

会话应答方法和装置、电子设备、存储介质

相关申请的交叉引用

本申请要求于 2023 年 07 月 24 日提交的申请号为 202310913820.9、名称为“会话应答方法和装置、电子设备、存储介质”的中国专利申请的优先权，该中国专利申请的全部内容通过引用全部并入全文。

技术领域

本公开涉及信息处理技术领域，具体而言，涉及一种会话应答方法和装置、电子设备、存储介质。

10 背景技术

智能客服是一种基于自然语言处理的技术，近年来，在各大互联网平台得到了广泛应用。相关技术中主要是基于自然语言处理技术，即对用户输入的文本内容进行理解和解答。但随着用户需求的多元化，用户输入呈多模态化，而单一的文本理解无法完整理解用户语境和意图，造成应答准确率低，用户体验差的问题。

15 需要说明的是，在上述背景技术部分公开的信息仅用于加强对本公开的背景的理解，因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

根据本公开的第一方面，提供了一种会话应答方法，包括：获取按照会话时序排列的多模态会话信息，多模态会话信息包括第一模态信息和第二模态信息；对第一模态信息中的至少两个模态特征进行多模态的语义表征，得到第一表征向量；对第二模态信息进行对应模态的语义表征，得到第二表征向量；将第一表征向量和第二表征向量及对应的会话时序分别输入会话匹配模型，获得匹配结果；根据匹配结果，对多模态会话信息进行应答。

20 可选地，对第一模态信息中的至少两个模态特征进行多模态的语义表征，包括：提取第一模态信息中的文本内容和文本位置信息；对第一模态信息进行分块，获得块信息及块位置坐标；对文本内容、文本位置信息以及块信息、块位置坐标分别进行向量化表示；通过多模态语义表征模型，对向量化表示结果进行多模态的语义表征，获得第一表征向量。

25 可选地，对文本内容、文本位置信息以及块信息、块位置坐标分别进行向量化表示，包括：对文本内容、文本位置信息和块位置坐标分别进行嵌入映射；对块信息进行线性映射。

30 可选地，多模态语义表征模型包括第一编码模型、第二编码模型和特征融合模型，对向量化表示结果进行多模态的语义表征，包括：通过第一编码模型，对块信息的向量表示结果和块位置坐标的向量表示结果分别进行编码处理，获得第一编码向量；通过第二编码模型，对文本内容的向量表示和文本位置信息的向量表示分别进行编码处理，获得第二编码向量；通过特征融合模型，对第一编码向量和第二编码向量的拼接向量进行特征融合，

获得第一表征向量；其中，第一编码模型和第二编码模型为基于对比学习的预训练过程获得的。

可选地，方法还包括：获取目标业务的业务语料库，业务语料库包括至少两个第一模态语料；将一第一模态语料的图像特征和文本特征作为正例，将一第一模态语料的图像特征和另一第一模态语料的文本特征作为负例，构建得到第一训练样本；利用第一训练样本对第一编码模型和第二编码模型进行对比学习训练，得到预训练的多模态编码模型。

可选地，利用第一训练样本对第一编码模型和第二编码模型进行对比学习训练，包括：将第一训练样本中的正例和负例对应的图像特征输入第一编码模型，输出图像表征向量；将第一训练样本中的正例和负例对应的文本特征输入第二编码模型，输出文本表征向量；根据图像表征向量和文本表征向量，计算对比损失；根据对比损失，对第一编码模型和第二编码模型进行训练，得到预训练的多模态编码模型。

可选地，对第二模态信息进行对应模态的语义表征，包括：通过语义编码模型，对第二模态信息进行语义编码，得到第二表征向量，语义编码模型为对预训练的第三编码模型进行训练获得。

可选地，方法还包括：将业务语料库中的各多模态会话语料作为第二训练样本；在检索知识库中检索出第二训练样本的标签，检索知识库为根据业务语料库构建的，检索知识库包括领域维度、领域内的意图维度及意图下的问题答案对维度，标签包括各维度下的标签；利用第二训练样本及对应的标签，对预训练的第一编码模型、第二编码模型、第三编码模型及特征融合模型、会话匹配模型分别进行训练。

可选地，会话匹配模型包括上下文表征模型和多个预测头，一个预测头对应检索知识库中的一个维度，获得匹配结果，包括：通过上下文表征模型，对第一表征向量和第二表征向量及对应的会话时序进行基于多头注意力机制的特征编码，得到第三表征向量；将第三表征向量分别输入各预测头，获得对应各维度的匹配结果。

可选地，获得各维度的匹配结果，包括：利用各预测头，对第三表征向量进行全连接处理和非线性激活处理，获得各预测头对应的预测概率分数；确定各预测头中预测概率分数位于前 n 位的预测结果为匹配结果， n 为正整数。

可选地，方法还包括：在领域关键词库中，对文本内容和第二模态信息分别进行关键词匹配；在意图关键词库中，对文本内容和第二模态信息分别进行关键词匹配；其中，领域关键词库和意图关键词库分别为基于检索知识库对应维度的知识信息构建的。

根据匹配结果，对多模态会话信息进行应答，包括：根据各预测头的匹配结果和关键词匹配结果，对问题答案对维度对应预测头的匹配结果进行重排，得到重排分数；将重排分数大于预设阈值的问题答案对作为应答内容进行会话应答。

根据本公开的第二方面，提供一种会话应答装置，装置包括：获取模块、第一表征模块、第二表征模块、和会话匹配模块和应答模块，其中：获取模块，被配置为获取按照会话时序排列的多模态会话信息，多模态会话信息包括第一模态信息和第二模态信息；第一

