



(21) 申请号 202011225091.0

G06F 40/35 (2020.01)

(22) 申请日 2020.11.05

G06F 40/295 (2020.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 112214593 A

(56) 对比文件

CN 111400470 A, 2020.07.10

CN 111506836 A, 2020.08.07

(43) 申请公布日 2021.01.12

审查员 孙鹏

(73) 专利权人 腾讯科技(深圳)有限公司

地址 518057 广东省深圳市南山区高新区

科技中一路腾讯大厦35层

(72) 发明人 程任清 邱东洋 李正兴 王智圣

魏重强

(74) 专利代理机构 北京派特恩知识产权代理有

限公司 11270

专利代理师 刘晖铭 张颖玲

(51) Int. Cl.

G06F 16/332 (2019.01)

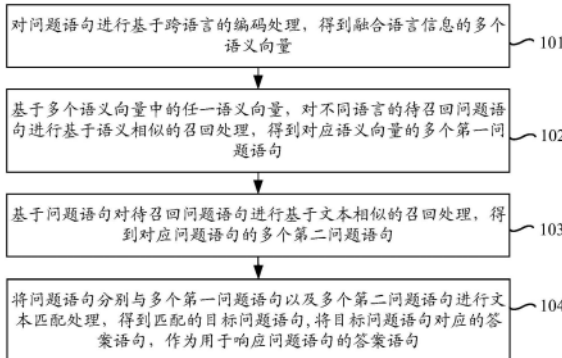
权利要求书4页 说明书19页 附图10页

(54) 发明名称

问答处理方法、装置、电子设备及存储介质

(57) 摘要

本申请提供了一种问答处理方法、装置、电子设备及计算机可读存储介质；涉及人工智能的自然语言处理技术；方法包括：对问题语句进行基于跨语言的编码处理，得到融合语言信息的多个语义向量；基于多个语义向量中的任一语义向量，对不同语言的待召回问题语句进行基于语义相似的召回处理，得到对应语义向量的多个第一问题语句；基于问题语句对待召回问题语句进行基于文本相似的召回处理，得到对应问题语句的多个第二问题语句；将问题语句分别与多个第一问题语句以及多个第二问题语句进行文本匹配处理，得到匹配的目标问题语句；将目标问题语句对应的答案语句，作为用于响应问题语句的答案语句。通过本申请，能够提高智能问答的准确性。



1. 一种问答处理方法,其特征在于,包括:

对待响应问题语句进行基于跨语言的改写处理,得到与所述待响应问题语句不同的问题语句,其中,对待响应问题语句进行基于跨语言的改写处理是基于领域词典进行匹配和基于多语言近义词词典进行匹配中至少之一实现的;

对与所述待响应问题语句不同的问题语句进行词形归一化处理,得到用于进行编码处理的所述问题语句;

对所述问题语句进行基于跨语言的编码处理,得到融合语言信息的多个语义向量;

基于所述多个语义向量中的任一语义向量,对不同语言的待召回问题语句进行基于语义相似的召回处理,得到对应所述语义向量的多个第一问题语句;

对所述待召回问题语句进行基于文本相似的召回处理,得到对应所述问题语句的多个第二问题语句;

将所述问题语句分别与所述多个第一问题语句以及所述多个第二问题语句进行文本匹配处理,得到匹配的目标问题语句;

将所述目标问题语句对应的答案语句,作为用于响应所述问题语句的答案语句。

2. 根据权利要求1所述的方法,其特征在于,所述对待响应问题语句进行基于跨语言的改写处理,得到与所述待响应问题语句不同的问题语句,包括:

将领域词典与所述待响应问题语句中的词语进行匹配处理,得到与所述待响应问题语句中的词语对应的第一匹配词;

基于所述第一匹配词对所述待响应问题语句中的词语进行替换处理,得到替换后的问题语句;

当多语言近义词词典匹配到与所述替换后的问题语句中的词语对应的第二匹配词时,基于所述第二匹配词对所述替换的问题语句进行替换处理,得到与所述待响应问题语句不同的问题语句。

3. 根据权利要求1所述的方法,其特征在于,所述对与所述待响应问题语句不同的问题语句进行词形归一化处理,得到用于进行编码处理的所述问题语句,包括:

确定与所述待响应问题语句不同的问题语句中的多个词语;

确定所述多个词语分别对应的词频;

将所述多个词语拆分为与所述多个词语一一对应的多个字符序列;

迭代执行以下处理:

基于所述多个词语分别对应的词频,确定所述多个字符序列中的连续字符对的出现频率;

将最高出现频率所对应的连续字符对进行合并,将合并结果作为新字符来更新所述字符序列;

当更新的字符序列中的连续字符对的出现频率为1时停止迭代处理。

4. 根据权利要求1所述的方法,其特征在于,所述对待响应问题语句进行基于跨语言的改写处理之前,所述方法还包括:

对所述待响应问题语句进行分词处理,得到对应所述待响应问题语句的词序列;

对所述词序列进行实体词识别处理,得到所述待响应问题语句的实体词;

基于所述待响应问题语句的实体词对所述词序列进行合并处理,将合并结果用于进行

所述改写处理。

5. 根据权利要求4所述的方法, 其特征在于,

所述对所述待响应问题语句进行分词处理, 得到对应所述待响应问题语句的词序列, 包括:

对所述待响应问题语句进行语言识别处理, 得到所述待响应问题语句的语言类型;

基于所述待响应问题语句的语言类型, 调用与所述语言类型对应的分词器进行分词处理, 得到对应所述待响应问题语句的词序列;

所述对所述词序列进行实体词识别处理, 得到所述待响应问题语句的实体词, 包括:

对所述词序列进行通用实体词识别处理, 得到通用实体词;

对所述词序列进行领域实体词识别处理, 得到领域实体词;

将所述通用实体词以及所述领域实体词作为所述待响应问题语句的实体词。

6. 根据权利要求1所述的方法, 其特征在于,

用于问答处理的神经网络模型包括多个编码网络;

所述对所述问题语句进行基于跨语言的编码处理, 得到融合语言信息的多个语义向量, 包括:

通过所述多个编码网络中的任一所述编码网络, 基于不同于所述问题语句的语言, 对所述问题语句进行编码处理, 得到对应所述编码网络的语义向量;

将多个所述对应所述编码网络的语义向量作为所述多个语义向量。

7. 根据权利要求6所述的方法, 其特征在于,

所述用于问答处理的神经网络模型包括解码网络;

所述方法还包括:

通过所述多个编码网络中的任一所述编码网络, 基于不同于问题语句样本的语言, 对所述问题语句样本进行编码处理, 得到对应所述编码网络的语义向量;

通过所述解码网络对对应所述编码网络的语义向量进行解码处理, 得到对应所述问题语句样本的跨语言文本;

基于所述跨语言文本以及所述问题语句样本的跨语言文本标签, 构建所述神经网络模型的损失函数;

更新所述神经网络模型的参数直至所述损失函数收敛, 将所述损失函数收敛时所述神经网络模型的更新的参数, 作为训练后的所述神经网络模型的参数。

8. 根据权利要求7所述的方法, 其特征在于,

所述方法还包括:

对对应所述编码网络的语义向量进行分类处理, 得到对应所述问题语句样本的预测类别;

基于所述预测类别以及所述问题语句样本的类别标签, 构建所述神经网络模型的类别损失函数;

所述基于所述跨语言文本以及所述问题语句样本的跨语言文本标签, 构建所述神经网络模型的损失函数, 包括:

基于所述跨语言文本以及所述问题语句样本的跨语言文本标签, 构建所述神经网络模型的跨语言损失函数;

对所述类别损失函数以及所述跨语言损失函数进行加权求和,得到所述神经网络模型的损失函数。

9.根据权利要求1所述的方法,其特征在于,所述对不同语言的待召回问题语句进行基于语义相似的召回处理,得到对应所述语义向量的多个第一问题语句,包括:

针对所述融合语言信息的多个语义向量中的任一所述语义向量执行以下处理:

确定所述语义向量分别与不同语言的待召回问题语句的语义向量之间的语义相似度;

基于所述语义向量与所述待召回问题语句的语义向量之间的语义相似度,对所述待召回问题语句进行降序排序,将降序排序结果中排序在前N个的待召回问题语句作为所述多个第一问题语句;

其中,N为自然数。

10.根据权利要求1所述的方法,其特征在于,所述对所述待召回问题语句进行基于文本相似的召回处理,得到对应所述问题语句的多个第二问题语句,包括:

将所述问题语句中的关键词与所述待召回问题语句进行匹配,得到所述关键词在所述待召回问题语句中的出现频率;

基于所述关键词在所述待召回问题语句中的出现频率,确定所述问题语句与所述待召回问题语句之间的文本相似度;

基于所述问题语句与所述待召回问题语句之间的文本相似度,对所述待召回问题语句进行降序排序,将降序排序结果中排序在前M个的待召回问题语句作为所述多个第二问题语句;

其中,M为自然数。

11.根据权利要求1所述的方法,其特征在于,所述对所述待召回问题语句进行基于文本相似的召回处理,得到对应所述问题语句的多个第二问题语句,包括:

基于所述问题语句中的词语,确定所述问题语句中的词语的权重;

对所述问题语句中的词语进行哈希处理,得到对应所述词语的哈希值;

基于所述词语的权重,对对应所述词语的哈希值进行加权求和,得到所述问题语句的哈希签名;

基于所述问题语句的哈希签名以及所述待召回问题语句的哈希签名,确定所述问题语句与所述待召回问题语句之间的文本相似度;

基于所述问题语句与所述待召回问题语句之间的文本相似度,对所述待召回问题语句进行降序排序,将降序排序结果中排序在前K个的待召回问题语句作为所述多个第二问题语句;

其中,K为自然数。

12.根据权利要求1所述的方法,其特征在于,所述将所述问题语句与所述多个第一问题语句以及所述多个第二问题语句进行文本匹配处理之前,所述方法还包括:

基于所述问题语句与所述第一问题语句之间的相似度、以及所述问题语句与所述第二问题语句之间的相似度,对所述第一问题语句以及所述第二问题语句进行降序排序,将降序排序结果中排序在前L个的第一问题语句或者所述第二问题语句作为用于进行文本匹配的召回问题语句;

其中,L为自然数。

13.一种问答处理装置,其特征在于,所述装置包括:

第一召回模块,用于对待响应问题语句进行基于跨语言的改写处理,得到与所述待响应问题语句不同的问题语句,其中,对待响应问题语句进行基于跨语言的改写处理是基于领域词典进行匹配和基于多语言近义词词典进行匹配中至少之一实现的;对与所述待响应问题语句不同的问题语句进行词形归一化处理,得到用于进行编码处理的所述问题语句;对所述问题语句进行基于跨语言的编码处理,得到融合语言信息的多个语义向量;基于所述多个语义向量中的任一语义向量,对不同语言的待召回问题语句进行基于语义相似的召回处理,得到对应所述语义向量的多个第一问题语句;

第二召回模块,用于对所述待召回问题语句进行基于文本相似的召回处理,得到对应所述问题语句的多个第二问题语句;

匹配模块,用于将所述问题语句分别与所述多个第一问题语句以及所述多个第二问题语句进行文本匹配处理,得到匹配的目标问题语句;将所述目标问题语句对应的答案语句,作为用于响应所述问题语句的答案语句。

14.一种电子设备,其特征在于,所述电子设备包括:

存储器,用于存储可执行指令;

处理器,用于执行所述存储器中存储的可执行指令时,实现权利要求1至12任一项所述的问答处理方法。

15.一种计算机可读存储介质,存储有计算机可执行指令或者计算机程序,其特征在于,所述计算机可执行指令或者计算机程序被处理器执行时实现权利要求1至12任一项所述的问答处理方法。

16.一种计算机程序产品,包括计算机可执行指令或计算机程序,其特征在于,所述计算机可执行指令或计算机程序被处理器执行时实现权利要求1至12任一项所述的问答处理方法。

问答处理方法、装置、电子设备及存储介质

技术领域

[0001] 本申请涉及人工智能的自然语言处理技术,尤其涉及一种问答处理方法、装置、电子设备及计算机可读存储介质。

背景技术

[0002] 人工智能(Artificial Intelligence, AI)是计算机科学的一个综合技术,通过研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。人工智能技术是一门综合学科,涉及领域广泛,例如自然语言处理技术以及机器学习/深度学习等几大方向,随着技术的发展,人工智能技术将在更多的领域得到应用,并发挥越来越重要的价值。

[0003] 智能问答系统是自然语言处理领域的重要应用之一,在对话系统、客服系统、智能硬件等中都有广泛的应用,即智能问答系统是这些复杂系统的基础组件。

[0004] 但是,智能问答系统面临各国语言所带来的问答压力,相关技术中缺乏基于人工智能对各国语言进行问答的有效方案。

发明内容

[0005] 本申请实施例提供一种问答处理方法、装置、电子设备及计算机可读存储介质,能够进行跨语言的召回处理,提高智能问答的准确性。

[0006] 本申请实施例的技术方案是这样实现的:

[0007] 本申请实施例提供一种问答处理方法,包括:

