College of Computer Science and Technology, Jilin University

– Oldsweet Group –

Knowledge Quiz System for Disease-Centred Knowledge Graph in Pharmaceuticals

Document Data:	Reference Persons:
- 2023-11-01 -	- Hangbo Yu, Yichen Li, Baisong Li -

© 2023 Jilin University
Jilin, China

BaH (internal) reports are for internal only use within the NULL Group. They describe preliminary or instrumental work which should not be disclosed outside the group. BaH reports cannot be mentioned or cited by documents which are not BaH reports. BaH reports are the result of the collaborative work of members of the NULL group. The people whose names are in this page cannot be taken to be the authors of this report, but only the people who can better provide detailed information about its contents. Official, citable material produced by the NULL group may take any of the official Academic forms, for instance: Master and PhD theses, DISI technical reports, papers in conferences and journals, or books.

Contents

Revision History:

Revision	**Date**	**Author**	**Description of Changes**
revision1.0	11.8.2023	Hangbo Yu,Yichen Li	Writing Documents

1 Introduction

Disease is a complex process. People's exploration of diseases is endless, whether it is the causes, symptoms, complications directly related to the disease or the therapeutic drugs, related foods, preventive measures, etc. related to the disease. All these have become hot topics in the field of diseases. This information is of great use to both medical professionals and the general public. For medical professionals, comprehensive and effective information allows them to better treat patients. For the general public, the information can play a role in the prevention and popularization of diseases. For the patients, they can better understand their conditions and help alleviate them through appropriate food treatment. Therefore, from the perspective of life and health, it is necessary to integrate and utilize disease-related information.

The iTelos methodology, provided by Knowledge Graph Engineering (KGE) develops a process to enhance the reuse of resources within a specific domain. Following the iTelos methodology, we conducted a project to integrate and utilize knowledge and data resources about diseases.

This report is organized as follows. Section 2 defines the purpose and specific domains of the project. It provides a high-level overview of low-quality resources created by producers and high-quality resources identified and composed by consumers. Section 3 defines the purpose of our study by defining a series of scenarios, roles, and competency questions (CQs). Questions (CQs) to identify a set of entities that are modeled in an Entity-Relationship (ER) model. Section 4 describes the data sources we used and the use of knowledge correlation. In Section 5, we extract entity types, data attributes and object properties from the aforementioned data. Section 6 is devoted to the formalization of the Purpose Theory of the data sources, which is then integrated into a comprehensive Purpose Theory. At the same time, we construct a disease-centered ontology and integrate it with the aforementioned purpose theory to form an ontology. Section 7 outlines a data formalization based on disease-centered teleology. Section 8 develops a downstream task(Q&A system) based on it. Finally, we draw conclusions and provide an outlook for the future in Section 9.

2 Project Description

2.1 Purpose and Domain of Interest

Purpose of this project: The purpose of this project is to establish a comprehensive and practical disease-centered knowledge map, including information directly related to the disease, such as disease symptoms, required tests, complications, but also drugs, food, prevention and other disease-related. The integration and utilization of this information can enable people to do a certain degree of popularization of the disease, while medical professionals can use this project to play a role in learning and assistance.

Project domain of interest(DoI):Our project takes disease as its starting point and focuses on data in terms of rigor and policy, our source is MedicineNet, a US medical website that provides information on diseases, conditions, medications, and general health conditions. nhsinform is the new National Health Information Service for Scotland. Part of the data for the Q&A system (downstream task) came from already integrated Q&A datasets.

