

问题应答方法和装置、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 08 月 03 日提交的申请号为 202310975117.0、名称为“问题应
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
5 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，涉及一种问题应答方法和装置、电子设
备、存储介质。

10 背景技术

文档问答技术（Document Question Answering, DocQA）是一项非常重要的技术，可
以从非结构化的文档信息中抽取答案。目前文档问答技术大都是基于深度神经网络对文档
内容进行语义匹配来确定应答内容，忽略了文档内容中的其他模态信息，导致应答准确性
低。

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，提供了一种问题应答方法，包括：获取多模态信息中各第一
模态信息对应的第一模态特征和第二模态特征；对于一第一模态信息，对第一模态特征和
20 用户问题特征，及第二模态特征和用户问题特征，分别进行语义相似度匹配，对应获得第
一相似向量表征和第二相似向量表征；对第一相似向量表征和第二相似向量表征进行信息
融合，获得多模态相似分数；响应于一多模态相似分数大于等于预设阈值，将对应的第一
模态信息作为应答内容进行问题应答。

可选地，对第一模态特征和用户问题特征进行语义相似度匹配，包括：通过多层注意
25 力层，对第一模态特征和用户问题特征进行语义相似度匹配；通过前馈网络层，对多层注
意力层的输出进行非线性激活和线性变换，获得第一相似向量表征。

可选地，通过多层注意力层，对第一模态特征和用户问题特征进行语义相似度匹配，
包括：对于每个注意力层，对用户问题特征进行第一线性处理，获得第一中间向量；对第
一模态特征分别进行第二线性处理和第三线性处理，获得第二中间向量和第三中间向量；
30 确定第一中间向量与第二中间向量之间的相似度；根据相似度，对第三中间向量进行加权
匹配，获得当前注意力层的输出；对各注意力层的输出进行向量拼接，并通过第一全连接
层对拼接后向量进行全连接处理，获得多层注意力层的输出。

可选地，对第二模态特征和用户问题特征进行语义相似度匹配，包括：将第二模态特
征和用户问题特征进行拼接，获得拼接向量；通过第二全连接层，将拼接向量映射到低维

空间，获得第二相似向量表征。

可选地，对第一相似向量表征和第二相似向量表征进行信息融合，包括：通过门控网络，对第一相似向量表征和第二相似向量表征进行信息融合；对融合后向量进行全连接处理和非线性激活处理，获得多模态相似分数。

5 可选地，第一模态信息包括图像信息，第二模态特征包括文本特征，获取各第一模态信息对应的第二模态特征，包括：提取一图像信息的标题信息；通过预训练的目标检测模型，识别该图像信息中的内容及其属性信息；通过光学字符识别，提取该图像信息中的候选文本信息；对标题信息、内容及其属性信息及候选文本信息进行向量化表示，获得该图像信息对应的文本特征。

10 可选地，获取各第一模态信息对应的第一模态特征，包括：通过图像预训练模型，提取各第一模态信息中的图像向量信息，获得对应的第一模态特征。

可选地，方法还包括：响应于接收到用户问题语句，通过预训练编码模型对用户问题语句进行编码，获得用户问题特征。

15 可选地，多模态信息包括第二模态信息，方法还包括：响应于各多模态相似分数分别小于预设阈值，基于第二模态信息确定应答内容。

可选地，基于第二模态信息确定应答内容，包括：对第二模态信息进行段落拆分，并确定拆分获得的各段落的关键词；基于关键词，将各段落中语义相似的段落进行聚合；通过预训练的摘要生成模型，对聚合后段落进行语句凝练，获得摘要文本；并将摘要文本作为应答内容。

20 可选地，方法还包括：响应于多个多模态相似分数大于等于预设阈值，确定最高的多模态相似分数对应的第一模态信息为应答内容。

根据本公开的第二方面，提供一种问题应答装置，装置包括：获取模块、语义匹配模块、融合模块和应答模块，其中：获取模块，被配置为获取多模态信息中各第一模态信息对应的第一模态特征和第二模态特征；语义匹配模块，被配置为对于一第一模态信息，对
25 第一模态特征和用户问题特征，及第二模态特征和用户问题特征，分别进行语义相似度匹配，对应获得第一相似向量表征和第二相似向量表征；融合模块，被配置为对第一相似向量表征和第二相似向量表征进行信息融合，获得多模态相似分数；应答模块，被配置为响应于一多模态相似分数大于等于预设阈值，将对应的第一模态信息作为应答内容进行问题应答。

