Computer Science and Application 计算机科学与应用, 2021, 11(12), 3117-3125
Published Online December 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.1112315

Hans汉斯

基于知识库的C语言问答系统的设计与实现

宁海荣[1]，尤建清[1*]，李英豪[2]，肖丝莹[2]，王雨晴[2]，纪慧敏[2]

[1]北京信息科技大学计算机学院，北京
[2]北京信息科技大学理学院，北京

收稿日期：2021年11月27日；录用日期：2021年12月23日；发布日期：2021年12月30日

摘　要

问答系统可以提供快速、准确的知识获取渠道。本文设计了一个基于知识库的C语言问答系统，首先构建了一个基于自定义词典的本体知识库，并设计了相关问题模板，再通过分类器实现了问句和模板的关联匹配，最后返回问答结果。实验结果表明，本文设计的系统能够实现C语言知识的问答。

关键词

问答系统，本体知识库，分类器，问题模板

The Design and Implementation of the C Language Question and Answer System Based on the Knowledge Base

Hairong Ning[1], Jianqing You[1*], Yinghao Li[2], Siying Xiao[2], Yuqing Wang[2], Huimin Ji[2]

[1]Computer School, Beijing Information Science & Technology University, Beijing
[2]School of Applied Science, Beijing Information Science & Technology University, Beijing

Received: Nov. 27th, 2021; accepted: Dec. 23rd, 2021; published: Dec. 30th, 2021

Abstract

Question answer system can provide fast and accurate knowledge acquisition channels. This paper designed a C language question and answer system based on knowledge. First, an ontology

*通讯作者。

文章引用: 宁海荣, 尤建清, 李英豪, 肖丝莹, 王雨晴, 纪慧敏. 基于知识库的 C 语言问答系统的设计与实现[J]. 计算机科学与应用, 2021, 11(12): 3117-3125. DOI: 10.12677/csa.2021.1112315

knowledge based on a custom dictionary was constructed, and question templates were designed. Then the classifier was used to match questions and the templates. Finally, the result was returned from knowledge. The experimental results show that the system designed in this paper can match the question and answer of C language knowledge.

Keywords

Question Answer System, Ontology Knowledge Base, Classifier, Question Templates

1. 引言

C 语言初学者都会受到语法不清、编译报错的困扰，无从下手。他们获取解决方法途径是利用传统的信息检索，在网上大海捞针般地查找答案。这种方式虽然能够快速地反馈给用户大量的相关信息，但是信息冗余高，有效信息少，实际效率低。

本文针对以上问题，利用自然语言处理、建立知识图谱和问题模板匹配等方法，设计了一个能够理解用户真实意图的学科智能问答系统。该系统能够提高信息获取效率，反馈给用户有效的答案。

2. 相关工作研究

问答系统通常由知识图谱和知识推理两个重要部分组成，很多研究者都做了卓越而有效的研究。

2.1. 知识图谱

在问答系统中，知识图谱通常是利用知识表示模型，对特定的领域进行概念抽象、知识抽取和内容表达，从而完成对领域知识的结构化表示。刘园园等人[1]对高考信息进行分析和整理，提出了基于语义的知识表示模型，构建了高考信息知识图谱，并实现了高考咨询问答系统，能较为准确反馈给用户正确结果。刘爽等人[2]以少数民族文化作为研究点，为这一领域构建了相应的知识图谱，并基于该知识图谱成功搭建了可视化的少数民族文化知识查询平台。韩涛等人[3]针对航空发动机故障信息进行了分析，并构建了相应的知识图谱，提出了一种融合字、词序列信息的 Lattice Transformer-CRF 实体抽取方法。吴岳忠等人[4]针对包装领域设计了一个基于知识图谱的包装领域智能问答系统，并实现了图像识别和自动问答。