2.2 Project Development

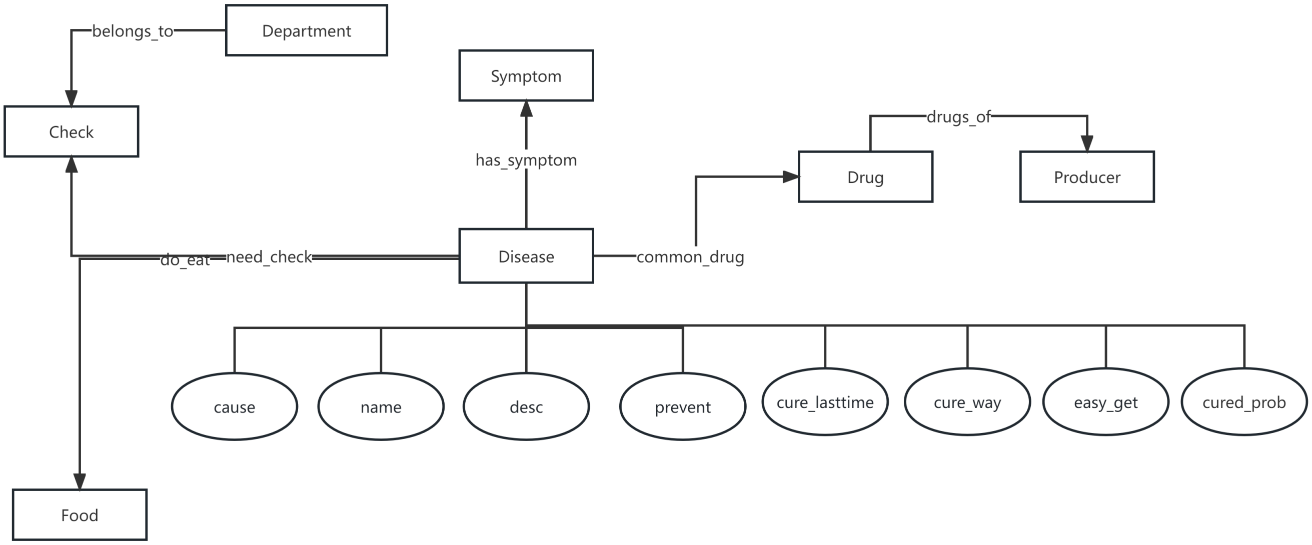
2.2.1 Data Production

In this phase, the goal of the producer is to generate datasets that satisfy the purpose of our project, if these resources do not yet exist or are of poor quality, we consider replacing the relevant information in the knowledge graph generation phase, if it is not possible to replace the information that is in the medical rigor we remove it, but in the downstream task phase we consider finding the most similar knowledge as an answer to be given back to the user. In our project, we need to create and formalize information about diseases. Symptoms of diseases are described in varying degrees of detail, and we distill them in simple terms such as: hypertrophy of breast tissue; deep hemorrhage of the brain parenchyma. Food items are usually substituted for food groups or related dishes that can assist in treatment. The information is basically in strings.

2.2.2 Data Composition

The goal of the consumer is that people can use our Knowledge Graph. Consumers may need to find, access and understand the information in the knowledge graph and use it as a basis for decision making or other activities. Therefore, the knowledge graph needs to provide them with information that meets their needs on the one hand and allows them to better convey information to them on the other hand. For this reason, we have tried to cover as many relevant fields as possible in the construction of the knowledge graph, and we have developed its downstream tasks based on this. The system correlates the relevant information in the knowledge graph and provides the required information to the consumer in the form of an answer.

Based on the above factors, the E-R diagram of our project is shown in Fig:



3 Project Formalization

The aim of our project is to focus on diseases and integrate all information related to them, so as to provide a certain degree of popularisation and querying in the field of medicine and health.

In order to describe the multiple aspects considered for the purpose of the project, we have listed a set of usage scenarios:

Scenario 1: The emergence of a disease is often accompanied by changes in the body's functioning that reflect problems, so the symptoms of the disease and the complications it can bring are a topic of interest to people.

Scenario 2: Once the disease appears, we need to examine and treat the disease, so the examination required for the disease, the recommended medication for a certain disease and the auxiliary nature of the food treatment will become more concerned about the problem.

Scenario 3: There are various kinds of diseases, such as colds and other diseases. For a certain type of disease, the suddenness of the disease makes healthy people want to know more about the disease and take preventive measures.

