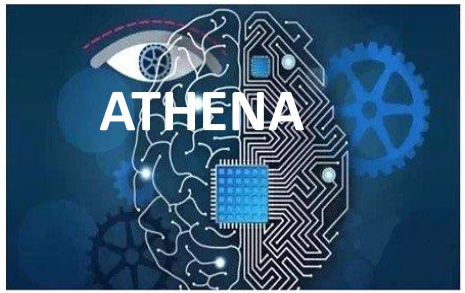
AI 开发者大会 NEXTCON 11月8日-9日 北京

CSDN CAMP

# 医疗知识图谱的敏捷构建和实践

刘升平 博士
资深技术专家，云知声 AI Labs

# 云知声的业务




Athena 智慧大脑

云
端
芯

产品技术架构

智慧生活
家居 车机 ...

智慧服务
医疗 教育 ...

两大商业实践领域

500万+
No.1
家居业务
产品出货量

1800万+
No.1
车载业务
产品出货量

400+
No.1
医疗业务
合作医院数

2.7亿
No.1
教育业务
口语评测日调用量

# 知性会话：基于**语境知识图谱**的人机对话系统


- 物理语境
  - 时间/地点/场所
  - 天气
  - 情绪和情感
  - 设备显示
  - 设备感知
- 言语语境
  - 上下文
  - 主题及焦点
  - 设备反馈
- 知识语境
  - 人类常识
  - 领域知识
  - 用户画像
  - Agent画像
  - 设备信息库

**语境的生命周期：**

- 请求级
- 会话级
- 长期

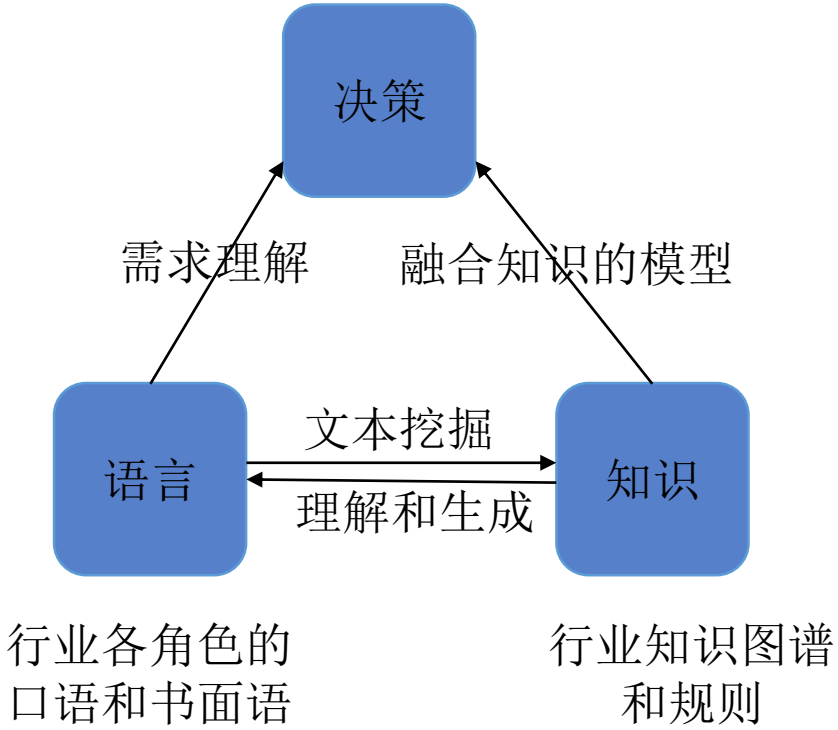
# 基于语境知识图谱的人机对话示例

# 行业认知智能

行业认知智能三要素:

- 语言
- 知识
- 决策

- 知识引导自然语言理解和生成
- 知识辅助决策



# 认知医疗-以医疗知识图谱为核心

决策

导诊分诊 | 预问诊 | 病历生成 | 病历质控 | 辅助诊断 | 智能随访

知识

医疗知识图谱
(疾病，症状，检查，药物，部位等）

+

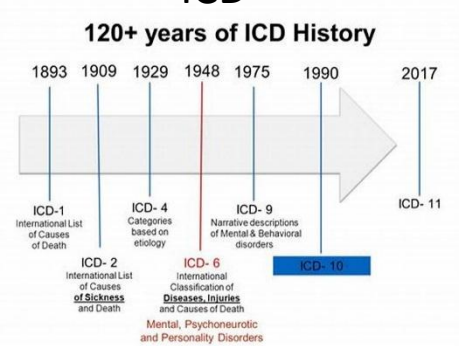
临床规则
(诊断规则，用药规则等）

语言

| 患者语言 | 临床医生语言 | 专家语言 |
| --- | --- | --- |
| 问诊对话，健康咨询(口语) | 问诊对话(口语)<br>病历(书面语) | 教材，文献，临床指南(书面语) |

# 医疗知识图谱的演进

ICD




MeSH

LOINC

SNOMED CT

UML-S

Gene Ontology

SNOMED CT OWL Ed.

