(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 113505209 A
(43)申请公布日 2021.10.15

(21)申请号 202110778221.1

(22)申请日 2021.07.09

(71)申请人 吉林大学
地址 130000 吉林省长春市前进大街2699号

(72)发明人 刘露 李春磊 彭涛 包铁

(74)专利代理机构 北京慕达星云知识产权代理事务所(特殊普通合伙) 11465
代理人 符继超

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/36(2019.01)
G06F 16/35(2019.01)
G06F 40/289(2020.01)

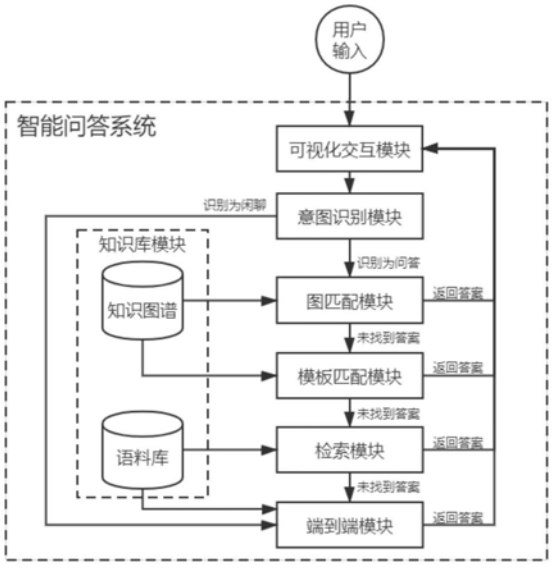
权利要求书3页 说明书9页 附图5页

(54)发明名称

一种面向汽车领域的智能问答系统

(57)摘要

本发明公开了一种面向汽车领域的智能问答系统,包括:知识库模块、可视化交互模块、意图识别模块、图匹配模块、模板匹配模块、检索模块和端到端模块;知识库模块存储有汽车领域的知识图谱和语料库;当用户输入问题后,对用户输入内容进行判断,根据判断出的不同的用户目的,调用系统的相应模块进行处理,得到问题的答案。将用户目的分为汽车领域提问和闲聊两类,针对汽车领域提问,使用基于汽车领域知识图谱的问答方法得到问题回答;针对闲聊则使用基于深度学习的端到端模块,生成回答。本发明能够提高分类精度,并能准确识别用户意图。



CN 113505209 A

1.一种面向汽车领域的智能问答系统,其特征在于,包括:知识库模块、可视化交互模块、意图识别模块、图匹配模块、模板匹配模块、检索模块和端到端模块;

所述知识库模块用于存储汽车领域的知识图谱和语料库;

所述可视化交互模块用于供用户输入问题并反馈问题答案;

所述意图识别模块用于判断用户输入问题类型为汽车领域问题或闲聊问题,将汽车领域问题输入至所述图匹配模块,将闲聊问题输入至所述端到端模块;

所述图匹配模块用于利用图匹配方法将汽车领域问题与所述知识库模块中的知识图谱进行匹配,若匹配成功,则将问题答案反馈至所述可视化交互模块;若匹配失败,则将汽车领域问题输入至所述模板匹配模块;

所述模板匹配模块,用于利用模板匹配方法将汽车领域问题与所述知识库模块中预先定义的问题模板进行匹配,若匹配成功,则将问题答案反馈至所述可视化交互模块;若匹配失败,则将汽车领域问题输入至所述检索模块;

所述检索模块用于在所述语料库中检索汽车领域问题的答案,若检索成功,则将问题答案反馈至所述可视化交互模块;若检索失败,则将汽车领域问题输入至所述端到端模块;

所述端到端模块用于根据预先构建的深度学习模型对汽车领域问题或闲聊问题进行识别,并生成问题答案,将问题答案反馈至所述可视化交互模块。

2.根据权利要求1所述的一种面向汽车领域的智能问答系统,其特征在于,所述知识图谱包括:汽车领域相关的结构化数据加工处理后生成的三元组知识;所述语料库至少包括闲聊语料、数据集以及汽车领域相关的非结构化文本数据。

3.根据权利要求1所述的一种面向汽车领域的智能问答系统,其特征在于,所述可视化交互模块为基于Django框架的web交互界面或基于itChat的微信交互界面。

4.根据权利要求1所述的一种面向汽车领域的智能问答系统,其特征在于,所述意图识别模块用于利用预先训练好的FastText文本分类模型对用户当前的输入问题进行标签化和分词操作,判断输入问题的类型为汽车领域问题或闲聊问题。

5.根据权利要求1所述的一种面向汽车领域的智能问答系统,其特征在于,所述图匹配模块包括:

字典模块,用于预先从所述知识图谱中提取并构建不同类型的字典;所述字典包括:实体字典、关系字典、属性字典、属性值字典、以及预先人工生成类型字典、疑问词字典和停用词字典;

字典树模块,用于根据各个所述字典的内容对应生成字典树;

预处理模块,用于将用户输入的问题进行去除标点符号和英文大小写转换处理;

匹配模块,用于对预处理后的用户输入的问题按照字符顺序从根节点出发向下匹配所述字典树的各个节点,直至无法匹配为止;

分词及词性标注模块,用于将所述匹配模块输出的匹配结果进行分词,并对分词结果进行词性标注;

依存树构建模块,用于分析所述分词结果中各部分之间的语义关联,并根据语义关联建立依存树;

节点类型判断模块,用于在各个类型的所述字典中遍历所述依存树的每个节点,确定每个节点的类型;如果节点所对应的词出现在停用词字典中,则将该节点设置成停用词节

点;如果节点所对应的词与实体字典或关系字典中的词的编辑距离相似度大于设定阈值,则将该节点设置成实体查询节点或关系查询节点;

代词消解模块,用于根据词性标注结果,找出所述依存树中节点类型为代词的节点,找到距离该节点最近的实体节点,将实体节点对应的词作为代词所指代的具体对象;

意图识别子模块,用于找出所述依存树中的疑问词节点,并找出距离该节点最近的属性节点或关系节点作为意图;

查询图构建模块,用于初始化一个查询图集合,然后遍历从所述依存树中找到的所有实体查询节点,对每个实体查询节点,找出所有距离该实体查询节点最近的其他实体查询节点,得到其他实体查询节点在依存树中的最短路径,最后将头尾节点和两节点间的路径添加到查询图集合中,得到查询图;

