(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111737424 A
(43)申请公布日 2020.10.02

(21)申请号 202010108370.2

(22)申请日 2020.02.21

(71)申请人 北京沃东天骏信息技术有限公司
地址 100176 北京市大兴区北京经济技术开发区科创十一街18号院2号楼4层A402室
申请人 北京京东世纪贸易有限公司

(72)发明人 刘航 陈蒙

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 孟金喆

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06K 9/62(2006.01)

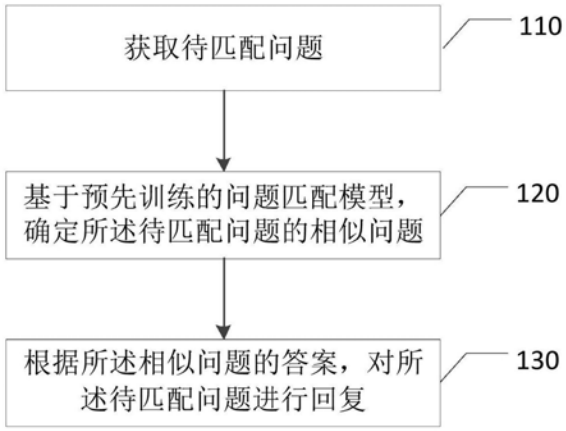
权利要求书2页 说明书9页 附图3页

(54)发明名称

一种问题的匹配方法、装置、设备及存储介质

(57)摘要

本发明公开了一种问题的匹配方法、装置、设备及存储介质，该问题的匹配方法包括：获取待匹配问题；基于预先训练的问题匹配模型，确定所述待匹配问题的相似问题；根据所述相似问题的答案，对所述待匹配问题进行回复；其中，所述问题匹配模型的训练数据包括基于历史对话语料挖掘的相似问题对。本发明实施例的技术方案，通过问题匹配模型进行用户问题的匹配，实现了用户问题的自动匹配和回复，同时，该问题匹配模型的训练数据包括基于历史对话语料挖掘的相似问题，提高了模型匹配的成功率和准确度。



CN 111737424 A

1.一种问题的匹配方法,其特征在于,包括:

获取待匹配问题;

基于预先训练的问题匹配模型,确定所述待匹配问题的相似问题;

根据所述相似问题的答案,对所述待匹配问题进行回复;

其中,所述问题匹配模型的训练数据包括基于历史对话语料挖掘的相似问题对。

2.根据权利要求1所述的方法,其特征在于,基于历史对话语料挖掘相似问题对,包括:

获取历史对话语料;

抽取所述历史对话语料的问题答案对;

确定所述问题答案对中的当前答案的相似答案,根据所述相似答案确定所述问题答案对中的当前问题对应的相似问题,将所述当前问题和所述当前问题对应的相似问题作为相似问题对,其中,所述当前问题为所述当前答案对应的问题。

3.根据权利要求2所述的方法,其特征在于,所述确定所述问题答案对中的当前答案的相似答案,根据所述相似答案确定所述问题答案对中的当前问题对应的相似问题,包括:

获取所述问题答案对中的当前答案的关键词;

根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度,确定所述当前答案的相似答案,其中,所述答案集合包括各个所述问题答案对中的答案;

将所述相似答案对应的问题,确定为所述问题答案对中的当前问题对应的相似问题。

4.根据权利要求3所述的方法,其特征在于,在根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度,确定所述当前答案的相似答案之前,还包括:

获取所述当前答案的长度;

根据所述当前答案的长度对所述答案集合进行筛选。

5.根据权利要求3所述的方法,其特征在于,在根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度,确定所述当前答案的相似答案之后,将所述相似答案对应的问题,确定为所述问题答案对中的当前问题对应的相似问题之前,还包括:

确定各个所述相似答案与所述当前答案的编辑距离;

当所述相似答案的编辑距离大于设定距离阈值时,则将所述相似答案移除。

6.根据权利要求2所述的方法,其特征在于,在抽取所述历史对话语料的问题答案对之前,还包括:

对所述历史对话语料进行预处理,其中,所处预处理包括:过滤长度小于设定长度的文本数据、根据预设话术库删除所述历史对话语料中的预设用语、以及对相同问题和答案进行去重处理中的至少一项。

7.根据权利要求1所述的方法,其特征在于,还包括:

获取人工标注的标注相似问题对以及基于所述标注相似问题对训练的初始匹配模型;

根据所述基于历史对话语料挖掘的相似问题对,生成所述问题匹配模型的正样本和负样本;

将所述正样本和负样本分别输入所述初始匹配模型,根据所述初始匹配模型的输出,得到各个正样本和负样本的样本分数;

将样本分数处于预设范围的正样本和负样本作为所述问题匹配模型的训练数据。

8.一种问题的匹配装置,其特征在于,包括:

