THUKC世界知识图谱 — XLORE

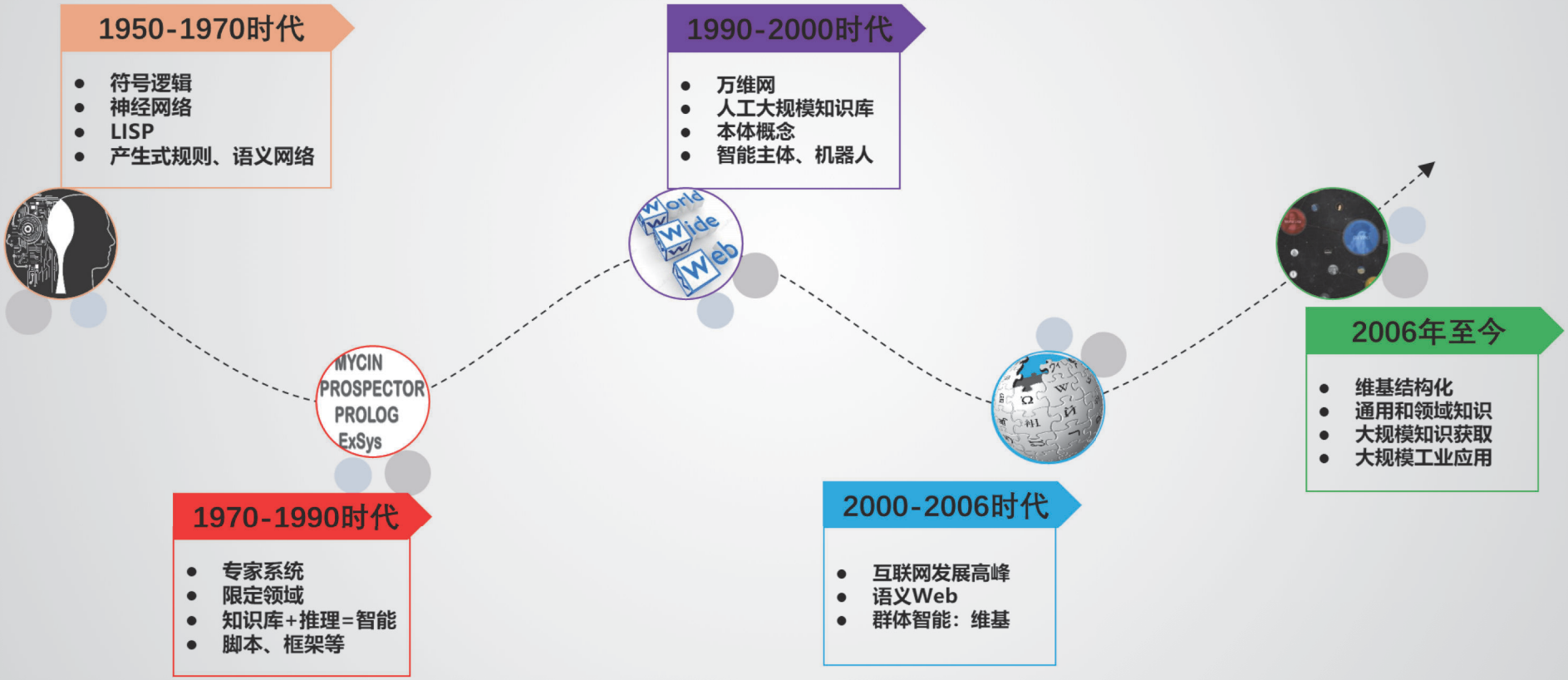
李涓子
清华大学计算机系
清华大学人工智能研究院

2019年1月21日

目录

目录

知识工程发展历程



高质量大规模知识图谱已经成为智能系统的重要基础设施

在线百科知识资源

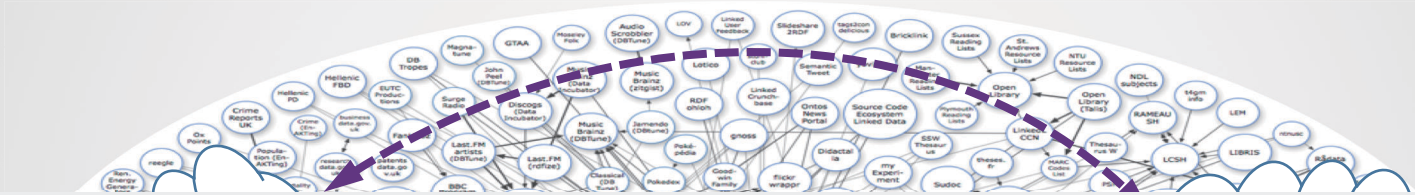
- 百科全书：概要介绍人类全部知识或某一特定领域或学科的工具书或纲要—wikipedia
- 大规模在线百科知识资源—以wikipedia为例
  - 优点
    - 互联网**最大**知识资源， **302种**语言**4800余万**词条，日访问超**5亿次**
    - 分类结构和条目知识**符合人类的知识结构**
    - 知识**动态更新**
    - **异构**资源之间**知识互补**
    - 包含丰富的不同结构类型的知识（结构化，半结构和非结构）
  - 缺点
    - **面向人**的知识资源服务，并**不是面向机器理解**的内容

多语言知识图谱现状

构建方式	知识图谱	实体数量	是否跨语言
基于人工构建	ResearchCyc	0.5	×
	WordNet	0.5	×
基于维基构建	DBpedia	38.3	√
	YAGO	10	√
	Freebase	23	√
	BabelNet	13.8	√
	Wikidata	24	√
开放信息抽取	NELL	-	×
	Knowitall	-	×
	Probase	5.4	×
中文知识图谱	百度知心	-	×
	搜狗知立方	-	×

存在问题：跨语言链接缺乏

XLORE构建目标



现实世界中**同一概念或实体**的**多语言融合**，实现对客观世界**多语言、多概念层次**语义建模：

- **促进知识共享**，支持跨语言自然语言处理等技术的突破
- **丰富世界知识**，解决单语言环境下知识规模不足的问题
- **提升知識精度**，剔除知识单语言描述中潜在存在的噪声

Baidu 百科　維基百科 自由的百科全書　WIKIPEDIA The Free Encyclopedia

技术挑战

多语言知识图谱形式化

$$O = \{ C, I, T, P \}$$

概念集　　实体集　　概念体系　　实体关系

挑战

- 概念上下位语义关系存在噪音
  - 约有**21.7%**的上下位从属关系是错误的
- 语义等价关系规模不足
  - 仅有**约6%**的英文维基词条存在等价的中文链接
- 实体语义关系大量缺失
  - 6个主要语言维基仅有**32.8%**的词条包括信息框