30 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，使得一个或多个处理器执行上述任一实施例的方法。

35 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，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问题应答方法的流程图。

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中的第二模态特征提取流程实体图。

图 3 示意性示出了根据本公开的一个实施例中第一模态信息对应图像。

图 4 示意性示出了根据本公开的一个实施例中第一模态特征和用户问题特征的语义相似度匹配过程的示意图。

图 5 示意性示出了根据本公开的一个实施例中多模态相似分数的确定过程的流程图。

图 6 示意性示出了根据本公开的一个实施例的问题应答方法的处理过程流程图。

图 7 示意性示出了根据本公开的一个实施例中问题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问题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器。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器的系统，并通过终端和服务器的交互实现。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，用户在用户客户端发送用户问题语句，服务器可以基于用户问题语句关联出多模态信息，并从知识库中获取该多模态信息中各第一模态信息对应的第一模态特征和第二模态特征；对于一第一模态信息，对第一模态特征和用户问题特征，及第二模态特征和用户问题特征，分别进行语义相似度匹配，对应获得第一相似向量表征和第二相似向量表征；对第一相似向量表征和第二相似向量表征进行信息融合，获得多模态相似分数；响应于一多模态相似分数大于等于预设阈值，确定对应的第一模态信息为应答内容，并将该应答内容通过客户客户端发送给用户终端。

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。当本实施例提供的问题应答方法通过终端和服务器的交互实现时，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问题应答方法可以在服务器执行，相应地，问题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问题应答方法进行介绍。

参考图 1 所示，本公开提供的一种示例实施方式的问题应答方法，可以应用于各种问答系统，可以包括以下步骤。

步骤 S110，获取多模态信息中各第一模态信息对应的第一模态特征和第二模态特征。

步骤 S120，对于一第一模态信息，对第一模态特征和用户问题特征，及第二模态特征和用户问题特征，分别进行语义相似度匹配，对应获得第一相似向量表征和第二相似向量表征。

步骤 S130，对第一相似向量表征和第二相似向量表征进行信息融合，获得多模态相似分数。

步骤 S140，响应于一多模态相似分数大于等于预设阈值，确定对应的第一模态信息为应答内容。

在本示例实施方式所提供的问题应答方法中，一方面，可以通过对第一模态特征和用户问题特征，及第二模态特征和用户问题特征，分别进行语义相似度匹配，并融合该两种语义相似匹配结果，挖掘出一种模态信息中的多个模态特征，并融合多个模态特征确定应答内容，提高了应答的准确性。另一方面，在多模态相似分数大于等于预设阈值的情况下，可以将原始的第一模态信息作为应答内容，避免了对文档内容的深度语义提取过程，可以实现应答系统的快速响应，提高应答效率。此外，可以对各第一模态信息

的第一模态特征和第二模态特征进行预先提取，进一步提高应答响应效率。

在步骤 S110 中，获取多模态信息中各第一模态信息对应的第一模态特征和第二模态特征。

在本示例实施方式中，多模态信息是指包括至少两种模态信息的文档信息。一种模态是指一种信息的来源或者形式；多模态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。第一模态信息是指多模态信息中包含至少两个模态特征的一种模态信息，例如图像信息或音频信息。第一模态特征与第二模态特征是不同的模态对应的两种特征。

示例性地，对于网页搜索场景，每个网页对应的所有信息可以认为是一种多模态信息，网页中包括图像信息和文本信息就是不同的模态信息，可以将一个图像信息作为一个第一模态信息，将一个图像信息中的图像特征作为第一模态特征，将图像信息中的文本特征作为第二模态特征。

在本示例实施方式中，可以对多模态信息进行预先处理，提取第一模态特征和第二模态特征，并将第一模态特征和第二模态特征保存于知识库中，以使在线应答时可以直接调用，减少响应时间。可以通过各种预训练模型提取第一模态特征和第二模态特征。

