



(12) 发明专利

(10) 授权公告号 CN 111737444 B

(45) 授权公告日 2020. 11. 20

(21) 申请号 202010822588.4

G06F 16/33 (2019.01)

(22) 申请日 2020.08.17

G06K 9/62 (2006.01)

(65) 同一申请的已公布的文献号

(56) 对比文件

申请公布号 CN 111737444 A

CN 110209778 A, 2019.09.06

CN 107463645 A, 2017.12.12

(43) 申请公布日 2020.10.02

审查员 石梦洁

(73) 专利权人 腾讯科技(深圳)有限公司

地址 518064 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 王宝林 马景林

(74) 专利代理机构 深圳市智圈知识产权代理事

务所(普通合伙) 44351

代理人 韩绍君

(51) Int. Cl.

G06F 16/332 (2019.01)

G06F 16/335 (2019.01)

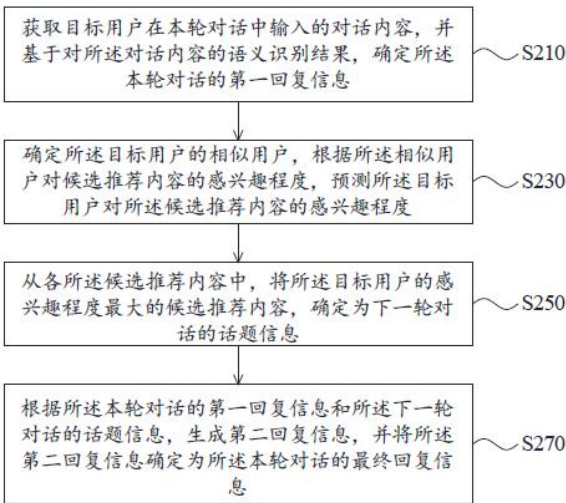
权利要求书3页 说明书18页 附图7页

(54) 发明名称

对话生成方法、装置及电子设备

(57) 摘要

本申请公开了一种对话生成方法、装置及电子设备,涉及人工智能技术领域。该方法包括:获取目标用户在本轮对话中输入的对话内容,并基于对对话内容的语义识别结果,确定本轮对话的第一回复信息;确定目标用户的相似用户,根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度;从各候选推荐内容中,将目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息;根据本轮对话的第一回复信息和下一轮对话的话题信息,生成第二回复信息,并将第二回复信息确定为本轮对话的最终回复信息。如此,可以自动开启下一轮对话的话题,并可以提升在下一轮对话中所推荐话题的丰富程度。



1. 一种对话生成方法,其特征在于,包括:

获取目标用户在本轮对话中输入的对话内容,并基于对所述对话内容的语义识别结果,确定所述本轮对话的第一回复信息;

基于所述语义识别结果,确定所述目标用户在所述本轮对话的对话意图和所述对话意图对应的每个槽位值,所述槽位值为生成所述对话意图对应的控制指令所需的信息;

从每个用户的历史对话中确定目标对话,所述目标对话的对话意图和槽位值与所述目标用户在本轮对话的对话意图和槽位值相同;

将在所述目标对话的下一轮对话中输入的对话内容确定为候选推荐内容;

确定所述目标用户的相似用户,根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度;

从各所述候选推荐内容中,将所述目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息;

根据所述本轮对话的第一回复信息和所述下一轮对话的话题信息,生成第二回复信息,并将所述第二回复信息确定为所述本轮对话的最终回复信息。

2. 根据权利要求1所述的方法,其特征在于,所述方法还包括:

基于所述语义识别结果,确定并记录所述目标用户在本轮对话的对话意图;

基于所述下一轮对话的话题信息,确定并记录所述目标用户在所述下一轮对话的对话意图。

3. 根据权利要求1所述的方法,其特征在于,在所述根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度之前,所述方法还包括:

从各用户的历史对话的对话内容中,将被输入次数符合第一条件的对话内容确定为所述候选推荐内容。

4. 根据权利要求1所述的方法,其特征在于,所述从各所述候选推荐内容中,将所述目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息,包括:

获取所述本轮对话的对话场景信息;

若所述本轮对话的对话场景信息与第一规则不匹配,则从各所述候选推荐内容中,将所述目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息。

5. 根据权利要求4所述的方法,其特征在于,所述方法还包括:

若所述本轮对话的对话场景信息与所述第一规则匹配,则按照所述第一规则生成所述下一轮对话的话题信息。

6. 根据权利要求1所述的方法,其特征在于,所述根据所述本轮对话的第一回复信息和所述下一轮对话的话题信息,生成第二回复信息,包括:

若所述本轮对话的对话内容、所述本轮对话的对话意图以及所述下一轮对话的话题信息与第二规则匹配,则按照所述第二规则对应的话术模板拼接所述第一回复信息和所述下一轮对话的话题信息,得到所述第二回复信息。

7. 根据权利要求6所述的方法,其特征在于,所述方法还包括:

若所述本轮对话的对话内容、所述本轮对话的对话意图以及所述下一轮对话的话题信息与所述第二规则不匹配,则获取机器学习模型根据所述第一回复信息和所述下一轮对话

的话题信息输出的信息,并将获取的信息确定为所述第二回复信息。

8.根据权利要求1所述的方法,其特征在于,所述目标用户为第一用户,所述确定所述目标用户的相似用户,包括:

基于所述第一用户输入的历史对话内容获取第一标签,所述第一标签是所述第一用户的用户画像标签;

基于第二用户的历史对话内容获取第二标签,所述第二标签是所述第二用户的用户画像标签;

将获取的第一标签与获取的第二标签中的相同标签的数量确定为第一数量,以及确定获取的第一标签的数量与获取的第二标签的数量的乘积,并将该乘积的平方根确定为第二数量,并将所述第一数量与所述第二数量的比值确定为所述第二用户与所述第一用户的用户相似度;

若所述用户相似度达到第二条件,则将所述第二用户确定为所述第一用户的相似用户。

9.根据权利要求1所述的方法,其特征在于,所述目标用户为第一用户,所述确定所述目标用户的相似用户,包括:

获取所述第一用户评价过的内容组成的第一历史内容集合,以及第二用户评价过的内容组成的第二历史内容集合,并确定所述第一历史内容集合与所述第二历史内容集合的交集;

根据所述第一用户和所述第二用户各自对所述交集中同一内容的感兴趣程度,以及所述第一用户和所述第二用户各自对所述交集中内容的平均感兴趣程度,确定所述第一用户对任一内容的感兴趣程度与所述第二用户对任一内容的感兴趣程度之间的皮尔逊相关系数;

将所述皮尔逊相关系数确定为所述第二用户与所述第一用户的用户相似度;

若所述用户相似度达到第三条件,则将所述第二用户确定为所述第一用户的相似用户。

10.根据权利要求1所述的方法,其特征在于,所述根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度,包括:

获取所述相似用户对所述候选推荐内容的最新感兴趣程度相较于所述相似用户对已推荐内容的平均感兴趣程度的波动范围,并将获取的波动范围作为第一波动范围;

根据所述相似用户与所述目标用户的用户相似度以及所述第一波动范围,确定所述目标用户对所述候选推荐内容的感兴趣程度相较于所述目标用户对已推荐内容的平均感兴趣程度的波动范围,并将所确定的波动范围作为第二波动范围;

根据所述目标用户对已推荐内容的平均感兴趣程度及所述第二波动范围,确定所述目标用户当前对所述候选推荐内容的感兴趣程度。

11.一种对话生成装置,其特征在于,包括:

获取模块,用于获取目标用户在本轮对话中输入的对话内容,并基于对所述对话内容的语义识别结果,确定所述本轮对话的第一回复信息;

确定模块,用于基于所述语义识别结果,确定所述目标用户在所述本轮对话的对话意图和所述对话意图对应的每个槽位值,所述槽位值为生成所述对话意图对应的控制指令所

需的信息；从每个用户的历史对话中确定目标对话，所述目标对话的对话意图和槽位值与所述目标用户在所述本轮对话的对话意图和槽位值相同；将在所述目标对话的下一轮对话中输入的对话内容确定为候选推荐内容；

预测模块，用于确定所述目标用户的相似用户，根据所述相似用户对候选推荐内容的感兴趣程度，预测所述目标用户对所述候选推荐内容的感兴趣程度；

所述确定模块，还用于从各所述候选推荐内容中，将所述目标用户的感兴趣程度最大的候选推荐内容，确定为下一轮对话的话题信息；

生成模块，用于根据所述本轮对话的第一回复信息和所述下一轮对话的话题信息，生成第二回复信息，并将所述第二回复信息确定为所述本轮对话的最终回复信息。

12. 一种电子设备，其特征在于，包括：

一个或多个处理器；

存储器；

一个或多个程序，其中所述一个或多个程序被存储于所述存储器中并被配置为由所述一个或多个处理器执行，所述一个或多个程序配置用于执行如权利要求1-10中任意一项所述的方法。

13. 一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有程序代码，所述程序代码可被处理器调用执行如权利要求1-10中任意一项所述的方法。

对话生成方法、装置及电子设备

技术领域

[0001] 本申请涉及人工智能技术领域,更具体地,涉及一种对话生成方法、装置及电子设备。

背景技术

[0002] 随着人工智能(Artificial Intelligence, AI)技术的飞速发展,电子产品通常具有人机对话功能,其可以识别用户输入的对话内容,执行该对话内容对应的指令,并针对该对话内容给出合适的回复。但是,现有电子产品通常只能按照一问一答的方式与用户进行交互,非常不便。

发明内容

[0003] 本申请提出了一种对话生成方法、装置及电子设备,可以改善上述问题。

[0004] 一方面,本申请实施例提供了一种对话生成方法,包括:获取目标用户在本轮对话中输入的对话内容,并基于对该对话内容的语义识别结果,确定本轮对话的第一回复信息;确定目标用户的相似用户,根据相似用户对候选推荐内容的感兴趣程度,预测目标用户对候选推荐内容的感兴趣程度;从各候选推荐内容中,将目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息;根据本轮对话的第一回复信息和下一轮对话的话题信息,生成第二回复信息,并将第二回复信息确定为本轮对话的最终回复信息。

[0005] 另一方面,本申请实施例提供了一种对话生成装置,包括:获取模块、预测模块以及确定模块。其中,获取模块用于获取目标用户在本轮对话中输入的对话内容,并基于对该对话内容的语义识别结果,确定本轮对话的第一回复信息。预测模块用于确定目标用户的相似用户,根据相似用户对候选推荐内容的感兴趣程度,预测目标用户对候选推荐内容的感兴趣程度。确定模块用于从各候选推荐内容中,将目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息。生成模块用于根据本轮对话的第一回复信息和下一轮对话的话题信息,生成第二回复信息,并将第二回复信息确定为本轮对话的最终回复信息。

[0006] 另一方面,本申请实施例提供了一种电子设备,包括:一个或多个处理器;存储器;一个或多个程序,其中所述一个或多个程序被存储于所述存储器中并被配置为由所述一个或多个处理器执行,所述一个或多个程序配置用于执行上述的方法。

