自然语言处理

Natural Language Processing

权小军 教授

中山大学数据科学与计算机学院

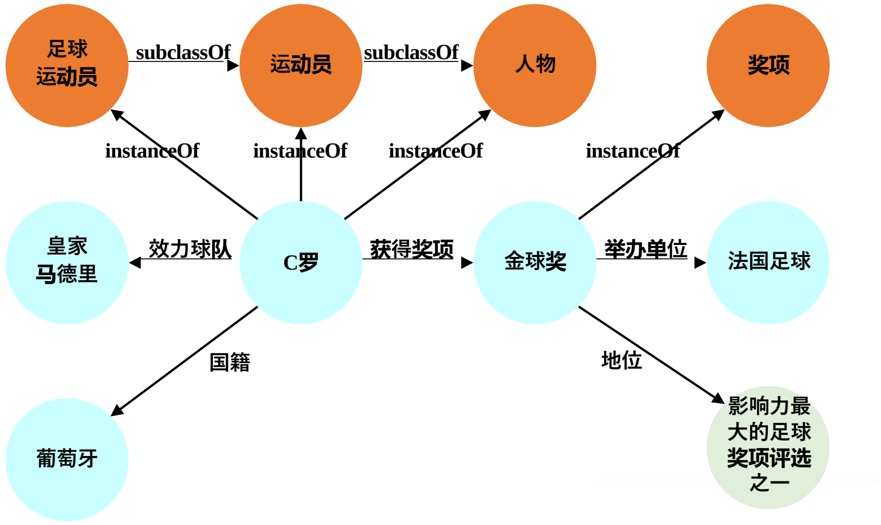
quanxj3@mail.sysu.edu.cn

Lecture 16：知识图谱（下）

课程回顾

知识图谱

- ❑ 知识图谱(Knowledge Graph)以结构化的方式描述客观世界中概念、实体及其之间的关系;
- ❑ 本质上是一种大规模语义网络(semantic network)



知识图谱

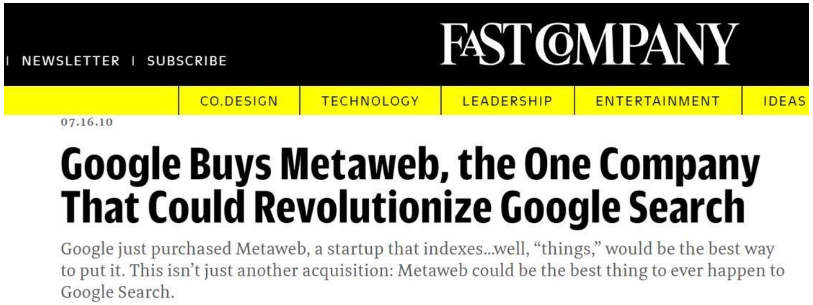
- □ 知识图谱通过对错综复杂的数据进行有效的加工、处理、整合，转化为简单、清晰的“实体-关系-实体”的三元组，最后聚合大量知识;

实体	关系	实体

实体	属性	值

诞生标志

- 2012年5月，Google收购Metaweb公司，并正式发布知识图谱



| NEWSLETTER | SUBSCRIBE

FAST COMPANY

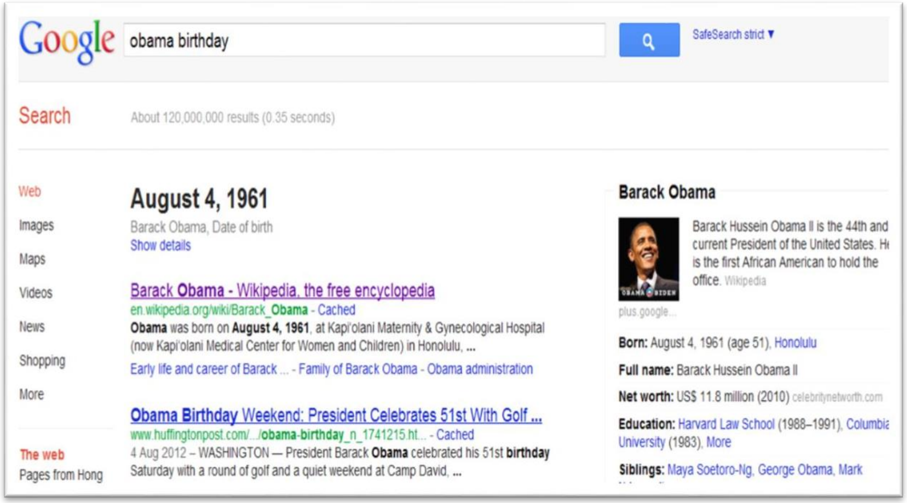
CO.DESIGN TECHNOLOGY LEADERSHIP ENTERTAINMENT IDEAS

07.16.10

Google Buys Metaweb, the One Company That Could Revolutionize Google Search

Google just purchased Metaweb, a startup that indexes...well, "things," would be the best way to put it. This isn't just another acquisition: Metaweb could be the best thing to ever happen to Google Search.

- 搜索核心需求：让搜索通往答案
  - 无法理解搜索关键词
  - 无法精准回答
- 根本问题
  - 缺乏大规模背景知识
  - 传统知识表示难以满足需求



About 120,000,000 results (0.35 seconds)

August 4, 1961

Barack Obama, Date of birth

Show details

Barack **Obama** - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Barack_**Obama** - Cached

Obama was born on **August 4, 1961**, at Kapi'olani Maternity & Gynecological Hospital (now Kapi'olani Medical Center for Women and Children) in Honolulu, ...

Early life and career of Barack ... - Family of Barack Obama - Obama administration

Obama Birthday Weekend: President Celebrates 51st With Golf ...

www.huffingtonpost.com/.../**obama-birthday**_n_1741215.ht... - Cached

4 Aug 2012 – WASHINGTON — President Barack **Obama** celebrated his 51st **birthday** Saturday with a round of golf and a quiet weekend at Camp David, ...

Barack Obama

Barack Hussein Obama II is the 44th and current President of the United States. H is the first African American to hold the office. Wikipedia

plus.google...

Born: August 4, 1961 (age 51), Honolulu

Full name: Barack Hussein Obama II

Net worth: US$ 11.8 million (2010) celebritynetworth.com

Education: Harvard Law School (1988–1991), Columbi University (1983), More

Siblings: Maya Soetoro-Ng, George Obama, Mark

KG优势1： large scale

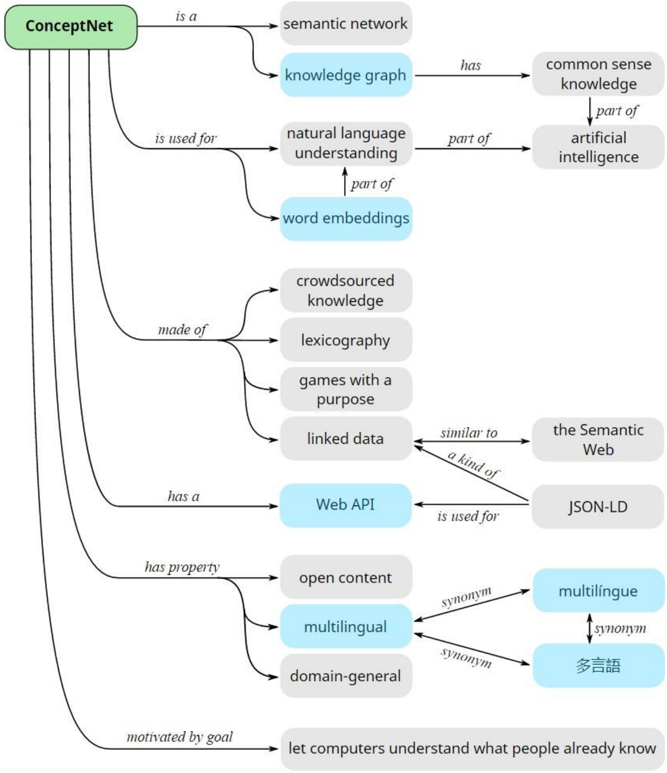
- ❑ Higher coverage over entities and concepts

KGs	# of Entities/Concepts	# of Relations
YAGO	10 Million	120 Million
DBpedia	28 Million	9.5 **Billion**
Probase	2.7 Million	70 **Billion**
BabelNet	14 Million	**5 Billion**
CN-DBpedia	17 Million	200 Million

KG优势2: semantically rich

❑ Higher coverage over numerous semantic relationships

KGs	# of Relations
DBpedia	1,650
YAGO1	14
YAGO3	74
CN-DBpedia	100 Thousands

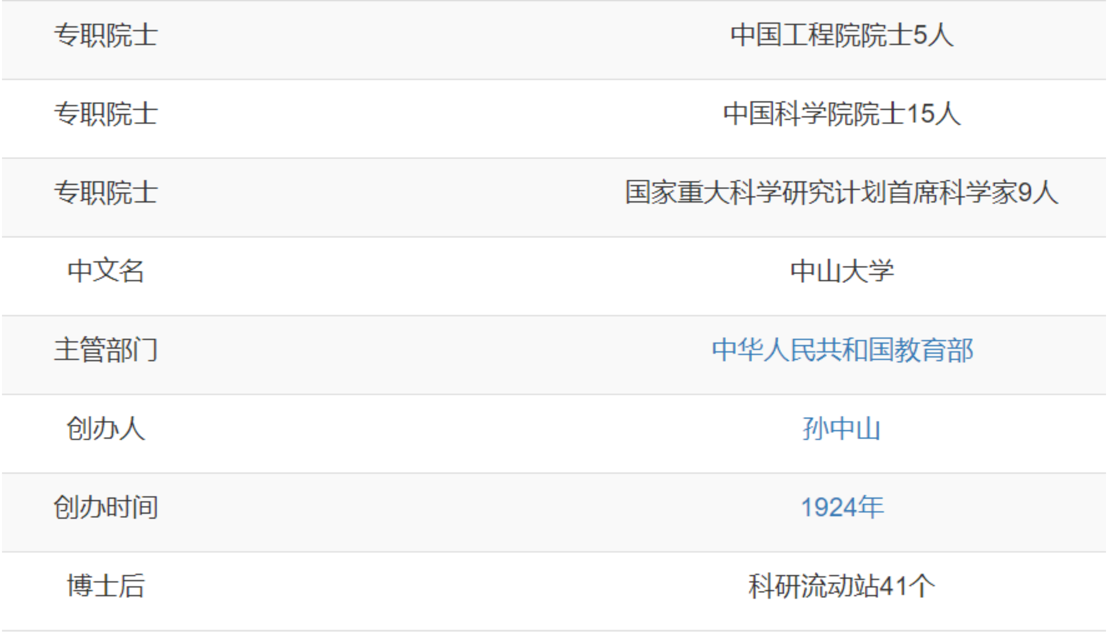


ConceptNet
is a
semantic network
knowledge graph
has
common sense knowledge
part of
artificial intelligence
is used for
natural language understanding
part of
part of
word embeddings
made of
crowdsourced knowledge
lexicography
games with a purpose
linked data
similar to
the Semantic Web
a kind of
JSON-LD
has a
Web API
is used for
has property
open content
multilingual
domain-general
synonym
multilíngue
synonym
synonym
多言語
motivated by goal
let computers understand what people already know

KG优势3：high quality

❑ High quality

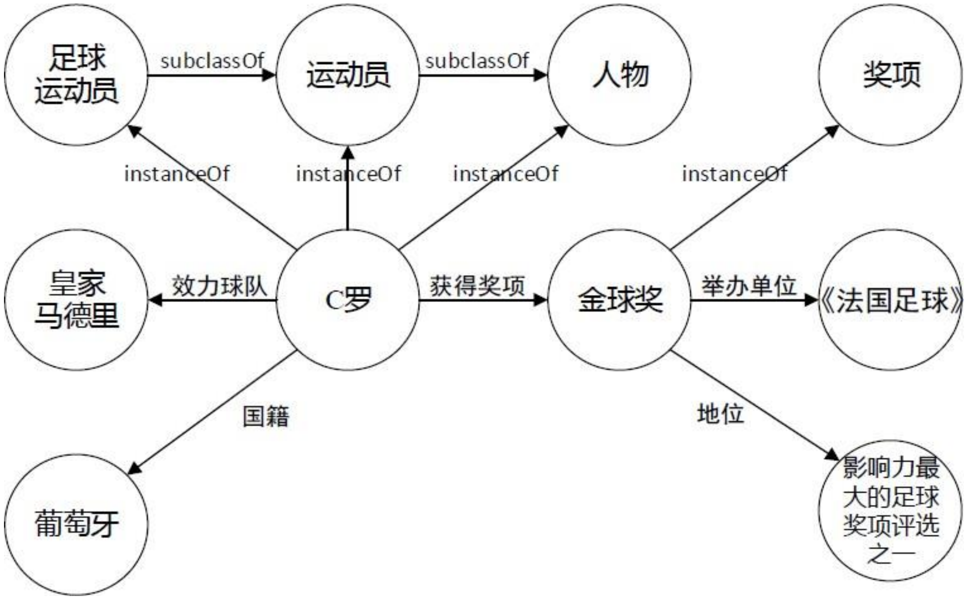
- Big data: Cross validation by multiple sources
- Crowd sourcing: quality guarantee



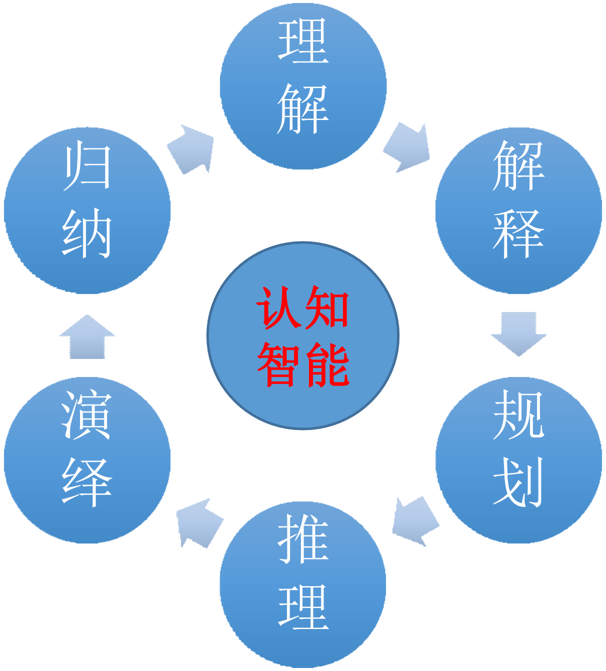
专职院士	中国工程院院士5人
专职院士	中国科学院院士15人
专职院士	国家重大科学研究计划首席科学家9人
中文名	中山大学
主管部门	中华人民共和国教育部
创办人	孙中山
创办时间	1924年
博士后	科研流动站41个

KG优势4: friendly structure

- Structured organization
  - By RDF
  - By graph



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

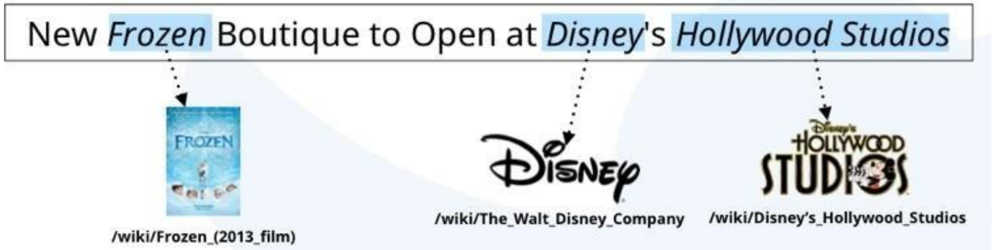
机器语言理解需要背景知识

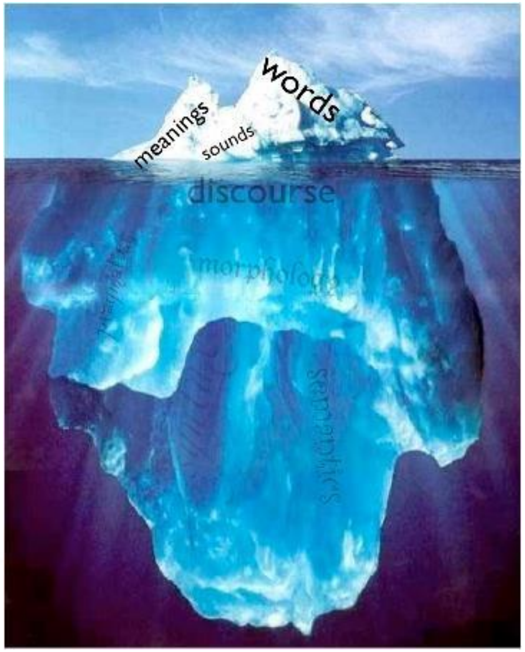
❑ Language is complicated

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

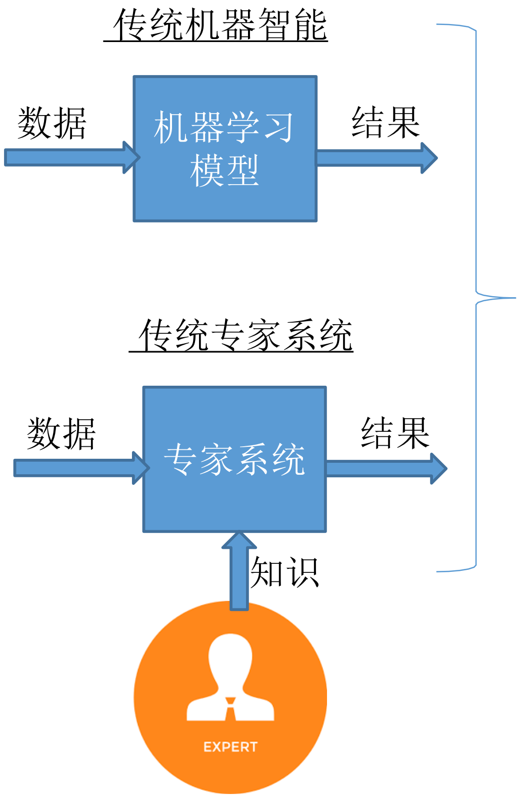
❑ Language understanding is difficult

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


New Frozen Boutique to Open at Disney's Hollywood Studios
FROZEN
/wiki/Frozen_(2013_film)
Disney
/wiki/The_Walt_Disney_Company
Disney's HOLLYWOOD STUDIOS
/wiki/Disney's_Hollywood_Studios



