



(21) 申请号 202110491454.3

G06F 16/35 (2019.01)

(22) 申请日 2021.05.06

(56) 对比文件

(65) 同一申请的已公布的文献号

CN 108804567 A, 2018.11.13

申请公布号 CN 113139039 A

CN 109446302 A, 2019.03.08

(43) 申请公布日 2021.07.20

审查员 郭坚

(73) 专利权人 京东科技控股股份有限公司

地址 100176 北京市大兴区经济技术开发区

区科创十一街18号C座2层221室

(72) 发明人 宋阳 陈蒙

(74) 专利代理机构 北京英赛嘉华知识产权代理

有限责任公司 11204

专利代理师 王达佐 马晓亚

(51) Int. Cl.

G06F 16/33 (2019.01)

G06F 16/332 (2019.01)

权利要求书3页 说明书11页 附图7页

(54) 发明名称

对话数据处理方法、装置、设备以及存储介

质

(57) 摘要

本申请公开了对话数据处理方法和装置,具体实施方案为:获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复;对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和标准答复;将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,其中,信息确定模型用于表征基于每个问题对应的问题答复、各个分类标识对应的各个问题答复和各个分类标识,确定输入的每个问题与各个标准答复之间的对应结果。该方案通过在训练信息确定模型时,将问题和答案之间的相关性考虑其中,提高了答案召回的效率。

100

获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复

101

对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复

102

将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型

103

1. 一种对话数据处理方法,所述方法包括:

获取对话数据,并对所述对话数据进行解析,得到所述对话数据对应的各个问题和各个问题对应的问题答复;

对各个所述问题答复进行分析,得到各个所述问题答复对应的各个分类标识、每个所述分类标识对应的各个问题答复和每个所述分类标识对应的标准答复,其中,所述各个分类标识基于对各个所述问题答复进行分类而得到,所述标准答复基于所述分类标识对应的各个问题答复而确定;

将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,其中,所述信息确定模型用于表征基于每个问题对应的所述问题答复、各个所述分类标识对应的各个问题答复和各个所述分类标识,确定输入的每个问题与各个所述标准答复之间的对应结果;

其中,所述信息确定模型包括:第一确定子模型和第二确定子模型,所述第一确定子模型用于表征基于每个问题对应的所述问题答复和各个所述分类标识对应的各个问题答复,确定输入的每个问题与各个所述分类标识之间的对应结果,所述第二确定子模型用于表征基于所述第一确定子模型输出每个问题对应的所述分类标识,确定输入的每个问题与各个所述标准答复之间的对应结果;

所述将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,包括:

将各个问题作为输入数据,将与输入的各个问题所对应的分类标识作为输出数据,训练得到所述第一确定子模型;

将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到所述第二确定子模型;

基于训练完成的所述第一确定子模型和训练完成的所述第二确定子模型,确定所述信息确定模型。

2. 根据权利要求1所述的方法,其中,所述对各个所述问题答复进行分析,得到各个所述问题答复对应的各个分类标识、每个所述分类标识对应的各个问题答复和每个所述分类标识对应的标准答复,包括:

将各个所述问题答复进行向量化;

对向量化后的各个所述问题答复进行聚类,得到各个所述问题答复对应的各个分类标识和每个所述分类标识对应的各个问题答复,其中,所述聚类用于表征利用聚类算法将内容相近的各个所述问题答复进行聚合;

对每个所述分类标识对应的各个问题答复进行选取,得到相应所述分类标识对应的标准答复。

3. 根据权利要求1所述的方法,在所述对各个所述问题答复进行分析,得到各个所述问题答复对应的各个分类标识、每个所述分类标识对应的各个问题答复和每个所述分类标识对应的标准答复之前,还包括:

对各个所述问题答复进行数据处理,得到处理后的各个所述问题答复,其中,所述数据处理包括:无效数据修正、数据合并、数据去重和数据归一化中的至少一项。

4. 一种用于生成信息的方法,所述方法包括:

获取用户发送的问题；

将所述问题输入至预先训练的信息确定模型,生成所述问题对应的标准答复,其中,所述信息确定模型通过如权利要求1-3之一所述的方法训练得到；

其中,所述信息确定模型包括标识确定子模型和答复确定子模型,所述标识确定子模型用于确定所述问题与分类标识之间的对应结果,所述答复确定子模型用于表征基于所述标识确定子模型输出的所述问题的分类标识,确定所述问题与标准答复之间的对应结果；

所述将所述问题输入至预先训练的信息确定模型,生成所述问题对应的标准答复,包括：

将所述问题输入至预先训练的所述标识确定子模型,生成所述问题对应的分类标识；

将生成的所述问题对应的分类标识输入至预先训练的答复确定子模型,生成所述问题对应的标准答复。

5.根据权利要求4所述的方法,还包括：

利用打分方法,对所述问题对应的标准答复进行打分,得到所述问题对应的标准答复的分值；

基于所述问题对应的各个标准答复的分值排序,确定所述问题的最终答复；

将所述最终答复信息发送至所述用户。

6.一种对话数据处理装置,所述装置包括：

获取单元,被配置成获取对话数据,并对所述对话数据进行解析,得到所述对话数据对应的各个问题和各个所述问题对应的问题答复；

分析单元,被配置成对各个所述问题答复进行分析,得到各个所述问题答复对应的各个分类标识、每个所述分类标识对应的各个问题答复和每个所述分类标识对应的标准答复,其中,所述各个分类标识基于对各个所述问题答复进行分类而得到,所述标准答复基于所述分类标识对应的各个问题答复而确定；

训练单元,被配置成将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,其中,所述信息确定模型用于表征基于每个问题对应的所述问题答复、各个所述分类标识对应的各个问题答复和各个所述分类标识,确定输入的每个问题与各个所述标准答复之间的对应结果；

其中,所述训练单元中的所述信息确定模型包括:第一确定子模型和第二确定子模型,所述第一确定子模型用于表征基于每个问题对应的所述问题答复和各个所述分类标识对应的各个问题答复,确定输入的每个问题与各个所述分类标识之间的对应结果,所述第二确定子模型用于表征基于所述第一确定子模型输出每个问题对应的所述分类标识,确定输入的每个问题与各个所述标准答复之间的对应结果；

所述训练单元,包括：

第一训练模块,被配置成将各个问题作为输入数据,将与输入的各个问题所对应的分类标识作为输出数据,训练得到所述第一确定子模型；

第二训练模块,被配置成将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到所述第二确定子模型；

确定模块,被配置成基于训练完成的所述第一确定子模型和训练完成的所述第二确定子模型,确定所述信息确定模型。

7. 根据权利要求6所述的装置,其中,所述分析单元,包括:

向量化模块,被配置成将各个所述问题答复进行向量化;

聚类模块,被配置成对向量化后的各个所述问题答复进行聚类,得到各个所述问题答复对应的各个分类标识和每个所述分类标识对应的各个问题答复,其中,所述聚类用于表征利用聚类算法将内容相近的各个所述问题答复进行聚合;

选取模块,被配置成对每个所述分类标识对应的各个问题答复进行选取,得到相应所述分类标识对应的标准答复。

8. 根据权利要求6所述的装置,还包括:

处理单元,被配置成对各个所述问题答复进行数据处理,得到处理后的各个所述问题答复,其中,所述数据处理包括:无效数据修正、数据合并、数据去重和数据归一化中的至少一项。

9. 一种用于生成信息的装置,所述装置包括:

问题获取单元,被配置成获取用户发送的问题;

信息生成单元,被配置成将所述问题输入至预先训练的信息确定模型,生成所述问题对应的标准答复,其中,所述信息确定模型通过如权利要求1-3之一所述的方法训练得到;

其中,所述信息生成单元中的所述信息确定模型包括标识确定子模型和答复确定子模型,所述标识确定子模型用于确定所述问题与分类标识之间的对应结果,所述答复确定子模型用于表征基于所述标识确定子模型输出的所述问题的分类标识,确定所述问题与标准答复之间的对应结果;

所述信息生成单元,包括:

第一生成模块,被配置成将所述问题输入至预先训练的所述标识确定子模型,生成所述问题对应的分类标识;

第二生成模块,被配置成将生成的所述问题对应的分类标识输入至预先训练的答复确定子模型,生成所述问题对应的标准答复。

10. 根据权利要求9所述的装置,还包括:

评分单元,被配置成利用评分方法,对所述问题对应的标准答复进行打分,得到所述问题对应的标准答复的分值;

确定单元,被配置成基于所述问题对应的各个标准答复的分值排序,确定所述问题的最终答复;

发送单元,被配置成将所述最终答复信息发送至所述用户。

11. 一种电子设备,其特征在于,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-3或权利要求4-5中任一项所述的方法。

12. 一种存储有计算机指令的非瞬时计算机可读存储介质,其特征在于,所述计算机指令用于使所述计算机执行权利要求1-3或权利要求4-5中任一项所述的方法。

对话数据处理方法、装置、设备以及存储介质

技术领域

[0001] 本申请的实施例涉及计算机技术领域,具体涉及人工智能技术领域,尤其涉及对话数据处理方法和装置。

背景技术

[0002] 目前工业界主流自动问答系统为检索式问答系统,检索式问答系统主要包含两个模块,答案召回模块和答案匹配模块。当用户输入一个问题后,检索式问答系统首先会通过答案召回模块从问答知识库当中召回n个候选答案,然后通过问题答案匹配模块进行打分排序得到最终答案并返回给用户。一个高质量的检索式问答系统非常依赖于高质量的问答知识库,目前工业界构建检索式问答系统知识库主要通过人工编辑和基于对话日志挖掘两种方式。基于人工编辑构建问答知识库的速度太慢,只适用于一些小规模领域的问答场景,而基于对话日志挖掘构建的大规模检索式问答系统知识库通过对问题建立索引,当用户输入一个问题后,通过倒排索引或语义索引等方式召回候选问答对,其中只考虑问题和问题之间的相关性,没有考虑问题和答案之间的相关性。

发明内容

[0003] 本申请提供了一种对话数据处理方法、装置、设备以及存储介质和一种用于生成信息的方法、装置、设备以及存储介质。

[0004] 根据本申请的第一方面,提供了一种对话数据处理方法,该方法包括:获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复;对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复,其中,各个分类标识基于对各个问题答复进行分类而得到,标准答复基于分类标识对应的各个问题答复而确定;将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,其中,信息确定模型用于表征基于每个问题对应的问题答复、各个分类标识对应的各个问题答复和各个分类标识,确定输入的每个问题与各个标准答复之间的对应结果。

[0005] 在一些实施例中,对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复,包括:将各个问题答复进行向量化;对向量化后的各个问题答复进行聚类,得到各个问题答复对应的各个分类标识和每个分类标识对应的各个问题答复,其中,聚类用于表征利用聚类算法将内容相近的各个问题答复进行聚合;对每个分类标识对应的各个问题答复进行选取,得到相应分类标识对应的标准答复。

[0006] 在一些实施例中,信息确定模型包括:第一确定子模型和第二确定子模型,第一确定子模型用于表征基于每个问题对应的问题答复和各个分类标识对应的各个问题答复,确定输入的每个问题与各个分类标识之间的对应结果,第二确定子模型用于表征基于第一确定子模型输出每个问题对应的分类标识,确定输入的每个问题与各个标准答复之间的对应

结果;将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,包括:将各个问题作为输入数据,将与输入的各个问题所对应的分类标识作为输出数据,训练得到第一确定子模型;将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到第二确定子模型;基于训练完成的第一确定子模型和训练完成的第二确定子模型,确定信息确定模型。

[0007] 在一些实施例中,在对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复之前,还包括:对各个问题答复进行数据处理,得到处理后的各个问题答复,其中,数据处理包括:无效数据修正、数据合并、数据去重和数据归一化中的至少一项。

[0008] 根据本申请的第二方面,提供了一种用于生成信息的方法,方法包括:获取用户发送的问题;将问题输入至预先训练的信息确定模型,生成问题对应的标准答复,其中,信息确定模型通过如上述对话数据处理方法中任一实施例的方法训练得到。

[0009] 在一些实施例中,信息确定模型包括标识确定子模型和答复确定子模型,标识确定子模型用于确定问题与分类标识之间的对应结果,答复确定子模型用于表征基于标识确定子模型输出的问题的分类标识,确定问题与标准答复之间的对应结果;将问题输入至预先训练的信息确定模型,生成问题对应的标准答复,包括:将问题输入至预先训练的标识确定子模型,生成问题对应的分类标识;将生成的问题对应的分类标识输入至预先训练的答复确定子模型,生成问题对应的标准答复。

[0010] 在一些实施例中,方法还包括:利用打分方法,对问题对应的标准答复进行打分,得到问题对应的标准答复的分值;基于问题对应的各个标准答复的分值排序,确定问题的最终答复;将最终答复信息发送至用户。

