

深度问答技术

Deep Question Answering

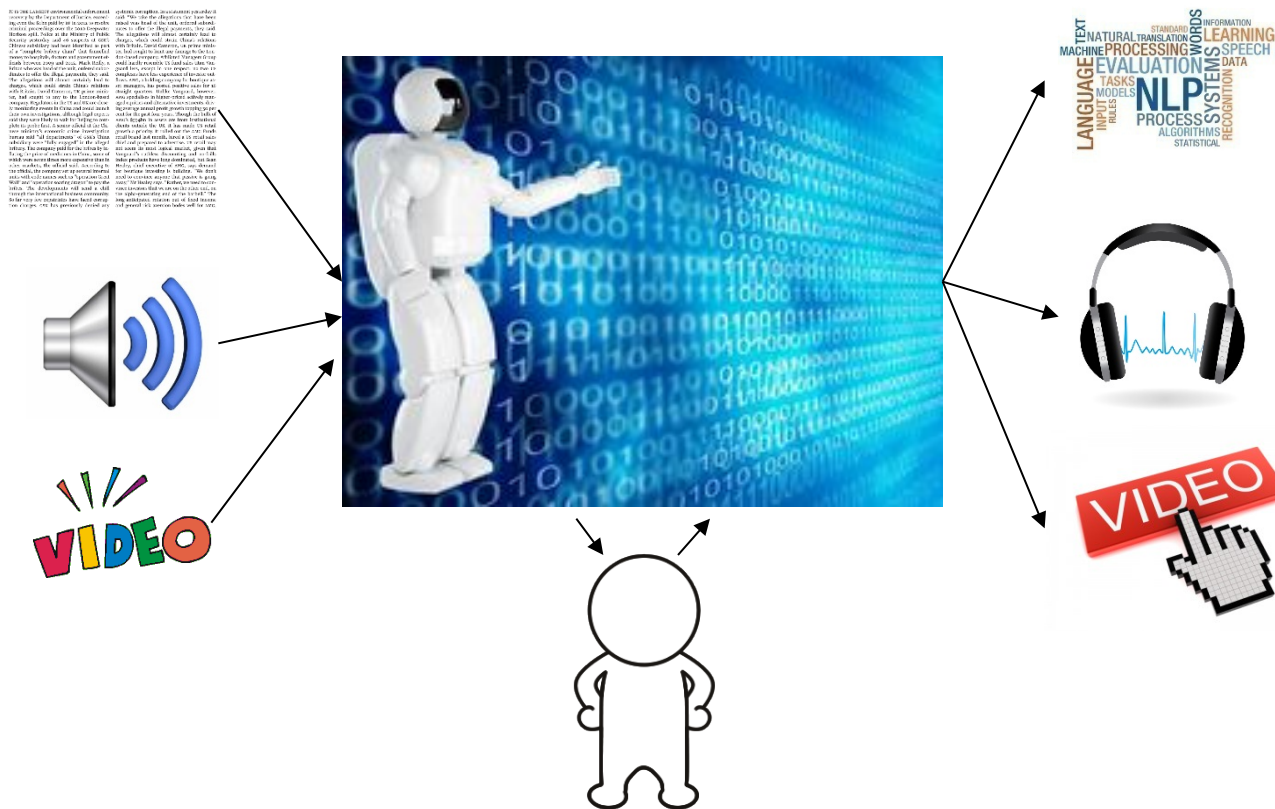
赵 军

模式识别国家重点实验室



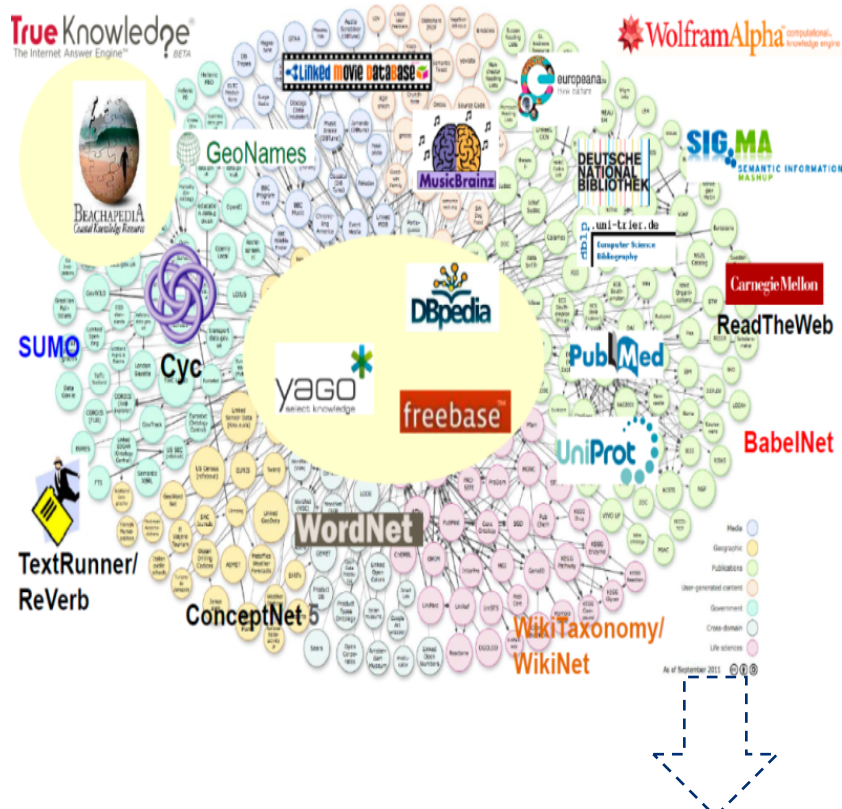
中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