待匹配问题获取模块，用于获取待匹配问题；

相似问题确定模块，用于基于预先训练的问题匹配模型，确定所述待匹配问题的相似问题；

问题回复模块，用于根据所述相似问题的答案，对所述待匹配问题进行回复；

其中，所述问题匹配模型的训练数据包括基于历史对话语料挖掘的相似问题对。

9.一种终端设备，其特征在于，所述设备包括：

一个或多个处理器；

存储器，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，使得所述一个或多个处理器实现如权利要求1-7中任一项所述的问题的匹配方法。

10.一种包含计算机可执行指令的存储介质，其特征在于，所述计算机可执行指令在由计算机处理器执行时用于执行如权利要求1-7中任一项所述的问题的匹配方法。

一种问题的匹配方法、装置、设备及存储介质

技术领域

[0001] 本发明实施例涉及文本匹配技术领域，尤其涉及一种问题的匹配方法、装置、设备及存储介质。

背景技术

[0002] 在智能客服的应用场景中，对于用户频繁问到的业务知识类问题的自动解答是一个非常关键的需求。尤其在电商行业，大促期间人工客服很难快速应对每一个用户，因此建立一个高质量的智能客服系统，可以显著地降低人工客服的数量与成本。

[0003] 目前主流的智能客服系统大多采用了检索式的方法，从知识库(问句-答案)中检索出与用户问题最相似的问句，然后返回该问句的答案来解答用户问题。该方法可以针对用户的问题进行专业性解答，避免出现插科打诨的答案。

[0004] 然而，在实现本发明过程中，发明人发现现有技术中至少存在如下问题：现有的用户问题匹配方法匹配错误率和失败率较高，无法满足用户的需求。

发明内容

[0005] 本发明提供了一种问题的匹配方法、装置、设备及存储介质，以提高问题匹配的成功率和准确率。

[0006] 第一方面，本发明实施例提供了一种问题的匹配方法，该方法包括：

[0007] 获取待匹配问题；

[0008] 基于预先训练的问题匹配模型，确定所述待匹配问题的相似问题；

[0009] 根据所述相似问题的答案，对所述待匹配问题进行回复；

[0010] 其中，所述问题匹配模型的训练数据包括基于历史对话语料挖掘的相似问题对。

[0011] 第二方面，本发明实施例还提供了一种问题的匹配装置，该装置包括：

[0012] 待匹配问题获取模块，用于获取待匹配问题；

[0013] 相似问题确定模块，用于基于预先训练的问题匹配模型，确定所述待匹配问题的相似问题；

[0014] 问题回复模块，用于根据所述相似问题的答案，对所述待匹配问题进行回复；

[0015] 其中，所述问题匹配模型的训练数据包括基于历史对话语料挖掘的相似问题对。

[0016] 第三方面，本发明实施例还提供了一种问题的匹配设备，该设备包括：

[0017] 一个或多个处理器；

[0018] 存储器，用于存储一个或多个程序；

[0019] 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现本发明任意实施例提供的问题的匹配方法。

[0020] 第四方面，本发明实施例还提供了一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行本发明任意实施例提供的问题的匹配方法。

[0021] 本发明实施例的技术方案，通过预先训练的模型进行待匹配问题的相似问题匹配，以根据相似问题的答案进行回复，实现了问题的自动回复，同时，通过模型匹配相似问题来进行问题回复，提高了问题回复的成功率；并且，该模型的训练数据中包括基于对话语料挖掘的相似问题对，提高了模型建立的效率，同时，训练数据的数据量大，从而提高了模型匹配的准确度。

附图说明

[0022] 图1是本发明实施例一中的一种问题的匹配方法的流程图；

[0023] 图2是本发明实施例二中的一种问题的匹配方法的流程图；

[0024] 图3是本发明实施例三中的一种问题的匹配装置的结构示意图；

[0025] 图4是本发明实施例四中的一种问题的匹配设备的结构示意图。

具体实施方式

[0026] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅用于解释本发明，而非对本发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部结构。

[0027] 实施例一

[0028] 图1是本发明实施例一提供的一种问题的匹配方法的流程图，本实施例可适用于对用户问题进行匹配和自动回复的情况，该方法可以由问题的匹配装置来执行，如图1所示，该方法具体包括如下步骤：

[0029] 步骤110、获取待匹配问题。

[0030] 其中，待匹配问题可以是用户输入的问题待回复的问题或者语句，其输入形式可以是文字形式、语音形式或者其他形式。待匹配问题可以是采用自然语言编辑的问题，如汉语、英语、日语等。待匹配问题可以是任意一个用户提出的问题，如“今天可以送货吗?”、“乡下支持送货安装吗”、“应该是包邮的，对吧”等，可以是以问号为结尾的问句，或者是未含有问号的问句，还可以是陈述句。需要注意的是，本发明实施例所提及的待匹配问题，可以是一个问句也可以是陈述句，用于表示需要回答的语句，不限于问句。