技术特色

百科类知识资源的结构化和融合

- 更合理的概念分类体系
- 更丰富的跨语言对齐信息
- 更准确的实体语义信息

概念体系构建

- 跨语言分类体系验证
- 事件话题概念体系学习

跨语言对齐

- 跨语言实体对齐

实体分类

- 知识图谱细粒度实体分类

实体链接

- 跨语言实体链接系统

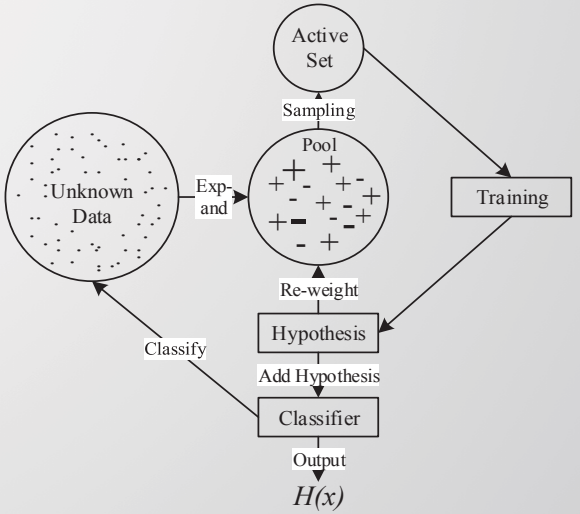
基于跨语言链接校验的概念体系构建

- 问题

概念体系是知识组织和推理的重要部分，维基类知识资源提供的上下位分类结构存在错误

- 效果
  - 概念-概念1,606,364个
  - 概念-实例4,936,423个

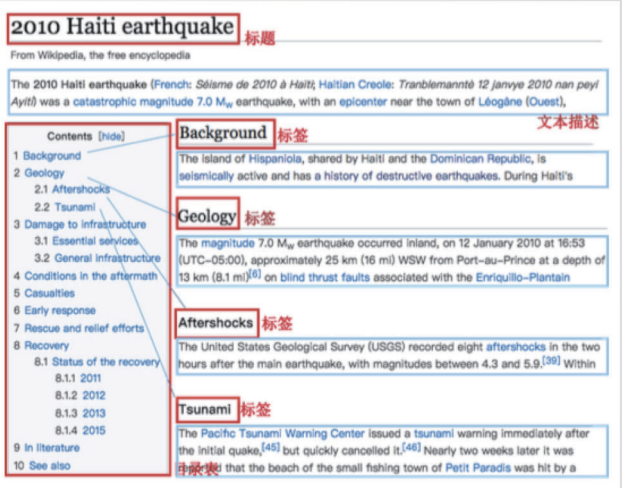
- 解决方法（DAB模型）
  - 基于跨语言链接校验的分类器
    - 标注样本：通过跨语言链接校验自动**增量扩充**
    - 特征定义：跨语言校验不同特征带来**优势互补**
  - 动态自适应增强学习模型
    - 标注样本扩充；样本权重更新



Methods	English SubClassOf			Chinese SubClassOf			English InstanceOf			Chinese InstanceOf		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HL	87.1	81.3	84.1	91.4	91.4	91.4	94.3	89.4	91.8	42.4	51.9	46.7
DT	88.7	86.9	87.8	90.9	92.0	91.4	91.9	95.6	93.7	46.8	58.1	51.8
AdaBoost	90.8	90.9	90.9	91.4	92.3	91.8	94.3	94.1	94.2	51.4	63.9	57.0
DAB	90.7	91.8	**91.2**	91.1	95.2	**93.1**	94.1	97.7	**95.9**	77.8	75.0	**76.4**

基于维基百科的事件话题概念体系学习

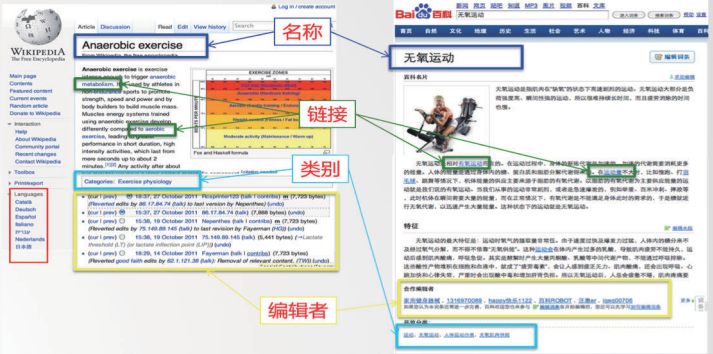
- 问题
  - 维基百科提供高质量相似事件知识
  - 半结构化与非结构化知识并存
  - 不同目录结构存在异构、关系不一致
- 解决方法
  - 话题语义图构建
    - 结构信息：概率贝叶斯网络
    - 文本信息：层次狄利克雷模型
  - 最大生成树
    - 话题贝叶斯网络→话题概念体系
    - Chu-Liu/Edmonds算法
- 实验结果
  - 人工构建了地震、选举事件数据集
  - 显著优于仅考虑文档级共现的基线方法
  - 结构和文本信息的融合效果最佳