对于学科类的问答系统，知识图谱具有鲜明的学科特性，其构建过程与学科的知识点极为密切。何政等人[5]结合“软件构造基础——C#程序设计”课程开发知识图谱，并以此为基础构建学习导航系统用于软件工程专业的本科课程教学实践。李家瑞等人[6]构建了计算机学科的知识图谱，实现了计算机学科相关资源的快速查询。廖子慧[7]构建了基于英语语法知识图谱，利用 Cypher 查询语句在知识库中查询答案。

2.2. 知识推理

知识推理是问答系统非常重要的一环，它主要是通过一定的知识推理机制，对问句涉及的内容进行推理，并基于已有知识进行答案反馈。张紫薇[8]利用图归纳，强化学习模型等来学习结点间的特征关系，

设计出了一套更优的知识推理解决方案,该系统解决困难问题的能力有明显的提升。刘焕勇等人[9]探索出了一套基于逻辑推理知识库的可解释性路径推理方法和金融实体影响生成系统。李诗轩等人[10]基于领域知识构建上市公司财务危机预警本体，实现上市公司财务危机预警。

很多时候，知识推理技术都涉及到知识分类、内容表示等方面的研究。王峻[11]在各个属性分组内通过添加有向边的方式表达属性间的相关性，从而简化扩展朴素贝叶斯分类器的结构，提高分类正确率。杨航等人[12]在 R-vine Copula 理论的基础上，通过 AIC 准则选取最合适的 Pair Copula 函数，用极大似然估计法确定其参数，使用这种方法改进了朴素贝叶斯分类器。贾宇彤[13]对基于知识图谱的知识推理技术进行研究，提出了融合知识图谱中书体描述信息的知识推理模型以及基于三元组结构信息，引入非线性建模方式的知识推理模型。

本文利用自然语言处理技术，结合图数据库查询速度快、处理复杂关系能力强的优点，设计了基于本体知识库的 C 语言问答系统。

3. 系统架构

问答系统的目标是获得用户的问句请求，返回恰当的结果，本文采用基于句型模板匹配的问句预处理方法[14]，将问句处理成含有特定含义的语义逻辑形式[15]，构建 C 语言领域本体知识库，利用贝叶斯分类器，对问题和模板最相应地匹配，最后利用知识推理反馈结果。本文的问答系统流程如图 1 所示。