[0031] 具体的，可以通过设定对话框获取用户输入的待匹配问题。

[0032] 进一步地，在获取待匹配问题之后，还包括：

[0033] 对所述待匹配问题进行预处理。

[0034] 其中，预处理包括过滤处理、分词处理、语义分析、特征提取、语句编码等其中的一项或者多项。

[0035] 步骤120、基于预先训练的问题匹配模型，确定所述待匹配问题的相似问题。

[0036] 其中，预先训练的问题匹配模型可以是深度特征语义模型(Deep Feature Semantic Model，DFSM)，具体可以是深度结构化语义模型(Deep Structured Semantic Model，DSSM)、双向编码器表征模型(Bidirectional Encoder Representations from Transformers，BERT)、增强序列推断模型(Enhanced Sequential Inference Model，ESIM)或者其他学习模型。

[0037] 其中，所述问题匹配模型的训练数据包括基于历史对话语料挖掘的相似问题对。

历史对话语料指的是设定平台或设定商铺所获取的各个用户与任意一个或多个客服的历史对话内容。相似问题对指的是问题或问句内容相似或相匹配的两个或两个以上的问题的组合，相似问题对中的各个问题对应的答案内容也是相似的，或者，相似问题对中任意一个问题的答案均可以作为其他问题的答案。

[0038] 可选的，基于历史对话语料挖掘相似问题对，包括：

[0039] 获取历史对话语料；抽取所述历史对话语料的问题答案对；确定所述问题答案对中的当前答案的相似答案，根据所述相似答案确定所述问题答案对中的当前问题对应的相似问题，将所述当前问题和所述当前问题对应的相似问题作为相似问题对，其中，所述当前问题为所述当前答案对应的问题。

[0040] 其中，问题答案对指的是问题及其对应的答案。相似答案指的是各个答案所对应的语句的内容相似或相近。

[0041] 具体的，可以根据预设标识确定当前语句是来自客户还是来自设定平台的客服，来自客户的语句即为问题，而与在该客户语句之后的来自客服的语句即为该问题的答案。预设标识可以是Q和A标识，Q标识表示来自用户的问题，A则表示来自客服的答案。当然也可以采用其他标识进行用户提出的问题和客服回复的答案的区分。

[0042] 具体的，抽取了历史对话语料中的全部问题答案对之后，根据各个问题对应的答案是否相似来判断相应的问题是否是相似问题对。即若两个问题所对应的答案为相似答案，那么，这两个问题即为相似问题。通过确定当前问题对应的答案的相似答案的方式，确定历史对话语料中与当前问题相似的问题，从而从对话语料中挖掘出各个问题的相似问题。

[0043] 具体的，所述问题匹配模型，根据接收到的所述待匹配问题，自动输出可选的候选相似问题及各个候选相似问题对应的分数，并将得分最高的候选相似问题确定为所输入的待匹配问题的相似问题。

[0044] 进一步地，所述问题匹配模型的训练数据还可以包括由人工标注的相似问题对组成的数据库。即该问题匹配模型的训练数据可以仅包括基于历史对话语料挖掘相似问题对，也可以包括由人工标注的相似问题对组成的数据库和基于历史对话语料挖掘相似问题对。

[0045] 步骤130、根据所述相似问题的答案，对所述待匹配问题进行回复。

[0046] 由于问题与答案在历史对话语料中是成对出现的，即一个问题对应一个答案。那么，在确定了待匹配问题的相似问题后，将该相似问题对应的答案作为待匹配问题的答案进行回复。

[0047] 本发明实施例的技术方案，通过预先训练的模型进行待匹配问题的相似问题匹配，以根据相似问题的答案进行回复，实现了问题的自动回复，同时，通过模型匹配相似问题来进行问题回复，提高了问题回复的成功率；并且，该模型的训练数据中包括基于对话语料挖掘的相似问题对，提高了模型建立的效率，同时，训练数据的数据量大，从而提高了模型匹配的准确度。

[0048] 实施例二

[0049] 图2为本发明实施例二提供的一种问题的匹配方法的流程图，本实施例是对上一实施例的进一步细化和补充，本实施例所提供的问题的匹配方法还包括问题匹配模型的训

练过程。

[0050] 如图2所示，该问题的匹配方法包括如下步骤：

[0051] 步骤210、获取历史对话语料。

[0052] 步骤220、对所述历史对话语料进行预处理。

[0053] 其中，所处预处理包括：过滤长度小于设定长度的文本数据、根据预设话术库删除所述历史对话语料中的预设用语、以及对相同问题和答案进行去重处理中的至少一项。

