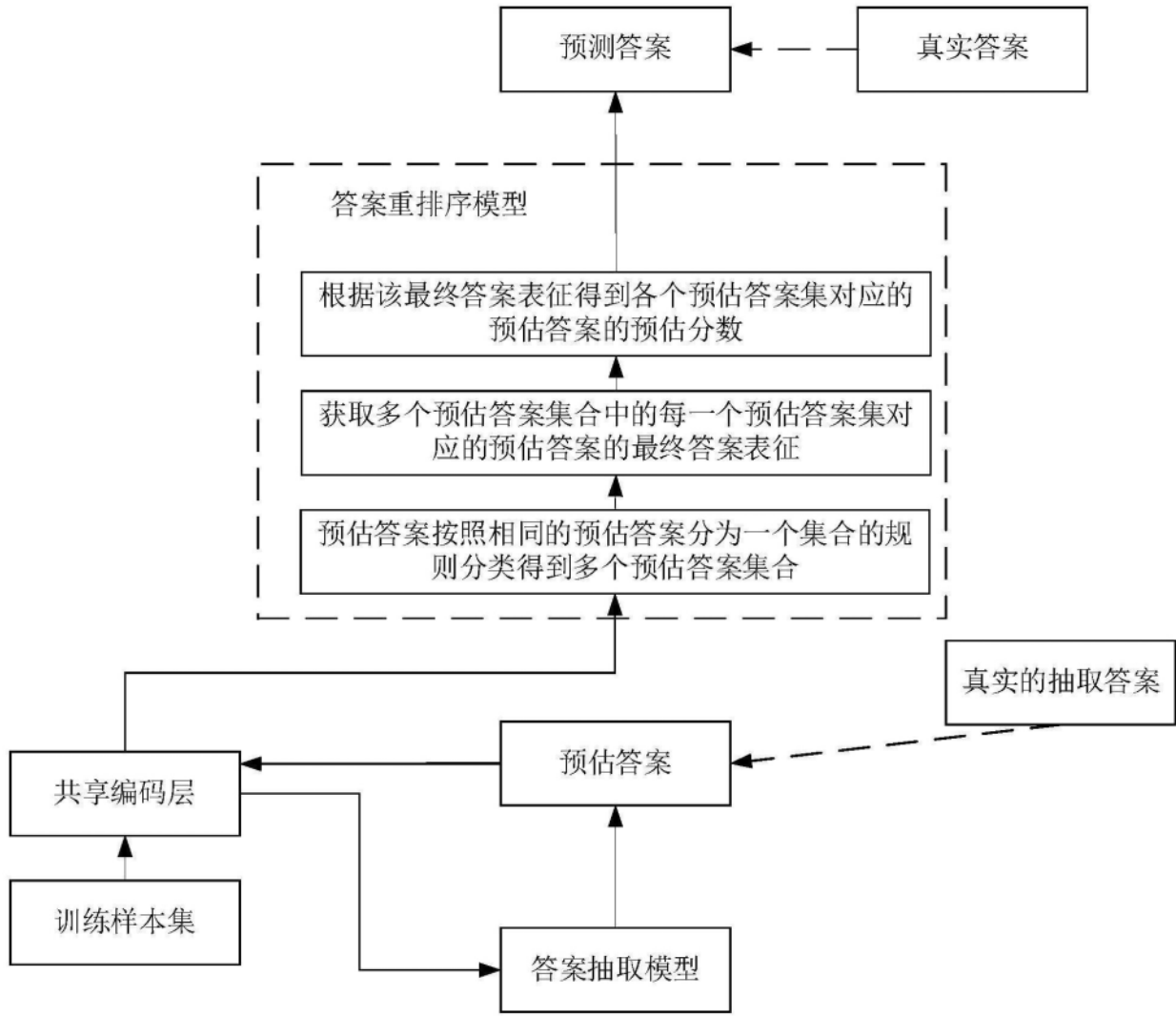


图5a

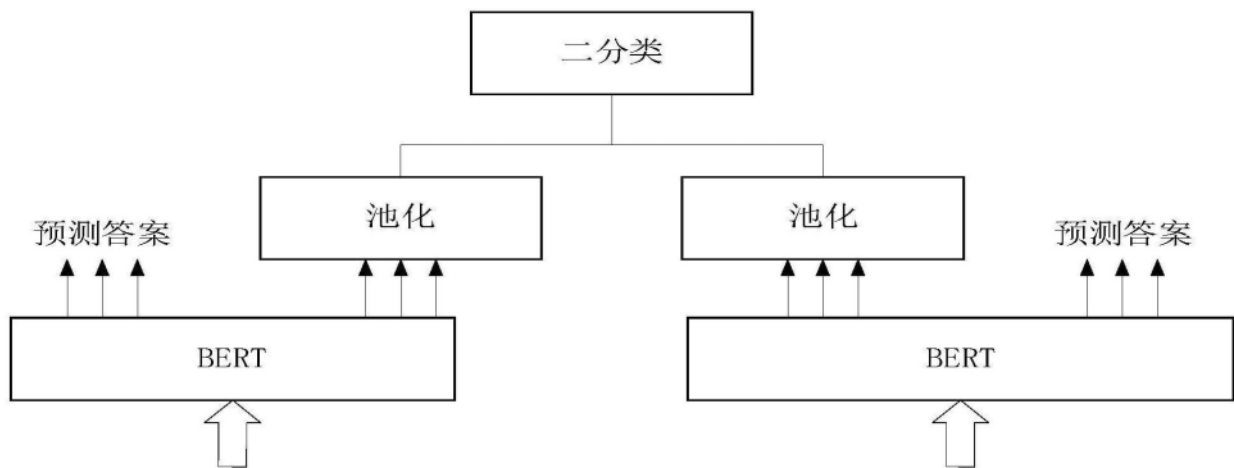


