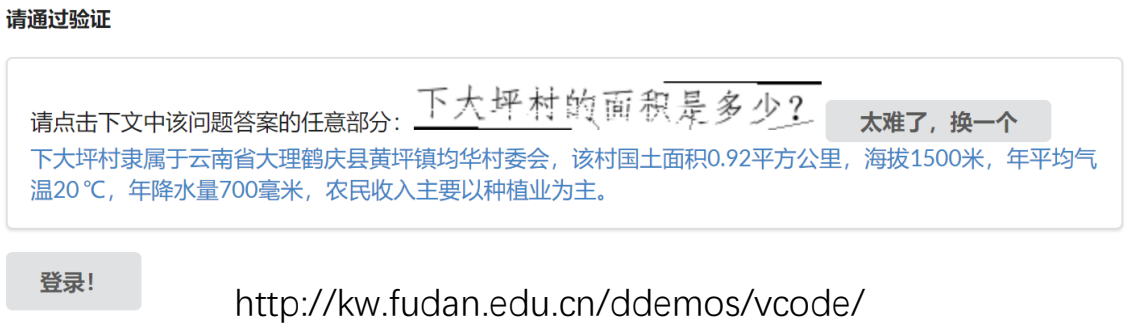
大数据时代的机遇—众包技术

- 众包与群智成为大规模知识获取的一条新路径

案例1: 基于知识问答验证码的知识获取

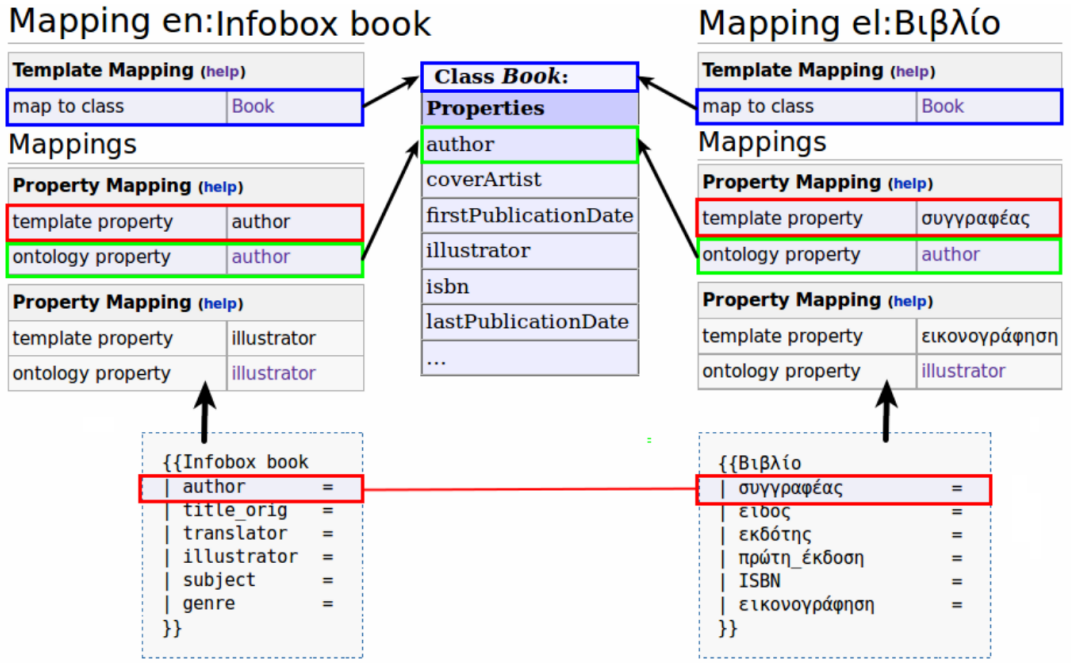
- 复旦大学知识工场实验室提供知识验证码服务，通过众包的方式对现有知识进行验证



http://kw.fudan.edu.cn/ddemos/vcode/

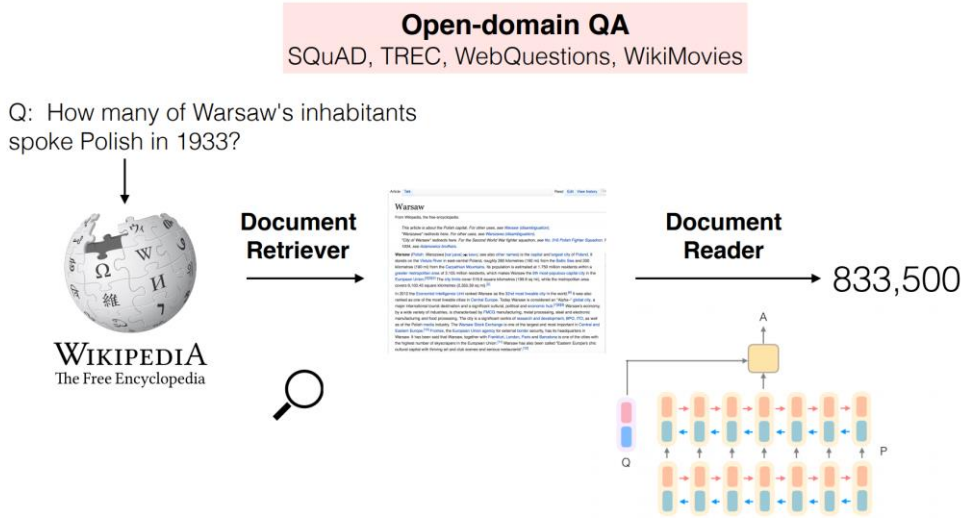
案例2:基于众包的Taxonomy构建

- DBpedia通过众包方式构建了DBpedia Ontology



大数据时代的机遇—高质量UGC

- Web2.0时代到来，产生大量的高质量UGC(User Generated Content)
  - 提供获得广大用户一致认可的高质量数据源
    - Wikipedia, 百度百科
  - 为自动挖掘知识提供了高质量数据源
  - 为构建抽取模型提供了高质量样本



Ref: Danqi Chen, etc.. Reading Wikipedia to Answer Open-Domain Questions

周杰伦 共被编辑4812次

版本对比	更新时间	全部版本	贡献者	修改原因	区块链信息
	2018-06-06 03:36	查看	w_ou	内链修复	查看
	2018-03-11 16:14	查看	海澜～～～天空	内容扩充 内链	查看
	2018-03-01 20:20	查看	爱锦瑟的年华	图片	查看
	2018-02-28 18:59	查看	爱锦瑟的年华	内容扩充 参考资料	查看
	2018-02-15 08:05	查看	紫雪510	内容扩充 参考资料	查看
	2018-02-11 20:54	查看	5ssax	更正错误 图片	查看
	2018-02-10 11:31	查看	Mini小北1992	完善作品信息	查看

Wiki和百科的编辑机制保证了UGC内容的质量

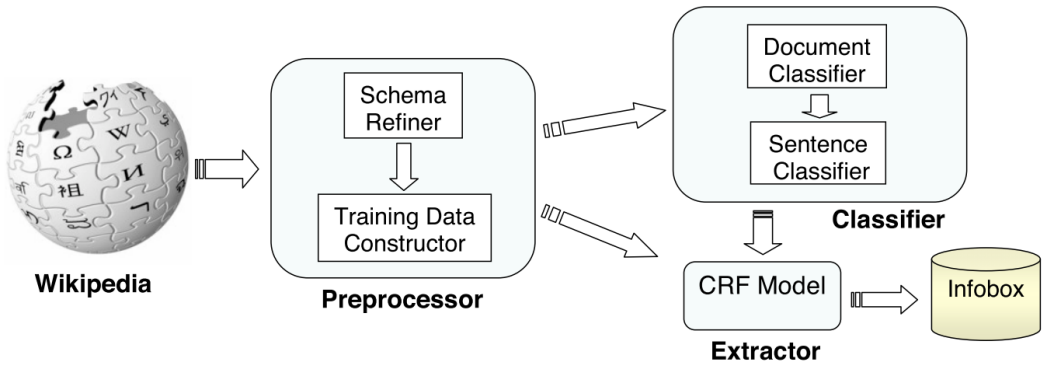


Figure 3: Architecture of KYLIN's infobox generator.

Ref: Fei Wu, etc.. Autonomously Semantifying Wikipedia

知识图谱的研究意义

未来已至：人类已经进入智能时代

- 大数据的日益积累、计算能力的快速增长为人类进入智能时代奠定了基础
- 大数据为智能技术的发展带来了前所未有的数据红利
- 机器计算智能、感知智能达到甚至超越人类

2012年，在图像识别的国际大赛ILSVRC(大型视觉辨识挑战竞赛)中，加拿大多伦多大学的研究团队基于深度卷积神经网络的模型[1]夺冠，把TOP5错误率降到15.3%，领先第二名超过十个百分比，震惊学术圈。

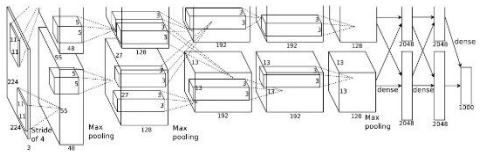
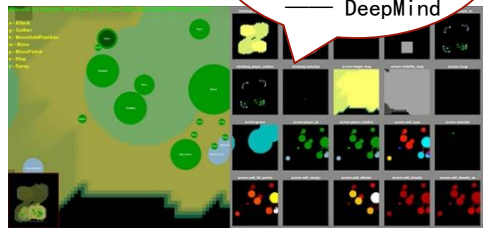


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

2016年，Google全资收购的DeepMind推出名为AlphaGo的围棋程序[2]，以4:1的总比分击败世界顶级职业围棋选手李世石，让全世界开始关注人工智能技术巨大的应用前景。

2017年，DeepMind联合游戏公司暴雪，宣布共同开发可以在“星际争霸2”中与人类玩家对抗的人工智能，并且发布了旨在加速即时战略游戏的人工智能应用的工具集[3]。

