

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2021 年 9 月 23 日 (23.09.2021)



(10) 国际公布号
WO 2021/184547 A1

(51) 国际专利分类号:
G06F 16/332 (2019.01)

(21) 国际申请号: PCT/CN2020/093043

(22) 国际申请日: 2020 年 5 月 28 日 (28.05.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010201001.8 2020年3月20日 (20.03.2020) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 陈亮(CHEN, Liang); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 李治根(LI, Zhigen); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 杨坤(YANG, Kun); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 许开河(XU, Kaihe); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 周琳(ZHOU, Lin); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 王少军(WANG, Shaojun); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(54) Title: CONVERSATION ROBOT INTENTION CORPUS GENERATION METHOD AND APPARATUS, MEDIUM, AND ELECTRONIC DEVICE

(54) 发明名称: 对话机器人意图语料生成方法、装置、介质及电子设备

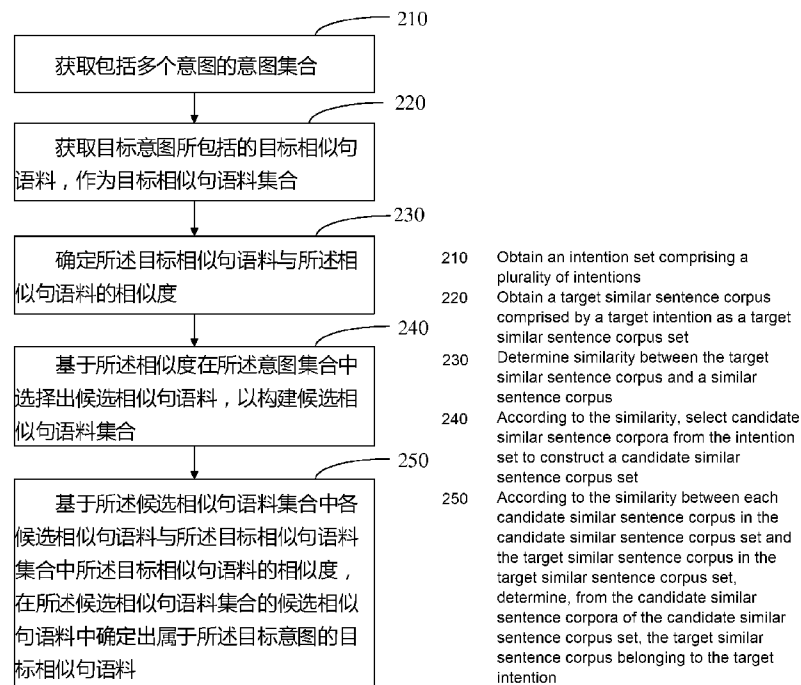


图 2

(57) Abstract: A conversation robot intention corpus generation method and apparatus, a medium, and an electronic device, relating to the field of data processing. The method comprises: obtaining an intention set comprising a plurality of intentions (210); obtaining a target similar sentence corpus of a target intention as a target similar sentence corpus set (220); determining similarity between the target similar sentence corpus and a similar sentence corpus (230); according to the similarity, selecting candidate similar sentence corpora from the intention set to construct a candidate similar sentence corpus set (240); and according to the similarity between each

王嘉雯(WANG, Jiawen); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司(SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路80号汇华商贸大厦1508室, Guangdong 510070 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

candidate similar sentence corpus in the candidate similar sentence corpus set and the target similar sentence corpus in the target similar sentence corpus set, determining, from the candidate similar sentence corpora of the candidate similar sentence corpus set, the target similar sentence corpus belonging to the target intention (250). The method achieves the automatic expansion of intention corpora, increases the number of the intention corpora, and can make the number of the corpora of each intention be more balanced, thereby further improving the accuracy of intention recognition, and further reducing the cost required for expanding the intention corpora.

(57) 摘要: 一种对话机器人意图语料生成方法、装置、介质及电子设备, 涉及数据处理领域。该方法包括: 获取包括多个意图的意图集合(210); 获取目标意图的目标相似句语料, 作为目标相似句语料集合(220); 确定目标相似句语料与相似句语料的相似度(230); 基于相似度在意图集合中选择出候选相似句语料, 以构建候选相似句语料集合(240); 基于候选相似句语料集合中各候选相似句语料与目标相似句语料集合中目标相似句语料的相似度, 在候选相似句语料集合的候选相似句语料中确定出属于目标意图的目标相似句语料(250)。此方法下, 实现了意图语料的自动扩充, 提高了意图语料的数量, 可以使各意图的语料数量更为均衡, 进而提高了意图识别的准确率, 还降低了扩展意图语料所需的成本。

对话机器人意图语料生成方法、装置、介质及电子设备

本申请要求于 2020 年 3 月 20 日提交中国专利局，申请号为 2020102010018、发明名称为“对话机器人意图语料生成方法、装置、介质及电子设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及数据处理技术领域，特别涉及一种对话机器人意图语料生成方法、装置、介质及电子设备。

背景技术

10 目前对话机器人，特别是任务型对话机器人普遍依赖意图识别算法进行意图识别，对话机器人一般根据识别后的意图执行相应的动作，如话术回复，信息查询等。然而，发明人意识到，对话机器人在进行对话时，若要保证对话质量，需要对每个意图下的相似句数量和质量都有很高的要求。不同对话机器人能够执行不同任务的对话，经常存在有些对话机器人积累的低频问题的意图语料较少、各意图语料间数量不均衡等问题，从而降低了意图识别的准确率，此外，如果安排标注人员进行标注，还会浪费大量的人工成本。

15 发明内容

在数据处理技术领域，为了解决上述技术问题，本申请的目的在于提供一种对话机器人意图语料生成方法、装置、介质及电子设备。

根据本申请的一方面，提供了一种对话机器人意图语料生成方法，所述方法包括：

20 获取包括多个意图的意图集合，其中，每一意图包括多个相似句语料，每一意图对应一个对话机器人，每一对话机器人具有至少一个意图；

获取目标意图所包括的目标相似句语料，作为目标相似句语料集合；

确定所述目标相似句语料与所述相似句语料的相似度；

基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合；

25 基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料。

根据本申请的另一方面，提供了一种对话机器人意图语料生成装置，所述装置包括：

30 第一获取模块，被配置为获取包括多个意图的意图集合，其中，每一意图包括多个相似句语料，每一意图对应一个对话机器人，每一对话机器人具有至少一个意图；

第二获取模块，被配置为获取目标意图所包括的目标相似句语料，作为目标相似句语料集合；

第一确定模块，被配置为确定所述目标相似句语料与所述相似句语料的相似度；

35 构建模块，被配置为基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合；

第二确定模块，被配置为基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料。

根据本申请的另一方面，提供了一种电子设备，所述电子设备包括：

40 处理器；

存储器，所述存储器上存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，实现：

获取包括多个意图的意图集合，其中，每一意图包括多个相似句语料，每一意图对应一个对话机器人，每一对话机器人具有至少一个意图；

获取目标意图所包括的目标相似句语料，作为目标相似句语料集合；

确定所述目标相似句语料与所述相似句语料的相似度；

基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合；

5 基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料。

根据本申请的另一方面，提供了一种计算机可读存储介质，其存储有计算机程序指令，当所述计算机程序指令被计算机执行时，使计算机执行如前所述的方法。

10 本申请可实现意图语料的自动扩充，提高了意图语料的数量，可以使各意图的语料数量更为均衡，进而在一定程度上提高了意图识别的准确率，还降低了扩展意图语料所需的成本。

附图说明

15 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。

图 1 是根据一示例性实施例示出的一种对话机器人意图语料生成方法的系统架构示意图；

图 2 是根据一示例性实施例示出的一种对话机器人意图语料生成方法的流程图；

图 3 是根据图 2 对应实施例示出的一实施例的步骤 210 及步骤 220 的细节的流程图；

20 图 4 是根据图 2 对应实施例示出的一实施例的步骤 240 的细节流程图；

图 5 是根据一示例性实施例示出的一种对话机器人意图语料生成装置的框图；

图 6 是根据一示例性实施例示出的一种实现上述对话机器人意图语料生成方法的电子设备示例框图；

25 图 7 是根据一示例性实施例示出的一种实现上述对话机器人意图语料生成方法的计算机可读存储介质。

具体实施方式

30 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

此外，附图仅为本申请的示意性图解，并非一定是按比例绘制。图中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。