表征模块，被配置为对第一模态信息中的至少两个模态特征进行多模态的语义表征，得到第一表征向量；第二表征模块，被配置为对第二模态信息进行对应模态的语义表征，得到第二表征向量；会话匹配模块，被配置为将第一表征向量和第二表征向量及对应的会话时序分别输入会话匹配模型，获得匹配结果；应答模块，被配置为根据匹配结果，对多模态
5 会话信息进行应答。

根据本公开的第三方面，提供一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现上述任一实施例的方法。

根据本公开的第四方面，提供一种电子设备，包括：一个或多个处理器；以及存储装置，用于一个或多个程序，当一个或多个程序被一个或多个处理器执行时，使得一个或多个
10 个处理器执行上述任一实施例的方法。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本公开。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本公开的实施例，
15 并与说明书一起用于解释本公开的原理。显而易见地，下面描述中的附图仅仅是本公开的一些实施例，对于本领域普通技术人员来讲，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。

图 1 示意性示出了根据本公开的一个实施例中会话应答方法的流程图。

图 2 示意性示出了根据本公开的一个实施例中第一模态信息的多模态表征过程流
20 程图。

图 3 示意性示出了根据本公开的一个实施例中对比学习训练过程流程图。

图 4 示意性示出了根据本公开的一个实施例中级联式检索知识库的示意图。

图 5 示意性示出了根据本公开的一个实施例中关键词库的关键词匹配过程的流程图。
25

图 6 示意性示出了根据本公开的一个实施例的会话应答过程的处理过程流程图。

图 7 示意性示出了根据本公开的一个实施例中会话应答装置的结构框图。

图 8 示出了适于用来实现本公开实施例的电子设备框图。

具体实施方式

现在将参考附图更全面地描述示例实施方式。然而，示例实施方式能够以多种形式
30 实施，且不应被理解为限于在此阐述的范例；相反，提供这些实施方式使得本公开将更加全面和完整，并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。在下面的描述中，提供许多具体细节从而给出对本公开的实施方式的充分理解。然而，本领域技术人员将意识到，可以实践本公开的技术方案而省略特定细节中的一个或更多，或者可以采用其

它的方法、组元、装置、步骤等。在其它情况下，不详细示出或描述公知技术方案以避免喧宾夺主而使得本公开的各方面变得模糊。

此外，附图仅为本公开的示意性图解，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。可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

本公开示例提供了一种会话应答方法和装置的应用场景示例性系统图，该系统包括终端和服务端。本实施例的方法可以应用于服务器，可以理解的是，该方法也可以应用于终端，还可以应用于包括终端和服务端的系统，并通过终端和服务端的交互实现。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，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN（Content Delivery Network，内容分发网络）、以及大数据和人工智能平台等基础云计算服务的云服务器，也可以是区块链中的节点。

例如，在客服服务过程中，用户在客户端输入多模态会话信息，服务器可以基于多模态会话信息，对其中的第一模态信息中的至少两个模态特征进行多模态的语义表征，得到第一表征向量；对其中的第二模态信息进行对应模态的语义表征，得到第二表征向量；将第一表征向量和第二表征向量及对应的会话时序分别输入会话匹配模型，获得匹配结果；根据匹配结果，确定应答内容，并将应答内容发送给用户终端。

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、车载设备等，但并不局限于此。当本实施例提供的会话应答方法通过终端和服务端的交互实现时，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，本公开在此不做限制。

本公开实施例所提供的会话应答方法可以在服务器执行，相应地，会话应答装置一般设置于服务器中。

下面，结合具体的实施例，对本说明书实施例披露的会话应答方法进行介绍。

参考图 1 所示，本公开提供的一种示例实施方式的会话应答方法，可以应用于各种智能客服系统，可以包括以下步骤：

步骤 S110，获取按照会话时序排列的多模态会话信息，多模态会话信息包括第一模态信息和第二模态信息。

步骤 S120，对第一模态信息中的至少两个模态特征进行多模态的语义表征，得到第一表征向量。

步骤 S130，对第二模态信息进行对应模态的语义表征，得到第二表征向量。

步骤 S140，将第一表征向量和第二表征向量及对应的会话时序分别输入会话匹配

模型，获得匹配结果。

步骤 S150，根据匹配结果，对多模态会话信息进行应答。

在本示例实施方式所提供的会话应答方法中，一方面，可以对多模态会话信息中的第一模态信息和第二模态信息分别进行语义表征，将表征结果及对应的会话时序输入会话匹配模型，获得匹配结果，基于匹配结果进行会话应答，能够融合多模态会话信息中的多个模态信息和会话时序进行应答，可以提高对用户上文语境的理解，进而提高应答的准确率。另一方面，通过对第一模态信息中的至少两个模态特征进行多模态的语义表征，可以挖掘第一模态信息中的多个模态特征，全面深入的理解第一模态信息中的语义，提高对第一模态信息的语义理解深度和全面性，从而提高应答的准确率。

在步骤 S110 中，获取按照会话时序排列的多模态会话信息。

在本示例实施方式中，多模态会话信息是指包括至少两种模态信息的会话信息，示例性地，会话信息可以是由用户与客服之间的交互对话形成的具有时序的信息，会话时序是指会话中的时间顺序。一种模态是指一种信息的来源或者形式；多模态会话信息可以包括图像信息、音频信息和文本信息等信息中的至少两种。例如，多模态会话信息包括第一模态信息和第二模态信息，第一模态信息为图像信息，第二模态信息为文本信息或音频信息。第一模态信息与第二模态信息是不同模态对应的信息。

