(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 118152547 B
(45)授权公告日 2024.08.09

(21)申请号 202410578260.0

(22)申请日 2024.05.11

(65)同一申请的已公布的文献号
申请公布号 CN 118152547 A

(43)申请公布日 2024.06.07

(73)专利权人 青岛网信信息科技有限公司
地址 266000 山东省青岛市崂山区松岭路169号软件外包中心202、216室

(72)发明人 周书田 于海洋 王炳文 彭晓彬

(74)专利代理机构 武汉聚信汇智知识产权代理有限公司 42258
专利代理师 刘丹

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 40/211(2020.01)
G06F 40/284(2020.01)
G06F 16/36(2019.01)

(56)对比文件
CN 114547342 A,2022.05.27
CN 116127095 A,2023.05.16

审查员 岳孟果

权利要求书3页 说明书11页 附图1页

(54)发明名称

一种根据提问者理解能力的机器人回答方法、介质及系统

(57)摘要

本发明提供了一种根据提问者理解能力的机器人回答方法、介质及系统,属于人工智能技术领域,该根据提问者理解能力的机器人回答方法包括以下步骤:收集用户提出问题的文本,并提取专业术语,得到专业术语集、以及专业句集;提取类专业术语,得到类专业术语集、以及类专业句集;对专业句集、类专业句集分析语气,判断用户是否提问,且将专业句集、所述类专业句集进行二元划分后,计算用户文本对专业术语的应用指数;根据应用指数,筛选对应的知识数据集,以及对提问者的理解能力进行分级,并将对应级别与新问题合并作为prompt,得到回答文本;本发明能够根据提问者理解能力输出提问者理解的回答。

CN 118152547 B

1.一种根据提问者理解能力的机器人回答方法,其特征在于,包括以下步骤:

S10、获取提问者与机器人的历史互动记录,收集用户提出的文本记录,记作用户文本;

S20、对所述用户文本基于预先设定好的知识库进行专业术语提取,得到专业术语集、以及专业术语所在语句集,记为专业句集;

S30、采用大语言模型根据预先设定好的知识库进行类专业术语提取,得到类专业术语集、以及类专业术语所在语句集,记为类专业句集;

S40、根据历史互动记录,对所述专业句集、所述类专业句集进行语气分析,判断用户是否提问,且将所述专业句集、所述类专业句集划分为已知专业句集、未知专业句集、已知类专业句集、未知类专业句集;

S50、根据所述专业术语集、所述类专业术语集、所述已知专业句集、所述未知专业句集、所述已知类专业句集、所述未知类专业句集,计算用户文本对专业术语的应用指数;

S60、根据计算得到的专业术语的应用指数,在所述知识库中筛选对应理解能力级别的知识数据集,作为机器人回答提问者的知识库参考;

S70、根据计算得到的专业术语应用指数,对提问者的理解能力进行分级,并将对应的理解能力级别与提问者的新提问合并作为prompt,输入大语言模型,得到的输出作为机器人的回答文本;

其中,用户文本词数表示为N;专业术语集合为T,术语数量为|T|,则专业术语密度为:

$$d = \frac{|T|}{N};$$

专业句子集合为S,句子数量为|S|;

专业句覆盖率为:

$$r = \frac{|S|}{N};$$

未知专业句子集合$Q_{unknown}$,该集合句子数量为$N_{unknown}$;

则未知句占比为:

$$p = \frac{N_{unknown}}{|S|};$$

专业术语应用指数I的计算公式为:

$I=w_1d+w_2r-w_3p$;

其中w_1,w_2,w_3为加权系数;

其中,所述步骤S10具体包括:

设置互动记录的收集范围;

从数据库中提取该用户的互动记录文本,记录以UTF-8格式存储;

对文本记录进行清洗预处理;

构建用户提问记录语料库;

构建机器人回答记录语料库;

其中,所述步骤S20具体包括:

建立人工标注的专业术语知识库;

利用N元语法模型提取候选专业术语;

在知识库中查找验证专业术语;

判断每个专业术语在文本中的上下文句子;

构成专业术语及上下文句子集合;

其中,所述步骤S30具体包括:

训练识别类专业术语的文本分类模型;

使用文本分类模型对用户文本进行类专业术语识别;

判断识别出的类专业术语所在句子;

计算句子与类专业术语向量的相似度;

剔除相似度较低句子;

其中,所述步骤S40具体包括:

构建表达不同语气的语气词库;

利用语气词库判断问句;

在专业句集和类专业句集中识别问句;

统计术语对应的问句数量,划分已知与未知集合,已知集合为用户陈述句占比大于阈值的集合,所述未知集合为用户疑问句占比大于阈值的集合;

经语气词库与统计分析,判断语气词在句子中的占比;

其中,训练识别类专业术语的文本分类模型,具体是:基于BERT预训练语言模型,训练识别类专业术语的文本分类模型,构建包含真实专业术语和非专业术语的训练数据集,带标注类别,利用迁移学习的技术进行模型训练;使用训练得到的文本分类模型对用户文本进行类专业术语识别,模型对每个词语和短语进行判断,判定为类专业术语或非类专业术语,得到一组类专业术语的集合。

2.根据权利要求1所述的一种根据提问者理解能力的机器人回答方法,其特征在于,所述步骤S50具体包括:

定义计算用户输入的文本词语和句法信息的专业术语应用指数计算公式;

计算用户文本中的专业术语密度、专业句覆盖率、未知专业句占比三个因素;

设定权重算法,基于三因素计算指数。

3.根据权利要求2所述的一种根据提问者理解能力的机器人回答方法,其特征在于,所述步骤S60具体包括:

基于指数区间划分用户知识理解能力级别;

构建知识梯度的文本文档集;

匹配用户指数对应文档子集作为回答知识库。

4.根据权利要求3所述的一种根据提问者理解能力的机器人回答方法,其特征在于,所述步骤S70具体包括:

构建表示不同文本难度的提示样本集;

训练序列到序列或T5模型实现问答映射;

选择用户理解级别对应的提示语句与问题合并成新prompt,输入问答映射模型;

评价并迭代优化输出回答质量;

得到的输出作为对用户问句的结果。

5.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质中存储有程序指

令,所述程序指令运行时,用于执行权利要求1-4任一项所述的一种根据提问者理解能力的机器人回答方法。

6.一种根据提问者理解能力的机器人回答系统,其特征在于,包含权利要求5所述的计算机可读存储介质。

一种根据提问者理解能力的机器人回答方法、介质及系统

技术领域

[0001] 本发明属于人工智能技术领域,具体而言,涉及一种根据提问者理解能力的机器人回答方法、介质及系统。

背景技术

[0002] 语义理解是人工智能领域中的核心技术之一。在人机交互场景中,机器需要识别用户的输入语义,才能生成高质量的响应。近年来,序列到序列(seq2seq)模型及其变体(如Transformer)在机器翻译、对话系统等领域取得了巨大成功。这些模型通过编码器-解码器的框架,建模输入和输出序列之间的语义映射,实现端到端的生成任务。

[0003] 然而,这些seq2seq模型都是数据驱动,它们的理解能力和输出质量高度依赖训练数据的覆盖面。对于那些训练数据中没有覆盖或者量不足的长尾分布领域,模型的生成效果仍然很难达到要求。此外,在实际应用中不同用户对语义理解的要求也有较大差异。对专业知识了解充分的用户,期望机器给出专业性更强的回答;而知识面有限的用户则需要简单易懂的响应结果。

[0004] 当前的Seq2Seq模型主要采用单一的数据集进行训练,所有的用户都共享相同的模型。这样得到的对话系统存在无法适应不同用户的个性化理解要求的问题。

发明内容

[0005] 有鉴于此,本发明提供一种根据提问者理解能力的机器人回答方法、介质及系统,能够根据提问者理解能力输出提问者理解的回答。

[0006] 本发明是这样实现的:

[0007] 本发明的第一方面提供一种根据提问者理解能力的机器人回答方法,其中,包括以下步骤:

[0008] S10、获取提问者与机器人的历史互动记录,收集用户提出的文本记录,记作用户文本;

[0009] S20、对所述用户文本基于预先设定好的知识库进行专业术语提取,得到专业术语集、以及专业术语所在语句集,记为专业句集;

[0010] S30、采用大语言模型根据预先设定好的知识库进行类专业术语提取,得到类专业术语集、以及类专业术语所在语句集,记为类专业句集;

[0011] S40、根据历史互动记录,对所述专业句集、所述类专业句集进行语气分析,判断用户是否提问,且将所述专业句集、所述类专业句集划分为已知专业句集、未知专业句集、已知类专业句集、未知类专业句集;

[0012] S50、根据所述专业术语集、所述类专业术语集、所述已知专业句集、所述未知专业句集、所述已知类专业句集、所述未知类专业句集,计算用户文本对专业术语的应用指数;

[0013] S60、根据计算得到的专业术语的应用指数,在所述知识库中筛选对应理解能力级别的知识数据集,作为机器人回答提问者的知识库参考;

[0014] S70、根据计算得到的专业术语应用指数,对提问者的理解能力进行分级,并将对应的理解能力级别与提问者的新提问合并作为prompt,输入大语言模型,得到的输出作为机器人的回答文本。

[0015] 在上述技术方案的基础上,本发明的一种根据提问者理解能力的机器人回答方法还可以做如下改进:

[0016] 其中,所述步骤S10具体包括:

[0017] 设置互动记录的收集范围;

[0018] 从数据库中提取该用户的互动记录文本,记录以UTF-8格式存储;

[0019] 对文本记录进行清洗预处理;

[0020] 构建用户提问记录语料库;

[0021] 构建机器人回答记录语料库。

[0022] 进一步的,所述步骤S20具体包括:

[0023] 建立人工标注的专业术语知识库;

[0024] 利用N元语法模型提取候选专业术语;

[0025] 在知识库中查找验证专业术语;

[0026] 判断每个专业术语在文本中的上下文句子;

[0027] 构成专业术语及上下文句子集合。

[0028] 进一步的,所述步骤S30具体包括:

[0029] 训练识别类专业术语的文本分类模型;

[0030] 使用文本分类模型对用户文本进行类专业术语识别;

[0031] 判断识别出的类专业术语所在句子;

[0032] 计算句子与了类专业术语向量的相似度;

[0033] 剔除相似度较低句子。

[0034] 进一步的,所述步骤S40具体包括:

[0035] 构建表达不同语气的语气词库;

[0036] 利用语气词库判断问句;