知识增强机器学习能力



传统机器智能

数据

机器学习模型

结果

传统专家系统

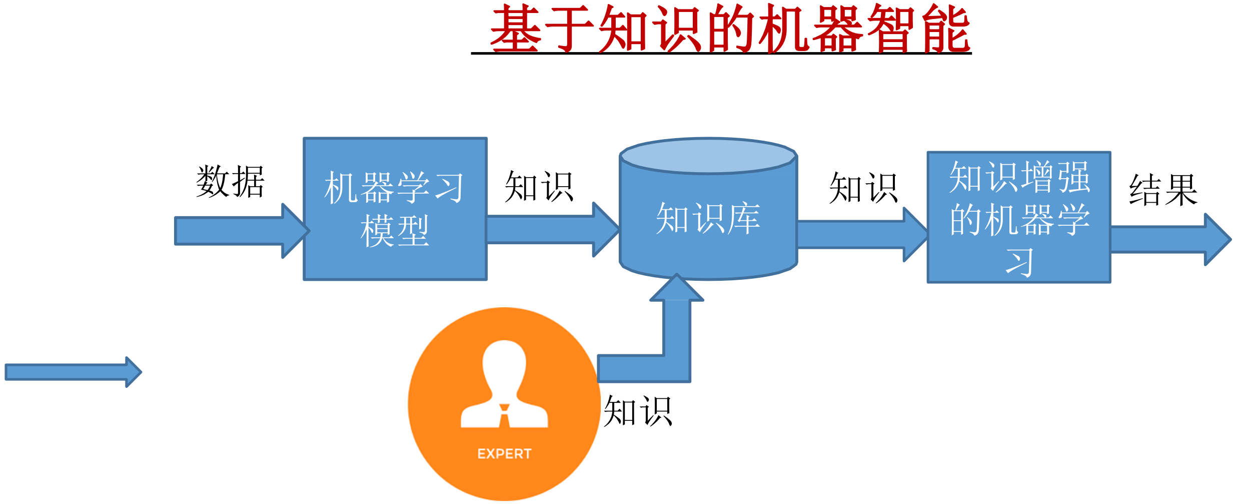
数据

专家系统

结果

知识

EXPERT



基于知识的机器智能

数据

机器学习模型

知识

知识库

知识

知识增强的机器学习

结果

知识

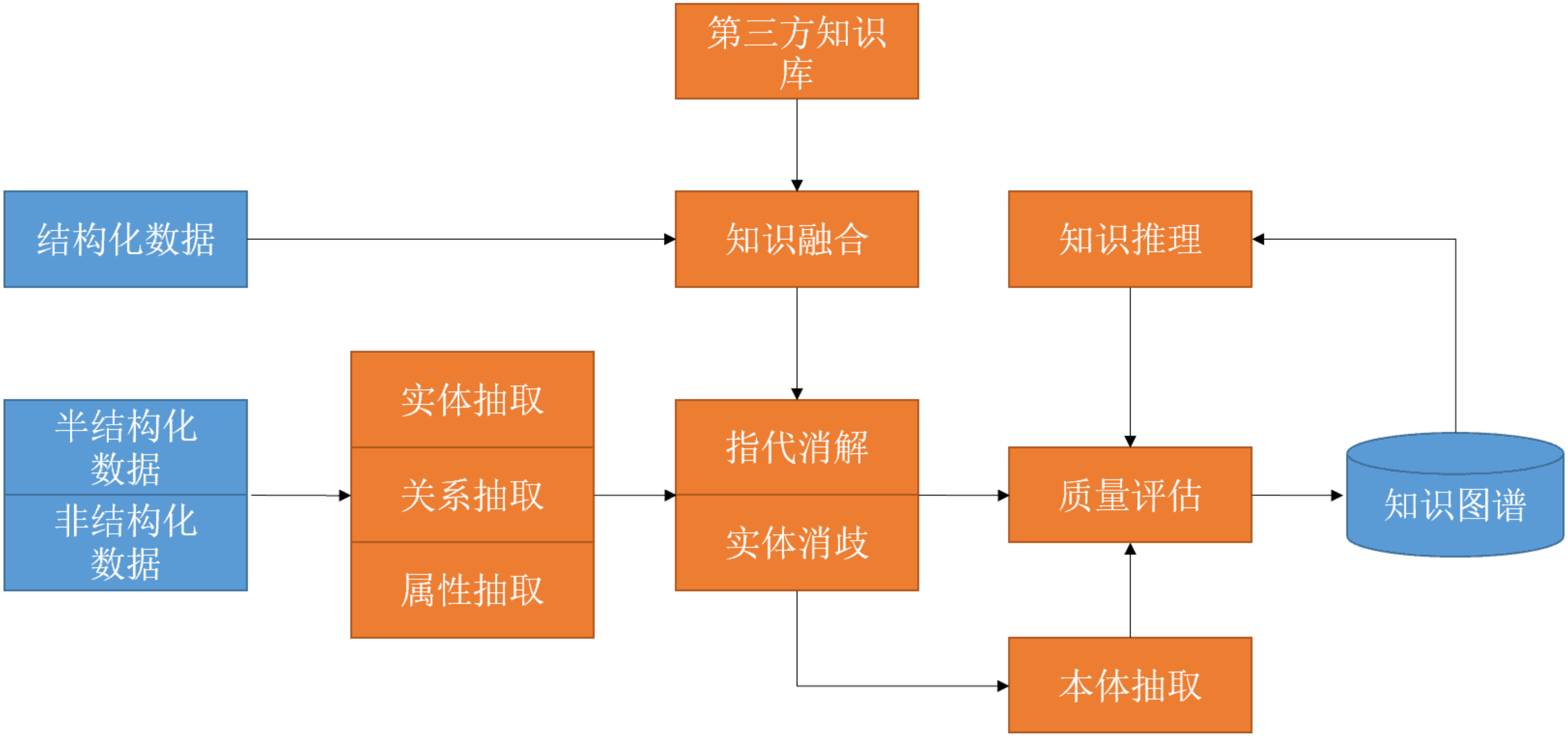
EXPERT

- 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

Lecture 16.1：知识图谱的构建

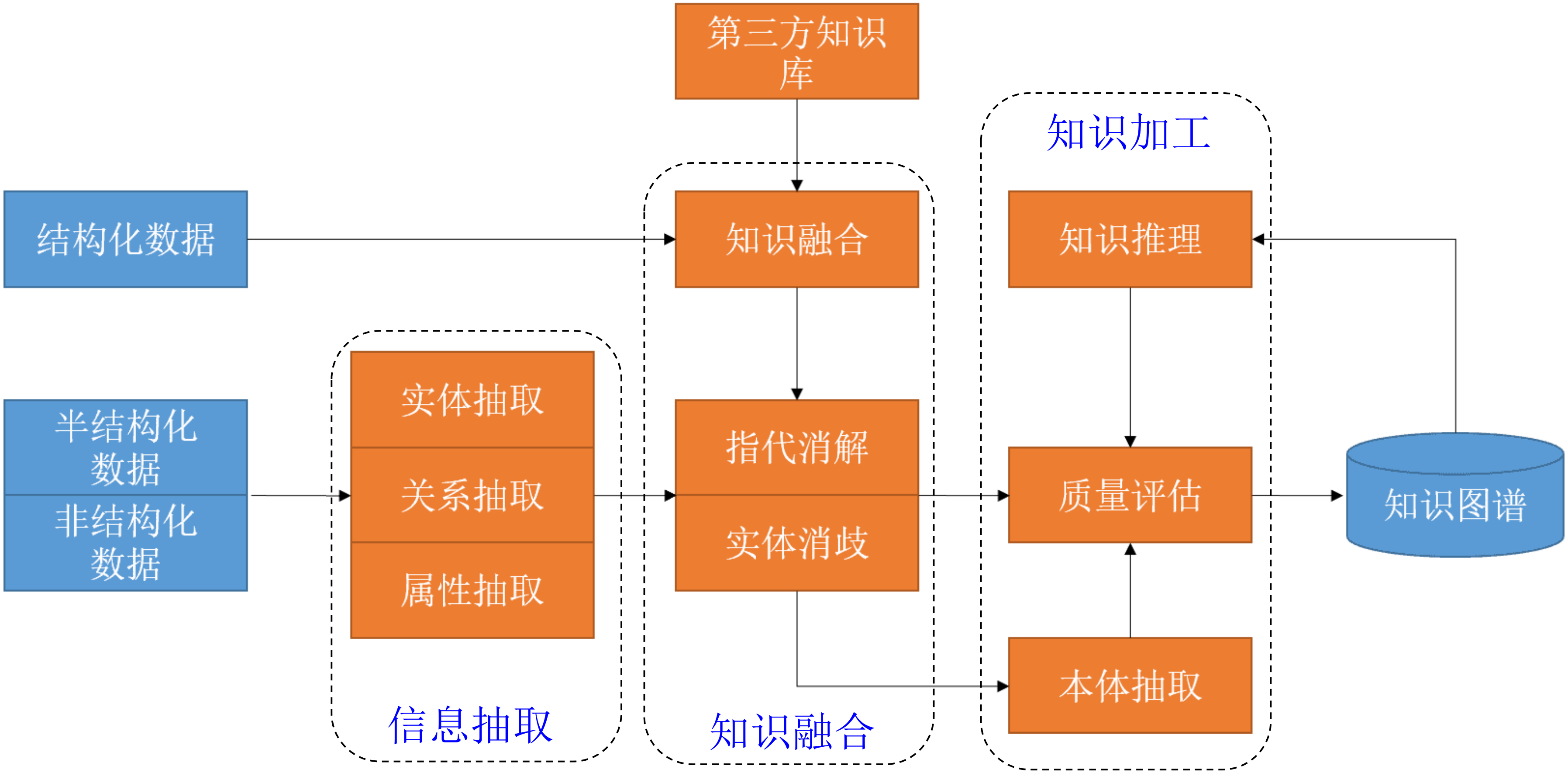
知识图谱的构建

知识图谱的技术架构：



知识图谱的构建

知识图谱的技术架构：



知识图谱的构建

知识图谱的技术架构：

- **信息抽取**：从各种类型的数据源中提取出实体、属性以及实体间的相互关系，在此基础上形成本体化的知识表达；
- **知识融合**：在获得新知识之后，需要对其进行整合，以消除矛盾和歧义，比如某些实体可能有多种表达，某个特定称谓也许对应于多个不同的实体等；
- **知识加工**：对于经过融合的新知识，需要经过质量评估之后（部分需要人工参与甄别），才能将合格的部分加入到知识库中，以确保知识库的质量。

知识图谱的构建：信息抽取

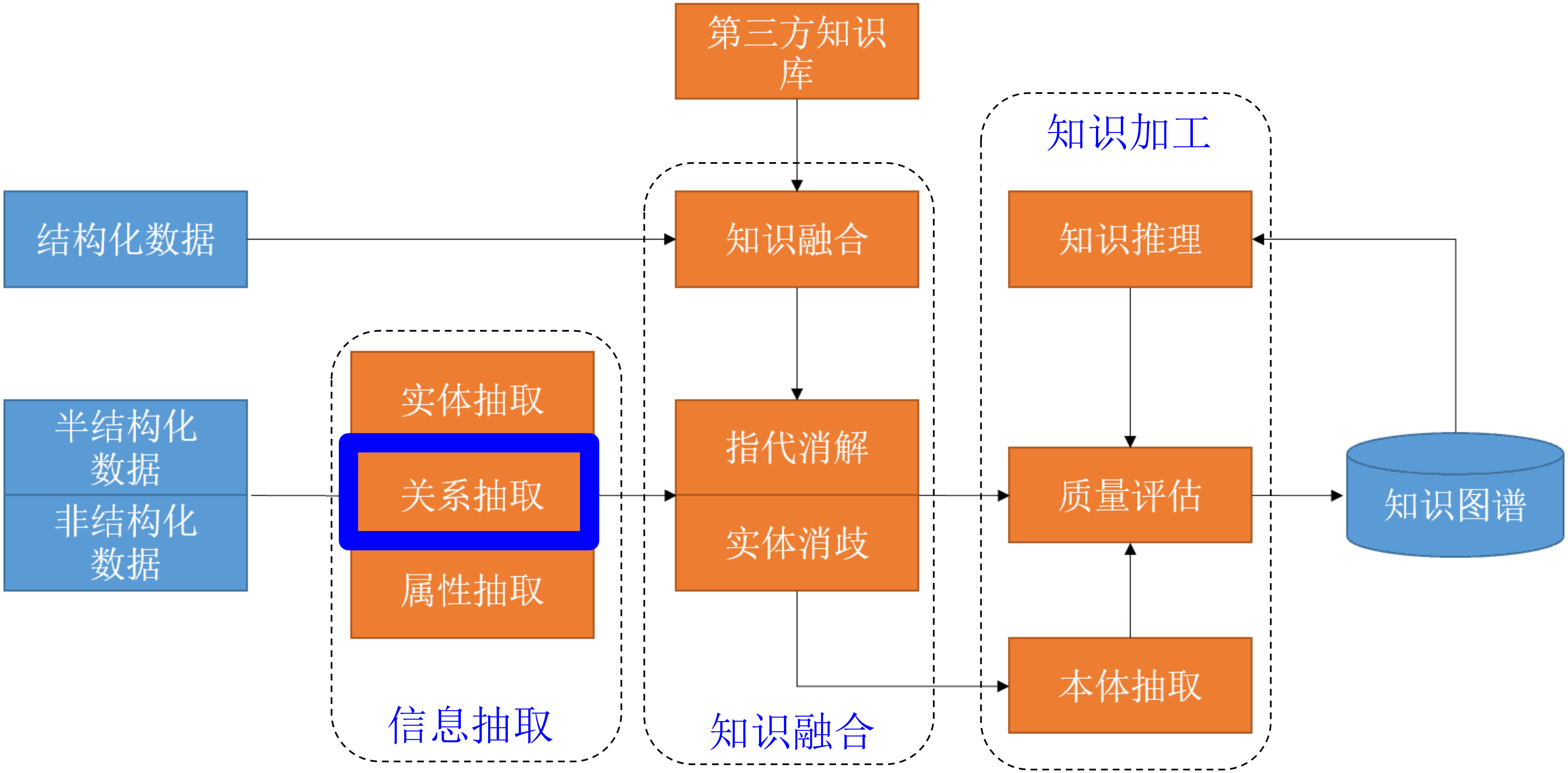
- ❑ 信息抽取（information extraction）是一种自动化地从半结构化和无结构数据中抽取实体、关系以及实体属性等结构化信息的技术
- ❑ 信息抽取是知识图谱构建的第一步，关键问题是：如何从异构数据源中自动抽取信息得到候选单元
- ❑ 涉及的关键技术包括：实体抽取、关系抽取和属性抽取

信息抽取：实体抽取

- ❑ 实体抽取又称为命名实体识别（named entity recognition，NER），是指从文本数据集中自动识别出命名实体
- ❑ 实体抽取的质量（准确率和召回率）对后续的知识获取效率和质量影响极大，因此是信息抽取中最为基础和关键的部分