示例性地，可以通过图像预训练模型，提取各第一模态信息中的图像向量信息，获得各第一模态信息对应的第一模态特征。

在本示例实施方式中，图像预训练模型用于识别各第一模态信息（图像信息），图像预训练模型可以是各种预训练的卷积神经网络，如 CNN（Convolutional Neural Network）、VGG16（Visual Geometry Group Network 16）、VGG19（Visual Geometry Group Network 19）、ResNet-50（Residual Neural Network 50）等，本示例对此不做限定。示例性地，通过 VGG16 的多个卷积层和池化层进行特征提取，再通过多个全连接层和 softmax 激活函数得到输出结果，即第一模态特征。

示例性地，第一模态信息包括图像信息，第二模态特征包括文本特征，如图 2 所示，可以通过以下步骤获取各第一模态信息对应的第二模态特征：

第一步，提取一图像信息的标题信息。

在本示例实施方式中，可以在图像信息具有标题的情况下，提取图像信息的标签即可；在图像信息没有标题的情况下，可以使用图像标题生成模型，将图像级别信息转换为标题信息，图像标题生成模型可以是现有的图像标题生成模型或工具，例如，视觉处理深度学习库 LAVIS，这是一个开源的库，用于在丰富的常见任务和数据集系列上训练和评估最先进的语言-视觉模型，并用于在定制的语言-视觉数据上进行现成的推理，例如，输入一张图像，输出对应的文字描述。

第二步，通过预训练的目标检测模型，识别该图像信息中的内容及其属性信息。

在本示例实施方式中，目标检测模型用于找出图像中所有感兴趣的目标对象，确定对象的类别和位置。目标检测模型可以是 R-CNN（Region Convolutional Neural Network）、

SPP-Net (Spatial Pyramid Pooling Net)、Fast R-CNN、Faster R-CNN 和 R-FCN (Region-based fully convolutional network) 等中的一个, 本示例对此不做限定。图像信息中的内容可以包括各种对象(如环境、人、物品等), 内容属性信息可以包括对象名称、性别等基本属性信息, 还可以包括对象职业、对象颜色、对象高度等属性信息, 本示例对此不做限定。

第三步, 通过光学字符识别, 提取该图像信息中的候选文本信息。

在本示例实施方式中, 可以通过光学字符识别(Optical Character Recognition, OCR)技术, 提取图像信息中的文本, 即提取图像中本身含有的文本信息(即候选文本信息)。

第四步, 对标题信息、内容及其属性信息及候选文本信息进行向量化表示, 获得该图像信息对应的文本特征。

在本示例实施方式中, 可以将上述步骤提取的标题信息、内容及属性信息和候选文本信息作为图像信息中提取的文本信息(即第一模态信息中提取的第二模态信息), 将该文本特征进行向量化表示, 获得对应的第二模态特征。可以采用预训练的文本向量化模型进行向量化表示, 例如可以采用预训练的 Bert 模型。

举例而言, 如图 3 所示, 对于该张图像提取的标题信息为“便民停车场”, 提取的内容及其属性信息包括“白色汽车”、“蓝色指示牌”、“树林”、“楼房”等, 提取的候选文本信息为“停车场出口; 温馨提示、监控覆盖区域 请有序停放车辆……”。将提取的以上信息进行向量化表示(使用预训练的 Bert 模型)即可得到该图像的第二模态特征(即该图像的文本特征)。

示例性地, 可以将上述实施例提取的第一模态特征和第二模态特征分别存储于两个知识库中, 上述第一模态特征和第二模态特征的内容解析过程可以通过预先对目标文档内的图像信息进行处理而获得, 并存储于知识库中, 便于提高后续应答响应速度。

本公开针对多模态的文档问答系统, 在一些实施例中, 方法还包括: 响应于接收到用户问题语句, 通过预训练编码模型对用户问题语句进行编码, 获得用户问题特征。

在本示例实施方式中, 用户问题语句可以是用户在问答系统中发送的问题语句, 例如客服系统接收到的用户问句, 可以对接收到的用户问题语句进行向量化表示, 预训练编码模型可以是与第二模态特征相同的向量化表示模型, 如预训练的 Bert 模型。

在步骤 S120 中, 对于一第一模态信息, 对第一模态特征和用户问题特征, 及第二模态特征和用户问题特征, 分别进行语义相似度匹配, 对应获得第一相似向量表征和第二相似向量表征。