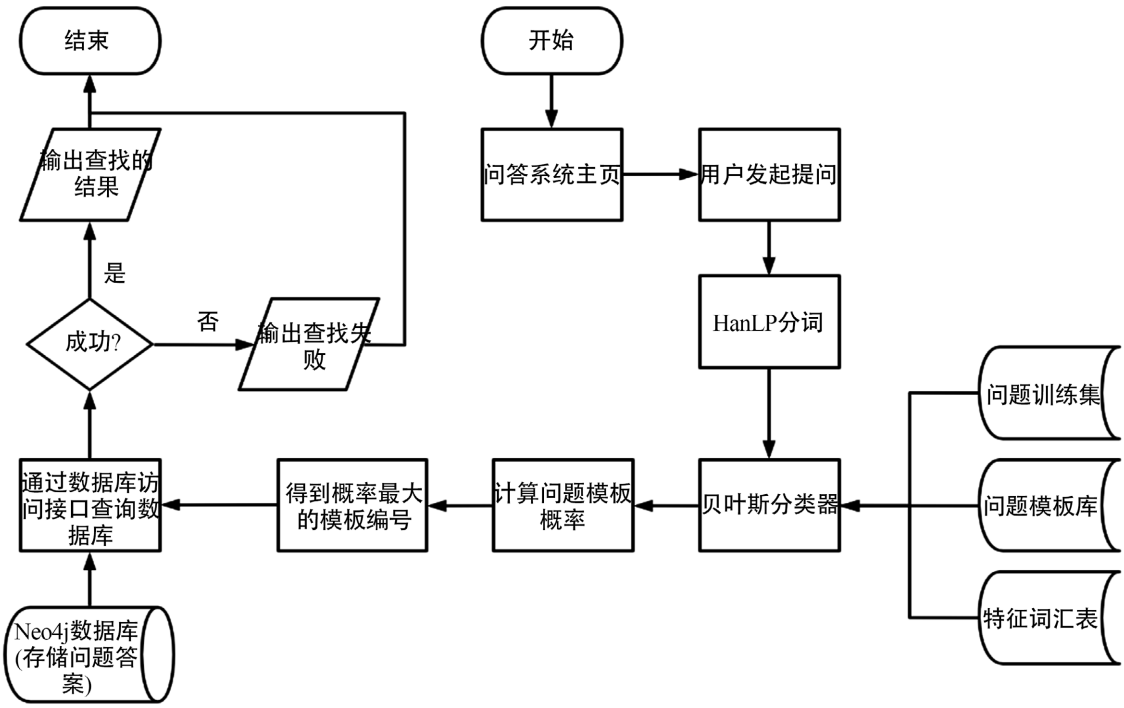

Figure 1. Question and answer system flow chart
图 1. 问答系统流程图

由于有关 C 语言的知识体系中大部分本体为专有名词，故本系统采用“问句分析——模板匹配”的方式来完成系统的构建。其主要过程为对用户发起的问句进行分词，将分词抽象后的句子匹配到对应的问题模板，通过数据库查询接口将问题模板对应到 Cypher 查询语句，最后将数据库查询到的结果返回给用户。

系统主要分为三大模块：本体与属性识别、句子抽象化、问题模板匹配。

问答系统的具体执行流程如图 2 所示。系统得到用户输入的问题后进行 HanLp 分词和词性标注的工作，在标注过程中，本文根据 C 语言的相关知识点，设置了一些自定义词典作为本体或本体的属性。在句子抽象化过程中，不属于本体及属性的关键字被还原到句子中，属于本体或属性的词用特定的标记替代，从而得到问句的抽象化后的语义逻辑表达，将抽象化后的句子通过朴素贝叶斯分类器匹配到对应的问题模板。最后根据匹配结果结合 Cypher 查询语句，通过数据库访问接口完成答案检索。

