

CONTENTS

目录

- 新冠开放知识图谱构建及关键技术
- 跨数据集关联与融合
- 数据规范与统一访问接口
- 新冠开放知识图谱潜在应用与发展方向
- 新冠开放知识图谱相关数据竞赛

AI 正在向“认知智能”演进

国务院新一代人工智能发展规划

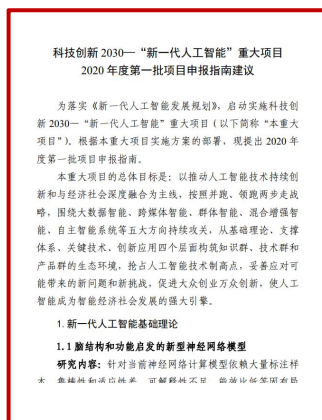
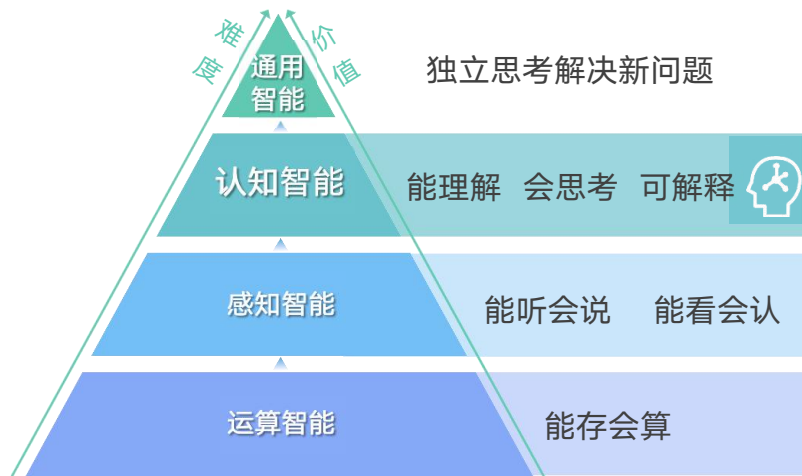
2017年7月8日 国务院明确提出了