在步骤 S120 中，对第一模态信息中的至少两个模态特征进行多模态的语义表征，得到第一表征向量。

在本示例实施方式中，模态特征是指文本特征、图像特征和音频特征中的至少一种。可以通过各种信息提取方式从第一模态信息中提取出不同的模态特征。可以通过多模态语义表征模型对提取的多模态特征进行语义表征，多模态语义表征模型可以包括多个并列的语义表征模块和模态融合模块，一个模态可以对应一个语义表征模块，语义表征模块用于进行对应模态的语义表征，模态融合模块用于对多个报表正结果进行融合。

示例性地，参考图 2 所示，多模态语义表征模型包括第一编码模型、第二编码模型和特征融合模型，可以通过以下步骤进行多模态的语义表征：

第一步，提取第一模态信息中的文本内容和文本位置信息。

在本示例实施方式中，可以利用光学字符识别模型提取第一模态信息中的文本内容和文本位置信息，即第一模态特征。本示例可以利用 OCR (Optical Character Recognition, 光学字符识别) 提取图像信息中的文本内容和文本框坐标。

第二步，对第一模态信息进行分块，获得块信息及块位置坐标。

在本示例实施方式中，分块是指对图像的分块 (patch)，块信息是指获得的图像块，块位置坐标是指每个图像块在图像中的位置坐标，本示例可以提取出图像块及其位置坐标即第一模态特征。

第三步，对文本内容、文本位置信息以及块信息、块位置坐标分别进行向量化表示。

在本示例实施方式中，可以对文本内容进行分词后转换为词向量，如对词进行

embedding，文本位置信息和块位置坐标也可以通过 embedding 进行嵌入映射，将文本位置向量与词向量叠加形成第一模态特征的向量表示。块信息（图像块）可以通过线性变换映射成低维向量，将块向量与块位置向量叠加形成第二模态特征的向量表示。块向量维度与词向量的维度相同。

5 第四步，通过第一编码模型，对块信息的向量表示结果和块位置坐标的向量表示结果分别进行编码处理，获得第一编码向量。

在本示例实施方式中，第一编码模型用于对第一模态信息中的图像块及其位置信息进行特征编码，第一编码模型可以是 transformer 架构，利用多头注意力机制对图像块进行编码。例如，第一编码模型的输入为对应 CLS token（分类特征/用于全局特征汇聚）
10 和图像 patch token（包括块位置坐标 embedding），通过加入块位置坐标来增强模型对于页面整体布局识别的能力，通过 CLS token 的输出作为第一模态信息的单模态表征。

第五步，通过第二编码模型，对文本内容的向量表示和文本位置信息的向量表示分别进行编码处理，获得第二编码向量。

在本示例实施方式中，第二编码模型用于对第一模态信息中的文本内容及其位置信息
15 进行特征编码，第二编码模型可以是 transformer 架构，利用多头注意力机制对 OCR 文本进行编码。例如，第二编码模型的输入为对应 CLS token 和 OCR 文本 token（包括文本位置信息 embedding），通过加入文本位置信息来增强模型对于文本结构和页面整体布局的识别能力，通过 CLS token 的输出作为第二模态信息的单模态表征。

第六步，通过特征融合模型，对第一编码向量和第二编码向量的拼接向量进行特征
20 融合，获得第一表征向量。

在本示例实施方式中，可以将第一编码向量和第二编码向量进行向量拼接后，输入特征融合模型，以进行特征维度的融合。特征融合模型可以是多层感知机，通过多层感知机对拼接向量进行线性变换和非线性处理，完成特征深度融合，输出图文融合的第一
25 表征向量。

以上实施例中的多模态语义表征模型是经过训练的，其中第一编码模型和第二编码模型是基于对比学习的预训练过程后再进行整体模型训练对参数微调获得的，特征融合模型是经整体的模型训练获得的。

示例性地，针对第一编码模型和第二编码模型的对比学习预训练过程可以包括以下步骤：

30 获取目标业务的业务语料库，业务语料库包括至少两个第一模态语料。

在本示例实施方式中，目标业务可以是客服业务或客服场景的其他业务（如售前业务或售后业务、金融业务等），业务语料库是针对目标业务的语料知识库，可以包括各种客服场景下的用户与客服间的交互语料。语料可以包括大量的文本、图像、音频等信息。第一模态语料可以是图像语料，一张图像可以作为一个第一模态语料。

35 将一第一模态语料的图像特征和文本特征作为正例，将一第一模态语料的图像特征

和另一第一模态语料的文本特征作为负例，构建得到第一训练样本。

在本示例实施方式中，可以将同一张图像的图像特征和文本特征作为正例，将一种图像的图像特征与另一种图像的文本特征作为负例，即负例的文本特征和图像特征来源于不同图像，构建第一训练样本。

- 5 利用第一训练样本对第一编码模型和第二编码模型进行对比学习训练，得到预训练的多模态编码模型。

在本示例实施方式中，采用对比学习训练第一编码模型和第二编码模型，学习正例之间的共同特征，区分负例之间的不同之处。经对比学习的两个编码模型对同类数据进行相似的编码，并使不同类的数据的编码结果尽可能的不同。对比学习不需要关注实例上繁琐的细节，只需要在抽象语义级别的特征空间上学会对数据的区分即可，模型及其
10 优化更加简单，泛化能力更强。

示例性地，参考图 3，可以通过以下步骤进行对比学习训练：

- 将第一训练样本中的正例和负例对应的图像特征输入第一编码模型，输出图像表征向量；将第一训练样本中的正例和负例对应的文本特征输入第二编码模型，输出文本表
15 征向量；根据图像表征向量和文本表征向量，计算对比损失；根据对比损失，对第一编码模型和第二编码模型进行训练，得到预训练的多模态编码模型。