例如：

1) 原始问句：while 是什么意思；

2) 分词：while/key 是/v 什么/r 意思/n；

3) 句子抽象化：key 是什么意思；

4) 套用问题模板：key 简介；

5) 本体替换：while 简介；

6) 生成的查询语句：match(n) where n.name Contains while return n.description；

7) 返回结果：循环语句，当满足条件时进入循环，进入循环后，当条件不满足时，跳出循环。while 语句的一般表达式为：while(表达式){循环体}。

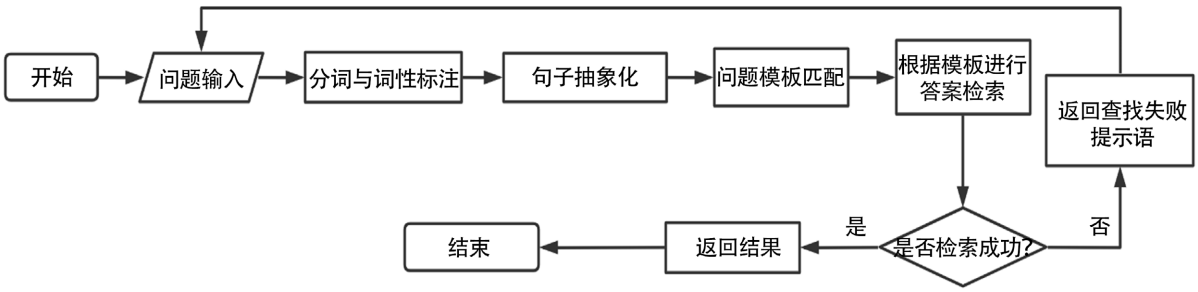

Figure 2. Question and answer processing flow
图 2. 问答处理流程

4. 本体与知识图谱的构建

4.1. 本体及其作用

本体是对客观存在的事物的结构化抽象[16]，抽象后的结果可以更好地被机器理解，能够为问答系统的知识描述提供帮助。在系统构建过程中，本体有不可或缺的作用：

1) 在爬取问答系统所需的数据时，利用本体可以快速确定网站页面内容是否与本领域相关。

2) 在问句处理过程中，本体与本体之间的相互关系，本体的属性及约束关系都可以为问句预处理提供很好的支持。

4.2. C 语言领域本体

本文对 C 语言的基础知识进行整理分析，确定本体的概念集合，完成本体库的构建。下面是 C 语言部分核心概念：

概念集合 = {标识符、常量、变量、整形数据、实型数据、字符数据、算术运算、自增自减运算、运算优先级、关系运算、逻辑运算、while 语句、for 语句、函数参数、指针、数组、二维数组、结构体、链表、枚举类型}。

4.3. 本体的属性

在本文所构建的 C 语言本体库中，本体属性是用来表示本体的特性和本体与本体之间的各种关系。用户在输入问句时，句子中通常包含有本体及其属性，下面是部分本体的属性集合：

字符串 ={长度、处理算法、字符类型……}；

数组 ={元素个数、元素类型……}；

链表 ={遍历、节点、链表插入……}；

数组 ={二维数组、一维数组……}。

4.4. 建立知识图谱

首先通过网络爬取、人工整理等方式获取到相关的资料并将其格式化成相应的本体概念或本体属性；然后将整理得到的本体概念、本体属性以及本体之间的关系存储在 neo4j 图数据库中，完成知识图谱的构建，其流程如图 3 所示。