[0008] 对问题语句进行基于跨语言的编码处理,得到融合语言信息的多个语义向量;

[0009] 基于所述多个语义向量中的任一语义向量,对不同语言的待召回问题语句进行基于语义相似的召回处理,得到对应所述语义向量的多个第一问题语句;

[0010] 基于所述问题语句对所述待召回问题语句进行基于文本相似的召回处理,得到对应所述问题语句的多个第二问题语句;

[0011] 将所述问题语句分别与所述多个第一问题语句以及所述多个第二问题语句进行文本匹配处理,将与匹配得到的问题语句对应的答案语句,作为用于响应所述问题语句的答案语句。

[0012] 在上述技术方案中,所述将所述问题语句分别与所述多个第一问题语句以及所述多个第二问题语句进行文本匹配处理,得到匹配的目标问题语句,包括:

[0013] 通过分类器对所述问题语句与所述多个第一问题语句进行分类处理,得到所述问题语句分别与所述多个第一问题语句之间的相似度;

[0014] 通过所述分类器对所述问题语句与所述多个第二问题语句进行分类处理,得到所述问题语句分别与所述多个第二问题语句之间的相似度;

[0015] 将最大相似度所对应的所述第一问题语句或者所述第二问题语句作为所述目标问题语句。

[0016] 在上述技术方案中,所述将所述目标问题语句对应的答案语句,作为用于响应所述问题语句的答案语句,包括:

[0017] 当在问题答案库中查询到与所述目标问题语句对应的答案语句,且所述对应的答案语句的语言与所述问题语句的语言一致时,将所述目标问题语句对应的答案语句,作为用于响应所述问题语句的答案语句;

[0018] 当在所述问题答案库中查询到与所述目标问题语句对应的答案语句,且所述对应的答案语句的语言与所述问题语句的语言不一致时,获取所述答案语句对应所述问题语句的语言的翻译结果,并作为用于响应所述问题语句的答案语句。

[0019] 本申请实施例提供一种问答处理装置,包括:

[0020] 第一召回模块,用于对问题语句进行基于跨语言的编码处理,得到融合语言信息的多个语义向量;

[0021] 基于所述多个语义向量中的任一语义向量,对不同语言的待召回问题语句进行基于语义相似的召回处理,得到对应所述语义向量的多个第一问题语句;

[0022] 第二召回模块,用于对所述待召回问题语句进行基于文本相似的召回处理,得到对应所述问题语句的多个第二问题语句;

[0023] 匹配模块,用于将所述问题语句分别与所述多个第一问题语句以及所述多个第二问题语句进行文本匹配处理,得到匹配的目标问题语句;将所述目标问题语句对应的答案语句,作为用于响应所述问题语句的答案语句。

[0024] 本申请实施例提供一种用于问答处理的电子设备,所述电子设备包括:

[0025] 存储器,用于存储可执行指令;

[0026] 处理器,用于执行所述存储器中存储的可执行指令时,实现本申请实施例提供的问答处理方法。

[0027] 本申请实施例提供一种计算机可读存储介质,存储有可执行指令,用于引起处理器执行时,实现本申请实施例提供的问答处理方法。

[0028] 本申请实施例具有以下有益效果:

[0029] 通过进行基于语义相似的召回处理以及基于文本相似的召回处理,得到不同语言的问题语句,从而通过不同语言的问题语句应对各国语言所带来的问答压力;并且通过基于语义相似的召回处理以及基于文本相似的召回处理,综合两种召回方式所得到的问题语句,还可以提高召回得到的问题语句的多样性,从而提高智能问答的准确性。

附图说明

[0030] 图1是本申请实施例提供的智能问答系统的应用场景示意图;

[0031] 图2是本申请实施例提供的用于问答处理的电子设备的结构示意图;

[0032] 图3A-3C是本申请实施例提供的问答处理方法的流程示意图;

[0033] 图4是本申请实施例提供的跨语言问答模块的结构示意图;

[0034] 图5是本申请实施例提供的预处理的流程示意图;

[0035] 图6是本申请实施例提供的跨语言语义相似召回的网络结构的结构示意图;

[0036] 图7是本申请实施例提供的精排模块的结构示意图;

[0037] 图8是本申请实施例提供的英文准确率的效果对比示意图;

- [0038] 图9是本申请实施例提供的小语种准确率的效果对比示意图；
- [0039] 图10是本申请实施例提供的手游中的问答界面图；
- [0040] 图11是本申请实施例提供的手游中的英文问答界面图；
- [0041] 图12是本申请实施例提供的手游中的中文问答界面图。

具体实施方式

[0042] 为了使本申请的目的、技术方案和优点更加清楚,下面将结合附图对本申请作进一步地详细描述,所描述的实施例不应视为对本申请的限制,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例,都属于本申请保护的范围。

[0043] 在以下的描述中,所涉及的术语“第一\第二”仅仅是是区别类似的对象,不代表针对对象的特定排序,可以理解地,“第一\第二”在允许的情况下可以互换特定的顺序或先后次序,以使这里描述的本申请实施例能够以除了在这里图示或描述的以外的顺序实施。

[0044] 除非另有定义,本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中所使用的术语只是为了描述本申请实施例的目的,不是旨在限制本申请。

[0045] 对本申请实施例进行进一步详细说明之前,对本申请实施例中涉及的名词和术语进行说明,本申请实施例中涉及的名词和术语适用于如下的解释。

[0046] 1) 卷积神经网络(CNN,Convolutional Neural Networks):一类包含卷积计算且具有深度结构的前馈神经网络(FNN,Feedforward Neural Networks),是深度学习(deep learning)的代表算法之一。卷积神经网络具有表征学习(representation learning)能力,能够按其阶层结构对输入图像进行平移不变分类(shift-invariant classification)。

[0047] 2) 循环神经网络(RNN,Recurrent Neural Network):一类以序列(sequence)数据为输入,在序列的演进方向进行递归(recursion)且所有节点(循环单元)按链式连接的递归神经网络。

[0048] 3) 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。

[0049] 4) 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“妻子”为实体词。其中,通用实体词是指常规的实体词,例如表示一些生活用品的词语,“洗发水”为通用实体词;领域实体词是指专业领域的实体词,例如游戏中的“诸葛连弩”为领域实体词。

[0050] 5) 召回(Recall):从文本库中检索出相关文本,例如基于文本1,对文本库中的各文本进行匹配,将匹配到的文本作为与文本1相关的召回文本。

[0051] 6) 智能问答系统:基于大量语料数据,通过数学模型,通过相关编程语言实现的一个能够和人类进行对话,并解决问题的一个软件系统。

[0052] 7) 常见问题解答(FAQ,Frequently Asked Question):一种在已有的“问题一答案”对集合中找到与用户提问相匹配的问题,并将其对应的答案返回给用户的问答式检索系统。

[0053] 8) 词汇不足(oov,out-of-vocabulary)问题:在自然语言处理或者文本处理的时

候,通常会有一个词库(vocabulary),当遇到某些词并不在现有词库中,则称这些词为oov词。

[0054] 9) 哈希:把任意长度的输入(又叫做预映射)通过散列算法变换成固定长度的输出,该输出就是散列值。这种变换是一种压缩映射,即散列值的空间通常远小于输入的空间,不同的输入可能会散列成相同的输出,所以不可能从散列值来确定唯一的输入值。即哈希处理是一种将任意长度的消息压缩到某一固定长度的消息摘要的处理过程。

[0055] 10) 领域词典:通过专业领域中的各词语所建立的知识图谱(即词语之间的对应关系),例如,“诸葛连弩”为游戏中的专业名词,在领域词典中可以查询到“诸葛”与“诸葛连弩”对应,即在领域词典中“诸葛”与“诸葛连弩”存在对应关系。

[0056] 11) 多语言近义词词典:通过不同语言的近义词所建立的知识图谱(即词语之间的对应关系),例如,“配件”为中文的名词,在多语言近义词词典中可以查询到“accessories”与“配件”对应,即在多语言近义词词典中“accessories”与“配件”存在对应关系。

[0057] 本申请实施例提供了一种问答处理方法、装置、电子设备及计算机可读存储介质,能够进行跨语言的召回处理,提高智能问答的准确性。

[0058] 本申请实施例所提供的问答处理方法,可以由终端/服务器独自实现;也可以由终端和服务器协同实现,例如终端独自承担下文所述的问答处理方法,或者,终端向服务器发送针对问题语句的问答请求(包括问题语句),服务器根据接收的针对问题语句的问答请求执行问答处理方法,并向终端发送响应问题语句的答案语句。

[0059] 本申请实施例提供的用于问答处理的电子设备可以是各种类型的终端设备或服务器,其中,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,还可以是提供云计算服务的云服务器;终端可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等,但并不局限于此。终端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。

[0060] 以服务器为例,例如可以是部署在云端的服务器集群,向用户开放人工智能云服务(AI as a Service,AIaaS),AIaaS平台会把几类常见的AI服务进行拆分,并在云端提供独立或者打包的服务,这种服务模式类似于一个AI主题商城,所有的用户都可以通过应用程序编程接口的方式来接入使用AIaaS平台提供的一种或者多种人工智能服务。

[0061] 例如,其中的一种人工智能云服务可以为问答服务,即云端的服务器封装有本申请实施例提供的问答处理的程序。用户通过终端(运行有客户端,例如智能问答客户端)调用云服务中的问答服务,以使部署在云端的服务器调用封装的问答处理的程序,进行基于语义相似的召回处理以及基于文本相似的召回处理,得到对应的第一问题语句以及第二问题语句,基于问题语句、第一问题语句以及第二问题语句,确定出响应问题语句的答案语句,例如,对于智能问答应用,进行基于语义相似的召回处理以及基于文本相似的召回处理,得到对应的第一问题语句以及第二问题语句,基于问题语句、第一问题语句以及第二问题语句,确定出对应问题语句的答案语句,以响应针对问题语句的问答请求,以便用户快速且准确地得到想要的答案。

[0062] 参见图1,图1是本申请实施例提供的智能问答系统10的应用场景示意图,终端200通过网络300连接服务器100,网络300可以是广域网或者局域网,又或者是二者的组合。

[0063] 终端200(运行有客户端,例如智能问答客户端)可以被用来获取针对问题语句的

问答请求,例如,用户在客户端中输入问题语句后,例如,中文问题:“为什么游戏这么卡顿?”,则终端自动获取针对问题语句的问答请求,其问答请求包括中文问题“为什么游戏这么卡顿?”。

[0064] 在一些实施例中,终端中运行的客户端中可以植入有问答插件,用以在客户端本地实现问答处理方法。例如,终端200获取针对问题语句的问答请求后,调用问答插件,以实现问答处理方法,进行基于语义相似的召回处理以及基于文本相似的召回处理,得到对应的第一问题语句以及第二问题语句,基于问题语句、第一问题语句以及第二问题语句,确定出响应问题语句的答案语句,例如,对于智能问答应用,用户输入中文问题:“为什么游戏这么卡顿?”后,则终端自动获取针对问题语句的问答请求,并得到响应问题语句的答案语句,例如中文答案:“玩游戏的人多了”;用户输入中文问题:“Why the game is lagging”后,则终端自动获取针对问题语句的问答请求,并得到响应问题语句的答案语句,例如英文答案:“There are more people playing games”。

[0065] 在一些实施例中,终端200获取针对问题语句的问答请求后,调用服务器100的问答接口(可以提供为云服务的形式,即问答服务),服务器100问答处理方法,进行基于语义相似的召回处理以及基于文本相似的召回处理,得到对应的第一问题语句以及第二问题语句,基于问题语句、第一问题语句以及第二问题语句,确定出响应问题语句的答案语句,以响应针对问题语句的问答请求,例如,对于智能问答应用,用户在智能问答应用中输入中文问题:“为什么游戏这么卡顿?”后,则终端自动获取针对问题语句的问答请求,并发送针对问题语句的问答请求,服务器接收到针对问题语句的问答请求后,进行基于语义相似的召回处理以及基于文本相似的召回处理,得到中文答案:“玩游戏的人多了”,并将该中文答案反馈至智能问答应用,以在终端上显示中文答案:“玩游戏的人多了”。

[0066] 下面说明本申请实施例提供的用于问答处理的电子设备的结构,参见图2,图2是本申请实施例提供的用于问答处理的电子设备500的结构示意图,以电子设备500是服务器为例说明,图2所示的用于问答处理的电子设备500包括:至少一个处理器510、存储器550以及至少一个网络接口520。电子设备500中的各个组件通过总线系统530耦合在一起。可理解,总线系统530用于实现这些组件之间的连接通信。总线系统530除包括数据总线之外,还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见,在图2中将各种总线都标为总线系统530。

[0067] 处理器510可以是一种集成电路芯片,具有信号的处理能力,例如通用处理器、数字信号处理器(DSP, Digital Signal Processor),或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等,其中,通用处理器可以是微处理器或者任何常规的处理器等。