[0054] 具体的，由于历史对话语料通常为用户与客服的对话内容，通常会包括“你好”、“谢谢”等客气用语，还会存在一个问题重复发送的现象，以及一些长度过段甚至是数据内容为空或标点符号、表情等无实质内容的语句，因此，在获取历史对话语料之后，在抽取问题答案对之前，为了提高处理效率，需要对历史对话语料进行一些预处理，具体过程为：首先，删除历史对话语料中的特殊符号和表情；其次，计算历史对话语料中的各个语句的文本长度，过滤历史对话语料中文本长度小于设定长度的文本数据，如长度小于3个汉字长度；接着，根据预设话术库删除历史对话语料中的预设用语，其中，预设用语包括客气用语，如“你好”、“谢谢”、“不客气”等，还包括惯用话术，如“在吗”、“亲亲”、“未能及时回复请见谅”等；最后，将历史对话语料中的相同问题和答案进行去重处理，即对于完全相同的问题答案对仅保留其中一个。通过上述预处理，有效减少了历史对话语料的数据量，提高了问题答案对的抽取效率。

[0055] 步骤230、抽取所述历史对话语料的问题答案对。

[0056] 步骤240、获取所述问题答案对中的当前答案的关键词。

[0057] 具体的，可以基于预设算法获取当前答案的关键词，其中，预设算法可以是分词算法，如基于字符串匹配的分词方法、基于语义分析的分词方法、基于字标注的中文分词方法或者基于神经网络的分词算法等。也可以简单的根据语句(当前答案)中各个词的词性进行关键词抽取，如仅获取当前答案中的名词和动词。

[0058] 步骤250、根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度，确定所述当前答案的相似答案。

[0059] 其中，所述答案集合包括各个所述问题答案对中的答案。重合度指的是两个答案所对应的关键词重合的个数，也可以指的是答案集合中的答案的关键词与当前答案的关键词的重合率，其表达式为：

[0060] $$P = \frac{M}{N} \times 100\%$$

[0061] 其中，P表示重合率，M表示两个答案(答案集合中的任意一个答案和当前答案)的关键词的重复个数，N表示当前答案的关键词的总个数。

[0062] 具体的，可以当答案集合中的目标答案与当前答案的关键词的重合度大于设定个数，或重合度大于设定百分比时，则确定所述目标答案为当前答案的相似答案。设定个数可以是3个、5个或者其他值，设定百分比可以是[80%，100%)之前的任意值，如90%、95%等。

[0063] 可选的，在根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度，确定所述当前答案的相似答案之前，还包括：

[0064] 获取所述当前答案的长度；根据所述当前答案的长度对所述答案集合进行筛选。

[0065] 假设当前答案的长度为l_0，长度阈值区间为$[l_{min}, l_{max}]$，其中，$l_{min}=a_1*l_0$，$l_{max}=$

a_2*l_0，其中，a_1为长度阈值下限系数，其值可以是0.7、0.8或者其他值，a_2为长度阈值上限系数，其值可以是1.2、1.3或者其他值。若答案集合中的答案的长度位于该长度阈值区间内则保留，反之则将该答案从答案集合中删除。

[0066] 可选的，在根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度，确定所述当前答案的相似答案之后，将所述相似答案对应的问题，确定为所述问题答案对中的当前问题对应的相似问题之前，还包括：

[0067] 确定各个所述相似答案与所述当前答案的编辑距离；当所述相似答案的编辑距离大于设定距离阈值时，则将所述相似答案移除。

[0068] 其中，编辑距离(Minimum Edit Distance，MED)指的是相似答案转换为当前答案所需要的最少编辑操作次数，如插入、删除、替换等。设定距离阈值可以是3、5、8或者其他值。

[0069] 进一步地，还可以根据预设编辑距离评分关系式，计算相似答案的编辑距离分数，其中，编辑距离分数越大，则表示相似答案与当前答案的编辑距离越小。当该编辑距离分数大于设定分数时，则保留该相似答案，反之，则将该相似答案移除。该编辑距离分数的满分可以是1分，当编辑距离分数为满分时，则表示相似答案与当前答案相同，无需编辑。

[0070] 具体的，设定分数可以是0.75、0.8、0.9或者其他值，编辑距离分数的表达式可以为：

[0071] $$S_{ij}=1-\frac{D_{ij}}{\max(A_i,A_j)}$$

[0072] 其中，S_{ij}表示当前答案A_i和相似答案A_j的编辑距离分数，D_{ij}表示当前答案A_i和相似答案A_j的编辑距离，$\max(A_i,A_j)$表示当前答案A_i和相似答案A_j的长度的最大值。

[0073] 示例性的，假设答案A_i和A_j为相似答案对，设定分数为0.7，其中，A_i为当前答案，其长度为18，A_j为相似答案，其长度为20，A_j和A_i的编辑距离为5，则编辑距离分数为0.75，大于设定分数，则保留该相似答案A_j。