[0037] 在专业句集和类专业句集中识别问句;

[0038] 统计术语对应的问句数量,划分已知与未知集合,已知集合为用户陈述句占比大于阈值的集合,未知集合为用户疑问句占比大于阈值的集合;

[0039] 经语气词库与统计分析,判断语气词在句子中的占比。

[0040] 进一步的,所述步骤S50具体包括:

[0041] 定义计算用户输入文本词语和句法信息的专业术语应用指数计算公式;

[0042] 计算用户文本中的专业术语密度、专业句覆盖率、未知专业句占比三个因素;

[0043] 设定权重算法,基于三因素计算指数。

[0044] 进一步的,所述步骤S60具体包括:

[0045] 基于指数区间划分用户知识理解能力级别;

[0046] 构建知识梯度的文本文档集;

[0047] 匹配用户指数对应文档子集作为回答知识库。

[0048] 进一步的,所述步骤S70具体包括:

[0049] 构建表示不同文本难度的提示样本集;

[0050] 训练序列到序列或T5模型实现问答映射;

[0051] 选择用户理解级别对应的提示语句与问题合并成新prompt,输入问答映射模型;

[0052] 评价并迭代优化输出回答质量;

[0053] 得到的输出作为对用户问句的结果。

[0054] 本发明的第二方面提供一种计算机可读存储介质,其中,所述计算机可读存储介质中存储有程序指令,所述程序指令运行时,用于执行上述的一种根据提问者理解能力的机器人回答方法。

[0055] 本发明的第三方面提供一种根据提问者理解能力的机器人回答系统,其中,包含上述的计算机可读存储介质。

[0056] 相比现有技术,本发明的技术效果主要体现在以下三个方面:

[0057] 1.通过构建领域知识库,改进了模型对长尾问题的语义理解和生成能力。知识库内容丰富,覆盖面广,有效扩展了训练数据集的语义表示。模型输出时也可以检索知识库以丰富回答内容;

[0058] 2.根据用户的理解水平和问句情况,实现了个性化地调整模型提示输入的技术。相比单一固定输入,本方案的prompts设计更加灵活丰富,可以引导模型生成不同难度、风格的输出结果;

[0059] 3.应用大规模预训练语言模型,极大拓展了文本生成的语义表达能力。这类模型对语义信息高度敏感,理解类比能力强,是当前最强大的语义表示框架。本方案充分利用其能力,生成的文本质量明显高于以往基于RNN的Seq2Seq模型。

附图说明

[0060] 为了更清楚地说明本发明实施例的技术方案,下面将对本发明实施例的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性的前提下,还可以根据这些附图获得其他的附图。

[0061] 图1为一种根据提问者理解能力的机器人回答方法的流程图。

具体实施方式

[0062] 为使本发明实施方式的目的、技术方案和优点更加清楚,下面将结合本发明实施方式中的附图,对本发明实施方式中的技术方案进行清楚、完整地描述。

[0063] 如图1所示,是本发明第一方面提供一种根据提问者理解能力的机器人回答方法的第一实施例流程图,在本实施例中,包括以下步骤:

[0064] S10、获取提问者与机器人的历史互动记录,收集用户提出的文本记录,记作用户文本;

[0065] S20、对用户文本基于预先设定好的知识库进行专业术语提取,得到专业术语集、以及专业术语所在语句集,记为专业句集;

[0066] S30、采用大语言模型根据预先设定好的知识库进行类专业术语提取,得到类专业术语集、以及类专业术语所在语句集,记为类专业句集;

[0067] S40、根据历史互动记录,对专业句集、类专业句集进行语气分析,判断用户是否提问,且将专业句集、类专业句集划分为已知专业句集、未知专业句集、已知类专业句集、未知类专业句集;

[0068] S50、根据专业术语集、类专业术语集、已知专业句集、未知专业句集、已知类专业句集、未知类专业句集,计算用户文本对专业术语的应用指数;

[0069] S60、根据计算得到的专业术语的应用指数,在知识库中筛选对应理解能力级别的知识数据集,作为机器人回答提问者的知识库参考;

[0070] S70、根据计算得到的专业术语应用指数,对提问者的理解能力进行分级,并将对应的理解能力级别与提问者的新提问合并作为prompt,输入大语言模型,得到的输出作为机器人的回答文本。

[0071] 针对步骤S10的具体实施方式,可以分为以下几个子步骤:

[0072] 1、设定互动记录的收集范围。具体可以设定最近一定时间段内的互动记录,例如最近1年内的记录。也可以设定最近与该用户的多少次互动记录,例如最近100次互动记录;

[0073] 2、从数据库中提取该用户的互动记录。包括用户提问的原始文本记录,以及机器人给出的回答文本记录。文本记录以UTF-8格式存储;

[0074] 3、对文本记录进行预处理。包括清除标点符号、转换为小写字母等,得到干净的文本记录。这一步主要是为了后续的文本分析做准备;

[0075] 4、构建用户提问记录的语料库。将预处理后的用户提问文本记录整合,按时间顺序构建语料库。这一步得到的语料库反映了该用户在一段时间内的提问习惯和兴趣爱好;

[0076] 5、构建机器人回答记录的语料库。将预处理后的机器人回答文本记录整合,按时间顺序构建语料库。这一步得到的语料库反映了机器人基于该用户提问的回答风格和知识覆盖范围。