[0007] 另一方面,本申请实施例提供了一种计算机可读存储介质,其上存储有程序代码,所述程序代码可被处理器调用执行上述的方法。

[0008] 本申请提供的方案,在获取到目标用户在本轮对话中输入的对话内容时,基于对该对话内容的语义识别结果确定本轮对话的第一回复信息,并确定目标用户的相似用户,从而根据相似用户对候选推荐内容的感兴趣程度预测目标用户对候选推荐内容的感兴趣程度,进而可以从各候选推荐内容中,将目标用户的感兴趣程度最大的候选推荐内容确定为下一轮对话的话题信息。基于本轮对话的第一回复信息和下一轮对话的话题信息,可以

生成第二回复信息,从而可以将第二回复信息确定为本轮对话的最终回复信息。如此,可以从候选推荐内容中选取出一轮对话的话题信息并随本轮对话的回复信息推荐给用户,以便自动开启下一轮对话,进而可以在用户目标未完全明确的情况下引导用户进行内容搜索或闲聊,从而明确用户需要执行的控制指令,提升使用效率。

[0009] 并且,该话题信息可以不必总是与目标用户的历史对话内容重复,提升了自动开启的对话话题的丰富程度。

[0010] 本申请的这些方面或其他方面在以下实施例的描述中会更加简明易懂。

附图说明

[0011] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域技术人员来讲,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。

[0012] 图1示出了一种适用于本申请实施例的应用环境示意图。

[0013] 图2示出了本申请一实施例提供的对话生成方法的流程示意图。

[0014] 图3示出了图2所示实施例中对话生成方法的另一流程示意图。

[0015] 图4示出了图2所示实施例中对话生成方法的又一流程示意图。

[0016] 图5示出了本申请另一实施例提供的对话生成方法的流程示意图。

[0017] 图6示出了图5所示步骤S520的一种子步骤示意图。

[0018] 图7示出了图5所示步骤S520的另一种子步骤示意图。

[0019] 图8示出了应用于本申请实施例的可控文本生成模型的架构示意图。

[0020] 图9a示出了用户设备执行本申请实施例提供的对话生成方法的流程。

[0021] 图9b示出了图9a中任务一的实现流程。

[0022] 图9c示出了图9a中任务二的实现流程。

[0023] 图10示出了本申请实施例提供的一种对话生成装置的框图。

[0024] 图11是本申请实施例的用于执行根据本申请实施例的对话生成方法的电子设备的框图。

[0025] 图12是本申请实施例的用于保存或者携带实现根据本申请实施例的对话生成方法的程序代码的存储单元。

具体实施方式

[0026] 为了使本技术领域的人员更好地理解本申请方案,下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述。

[0027] 随着人工智能技术的飞速发展,越来越多的电子产品具有智能对话功能,以便更好地为用户服务。这里的电子产品可以是硬件设备(如,智能音箱、车载智能语音设备、智能电视盒子)或软件产品,软件产品例如可以是智能语音助手、在线语音交互系统、智能虚拟人等。

[0028] 一些实施方式中,可以选取用户感兴趣的话题推送给用户,进而开启对话。为了确保推送的是用户感兴趣的内容,通常会从用户的历史对话内容中选择用于推荐的话题。然

而,这种方式会导致推荐给用户的总是重复过的话题,非常单一。

[0029] 发明人经过长期的研究,提出了一种对话生成方法、装置及电子设备,可以改善上述问题。下面对该内容进行详细描述。

[0030] 请参照图1,图1是一种适用于本申请实施例的架构示意图。其中,用户设备(User Equipment, UE) 100可以用于执行本申请实施例提供的对话生成方法,或可以部署有本申请实施例提供的对话生成装置,用户设备100与服务器200通信连接,服务器200可以为用户设备100提供候选推荐内容、用户的历史对话内容、用于生成话题信息的规则、用于生成回复信息的规则等。

[0031] 其中,用户设备100可以是,但不限于个人计算机(Personal Computer, PC)、笔记本电脑、智能设备、掌上电话、便携式穿戴设备、平板电脑等。这里的智能设备例如可以是智能手机、智能音箱、智能电视盒子、智能语音设备等。服务器200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、大数据和人工智能平台等基础云计算服务的云服务器。本申请实施例对此没有限制。

[0032] 请参照图2,图2示出了本申请一实施例提供的对话生成方法的流程示意图,该方法可以应用于图1所示的用户设备100。下面对该方法的步骤进行详细介绍。

[0033] S210,获取目标用户在本轮对话中输入的对话内容,并基于对所述对话内容的语义识别结果,确定所述本轮对话的第一回复信息。

[0034] 其中,目标用户可以是用户设备100上当前登录的用户帐号信息所对应的用户。比如,用户设备100上当前登录的用户帐号信息为“aaa”,则“aaa”所表示的用户即为目标用户。

[0035] 目标用户输入的对话内容可以包括目标用户主动输入的内容。一种实施方式中,该对话内容可以是目标用户主动发起的表示特定控制指令的语句,比如,目标用户可以主动输入“开启客厅空调”,即为表示控制指令的语句。另一种实施方式中,该对话内容可以是目标用户主动发起的信息查询内容,比如,目标用户可以主动输入“今天天气咋样”,这可以视为信息查询内容。又一种实施方式中,该对话内容可以是目标用户主动发起的闲聊内容,比如,目标用户主动输入“你真是讨厌”,就可以视为闲聊内容。

[0036] 此外,目标用户输入的对话内容还可以包括针对用户设备100在本轮对话中输出的询问信息的回答。一些例子中,目标用户输入的可能是表示控制指令的语句,而明确该控制指令可能存在一些必需的信息。

[0037] 比如,目标用户输入语句1,其内容为“帮我打个车”,通过语义识别可以确定语句1的对话意图是“打车”。用户设备100可以从服务器200查询生成“打车”这一对话意图对应的控制指令所需的信息项,信息项也可以称为槽位(slot)。可以理解,“打车”通常需要确定的信息项包括:出发地点、出发时间和目的地点。出发地点、出发时间、目的地点即为前述的信息项,也就是槽位。目标用户提供的具体出发地点、具体出发时间和具体目的地点,即为信息项的值,也可以称为槽位值。

[0038] 实施过程中,用户设备100可以尝试从语句1的语义识别结果中提取上述的三个槽位值(即,具体出发地点、具体出发时间和具体目的地点),由于语句1并没有给出这三个槽位值,提取失败,可以通过配置的获取方式,获取这三个槽位值。

[0039] 比如,可以通过GPS(Global Positioning System,全球定位系统)定位获取目标

用户的当前定位信息作为出发地点。或者,可以从前几轮对话的对话内容中查找出发地点。或者,还可以通过主动询问目标用户的方式,促使目标用户输入具体的出发地点。可以理解,当针对一个槽位值配置有多种获取方式时,通常还可以配置有所述多种获取方式的优先级,比如前述的具体出发地点的三种获取方式的优先级从高到低可以为:主动询问、查找前几轮对话、GPS定位。实施过程中,可以将通过优先级最高的获取方式获取到的槽位值确定为实际槽位值,对应地,在生成控制指令时使用是该实际槽位值。

[0040] 又比如,可以通过主动询问和查询前几轮对话这两种方式中的至少一者,获取目标用户的具体目的地点和具体出发时间。

[0041] 上述过程中,用户设备100如果通过主动询问方式来获取槽位值,则目标用户针对用户设备100的主动询问信息给出的回答也属于是目标用户在本轮对话中输入的对话内容的一部分。

[0042] 本实施例中,对话内容的语义识别结果可以包括对话内容的对话意图以及该对话意图对应的槽位值。这里的对话意图可以理解成对话内容的目的。比如,上述的语句1的对话意图是“打车”。通过后续处理,可以进一步获得“打车”对应的具体出发地点、具体出发时间和具体目的地点。又比如语句2,其内容是“放一首周杰伦的歌”,则通过识别可以确定语句2的对话意图是“播放音乐”,其对应的槽位可以有:音乐名称、歌手、音乐类型等。其中,每个槽位可以具有对应的属性,该属性可以用于表示该槽位对应的槽位值是否为必填值。一个例子中,在给出具体音乐名称的情况下,歌手和音乐类型可以不是必填的。另一个例子中,在给出具体歌手的情况下,音乐名称和音乐类型可以不是必填的。

[0043] 在获得对话内容的语义识别结果之后,用户设备100可以按照服务处理逻辑,确定所述对话内容的语义识别结果对应的回复信息。示例性地,可以根据对话内容的对话意图及各槽位值,查找命中的话术模板,进而依照该话术模板生成相应的回复信息。这里的回复信息是指与目标用户输入的对话内容完全对应的回复信息,其可以充当S210中的第一回复信息。比如,上述语句2对应的第一回复信息可以是“已经为您选择了周杰伦的七里香,马上为您播放”。可以理解,这里的服务处理逻辑可以是预先定义的生成规则,如上述的话术模板,也可以是预先训练并配置的用于生成话术的机器学习模型,本实施例对此没有限制。

[0044] 值得说明的是,本实施例中,第一回复信息可以是针对目标用户给出的问题的回答,也可以是为了明确用户的控制指令而输出的询问信息。

[0045] 本实施例中,在获取到目标用户在本轮对话输入的对话内容时,除了确定第一回复信息,还可以通过步骤S230和S250确定用于开启本轮对话的下一轮对话的话题信息。

[0046] S230,确定所述目标用户的相似用户,根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度。

[0047] 本申请实施例中,为了便于描述,将目标用户约定为第一用户,将目标用户之外的其他用户约定为第二用户。实施过程中,用户设备100可以分别确定每个第二用户与第一用户的用户相似度,进而从各第二用户中确定与第一用户的用户相似度符合条件的第二用户,并将所确定的第二用户作为第一用户的相似用户,也就是S230中描述的目标用户的相似用户。这里的用户相似度可以是基于第一用户和第二用户各自的历史对话内容确定的,历史对话内容可以从服务器200获取。

[0048] 可以理解,上述的确定相似用户的过程,也可以由用户设备100向服务器200发送

请求,服务器200响应该请求,确定目标用户的相似用户,并向用户设备100返回相似用户的标识信息(如,帐号信息、昵称、用户名等)。本实施例对此没有限制。

[0049] 服务器200可以存储有一内容池,该内容池中的内容为候选推荐内容。由于相似的用户之间具有类似的喜好,可以基于目标用户的相似用户对内容池中候选推荐内容的感兴趣程度,来预测目标用户对这个候选推荐内容的感兴趣程度。比如,可以通过将目标用户的相似用户对于一个候选推荐内容 i 的感兴趣程度做加权平均,并将加权平均的结果作为目标用户对候选推荐内容 i 的预测感兴趣程度。可选地,每个相似用户的权重可以相同(如,均为1),也可以是其与目标用户的相似度,本实施例对此没有限制。

[0050] 示例性地,感兴趣程度可以通过用户对内容的评分表示,也可以通过用户输入该内容的频次来表示。本实施例对此没有限制。

[0051] S250,从各所述候选推荐内容中,将所述目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息。

[0052] 通过S230,基本可以确定目标用户对内容池中每一候选推荐内容的预测感兴趣程度。在S250中,用户设备100可以对比目标用户对内容池中各候选推荐内容的预测感兴趣程度,进而基于比较结果从内容池中确定一候选推荐内容 y ,目标用户对内容池中其他所有候选推荐内容的预测感兴趣程度都不大于目标用户对候选推荐内容 y 的预测感兴趣程度。则候选推荐内容 y 可以被作为下一轮对话的话题信息。