[0074] 进一步地，在进行确定编辑距离之前，还包括：获取相似答案的个数，若相似答案的个数小于预设个数，则跳过确定该相似答案的编辑距离以及根据编辑距离进行相似答案移除的步骤，直接将相似答案确定为当前答案的相似答案。其中，预设个数可以是3个、2个、1个或者其他值。

[0075] 步骤260、将所述相似答案对应的问题，确定为所述问题答案对中的当前问题对应的相似问题，将所述当前问题和所述当前问题对应的相似问题作为相似问题对。

[0076] 其中，所述当前问题为所述当前答案对应的问题。

[0077] 通过上述步骤确定了历史对话语料中的相似问题对，将挖掘的相似问题对作为问题匹配模型训练数据中的正样本，将历史对话语料中的不相似的问题对作为模型训练数据中的负样本，进行模型训练。

[0078] 进一步地，还包括：

[0079] 获取人工标注的标注相似问题对以及基于所述标注相似问题对训练的初始匹配模型；根据所述基于历史对话语料挖掘的相似问题对，生成所述问题匹配模型的正样本和负样本；将所述正样本和负样本分别输入所述初始匹配模型，根据所述初始匹配模型的输出，得到各个正样本和负样本的样本分数；将样本分数处于预设范围的正样本和负样本作

为所述问题匹配模型的训练数据。

[0080] 其中，初始匹配模型可以是与问题匹配模型采用相同模型结构建立的模型，初始匹配模型也可以是现有的任意一种训练好的模型。正样本指的是由各个相似问题对组成的样本，负样本则是由各个非相似问题对组成的样本。正样本的预设范围可以是[0.7,0.9]、[0.7,0.8]或者其他范围区间，负样本的预设范围可以是[0.1,0.3]、[0.1,0.2]或者其他范围区间，其中1分为满分，表示两个问题完全相同。通过根据由人工标注数据训练的初始匹配模型进行训练数据的筛选，减少了训练数据的噪声数据，进一步提高了训练数据的精准度。同时，当问题匹配模型的训练数据同时采用基于历史对话语料挖掘的数据和人工标注数据时，有效避免了两种数据的数据冗余，提高了模型训练的效率。

[0081] 步骤270、根据训练数据，进行问题匹配模型训练，以获取训练好的问题匹配模型。

[0082] 具体的，将通过上述步骤从历史对话语料中挖掘的相似问题对和非相似问题对作为训练数据进行模型训练，获取由人工标注的标注数据库，其中，包括各个标注相似问题对和标注非相似问题对，并将该标注数据库中的数据按一定比例生成验证集和测试集，如1:1,2:1等，以进行模型的验证和测试。

[0083] 通过上述的模型训练、验证和测试，最终获取训练完成的问题匹配模型。

[0084] 步骤280、获取待匹配问题。

[0085] 步骤290、基于预先训练的所述问题匹配模型，确定所述待匹配问题的相似问题。

[0086] 步骤300、根据所述相似问题的答案，对所述待匹配问题进行回复。

[0087] 本发明实施例的技术方案，通过对历史对话语料进行预处理，去除了数据中的客气用语、重复问题和无实质意义的语句，减少了历史对话语料的数据量；通过抽取历史对话语料中的问题答案对，并根据答案的长度、关键词以及编辑距离三方面进行相似答案的确定，提高了相似答案确定的效率和准确度，并根据相似答案确定相似问题，从而实现了基于历史对话语料挖掘相似问题对，并将其作为训练数据，进行模型训练，增加了模型训练所采用的数据量，提高了模型的匹配准确度，以及提高了问题匹配的成功率。同时，在确定训练数据时，结合人工标注数据训练的初始匹配模型进行数据筛选，提高了训练数据的质量，同时避免了挖掘数据与标注数据的冗余，提高了模型训练的效率。

[0088] 实施例三

[0089] 图3是本发明实施例三提供的一种问题的匹配装置的示意图，如图3所示，该装置包括：待匹配问题获取模块310、相似问题确定模块320和问题回复模块330。

[0090] 其中，待匹配问题获取模块310，用于获取待匹配问题；相似问题确定模块320，用于基于预先训练的问题匹配模型，确定所述待匹配问题的相似问题；问题回复模块330，用于根据所述相似问题的答案，对所述待匹配问题进行回复；其中，所述问题匹配模型的训练数据包括基于历史对话语料挖掘的相似问题对。

[0091] 本发明实施例的技术方案，通过预先训练的模型进行待匹配问题的相似问题匹配，以根据相似问题的答案进行回复，实现了问题的自动回复，同时，通过模型匹配相似问题来进行问题回复，提高了问题回复的成功率；并且，该模型的训练数据中包括基于对话语料挖掘的相似问题对，提高了模型建立的效率，同时，训练数据的数据量大，从而提高了模型匹配的准确度。

