(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 107704521 A
(43)申请公布日 2018.02.16

(21)申请号 201710801967.3

(22)申请日 2017.09.07

(71)申请人 北京零秒科技有限公司
地址 100000 北京市海淀区领秀新硅谷C区22号楼601

(72)发明人 许宇航 黄丽辉

(74)专利代理机构 北京卓唐知识产权代理有限公司 11541
代理人 龚洁

(51)Int.Cl.
G06F 17/30(2006.01)

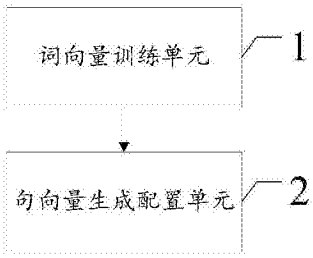
权利要求书1页 说明书5页 附图3页

(54)发明名称

一种问答处理服务器、客户端以及实现方法

(57)摘要

本发明公开了一种问答处理服务器、客户端以及实现方法，所述服务器包括：词向量训练单元和句向量生成配置单元，所述词向量训练单元，用以根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，所述句向量生成配置单元，用以根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口。本发明训练词向量的语料能够让使用者替换，对于句向量中的关键词可进行权重的调整，从而能够根据不同的对话场景配置训练不同的词向量模型，使得模型的准确度更高，从而提高问答处理服务器的匹配准确率。此外，本发明还能够满足用户特殊要求的词汇处理，匹配出更符合的问答答案。



CN 107704521 A

1.一种问答处理服务器，其特征在于，包括：词向量训练单元和句向量生成配置单元，

所述词向量训练单元，用以根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，

所述句向量生成配置单元，用以根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口。

2.根据权利要求1所述的问答处理服务器，其特征在于，根据不同问答场景配置出自然语言的语料的方法进一步为：

针对不同问答场景的训练得到对应的词向量模型，

通过上传/导入与所述问答场景匹配的自然语言的语料。

3.根据权利要求1所述的问答处理服务器，其特征在于，自定义权重配置接口通过句子中词属性进行手动赋权。

4.根据权利要求1所述的问答处理服务器，其特征在于，还包括：问答库预处理单元，用以存放根据不同问答场景问答的答案。

5.根据权利要求1或4所述的问答处理服务器，其特征在于，还包括：匹配单元，用以基于句向量空间匹配出问答不同问答场景问答的答案。

6.根据权利要求1所述的问答处理服务器，其特征在于，还用以：提供一API服务接口。

7.一种客户端，其特征在于，包括：

根据不同问答场景配置出自然语言的语料接口，

以及，提供句向量模型中的自定义权重配置接口，

通过所述语料训练得到词向量模型，再根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，

在所述客户端通过web服务器，提供共上述调用的接口。

8.根据权利要求7所述的客户端，其特征在于，通过访问API服务器获得问答处理结果。

9.根据权利要求7所述的客户端，其特征在于，还包括：手动配置接口，用以提供句向量模型中的特定词汇的配置权值接口。

10.一种问答处理的实现方法，其特征在于，包括如下步骤：

根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，

根据所述词向量模型，将所述语料的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口。

一种问答处理服务器、客户端以及实现方法

技术领域

[0001] 本发明涉及计算机软件领域、自然语言处理领域，特别涉及一种问答处理服务器、客户端以及实现方法。

背景技术

[0002] 问答系统指的是这样一种场景：对于一个计算机系统，当用户输入一个自然语言(Natural Language)的问题时，该问答系统能够在预设的问答库中找到和所问问题的意思(涉及自然语言处理)比较贴合的答案并返回。

[0003] 现阶段一般的做法都是将问句通过自然语言模型转化为高维向量空间的向量(即句向量)，然后和系统中的问句进行比较最终给出答案。但对于问答系统的不用的应用场景，生成句向量的算法往往存在差异，不能够动态的适应各种不同的场景，使得最终效果往往不是特别好。

发明内容

[0004] 本发明要解决的技术问题是，提供一种问答处理服务器，通过在现有传统方法的基础上，将训练句向量的各个环节变成可配置的变量，从而实现该系统对不同的场景都具有较好的效果。