[0068] 存储器550包括易失性存储器或非易失性存储器,也可包括易失性和非易失性存储器两者。其中,非易失性存储器可以是只读存储器(ROM, Read Only Memory),易失性存储器可以是随机存取存储器(RAM, Random Access Memory)。本申请实施例描述的存储器550旨在包括任意适合类型的存储器。存储器550可选地包括在物理位置上远离处理器510的一个或多个存储设备。

[0069] 在一些实施例中,存储器550能够存储数据以支持各种操作,这些数据的示例包括程序、模块和数据结构或者其子集或超集,下面示例性说明。

[0070] 操作系统551,包括用于处理各种基本系统服务和执行硬件相关任务的系统程序,例如框架层、核心库层、驱动层等,用于实现各种基础业务以及处理基于硬件的任务;

[0071] 网络通信模块553,用于经由一个或多个(有线或无线)网络接口520到达其他计算设备,示例性的网络接口520包括:蓝牙、无线相容性认证(WiFi)、和通用串行总线(USB, Universal Serial Bus)等;

[0072] 在一些实施例中,本申请实施例提供的问答处理装置可以采用软件方式实现,例如,可以是上文所述的终端中的问答插件,可以是上文所述的服务器中问答服务。当然,不局限于此,本申请实施例提供的问答处理装置可以提供为各种软件实施例,包括应用程序、软件、软件模块、脚本或代码在内的各种形式。

[0073] 图2示出了存储在存储器550中的问答处理装置555,其可以是程序和插件等形式的软件,例如问答插件,并包括一系列的模块,包括第一召回模块5551、第二召回模块5552、匹配模块5553、预处理模块5554以及训练模块5555;其中,第一召回模块5551、第二召回模块5552、匹配模块5553、预处理模块5554用于实现本申请实施例提供的问答处理功能,训练模块5555用于训练用于问答处理的神经网络模型。

[0074] 如前,本申请实施例提供的问答处理方法可以由各种类型的电子设备实施。参见图3A,图3A是本申请实施例提供的问答处理方法的流程示意图,结合图3A示出的步骤进行说明。

[0075] 在下面步骤中,待召回问题语句可以是文本库中的文本,该文本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的文本,例如网页上最新报道的新闻、公众号上的文本。

[0076] 在下面步骤中,多个表示至少两个;语义向量的概念是将一个单词映射到多维语义空间,成为多维语义空间中的一个点(值),该点包括该单词的多维特征,即编码过程就是从多维语义空间查询到与单词对应的点。

[0077] 在步骤101中,对问题语句进行基于跨语言的编码处理,得到融合语言信息的多个语义向量。

[0078] 其中,终端自动向服务器发送问题语句,服务器接收到问题语句后,对问题语句进行基于跨语言的编码处理,得到融合语言信息的多个语义向量,以便后续基于语义向量进行基于语义相似的召回处理。

[0079] 在一些实施例中,对问题语句进行基于跨语言的编码处理之前,方法还包括:对待响应问题语句进行变换处理,得到用于进行编码处理的问题语句。

[0080] 作为获取待响应问题语句的示例,在用户通过终端的输入界面输入待响应问题语句(即用户输入的原始的问题语句),输入完成后,终端基于用户输入的待响应问题语句,自动生成针对待响应问题语句的问答请求(包括待响应问题语句),并向服务器发送该针对待响应问题语句的问答请求。服务器接收到该针对待响应问题语句的问答请求后,先对待响应问题语句进行预处理,将预处理结果作为用于进行编码处理的问题语句,以进行后续的编码处理。

[0081] 在一些实施例中,对待响应问题语句进行预处理,将预处理结果作为用于进行编码处理的问题语句,包括:对待响应问题语句进行基于跨语言的改写处理,得到与待响应问题语句不同的问题语句;对与待响应问题语句不同的问题语句进行词形归一化处理,得到

用于进行编码处理的问题语句。

[0082] 其中,对待响应问题语句进行变换处理包括改写处理以及词形归一化处理。通过改写处理,以便在待响应问题语句中融入不同语言的词语,使得得到的问题语句的通用性更强;通过词形归一化处理,提高多语言在向量空间的对齐效果,以便进行后续的编码处理。

[0083] 在一些实施例中,基于领域词典进行匹配和/或基于多语言近义词词典进行匹配,以得到与待响应问题语句不同的问题语句。

[0084] 例如,将领域词典与待响应问题语句中的词语进行匹配处理,得到与待响应问题语句中的词语对应的匹配词,基于匹配词对待响应问题语句中的词语进行替换处理,得到替换后的问题语句,将替换后的问题语句作为与待响应问题语句不同的问题语句,例如,待响应语句为“akm的配件?”,在领域词典中匹配到与“akm”匹配的“ak”,则基于“ak”,得到替换后的问题语句为“ak的配件?”。

[0085] 例如,将多语言近义词词典与待响应问题语句中的词语进行匹配处理,得到与待响应问题语句中的词语对应的匹配词,基于匹配词对待响应问题语句中的词语进行替换处理,得到替换后的问题语句,将替换后的问题语句作为与待响应问题语句不同的问题语句,例如,待响应语句为“akm的配件?”,在多语言近义词词典中匹配到与“配件”匹配的“accessories”,则基于“accessories”,得到替换后的问题语句为“akm的accessories?”。

[0086] 例如,对待响应问题语句进行基于跨语言的改写处理,得到与待响应问题语句不同的问题语句,包括:将领域词典与待响应问题语句中的词语进行匹配处理,得到与待响应问题语句中的词语对应的第一匹配词;基于第一匹配词对待响应问题语句中匹配的词语进行替换处理,得到替换后的问题语句;当在多语言近义词词典匹配到与替换后的问题语句中的词语对应的第二匹配词时,基于第二匹配词对替换的问题语句中匹配的词语进行替换处理,得到与待响应问题语句不同的问题语句。

[0087] 在一些实施例中,对与待响应问题语句不同的问题语句进行词形归一化处理(对语句中的词形进行归一化处理,以统一词形,实现词形标准化),得到用于进行编码处理的问题语句,包括:确定与待响应问题语句不同的问题语句中的多个词语分别对应的词频;将多个词语拆分为与多个词语一一对应的多个字符序列;迭代执行以下处理:基于多个词语分别对应的词频,确定多个字符序列中的连续字符对的出现频率;将最高出现频率所对应的连续字符对进行合并(即连接、拼接),将合并结果作为新字符来更新字符序列;当更新的字符序列中的连续字符对的出现频率为1时停止迭代处理。

[0088] 例如,待响应问题语句为“The building A is lower than the building B, but it is the latest development, it has the widest view……”,对待响应问题语句中的词语进行统计,得到词语分别对应的词频,例如,{‘lower’:2, ‘newest’:6, ‘widest’:3, ‘low’:5},先该多个词语拆分为与多个词语一一对应的多个字符序列,即‘lower’拆分为字符序列‘l o w e r’、‘newest’拆分为字符序列‘n e w e s t’、‘widest’拆分为字符序列‘w i d e s t’、‘low’拆分为字符序列‘l o w’;基于‘lower’、‘newest’、‘widest’、‘low’分别对应的词频,确定‘l o w e r’、‘n e w e s t’、‘w i d e s t’、‘l o w’中的连续字符对(子字符序列中连续出现的两个字符)的出现频率,其中,连续字符对(‘s’与‘t’)的出现频率最高为9;将最高出现频率所对应的连续字符对进行合并(‘s’与‘t’),将合并得到的

结果(‘st’)作为新字符来更新字符序列,得到{‘n e w e st’:6,‘l o w e r’:2,‘w i d e st’:3,‘l o w’:5},继续上述的更新字符序列的操作,直至更新的字符序列中的连续字符对的出现频率为1,最终更新的字符序列即为用于进行编码处理的问题语句,通过连续字符对进行更新迭代,可以通过高词频的字符(即较少的字符)准确表征词语,例如通过“wi d”字符即可表征“widest”词语。

[0089] 在一些实施例中,对待响应问题语句进行基于跨语言的改写处理之前,方法还包括:对待响应问题语句进行分词处理,得到对应待响应问题语句的词序列(词语所构成的序列);对词序列进行实体词识别处理,得到待响应问题语句的实体词(表示实体的词语);基于待响应问题语句的实体词对词序列进行合并处理,将合并结果用于进行改写处理。

[0090] 例如,在对待响应问题语句进行基于跨语言的改写处理之前,还可以对待响应问题语句进行通用预处理,例如去除待响应问题语句中的无用字符(对待响应问题语句的内容没有影响的字符),例如标点符号、表情符、乱码字符等停用词,从而去除待响应问题语句中的不重要信息。在去除无用字符后,还可以通过分词器(用于分词的模型)进行分词处理,得到对应待响应问题语句的词序列,例如,待响应问题语句为“我喜欢玩诸葛连弩和貂蝉”,则对应的词序列为“我喜欢玩诸葛连弩和貂蝉”;通过识别器(用于实体词识别的模型)对词序列进行实体词识别,得到待响应问题语句的实体词,例如,实体词为“诸葛连弩”和“貂蝉”,然后,对词序列进行合并处理,得到合并结果,例如,合并结果为“我喜欢玩【诸葛连弩】和【貂蝉】”,以便经过通用预处理后的待响应问题语句进行后续改写处理。

[0091] 在一些实施例中,对待响应问题语句进行分词处理,得到对应待响应问题语句的词序列,包括:对待响应问题语句进行语言识别处理,得到待响应问题语句的语言类型;基于待响应问题语句的语言类型,调用与语言类型对应的分词器进行分词处理,得到对应待响应问题语句的词序列;对词序列进行实体词识别处理,得到待响应问题语句的实体词,包括:对词序列进行通用实体词识别处理,得到通用实体词;对词序列进行领域实体词识别处理,得到领域实体词;将通用实体词以及领域实体词作为待响应问题语句的实体词。

[0092] 承接上述示例,为了进行个性化分词处理,可以先识别待响应问题语句的语言类型,然后基于待响应问题语句的语言类型,调用与语言类型对应的分词器进行个性化分词,从而提高分词的准确性,例如,中文使用结巴(jieba)分词器,日语使用mecab分词器,其余语言使用moses分词器。

[0093] 其中,为了准确识别实体词,可以通过两种实体词识别方法对词序列进行识别,例如,对词序列进行通用实体词识别处理,得到通用实体词(包括属性的词语),例如人名、地名、机构名等词语;对词序列进行领域实体词识别处理,得到领域实体词,例如游戏领域中的领域实体词包括“诸葛连弩”等固定领域出现的词语。

[0094] 参见图3B,图3B是本发明实施例提供的一个可选的流程示意图,在一些实施例中,图3B示出步骤101可以通过图3B示出的步骤1011至步骤1012实现:在步骤1011中,通过多个编码网络中的任一编码网络,基于不同于问题语句的语言,对问题语句进行编码处理,得到对应编码网络的语义向量;在步骤1012中,将多个对应编码网络的语义向量作为多个语义向量。

[0095] 其中,用于问答处理的神经网络模型包括多个编码网络。例如神经网络模型包括编码网络1、编码网络2、编码网络3,则编码网络1基于中文,对问题语句进行编码,得到对应

编码网络1的语义向量;编码网络2基于日文,对问题语句进行编码,得到对应编码网络2的语义向量;编码网络3基于英文,对问题语句进行编码,得到对应编码网络3的语义向量,从而得到多种语义向量,每个语义向量对应一个编码网络。

[0096] 在一些实施例中,用于问答处理的神经网络模型包括解码网络;神经网络模型的训练方法包括:通过多个编码网络中的任一编码网络,基于不同于问题语句样本的语言,对问题语句样本进行编码处理,得到对应编码网络的语义向量;通过解码网络对对应编码网络的语义向量进行解码处理,得到对应问题语句样本的跨语言文本(与问题语句样本的语言不通的文本,例如问题语句是中文文本,即“我喜欢看书”,则跨语言文本是对应问题语句的英文文本,即“I like read”);基于跨语言文本以及问题语句样本的跨语言文本标签(即需要基于问题语句样本所生成的跨语言文本的语言标签,例如跨语言文本标签为英文,而实际基于问题语句样本所生成的跨语言文本为日文),构建神经网络模型的损失函数;更新神经网络模型的参数直至损失函数收敛,将损失函数收敛时神经网络模型的更新的参数,作为训练后的神经网络模型的参数。

[0097] 例如,根据跨语言文本以及跨语言文本标签,确定神经网络模型的损失函数的值后,可以判断神经网络模型的损失函数的值是否超出预设阈值,当神经网络模型的损失函数的值超出预设阈值时,基于神经网络模型的损失函数确定神经网络模型的误差信号,将误差信息在神经网络模型中反向传播,并在传播的过程中更新各个层的模型参数。

[0098] 这里,对反向传播进行说明,将训练样本数据输入到神经网络模型的输入层,经过隐藏层,最后达到输出层并输出结果,这是神经网络模型的前向传播过程,由于神经网络模型的输出结果与实际结果有误差,则计算输出结果与实际值之间的误差,并将该误差从输出层向隐藏层反向传播,直至传播到输入层,在反向传播的过程中,根据误差调整模型参数的值;不断迭代上述过程,直至收敛。