图5b

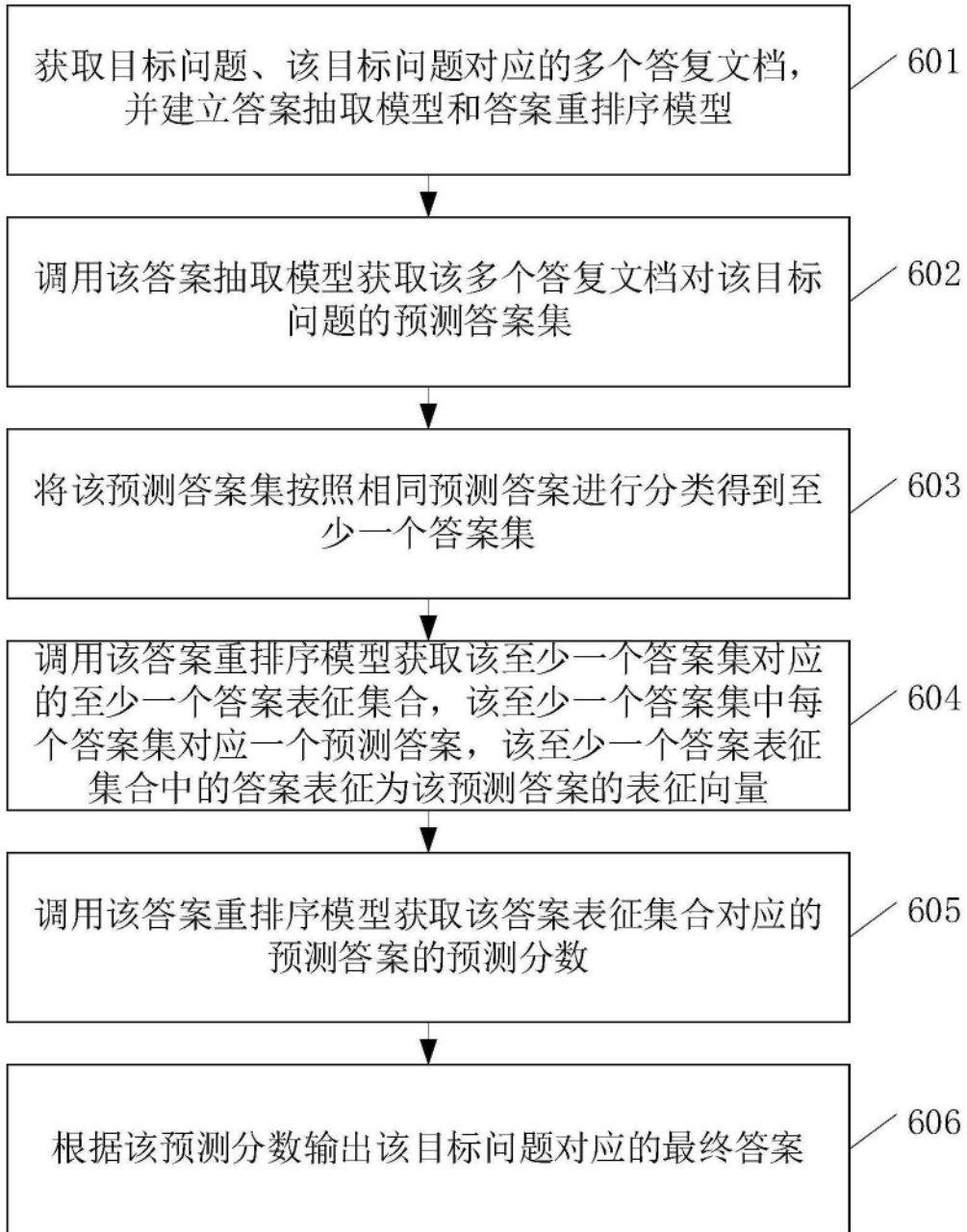


图6

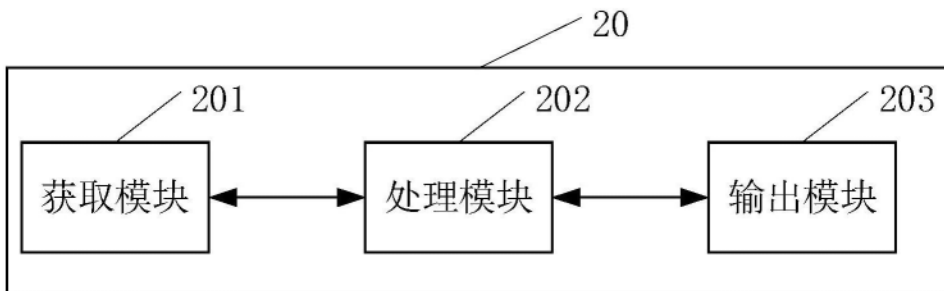