[0053] 可选地,一些情况下,可能存在至少两个候选推荐内容如 y_1 、 y_2 的预测感兴趣程度基本相同(如差值处于预设的误差范围内),此时,可以首先判断所述至少两个候选推荐内容 y_1 、 y_2 是否是可拼接的内容。如果所述至少两个候选推荐内容是可拼接的内容,则将所述至少两个候选推荐内容均作为下一轮对话的话题信息。如果所述至少两个候选推荐内容不是可拼接的内容,则可以分别获取所述至少两个候选推荐内容与本轮对话中输入的对话内容的可拼接程度,从所述至少两个候选推荐内容中选择可拼接程度高的一者作为下一轮对话的话题信息。

[0054] 其中,对于至少两个候选推荐内容是否是可拼接的内容的判断,以及候选推荐内容与对话内容的可拼接程度的确定,可以采用神经网络模型(在此描述为模型A)实现。详细地,可以选取多个样本对,每个样本对包括两个样本内容及一标签,标签表示该样本对中两个样本内容的拼接分数。样本对的拼接分数越高,表示样本对中的两个样本内容的可拼接程度越高。训练过程中,可以将每个样本对中的两个样本内容输入神经网络模型,神经网络模型A分别从两个样本内容中提取特征,并计算提取的特征之间的相似度作为两个样本内容的预测拼接分数。通过损失函数计算样本对的预测拼接分数与标签之间的损失值,基于损失值调整神经网络模型A的模型参数,以使损失函数收敛。如此,可以得到经过训练的神经网络模型A'。

[0055] 实施过程中,可以将任意两个候选推荐内容输入经过训练的神经网络模型A',从而得到这两个候选推荐内容的预测拼接分数,如果预测拼接分数达到阈值,则可以确定这两个候选推荐内容是可拼接的内容。类似地,可以将任一候选推荐内容与本轮对话中输入的对话内容输入经过训练的神经网络模型A',从而得到该候选推荐内容与对话内容的预测拼接分数,该预测拼接分数可以视为上文描述的可拼接程度。

[0056] 鉴于通过S230和S250可以在一定程度上确保下一轮对话的话题信息是目标用户

感兴趣的,内容池中的候选推荐内容可以从所有用户(包括第一用户和各第二用户)的历史对话内容中选取,而不再局限于从目标用户的历史对话内容中选取,即可以扩大主动开启的对话话题的内容范围,从而可以提高主动开启的对话话题的丰富程度。

[0057] S270,根据所述本轮对话的第一回复信息和所述下一轮对话的话题信息,生成第二回复信息,并将所述第二回复信息确定为所述本轮对话的最终回复信息。

[0058] 在获得本轮对话的第一回复信息和下一轮对话的话题信息之后,可以将两者拼接成第二回复信息。如此,第二回复信息既包含了针对本轮对话的实际回复信息,又包含了用于开启下一轮对话的话题。将第二回复信息作为本轮对话的最终回复信息展示给目标用户,可以自动开启下一轮对话。并且,下一轮对话的对话意图可以不再局限于本轮对话的对话意图,即,可以实现跨意图的多轮对话的自动开启。

[0059] 请一并参照图2和图3,下面将对图2所示实施例中的对话生成方法做进一步的详细介绍。实施过程中,为了提升相邻两轮对话之间的连续性,在本申请实施例提供的对话生成方法中,可以通过图3所示的流程来确定目标用户当前的候选推荐内容,也即,目标用户对应的内容池中的当前内容。可以理解,图3所示的步骤可以是在S220中的根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度之前执行。详细实现过程如下。

[0060] S221,基于目标用户的语义识别结果,确定所述目标用户在本轮对话的对话意图和所述对话意图对应的每个槽位值,所述槽位值为生成所述对话意图对应的控制指令所需的信息。

[0061] 参照上文的描述,对话意图表示对话内容的目的,而为了明确对话意图的目的,存在一些必需的信息项,这些信息项就是槽位,这些信息项的具体值就是槽位值。本实施例中,对话内容的语义识别结果至少可以包括对话内容的分词信息,以及对话意图。基于对话意图,用户设备100可以从本设备或者从服务器200查找对话意图对应的槽位,进而从分词信息中查找槽位对应的槽位值,如果没有查找到则按照配置的获取方式进行获取,进而可以得到对话内容的所有槽位值。如此,可以获得目标用户在本轮对话的对话内容所对应的对话意图和槽位值。

[0062] S222,从每个用户的历史对话中确定目标对话,所述目标对话的对话意图和槽位值与所述目标用户在本轮对话的对话意图和槽位值相同。

[0063] S223,将在所述目标对话的下一轮对话中输入的对话内容确定为候选推荐内容。

[0064] 本实施例中,服务器200可以记录有每个用户的历史对话,及每一轮历史对话的对话内容、对话意图和槽位值。其中,可能有一些用户,他们在某一轮对话 D_i 的对话意图和槽位值与目标用户在本轮对话的对话意图和槽位值分别对应相同,此时,可以获取他们在对话 D_i 的下一轮对话的对话内容,并将该对话内容确定为目标用户的一个候选推荐内容。这样,为目标用户确定的候选推荐内容,都是目标用户或其他用户在本轮对话的对话意图和槽位值之下主动输入的内容,与本轮对话的对话内容具有较好的连续性,更加拟人化。对应地,从候选推荐内容中确定的话题信息,也与本轮对话具有更好的连续性,更加拟人化,用户体验更好。

[0065] 可选地,服务器200中记录的历史对话可以是每个用户在指定时间段内的对话。这里的指定时间段例如可以是当前时刻之前的预设时间段,预设时间段例如可以是一个周、

半个月、一个月或者半年,本实施例对此没有限制。

[0066] 可选地,本实施例中,还可以从各个用户的历史对话的对话内容中,将被输入次数符合第一条件的对话内容确定为目标用户当前的候选推荐内容,即,上述的S230中提及的候选推荐内容。这里的第一条件可以是被输入频次达到特定次数,如20次、30次、45次等。第一条件也可以是将各对话内容按被输入频次从大到小排序后,排序靠前的预设数量个对话内容。这里的预设数量可以灵活设置,比如可以是5、8或10等。被输入频次表示的是对话内容的受欢迎程度或是热度,基于此选出的候选推荐内容相应也更容易符合用户需求。

[0067] 进一步地,也可以是从各个用户在所述指定时间段内的历史对话的对话内容中,将被输入频次符合第一条件的对话内容确定为目标用户当前的候选推荐内容,本实施例对此没有限制。如此,所确定的候选推荐内容是指定时间段内热度较高的内容,即近期热度较高的内容,更容易被目标用户接受。

[0068] 本实施例中,为了使得目标用户在本轮对话和下一轮对话的对话内容均可以作为确定各用户的候选推荐内容的依据,本实施例提供的对话生成方法还可以包括图4所示的S211和S251。

[0069] S211,基于目标用户在本轮对话中输入的对话内容的语义识别结果,确定并记录所述目标用户在所述本轮对话的对话意图。

[0070] S251,基于下一轮对话的话题信息,确定并记录所述目标用户在所述下一轮对话的对话意图。

[0071] 其中,S211可以是在S210之后执行,S251可以是在S250之后执行。除了对话意图之外,还可以记录槽位值及对话内容本身。对应地,每一轮对话都具有对应的会话状态,当本轮对话没有结束时,会话状态表示本轮对话处于进行中,当本轮对话结束时,会话状态变为结束,并在下一轮对话开始时被更新为进行中。当一轮对话结束时,用户设备100可以获取本轮对话的对话内容、对话意图和槽位值并进行记录。比如,下一轮对话的话题信息,一些情况下,可能只能用于确定对话意图,在经过与目标用户的后续交互之后,才可以获得完整的对话内容及槽位值。

[0072] 请参照图5,图5为本申请另一实施例提供的对话生成方法的流程示意图,该方法也可以应用于图1所示的用户设备100。下面对该方法的步骤进行详细介绍。

[0073] S510,获取目标用户在本轮对话中输入的对话内容,并基于对所述对话内容的语义识别结果,确定所述本轮对话的第一回复信息。

[0074] 其中,S510的详细实现过程与S210类似,在此不再赘述。

[0075] S520,确定所述目标用户的相似用户。

[0076] 本实施例中,S520可以有多种实施方式。一种实施方式中,S520可以通过图6所示步骤实现。详细描述如下。

[0077] S521,基于所述第一用户输入的历史对话内容获取第一标签,所述第一标签是所述第一用户的用户画像标签。

[0078] S522,基于所述第二用户的历史对话内容获取第二标签,所述第二标签是所述第二用户的用户画像标签。

[0079] 本实施例中,第一用户即为上文提及的目标用户,第二用户则是目标用户之外的其他用户。每一用户的用户画像标签可以通过一分类模型确定。将任意用户(如,U1)的历史

对话内容输入分类模型,分类模型可以输出用户U1可能的画像标签以及每个可能的画像标签对应的置信度。实施过程中,可以将置信度符合要求(如,达到60以上)的画像标签确定为用户U1的用户画像标签。基于此,每个用户可以具有一个或多个用户画像标签。换句话说,基于第一用户的历史对话内容所确定的第一标签可能有一个或多个,基于每个第二用户的历史对话内容所确定的第二标签可能有一个或多个。

[0080] 可以理解,用户画像标签表示的是用户类别,该用户类别可以是在训练数据中预先标注的类别,如,影视爱好者、篮球爱好者、新闻爱好者、音乐爱好者等。基于该训练数据训练后的分类模型则可以区分出所述预先标注的类别。

[0081] S523,将获取的第一标签与获取的第二标签中的相同标签的数量确定为第一数量,以及确定获取的第一标签的数量与获取的第二标签的数量的乘积,并将该乘积的平方确定为第二数量,并将所述第一数量与所述第二量的比值确定为所述第二用户与所述第一用户的用户相似度。

[0082] 实施过程中,可以将第一用户的第一标签与第二用户的第二标签一一对比,进而可以确定第一用户的第一标签与第二用户的第二标签中相同标签的数量。比如:第一用户U1的用户画像标签包括tag1、tag2、tag3,则获取的第一标签有3个,分别是tag1、tag2、tag3;某一第二用户U2的用户画像标签包括tag2、tag3、tag4、tag5,则获取的第二标签有4个,分别是tag2、tag3、tag4、tag5。并且相同标签有2个,分别是tag2和tag3。即,第一数量为2。获取的第一标签的数量3和第二标签的数量4的乘积为12。如此,可以确定第二用户U2与第一用户U1的用户相似度为 $2/\sqrt{12}$,约为0.577。

[0083] 示例性地,S523的实现过程可以表示为如下表达式:

$$[0084] \quad s1(u,v) = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)||N(v)|}}, (1)$$

[0085] 其中,u表示第一用户,v表示一个第二用户。 $N(u)$ 表示第一用户u的用户画像标签集合,即获取的第一用户u的第一标签所组成的集合。 $N(v)$ 表示第二用户v的用户画像标签集合,即获取的第二用户v的第二标签所组成的集合。其中, $N(u) \cap N(v)$ 表示用户u的第一标签与用户v的第二标签中相同标签的数量。 $|N(u)||N(v)|$ 表示的是用户u的第一标签的数量与用户v的第二标签的数量的乘积。 $s1(u,v)$ 表示的是S523里的用户相似度。这里的 $s1(u,v)$ 也可以理解为集合 $N(u)$ 与 $N(v)$ 的Ochiai系数,这个系数和余弦相似性基本是相同的。