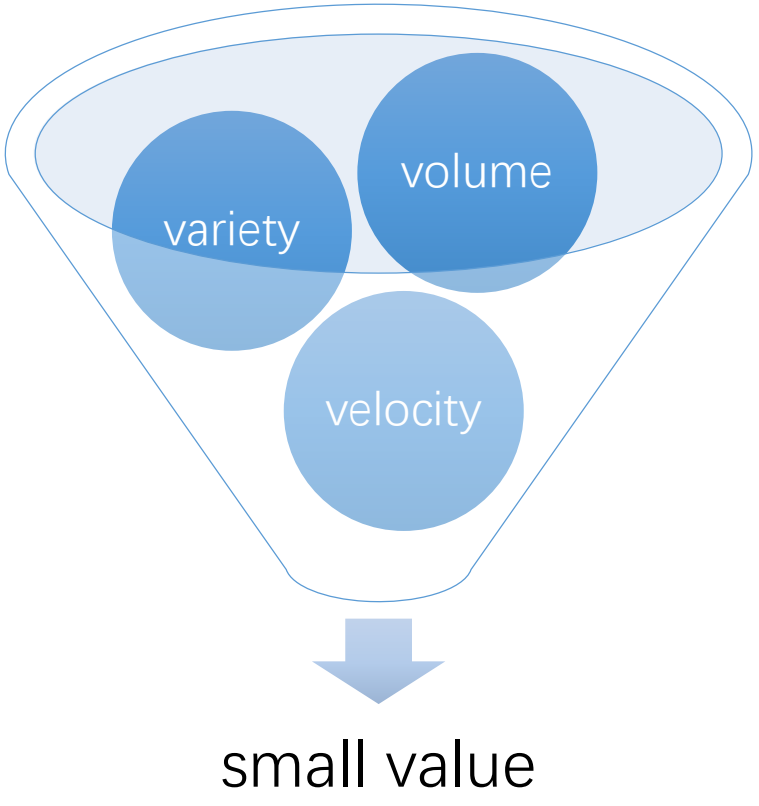


各行业智能化升级与转型

- 宏观形势
  - 人口红利消失
  - 专家成本高昂
  - 实体经济结构转型
  - 传统行业发展内涵升级
- 技术发展态势
  - 数据丰富
  - 场景丰富
  - 丰富AI技术积累

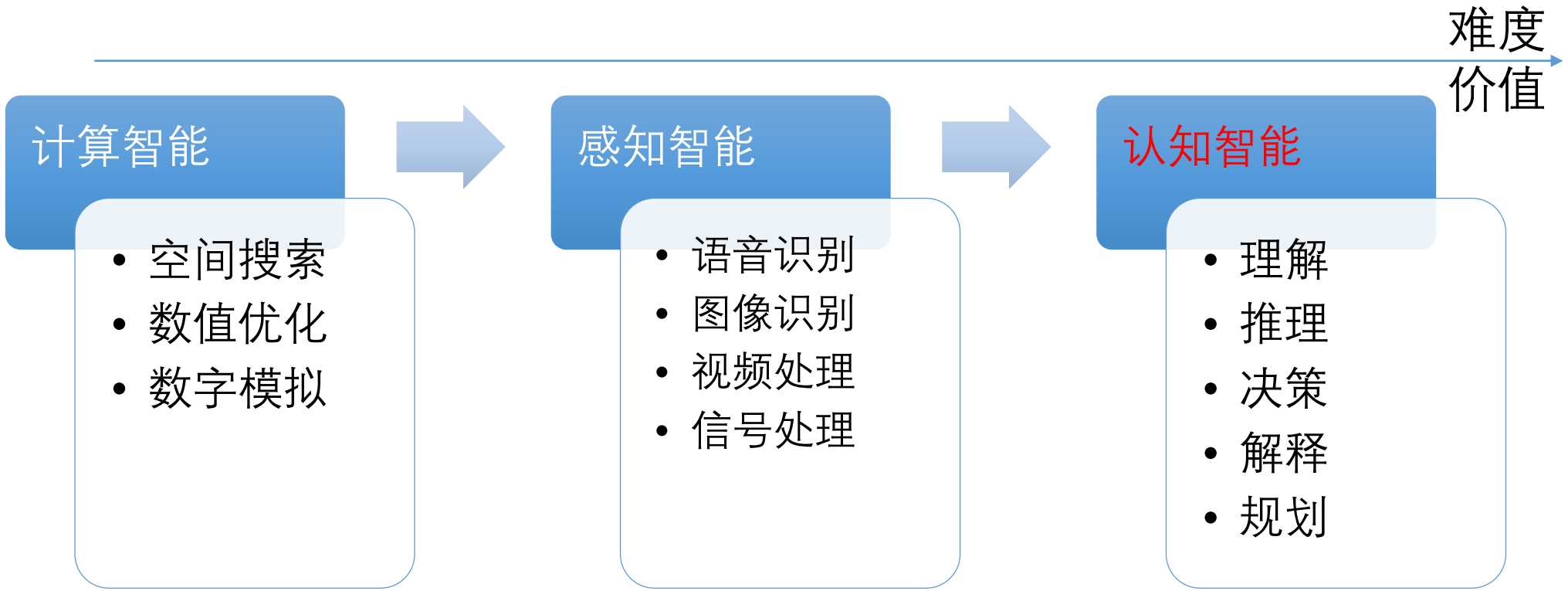
大数据价值变现困难倒逼智能化转型

- 《中国经济周刊》
  - “**投上百亿建大数据中心 内部称产出十分微小”
- 英特尔中国研究院院长吴甘沙
  - “鉴于大数据信息密度低，大数据是贫矿，投入产出比不见得好。
- 李国杰院士
  - “实际上，大数据的价值，主要体现在它的驱动效应上，大数据对经济的贡献，并不完全反映在大数据公司的直接收入上，应考虑对其他行业效率和质量提高的贡献。”



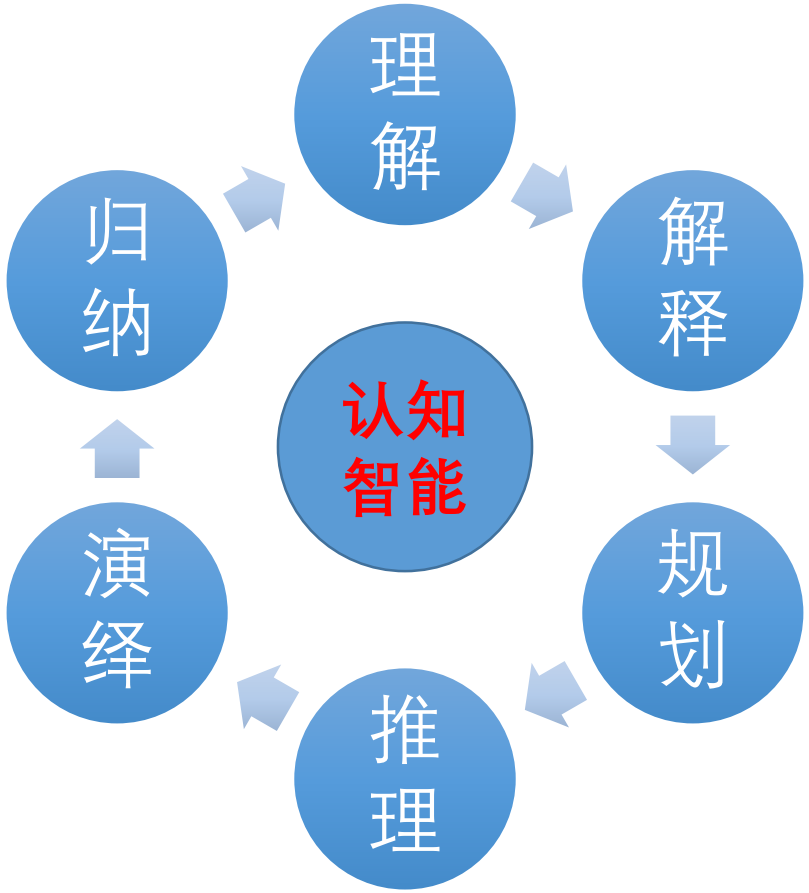
大数据价值变现的尴尬现状：高射炮打蚊子，大材小用

智能化需要机器智能，特别是认知智能



- 随着数据红利消耗殆尽，以深度学习为代表的感知智能遇到天花板
- 认知智能将是未来一段时期内AI发展的焦点，是进一步释放AI产能的关键

认知智能是智能化的关键

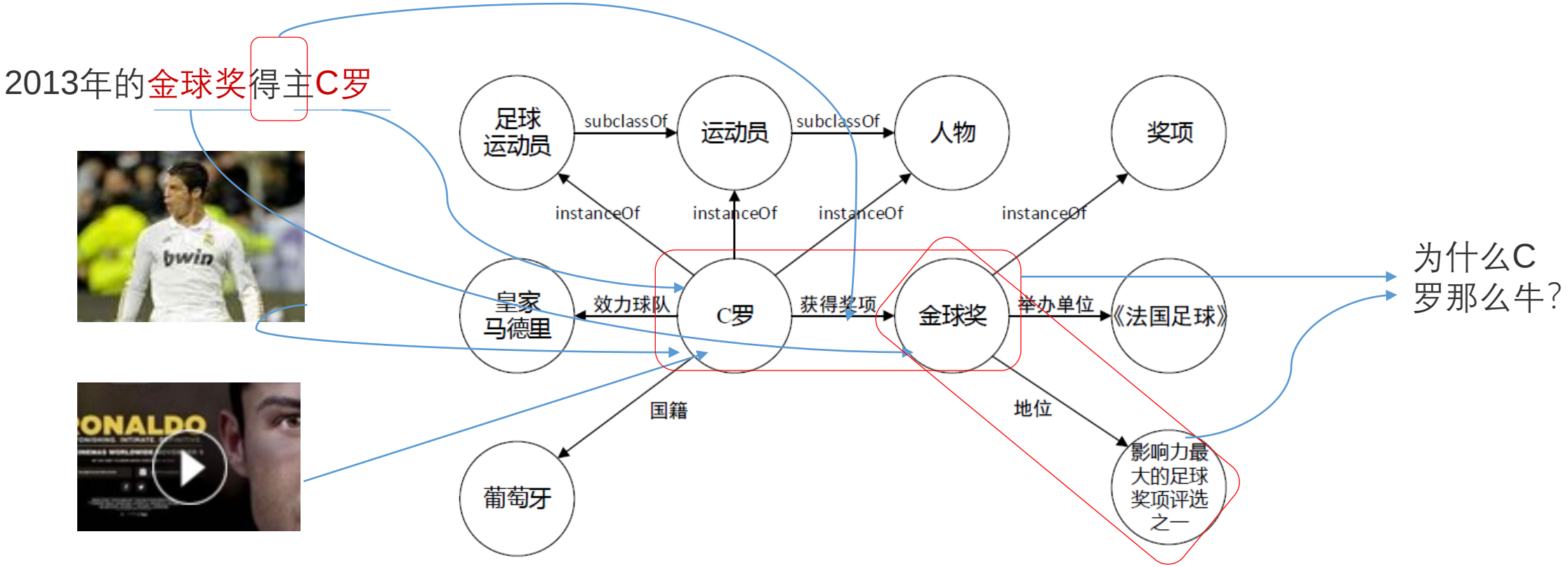


Can machine ***think like humans***?