知识图谱的构建

知识图谱的技术架构：



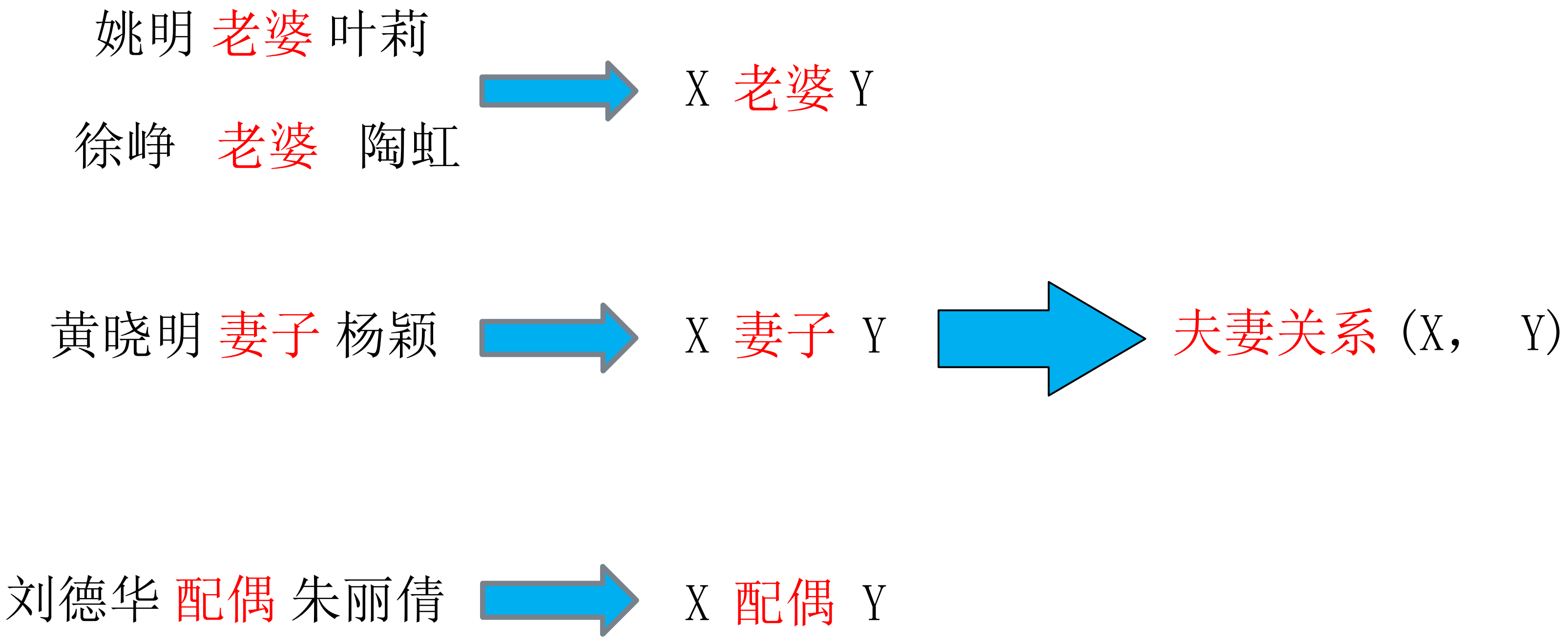
信息抽取：关系抽取

- 文本语料经过实体抽取，得到的是离散的命名实体，为了得到语义信息，还需要从相关的语料中提取出实体之间的关联关系，通过关联关系将实体（概念）联系起来，形成网状的知识结构，研究关系抽取技术的目的，就是解决如何从文本语料中抽取实体间的关系这一基本问题

信息抽取：关系抽取

- □ 基于模板的方法
  - ➢ 基于触发词的Pattern
  - ➢ 基于依存句法分析的Pattern
- □ 监督学习方法
  - ➢ 机器学习方法
  - ➢ 深度学习方法
- □ 弱监督学习方法
  - ➢ 远程监督
  - ➢ Bootstrapping

基于模板的方法—基于触发词的Pattern



姚明 老婆 叶莉

徐峥　老婆　陶虹

X 老婆 Y

黄晓明 妻子 杨颖

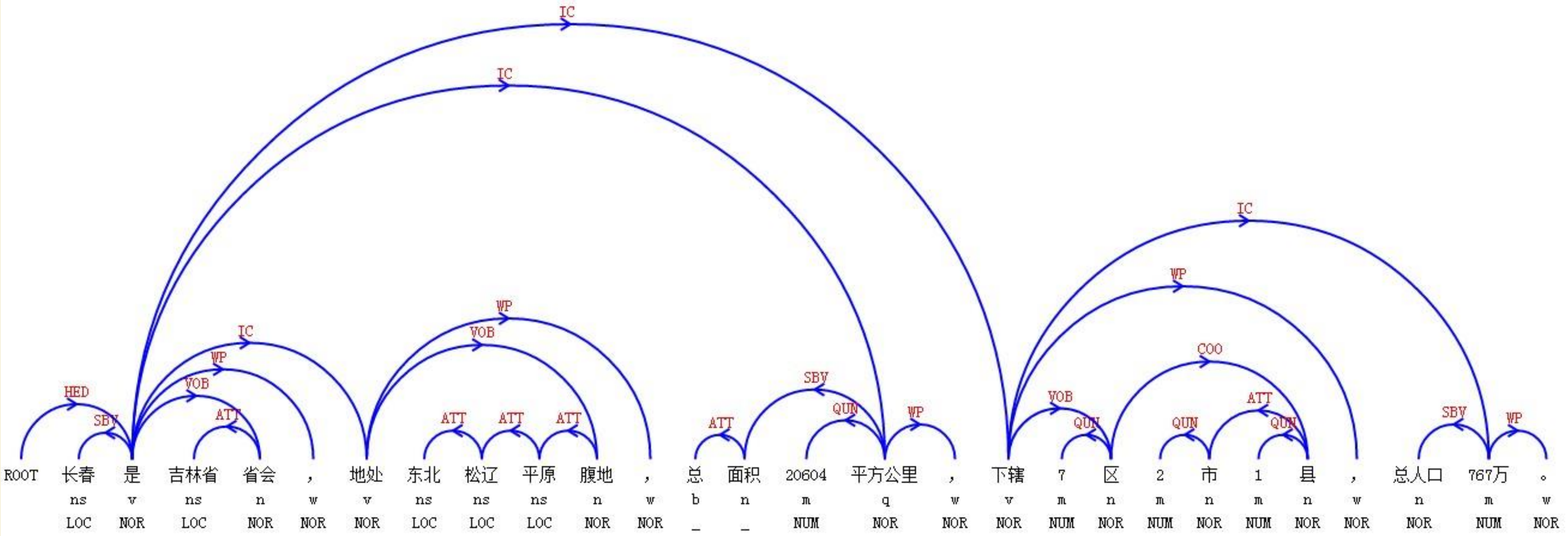
X 妻子 Y

夫妻关系（X，　Y）

刘德华 配偶 朱丽倩

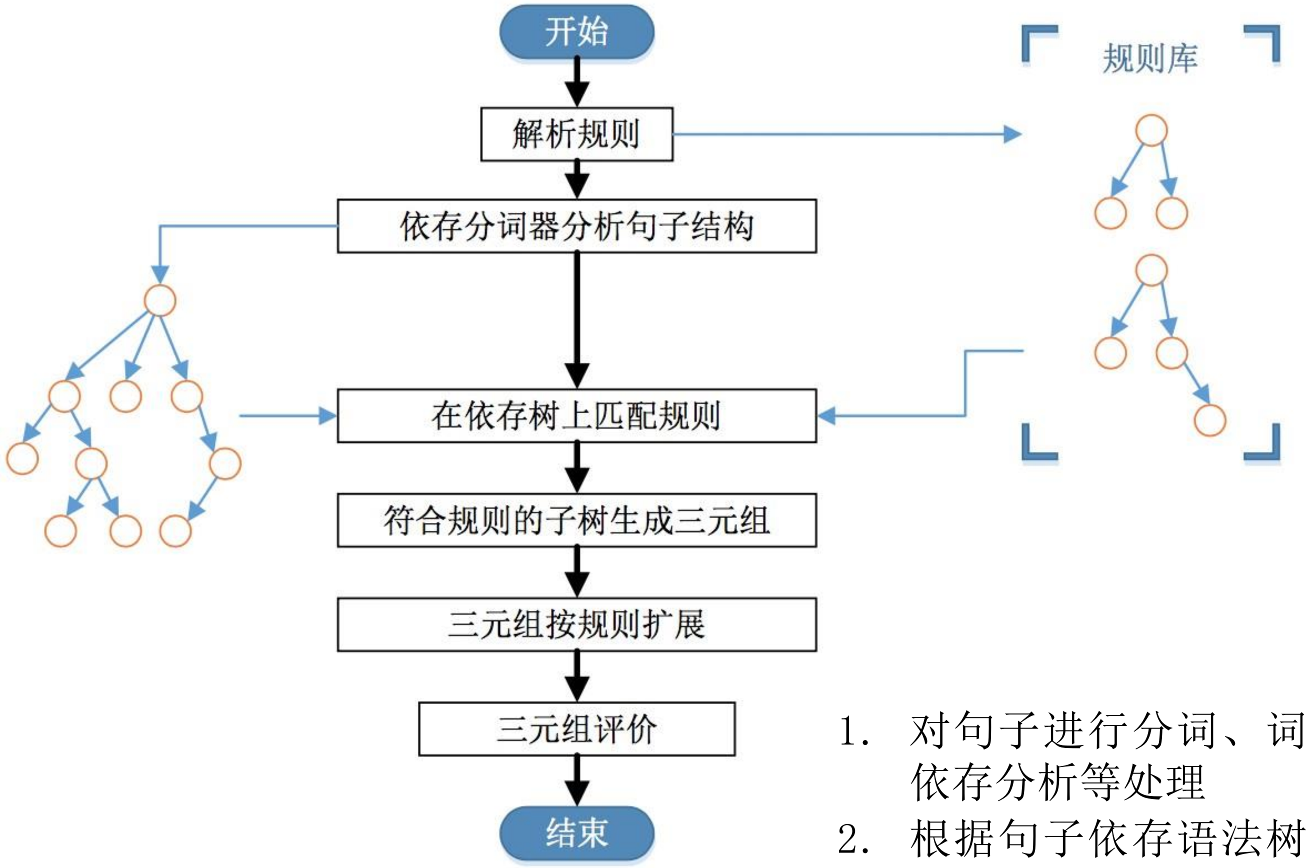
X 配偶 Y

基于模板的方法—基于依存句法分析的Pattern



- 依存句法分析句子的句法结构
- 以动词为起点，构建规则，对节点上的词性和边上的依存关系进行限定

基于模板的方法—基于依存句法分析的Pattern



1. 对句子进行分词、词性标注、命名实体识别、依存分析等处理
2. 根据句子依存语法树结构上匹配规则，每匹配一条规则就生成一个三元组
3. 根据扩展规则对抽取到的三元组进行扩展
4. 对三元组实体和触发词进一步处理抽取出关系

基于模板的方法—基于依存句法分析的Pattern

董卿现身国家博物馆看展优雅端庄大方

依存分析结果

词顺序	词	词性	依存关系
0	董卿	人名	定语
1	现身	动词	核心词
2	国家博物馆	地名	宾语
3	看	动词	顺承
4	展	动词	补语
5	优雅	形容词	定语
6	端庄	形容词	定语
7	大方	形容词	宾语

规则抽取结果

（董卿，现身，国家博物馆） ➡ 位于（董卿，国家博物馆）

基于模板的方法—优劣

- 优点
  - 在小规模数据集上容易实现
  - 构建简单
- 缺点
  - 特定领域的模板需要专家构建
  - 难以维护
  - 可移植性差
  - 规则集合小的时候，召回率很低

监督学习

确定实体对的情况下，根据句子上下文对实体关系进行预测，构建一个监督学习应该怎么做？

- 预先定义好关系的类别
- 人工标注一些数据
- 设计特征表示
- 选择一个分类方法（SVM、NN、Naive Bayes）
- 评估结果

监督学习-特征

- 轻量级特征
  - 实体前后的词
  - 实体的类型
  - 实体之间的距离
- 中等量级特征
  - Chunk序列
- 重量级特征
  - 实体间的依存关系路径
  - 实体间树结构的距离
  - 特定的结构信息

监督学习—深度学习方法

❑ Pipeline

- 识别实体和关系分类是完全分离的两个过程，不会相互影响，关系的识别依赖于实体识别的效果

❑ Joint Model

- 实体识别和关系分类的过程是共同优化的

监督学习—优劣

- 优点
  - 准确率高，标注数据越多越准确

- 缺点
  - 标注数据成本太高
  - 不能扩展新的关系

半监督学习

- 没有足够多标注数据的情况下，怎么办?
- 数据量特别大的情况下，如何抽取实体间关系?

➢ 远程监督方法

知识库与非结构化文本对齐来自动构建大量训练数据，减少模型对人工标注数据的依赖，增强模型跨领域适应能力

➢ Bootstrapping

通过在文本中匹配实体对和表达关系短语模式，寻找和发现新的潜在关系三元组

半监督学习—远程监督

- 两个实体如果在知识库中存在某种关系，则包含该两个实体的非结构化句子均能表示出这种关系。

  在某知识库中存在： 创始人（乔布斯，苹果公司）

  则可构建训练正例：乔布斯是苹果公司的联合创始人和CEO

- 具体步骤

  1. 从知识库中抽取存在关系的实体对
  2. 从非结构化文本中抽取含有实体对的句子作为训练样例

远程监督—优劣

- 优点
  - 可以利用丰富的知识库信息，减少一定的人工标注

- 缺点
  - 假设过于肯定，引入大量噪声，存在语义漂移现象
  - 很难发现新的关系

半监督学习—Bootstrapping

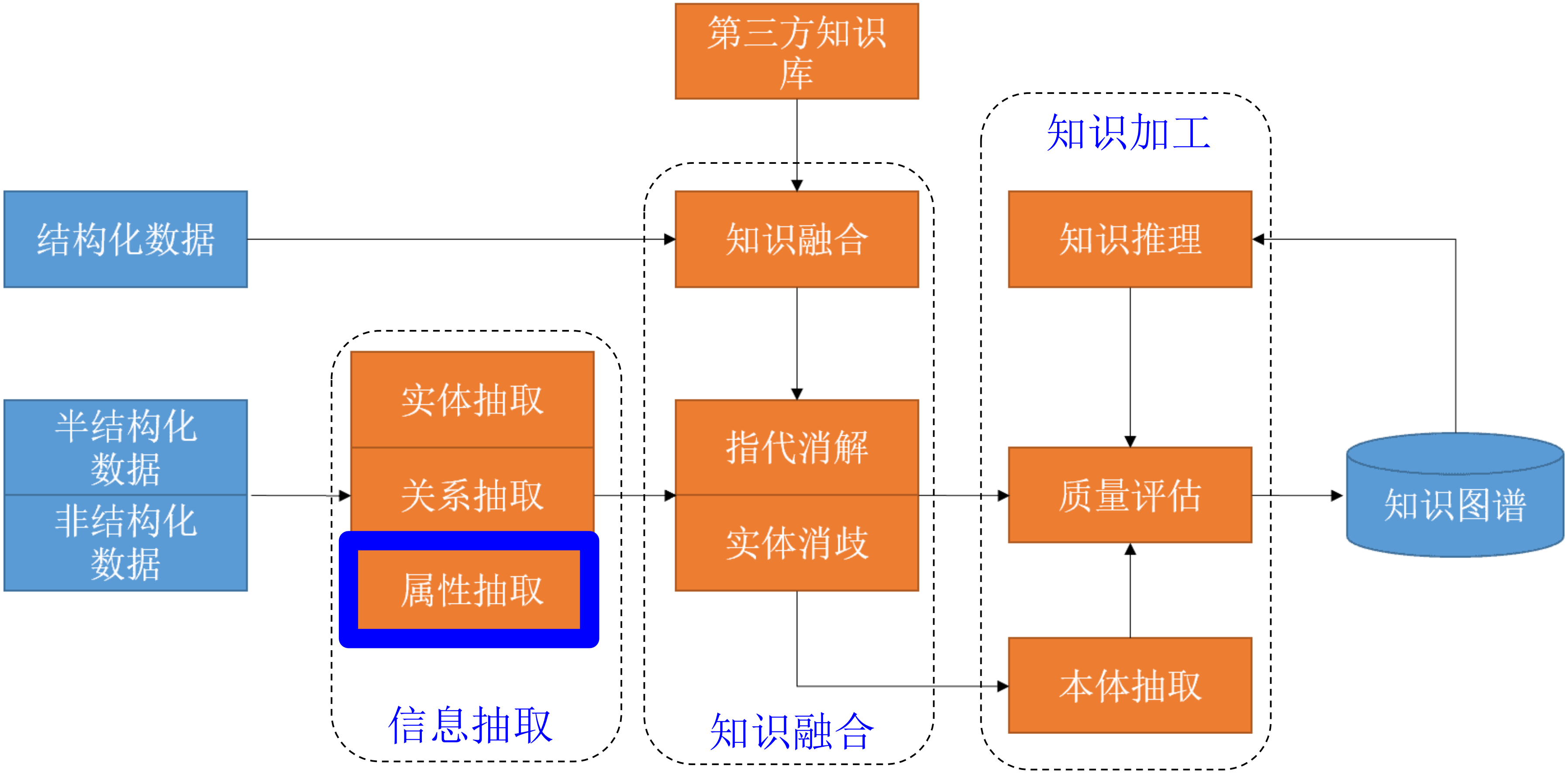
- 给定种子集合，如：<姚明，叶莉>
  1. 从文档中抽取出包含种子实体的新闻，如
     - 姚明 老婆 叶莉 简历身高曝光
       X 老婆 Y 简历身高曝光
     - 姚明 与妻子 叶莉 外出赴约
       X 与妻子 Y 外出赴约
     - 姚明 携爱妻 叶莉 外出赴约
       X 携爱妻 Y 外出赴约
  2. 将抽取出的Pattern去文档集中匹配
     - 小猪 与妻子 伊万 外出赴约
  3. 根据Pattern抽取出的新关系加入种子库，迭代多轮直到不符合条件

Bootstrapping—优劣

- □ 优点
  - ○ 构建成本低，适合大规模构建
  - ○ 可以发现新的关系 (隐含的)
- □ 缺点
  - ○ 对初始给定的种子集敏感
  - ○ 存在语义漂移问题
  - ○ 结果准确率较低
  - ○ 缺乏对每一个结果的置信度的计算

知识图谱的构建

知识图谱的技术架构：



信息抽取：属性抽取

- ❑ 属性抽取(Attribute Extraction)的目标是从不同信息源中采集特定实体的属性信息
- ❑ 例如针对某个公众人物，可以从网络公开信息中得到其昵称、生日、国籍、教育背景等信息