[0086] S524,若所述用户相似度达到第二条件,则将所述第二用户确定为所述第一用户的相似用户。

[0087] 这里的第二条件可以灵活设置,比如可以是达到第一阈值,第一阈值可以灵活设置,比如可以是0.5-0.8之间的任意值,例如0.65。当通过S523计算出的任意一个第二用户与第一用户的用户相似度达到了上述第一阈值时,可以确定该第二用户是第一用户的相似用户。

[0088] 另一种实施方式中,S520可以通过图7所示的步骤实现,详细描述如下。

[0089] S525,获取所述第一用户评价过的内容组成的第一历史内容集合,以及第二用户评价过的内容组成的第二历史内容集合,并确定所述第一历史内容集合与所述第二历史内容集合的交集。

[0090] S526,根据所述第一用户和所述第二用户各自对所述交集中同一内容的感兴趣程度,以及所述第一用户和所述第二用户各自对所述交集中内容的平均感兴趣程度,确定所述第一用户对任一内容的感兴趣程度与所述第二用户对任一内容的感兴趣程度之间的皮尔逊相关系数。

[0091] S527,将所述皮尔逊相关系数确定为所述第二用户与所述第一用户的用户相似度。

[0092] 本实施例中,第一用户u对任一内容的感兴趣程度可以视为一个随机变量,一个第二用户v对任一内容的感兴趣程度也可以视为一个随机变量。两个随机变量的皮尔逊相关系数可以用于表示这两个随机变量的相关程度,对应地,第一用户u对任一内容的感兴趣程度与第二用户v对任一内容的感兴趣程度之间的皮尔逊相关系数可以用于表示第一用户u与第二用户v之间的用户相似度。

[0093] 详细地,第一用户u对任一内容的感兴趣程度与第二用户v对任一内容的感兴趣程度之间的皮尔逊相关系数 $s2(u,v)$ 可以如下表达式计算获得:

$$[0094] \quad s2(u,v) = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_u \cap I_v} (r_{v,i} - \bar{r}_v)^2}}, \quad (2)$$

[0095] 其中,i表示内容池中的第i个候选推荐内容,i为正整数。 $r_{u,i}$ 表示第一用户u对候选推荐内容i的感兴趣程度, $\bar{r}_u$ 表示第一用户u对已推荐内容的平均感兴趣程度。这里的已推荐内容可以是在上文描述的指定时间段内容推荐给第一用户u的内容,感兴趣程度可以是第一用户u对已推荐内容的评分。 $r_{v,i}$ 表示第二用户v对候选推荐内容i的感兴趣程度, $\bar{r}_v$ 表示第二用户v对已推荐内容的平均感兴趣程度。这里的已推荐内容可以是在上述指定时间段内推荐给第二用户v的内容,感兴趣程度可以是第二用户对已推荐内容的评分。

[0096] 在获得皮尔逊相关系数 $s2(u,v)$ 之后,可以将 $s2(u,v)$ 确定为第一用户u与第二用户v之间的用户相似度。

[0097] S528,若所述用户相似度达到第三条件,则将所述第二用户确定为所述第一用户的相似用户。

[0098] 其中,第三条件是针对于皮尔逊相关系数而设置的,比如可以是皮尔逊相关系数达到第二阈值。第二阈值可以根据经验灵活设置,本实施例对此没有限制。在此情况下,如果某一第二用户对内容的感兴趣程度与第一用户对内容的感兴趣程度之间的皮尔逊相关系数达到了第二阈值,则可以确定该第二用户是第一用户的相似用户。

[0099] 本实施例中,在确定目标用户的相似用户之后,可以基于相似用户对目标用户当前的每一候选推荐内容的感兴趣程度,来预测目标用户对该候选推荐内容的感兴趣程度,具体可以通过如下所述的S530-S550实现。值得说明的是,上述实施例中的S230中根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度的步骤也可以通过S530-S550实现。此外,目标用户当前的候选推荐内容的确定方式,可以参照上述实施例对相关流程的描述,在此不再赘述。

[0100] S530,获取所述相似用户对所述候选推荐内容的最新感兴趣程度相较于所述相似用户对已推荐内容的平均感兴趣程度的波动范围,并将获取的波动范围作为第一波动范

围。

[0101] 本实施例中,一个用户对于已推荐给他的内容存在一平均感兴趣程度,而这个用户对某个具体推荐内容的感兴趣程度,总是相对其平均感兴趣程度波动的,即,这个用户对某个具体推荐内容的感兴趣程度相对于其平均感兴趣程度存在一个波动范围。实施过程中,可以根据目标用户的相似用户对某一候选推荐内容*i*的感兴趣程度相较于其平均感兴趣程度的波动范围,来预测目标用户对这个候选推荐内容*i*的感兴趣程度相较于目标用户的平均感兴趣程度的波动范围。

[0102] 其中,第一波动范围是指相似用户对一个候选推荐内容*i*的感兴趣程度相对于其对已推荐内容的平均感兴趣程度的波动范围。本实施例中,可以基于相似用户对候选推荐内容*i*的最新感兴趣程度和平均感兴趣程度来确定相似用户针对候选推荐内容*i*的第一波动范围。这样,第一波动范围反映的是相似用户针对候选推荐内容*i*的感兴趣程度的最新变化趋势,对应地,基于该第一波动范围预测的第二波动范围也就更能反映目标用户对候选推荐内容*i*的感兴趣程度的当前变化趋势。

[0103] 示例性地,本实施例中,相似用户对候选推荐内容的最新感兴趣程度可以从服务器200存储的历史数据中获得。示例性地,服务器200可以存储有每个用户的历史数据,该历史数据可以包括上文描述的指定时间段内的每轮对话的对话内容、对话意图和槽位值。

[0104] 此外,用户的历史数据还可以包括用户对已推荐的每个候选推荐内容的指定数量个评分及每个评分的产生时间。产生时间例如可以是用户设备100采集到该评分的时间或者用户设备100向服务器200发送该评分的时间。以指定数量是*M*(大于或等于1的正整数)为例,对于已推荐给用户*u*的每个内容*j*,可以记录用户*u*最近给出的针对内容*j*的*M*个评分。这里的最近给出的针对内容*j*的*M*个评分,是指将用户*u*针对内容*j*的评分按产生时间的先后顺序进行排序后,排列顺序靠前的*M*个评分。在此情况下,可以从相似用户的历史数据中,查找相似用户针对候选推荐内容的*M*个评分,并从查找到的*M*个评分中选取产生时间最新的一个评分,作为相似用户对候选推荐内容的最新感兴趣程度。

[0105] 可选地,相似用户对已推荐内容的平均感兴趣程度,可以通过相似用户的历史数据中的所有评分得到。一种实施方式中,可以直接计算相似用户的历史数据中的所有评分的平均值,并将该平均值确定为相似用户对已推荐内容的平均感兴趣程度。另一种实施方式中,可以按照每个评分的产生时间为该评分确定一权重,评分的权重可以与评分的产生时间成反比。然后,可以按照每个评分的权重对相似用户的历史数据中的所有评分做加权平均,得到的平均值可以作为相似用户对已推荐内容的平均感兴趣程度。

[0106] S540,根据所述相似用户与所述目标用户的用户相似度以及所述第一波动范围,确定所述目标用户对所述候选推荐内容的感兴趣程度相较于所述目标用户对已推荐内容的平均感兴趣程度的波动范围,并将所确定的波动范围作为第二波动范围。

[0107] 其中,第二波动范围是指目标用户对一个候选推荐内容*i*的感兴趣程度相对于其对已推荐内容的平均感兴趣程度的波动范围。值得说明的是,目标用户对于已推荐内容的平均感兴趣程度,也可以按照相似用户的平均感兴趣程度的获取方式得到,在此不再赘述。

[0108] 详细地,以感兴趣程度是用户对内容的评分为例,第一波动范围可以是相似用户对候选推荐内容的最新评分与相似用户对已推荐内容的平均评分的差值(后文描述为“第一波动分值”)。可以将每个相似用户与目标用户的用户相似度确定为该相似用户的权重。

针对每一候选推荐内容,可以按照每个相似用户的权重,将各相似用户与该候选推荐内容对应的第一波动分值做加权平均,得到的平均值可以作为目标用户与该候选推荐内容的第二波动范围(也可以称为“第二波动分值”)。

[0109] 上述确定第二波动分值的过程,可以通过如下表达式表示:

$$[0110] \quad \Delta r_{u,i} = \frac{\sum_{u' \in N} s(u,u') (r_{u',i} - \bar{r}_{u'})}{\sum_{u' \in N} |s(u,u')|}, (3)$$

[0111] 其中, u 表示目标用户(如,上文中的第一用户), u' 表示目标用户的一个相似用户(如,上文中的一个第二用户)。 $\Delta r_{u,i}$ 目标用户 u 对候选推荐内容 i 的第二波动分值, $s(u,u')$ 表示目标用户 u 与相似用户 u' 的用户相似度, $r_{u',i}$ 表示相似用户 u' 对候选推荐内容 i 的最新评分, $\bar{r}_{u'}$ 表示相似用户 u' 对已推荐内容的平均评分。

[0112] S550,根据所述目标用户对已推荐内容的平均感兴趣程度及所述第二波动范围,确定所述目标用户当前对所述候选推荐内容的感兴趣程度。

[0113] 在确定目标用户对候选推荐内容的第二波动范围之后,可以按照第二波动范围对目标用户的平均感兴趣程度进行调整,从而得到目标用户对候选推荐内容的预测感兴趣程度。

[0114] 以感兴趣程度是评分为例,当确定了目标用户对候选推荐内容 i 的第二波动分值之后,目标用户当前对候选推荐内容 i 的感兴趣程度 $p_{u,i}$,可以通过如下表达式获得:

$$[0115] \quad p_{u,i} = \bar{r}_u + \Delta r_{u,i}, (4)$$

[0116] 其中, $\bar{r}_u$ 表示目标用户 u 对已推荐内容的平均评分。

[0117] S560,获取所述本轮对话的对话场景信息。

[0118] S570,若所述本轮对话的对话场景信息与第一规则匹配,则按照所述第一规则生成所述下一轮对话的话题信息。

[0119] 本实施例中,对话场景信息可以是任意能够描述当前对话所处场景的信息,比如可以是对话发生时的时间、进行对话的目标用户的用户画像、当前对话的对话意图等等,本实施例对此没有限制。

[0120] 第一规则可以包括一个或多个规则项。每个规则项可以包括匹配域和动作域,匹配域为对话场景信息需要满足的条件,动作域为用于确定与匹配域对应的话题信息的动作指令,即,当对话的对话场景信息满足匹配域中的条件时,需要执行的动作指令。

[0121] 一个例子中,第一规则的一个规则项的匹配域中的条件可以是:当前时间是早上6点至9点之间,且当前对话是用户与用户设备100的首次交互;该规则项的动作域中的动作指令可以表示:将今天的至少一个热点新闻确定为下一轮对话的话题信息。

[0122] 另一个例子中,第二规则的一个规则项的匹配域中的条件可以是:对话发生的日期与母亲节或父亲节的日期之间的间隔小于预设天数(比如,1天、2天、3天等);该规则项的动作域中的动作指令可以表示:将用于询问是否设置日程提醒自己给妈妈/爸爸打电话的询问信息,确定为下一轮对话的话题信息。