建立新一代人工智能关键共性技术体系重点任务，

特别强调了研究跨媒体统一表征、关联理解与知识挖掘、知识图谱构建与学习、知识演化与推理等技术。

新一代人工智能重大项目2020年度第一批项目申报指南

“认知”出现 16 次，“知识图谱”出现 13 次
鼓励在金融、客服、教育、工业、医疗等领域构建行业知识图谱。



新冠开放知识图谱构建及关键技术

新冠开放知识图谱助力疫情防控



新冠开放
知识图谱

我想对新型冠状病毒和新冠肺炎有一个基本的认识...<百科>

1

得了新冠肺炎会有哪些症状，现在都怎么治疗的...<临床><健康>

2

个人防护、场所防控都权威的指导手册...<防控>

3

疫情中，物资现在是什么状态...<物资>

4

现在有没有研发出疫苗，有没有特效药呢...<科研>

5

有没有技术可以追溯传染源，探索被接触者...<流行病学>

6

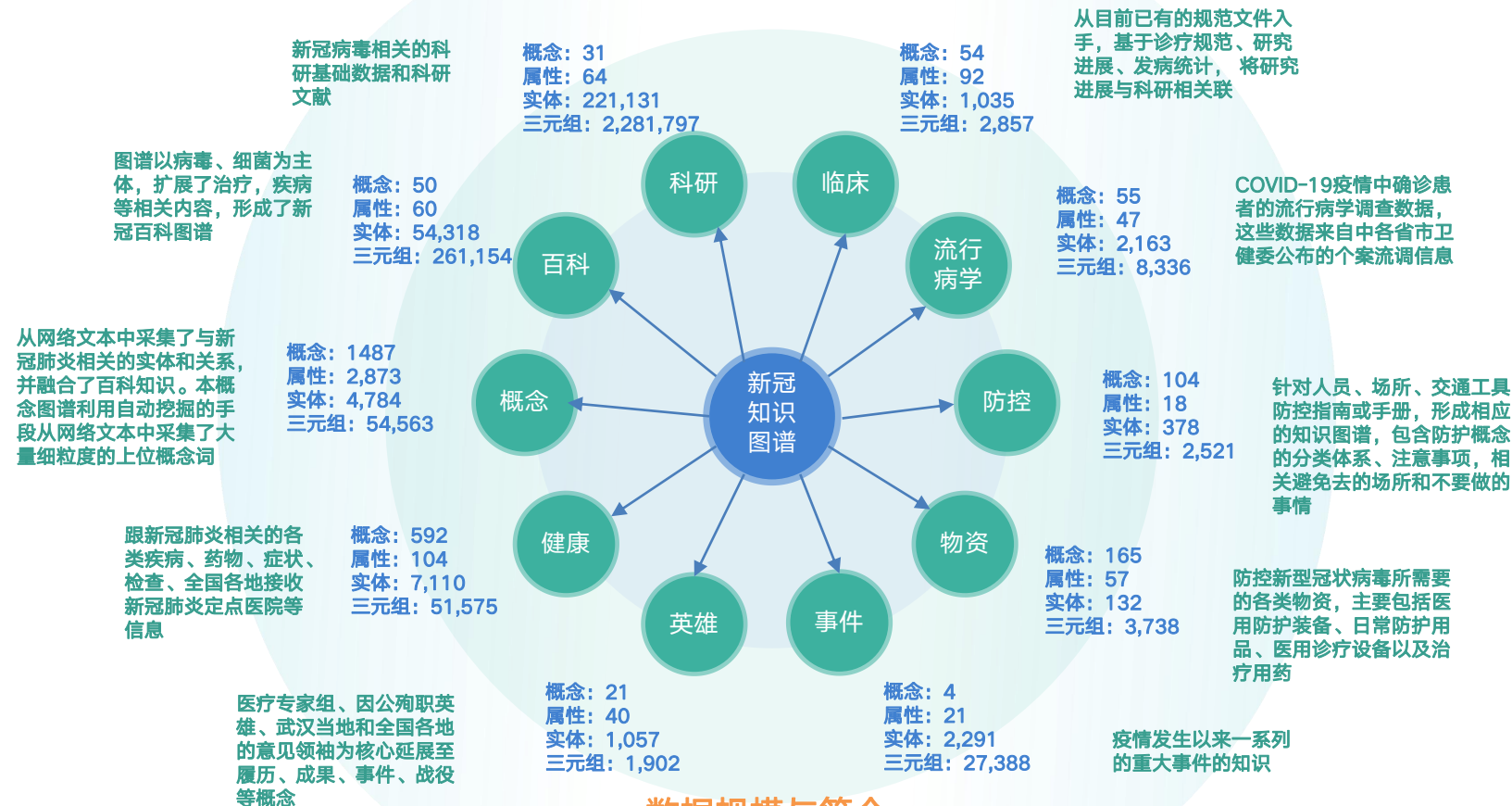
这场“战疫”中，有哪些伟大的英雄...<英雄>

7

有关疫情的热点事件有哪些...<事件>

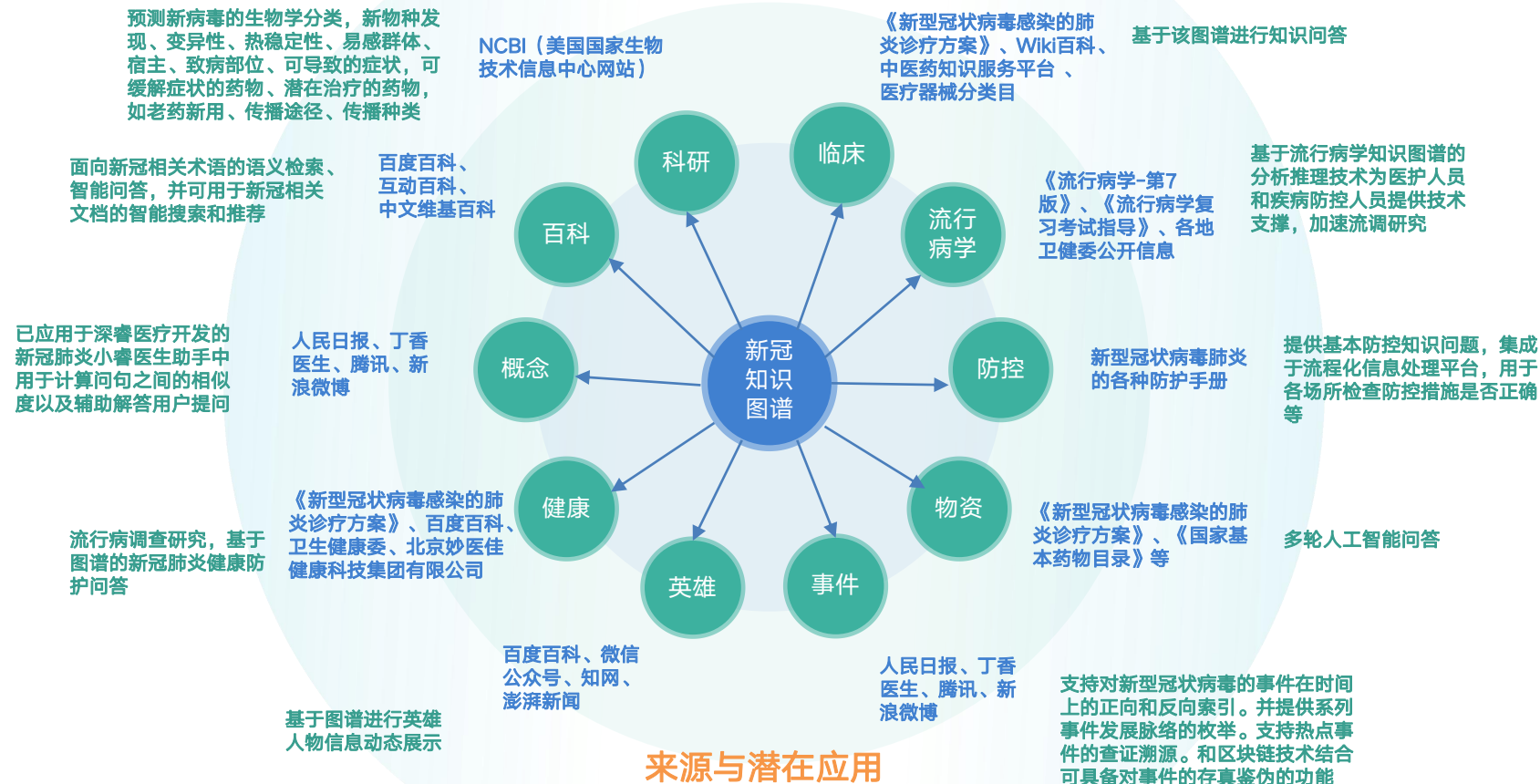
8

新冠开放知识图谱助力疫情防控

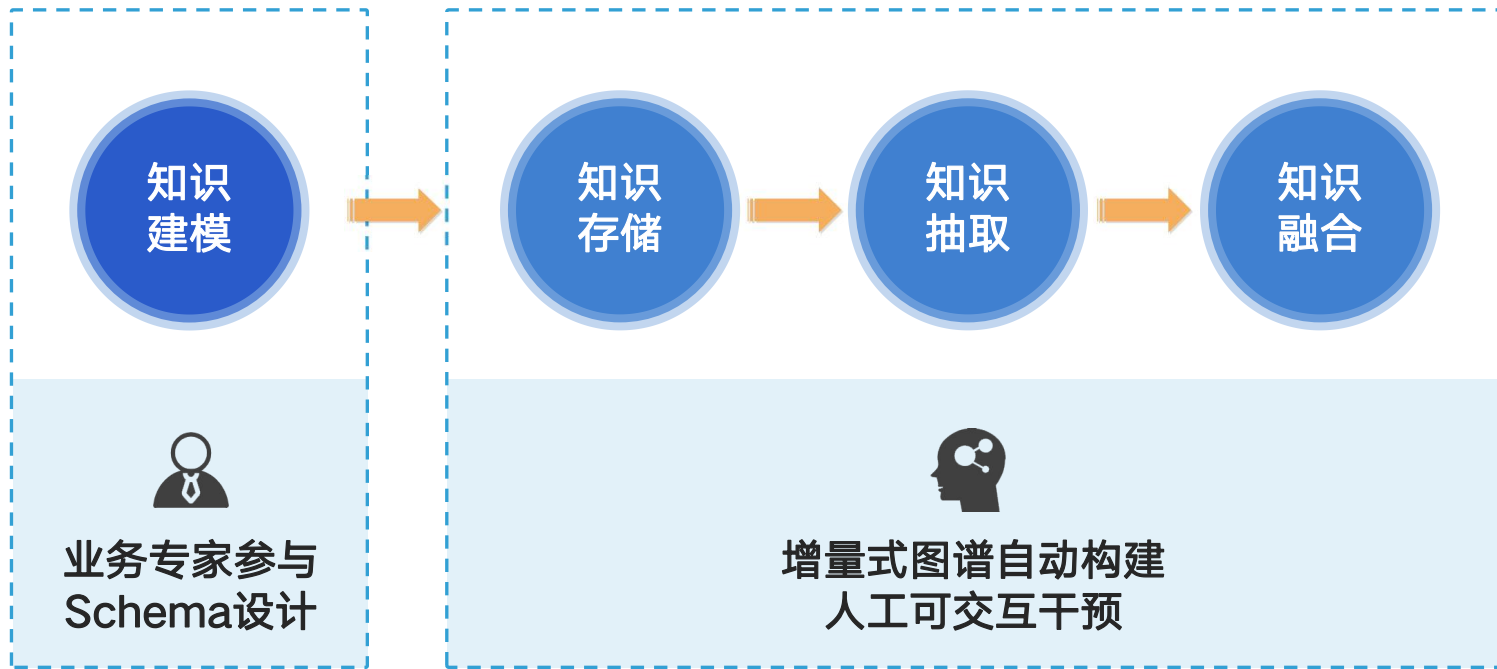


数据规模与简介

新冠开放知识图谱助力疫情防控

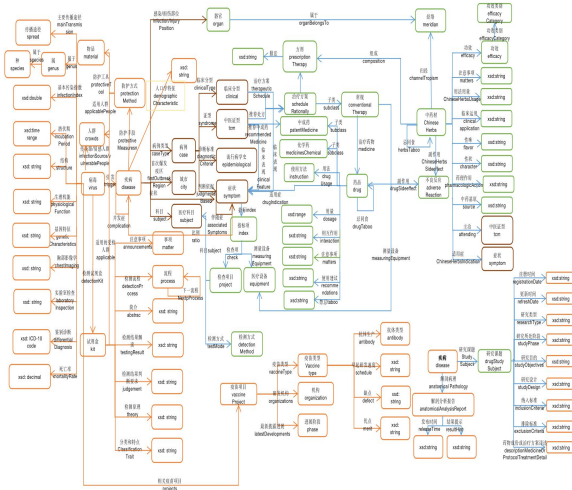


新冠开放知识图谱一构建流程



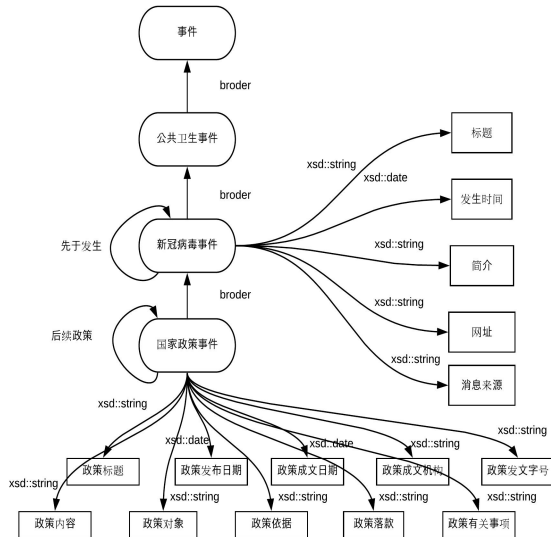
新冠临床图谱

以新冠肺炎疾病的患者病例、治疗方案、药物研究、中西药物作用等为重点



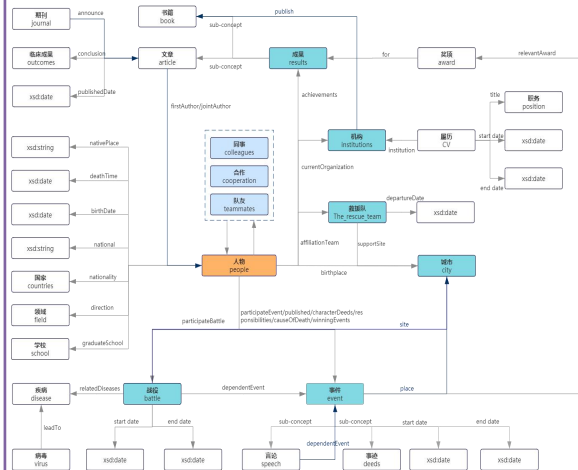
新冠事件图谱

以政策、发文机构、发布日期、依据、标题、对象为重点概念，包括顺承、组成、前序关系



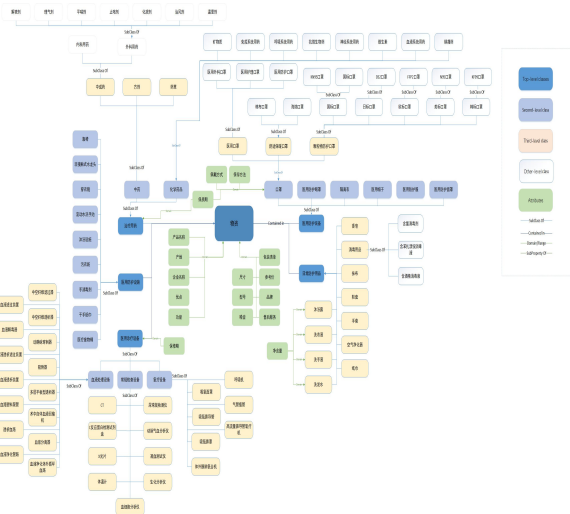
新冠英雄图谱

以人物、战役、事件、成果、文章、履历等为重点概念