在本示例实施方式中, 第一模态信息可以是一张图像, 可以对目标文档内的每张图像分别与用户问题特征进行语义相似度匹配。第一模态特征可以对应图像特征向量, 第二特征向量可以对应图像中提取的文本特征向量, 将两种模态特征分别与用户问题特征进行语义相似度匹配。例如, 可以基于待匹配的两个特征向量建模两者匹配关系的特征向量, 再用其他模型(比如多层感知机 MLP)来学习通用的文本关系函数映射。

示例性地，可以通过多模态的 transformer 模型对第一模态特征向量和用户问题特征进行语义相似度匹配，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，即解码器中的自注意力机制输出与编码器的输出再计算注意力得分，之后，再进入解码器的前馈神经网络。编码器可以包括多个上述结构的编码模块，解码器也可以包括多个相同的上述解码模块。具体可以包括以下步骤。

通过多层注意力层，对第一模态特征和用户问题特征进行语义相似度匹配。通过前馈网络层，对多层注意力层的输出进行非线性激活和线性变换，获得第一相似向量表征。

在本示例实施方式中，多层注意力层用于进行基于自注意力机制的匹配，可以使用用户问题特征映射到 Q，使用第一模态特征（图像向量）作为 K 和 V 向量进行基于自注意力机制的处理。前馈网络层用于对多层注意力层的输出进行非线性激活和线性变换，如进行多个线性变换和 ReLU 激活函数处理。其中编码器和解码器的输入还包括向量的位置信息。前馈神经网络层的输出可以是基于自注意力机制的上下文向量，对该向量进行全连接处理就可以得到第一相似向量表征（两个输入向量的相似度表征向量）。

示例性地，如图 4 所示，可以通过多层注意力层，对第一模态特征和用户问题特征进行语义相似度匹配。

对于每个注意力层，采用第一线性层对用户问题特征进行第一线性处理，获得第一中间向量；采用第二线性层和第三线性层分别对第一模态特征进行第二线性处理和第三线性处理，获得第二中间向量和第三中间向量；确定第一中间向量与第二中间向量之间的相似度；根据相似度，对第三中间向量进行加权匹配，获得当前注意力层的输出；对各注意力层的输出进行向量拼接，并通过第一全连接层对拼接后向量进行全连接处理，获得多层注意力层的输出。

在本示例实施方式中，第一中间向量可以对应 Q（Query），第二中间向量可以对应 K（Key），第三中间向量可以对应 V（Value），可以通过计算 Q 和 K 之间的点积，确定两者之间的相似度；还可以采用规模化注意力层对相似度进行规模化处理。然后利用 Softmax 操作将其结果归一化为概率分布，再乘以 V 为加权权重进行加权求和，得到当前注意力层的输出。每个注意力层进行同样的处理，不同层的模型参数可以不同，获得多个自注意力机制的向量表示。将多个注意力层的输出进行向量拼接后再利用全连接层（第一全连接层）进行全连接处理，获得多层注意力层的输出。

例如，一个注意力层的处理过程可以用以下公式表示：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

$$Q = W_q Q_i + b_q$$

$$K = W_k I_n + b_k$$

$$V = W_v I_n + b_v$$

式中, $Attention(Q, K, V)$ 表示该注意力层的输出, $softmax$ 表示 $softmax$ 函数, Q 、 K 、 V 分别表示第一中间向量、第二中间向量和第三中间向量; d 表示 K 向量的维度, 此处采用 $\sqrt{d}$ 进行规模化, 用于防止结果过大; W_q 、 b_q 表示第一线性处理的模型参数, W_k 、 b_k 表示第二线性处理的模型参数, W_v 、 b_v 表示第三线性处理的模型参数, Q_i 表示用户问题特征, I_n 表示第一模态特征, 上标 T 表示矩阵转置操作。