[0123] 可以理解,上述的第一规则的规则项仅为举例说明,还可以根据需要为第一规则

灵活设置更多的规则项。本实施例对此没有限定。

[0124] 本实施例中,第一规则可以被配置于用户设备100或服务器200中,如果被配置于用户设备100中,用户设备100直接从本设备访问第一规则。如果第一规则被配置于服务器200中,一种方式中,用户设备100可以从服务器200将第一规则拉取到本地,再对第一规则进行访问,以确定本轮对话的对话场景信息是否与第一规则匹配;另一种方式中,用户设备100可以将本轮对话的对话场景信息发送给服务器200,由服务器200确定该对话场景信息是否与第一规则匹配,并将匹配结果返回给用户设备100。

[0125] 以用户设备100执行上述对话场景信息与第一规则的匹配过程为例,用户设备100可以将对话场景信息与第一规则的各个规则项一一对比,其中,第一规则的各个规则项之间可以设置优先级,对比可以是按照优先级从高到低的顺序进行,一旦确定对话场景信息与某一个规则项匹配,则可以停止对比,从而按照该规则项的动作域中的动作确定话题信息。如果确定对话场景信息与第一规则的所有规则项均不匹配,则可以确定对话场景信息与第一规则不匹配,进而可以按照S580进行处理。

[0126] 服务器200执行匹配的过程与用户设备100类似,在此不再赘述。区别在于,服务器200执行匹配之后,会将匹配结果返回给用户设备100,该匹配结果可以含有表征是否匹配的标识信息。特别地,如果该标识信息表征匹配,匹配结果还可以包括本轮对话的对话场景信息所匹配的规则项的动作域中的动作指令。

[0127] 用户设备100识别该标识信息,可以确定本轮对话的对话场景信息是否与第一规则匹配,如匹配,则按照匹配结果携带的动作指令确定下一轮对话的话题信息。如不匹配,则可以按照S580进行处理。

[0128] S580,若所述本轮对话的对话场景信息与第一规则不匹配,则从各所述候选推荐内容中,将所述目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息。

[0129] 用户设备100可以将目标用户对当前的各候选推荐内容的感兴趣程度(上文提及的预测感兴趣程度)进行排序,并根据排序结果选择感兴趣程度最大的一个候选推荐内容作为下一轮对话的话题信息。

[0130] S590,若所述本轮对话的对话内容、所述本轮对话的对话意图以及所述下一轮对话的话题信息与第二规则匹配,则按照所述第二规则对应的话术模板拼接所述第一回复信息和所述下一轮对话的话题信息,得到所述第二回复信息。

[0131] 其中,第二规则也可以包括一个或多个规则项,每个规则项包括匹配域和动作域,匹配域为需要满足的条件,动作域则是满足匹配域中条件时需要执行的动作指令,该动作指令包含需要使用的話术模板的索引信息或存储位置信息。第二规则可以被配置于用户设备100或服务器200中,与第一规则的匹配过程类似地,可以由用户设备100执行匹配过程,也可以由服务器200执行匹配过程并返回匹配结果。

[0132] 以用户设备100执行第二规则的匹配过程为例,其也可以按照第二规则中各规则项的优先级从高到低的顺序,依次确定本轮对话的对话内容、本轮对话的对话意图以及下一轮对话的话题信息是否与每一规则项匹配,一旦确定与某一规则项匹配,则可以停止匹配过程。

[0133] 可选地,第二规则的每个规则项可以包括针对本轮对话的对话内容、本轮对话的

对话意图以及下一轮对话的话题信息中的一者或多者的条件。本实施例对此没有限制。比如,当某个规则项仅包含针对本轮对话的对话内容的条件时,可以只将本轮对话的对话内容与该规则项的匹配域中内容对比。

[0134] 一个例子中,第二规则的一个规则项的匹配域可以为:对话内容为“早上好”,且下一轮对话的话题信息为“当前热点新闻”;该规则项的动作域 可以包括话术模板t1的索引信息,话术模板t1的内容是:“{xxx},为你精选了今天的热点新闻,先来听听吧”。其中“xxx”用于填充服务处理结果,也就是上文提及的第一回复信息。

[0135] 另一个例子中,第二规则的一个规则项的匹配域可以是:对话意图为“百科且百科对象为[歌手A]”;该规则项的动作域可以包含话术模板t2的索引信息。其中,话术模板t2的内容是:“{xxx},[歌手A/歌手A的昵称]最近刚出了首单曲,要不要听一下?”。其中,“xxx”用于填充上文提及的第一回复信息,歌手A与歌手A的昵称的选取方式可以是,当歌手A存在昵称时,选用歌手A的昵称,否则选用歌手A的真实姓名。比如,歌手A为“周杰伦”,则通过搜索可以确定歌手A的昵称是“周董”。

[0136] 上述的第二规则的规则项仅为举例,还可以根据需要灵活地设置更多的规则项,本实施例对此没有限定。

[0137] S5100,若所述本轮对话的对话内容、所述本轮对话的对话意图以及所述下一轮对话的话题信息与第二规则不匹配,则获取机器学习模型根据所述第一回复信息和所述下一轮对话的话题信息输出的信息,并将获取的信息确定为所述第二回复信息。

[0138] 如果通过匹配过程确定本轮对话的对话内容、本轮对话的对话意图以及下一轮对话的话题信息与第二规则的所有规则项都不匹配,则可以将第一回复信息和下一轮对话的话题信息输入到机器学习模型,并将机器学习模型输出的信息作为第二回复信息。

[0139] 这里的机器学习模型是经过训练的机器学习模型,经过训练的机器学习模型又可以称为自然语言生成引擎。本实施例中,可以在模型仓库中部署多个经过训练的机器学习模型,实施时,可以从模型仓库中选取相应的机器学习模型,以用于实现S5100。

[0140] 以机器学习模型是可控文本生成(Controlled Text Generation,CTG)模型为例,对S5100的实现过程进行介绍。

[0141] 请参照图8,其中示出了应用于本实施例的CTG模型的一个架构示意图。CTG模型可以包括编码器(Encoder)801、生成器(Generator)802和鉴定器(Discriminator)803。其中,编码器和生成器可以视为一个变分自动编码器(Variational Autoencoder,VAE)810。

[0142] 本实施例中,输入信息x可以具有多项句子属性,变分自动编码器 810可以预测输入信息x的各项属性的分布区间,进而按照所预测的分布区间生成对应的文本信息。其中,输入信息x可以包括第一回复信息以及下一轮对话的话题信息,z表示编码器801基于输入信息x输出的隐向量。隐向量z是非结构化编码信息,输入信息x的各项属性的对应编码信息是纠缠在一起的,无法单独地控制某项属性进行定向生成。基于此,可以将隐向量z修正为结构化编码信息c。这里,结构化编码信息c关联着需要控制的特定属性,而非结构化编码信息z则关联着特定属性之外的其他属性。如此,生成器802可以基于(z,c)输出句子x',句子x'可以作为鉴定器803的输入,进而可以基于鉴定器803的输出去训练改变结构化编码信息的分布。经过训练之后,可以直接使用变分自动编码器810来输入的第一回复信息和下一轮对话的话题信息进行处理,继而输出第二回复信息。

[0143] 可以理解,上述的CTG模型仅为举例说明,S5100中的机器学习模型还可以是其他模型,比如,可以是即插即用语言模型(Plug and Play Language Models,PPLM)。本实施例对此没有限制。

[0144] S5110,将所述第二回复信息确定为本轮对话的最终回复信息。

[0145] 本实施例中,用户设备100可以记录有每轮对话的会话状态,该会话状态可以包括每轮对话的会话标识以及当前进度。会话标识例如可以是每轮对话的时间戳信息。示例性地,用户设备100可以在确定本轮对话的对话意图所对应的所有槽位值之后,将会话状态更新为表示待结束的标识符(或者也可以是表示结束的标识符,在这里含义是类似的,都表示可以开启下一轮对话)。

[0146] 在获得第二回复信息之后,可以获取本轮对话的会话状态,如果确定会话状态为待结束(或结束),则可以将第二回复信息确定为最终回复信息,并展示给用户。这里的展示可以是显示或语音播放,本实施例不以此为限制。展示本轮对话的最终回复信息之后,由于最终回复信息实际包含了下一轮对话的话题信息,如果能够接收到用户针对该话题信息输入的对话内容,则可以将会话状态更新为下一轮对话的会话标识,且当前进度为进行中。

[0147] 通过本实施例提供的对话生成方法,可以从丰富的候选推荐内容中为目标用户选择适用于下一轮对话的话题信息,并在回复本轮对话时将该话题信息推荐给目标用户,以便主动开启下一轮对话,从而实现了跨意图的多轮对话的自动开启。并且,通过选用合适的候选推荐内容,使得相邻两轮对话之间的对话内容更加具有连续性,使得用户设备的对话功能更加拟人化,从而可以有效改善用户体验。

[0148] 为了使本领域技术人员更好地理解本申请的方案,下面参照图9a、图9b和图9c所示的处理流程,对用户设备100执行本申请实施例提供的对话生成方法的详细过程进行示例性的介绍。

[0149] 详细地,目标用户在使用用户设备100可以按照图9a所示的整体流程进行处理。目标用户可以通过用户设备100输入对话指令,这个对话指令可以理解成在一轮对话中输入的对话内容,该轮对话可以理解为上述实施例中提及的本轮对话。用户设备100基于输入的对话指令可以进行对话管理,以确定需要执行的任务。具体地,可以对对话指令进行语义理解,并分别根据语义理解结果执行服务处理和任务一,这里的服务处理可以是基于语义理解结果确定对话指令对应的第一回复信息,可以参照上述实施例对S210或S510的详细描述。

[0150] 任务一是进行主动对话内容推荐,也就是确定下一轮对话的话题信息,其实现过程可以参照图9b所示的流程。详细地,可以分别通过推荐引擎和规则引擎对语义理解结果进行处理,其中,语义理解结果包括本轮对话的对话意图和槽位值。推荐引擎可以利用模型仓库中的推荐算法来基于目标用户的用户画像和用户历史数据确定目标用户的相似用户,基于对话意图和槽位值确定内容池中的候选推荐内容,从候选推荐内容中确定下一轮对话的话题信息。规则引擎可以按照规则库中的第一规则来确定下一轮对话的话题信息。其中,对于规则引擎和推荐引擎各自确定的话题信息的排序,可以是:如果规则引擎有确定话题信息,则使用规则引擎确定的话题信息,否则使用推荐引擎确定的话题信息。当然,也可以按照其他排序方式处理。任务一的具体实现过程可以参照上述实施例中对S230-S250的详细描述或者是对S520-S580的详细描述,这里不再继续展开说明。

[0151] 在获得服务处理结果和通过任务一确定下一轮对话的话题信息之后,可以执行任务二。任务二是进行主动对话话术生成与会话管理,比如可以包括生成第二回复信息、更新会话状态、保存历史对话。下面参照图9c所示流程对任务二的执行过程进行介绍。