图7

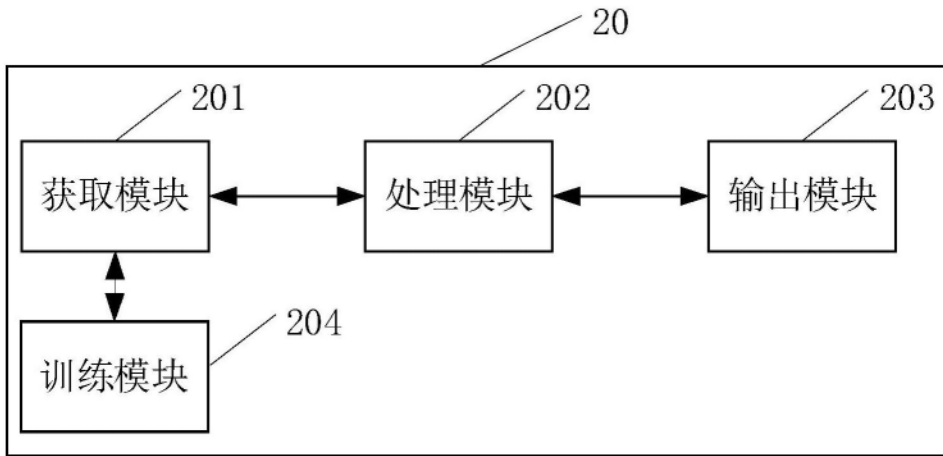


图8