[0011] 根据本申请的第三方面,提供了一种对话数据处理装置,装置包括:获取单元,被配置成获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复;分析单元,被配置成对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复,其中,各个分类标识基于对各个问题答复进行分类而得到,标准答复基于分类标识对应的各个问题答复而确定;训练单元,被配置成将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,其中,信息确定模型用于表征基于每个问题对应的问题答复、各个分类标识对应的各个问题答复和各个分类标识,确定输入的每个问题与各个标准答复之间的对应结果。

[0012] 在一些实施例中,分析单元,包括:向量化模块,被配置成将各个问题答复进行向量化;聚类模块,被配置成对向量化后的各个问题答复进行聚类,得到各个问题答复对应的各个分类标识和每个分类标识对应的各个问题答复,其中,聚类用于表征利用聚类算法将内容相近的各个问题答复进行聚合;选取模块,被配置成对每个分类标识对应的各个问题答复进行选取,得到相应分类标识对应的标准答复。

[0013] 在一些实施例中,训练单元中的信息确定模型包括:第一确定子模型和第二确定子模型,第一确定子模型用于表征基于每个问题对应的问题答复和各个分类标识对应的各个问题答复,确定输入的每个问题与各个分类标识之间的对应结果,第二确定子模型用于表征基于第一确定子模型输出每个问题对应的分类标识,确定输入的每个问题与各个标准

答复之间的对应结果;训练单元,包括:第一训练模块,被配置成将各个问题作为输入数据,将与输入的各个问题所对应的分类标识作为输出数据,训练得到第一确定子模型;第二训练模块,被配置成将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到第二确定子模型;确定模块,被配置成基于训练完成的第一确定子模型和训练完成的第二确定子模型,确定信息确定模型。

[0014] 在一些实施例中,装置还包括:处理单元,被配置成对各个问题答复进行数据处理,得到处理后的各个问题答复,其中,数据处理包括:无效数据修正、数据合并、数据去重和数据归一化中的至少一项。

[0015] 根据本申请的第四方面,提供了一种用于生成信息的装置,装置包括:问题获取单元,被配置成获取用户发送的问题;信息生成单元,被配置成将问题输入至预先训练的信息确定模型,生成问题对应的标准答复,其中,信息确定模型通过如上述对话数据处理方法中任一实施例的方法训练得到。

[0016] 在一些实施例中,信息生成单元中的信息确定模型包括标识确定子模型和答复确定子模型,标识确定子模型用于确定问题与分类标识之间的对应结果,答复确定子模型用于表征基于标识确定子模型输出的问题的分类标识,确定问题与标准答复之间的对应结果;信息生成单元,包括:第一生成模块,被配置成将问题输入至预先训练的标识确定子模型,生成问题对应的分类标识;第二生成模块,被配置成将生成的问题对应的分类标识输入至预先训练的答复确定子模型,生成问题对应的标准答复。

[0017] 在一些实施例中,装置还包括:评分单元,被配置成利用评分方法,对问题对应的标准答复进行打分,得到问题对应的标准答复的分值;确定单元,被配置成基于问题对应的各个标准答复的分值排序,确定问题的最终答复;发送单元,被配置成将最终答复信息发送至用户。

[0018] 根据本申请的第五方面,提供了一种电子设备,包括:至少一个处理器;以及与至少一个处理器通信连接的存储器;其中,存储器存储有可被至少一个处理器执行的指令,指令被至少一个处理器执行,以使至少一个处理器能够执行如第一方面或第二方面中任一实施方式描述的方法。

[0019] 根据本申请的第六方面,本申请提供了一种存储有计算机指令的非瞬时计算机可读存储介质,其特征在于,计算机指令用于使计算机执行如第一方面或第二方面中任一实施方式描述的方法。

[0020] 根据本申请的技术采用获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复,对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复,将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,通过在训练信息确定模型时,将问题和答案之间的相关性考虑其中,解决了因为挖掘出的问答对过多对索引系统造成的压力。方法将QA挖掘和答案召回合并为一个先聚类后分类的过程,极大的减少了内存的使用,提高了答案召回的效率。

[0021] 应当理解,本部分所描述的内容并非旨在标识本申请的实施例的关键或重要特征,也不用于限制本申请的范围。本申请的其它特征将通过以下的说明书而变得容易理解。

附图说明

- [0022] 附图用于更好地理解本方案,不构成对本申请的限定。
- [0023] 图1是根据本申请的对话数据处理方法的第一实施例的示意图;
- [0024] 图2-1是根据本申请的对话数据处理方法的对话数据格式的示意图;
- [0025] 图2-2是根据本申请的对话数据处理方法的数据分析结果的第一示意图;
- [0026] 图2-3是根据本申请的对话数据处理方法的数据分析结果的第二示意图;
- [0027] 图2-4是根据本申请的对话数据处理方法的模型训练结果的示意图;
- [0028] 图3是可以实现本申请实施例的对话数据处理方法的场景图;
- [0029] 图4是根据本申请的对话数据处理方法的第二实施例的示意图;
- [0030] 图5是根据本申请的用于生成信息的方法的第一实施例的示意图;
- [0031] 图6是根据本申请的对话数据处理装置的一个实施例的结构示意图;
- [0032] 图7是根据本申请的用于生成信息的装置的一个实施例的结构示意图;
- [0033] 图8是用来实现本申请实施例的电子设备的框图。

具体实施方式

[0034] 以下结合附图对本申请的示范性实施例做出说明,其中包括本申请实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本申请的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0035] 需要说明的是,在不冲突的情况下,本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0036] 图1示出了根据本申请的对话数据处理方法的第一实施例的示意图100。该对话数据处理方法,包括以下步骤:

[0037] 步骤101,获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复。

[0038] 在本实施例中,执行主体(例如服务器或终端设备)可以通过有线连接或无线连接的方式从其他电子设备或者本地获取对话数据,并基于语义相似度对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复。对话数据可以用于指示用户与系统或客服人员的问答信息,对话数据的数据格式可以以会话为单位,例如图2-1所示。各个问题和各个问题答复可以以字符形式的文本、二进制数据等各种格式表示。各个问题和各个问题答复可以基于对对话数据中的问题和问题答复进行标识而得到。例如,问题是“在么”,问题答复是“您好,在的哦,请问有什么可以帮您?”。

[0039] 需要说明的是,上述无线连接方式可以包括但不限于3G、4G、5G连接、WiFi连接、蓝牙连接、WiMAX连接、Zigbee连接、UWB(ultra wideband)连接、以及其他现在已知或将来开发的无线连接方式。

[0040] 步骤102,对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复。

[0041] 在本实施例中,执行主体可以利用预设的分析法对步骤101中获取到的各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题