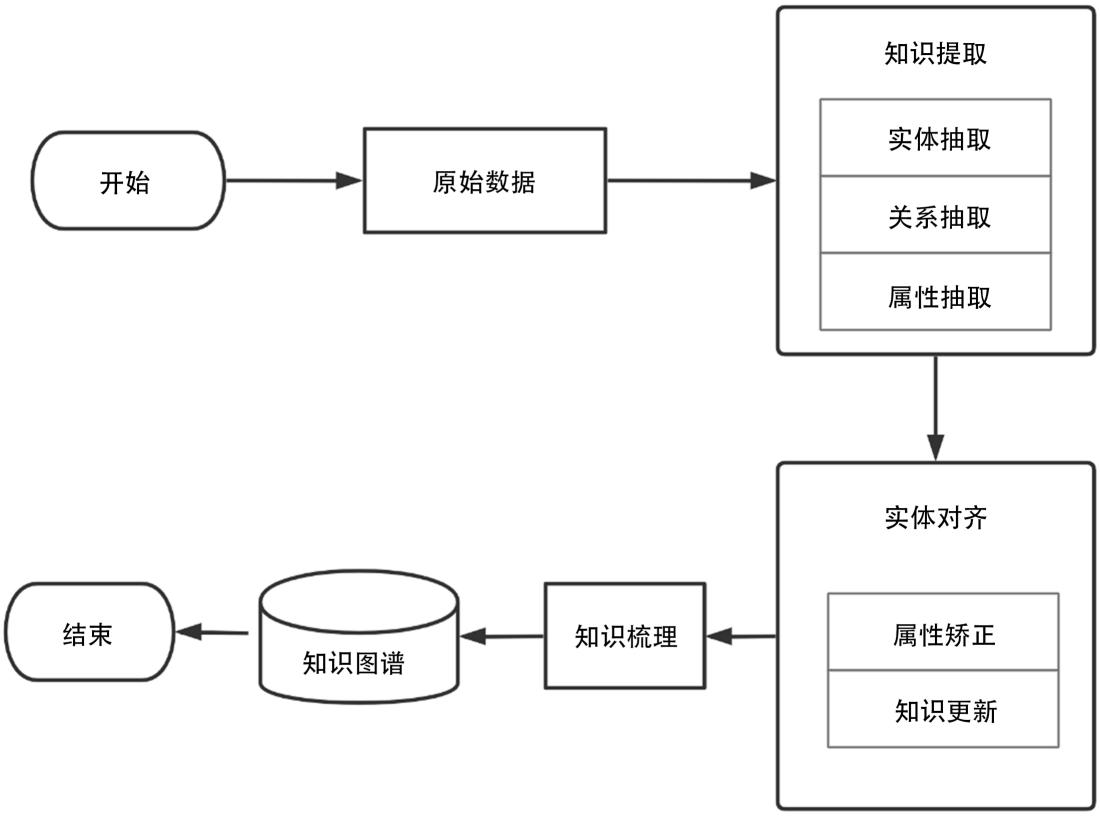

Figure 3. Construction a knowledge graph
图 3. 知识图谱的构建

在知识提取阶段，首先利用分词工具对原始数据做分词处理，然后人工标注实体、属性以及关系。在实体对齐阶段，检查实体对应的属性是否存在问题以及实体属性的完备性。在知识梳理阶段，通过人工验证的方式结合 C 语言知识点对本体知识库进行梳理。