[0152] 实施过程中,可以根据语义理解与服务响应结果(即,上述的服务处理结果)以及所确定的下一轮对话的话题信息,进行会话更新。这里的会话更新可以是判断本轮对话是否结束,若确定本轮对话可以结束,则将会话状态更新为下一轮对话的会话标识及当前进度(比如为“进行中”),详细实现过程可以参照上文对S5110的描述。如果在会话更新过程中确定本轮对话结束,可以进行对话管理,比如记录本轮对话的对话内容、对话意图和槽位值。

[0153] 然后,可以根据第一回复信息(服务处理结果)以及下一轮对话的话题信息(主动对话推荐结果)进行话术生成。这里的话术生成,可以从规则引擎和自然语言生成引擎中选择一个用于生成第二回复信息,并在话术合成阶段,通过选定的引擎对第一回复信息和下一轮对话的话题信息进行处理,从而生成第二回复信息,并将第二回复信息作为主动对话结果输出。

[0154] 请参阅图10,其示出了本申请实施例提供的一种对话生成装置的结构框图,该装置1000可以包括:获取模块1010、预测模块1020、确定模块1030以及生成模块1040。

[0155] 其中,获取模块1010用于获取目标用户在本轮对话中输入的对话内容,并基于对所述对话内容的语义识别结果,确定所述本轮对话的第一回复信息。

[0156] 预测模块1020用于确定所述目标用户的相似用户,根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度。

[0157] 确定模块1030用于从各所述候选推荐内容中,将所述目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息。

[0158] 生成模块1040用于根据所述本轮对话的第一回复信息和所述下一轮对话的话题信息,生成第二回复信息,并将所述第二回复信息确定为所述本轮对话的最终回复信息。

[0159] 可选地,本申请实施例中,在预测模块1020根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度之前,所述确定模块1030还可以用于:

[0160] 基于所述语义识别结果,确定所述目标用户在所述本轮对话的对话意图和所述对话意图对应的每个槽位值,所述槽位值为生成所述对话意图对应的控制指令所需的信息;从每个用户的历史对话中确定目标对话,所述目标对话的对话意图和槽位值与所述目标用户在所述本轮对话的对话意图和槽位值相同;将在所述目标对话的下一轮对话中输入的对话内容确定为所述候选推荐内容。

[0161] 可选地,本申请实施例中,装置1000还可以包括记录模块。记录模块可以用于:基于所述语义识别结果,确定并记录所述目标用户在所述本轮对话的对话意图;基于所述下一轮对话的话题信息,确定并记录所述目标用户在所述下一轮对话的对话意图。

[0162] 可选地,本申请实施例中,在预测模块1020根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度之前,确定模块1030还可以用于:从各用户的历史对话的对话内容中,将被输入次数符合第一条件的对话内容确定为所述候选推荐内容。

[0163] 可选地,确定模块1030可以通过以下方式从各所述候选推荐内容中,将所述目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息:获取所述本轮对话的对话场景信息;若所述本轮对话的对话场景信息与第一规则匹配,则从各所述候选推荐内容中,将所述目标用户的感兴趣程度最大的候选推荐内容,确定为下一轮对话的话题信息;若所述本轮对话的对话场景信息与所述第一规则不匹配,则按照所述第一规则生成所述下一轮对话的话题信息。

[0164] 可选地,生成模块1040根据所述本轮对话的第一回复信息和所述下一轮对话的话题信息,生成第二回复信息的方式可以是:若所述本轮对话的对话内容、所述本轮对话的对话意图以及所述下一轮对话的话题信息与第二规则匹配,则按照所述第二规则对应的话术模板拼接所述第一回复信息和所述下一轮对话的话题信息,得到所述第二回复信息;若所述本轮对话的对话内容、所述本轮对话的对话意图以及所述下一轮对话的话题信息与第二规则不匹配,则获取机器学习模型根据所述第一回复信息和所述下一轮对话的话题信息输出的信息,并将获取的信息确定为所述第二回复信息。这里的机器学习模型可以是可控文本生成(Controlled Text Generation,CTG)模型、即插即用语言模型(Plug and Play Language Models,PPLM)等。

[0165] 可选地,所述目标用户为第一用户,预测模块1020可以通过以下方式确定所述目标用户的相似用户:

[0166] 基于所述第一用户输入的历史对话内容获取第一标签,所述第一标签是所述第一用户的用户画像标签;基于所述第二用户的历史对话内容获取第二标签,所述第二标签是所述第二用户的用户画像标签;将获取的第一标签与获取的第二标签中的相同标签的数量确定为第一数量,以及确定获取的第一标签的数量与获取的第二标签的数量的乘积,并将该乘积的平方确定为第二数量,并将所述第一数量与所述第二数量的比值确定为所述第二用户与所述第一用户的用户相似度;若所述用户相似度达到第二条件,则将所述第二用户确定为所述第一用户的相似用户。

[0167] 可选地,预测模块1020还可以通过以下方式确定所述目标用户的相似用户:获取所述第一用户评价过的内容组成的第一历史内容集合,以及第二用户评价过的内容组成的第二历史内容集合,并确定所述第一历史内容集合与所述第二历史内容集合的交集;根据所述第一用户和所述第二用户各自对所述交集中同一内容的感兴趣程度,以及所述第一用户和所述第二用户各自对所述交集中内容的平均感兴趣程度,确定所述第一用户对任一内容的感兴趣程度与所述第二用户对任一内容的感兴趣程度之间的皮尔逊相关系数;将所述皮尔逊相关系数确定为所述第二用户与所述第一用户的用户相似度;若所述用户相似度达到第三条件,则将所述第二用户确定为所述第一用户的相似用户。

[0168] 可选地,预测模块1020根据所述相似用户对候选推荐内容的感兴趣程度,预测所述目标用户对所述候选推荐内容的感兴趣程度的方式可以是:

[0169] 获取所述相似用户对所述候选推荐内容的最新感兴趣程度相较于所述相似用户对已推荐内容的平均感兴趣程度的波动范围,并将获取的波动范围作为第一波动范围;根据所述相似用户与所述目标用户的用户相似度以及所述第一波动范围,确定所述目标用户对所述候选推荐内容的感兴趣程度相较于所述目标用户对已推荐内容的平均感兴趣程度的波动范围,并将所确定的波动范围作为第二波动范围;根据所述目标用户对已推荐内容

的平均感兴趣程度及所述第二波动范围,确定所述目标用户当前对所述候选推荐内容的感兴趣程度。

[0170] 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,上述描述装置和模块的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0171] 在本申请所提供的几个实施例中,所显示或讨论的模块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,装置或模块的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0172] 另外,在本申请各个实施例中的各功能模块可以集成在一个处理模块中,也可以是各个模块单独物理存在,也可以两个或两个以上模块集成在一个模块中。上述集成的模块既可以采用硬件的形式实现,也可以采用软件功能模块的形式实现。

[0173] 请参照图11,其示出了本申请实施例提供的一种电子设备的框图,该电子设备1100可以是图1所示的用户设备100。电子设备1100可以包括一个或多个如下部件:处理器1110、存储器1120、以及一个或多个应用程序,其中一个或多个应用程序可以被存储在存储器1120中并被配置为由一个或多个处理器1110执行,一个或多个程序配置用于执行如前述方法实施例所描述的方法。

[0174] 处理器1110可以包括一个或者多个处理核。处理器1110利用各种接口和线路连接整个电子设备1100内的各个部分,通过运行或执行存储在存储器1120内的指令、程序、代码集或指令集,以及调用存储在存储器1120内的数据,执行电子设备1100的各种功能和处理数据。可选地,处理器1110可以采用数字信号处理(Digital Signal Processing,DSP)、现场可编程门阵列(Field-Programmable Gate Array,FPGA)、可编程逻辑阵列(Programmable Logic Array,PLA)中的至少一种硬件形式来实现。处理器1110可集成中央处理器(Central Processing Unit,CPU)、图像处理器(Graphics Processing Unit,GPU)和调制解调器等中的一种或几种的组合。其中,CPU主要处理操作系统、用户界面和应用程序等;GPU用于负责显示内容的渲染和绘制;调制解调器用于处理无线通信。可以理解的是,上述调制解调器也可以不集成到处理器1110中,单独通过一块通信芯片进行实现。

[0175] 存储器1120可以包括随机存储器(Random Access Memory,RAM),也可以包括只读存储器(Read-Only Memory)。存储器1120可用于存储指令、程序、代码、代码集或指令集。存储器1120可包括存储程序区和存储数据区,其中,存储程序区可存储用于实现操作系统的指令、用于实现至少一个功能的指令(比如触控功能、声音播放功能、图像播放功能等)、用于实现下述各个方法实施例的指令等。存储数据区还可以存储终端100在使用中所创建的数据(比如第一规则、第二规则、候选推荐内容)等。

[0176] 请参照图12,其中示出了本申请实施例提供的一种计算机可读存储介质1200的结构框图。该计算机可读介质1200中存储有程序代码,所述程序代码可被处理器调用执行上述方法实施例中所描述的方法。

[0177] 计算机可读存储介质1200可以是诸如闪存、EEPROM(电可擦除可编程只读存储器)、EPROM、硬盘或者ROM之类的电子存储器。可选地,计算机可读存储介质1200包括非瞬态性计算机可读介质(non-transitory computer-readable storage medium)。计算机可读存储介质1200具有执行上述方法中的任何方法步骤的程序代码1210的存储空间。这些程序代码可以从一个或者多个计算机程序产品中读出或者写入到这一个或者多个计算机程序

产品中。程序代码1210可以例如以适当形式进行压缩。

[0178] 最后应说明的是：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，并不驱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。