35 本申请提供的对话机器人意图语料生成方法，可同样适用于人工智能领域，基于机器学习、深度学习等方面增强对话机器人的普适度。对话机器人可以是各种能够与人类进行人机对话的机器人，对话机器人可以包括多种模型或者算法，比如可以包括语言模型、声学模型等，对话机器人可以与人类进行文本、语音或者视频语音对话。对话机器人的意图和语料的关系就是一种含义的不同表达方式，含义就是意图，而一种具体的表达方式就是一种语料，因此，对话机器人的一个意图通常对应多个相似的语料，不同对话机器人也有不同的意图和语料。对话机器人在进行人机对话时采用意图识别的一般方式为基于统计学习或深度学习的分类算法来学习每个意图下对应了哪些相似的语料来进行意图分类。而意图语料生成就是为对话机器人的某一意图增加语料的过程，也就是说，本申请提供的对话机器人意图语料生成方法可以增加对话机器人的某一意图的语料。

本申请的实施终端可以是任何具有运算、处理以及存储功能的设备，该设备可以与外

部设备相连，用于接收或者发送数据，具体可以是便携移动设备，例如智能手机、平板电脑、笔记本电脑、PDA (Personal Digital Assistant) 等，也可以是固定式设备，例如，计算机设备、现场终端、台式电脑、服务器、工作站等，还可以是多个设备的集合，比如云计算的物理基础设施或者服务器集群。

5 可选地，本申请的实施终端可以为服务器或者云计算的物理基础设施。

图 1 是根据一示例性实施例示出的一种对话机器人意图语料生成方法的系统架构示意图。如图 1 所示，该系统架构包括服务器 110、多个机器人终端 120 以及与每一机器人终端 120 对应的数据库 130，各机器人终端 120 与服务器 110 之间、每一机器人终端 120 与对应的数据库 130 之间均通过通信链路相连，从而可以进行数据的接收和发送。每一机器人终端 120 上固设有对话机器人，而机器人终端 120 对应的数据库 130 则存储有对话机器人进行对话所使用的数据，比如，可以包括意图和对应的语料数据，语料数据比如可以是文本等类型的数据，每一机器人终端 120 对应的数据库 130 可以存储多个意图对应的多个语料数据。在图 1 所示实施例中，服务器 110 为本申请的实施终端，服务器 110 可以经由各机器人终端 120 对各机器人终端 120 对应的数据库 130 中的语料数据进行操作，比如可以从一个机器人终端 120 对应的数据库 130 获取语料数据并将获取的该语料数据迁移至其他机器人终端 120 对应的数据库 130，这样就可以增加某一对话机器人对应意图的语料。

值得一提的是，图 1 仅为本申请的一个实施例。虽然在本实施例中的实施终端为服务器，但在其他实施例中，实施终端可以为如前所述的各种终端或设备；虽然在本实施例中，不同对话机器人固设在不同的终端上，不同对话机器人对应的意图的语料也分别存储于不同的数据库中，但在其他实施例或者具体应用中，各对话机器人和/或各对话机器人对应的意图的语料可以存储于同一终端或不同的终端上，各对话机器人以及对应的意图的语料还可以存储于本申请的实施终端本地，本申请对此不作任何限定，本申请的保护范围也不应因此而受到任何限制。

图 2 是根据一示例性实施例示出的一种对话机器人意图语料生成方法的流程图。本实施例提供的对话机器人意图语料生成方法可以由服务器执行，如图 2 所示，包括以下步骤：步骤 210，获取包括多个意图的意图集合。

其中，每一意图包括多个相似句语料，每一意图对应一个对话机器人，每一对话机器人具有至少一个意图。

每一意图对应一个对话机器人是指，意图是对话机器人的意图，对话机器人可以利用意图对人类进行对话。

在一个实施例中，每一意图包括对话机器人的标识，所述意图通过包括的对话机器人的标识与对话机器人相对应。

如前所述，意图和语料的关系就是一种含义和该含义对应的不同表达方式的关系，一种含义就相当于一种意图，该含义对应的一种具体表达方式就相当于一种语料。同一意图包括的语料之间通常是相似的，所以称为相似句语料。比如，在保险领域，“我不知道医疗险”和“医疗险是什么意思”这两个语料是相似句语料，都属于“我想知道关于医疗险的详细介绍”这一意图。

在一个实施例中，包括多个意图的意图集合 W 可以利用如下表达式来表示：

$$W = [(I_1 \rightarrow S_{11}), (I_1 \rightarrow S_{12}), \dots, (I_x \rightarrow S_{xi})],$$

其中属于同一括号对内的分别为一个意图 I_x 和该意图包括的一个相似句语料 S_{xi} ，比如， I_1 可以代表编号为 1 的意图，而 S_{11} 可以代表该意图包括的第一个相似句语料， S_{12} 可以代表该意图包括的第二个相似句语料，以此类推。

在一个实施例中,所述意图集合预先存储在本地,所述获取包括多个意图的意图集合,包括:从本地读取包括多个意图的意图集合。

在一个实施例中,所述意图集合预先存储在数据库中,所述获取包括多个意图的意图集合,包括:通过查询数据库获取包括多个意图的意图集合。

5 在一个实施例中,所述意图集合预先存储在本端之外的目标终端中,所述获取包括多个意图的意图集合,包括:

向目标终端发送意图集合获取请求;

接收目标终端根据所述意图集合获取请求做出响应时返回的包括多个意图的意图集合。

10 在一个实施例中,所述意图集合还可预先存储至区块链网络节点中,通过区块链存储,实现意图集合在不同平台之间的共享,也可防止数据被篡改。当需要获取意图集合时,可通过调用智能合约的方式直接从区块链中获取。

其中,区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链本质上是一个去中心化的数据库,是一串使用密码学方法相关联产生的数据块,每一个数据块中包含了一批网络交易的信息,用于验证其信息的有效性(防
15 伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层。

步骤 220,获取目标意图所包括的目标相似句语料,作为目标相似句语料集合。

在一个实施例中,所述获取目标意图所包括的目标相似句语料,作为目标相似句语料集合,包括:从在本地的预设路径读取目标意图所包括的目标相似句语料,作为目标相似句语料集合。

20 在一个实施例中,步骤 210 以及步骤 220 的具体步骤可以如图 3 所示。图 3 是根据图 2 对应实施例示出的一实施例的步骤 210 及步骤 220 的细节的流程图。如图 3 所示,包括以下步骤:

步骤 211,基于第一预定规则从包括多个意图的意图总集合中选择出多个意图组成意图子集合。

25 其中,每一意图包括多个相似句语料,所述意图总集合中的每一意图对应一个对话机器人。

可以基于各种方式或规则从意图总集合中选择出组成意图子集合的多个意图,比如,第一预定规则可以是意图总集合中随机选取多个意图组成意图子集合,还可以是按照各意图的生成顺序从包括多个意图的意图总集合中依次选取预定数目个意图组成意图子集合。

30 步骤 221,基于第二预定规则在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中选择出目标意图。

在一个实施例中,所述基于第二预定规则在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中选择出目标意图,包括:

35 在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中选择出包括的相似句语料最少的意图,作为目标意图。

在本实施例中,通过将包括的相似句语料最少的意图作为目标意图,可以为包括的相似句语料最少的意图这类低频意图优先生成语料。

在一个实施例中,所述基于第二预定规则在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中选择出目标意图,包括:

40 在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中确定出包括的相似句语料的数量小于第一预定数目的意图,作为第一候选目标意图;

在所述第一候选目标意图中任取一个,作为目标意图。

在本实施例中,使包括的相似句语料的数量小于第一预定数目的意图均有相同的被选为目标意图的可能性,提高了公平性,并且由于选出的目标意图包括的相似句语料的数量

小于第一预定数目，可以为低频意图优先生成语料。

在一个实施例中，所述基于第二预定规则在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中选择出目标意图，包括：

确定所述意图子集合内各意图所包括的相似句语料的数目的最小值；

5 在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中确定出包括的相似句语料的数目小于所述最小值的意图，作为第二候选目标意图；

在所述第二候选目标意图中任取一个，作为目标意图。

10 当一个意图包括的相似句语料的数目比意图子集合内各意图所包括的相似句语料的数目的最小值还要小时，说明该意图包括的相似句语料的数目是足够小的，在本实施例中通过在这类意图中任选一个作为目标意图，保证了选择的目标意图的合理性。