图匹配子模块,用于利用广度优先方式遍历所述查询图中的每个节点,将由所述查询图中每个节点出发的边转换成图数据库查询语句,若能在图数据库中查询出结果则表明这条边匹配成功;如果边匹配成功,则将该边加入到查询子图中;判断查询子图是否是所述查询图的生成子图;如果是生成子图,则说明查询子图与知识图谱匹配成功,将所述意图识别子模块识别出的关系结点所对应的边转换为查询语句,将查询结果作为查询子图的匹配答案,存入到结果集合中;其中,生成子图为包含了查询图中的全部顶点子图;

答案排序模块,用于对图匹配子模块得到的匹配答案进行排序,将所述查询图边的数目与所述查询图全部节点的相似度之和相加算出得分,并将前N个答案作为最终结果。

6.根据权利要求1所述的一种面向汽车领域的智能问答系统,其特征在于,所述模板匹配模块包括:

模板预定义模块,用于预先定义好一些问题的模板,并为每个模板设定触发词,同时设定好待填充的模板槽位,如果在用户输入的问题中识别到了触发词,则认为匹配到了对应的模板;

触发词及关键信息识别模块,用于对用户输入的问题进行分词、语义解析操作,识别其中的触发词及关键信息,以匹配到相应的模板;

模板槽位填充模块,用于根据触发词匹配到的模板,将所述触发词及关键信息识别模块识别到的关键信息填入到预先设定好的模板槽位中,生成完整的模板查询语句;

查询执行模块,用于使用完整的模板查询语句生成图数据库查询语句,并利用图数据库查询语句匹配所述知识图谱,得到问题答案。

7.根据权利要求1所述的一种面向汽车领域的智能问答系统,其特征在于,所述检索模块包括:

索引文件建立模块,用于对语料库中的原始数据进行拆分,将拆分出的数据存入词汇表中,并拆分的过程中得到数据的索引值,根据索引值建立索引文件;

分词模块,用于采用IKAnalyzer分词器按照中文词语对用户输入的问题进行分词,得到问题中的关键词;

检索排序模块,用于采用Lucene全文信息检索引擎在索引文件中搜索分词结果得到的关键词,搜索时将先从词汇表中进行检索,检索成功后再进入到原始数据中进行检索;再根据相关度得分,对不同的搜索结果进行排序,将得分最高的前5个结果作为问题答案,并返回至所述可视化交互模块。

8.根据权利要求1所述的一种面向汽车领域的智能问答系统，其特征在于，所述端到端模块包括：

模型搭建模块，用于采用词嵌入层+双向门控循环单元网络搭建深度学习模型；

词频字典生成模块，用于在所述深度学习模型的训练阶段，将所述语料库作为训练数据集，并对训练数据集依次进行分词和词频统计，生成词频字典；

序列化操作模块，用于根据词频字典生成序列化字典和反序列化字典，并用于利用所述序列化字典对训练数据进行序列化操作，将训练数据的文本转化为数字序列，将训练数据的数字序列输入至所述深度学习模型；或在所述深度学习模型的测试阶段，将用户输入的问题进行序列化操作，将问题的文本转化为数字序列，将问题的数字序列输入所述深度学习模型进行识别；

反序列化操作模块，用于将经所述深度学习模型识别输出的问题答案序列进行反序列化操作，将数字序列转化为文本序列，得到文本形式的问题答案。

9.根据权利要求8所述的一种面向汽车领域的智能问答系统，其特征在于，所述深度学习模型包括编码器和解码器；编码器和解码器均基于词嵌入层+双向门控循环单元网络构建；所述编码器中还引入了注意力机制。

10.根据权利要求8所述的一种面向汽车领域的智能问答系统，其特征在于，所述深度学习模型在训练阶段，通过损失函数计算生成当前输出的回答与真实答案之间的损失值，采用优化算法调节深度学习模型的参数，并结合Teacher Forcing机制，加快所述深度学习模型的收敛速度，直至所述损失值小于设定阈值，则所述深度学习模型的训练结束。

一种面向汽车领域的智能问答系统

技术领域

[0001] 本发明涉及智能技术领域，更具体的说是涉及一种面向汽车领域的智能问答系统。

背景技术

[0002] 现阶段，人工智能无疑是最热门的研究内容之一，作为人工智能一个主要方向的自然语言处理，其应用方式之一就是智能问答。智能问答是指计算机通过分析用户提问，自动回答用户所提出的问题，是一种高级形式的信息服务。

[0003] 近年来，各大互联网公司都在逐步推出和完善自己的智能问答系统，如苹果公司的“Siri”、Google的“Assistant”、亚马逊的“Alexa”、阿里巴巴的“阿里小蜜”、小米的“小爱同学”等。虽然现阶段已有的问答方法已经能够较好的完成人机交互问题，但仍存在诸如语义识别不清、生成错误或无意义的回答、系统响应时间过长等问题。

[0004] 知识图谱本质上是结构化的语义知识库，是一种由节点和边组成的图数据结构。智能问答技术中涉及的语言理解、信息查询、语言组织等诸多重要环节，都需要语言知识、常识知识以及领域知识的指导。而知识图谱非常适合作为外部知识源注入到智能问答技术中，例如，可以利用知识图谱辅助问句理解，借助知识图谱中节点的属性及关系，通过相应技术发现问句中的实体，进而更好理解用户问题。另外，在端到端模型的解码阶段，可以从知识图谱中检索相关的实体作为应答，并将实体应答与文本应答拼接形成回复。

[0005] 目前，在汽车领域中的智能问答系统对用户输入词的词性识别条件约束单一，识别准确性较低，且没有考虑输入问题的语义结构信息以及词语间的依存关系，仅实现浅层语义解析，难以回答较为复杂的问题。

[0006] 因此，如何提供一种引入知识图谱，能够显著提升智能问答的质量和效率的面向汽车领域的智能问答系统，是本领域技术人员亟需解决的技术问题。

发明内容

[0007] 有鉴于此，本发明提供了一种面向汽车领域的智能问答系统，将用户输入问题分为汽车领域提问和闲聊两类，针对汽车领域提问，使用基于汽车领域知识图谱的问答方法得到问题回答；针对闲聊则使用基于深度学习的端到端模块，生成回答，能够提高分类精度，并能准确识别用户意图。

[0008] 为了实现上述目的，本发明采用如下技术方案：

[0009] 一种面向汽车领域的智能问答系统，包括：知识库模块、可视化交互模块、意图识别模块、图匹配模块、模板匹配模块、检索模块和端到端模块；

[0010] 所述知识库模块用于存储汽车领域的知识图谱和语料库；

[0011] 所述可视化交互模块用于供用户输入问题并反馈问题答案；

[0012] 所述意图识别模块用于判断用户输入问题类型为汽车领域问题或闲聊问题，将汽车领域问题输入至所述图匹配模块，将闲聊问题输入至所述端到端模块；

[0013] 所述图匹配模块用于利用图匹配方法将汽车领域问题与所述知识库模块中的知识图谱进行匹配，若匹配成功，则将问题答案反馈至所述可视化交互模块；若匹配失败，则将汽车领域问题输入至所述模板匹配模块；