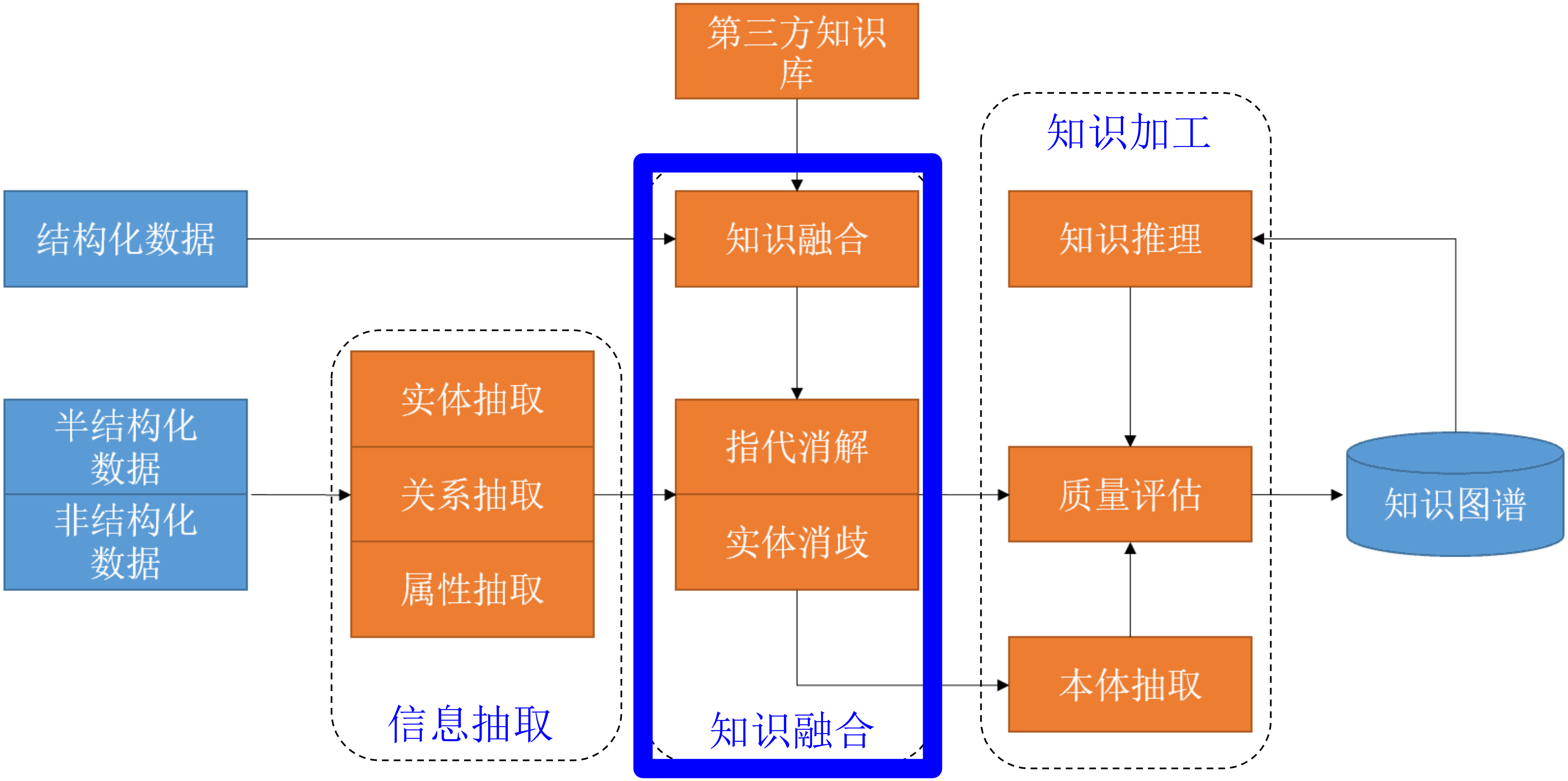
信息抽取：属性抽取

❑ 属性抽取(Attribute Extraction):

1. 将实体的属性视作实体与属性值之间的一种名词性关系，将属性抽取任务转化为关系抽取任务
2. 基于规则和启发式算法，抽取结构化数据
3. 基于半结构化数据，通过自动抽取生成训练语料，用于训练实体属性标注模型，然后将其应用于对非结构化数据的实体属性抽取
4. 采用数据挖掘的方法直接从文本中挖掘实体属性和属性值之间的关系模式，据此实现对属性名和属性值在文本中的定位。

知识图谱的构建

知识图谱的技术架构：



知识图谱的构建：知识融合

- ❑ 通过信息抽取，我们就从原始的非结构化和半结构化数据中获取到了实体、关系以及实体的属性信息
- ❑ 如果我们将接下来的过程比喻成拼图的话，那么这些信息就是拼图碎片，散乱无章，甚至还有从其他拼图里跑来的碎片、本身就是用来干扰我们拼图的错误碎片
- ❑ 拼图碎片（信息）之间的关系是扁平化的，缺乏层次性和逻辑性；拼图（知识）中还存在大量冗杂和错误的拼图碎片

知识图谱的构建：知识融合

❑ 知识融合包括：实体链接和知识合并

- ○ **实体链接**（entity linking）：是指对于从文本中抽取得到的实体对象，将其链接到知识库中对应的正确实体对象的操作
- ○ 其基本思想是首先根据给定的实体指称项，从知识库中选出一组候选实体对象，然后通过**相似度计算**将指称项链接到正确的实体对象

知识图谱的构建：知识融合

❑ 实体链接的流程：

1. 从文本中通过实体抽取得到实体指称项
2. 进行**实体消歧**和**共指消解**，判断知识库中的同名实体与之是否代表不同的含义，以及是否有其他实体与之表示相同的含义
3. 在确认知识库中对应的正确实体对象之后，将该实体指称项链接到知识库中对应实体
4. **实体消歧**：专门用于解决同名实体产生歧义问题的技术，通过实体消歧，就可以根据当前的语境，准确建立实体链接，实体消歧主要采用聚类法
5. **共指消解**：**主要用于解决多个指称对应同一实体对象的问题。**在一次会话中，多个指称可能指向的是同一实体对象。利用共指消解技术，可以将这些指称项关联到正确的实体对象

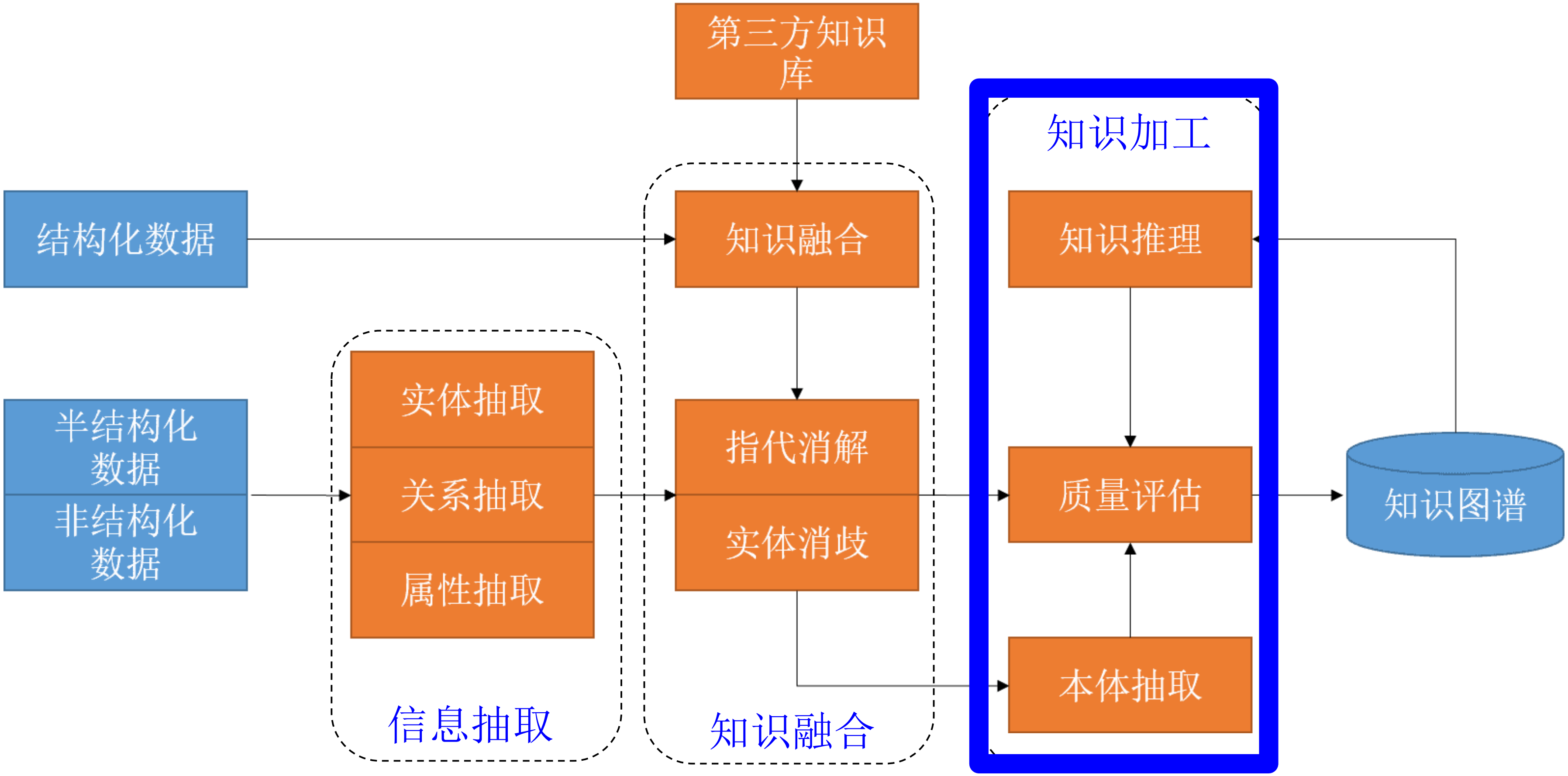
知识图谱的构建：知识融合

❑知识合并

- ○ 构建知识图谱时，可从第三方知识库或结构化数据获取输入
- ○ 将外部知识库融合到本地知识库需要处理两个层面的问题：
  - 数据层的融合，包括实体的指称、属性、关系以及所属类别等，主要的问题是如何避免实例以及关系的冲突问题，造成不必要的冗余
  - 通过模式层的融合，将新得到的本体融入已有的本体库中
- ○ 然后是合并关系数据库，在知识图谱构建过程中，一个重要的高质量知识来源是企业或者机构自己的关系数据库。为了将这些结构化的历史数据融入到知识图谱中，可以采用资源描述框架（RDF）作为数据模型，其实质就是将关系数据库的数据换成RDF的三元组数据

知识图谱的构建

知识图谱的技术架构：



知识图谱的构建：知识加工

- ❑ 通过信息抽取，从原始语料中提取出了实体、关系与属性等知识要素；经过知识融合，消除实体指称项与实体对象之间的歧义，得到一系列基本的事实表达
- ❑ 然而事实本身并不等于知识。要想最终获得结构化、网络化的知识体系，还需要经历知识加工的过程
- ❑ 知识加工主要包括：**本体构建**、**知识推理**和**质量评估**

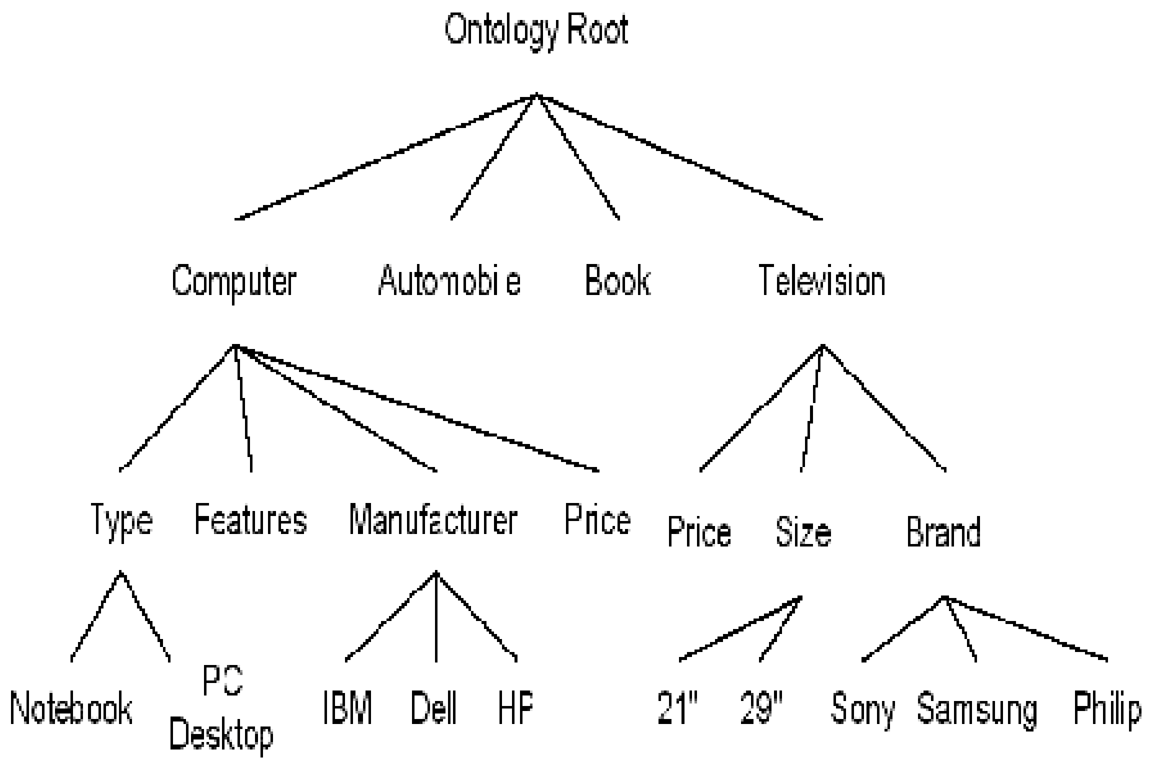
知识加工：本体构建

- ❑ 本体（ontology）是指人工的概念集合、概念框架，如“人”、“事”、“物”等

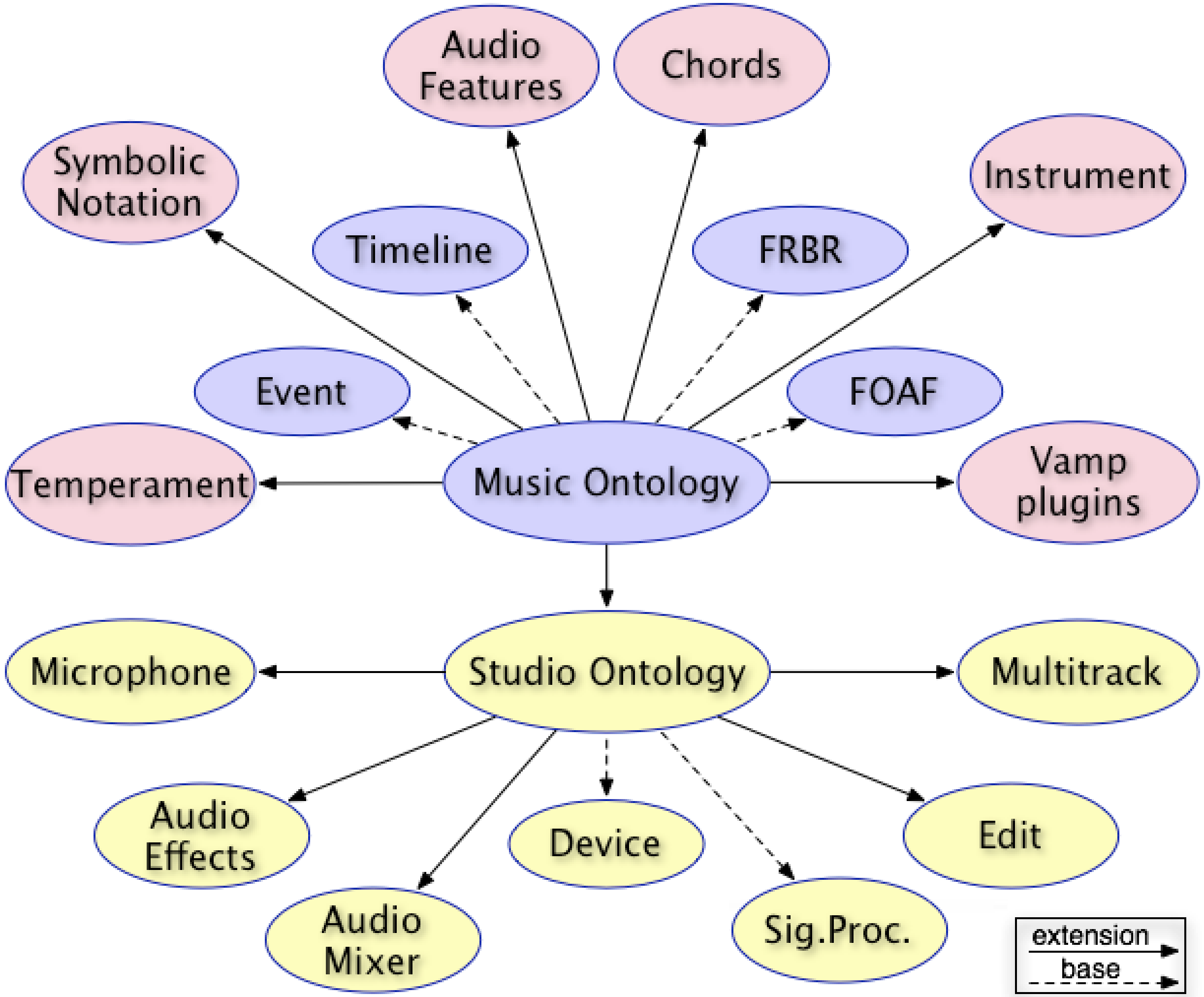
- ❑ 本体可以采用人工编辑的方式手动构建，也可以以数据驱动的自动化方式构建本体。因为人工方式工作量巨大，且很难找到符合要求的专家，因此当前主流的全局本体库产品，都是从一些面向特定领域的现有本体库出发，采用自动构建技术逐步扩展得到的