步骤 222，获取目标意图包括的相似句语料作为目标相似句语料，得到目标相似句语料集合。

本实施例即为从意图集合之外的意图获取相似句语料的示例。

步骤 230，确定所述目标相似句语料与所述相似句语料的相似度。

15 可以利用各种算法或公式来计算两个相似句语料之间的相似度。

在一个实施例中，所述目标相似句语料和所述相似句语料分别由多个词元素组成，所述确定所述目标相似句语料与所述相似句语料的相似度，包括：

利用如下公式确定所述目标相似句语料与所述相似句语料的相似度：

$$f_{score}(s_1, s_2) = \frac{Len(s_1 \cap s_2)}{Len(s_1 \cup s_2)},$$

20 其中， s_1 代表所述目标相似句语料， s_2 代表所述相似句语料， Len 用于求取集合内词元素的个数， $f_{score}(s_1, s_2)$ 为所述目标相似句语料与所述相似句语料的相似度。

比如， $Len(s_1 \cap s_2)$ 用于计算所述目标相似句语料与所述相似句语料共同包含的词元素的个数，而 $Len(s_1 \cup s_2)$ 用于计算所述目标相似句语料与所述相似句语料所包含的所有词元素的个数。

在一个实施例中，所述确定所述目标相似句语料与所述相似句语料的相似度，包括：

25 针对每一所述目标相似句语料，确定该目标相似句语料与每一所述相似句语料的相似度。

在本实施例中，使得确定出的目标相似句语料与相似句语料的相似度的数量最大，从而可以使建立的候选相似句语料集合的规模最大。

在一个实施例中，所述确定所述目标相似句语料与所述相似句语料的相似度，包括：

30 在目标意图所包括的目标相似句语料中任取一个目标相似句语料；

确定任取的该目标相似句语料与每一所述相似句语料的相似度。

步骤 240，基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合。

35 在一个实施例中，步骤 240 的具体步骤可以如图 4 所示。图 4 是根据图 2 对应实施例示出的一实施例的步骤 240 的细节流程图。参照图 4 所示，步骤 240 可以包括以下步骤：

步骤 241，针对所述意图集合中每一意图，若该意图包括的相似句语料中存在一个相似句语料与所述目标相似句语料的相似度大于预定相似度阈值，则获取该意图包括的所有相似句语料作为候选相似句语料。

预定相似度阈值可以是属于[0,1]范围内的浮点数。

40 步骤 242，利用获取的所有候选相似句语料构建候选相似句语料集合。

在本实施例中，在只要一个意图中的相似句语料与目标相似句语料的相似度大于预定相似度阈值的情况下，就选取该意图包括的所有相似句语料作为候选相似句语料来构建候

选相似句语料集合，不仅保证了构建的候选相似句语料集合中候选相似句语料的数量，而且对于一个意图来说，如果确定该意图包括的相似句语料中有一个相似句语料与目标相似句语料的相似度大于预定相似度阈值，就不需要对该意图的其他相似句语料进行判断，还可以减少计算量。

- 5 在一个实施例中，所述确定所述目标相似句语料与所述相似句语料的相似度，包括：
针对每一所述目标相似句语料，确定该目标相似句语料与每一所述相似句语料的相似度；

所述基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合，包括：

- 10 针对每一所述相似句语料，确定各目标相似句语料与该相似句语料的相似度的平均值；
获取所述平均值大于预定相似度平均值阈值的相似句语料所属意图包括的所有相似句语料作为候选相似句语料，并利用获取的所有候选相似句语料构建候选相似句语料集合。

在一个实施例中，所述确定所述目标相似句语料与所述相似句语料的相似度，包括：
针对每一所述目标相似句语料，确定该目标相似句语料与每一所述相似句语料的相似

- 15 度；
所述基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合，包括：

针对每一所述相似句语料，确定各目标相似句语料与该相似句语料的相似度的最大值；
获取所述最大值大于预定相似度最大值阈值的相似句语料作为候选相似句语料，并利

- 20 用获取的所有候选相似句语料构建候选相似句语料集合。

在一个实施例中，所述确定所述目标相似句语料与所述相似句语料的相似度，包括：
针对每一所述目标相似句语料，确定该目标相似句语料与每一所述相似句语料的相似

- 25 度；
所述基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合，包括：

针对每一所述相似句语料，确定各目标相似句语料与该相似句语料的相似度的最小值；
获取所述最小值大于预定相似度最小值阈值的相似句语料所属意图包括的所有相似句语料作为候选相似句语料，并利用获取的所有候选相似句语料构建候选相似句语料集合。

- 30 对于一个相似句语料来说，当各目标相似句语料与该相似句语料的相似度的最小值还比预定相似度最小值阈值大，说明该相似句语料与各目标相似句语料的整体相比是足够相似的，在本实施例中，提高了获取候选相似句语料的标准。

步骤 250，基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料。

- 35 在一个实施例中，步骤 250 可以包括：

基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，利用如下公式计算所述候选相似句语料集合中各候选相似句语料的得分，并基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料：

$$selectSen = \max_{k=1, k \in C}^n \left(\alpha \frac{\sum_{i=1, i \in O}^m f_{score}(S_i, S_k)}{m} - (1 - \alpha) \max_{j=1, j \in O}^m (f_{score}(S_j, S_k)) \right),$$

其中, s_i 和 s_j 代表所述目标相似句语料, s_k 代表所述候选相似句语料, Len 用于求取集合内词元素的个数, $f_{score}(s_1, s_2)$ 为所述目标相似句语料与所述候选相似句语料的相似度, C 为所述候选相似句语料集合, O 为所述目标相似句语料集合, n 为所述候选相似句语料集合中所述候选相似句语料的数目, m 为所述目标相似句语料集合中所述目标相似句语料的数目, α 为权重因子, $selectSen$ 为所述候选相似句语料集合中候选相似句语料的得分。

比如, α 可以为 0.7, 那么 $1 - \alpha$ 为 0.3。

在上述公式中, $\frac{\sum_{i=1}^m, i \in O} f_{score}(s_i, s_k)}$ 这一部分计算了所述目标相似句语料集合中所述目标相似句语料与所述候选相似句语料集合中所述候选相似句语料的相似度的平均值, 即衡量了所述目标相似句语料与所述候选相似句语料的平均相似程度; $Max_{j=1, j \in O}^m (f_{score}(s_j, s_k))$ 这一部分计算了所述目标相似句语料集合中所述目标相似句语料与所述候选相似句语料集合中所述候选相似句语料的相似度的最大值。

因此, 上述公式一方面考虑了选择平均相似度高的候选相似句语料, 可以保证目标相似句语料与原目标意图的目标相似句语料含义相似; 同时也计算了相似性总分上减掉一定权重的候选相似句语料与已有目标相似句语料中某条目标相似句语料最相似的相似度, 可以保证生成的目标相似句语料是对已有目标相似句语料的语义补充。

在一个实施例中, 所述基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度, 利用如下公式计算所述候选相似句语料集合中各候选相似句语料的得分, 并基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料, 包括:

迭代执行目标相似句语料选取步骤, 所述目标相似句语料选取步骤包括:

执行确定候选相似句语料得分步骤, 所述确定候选相似句语料得分步骤包括: 基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度, 利用如下公式计算所述候选相似句语料集合中各候选相似句语料的得分:

$$selectSen = Max_{k=1, k \in C}^n \left(\alpha \frac{\sum_{i=1}^m, i \in O} f_{score}(s_i, s_k)}{m} - (1 - \alpha) Max_{j=1, j \in O}^m (f_{score}(s_j, s_k)) \right),$$

其中, s_i 和 s_j 代表所述目标相似句语料, s_k 代表所述候选相似句语料, Len 用于求取集合内词元素的个数, $f_{score}(s_1, s_2)$ 为所述目标相似句语料与所述候选相似句语料的相似度, C 为所述候选相似句语料集合, O 为所述目标相似句语料集合, n 为所述候选相似句语料集合中所述候选相似句语料的数目, m 为所述目标相似句语料集合中所述目标相似句语料的数目, α 为权重因子, $selectSen$ 为所述候选相似句语料集合中候选相似句语料的得分;