GALEN

FMA

Google Medical KG

DrugBank

Bio2RDF

LinkedLifeData

医学术语系统

医学本体

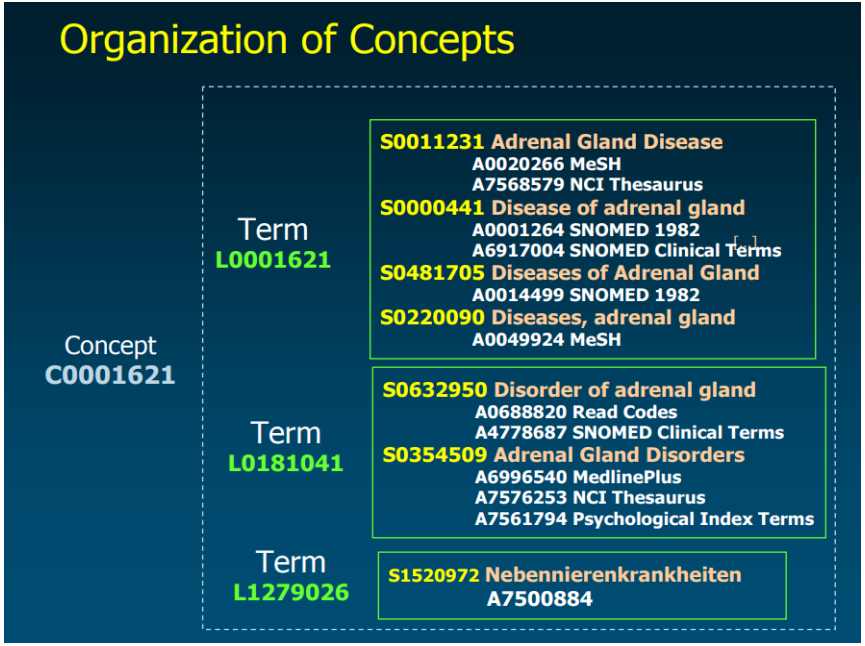
医学知识图谱

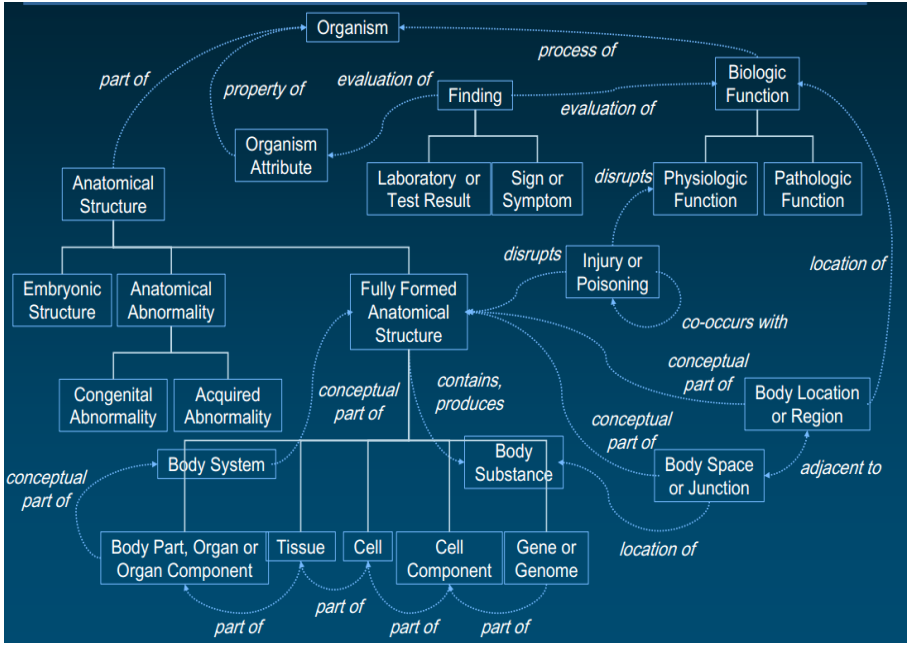
1900

2000

2012

# 术语系统（Terminology System）



# 医学术语系统本体化：以SNOMED CT为例

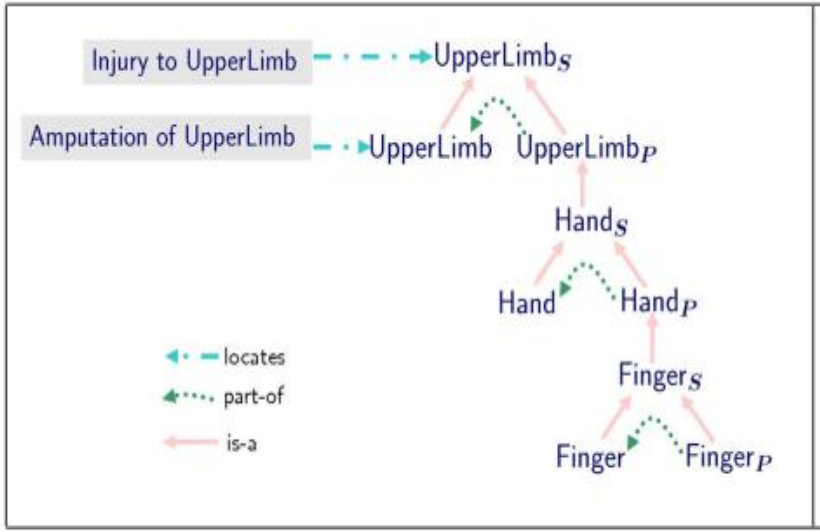
**Table 1 – Characteristics of description logic $EL^{++}$**