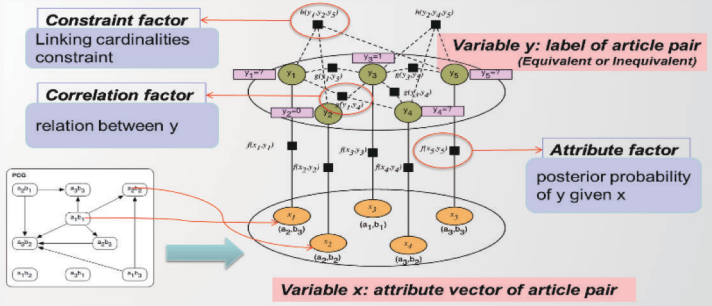


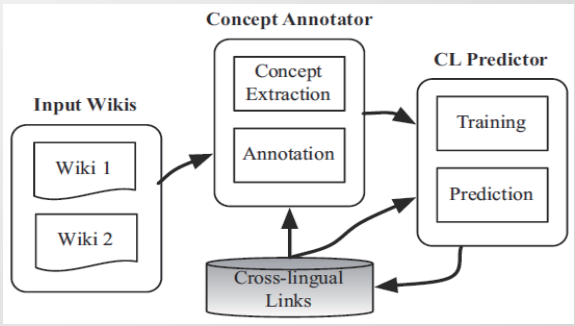
	Earth.(En)	Elect.(En)	Earth.(Ch)
Basic	0.5965	0.7719	0.7143
Pro+S	0.8971	0.8596	0.9017
Pro+ST	0.9543	0.9298	0.9286

跨语言知识链接

- 问题
  - 跨语言实体对齐关系稀疏
  - 不同实体知识异构情况严重
  - 实体概念信息缺失
- 解决方法
  - 概率因子图模型
    - 局部因子：实体对特征→对齐关系
    - 相关因子：多实体对相关性
    - 约束因子：链接候选间的约束
  - 迭代对齐框架
    - 已有对齐→概念标注
    - 跨语言对齐关系分类
- 实验结果
  - 基于维基和百度人工构建了标准评测集
  - 初步发现英文维基和百度百科间约20万跨语言链接，迭代扩展至40万