自然语言处理



人与计算机之间用自然语言进行通信的理论和方法

问答系统



自然语言问句

我国有两个省接壤的省份最多，
与它们接壤的省分别有哪些？

涉及到自然语言理解、知识表示、知识推理等人工智能的难点问题

问答发展历程

图灵测试

专家系统

问答式检索

基于知识图谱的深度问答

社区问答

阅读理解式问答

1950

1960

1990

2000

2011

2013



MIT
BaseBall

MIT
MACSYMA

哈佛
LUNAR

TREC



COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

2011年IBM WATSON

提纲

- 深度问答中的知识表示
 - 符号表示
 - 分布表示
- 关键技术
 - 问句语义解析
 - 知识抽取
 - 推理规则的学习
- 资源融合
 - 文本、抽取知识库与知识图谱
- 展望

提纲

●深度问答中的知识表示

- 符号表示
- 分布表示

●关键技术

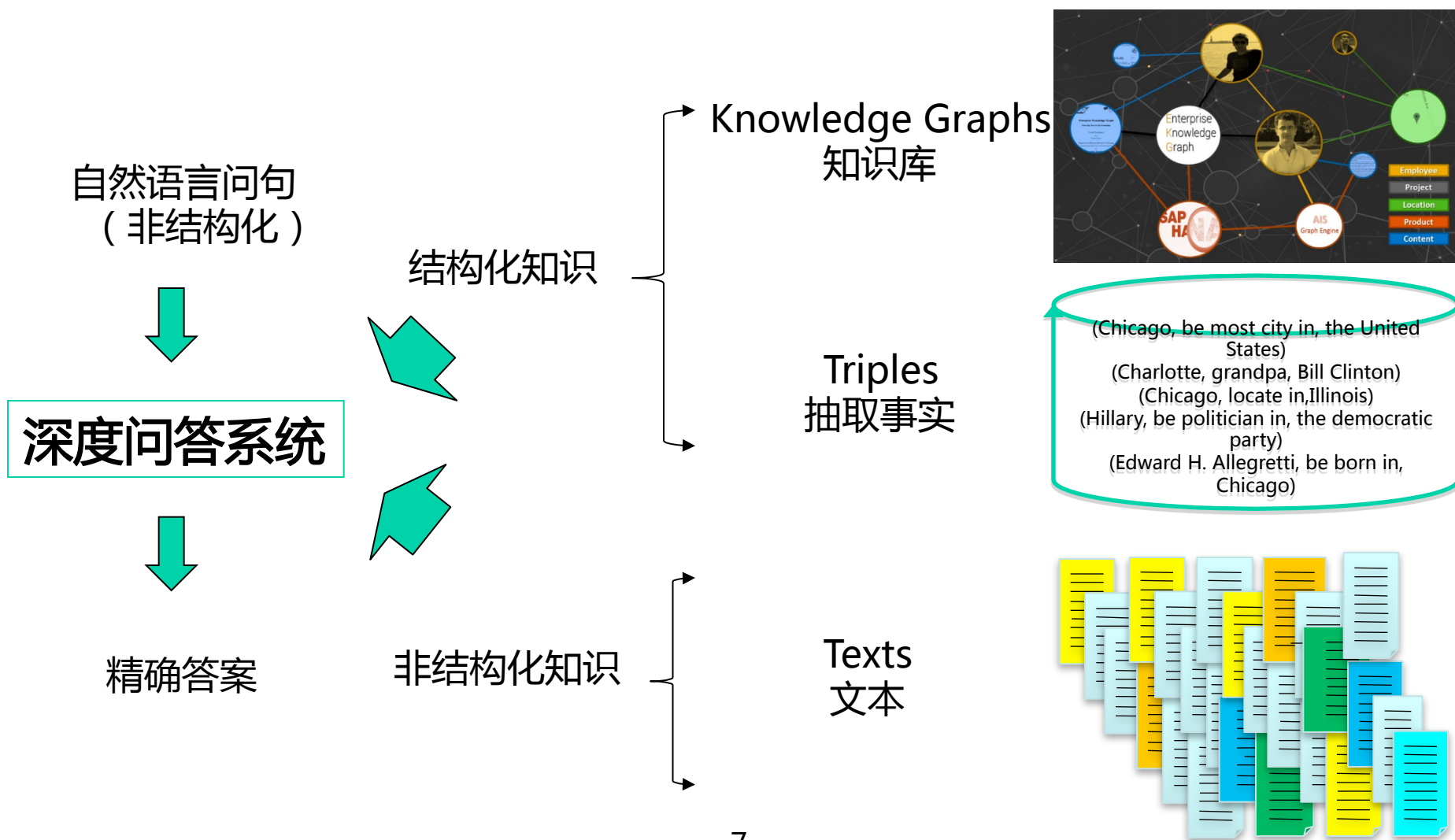
- 问句语义解析
- 知识抽取
- 推理规则学习

●资源融合

- 文本、抽取知识库与知识图谱

●展望

答案的三种来源



人的问答过程

两岸媒体四川采访结束 李维一吁持续关注重建
国际社会向中国地震灾区提供现金援助44亿多元

汶川地震共造成四川68712人遇难 17921人失踪

华夏经纬网 2009-05-07 12:42:50

字号: 小 大

中新网5月7日电 四川省人民政府新闻办公室今日通报“5.12”汶川特大地震灾后恢复重建情况。通报称,在“5.12”汶川特大地震抢险救援阶段,经各级抗震救灾指挥部统计,四川省共有8712名同胞遇难,17921名同胞失踪。