- 理解与解释是后深度学习时代人工智能的核心使命之一

知识图谱使能认知智能

- 机器理解数据的本质：建立从数据到知识库中实体、概念、关系的映射
- 机器解释现象的本质：利用知识库中实体、概念、关系解释现象的过程



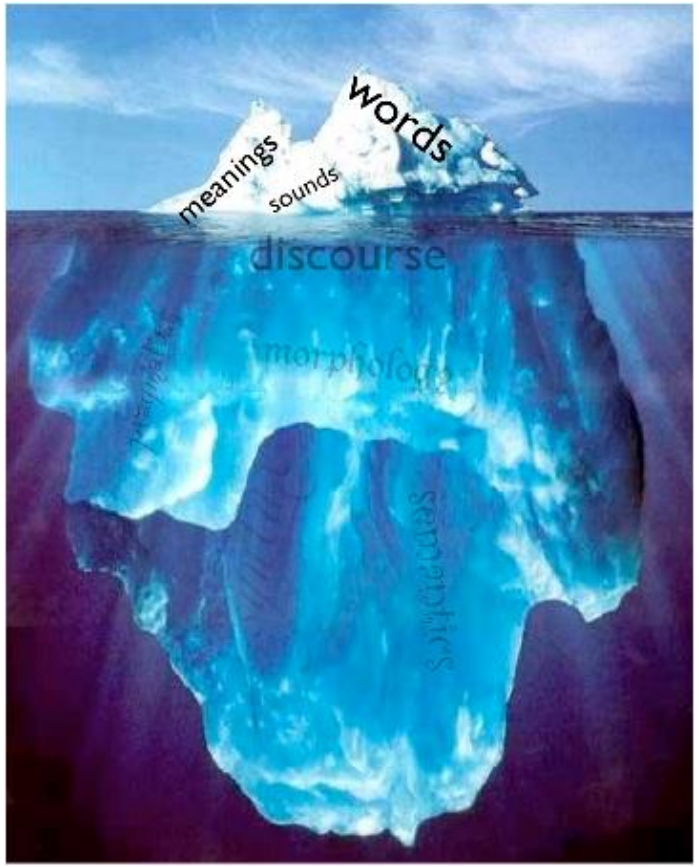
机器语言理解需要背景知识

Language is complicated

- **Ambiguous**, **contextual** and **implicit**
- Seemingly **infinite** number of ways to express the same meaning

Language understanding is difficult

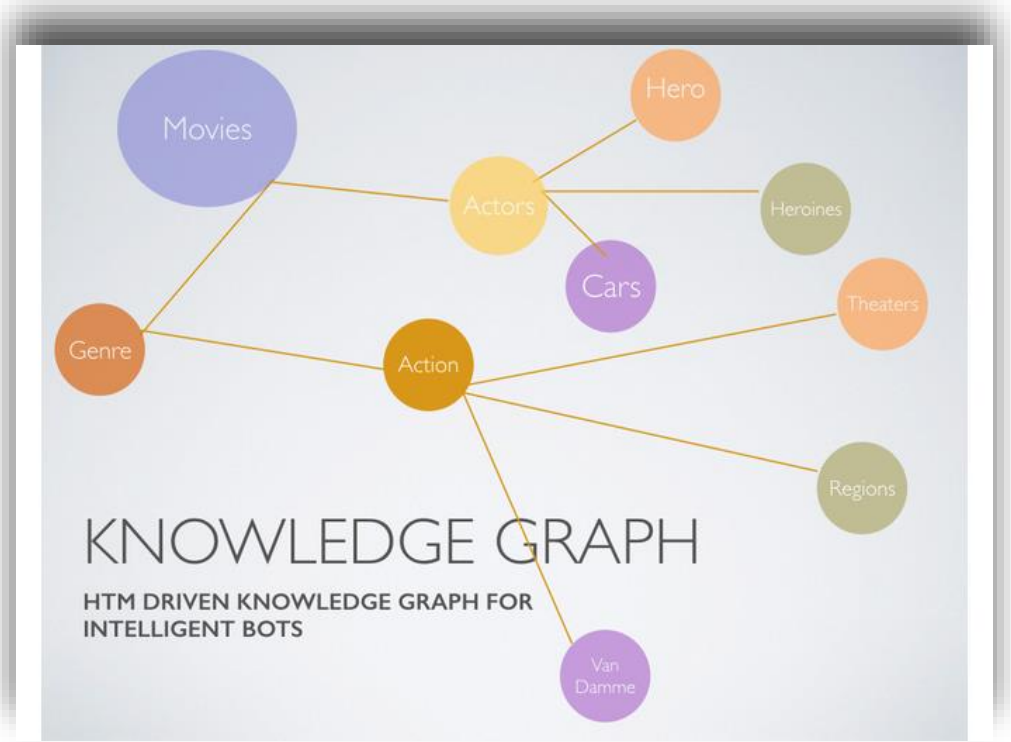
- Grounded only in **human cognition**
- Needs significant **background knowledge**



知识图谱*使能*(Enable)机器语言认知

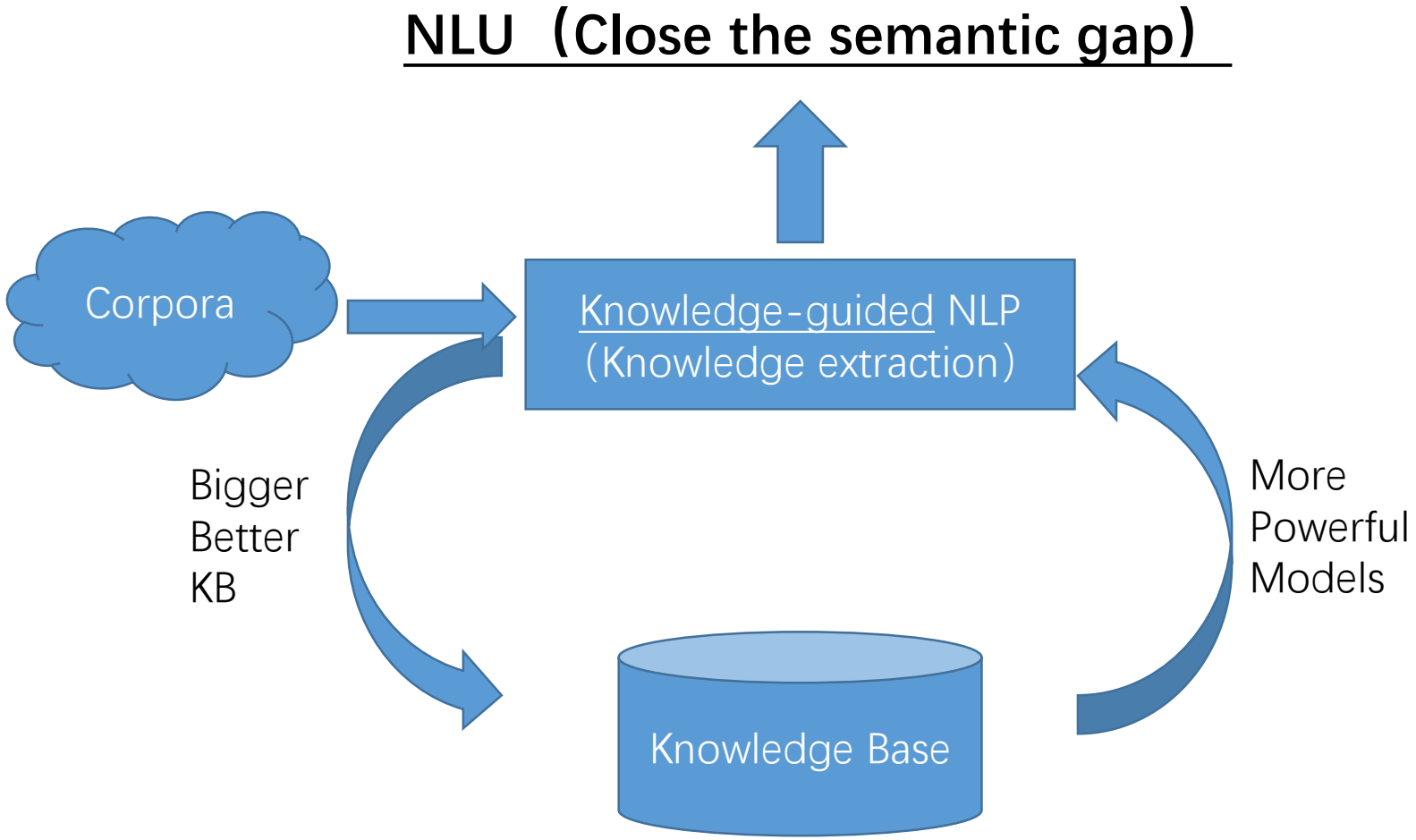
- Language understanding of machines needs knowledge bases
  - Large scale
  - Semantically rich
  - Friendly structure
  - High quality

- Traditional knowledge representations can not satisfy these requirements, but KG can
  - Ontology
  - Semantic network / frame
  - Texts