In the scenarios defined above, we represents a set of real users with specific features included in the project purpose, whhich are listed as follows:

Personas 1: David, a 60 year old man who has maintained his health well and believes in therapeutic food, so he is reluctant to go to the hospital.

Personas 2: Lily, a 30-year-old patient with poor health who is a frequent visitor to the hospital. Although she often goes to the hospital, she is not interested in how she got sick or the follow-up of her illness.

Personas 3: Li Ming, a 20 year old medical student with good grades and hard work, who wants to use technology to help him learn more about the disease and to prepare for the tests he will have to take later in his studies.

Personas 4: Dahua, a 33-year-old doctor, a relatively inexperienced doctor who likes to use the Internet and his own medical experience to help him better analyse his patients' conditions.

Taking into account the personas in the scenarios defined, we create Competency Questions (CQs):

CQ1: Please recommend foods and medicines for tinnitus in David's old age.

CQ2: Lily has been having a runny nose lately and she wants to know what to do about it and what kidney disease she might be suffering from if she has these symptoms.

CQ3: Li Ming is doing a research assignment related to liver disease, he needs to understand the causes, complications and treatment drugs of liver disease.

CQ4: Dahua is giving a prescription to a patient suffering from a cold and is interested in the new medicines that have been introduced to the hospital. He would like to know what the medicines are for and what points he needs to raise when advising the patient on prevention.

From the CQs, referring to Personas and Scenarios, we extract Entities with properties.These entities are categorized as either *Common, Core, or Contextual* entities by considering Focus classification and Popularity classification. The details of this work are outlined in Table 1.

Scenarios	Personas	CQs	Entities	Properties	Focus classification	Popularity classifiction
2	1	1	Drug	string	Common	Common
1	2	2	Symptom	string	Core	Core
1,2	3	3	Check,Disease,Symptom	string	Contextual	Contextual
3	4	4	Drug,Check	string	Contextual	Contextual

4 Information Gathering

In this section, the producers aim to identify informal sources of data and knowledge and integrate them to achieve the project objectives. The producer then collects informal data and knowledge from informal sources and processes them to extract the resources to be used. And the consumers aim to identify formal data and knowledge sources and then capture formal data and knowledge sources to integrate all resources.

4.1 Data and Knowledge Source

The data sources for this project are mainly informal data and knowledge from producers. Searching for Medicine is a medical information retrieval site for common diseases, containing a large amount of data on departments, diseases, medicines, and diagnostic and treatment information. We crawled the information related to diseases by python and used it as informal data for medical knowledge graph construction. The quantity volume is about 45M in total.

Resource name	Medical Information
Domain	Common medical conditions and treatment options
Language	Chinese
Date URL	https://jib.xywy.com/
Date format	.json file
Date description	Each piece of information in the dataset is related to a disease, including the name of the disease, symptoms, tests, medication, treatment costs, etc.
Knowledge URL	N/A
Knowledge description	N/A

medicinenet is an English-language database in the medical field operated by MedicineNet, Inc. storing about 20,000 medical terms and their general explanations in a dictionary order, which can be combined with Searching for Medicine as an initial data source to enrich our medical knowledge.

Resource name	Medical-Dictionary
Domain	Common medical terms and their corresponding explanations
Language	English
Date URL	https://www.medicinenet.com/medterms-medical-dictionary/article.htm
Date format	.csv file
Date description	Define difficult medical language in easy-to-understand explanations of over 19,000 medical terms.
Knowledge URL	N/A
Knowledge description	N/A

DBpedia is one of the world's largest multi-domain knowledge ontologies, from which we extracted the structure of medical, disease, and drug-related ontologies as a reference for our knowledge graph ontology construction.

Resource name	Ontological reference
Domain	Medical Field Ontology
Language	English
Provider	https://www.dbpedia.org/
Knowledge URL	https://databus.dbpedia.org/sparql
Knowledge description	Sparql query statements were used to query for medical-related concepts and attributes in dbpedia, as well as the relevance of each concept, which were used to assist in the construction of the knowledge ontology for this project.