通报说,核实公布遇难人员名单,是政府对历史负责、对人民负责、对遇难同胞负责的具体体现。四川省委、省政府高度重视这项工作,确定了实事求是、依法办事、分类处理的原则和核实一批、公布一批的办法,制订了依法规范的工作方案。遇难人员名单公布,须由利害关系人或所在单位、基层组织向乡(镇)政府或街道办事处申报,并提供死亡证明、火化通知书或宣告死亡判决书等相关材料,报经县(市、区)政府审核后,由乡(镇)政府或街道办事处张榜公布;失踪人员名单公布,须由利害关系人或所在单位、基层组织向乡(镇)政府或街道办事处申报,并提供宣告失踪判决书,报经县(市、区)政府审核后,由乡(镇)政府或街道办事处张榜公布。

“5.12”汶川特大地震共造成全省181个县(市、区)中的98个县(市、区)有遇难或失踪人员,目前80%以上的县(市、区)已经完成公布工作。

“5.12”汶川特大地震是新中国成立以来破坏性最强、波及范围最广、救灾难度最大的一次地震,公布遇难失踪人员名单涉及面广、情况复杂、政策性强,必须依法依规进行。

通报说,当前核实公布工作面临一些困难和问题,如遇难失踪人员中的流动人口身份核实困难;部分遗体发现时已无法辨认身份;紧急状态下为防止发生疫情,对部分遗体实行集中掩埋或就地掩埋,没有留下身份资料;部分遇难失踪人员的利害关系人由于各种原因没有或不申报公布名单;宣告死亡或失踪必须由利害关系人向基层法院申请,并履行相关法律程序,所需时间较长。

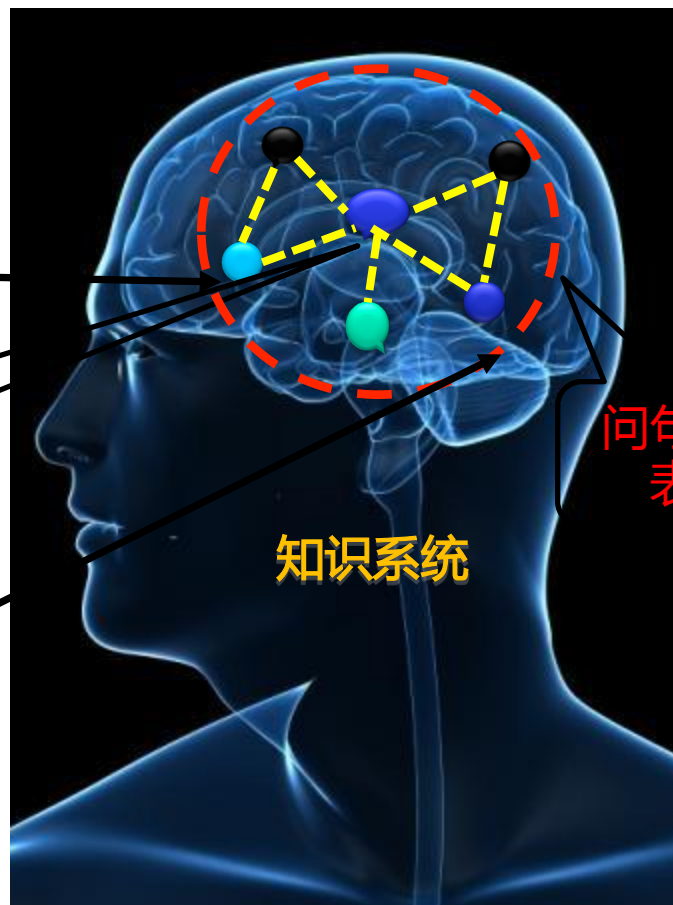
通报称,灾区各级政府将继续本着高度负责的精神,严格依法办事,郑重稳妥地做好“5.12”汶川地震遇难失踪人员名单公布工作和最终遇难人数核定工作。

最近一次死亡受伤人数在十万人以上的地震发生在哪个国家?