一些实施例中, 可以采用现有的各种语义相似度匹配模型对第二模态特征和用户问题特征进行语义相似度匹配。

示例性地, 可以将第二模态特征和用户问题特征进行拼接, 获得拼接向量; 通过第二全连接层, 将拼接向量映射到低维空间, 获得第二相似向量表征。

在本示例实施方式中, 第二模态特征可以是图像信息中提取的文本特征, 可以通过一个全连接层将向量映射到低维空间。

在步骤 S130 中, 对第一相似向量表征和第二相似向量表征进行信息融合, 获得多模态相似分数。

在本示例实施方式中, 可以通过门控机制对第一相似向量表征和第二相似向量表征进行信息融合。

示例性地, 可以通过门控网络, 对第一相似向量表征和第二相似向量表征进行信息融合; 对融合后向量进行全连接处理和非线性激活处理, 获得多模态相似分数。

在本示例实施方式中, 门控网络可以为简单的门控机制 (如门限超参 g 为 0.3), 也可以是学习的门控网络模型, 本示例对此不做限定。

例如, 信息融合过程可以表示为:

$$\begin{aligned} g &= \delta(W_i V_{iq} + W_t V_{tq} + b) \\ V_m &= g \circ V_{iq} + (1 - g) \circ V_{tq} \\ y &= \text{softmax}(\text{FNN}(V_m)) \end{aligned}$$

式中, $\circ$ 表示点积运算, V_m 表示门控网络的输出, V_{iq} 表示第一相似向量表征, V_{tq} 表示第二相似向量表征, W_i 、 W_t 、 b 表示门控网络的模型参数, δ 表示归一化激活函数, FNN 表示全连接处理, y 表示输出的多模态相似分数。

在上述实施例中, 如图 5 所示, 多模态相似分数的确定过程可以使用多模态匹配模型, 可以基于不同的应用领域, 利用该应用领域的训练样本对多模态匹配模型进行模型训练, 如客服领域、网页搜索领域等。该模型可以包括第一相似度匹配模块、第二相似度匹配模块、门控网络、全连接层、输出层, 第一相似度匹配模块为基于 **transformer** 的匹配模型, 第二相似度匹配模型可以包括线性映射层, 将用户问句特征和第一模态特征输入第一相似度匹配模型, 将用户问句特征和第二模态特征输入第二相似度匹配模型, 输出层输出问句与图像的多模态相似分数。

在步骤 S140 中, 响应于一多模态相似分数大于等于预设阈值, 将对应的第一模态

信息作为应答内容进行问题应答。

在本示例实施方式中，预设阈值是第一模态应答的判断条件，可以根据经验和实际情况来设置该阈值（如可以设置为 0.9），也可以根据不同情况调整该阈值，本示例对此不做限定。在一个图像与用户问句的多模态相似分数大于等于该阈值时，可以优先使用第一模态信息（即图像）作为应答语句，以增加应答内容的信息丰富度和多样性。例如，对于用户问题语句为“xx 对象的外观或者样貌”等时，采用第一模态信息（含图像的信息）进行应答能给用户带来更直观的答案，提高用户满意度。

示例性地，在存在多个多模态相似分数大于等于预设阈值的情况下，确定最高的多模态相似分数对应的第一模态信息为应答内容。例如，多个图像与用户问句的多模态相似分数满足匹配条件（大于等于预设阈值），可以将相似分数最高的图像作为应答内容进行问题应答。

在一些实施例中，多模态信息包括第二模态信息，方法还包括：响应于各多模态相似分数分别小于预设阈值，基于第二模态信息确定应答内容。

在本示例实施方式中，第二模态信息可以是文本类信息，在未达到图像匹配阈值的情况下，可以基于文本类信息进行匹配应答。

示例性地，可以对第二模态信息进行段落拆分，并确定拆分获得的各段落的关键词；基于关键词，将各段落中语义相似的段落进行聚合；通过预训练的摘要生成模型，对聚合后段落进行语句凝练，获得摘要文本；并将摘要文本作为应答内容。

在本示例实施方式中，第二模态信息是指多模态信息中除第一模态信息以外的文档信息，例如，对于含有文本和图像的文档，第二模态信息可以是除图像以外的文本信息。可以使用段落拆分工具进行段落拆分，可以采用现有的关键词提取算法进行关键词提取，例如 TF-IDF（Term-Frequency Inverse Document Frequency，词频倒文本频率）算法、TextRank 算法或者两者的结合算法。关键词也可以是各段落的标题，当前后段落/标题共享指定数量的关键词的情况下，将对应的段落聚成一个大段，以使同一段落描述相同的内容。可以使用预训练的摘要生成模型（如基于 CNN 的 Sequence-to-Sequence 模型、基于长短时记忆网络（LSTM）的 Sequence-to-Sequence 模型或基于注意力机制的 Sequence-to-Sequence 模型）对每个聚合的长段落将文本缩写成语义凝练的短句。例如，将 100 句的长段落凝练为 10 句，也可以对语句进行改述，本示例对此不做限定。