在所述候选相似句语料集合的各候选相似句语料中获取所述得分最高的候选相似句语料, 作为目标候选相似句语料;

若该目标候选相似句语料的得分达到预定得分阈值, 则将该目标候选相似句语料作为目标相似句语料加入至所述目标相似句语料集合, 并将该目标候选相似句语料从所述候选相似句语料集合删除;

转至所述确定候选相似句语料得分步骤, 直至所述目标相似句语料集合中包括的目标相似句语料的数目达到第二预定数目或者对所述候选相似句语料集合的所有候选相似句语料均经过了判断。

在本实施例中, 一方面, 通过在将目标候选相似句语料作为目标相似句语料加入至目

标相似句语料集合后，重新转至确定候选相似句语料得分步骤，利用扩增后的目标相似句语料集合重新计算候选相似句语料集合中各候选相似句语料的得分，使得确定出的候选相似句语料的得分越来越准确，从而保证了加入至目标相似句语料集合的目标相似句语料的质量；另一方面，通过每次选取得分最高并且得分达到预定得分阈值的候选相似句语料加入至目标相似句语料集合，使得加入至目标相似句语料集合的候选相似句语料总是候选相似句语料集合中得分最高的，从而进一步保证了所迁移的目标相似句语料的质量。

在一个实施例中，通过以下方式确定对所述候选相似句语料集合的所有候选相似句语料均经过了判断：

每当对所述候选相似句语料集合的一个候选相似句语料进行了判断，则为该候选相似句语料打上标签，当所述候选相似句语料集合的所有候选相似句语料均被打上标签时，则确定对所述候选相似句语料集合的所有候选相似句语料均经过了判断。

在一个实施例中，所述基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料，包括：

获取所述得分达到预定得分阈值的候选相似句语料，作为属于所述目标意图的目标相似句语料。

在本实施例中，通过得分与预定得分阈值的比较来确定目标相似句语料，保证了选择出的目标相似句语料的合理性。

在一个实施例中，所述基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料，包括：

若所述得分达到预定得分阈值的候选相似句语料的数目达到第三预定数目，则在所述得分达到预定得分阈值的候选相似句语料中任取第三预定数目个候选相似句语料，作为属于所述目标意图的目标相似句语料；

若所述得分达到预定得分阈值的候选相似句语料的数目未达到第三预定数目，则获取所述得分达到预定得分阈值的候选相似句语料，作为属于所述目标意图的目标相似句语料。

在本实施例中，当得分达到预定得分阈值的候选相似句语料的数目过多时，对最终选择的目標相似句语料的数量进行了限制。

在一个实施例中，所述基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料，包括：

执行确定目标候选相似句语料步骤，所述确定目标候选相似句语料步骤包括：在所述候选相似句语料集合的各候选相似句语料中获取所述得分最高的候选相似句语料，作为目标候选相似句语料；

若该目标候选相似句语料的得分达到预定得分阈值，则将该目标候选相似句语料作为目标相似句语料加入至所述目标相似句语料集合，并将该目标候选相似句语料从所述候选相似句语料集合删除；

转至所述确定目标候选相似句语料步骤，直至所述目标相似句语料集合中包括的目标相似句语料的数目达到第二预定数目或者对所述候选相似句语料集合的所有候选相似句语料均经过了判断。

在本实施例中，通过每次所述选取得分最高的候选相似句语料，并在判断该候选相似句语料的得分达到预定得分阈值时就将该候选相似句语料加入至目标相似句语料集合，使得加入至目标相似句语料集合的候选相似句语料得分是最高的，从而保证了所迁移的目标相似句语料的质量。

在一个实施例中，所述基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料，包括：

对所述候选相似句语料集合的各候选相似句语料按照所述得分从高到低的顺序进行排

序；

按照所述排序的顺序，每次选取一个候选相似句语料，若该候选相似句语料的得分达到预定得分阈值，则将该候选相似句语料作为目标相似句语料加入至所述目标相似句语料集合，并将该候选相似句语料从所述候选相似句语料集合删除，直至所述目标相似句语料集合中包括的目标相似句语料的数目达到第二预定数目或者选取的候选相似句语料的得分未达到预定得分阈值。

综上所述，根据图 2 实施例提供的对话机器人意图语料生成方法，通过利用知识迁移的方式，将其他意图的语料迁移到需要扩展的意图中，从而实现意图语料的自动扩充，提高了意图语料的数量，可以使各意图的语料数量更为均衡，进而在一定程度上提高了意图识别的准确率，还降低了扩展意图语料所需的成本。

本申请还提供了一种对话机器人意图语料生成装置，以下是本申请的装置实施例。

图 5 是根据一示例性实施例示出的一种对话机器人意图语料生成装置的框图。如图 5 所示，装置 500 包括：

第一获取模块 510，被配置为获取包括多个意图的意图集合，其中，每一意图包括多个相似句语料，每一意图对应一个对话机器人，每一对话机器人具有至少一个意图；

第二获取模块 520，被配置为获取目标意图所包括的目标相似句语料，作为目标相似句语料集合；

第一确定模块 530，被配置为确定所述目标相似句语料与所述相似句语料的相似度；

构建模块 540，被配置为基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合；

第二确定模块 550，被配置为基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料。

根据本申请的第三方面，还提供了一种能够实现上述方法的电子设备。

所属技术领域的技术人员能够理解，本申请的各个方面可以实现为系统、方法或程序产品。因此，本申请的各个方面可以具体实现为以下形式，即：完全的硬件实施方式、完全的软件实施方式（包括固件、微代码等），或硬件和软件方面结合的实施方式，这里可以统称为“电路”、“模块”或“系统”。

下面参照图 6 来描述根据本申请的这种实施方式的电子设备 600。图 6 显示的电子设备 600 仅仅是一个示例，不应对本申请实施例的功能和使用范围带来任何限制。

如图 6 所示，电子设备 600 以通用计算设备的形式表现。电子设备 600 的组件可以包括但不限于：上述至少一个处理单元 610、上述至少一个存储单元 620、连接不同系统组件（包括存储单元 620 和处理单元 610）的总线 630。

其中，所述存储单元存储有程序代码，所述程序代码可以被所述处理单元 610 执行，使得所述处理单元 610 执行本说明书上述“实施例方法”部分中描述的根据本申请各种示例性实施方式的步骤。

存储单元 620 可以包括易失性存储单元形式的可读存储介质，例如随机存取存储单元（RAM）621 和/或高速缓存存储单元 622，还可以进一步包括只读存储单元（ROM）623。

存储单元 620 还可以包括具有一组（至少一个）程序模块 625 的程序/实用工具 624，这样的程序模块 625 包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

总线 630 可以为表示几类总线结构中的一种或多种，包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

电子设备 600 也可以与一个或多个外部设备 800 (例如键盘、指向设备、蓝牙设备等) 通信, 还可与一个或者多个使得用户能与该电子设备 600 交互的设备通信, 和/或与使得该电子设备 600 能与一个或多个其它计算设备进行通信的任何设备 (例如路由器、调制解调器等等) 通信。这种通信可以通过输入/输出 (I/O) 接口 650 进行。并且, 电子设备 600 还可以通过网络适配器 660 与一个或者多个网络 (例如局域网 (LAN), 广域网 (WAN) 和/或公共网络, 例如因特网) 通信。如图所示, 网络适配器 660 通过总线 630 与电子设备 600 的其它模块通信。应当明白, 尽管图中未示出, 可以结合电子设备 600 使用其它硬件和/或软件模块, 包括但不限于: 微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID 系统、磁带驱动器以及数据备份存储系统等。

通过以上的实施方式的描述, 本领域的技术人员易于理解, 这里描述的示例实施方式可以通过软件实现, 也可以通过软件结合必要的硬件的方式来实现。因此, 根据本申请实施方式的技术方案可以以软件产品的形式体现出来, 该软件产品可以存储在一个非易失性存储介质 (可以是 CD-ROM, U 盘, 移动硬盘等) 中或网络上, 包括若干指令以使得一台计算设备 (可以是个人计算机、服务器、终端装置、或者网络设备等) 执行根据本申请实施方式的方法。

