(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117951269 A
(43)申请公布日 2024.04.30

(21)申请号 202311818708.3

(22)申请日 2023.12.27

(71)申请人 北京易华录信息技术股份有限公司
地址 100043 北京市石景山区阜石路165号院1号楼1001室

(72)发明人 韩璐鑫

(74)专利代理机构 北京智沃律师事务所 11620
专利代理师 梁晨

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 16/31(2019.01)
G06F 40/151(2020.01)
G06N 5/04(2023.01)
G06N 5/022(2023.01)
G06N 20/00(2019.01)

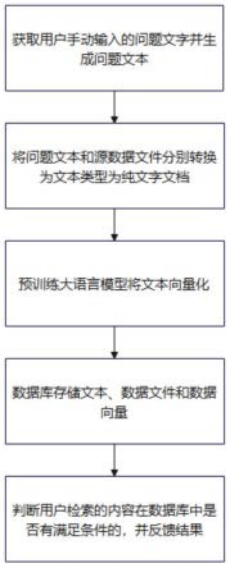
权利要求书2页 说明书4页 附图1页

(54)发明名称

一种基于大语言模型的问答系统

(57)摘要

本发明涉及数据处理技术领域,公开了一种基于大语言模型的问答系统,具体包括用户端模块,用于获取用户的问题并生成问题文本;多模态转换模块,用于将文本文件转换为文本类型为纯文字文档;数据处理模块,用于将文本拆分成若干个文本块,还用于将文本块转换为数据向量;数据库,用于存储文本块、数据向量和数据文件;应答输出模块,用于判断数据向量的相似度,调取文本块并生成自然语言文本;本发明通过通过设置多模态转换模块,支持多模态知识录入,允许用户以不同的方式输入知识,同时结合大语言模型通过对知识数据向量化,实现自动化知识点拆分和向量化知识库构建,大幅提升了系统的知识检索效率和准确率。



CN 117951269 A

1.一种基于大语言模型的问答系统,其特征在于,包括:

用户端模块,用于获取用户手动输入的问题文字并生成问题文本,问题文本输入预训练大语言模型并输出包含文本数据特征的第一数据向量;

多模态转换模块,用于识别问题文本和从外部采集的源数据文件的文件类型,将问题文本和源数据文件分别转换为文本类型为纯文字文档的第一文本和第二文本;

数据处理模块,用于将第一文本和第二文本按照预设文本长度分别拆分成若干个第一文本块和第二文本块,源数据文件的链接嵌入第二文本块中,按照顺序给每个文本块标记序号;还用于将文本块输入预训练的大语言模型并输出包含文本数据特征的第二数据向量,其中,文本块和第二数据向量互为对应关系;

数据库,用于存储第二文本块、第二数据向量、元数据文件和源数据文件,元数据文件包含第二数据向量在数据库中的文件位置信息和文件名称;

应答输出模块,用于查询在数据库中与第一数据向量最相似的第二数据向量,计算第一数据向量和第二数据向量的相似度并判断相似度是否超过预设相似度第一阈值,若是,则依据元数据文件调取第二数据向量对应的第二文本块以及第二文本块序号顺序前后的文本块输入所述的预训练大语言模型并生成自然语言文本,自然语言文本与源数据文件的链接传输至用户端模块,否则反馈无搜索结果至用户端模块。

2.根据权利要求1所述的一种基于大语言模型的问答系统,其特征在于,多模态处理模块还包括用于转换文件文本类型的文本转换单元,文本转换单元获取数据类型并执行以下命令:

a.若文件的类型为文本,则提取源数据文件中的文本信息并生成文字文本,剔除文字文本中的空格、停用词和标点符号并获得若干个关键句,利用分词算法对关键句进行分词处理得到若干个关键词并生成文本;

b.若文件的类型为图片,则利用OCR技术识别并提取源数据文件中的文本信息并执行a命令;

c.若文件的类型为语音,则利用语音识别技术将源数据文件中的语音内容转换为文本信息并执行a命令;

d.若文件的类型为视频,则将源数据文件拆分为音频数据文件和视频数据文件,音频数据执行c命令,视频数据拆分为若干帧图像执行b命令;

e.若文件的类型无法识别,则停止执行命令。

3.根据权利要求1所述的一种基于大语言模型的问答系统,其特征在于,用户手动输入的问题文字并生成问题文本,问题文本输入预训练大语言模型的过程还包括:

将文字文本中的终止符号替换为空格,对每两个空格间的文字进行分词处理并生成若干个目标词,目标词逐个输入预训练大语言模型。

4.根据权利要求1所述的一种基于大语言模型的问答系统,其特征在于,用户端模块还包括可视化页面,可视化页面用于接收应答输出模块中的自然语言文本并显示于可视化页面上,还用于接收源数据文件的链接并根据类型以显示、播放或二者组合的方式响应。