[0005] 本发明还提供了客户端，对于有特殊需求的用户能够让其有足够的灵活度来配置整个系统的特性，从而能够更好的满足用户的特殊要求。

[0006] 解决上述技术问题，本发明提供了一种问答处理服务器，包括：词向量训练单元和句向量生成配置单元，

[0007] 所述词向量训练单元，用以根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，

[0008] 所述句向量生成配置单元，用以根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口。

[0009] 更进一步，根据不同问答场景配置出自然语言的语料的方法为：

[0010] 针对不同问答场景的训练得到对应的词向量模型，

[0011] 通过上传/导入与所述问答场景匹配的自然语言的语料。

[0012] 更进一步，自定义权重配置接口通过句子中词属性进行手动赋权。

[0013] 更进一步，服务器还包括：问答库预处理单元，用以存放根据不同问答场景问答的答案。

[0014] 更进一步，服务器还包括：匹配单元，用以基于句向量空间匹配出问答不同问答场景问答的答案。

[0015] 更进一步，服务器还用以：提供一API服务接口。

[0016] 本发明提供了一种客户端，包括：

[0017] 根据不同问答场景配置出自然语言的语料接口，

[0018] 以及，提供句向量模型中的自定义权重配置接口，

[0019] 通过所述语料训练得到词向量模型，再根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，

[0020] 在所述客户端通过web服务器，提供共上述调用的接口。

[0021] 更进一步，通过访问API服务器获得问答处理结果。

[0022] 更进一步，客户端还包括：手动配置接口，用以提供句向量模型中的特定词汇的配置权值接口。

[0023] 本发明提供了一种问答处理的实现方法，包括如下步骤：

[0024] 根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，

[0025] 根据所述词向量模型，将所述语料的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口。

[0026] 本发明的有益效果：

[0027] 1)本发明中的问答处理服务器，由于包括词向量训练单元，可根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，能够根据不同的对话场景配置训练不同的词向量模型，使得模型的准确度更高。由于包括句向量生成配置单元，可根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口，句向量生成算法中的权重可调节，使得整个问答系统的匹配准确率更高。从而本发明中的问答处理服务器，将训练句向量的各个环节变成可配置的变量，从而实现该系统对不同的场景都有比较不错的效果。

[0028] 2)更进一步，由于在客户端提供了提供句向量模型中的自定义权重配置接口，从而对于有特殊需求的使用者能够让其有足够的灵活度来配置整个系统的特性，能够更好的满足使用者的特殊要求。比如，可以通过词的词性，出现频率甚至一些特殊词汇进行手动赋权，从而达到最好的效果和用户体验。

[0029] 3)本发明训练出问答数据模型文件，然后以API服务器的形式向外提供服务，可应用于多种场景、不同客户端的调用。

附图说明

[0030] 图1是本发明一实施例中的问答处理服务器示意图；

[0031] 图2是本发明一实施例中的客户端结构示意图；

[0032] 图3是本发明一实施例中的实现方法流程示意图；

[0033] 图4是本发明一优选实施例中的问答处理服务器结构示意图；

[0034] 图5是图4中的问答处理服务器上的具体操作示意图。

具体实施方式

[0035] 现在将参考一些示例实施例描述本公开的原理。可以理解，这些实施例仅出于说明并且帮助本领域的技术人员理解和实施例本公开的目的而描述，而非建议对本公开的范围的任何限制。在此描述的本公开的内容可以以下文描述的方式之外的各种方式实施。

[0036] 如本文中所述，术语“包括”及其各种变体可以被理解为开放式术语，其意味着“包

括但不限于”。术语“基于”可以被理解为“至少部分地基于”。术语“一个实施例”可以被理解为“至少一个实施例”。术语“另一实施例”可以被理解为“至少一个其它实施例”。

[0037] 请参考图1是本发明一实施例中的问答处理服务器示意图，本实施例中的一种问答处理服务器，包括：词向量训练单元1和句向量生成配置单元2，所述词向量训练单元1，用以根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，所述句向量生成配置单元2，用以根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口。具体地，在词向量训练单元1中需要大量自然语言的语料，计算词向量模型。自然语言理解的问题要转化为机器学习的问题，在自然语言处理(NLP，Natural Language Processing)中把每个词表示为一个很长的向量。这个向量的维度是词表大小，其中绝大多数元素为0，只有一个维度的值为1，这个维度就代表了当前的词。词向量的训练可包括但不限于，word2vec算法或者GLOVE算法。该些算法的具体实现都已经有相应的开源项目如gensim提供了。做为问答处理服务器只需要调用这些项目的接口即可。进一步，在所述词向量训练单元1中训练好的词向量模型其实是一个记录了所有的单词到一个多维空间的点的文件，文件中每一行就是{单词，向量}的格式。所述词向量模型的主要目的是将自然语言中的词汇映射到一个多维空间。具体地，整个映射的过程是在词向量模型文件中找到以目标词开头的一行，则该行的向量即为该单词映射到的向量。词汇在空间中对应的向量之间能够在一定程度上表达词汇之间表意的联系。