[0014] 所述模板匹配模块，用于利用模板匹配方法将汽车领域问题与所述知识库模块中预先定义的问题模板进行匹配，若匹配成功，则将问题答案反馈至所述可视化交互模块；若匹配失败，则将汽车领域问题输入至所述检索模块；

[0015] 所述检索模块用于在所述语料库中检索汽车领域问题的答案，若检索成功，则将问题答案反馈至所述可视化交互模块；若检索失败，则将汽车领域问题输入至所述端到端模块；

[0016] 所述端到端模块用于根据预先构建的深度学习模型对汽车领域问题或闲聊问题进行识别，并生成问题答案，将问题答案反馈至所述可视化交互模块。

[0017] 优选的，在上述一种面向汽车领域的智能问答系统中，所述知识图谱包括：汽车领域相关的结构化数据加工处理后生成的三元组知识；所述语料库至少包括闲聊语料、数据集以及汽车领域相关的非结构化文本数据。

[0018] 优选的，在上述一种面向汽车领域的智能问答系统中，所述可视化交互模块为基于Django框架的web交互界面或基于itChat的微信交互界面。

[0019] 优选的，在上述一种面向汽车领域的智能问答系统中，所述意图识别模块用于利用预先训练好的FastText文本分类模型对用户当前的输入问题进行标签化和分词操作，判断输入问题的类型为汽车领域问题或闲聊问题。

[0020] 优选的，在上述一种面向汽车领域的智能问答系统中，所述图匹配模块包括：

[0021] 字典模块，用于预先从所述知识图谱中提取并构建不同类型的字典；所述字典包括：实体字典、关系字典、属性字典、属性值字典、以及预先人工生成类型字典、疑问词字典和停用词字典；

[0022] 字典树模块，用于根据各个所述字典的内容对应生成字典树；

[0023] 预处理模块，用于将用户输入的问题进行去除标点符号和英文大小写转换处理；

[0024] 匹配模块，用于对预处理后的用户输入的问题按照字符顺序从根节点出发向下匹配所述字典树的各个节点，直至无法匹配为止；

[0025] 分词及词性标注模块，用于将所述匹配模块输出的匹配结果进行分词，并对分词结果进行词性标注；

[0026] 依存树构建模块，用于分析所述分词结果中各部分之间的语义关联，并根据语义关联建立依存树；

[0027] 节点类型判断模块，用于在各个类型的所述字典中遍历所述依存树的每个节点，确定每个节点的类型；如果节点所对应的词出现在停用词字典中，则将该节点设置成停用词节点；如果节点所对应的词与实体字典或关系字典中的词的编辑距离相似度大于设定阈值，则将该节点设置成实体查询节点或关系查询节点；

[0028] 代词消解模块，用于根据词性标注结果，找出所述依存树中节点类型为代词的节点，找到距离该节点最近的实体节点，将实体节点对应的词作为代词所指代的具体对象；

[0029] 意图识别子模块，用于找出所述依存树中的疑问词节点，并找出距离该节点最近的属性节点或关系节点作为意图；

[0030] 查询图构建模块，用于初始化一个查询图集合，然后遍历从所述依存树中找到的所有实体查询节点，对每个实体查询节点，找出所有距离该实体查询节点最近的其他实体查询节点，得到其他实体查询节点在依存树中的最短路径，最后将头尾节点和两节点间的路径添加到查询图集合中，得到查询图；

[0031] 图匹配子模块，用于利用广度优先方式遍历所述查询图中的每个节点，将由所述查询图中每个节点出发的边转换成图数据库查询语句，若能在图数据库中查询出结果则表明这条边匹配成功；如果边匹配成功，则将该边加入到查询子图中；判断查询子图是否是所述查询图的生成子图；如果是生成子图，则说明查询子图与知识图谱匹配成功，将所述意图识别子模块识别出的关系结点所对应的边转换为查询语句，将查询结果作为查询子图的匹配答案，存入到结果集合中；其中，生成子图为包含了查询图中的全部顶点子图；

[0032] 答案排序模块，用于对图匹配子模块得到的匹配答案进行排序，将所述查询图边的数目与所述查询图全部节点的相似度之和相加算出得分，并将前N个答案作为最终结果。

[0033] 优选的，在上述一种面向汽车领域的智能问答系统中，所述模板匹配模块包括：

[0034] 模板预定义模块，用于预先定义好一些问题的模板，并为每个模板设定触发词，同时设定好待填充的模板槽位，如果在用户输入的问题中识别到了触发词，则认为匹配到了对应的模板；

[0035] 触发词及关键信息识别模块，用于对用户输入的问题进行分词、语义解析操作，识别其中的触发词及关键信息，以匹配到相应的模板；

[0036] 模板槽位填充模块，用于根据触发词匹配到的模板，将所述触发词及关键信息识别模块识别到的关键信息填入到预先设定好的模板槽位中，生成完整的模板查询语句；

[0037] 查询执行模块，用于使用完整的模板查询语句生成图数据库查询语句，并利用图数据库查询语句匹配所述知识图谱，得到问题答案。

[0038] 优选的，在上述一种面向汽车领域的智能问答系统中，所述检索模块包括：

[0039] 索引文件建立模块，用于对语料库中的原始数据进行拆分，将拆分出的数据存入词汇表中，并拆分的过程中得到数据的索引值，根据索引值建立索引文件；

[0040] 分词模块，用于采用IKAnalyzer分词器按照中文词语对用户输入的问题进行分词，得到问题中的关键词；

[0041] 检索排序模块，用于采用Lucene全文信息检索引擎在索引文件中搜索分词结果得到的关键词，搜索时将先从词汇表中进行检索，检索成功后再进入到原始数据中进行检索；再根据相关度得分，对不同的搜索结果进行排序，将得分最高的前5个结果作为问题答案，并返回至所述可视化交互模块。

[0042] 优选的，在上述一种面向汽车领域的智能问答系统中，所述端到端模块包括：

[0043] 模型搭建模块，用于采用词嵌入层+双向门控循环单元网络搭建深度学习模型；

[0044] 词频字典生成模块，用于在所述深度学习模型的训练阶段，将所述语料库作为训练数据集，并对训练数据集依次进行分词和词频统计，生成词频字典；

[0045] 序列化操作模块，用于根据词频字典生成序列化字典和反序列化字典，并用于利用所述序列化字典对训练数据进行序列化操作，将训练数据的文本转化为数字序列，将训练数据的数字序列输入至所述深度学习模型；或在所述深度学习模型的测试阶段，将用户输入的问题进行序列化操作，将问题的文本转化为数字序列，将问题的数字序列输入所述