- ❑ 自动化本体构建过程包含三个阶段：
  - ○ 实体并列关系相似度计算
  - ○ 实体上下位关系抽取
  - ○ 本体的生成

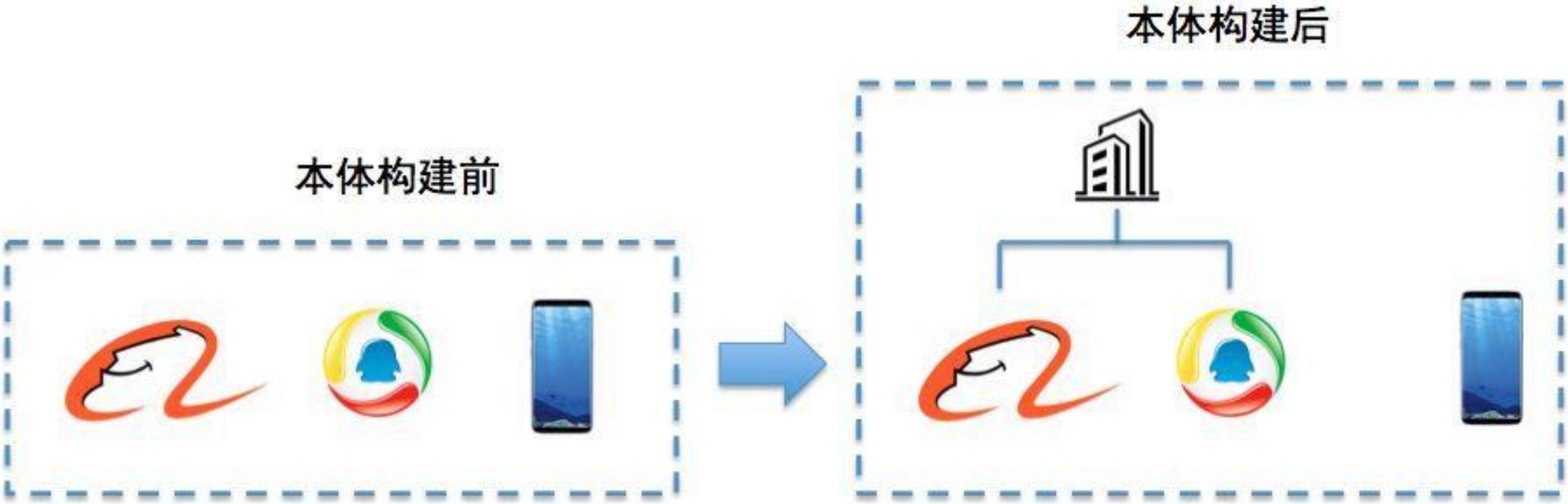
知识加工：本体构建


Ontology Root
Computer
Automobi e
Book
Television
Type
Features
Manufacturer
Price
Price
Size
Brand
Notebook
PC
Desktop
IBM
Dell
HP
21"
29"
Sony
Samsung
Philip

知识加工：本体构建

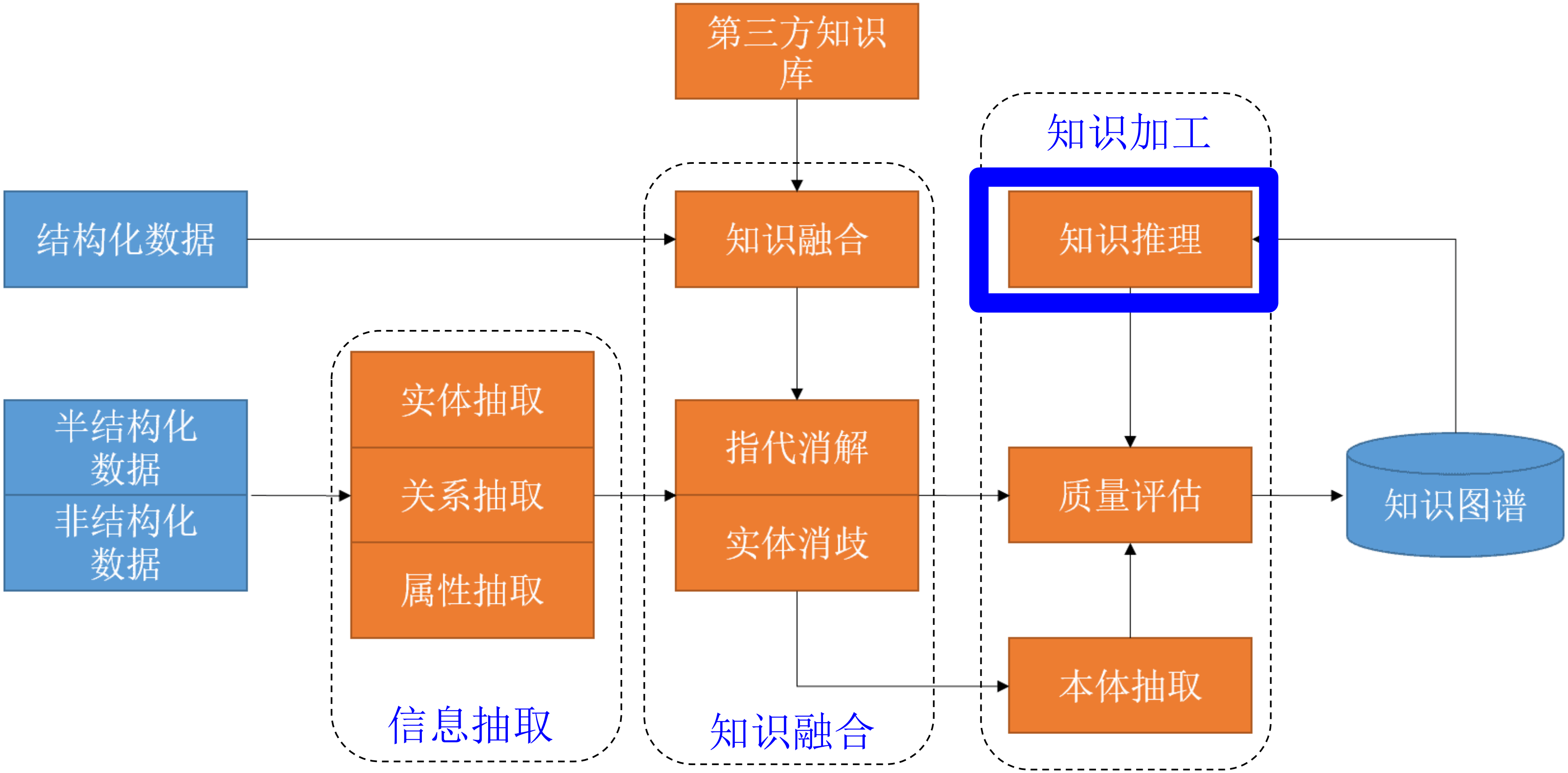

Audio Features
Chords
Symbolic Notation
Instrument
Timeline
FRBR
Event
FOAF
Temperament
Music Ontology
Vamp plugins
Microphone
Studio Ontology
Multitrack
Audio Effects
Device
Edit
Audio Mixer
Sig.Proc.
extension
base

知识加工：本体构建



知识图谱的构建

知识图谱的技术架构：



知识加工：知识推理

❑ 本体构建之后，一个知识图谱的雏形便已经搭建好了。但此时知识图谱之间的关系大多是残缺的，可以使用知识推理技术完成知识发现

❑ 我们可以发现：如果A是B的配偶，B是C的主席，C坐落于D，那么我们就可以认为，A生活在D这个城市

❑ 当然知识推理的对象也并不局限于实体间的关系，也可以是实体的属性值，本体的概念层次关系等

❑ 推理属性值：已知某实体的生日属性，可以通过推理得到该实体的年龄属性

❑ 推理概念：已知(老虎，科，猫科)和（猫科，目，食肉目）可以推出（老虎，目，食肉目）

知识加工：质量评估

- 质量评估也是知识库构建技术的重要组成部分，这一部分存在的意义在于：可以对知识的可信度进行量化，通过舍弃置信度较低的知识来保障知识库的质量

Lecture 16.2：知识图谱的管理

知识图谱的架构

知识图谱在逻辑上可分为**模式层**与**数据层**两个层次。

模式层：

- ○ **模式层**构建在数据层之上，是知识图谱的核心，通常采用本体库来管理知识图谱的模式层
- ○ 本体是结构化知识库的概念模板，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小

模式层：实体-关系-实体，实体-属性-值

知识图谱的架构

知识图谱在逻辑上可分为**模式层**与**数据层**两个层次。

数据层：

○ **数据层**主要是由一系列的事实组成，而知识将以事实为单位进行存储

数据层：比尔盖茨-妻子-梅琳达盖茨，比尔盖茨-总裁-微软

知识图谱管理

知识图谱的管理，主要是知识库的更新，包括**概念层的更新**和**数据层的更新**：

- ○ 概念层的更新是指新增数据后获得了新的概念，需要自动将新的概念添加到知识库的概念层中
- ○ 数据层的更新主要是新增或更新实体、关系、属性值，对数据层进行更新需要考虑数据源的可靠性、数据的一致性（是否存在矛盾或冗杂等问题）等可靠数据源，并选择在各数据源中出现频率高的事实和属性加入知识库

知识图谱管理

知识图谱的内容更新有两种方式:

- **全面更新**：指以更新后的全部数据为输入，从零开始构建知识图谱。这种方法比较简单，但资源消耗大，而且需要耗费大量人力资源进行系统维护
- **增量更新**：以当前新增数据为输入，向现有知识图谱中添加新增知识。这种方式资源消耗小，但目前仍需要大量人工干预（定义规则等），因此实施起来十分困难

Lecture 16.3：知识图谱与智能问答

问答系统历史

基于信息检索的问答

基于关键词匹配+信息抽取，基于浅层语义分析

基于社区的问答

依赖于网民贡献，问答过程依赖于关键词检索技术

基于知识库的问答

知识库
语义解析

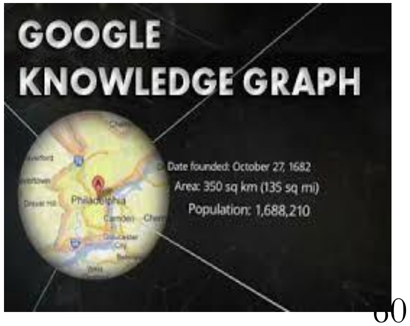
Text REtrieval Conference (TREC)
Overview
Publications
Other Evaluations
Information for Active Participants
Frequently Asked Questions
Tracks
Data
Past TREC Results
Contact Information

YAHOO! ANSWERS

知乎

与世界分享你的知识、经验和见解

WolframAlpha computational knowledge engine

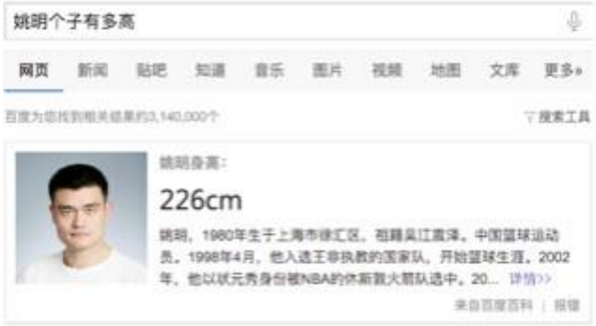


GOOGLE KNOWLEDGE GRAPH

Date founded: October 27, 1682
Area: 350 sq km (135 sq mi)
Population: 1,688,210

根据问答形式分类

- □ 一问一答
- □ 交互式问答
- □ 阅读理解



姚明个子有多高

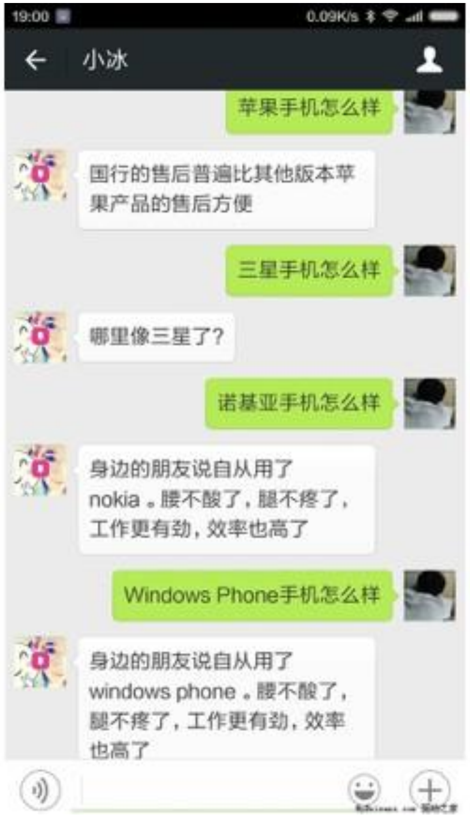
百度为您找到相关结果约3,140,000个

姚明身高：

226cm

姚明，1980年生于上海市徐汇区，祖籍吴江震泽，中国篮球运动员。1998年4月，他入选王非执教的国家队，开始篮球生涯。2002年，他以状元秀身份被NBA的休斯敦火箭队选中。20... 详情>>

来自百度百科 | 报错



苹果手机怎么样

国行的售后普遍比其他版本苹果产品的售后方便

三星手机怎么样

哪里像三星了？

诺基亚手机怎么样

身边的朋友说自从用了nokia。腰不酸了，腿不疼了，工作更有劲，效率也高了

Windows Phone手机怎么样

身边的朋友说自从用了windows phone。腰不酸了，腿不疼了，工作更有劲，效率也高了

Mary journeyed to the den.
Mary went back to the kitchen.
John journeyed to the bedroom.
Mary discarded the milk.
Q: Where was the milk before the den?
A. Hallway

Brian is a lion.
Julius is a lion.
Julius is white.
Bernhard is green.
Q: What color is Brian?
A. White

Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.
Q: Where is the apple?
A. Bedroom

基于知识图谱的问答

机器人及IOT设备的智能化：给万物都挂接一个背景知识库

对话式的信息获取更更加需要精准度和可靠度，知识图谱对于提升用户体验更更加不不可少。

IBM Watson

DBpedia　Yago

Wordnet

amazon alexa

True Knowledge/Evi

Siri

WolframAlpha

DBpedia

holoera

AI全息3D虚拟养成偶像

琥珀·虚颜

CHINAJOY

2016.7.28-7.31　上海新国际博览中心E2号馆 S109

zhishi.me

……

……

……

DUER OS

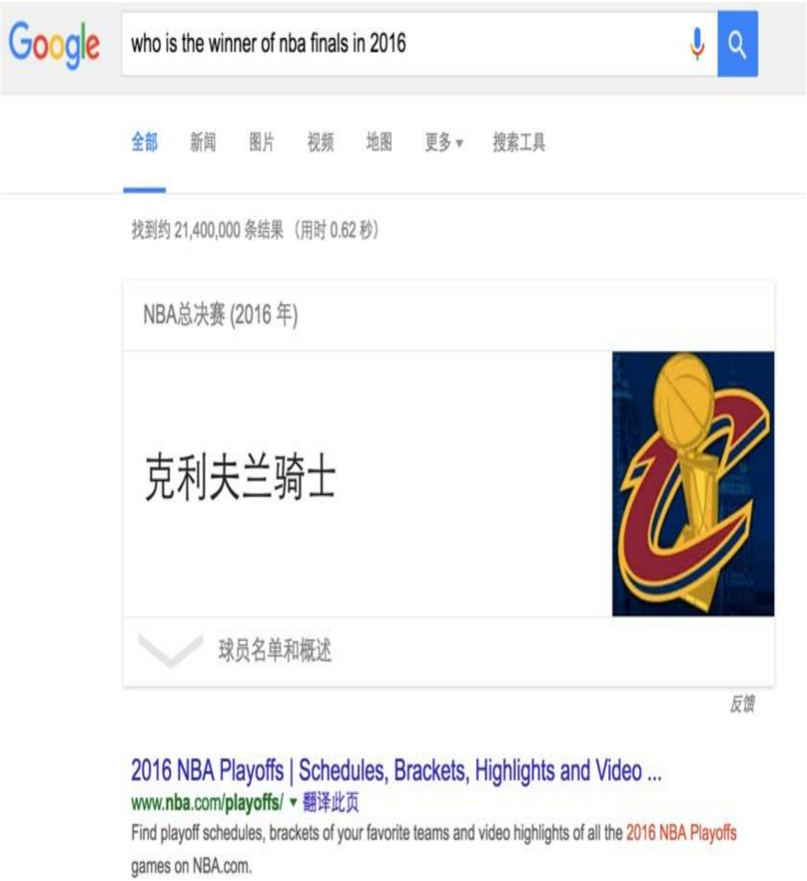
Conversational AI OS for everyone and everywhere

VIV

THE GLOBAL BRAIN

Viv radically simplifies the world by providing an intelligent interface to everything.