本文构建的知识图谱中，存储了一系列的 C 语言知识本体、对应属性以及本体之间的关系，为该系统提供良好的数据支持，部分知识图谱如图 4 所示。

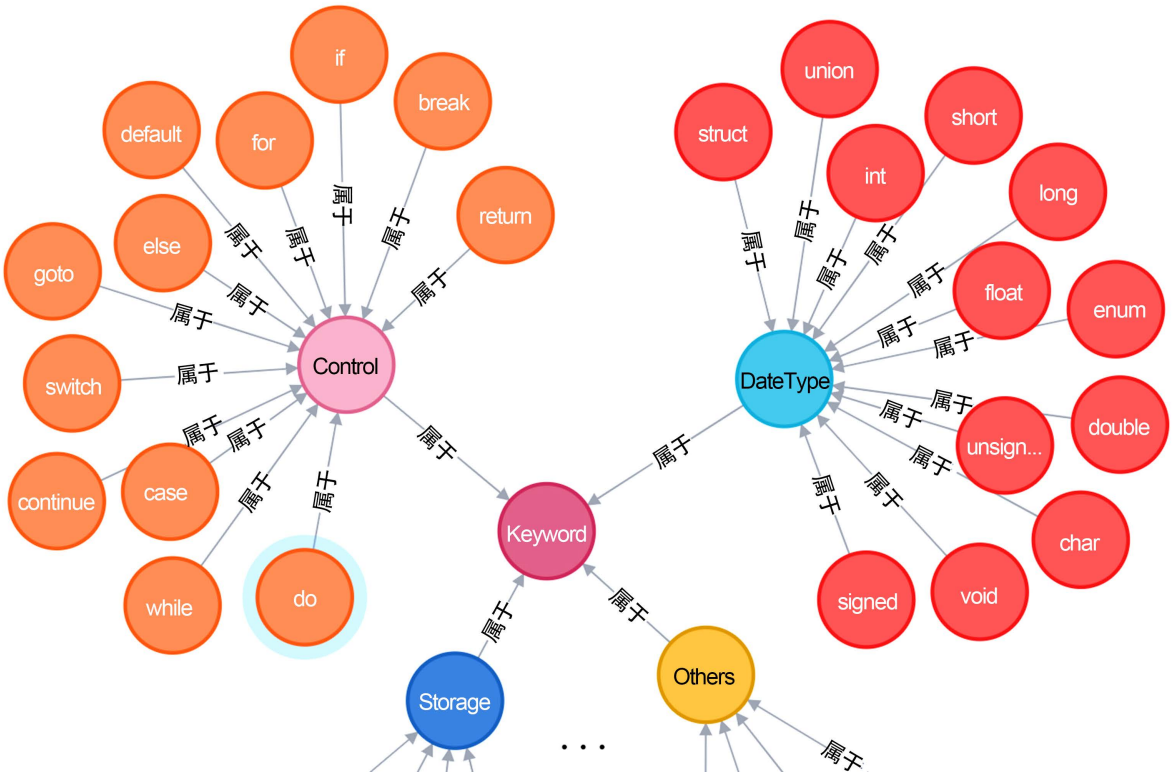

Figure 4. Partial knowledge graph
图 4. 部分知识图谱

5. 问题模板设计

5.1. 问题模板简介

问题模板是对同一句式的问句的抽象，其作用是将用户各种各样的提问，通过语义解析手段转化成为机器可理解的真实意图，例如：

1) “int 在 C 语言中的作用是什么？”
2) “C 语言中的 int 是啥意思？”
3) “int 是干什么用的？”

上述三个问句再经过语义解析之后都能匹配到“<key>简介”模板，即上述问句都是表达“int 简介”的问句。

5.2. 建立问题模板

在设计问题模板时，需要考虑到问句中可能包含有多个本体及属性，以及对本体相关操作的询问。经过分析，在 C 语言基础知识邻域问题模板可以设计为以下 4 种：

1) 询问本体的概念；
2) 询问多个本体间的关系；
3) 询问本体的属性；
4) 询问本体的相关操作；

本系统中设计的问题模板及部分问题实例如表 1 所示。

Table 1. Part of the question template
表 1. 部分问题模板

问题模板类型	问题实例
本体概念	while 是什么意思？
多个本体间的关系	快速排序和堆排序有什么联系？
本体及其属性	快速排序的复杂度是多少？
本体及操作	对链表求和。

5.3. 问题模板匹配原理

问题模板匹配的任务主要是由贝叶斯分类器完成的，其工作流程如图 5 所示。

首先使用现有的样本集来训练获得已知事件的概率，然后通过贝叶斯公式来获得所求事件的概率，以得到概率最大的问题模板标号。

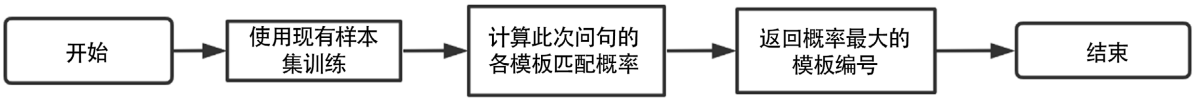

Figure 5. Naive Bayes classifier workflow
图 5. 朴素贝叶斯分类器工作流程

贝叶斯分类器的理论依据是贝叶斯定理，由公式 1 给出：

$$p(Y|X)=\frac{p(X|Y)p(Y)}{p(X)} \tag{1}$$

其中： $p(Y)$ 为先验概率， $p(X|Y)$ 为条件概率， $p(Y|X)$ 是进行分类预测的后验概率。

在本系统中 $X=(x_1,x_2\cdots,x_n)$ 是问句的 n 维特征词的； $Y=(y_1,y_2\cdots,y_k)$ 是 k 个问题模板。

本文假设每个关键词之间为相互独立的情况下，对于给定的 k 个问题模板可以采用公式 2 来分别求解 $y_1,y_2\cdots,y_k$ 的概率

$$P(y_k|x_1,\cdots,x_n)=\frac{P(x_1|y_k)P(x_2|y_k)\cdots P(x_n|y_k)P(y_k)}{P(x_1)P(x_2)\cdots P(x_n)} \tag{2}$$

其中：

- $x_1,x_2\cdots,x_n$ 为已经发生的事件，在本系统中它们是句子中被分解出来的每一个特征词；
- y_k 为一个待求解事件，在本系统中为已知的 k 个问题模板；
- $P(x_n|y_k)$ 以及其余的变量则是通过训练集训练得到的已知的概率。