根据本申请的第四方面, 还提供了一种计算机可读存储介质, 其上存储有能够实现本说明书上述方法的程序产品。其中, 该计算机可读存储介质可以是非易失性, 也可以是易失性。在一些可能的实施方式中, 本申请的各个方面还可以实现为一种程序产品的形式, 其包括程序代码, 当所述程序产品在终端设备上运行时, 所述程序代码用于使所述终端设备执行本说明书上述“示例性方法”部分中描述的根据本申请各种示例性实施方式步骤。

参考图 7 所示, 描述了根据本申请的实施方式的用于实现上述方法的程序产品 700, 其可以采用便携式紧凑盘只读存储器 (CD-ROM) 并包括程序代码, 并可以在终端设备, 例如个人电脑上运行。然而, 本申请的程序产品不限于此, 在本文件中, 可读存储介质可以是任何包含或存储程序的有形介质, 该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

所述程序产品可以采用一个或多个可读存储介质的任意组合。可读存储介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件, 或者任意以上的组合。可读存储介质的更具体的例子 (非穷举的列表) 包括: 具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器 (RAM)、只读存储器 (ROM)、可擦式可编程只读存储器 (EPROM 或闪存)、光纤、便携式紧凑盘只读存储器 (CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号, 其中承载了可读程序代码。这种传播的数据信号可以采用多种形式, 包括但不限于电磁信号、光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读存储介质, 该可读存储介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

可读存储介质上包含的程序代码可以用任何适当的介质传输, 包括但不限于无线、有线、光缆、RF 等等, 或者上述的任意合适的组合。

可以以一种或多种程序设计语言的任意组合来编写用于执行本申请操作的程序代码, 所述程序设计语言包括面向对象的程序设计语言—诸如 Java、C++ 等, 还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程

计算设备的情形中，远程计算设备可以通过任意种类的网络，包括局域网（LAN）或广域网（WAN），连接到用户计算设备，或者，可以连接到外部计算设备（例如利用因特网服务提供商来通过因特网连接）。

- 应当理解的是，本申请并不局限于上面已经描述并在附图中示出的精确结构，并且可以
5 在不脱离其范围执行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

权利要求

1、一种对话机器人意图语料生成方法，其中，所述方法包括：

获取包括多个意图的意图集合，其中，每一意图包括多个相似句语料，每一意图对应一个对话机器人，每一对话机器人具有至少一个意图；

5 获取目标意图所包括的目标相似句语料，作为目标相似句语料集合；

确定所述目标相似句语料与所述相似句语料的相似度；

基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合；

10 基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料。

2、根据权利要求1所述的方法，其中，所述意图集合为意图子集合，所述获取包括多个意图的意图集合，包括：

15 基于第一预定规则从包括多个意图的意图总集合中选择出多个意图组成意图子集合，其中，每一意图包括多个相似句语料，所述意图总集合中的每一意图对应一个对话机器人；

所述获取目标意图所包括的目标相似句语料，作为目标相似句语料集合，包括：

基于第二预定规则在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中选择出目标意图；

获取目标意图包括的相似句语料作为目标相似句语料，得到目标相似句语料集合。

20 3、根据权利要求1或2所述的方法，其中，所述目标相似句语料和所述相似句语料分别由多个词元素组成，所述确定所述目标相似句语料与所述相似句语料的相似度，包括：

利用如下公式确定所述目标相似句语料与所述相似句语料的相似度：

$$f_{score}(s_1, s_2) = \frac{Len(s_1 \cap s_2)}{Len(s_1 \cup s_2)},$$

其中， s_1 代表所述目标相似句语料， s_2 代表所述相似句语料， Len 用于求取集合内词元素的个数， $f_{score}(s_1, s_2)$ 为所述目标相似句语料与所述相似句语料的相似度。

25 4、根据权利要求1或2所述的方法，其中，所述基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合，包括：

针对所述意图集合中每一意图，若该意图包括的相似句语料中存在一个相似句语料与所述目标相似句语料的相似度大于预定相似度阈值，则获取该意图包括的所有相似句语料作为候选相似句语料；

30 利用获取的所有候选相似句语料构建候选相似句语料集合。

5、根据权利要求1或2所述的方法，其中，所述基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料，包括：

35 基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，利用如下公式计算所述候选相似句语料集合中各候选相似句语料的得分，并基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料：

$$selectSen = \max_{k=1, k \in C}^n \left(\alpha \frac{\sum_{i=1}^m f_{score}(S_i, S_k)}{m} - (1 - \alpha) \max_{j=1, j \in C}^m (f_{score}(S_j, S_k)) \right),$$

其中, s_i 和 s_j 代表所述目标相似句语料, s_k 代表所述候选相似句语料, Len 用于求取集合内词元素的个数, $f_{score}(s_1, s_2)$ 为所述目标相似句语料与所述候选相似句语料的相似度, C 为所述候选相似句语料集合, O 为所述目标相似句语料集合, n 为所述候选相似句语料集合中所述候选相似句语料的数目, m 为所述目标相似句语料集合中所述目标相似句语料的数目, α 为权重因子, $selectSen$ 为所述候选相似句语料集合中候选相似句语料的得分。

6、根据权利要求 5 所述的方法, 其中, 所述基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度, 利用如下公式计算所述候选相似句语料集合中各候选相似句语料的得分, 并基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料, 包括:

迭代执行目标相似句语料选取步骤, 所述目标相似句语料选取步骤包括:
执行确定候选相似句语料得分步骤, 所述确定候选相似句语料得分步骤包括: 基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度, 利用如下公式计算所述候选相似句语料集合中各候选相似句语料的得分:

$$selectSen = \max_{k=1, k \in C}^n \left(\alpha \frac{\sum_{i=1}^m, i \in O} f_{score}(s_i, s_k)}{m} - (1 - \alpha) \max_{j=1, j \in O}^m (f_{score}(s_j, s_k)) \right),$$

其中, s_i 和 s_j 代表所述目标相似句语料, s_k 代表所述候选相似句语料, Len 用于求取集合内词元素的个数, $f_{score}(s_1, s_2)$ 为所述目标相似句语料与所述候选相似句语料的相似度, C 为所述候选相似句语料集合, O 为所述目标相似句语料集合, n 为所述候选相似句语料集合中所述候选相似句语料的数目, m 为所述目标相似句语料集合中所述目标相似句语料的数目, α 为权重因子, $selectSen$ 为所述候选相似句语料集合中候选相似句语料的得分;

在所述候选相似句语料集合的各候选相似句语料中获取所述得分最高的候选相似句语料, 作为目标候选相似句语料;

若该目标候选相似句语料的得分达到预定得分阈值, 则将该目标候选相似句语料作为目标相似句语料加入至所述目标相似句语料集合, 并将该目标候选相似句语料从所述候选相似句语料集合删除;

转至所述确定候选相似句语料得分步骤, 直至目标相似句语料集合中包括的目标相似句语料的数目达到第二预定数目或者对所述候选相似句语料集合的所有候选相似句语料均经过了判断。

7、根据权利要求 5 所述的方法, 其中, 所述基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料, 包括:

执行确定目标候选相似句语料步骤, 所述确定目标候选相似句语料步骤包括: 在所述候选相似句语料集合的各候选相似句语料中获取所述得分最高的候选相似句语料, 作为目标候选相似句语料;

若该目标候选相似句语料的得分达到预定得分阈值, 则将该目标候选相似句语料作为目标相似句语料加入至所述目标相似句语料集合, 并将该目标候选相似句语料从所述候选相似句语料集合删除;

转至所述确定目标候选相似句语料步骤, 直至所述目标相似句语料集合中包括的目标相似句语料的数目达到第二预定数目或者对所述候选相似句语料集合的所有候选相似句语料均经过了判断。

8、根据权利要求 2 所述的方法, 其中, 所述基于第二预定规则在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中选择出目标意图, 包括:

在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中确定出包括的相似句语料的数量小于第一预定数目的意图，作为第一候选目标意图；

在所述第一候选目标意图中任取一个，作为目标意图。

9、根据权利要求 5 所述的方法，其中，所述基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料，包括：

若所述得分达到预定得分阈值的候选相似句语料的数目达到第三预定数目，则在所述得分达到预定得分阈值的候选相似句语料中任取第三预定数目个候选相似句语料，作为属于所述目标意图的目标相似句语料；