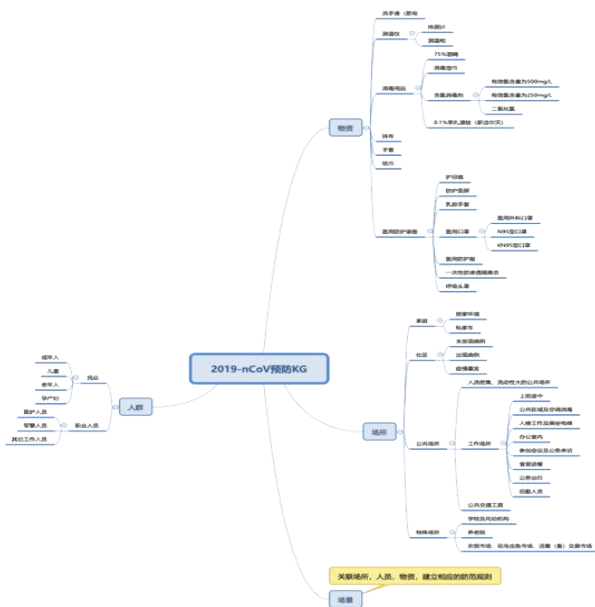
新冠物资图谱

以医用防护装备、日常防护用品、医用诊疗设备以及治疗用药为重点



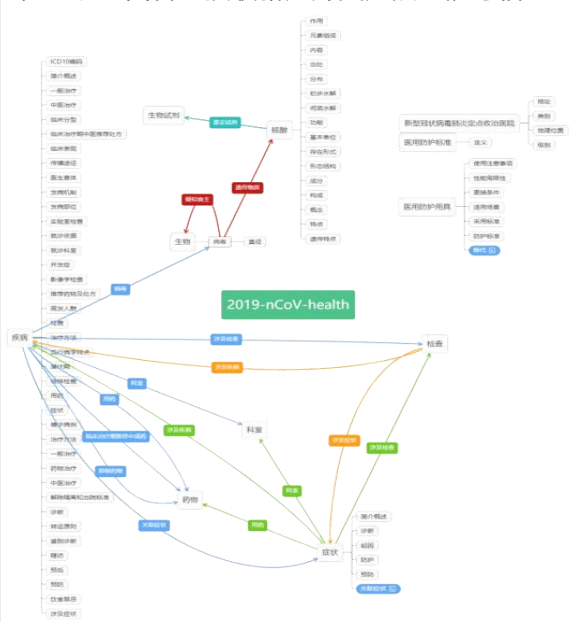
新冠防控图谱

以医用防护装备、日常防护用品、医用诊疗设备以及治疗用药为重点

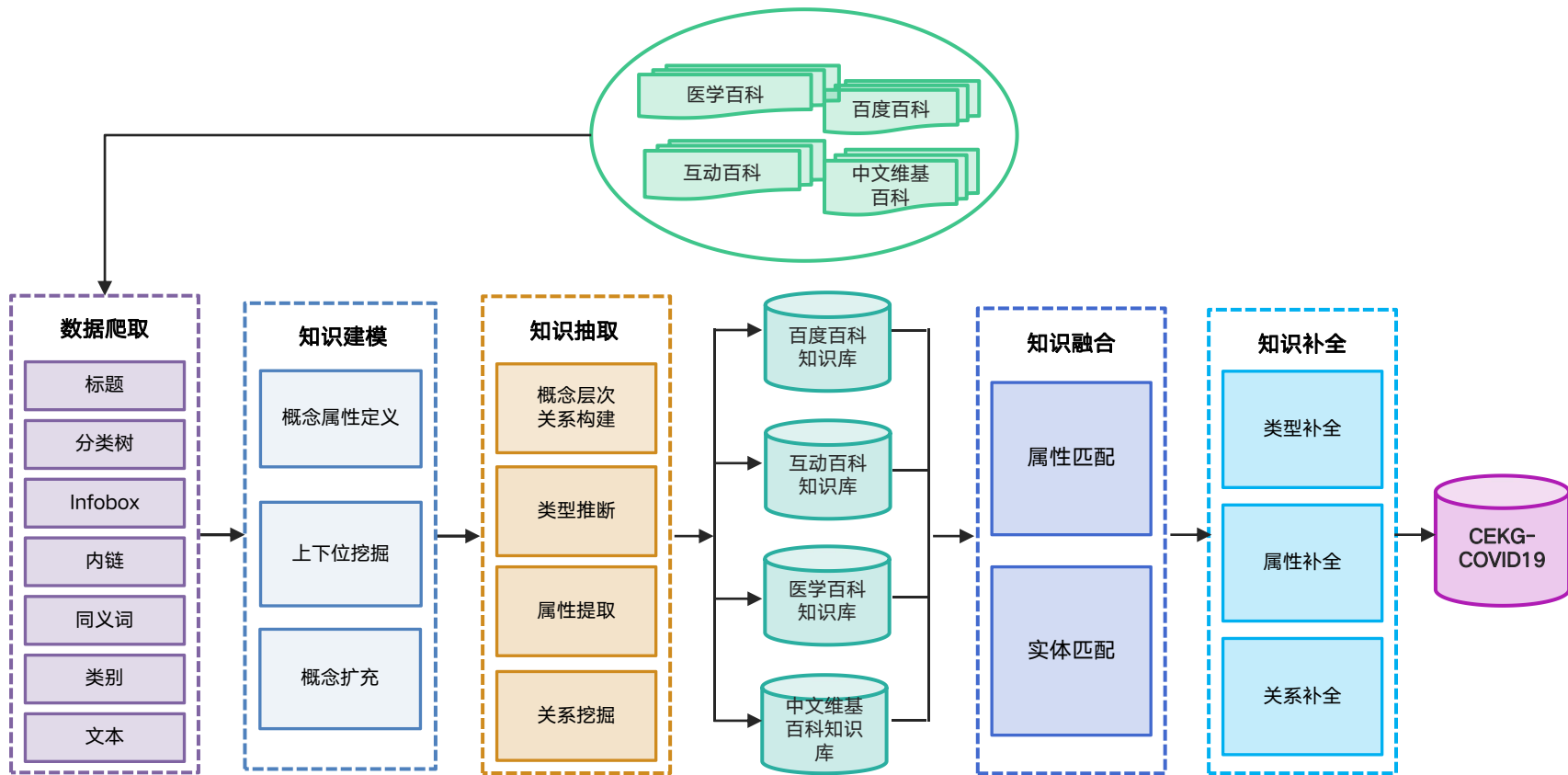


新冠健康图谱

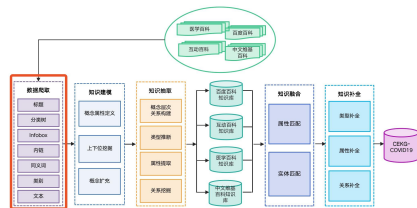
包含和新冠肺炎相关的各类疾病、药物、症状、检查、全国各地接收新冠肺炎定点医院等信息



新冠开放知识图谱—构建流程（以新冠百科图谱构建过程为例）



新冠百科知识图谱—数据爬取



1 内部链接 (InnerLink)

在本病的诊断中病史极为重要，往往曾反复发生 [鼻窦炎](#)、[支气管炎](#) 或肺炎，或曾患 [麻疹](#)、[百日咳](#)、[流行性感](#)或 [腺病毒肺炎](#)。确定诊断 [需要](#) [结合](#) [病史](#)、[症状](#) 和 [丝](#) [检查](#)。

2 分类树

百科分类树 知识地图

搜分类

病毒

显示树型结构

- 病毒 [收起](#)
 - DNA病毒 [收起](#)
 - + 双链DNA病毒 [展开](#)
 - + 单链DNA病毒 [展开](#)
 - RNA病毒 [收起](#)
 - + 双链RNA病毒 [展开](#)
 - + 正链RNA病毒 [展开](#)
 - + 负链RNA病毒 [展开](#)
 - + 逆转录病毒 [展开](#)

3 类别

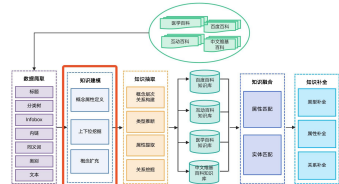
异烟肼磺酸钠

开放分类: 医学 | 处方药 | 药品 | 西药

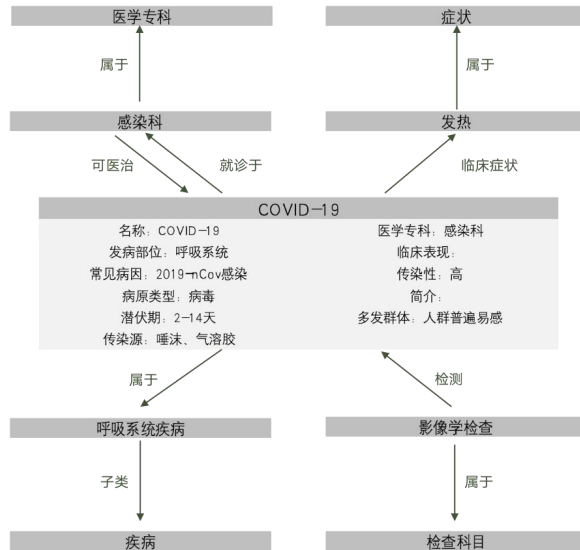
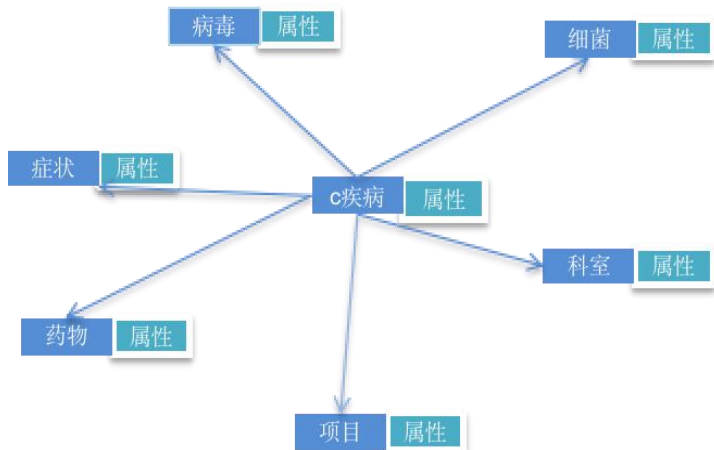
异烟肼磺酸钠，别名 甲磺烟肼钠、异烟肼磺酸钠、异烟肼甲磺酸钠，用于各型结核病的治疗。口服或肌注：0.1~0.3g/次，2~3次/日。