深度学习模型进行识别；

[0046] 反序列化操作模块，用于将经所述深度学习模型识别输出的问题答案序列进行反序列化操作，将数字序列转化为文本序列，得到文本形式的问题答案。

[0047] 优选的，在上述一种面向汽车领域的智能问答系统中，所述深度学习模型包括编码器和解码器；编码器和解码器均基于词嵌入层+双向门控循环单元网络构建；所述编码器中还引入了注意力机制。

[0048] 优选的，在上述一种面向汽车领域的智能问答系统中，所述深度学习模型在训练阶段，通过损失函数计算生成当前输出的回答与真实答案之间的损失值，采用优化算法调节深度学习模型的参数，并结合Teacher Forcing机制，加快所述深度学习模型的收敛速度，直至所述损失值小于设定阈值，则所述深度学习模型的训练结束。

[0049] 经由上述的技术方案可知，与现有技术相比，本发明公开提供了一种面向汽车领域的智能问答系统，具有以下有益效果：

[0050] 1、本发明通过深度学习模型来对用户问题进行意图识别，相较于传统方法，深度学习模型分类精度更高，能够提高对用户意图判断的准确性。

[0051] 2、本发明使用的图匹配方法首先将用户问句转化为依存树，然后将依存树结构转化为查询图结构，最后进行知识图谱和查询图的匹配，在知识图谱中得到问题的答案。转化为依存树这一步骤能够充分考虑用户的语义信息，对复杂问题可以做出有效回答，而转化为查询图以及后续的图匹配则能够在保留用户语义信息的同时，利用知识图谱庞大的知识信息准确地得到用户问题的答案。

[0052] 3、本发明的图匹配模块和模板匹配模块依托于知识图谱，需从知识图谱中得到答案，检索模块从语料库中检索答案，端到端模块使用训练好的深度学习模型，直接根据用户输入内容生成回答。四个模块在问答方法和问答依托的知识库上各不相同，互为补充，大大提高了系统回答问题的能力和范围。

附图说明

[0053] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的实施例，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，还可以根据提供的附图获得其他的附图。

[0054] 图1附图为本发明提供的面向汽车领域的智能问答系统的整体结构框图；

[0055] 图2附图为本发明提供的知识图谱的结构示意图；

[0056] 图3附图为本发明提供的web交互界面示例图；

[0057] 图4附图为本发明提供的微信交互界面示例图；

[0058] 图5附图为本发明提供的意图识别模块的识别流程图；

[0059] 图6附图为本发明提供的图匹配模块的图匹配流程图；

[0060] 图7附图为本发明提供的字典树的结构示意图；

[0061] 图8附图为本发明提供的依存树的结构示意图；

[0062] 图9附图为本发明提供的查询图的结构示意图；

[0063] 图10附图为本发明提供的模板匹配模块的匹配流程图；

[0064] 图11附图为本发明提供的检索模块的检索流程图；

[0065] 图12附图为本发明提供的端到端模块的训练及测试流程图。

具体实施方式

[0066] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0067] 如图1所示，本发明实施例公开了一种面向汽车领域的智能问答系统，包括：知识库模块、可视化交互模块、意图识别模块、图匹配模块、模板匹配模块、检索模块和端到端模块；

[0068] 知识库模块用于存储汽车领域的知识图谱和语料库；

[0069] 可视化交互模块用于供用户输入问题并反馈问题答案；

[0070] 意图识别模块用于判断用户输入问题类型为汽车领域问题或闲聊问题，将汽车领域问题输入至图匹配模块，将闲聊问题输入至端到端模块；

[0071] 图匹配模块用于利用图匹配方法将汽车领域问题与知识库模块中的知识图谱进行匹配，若匹配成功，则将问题答案反馈至可视化交互模块；若匹配失败，则将汽车领域问题输入至模板匹配模块；

[0072] 模板匹配模块，用于利用模板匹配方法将汽车领域问题与知识库模块中预先定义的问题模板进行匹配，若匹配成功，则将问题答案反馈至可视化交互模块；若匹配失败，则将汽车领域问题输入至检索模块；

[0073] 检索模块用于在语料库中检索汽车领域问题的答案，若检索成功，则将问题答案反馈至可视化交互模块；若检索失败，则将汽车领域问题输入至端到端模块；

[0074] 端到端模块用于根据预先构建的深度学习模型对汽车领域问题或闲聊问题进行识别，并生成问题答案，将问题答案反馈至可视化交互模块。

[0075] 具体的，知识库模块包含两个部分，知识图谱和语料库。数据主要通过爬虫技术爬取汽车领域垂直领域网站获得，具体包括汽车领域相关的结构化数据及非结构化的文本语料。其中，对结构化数据进行加工处理，生成三元组知识，存入图数据库中，构成知识图谱，如图2所示，知识图谱供图匹配模块及模板匹配模块使用。非结构化文本数据存入关系型数据库中，作为文本语料，用在检索模块以及端到端模块的训练中。语料库中还包含有一些开源的闲聊语料、数据集等，这部分数据主要用于端到端模块和意图识别模块的训练。

[0076] 本发明实施例设计了两种可视化界面进行交互，分别是基于Django框架的web交互界面以及基于itChat的微信交互界面。web交互界面可以让用户输入问题，并将系统生成的回答呈现给用户，如图3所示；微信交互界面通过调用itChat接口，可以将个人微信接入智能问答系统，提供类似聊天机器人的功能，如图4所示。

[0077] 意图识别模块用于利用预先训练好的FastText文本分类模型对用户当前的输入问题进行标签化和分词操作，判断输入问题的类型为汽车领域问题或闲聊问题。其训练和测试过程如图5所示，具体为：

[0078] 1、在文本分类模型的训练过程，使用汽车领域网站问答板块的问题作为正例数

据，使用闲聊语料作为反例数据，对数据进行标签化、分词后，将数据输入FastText文本分类模型进行训练，得到训练好的模型；

[0079] 2、在测试阶段，当用户输入问句时，使用训练好的模型对用户输入进行意图识别，根据分类结果将用户问句提交给图匹配模块或可视化交互模块处理。

[0080] 图匹配模块的对用户输入问题的匹配流程如图6所示，图匹配模块包括：字典模块、字典树模块、预处理模块、匹配模块、分词及词性标注模块、依存树构建模块、节点类型判断模块、代词消解模块、意图识别子模块、查询图构建模块、图匹配子模块和答案排序模块。