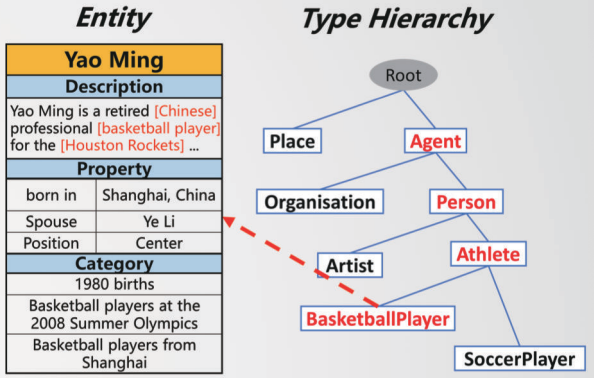




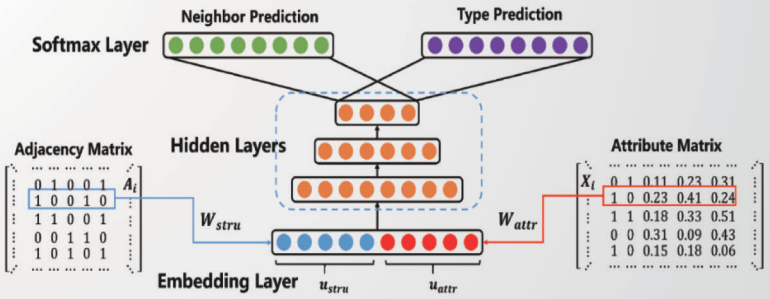


知识图谱细粒度实体分类

- 问题
  - 实体细粒度类型提供更准确的语义信息
  - 实体信息异构
  - 类型层次细粒度



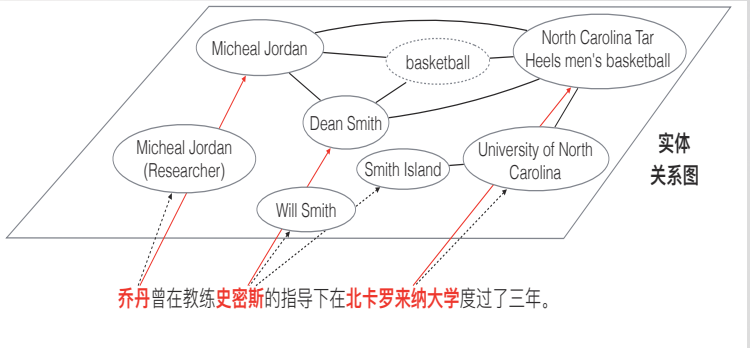
- 解决方法
  - 异构网络表示学习
    - 属性建模：与类别高度相关
    - 半监督：融合部分标注信息
  - 排序学习
    - 类型层次建模



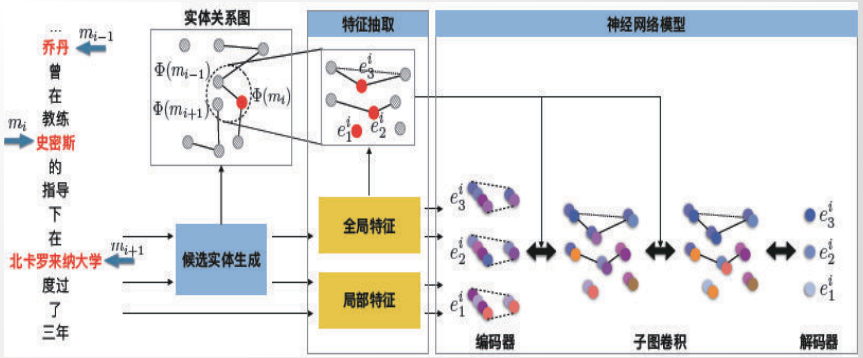
- 实验结果
  - 构建了包含5个层次、451个类型、300万实体的评测集
  - Mi-F1和Ma-F1值分别提高2.9%和3.4%

跨语言协同实体链接系统

- 问题
  - 翻译工具错误累积
  - 无法有效利用知识图谱语义关联



- 解决方法
  - 跨语言词和实体联合表示学习
    - 弱监督：维基资源→相似句对
    - 注意力机制：噪音控制
  - 协同实体链接
    - “提及-实体”词典→实体识别
    - 图卷积：语义关联图→实体消歧



- 实验结果
  - GERBIL评测框架下均优于基线方法
  - 开发了实体链接服务系统XLink

XLORE
中英文跨语言百科知识图谱

XLink KEG Entity Linking System based on XLore

XLORE系统框架

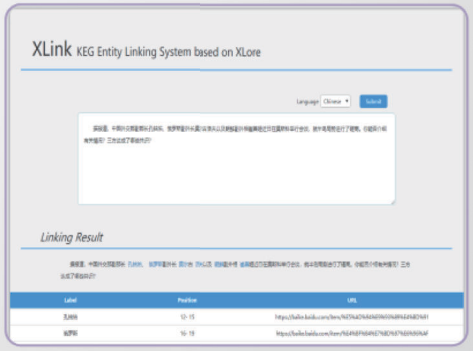
百度百科　中文维基　英文维基

基于跨语言知识校验的上下位语义关系识别

基于异构网络表示学习的跨语言实例匹配

基于表示学习的细粒度实体分类

基于因子图模型的跨语言属性对齐



与国际著名知识图谱数据比较

- DBpedia — 最早的维基类知识图谱
- Freebase — Google知识图谱的核心
- YAGO 和 BabelNet — 2017年IJCAI卓越论文奖