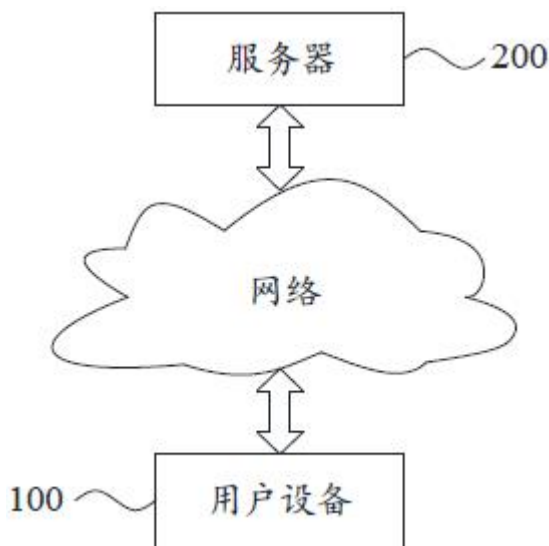


图1

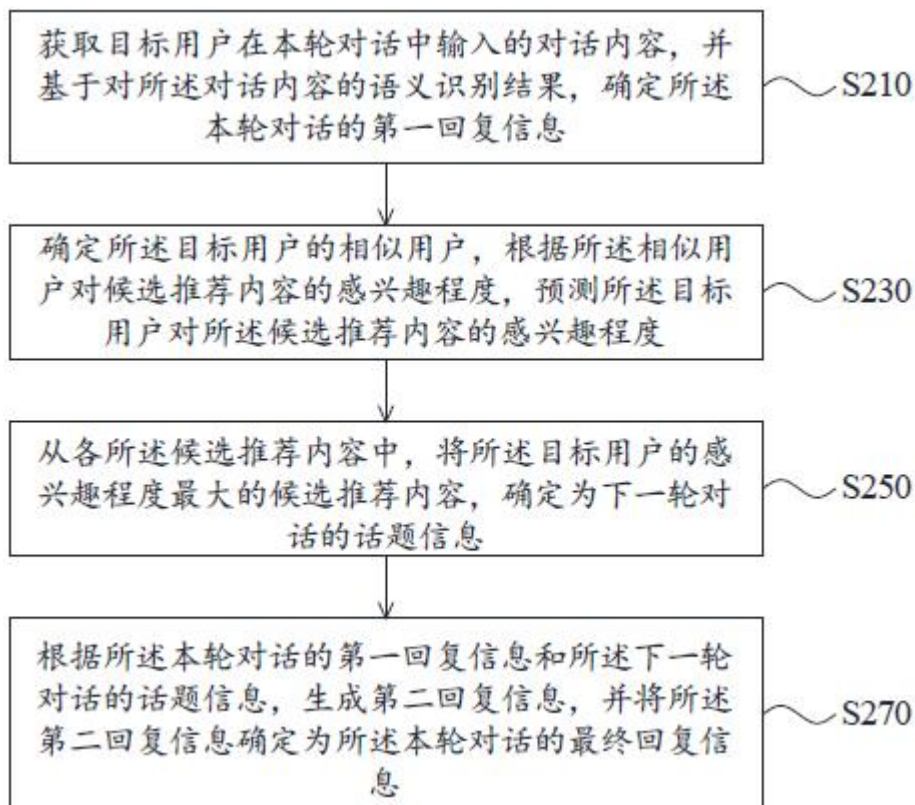


图2

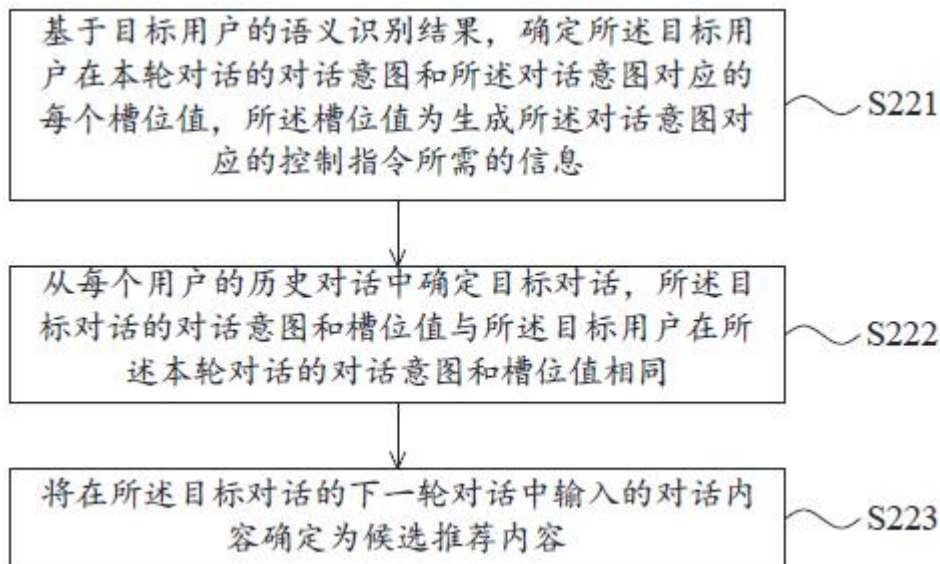


图3

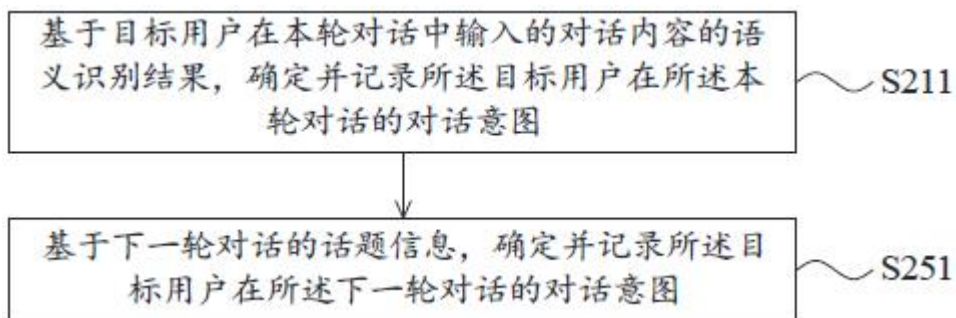


图4

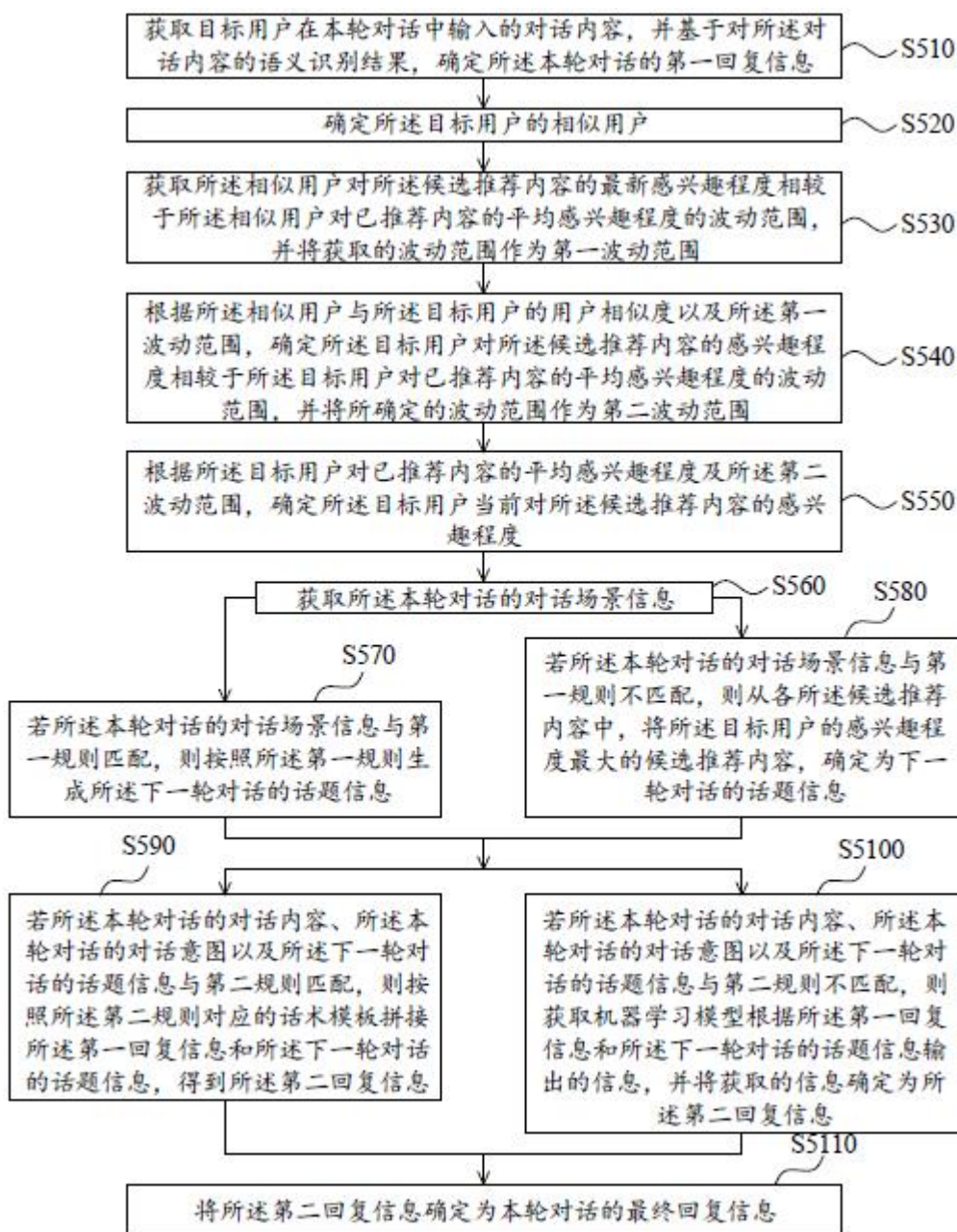


图5

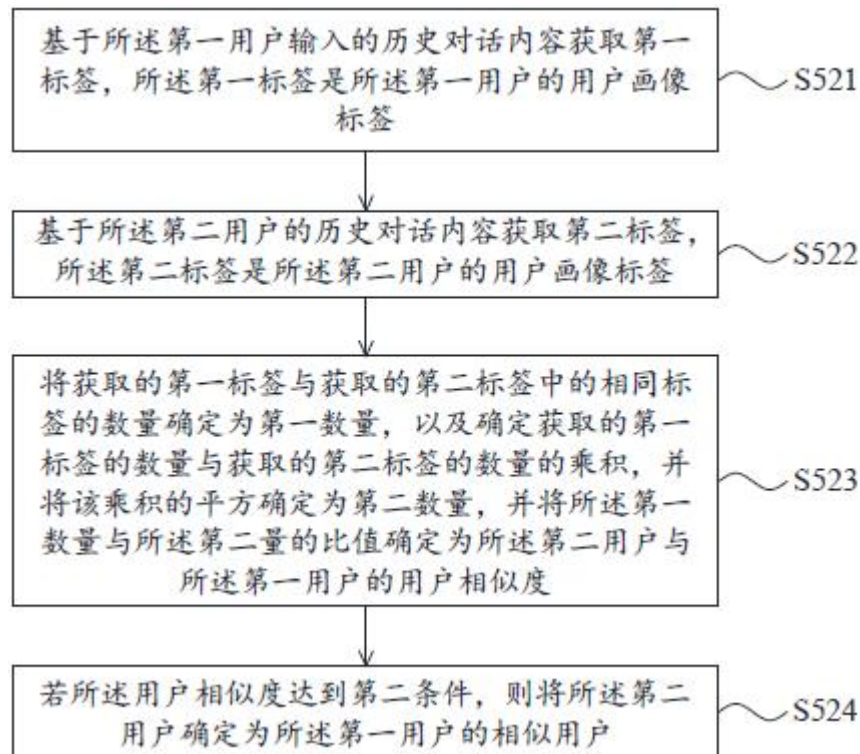


图6

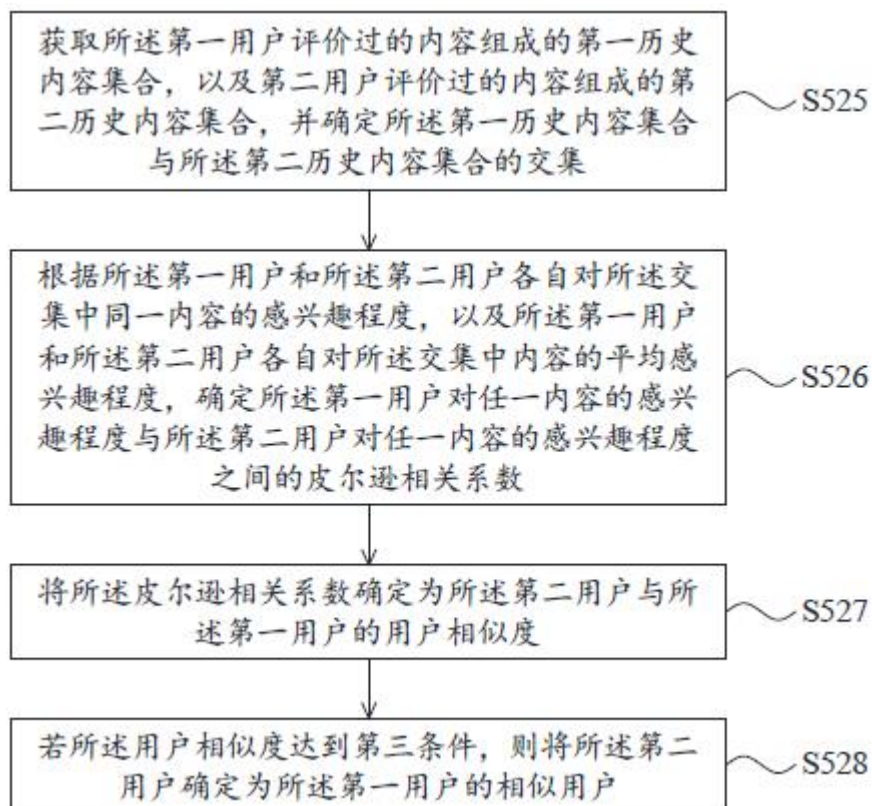


图7