本公开方法中，对多模态信息中的第一模态信息的第一模态特征、第二模态特征的提取过程以及第二模态信息的处理过程，可以预先完成，并将提取/处理结果存储于对应的知识库中，在线问答系统可以从知识库获取相应信息进行应答响应。

举例而言，本公开的问题应答方法的实现过程如图 6 所示，具体可以通过以下步骤实现：

步骤 S601，用户向应答系统发送用户问题语句。

步骤 S602，应答系统通过预训练编码模型对用户问题语句进行编码，获得用户问

题特征。

步骤 S603，应答系统基于接收的用户问题语句，确定目标文档对应的多模态信息。

本示例中，可以通过用户问题语句关联出相关的文档内容作为目标文档，对于目标文档具有多模态信息的情况，可以转至 S604。对于目标文档具有多模态信息的情况，
5 可以对目标文档进行语句凝练并应答。

步骤 S604，应答系统从知识库分别获取多模态信息中各第一模态信息对应的第一模态特征和第二模态特征。

步骤 S605，通过预训练的摘要生成模型，对多模态信息中的第二模态信息进行语句凝练，获得摘要文本。

10 步骤 S606，应答系统对于一第一模态信息，通过第一相似度匹配模型对第一模态特征和用户问题特征进行语义相似度匹配。

本示例中，第一相似度匹配模型为 transformer 模型，transformer 模型中使用用户问题特征映射到 Q，使用第一模态特征作为 K 和 V 信息。

步骤 S607，应答系统通过第二相似度匹配模型，对第二模态特征和用户问题特征
15 进行语义相似度匹配。

本示例中，第二相似度匹配模型为全连接网络。

步骤 S608，应答系统通过门控网络，对第一相似向量表征和第二相似向量表征进行信息融合，获得多模态相似分数。

步骤 S609，应答系统判断是否存在至少一个多模态相似分数大于等于预设阈值，
20 若是，转至 S610，否则转至步骤 S611。

步骤 S610，应答系统将多模态相似分数最高的第一模态信息作为应答内容发送给用户。

步骤 S611，应答系统将步骤 S605 中的摘要文本作为应答内容发送给用户。

以上实施例中的问题应答过程中，第一相似度匹配模型、第二相似度匹配模型、门
25 控网络等与语义相似度匹配相关的模型需要进行训练，其他模型均可以使用预训练模型直接使用。

本公开针对现实应用中多数文档存在文本内容、图片内容、音频内容等并存的多模态信息表达的情况，一方面提出基于多模态的文档内容解析方法，对于含图像内容的第一模态信息进行多方面的文本信息提取，融合图像中的视觉信息和文本信息，提高图像的
30 匹配准确度。同时，基于文档内容的提前上传的现状，可以对文档中的多模态信息进行预先解析，以提高在应答过程中的系统响应速度。另一方面，通过对图像内容的两个维度的语义匹配，即第一模态特征和用户问题特征，及第二模态特征和用户问题特征的语义相似度匹配，提高匹配准确度，从而提高应答准确性。此外，对图像内容的相似分数较高的情况，直接使用图像进行应答，可以提高应答效率和应答效果，提升用户体验。

35 进一步的，本示例实施方式中，还提供了一种问题应答装置 800。该问题应答装置

700 可以应用于服务器。参考图 7 所示, 该问题应答装置 700 可以包括: 获取模块 710、语义匹配模块 720、融合模块 730 和确定模块 740, 获取模块 710, 被配置为获取多模态信息中各第一模态信息对应的第一模态特征和第二模态特征; 语义匹配模块 720, 被配置为对于一第一模态信息, 对第一模态特征和用户问题特征, 及第二模态特征和用户问题特征, 分别进行语义相似度匹配, 对应获得第一相似向量表征和第二相似向量表征; 融合模块 730, 被配置为对第一相似向量表征和第二相似向量表征进行信息融合, 获得多模态相似分数; 应答模块 740, 被配置为响应于一多模态相似分数大于等于预设阈值, 将对应的第一模态信息作为应答内容进行问题应答。