[0077] 这一步骤S10的主要作用是收集用户与机器人过去的互动文本记录,为后续分析用户对专业知识的理解程度打下基础。通过构建提问语料库和回答语料库,可以更全面地反映用户的知识结构。

[0078] 针对步骤S20的具体实施方式,可以分为以下几个子步骤:

[0079] 1、建立专业术语知识库。需要人工识别各个专业领域的专业术语,整理建立专业术语知识库。知识库可采用关系数据库存储,每个专业术语记录包括名称、定义、所属专业领域等信息。其中,知识库的建立步骤如下:

[0080] (1)确定知识库的覆盖领域和体系:根据方法的应用场景和目标用户群体,确定知识库需要覆盖的专业领域,比如计算机、医学、法律等;并确定知识点之间的上下级与关联关系,设计出树状的知识本体结构;

[0081] (2)文本数据采集与知识抽取:针对选择的专业领域,通过网络爬虫、信息检索等技术从大量半结构化和非结构化数据源中采集学术论文、新闻报道、百科全书等相关文本;利用自然语言处理技术从中抽取概念、实体以及事实,构建初始知识图谱;

[0082] (3)知识融合与质量控制:将从不同来源获取的重复或相似知识进行去重、关联与融合,消除冲突,并结合领域专家进行审核,删除错误知识,不断优化知识框架与内容质量;

[0083] (4)应用通用模型插件知识:利用预训练语言模型等技术,快速对语言知识网络模型进行微调,使其既具有泛化语义理解能力,又掌握领域所需的专业常识,将其中的知识表

示剥离嵌入知识库,实现知识的适配与补全;

[0084] (5)持续迭代更新:建立知识库的维护与迭代机制,持续使用类似流程不断丰富知识库内容,扩展覆盖面;并监测使用过程中的知识缺失,进行针对性知识补充。

[0085] 2、利用N-gram模型对用户文本进行N元语法分析,提取文本中的N-gram词组作为候选专业术语。N值一般取2-5,即考虑2-5个词组成的候选词组;其中N-Gram是大词汇连续语音识别中常用的一种语言模型,对中文而言,称之为汉语语言模型(CLM,Chinese Language Model)。汉语语言模型利用上下文中相邻词间的搭配信息,可以实现到汉字的自动转换,汉语语言模型利用上下文中相邻词间的搭配信息,在需要把连续无空格的拼音、笔划,或代表字母或笔划的数字,转换成汉字串(即句子)时,可以计算出具有最大概率的句子,从而实现到汉字的自动转换,无需用户手动选择,避开了许多汉字对应一个相同的拼音(或笔划串,或数字串)的重码问题。该模型基于这样一种假设,第N个词的出现只与前面N-1个词相关,而与其它任何词都不相关,整句的概率就是各个词出现概率的乘积。这些概率可以通过直接从语料中统计N个词同时出现的次数得到。常用的是二元的Bi-Gram和三元的Tri-Gram。

[0086] 3、对每个候选专业术语,在专业术语知识库中查找,如果知识库中存在该词组记录,则判定该词组为真正的专业术语,加入专业术语集。

[0087] 4、进一步判断每个专业术语在用户文本中所在的句子,将这些句子提取出来,构成专业术语所在句子的集合,即专业句集。

[0088] 5、对专业句集进行语义分析,剔除与专业术语意义无关的句子。这一步采用词向量等自然语言理解技术判断句子语义。

[0089] 6、得到精炼的专业术语集合以及专业术语所在句子集合。

[0090] 这一步骤S20主要通过N元语法模型抽取候选专业术语,结合人工构建的专业术语知识库判断真伪,从而自动化地从文本中提取出专业术语及其上下文句子,为判断用户对专业知识的理解程度奠定基础。

[0091] 针对步骤S30的具体实施方式,可以分为以下几个子步骤:

[0092] 1、基于BERT等预训练语言模型,训练识别类专业术语的文本分类模型。构建包含真实专业术语和非专业术语的训练数据集,带标注类别,利用迁移学习等技术进行模型训练;

[0093] 2、使用训练得到的文本分类模型对用户文本进行类专业术语识别。模型对每个词语和短语进行判断,判定为类专业术语或非类专业术语,得到一组类专业术语的集合;

[0094] 3、判断每个识别出的类专业术语在用户文本中所在的句子,将这些句子提取出来,构成类专业术语所在句子的集合,即类专业句集;

[0095] 4、采用词向量技术计算每个句子向量的均值,与类专业术语向量的余弦相似度。设定相似度阈值,剔除相似度较低的句子;

[0096] 5、得到精炼的类专业术语集合以及类专业术语所在句子集合。

[0097] 这一步骤S30通过迁移学习方法训练专业术语识别模型,实现从文本中自动提取类专业术语。区别于S20中的N元语法方法,这种方法可以识别新出现的类专业术语。结合向量相似度技术剔除无关句子,为后续判断用户对类专业知识的理解程度奠定基础。

[0098] 针对步骤S40的具体实施方式,可以分为以下几个子步骤:

[0099] 1、构建语气词库。收集表达命令、疑问、感叹等语气的词语和短语，建立语气词库。语气词库存储为关系型数据库，包括语气词及其所表达语气类别等信息；

[0100] 2、利用文本特征提取技术，从用户的历史互动记录中提取问句。针对每个句子，计算其与语气词库中各类别语气词的覆盖率，即句子中语气词数量与总词数的比例。如果覆盖率超过设定阈值，则判断该句子为问句；