答复和每个分类标识对应的标准答复,例如图2-2和图2-3所示。各个分类标识和每个分类标识对应的各个问题答复基于对各个问题答复进行分类而得到,标准答复利用标准答复确定方法基于分类标识对应的各个问题答复,确定每个分类标识对应的标准答复。标准答复用于表征同类问题答复中作为标准的唯一的问题答复。标准答复可以基于每个分类标识对应的各个问题答复通过确定算法计算得到,也可以通过对每个分类标识对应的各个问题答复进行选取而得到,本实施例不对此方面内容做任何限定。

[0042] 在本实施例的一些可选的实现方式中,在对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复之前,还包括:对各个问题答复进行数据处理,得到处理后的各个问题答复,其中,数据处理包括:无效数据修正、数据合并、数据去重和数据归一化中的至少一项。很多相同内容的问题答案因为多一两个字就会变成不同答案,对答案归一化后,相同内容答案只保留一个使用频率最高的,减少了问题答复的重复数量,增加了问题答复召回的多样性,提升了问题答复的精度和准度,进一步使模型的训练更加精准而有效。

[0043] 步骤103,将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型。

[0044] 在本实施例中,执行主体可以利用机器学习算法,将步骤101中得到的各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型。信息确定模型用于表征基于每个问题对应的问题答复、各个分类标识对应的各个问题答复和各个分类标识,利用问题答复确定算法确定输入的每个问题与各个标准答复之间的对应结果。信息确定模型可以基于各个问题与分类标识之间的对应关系,确定输入的每个问题与各个标准答复之间的对应结果,例如图2-4所示。

[0045] 该信息确定模型可以用于表征问题与标准答复之间的对应关系。需要指出的是,该信息确定模型例如可以是数据表或计算公式等,本实施例不对此方面内容做任何限定。该信息确定模型的模型结构可以基于现有的各种逻辑回归模型而构建,例如但不限于:BERT、FastText、TextCNN等。

[0046] 继续参见图3,本实施例的对话数据处理方法300运行于服务器301中。服务器301首先获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复302,然后服务器301对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复303,其中,各个分类标识基于对各个问题答复进行分类而得到,标准答复基于分类标识对应的各个问题答复而确定,最后服务器301将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型304,其中,信息确定模型用于表征基于每个问题对应的问题答复、各个分类标识对应的各个问题答复和各个分类标识,确定输入的每个问题与各个标准答复之间的对应结果。

[0047] 本申请的上述实施例提供的对话数据处理方法采用获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复,对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复,将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,通过在训练信息确定模型时,将问题和答案

之间的相关性考虑其中,解决了因为挖掘出的问答对过多对索引系统造成的压力。方法将QA挖掘和答案召回合并为一个先聚类后分类的过程,极大的减少了内存的使用,提高了答案召回的效率。

[0048] 进一步参考图4,其示出了对话数据处理方法的第二实施例的示意图400。该方法的流程包括以下步骤:

[0049] 步骤401,获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复。

[0050] 步骤402,将各个问题答复进行向量化。

[0051] 在本实施例中,执行主体可以利用向量化算法将各个问题答复进行向量化。向量化算法可以包括但不限于:TF-IDF(term frequency-inverse document frequency)技术、Sentence2Vec模型、BERT等。

[0052] 步骤403,对向量化后的各个问题答复进行聚类,得到各个问题答复对应的各个分类标识和每个分类标识对应的各个问题答复。

[0053] 在本实施例中,执行主体可以利用聚类算法将内容相近的各个问题答复进行聚合,得到各个问题答复对应的各个分类标识和每个分类标识对应的各个问题答复。聚类算法可以包括但不限于:k均值聚类算法、AP_Cluster聚类算法。

[0054] 步骤404,对每个分类标识对应的各个问题答复进行选取,得到相应分类标识对应的标准答复。

[0055] 在本实施例中,执行主体可以通过选取方法对每个分类标识对应的各个问题答复进行选取,得到相应分类标识对应的标准答复。例如,选取方法为选取每个分类标识对应的各个问题答复中出现次数最多的问题答复为该分类标识的标准答复。选取方法还可以基于问题答复的字数多少、问题答复的具体内容等来设定。

[0056] 步骤405,将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型。

[0057] 在本实施例中,执行主体可以利用机器学习算法,将步骤401中得到的各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型。信息确定模型用于表征基于每个问题对应的问题答复、各个分类标识对应的各个问题答复和各个分类标识,利用问题答复确定算法确定输入的每个问题与各个标准答复之间的对应结果。

[0058] 在本实施例的一些可选的实现方式中,信息确定模型包括:第一确定子模型和第二确定子模型,第一确定子模型用于表征基于每个问题对应的问题答复和各个分类标识对应的各个问题答复,确定输入的每个问题与各个分类标识之间的对应结果,第二确定子模型用于表征基于第一确定子模型输出每个问题对应的分类标识,确定输入的每个问题与各个标准答复之间的对应结果;将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,包括:将各个问题作为输入数据,将与输入的各个问题所对应的分类标识作为输出数据,训练得到第一确定子模型;将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到第二确定子模型;基于训练完成的第一确定子模型和训练完成的第二确定子模型,确定信息确定模型。通过多个子模型分别进行模型训练,实现分步、递进式的模型训练,提升了模型训练的精准

度。

[0059] 需要说明的是,第一确定子模型和第二确定子模型均可以是数据表或计算公式等,本实施例不对此方面内容做任何限定,并且子模型结构可以基于现有的各种逻辑回归模型而构建。

[0060] 在本实施例中,步骤401具体操作与图1所示的实施例中的步骤101的操作基本相同,在此不再赘述。

[0061] 从图4中可以看出,与图1对应的实施例相比,本实施例中的对话数据处理方法的示意图400采用将各个问题答复进行向量化,对向量化后的各个问题答复进行聚类,得到各个问题答复对应的各个分类标识和每个分类标识对应的各个问题答复,对每个分类标识对应的各个问题答复进行选取,得到相应分类标识对应的标准答复,通过对答案进行聚类,减少了问题答复重复出现的可能,针对分类后的各个问题答复进行选取,提升了系统处理效率,进一步提升了模型训练效率。

[0062] 进一步参考图5,其出了根据本申请的用于生成信息的方法的第一实施例的示意图500。该用于生成信息的方法,包括以下步骤:

[0063] 步骤501,获取用户发送的问题。

[0064] 在本实施例中,执行主体(例如服务器或终端设备)可以通过有线连接或无线连接的方式从其他电子设备或者本地获取用户发送的问题。

[0065] 步骤502,将问题输入至预先训练的信息确定模型,生成问题对应的标准答复。

[0066] 在本实施例中,执行主体可以将步骤501获取到的问题输入至预先训练的信息确定模型,生成问题对应的标准答复。信息确定模型通过如上述对话数据处理方法中任一实施例的方法训练得到。