编辑摘要

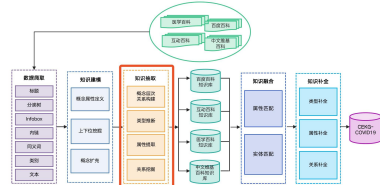
新冠百科知识图谱—知识建模



从COVID-19的基本信息、症状、致病原因、治疗药物、检查项目、诊治科室六个切入点，选取疾病、症状、细菌等七个顶级概念，获取四大百科category system中这七大概念的所有子概念，上下位关系取四大百科的交集部分，对齐BabelNet，并以此扩充概念，构建新冠百科知识图谱Schema。



新冠百科知识图谱—知识抽取



类型推断(Type Inference)

- 句法分析
- Infobox 模板
- 外部知识库
- 观察特征

2019冠状病毒病^{[9][10]}（英语：Coronavirus disease 2019，缩写：COVID-19^{[11][12][13]}）是由严重急性呼吸系统综合征冠状病毒2（缩写：SARS-CoV-2）引发的传染病。

维基百科的Infobox模板拥有**class**属性

CDD+ 临床医学知识服务系统

病毒资源数据库
Database of Virus Resource

微生物资源数据库共享平台

附桂紫金膏 <附桂紫金膏, TypeOf, 药物>

<肺炭疽, 潜伏期, 一般为1~5天> ➡ <肺炭疽, TypeOf, 疾病>

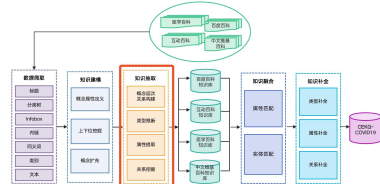
异烟肼磺酸钠

开放分类：医学 | 处方药 | 药品 | **西药**

异烟肼磺酸钠，别名 甲磺烟肼钠、异烟肼磺酸钠、异烟肼甲磺酸钠，用于各型结核病的**治疗**。口服或肌注：0.1~0.3g/次，2~3次/日。

编辑报

新冠百科知识图谱—知识抽取



属性与属性值提取



结构化数据

Infobox获取属性和属性值对

传播途径:

近距离飞沫传播或接触患者呼吸道分泌物传播

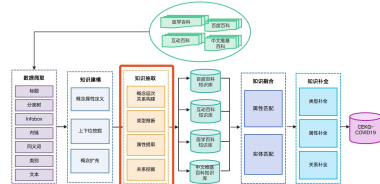
属性与
属性值



非结构化数据

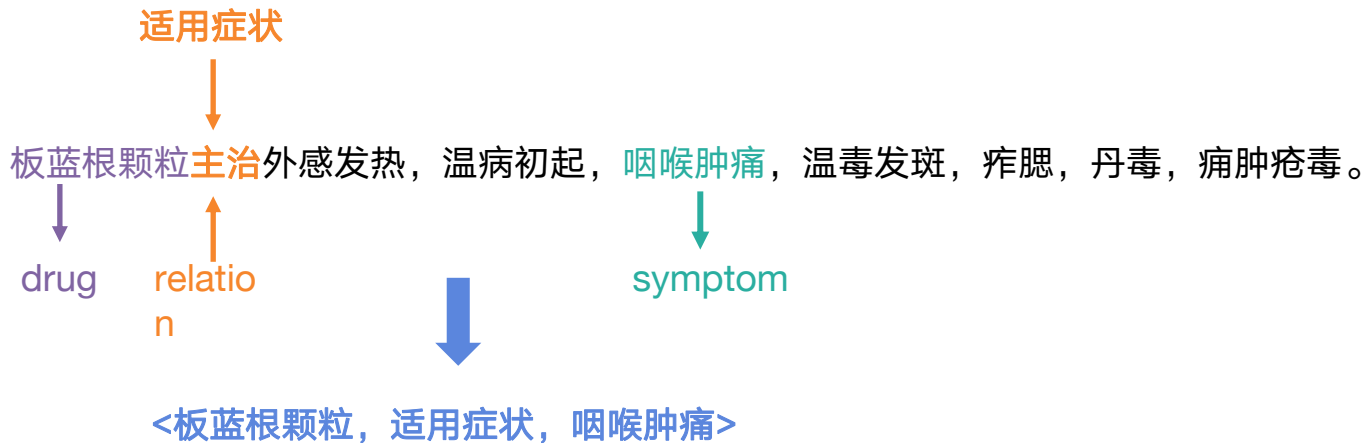
- 1.网页抽取特定属性和属性值同时出现的句子，“该疾病的传播途径包括近距离飞沫传播、呼吸道分泌物传播等。”
- 2.把1.中抽取的句子作为正例，其他作为负例，训练最大熵模型
- 3.通过模型判断出潜在正例句子，使用CRF标注属性值，如“COVID主要传播途径有唾液，呼吸道，密切接触。”

新冠百科知识图谱—知识抽取

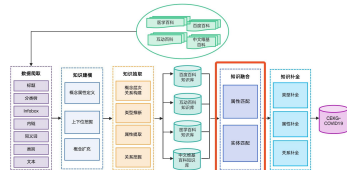


关系提取

1. 抽取潜在relation语义的句子 (Regex Matchig)
2. Bert+BiLSTM+CRF进行实体识别
3. 三元组构建



新冠百科知识图谱—知识融合



属性融合

- 属性的同义词库
- 谓语句频数统计与归纳

实例：重症急性呼吸综合征

百度百科

别称：严重急性呼吸综合征、SARS、非典、非典型肺炎
英文别名：severe acute respiratory syndrome, SARS
就诊科室：感染科
常见病因：SARS冠状病毒感染
常见症状：发热、头痛、肌肉酸痛、呼吸衰竭
传染性：有
传播途径：近距离飞沫传播或接触患者呼吸道分泌物传播

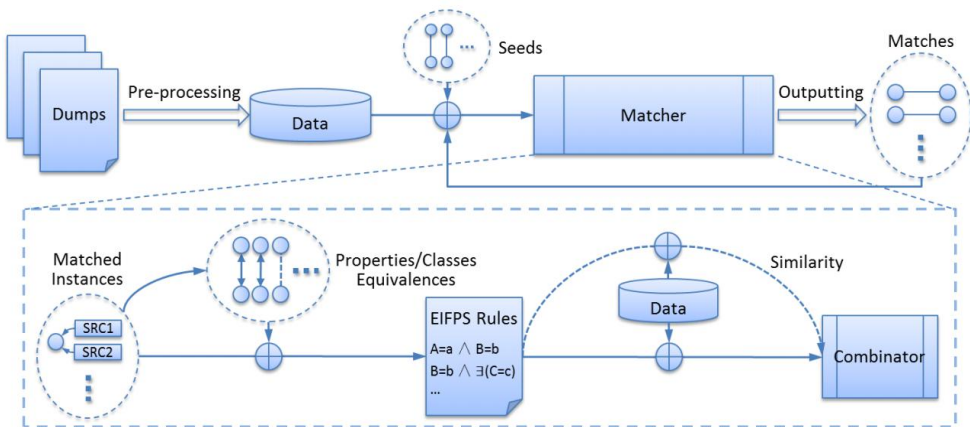
互动百科

中文名：重症急性呼吸综合征
外文名：severe acute respiratory syndrome
别名：严重急性呼吸综合征、SARS、非典、
季节分布：四季
传染病：是
传播途径：近距离飞沫传播或接触患者呼吸道分泌物传播
临床表现：发热、头痛、肌肉酸痛、呼吸衰竭
就诊科室：传染科
常见病因：SARS冠状病毒感染
英文别名：severe acute respiratory syndrome, SARS

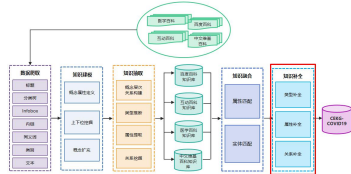
实体融合

- 挖掘等价规则

首先从种子实例中挖掘出等价属性，如“别名”和“别称”、“常见症状”和“临床表现”等是等价属性，然后根据已知的等价实体的等价属性的出现频率情况，挖掘匹配规则。



新冠百科知识图谱—知识补全



属性值补全

- 从Infobox中抽取特定属性与属性值
- 在页面中抽取特定属性和属性值同时出现的句子
- 这些句子作为正例，其他句子作为负例训练分类器，以预测某一句子是否可能包含特定属性的属性值。
- 对于预测包含属性值的句子，使用标注方法对属性值进行预测