知识抽取
和表示

知识推理

问句语义
解析



问句语义
表示

知识系统

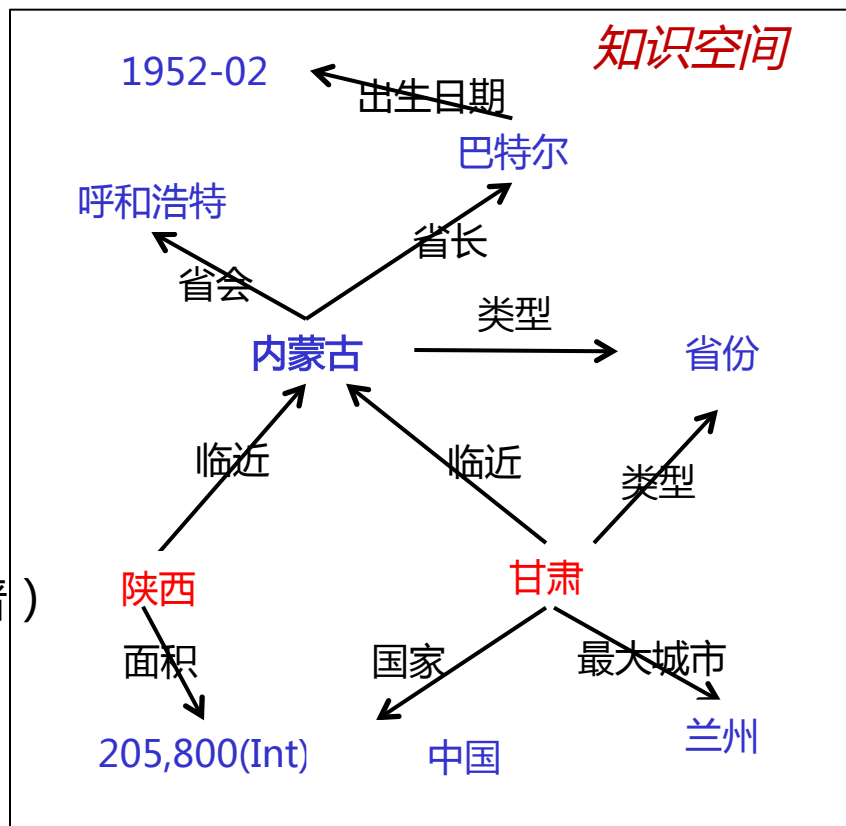
- 1) 语义单元识别
- 2) 知识激活与关联
- 3) 知识推理

符号化的知识表示：结构化知识

形式化语义表示方法：

- 1) 产生式系统
- 2) 谓词逻辑
- 3) 框架
- 4) 脚本
- 5) 语义网络 (知识图谱)

...



数理演算：

- 1) lambda 演算

...

近似推理：

- 1) 归纳逻辑编程
- 2) 马尔科夫逻辑网
- 3) 概率软逻辑

...

概念或关系的语义蕴含在形式化知识结构之中
通过数理演算和近似推理进行语义计算 (问句解析和知识推理)

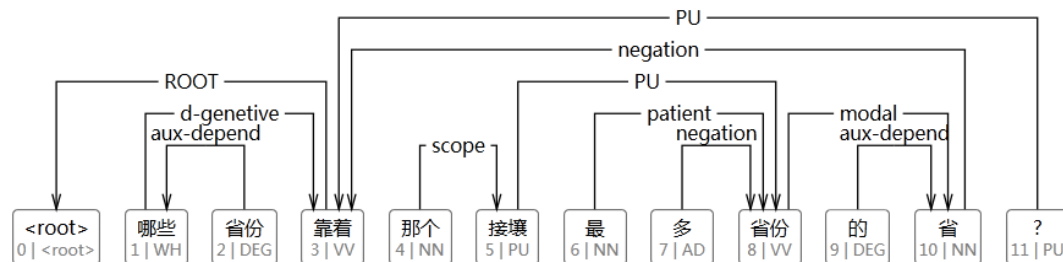
符号化的知识表示：非结构化知识

哪些省份靠着那个接壤最多省份的省？

- 关键词组合

{靠着、省份、接壤、最多}

- 语义树(图)结构



- 逻辑表达式结构

$\lambda x. \text{省}(x) \wedge \text{临近}(x, \text{argmax}(\lambda y. \text{省}(y), \lambda y. \text{count}(\lambda z. \text{省}(z) \wedge \text{临近}(y, z))))$

易于数值计算的知识表示：分布表示

- 语言符号和知识符号

- 符号逻辑系统

- 计算机擅长处理什么

- 二进制：进行布尔运算
- 数字：进行数值计算

- 自然语言处理

- 用计算机能处理的方式表示和处理自然语言



使用数值表示自然语言中的概念和结构
通过数值计算进行语义计算

易于数值计算的知识表示：分布表示

●自然语言处理任务

分类：情感分析、词性标注

预测：语言模型、文本生成

匹配：搜索、问答

翻译：机器翻译

结构预测：信息抽取、句法分析



内容 $\rightarrow$ 表示 $\rightarrow$ 类标

内容 $\rightarrow$ 内容 + 表示

$\langle$ 内容, 内容 $\rangle \rightarrow$ 表示 $\rightarrow$ 类标(数值)