4.2 Resource Collection, Processing and Scraping

Informal resources are collected and processed from producers as follows. Based on the above data and knowledge sources, we separated different fields from the collected json files to get different categories of data. In total, we get about 44,000 entity counts with about 30w relationships.

Files	check.txt,deny.txt,department.txt,disease.txt,drug.txt, food.txt,producer.txt,symptom.txt
Description	Each file contains information on a single category of entities
Source	Medical Information

5 Language Definition

In the previous section, we extracted entity types, and relationship types from healthcare information. Next, we formalize the language of these concepts by mapping them to Global Identifiers (GIDs) in the Universal Knowledge Core (UKC). In UKC, each GID corresponds to a unique definition of a concept.

5.1 Entity and Relationship Type

Concept labels	Description
Name	疾病名称，如“喘息样支气管炎”
Desc	疾病简介，如“又称哮喘性支气管炎...”
Cause	疾病病因，如“常见的有合胞病毒等...”
Prevent	预防措施，如“注意家族与患儿自身过敏史...”
Cure_lasttime	治疗周期，如“6-12个月”
Cure_way	治疗方式，如”药物治疗”,”支持性治疗”
Cured_prob	治愈概率
Easy_get	疾病易感人群，如“无特定的人群”

Concept labels	Description
Check	诊断检查项目，如支气管造影、关节镜检查等
Desc	医疗科目，如整形美容科、烧伤科等
Drug	药物，京万红痔疮膏、布林佐胺滴眼液等
Food	食品，番茄冲菜牛肉丸汤、竹笋炖羊肉等
Producer	在售药品，通药制药青霉素V钾片、青阳醋酸地塞米松片等
Symptom	疾病症状，乳腺组织肥厚;、脑实质深部出血等

Concept labels	Relationship Description
Belongs_to	《妇科,属于,妇产科》
Common_drug	《阳强,常用,甲磺酸酚妥拉明分散片》
Do_eat	《胸椎骨折,宜吃,黑鱼》
Drugs_of	《青霉素V钾片,在售,通药制药青霉素V钾片》
Need_check	《单侧肺气肿,所需检查,支气管造影》
No_eat	《唇病,忌吃,杏仁》
Recommend_drug	《混合痔,推荐用药,京万红痔疮膏》
Recommend_eat	《鞘膜积液,推荐食谱,番茄冲菜牛肉丸汤》
Has_symptom	《早期乳腺癌,疾病症状,乳腺组织肥厚》
Accompany_with	《下肢交通静脉瓣膜关闭不全,并发疾病,血栓闭塞性脉管炎》

5.2 Data Type

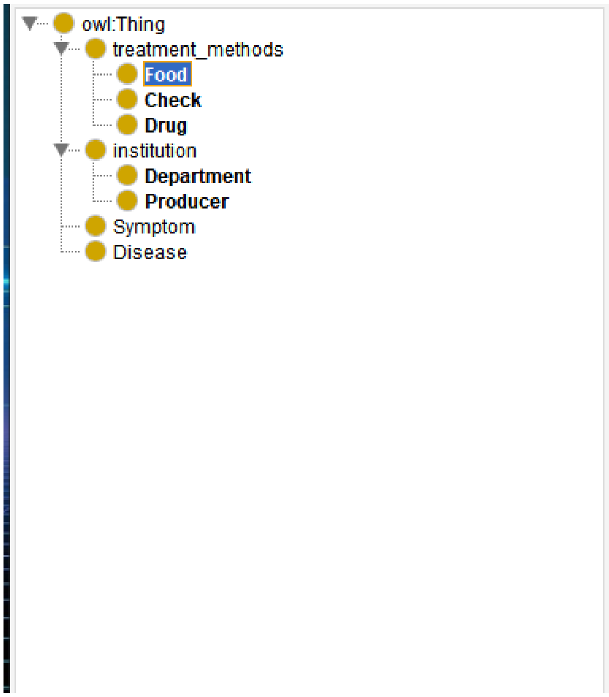
Concept labels	Description
Name	疾病名称，如“喘息样支气管炎 “
Desc	疾病简介，如“又称哮喘性支气管炎...”
Cause	疾病病因，如“常见的有合胞病毒等...”
Prevent	预防措施，如“注意家族与患儿自身过敏史... “
Cure_lasttime	治疗周期，如“6-12个月”
Cure_way	治疗方式，如”药物治疗”,”支持性治疗”
Cured_prob	治愈概率
Easy_get	疾病易感人群，如“无特定的人群”