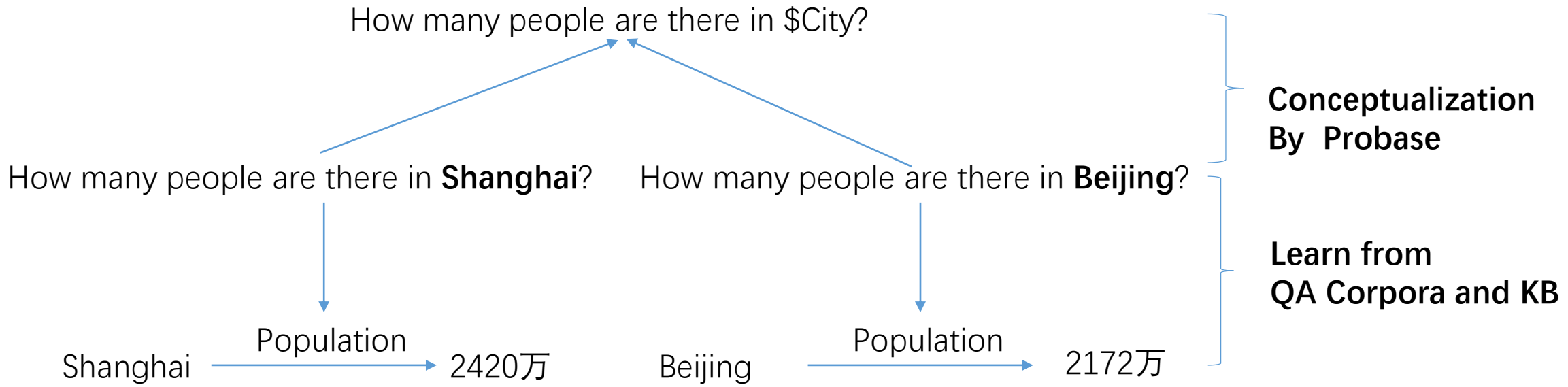
NLP+KB= NLU， NLP=Natural language processing， NLU=natural language understanding

The roadmap of knowledge-guided NLP



Example: Using concepts to understand a natural language?

- Representation: concept based templates.
  - Questions are asking about entities. The semantic of the question is reflected by its corresponding concept.
  - Advantage: Interpretable, user-controllable
- Learn templates from QA corpus, instead of manfully construction.



[Wanyun Cui et al. 2017]

知识图谱*使能*可解释人工智能

鲨鱼为什么那么可怕?
因为它们是食肉动物

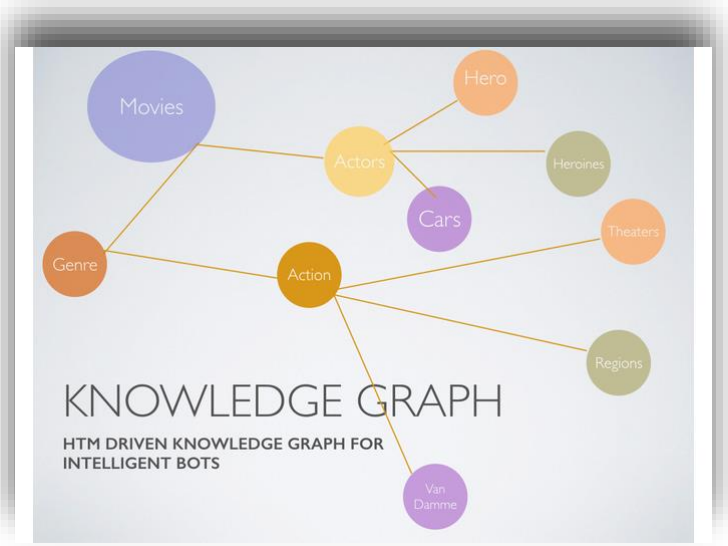
鸟儿为何能够飞翔?
因为它们有翅膀

鹿晗关晓彤最近为何刷屏?
因为关晓彤是鹿晗女朋友

概念

属性

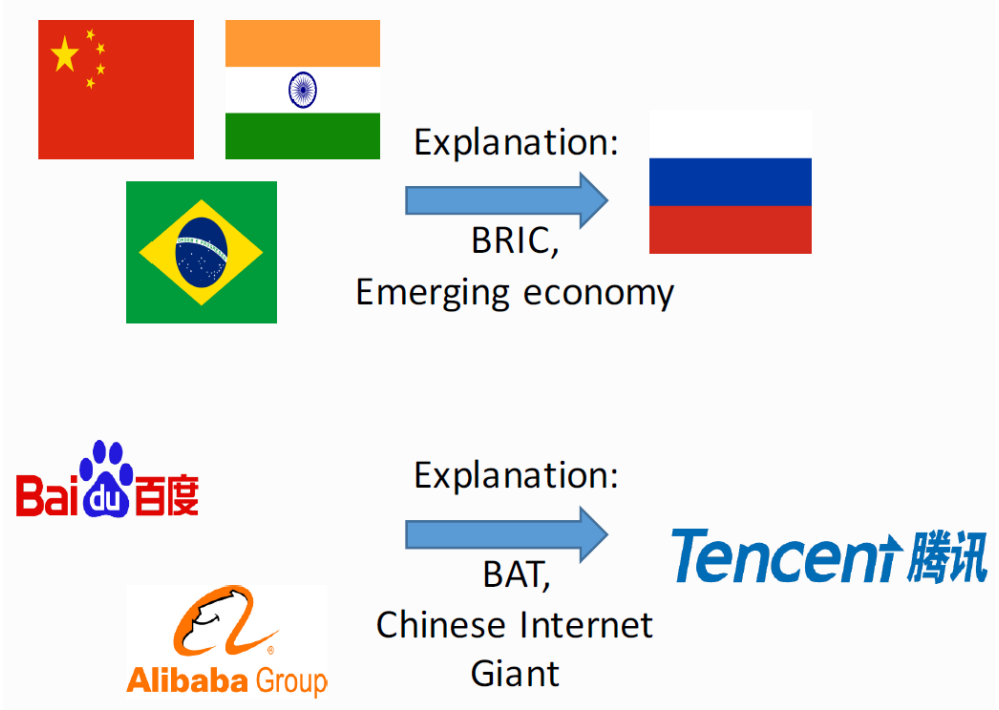
关系



解释取决于人类认知的基本框架;
概念、属性、关系是认知的基石

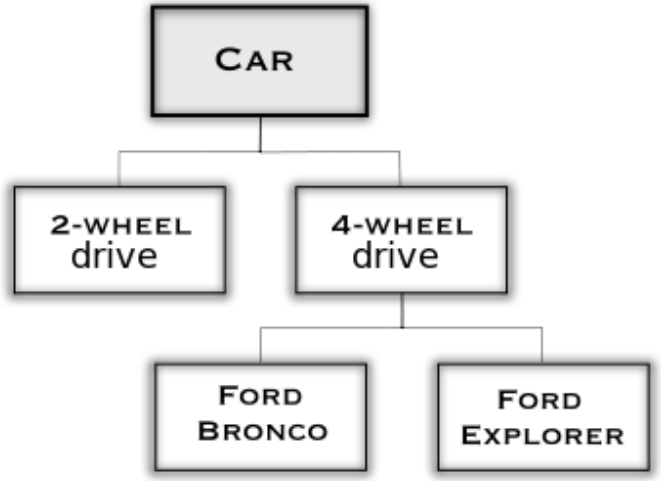
"Concepts are the glue that holds our mental world together"
--Gregory Murphy

Example 1: Explainable entity recommendation using taxonomy



Problem:
Given a set of entities, can we understand its concept and recommend a most related entity?

Applications:
E-commerce: if users are searching samsung s6, and iPhone 6, what should we recommend and why?



Taxonomy

[Yi Zhang, et al, 2017]

Example 2: Explain a Concept/Category using Properties

Problem:

How do we understand a concept/category?

Example:

How to understand “Bachelor”

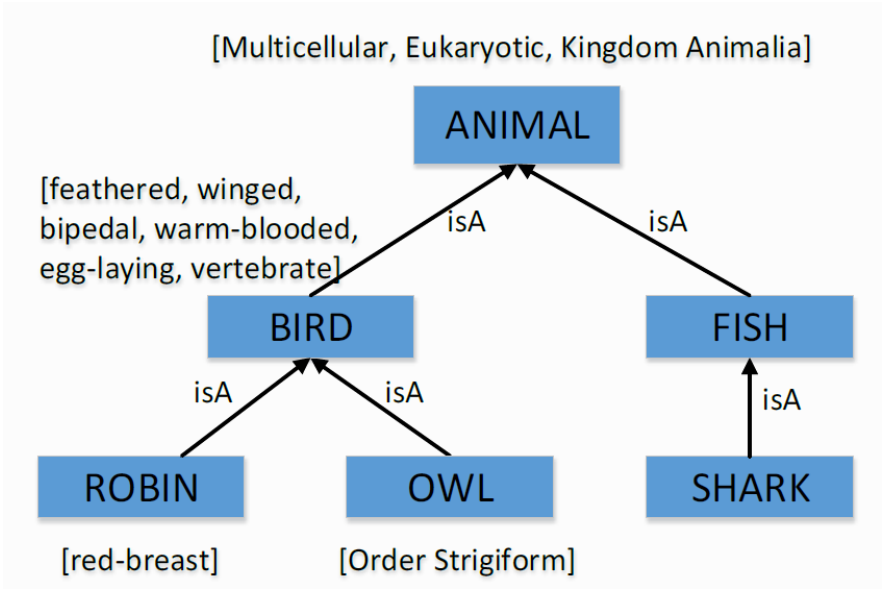
=> (Sex=man, Marriage status=unmarried)

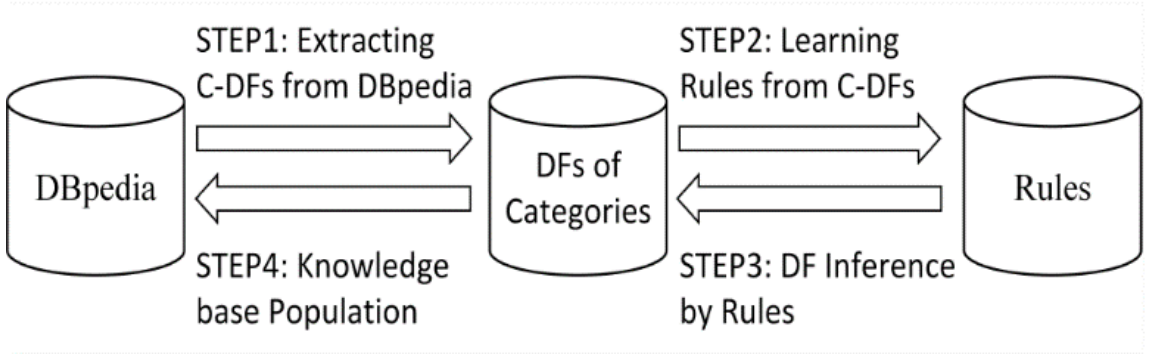
Basic Idea:

Mining Dbpedia, using properties to explain a category

Model:

Mining **Defining Features** from DBpeida

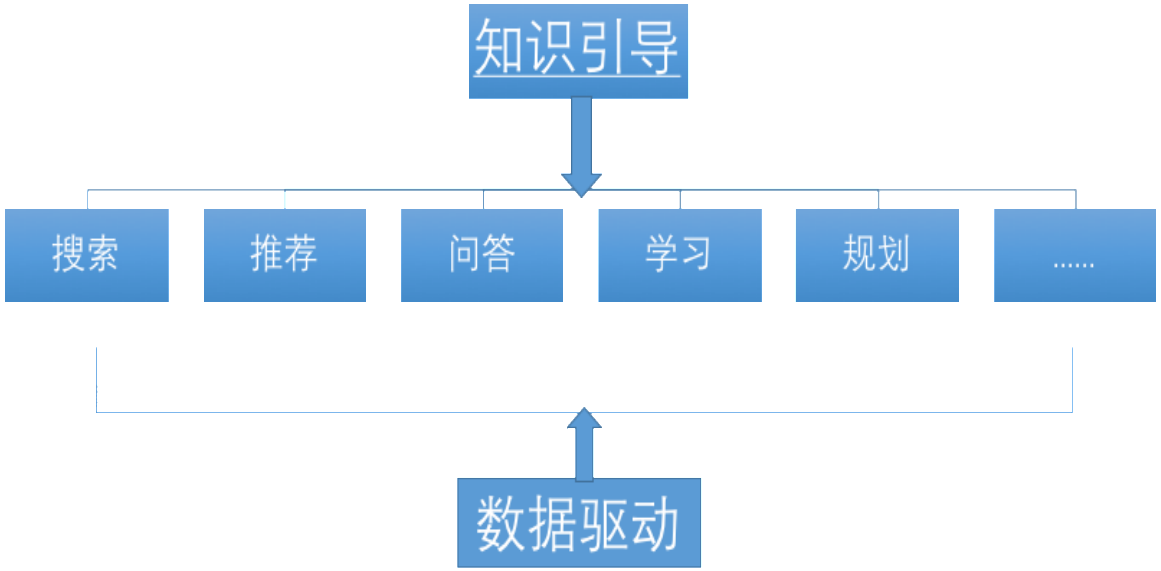


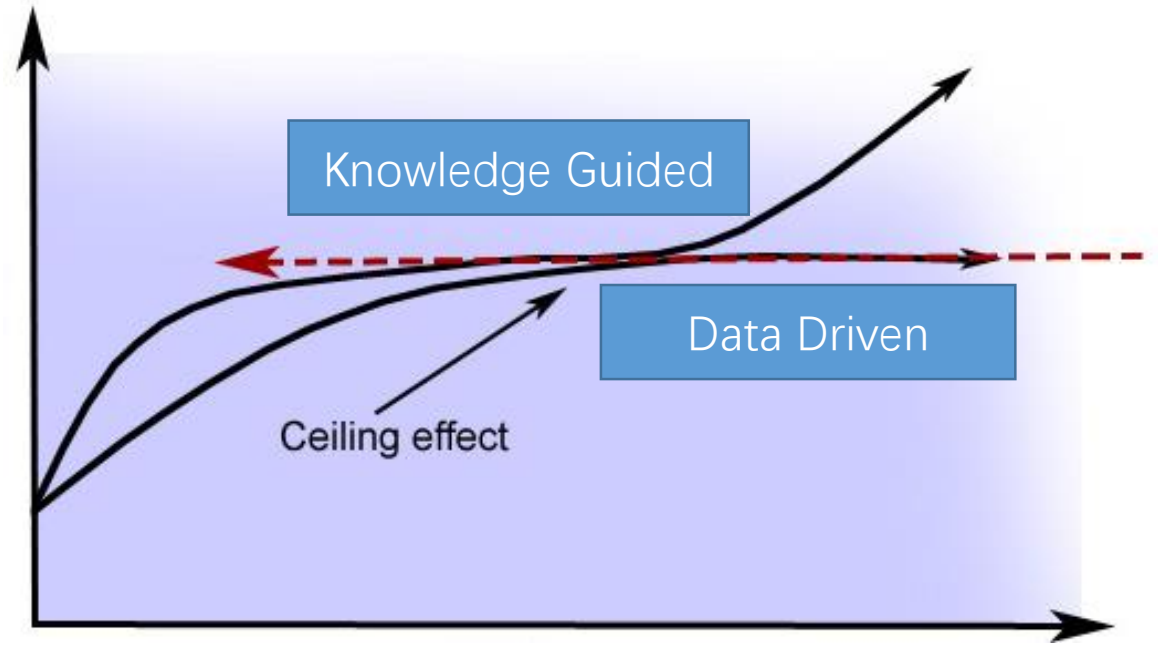


Solution Framework

[Bo Xu, et al, 2016]

知识引导将成为解决问题的主要方式





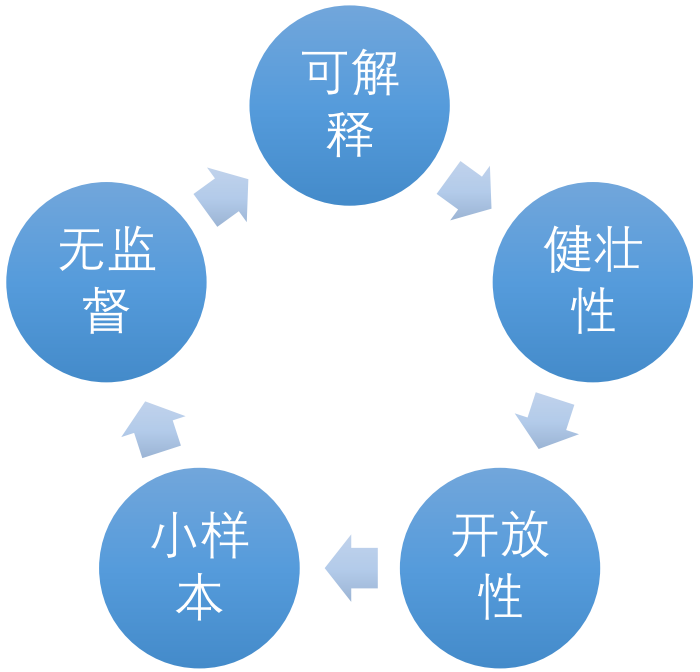
- “数据驱动”利用统计模式解决问题
- 单纯依赖统计模式难以有效解决很多实际问题