在本示例实施方式中，第一编码模型用于进行图像表征，第二编码模型用于进行文本表征。可以采用 InfoNCE 损失进行训练。具体地对比损失可以表示为：

$$L_{\text{InfoNCE}} = L_{v2l} + L_{l2v}$$

其中， L_{v2l} 对应给定图像下的对比学习损失：

$$L_{v2l} = - \sum_i \log \frac{\exp(f_i^{vT} \cdot f_i^l / \tau)}{\exp(f_i^{vT} \cdot f_i^l / \tau) + \sum_{j \neq i} \exp(f_i^{vT} \cdot f_j^l / \tau)}$$

- 20 L_{l2v} 对应给定文本下的对比学习损失：

$$L_{l2v} = - \sum_i \log \frac{\exp(f_i^{vT} \cdot f_i^l / \tau)}{\exp(f_i^{vT} \cdot f_i^l / \tau) + \sum_{j \neq i} \exp(f_j^{vT} \cdot f_i^l / \tau)}$$

其中， f_i^v 为图像表征向量， f_i^l 为文本表征向量，* 可以是 i 或 j，i、j 分别表示图像编号， τ 为温度超参数（如设为 0.05），上标 T 表示转置操作。

- 本示例中，通过大量的正负样本进行无监督对比学习预训练，第一编码模型和第二编码模型可以学习到两个维度的信息：其一，图像的整体页面布局和 OCR 文本框的布局信息，该类信息有助于进行领域的判断；其二，OCR 文本的语义信息，借助语义可以
25 从另一个维度对领域进行判断，部分图像也可以根据语义进行意图判断。

在以上对第一编码模型和第二编码模型进行预训练的基础上，还可以对整个模型进行整体训练，整个模型可以包括多模态语义表征模型、预训练第三编码模型和会话匹配模型，整体训练过程可以包括以下步骤。

根据业务语料库，构建检索知识库。将业务语料库中的各多模态会话语料作为第二训练样本；在检索知识库中检索出第二训练样本的标签；利用第二训练样本及对应的标签，对预训练的第一编码模型和第二编码模型、预训练的第三编码模型、特征融合模型、会话匹配模型分别进行训练。

在本示例实施方式中，由于本方法可以理解为基于检索式的应答方式，为尽可能的覆盖用户的所有问题和需求，需要基于客服场景中的每个领域、用户意图和可能的 FAQ（Frequently Asked Questions，常见问题）和答案形成检索知识库，使其能够覆盖到全部的客服业务场景。如图 4 所示，检索知识库可以包括领域维度（领域 1，领域 2，…，领域 N）、领域内的意图维度（如领域 1 内的意图 1-2，…，意图 1-m1）及意图下的问题答案对维度（如意图 1-1 下的问题答案对 1-1-1，…，1-1-p1），N 为领域数量、m1 为领域 1 内的意图数量、p1 为意图 1-1 下的问题答案对数量，图 4 中的“……”表示此处省略具体示例。检索知识库的构建过程可以是：可以先根据用户入口、用户来源、产品、功能等维度将客服场景划分成多个领域，划分原则为确保领域和领域之间无重叠且尽可能的覆盖全部客服场景；其次，在每个领域内，对用户提问的意图进行枚举，例如可以根据用户问题的类型进行粗略划分，同样确保意图与意图之前无重叠且尽可能覆盖领域内全部的用户提问；最后，在给定的领域和意图下，总结出大量的 FAQ 极其对应的答案，尽可能的覆盖到用户提出的所有问题，最终形成一个领域-意图-FAQ 答案对的级联式检索知识库。例如，对于“怎么进行还款”这一 FAQ，在检索知识库中可以对应金融领域下的还款方式提问意图下的 FAQ。级联式检索知识库可以提高匹配准确率。

在本示例实施方式中，可以将一段会话语料作为一个训练样本，一个训练样本可以包括多个文本语句和图像信息。基于该多维度的检索知识库，样本标签也包括各维度下的标签，即一个维度对应一个标签，每个样本可以有领域标签、意图标签和问题答案对标签。可以将样本输入模型，以对预训练的模型（第一编码模型、第二编码模型、第三编码模型）进行参数微调，对未训练的模型（特征融合模型和会话匹配模型）进行训练。其中样本中的图像信息输入多模态语义表征模型，样本中的文本信息输入第三编码模型。通过每个维度对应预测头的输出与标签计算损失函数，通过最小化损失函数训练模型。此处的损失函数可以采用交叉熵损失函数。模型训练结束之后即可用于在线会话应答。

在步骤 S130 中，对第二模态信息进行对应模态的语义表征，得到第二表征向量。

在本示例实施方式中，可以通过语义编码模型，对第二模态信息进行语义编码，得到第二表征向量。语义编码模型是对预训练的第三编码模型进行训练获得的，即经上述模型整体训练之后获得的。第三编码模型可以是 Bert 模型，也可以是其他的文本语义表征模型，本示例对此不做限定。

在步骤 S140 中，将第一表征向量和第二表征向量及对应的会话时序分别输入会话匹配模型，获得匹配结果。

在本示例实施方式中，会话匹配模型可以包括上下文表征模型和多个预测头，一个预测头对应检索知识库中的一个维度，即一个维度有一个预测头，用于预测会话的该维度信息。上下文表征模型可以是上下文语义理解类模型，如 transformer 模型。预测头可以是各类分类器，此处可以是全连接网络。

5 示例性地，通过上下文表征模型，对第一表征向量和第二表征向量及对应的会话时序进行基于多头注意力机制的特征编码，得到第三表征向量；将第三表征向量分别输入各预测头，获得对应各维度的匹配结果。