| | |
|---|---|
| Existential quantification | $\exists FindingSite.AppendixStructure$ |
| Conjunction | $\exists AssociatedMorphology.Inflammation \sqcap \exists FindingSite.AppendixStructure$ |
| Necessary condition | $Acute\ Appendicitis \sqsubseteq Appendicitis$ |
| Necessary and sufficient conditions | $Appendicitis \equiv \exists AssociatedMorphology.Inflammation \sqcap \exists FindingSite.AppendixStructure$ |
| General inclusions | $Ulcer \sqcap \exists has\text{-}location.Stomach \sqsubseteq Ulcer \sqcap \exists has\text{-}location.(Lining \sqcap \exists part\text{-}of.Stomach)$ |
| Class disjointness | $BodyPart \sqcap Organism \sqsubseteq \bot$ |
| Domain restrictions | $\exists has\text{-}location.\top \sqsubseteq Disease$ |
| Range restrictions | $\top \sqsubseteq \forall has\text{-}location.BodyPart$ |
| Role hierarchy | $Proper\text{-}part\text{-}of \sqsubseteq part\text{-}of$ |
| Role reflexivity | $\varepsilon \sqsubseteq part\text{-}of$ |
| Role transitivity | $part\text{-}of \circ part\text{-}of \sqsubseteq part\text{-}of$ |
| Right identity on roles | $has\text{-}location \circ part\text{-}of \sqsubseteq has\text{-}location$ |
| Concrete domains | $Minor \equiv Person \sqcap <_{18year}(age)$ |
| Nominals | $Kangaroo \sqsubseteq \exists has\text{-}origin.\{Australia\}$ |
| Class assertions | $London \in GeographicLocation$ |
| Role assertions | $(London, England) \in has\text{-}location$ |

SNOMED CT
==
OWL EL++

Schulz S, Markó K, Suntisrivaraporn A B. Formal representation of complex SNOMED CT expressions[J]. Bmc Medical Informatics & Decision Making, 2008, 8(Suppl 1):S9-S9.

# SNOMED CT中的身体部位： ***SEP-Triplets***




$$
\begin{aligned}
\text{UpperLimb} &\sqsubseteq \text{UpperLimb}_S \\
\text{UpperLimb}_P &\sqsubseteq \text{UpperLimb}_S \sqcap \exists \text{part-of}.\text{UpperLimb} \\
\text{Hand}_S &\sqsubseteq \text{UpperLimb}_P \\
\text{Hand} &\sqsubseteq \text{Hand}_S \\
\text{Hand}_P &\sqsubseteq \text{Hand}_S \sqcap \exists \text{part-of}.\text{Hand} \\
\text{Finger}_S &\sqsubseteq \text{Hand}_P \\
\text{Finger} &\sqsubseteq \text{Finger}_S \\
\text{Finger}_P &\sqsubseteq \text{Finger}_S \sqcap \exists \text{part-of}.\text{Finger}
\end{aligned}
$$

**Fig. 1.** Complete SEP-triplets in SNOMED CT.

# SNOMED CT本体化： *EL++ for SEP-Triplets*

$$\text{Finger} \sqsubseteq \text{BodyPart} \sqcap \exists \text{proper-part-of}.\text{Hand} \quad (1)$$
$$\text{Hand} \sqsubseteq \text{BodyPart} \sqcap \exists \text{proper-part-of}.\text{UpperLimb} \quad (2)$$
$$\text{UpperLimb} \sqsubseteq \text{BodyPart} \quad (3)$$
$$\text{AmputationOfFinger} \equiv \text{Amputation} \sqcap \exists \text{has-exact-location}.\text{Finger} \quad (4)$$
$$\text{AmputationOfHand} \equiv \text{Amputation} \sqcap \exists \text{has-exact-location}.\text{Hand} \quad (5)$$
$$\text{AmputationOfUpperLimb} \equiv \text{Amputation} \sqcap \exists \text{has-exact-location}.\text{UpperLimb} \quad (6)$$
$$\text{InjuryToFinger} \equiv \text{Injury} \sqcap \exists \text{has-location}.\text{Finger} \quad (7)$$
$$\text{InjuryToHand} \equiv \text{Injury} \sqcap \exists \text{has-location}.\text{Hand} \quad (8)$$
$$\text{InjuryToUpperLimb} \equiv \text{Injury} \sqcap \exists \text{has-location}.\text{UpperLimb} \quad (9)$$
$$\text{proper-part-of} \circ \text{proper-part-of} \sqsubseteq \text{proper-part-of} \quad (10)$$
$$\text{proper-part-of} \sqsubseteq \text{part-of} \quad (11)$$
$$\text{part-of} \circ \text{part-of} \sqsubseteq \text{part-of} \quad (12)$$
$$\epsilon \sqsubseteq \text{part-of} \quad (13)$$
$$\text{has-exact-location} \sqsubseteq \text{has-location} \quad (14)$$
$$\text{has-location} \circ \text{proper-part-of} \sqsubseteq \text{has-location} \quad (15)$$

**Fig. 2.** A re-engineered extract of SNOMED CT without SEP-triplets.

Suntisrivaraporn B, Baader F, Schulz S, et al. Replacing SEP-Triplets in SNOMED CT Using Tractable Description Logic Operators, Conference on Artificial Intelligence in Medicine. Springer-Verlag, 2007:287-291.