[0038] 在一些实施例中，使用word2vec算法中的CBOW(CBoW模型Continuous Bag-of-Words Model)计算词向量模型。

[0039] 在一些实施例中，使用word2vec算法中的Skip-gram计算词向量模型。Skip-gram模型的本质是计算输入word的input vector与目标word的output vector之间的余弦相似度，并进行softmax归一化。

[0040] 在一些实施例中，使用GLOVE算法计算词向量模型。

[0041] 作为本实施例中的优选，根据不同问答场景配置出自然语言的语料的方法进一步为：针对不同问答场景的训练得到对应的词向量模型，通过上传/导入与所述问答场景匹配的自然语言的语料。在实践使用时，针对不同场景的词向量空间的训练，若使用更加符合当前场景的语料，效果会有非常大的提升。所述上传/导入包括但不限于，用户上传语料库的内容或者导入语料包。比如，采用本实施例中的问答处理服务器的需求是基于问答系统是语言组织严谨的知识型问答系统，若使用普通聊天语料生成的词向量模型的效果就非常差，而使用接近于知识介绍、百科之类严谨文本做为语聊效果就会更好。所以通过上传/导入专业的语料文本，能够提升匹配回答的正确率。由于词向量的训练的语料是可以进行配置的，所以用户可以通过上传/导入更加符合其场景的语料来达到更好的效果如下表1。

[0042] 表1

[0043]

场景	语料	方式
零售场景	沃尔玛百科语料	语料上传
导购场景	日常购物语料	语料API
虚拟现实场景	VR语料	语料上传

SUV汽车场景	汽车选购语料	语料导入
比特币场景	区块链语料	语料上传
探月场景	航空航天语料	语料导入

[0044] 在本实施例中的所述句向量生成配置单元2，基于所述词向量训练单元1的生成结果，将所述语料中的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口。具体地，基于词向量的生成结果，对于一个自然语言的句子，并将其转换为向量空间中的一个向量。本领域技术人员能够明了，一种可实施的方法是对于句子中的每一个词，计算其词向量，在通过某种方式加权称为整个句子的句向量。

[0045] 在一些实施例中的具体的实现方式为：

[0046] 若自然语言的句子s通过分词可以分为n个词，word1、word2……wordn，通过所述词向量训练单元1中的词向量模型，

[0047] 该些词可以对应成n个向量：v1、v2……vn，在句向量生成模块会生成一个用户可以自行定义的权值函数W(word)，则最终的句向量就等于vi*W(wordi)对i＝1到n相加。本实施例中可以通过提供该句向量模型中的自定义权重配置接口，可以通过词的词性，出现频率甚至一些特殊词汇进行手动赋权即所述W(word)，从而达到最好的效果。

[0048] 作为本实施例中的优选，自定义权重配置接口通过句子中词属性进行手动赋权。

[0049] 请参考图2是本发明一实施例中的客户端结构示意图，本实施例中的一种客户端，包括：根据不同问答场景配置出自然语言的语料接口5，以及，提供句向量模型中的自定义权重配置接口6，通过所述语料训练得到词向量模型，再根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，在所述客户端通过web服务器，提供共上述调用的接口。由于在服务器中包括词向量训练单元，可根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，能够根据不同的对话场景配置训练不同的词向量模型，使得模型的准确度更高。此外，由于在服务器中还包括句向量生成配置单元，可根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口，句向量生成算法中的权重可调节，使得整个问答系统的匹配准确率更高。从而本发明中的问答处理服务器，将训练句向量的各个环节变成可配置的变量，从而实现该系统对不同的场景都有比较不错的效果。在客户端通过短链接可访问上述的问答处理服务器。

[0050] 所述客户端可以是，PC、安卓、iPhone、WP、iPad、Mac等客户端。也可以基于HTML5二次开发的访问窗口。用户通过所述客户端，能够根据不同问答场景配置出自然语言的语料同时通过提供该句向量模型中的自定义权重配置接口可以进行权重调整。