若所述得分达到预定得分阈值的候选相似句语料的数目未达到第三预定数目，则获取所述得分达到预定得分阈值的候选相似句语料，作为属于所述目标意图的目标相似句语料。

10、一种对话机器人意图语料生成装置，其中，所述装置包括：

第一获取模块，被配置为获取包括多个意图的意图集合，其中，每一意图包括多个相似句语料，每一意图对应一个对话机器人，每一对话机器人具有至少一个意图；

第二获取模块，被配置为获取目标意图所包括的目标相似句语料，作为目标相似句语料集合；

第一确定模块，被配置为确定所述目标相似句语料与所述相似句语料的相似度；

构建模块，被配置为基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合；

第二确定模块，被配置为基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料。

11、一种电子设备，其中，所述电子设备包括：

处理器；

存储器，所述存储器上存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，实现：

获取包括多个意图的意图集合，其中，每一意图包括多个相似句语料，每一意图对应一个对话机器人，每一对话机器人具有至少一个意图；

获取目标意图所包括的目标相似句语料，作为目标相似句语料集合；

确定所述目标相似句语料与所述相似句语料的相似度；

基于所述相似度在所述意图集合中选择出候选相似句语料，以构建候选相似句语料集合；

基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度，在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料。

12、根据权利要求 11 所述的电子设备，其中，所述意图集合为意图子集合，所述计算机可读指令被所述处理器执行时，实现：

基于第一预定规则从包括多个意图的意图总集合中选择出多个意图组成意图子集合，其中，每一意图包括多个相似句语料，所述意图总集合中的每一意图对应一个对话机器人；

所述获取目标意图所包括的目标相似句语料，作为目标相似句语料集合，包括：

基于第二预定规则在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中选择出目标意图；

获取目标意图包括的相似句语料作为目标相似句语料，得到目标相似句语料集合。

13、根据权利要求 11 或 12 所述的电子设备，其中，所述目标相似句语料和所述相似句语料分别由多个词元素组成，所述计算机可读指令被所述处理器执行时，实现：

利用如下公式确定所述目标相似句语料与所述相似句语料的相似度：

$$f_{score}(s_1, s_2) = \frac{Len(s_1 \cap s_2)}{Len(s_1 \cup s_2)},$$

其中, s_1 代表所述目标相似句语料, s_2 代表所述相似句语料, Len 用于求取集合内词元素的个数, $f_{score}(s_1, s_2)$ 为所述目标相似句语料与所述相似句语料的相似度。

14、根据权利要求 11 或 12 所述的电子设备, 其中, 所述计算机可读指令被所述处理器执行时, 实现:

针对所述意图集合中每一意图, 若该意图包括的相似句语料中存在一个相似句语料与所述目标相似句语料的相似度大于预定相似度阈值, 则获取该意图包括的所有相似句语料作为候选相似句语料;

利用获取的所有候选相似句语料构建候选相似句语料集合。

15、根据权利要求 11 或 12 所述的电子设备, 其中, 所述计算机可读指令被所述处理器执行时, 实现:

基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度, 利用如下公式计算所述候选相似句语料集合中各候选相似句语料的得分, 并基于所述得分在所述候选相似句语料集合的候选相似句语料中确定出属于所述目标意图的目标相似句语料:

$$selectSen = \max_{k=1, k \in C}^n \left(\alpha \frac{\sum_{i=1, i \in O}^m f_{score}(s_i, s_k)}{m} - (1 - \alpha) \max_{j=1, j \in O}^m (f_{score}(s_j, s_k)) \right),$$

其中, s_i 和 s_j 代表所述目标相似句语料, s_k 代表所述候选相似句语料, Len 用于求取集合内词元素的个数, $f_{score}(s_1, s_2)$ 为所述目标相似句语料与所述候选相似句语料的相似度, C 为所述候选相似句语料集合, O 为所述目标相似句语料集合, n 为所述候选相似句语料集合中所述候选相似句语料的数目, m 为所述目标相似句语料集合中所述目标相似句语料的数目, α 为权重因子, $selectSen$ 为所述候选相似句语料集合中候选相似句语料的得分。

16、根据权利要求 15 所述的电子设备, 其中, 所述计算机可读指令被所述处理器执行时, 实现:

迭代执行目标相似句语料选取步骤, 所述目标相似句语料选取步骤包括:

执行确定候选相似句语料得分步骤, 所述确定候选相似句语料得分步骤包括: 基于所述候选相似句语料集合中各候选相似句语料与所述目标相似句语料集合中所述目标相似句语料的相似度, 利用如下公式计算所述候选相似句语料集合中各候选相似句语料的得分:

$$selectSen = \max_{k=1, k \in C}^n \left(\alpha \frac{\sum_{i=1, i \in O}^m f_{score}(s_i, s_k)}{m} - (1 - \alpha) \max_{j=1, j \in O}^m (f_{score}(s_j, s_k)) \right),$$

其中, s_i 和 s_j 代表所述目标相似句语料, s_k 代表所述候选相似句语料, Len 用于求取集合内词元素的个数, $f_{score}(s_1, s_2)$ 为所述目标相似句语料与所述候选相似句语料的相似度, C 为所述候选相似句语料集合, O 为所述目标相似句语料集合, n 为所述候选相似句语料集合中所述候选相似句语料的数目, m 为所述目标相似句语料集合中所述目标相似句语料的数目, α 为权重因子, $selectSen$ 为所述候选相似句语料集合中候选相似句语料的得分;

在所述候选相似句语料集合的各候选相似句语料中获取所述得分最高的候选相似句语料, 作为目标候选相似句语料;

若该目标候选相似句语料的得分达到预定得分阈值，则将该目标候选相似句语料作为目标相似句语料加入至所述目标相似句语料集合，并将该目标候选相似句语料从所述候选相似句语料集合删除；

5 转至所述确定候选相似句语料得分步骤，直至目标相似句语料集合中包括的目标相似句语料的数目达到第二预定数目或者对所述候选相似句语料集合的所有候选相似句语料均经过了判断。

17、根据权利要求 15 所述的电子设备，其中，所述计算机可读指令被所述处理器执行时，实现：

10 执行确定目标候选相似句语料步骤，所述确定目标候选相似句语料步骤包括：在所述候选相似句语料集合的各候选相似句语料中获取所述得分最高的候选相似句语料，作为目标候选相似句语料；

若该目标候选相似句语料的得分达到预定得分阈值，则将该目标候选相似句语料作为目标相似句语料加入至所述目标相似句语料集合，并将该目标候选相似句语料从所述候选相似句语料集合删除；

15 转至所述确定目标候选相似句语料步骤，直至所述目标相似句语料集合中包括的目标相似句语料的数目达到第二预定数目或者对所述候选相似句语料集合的所有候选相似句语料均经过了判断。

18、根据权利要求 12 所述的电子设备，其中，所述计算机可读指令被所述处理器执行时，实现：

20 在所述意图子集合中意图对应的对话机器人之外的其他所有对话机器人对应的意图中确定出包括的相似句语料的数目小于第一预定数目的意图，作为第一候选目标意图；

在所述第一候选目标意图中任取一个，作为目标意图。

19、根据权利要求 15 所述的电子设备，其中，所述计算机可读指令被所述处理器执行时，实现：

25 若所述得分达到预定得分阈值的候选相似句语料的数目达到第三预定数目，则在所述得分达到预定得分阈值的候选相似句语料中任取第三预定数目个候选相似句语料，作为属于所述目标意图的目标相似句语料；