[0081] 其中构建字典和构建字典树在问题输入前进行，属于离线工作部分，其他模块均在用户输入问题后进行，属于在线工作部分。具体步骤如下：

[0082] 1、构建字典：字典模块从知识图谱中提取并构建不同类型的字典；字典包括：实体字典、关系字典、属性字典、属性值字典、以及预先人工生成类型字典、疑问词字典和停用词字典。

[0083] 2、构建字典树：字典树模块根据第1步中构建的不同类型的字典中的内容生成字典树，如图7所示，为由字符串“奔驰S级”、“奔驰E级”、“奔腾”、“宝马X5”、“宝马X6”组成的字典树。

[0084] 3、问题预处理：从这里开始进入在线部分，即需要用户的问题输入。当用户问句输入时，预处理模块对输入的问题进行预处理，如去除标点符号，英文转换为小写等。

[0085] 4、基于字典树最长匹配：匹配模块对用户问句进行基于第2步得到的字典树的最长匹配。具体是指从根节点出发按照字符顺序向下匹配字典树的节点，直至无法匹配为止。

[0086] 5、分词及词性标注：为了后续图匹配方法的实施，需要对用户问句进行分词，过程中分词及词性标注模块将第4步的匹配结果作为额外的分词字典传入分词器，提高分词准确度，分词后对分词结果进行词性标注。

[0087] 6、建立依存树：依存树又叫做语义依存分析，依存树构建模块分析句子各个部分之间的语义关联，并将语义关联以树型结构呈现。为第5步得到的结果建立依存树。如图8所示，为问题“奔驰车的排量是多少”对应的依存树。

[0088] 7、判断节点类型：节点类型判断模块在第1步得到的字典中遍历依存树的每个节点，如果节点所对应的词出现在停用词字典中，则相应地将该节点设置成停用词节点；如果节点所对应的词与实体字典或关系字典中的词的编辑距离相似度大于设定阈值，则相应地将该节点设置成实体查询节点或关系查询节点。

[0089] 8、代词消解：利用依存树找出代词所指代的具体对象。如果输入问句中包含代词则执行该步骤，否则跳过。具体方法为：首先分词及词性标注模块得到的词性标注结果找出依存树中节点类型为代词的节点，然后找到距离该节点最近的实体节点，将实体节点对应的词作为代词所指代的具体对象；

[0090] 9、意图识别：用于识别出用户输入问题的意图。具体方法为：意图识别模块首先找出依存树中的疑问词节点，然后找出距离该节点最近的属性节点或者关系节点作为意图；

[0091] 10、构建查询图：查询图构建模块初始化一个查询图集合。首先，遍历从依存树中找到的所有实体查询节点，对每个实体查询节点，找出所有距离该节点最近的其他实体查询节点，然后得到它们在依存树中的最短路径。最后将头尾结点和两节点间的路径添加到

查询图集合中。如图9所示，为问句“奔驰车的排量是多少”对应的依存树转化为查询图的结果；

[0092] 11、图匹配：图匹配模块利用广度优先方式遍历查询图中的节点，将由图中每个节点出发的边转换成图数据库查询语句，若能在数据库中查询出结果则表明这条边匹配成功。如果边匹配成功，则将边加入到查询子图中。判断查询子图是否是查询图的生成子图，生成子图是指子图中包含了查询图中的全部顶点。如果是生成子图，则说明查询子图与知识图谱匹配成功。将步骤9中意图识别得到的关系结点所对应的边转换为查询语句，将查询结果作为查询子图的匹配答案，存入到结果集合中；

[0093] 12、答案排序：答案排序模块对图匹配模块生成的答案进行排序，将查询图边的数目与查询图全部节点的相似度之和相加算出得分，并将前N个答案作为最终结果。

[0094] 模板匹配模块包括：模板预定义模块、触发词及关键信息识别模块、模板槽位填充模块和查询执行模块，各模块的实施步骤如图10所示，具体为：

[0095] 1、预定义模板：利用模板预定义模块预先定义好一些问题的模板，为每个模板设定触发词，同时设定好待填充的模板槽位。如果在用户问句中识别到了触发词，则认为匹配到了对应的模板。例如“[品牌]的车系有哪些?”、“[数字]万到[数字]万的[汽车]有哪些?”，分别对应于询问品牌车系的问题和汽车售价的问题；

[0096] 2、触发词及关键信息识别：利用触发词及关键信息识别模块对用户问句进行分词、语义解析等操作，识别其中的触发词及关键信息，进而匹配到相应的模板，其中关键信息还将用于模板槽位填充。对于问句“奔驰的车系有哪些?”，将识别出触发词“车系”和关键信息“奔驰”；

[0097] 3、模板槽位填充：模板槽位填充模块根据触发词匹配到的模板，将第2步中识别的关键信息填入到预设定好的模板槽位中，从而生成完整的模板查询语句；

[0098] 4、查询语句执行：查询执行模块使用完整的模板查询语句生成图数据库查询语句并执行，在知识图谱中得到问题答案。

[0099] 检索模块包括：索引文件建立模块、分词模块和检索排序模块。检索模块采用Lucene全文信息检索引擎，Lucene是Apache软件基金会发布的一个开放源代码的全文检索引擎工具包。如图11所示，各模块的具体实施步骤如下：

[0100] 1、建立索引文件：索引文件建立模块对语料库中的原始数据进行拆分，将拆分出的数据存入词汇表中，拆分的过程中得到数据的索引值，建立索引文件。在搜索时将先从词汇表中进行检索，检索成功后再进入到原始数据中进行检索；

[0101] 2、分词：分词模块对用户问题进行分词，包括去除停用词，英文字母转换为小写等操作，具体使用IKAnalyzer分词器，它能够按照中文词语进行分词，分词后得到问题中的关键词；

[0102] 3、检索排序：检索排序模块在Lucene中可以设置相关度得分来对不同的结果进行排序，本发明设定为按照单个字段进行排序。在索引文件中搜索分词结果得到的关键词，将得分最高的前5个结果作为答案返回。

[0103] 端到端模块包括：模型搭建模块、词频字典生成模块、序列化操作模块和反序列化操作模块。

[0104] 其中，模型搭建模块用于采用词嵌入层+双向门控循环单元网络搭建深度学习模

型。深度学习模型包括编码器和解码器;编码器和解码器均基于词嵌入层+双向门控循环单元网络构建;编码器中还引入了注意力机制。

[0105] 词频字典生成模块用于在深度学习模型的训练阶段,将语料库作为训练数据集,并对训练数据集依次进行分词和词频统计,生成词频字典。

[0106] 序列化操作模块用于根据词频字典生成序列化字典和反序列化字典,并用于利用序列化字典对训练数据进行序列化操作,将训练数据的文本转化为数字序列,将训练数据的数字序列输入至深度学习模型;或在深度学习模型的测试阶段,将用户输入的问题进行序列化操作,将问题的文本转化为数字序列,将问题的数字序列输入深度学习模型进行识别。