[0101] 3、在步骤S20中得到的专业句集和步骤S30中得到的类专业句集中，筛选出被判断为问句的语句，分别构成专业问句集和类专业问句集。对应地可以得到已知陈述句集和类专业陈述句集；

[0102] 4、统计用户历史互动记录中已知专业术语和类专业术语的出现次数，以及这些术语对应问句和陈述句的数量。如果某专业术语的问句数量占比超过阈值，则将这个术语及对应的句子判定为“未知”类别，加入未知专业句集和未知类专业句集；

[0103] 5、经过以上处理，得到用户文本中专业句集和类专业句集的二元分类，即已知和未知两类，表示用户对这些专业知识的理解程度。

[0104] 这一步骤S40通过语气词库判定问句，再结合用户历史记录，实现对专业句集和类专业句集的二元划分，为后续评估用户对专业知识的理解程度打下基础。

[0105] 针对步骤S50的具体实施方式，可以分为以下几个子步骤：

[0106] 1、定义专业术语应用指数计算公式。该指数综合考虑三个因素的计算结果，包括：用户文本中专业术语密度、专业句覆盖率、未知专业句占比；

[0107] 2、专业术语密度指用户文本中专业术语数量与总词数的比例。直接利用词频统计方法计算；

[0108] 3、专业句覆盖率指专业句集中句子数量与用户文本总句子数量的比例。利用句子划分技术获取用户文本句子总数；

[0109] 4、未知专业句占比指未知专业句集中句子数量与专业句集总句子数量的比例。通过步骤S40得到的划分结果计算；

[0110] 5、综合以上三个因素，设定权重加权算法计算专业术语应用指数。指数值范围为0-1，值越大表示用户文本对专业知识的应用程度越高。

[0111] 这一步骤S50通过从词语、句子两个层面评估用户文本对专业知识的使用，定义量化的评价指数，为后续确定用户知识理解能力水平及机器人回答策略提供参考依据。

[0112] 针对步骤S60的具体实施方式，可以分为以下几个子步骤：

[0113] 1、基于专业术语应用指数的数值范围，划分多个用户知识理解能力级别。例如可以划分为初级、中级、高级三个级别。每个级别对应指数值的一个区间范围；

[0114] 2、构建知识库文档集。文本文档表示一定知识量的知识点，文档数量代表知识量。按从低到高的顺序组织文档，形成知识梯度；

[0115] 3、根据用户文本得到的专业术语应用指数值，匹配知识库中不同知识理解能力级别对应的文档子集。如果指数值为初级区间，则选择知识库底层文档构成回答知识库；

[0116] 4、回答知识库文档数量与用户指数值正相关。指数值越高，选择文档子集的上限索引值越大，即知识量级别越高。相应地，文档集子集与用户知识理解能力水平相适应。

[0117] 这一步骤S60基于前面评估得到的定量指数，实现自动化构建与个性化用户知识水平相匹配的机器人回答知识库，为后续生成适合用户理解能力的回答文本提供知识来

源。

[0118] 针对步骤S70的具体实施方式,可以分为以下几个子步骤:

[0119] 1、代表不同文本难度的提示样本数据集。根据语言模型困惑度等指标,手工标注文本难度级别;

[0120] 2、训练seq2seq模型或谷歌T5模型,建立用户提问和回答文本之间的映射模型。模型综合了上下文和知识库支持,实现高质量的问答;

[0121] 3、根据前面步骤提问者的理解能力级别,从提示样本集中选择对应等级的提示语句,与用户问题拼接成新prompt,输入问答映射模型;

[0122] 4、模型输出答案,如果质量不够好,则调节prompt的难度级别,再次生成。评价维度包括语法、语义的正确性,以及答案文本的难度与用户知识水平的匹配程度;

[0123] 5、迭代优化找到最佳prompt,对应的输出作为对用户提问的回答结果。

[0124] 通过上述流程,实现了根据每个用户具体的知识水平状况,动态调整问答模型的提示输入,生成个性化、适配性强的回答文本。

[0125] 这一步骤S70借助大规模预训练语言模型实现对不同知识水平用户的自然语言问答,是本方法的关键输出。

[0126] 如图1所示,是本发明第一方面提供一种根据提问者理解能力的机器人回答方法的第二实施例流程图,在本实施例中,包括以下步骤:

[0127] S10、获取提问者与机器人的历史互动记录,收集用户提出的文本记录,记作用户文本;

[0128] S20、对用户文本基于预先设定好的知识库进行专业术语提取,得到专业术语集、以及专业术语所在语句集,记为专业句集;

[0129] S30、采用大语言模型根据预先设定好的知识库进行类专业术语提取,得到类专业术语集、以及类专业术语所在语句集,记为类专业句集;

[0130] S40、根据历史互动记录,对专业句集、类专业句集进行语气分析,判断用户是否提问,且将专业句集、类专业句集划分为已知专业句集、未知专业句集、已知类专业句集、未知类专业句集;

[0131] S50、根据专业术语集、类专业术语集、已知专业句集、未知专业句集、已知类专业句集、未知类专业句集,计算用户文本对专业术语的应用指数;