最终通过求解每一个模板在已存在 $x_1,x_2\cdots,x_n$ 关键字的情况下匹配到 y_k 的概率 $P(y_k|x_1,\cdots,x_n)$，取其中概率最大者作为该情况下匹配到的结果，并返回该结果所对应的问题模板编号。

6. 系统实现与结果展示

本系统的实现使用了 Spring Boot 微服务框架，结合 HanLp 分词工具以及基于 Spark 集群计算环境的朴素贝叶斯分类器，完成系统搭建。

在样本训练阶段，系统读取训练集文件，将文本转化为 n 维词向量集合，根据此向量集合由 SparkContext 创建出可以并行执行的数据集 RDD 并将该数据集经过转化后交给朴素贝叶斯分类器训练。

在问题匹配阶段，系统接收由前端传入的问句，在经过分词到构造词向量的一系列操作后，将构造好的词向量交给贝叶斯分类器，由贝叶斯分类器将构造好的词向量与训练数据做概率预测，最后返回问句与各个问题模板的匹配概率。

在答案检索阶段，系统将原句中的本体、属性等作为参数，传入概率最大的问题模板对应的 Cypher 查询语句中，从数据库中查询结果。

例如“while 是什么意思”这个问题，对应的问题模板为“本体概念”，其所对应的查询语句为：“match(n) where n.name Contains $name return n.description”，对其传入$name(while)，利用 Cypher 语句在知识库中查询即可得到问题的答案。

在采用用户自定义词典协同分词且每条模板对应 100 条训练样本的情况下，对该系统的匹配准确率进行验证，表 2 为测试样本的数量一定的情况下，该系统能够成功匹配到对应问题模板的概率。其中，正确率为匹配成功数与测试样本数之比。

Table 2. Matching accuracy of different templates under the same test volume
表 2. 不同模板在同一测试量下的匹配正确率

问题模板编号	测适量	正确率
1	50	0.84
2	50	0.82
3	50	0.80
4	50	0.80

由测试结果可见，对于常规的提问方式，本文所设计的系统能够较好地将正确的结果反馈给用户，基本能够满足 C 语言初学者的使用需求。

系统运行效果如图 6 所示。

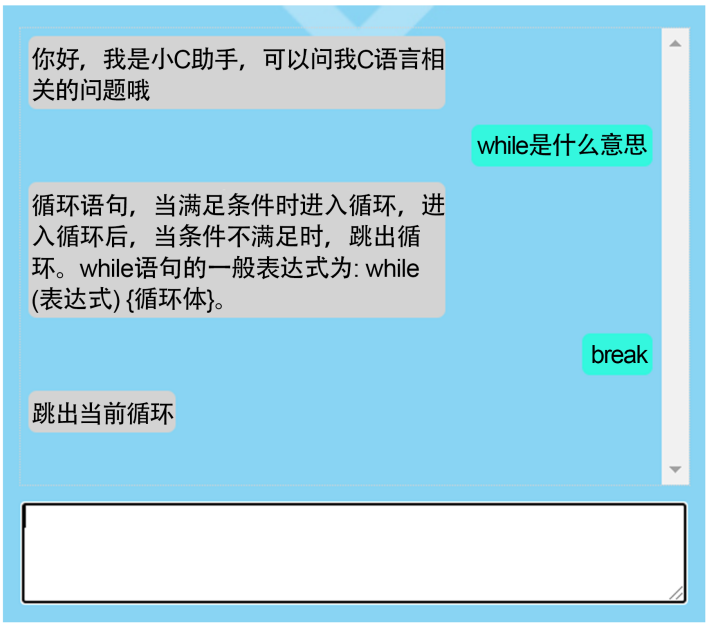

Figure 6. System operation renderings
图 6. 系统运行效果图

7. 结束语

本文针对 C 语言知识领域设计了一套问答系统，通过用自然语言处理、问题模板匹配、知识库检索等技术使用问答的方式进行用户交互。在与以往的知识获取方式的对比下，该系统能够较为准确地理解到用户的意图，反馈给用户更加准确的答案。