# SNOMED CT to OWL

- SNOMED CT OWL Reference Set： 2018 July International Edition
  - Draft： 2018 Jan Edition
- [SNOMED CT to OWL toolkit]


## 3.1. Feature Overview

SNOMED CT supports the following types of class restrictions:

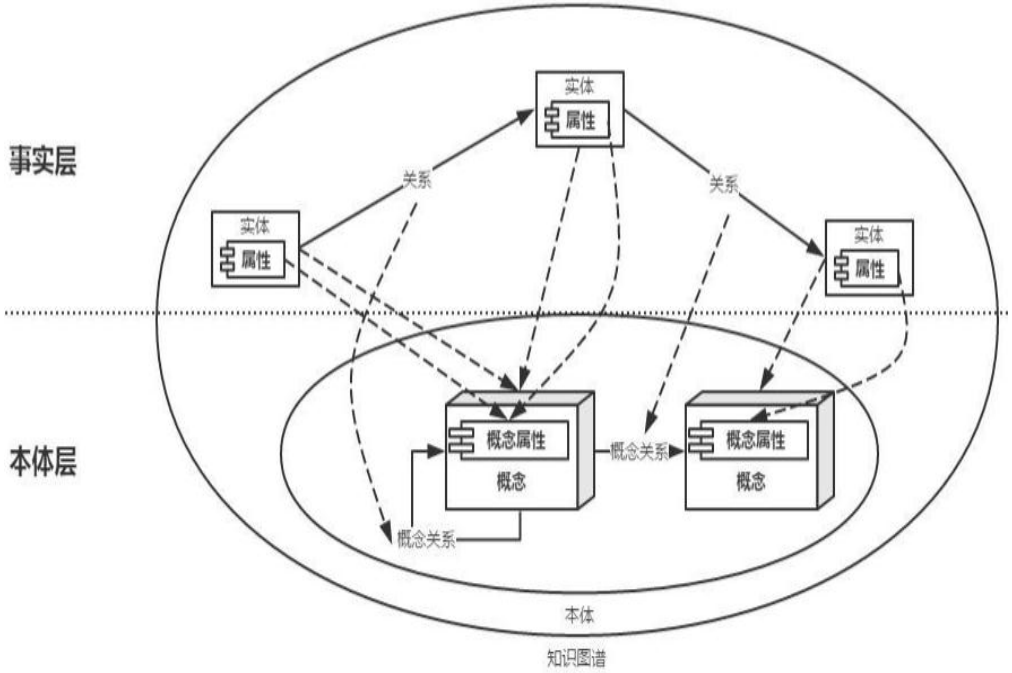
- existential quantification to a class expression (ObjectSomeValuesFrom)
- intersection of classes (ObjectIntersectionOf)
- existential quantification to a literal (DataHasValue)

SNOMED CT supports the following axioms, all of which are restricted to the allowed set of class expressions:

- class inclusion (SubClassOf)
- class equivalence (EquivalentClasses)
- object property inclusion (SubObjectPropertyOf) with or without property chains, and data property inclusion (SubDataPropertyOf)
- class disjointness (DisjointClasses)
- property equivalence (EquivalentObjectProperties and EquivalentDataProperties),
- transitive object properties (TransitiveObjectProperty)
- reflexive object properties (ReflexiveObjectProperty)

# 术语系统，本体，知识图谱，知识库

- 术语系统
  - 偏重term，不够形式化
  - Language-centered view
- 本体
  - 偏重概念模型，可以有规则(公理）和实例
  - is the study of *what there is*
- 知识图谱
  - 偏重实体及关系, 图模型
  - 常包含一个本体(schema)
- 知识库
  - 泛指知识的一种组织形式