5.根据权利要求1所述的一种基于大语言模型的问答系统,其特征在于,预训练的大语言模型的训练方法为:

以ChatGPT作为大语言模型的基础模型,利用预先采集的语料数据对初始模型预训练。

6.根据权利要求4所述的一种基于大语言模型的问答系统,其特征在于,还包括迭代训练方法:

在用户端模块设置用于用户反馈准确率的选项,获取问题文本,采集自然语言文本,将问题文本内容和自然语言文本内容构成映射关系,并以选项中的准确率作为权重系数对大语言模型进行深度学习。

7.根据权利要求1所述的一种基于大语言模型的问答系统,其特征在于,计算相似度的方法为利用余弦相似度计算第一数据向量和第二数据向量的相似度,若相似度超过预设相似度第一阈值,设定序号长度为X,标记第二数据向量对应的第二文本块为目标文本块,获取目标文本块前后各X个文本块,计算每个文本块对应数据向量的相似度,剔除相似度低于预设相似度第二阈值的数据向量及其对应的文本块并整合输入预训练大语言模型。

一种基于大语言模型的问答系统

技术领域

[0001] 本发明涉及数据处理技术领域,具体地说,涉及一种基于大语言模型的问答系统。

背景技术

[0002] 在信息爆炸的时代,我们每天都会都通过互联网接收到海量的信息,无论是来自个人工作过程中的知识积累、社交媒体的信息传递、团队内部的知识共享还是其他渠道,同时我们存储、传递和获取知识信息的能力变得前所未有地强大,但如何高效地管理和利用这些海量的知识也成为巨大的挑战。

[0003] 在现有技术中,知识管理通常仅限于单一模态,如文本、图片或表格等,而且往往缺乏自动化和智能化的处理能力,对于导入和识别的知识内容,需要大量的人工操作和干预,这大大降低了知识管理的效率和效果;同时,现有的知识库构建方法通常基于传统的数据库或文件系统,这使得知识检索的效率和准确率受到限制。

发明内容

[0004] 本发明提供了一种基于大语言模型的问答系统,其能够克服现有技术的某种或某些缺陷。

[0005] 为实现以上目的,本发明通过以下技术方案予以实现,其包括:

[0006] 用户端模块,用于获取用户手动输入的问题文字并生成问题文本,问题文本输入预训练大语言模型并输出包含文本数据特征的第一数据向量;

[0007] 多模态转换模块,用于识别问题文本和从外部采集的源数据文件的文件类型,将问题文本和源数据文件分别转换为文本类型为纯文字文档的第一文本和第二文本;

[0008] 数据处理模块,用于将第一文本和第二文本按照预设文本长度分别拆分成若干个第一文本块和第二文本块,源数据文件的链接嵌入第二文本块中,按照顺序给每个文本块标记序号;还用于将文本块输入预训练的大语言模型并输出包含文本数据特征的第二数据向量,其中,文本块和第二数据向量互为对应关系;

[0009] 数据库,用于存储第二文本块、第二数据向量、元数据文件和源数据文件,元数据文件包含第二数据向量在数据库中的文件位置信息和文件名称;

[0010] 应答输出模块,用于查询在数据库中与第一数据向量最相似的第二数据向量,计算第一数据向量和第二数据向量的相似度并判断相似度是否超过预设相似度第一阈值,若是,则依据元数据文件调取第二数据向量对应的第二文本块以及第二文本块序号顺序前后的文本块输入所述的预训练大语言模型并生成自然语言文本,自然语言文本与源数据文件的链接传输至用户端模块,否则反馈无搜索结果至用户端模块。

[0011] 作为优选,多模态处理模块还包括用于转换文件文本类型的文本转换单元,文本转换单元获取数据类型并执行以下命令:

[0012] a.若文件的类型为文本,则提取源数据文件中的文本信息并生成文字文本,剔除文字文本中的空格、停用词和标点符号并获得若干个关键句,利用分词算法对关键句进行

分词处理得到若干个关键词并生成文本；

[0013] b.若文件的类型为图片，则利用OCR技术识别并提取源数据文件中的文本信息并执行a命令；

[0014] c.若文件的类型为语音，则利用语音识别技术将源数据文件中的语音内容转换为文本信息并执行a命令；

[0015] d.若文件的类型为视频，则将源数据文件拆分为音频数据文件和视频数据文件，音频数据执行c命令，视频数据拆分为若干帧图像执行b命令；

[0016] e.若文件的类型无法识别，则停止执行命令。

[0017] 作为优选，用户手动输入的问题文字并生成问题文本，问题文本输入预训练大语言模型的过程还包括：

[0018] 将文字文本中的终止符号替换为空格，对每两个空格间的文字进行分词处理并生成若干个目标词，目标词逐个输入预训练大语言模型。

[0019] 作为优选，用户端模块还包括可视化页面，可视化页面用于接收应答输出模块中的自然语言文本并显示于可视化页面上，还用于接收源数据文件的链接并根据类型以显示、播放或二者组合的方式响应。