[0067] 在本实施例的一些可选的实现方式中,信息确定模型包括标识确定子模型和答复确定子模型,标识确定子模型用于确定问题与分类标识之间的对应结果,答复确定子模型用于表征基于标识确定子模型输出的问题的分类标识,确定问题与标准答复之间的对应结果;将问题输入至预先训练的信息确定模型,生成问题对应的标准答复,包括:将问题输入至预先训练的标识确定子模型,生成问题对应的分类标识;将生成的问题对应的分类标识输入至预先训练的答复确定子模型,生成问题对应的标准答复。通过多个模型得到最终的输出结果,提升了信息生成的效率和精准度。

[0068] 在本实施例的一些可选的实现方式中,方法还包括:利用打分方法,对问题对应的标准答复进行打分,得到问题对应的标准答复的分值;基于问题对应的各个标准答复的分值排序,确定问题的最终答复;将最终答复信息发送至用户。实现了一种全新的检索式对话系统知识库构建和答案召回方案,不需要从对话日志中挖掘QA对,也不需要维护问题索引库。通过将结果反馈给用户,使用户快速而精准的得到问题的答案。

[0069] 从图5中可以看出,与图1对应的实施例相比,本实施例中的用于生成信息的方法的流程500突出了采用训练得到的信息确定模型,来生成问题对应的标准答复的步骤。由此,本实施例描述的方案可以实现针对用户问题的答案获取。

[0070] 进一步参考图6,作为对上述图1~4所示方法的实现,本申请提供了一种对话数据处理装置的一个实施例,该装置实施例与图1所示的方法实施例相对应,除下面所记载的特征外,该装置实施例还可以包括与图1所示的方法实施例相同或相应的特征,以及产生与图

1所示的方法实施例相同或相应的效果,该装置具体可以应用于各种电子设备中。

[0071] 如图6所示,本实施例的对话数据处理装置600包括:获取单元601分析单元602和训练单元603,其中,获取单元,被配置成获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复;分析单元,被配置成对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复,其中,各个分类标识基于对各个问题答复进行分类而得到,标准答复基于分类标识对应的各个问题答复而确定;训练单元,被配置成将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,其中,信息确定模型用于表征基于每个问题对应的问题答复、各个分类标识对应的各个问题答复和各个分类标识,确定输入的每个问题与各个标准答复之间的对应结果。

[0072] 在本实施例中,对话数据处理装置600的获取单元601分析单元602和训练单元603的具体处理及其所带来的技术效果可分别参考图1对应的实施例中的步骤101到步骤103的相关说明,在此不再赘述。

[0073] 在本实施例的一些可选的实现方式中,分析单元,包括:向量化模块,被配置成将各个问题答复进行向量化;聚类模块,被配置成对向量化后的各个问题答复进行聚类,得到各个问题答复对应的各个分类标识和每个分类标识对应的各个问题答复,其中,聚类用于表征利用聚类算法将内容相近的各个问题答复进行聚合;选取模块,被配置成对每个分类标识对应的各个问题答复进行选取,得到相应分类标识对应的标准答复。

[0074] 在本实施例的一些可选的实现方式中,训练单元中的信息确定模型包括:第一确定子模型和第二确定子模型,第一确定子模型用于表征基于每个问题对应的问题答复和各个分类标识对应的各个问题答复,确定输入的每个问题与各个分类标识之间的对应结果,第二确定子模型用于表征基于第一确定子模型输出每个问题对应的分类标识,确定输入的每个问题与各个标准答复之间的对应结果;训练单元,包括:第一训练模块,被配置成将各个问题作为输入数据,将与输入的各个问题所对应的分类标识作为输出数据,训练得到第一确定子模型;第二训练模块,被配置成将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到第二确定子模型;确定模块,被配置成基于训练完成的第一确定子模型和训练完成的第二确定子模型,确定信息确定模型。

[0075] 在本实施例的一些可选的实现方式中,装置还包括:处理单元,被配置成对各个问题答复进行数据处理,得到处理后的各个问题答复,其中,数据处理包括:无效数据修正、数据合并、数据去重和数据归一化中的至少一项。

[0076] 本公开的上述实施例提供对话数据处理装置,通过获取单元,获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复,利用分析单元,对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复,其中,各个分类标识基于对各个问题答复进行分类而得到,标准答复基于分类标识对应的各个问题答复而确定,最后通过训练单元,将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,通过在训练信息确定模型时,将问题和答案之间的相关性考虑其中,解决了因为挖掘出的问答对过多对索引系统造成的压力,实现了一种针对用户问题-答案的信息确定模型训练方法。

[0077] 继续参考参考图7,作为对上述图5所示方法的实现,本申请提供了一种用于生成信息的装置的一个实施例,该装置实施例与图5所示的方法实施例相对应,除下面所记载的特征外,该装置实施例还可以包括与图5所示的方法实施例相同或相应的特征,以及产生与图5所示的方法实施例相同或相应的效果,该装置具体可以应用于各种电子设备中。

[0078] 如图7所示,本实施例的用于生成信息的装置700包括:问题获取单元701和信息生成单元702,其中,问题获取单元,被配置成获取用户发送的问题;信息生成单元,被配置成将问题输入至预先训练的信息确定模型,生成问题对应的标准答复,其中,信息确定模型通过如上述对话数据处理方法中任一实施例的方法训练得到。

[0079] 在本实施例中,用于生成信息的装置700的问题获取单元701和信息生成单元702的具体处理及其所带来的技术效果可分别参考图4对应的实施例中的步骤501到步骤502的相关说明,在此不再赘述。

[0080] 在本实施例的一些可选的实现方式中,信息生成单元中的信息确定模型包括标识确定子模型和答复确定子模型,标识确定子模型用于确定问题与分类标识之间的对应结果,答复确定子模型用于表征基于标识确定子模型输出的问题的分类标识,确定问题与标准答复之间的对应结果;信息生成单元,包括:第一生成模块,被配置成将问题输入至预先训练的标识确定子模型,生成问题对应的分类标识;第二生成模块,被配置成将生成的问题对应的分类标识输入至预先训练的答复确定子模型,生成问题对应的标准答复。

[0081] 在本实施例的一些可选的实现方式中,装置还包括:评分单元,被配置成利用评分方法,对问题对应的标准答复进行打分,得到问题对应的标准答复的分值;确定单元,被配置成基于问题对应的各个标准答复的分值排序,确定问题的最终答复;发送单元,被配置成将最终答复信息发送至用户。