在本示例实施方式中，可以采用 transformer 模型（上下文表征模型）对第一表征向量和第二表征向量及对应的会话时序（即位置编码）进行基于多头注意力机制的特征
10 编码，transformer 模型可以包括编码器和解码器，编码器采用自注意力机制和前馈神经网络的结构，解码器采用与编码器相同的结构，不同之处在于，解码器在两个结构之间还有一个注意力结构，用于使解码器关注输入特征的相关部分。将 transformer 模型输出的表征结果分别输入三个预测头（即领域预测头、意图预测头和问题答案对预测头）进行各维度的匹配预测，每个预测头输出对应匹配结果。

15 在一些实施例中，可以利用各预测头，对第三表征向量进行全连接处理和非线性激活处理，获得各预测头对应的预测概率分数。

在本示例实施方式中，每个预测头可以是全连接层，通过全连接层对表征结果进行全连接处理，再通过 softmax 激活函数映射为各类别概率，每个领域或意图或问题答案对可以对应一个类别，可以预测会话涉及的领域、意图和问题及相应答案。示例性地，
20 可以选择各预测头中预测概率分数位于前 n（n 为正整数）位的预测结果为对应的匹配结果，不同预测头对应的 n 可以相同、也可以不同，本示例对此不做限定。例如，可以确定概率分数最高的领域为领域匹配结果，概率分数位于前 5 位的意图为意图匹配结果，概率分数位于前 20 位的问题及相应答案为问题答案对匹配结果。

在步骤 S150 中，根据匹配结果，对多模态会话信息进行应答。
25 在本示例实施方式中，可以将匹配结果作为应答内容对多模态会话信息进行应答，也可以在匹配结果中选择部分内容进行应答，本示例对此不做限定。

在一些实施例中，还可以构建关键词库与匹配结果共同进行答案预测。可以在检索知识库的基础上，构建针对领域和意图的关键词库，即基于检索知识库领域维度和意图维度的知识信息构建对应关键词库。用于与用户提问文本或图像中文字进行字段匹配，
30 辅助进行领域和意图的预测。由于关键词的信息含量较为有限，因此只能进行领域或意图这种粗粒度的判断，可以与上述多模态模型的匹配结果进行融合，提高应答准确率。

在此基础上，参考图 5，可以在领域关键词库中，对文本内容和第二模态信息分别进行关键词匹配；在意图关键词库中，对文本内容和第二模态信息分别进行关键词匹配。若文本内容和第二模态信息在领域关键词库或意图关键词库中出现即视为命中，输出命
35 中的领域和意图。

在一些实施例中，在关键词匹配的基础上，可以对模型输出的匹配结果进行重排，以提高应答准确率，具体可以包括以下步骤。

根据各预测头的匹配结果和关键词匹配结果，对问题答案对维度对应预测头的匹配结果进行重排，得到重排分数；将重排分数大于预设阈值的问题答案对作为应答内容进行会话应答。

在本示例实施方式中，可以对每个匹配到的问题答案对按照各预测头的匹配结果和关键词匹配结果进行加权处理，得到加权后的重排分数，将重排分数大于预设阈值的问题答案对作为应答内容。预设阈值可以根据实际情况进行设置。

示例性地，针对模型输出的每个匹配到的 FAQ，重排分数的计算方式为：FAQ 重排分数 = 模型预测 FAQ 分数 + k_1 * 模型预测该 FAQ 对应领域的分数 + k_2 * 模型预测该 FAQ 对应意图的分数 + k_3 * (flag_keyword)，其中， k_1 , k_2 , k_3 分别表示三部分对应的权重系数，flag_keyword 为布尔变量，表示关键词是否成功匹配，若是，其值取 1，否则取 0。

举例而言，本公开的会话应答方法的实现过程如图 6 所示，可以基于会话应答模型实现，会话应答模型可以包括多模态表征模型、语义编码模型、上下文表征模型和多个预测头（领域预测头、意图预测头和问题答案对预测头），具体可以通过以下步骤实现。

第一步，获取业务语料库，构建检索知识库和关键词库（领域关键词库和意图关键词库）。

第二步，基于业务语料库和检索知识库构建训练样本，利用训练样本对会话应答模型进行训练。主要包括对多模态表征模型中的第一编码模型和第二编码模型的预训练（对比训练）和对整个会话应答模型的训练。

第三步，用户输入问题相关内容，可以包括文本（如文本 1、文本 2、文本 3）和图像（图像 1 和图像 2），会话应答系统获取按照会话时序排列的多模态会话信息。

第四步，通过多模态表征模型，对多模态会话信息中的第一模态信息（如图像 1 和图像 2）中的多个模态进行融合表征，得到第一表征向量（图像 1 表征和图像 2 表征）。

第五步，通过语义编码模型，对多模态会话信息中的第二模态信息（如文本 1、文本 2、文本 3）进行语义表征，得到第二表征向量（文本 1 表征、文本 2 表征、文本 3 表征）。

第六步，通过上下文表征模型，对第一表征向量和第二表征向量及对应的会话时序（会话时序编码）进行基于多头注意力机制的特征编码，得到第三表征向量。

第七步，将第三表征向量分别输入领域预测头、意图预测头和问题答案对预测头，获得预测领域、预测意图、预测问题答案对。

第八步，在领域关键词库中，对第一模态信息中的文本内容和第二模态信息分别进行关键词匹配；在意图关键词库中，对第一模态信息中的文本内容和第二模态信息分别进行关键词匹配。