# 医疗**知识图谱**应用(1): 问诊对话

APP


在候诊区使用

生成病历中的
病史信息

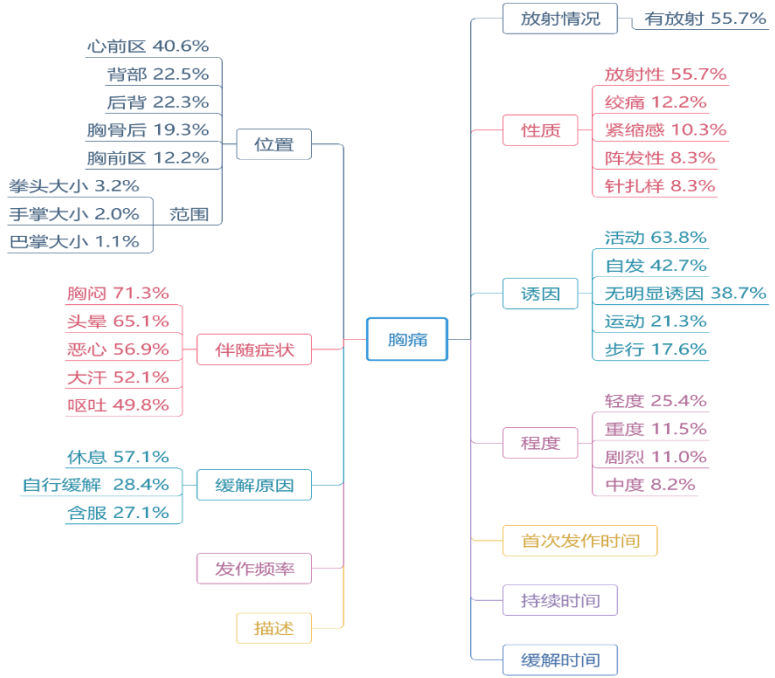
生成辅助诊断

机器人会话交互

# 问诊对话依赖的知识

以症状作为入口
把症状的发生看成是事件

- 发生部位
- 发生时间和频率
- 缓解，加重因素
- 程度
- 性状
- 伴随症状




胸痛 is a Class

# 医疗知识图谱应用(2): 病历质控

| | 形式质控 | 内涵质控 |
| --- | --- | --- |
| 主诉 | 不能缺项<br>长度不超过20个字 | 包括主要症状和持续时间<br>能导出第一诊断<br>主诉和现病史相关相符 |
| 入院诊断 | 不能缺项 | 诊断确切，依据充分<br>主次排列有序<br>无漏诊，误诊 |

# 病历质控依赖的知识

- 以疾病诊断为入口，判断诊断相关信息是否完整和一致
  - 疾病相关症状
  - 疾病相关个人史，既往史，家族史等
  - 疾病的诊断标准
  - 疾病是否特点人群相关
  - 疾病发生部位

胸痛 is an Instance

# 如何对知识图谱做测试

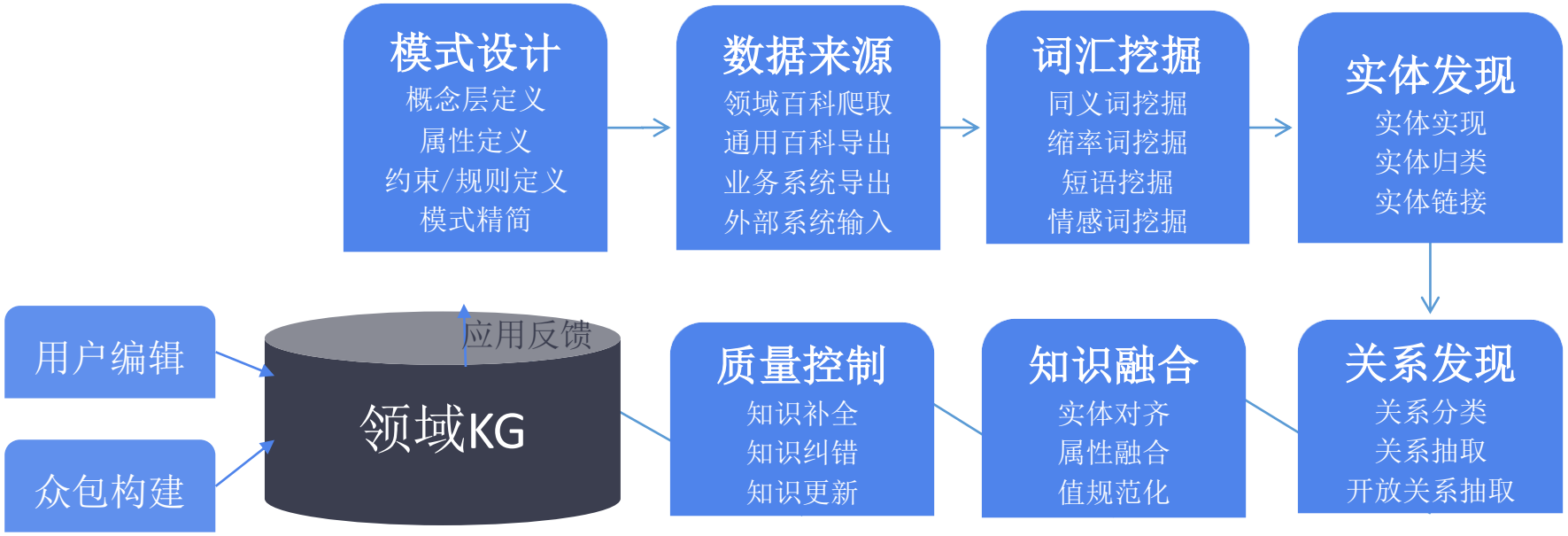
| 评估方法 | 方法说明 | 评级层次 |
| --- | --- | --- |
| 基于黄金标准评估 | 将所构建的本体与黄金标准（一个公认的比较成熟的本体或是人工标注术语集）进行比较，罗列出其不足并进行改进。 | 词汇数据层，层级分类层，语义关系层 |
| 基于本体任务/应用的本体评估 | 一个特定应用环境中，测试一组本体，看哪个本体最适合该应用，这些应用包括搜索、问答、推荐、决策等。 | 词汇数据层，层级分类层，语义关系层，应用层 |
| 数据驱动评估 | 通过衡量本体与领域语料的匹配度或本体的领域覆盖度来评估本体，或使用其他参考数据来辅助本体评估过程，这种方法常与文本分析、机器学习技术结合 | 词汇数据层，层级分类层，语义关系层 |
| 基于指标的评估（人工评估） | 基于一套预先定义好的原则、准则、标准等进行评估的方法，其多是从构建本体的原则来评估本体。 | 词汇数据层，层级分类层，语义关系层，应用层 |

袁凯琦，邓扬，陈道源，张冰，雷凯，沈颖. 医学知识图谱构建技术与研究进展. 2018，35(7).