在本公开的一种示例性实施例中, 语义匹配模块 720 包括第一匹配子模块, 第一匹配子模块被配置为: 通过多层注意力层, 对第一模态特征和用户问题特征进行语义相似度匹配; 通过前馈网络层, 对多层注意力层的输出进行非线性激活和线性变换, 获得第一相似向量表征。

在本公开的一种示例性实施例中, 第一匹配子模块还被配置为: 对于每个注意力层, 对用户问题特征进行第一线性处理, 获得第一中间向量; 对第一模态特征分别进行第二线性处理和第三线性处理, 获得第二中间向量和第三中间向量; 确定第一中间向量与第二中间向量之间的相似度; 根据相似度, 对第三中间向量进行加权匹配, 获得当前注意力层的输出; 对各注意力层的输出进行向量拼接, 并通过第一全连接层对拼接后向量进行全连接处理, 获得多层注意力层的输出。

在本公开的一种示例性实施例中, 语义匹配模块 720 还包括第二匹配子模块, 第二匹配子模块被配置为: 将第二模态特征和用户问题特征进行拼接, 获得拼接向量; 通过第二全连接层, 将拼接向量映射到低维空间, 获得第二相似向量表征。

在本公开的一种示例性实施例中, 融合模块 730 还被配置为: 通过门控网络, 对第一相似向量表征和第二相似向量表征进行信息融合; 对融合后向量进行全连接处理和非线性激活处理, 获得多模态相似分数。

在本公开的一种示例性实施例中, 第一模态信息包括图像信息, 第二模态特征包括文本特征, 获取模块 710 包括第一提取子模块, 第一提取子模块被配置为: 提取一图像信息的标题信息; 通过预训练的目标检测模型, 识别该图像信息中的内容及其属性信息; 通过光学字符识别, 提取该图像信息中的候选文本信息; 对标题信息、内容及其属性信息及候选文本信息进行向量化表示, 获得该图像信息对应的文本特征。

在本公开的一种示例性实施例中, 获取模块 710 还包括第二提取子模块, 第二提取子模块被配置为: 通过图像预训练模型, 提取各第一模态信息中的图像向量信息, 获得对应的第一模态特征。

在本公开的一种示例性实施例中, 装置 700 还包括编码模块, 被配置为响应于接收到用户问题语句, 通过预训练编码模型对用户问题语句进行编码, 获得用户问题特征。

在本公开的一种示例性实施例中, 多模态信息包括第二模态信息, 应答模块 740

还被配置为：响应于各多模态相似分数分别小于预设阈值，基于第二模态信息确定应答内容。

在本公开的一种示例性实施例中，应答模块 740 还被配置为：对第二模态信息进行段落拆分，并确定拆分获得的各段落的关键词；基于关键词，将各段落中语义相似的段落进行聚合；通过预训练的摘要生成模型，对聚合后段落进行语句凝练，获得摘要文本；
5 并将摘要文本作为应答内容。

在本公开的一种示例性实施例中，应答模块 740 还被配置为：响应于多个多模态相似分数大于等于预设阈值，确定最高的多模态相似分数对应的第一模态信息为应答内容。

上述问题应答装置中各模块或单元的具体细节已经在对应的问题应答方法中进行了详细的描述，因此此处不再赘述。
10

作为另一方面，本申请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电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。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电子设备执行时，使得该电子设备实现如下述实施例中的方法。例如，电子设备可以实现如
15 图 1、图 2 及图 4-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）、
20 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。
25 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
30 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电子设备 800。图 8 显示的电子设备 800 仅仅是一个示例，不应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，电子设备 800 以通用计算设备的形式表现。电子设备 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
35 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，并且将使本领域技术人员能够利用本公开。