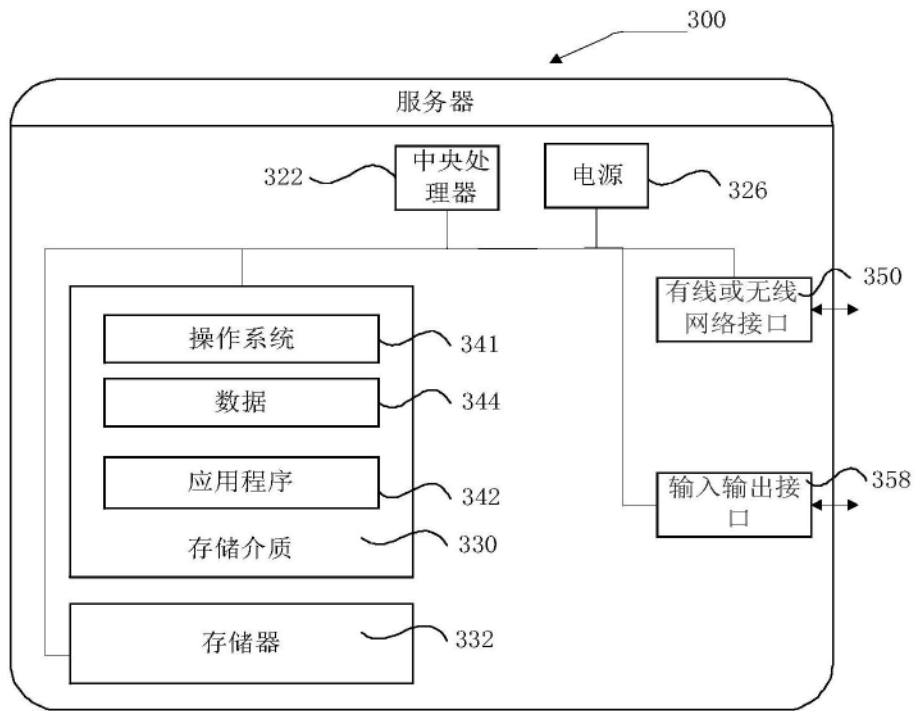


图9

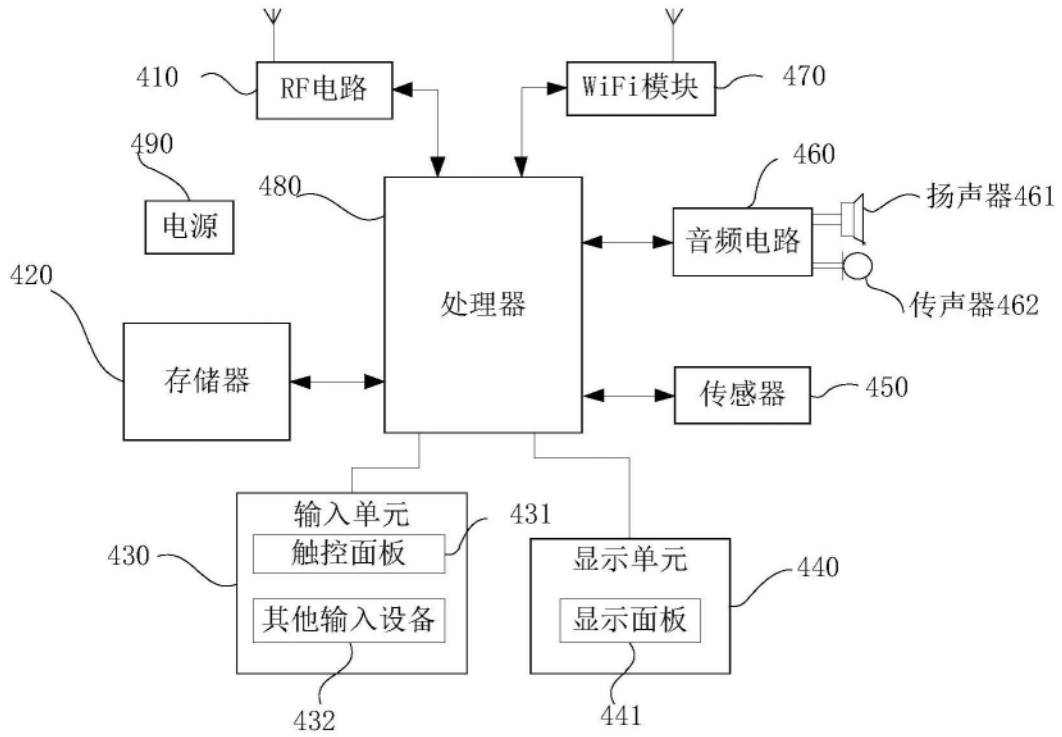


图10