[0099] 在一些实施例中,训练方法还包括:对对应编码网络的语义向量进行分类处理(语义向量为多维向量,通过映射整合输出一个值,然后根据该值可以确定类别),得到对应问题语句样本的预测类别;基于预测类别以及问题语句样本的类别标签,构建神经网络模型的类别损失函数;基于跨语言文本以及问题语句样本的跨语言文本标签,构建神经网络模型的损失函数,包括:基于跨语言文本以及问题语句样本的跨语言文本标签,构建神经网络模型的跨语言损失函数;对类别损失函数以及跨语言损失函数进行加权求和,得到神经网络模型的损失函数。

[0100] 承接上述示例,基于不同于问题语句样本的语言,对问题语句样本进行编码处理,得到对应编码网络的语义向量,通过解码网络对对应编码网络的语义向量进行解码处理,得到对应问题语句样本的跨语言文本,基于跨语言文本以及问题语句样本的跨语言文本标签,构建神经网络模型的跨语言损失函数,对对应编码网络的语义向量进行分类处理,得到对应问题语句样本的预测类别,基于预测类别以及问题语句样本的类别标签,构建神经网络模型的类别损失函数,对类别损失函数以及跨语言损失函数进行加权求和,得到神经网络模型的损失函数,更新神经网络模型的参数直至损失函数收敛,将损失函数收敛时神经网络模型的更新的参数,作为训练后的神经网络模型的参数,使得训练后的神经网络中的编码网络能够进行更加准确的编码处理。

[0101] 在步骤102中,基于多个语义向量中的任一语义向量,对不同语言的待召回问题语

句进行基于语义相似的召回处理,得到对应语义向量的多个第一问题语句。

[0102] 例如,在服务器得到多个语义向量后,可以对问答库进行召回处理,其中,问答库中包括不同语言的待召回问题语句以及对应的答案语句,从而可以进行语义相似召回,以召回与问题语句相似的第一问题语句。

[0103] 参见图3C,图3C是本发明实施例提供的一个可选的流程示意图,在一些实施例中,图3C示出步骤102可以通过图3C示出的步骤1021至步骤1022实现:在步骤1021中,确定语义向量分别与不同语言的待召回问题语句的语义向量之间的语义相似度;在步骤1022中,基于语义向量与待召回问题语句的语义向量之间的语义相似度,对待召回问题语句进行降序排序,将降序排序结果中排序在前的多个待召回问题语句作为多个第一问题语句。

[0104] 例如,通过各种相似度算法(例如余弦相似度、欧几里得距离、汉明距离等),确定问题语句的语义向量与待召回问题语句1的语义向量之间的语义相似度为0.9、语义向量与待召回问题语句2的语义向量之间的语义相似度为0.7、语义向量与待召回问题语句3的语义向量之间的语义相似度为0.8、语义向量与待召回问题语句4的语义向量之间的语义相似度为0.4,对待召回问题语句进行降序排序,选取前2个待召回问题语句作为多个第一问题语句,即待召回问题语句1和待召回问题语句3作为第一问题语句。

[0105] 在步骤103中,基于问题语句对待召回问题语句进行基于文本相似的召回处理,得到对应问题语句的多个第二问题语句。

[0106] 例如,服务器还可以直接根据问题语句,对问答库中的待召回问题语句进行基于文本相似的召回处理,以得到不同于第一问题语句的第二问题语句,从而从不同角度召回得到多种问题语句,以进行后续的匹配处理。

[0107] 在一些实施例中,基于问题语句对待召回问题语句进行基于文本相似的召回处理,得到对应问题语句的多个第二问题语句,包括:对问题语句中的关键词(即用于影响相似度计算的预先设置的词语)以及待召回问题语句进行匹配,得到关键词在待召回问题语句中的出现频率;基于关键词在待召回问题语句中的出现频率,确定问题语句与待召回问题语句之间的文本相似度;基于问题语句与待召回问题语句之间的文本相似度,对待召回问题语句进行降序排序,将降序排序结果中排序在前的多个待召回问题语句作为多个第二问题语句。

[0108] 例如,可以先识别问题语句中的关键词,将问题语句中的关键词与待召回问题语句进行匹配,得到关键词在待召回问题语句中的出现频率,例如关键词1在待召回问题语句1的出现频率为6、关键词2在待召回问题语句1的出现频率为10,待召回问题语句中有26个词语,则综合关键词1和关键词2,得到问题语句与待召回问题语句1之间的文本相似度为 $0.6 = (6+10) / 26$,当问题语句与待召回问题语句2之间的文本相似度为0.2、问题语句与待召回问题语句3之间的文本相似度为0.7,选取前2个待召回问题语句作为多个第二问题语句,即待召回问题语句1和待召回问题语句3作为第二问题语句。即仅通过关键词,即可快速获取问题语句与待召回问题语句之间的相似度,从而能够快速召回语句(第二问题语句),以进行后续的文本匹配任务。

[0109] 在一些实施例中,基于问题语句对待召回问题语句进行基于文本相似的召回处理,得到对应问题语句的多个第二问题语句,包括:基于问题语句中的词语,确定问题语句中的词语的权重;对问题语句中的词语进行哈希处理,得到对应词语的哈希值;基于词语的

权重,对对应词语的哈希值进行加权求和,得到问题语句的哈希签名;基于问题语句的哈希签名以及待召回问题语句的哈希签名,确定问题语句与待召回问题语句之间的文本相似度;基于问题语句与待召回问题语句之间的文本相似度,对待召回问题语句进行降序排序,将降序排序结果中排序在前的多个待召回问题语句作为多个第二问题语句。

[0110] 例如,为每一个问题语句中的词语设置5个级别(1-5等)的权重(权重可以是这个词语在问题语句中出现的次数)。例如,问题语句为“CSDN博客结构之法算法之道的作者July”,分词后为:“CSDN博客结构之法算法之道的作者July”,然后为每个词语赋予权值:“CSDN(4)博客(5)结构(3)之(1)法(2)算法(3)之(1)道(2)的(1)作者(5)July(5)”,其中,括号里的数字代表这个单词在整条语句中的重要程度,即权重,数字越大代表越重要。

[0111] 通过哈希函数计算各个词语的哈希值,哈希值为二进制数0、1组成的字节签名。例如“CSDN”的哈希值为Hash(CSDN):100101,“博客”的哈希值为Hash(博客):“101011”。在哈希值的基础上,给所有词语进行加权,即 $W = \text{Hash} * \text{权重}$,且遇到1则哈希值和权值正相乘,遇到0则hash值和权值负相乘。例如“CSDN”的哈希值“100101”加权得到: $W(\text{CSDN}) = 100101 * 4 = 4 -4 -4 4 -4 4$,“博客”的哈希值“101011”加权得到: $W(\text{博客}) = 101011 * 5 = 5 -5 5 -5 5 5$ 。将上述各个词语的加权结果累加,变成只有一个序列串,即哈希签名,以前两个特征向量举例,例如“CSDN”的“4 -4 -4 4 -4 4”和“博客”的“5 -5 5 -5 5 5”进行累加,得到“4+5 -4+-5 -4+5 4+-5 -4+5 4+5”,得到“9 -9 1 -1 1”。

[0112] 其中,与上述哈希签名的获取方式类似,待召回问题语句的哈希签名的计算方式如下:基于待召回问题语句中的词语,确定待召回问题语句中的词语的权重;对待召回问题语句中的词语进行哈希处理,得到对应词语的哈希值;基于词语的权重,对对应词语的哈希值进行加权求和,得到待召回问题语句的哈希签名。

[0113] 最后,在服务器计算得到问题语句的哈希签名以及待召回问题语句的哈希签名后,基于问题语句的哈希签名以及待召回问题语句的哈希签名,通过各种相似度算法(例如余弦相似度、欧几里得距离等)确定问题语句与待召回问题语句之间的文本相似度,基于问题语句与待召回问题语句之间的文本相似度,对待召回问题语句进行降序排序,将降序排序结果中排序在前的多个待召回问题语句作为多个第二问题语句,例如,将前两个待召回问题语句作为第二问题语句。

[0114] 在步骤104中,将问题语句分别与多个第一问题语句以及多个第二问题语句进行文本匹配处理,得到匹配的目标问题语句,将目标问题语句对应的答案语句,作为用于响应问题语句的答案语句。

[0115] 其中,通过分类器对问题语句与多个第一问题语句进行分类(映射整合)处理,输出多个概率值,该多个概率值表征问题语句分别与多个第一问题语句之间的相似度,通过分类器对问题语句与多个第二问题语句进行分类处理,得到问题语句分别与多个第二问题语句之间的相似度,将最大相似度所对应的第一问题语句或第二问题语句作为目标问题语句,并将目标问题语句对应的答案语句,作为用于响应问题语句的答案语句,仅通过分类处理,即可从多个第一问题语句以及多个第二问题语句中挑选出目标问题语句,从而快速查找到用于响应问题语句的答案语句,用户即可快速得到想要的答案,无需浪费时间等待回复。

[0116] 在一些实施例中,在将问题语句与多个第一问题语句以及多个第二问题语句进行

文本匹配处理之前,方法还包括:基于问题语句与第一问题语句之间的相似度、以及问题语句与第二问题语句之间的相似度,对第一问题语句以及第二问题语句进行降序排序,将降序排序结果中排序在前的第一问题语句或者第二问题语句作为用于进行文本匹配的召回问题语句。

[0117] 例如,服务器可以基于语义相似召回确定问题语句与多个第一问题语句之间的第一相似度,基于文本相似召回确定问题语句与多个第二问题语句之间的第二相似度,并基于第一问题语句对应的权重以及第二问题语句对应的权重,对第一相似度以及第二相似度进行加权,确定问题语句与第一问题语句之间的综合得分(综合相似度)、问题语句与第二问题语句之间的综合得分,基于问题语句与第一问题语句之间的综合得分、问题语句与第二问题语句之间的综合得分,对第一问题语句以及第二问题语句进行降序排序,将降序排序结果中排序在前的第一问题语句或者第二问题语句作为用于进行文本匹配的召回问题语句,从而减小后续需要进行文本匹配的问题语句,以加快匹配进程。

[0118] 在一些实施例中,将与匹配得到的问题语句对应的答案语句,作为用于响应问题语句的答案语句,包括:当在问题答案库(问答库)中查询到与匹配得到的召回问题语句对应的答案语句,且对应的答案语句的语言与问题语句的语言一致时,将匹配得到的召回问题语句对应的答案语句,作为用于响应问题语句的答案语句;当在问题答案库中查询到与匹配得到的召回问题语句对应的答案语句,且对应的答案语句的语言与问题语句的语言不一致时,获取答案语句对应问题语句的语言的翻译结果,并作为用于响应问题语句的答案语句。

[0119] 例如,在服务器获得匹配得到的问题语句后,当在问答库中查询到与匹配得到的召回问题语句对应的答案语句,且对应的答案语句的语言与问题语句的语言一致时,将匹配得到的召回问题语句对应的答案语句,作为用于响应问题语句的答案语句,例如,问题语句的语言为中文,召回问题语句对应的答案语句为中文,则直接将召回问题语句对应的答案语句作为用于响应问题语句的答案语句;当在问题答案库中查询到与匹配得到的召回问题语句对应的答案语句,且对应的答案语句的语言与问题语句的语言不一致时,则可以翻译对应的答案语句,使得翻译结果的语言与问题语句的语言一致,例如,问题语句的语言为中文,召回问题语句对应的答案语句为英文,则先将召回问题语句对应的答案语句翻译为中文,并将翻译后的答案语句作为用于响应问题语句的答案语句。

[0120] 下面,将说明本申请实施例在一个实际的应用场景中的示例性应用。

[0121] 在智能问答系统中,FAQ是其中重要的一个分支,其通过检索问答(QA, Question Answer)库里与用户提问最相近的问题来完成问答。例如,不同国家的用户在驾驶的过程,需要智能问答系统进行问题解答,FAQ将要面临各国语言所带来的问答压力;随着游戏的出海,FAQ也将要面临各国语言所带来的问答压力。

[0122] 为解决智能问答中遇到的多语言问题,相关技术提出两种解决方案,分别为:1)通过翻译的方案,即通过将用户提问的问题进行翻译,以使得单语言问答机器人支持到多个语言;2)通过为每个语言建设一套FAQ检索系统及QA库,从而实现多语言的支持。

[0123] 虽然上述现有技术能够实现多语言的问答,但是上述方案1)依赖于翻译的质量,小语种、口语化的问题翻译的误差会不断积累,导致最终模型的训练和预测偏差较大;上述方案2)需要为每一个语言建设一套FAQ检索系统,从而加大开发周期和维护成本,同时还带

来为每个语言建设独立QA库的标注成本,且无法提供快速接入新语言的能力。

[0124] 为解决上述技术问题,本申请实施例基于文本匹配的技术,提出一种统一多语言问答解决方案,通过多语言预处理、跨语言的语义相似召回及文本相似召回、跨语言文本匹配等技术方案实现单模型跨语言QA能力,无需为每个语言构建一种QA能力及QA库,从而减少为每个语言建设独立QA库的标注成本,减少为每一个语言建设一套FAQ检索能力所带来的开发周期和维护成本,并解决低资源及冷启动的问题。