[0107] 反序列化操作模块用于将经深度学习模型识别输出的问题答案序列进行反序列化操作,将数字序列转化为文本序列,得到文本形式的问题答案。

[0108] 在使用端到端模块前,首先需要完成深度学习模型的搭建以及模型的训练,对深度学习模型的训练及测试过程如图12所示,具体如下:

[0109] 1、深度学习模型的搭建:深度学习模型(以下简称模型)的编码器和解码器均使用词嵌入层+双向门控循环单元构建,解码器中引入注意力机制,具体使用Luong注意力机制,注意力权重计算使用点乘方法,模型的优化算法使用Adam优化算法,损失函数使用负对数似然损失。

[0110] 2、在深度学习模型的训练阶段,使用知识库模块中的语料库作为训练数据集。首先对数据集中的数据进行预处理,包括对噪声的过滤,去除数据中的空行以及特殊符号,然后对数据进行分词,设计使用加载停用词字典的jieba模块进行中文分词,将语料库中的语句分成一个个词的形式,便于后续处理。

[0111] 3、分词后对分词结果进行词频统计,生成词频字典,根据词频字典生成序列化字典和反序列化字典。具体来说,词频字典以语料中出现的词语为键,以对应词语出现的次数为值,序列化字典则是选择词频字典中词语数量大于一定阈值的词,以这些词为字典的键,以数字序号作为字典的值,通过序列化字典,即可将文本转化为数字序列。将序列化字典的键值颠倒,即可得到反序列化字典,使用反序列化字典可将数字序列转化为文本。

[0112] 4、使用序列化字典对处理后的训练数据进行序列化操作,将文本转化为数字序列,随后输入模型进行训练。模型的输入为序列化后的问句序列,输出为生成的回答序列。通过损失函数计算生成回答与真实答案之间的损失值,使用优化算法调节模型的参数。训练过程中引入Teacher Forcing机制,加速模型的收敛,当模型的损失值小于设定阈值时,结束模型的训练。相关计算公式如下:

[0113] 注意力权重计算公式:

[0114] $$score\left(h_t,\bar{h}_s\right)=h_t^T\bar{h}_s \tag{1}$$

[0115] 公式中h_t^T表示解码器的隐藏状态,$\bar{h}_s$表示编码器的输出。

[0116] Adam优化算法:

[0117] $$w=w-\frac{\alpha}{\sqrt{s_w+\delta}}v_w \tag{2}$$

[0118] w表示模型的参数，α表示学习率，v_w表示计算衰减后的动量，s_w表示衰减的历史平方梯度的平均，δ是平滑项，用于防止除数为0。

[0119] 负对数似然损失：

[0120] $loss(p,x) = -\sum x*\log(p)$ (3)

[0121] x表示对应样本的真实值，p表示对应样本分类的概率。

[0122] 模型训练完毕后，端到端模块即可开始使用。当用户问句输入时，对用户问句进行预处理和分词操作，对处理后的数据进行序列化，输入模型，模型输出答案序列，将输出序列反序列化即得到文本形式的问题答案。

[0123] 本发明使用深度学习模型进行用户意图识别，提高对用户意图判断的准确性。本发明使用的图匹配方法能够充分利用用户问题中的语义信息，提高回答的质量；本发明将多种问答方法相结合，互为补充，提高了系统回答问题的能力和范围。

[0124] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

[0125] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

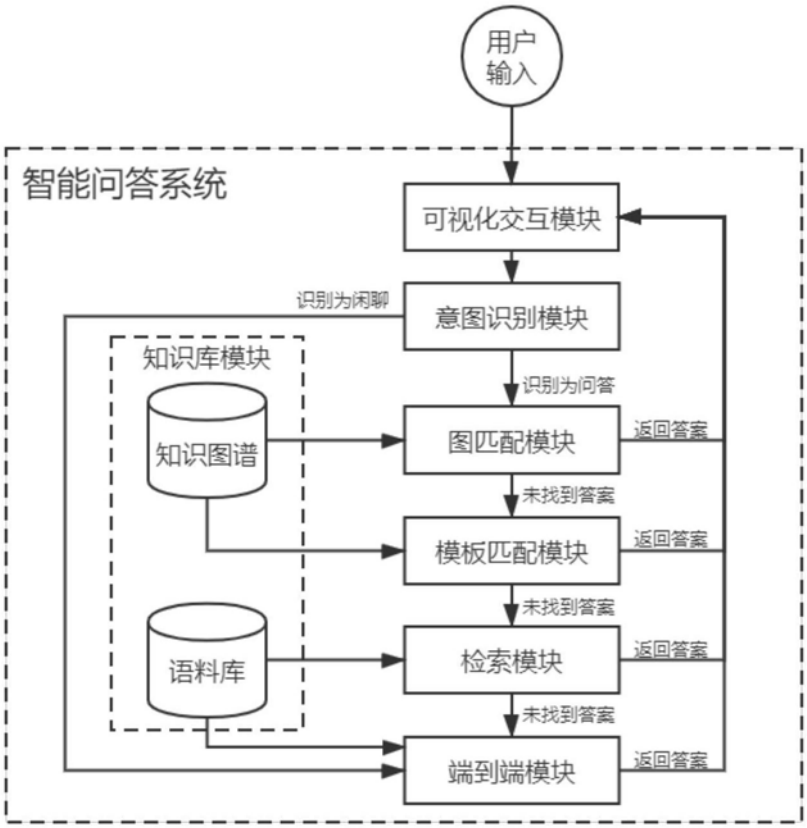

用户输入
智能问答系统
可视化交互模块
意图识别模块
识别为闲聊
识别为问答
知识库模块
知识图谱
图匹配模块
返回答案
未找到答案
模板匹配模块
返回答案
未找到答案
语料库
检索模块
返回答案
未找到答案
端到端模块
返回答案