基于知识图谱的问答



who is the winner of nba finals in 2016

找到约 21,400,000 条结果（用时 0.62 秒）

NBA总决赛 (2016 年)

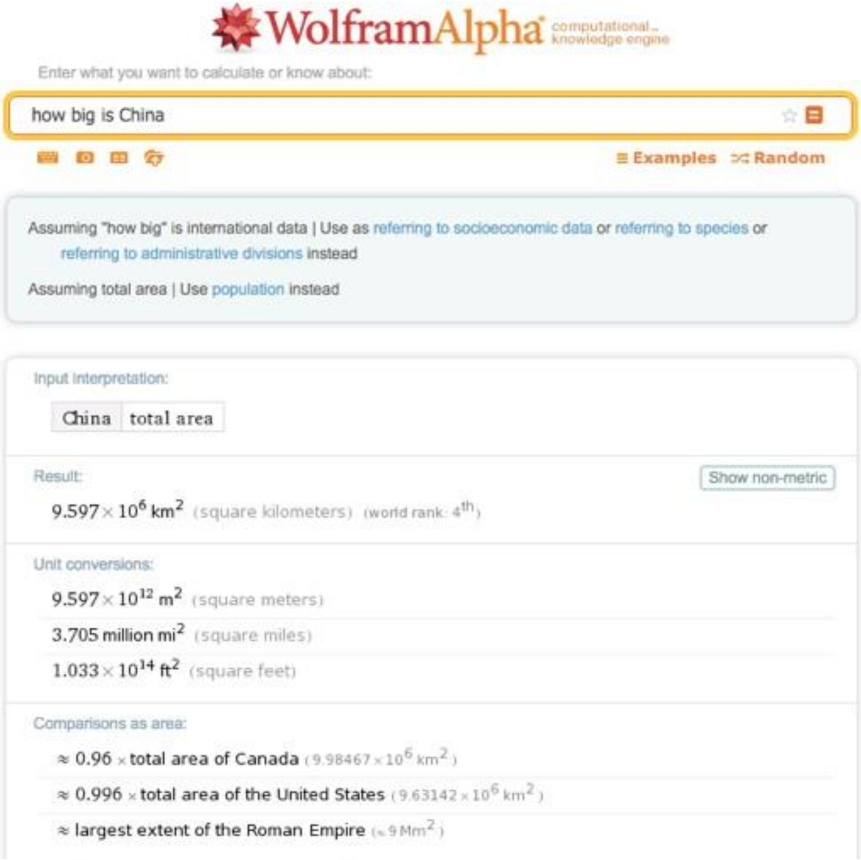
克利夫兰骑士

球员名单和概述

2016 NBA Playoffs | Schedules, Brackets, Highlights and Video ...

www.nba.com/playoffs/ ▾ 翻译此页

Find playoff schedules, brackets of your favorite teams and video highlights of all the 2016 NBA Playoffs games on NBA.com.


WolframAlpha computational knowledge engine
Enter what you want to calculate or know about:
≡ Examples
⤨ Random

how big is China

Assuming "how big" is international data | Use as referring to socioeconomic data or referring to species or referring to administrative divisions instead

Assuming total area | Use population instead

Input interpretation:

China	total area

Result:

9.597×10^6 km^2 (square kilometers) (world rank: 4th)

Unit conversions:

9.597×10^{12} m^2 (square meters)

3.705 million mi^2 (square miles)

1.033×10^{14} ft^2 (square feet)

Comparisons as area:

≈ 0.96 × total area of Canada (9.98467×10^6 km^2)

≈ 0.996 × total area of the United States (9.63142×10^6 km^2)

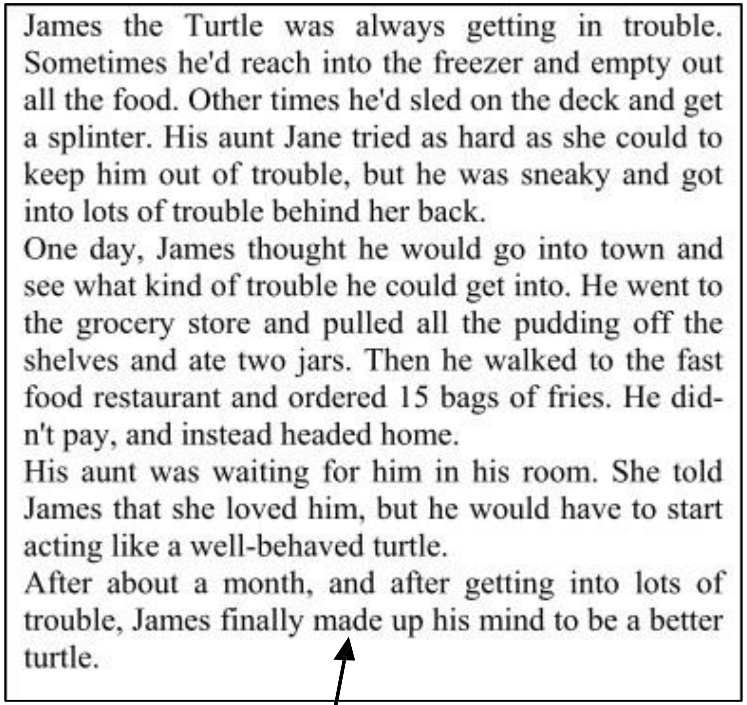
≈ largest extent of the Roman Empire (≈ 9 Mm^2)

Lecture 16.4：问答与对话

本节内容

- 阅读理解式问答系统
- 对话系统

阅读理解式问答



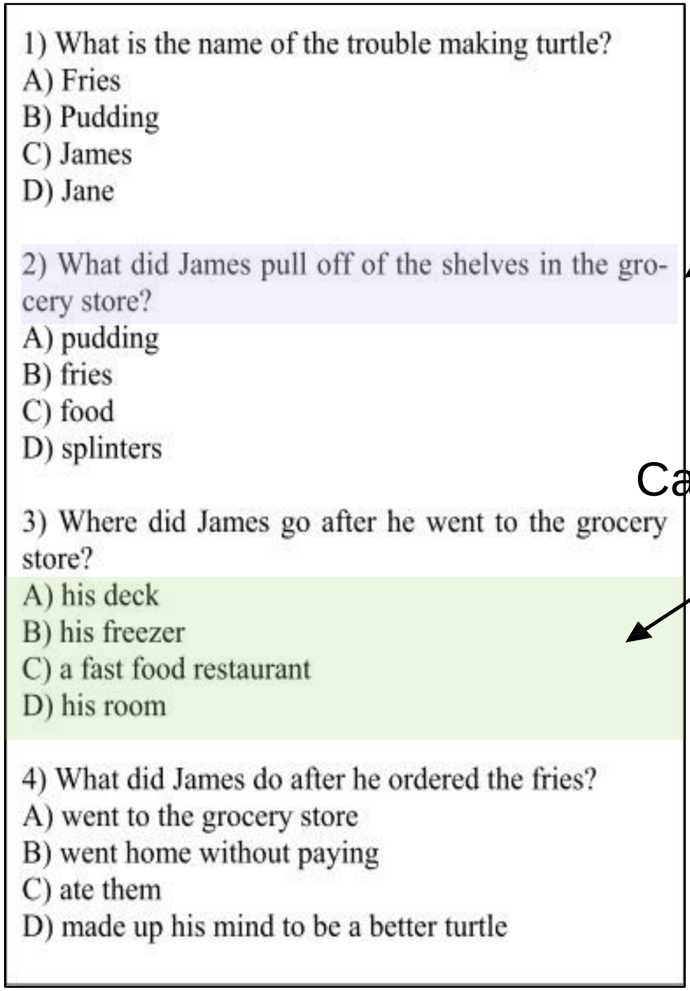
James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

Document



1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

4) What did James do after he ordered the fries?
A) went to the grocery store
B) went home without paying
C) ate them
D) made up his mind to be a better turtle

Question

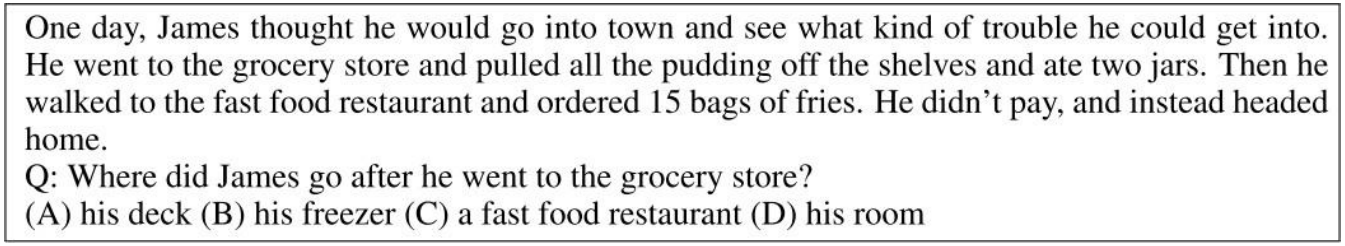
Candidate Answer

阅读理解式问答

- 给出的信息来源只有一篇相关文档
- 问题答案候选已经给出，一般由几个选项构成
- 问题形式多种多样，主要考察语义理解和推理

数据集：MCTest (EMNLP 2013)

- 早期阅读理解任务的一个典型的数据集
  - 文档：660个短故事
  - 问题：2640个人工提出的问题（每个文档对应4个问题）
  - 答案：四选一的选择题形式

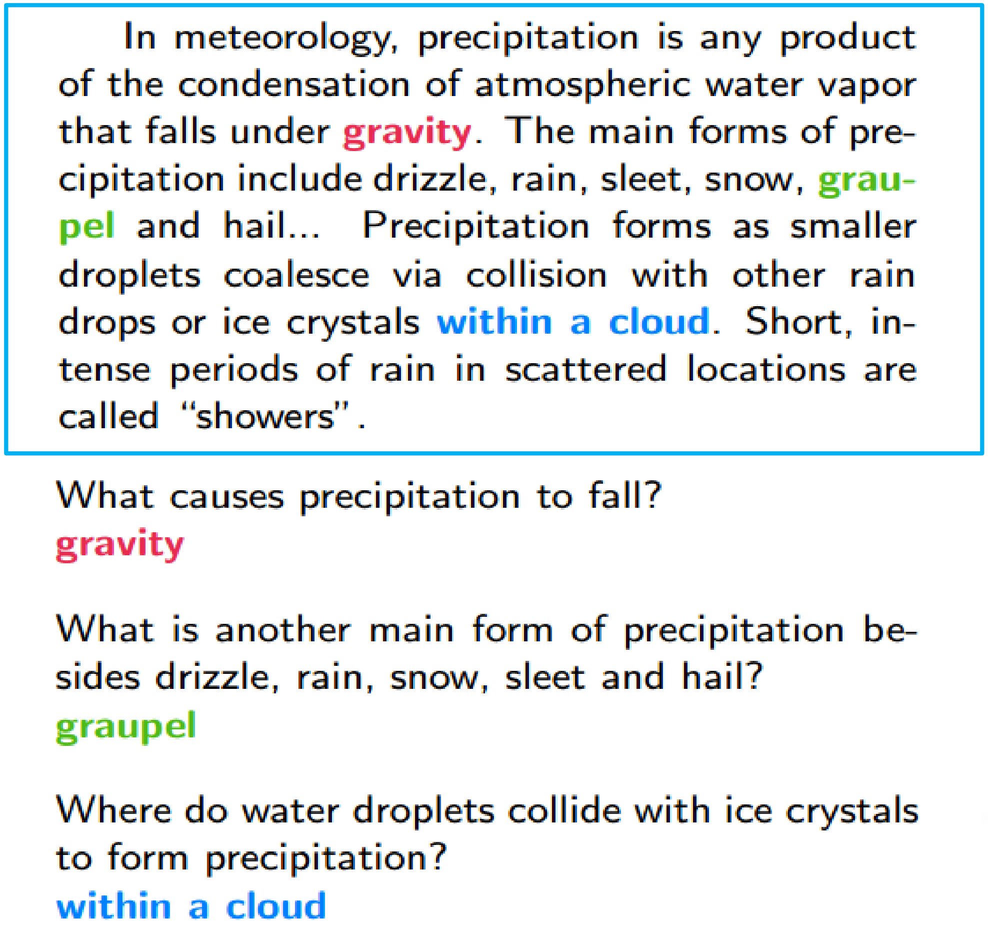


One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
Q: Where did James go after he went to the grocery store?
(A) his deck (B) his freezer (C) a fast food restaurant (D) his room

- 问题被限定在7岁儿童可回答的范围，考察该层次机器的推理能力
  - 优点：问题难度较大，包含很多常识性问题
  - 缺点：数据规模非常小，深度学习模型难以应用

数据集： SQuAD (EMNLP 2016)

- 近两年来最经典且最被广泛关注的阅读理解数据集



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail… Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

Document

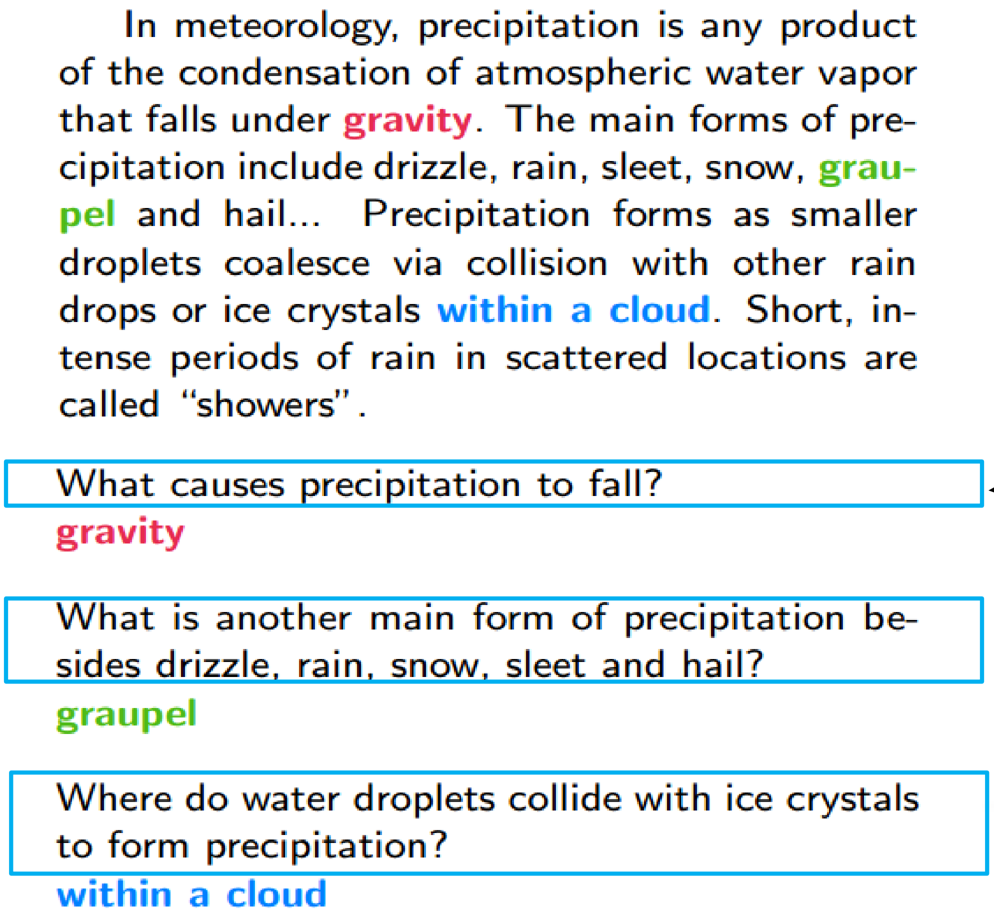
What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

数据集： SQuAD (EMNLP 2016)