[0125] 下面具体说明本申请实施例提出的跨语言问答技术,如图4所示,该跨语言问答技术包括三个部分:跨语言预处理(预处理层)、跨语言召回(召回层)、跨语言精排(精排层)。下面具体介绍该三个部分:

[0126] A、预处理部分(预处理层)

[0127] 其中,预处理部分包括通用预处理组件、搜索词(Query)改写、通用词形归一化。其中,Query即为用户输入的问题。

[0128] 1) 通用预处理

[0129] 如图5所示,此部分提供对文本的预处理,处理流程如下:

[0130] a) 无用字符(对Query内容没有影响的字符)包括:标点符号、表情符、乱码字符等。

[0131] b) 分词部分针对中文等连写语言使用特有分词器,本申请实施例中的中文使用结巴(jieba)分词器,日语使用mecab分词器,其余语言使用moses分词器。其中,moses分词器支持任何语言的分词,jieba分词器则是一款在中文分词上表现优异的分词组件,而mecab分词器是基于条件随机场(CRF,Conditional Random Field)的一个日文分词系统。

[0132] c) 通用实体词识别:识别的通用实体词包括:人名、地名、机构名等。本申请实施例调用Stanford stanza提供的多种语言的通用实体识别。其中,stanza是目前涵盖语言数量最多,且当前最优结果(SOTA,State of the Art)完全基于神经网络框架的自然语言处理开源库。

[0133] d) 领域实体词识别:通过前缀(trie)树识别领域的实体词。例如,基于整理得到的游戏领域的实体词,并通过trie树匹配得到问题中的领域词。

[0134] e) 混合粒度分词结果生成:结合通用实体词和领域实体词的结果,对分词结果中被切分开的实体词进行合并,以形成混合粒度分词的结果。

[0135] 2) Query改写(通过Query改写以实现Query扩展)

[0136] 利用领域别名词典、多语言近义词词典对Query进行改写。以Query为“a km的配件?”为例,通过领域别名词典将“akm的配件?”改写为“ak的配件?”,通过多语言近义词词典将“akm的配件?”改写为“akm的accessories?”。

[0137] 3) 通用词形归一化

[0138] 为解决多语言情况下词形多样性导致的词典增大及oov问题,采用字节对编码(BPE,Byte Pair Encoder)算法,将分词后得到的词语拆分为字符序列,并在末尾添加后缀</w>,然后进行以下处理:

[0139] a、统计词语的词频,例如,词语“low”的词频是5,则将其改写为“l o w</w>”:5,其中,“l o w”为字符序列;

[0140] b、统计每一个连续字符对的出现频率,选择最高频率所对应的连续字符对合并成新的子词(subword),作为字符序列中的一个新字符。

[0141] c、重复步骤a和步骤b,直到subword词典大小达到预设值或连续字符对的出现频率的最高频为1。

[0142] 因此,通过该统计方法将单词分解后并合并成subword的方法,可以将词语拆解成常见的前缀、后缀等,使得能对曲折变化多的语言用较小的词语来表示,以提高多语言在向量空间的对齐效果。

[0143] B、召回部分

[0144] 其中,召回部分包括跨语言语义相似召回、跨语言文本相似召回以及多路召回合并策略。

[0145] 1) 跨语言语义相似召回

[0146] 如图6所示,跨语言语义相似召回的网络结构包括编码网络(Encoder)、解码网络(Decoder)以及分类器(classification)。其中,编码网络可不固定,可以为CNN、长短期记忆网络(LSTM,Long Short-Term Memory)、RNN、变换器(Transformer)等。为实现编码网络的跨语言能力,通过两个下游任务来完成。第一个下游任务为翻译任务,训练集采用平行语料,即在解码网络加入语种特征(语言标签),而编码网络则没有加入语种特征,这样做可以使得编码网络学习到的句向量不受语言限制。第二个下游任务是分类任务,QA库里的语料天然簇形式的,把每个簇作为一种类别来进行分类,从而使得编码网络学习到簇信息。本申请实施例将Encoder通过不同的网络结构来学习不同的特征,从而实现多路语义相似召回(一种编码网络输出的语义向量对应一路语义召回)。

[0147] 其中,在进行跨语言语义相似召回应用时,将使用编码网络;在进行训练时,使用编码网络、解码网络以及分类器。

[0148] 2) 跨语言文本相似召回

[0149] 利用预处理部分的词形归一化进行扩展后的Query,采用局部敏感哈希(simhash)的方法,使用simhash构建问题的哈希签名,即基于问题中的词语,确定问题中的词语的权重,对问题中的词语进行哈希处理,得到对应词语的哈希值,基于词语的权重,对对应词语的哈希值进行加权求和,得到问题的哈希签名;使用simhash构建问答库中的待召回问题的哈希签名,即基于待召回问题中的词语,确定待召回问题中的词语的权重,对待召回问题中的词语进行哈希处理,得到对应词语的哈希值,基于词语的权重,对对应词语的哈希值进行加权求和,得到待召回问题的哈希签名,利用faiss进行汉明距离检索,确定问题与待召回问题之间的文本相似度,基于问题与待召回问题之间的文本相似度,从而实现文本相似召回。其中,faiss是针对聚类和相似性搜索库,为稠密向量提供高效相似度搜索和聚类的近似近邻搜索库,包含多种搜索任意大小向量集的算法。

[0150] 3) 多路召回合并策略

[0151] 对多路召回的结果(包括语义相似召回的结果和文本相似召回的结果),将各路召回的结果进行加权以构建综合得分(merge_score),根据merge_score的排序结果取前16个结果。假设有n路召回,则使用到的特征为score_1,rank_1,...,score_n,rank_n,其中score_i表示第i路的得分,rank_i表示第i路的顺序。最后利用逻辑回归进行训练得到权重。

[0152] C、精排部分(对多个文本进行筛选的过程,例如从多个问题中筛选出一个目标问题)

[0153] 如图7所示,该部分包括编码器以及分类器,输入为句子对(sentence1,sentence2),其中,句子1(sentence1)为Query,句子2(sentence2)为召回的问题(召回问题),此任务采用分类任务进行,即通过映射整合,输出一个概率值,该概率值表征该句子对相似度得分。

[0154] 关于训练数据构建部分,对于FAQ的QA库主要记录的是很多高频的知识点(知识点包括问题和答案),这些高频的知识点对应的问法并不唯一,即问答库的结构为一个问题集合对应一个答案,即有以下三种数据类型,如表1所示:

[0155] 表1

[0156]	类型	符号	描述
	标准问	Std_q	经过人工整理针对某一问题的标准问法
	答案	Answer	FAQ针对某一个标准问的标准答案
	相似问	q1,q2,……	跟标准问相似的query集合

[0157] 如上所述,文本匹配需要的数据格式为句子对的形式(sentence1,sentence2),因此在构建训练数据集时,需要将FAQ知识点形式的数据转化为句子对形式。其中,训练集包含两个部分,一部分是正例的构建,另一部分是负例的构建,正例的句子对可以理解为这两个句子是相似的,负例则反之。这两部分数据的构建方式直接影响到最终的结果。

[0158] 其中,正例的构建方式是:对于知识点中的每条相似问,会在同语言的相似问集合中两两组合成句子对,对同语言相似问多的知识点进行裁剪(即当同语言相似问题很多时,从很多的同语言相似问题中挑选出几个相似问题),另外为了支持跨语言匹配,会在其他语言的相似问集合中召回部分相似的和随机选取部分相似问来构建句子对。负例的构建方式是:对于知识点(簇)的每条相似问主要有两类的数据来构建句子对,一是从按照各知识点的样本比例从其他知识点随机选择的相似问;二是从其他知识点通过句向量的跨语言召回的相似问来构建数据集,其中,跨语言的召回的时候会控制各语言的召回比例。

[0159] 综上,本申请实施例提供的基于文本匹配的跨语言问法技术,在游戏问答实际场景中,效果均优于翻译方案、单语言引擎方案。关于低资源/冷启动情况:只提供1个语种语料库,在6语种上效果比单语种各自建设语料库好,与翻译基线相当。关于支持定向语种效果提升:提供小语种15%语料,准确率提升8%~12%。

[0160] 如图8所示,在英文准确率测评上,本申请实施例提供的基于文本匹配的跨语言问法技术(多语引擎),明显优于竞品以及翻译引擎。如图9所示,在小语种准确率测评上,本申请实施例提供的基于文本匹配的跨语言问法技术(跨语引擎以及多语引擎),优于翻译引擎以及单语引擎。其中,跨语引擎仅适用单语语料(如英文语料)进行训练,而多语引擎是在单语语料的基础上,提供少量的其他语种语料进行训练。

[0161] 如图10所示,图10是本申请实施例提供的手游中的问答界面图,点击图10中的入口1001,可进入如图11所示的英文问答界面或者如图12所示的中文问答界面,用户可以在如图11所示的英文问答界面的输入框1101中输入问题,例如“game is lagging and freezing”,则智能问答系统返回英文的答案,例如“Please select an option below to proceed:”;用户可以在如图12所示的中文问答界面的输入框1201中输入问题,例如“玩的时候卡顿好严重”,则智能问答系统返回中文的答案,例如“请在下面选择一个选项以继续:”。

[0162] 本申请实施例中的智能问答系统并不局限于游戏,还可以应用于车联网等应用场景中,例如用户在驾驶的过程中,通过语音进行提问,智能问答系统基于用户提出的某语言的问题进行相应语言的解答。

[0163] 至此已经结合本申请实施例提供的服务器的示例性应用和实施,说明本申请实施例提供的问答处理方法。本申请实施例还提供问答处理装置,实际应用中,问答处理装置中的各功能模块可以由电子设备(如终端设备、服务器或服务器集群)的硬件资源,如处理器等计算资源、通信资源(如用于支持实现光缆、蜂窝等各种方式通信)、存储器协同实现。图2示出了存储在存储器550中的问答处理装置555,其可以是程序和插件等形式的软件,例如,软件C/C++、Java等编程语言设计的软件模块、C/C++、Java等编程语言设计的应用软件或大型软件系统中的专用软件模块、应用程序接口、插件、云服务等实现方式,下面对不同的实现方式举例说明。

[0164] 示例一、问答处理装置是移动端应用程序及模块

[0165] 本申请实施例中的问答处理装置555可提供为使用软件C/C++、Java等编程语言设计的软件模块,嵌入到基于Android或iOS等系统的各种移动端应用中(以可执行指令存储在移动端的存储介质中,由移动端的处理器执行),从而直接使用移动端自身的计算资源完成相关的箱式结构检测任务,并且定期或不定期地通过各种网络通信方式将处理结果传送给远程的服务器,或者在移动端本地保存。

[0166] 示例二、问答处理装置是服务器应用程序及平台

[0167] 本申请实施例中的问答处理装置555可提供为使用C/C++、Java等编程语言设计的应用软件或大型软件系统中的专用软件模块,运行于服务器端(以可执行指令的方式在服务器端的存储介质中存储,并由服务器端的处理器运行),服务器使用自身的计算资源完成相关的信息推荐任务。

[0168] 本申请实施例还可以提供为在多台服务器构成的分布式、并行计算平台上,搭载定制的、易于交互的网络(Web)界面或其他各用户界面(UI, User Interface),形成供个人、群体或单位使用的信息推荐平台等。

[0169] 示例三、问答处理装置是服务器端应用程序接口(API, Application Program Interface)及插件

[0170] 本申请实施例中的问答处理装置555可提供为服务器端的API或插件,以供用户调用,以执行本申请实施例的问答处理方法,并嵌入到各类应用程序中。

[0171] 示例四、问答处理装置是移动设备客户端API及插件

[0172] 本申请实施例中的问答处理装置555可提供为移动设备端的API或插件,以供用户调用,以执行本申请实施例的问答处理方法。

[0173] 示例五、问答处理装置是云端开放服务

[0174] 本申请实施例中的问答处理装置555可提供为向用户开发的目标检测云服务,供个人、群体或单位获取推荐列表。

[0175] 其中,问答处理装置555包括一系列的模块,包括第一召回模块5551、第二召回模块5552、匹配模块5553、预处理模块5554以及训练模块5555。下面继续说明本申请实施例提供的问答处理装置555中各个模块配合实现问答方案。

[0176] 第一召回模块5551,用于对问题语句进行基于跨语言的编码处理,得到融合语言

信息的多个语义向量;基于所述多个语义向量中的任一语义向量,对不同语言的待召回问题语句进行基于语义相似的召回处理,得到对应所述语义向量的多个第一问题语句;第二召回模块5552,用于对所述待召回问题语句进行基于文本相似的召回处理,得到对应所述问题语句的多个第二问题语句;匹配模块5553,用于将所述问题语句分别与所述多个第一问题语句以及所述多个第二问题语句进行文本匹配处理,得到匹配的目标问题语句;将所述目标问题语句对应的答案语句,作为用于响应所述问题语句的答案语句。