内容 $\rightarrow$ 表示 $\rightarrow$ 内容

内容 $\rightarrow$ 结构 + 表示

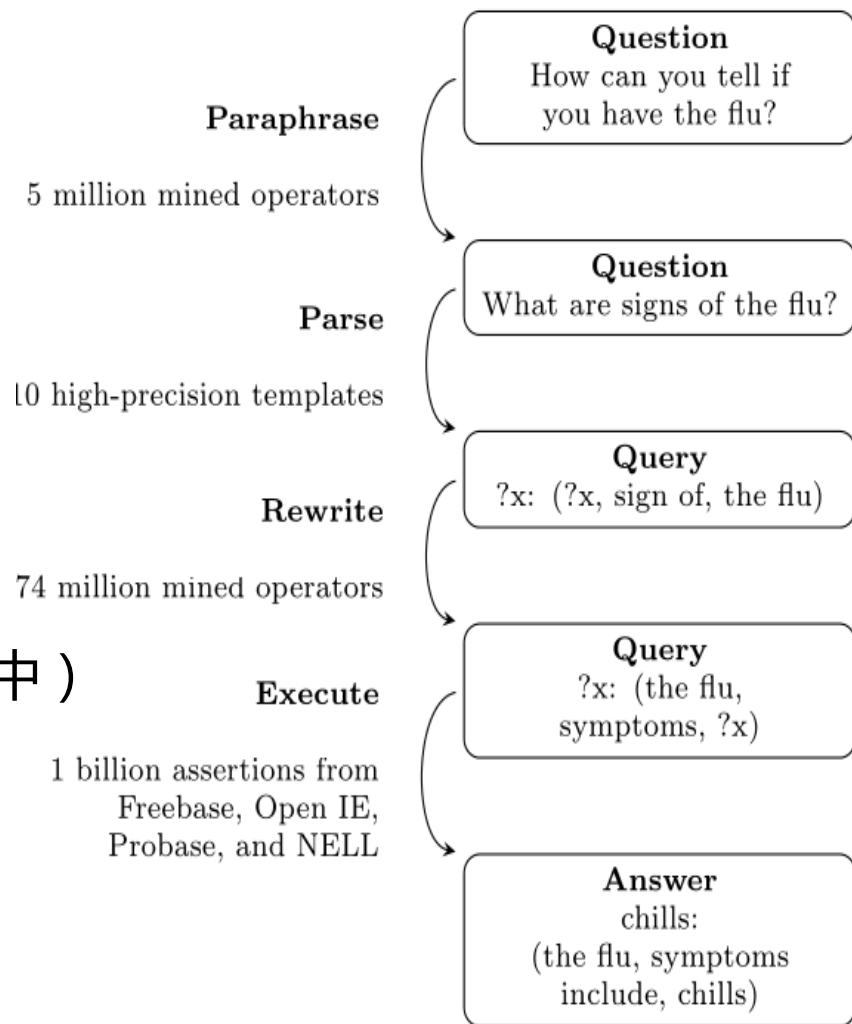
表示：特征工程 $\rightarrow$ 特征学习（表示学习）

基于符号语义的深度问答

基于复述的语义匹配：事实库问答

[Fader et al. KDD 2014]

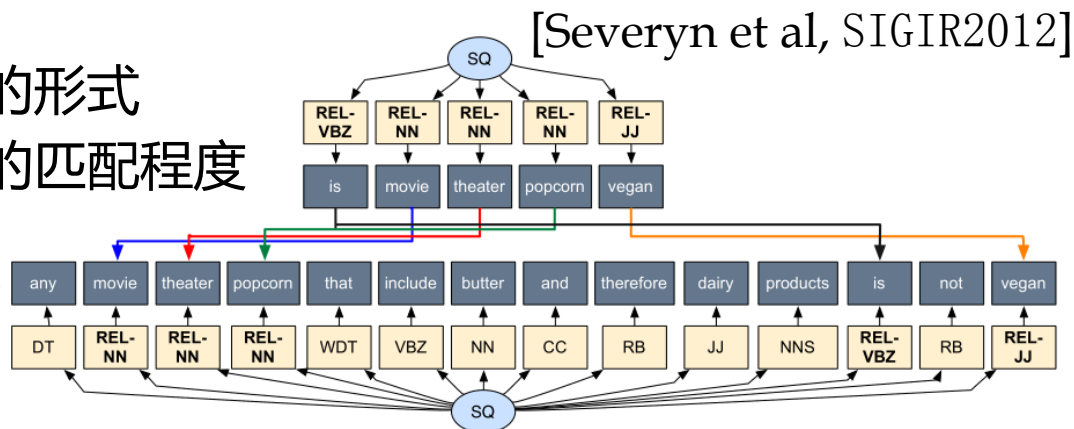
- 回答事实性问题
- 信息抽取工具抽取出事实三元组
- 如何利用这些事实三元组回答问题
- 步骤1：问句复述（问句的不同问法）
- 步骤2：问句解析（问句映射到事实库中）
- 步骤3：查询转换（事实的不同表示）
- 步骤4：执行查询，回答问题