[0132] S60、根据计算得到的专业术语的应用指数,在知识库中筛选对应理解能力级别的知识数据集,作为机器人回答提问者的知识库参考;

[0133] S70、根据计算得到的专业术语应用指数,对提问者的理解能力进行分级,并将对应的理解能力级别与提问者的新提问合并作为prompt,输入大语言模型,得到的输出作为机器人的回答文本。

[0134] 在步骤S10中,主要目的是收集用户与机器人的历史互动记录,为后续分析用户对专业知识的理解程度打下基础。其中关键的是构建用户提问文本集合D_{query}以及机器人回答文本集合$D_{response}$。具体实施方式如下:

[0135] 定义最近互动记录的时间范围,例如最近1年内的互动记录。

[0136] 从数据库中检索出用户U在时间范围内的提问记录集合

$Q=\{q_1,\ q_2,\ \ldots,\ q_i\ldots,\ q_n\}$,其中$q_i$表示第$i$条提问语句。

[0137] 从数据库中检索出机器人对应Q中每条提问的回答记录集合$A=\{a_1,\ a_2,\ \ldots,\ a_i\ldots,\ a_n\}$,其中$a_i$表示对应第$i$条提问$q_i$的回答语句。

[0138] 对提问记录集合Q中的每条提问q_i进行文本清洗预处理,得到清洗后的提问集合$Q'=\{q_1',\ q_2',\ \ldots,\ q_n'\}$。文本清洗预处理函数定义为$f_{\text{clean}}(x)$,具体可以包括清除标点符号,皮词去除等操作。

[0139] 对回答记录集合A进行相同的清洗预处理,得到清洗后的回答集合$A'=\{a_1',\ a_2',\ \ldots,\ a_n'\}$。

[0140] 将预处理后的提问集合Q'整合构建用户提问语料库$D_{query}=\bigcup_{i=1}^{n}q_i'$。

[0141] 将预处理后的回答集合A'整合构建机器人回答语料库$D_{response}=\bigcup_{i=1}^{n}a_i'$。

[0142] 在步骤S20中,目标是从用户文本中提取出专业术语及其所在句子,为判断用户对专业知识的理解程度奠定基础。提取过程中采用N元语法模型识别候选专业术语,并结合人工构建的专业术语知识库进行验证。具体实施方式如下:

[0143] 构建专业术语知识库KB_{term},存储格式为关系型数据库。每个专业术语条目包括:

[0144] 术语名称t_{name};

[0145] 术语定义t_{def};

[0146] 术业所属领域t_{domain};

[0147] 输入用户文本D_{input},利用N元语法模型提取候选专业术语:

[0148] 执行N元切分,生成词序列$W=\{w_1,\ w_2,\ \ldots,\ w_m\}$;

[0149] 滑窗扫描,提取所有长度为N的词组作为候选专业术语集合$C=\{c_1,\ c_2,\ \ldots,\ c_k\}$;

[0150] 对每个候选术语c_i,在知识库KB_{term}中搜索,如果存在匹配的术语条目,则判定c_i为真实专业术语,加入集合T,作为最终提取的专业术语集合。

[0151] 对每个识别出的专业术语$t_j\in T$,进一步判断其在原文本D_{input}中的上下文窗口,提取窗口句子构成集合S。

[0152] 对S中每个句子s_i,利用词向量技术计算其与包含专业术语的句子的语义相似度$sim(s_i,\ s_j)$。设定相似度阈值δ,剔除相似度较低的无关句子。

[0153] 最终得到精炼的专业术语集合T以及专业术语上下文句子集合S'。

[0154] 在步骤S30中,关键是训练一个文本分类模型,用于从用户文本中识别出类专业术语。实现上采用基于BERT等预训练语言模型的迁移学习方法。具体实施方式如下:

[0155] 构建训练数据集$D=\{(x_1,\ y_1),\ (x_2,\ y_2),\ \ldots,\ (x_n,\ y_n)\}$,其中$x_i$表示文本片段,$y_i\in[0,1]$表示是否为类专业术语。

[0156] 基于BERT，添加分类层，构建文本分类模型，损失函数为交叉熵损失$L=-\sum y\,log(\hat{y})$。其中$\hat{y}$是模型对样本的分类预测概率。

[0157] 训练模型参数θ，最小化损失函数L。优化算法为Adam等，可以调节超参数进行正则化，防止过拟合。

[0158] 输入用户文本x，利用训练好的模型进行类专业术语判断：$\hat{y}=f_\theta(x)$。其中f_θ表示训练得到的文本分类模型，θ为训练的参数。