[0177] 在一些实施例中,所述问答处理装置555还包括:预处理模块5554,用于对待响应问题语句进行变换处理,得到用于进行所述编码处理的所述问题语句。

[0178] 在一些实施例中,所述预处理模块5554还用于对所述待响应问题语句进行基于跨语言的改写处理,得到与所述待响应问题语句不同的问题语句;对与所述待响应问题语句不同的问题语句进行词形归一化处理,得到用于进行所述编码处理的所述问题语句。

[0179] 在一些实施例中,所述预处理模块5554还用于将领域词典与所述待响应问题语句中的词语进行匹配处理,得到与所述待响应问题语句中的词语对应的第一匹配词;基于所述第一匹配词对所述待响应问题语句中的词语进行替换处理,得到替换后的问题语句;当多语言近义词词典匹配到与所述替换后的问题语句中的词语对应的第二匹配词时,基于所述第二匹配词对所述替换的问题语句进行替换处理,得到与所述待响应问题语句不同的问题语句。

[0180] 在一些实施例中,所述预处理模块5554还用于确定与所述待响应问题语句不同的问题语句中的多个词语;确定所述多个词语分别对应的词频;将所述多个词语拆分为与所述多个词语一一对应的多个字符序列;迭代执行以下处理:基于所述多个词语分别对应的词频,确定所述多个字符序列中的连续字符对的出现频率;将最高出现频率所对应的连续字符对进行合并,将合并结果作为新字符来更新所述字符序列;当所述更新的字符序列中的连续字符对的出现频率为1时停止所述迭代处理。

[0181] 在一些实施例中,所述预处理模块5554还用于对所述待响应问题语句进行分词处理,得到对应所述待响应问题语句的词序列;对所述词序列进行实体词识别处理,得到所述待响应问题语句的实体词;基于所述待响应问题语句的实体词对所述词序列进行合并处理,将合并结果用于进行所述改写处理。

[0182] 在一些实施例中,预处理模块5554还用于对所述待响应问题语句进行语言识别处理,得到所述待响应问题语句的语言类型;基于所述待响应问题语句的语言类型,调用与所述语言类型对应的分词器进行分词处理,得到对应所述待响应问题语句的词序列;对所述词序列进行通用实体词识别处理,得到通用实体词;对所述词序列进行领域实体词识别处理,得到领域实体词;将所述通用实体词以及所述领域实体词作为所述待响应问题语句的实体词。

[0183] 在一些实施例中,用于问答处理的神经网络模型包括多个编码网络;所述第一召回模块5551还用于通过所述多个编码网络中的任一所述编码网络,基于不同于所述问题语句的语言,对所述问题语句进行编码处理,得到对应所述编码网络的语义向量;将多个所述对应所述编码网络的语义向量作为所述多个语义向量。

[0184] 在一些实施例中,所述用于问答处理的神经网络模型包括解码网络;所述问答处理装置555还包括:训练模块5555,用于通过所述多个编码网络中的任一所述编码网络,基

于不同于问题语句样本的语言,对所述问题语句样本进行编码处理,得到对应所述编码网络的语义向量;通过所述解码网络对对应所述编码网络的语义向量进行解码处理,得到对应所述问题语句样本的跨语言文本;基于所述跨语言文本以及所述问题语句样本的跨语言文本标签,构建所述神经网络模型的损失函数;更新所述神经网络模型的参数直至所述损失函数收敛,将所述损失函数收敛时所述神经网络模型的更新的参数,作为训练后的所述神经网络模型的参数。

[0185] 在一些实施例中,训练模块5555还用于对对应所述编码网络的语义向量进行分类处理,得到对应所述问题语句样本的预测类别;基于所述预测类别以及所述问题语句样本的类别标签,构建所述神经网络模型的类别损失函数;基于所述跨语言文本以及所述问题语句样本的跨语言文本标签,构建所述神经网络模型的跨语言损失函数;对所述类别损失函数以及所述跨语言损失函数进行加权求和,得到所述神经网络模型的损失函数。

[0186] 在一些实施例中,所述第一召回模块5551还用于针对所述融合语言信息的多个语义向量中的任一所述语义向量执行以下处理:确定所述语义向量分别与不同语言的待召回问题语句的语义向量之间的语义相似度;基于所述语义向量与所述待召回问题语句的语义向量之间的语义相似度,对所述待召回问题语句进行降序排序,将降序排序结果中排序在前N个的待召回问题语句作为所述多个第一问题语句;其中,N为自然数。

[0187] 在一些实施例中,所述第二召回模块5552还用于将所述问题语句中的关键词与所述待召回问题语句进行匹配,得到所述关键词在所述待召回问题语句中的出现频率;基于所述关键词在所述待召回问题语句中的出现频率,确定所述问题语句与所述待召回问题语句之间的文本相似度;基于所述问题语句与所述待召回问题语句之间的文本相似度,对所述待召回问题语句进行降序排序,将降序排序结果中排序在前M个的待召回问题语句作为所述多个第二问题语句;其中,M为自然数。

[0188] 在一些实施例中,所述第二召回模块5552还用于基于所述问题语句中的词语,确定所述问题语句中的词语的权重;对所述问题语句中的词语进行哈希处理,得到对应所述词语的哈希值;基于所述词语的权重,对对应所述词语的哈希值进行加权求和,得到所述问题语句的哈希签名;基于所述问题语句的哈希签名以及所述待召回问题语句的哈希签名,确定所述问题语句与所述待召回问题语句之间的文本相似度;基于所述问题语句与所述待召回问题语句之间的文本相似度,对所述待召回问题语句进行降序排序,将降序排序结果中排序在前K个的待召回问题语句作为所述多个第二问题语句;其中,K为自然数。

[0189] 在一些实施例中,所述匹配模块5553还用于基于所述问题语句与所述第一问题语句之间的相似度、以及所述问题语句与所述第二问题语句之间的相似度,对所述第一问题语句以及所述第二问题语句进行降序排序,将降序排序结果中排序在前L个的第一问题语句或者所述第二问题语句作为用于进行文本匹配的召回问题语句;其中,L为自然数。

[0190] 在一些实施例中,所述匹配模块5553还用于通过分类器对所述问题语句与所述多个第一问题语句进行分类处理,得到所述问题语句分别与所述多个第一问题语句之间的相似度;通过所述分类器对所述问题语句与所述多个第二问题语句进行分类处理,得到所述问题语句分别与所述多个第二问题语句之间的相似度;将最大相似度所对应的召回问题语句作为所述文本匹配处理的结果。

[0191] 在一些实施例中,所述匹配模块5553还用于当在问题答案库中查询到与所述匹配

得到的召回问题语句对应的答案语句,且所述对应的答案语句的语言与所述问题语句的语言一致时,将所述匹配得到的召回问题语句对应的答案语句,作为用于响应所述问题语句的答案语句;当在所述问题答案库中查询到与所述匹配得到的召回问题语句对应的答案语句,且所述对应的答案语句的语言与所述问题语句的语言不一致时,获取所述答案语句对应所述问题语句的语言的翻译结果,并作为用于响应所述问题语句的答案语句。

[0192] 本申请实施例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本申请实施例上述的问答处理方法。

[0193] 本申请实施例提供一种存储有可执行指令的计算机可读存储介质,其中存储有可执行指令,当可执行指令被处理器执行时,将引起处理器执行本申请实施例提供的目标检测方法,例如,如图3A-3C示出的问答处理方法。

[0194] 在一些实施例中,计算机可读存储介质可以是FRAM、ROM、PROM、EP ROM、EEPROM、闪存、磁表面存储器、光盘、或CD-ROM等存储器;也可以是包括上述存储器之一或任意组合的各种设备。

[0195] 在一些实施例中,可执行指令可以采用程序、软件、软件模块、脚本或代码的形式,按任意形式的编程语言(包括编译或解释语言,或者声明性或过程性语言)来编写,并且其可按任意形式部署,包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

[0196] 作为示例,可执行指令可以但不一定对应于文件系统中的文件,可以可被存储在保存其它程序或数据的文件的一部分,例如,存储在超文本标记语言(H TML,Hyper Text Markup Language)文档中的一个或多个脚本中,存储在专用于所讨论的程序的单个文件中,或者,存储在多个协同文件(例如,存储一个或多个模块、子程序或代码部分的文件)中。

[0197] 作为示例,可执行指令可被部署为在一个计算设备上执行,或者在位于一个地点的多个计算设备上执行,又或者,在分布在多个地点且通过通信网络互连的多个计算设备上执行。

[0198] 以上所述,仅为本申请的实施例而已,并非用于限定本申请的保护范围。凡在本申请的精神和范围之内所作的任何修改、等同替换和改进等,均包含在本申请的保护范围之内。