权利要求

1. 一种问题应答方法，包括：
获取多模态信息中各第一模态信息对应的第一模态特征和第二模态特征；
对于一所述第一模态信息，对所述第一模态特征和用户问题特征，及所述第二模态特征和所述用户问题特征，分别进行语义相似度匹配，对应获得第一相似向量表征和第二相似向量表征；
对所述第一相似向量表征和所述第二相似向量表征进行信息融合，获得多模态相似分数；
响应于一所述多模态相似分数大于等于预设阈值，将对应的第一模态信息作为应答内容10 容进行问题应答。
2. 根据权利要求 1 所述的问题应答方法，其中，所述对第一模态特征和用户问题特征进行语义相似度匹配，包括：
通过多层注意力层，对所述第一模态特征和所述用户问题特征进行语义相似度匹配；
通过前馈网络层，对所述多层注意力层的输出进行非线性激活和线性变换，获得所述15 第一相似向量表征。
3. 根据权利要求 2 所述的问题应答方法，其中，所述通过多层注意力层，对所述第一模态特征和所述用户问题特征进行语义相似度匹配，包括：
对于每个注意力层，对所述用户问题特征进行第一线性处理，获得第一中间向量；
对所述第一模态特征分别进行第二线性处理和第三线性处理，获得第二中间向量和第三中间向量；
20 确定所述第一中间向量与所述第二中间向量之间的相似度；
根据所述相似度，对所述第三中间向量进行加权匹配，获得当前注意力层的输出；
对各注意力层的输出进行向量拼接，并通过第一全连接层对拼接后向量进行全连接处理，获得所述多层注意力层的输出。
- 25 4. 根据权利要求 1 所述的问题应答方法，其中，所述对第二模态特征和用户问题特征进行语义相似度匹配，包括：
将所述第二模态特征和所述用户问题特征进行拼接，获得拼接向量；
通过第二全连接层，将所述拼接向量映射到低维空间，获得所述第二相似向量表征。
5. 根据权利要求 1 所述的问题应答方法，其中，所述对所述第一相似向量表征和所述30 第二相似向量表征进行信息融合，包括：
通过门控网络，对所述第一相似向量表征和所述第二相似向量表征进行信息融合；
对融合后向量进行全连接处理和非线性激活处理，获得所述多模态相似分数。
6. 根据权利要求 1 所述的问题应答方法，其中，所述第一模态信息包括图像信息，所述35 第二模态特征包括文本特征，所述获取各第一模态信息对应的第二模态特征，包括：
提取一所述图像信息的标题信息；

通过预训练的目标检测模型，识别该图像信息中的内容及其属性信息；

通过光学字符识别，提取该图像信息中的候选文本信息；

对所述标题信息、所述内容及其属性信息及所述候选文本信息进行向量化表示，获得该图像信息对应的文本特征。

5 7. 根据权利要求 1 所述的问题应答方法，其中，所述获取各第一模态信息对应的第一模态特征，包括：

通过图像预训练模型，提取各所述第一模态信息中的图像向量信息，获得对应的第一模态特征。

8. 根据权利要求 6 所述的问题应答方法，其中，所述方法还包括：

10 响应于接收到用户问题语句，通过预训练编码模型对所述用户问题语句进行编码，获得所述用户问题特征。

9. 根据权利要求 1 至 8 任一项所述的问题应答方法，其中，所述多模态信息包括第二模态信息，所述方法还包括：

15 响应于各所述多模态相似分数分别小于预设阈值，基于所述第二模态信息确定应答内容。

10. 根据权利要求 9 所述的问题应答方法，其中，所述基于所述第二模态信息确定应答内容，包括：

对所述第二模态信息进行段落拆分，并确定拆分获得的各段落的关键词；

基于所述关键词，将各段落中语义相似的段落进行聚合；

20 通过预训练的摘要生成模型，对聚合后段落进行语句凝练，获得摘要文本；并将所述摘要文本作为应答内容。

11. 根据权利要求 1 所述的问题应答方法，其中，所述方法还包括：

响应于多个所述多模态相似分数大于等于预设阈值，确定最高的多模态相似分数对应的第一模态信息为应答内容。

25 12. 一种问题应答装置，所述装置包括：

获取模块，被配置为获取多模态信息中各第一模态信息对应的第一模态特征和第二模态特征；

30 语义匹配模块，被配置为对于一所述第一模态信息，对所述第一模态特征和用户问题特征，及所述第二模态特征和所述用户问题特征，分别进行语义相似度匹配，对应获得第一相似向量表征和第二相似向量表征；

融合模块，被配置为对所述第一相似向量表征和所述第二相似向量表征进行信息融合，获得多模态相似分数；

应答模块，被配置为响应于一所述多模态相似分数大于等于预设阈值，将对应的第一模态信息作为应答内容进行问题应答。

35 13. 一种电子设备，包括：一个或多个处理器；以及

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。