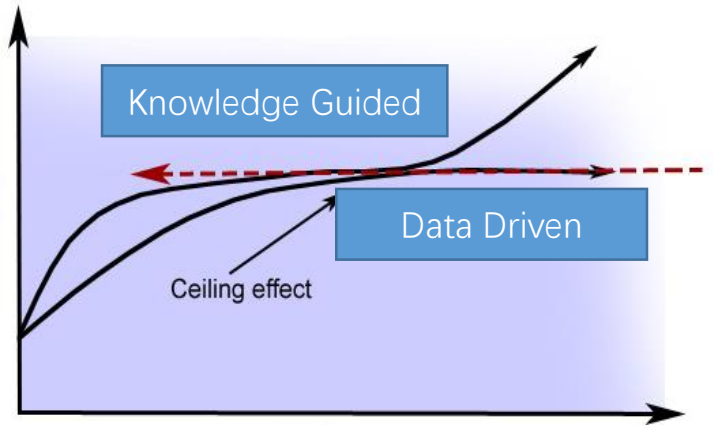
张三把李四打了，他进医院了

张三把李四打了，他进监狱了

知识引导突破统计学习的天花板

- 依赖数据驱动的统计学习基于统计模式解决问题
- 单纯统计模式日益面临性能的天花板



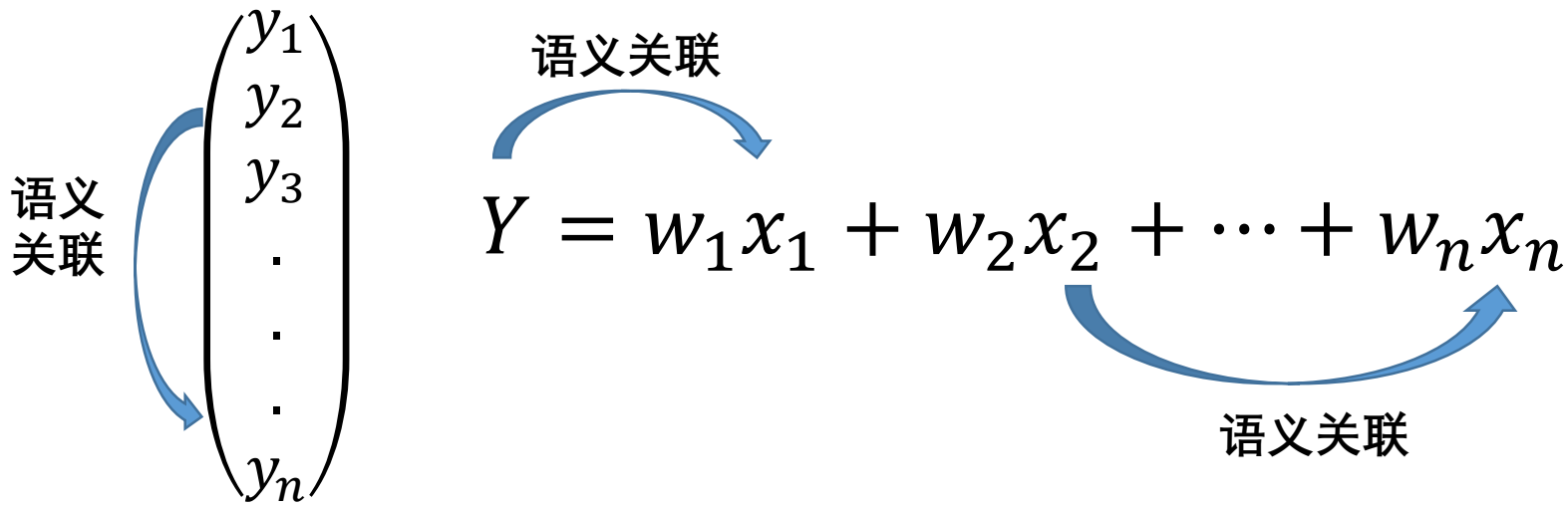


张三把李四打了，他进医院了

张三把李四打了，他进监狱了

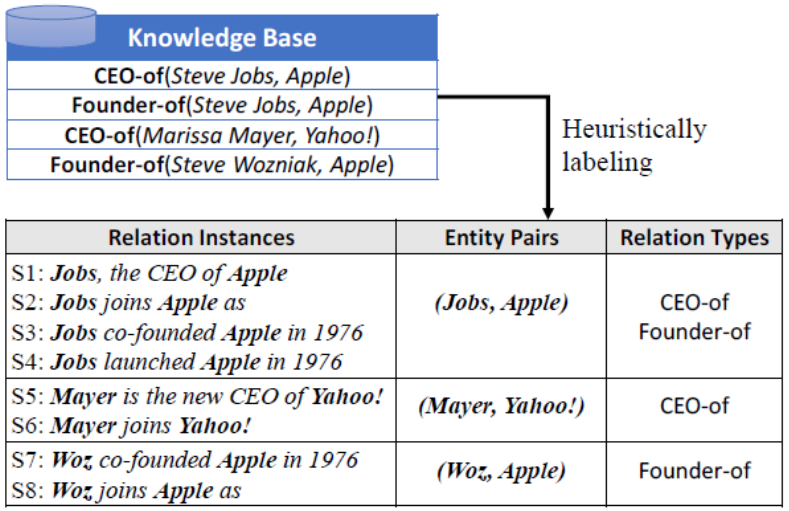
开放性：提升机器学习的开放性

- 难以处理开放性问题
  - Zero-shot learning: Unknown Labels
  - One-shot learning: Rare labels
- 忽略解释变量、响应变量及其之间的各类语义关联



无监督：知识库给机器学习提供丰富的样本

- 远程监督（Distant Supervision）提供大规模自动化弱标注样本
- 领域专家构建的知识库质量精良，提供高质量的种子样本



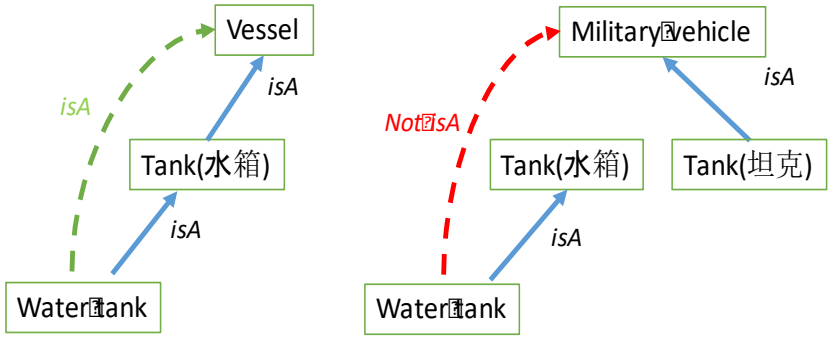
Relation Instances	Entity Pairs	Relation Types
S1: *Jobs, the CEO of Apple* S2: *Jobs joins Apple as* S3: *Jobs co-founded Apple in 1976* S4: *Jobs launched Apple in 1976*	*(Jobs, Apple)*	CEO-of Founder-of
S5: *Mayer is the new CEO of Yahoo!* S6: *Mayer joins Yahoo!*	*(Mayer, Yahoo!)*	CEO-of
S7: *Woz co-founded Apple in 1976* S8: *Woz joins Apple as*	*(Woz, Apple)*	Founder-of

Idea: 通过结构化知识库与文本比对，完成大规模弱标注，广泛应用于实体识别、关系抽取等任务

Ref: Distant supervision for relation extraction without labeled data. ACL09

Lexiacal Taxonomy isA传递性判定问题

Example：Einstein isA Physicist, Physicist isA Job, is Einstein a Job?



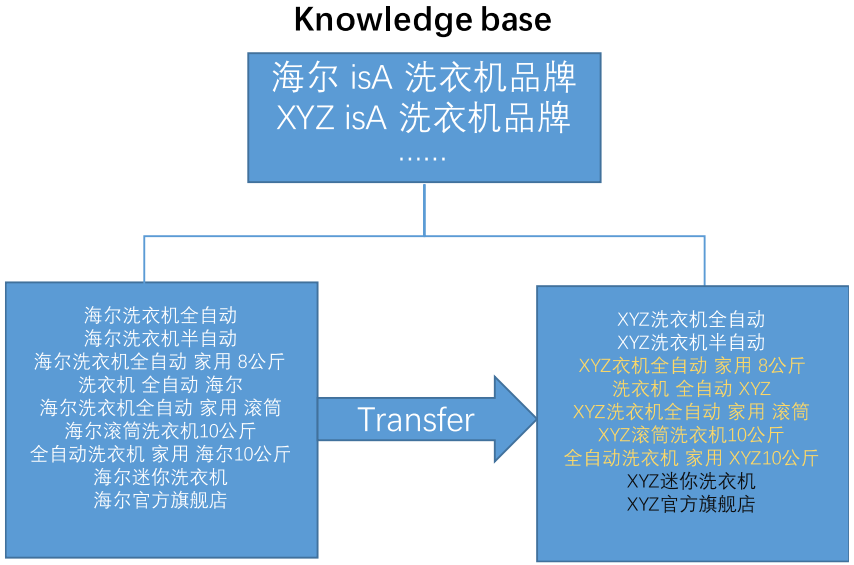
正例: water tank - tank - vessel 负例: water tank - tank - military vehicle

Idea: 使用专家构建的WordNet，自动化构造判断isA传递性的标注样本

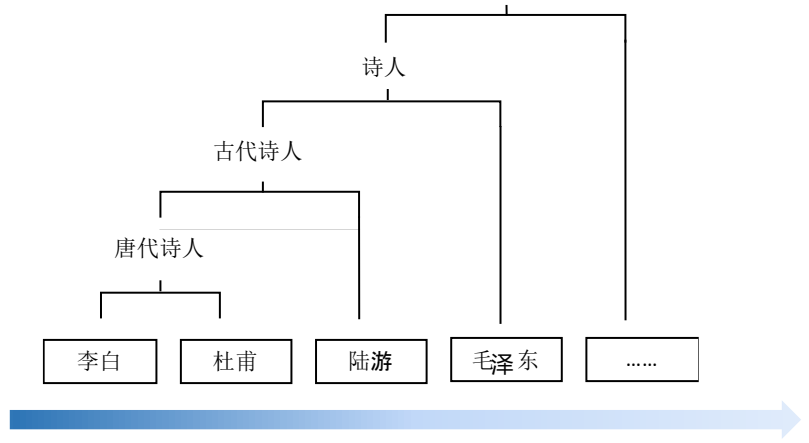
Ref：On the Transitivity of Hypernym-hyponym Relations in Data-Driven Lexical Taxonomies, (**AAAI 2017**)

小样本：知识引导下的样本增强

- 符号知识可以广泛用于指导样本的生成、选择、增强、优化
  - 边界样本选择、负样本选择、代表性样本选择、Unknown Unknowns识别
- 有效应对样本稀缺、有效处理长尾对象



Idea: 使用Taxonomy指导样本生成，提升长尾品牌词的embedding效果



Positive Sample:
Q：李白被誉为什么？ A：诗仙
Negative Samples:
Q：杜甫的中文名是?
Q：陆游经典作品有哪些?
……

Idea: 使用Taxonomy指导负样本生成，提升问答准确性；落地在知识工场知识问答系统（不倒翁问答）在通用领域做到85%的准确率

健壮性：符号知识优化机器学习模型

- 符号知识被广泛用于，提升机器学习对于有偏样本的健壮性
  - 构建正则项、约束、事后检验、注意力机制

$$\text{Maximize} \sum_{t \in T} \left(\max_{m \in M_e} P(t|m) - \theta \right) \times x_{e,t}$$

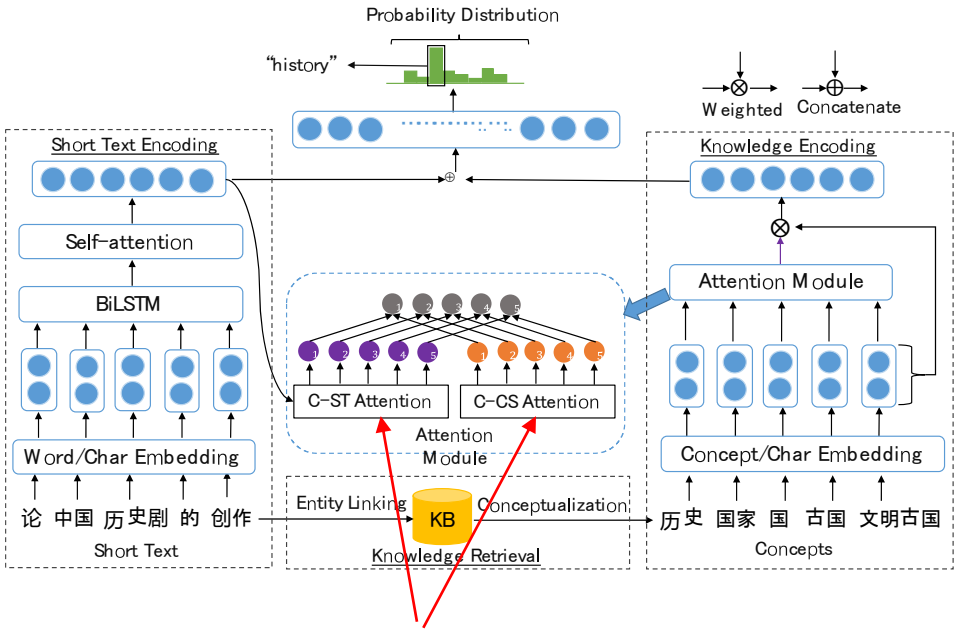
Subject to

$$\forall_{ME(t_1,t_2)}\ x_{e,t_1} + x_{e,t_2} \leq 1$$

$$\forall_{ISA(t_1,t_2)}\ x_{e,t_1} - x_{e,t_2} \leq 0$$

Type Disjointness Constraint

Type Hierarchy Constraint



Idea: 使用Type之间的语义约束对于结果进行筛选

Ref, METIC: Multi-Intance Entity Typying from Corpus, CIKM 2018

Idea: 使用CN-Probase中的概念关系构建Attention，优化端文本分类模型

Ref; Deep Short Text Classification with Knowledge Powered Attentions, (**AAAI 2019**)