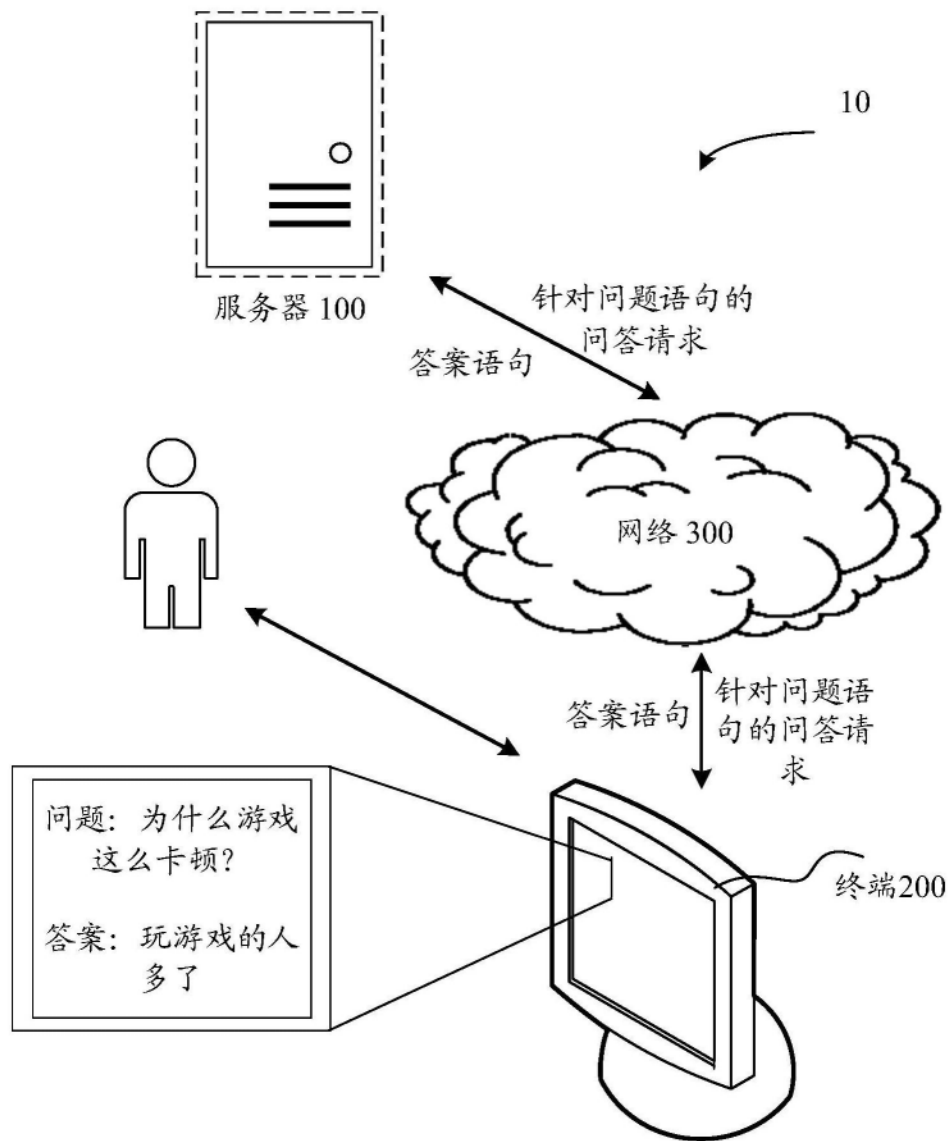


图1

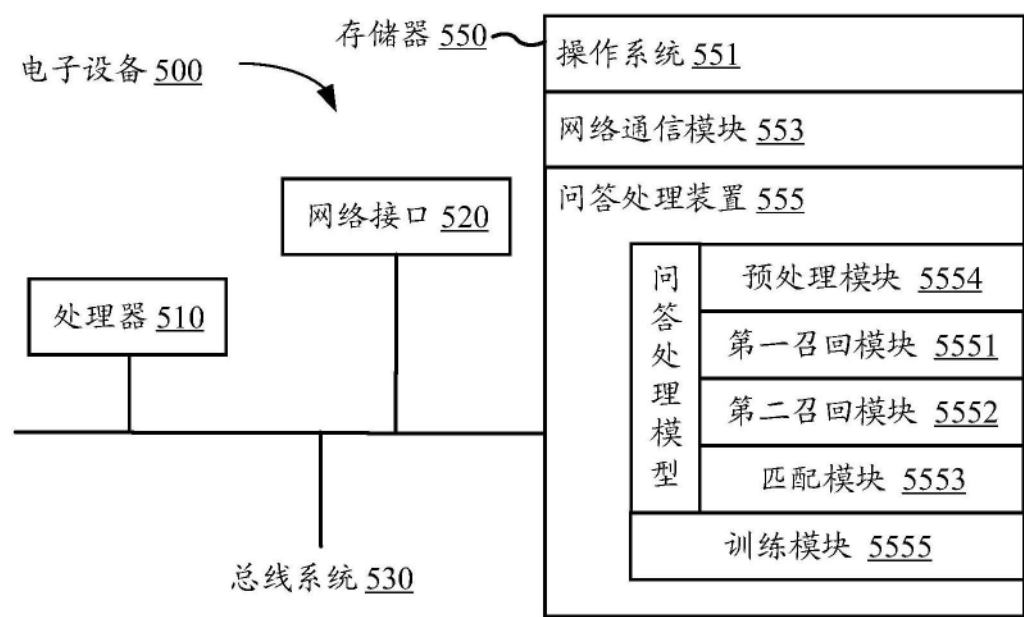


图2

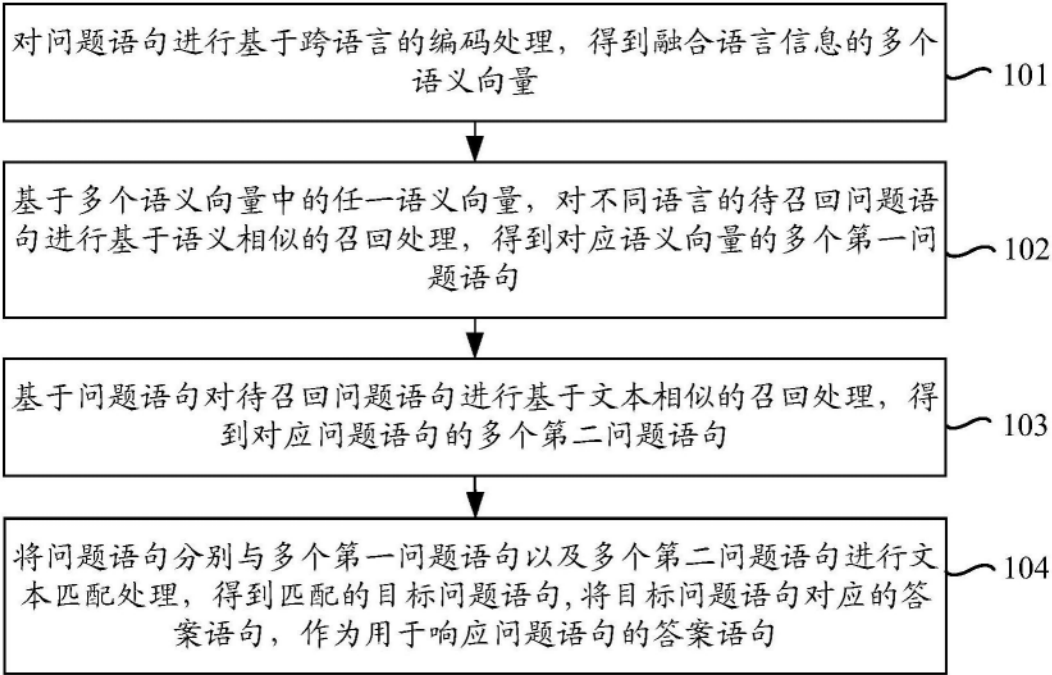


图3A

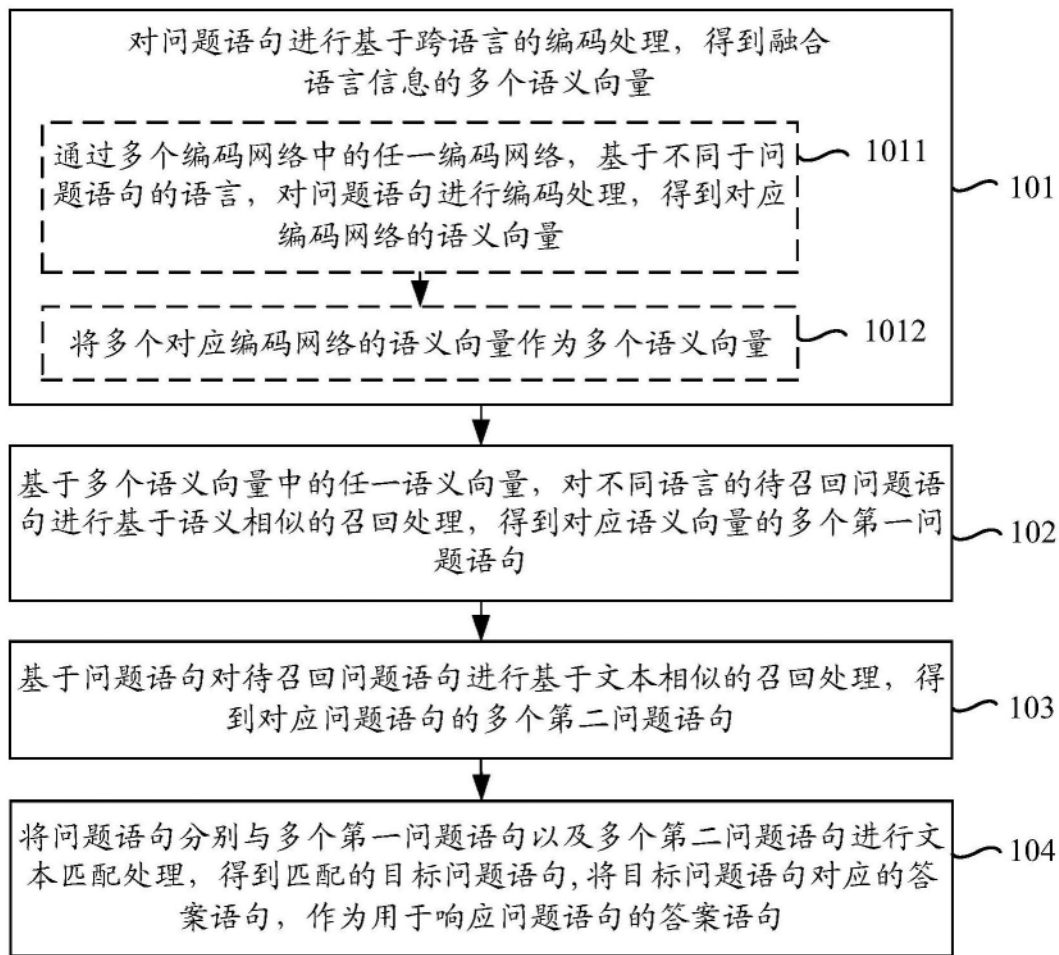


图3B

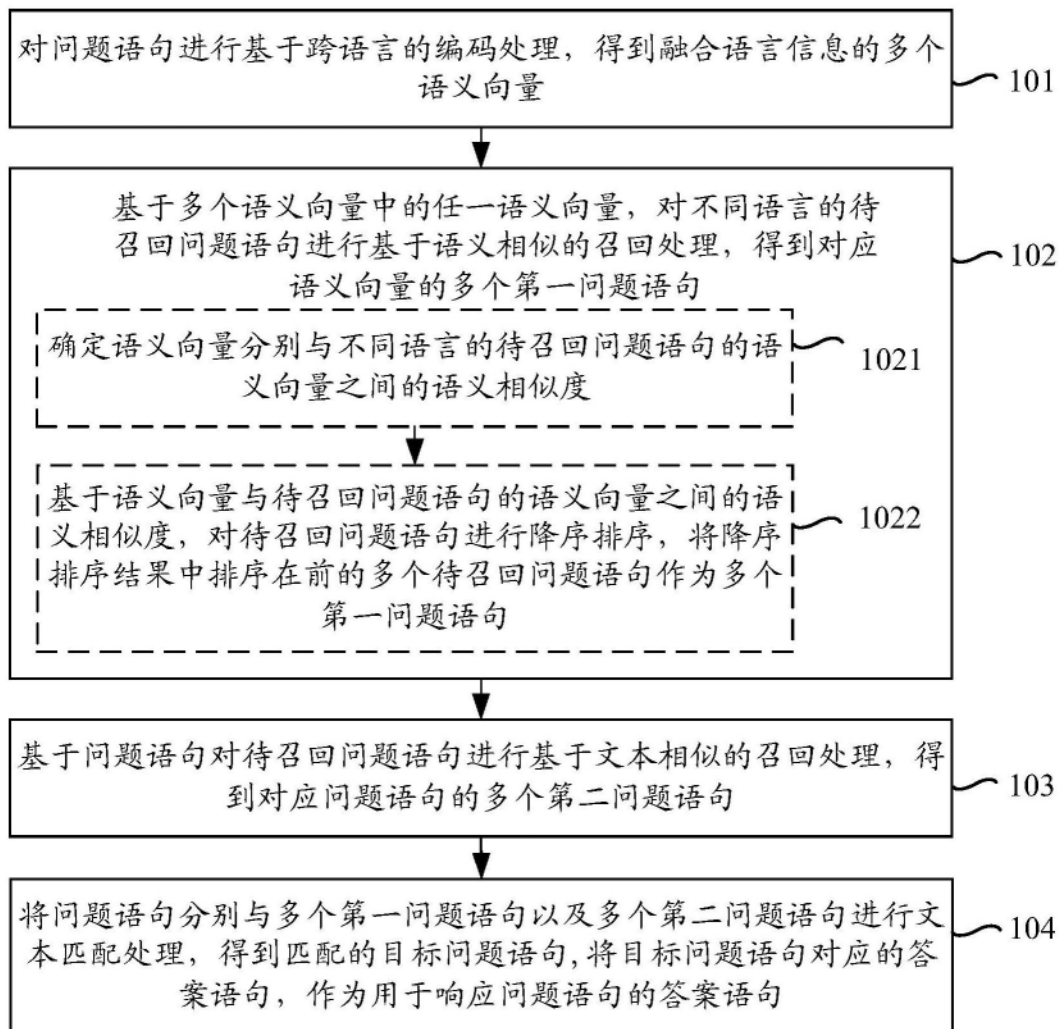


图3C

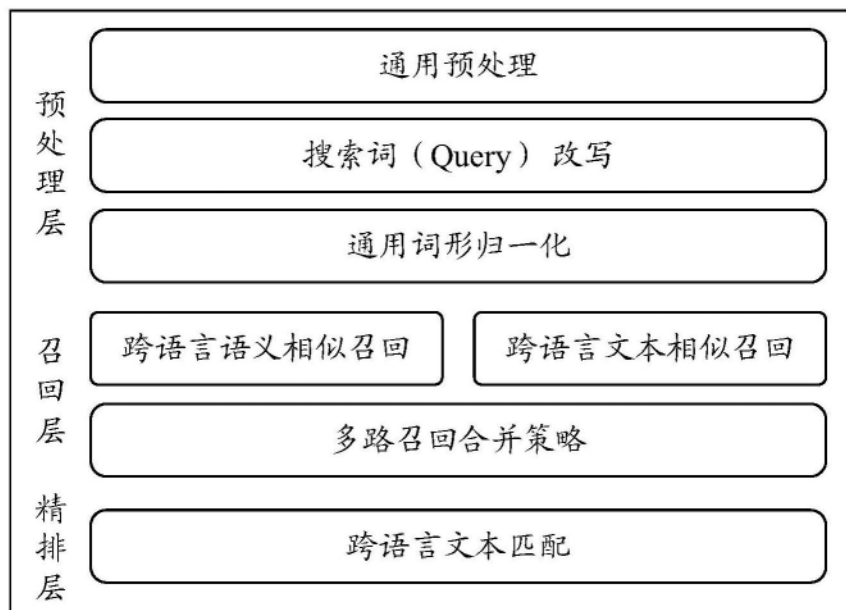


图4

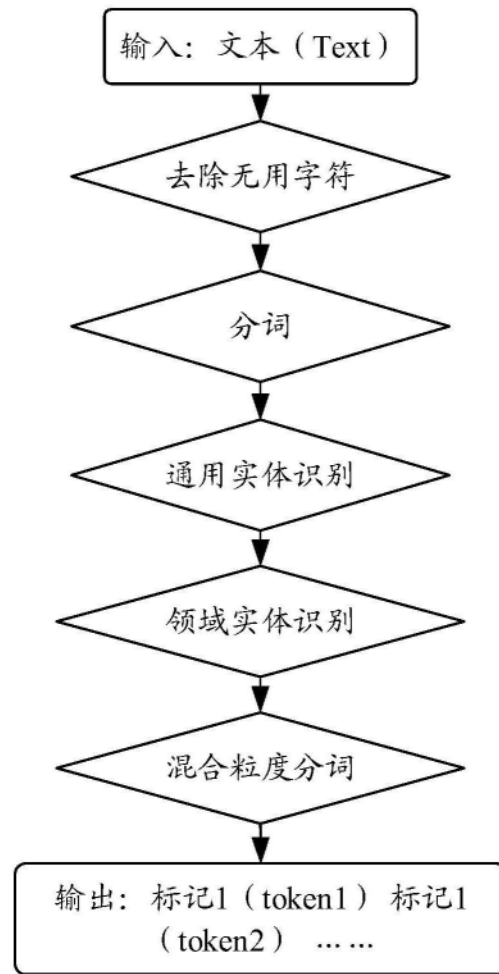


图5