在未来，我们将继续完善该系统，进一步扩充知识图谱，增加训练集的样本量，让该系统能够准确地解决更多问题。

基金项目

北京信息科技大学 2021 年大学生创新创业训练计划项助，项目号 5102110805；

北京信息科技大学高教研究课题，项目号 2020GJYB13；

促进高校分类发展－专业建设与人才培养模式改革－计算机学院人才培养模式改革，项目号：5112110852。

参考文献

[1] 刘园园, 李劲华, 赵俊莉. 基于语义解析的领域问答系统的设计与实现[J]. 计算机应用与软件, 2021, 38(11): 42-48+97.

[2] 刘爽, 杨辉, 李佳宜, 谭楠楠. 面向多数据源的少数民族文化知识图谱构建[J]. 计算机技术与发展, 2021, 31(8): 191-197+203.

[3] 韩涛, 黄海松, 姚立国. 面向航空发动机故障知识图谱构建的实体抽取[J]. 组合机床与自动化加工技术, 2021(10): 69-73+78.

[4] 吴岳忠, 沈雪豪, 肖发龙, 邓芝一, 李长云. 基于知识图谱的包装领域智能问答系统的设计[J]. 包装工程, 2021, 42(15): 203-210.

[5] 何政, 叶刚. 基于知识图谱的 C#课程学习导航系统研究[J]. 太原城市职业技术学院学报, 2021(9): 97-99.

[6] 李家瑞, 李华昱, 闫阳. 面向多源异质数据源的学科知识图谱构建方法[J]. 计算机系统应用, 2021, 30(10): 59-67.

[7] 廖子慧. 基于知识图谱的英语语法智能题库系统研建[D]: [硕士学位论文]. 北京: 北京林业大学, 2020.

[8] 张紫薇. 基于机器学习的知识推理的研究与设计[D]: [硕士学位论文]. 北京: 北京邮电大学, 2021.

[9] 刘焕勇, 薛云志, 李瑞, 任红萍, 陈贺, 张鹏. 面向开放文本的逻辑推理知识抽取与事件影响推理探索[J]. 中文信息学报, 2021, 35(10): 56-63.

[10] 李诗轩, 陈烨, 石文萱, 杨达森. 基于知识推理的上市公司财务危机预警研究[J]. 武汉理工大学学报(信息与管理工程版), 2021, 43(4): 322-329.

[11] 王峻. 基于属性分组的扩展朴素贝叶斯分类器[J]. 洛阳理工学院学报(自然科学版), 2021, 31(3): 85-88+93.

[12] 杨航, 刘赫, 夏美美, 范元静. 基于 R-vine Copula 理论的改进朴素贝叶斯分类器[J]. 甘肃科学学报, 2021, 33(3): 12-16.

[13] 贾宇彤. 基于知识图谱的知识推理研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2021.

[14] Wang, L., Liu, H., Zhou, T., Liang, W. and Shan, M. (2021) Multidimensional Emotion Recognition Based on Semantic Analysis of Biomedical EEG Signal for Knowledge Discovery in Psychological Healthcare. *Applied Sciences*, **11**, Article No. 1338. https://doi.org/10.3390/app11031338

[15] 关慧, 吕颖, 贾成真. 基于句法和语义的需求依赖关系自动获取[J]. 计算机技术与发展, 2021, 31(2): 20-26.

[16] Chen, K., Shen, G., Huang, Z. and Wang, H. (2021) Improved Entity Linking for Simple Question Answering Over Knowledge Graph. *International Journal of Software Engineering and Knowledge Engineering*, **31**, 55-80. https://doi.org/10.1142/S0218194021400039