图1

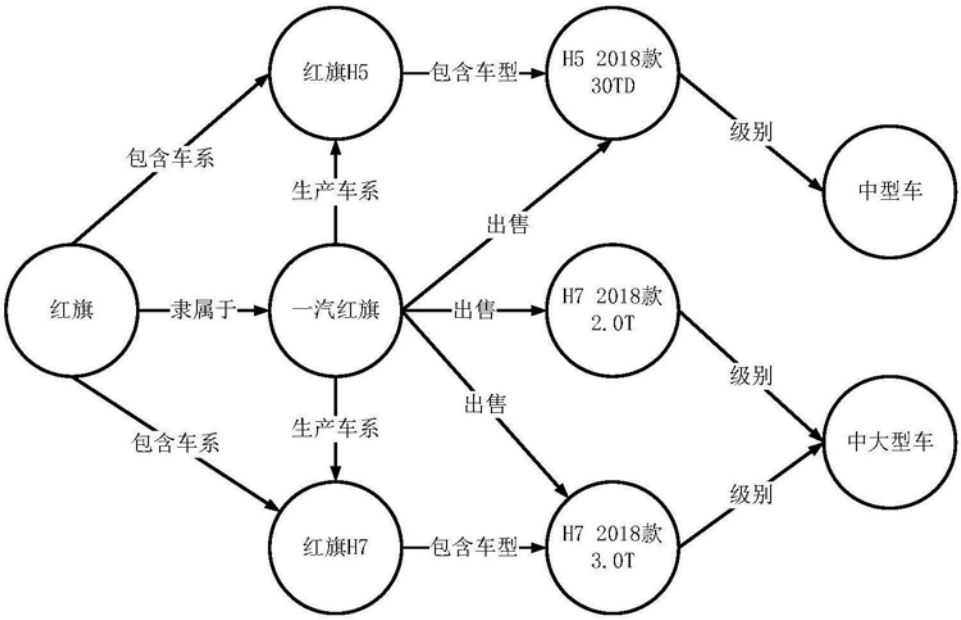

红旗H5
包含车型
H5 2018款 30TD
级别
中型车
包含车系
生产车系
出售
红旗
隶属于
一汽红旗
出售
H7 2018款 2.0T
级别
出售
中大型车
包含车系
生产车系
级别
红旗H7
包含车型
H7 2018款 3.0T

图2


http://127.0.0.1:8000/qa/question
问题：
红旗h5价格多少
搜索
答案：红旗H5的价格是14.58-19.08万

图3


文件传输助手
文件传输助手
文件传输助手

图4

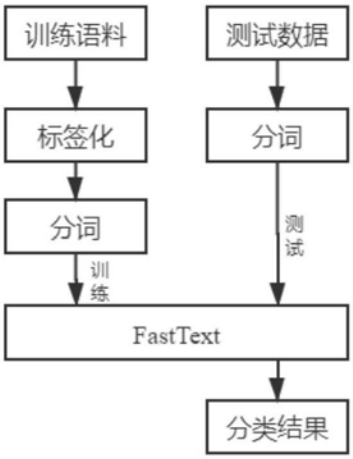

训练语料
测试数据
标签化
分词
分词
训练
测试
FastText
分类结果

图5

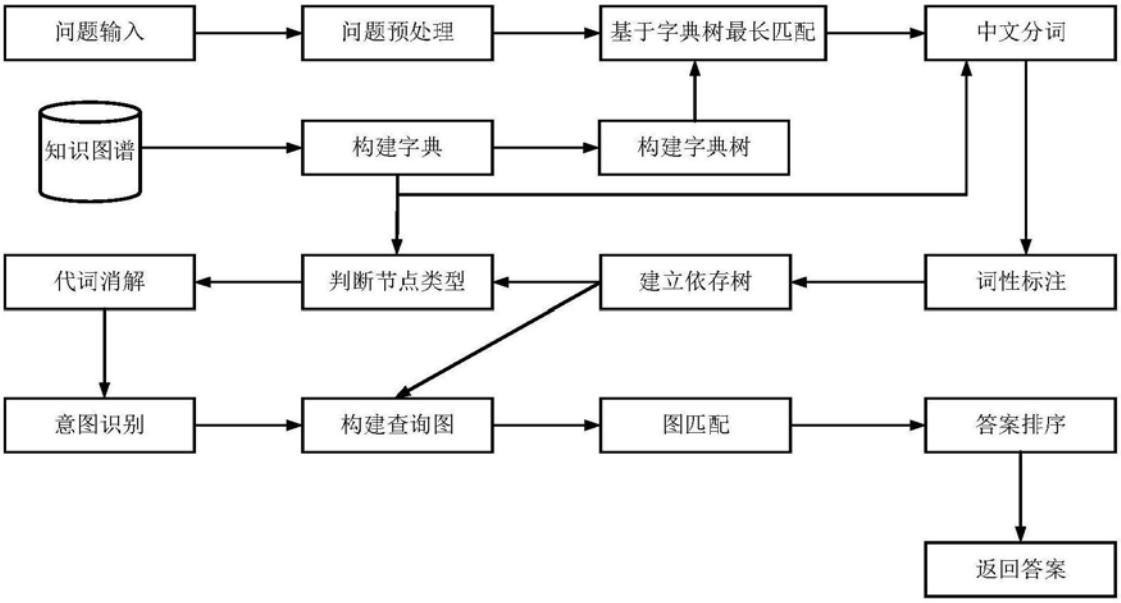

问题输入
问题预处理
基于字典树最长匹配
中文分词
知识图谱
构建字典
构建字典树
代词消解
判断节点类型
建立依存树
词性标注
意图识别
构建查询图
图匹配
答案排序
返回答案