# 我们的评估指标

| 评估方法 | 评估指标 |
| --- | --- |
| 基于指标的评估 | 一致性：是否存在一个term用在多个不同的地方；<br>精确性（人工）：是否存在多个实体表示同一个意思；<br>正确性（人工）：实体的属性，关系是否正确；<br>相关性（人工）：是否跟领域紧密相关 |
| 基于黄金标准的评估 | Term覆盖率，关系的准确率和覆盖率<br>（如：以CCKS 2017和2018医疗实体评测为金标准） |
| 基于应用的评估 | 基于知识图谱的NER效果和病历质控等应用效果 |

# 领域知识图谱的一般构建方法




肖仰华：领域知识图谱落地实践中的问题与对策，2018

# 医疗领域知识图谱构建的挑战

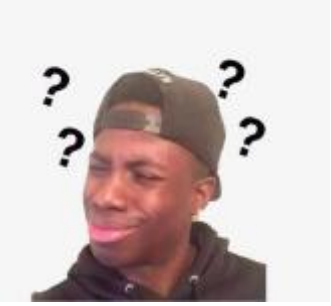
# 医疗知识图谱现状

## 国外

- UML-S
- FMA
- DrugBank
- SNOMED CT
- ICD-10
- LinkedLife Data
- LOINC
- GALEN
- Bio2RDF
- RxNorm
- Gene Ontology
- More…