第九步，根据各预测头的匹配结果和关键词匹配结果，对问题答案对维度对应预测头的匹配结果进行重排，得到重排分数。

第十步，将重排分数大于预设阈值的问题答案对作为应答内容进行会话应答。

以上实施例中的会话应答过程中，整个模型的训练主要是针对未进行预训练的模型，例如特征融合模型、上下文表征模型和多个预测头，其他预训练模型只是进行参数微调。

现有智能客服技术主要是基于自然语言处理技术，对文本类的用户输入（如输入文本或输入语音转换成文本）可以做到较好的理解并输出相应的解答。但随着智能客服技术的发展和用户需求的多元化，目前大多数智能客服场景，用户可以更为方便的输入图片，辅以文字以更为准确的描述问题和需求。也就是说，用户输入模态从单一的文本模态变成了文本/音频和图像的多模态信息，这给基于自然语言处理的智能客服技术带来了巨大的挑战。相关技术在遇到图像信息时往往无法正确地理解和判断，而若忽略会话中的图片会造成无法正确且完整的理解用户上文的语境和意图，容易得出不符合当前语境和用户需求的解答，造成应答准确率和用户体验的降低。

本公开针对用户的多模态输入（文本+图片）所面临的问题，设计了基于图文多模态模型的检索式智能客服应答方法，通过多模态表征模型融合图像信息中的多个模态信息，同时融合了图像和文本的位置坐标信息，强化模型对文本结构和页面整体布局识别的能力。结合语义编码模型和上下文表征模型，能够同时处理用户输入的文本和图像模态的信息，通过多个预测头，准确理解语境和用户意图，即用户当前问题所属的意图/领域，并在此意图/领域范围内，筛选出符合当前语境和用户需求的解答，能够显著提升应答准确率和用户体验，更好的应对多元化智能客服场景。

本公开通过构建级联式的检索知识库，可以对会话语义进行由粗到细逐层理解匹配，能够模型预测的准确性。本公开在模型预测的基础上，结合关键词库匹配，增加了对问题应答对粒度的应答结果的准确性。

进一步的，本示例实施方式中，还提供了一种会话应答装置 700。该会话应答装置 700 可以应用于服务器。参考图 7 所示，该会话应答装置 700 可以包括：获取模块 710、第一表征模块 720、第二表征模块 730、会话匹配模块 740 和应答模块 750，其中：获取模块 710，被配置为获取按照会话时序排列的多模态会话信息，多模态会话信息包括第一模态信息和第二模态信息；第一表征模块 720，被配置为对第一模态信息中的至少两个模态特征进行多模态的语义表征，得到第一表征向量；第二表征模块 730，被配置为对第二模态信息进行对应模态的语义表征，得到第二表征向量；会话匹配模块 740，被配置为将第一表征向量和第二表征向量及对应的会话时序分别输入会话匹配模型，获得匹配结果；应答模块 750，被配置为根据匹配结果，对多模态会话信息进行应答。

在本公开的一种示例性实施例中，第一表征模块 720 包括：提取子模块，被配置为提取第一模态信息中的文本内容和文本位置信息；分块子模块，被配置为对第一模态信息进行分块，获得块信息及块位置坐标；向量化子模块，被配置为对文本内容、文本位

置信息以及块信息、块位置坐标分别进行向量化表示；语义表征子模块，被配置为通过多模态语义表征模型，对向量化表示结果进行多模态的语义表征，获得第一表征向量。

在本公开的一种示例性实施例中，向量化子模块还被配置为：对文本内容、文本位置信息和块位置坐标分别进行嵌入映射；对块信息进行线性映射。

5 在本公开的一种示例性实施例中，多模态语义表征模型包括第一编码模型、第二编码模型和特征融合模型，语义表征子模块还被配置为：通过第一编码模型，对块信息的向量表示结果和块位置坐标的向量表示结果分别进行编码处理，获得第一编码向量；通过第二编码模型，对文本内容的向量表示和文本位置信息的向量表示分别进行编码处理，获得第二编码向量；通过特征融合模型，对第一编码向量和第二编码向量的拼接向量进行特征融合，获得第一表征向量；其中，第一编码模型和第二编码模型为基于对比学习的预训练过程获得的。

10 在本公开的一种示例性实施例中，装置 700 还包括第一训练模块，第一训练模块被配置为：获取目标业务的业务语料库，业务语料库包括至少两个第一模态语料；将一第一模态语料的图像特征和文本特征作为正例，将一第一模态语料的图像特征和另一第一模态语料的文本特征作为负例，构建得到第一训练样本；利用第一训练样本对第一编码模型和第二编码模型进行对比学习训练，得到预训练的多模态编码模型。

15 在本公开的一种示例性实施例中，第一训练模块还被配置为：将第一训练样本中的正例和负例对应的图像特征输入第一编码模型，输出图像表征向量；将第一训练样本中的正例和负例对应的文本特征输入第二编码模型，输出文本表征向量；根据图像表征向量和文本表征向量，计算对比损失；根据对比损失，对第一编码模型和第二编码模型进行训练，得到预训练的多模态编码模型。

20 在本公开的一种示例性实施例中，第二表征模块 730 还被配置为：通过语义编码模型，对第二模态信息进行语义编码，得到第二表征向量，语义编码模型为对预训练的第三编码模型进行训练获得。

25 在本公开的一种示例性实施例中，装置 700 还包括第二训练模块，第二训练模块还被配置为：将业务语料库中的各多模态会话语料作为第二训练样本；在检索知识库中检索出第二训练样本的标签，检索知识库为根据业务语料库构建的，检索知识库包括领域维度、领域内的意图维度及意图下的问题答案对维度，标签包括各维度下的标签；利用第二训练样本及对应的标签，对预训练的第一编码模型、第二编码模型、第三编码模型及特征融合模型、会话匹配模型分别进行训练。