若所述得分达到预定得分阈值的候选相似句语料的数目未达到第三预定数目，则获取所述得分达到预定得分阈值的候选相似句语料，作为属于所述目标意图的目标相似句语料。

30 20、一种计算机可读存储介质，其中，其存储有计算机程序指令，当所述计算机程序指令被计算机执行时，使计算机执行根据权利要求 1 至 9 中任一项所述的方法。

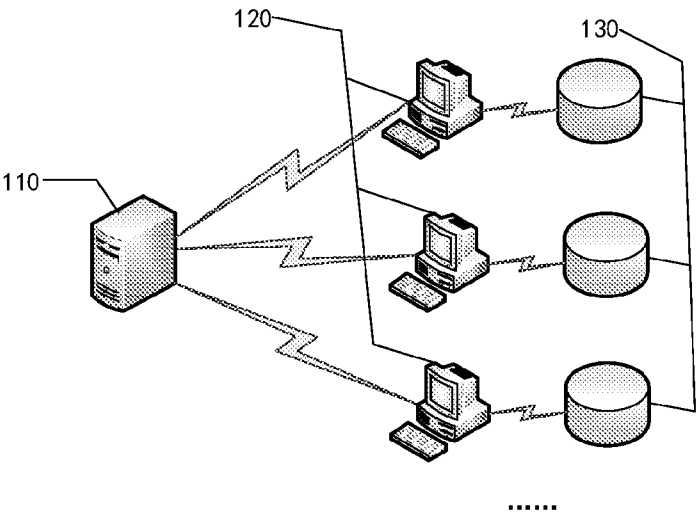


图 1

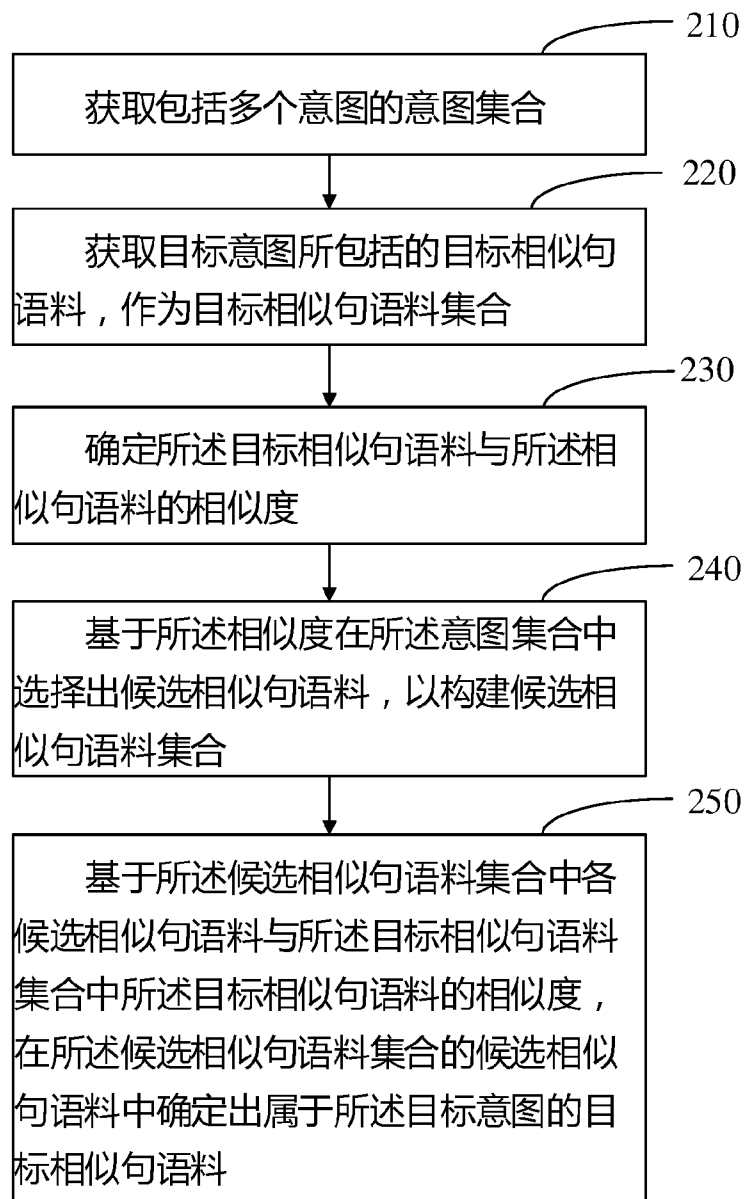


图 2

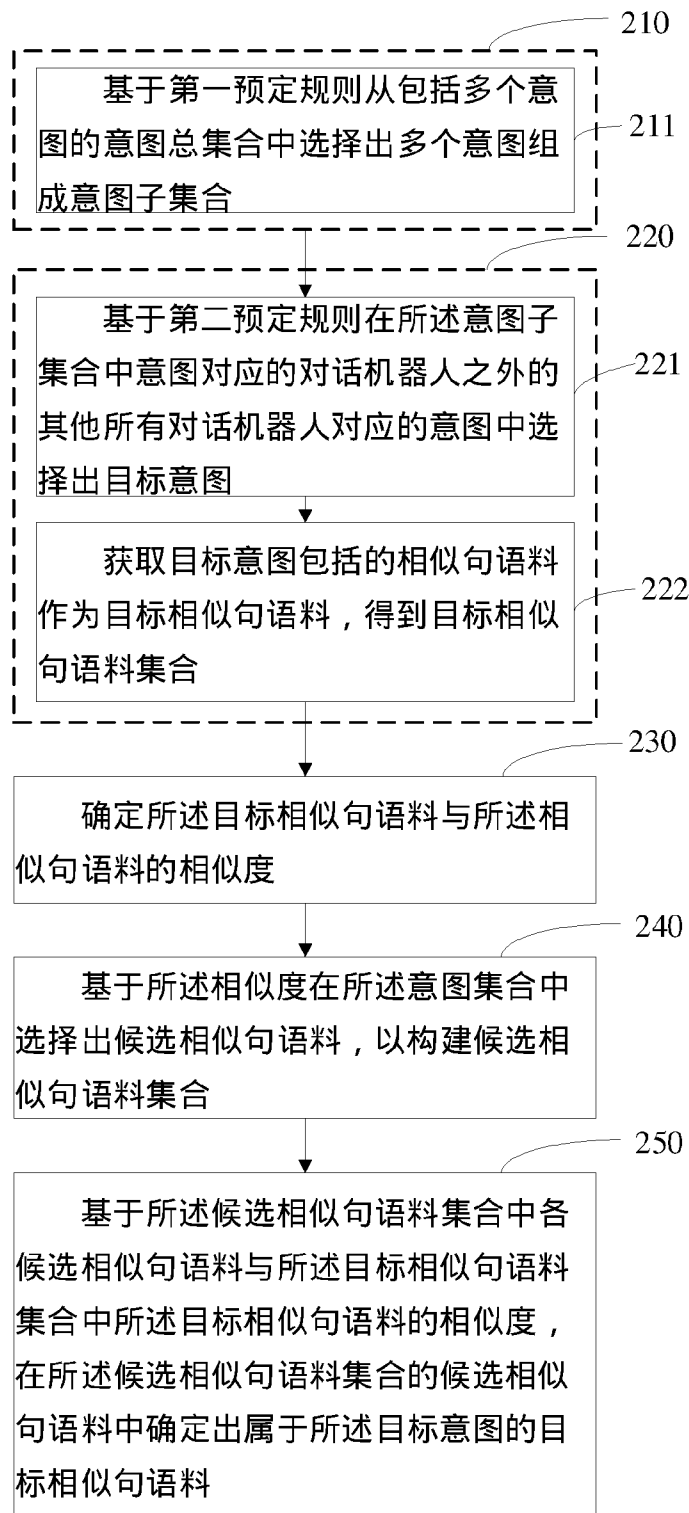


图 3

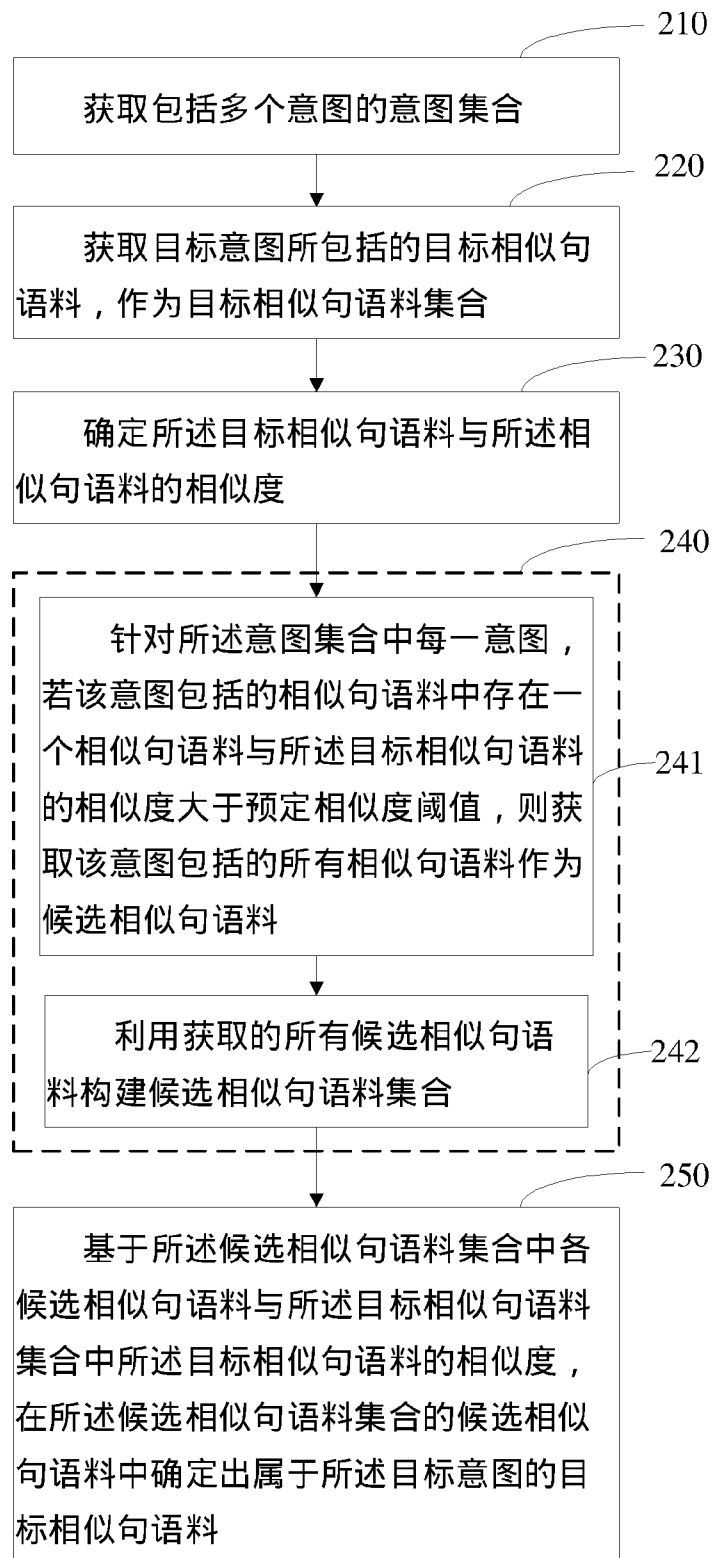


图 4



图 5

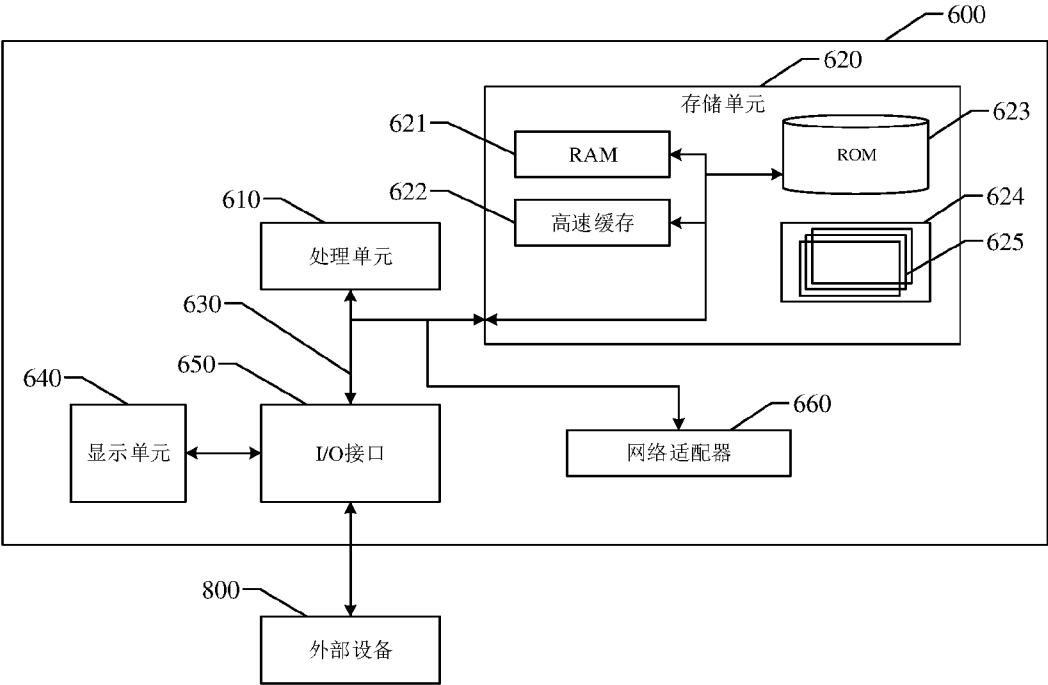


图 6

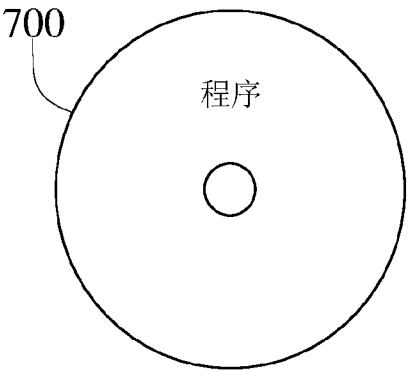


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/093043

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/332(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, EPODOC, WPI, CNKI, ISI: 对话, 机器人, 意图, 语料, 生成, 集合, 候选, 选择, 相似, 句子, 相似度, dialog, robot, intent, corpus, generate, set, candidate, select, similar, sentence, similarity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 110765759 A (PUXIN HENGYE TECHNOLOGY DEVELOPMENT (BEIJING) CO., LTD.) 07 February 2020 (2020-02-07) description, paragraphs [0069]-[0125] and figures 1-3	1-20
A	CN 109710939 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 03 May 2019 (2019-05-03) entire document	1-20
A	CN 109597873 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 09 April 2019 (2019-04-09) entire document	1-20
A	CN 110390006 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 29 October 2019 (2019-10-29) entire document	1-20
A	US 2007073534 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 29 March 2007 (2007-03-29) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

- “A” document defining the general state of the art which is not considered to be of particular relevance
- “E” earlier application or patent but published on or after the international filing date
- “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- “O” document referring to an oral disclosure, use, exhibition or other means
- “P” document published prior to the international filing date but later than the priority date claimed