[0051] 作为本实施例中的优选，通过访问API服务器获得问答处理结果。通过调用API接口，获取问答问题的答案/回复。

[0052] 作为本实施例中的优选，客户端还包括：手动配置接口，用以提供句向量模型中的特定词汇的配置权值接口。

[0053] 请参考图3是本发明一实施例中的实现方法流程示意图，本实施例中的实现方法包括：

[0054] 步骤S100根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，

[0055] 步骤S101根据所述词向量模型，将所述语料的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口

[0056] 请参考图4和图5，本实施例中的问答处理服务器，其包括：词向量训练单元1和句向量生成配置单元2，所述词向量训练单元1，用以根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型，所述句向量生成配置单元2，用以根据所述词向量模型，将所述语料中的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口。作为本实施例中的优选，问答处理服务器还包括：问答库预处理单元3，用以存放根据不同问答场景问答的答案。作为本实施例中的优选，问答处理服务器还包括：匹配单元4，用以基于句向量空间匹配出问答不同问答场景问答的答案。通过词向量训练单元1、句向量生成配置单元2以及问答库预处理单元3，训练出问答数据模型文件，在提供一API(调用接口)服务接口向外提供服务，可以基于微信、qq等聊天软件。

[0057] 应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列(PGA)，现场可编程门阵列(FPGA)等。

[0058] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

[0059] 总体而言，本公开的各种实施例可以以硬件或专用电路、软件、逻辑或其任意组合实施。一些方面可以以硬件实施，而其它一些方面可以以固件或软件实施，该固件或软件可以由控制器、微处理器或其它计算设备执行。虽然本公开的各种方面被示出和描述为框图、流程图或使用其它一些绘图表示，但是可以理解本文描述的框、设备、系统、技术或方法可以以非限制性的方式以硬件、软件、固件、专用电路或逻辑、通用硬件或控制器或其它计算设备或其一些组合实施。

[0060] 此外，虽然操作以特定顺序描述，但是这不应被理解为要求这类操作以所示的顺序执行或是以顺序序列执行，或是要求所有所示的操作被执行以实现期望结果。在一些情形下，多任务或并行处理可以是有利的。类似地，虽然若干具体实现方式的细节在上面的讨论中被包含，但是这些不应被解释为对本公开的范围的任何限制，而是特征的描述仅是针对具体实施例。在分离的一些实施例中描述的某些特征也可以在单个实施例中组合地执行。相反对，在单个实施例中描述的各种特征也可以在多个实施例中分离地实施或是以任何合适的子组合的方式实施。