[0082] 本公开的上述实施例提供对话数据处理装置,通过问题获取单元,获取用户发送的问题,利用信息生成单元,将问题输入至预先训练的信息确定模型,生成问题对应的标准答复,其中,信息确定模型通过如上述对话数据处理方法中任一实施例的方法训练得到,实现了针对用户问题的答案获取。

[0083] 根据本申请的实施例,本申请还提供了一种电子设备和一种可读存储介质。

[0084] 如图8所示,是根据本申请实施例的对话数据处理方法的电子设备的框图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0085] 如图8所示,该电子设备包括:一个或多个处理器801、存储器802,以及用于连接各部件的接口,包括高速接口和低速接口。各个部件利用不同的总线互相连接,并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理,包括存储在存储器中或者存储器上以在外部输入/输出装置(诸如,耦合至接口的显示设备)上显示GUI的图形信息的指令。在其它实施方式中,若需要,可以将多个处理器和/或多条总线与多个存储器和多个存储器一起使用。同样,可以连接多个电子设备,各个设备提供部分必要的操作(例如,作为服务器阵列、一组刀片式服务器、或者多处理器系

统)。图8中以一个处理器801为例。

[0086] 存储器802即为本申请所提供的非瞬时计算机可读存储介质。其中,存储器存储有可由至少一个处理器执行的指令,以使至少一个处理器执行本申请所提供的对话数据处理方法。本申请的非瞬时计算机可读存储介质存储计算机指令,该计算机指令用于使计算机执行本申请所提供的对话数据处理方法。

[0087] 存储器802作为一种非瞬时计算机可读存储介质,可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及模块,如本申请实施例中的对话数据处理方法对应的程序指令/模块(例如,附图6所示的获取单元601分析单元602和训练单元603)。处理器801通过运行存储在存储器802中的非瞬时软件程序、指令以及模块,从而执行服务器的各种功能应用以及数据处理,即实现上述方法实施例中的对话数据处理方法。

[0088] 存储器802可以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、至少一个功能所需要的应用程序;存储数据区可存储根据对话数据处理电子设备的使用所创建的数据等。此外,存储器802可以包括高速随机存取存储器,还可以包括非瞬时存储器,例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中,存储器802可选包括相对于处理器801远程设置的存储器,这些远程存储器可以通过网络连接至对话数据处理电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0089] 对话数据处理方法的电子设备还可以包括:输入装置803和输出装置804。处理器801、存储器802、输入装置803和输出装置804可以通过总线或者其他方式连接,图8中以通过总线连接为例。

[0090] 输入装置803可接收输入的数字或字符信息,以及产生与对话数据处理电子设备的用户设置以及功能控制有关的键信号输入,例如触摸屏、小键盘、鼠标、轨迹板、触模板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置804可以包括显示设备、辅助照明装置(例如,LED)和触觉反馈装置(例如,振动电机)等。该显示设备可以包括但不限于,液晶显示器(LCD)、发光二极管(LED)显示器和等离子体显示器。在一些实施方式中,显示设备可以是触摸屏。

[0091] 此处描述的系统和技术各种实施方式可以在数字电子电路系统、集成电路系统、专用ASIC(专用集成电路)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0092] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令,并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的,术语“机器可读介质”和“计算机可读介质”指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如,磁盘、光盘、存储器、可编程逻辑装置(PLD)),包括,接收作为机器可读信号的机器指令的机器可读介质。术语“机器可读信号”指的是用于将机器指令和/或数据提供给可编程处理器的任何

信号。

[0093] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0094] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0095] 计算机系统可以包括客户端和服务端。客户端和服务端一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务端关系的计算机程序来产生客户端和服务端的关系。

[0096] 根据本申请实施例的技术方案采用获取对话数据,并对对话数据进行解析,得到对话数据对应的各个问题和各个问题对应的问题答复,对各个问题答复进行分析,得到各个问题答复对应的各个分类标识、每个分类标识对应的各个问题答复和每个分类标识对应的标准答复,将各个问题作为输入数据,将与输入的各个问题所对应的标准答复作为输出数据,训练得到信息确定模型,通过在训练信息确定模型时,将问题和答案之间的相关性考虑其中,解决了因为挖掘出的问答对过多对索引系统造成的压力。方法将QA挖掘和答案召回合并为一个先聚类后分类的过程,极大的减少了内存的使用,提高了答案召回的效率。

[0097] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本申请公开的技术方案所期望的结果,本文在此不进行限制。

[0098] 上述具体实施方式,并不构成对本申请保护范围的限制。本领域技术人员应该明白的是,根据设计要求和因素,可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等,均应包含在本申请保护范围之内。

100

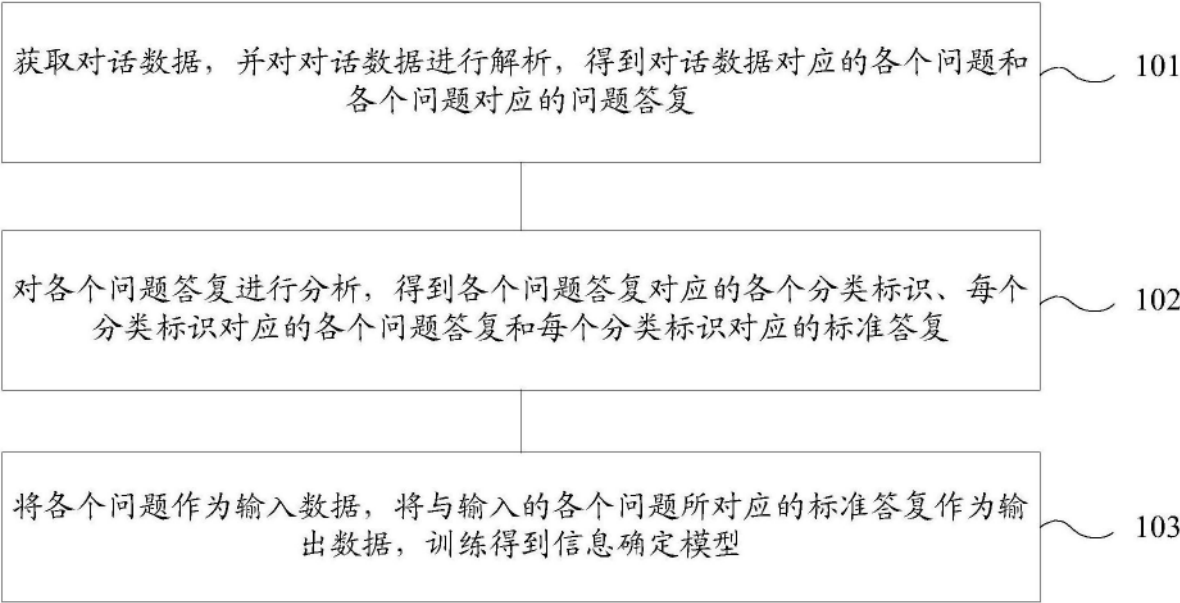


图1

会话 ID	问题或答案标志	文本
0	Q	在吗?
0	A	在的，请问有什么可以帮您
0	Q	包安装吗?
0	A	我们包安装、包配送的
1	Q	在吗
1	A	您好，在的哦，请问有什么可以帮您?
1	Q	可以开发票吗
1	A	稍等帮你查询下
1	A	我们可以开电子发票的