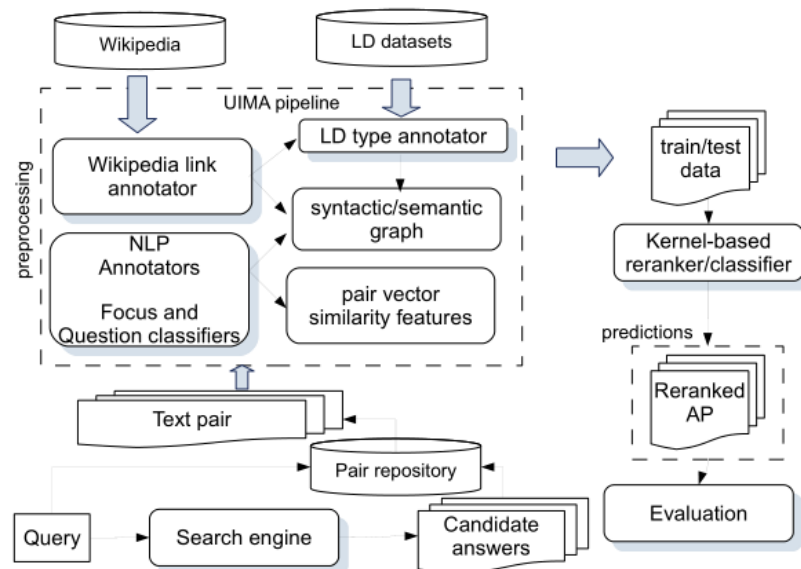
基于句法树的语义匹配：文本库问答

- 问句和答案表示为句法树的形式
- 基于树核函数计算树结构的匹配程度

$$TK(T_1, T_2) = \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \Delta(n_1, n_2)$$



- 答案篇章重排序
- 利用基于浅层和深度句法分析器的句法和语义结构的树核函数
- 在表示中利用链接到开放知识库中的资源进行匹配



基于篇章语义图的语义演算：篇章阅读理解

[Berant et al. EMNLP 2014]

- 从文本中抽取事实、事件及他们的结构，用符号进行表示
- 从问句中抽取关键元素，生成相应的形式化查询
- 把查询在知识结构中匹配和推理，确定最终答案

“... **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

需要对文本进行深度理解

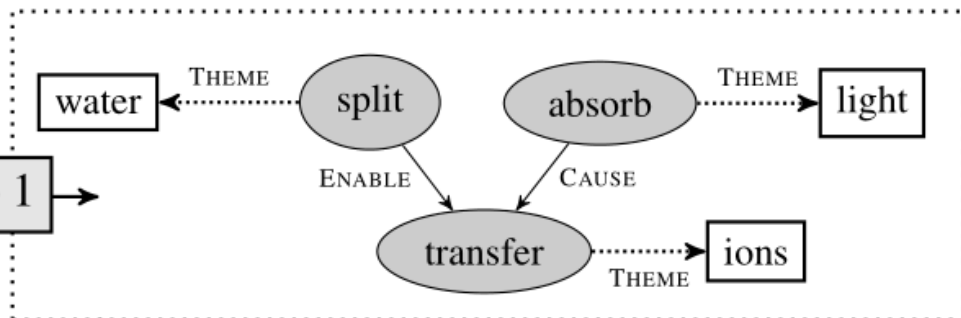
基于篇章语义图的语义演算

[Berant et al. EMNLP 2014]

“... **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+ ...”

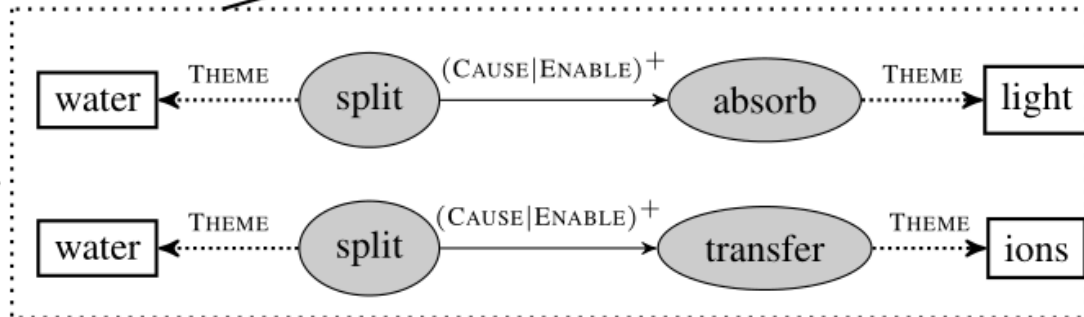
Q What can the splitting of water lead to?
a Light absorption
b Transfer of ions