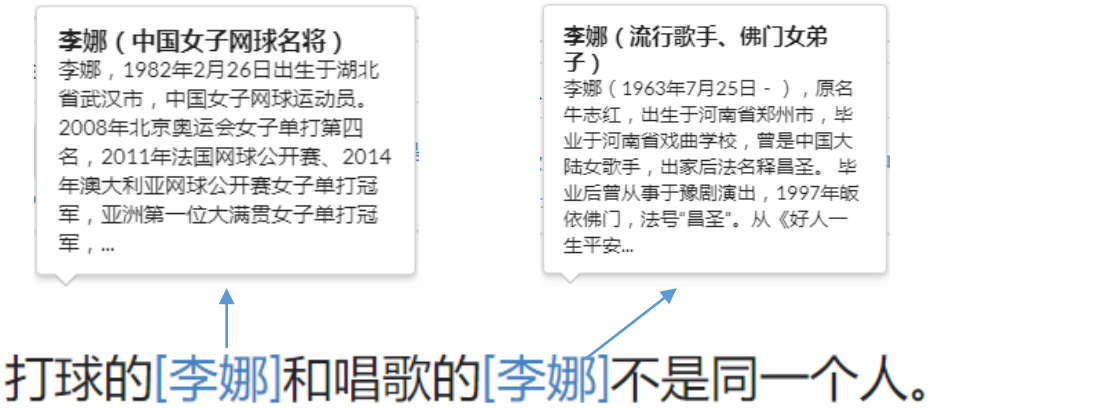
Example 1: Use Concepts for Chinese Entity Linking

- Entity linking： P(e|C),
  - where C is context and e is candidate entity
- Basic idea: using concepts (t) in knowledge base

$$P(e_i|C) = \sum_t P(e_i|t) \times P(t|C)$$

Typicality of an entity within a concept

The probability to observe an entity of t given context C



李娜（中国女子网球名将）：人物、体育人物、运动员、名将

李娜（流行歌手、佛门女弟子）：人物、演员、歌手、弟子

	** Entity Annotation API	Our Method
Precision	56.7%	86.1%
Recall	67.8%	84.5%
F1	61.7%	85.3%

Example 2: Using knowledge to prevent semantic drift in pattern based IE

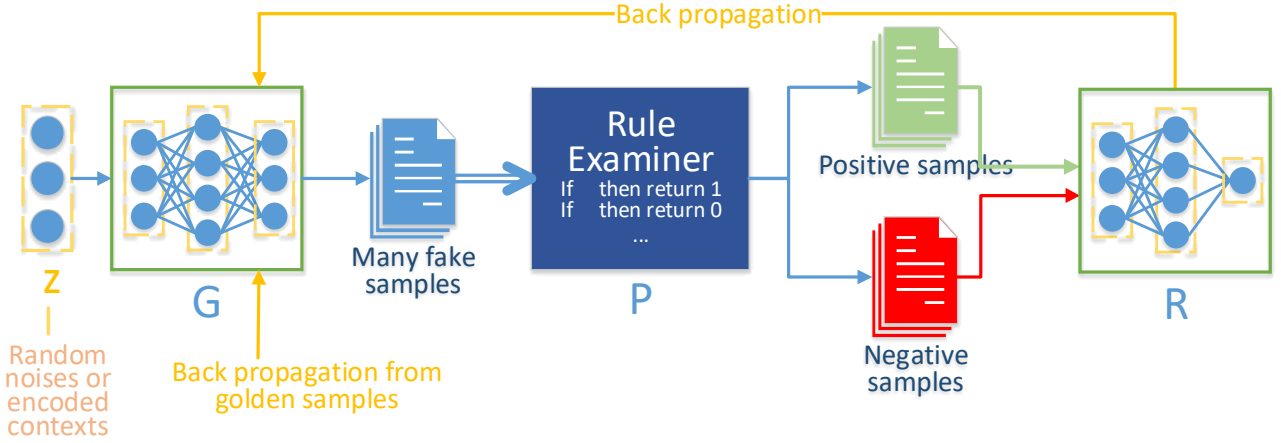
- Pattern based bootstrapping is popular
- Problem: semantic drift
  - <China isA country> =>
  - ‘occupation of $’, =>
  - ‘occupation of Planet earth’=>
  - <Planet Earch isA coutntry>
- Principles: no bad patterns, only wrong applications
- Our idea
  - Run a pattern on the text for an appropriate entity
  - Using knowledge to guide the execution of the learned pattern
  - 95%+ accuracy

#复旦大学（Fudan University），简称“复旦”，位于中国上海，由中华人民共和国教育部直属，中央直管副部级建制，位列211工程、985工程、双一流A类，入选“珠峰计划”、“111计划”、“2011计划”、“卓越医生教育培养计划”，为“九校联盟”成员、中国大学校长联谊会成员、东亚研究型大学协会成员、环太平洋大学协会成员，是一所世界知名、国内顶尖的综合性研究型的全国重点大学。\n复旦大学创建于1905年，原名复旦公学，是中国人自主创办的第一所高等院校，创始人为中国近代知名教育家马相伯，首任校董

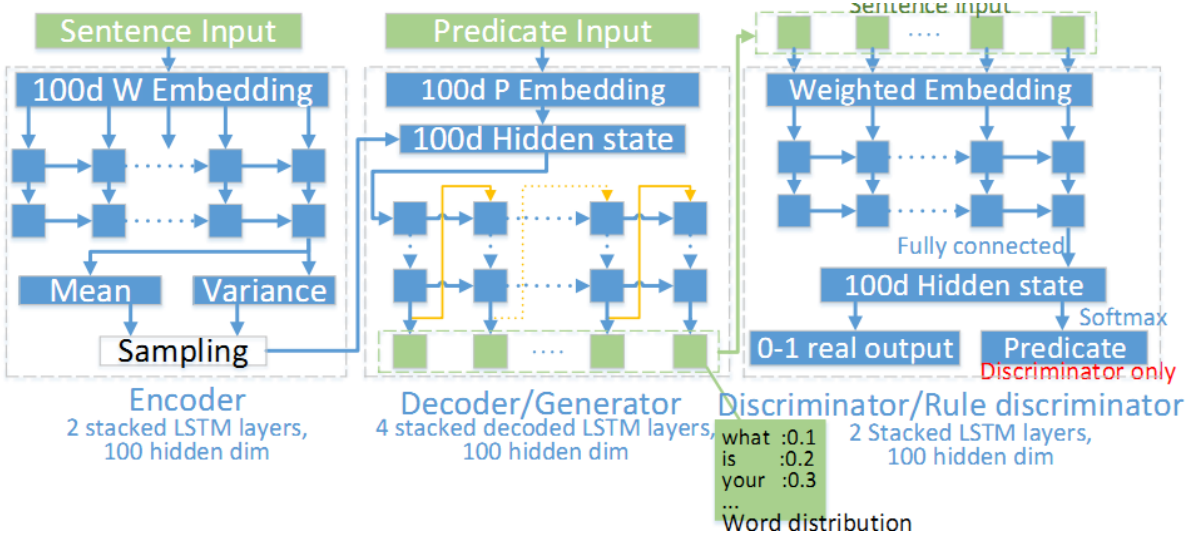
<复旦大学 - 简称 - 复旦>
<复旦大学 - 创始人 - 马相伯>
……

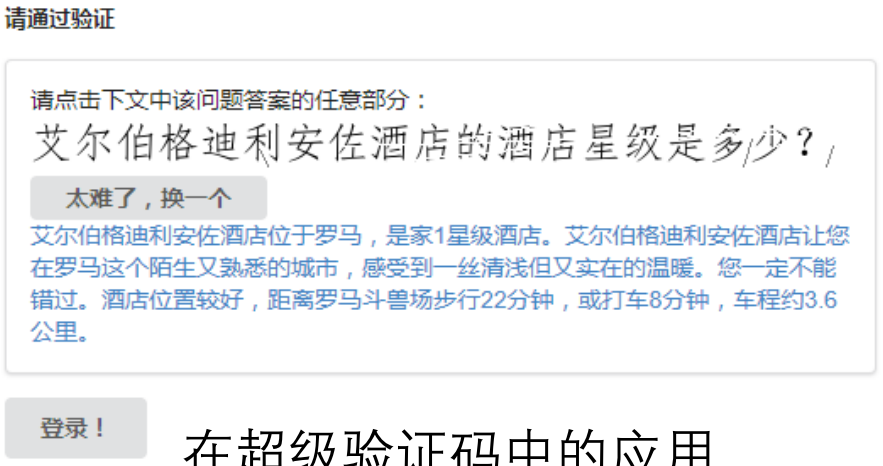
鹿晗	外文名称	LU HAN
鹿晗	出生日期	1990年4月20日
刘诗诗	职业	影视出品人
张艺兴	外文名称	LAY
张艺兴	出生日期	1991年10月7日
angelababy	出生日期	1989年2月28日
赵丽颖	出生地	河北省廊坊市
杨幂	出生日期	1986年9月12日
郑爽（中国内地90后女演员）	出生地	辽宁省沈阳市
宋茜	外文名称	Victoria
宋茜	出生日期	1987年2月2日
宋茜	职业	广告模特在亚洲地区正式开始演艺活动
刘德华（中国香港男演员、歌手、词作人）	出生日期	1961年9月27日
刘德华（中国香港男演员、歌手、词作人）	代表作品	只知道此刻爱你
李易峰	出生日期	1987年5月4日
李易峰	出生地	四川成都
李易峰	代表作品	小先生
周杰伦（华语流行男歌手）	出生日期	1979年1月18日
周杰伦（华语流行男歌手）	主要成就	台湾电影金马奖年度台湾杰出电影
周杰伦（华语流行男歌手）	代表作品	Jay

Example 3: Deep language generation with prior knowledge



Rules for Chinese question generation
1 Sentences should end with '#' (a special character).
2 The subject should appear only once in a sentence.
3 There are no continuously repeated characters in the sentence.
4 The length of sentences should be more than 4 characters. *(A Chinese question should not be too short.)*
5 The number of low frequency words should less than half of sentence length.