□ 近两年来最经典且最被广泛关注的阅读理解数据集



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

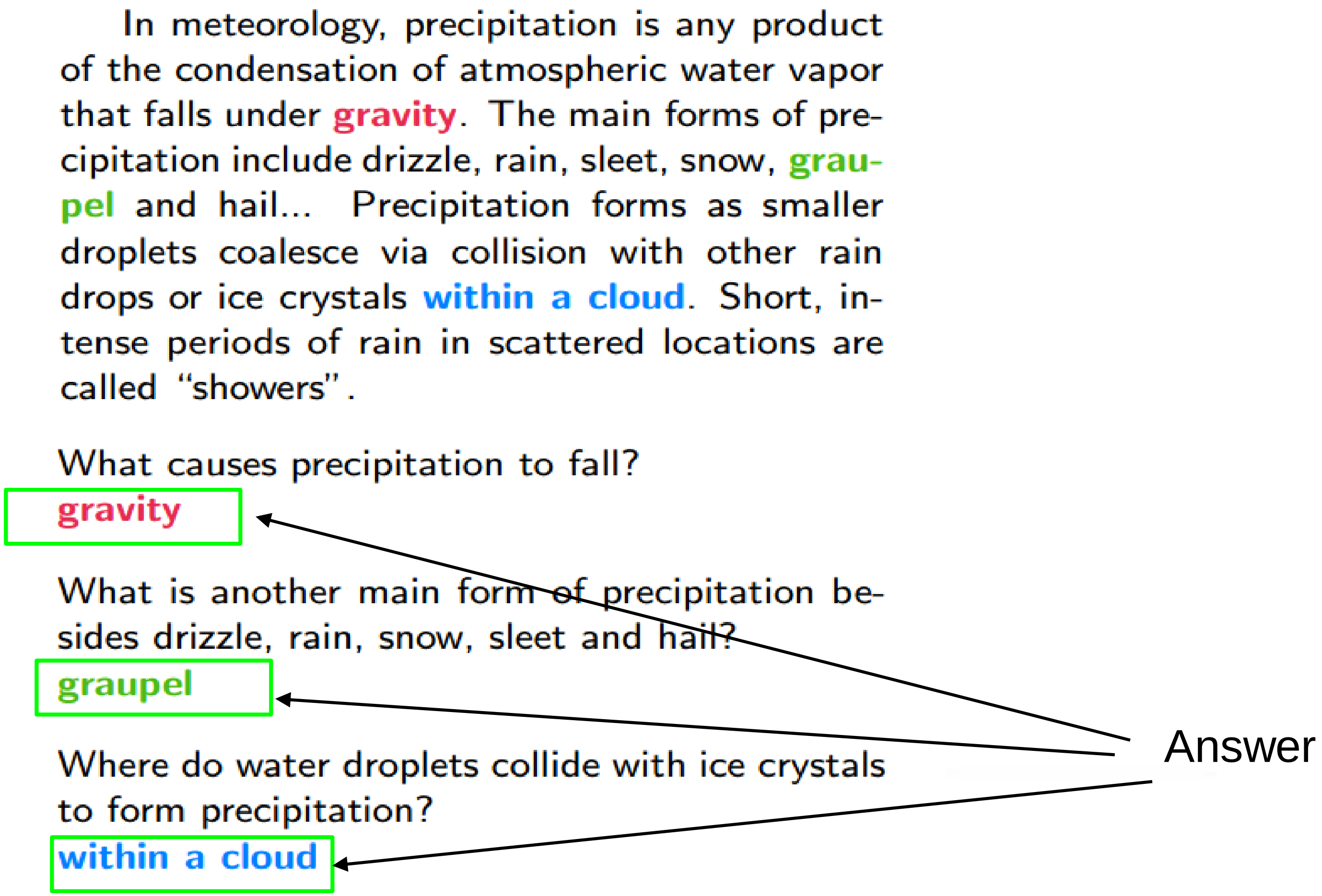
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Question

数据集： SQuAD (EMNLP 2016)

- 近两年来最经典且最被广泛关注的阅读理解数据集



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail… Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

数据集： SQuAD (EMNLP 2016)

❑ 近两年来最经典且最被广泛关注的阅读理解数据集

- 文档：从536篇Wikipedia的文章中抽取2万多个段落
- 问题：基于文档，人工生成的10万个问题
- 答案：原文中的一个区间（一个词或几个词组成）

❑ SQuAD引领了近两年阅读理解任务的发展

- 优点：数据规模较大，适用于深度学习方法;并且数据质量较高
- 缺点：
  - ○ 段落文档均出自536篇Wikipedia文章，词汇和表达的多样性不足
  - ○ 问题被限定为必须能被原文区间回答，导致可提问的角度受限

数据集： MS MACRO (NIPS2016)

❑代表了阅读理解任务向开放域发展的新趋势

- 文档：100万篇由搜索引擎搜索得到的文章
- 问题：10万个Bing搜索中出现的真实问题
- 答案：根据文档，人工总结得到(不一定是原文区间)

数据集： MS MACRO (NIPS2016)

❑ 所谓“开放域”

	给定	要求
SQuAD等任务	问题，文档	答案
开放域	问题	答案

❑ 数据集特点

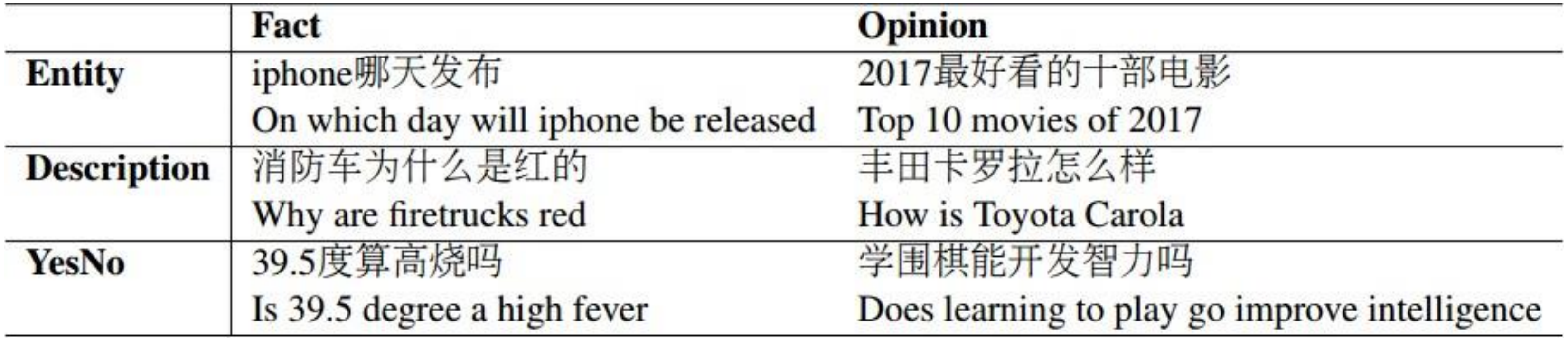
- 更加贴近真实的问答场景，即只有问题，事先没有准备好的文档；该数据集的文档是根据已有的问题去搜索并整理出文档，而不是根据文档提出问题

数据集：DuReader (2017,Baidu)

- DuReader数据集是一个比较有代表性的中文数据集
  - 文档：100万篇由搜索引擎搜索得到的文章
  - 问题：20万个Baidu搜索中出现的真实问题
  - 答案：根据文档，人工总结得到

数据集：DuReader (2017,Baidu)

- DuReader形式上与MS_MARCO很相似，最大的特点是引入观点型(opinion)问题：



	Fact	Opinion
Entity	iphone哪天发布 On which day will iphone be released	2017最好看的十部电影 Top 10 movies of 2017
Description	消防车为什么是红的 Why are firetrucks red	丰田卡罗拉怎么样 How is Toyota Carola
YesNo	39.5度算高烧吗 Is 39.5 degree a high fever	学围棋能开发智力吗 Does learning to play go improve intelligence

- 将问题按照事实性(fact)和观点型(opinion)两个角度划分，之前其他数据集只有这里的 fact类型问题，此处引入opinion型。但是，该类型问题的引入使得答案评判更难

机器阅读理解的方法

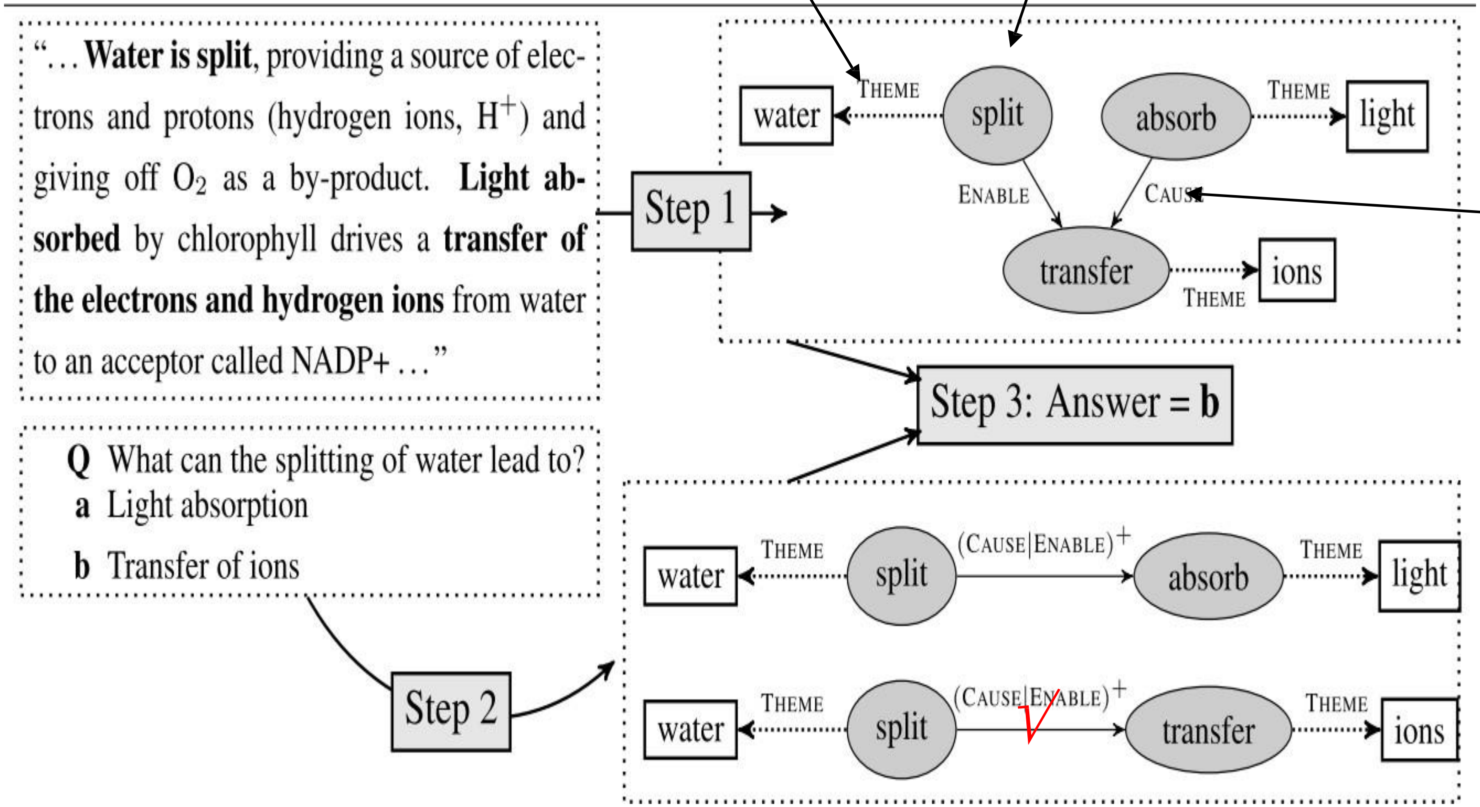
- 传统特征工程的方法
  - 文本分析
  - 问句解析
  - 匹配答案
- 神经网络的方法
  - 文档和问句的表示学习
  - 文档和问句的匹配计算
  - 深度推理机制

基于传统特征工程的方法

事件元素的分类提取 |
语义角色标注（SRL）

事件抽取、分类

事件之间的
关系分类



"… **Water is split**, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O_2 as a by-product. **Light absorbed** by chlorophyll drives a **transfer of the electrons and hydrogen ions** from water to an acceptor called NADP+ …"

Q What can the splitting of water lead to?
a Light absorption
b Transfer of ions

Modeling Biological Processes for Reading Comprehension, by Jonathan Berant et al. EMNLP14

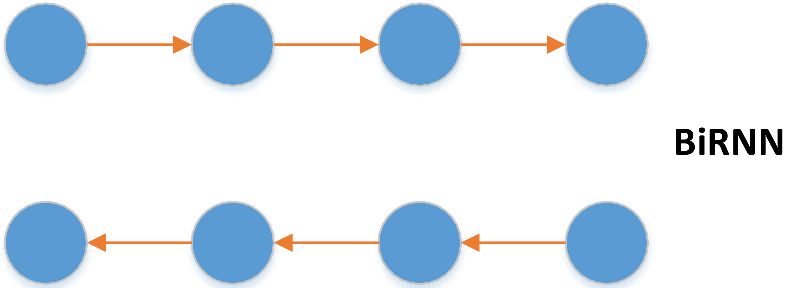
基于传统特征工程的方法

- 优点
  - □ 对过程进行建模，清晰明了，各个部分的作用可以显式地表示
  - □ 对问题进行了同样的过程处理，最后的结果是确定的
  - □ 语义建模方式非常明显，类似于标准的Semantic Parsing，每一部分的语义都能很直观地表示出来
- 缺点
  - □ 由于是基于传统特征工程的方法，非常耗时耗力，训练样本有限
  - □ 领域适应能力差，在这个训练集上训练的模型会有领域倾向性

基于神经网络的方法

■ 文章表示

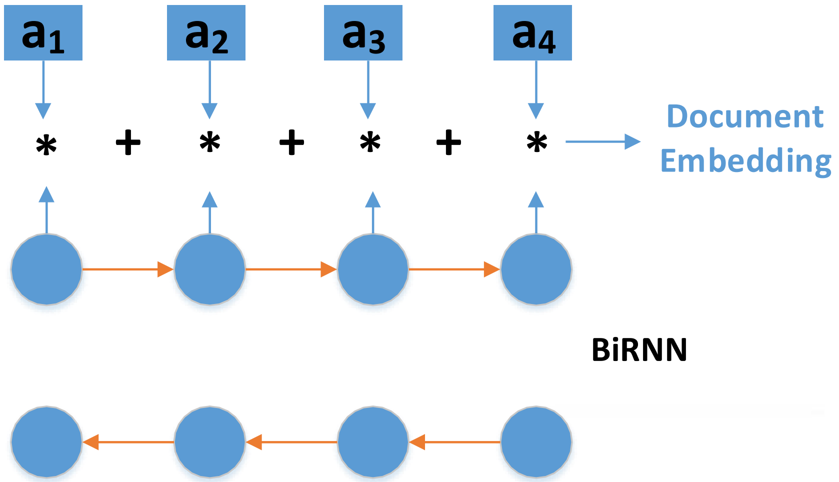
- □ 第一种方法：将文章看作单词序列，在这个序列上使用RNN对文章进行建模， 每个单词对应RNN序列中的一个时刻t的输入，RNN的隐层状态是融合了当前词义和上下文语义的编码



基于神经网络的方法

■文章表示

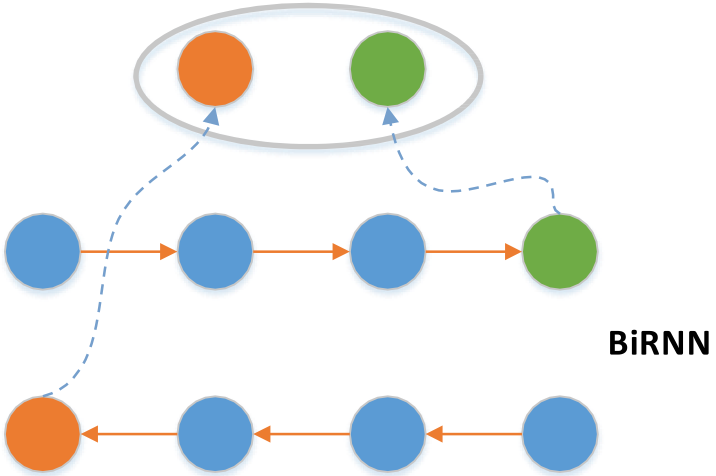
- □ 第二种方法：引入Attention机制，该方法也是采用双向RNN对每个单词及其上下文信息进行建模，得到隐层状态表示。不同点在于，得到隐层表示向量的每一维都要乘以某个系数，该系数代表该单词对于整个文章语义表达的重要程度。



基于神经网络的方法

问句表示：

- 有与文章表示方法类似的两种建模方法
- 还有另外一种建模方法：首先使用双向RNN对其进行表示。将这双向RNN首尾词节点的隐层状态拼接起来，就得到了整个句子的最终表示



基于神经网络的方法

■问句表示：

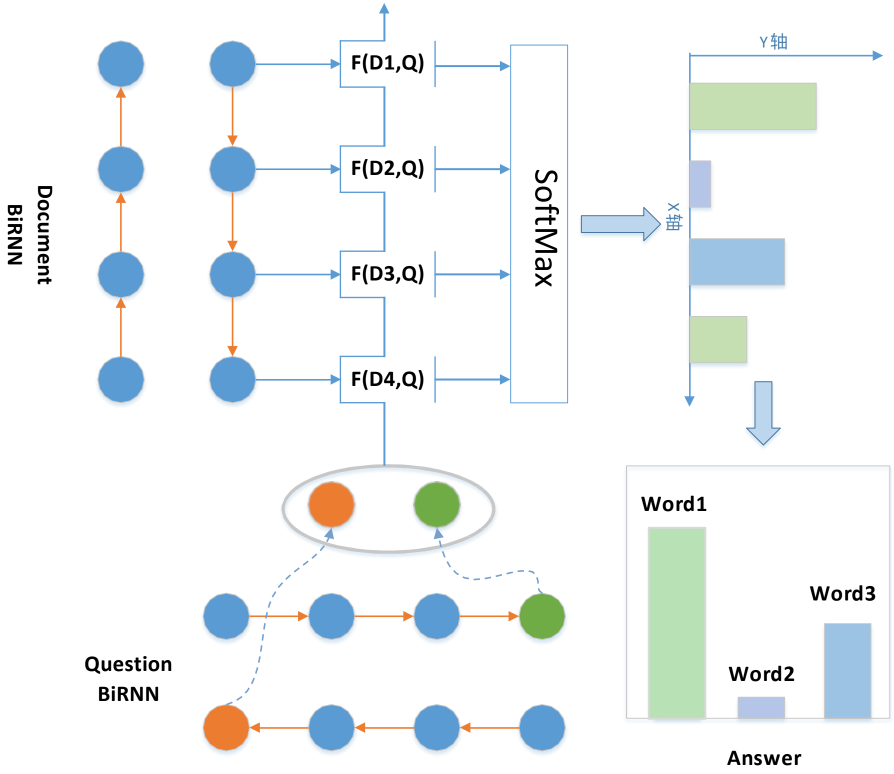
- □ 有与文章表示方法类似的两种建模方法
- □ 还有另外一种建模方法：首先使用双向RNN对其进行表示。将这双向RNN首尾词节点的隐层状态拼接起来，就得到了整个句子的最终表示