		DBpedia	YAGO	Freebase	BabelNet	XLORE
概念数	中文	760	352K	53.1K	2,355K	
	英文				3,229K	
实体数	中文	859K	325K	49.9M	954K	9,166K
	英文	4,298K	3,420K		5,626K	
属性数	中文	2,859	77	70.9K	-	
	英文					
三元组数	中文	45.4M	706K	3,129M	1,959M (271种语言)	141.6M
	英文	438.3M	6,587K			
上下位关系数	中文	-	1.5M	24.4M	-	
	英文	30.1M	10.3M			
跨语言链接数		545K	325K	-	-	758K

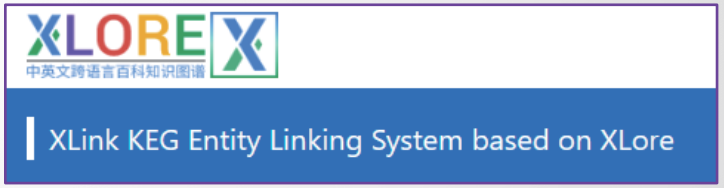
9.6倍

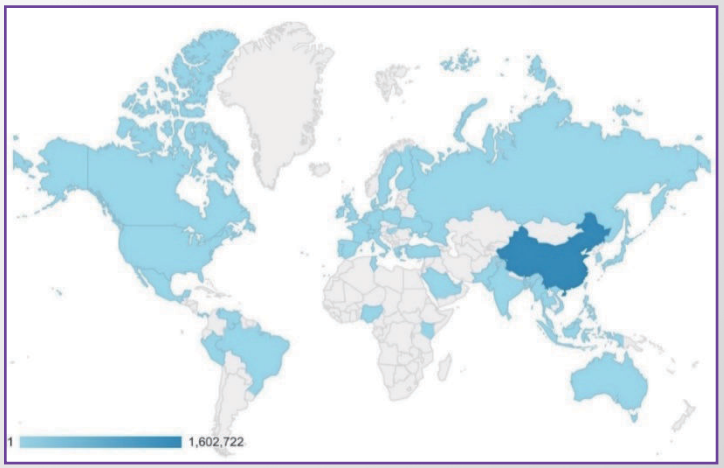
3.1倍

+39%

XLORE应用服务

- XLORE数据服务API
  - 关键词查询、词条精确检索
  - 概念、属性、实例检索
  - 实体关系预测
- 系统访问情况统计
  - 系统访问量上亿次
  - 近一年API调用174余万次
  - 地域覆盖
    - 国际：58个国家或地区
    - 国内：129个主要城市





XLink跨语言实体链接服务

XLORE 中英文跨语言百科知识图谱

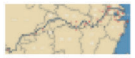
首页 图谱 APIs XLink 成果 关于我们 CN

XLink KEG Entity Linking System based on XLore

党的十九大以来，从长江之行，到东北之行，再到京津冀之行，习近平国内考察调研跨多省市、着眼特定区域发展的新特点越来越鲜明。 常常在几天时间内行程上千公里甚至几千公里，充分表明区域协调发展这件事在习近平心中的分量越来越重，在党和国家发展全局中的地位越来越凸显。 沿着习近平的足迹，中国经济高质量发展的区域“新版图”愈加清晰，区域协调发展战略愈加走向深入。 定方向 理思路

2018年4月，习近平深入湖北宜昌市和荆州市、湖南岳阳市以及三峡坝区等地，实地了解长江经济带发展战略实施情况。

长江 1



长江发源于“世界屋脊”——青藏高原的唐古拉山脉各拉丹冬峰西南侧，是中华民族的母亲河。干流流经青海、西藏、四川、云南、重庆、湖北、湖南、江西、安徽、江苏、上海11个省、自治区、直辖市，于崇明岛

京津冀 1

京津冀是中国的“首都经济圈”，包括北京市、天津市以及河北省的保定、唐山、廊坊、石家庄、秦皇岛、张家口、承德、沧州、邯郸、邢台、衡水等11个地级市。其中北京、天津、保定、廊坊为中