[0092] 可选的，该问题的匹配装置，还包括：

[0093] 相似问题对挖掘模块，用于基于历史对话语料挖掘相似问题对；模型训练模块，用于基于历史对话语料挖掘的相似问题对进行所述问题匹配模型训练。

[0094] 可选的，相似问题对挖掘模块，包括：

[0095] 历史对话语料获取单元，用于获取历史对话语料；问题答案对抽取单元，用于抽取所述历史对话语料的问题答案对；相似问题确定单元，用于确定所述问题答案对中的当前答案的相似答案，根据所述相似答案确定所述问题答案对中的当前问题对应的相似问题，将所述当前问题和所述当前问题对应的相似问题作为相似问题对，其中，所述当前问题为所述当前答案对应的问题。

[0096] 可选的，相似问题确定单元，具体用于：

[0097] 获取所述问题答案对中的当前答案的关键词；根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度，确定所述当前答案的相似答案，其中，所述答案集合包括各个所述问题答案对中的答案；将所述相似答案对应的问题，确定为所述问题答案对中的当前问题对应的相似问题。

[0098] 可选的，该问题的匹配装置，还包括：

[0099] 长度筛选单元，用于在根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度，确定所述当前答案的相似答案之前，获取所述当前答案的长度；根据所述当前答案的长度对所述答案集合进行筛选。

[0100] 可选的，该问题的匹配装置，还包括：

[0101] 编辑距离筛选单元，用于在根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度，确定所述当前答案的相似答案之后，将所述相似答案对应的问题，确定为所述问题答案对中的当前问题对应的相似问题之前，确定各个所述相似答案与所述当前答案的编辑距离；当所述相似答案的编辑距离大于设定距离阈值时，则将所述相似答案移除。

[0102] 可选的，该问题的匹配装置，还包括：

[0103] 预处理模块，用于在抽取所述历史对话语料的问题答案对之前，对所述历史对话语料进行预处理，其中，所处预处理包括：过滤长度小于设定长度的文本数据、根据预设话术库删除所述历史对话语料中的预设用语、以及对相同问题和答案进行去重处理中的至少一项。

[0104] 可选的，该问题的匹配装置，还包括：

[0105] 训练数据筛选模块，用于获取人工标注的标注相似问题对以及基于所述标注相似问题对训练的初始匹配模型；根据所述基于历史对话语料挖掘的相似问题对，生成所述问题匹配模型的正样本和负样本；将所述正样本和负样本分别输入所述初始匹配模型，根据所述初始匹配模型的输出，得到各个正样本和负样本的样本分数；将样本分数处于预设范围的正样本和负样本作为所述问题匹配模型的训练数据。

[0106] 本发明实施例所提供的问题的匹配装置可执行本发明任意实施例所提供的问题的匹配方法，具备执行方法相应的功能模块和有益效果。

[0107] 实施例四

[0108] 图4为本发明实施例四提供的一种问题的匹配设备的结构示意图，如图4所示，该设备包括处理器410、存储器420、输入装置430和输出装置440；设备处理器410的数量可以

是一个或多个，图4中以一个处理器410为例；设备中的处理器410、存储器420、输入装置430和输出装置440可以通过总线或其他方式连接，图4中以通过总线连接为例。

[0109] 存储器420作为一种计算机可读存储介质，可用于存储软件程序、计算机可执行程序以及模块，如本发明实施例中的问题的匹配方法对应的程序指令/模块(例如，问题的匹配装置中的待匹配问题获取模块310、相似问题确定模块320和问题回复模块330)。处理器410通过运行存储在存储器420中的软件程序、指令以及模块，从而执行设备的各种功能应用以及数据处理，即实现上述的问题的匹配方法。

[0110] 存储器420可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序；存储数据区可存储根据终端的使用所创建的数据等。此外，存储器420可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实例中，存储器420可进一步包括相对于处理器410远程设置的存储器，这些远程存储器可以通过网络连接至设备/终端/服务器。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0111] 输入装置430可用于接收输入的数字或字符信息，以及产生与设备的用户设置以及功能控制有关的键信号输入。输出装置440可包括显示屏等显示设备。

[0112] 实施例五

[0113] 本发明实施例五还提供一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行一种问题的匹配方法，该方法包括：

[0114] 获取待匹配问题；

[0115] 基于预先训练的问题匹配模型，确定所述待匹配问题的相似问题；

[0116] 根据所述相似问题的答案，对所述待匹配问题进行回复；

[0117] 其中，所述问题匹配模型的训练数据包括基于历史对话语料挖掘的相似问题对。

[0118] 当然，本发明实施例所提供的一种包含计算机可执行指令的存储介质，其计算机可执行指令不限于如上所述的方法操作，还可以执行本发明任意实施例所提供的问题的匹配方法中的相关操作。

[0119] 通过以上关于实施方式的描述，所属领域的技术人员可以清楚地了解到，本发明可借助软件及必需的通用硬件来实现，当然也可以通过硬件实现，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如计算机的软盘、只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、闪存(FLASH)、硬盘或光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述的方法。