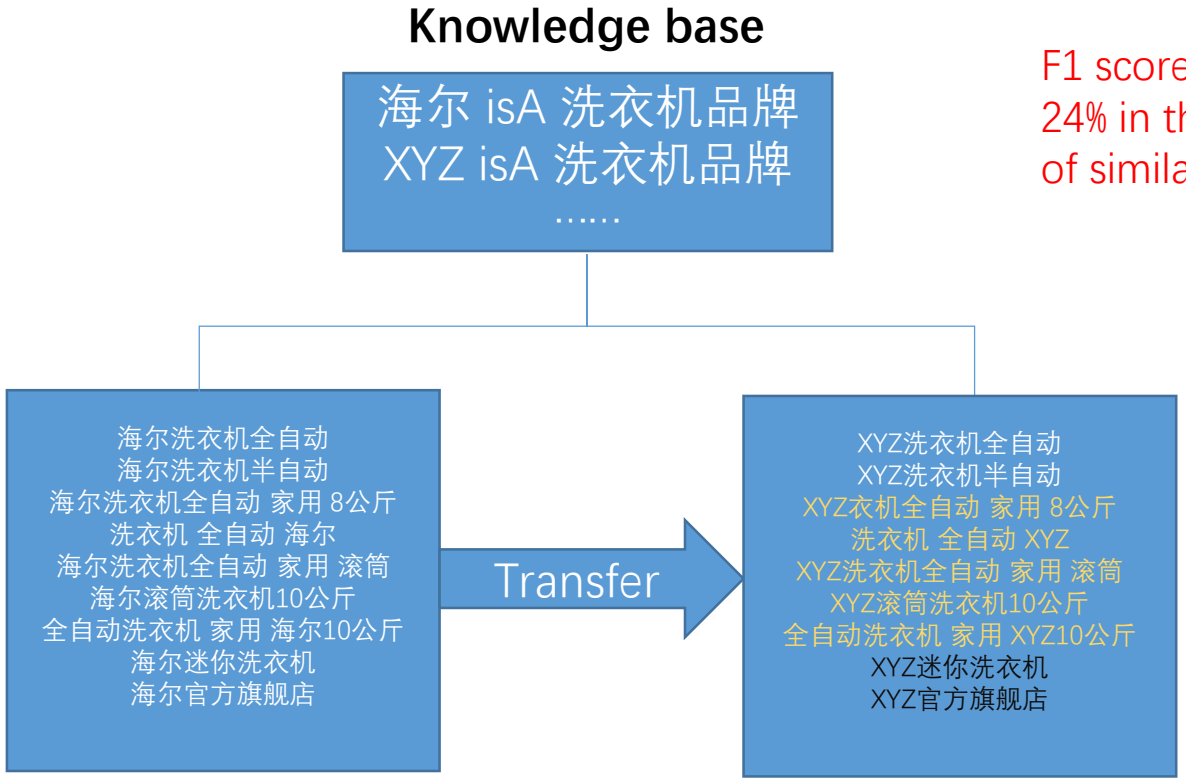
在超级验证码中的应用

Demo地址：http://kw.fudan.edu.cn/ddemos/vcode/
API地址：http://kw.fudan.edu.cn/apis/supervcode/

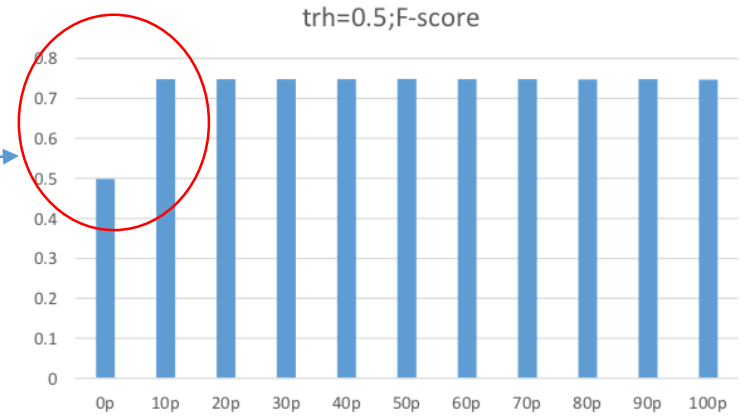
Incorporating Complicated Rules in Deep Generative Models, under review

Example 4: Long-tailed query term embedding guided by knowledge

- In Deep IR, it's hard to train effective word embedding for long tailed query terms



F1 score increases by 24% in the evaluation of similar queries



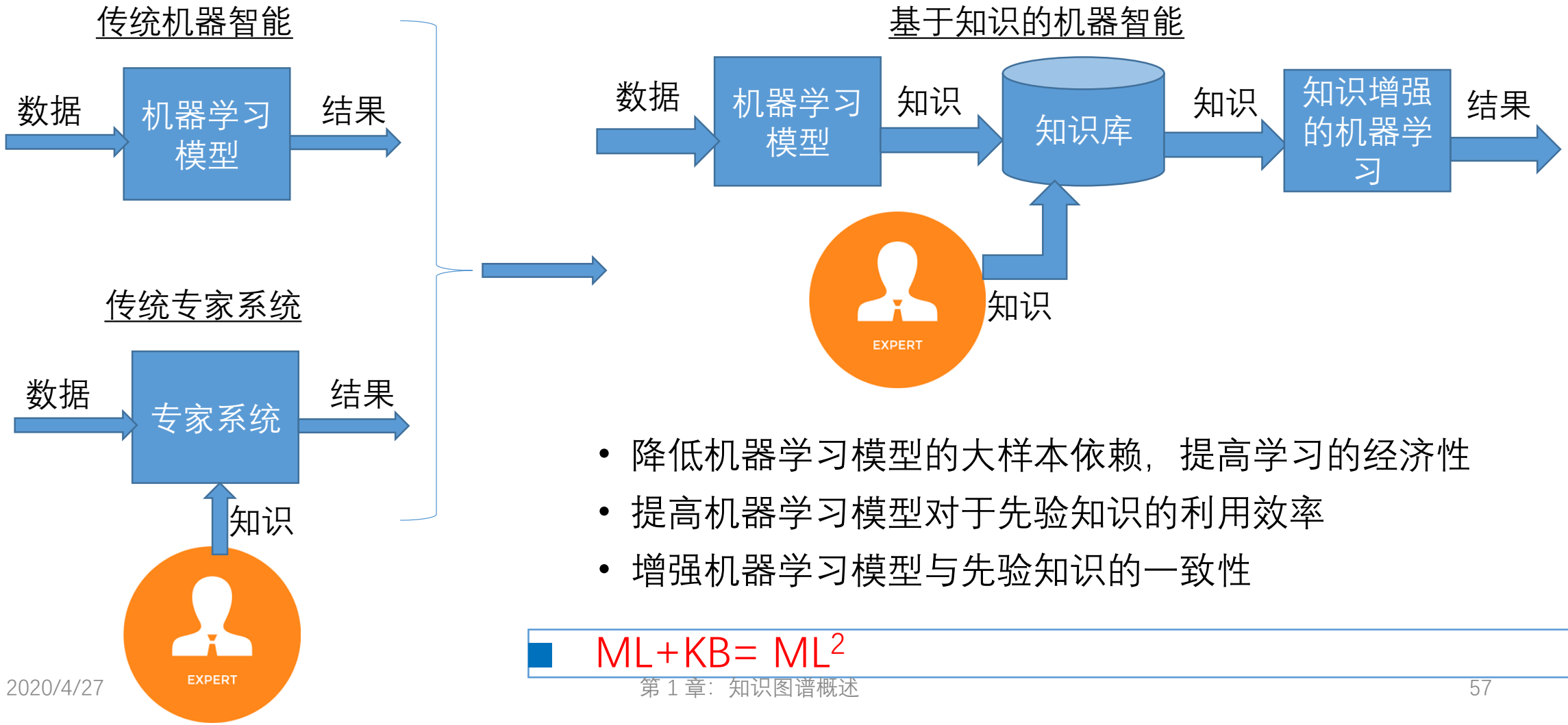
raw

日式 -1	
0	日式 1.0
1	剑林 0.9090448594528984
2	山田烧 0.9076092872105608
3	摩登主妇 0.9041964983257302
4	lototo 0.903521871998948
5	朵颐 0.9018902344911408
6	川岛屋 0.8992372679673781
7	一人食 0.8990165065859497
8	手绘碗 0.8966848604326612
9	二人食 0.8946378874188308

new

日式 -1	
0	日式 1.0
1	纯白色 0.9288789768835618
2	韩式 0.9282034998043911
3	法式 0.927806029695622
4	风格 0.9275196253763414
5	田园 0.9249904565619058
6	禅意 0.9235103898321229
7	素雅 0.9199717204810847
8	欧式 0.9188282247364457
9	田园风 0.9182616342150595

知识将显著增强机器学习能力



- 降低机器学习模型的大样本依赖，提高学习的经济性
- 提高机器学习模型对于先验知识的利用效率
- 增强机器学习模型与先验知识的一致性

$ML+KB= ML^2$

知识将成为比数据更为重要的资产

- 大数据时代是得“数据者” 得天下
- 人工智能时代是得“知识者”得天下

- 数据是石油，知识就是石油的萃取物

知识加工与石油萃取

“Knowledge is power in AI”，　Edward Feigenbaum

知识图谱的应用价值

知识图谱应用

- 认知智能应用需求广泛多样，需要对传统信息化手段的全面而彻底的革新
- 认知智能：人类脑力解放，机器生产力显著提高