习近平 4

习近平，男，汉族，1953年6月生，陕西富平人，1969年1月参加工作，1974年1月加入中国共产党，清华大学人文社会学院马克思主义理论与思想政治教育专业毕业，在职研究

实例

战略实施[Strategy i... BD | EN

简介

战略实施是战略管理过程第三阶段活动。把战略制定阶段所确定的意图性战略转化为具体的组织行动，保障战略实现预定目标。新战略的实施常常要求一个组织在组织结构、经营过程、能力建设、资源配置、企业文化、激励制度、治理机制等方面做出相应的变化和采取相应的行动。也涉及对被实施的战略进行评估。

中文名 战略实施
释义 对战略规划的实施与执行
又名 战略执行
实现 企业战略目标

相关实例

绩效 战略制定 战略目标 战略方针 战略分析

来源

https://baike.baidu.com/item/战略实施

XLORE数据

列表	**xlore.concept.list.ttl** 精简版XLORE概念列表 **xlore.instance.list.ttl** 精简版XLORE实例列表 **xlore.property.list.ttl** 精简版XLORE属性列表
信息框	**xlore.infobox.ttl** 精简版XLORE实例的信息框，描述实例的某一属性对
上下位关系	**xlore.instanceOf.ttl** 精简版XLORE实例与概念间的InstanceOf关系 **xlore.subclassOf.ttl** 精简版XLORE概念与概念间的SubClassOf关系
文本	**xlore.instance.text.ttl** 精简版XLORE实例的摘要（文本中包含锚文本）
相关关系	**xlore.concept.related.ttl** 精简版XLORE与给定概念相关的实例和概念 **xlore.instance.related.ttl** 精简版XLORE与给定实例相关的实例和概念
跨语言链接	**xlore.concept.sameAs.ttl** 精简版XLORE不同源概念间的等价跨语言链接 **xlore.instance.sameAs.ttl** 精简版XLORE不同源实例间的等价跨语言链接
URL	**xlore.concept.url.ttl** 精简版XLORE概念对应实际词条页面的URL **xlore.instance.url.ttl** 精简版XLORE实例对应实际词条页面的URL

概念	实例	属性
跨语言链接 上位概念 下位概念 相关概念 包含实例 URL	名字 别名 提及 跨语言链接 文本描述 所属概念 相关概念 相关实体 图片 URL	– 名字相关 rdfs:label, xlore:alias, xlore:supplement, xlore:hasMention – 上下位关系 owl:SubClassOf, owl:InstanceOf – 相关关系 xlore:isRelatedTo – 等价关系 owl:sameAs – 实例基本信息 rdf:type, rdfs:comment, xlore:hasURL, xlore:hasImage – 信息框属性

XLORE Core数据统计

整体

	百度百科	中文维基	英文维基	总数
概念	32,009	150,241	326,518	508,768
实例	1,629,591	640,622	1,235,178	3,505,391
属性	157,370	45,190	26,723	229.283

上下位关系

		百度百科	中文维基	英文维基	总数
上下位关系	InstanceOf	7,584,931	1,449,925	3,032,515	12,067,371
	SubClassOf	2,784	191,577	555,538	749,899

跨语言链接

	百度百科	中文维基	英文维基
百度百科	-	10,216/336,890	4,846/303,108
中文维基	10,216/336,890	-	28,921/454,579
英文维基	4,846/303,108	28,921/454,579	-

目录

总结与展望

- 研发了基于异构维基百科知识资源的知识图谱构建技术
- 构建了包含**千万级概念实体和亿级事实**的大规模跨语言知识图谱XLORE，并提供知识服务
- 发布了高质量核心数据XLORE Core
- 进一步工作
  - 优化图谱质量
  - 融合语言知识
  - 知识动态更新
  - 知识推理挖掘

用知识为AI赋能

清华大学人工智能研究院
知识智能研究中心
Knowledge Intelligence Research Center

谢谢大家

ai.tsinghua.edu.cn/kirc/