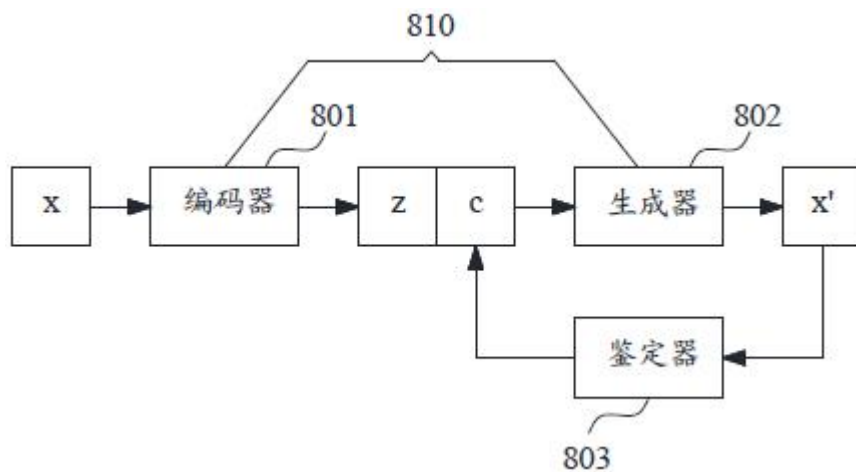


图8

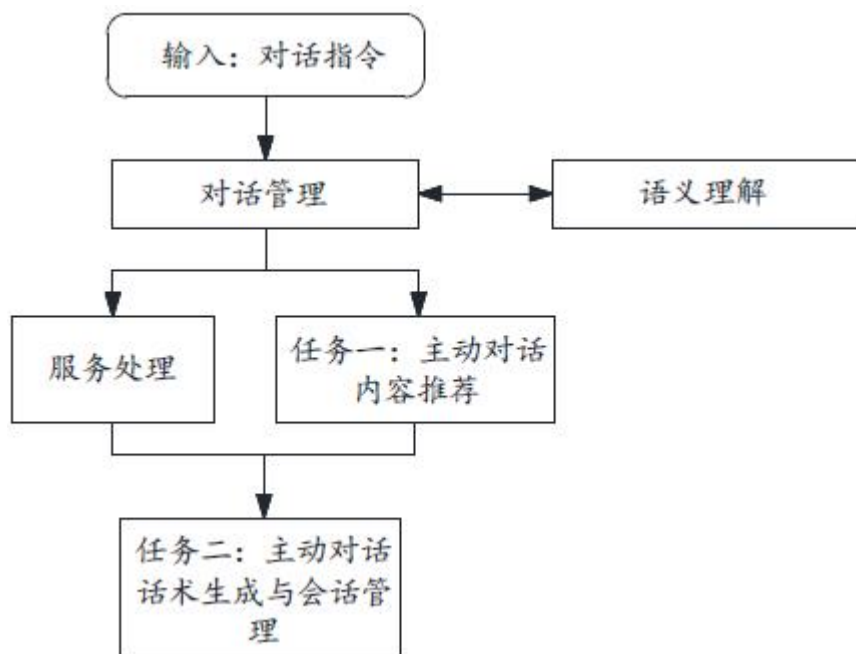


图9a

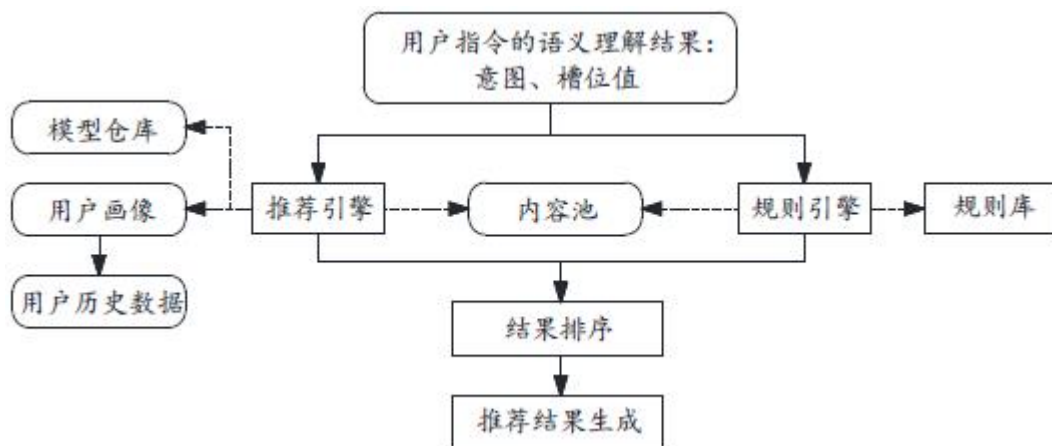


图9b

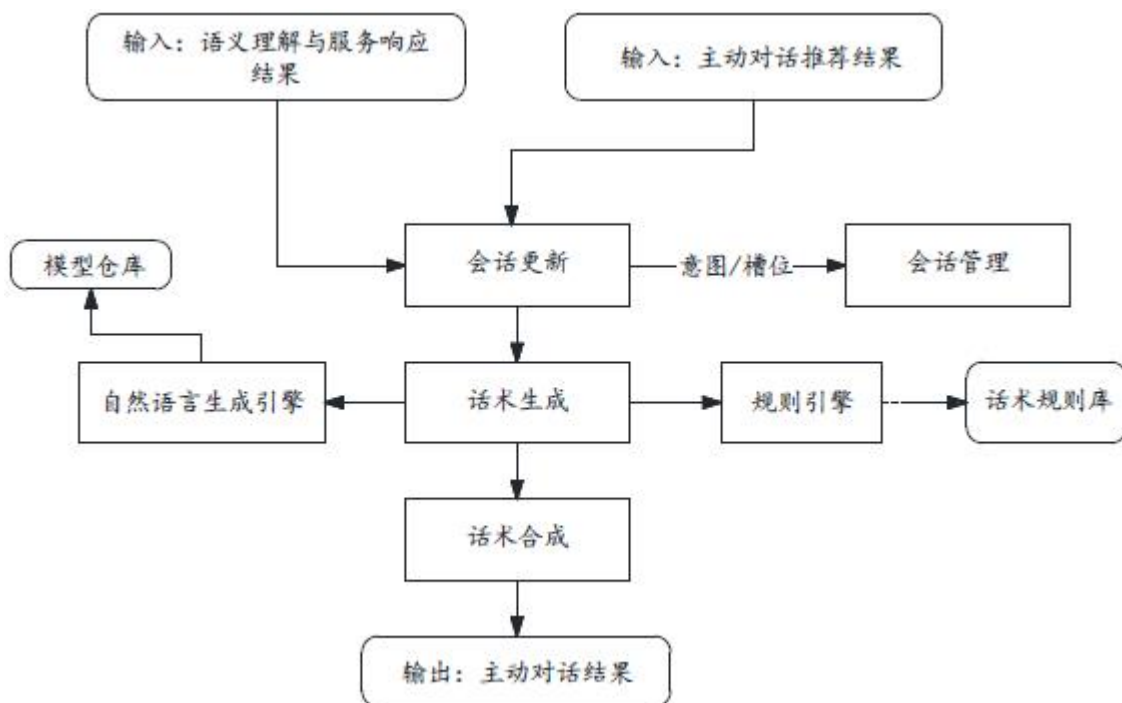


图9c

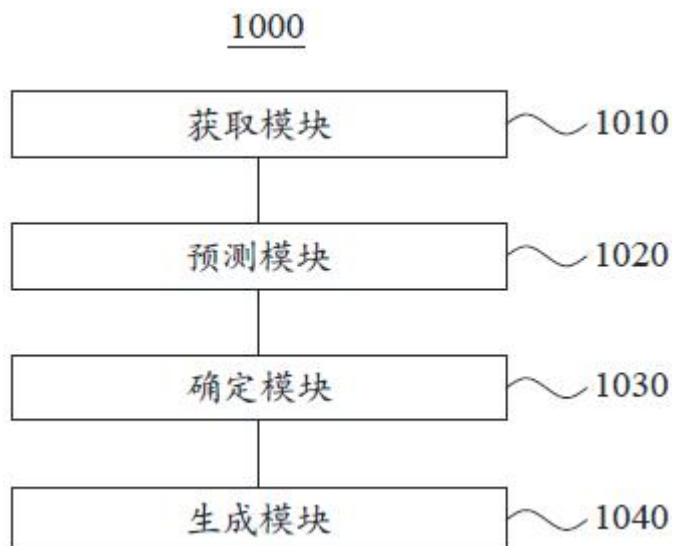


图10

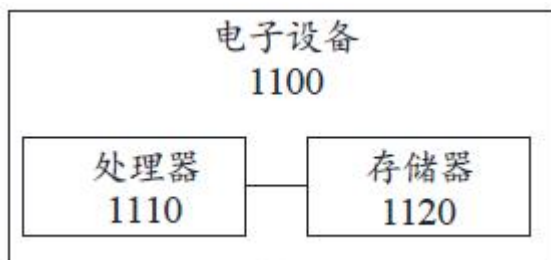


图11

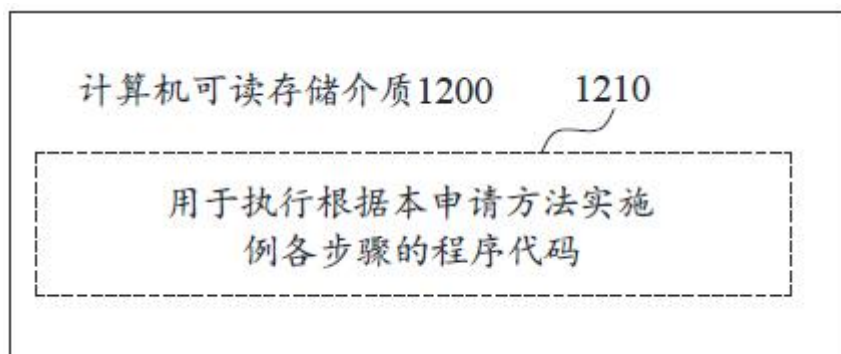


图12