Step 1



Step 3: Answer = b

Step 2

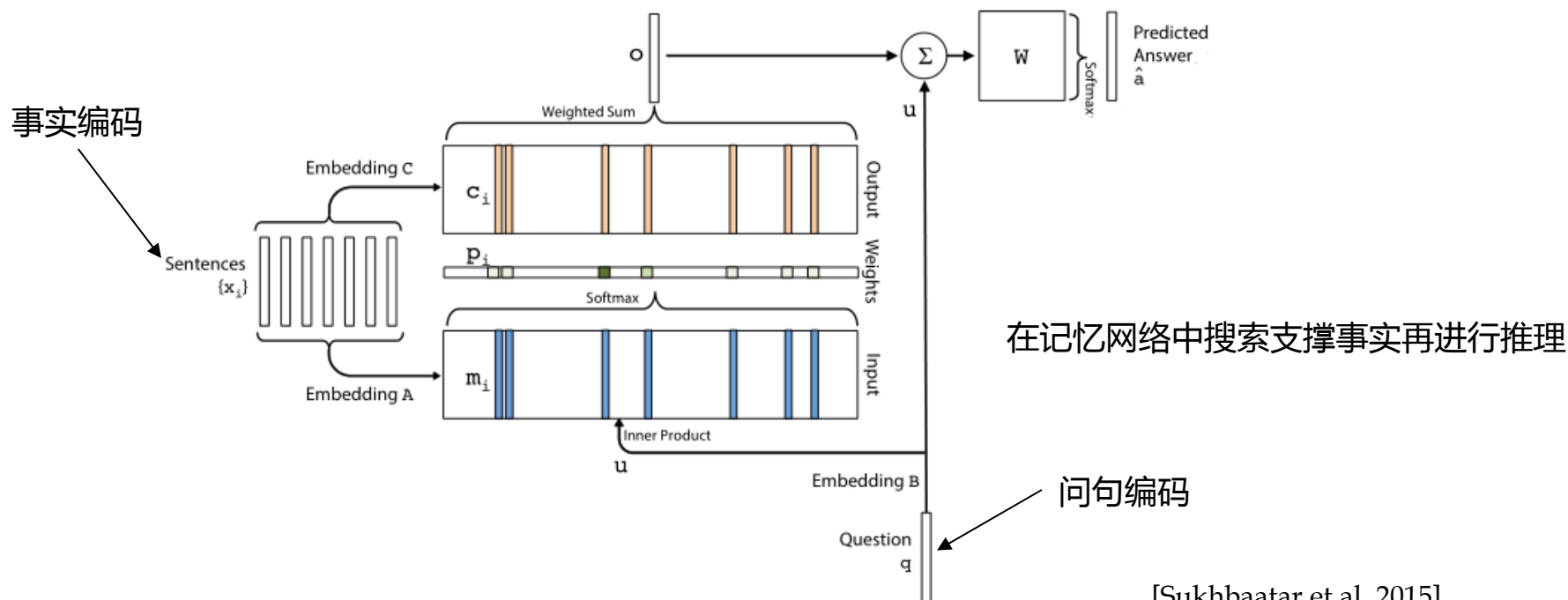


- 步骤1：分析输入段落的结构，用实体、事件关系图表示其含义
- 步骤2：对于每个答案，把问题-答案组合分析为同样的结构，形成查询
- 步骤3：在语义结构图中进行匹配，得到最终答案

基于分布语义的深度问答

在潜在空间中表示含义

- 基于神经网络对问句、文本、知识图谱和答案直接编码，表示为向量、矩阵等形式
 - 不需要NLP操作
 - 不需要人工提取特征
 - 在潜在空间中学习问句和答案的匹配



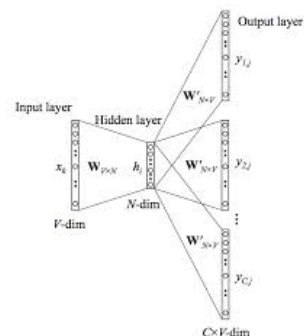
文本的向量化

- 词的向量化

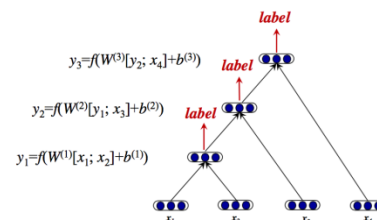
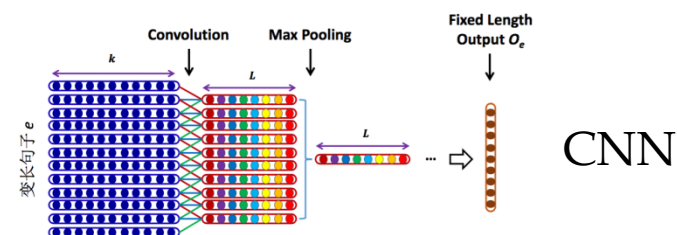
- NNLM, C&W, Skip-gram

- 句子 (文本) 的向量化

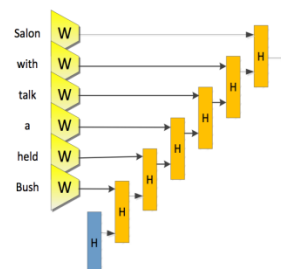
- 直接累加
- 卷积神经网络
(*Convolutional Neural Network*)
- 递归神经网络
(*Recursive neural network*)
- 循环神经网络
(*Recurrent neural network*)



Word Embedding



Recursive NN

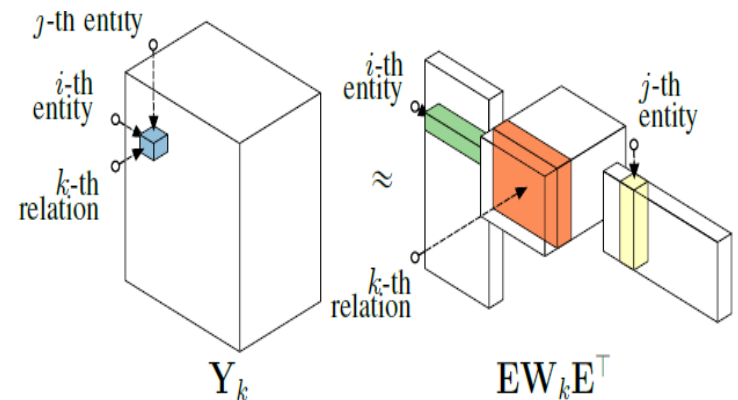
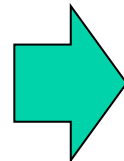