传播途径：

近距离飞沫传播或接触患者呼吸道分泌物传播

match

该疾病的传播途径包括近距离飞沫传播、呼吸道分泌物传播等。

train

train

分类器

标注器

predict

tag

类型补全

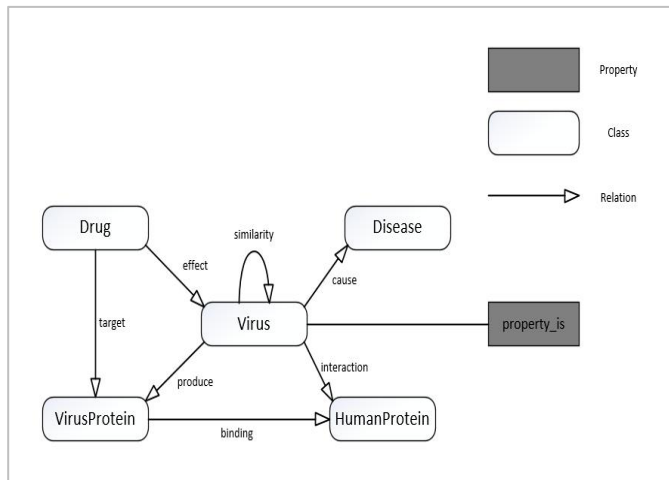
- 特定关系所链接的两个实体类型可预测。

关系补全

- 基于title和inner link（内链）之间的context。

COVID-19主要传播途径有唾液，呼吸道，密切接触。

其他典型图谱知识抽取举例（科研图谱）

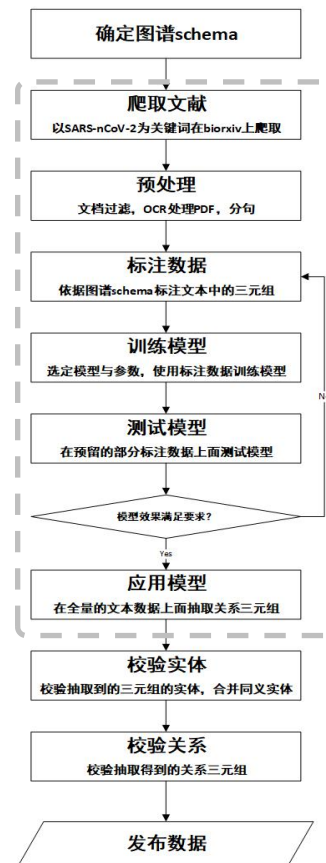


数据源是期刊论文，从病毒及与病毒相关的多种类型生物分子/化学分子出发，梳理其关系，构建新冠科研-基本信息图谱。

```

{
  "text": "Binding interactions of 2019-nCoV S-RBD with human ACE2 With the complex model in Fig.",
  "spo_list": [
    {
      "predicate": "binding",
      "subject_type": "VirusProtein",
      "object_type": "HumanProtein",
      "subject": "2019-nCoV S-RBD",
      "subject_position": 24,
      "object": "ACE2",
      "object_position": 51
    }
  ]
}
  
```

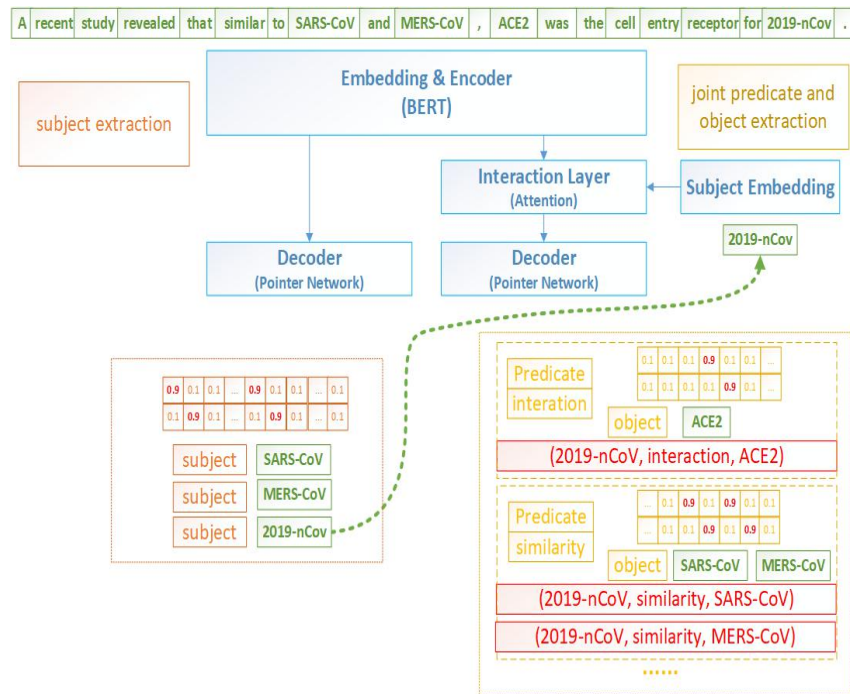
标注数据样例



其他典型图谱知识抽取举例（科研图谱）

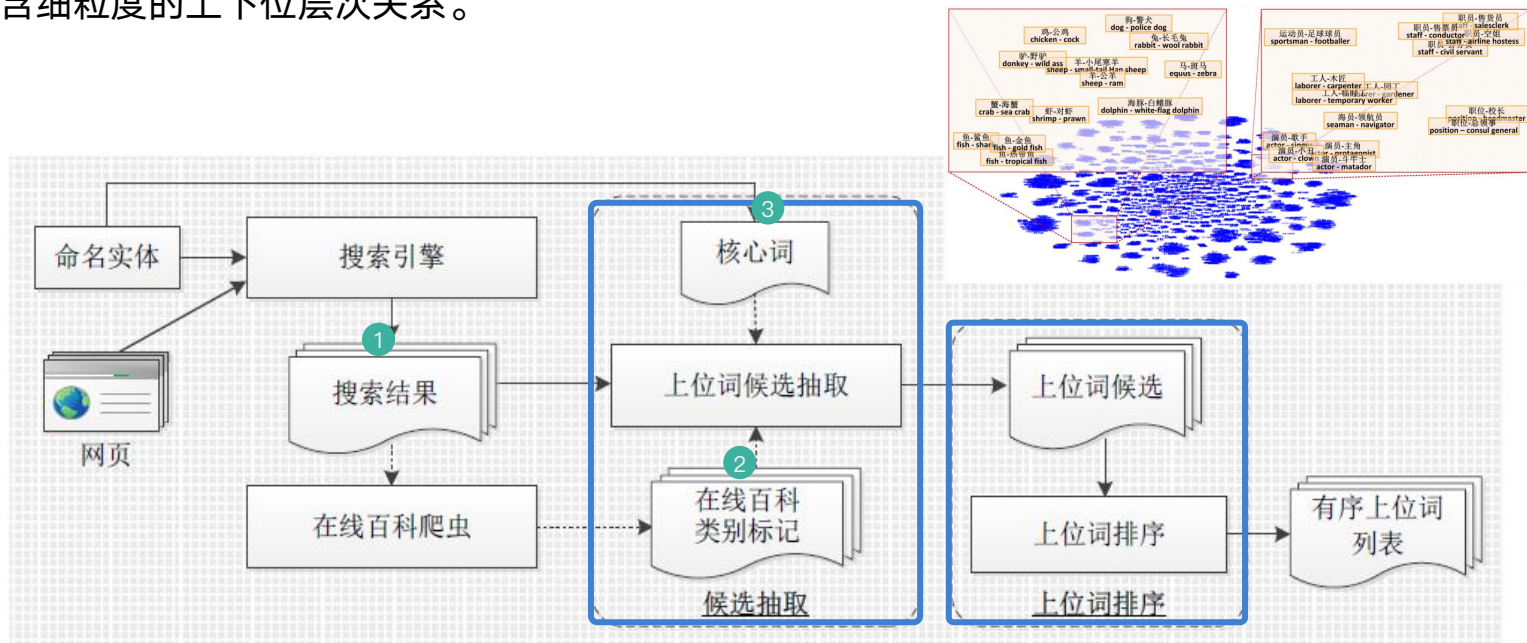
基于《A Unified MRC Framework for Named Entity Recognition》，提出了一个从医疗文献中抽取实体关系的模型。这个模型同样由两部分组成：**主语抽取部分**(subject extract)和**谓词宾语联合抽取部分**(joint predicate and object extraction)，如右图所示。

- 在主语抽取部分，模型将会抽取能够作为三元组主语的实体，输出部分是一个**双层指针网络**，标识出实体的头尾位置；
- 在谓词宾语联合抽取部分，模型将基于前一部分抽取到的主语，抽取出对应的宾语并**判断关系**。一个**注意力层**，一边的输入是**句子的BERT输出**，另一半的输入是这个**主语在句子中的BERT输出**。最后输出层是2*N层的指针网络（N是所需抽取的关系数量）。每一类关系与2层指针网络相对应，标识出宾语的头尾位置。



其他典型图谱知识抽取举例（概念图谱）

从网络文本中自动挖掘大量细粒度的概念词，自动构建概念的is-a层次结构（schema），富含细粒度的上下位层次关系。



新冠开放知识图谱—质量评估

对新冠开放知识图谱每个数据集进行了数据质量评测，针对每种关系随机采样一定数目的三元组，由对数据熟悉的人员或者专业人员评测。为了将采样评测结果扩展到整个数据集，我们计算了置信度为95%的威尔逊区间，评测结果如下：