6 Knowledge Definition

In this section, the goal of the producer is to generate the purpose theory for each dataset, always considering the reusability of the knowledge enhancement. The goal of the consumer is to integrate the purposive theories from the producers and determine an ontology for the entire healthcare knowledge graph.

6.1 Teleologies from Producer

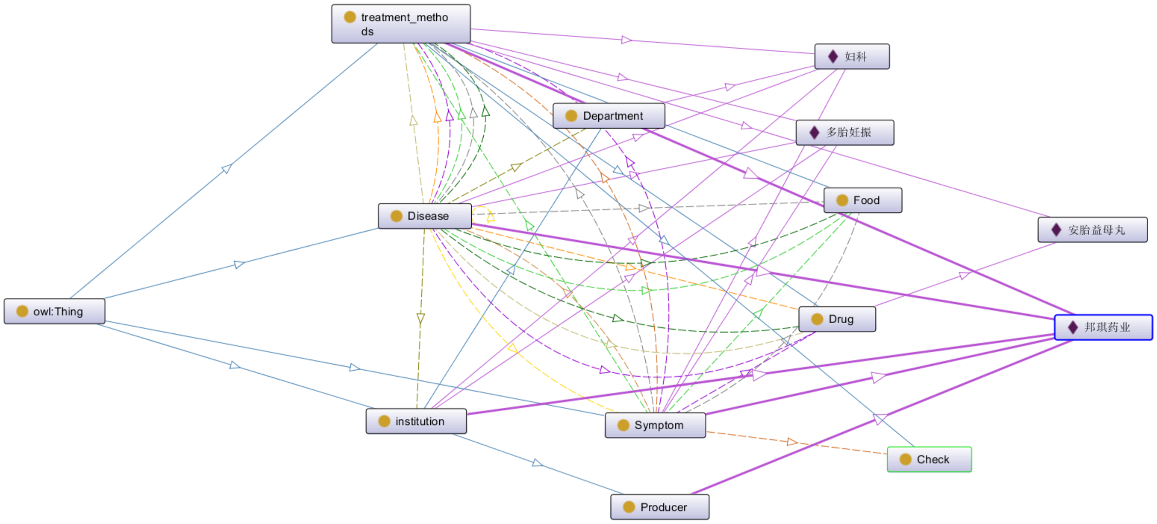
Based on the medical domain ontology referenced in the previous section, the following ontology information is obtained for the different datasets:



- owl:Thing
  - treatment_methods
    - Food
    - Check
    - Drug
  - institution
    - Department
    - Producer
  - Symptom
  - Disease

6.2 Teleologies from Consumer

Integrating the knowledge and data purposively enables to obtain the final constructed ontology of the whole healthcare knowledge graph:

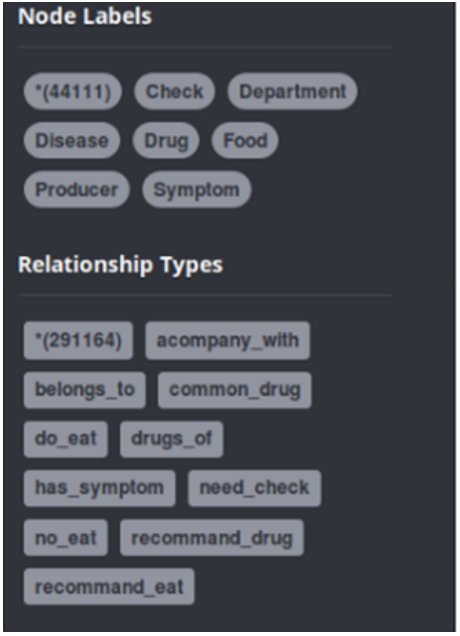


7 Data Definition

In this section, the goal of the producer is to formalize each dataset and map each dataset to its respective schema. Meanwhile, the consumer's goal is to combine all datasets and merge the combined data with the final purposive schema.

7.1 Dataset formatting by Consumer

Consumers formalize the data to instantiate the entity types, relationship types, and attributes therein as defined by the Healthcare Knowledge Graph ontology. For identical entities that appear in different documents, they are integrated through entity disambiguation. For example, some entities can be used as both medicines and food, we query the proximity table based on their names to merge them and merge the original attributes accordingly.



Node Labels

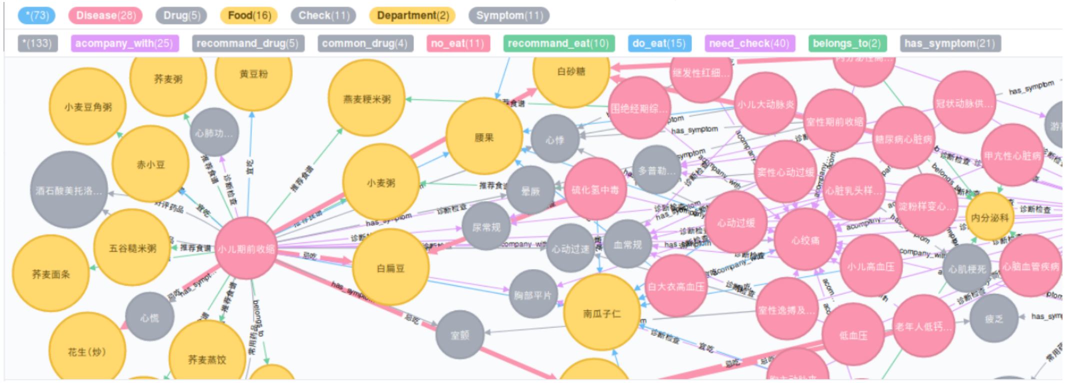
*(44111) Check Department
Disease Drug Food
Producer Symptom

Relationship Types

*(291164) acompany_with
belongs_to common_drug
do_eat drugs_of
has_symptom need_check
no_eat recommand_drug
recommand_eat

7.2 Data mapping

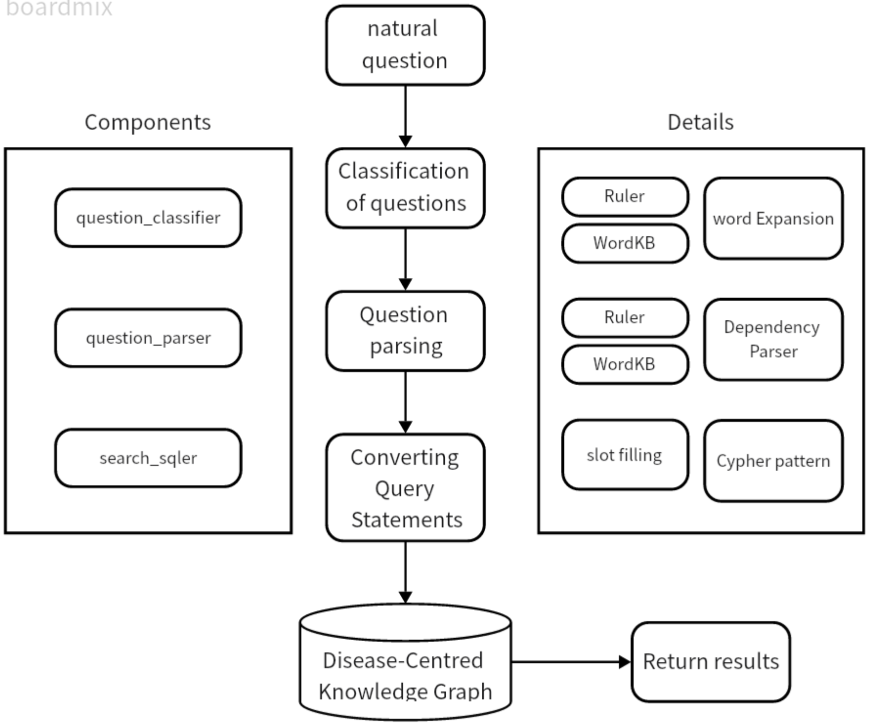
We uploaded the comprehensive data into the database of the final medical knowledge graph. A set of Entity Graphs (EGs) was generated, in which the disease entities are centered on the body part entities involved in the disease episodes, the body parts are in turn linked to the relevant department entities in the hospitals that perform the medical treatment, and on the other hand, according to the method of diagnosis and treatment, to the drug entities of the medication with the complementary types of ingredients such as dietary therapies.