图6

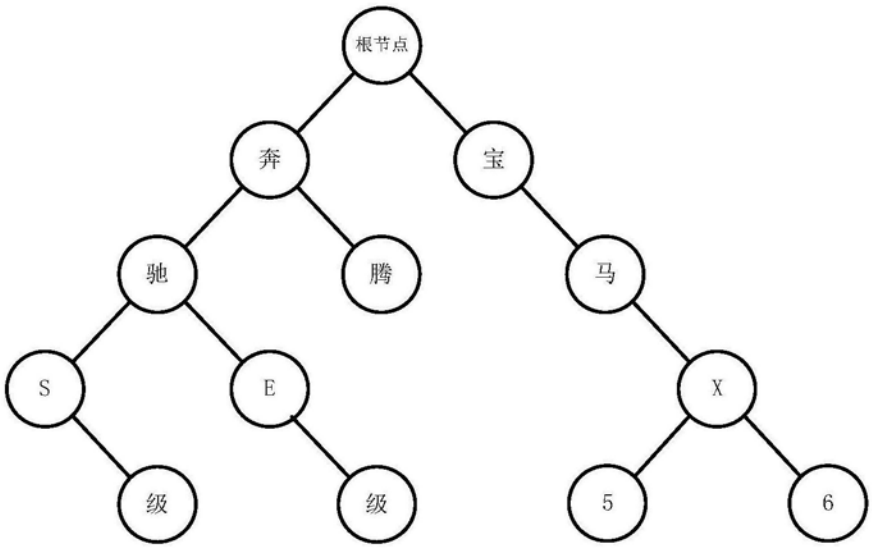

根节点
奔
宝
驰
腾
马
S
E
X
级
级
5
6

图7

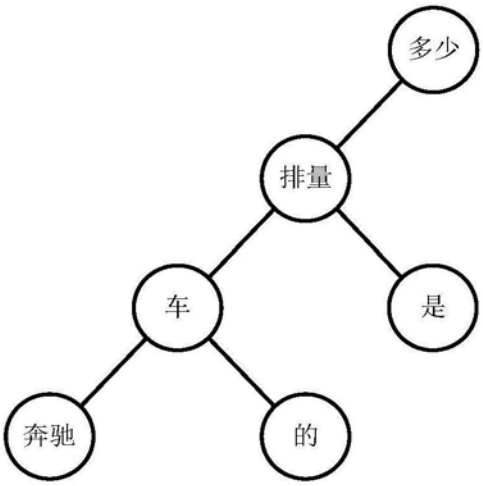

多少
排量
车
是
奔驰
的

图8

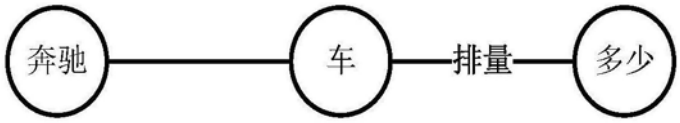

奔驰
车
排量
多少

图9

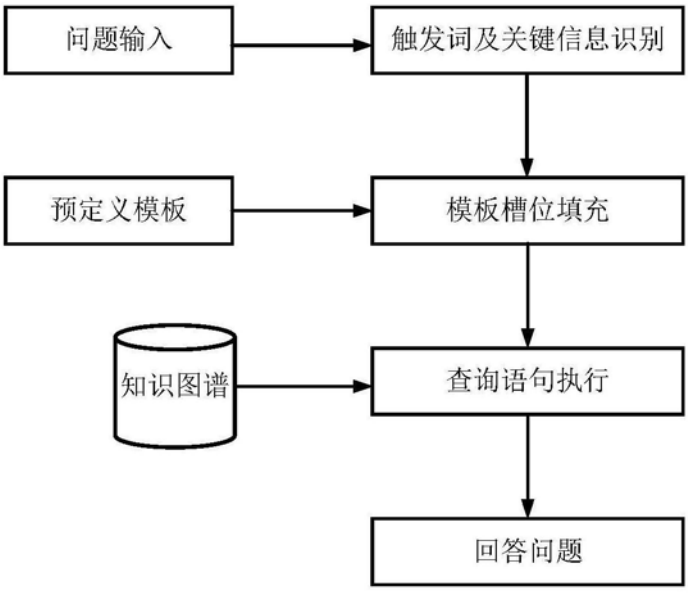

问题输入
触发词及关键信息识别
预定义模板
模板槽位填充
知识图谱
查询语句执行
回答问题

图10

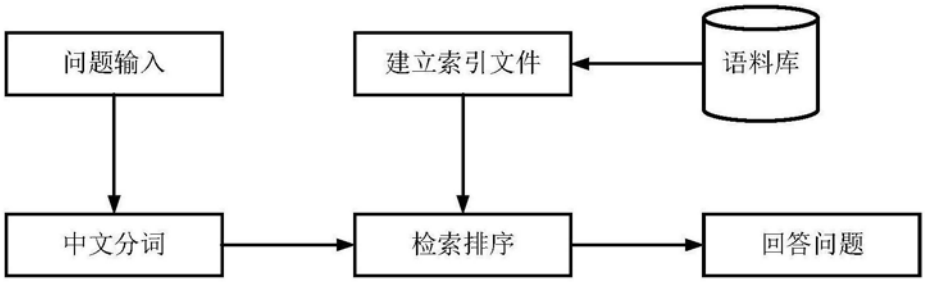

问题输入
建立索引文件
语料库
中文分词
检索排序
回答问题

图11

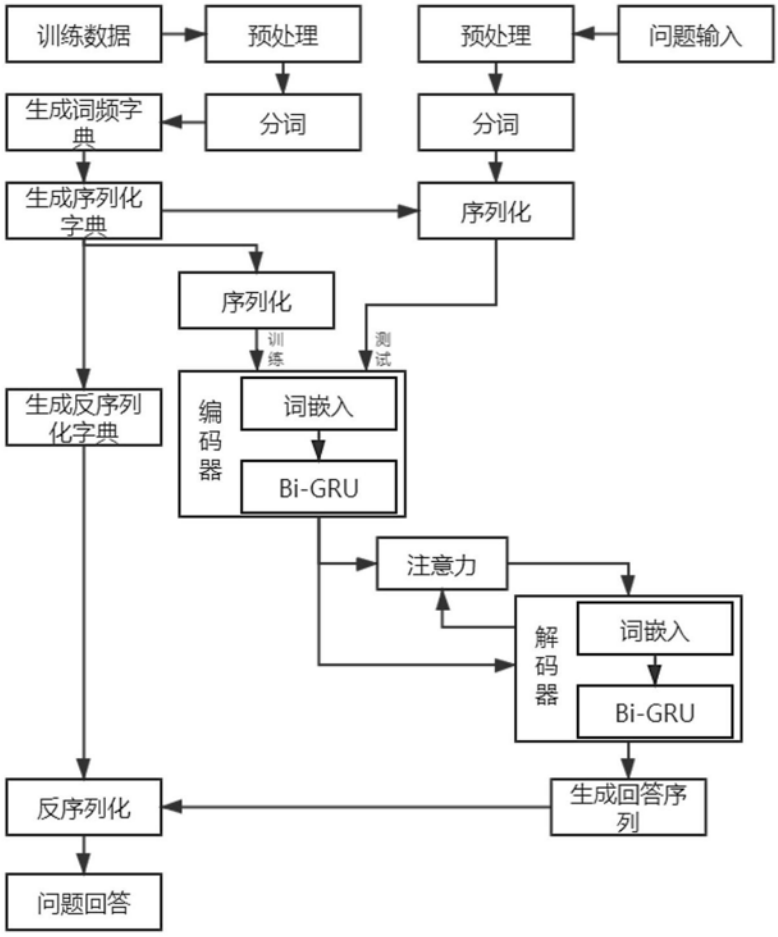

训练数据
预处理
预处理
问题输入
生成词频字典
分词
分词
生成序列化字典
序列化
序列化
训练
测试
生成反序列化字典
编码器
词嵌入
Bi-GRU
注意力
解码器
词嵌入
Bi-GRU
反序列化
生成回答序列
问题回答

图12