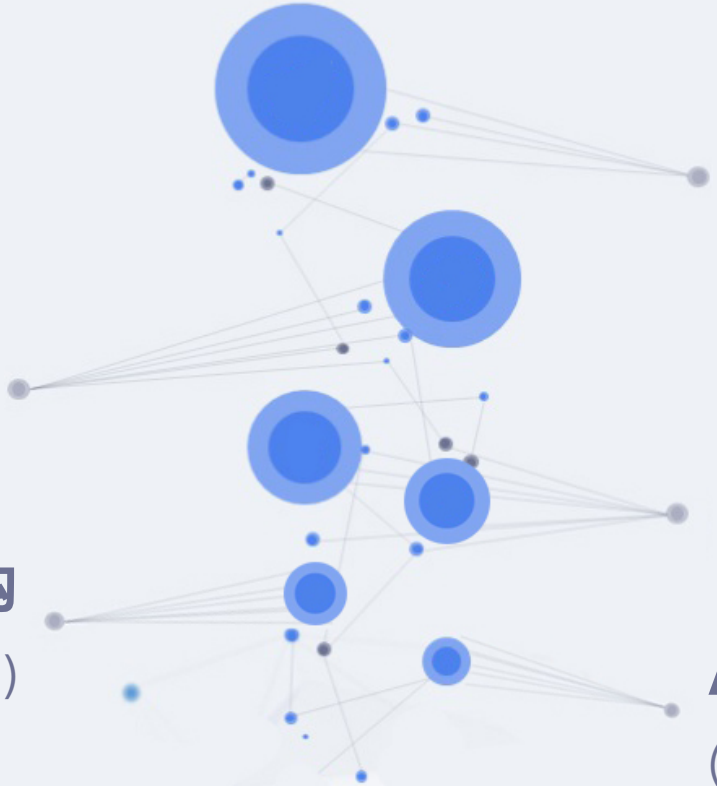
大规模医疗知识图谱的构建与应用

刘升平 博士
资深技术专家，云知声 AI Labs

云知声 一家全栈AI技术公司



13亿人民币C轮融资

语音语言领域最大

语音芯片

独立IC设计

产品探索与创新

（医疗，家居，车载，教育）

AI基础架构

（算法+算力+数据）

AI技术全栈

（语音，语言与知识，图像）

语言与知识

语用计算（CCKS 2016）

语用=语义+语境（环境，上下文，知识）

知性会话（CCKS 2017）

以知识图谱为中心的跨领域，跨交互形式的人机对话系统架构

领域知识图谱（CCKS 2018）

医疗知识图谱的敏捷构建，驱动智慧医疗应用

应用：基于知识图谱的诊疗全流程辅助

·诊前：导诊，预问诊

·诊中：辅助诊断

·诊后：随访和宣教

典型知识图谱评估方法分类

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； 精确性（人工）：是否存在多个实体表示同一个意思； 正确性（人工）：实体的属性，关系是否正确； 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 （如：以CCKS 2017和2018医疗实体评测为金标准）
基于数据驱动的评估	知识图谱的领域Fit程度
基于应用的评估	基于知识图谱的智慧医疗应用效果

领域知识图谱的构建方法

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

·冷启动

·敏捷构建

·缺少医疗专家

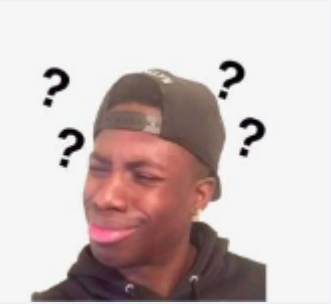
医疗知识图谱现状

国外

UML-S	FMA	MeSH
SNOMED CT	ICD-10	ICD-9-CM
LOINC	GALEN	WHO-ART
RxNorm	Gene Ontology	More…