8 Outcome Exploitation

In the results utilization session, we implemented a multi-question disease query system based on the constructed knowledge graph. We supported the Q&A service with cypher query statements as Q&A search sql. The following is the framework diagram of our Q&A system:

boardmix



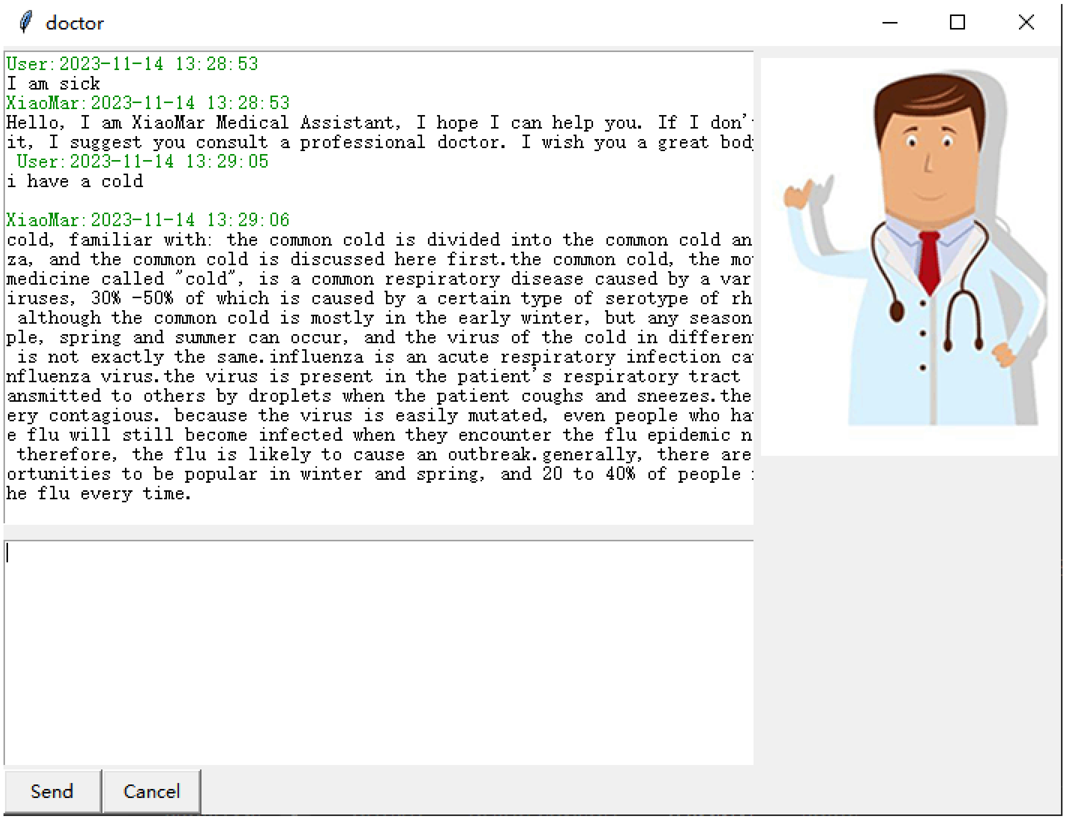
Knowledge graph-based Q&A framework

boardmix

For the files,question_classifier.py is used to perform the task of question type classification, question_parser.py performs the task of question parsing, and chatbot_graph.py is responsible for the operation of the entire question and answer program. The following table summarizes the medical issues supported by our program and the corresponding examples:

Type of question	Chinese Meaning	example
disease_symptom	疾病症状	乳腺癌的症状有哪些?
symptom_disease	已知症状找可能疾病	最近老流鼻涕怎么办?
disease_cause	疾病病因	为什么有的人会失眠?
disease_acompany	疾病的并发症	失眠有哪些并发症?
disease_not_food	疾病需要忌口的食物	失眠的人不要吃啥?
disease_do_food	疾病建议吃什么食物	耳鸣了吃点啥?
food_not_disease	疾病最好不要吃某事物	哪些人最好不好吃蜂蜜?
food_do_disease	食物对什么病有好处	鹅肉有什么好处?
disease_drug	啥病要吃啥药	肝病要吃啥药?
drug_disease	药品能治啥病	板蓝根颗粒能治啥病?
disease_check	疾病需要做什么检查	脑膜炎怎么才能查出来?
check_disease	检查能查什么病	全血细胞计数能查出啥来?
disease_prevent	预防措施	怎样才能预防肾虚?
disease_lasttime	治疗周期	感冒要多久才能好?
disease_cureway	治疗方式	高血压要怎么治?
disease_cureprob	治愈概率	白血病能治好吗?
disease_easyget	疾病易感人群	什么人容易得高血压?
disease_desc	疾病描述	糖尿病

The final results are presented in a visual interface, where the knowledge of the disease of interest is entered and the system gives feedback after querying. An example of its operation is shown below:



```
User:2023-11-14 13:28:53
I am sick
XiaoMar:2023-11-14 13:28:53
Hello, I am XiaoMar Medical Assistant, I hope I can help you. If I don'
it, I suggest you consult a professional doctor. I wish you a great bod
 User:2023-11-14 13:29:05