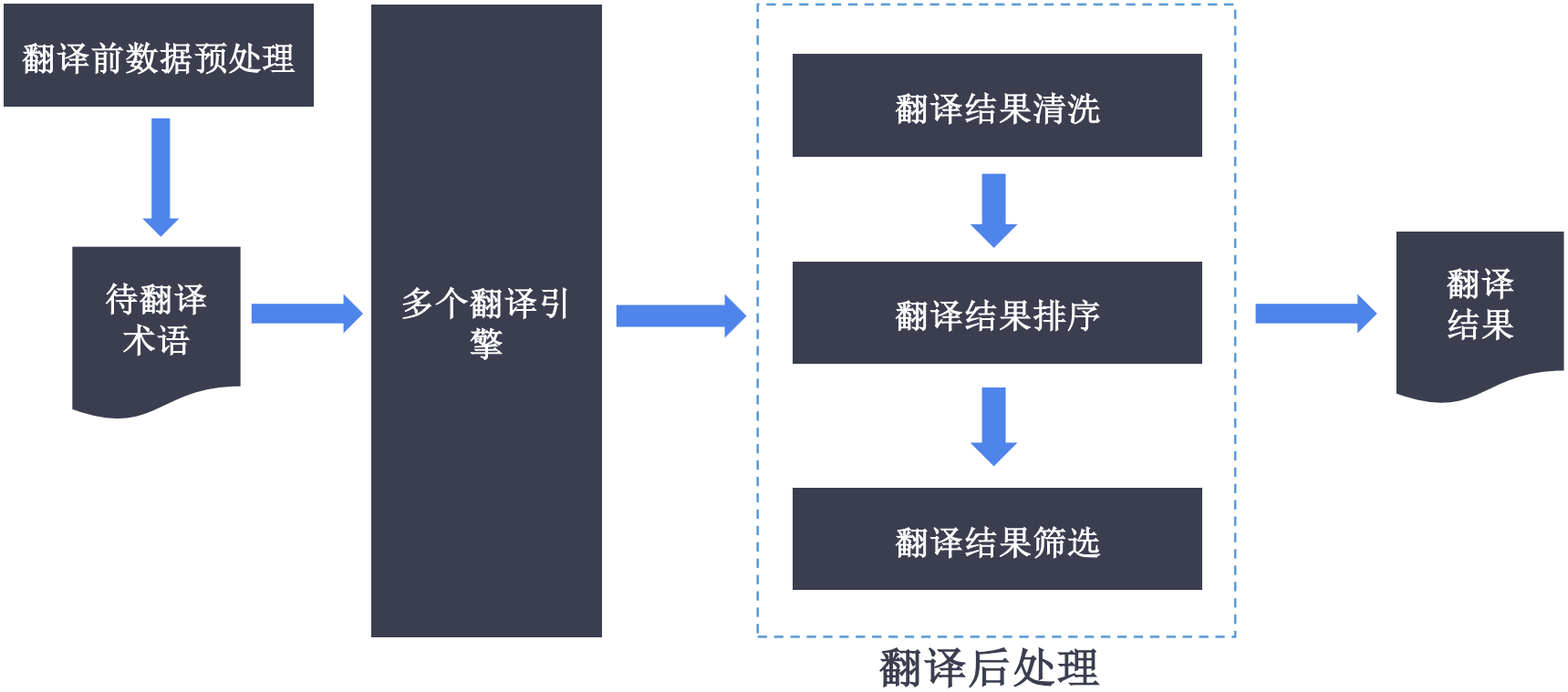
## 国内（中文版）

- MESH
- ICD-9-CM
- ICD-10
- 症状知识图谱 @OpenKG



The Open Biological and Biomedical Ontology (OBO) Foundry

# 冷启动-基于UMLS的中文汉化

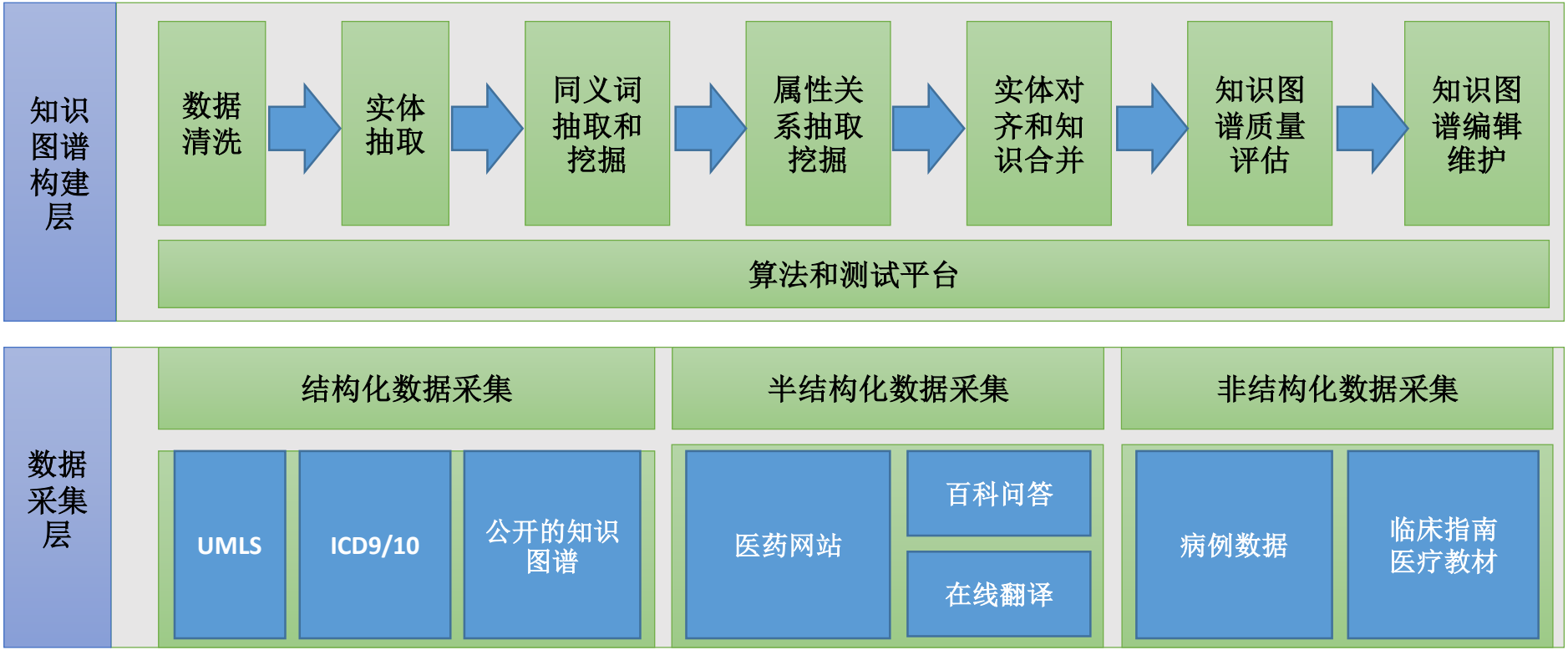


# 翻译结果排序和筛选

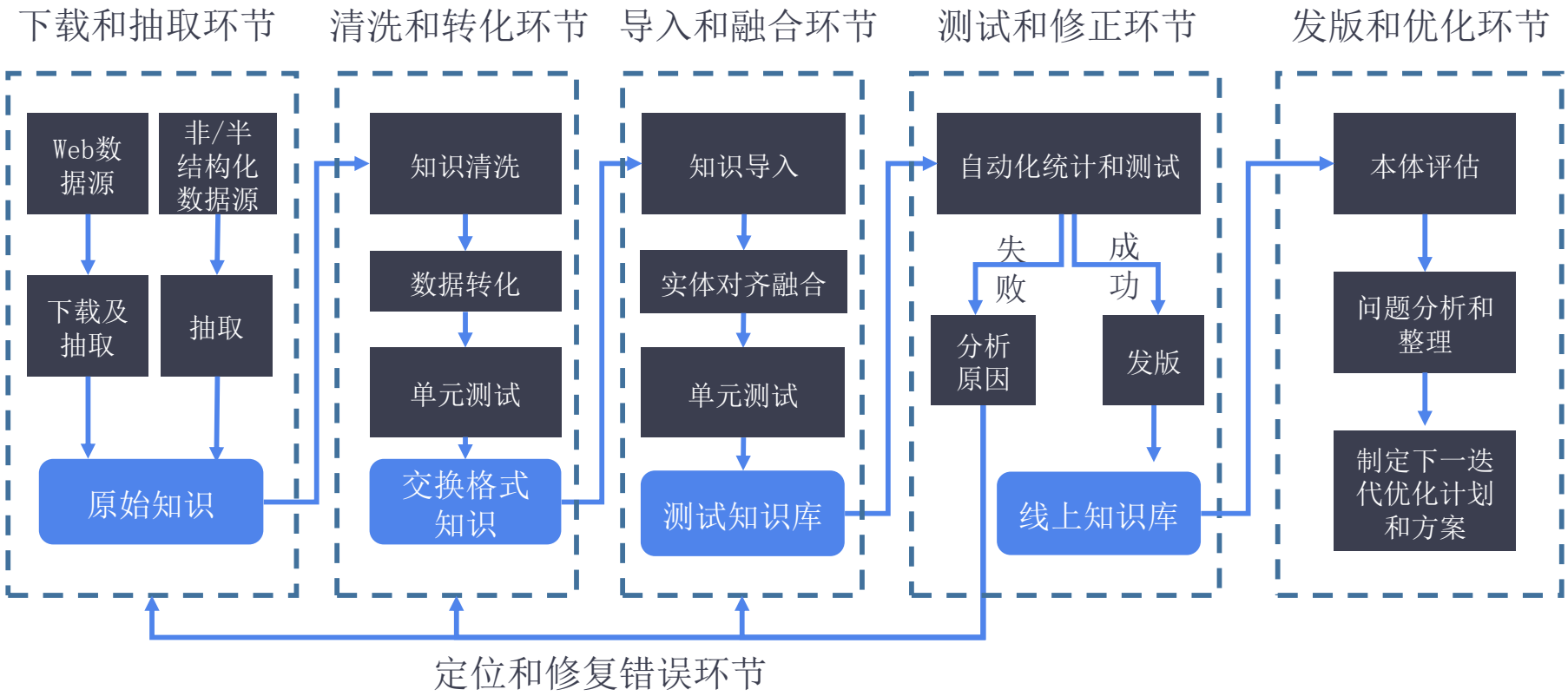
- **排序方法（按照以下依据依次比较）**
  - A. 是否标记为医疗专业术语［医］
  - B. 在病历文本中出现的频次
  - C. 在海量医疗文本中出现的频次
  - D. 在海量通用语料中出现的频次
  - E. 候选翻译术语的最大长度

- **筛选依据**
  - A. 满足A，或B>k，或C>k，都认为是正确的翻译结果
  - B. 其他需要人工处理

# 知识图谱构建总体架构


知识图谱构建层
数据清洗
实体抽取
同义词抽取和挖掘
属性关系抽取挖掘
实体对齐和知识合并
知识图谱质量评估
知识图谱编辑维护
算法和测试平台
数据采集层
结构化数据采集
半结构化数据采集
非结构化数据采集
UMLS
ICD9/10
公开的知识图谱
医药网站
百科问答
在线翻译
病例数据
临床指南
医疗教材

# 敏捷构建


下载和抽取环节
清洗和转化环节
导入和融合环节
测试和修正环节
发版和优化环节
Web数据源
非/半结构化数据源
下载及抽取
抽取
原始知识
知识清洗
数据转化
单元测试
交换格式知识
知识导入
实体对齐融合
单元测试
测试知识库
自动化统计和测试
失败
成功
分析原因
发版
线上知识库
本体评估
问题分析和整理
制定下一迭代优化计划和方案
定位和修复错误环节