图谱名称	数据规模	采样比例	采样数目	正确数目	准确率预估
百科	261,154	1.91%	5,000	4,778	95.52%±0.58%
临床	2,857	22.82%	652	620	94.81%±1.71%
科研	2,281,797	0.37%	8556	8555	99.96%±0.03%
事件	27,388	0.73%	200	198	96.35%±1.69%
英雄	1,902	29.97%	570	570	99.65%±0.35%
防控	2,521	17.29%	436	434	99.09%±0.79%
物资	3,738	9.76%	365	359	97.83%±1.42%
健康	51,575	0.94%	487	483	98.78%±0.91%
概念	54,563	1.20%	100	96	92.31%±4.96%
流行病学	8,336	2.4%	200	200	98.08%±1.92%

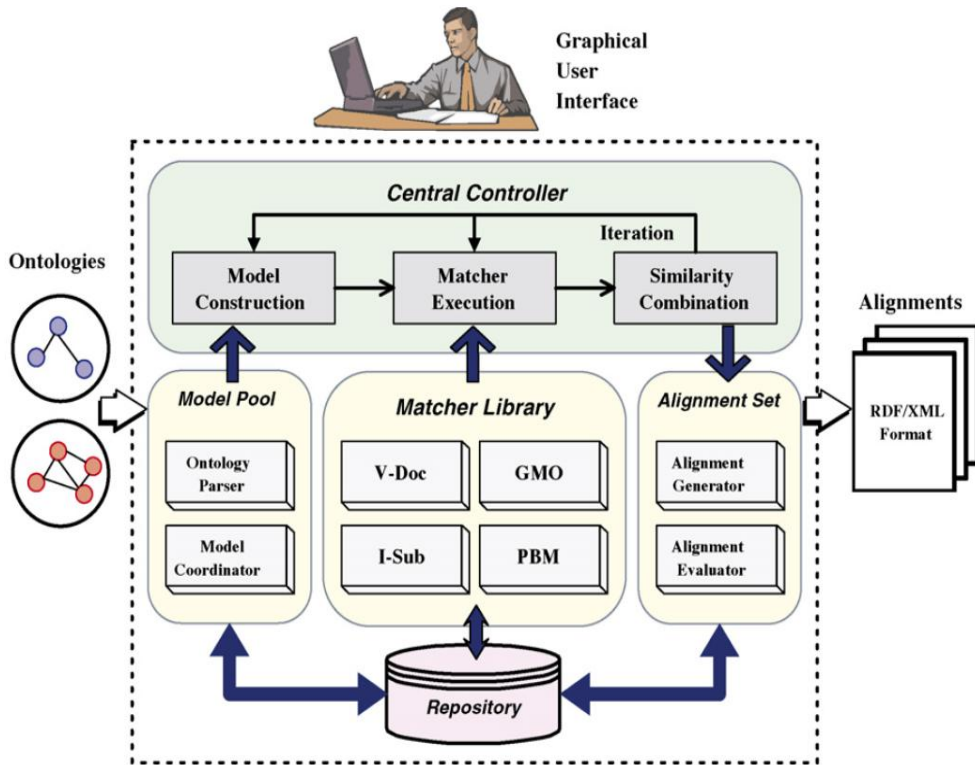
跨数据集关联与融合

新冠开放知识图谱一本体匹配

由于没有中心本体，我们对各个图谱的本体成对匹配。采用了南京大学研发的Falcon-AO工具，利用文本和结构特征进行自动化本体匹配

匹配器：

- V-Doc：采用了语言学的方法来做本体匹配，其最大的亮点在于，构建了虚拟文档
- I-Sub：将字符串的共性和差异结合考虑，使用字符串距离的度量方法比较本体的类名和属性名
- GMO是一种迭代式的结构的匹配方法，它使用RDF二部图来表示本体，计算类或属性以及三元组之间的结构相似度
- PBM使用分而治之的方法来寻找大规模本体之间的块映射



表中结果斜线左侧是类的个数，右侧是属性的个数，结果显示各个新冠知识图谱本体有重叠，但是重叠数量较少。

类/属性	百科	防控	概念	健康	科研	临床	流行病	事件	物资	英雄
百科	0/0	37/0	9/9	2/4	4/6	4/1	0/0	0/0	0/0	0/0
防控										
概念										
健康										
科研										
临床										
流行病										
事件										
物资										
英雄										

新冠开放知识图谱—实体对齐

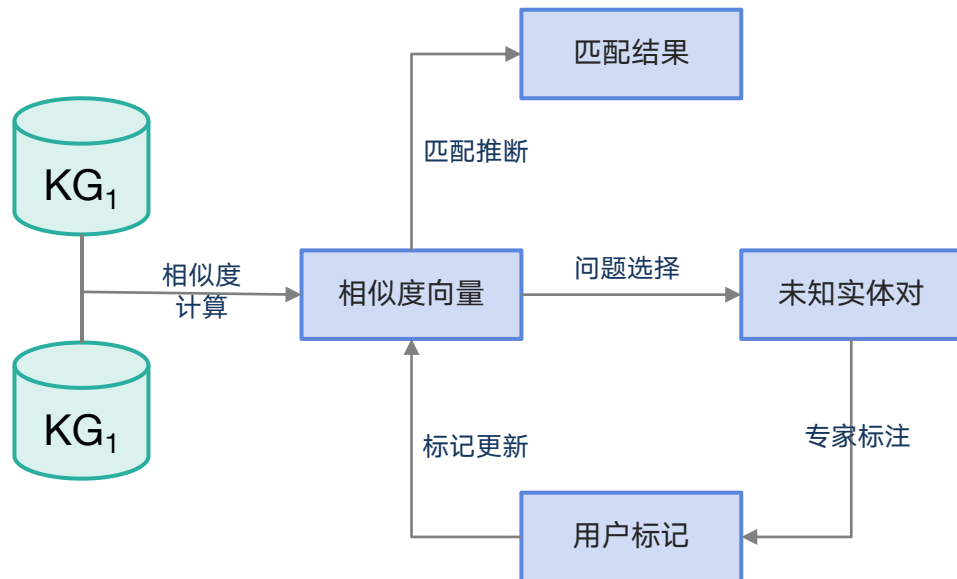
基于实体对相似度向量的主动学习方法

动机

- 实例数据属性以Datatype Property为主
- 缺少训练数据

方法

- 基于属性对齐的相似度计算
- 基于偏序的问题选择
- 基于偏序的匹配推断

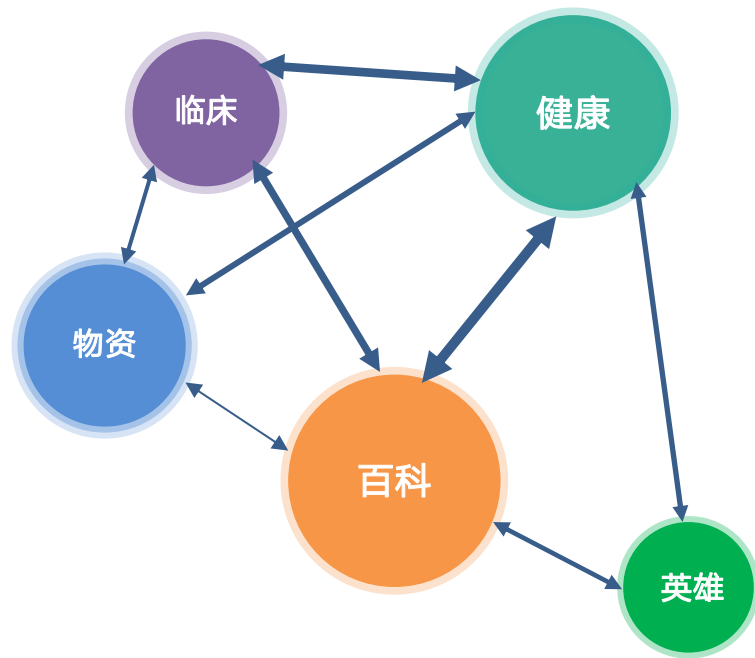


新冠开放知识图谱—实体对齐结果

实体对齐结果：

- 总计**1055**个匹配
- 绝大多数匹配的类是**药物**
- 百科数据集自身存在匹配（已有owl:sameAS）
- **百科**和其他数据集之间的匹配较多
- 百科和物资图谱匹配较少
- 健康和英雄图谱存在匹配（医疗方向）

	健康	临床	英雄	物资
百科	836	55	11	2
健康		95	19	28
临床				9



数据规范与质量评估

新冠开放知识图谱—命名规范

资源URI格式:

- 概念: `http://{namespace}/{graphname}/class/{localname}`
- 实体: `http://{namespace}/{graphname}/resource/{localname}`
- 属性: `http://{namespace}/{graphname}/property/{localname}`
- Namespace统一采用: `http://www.openkg.cn/2019-nCoV/`

LocalName采用类Wikidata编码方式:

- 概念: C + ID (Integer)
- 实体: R + ID (Integer)
- 属性: P + ID (Integer)
- 概念: `http://www.openkg.cn/2019-nCoV/character/class/C10`
- 实体: `http://www.openkg.cn/2019-nCoV/character/resource/R101`
- 属性: `http://www.openkg.cn/2019-nCoV/character/property/P5`

RDF文件格式:

支持RDF/XML、N-TRIPLE、TURTLE、N3、JSON-LD

新冠开放知识图谱命名:

新冠临床	→	medical
新冠英雄	→	character
新冠研究	→	research
新冠百科	→	wiki
新冠防控	→	control
新冠事件	→	event
新冠流行病	→	epidemiology
新冠健康	→	health
新冠物资	→	goods
新冠概念	→	concept

新冠开放知识图谱—三元组声明规范

01 声明声称

谓语: <http://www.w3.org/2000/01/rdf-schema#label>

02 声明概念

谓语: <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

宾语: <http://www.w3.org/2000/01/rdf-schema#Class>

03 概念继承

谓语: <http://www.w3.org/2000/01/rdf-schema#subClassOf>

04 声明实体

谓语: <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

05 声明属性

谓语: <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

06 声明实体间同义

谓语: <http://www.w3.org/2002/07/owl#sameAs>

07 声明单实体同义词

谓语: <http://www.openkg.cn/2019-nCoV/ontology/alias>

08 声明属性domain

谓语: <http://www.w3.org/2000/01/rdf-schema#domain>

09 声明属性range

谓语: <http://www.w3.org/2000/01/rdf-schema#range>

新冠开放知识图谱—三元组声明规范

对于对象属性，宾语使用概念URI

对于数值属性，宾语使用如下：

文本	http://www.w3.org/2001/XMLSchema#string	✓
整数	http://www.w3.org/2001/XMLSchema#integer	✓
浮点数	http://www.w3.org/2001/XMLSchema#double	✓
日期	http://www.w3.org/2001/XMLSchema#date	✓
时间	http://www.w3.org/2001/XMLSchema#time	✓
日期时间	http://www.w3.org/2001/XMLSchema#datetime	✓

新冠开放知识图谱统一访问接口

新冠开放知识图谱—OpenKG

OpenKG 发布的所有新冠知识图谱采用 CC-by SA 相似署名开放许可协议，供大家免费下载使用。

欢迎大家访问新冠图谱专题链接，获取更多新冠知识图谱：

<http://openkg.cn/group/coronavirus>

同时可以通过访问下面链接，得到新冠图谱数据规范以及任何新冠图谱schema：

<http://openkg.cn/dataset/covid-19-schema>



The screenshot shows the OpenKG website interface. The top navigation bar includes links for 图谱 (Graphs), 工具 (Tools), 机构 (Institutions), 分类 (Categories), 加入 (Join), and 关于 (About), along with a search bar. The main content area displays search results for '新冠开放知识图谱.科研' (COVID-19 Open Knowledge Graph. Research). The results list two datasets: '新冠基本分类图谱v2.0' and '新冠基本分类图谱v1.0'. The first dataset is described as a virus classification graph based on NCBI data, and the second is a virus classification graph based on DrugBank data. The results are displayed in a table with columns for 资源集 (Dataset), 活动流 (Activity Stream), and 关于 (About). The table shows 11 datasets found, with a search bar and a dropdown menu for sorting by 相关性 (Relevance).

新冠开放知识图谱—OpenBase



ELSEVIER

在OpenBase可以搜索任何粒度的实体，
搜索结果会展示出该实体的属性和关系

欢迎大家使用OpenBase：
<http://openbase.openkg.cn/>



介绍

愿景

数据概况

OpenBase旨在促进中文知识图谱数据的开放与
互联，促进知识图谱和语义技术的普及和广泛应
用。

三元组数
107,264,094

实体数
11,513,944

属性数
92,247,120

任务数
47,824

功能

你可以通过简单的描述来创建图谱需求、编辑实体条目。由知识图谱专家与行业专家来共同定义数据模式、审核项目，确保图谱的
正确性与可用性。同时你也可以搜索或下载图谱。



创建图谱



编辑图谱



审核图谱



下载图谱

游客

图谱搜索

图谱下载

浏览图谱

KG4AI

验收图谱

KG4AI

新冠知识图谱

实体信息-钟南山

summary

钟南山，1936年10月出生，福建 [b4179626] (厦门) 人，生于 [b4169726] (南京)，出身医学世家。 [b4203565] (呼吸病学) 专家。于1960年毕业于北京医学院 (今北京大学医学部)，2007年获美国爱丁堡大学荣誉博士学位。中国工程院院士。 [b4278616] (教授)，博士生导师。2003年赴任“非典”先进个人。 [b4278616] (中华医学会) 顾问、广州呼吸疾病研究所所长、广州市科委主席、广东省科协副主席、香港中文大学 (深圳) 理事会成员等职。主要从事高危/低危与肿瘤关系研究，曾由国家委有突出贡献专家，先后担任中华医学会呼吸分会主任委员、联合国 世界卫生组织的感染与健康医学顾问、国际胸科学会特别会员、亚太地区执委会理事。 [b4278616] (2016年6月1日)，中国工程院院士、著名呼吸病学专家钟南山 获颁第十一届中国工程科技成就奖。

中文名	钟南山
职业	医生
信仰	共产主义
主要成就	中国工程院院士
出生地	江苏南京
@name	钟南山
民族	汉族
appleID	b4b7486156
出生日期	19
毕业院校	北京医学院
国籍	福建厦门
uuid	a24bc84b302c4824b4b9b993b94fc2
国籍	中国

新冠开放知识图谱—OpenBase

新冠的数据还接入了OpenBase的**审核功能**，你可以联系我们**成为**OpenBase的数据审核员，可以在Openbase的网站上查看数据的审核，我们有**关系类审核**和**实体审核**两种众包任务。

尚忆秋
游客,超级管理员,审核员

个人信息

图谱搜索

图谱下载

浏览图谱
KG4AI

审核图谱
KG4AI

农业知识图谱

百科人物知识图谱

佛学知识图谱

新冠知识图谱

验收图谱
KG4AI

新冠知识图谱

人员管理

正在审核 —— 新冠知识图谱 关系

请审核以下实体之间的关系：

00:00:10

序号	实体1	关系 / 属性	实体2	操作
6f7ca7e828b546 8ca3dee9d9e8c9 bb8a_instanceOf	脊髓灰质炎后 遗症	instanceOf	疾病	正确 错误 关系错误

不知道

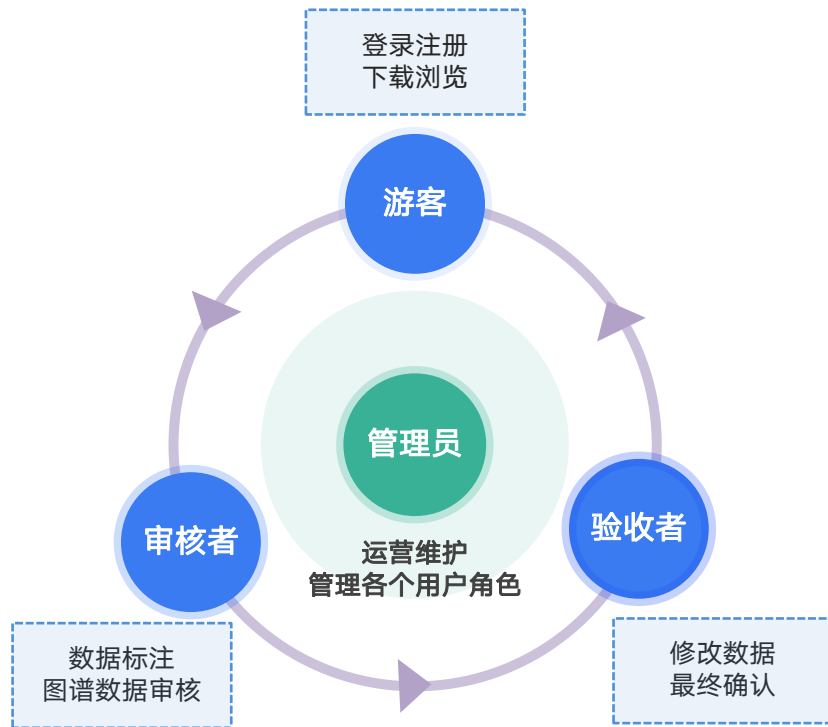
| 6f7ca7e828b546 8ca3dee9d9e8c9 bb8a_来源 | 脊髓灰质炎后 遗症 | 来源 | 互动百科 | 正确 错误 关系错误 |

不知道

| 6f7ca7e828b546 8ca3dee9d9e8c9 bb8a_名称 | 脊髓灰质炎后 遗症 | 名称 | 脊髓灰质炎后遗症 | 正确 错误 关系错误 |
 不知道