30 在本公开的一种示例性实施例中，会话匹配模型包括上下文表征模型和多个预测头，一个预测头对应检索知识库中的一个维度，会话匹配模块 740 还被配置为：通过上下文表征模型，对第一表征向量和第二表征向量及对应的会话时序进行基于多头注意力机制的特征编码，得到第三表征向量；将第三表征向量分别输入各预测头，获得对应各维度的匹配结果。

在本公开的一种示例性实施例中，会话匹配模块 740 还被配置为：利用各预测头，对第三表征向量进行全连接处理和非线性激活处理，获得各预测头对应的预测概率分数；确定各预测头中预测概率分数位于前 n 位的预测结果为匹配结果，n 为正整数。

在本公开的一种示例性实施例中，装置 700 还包括关键词匹配模块，关键词匹配模块被配置为：在领域关键词库中，对文本内容和第二模态信息分别进行关键词匹配；在意图关键词库中，对文本内容和第二模态信息分别进行关键词匹配；其中，领域关键词库和意图关键词库分别为基于检索知识库对应维度的知识信息构建的。应答模块 750 还被配置为：根据各预测头的匹配结果和关键词匹配结果，对问题答案对维度对应预测头的匹配结果进行重排，得到重排分数；将重排分数大于预设阈值的问题答案对作为应答内容进行会话应答。

上述会话应答装置中各模块或单元的具体细节已经在对应的会话应答方法中进行了详细的描述，因此此处不再赘述。

作为另一方面，本公开还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备中所包含的；也可以是单独存在，而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被一个该设备执行时，使得该设备实现如下实施例中的方法。例如，设备可以实现如图 1~图 3 及图 5~图 6 所示的各个步骤等。

需要说明的是，本公开所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——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。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中，计算机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。而在本公开中，计算机可读的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，其中承载了计算机可读的程序代码。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。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、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。计算机可读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、电线、光缆、RF 等等，或者上述的任意合适的组合。

下面参照图 8 来描述根据本公开的这种实施例的设备 800。图 8 显示的设备 800 仅仅是一个示例，不对本公开实施例的功能和使用范围带来任何限制。

如图 8 所示，设备 800 以通用计算设备的形式表现。设备 800 的组件可以

包括但不限于：上述至少一个处理单元 810、上述至少一个存储单元 820、连接不同系统组件（包括存储单元 820 和处理单元 810）的总线 830、显示单元 840。

其中，存储单元存储有程序代码，程序代码可以被处理单元 810 执行，使得处理单元 810 执行本说明书上述“示例性方法”部分中描述的根据本公开各种示例性实施例的步骤。

存储单元 820 可以包括易失性存储单元形式的可读介质，例如随机存取存储单元（RAM）8201 和/或高速缓存存储单元 8202，还可以进一步包括只读存储单元（ROM）8203。

存储单元 820 还可以包括具有一组（至少一个）程序模块 8205 的程序/实用工具 8204，这样的程序模块 8205 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 830 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 800 也可以与一个或多个外部设备 870（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 800 交互的设备通信，和/或与使得该电子设备 800 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 850 进行。并且，电子设备 800 还可以通过网络适配器 860 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 860 通过总线 830 与电子设备 800 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 800 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、RA 标识系统、磁带驱动器以及数据备份存储系统等。

通过以上的实施例的描述，本领域的技术人员易于理解，这里描述的示例实施例可以通过软件实现，也可以通过软件结合必要的硬件的方式来实现。因此，根据本公开实施例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）执行根据本公开实施例的方法。

此外，上述附图仅是根据本公开示例性实施例的方法所包括的处理的示意性说明，而不是限制目的。易于理解，上述附图所示的处理并不表明或限制这些处理的时间顺序。另外，也易于理解，这些处理可以是例如在多个模块中同步或异步执行的。

需要说明的是，尽管在附图中以特定顺序描述了本公开中方法的各个步骤，但是，这并非要求或者暗示必须按照该特定顺序来执行这些步骤，或是必须执行全部所示的步骤才能实现期望的结果。附加的或备选的，可以省略某些步骤，将多个步骤合并为一个

步骤执行，以及/或者将一个步骤分解为多个步骤执行等，均应视为本公开的一部分。

应可理解的是，本说明书公开和限定的本公开延伸到文中和/或附图中提到或明显的两个或两个以上单独特征的所有可替代组合。所有这些不同的组合构成本公开的多个可替代方面。本说明书的实施方式说明了已知用于实现本公开的最佳方式，并且将使本