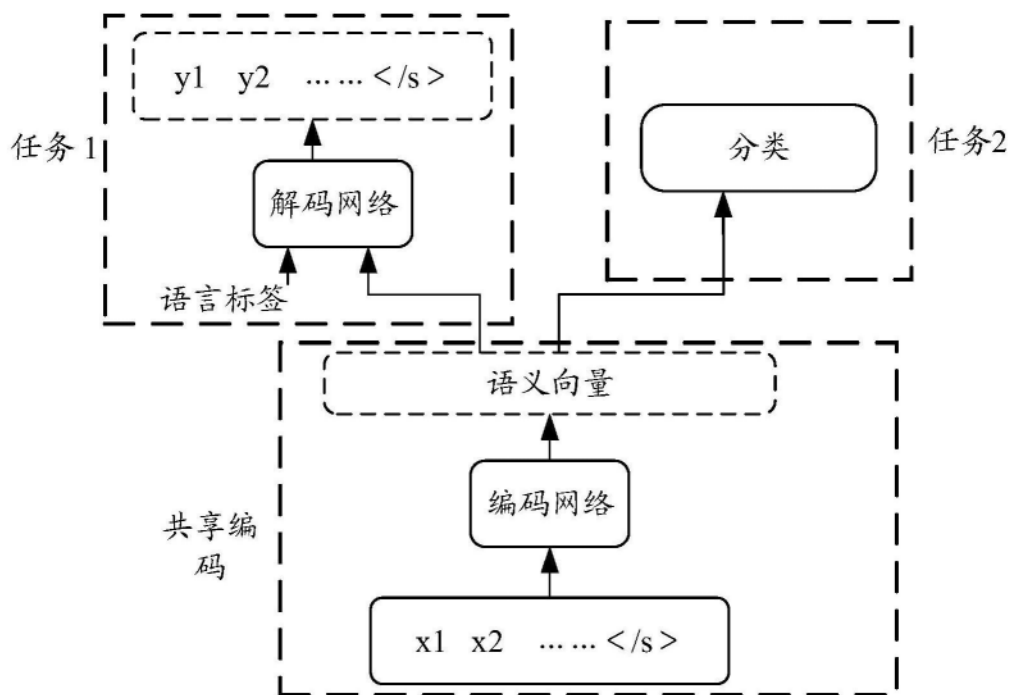


图6

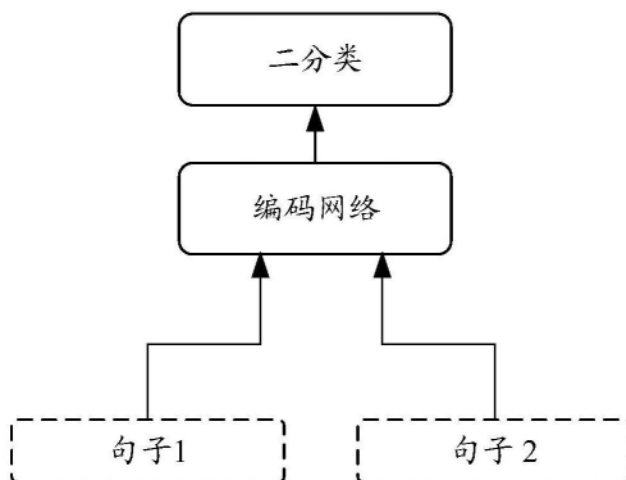


图7

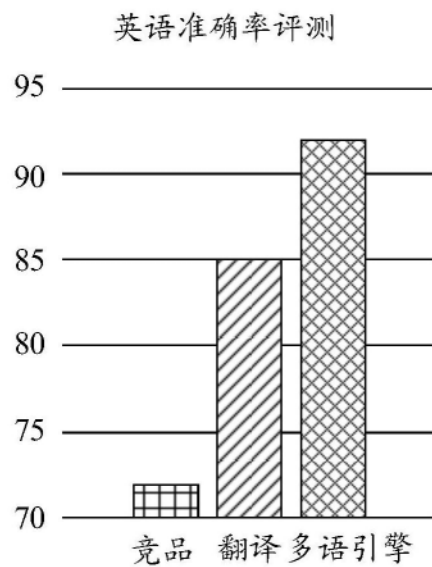


图8

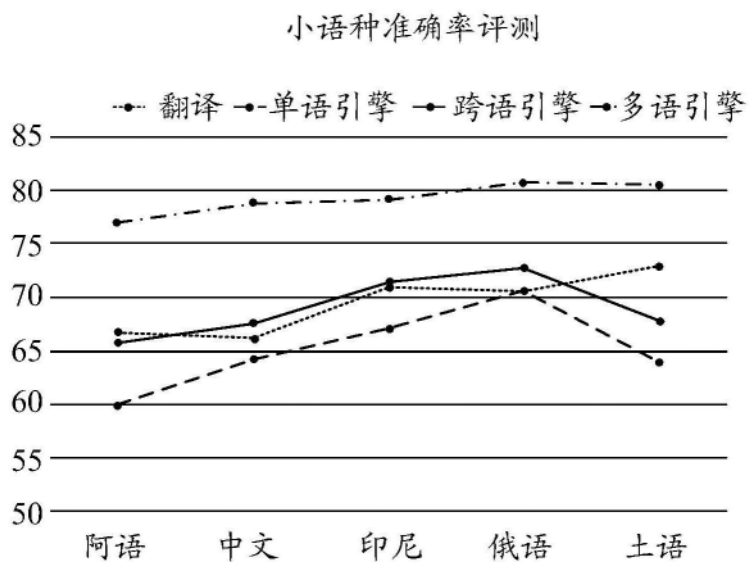


图9

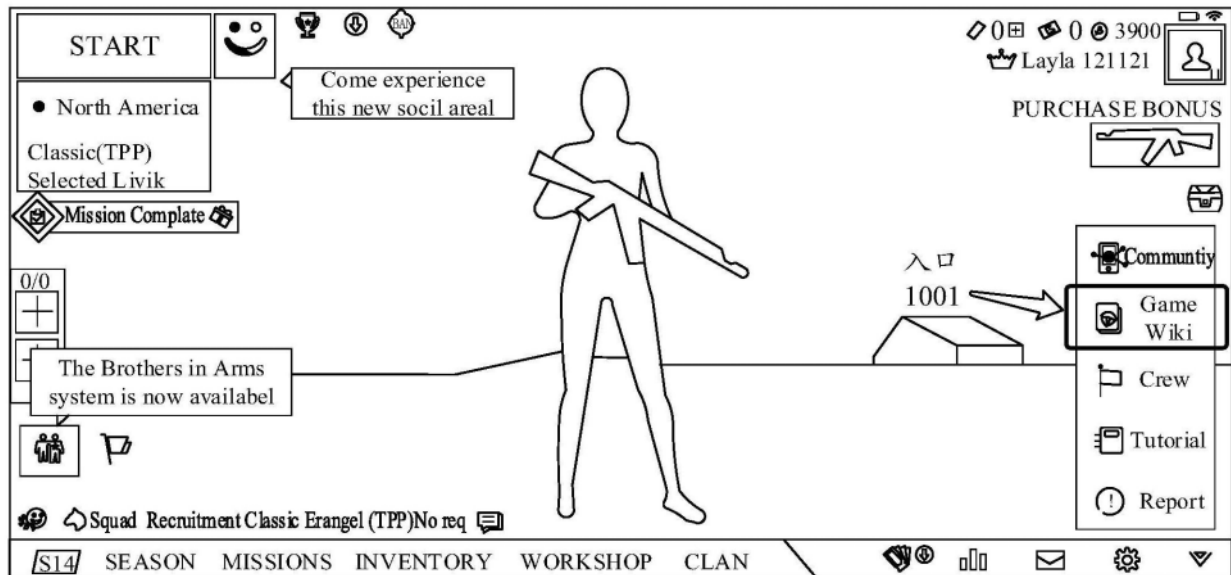


图10

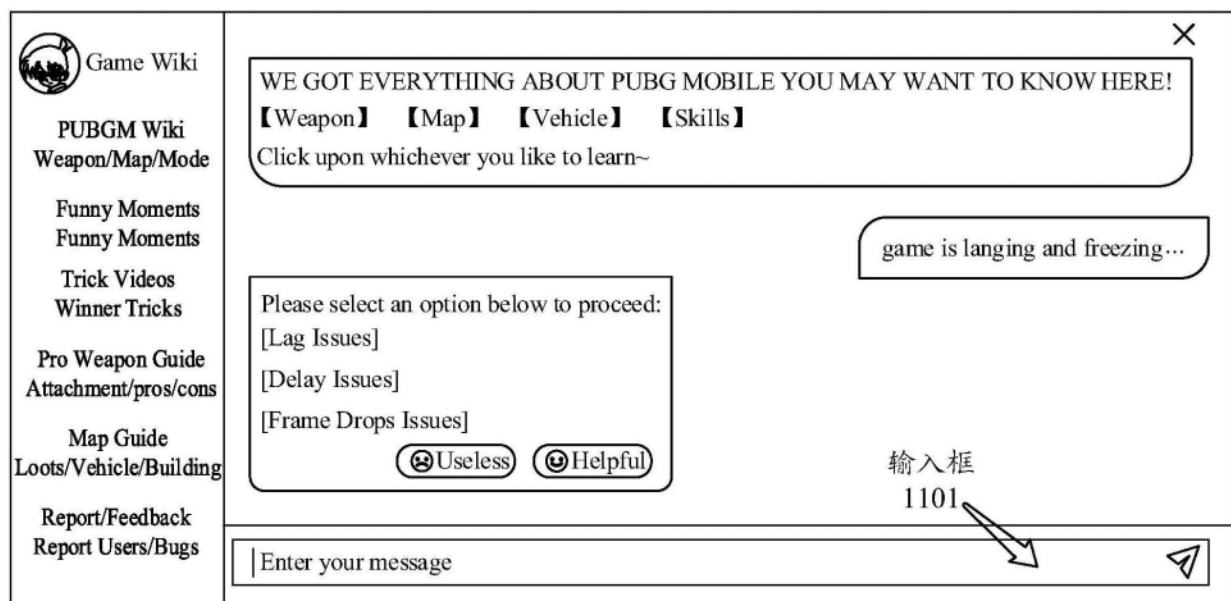


图11

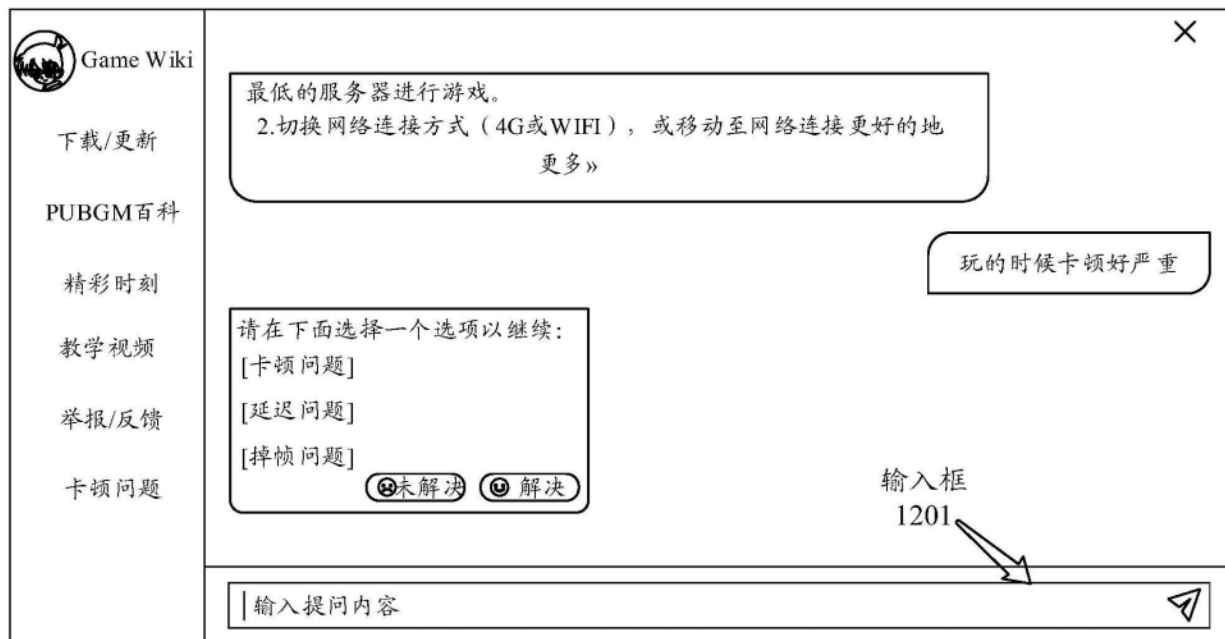


图12