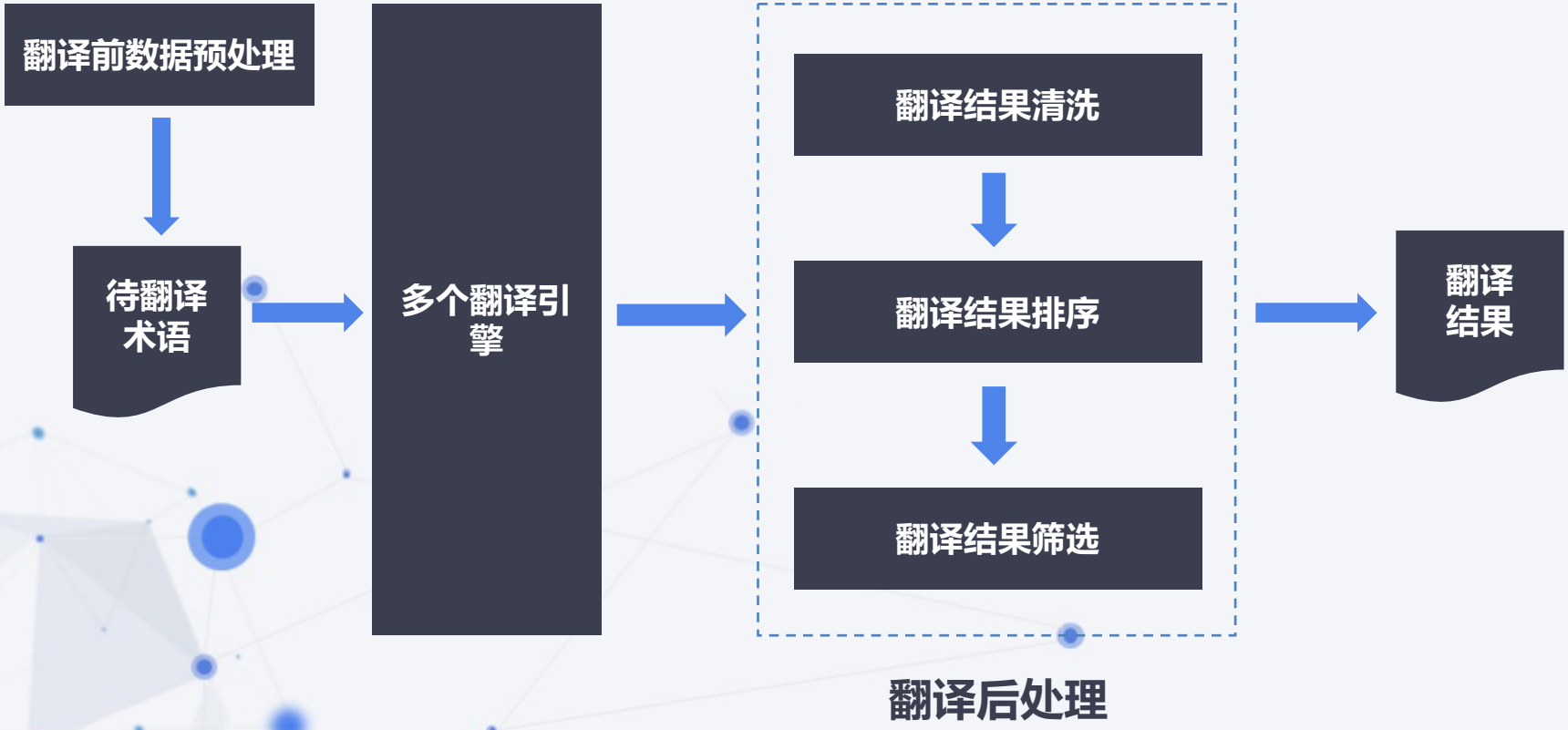
国内(中文版)

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



翻译前数据预处理

翻译前数据预处理规则	清洗前	清洗后
括号里包含常用描述词，则去掉括号及其内部内容	Aneurysm of thoracic aorta (disorder)	Aneurysm of thoracic aorta
	Accidental injury (finding)	Accidental injury
NOS在词表尾部，去掉NOS	Spondylolisthesis, NOS	Spondylolisthesis
	Catatonic schizophrenia NOS	Catatonic schizophrenia
以中括号开头，中间包着单个字母，则去掉中括号及其内部内容	[X]Congenital syphilis	Congenital syphilis
	[D]Snoring	Snoring
如果中划线连接缩写和全名，则保留全名	SAS - Sleep apnoea syndrome	Sleep apnoea syndrome