5 领域技术人员能够利用本公开。

权利要求

1. 一种会话应答方法，包括：
- 获取按照会话时序排列的多模态会话信息，所述多模态会话信息包括第一模态信息和第二模态信息；
- 5 对所述第一模态信息中的至少两个模态特征进行多模态的语义表征，得到第一表征向量；
- 对所述第二模态信息进行对应模态的语义表征，得到第二表征向量；
- 将所述第一表征向量和所述第二表征向量及对应的会话时序分别输入会话匹配模型，获得匹配结果；
- 10 根据所述匹配结果，对所述多模态会话信息进行应答。
2. 根据权利要求 1 所述的会话应答方法，其中，所述对第一模态信息中的至少两个模态特征进行多模态的语义表征，包括：
- 提取所述第一模态信息中的文本内容和文本位置信息；
- 对所述第一模态信息进行分块，获得块信息及块位置坐标；
- 15 对所述文本内容、所述文本位置信息以及所述块信息、所述块位置坐标分别进行向量化表示；
- 通过多模态语义表征模型，对向量化表示结果进行多模态的语义表征，获得所述第一表征向量。
3. 根据权利要求 2 所述的会话应答方法，其中，所述对所述文本内容、所述文本位置信息以及所述块信息、所述块位置坐标分别进行向量化表示，包括：
- 20 对所述文本内容、所述文本位置信息和所述块位置坐标分别进行嵌入映射；
- 对所述块信息进行线性映射。
4. 根据权利要求 2 或 3 所述的会话应答方法，其中，所述多模态语义表征模型包括第一编码模型、第二编码模型和特征融合模型，所述对向量化表示结果进行多模态的语义表征，包括：
- 25 通过第一编码模型，对所述块信息的向量表示结果和所述块位置坐标的向量表示结果分别进行编码处理，获得第一编码向量；
- 通过第二编码模型，对所述文本内容的向量表示和所述文本位置信息的向量表示分别进行编码处理，获得第二编码向量；
- 30 通过特征融合模型，对所述第一编码向量和所述第二编码向量的拼接向量进行特征融合，获得所述第一表征向量；
- 其中，所述第一编码模型和所述第二编码模型为基于对比学习的预训练过程获得的。
5. 根据权利要求 4 所述的会话应答方法，其中，所述方法还包括：
- 获取目标业务的业务语料库，所述业务语料库包括至少两个第一模态语料；
- 35 将一所述第一模态语料的图像特征和文本特征作为正例，将一所述第一模态语料的图

像特征和另一所述第一模态语料的文本特征作为负例，构建得到第一训练样本；

利用所述第一训练样本对所述第一编码模型和所述第二编码模型进行对比学习训练，得到预训练的多模态编码模型。

6. 根据权利要求 5 所述的会话应答方法，其中，所述利用所述第一训练样本对所述第一编码模型和所述第二编码模型进行对比学习训练，包括：

将所述第一训练样本中的正例和负例对应的图像特征输入第一编码模型，输出图像表征向量；将所述第一训练样本中的正例和负例对应的文本特征输入所述第二编码模型，输出文本表征向量；

根据所述图像表征向量和所述文本表征向量，计算对比损失；

10 根据所述对比损失，对所述第一编码模型和所述第二编码模型进行训练，得到预训练的多模态编码模型。

7. 根据权利要求 1 所述的会话应答方法，其中，所述对所述第二模态信息进行对应模态的语义表征，包括：

15 通过语义编码模型，对所述第二模态信息进行语义编码，得到所述第二表征向量，所述语义编码模型为对预训练的第三编码模型进行训练获得。

8. 根据权利要求 5 所述的会话应答方法，其中，所述方法还包括：

将所述业务语料库中的各多模态会话语料作为第二训练样本；

20 在检索知识库中检索出所述第二训练样本的标签，所述检索知识库为根据所述业务语料库构建的，所述检索知识库包括领域维度、领域内的意图维度及意图下的问题答案对维度，所述标签包括各维度下的标签；

利用所述第二训练样本及对应的标签，对所述预训练的第一编码模型、第二编码模型、第三编码模型及所述特征融合模型、所述会话匹配模型分别进行训练。

25 9. 根据权利要求 8 所述的会话应答方法，其中，所述会话匹配模型包括上下文表征模型和多个预测头，一个预测头对应所述检索知识库中的一个维度，所述获得匹配结果，包括：

通过所述上下文表征模型，对所述第一表征向量和所述第二表征向量及对应的会话时序进行基于多头注意力机制的特征编码，得到第三表征向量；

将所述第三表征向量分别输入各所述预测头，获得对应各维度的匹配结果。

30 10. 根据权利要求 9 所述的会话应答方法，其中，所述获得各维度的匹配结果，包括：

利用各所述预测头，对所述第三表征向量进行全连接处理和非线性激活处理，获得各预测头对应的预测概率分数；

确定各预测头中所述预测概率分数位于前 n 位的预测结果为所述匹配结果， n 为正整数。

11. 根据权利要求 9 或 10 所述的会话应答方法，其中，所述方法还包括：

35 在领域关键词库中，对所述文本内容和所述第二模态信息分别进行关键词匹配；

在意图关键词库中，对所述文本内容和所述第二模态信息分别进行关键词匹配；

其中，所述领域关键词库和所述意图关键词库分别为基于所述检索知识库对应维度的知识信息构建的；

所述根据匹配结果，对所述多模态会话信息进行应答，包括：

5 根据各预测头的匹配结果和所述关键词匹配结果，对问题答案对维度对应预测头的匹配结果进行重排，得到重排分数；

将所述重排分数大于预设阈值的问题答案对作为所述应答内容进行会话应答。

12. 一种会话应答装置，所述装置包括：

10 获取模块，被配置为获取按照会话时序排列的多模态会话信息，所述多模态会话信息包括第一模态信息和第二模态信息；

第一表征模块，被配置为对所述第一模态信息中的至少两个模态特征进行多模态的语义表征，得到第一表征向量；

第二表征模块，被配置为对所述第二模态信息进行对应模态的语义表征，得到第二表征向量；

15 会话匹配模块，被配置为将所述第一表征向量和所述第二表征向量及对应的会话时序分别输入会话匹配模型，获得匹配结果；

应答模块，被配置为根据所述匹配结果，对所述多模态会话信息进行应答。

13. 一种电子设备，包括：一个或多个处理器；以及

20 存储装置，用于存储一个或多个程序，当所述一个或多个程序被所述一个或多个处理器执行时，使得所述一个或多个处理器实现如权利要求 1 至 11 中任一项所述的方法。

14. 一种计算机可读存储介质，其上存储有计算机程序，所述程序被处理器执行时实现如权利要求 1 至 11 中任一项所述的方法。