其中，正向RNN的尾部单词隐层节点（图中绿色节点）正向融合了整个句子语义信息；相对的，反向RNN的尾部单词（局首词）隐层节点（图中橙色节点）则逆向融合了整个句子的语义信息。

基于神经网络的方法

文章与问句的匹配：一维匹配模型

- 双向RNN对文章和问题表示后，通过某种匹配函数来计算文章中每个单词D_i和问题Q的语义信息的匹配程度。
- 之后，对每个单词的匹配程度通过SoftMax函数进行归一化，这样就形成了一种Attention的操作，将更可能是问题答案的单词凸显出来。



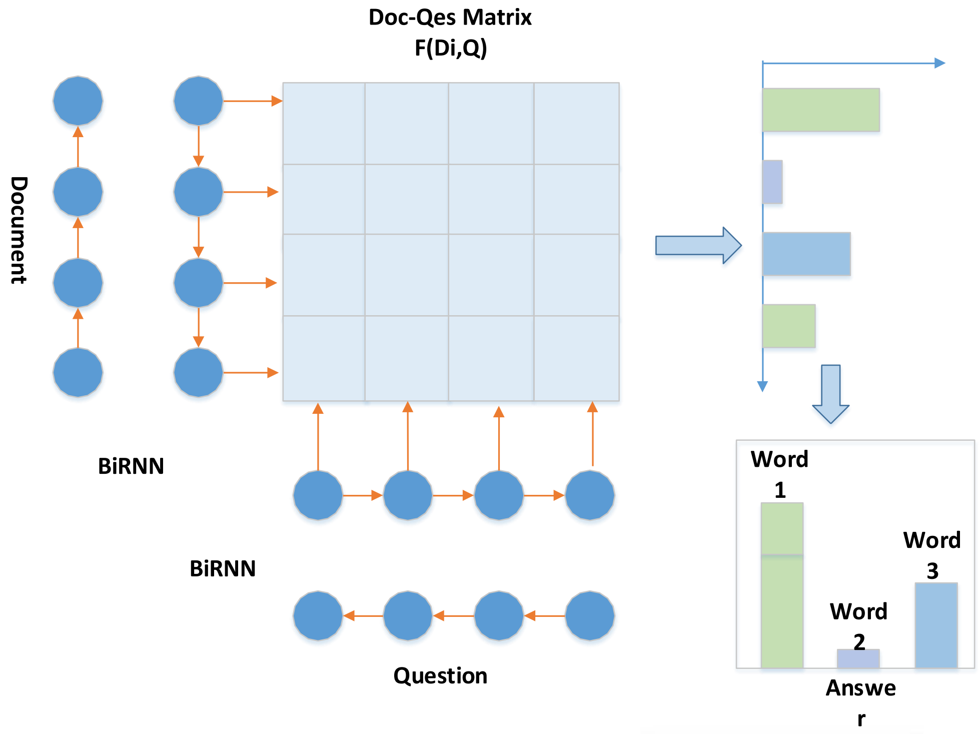
基于神经网络的方法

■ 文章与问句的匹配：二维匹配模型

- ❖ 其整体结构和一维匹配模型是类似的，最主要的区别体现在如何计算文章和问题的匹配程度上
- ❖ 一维匹配模型在匹配过程中，形成一维线性结构模型；而二维匹配模型在进行问题和文章的匹配时，形成二维矩阵结构
- ❖ 二维匹配模型使得问题Q中的每个词都可以和文档D中的词进行交互，粒度更细。

基于神经网络的方法

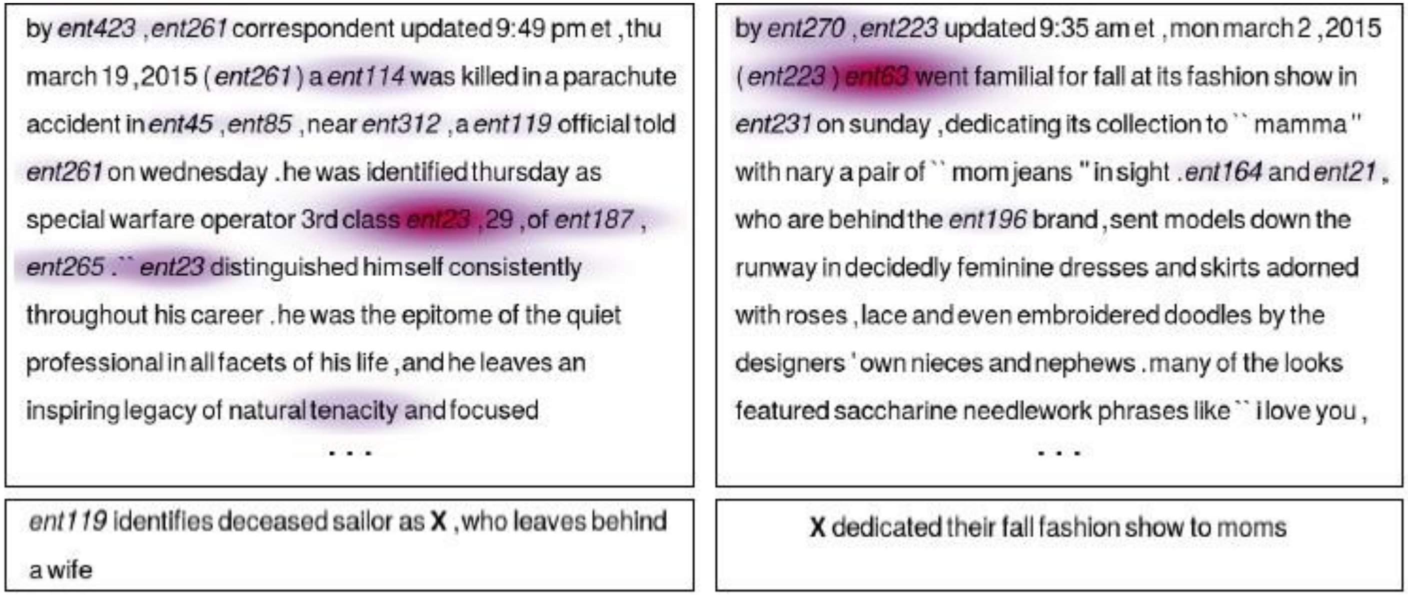
■文章与问句的匹配：二维匹配模型



基于神经网络的方法

❑ 可视化：

- 通过可视化的方法可以观察到模型在推理过程中哪些词被重点关注了，一定程度上为模型提供可解释性：



by *ent423* ,*ent261* correspondent updated 9:49 pm et ,thu march 19,2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* ,*ent85* ,near *ent312* ,a *ent119* official told *ent261* on wednesday .he was identified thursday as special warfare operator 3rd class *ent23* ,29 ,of *ent187* , *ent265* .`` *ent23* distinguished himself consistently throughout his career .he was the epitome of the quiet professional in all facets of his life ,and he leaves an inspiring legacy of natural tenacity and focused

. . .

by *ent270* ,*ent223* updated 9:35 am et ,mon march 2 ,2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday ,dedicating its collection to `` mamma'' with nary a pair of `` mom jeans '' in sight .*ent164* and *ent21* , who are behind the *ent196* brand ,sent models down the runway in decidedly feminine dresses and skirts adorned with roses ,lace and even embroidered doodles by the designers ' own nieces and nephews .many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

ent119 identifies deceased sailor as **X** ,who leaves behind a wife

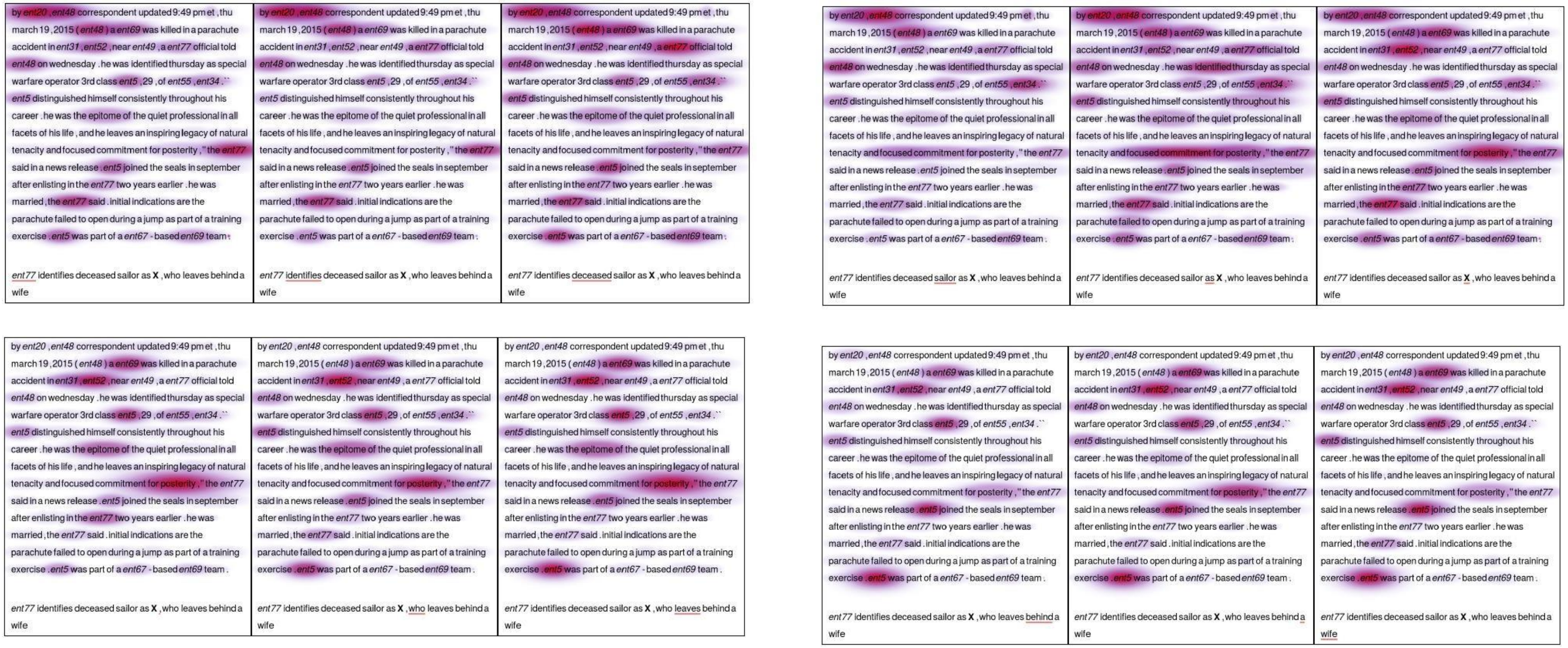
X dedicated their fall fashion show to moms

Teaching machines to read and comprehend, NIPS 2015

基于神经网络的方法

可视化：

- 问句不同，文档被关注的部分也不同：



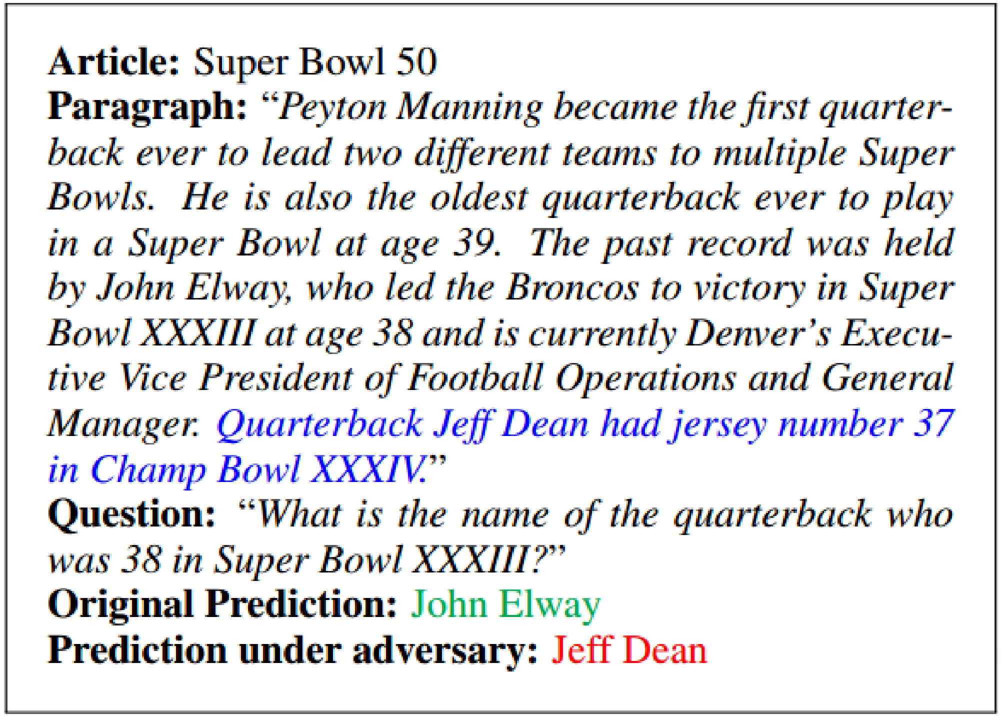
by *ent20* , *ent48* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent48*) a *ent69* was killed in a parachute accident in *ent31* , *ent52* , near *ent49* , a *ent77* official told *ent48* on wednesday . he was identified thursday as special warfare operator 3rd class *ent5* , 29 , of *ent55* , *ent34* . `` *ent5* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused commitment for posterity , '' the *ent77* said in a news release . *ent5* joined the seals in september after enlisting in the *ent77* two years earlier . he was married , the *ent77* said . initial indications are the parachute failed to open during a jump as part of a training exercise . *ent5* was part of a *ent67* - based *ent69* team .

ent77 identifies deceased sailor as **X** , who leaves behind a wife

基于神经网络的方法

■ 插入对抗负样本：

- □ Adversarial Examples for Evaluating Reading Comprehension Systems（Percy Liang, EMNLP2017)



Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

- □ 原本模型能够得到正确答案。在文档后面插入一句和包含答案的线索句子形式上相似的负样本句子(图中蓝色部 分)。结果，模型受到干扰，得出错误答案。

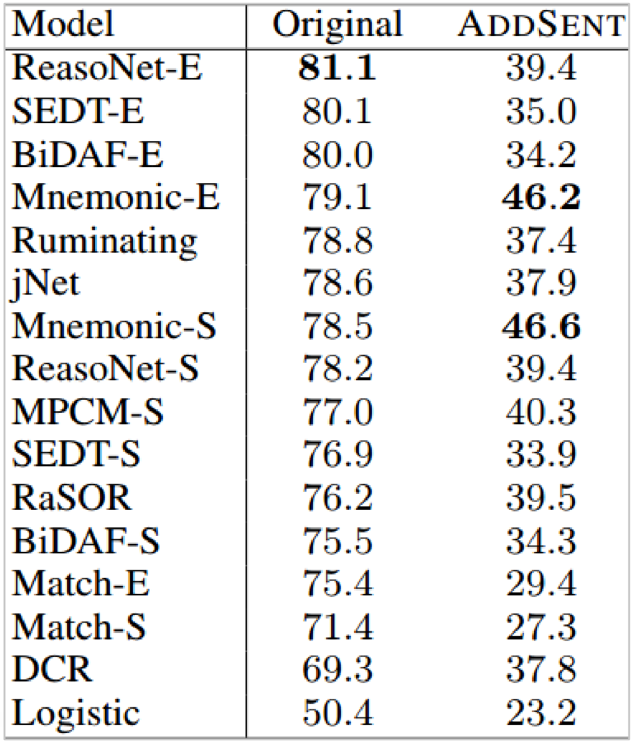
- □ 然而，这种负样本句子并不能对人的判断造成影响。

基于神经网络的方法

插入对抗负样本：

- Adversarial Examples for Evaluating Reading Comprehension Systems（Percy Liang, EMNLP2017)

- 测试16个之前在SQuAD数据集上表现良好的模型
- 在文章中插入负样本后，模型的性能都大幅下降
- **这篇工作表明：当前阅读理解模型只具有识别浅层模式的能力，而并没有真正的语言理解能力。因此要达到真正理解语言的目标，研究者们还有很长的路要走。**



Model	Original	ADDSENT
ReasoNet-E	**81.1**	39.4
SEDT-E	80.1	35.0
BiDAF-E	80.0	34.2
Mnemonic-E	79.1	**46.2**
Ruminating	78.8	37.4
jNet	78.6	37.9
Mnemonic-S	78.5	**46.6**
ReasoNet-S	78.2	39.4
MPCM-S	77.0	40.3
SEDT-S	76.9	33.9
RaSOR	76.2	39.5
BiDAF-S	75.5	34.3
Match-E	75.4	29.4
Match-S	71.4	27.3
DCR	69.3	37.8
Logistic	50.4	23.2

Thank you!

权小军　中山大学数据科学与计算机学院