[0159] 对每个被判断为类专业术语的文本拆分，提取其上下文句子构成集合S''。

[0160] 对S''中每个句子s_k'，利用词向量技术计算其与包含类专业术语的句子的语义相似度，设定相似度阈值进行句子筛选。

[0161] 最终得到文本分类模型识别出的类专业术语集合T'以及类专业术语上下文句子集合S'''。

[0162] 在步骤S40中，利用语气词库和统计分析的方法，实现对前面步骤中提取出的专业句集合S和类专业句集合S'''的进一步区分，具体实施如下：

[0163] 构建语气词库KB_{mood}，存储各类语气词，格式为关系型数据库。主要包括：

[0164] 疑问语气词$qw=\{qw_1,\ qw_2,\ \ldots\}$；

[0165] 感叹语气词$ew=\{ew_1,\ ew_2,\ \ldots\}$；

[0166] 命令语气词$cw=\{cw_1,\ cw_2,\ \ldots\}$；

[0167] 对专业句集合S中的每句s_i，统计句子中包含的语气词类别比例：

[0168] 疑问词占比：$r_{qw}(s_i)=\frac{|qw\cap s_i|}{|s_i|}$；

[0169] 感叹词占比：$r_{ew}(s_i)=\frac{|ew\cap s_i|}{|s_i|}$；

[0170] 命令词占比：$r_{cw}(s_i)=\frac{|cw\cap s_i|}{|s_i|}$；

[0171] 如果$r_{qw}(s_i)$超过阈值τ，则判断句子s_i为问句，加入疑问句集合Q；如果$r_{ew}(s_i)$超过阈值τ，则判断句子s_i为感叹句，加入感叹句集合Q_e；如果$r_{cw}(s_i)$超过阈值τ，则判断句子s_i为问句，加入命令句集合Q_c。其余构成陈述句集合D_S。类专业句集合S'''也同理处理。

[0172] 统计用户历史记录中，专业术语$\in T$对应的问句数量$n_q(t)$和陈述句数量$n_d(t)$。

[0173] 如果$\frac{n_q(t)}{n_q(t)+n_d(t)}>\sigma$，则判定该术语为“未知”概念，加入集合$T_{unknown}$。每个未知术语对应的问句也划分为未知集合$Q_{unknown}$。

[0174] 至此，获得了用户对已知与未知专业知识的划分结果。

[0175] 在步骤S50中，目标是定义一个专业术语应用指数I，评价用户文本对专业知识的

使用和理解程度。具体方法如下：

[0176] 用户文本词数表示为N。

[0177] 专业术语集合为T，术语数量为$|T|$。则专业术语密度为：

[0178] $$d = \frac{|T|}{N};$$

[0179] 专业句子集合为S，句子数量为$|S|$。专业句覆盖率为：

[0180] $$r = \frac{|S|}{N};$$

[0181] 未知专业句子集合$Q_{unknown}$，该集合句子数量为$N_{unknown}$。则未知句占比为：

[0182] $$p = \frac{N_{unknown}}{|S|};$$

[0183] 专业术语应用指数I的计算公式为：

[0184] $$I = w_1 d + w_2 r - w_3 p;$$

[0185] 其中w_1，w_2，w_3为加权系数。

[0186] 在步骤S60中，需要实现根据前面计算的专业术语应用指数I，动态构建与用户知识水平相匹配的回答知识库。具体方法如下：

[0187] 根据指数I的数值，将用户知识理解能力划分为m个等级，表示为集合$E = \{e_1, e_2, \ldots, e_j \ldots, e_m\}$。

[0188] 构建知识库文档集合$H = \{h_1, h_2, \ldots, h_n\}$，其中文档按知识量从低到高排列。

[0189] 计算出索引区间：

[0190] $$Range(e_j) = [l_j, u_j],\ l_j = \lfloor (j-1)\frac{n}{m} \rfloor,\ u_j = \lfloor j\frac{n}{m} \rfloor;$$

[0191] 其中$Range(e_j)$代表等级e_j对应的文档索引区间。

[0192] 对于应用指数I，计算其对应知识级别$e(I)$，然后在知识库文档集合H中选择子集$H' \subseteq H$，其中每个文档$h_i \in H'$满足$i \in Range(e(I))$。

[0193] 获得的文档子集H'即构成了根据用户水平匹配的回答知识库。

[0194] 在步骤S70中，利用seq2seq模型实现根据不同用户的个性化理解能力对其提问进行自然语言回答。具体实施方式如下：

[0195] 构建多级提示模板集合$P = \{p_1, p_2, \ldots, p_l\}$，根据文本难度进行了静态手工标定。

[0196] 训练序列到序列模型f_θ，输入为用户提问关联提示模板，输出为回答文本。损失函数为生成损失，进行模型训练。

[0197] 对用户u的提问q，获取其前面步骤计算出的知识级别$e(u)$。

[0198] 选择对应的提示模板$p \in P$，拼接序列$[q;\ p]$作为模型输入。

[0199] 模型生成回答$\hat{a} = f_\theta([q;\ p])$。评估回答效果,如果不满意迭代更改提示模板级别,直到生成质量达到要求。

[0200] 得到的$\hat{a}$即为针对该用户个性化水平的问答结果。

[0201] 本发明第二方面提供一种计算机可读存储介质的第一实施例,在本实施例中,计算机可读存储介质中存储有程序指令,程序指令运行时,用于执行上述的一种根据提问者理解能力的机器人回答方法。

[0202] 本发明第三方面提供一种根据提问者理解能力的机器人回答系统的第一实施例,在本实施例中,包含上述的计算机可读存储介质。

[0203] 本发明能够有效解决上述技术问题,主要原理在于:

[0204] 1.构建高质量结构化知识库,为Seq2Seq模型提供外部知识支持,弥补训练数据的不足;

[0205] 2.精细评估用户文本的语义特征,进行知识水平分级,实现个性化回答;