[0020] 作为优选，预训练的大语言模型的训练方法为：

[0021] 以ChatGPT作为大语言模型的基础模型，利用预先采集的语料数据对初始模型预训练。

[0022] 作为优选，还包括迭代训练方法：

[0023] 在用户端模块设置用于用户反馈准确率的选项，获取问题文本，采集自然语言文本，将问题文本内容和自然语言文本内容构成映射关系，并以选项中的准确率作为权重系数对大语言模型进行深度学习。

[0024] 作为优选，计算相似度的方法为利用余弦相似度计算第一数据向量和第二数据向量的相似度，若相似度超过预设相似度第一阈值，设定序号长度为X，标记第二数据向量对应的第二文本块为目标文本块，获取目标文本块前后各X个文本块，计算每个文本块对应数据向量的相似度，剔除相似度低于预设相似度第二阈值的数据向量及其对应的文本块并整合输入预训练大语言模型。

[0025] 与现有技术相比，本发明的有益效果是：通过设置多模态转换模块，支持多模态知识录入，允许用户以不同的方式输入知识或需要检索的问题，如通过键盘输入文字、通过语音输入、上传图片或视频等；同时结合大语言模型通过对知识数据向量化，实现自动化知识点拆分和向量化知识库构建，大幅提升了系统的知识检索效率和准确率。

附图说明

[0026] 图1为实施例1中本发明的流程图。

具体实施方式

[0027] 为进一步了解本发明的内容，结合实施例对本发明作详细描述。应当理解的是，实施例仅仅是对本发明进行解释而并非限定。

[0028] 实施例1

[0029] 如图1所示,一种基于大语言模型的问答系统,其包括:

[0030] 用户端模块,用于获取用户手动输入的问题文字并生成问题文本,问题文本输入预训练大语言模型并输出包含文本数据特征的第一数据向量;

[0031] 多模态转换模块,用于识别问题文本和从外部采集的源数据文件的文件类型,将问题文本和源数据文件分别转换为文本类型为纯文字文档的第一文本和第二文本;

[0032] 数据处理模块,用于将第一文本和第二文本按照预设文本长度分别拆分成若干个第一文本块和第二文本块,源数据文件的链接嵌入第二文本块中,按照顺序给每个文本块标记序号;还用于将文本块输入预训练的大语言模型并输出包含文本数据特征的第二数据向量,其中,文本块和第二数据向量互为对应关系;

[0033] 数据库,用于存储第二文本块、第二数据向量、元数据文件和源数据文件,元数据文件包含第二数据向量在数据库中的文件位置信息和文件名称;

[0034] 应答输出模块,用于查询在数据库中与第一数据向量最相似的第二数据向量,计算第一数据向量和第二数据向量的相似度并判断相似度是否超过预设相似度第一阈值,若是,则依据元数据文件调取第二数据向量对应的第二文本块以及第二文本块序号顺序前后的文本块输入所述的预训练大语言模型并生成自然语言文本,自然语言文本与源数据文件的链接传输至用户端模块,否则反馈无搜索结果至用户端模块。

[0035] 当用户提出一个问题时,首先将问题转换成向量表示。然后,遍历知识库中所有知识的向量表示,并计算它们与问题向量的相似度。在检索过程中可以通过控制相似度阈值控制回答的准确性和相关性,设置一个相似度阈值,只有相似度高于该阈值的知识才会被返回作为回答。

[0036] 本发明中,多模态处理模块还包括用于转换文件文本类型的文本转换单元,文本转换单元获取数据类型并执行以下命令:

[0037] a.若文件的类型为文本,则提取源数据文件中的文本信息并生成文字文本,剔除文字文本中的空格、停用词和标点符号并获得若干个关键句,利用分词算法对关键句进行分词处理得到若干个关键词并生成文本;

[0038] b.若文件的类型为图片,则利用OCR技术识别并提取源数据文件中的文本信息并执行a命令;

[0039] c.若文件的类型为语音,则利用语音识别技术将源数据文件中的语音内容转换为文本信息并执行a命令;

[0040] d.若文件的类型为视频,则将源数据文件拆分为音频数据文件和视频数据文件,音频数据执行c命令,视频数据拆分为若干帧图像执行b命令;