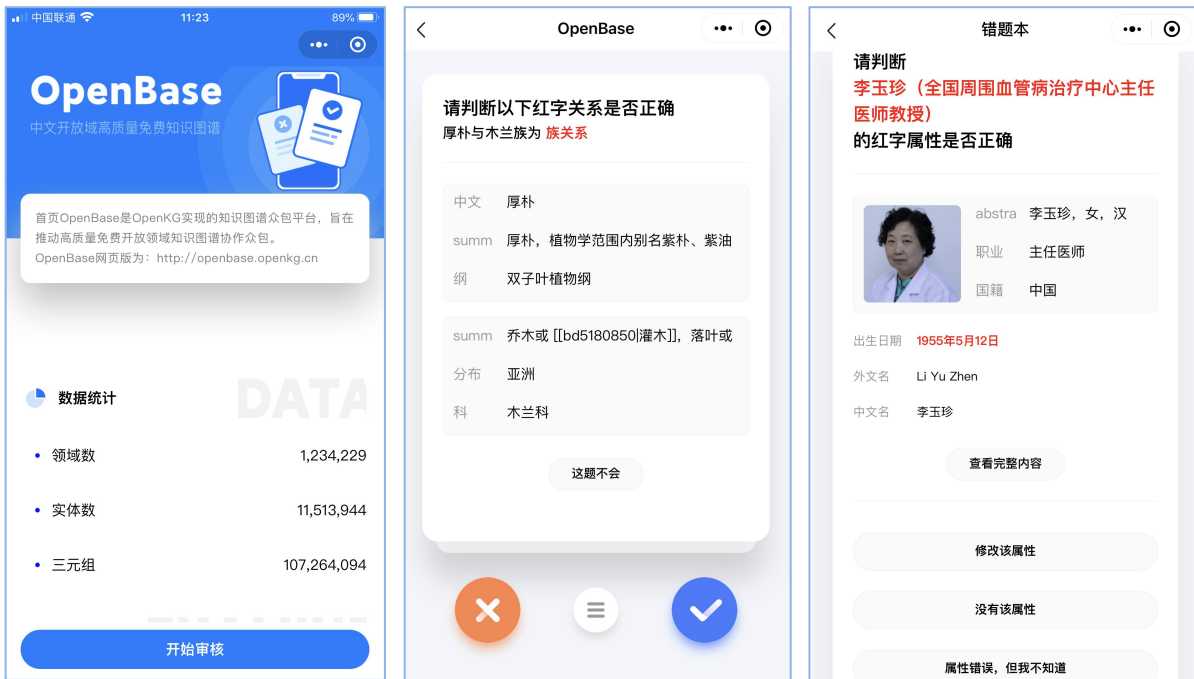
新冠开放知识图谱—OpenBase

OpenBase整体的众包，我们会抽样选择出新冠数据可能不准确的部分，然后切割成最简单的任务供志愿者来一起审核，分为游客、审核者、验收者、管理员四种角色。



新冠开放知识图谱—OpenBase

为了方便大家碎片化时间也可以做众包任务，我们还做了微信小程序版本，在这里大家可以随时在手机上进行标注。



新冠开放知识图谱—SPARQL Endpoint

SPARQL Endpoint提供在线查询，同时也可以点击“Click to Download”下载json格式的查询结果。

访问链接：

<http://covid.gstore-pku.com/>

COVID Query Endpoint


[About API](#)

Database Name

Query Examples

Query Text

```
select ?x ?y
where
{
  <新型肺炎> ?x ?y.
}
```

Results Format

Run Query

[Click to Return](#) [Click to Download](#)

Total answers: 3

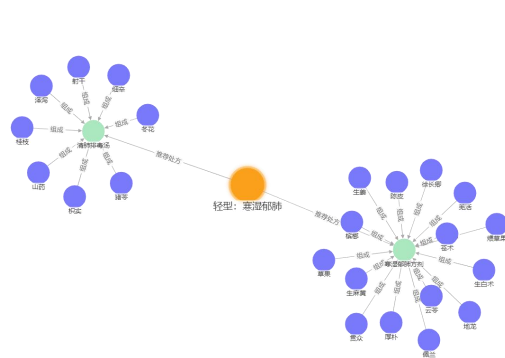
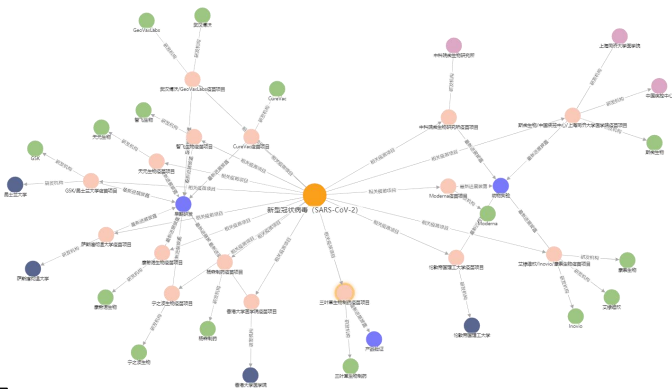
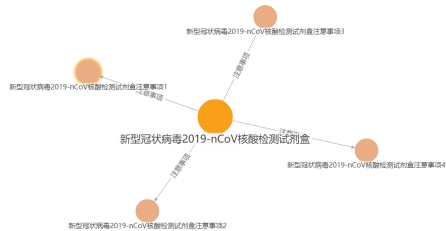
Query time: 3ms

?x	?y
<标签>	"新型肺炎"
<类型>	<疾病>
<openkg_uri>	"http://www.openkg.cn/COVID-19/character/resource/R59"

新冠知识图谱潜在应用与发展方向

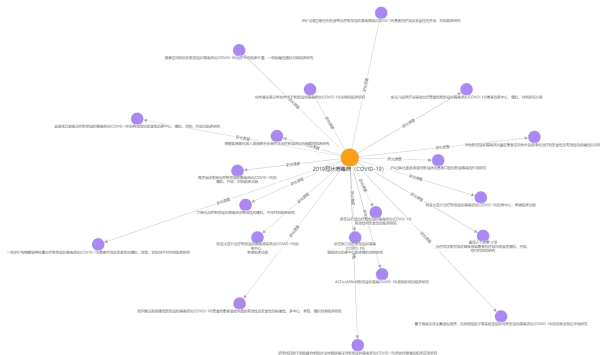
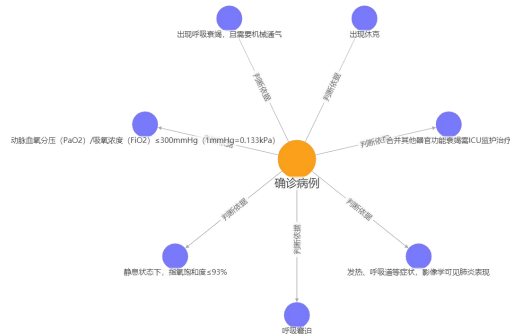
新冠肺炎疫苗项目进展情况

中医证型推荐处方及组成中药材



确诊病例判断依据

疾病相关研究课题

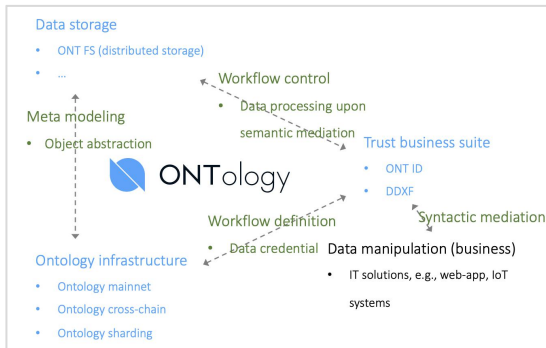


工具和平台建设

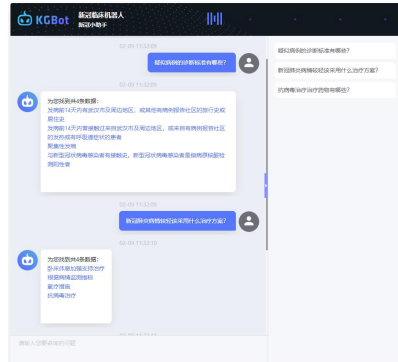
知识众包



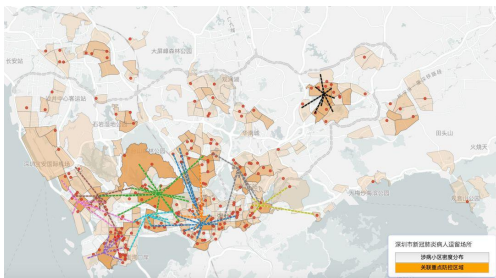
知识上链



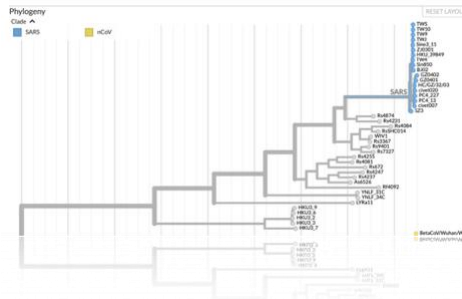
交互问答



辅助决策



预测分析



新冠开放图谱相关数据竞赛

新冠开放图谱相关数据竞赛

Task1 新冠百科知识图谱类型推断：围绕新冠百科知识图谱构建中的**实体类型推断**（Entity Type Inference）展开，从实体**百科页面**出发，从给定的数据推断相关实体的类型。

Task2 新冠概念图谱的上下位关系预测：利用自动挖掘的手段从网络文本中采集的细粒度上位概念词，实体和上位词之间以及不同上位词之间复杂的层次关系，自动准确的构建**细粒度的上下位层次关系**

Task3 新冠科研抗病毒药物图谱的链接预测：依据抗病毒药物图谱Schema及知识图谱的实体、实体属性、实体之间的关系，预测新的两个实体的关系，以进行关系预测，如**药物和病毒的靶向作用、蛋白质间的交互作用**等。

Task4 新冠知识图谱问答评测：面向新型冠状病毒构造了针对**健康、医药、疾病防控**等特定主旨的问答数据，输入中文问题后，期望问答系统从给定知识库中选择若干实体或属性值作为该问题的答案，并且既能处理百科类的**浅层**问题，也能处理具备一定领域知识（如流行疾病等）的**较深层**问题。

相关报名网站：<https://www.biendata.com/competition/>



THANK YOU



请扫码关注OpenKG,查看更多行业内容