[0206] 3.应用基于 Transformer 等注意力机制的预训练语言模型,其中自注意力结构能高效学习文本的内在语义信息,从而有效表达复杂语义。

[0207] 通过这三者的有机结合,即知识引导、个性化学习和超强表达模型,本方法全面提升了 Seq2Seq 模型的语义理解和生成能力。既能覆盖长尾问题,又适应不同用户,性能显著优于传统技术。这是本方法有效的技术原理所在。

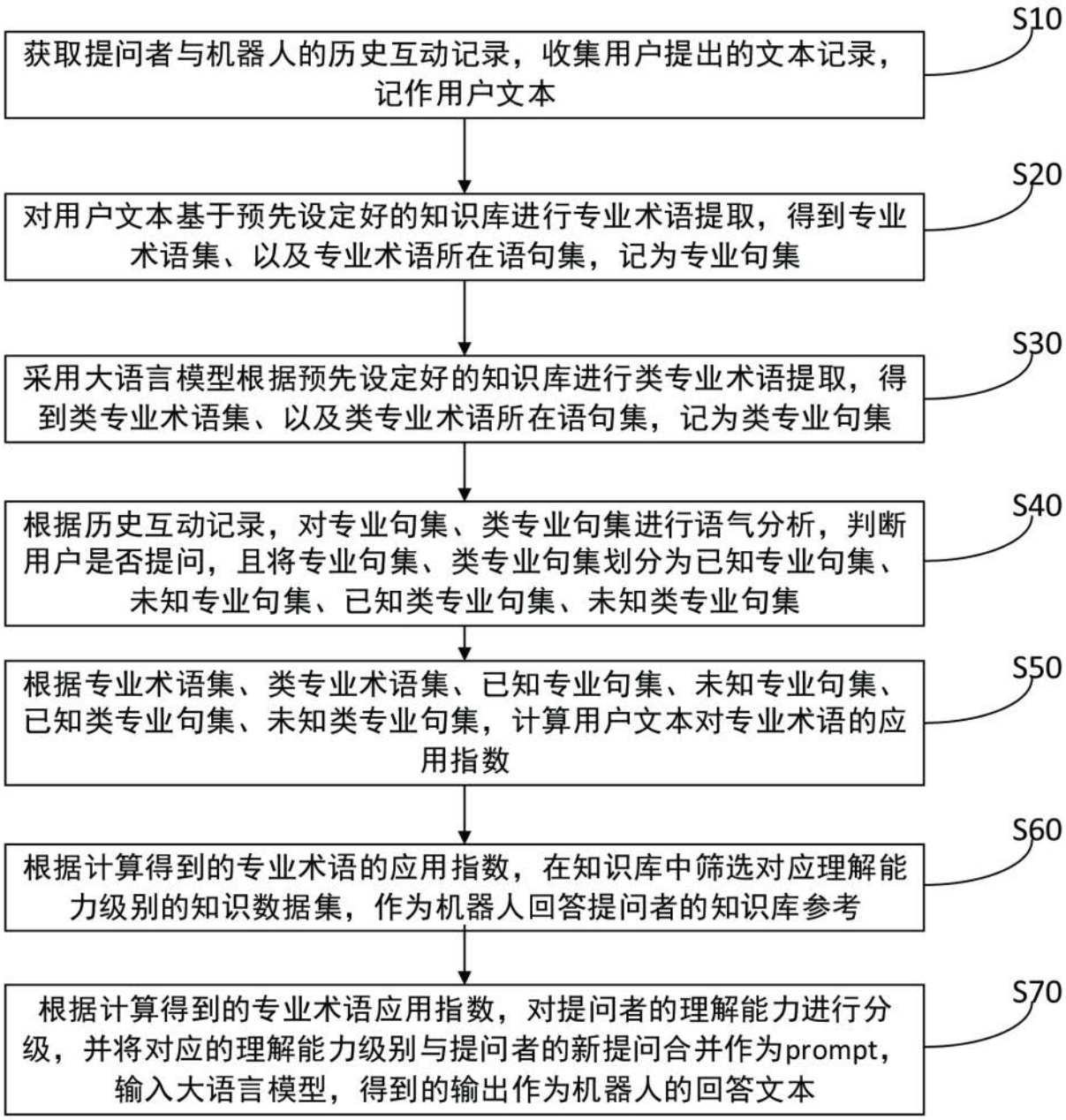
S10
获取提问者与机器人的历史互动记录，收集用户提出的文本记录，记作用户文本
S20
对用户文本基于预先设定好的知识库进行专业术语提取，得到专业术语集、以及专业术语所在语句集，记为专业句集
S30
采用大语言模型根据预先设定好的知识库进行类专业术语提取，得到类专业术语集、以及类专业术语所在语句集，记为类专业句集
S40
根据历史互动记录，对专业句集、类专业句集进行语气分析，判断用户是否提问，且将专业句集、类专业句集划分为已知专业句集、未知专业句集、已知类专业句集、未知类专业句集
S50
根据专业术语集、类专业术语集、已知专业句集、未知专业句集、已知类专业句集、未知类专业句集，计算用户文本对专业术语的应用指数
S60
根据计算得到的专业术语的应用指数，在知识库中筛选对应理解能力级别的知识数据集，作为机器人回答提问者的知识库参考
S70
根据计算得到的专业术语应用指数，对提问者的理解能力进行分级，并将对应的理解能力级别与提问者的新提问合并作为prompt，输入大语言模型，得到的输出作为机器人的回答文本

图 1