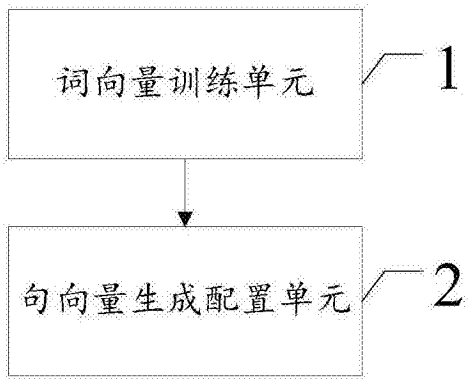

词向量训练单元
1
句向量生成配置单元
2

图1

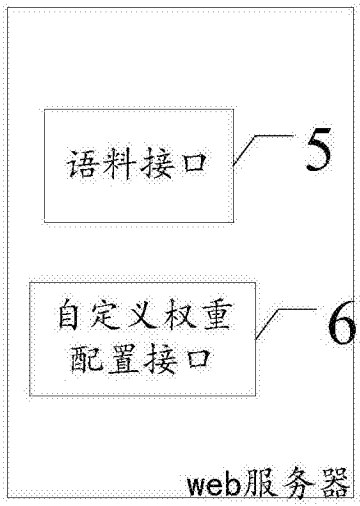

语料接口
5
自定义权重配置接口
6
web服务器

图2

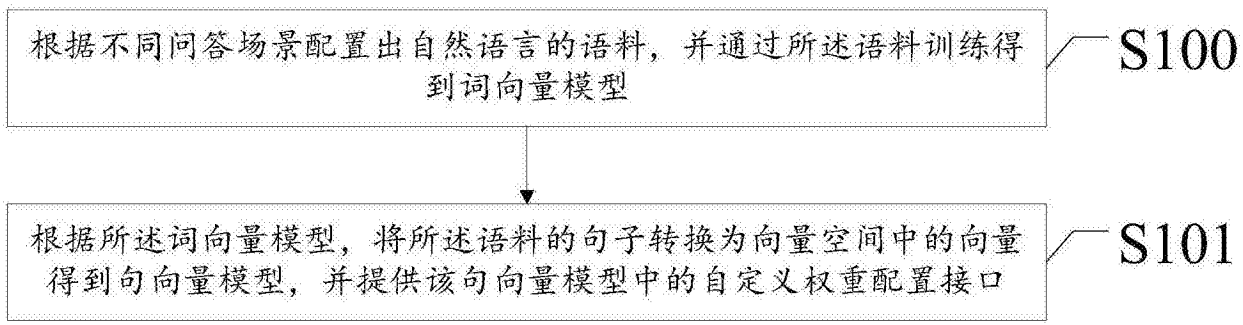

根据不同问答场景配置出自然语言的语料，并通过所述语料训练得到词向量模型
S100
根据所述词向量模型，将所述语料的句子转换为向量空间中的向量得到句向量模型，并提供该句向量模型中的自定义权重配置接口
S101

图3

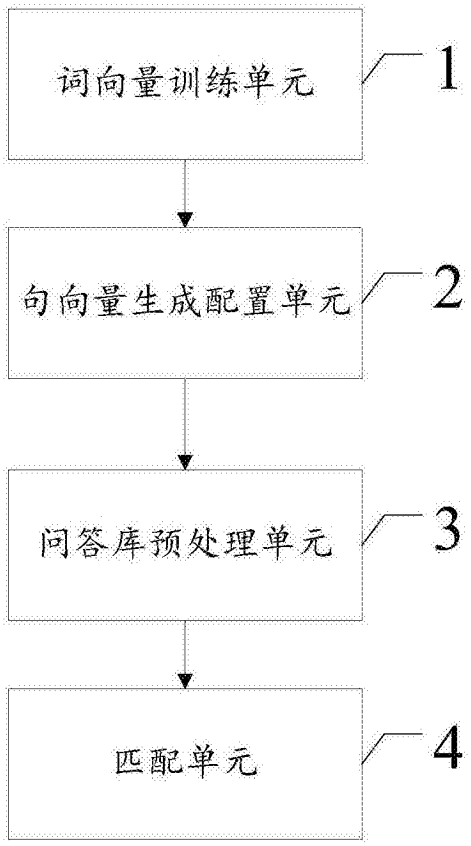


图4

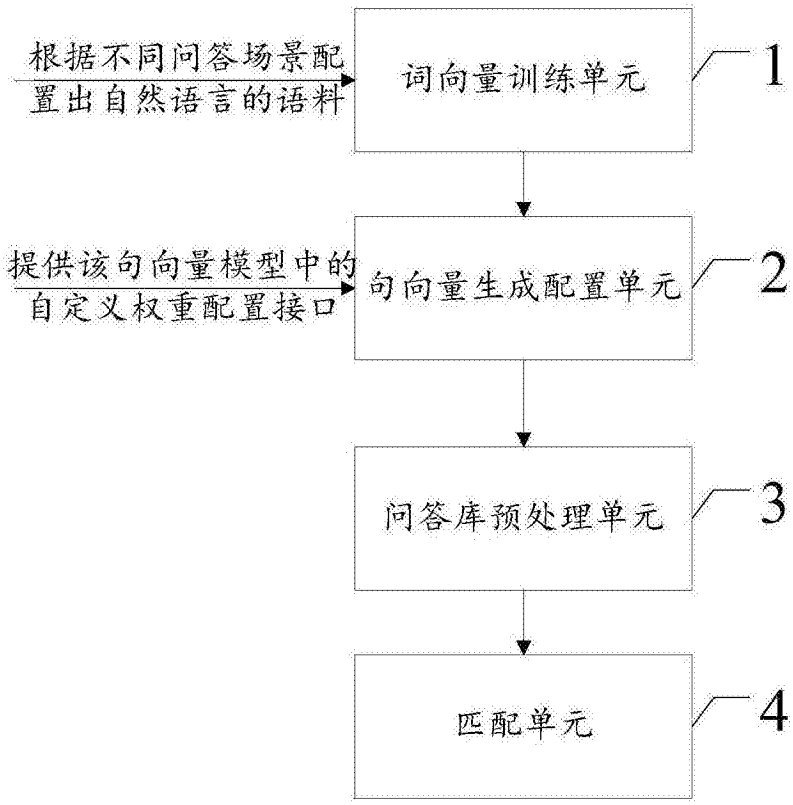

根据不同问答场景配置出自然语言的语料
词向量训练单元
1
提供该句向量模型中的自定义权重配置接口
句向量生成配置单元
2
问答库预处理单元
3
匹配单元
4

图5