翻译后数据清洗

翻译后数据清洗规则举例	清洗前	清洗后
如果括号内包含的数据与括号外的内容不能组成一个术语，则删除掉括号中的内容	螨虫引起的皮肤病（病症）	螨虫引起的皮肤病
	下腔静脉综合征[利]（障碍）	下腔静脉综合征
如果颠倒一下词序可以组成一个完整的术语T，则取T为最后结果	心力衰竭，充血性	充血性心力衰竭
	纤维化，肺	肺纤维化
如果包含"和/或"，则将其转成两条术语	妊娠和/或产褥期静脉并发症	妊娠和产褥期静脉并发症 妊娠或产褥期静脉并发症
如果翻译结果包含中英文或其他非中文字符，需要人工处理	Croupy breathing 翻译成Croupy的呼吸	直接交给人工处理

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

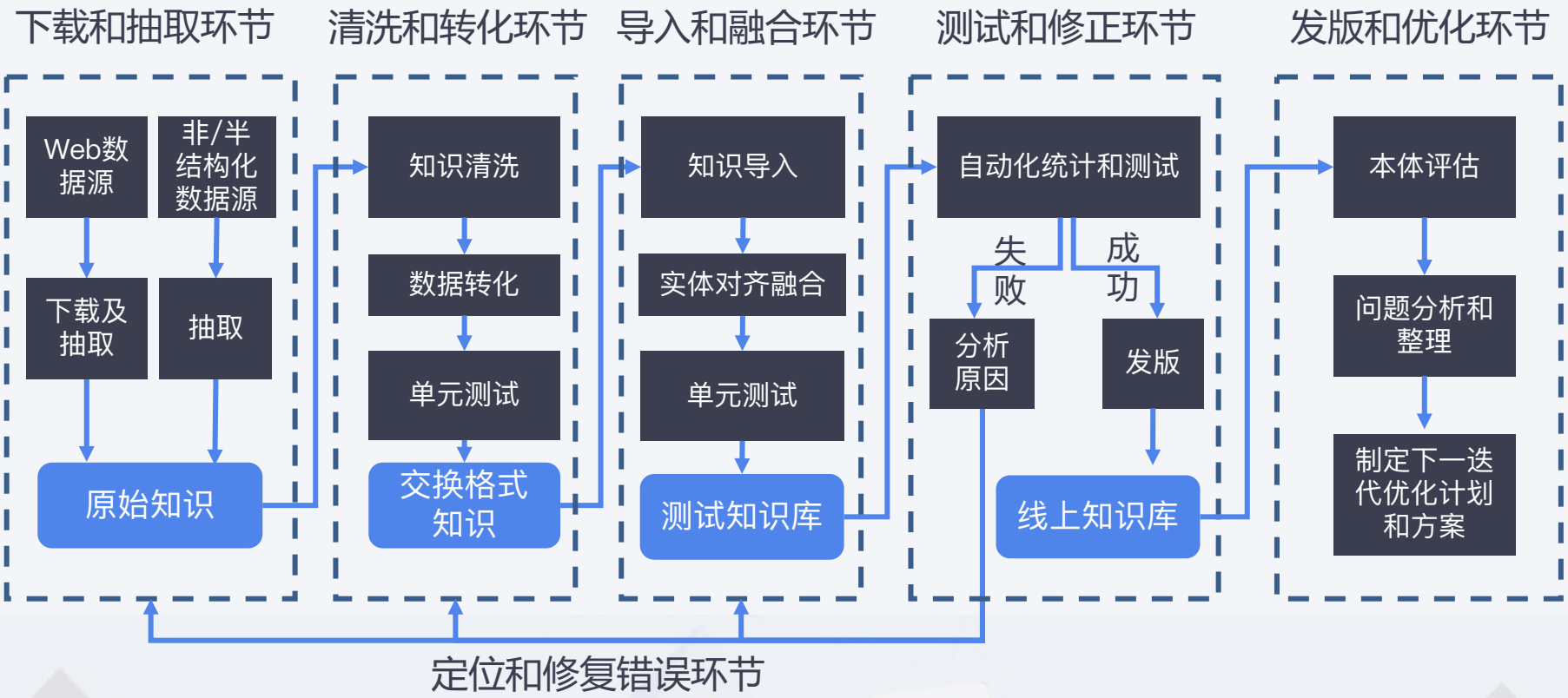
E. 候选翻译术语的最大长度

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

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

知识库的迭代

发现和梳理知识库中存在的问题

- 以应用为导向，根据知识库的应用效果，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
- 根据知识库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、知识库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
- 中间知识库要快速生成，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：病历和医疗百科等

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

云知声 Unisound

hr@unisound.com

微信/QQ: 357638