- “T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
- “&” document member of the same patent family

Date of the actual completion of the international search

08 December 2020

Date of mailing of the international search report

21 December 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/093043

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110765759	A	07 February 2020	None			
CN	109710939	A	03 May 2019	US	2020210522	A1	02 July 2020
CN	109597873	A	09 April 2019	None			
CN	110390006	A	29 October 2019	None			
US	2007073534	A1	29 March 2007	CN	1940915	A	04 April 2007
				US	2008250015	A1	09 October 2008

国际检索报告

国际申请号

PCT/CN2020/093043

A. 主题的分类

G06F 16/332(2019.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT, EPDOC, WPI, CNKI, ISI: 对话, 机器人, 意图, 语料, 生成, 集合, 候选, 选择, 相似, 句子, 相似度, dialog, robot, intent, corpus, generate, set, candidate, select, similar, sentence, similarity

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 110765759 A (普信恒业科技发展北京有限公司) 2020年 2月 7日 (2020 - 02 - 07) 说明书第[0069]-[0125]段及图1-3	1-20
A	CN 109710939 A (北京百度网讯科技有限公司) 2019年 5月 3日 (2019 - 05 - 03) 全文	1-20
A	CN 109597873 A (腾讯科技深圳有限公司) 2019年 4月 9日 (2019 - 04 - 09) 全文	1-20
A	CN 110390006 A (腾讯科技深圳有限公司) 2019年 10月 29日 (2019 - 10 - 29) 全文	1-20
A	US 2007073534 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2007年 3月 29日 (2007 - 03 - 29) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 12月 8日

国际检索报告邮寄日期

2020年 12月 21日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

李婉怡

电话号码 86-(10)-53961408

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/093043

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	110765759	A	2020年 2月 7日	无			
CN	109710939	A	2019年 5月 3日	US	2020210522	A1	2020年 7月 2日
CN	109597873	A	2019年 4月 9日	无			
CN	110390006	A	2019年 10月 29日	无			
US	2007073534	A1	2007年 3月 29日	CN	1940915	A	2007年 4月 4日
				US	2008250015	A1	2008年 10月 9日