[0120] 值得注意的是，上述问题的匹配装置的实施例中，所包括的各个单元和模块只是按照功能逻辑进行划分的，但并不局限于上述的划分，只要能够实现相应的功能即可；另外，各功能单元的具体名称也只是为了便于相互区分，并不用于限制本发明的保护范围。

[0121] 注意，上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解，本发明不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此，虽然通过以上实施例对本发明进行

了较为详细的说明,但是本发明不仅仅限于以上实施例,在不脱离本发明构思的情况下,还可以包括更多其他等效实施例,而本发明的范围由所附的权利要求范围决定。

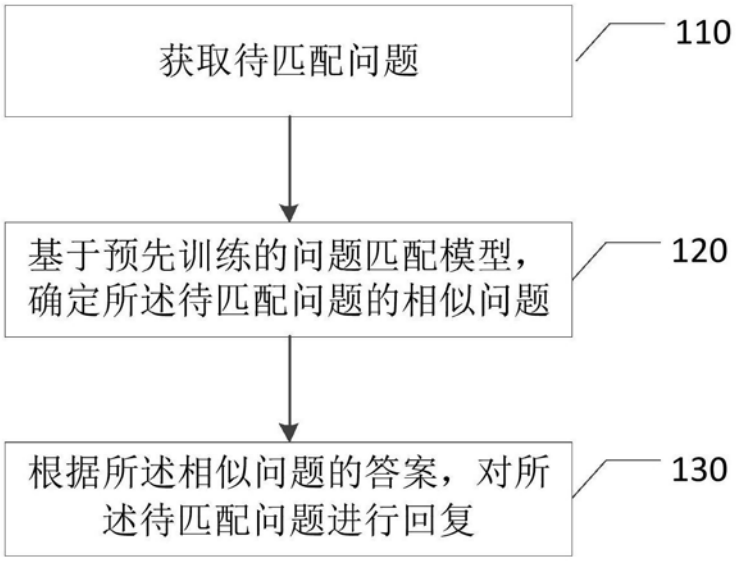

获取待匹配问题
110
基于预先训练的问题匹配模型，确定所述待匹配问题的相似问题
120
根据所述相似问题的答案，对所述待匹配问题进行回复
130

图1

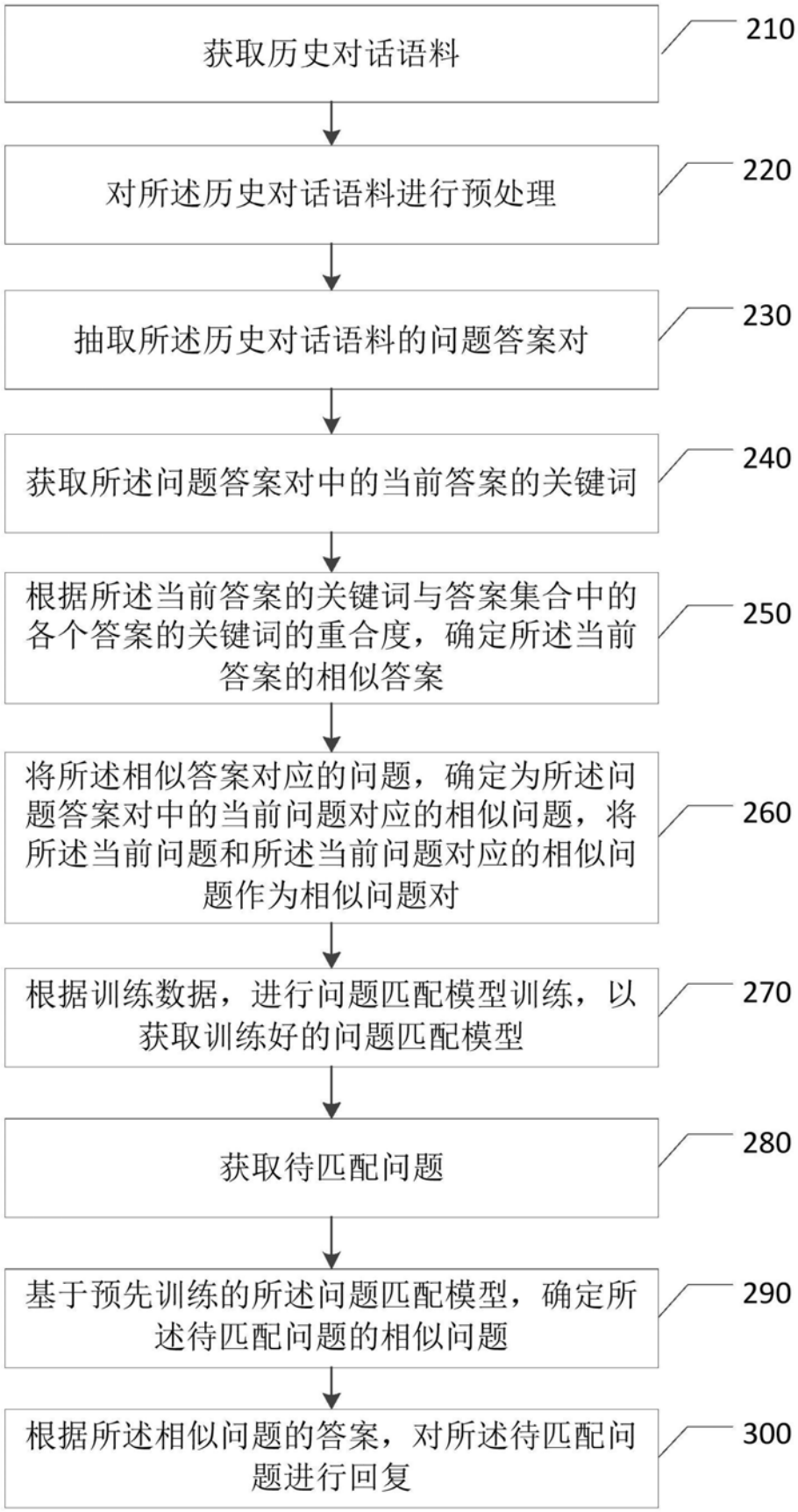