i have a cold

XiaoMar:2023-11-14 13:29:06
cold, familiar with: the common cold is divided into the common cold an
za, and the common cold is discussed here first.the common cold, the mo
medicine called "cold", is a common respiratory disease caused by a var
iruses, 30% -50% of which is caused by a certain type of serotype of rh
 although the common cold is mostly in the early winter, but any season
ple, spring and summer can occur, and the virus of the cold in differen
 is not exactly the same.influenza is an acute respiratory infection ca
nfluenza virus.the virus is present in the patient's respiratory tract
ansmitted to others by droplets when the patient coughs and sneezes.the
ery contagious. because the virus is easily mutated, even people who ha
e flu will still become infected when they encounter the flu epidemic n
 therefore, the flu is likely to cause an outbreak.generally, there are
ortunities to be popular in winter and spring, and 20 to 40% of people
he flu every time.
```

9 Conclusion and Open Issues

In the construction of this project,we followed the iTelos methodology technology route to build a disease-centered knowledge graph.And based on the knowledge graph,we discussed the downstream tasks related to it,and used the question and answer system as the project expansion point to finally realize our expected goal.

In the process of construction,we followed the technical framework design,and constructed many data into the knowledge graph,of course, there are still some problems that we have not realized in the process of implementation,in the future, we can talk about more sources of information in the health field to be fused to make the scale of the knowledge graph richer.We also hope that our knowledge graph can be used for more downstream tasks besides Q&A systems.