[0041] e.若文件的类型无法识别,则停止执行命令。

[0042] 通过本实施例中设置的多模态处理模块,实现了确保知识内容按照统一的格式和规范进行组织,这样在转换成向量时可以保持一致性。

[0043] 本发明中,用户手动输入的问题文字并生成问题文本,问题文本输入预训练大语言模型的过程还包括:

[0044] 将文字文本中的终止符号替换为空格,对每两个空格间的文字进行分词处理并生成若干个目标词,目标词逐个输入预训练大语言模型。

[0045] 本发明中,用户端模块还包括可视化页面,可视化页面用于接收应答输出模块中

的自然语言文本并显示于可视化页面上,还用于接收源数据文件的链接并根据类型以显示、播放或二者组合的方式响应。

[0046] 通过本实施例中设置的可视化页面,能够将不同形态的知识以最合适的形式展示出去。

[0047] 本发明中,预训练的大语言模型的训练方法为:

[0048] 以ChatGPT作为大语言模型的基础模型,利用预先采集的语料数据对初始模型预训练。

[0049] 本发明中,还包括迭代训练方法:

[0050] 在用户端模块设置用于用户反馈准确率的选项,获取问题文本,采集自然语言文本,将问题文本内容和自然语言文本内容构成映射关系,并以选项中的准确率作为权重系数对大语言模型进行深度学习。

[0051] 本发明中,计算相似度的方法为利用余弦相似度计算第一数据向量和第二数据向量的相似度,若相似度超过预设相似度第一阈值,设定序号长度为X,标记第二数据向量对应的第二文本块为目标文本块,获取目标文本块前后各X个文本块,计算每个文本块对应数据向量的相似度,剔除相似度低于预设相似度第二阈值的数据向量及其对应的文本块并整合输入预训练大语言模型。

[0052] 通过本实施例中第一数据向量和第二数据向量的相似度可以用余弦相似度等方法来衡量,值越大表示两个向量越相似。最后,选择与问题向量相似度最高的知识,作为后续进行提问的知识内容。

[0053] 通过本发明中通过设置的多模态转换模块,支持多模态知识录入,允许用户以不同的方式输入知识或需要检索的问题,如通过键盘输入文字、通过语音输入、上传图片或视频等;同时结合大语言模型通过对知识数据向量化,实现自动化知识点拆分和向量化知识库构建,大幅提升了系统的知识检索效率和准确率。

[0054] 容易理解的是,本领域技术人员在本申请提供的一个或几个实施例的基础上,可以对本申请的实施例进行结合、拆分、重组等得到其他实施例,这些实施例均没有超出本申请的保护范围。

[0055] 以上示意性的对本发明及其实施方式进行了描述,该描述没有限制性,实施例所示的也只是本发明的实施方式的部分,实际的结构并不局限于此。所以,如果本领域的普通技术人员受其启示,在不脱离本发明创造宗旨的情况下,不经创造性的设计出与该技术方案相似的结构方式及实施例,均应属于本发明的保护范围。

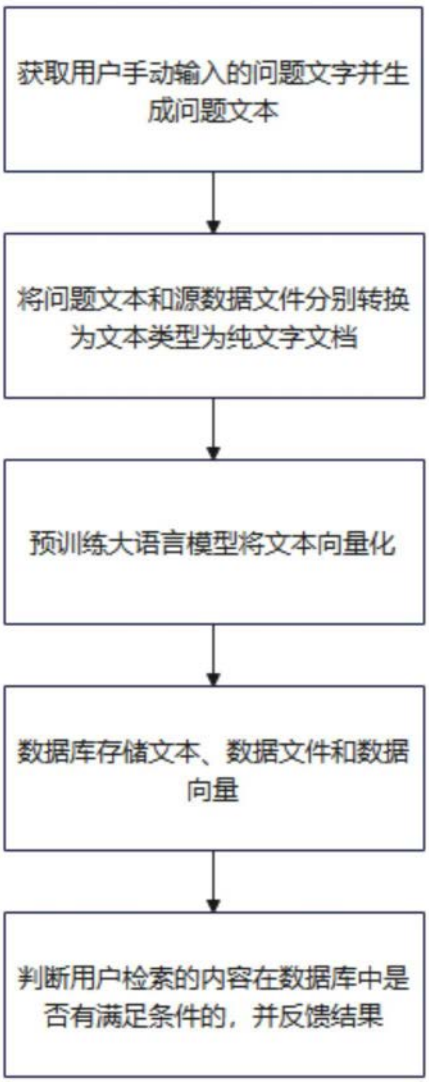

获取用户手动输入的问题文字并生成问题文本
将问题文本和源数据文件分别转换为文本类型为纯文字文档
预训练大语言模型将文本向量化
数据库存储文本、数据文件和数据向量
判断用户检索的内容在数据库中是否有满足条件的，并反馈结果

图1