获取历史对话语料
210
对所述历史对话语料进行预处理
220
抽取所述历史对话语料的问题答案对
230
获取所述问题答案对中的当前答案的关键词
240
根据所述当前答案的关键词与答案集合中的各个答案的关键词的重合度，确定所述当前答案的相似答案
250
将所述相似答案对应的问题，确定为所述问题答案对中的当前问题对应的相似问题，将所述当前问题和所述当前问题对应的相似问题作为相似问题对
260
根据训练数据，进行问题匹配模型训练，以获取训练好的问题匹配模型
270
获取待匹配问题
280
基于预先训练的所述问题匹配模型，确定所述待匹配问题的相似问题
290
根据所述相似问题的答案，对所述待匹配问题进行回复
300

图2

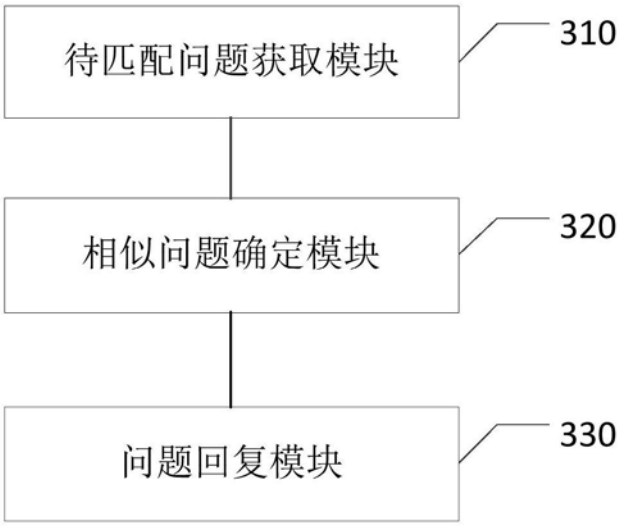

待匹配问题获取模块
310
相似问题确定模块
320
问题回复模块
330

图3

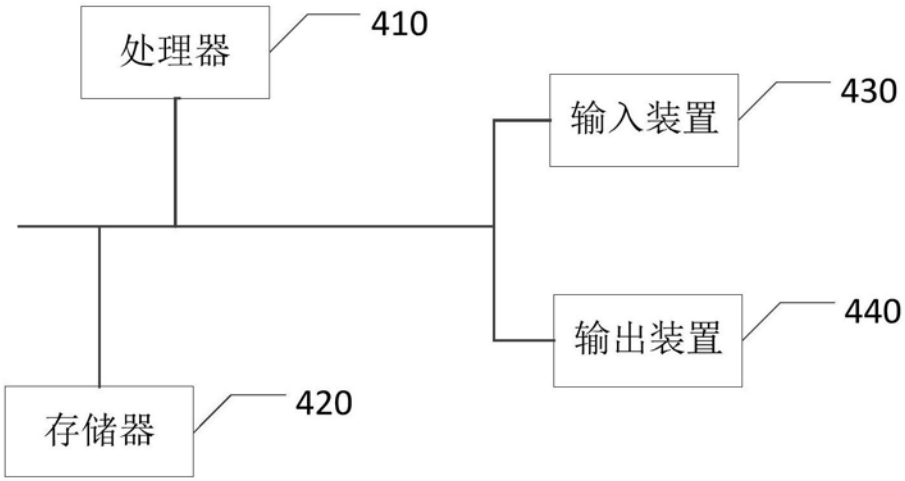

处理器
410
输入装置
430
输出装置
440
存储器
420

图4