图2-1

分类标识 A-ID	问题答复 A
0	您好，在的哦，请问有什么可以帮您？
0	在的，请问有什么可以帮您？
0	在的，请问有什么可以帮您
1	我们包安装、包配送的
1	我们包安装、包配送的哦
2	稍等帮你查询下
2	我们可以看电子发票的

图2-2

分类标识 A-ID	各个问题答复 A	标准答复 A'
0	您好，在的哦，请问有什么可以帮您？ 在的，请问有什么可以帮您？ 在的，请问有什么可以帮您	您好，在的，请问有什么可以帮您？
1	我们包安装、包配送的 我们包安装、包配送的哦	我们包安装、包配送的
2	稍等我帮你查询下 我们可以开电子发票的	我们可以开电子发票的

图2-3

问题 Q	A_ID	标准答复 A'
在吗?	0	您好, 在的, 请问有什么可以帮您?
包安装吗?	1	我们包安装、包配送的
在吗	0	您好, 在的, 请问有什么可以帮您?
包安装吗	1	我们包安装、包配送的
可以开发票吗	2	我们可以开电子发票的

图2-4

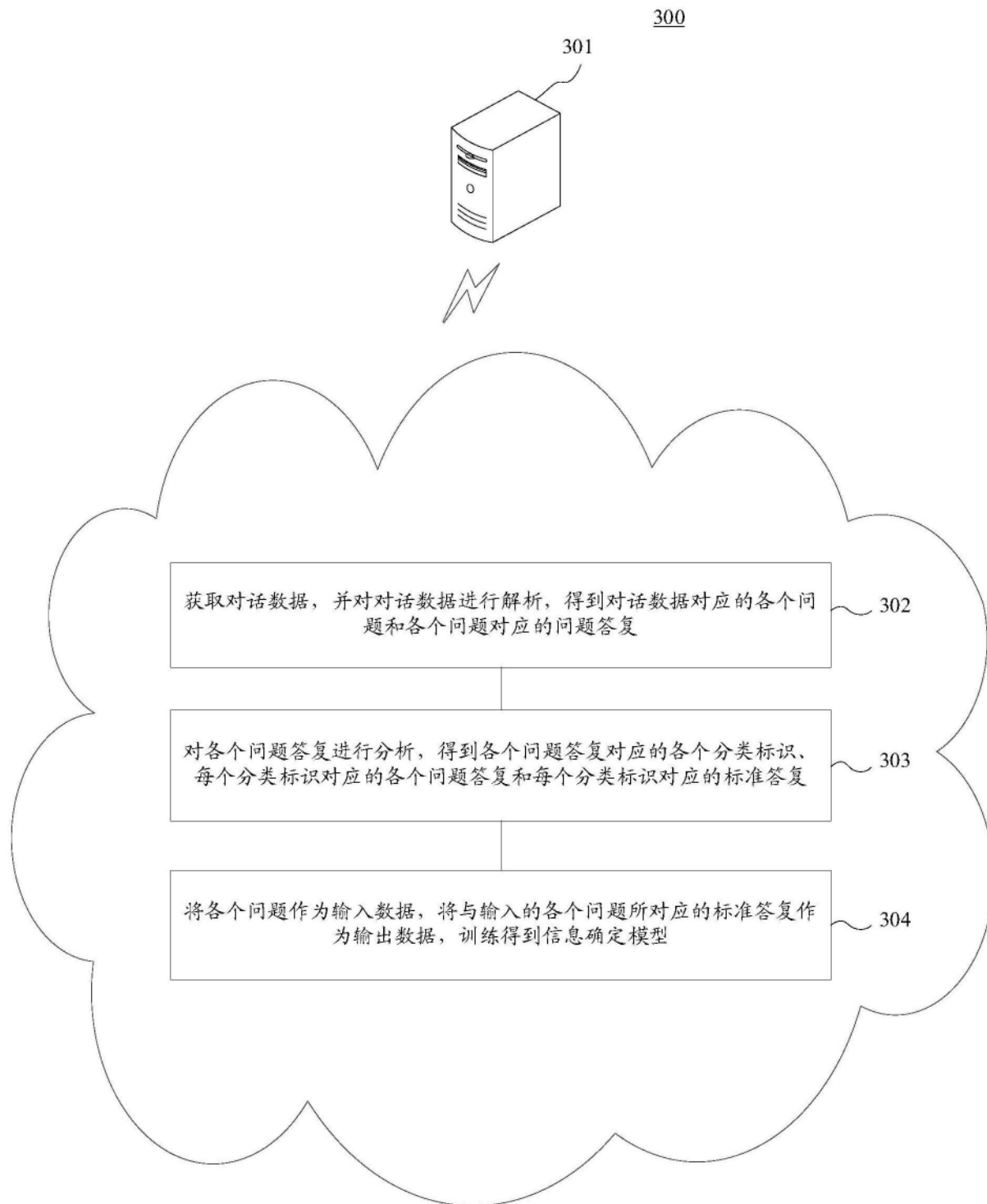


图3

400

图4

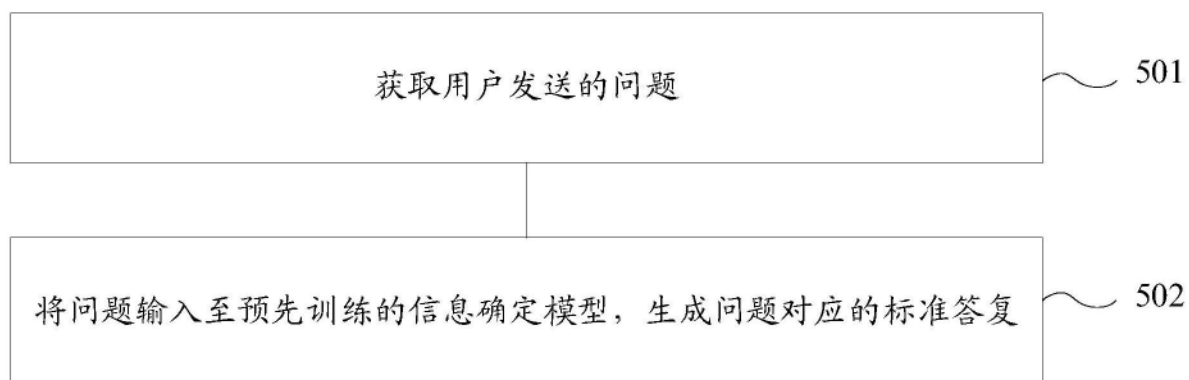
500

图5

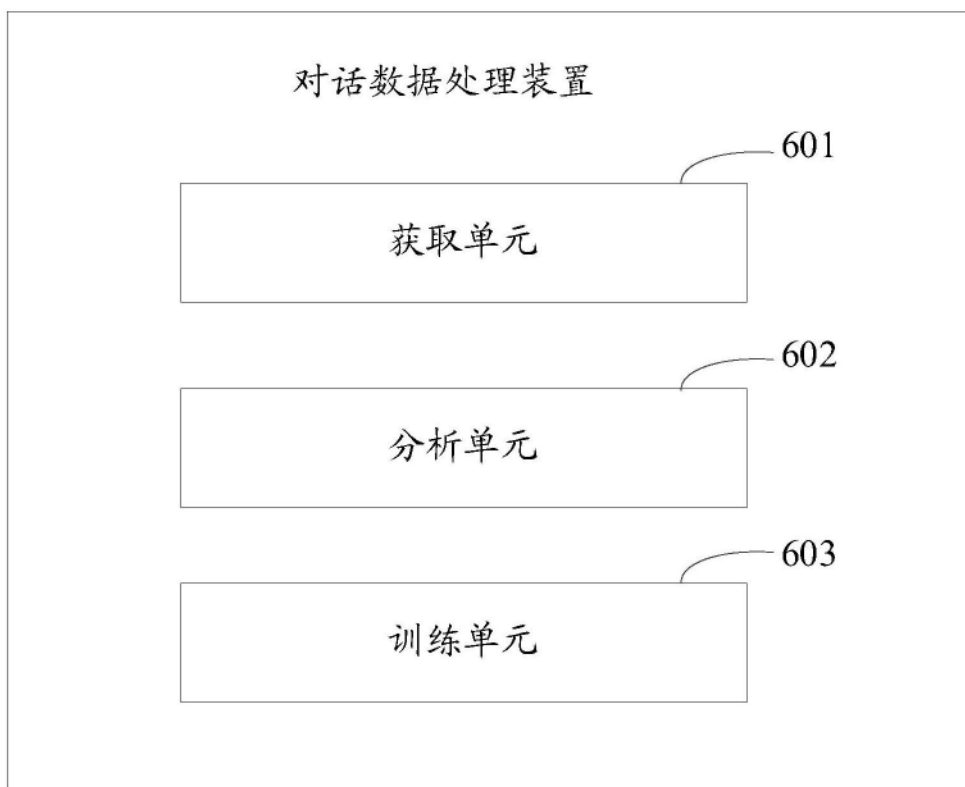
600

图6

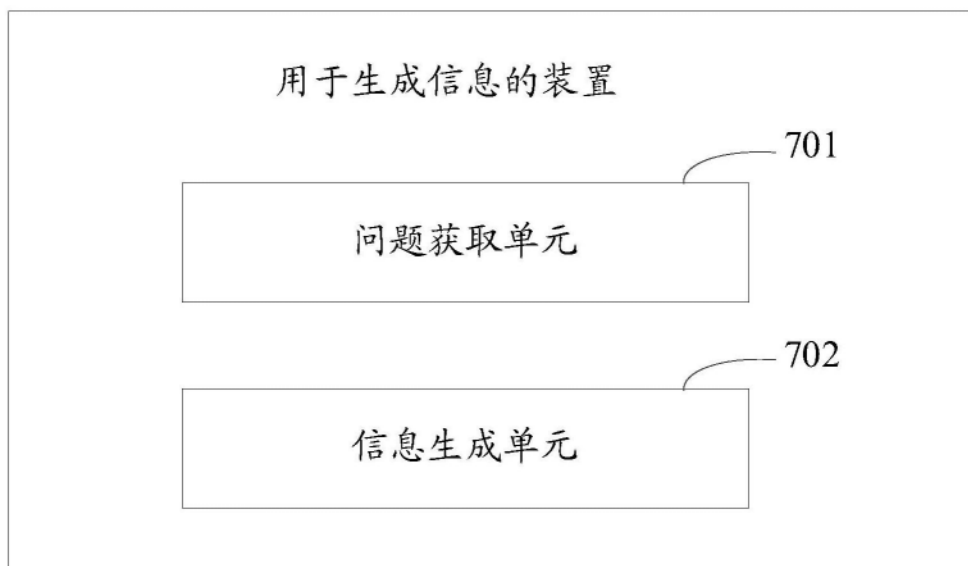
700

图7

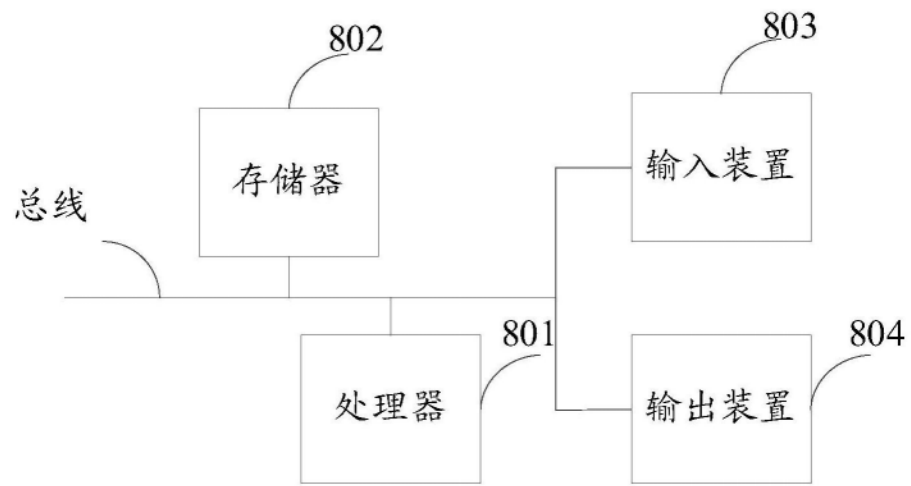


图8