# 演化和版本管理

- **版本描述**
  - 版本号及改进点描述
  - 本版本与上一版本的差异比较结果文件：增加或减少了哪些概念，属性，实体，关系等
- **版本比较**
  - 根据不同版本的知识库文件，生成比较结果文件
- **版本恢复**
  - 按照操作日志回滚（轻量级）
  - 根据版本差异比较文件恢复
  - 根据版本备份文件恢复（回溯到某个发布版本）
- **版本发布**
  - 每一次开发迭代都要完成发版
  - 发版要生成完整的版本描述和备份文件并存档

# 知识图谱的迭代

- **发现和梳理知识图谱存在的问题**
  - 以应用为导向，根据知识图谱的应用效果，提出改进意见
  - 领域专家通过可视化操作平台，分析知识库，并找出存在的问题
  - 根据知识图谱构建各环节缺陷提出待优化的问题
- **版本规划**
  - 列出所有需要改进的点
  - 综合考虑实际应用需求、知识图谱质量要求以及开发成本等因素排优先级
  - 确定下一版本发版计划和开发方案，并给出发版号
- **对知识图谱各个环节展开开发**
  - 按照预先讨论的方案开发
  - 中间知识图谱要快速生成，以确保其他环节有可用知识
  - 做好单元测试，最大限度减少传递错误的次数

# 缺少医疗领域专家，怎么办

- 利用少数的医疗专家做人工抽查评测
- 利用网上公开的知识
- 从文本中挖掘知识：病历，教材，百科等

# 医疗知识图谱敏捷构建的体会总结

- 应用驱动
- 敏捷构建最关键的两个技术点
  - 自动化评测
  - 知识融合
- 医疗知识图谱的形式化是下一步的重心
  - 知识本体的表示与推理：OWL EL++ with Meta-Modelling support
  - 病历大数据结合下的海量数据查询与推理

AI 开发者大会 NEXTCON
11月8日-9日 北京

CSDN ACAMP

# 谢谢！