Recurrent NN

知识库的向量化

- 把知识库中的实体和关系表示为低维空间的对象及其它们的操作，该表示能够蕴涵其在知识库中的性质。
 - Subjects and objects are represented by **vectors in R^d** : $\{\mathbf{E}_k\}_{k \in [1..n_e]}$
 - Relations are represented by **operator between vectors**: $\{\mathbf{R}_k\}_{k \in [1..n_r]}$
 - The learning depending on the similarities of triplets: **$\text{sim}(\text{sub}, \text{rel}, \text{obj})$** , it is parameterized by **$\mathbf{E}, \mathbf{R}$**

Subject(Arg1)	Relation(Property)	Object(Arg2)
Politician	subClassOf	Person
Hillary Clinton	type	Politician
Hillary Clinton	spouse	Bill Clinton
Hillary Clinton	birth_place	Chicago
Hillary Clinton	height	1.69m
birth_place	domain	Person
birth_place	range	Location



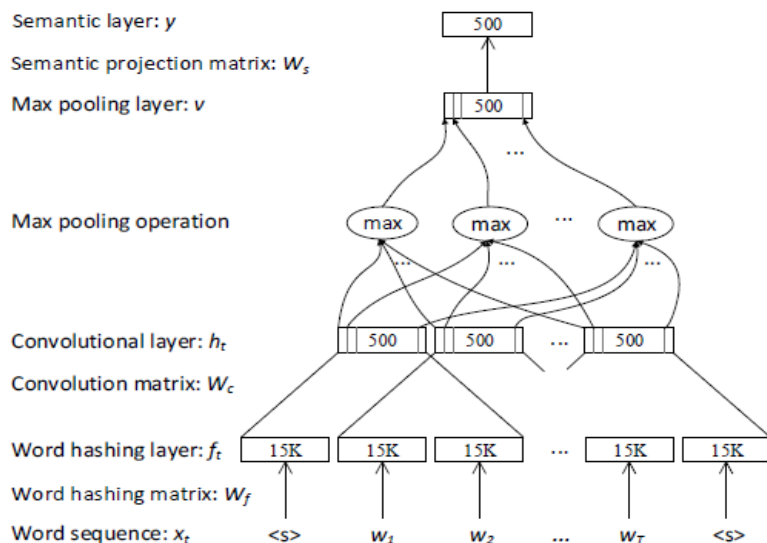
symbol VS vector (matrix)

RESCAL (Nickle, Tresp, and Kriegel 2011)

基于CNN的单关系语义解析

[Yih et al, ACL 2014]

- 使用规则把问句拆成实体描述和关系描述两部分
- 对于实体描述，使用CNN学习它与目标知识库中实体的匹配
- 对于关系描述，使用CNN学习它与目标知识库中关系的匹配
 - When were DVD players invented
 - » Question $\rightarrow$ Relation Pattern $\wedge$ Entity Mention
 - » Relation Pattern $\rightarrow$ when were X invented (*be-invent-in*)
 - » Entity Mention $\rightarrow$ dvd players (*dvd-player*)



$$f(cat) = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Indices of #-c-a, c-a-t, a-t-# in the letter-tri-gram list, respectively.

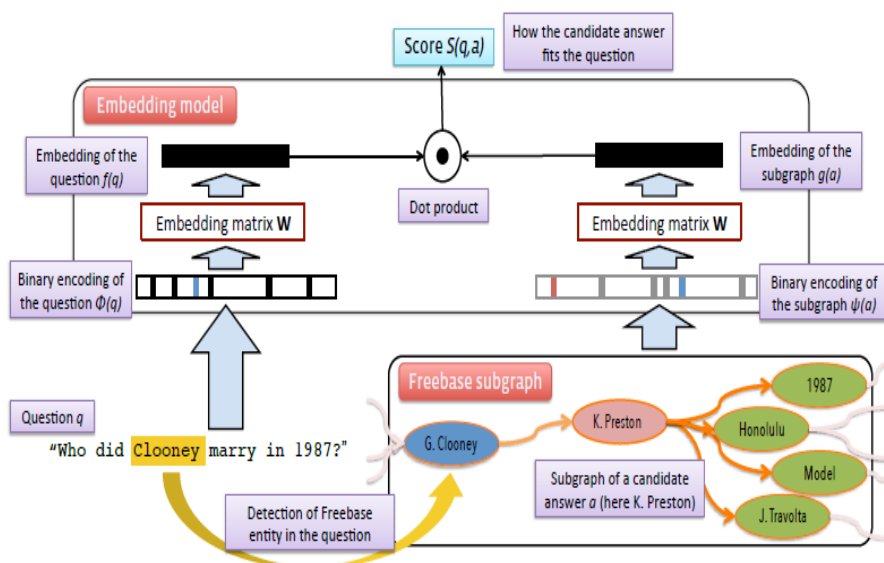
对于实体和关系建立两个独立的模型，没有考虑实体和关系之间的联系

问句和知识库的联合分布表示学习

[Borders et al. EMNLP 2014]

● 核心思想

- 对问句和答案进行联合表示
- 问句表示为其中词表示的组合
- 答案表示为它在知识图谱中所对应子图的表示
- 把它们表示在同一个空间中，学习问句中的词语与知识库中实体和关系的匹配



– 问句和子图进行向量表示

» 问句： $f(q) = \sum_{w \in q} \text{vec}(w)$

» 子图： $g(a) = \sum_{l \in g} \text{vec}(l)$

– 计算向量相似度对子图进行排序

» $s(q, a) = f(q)^T \cdot g(a)$

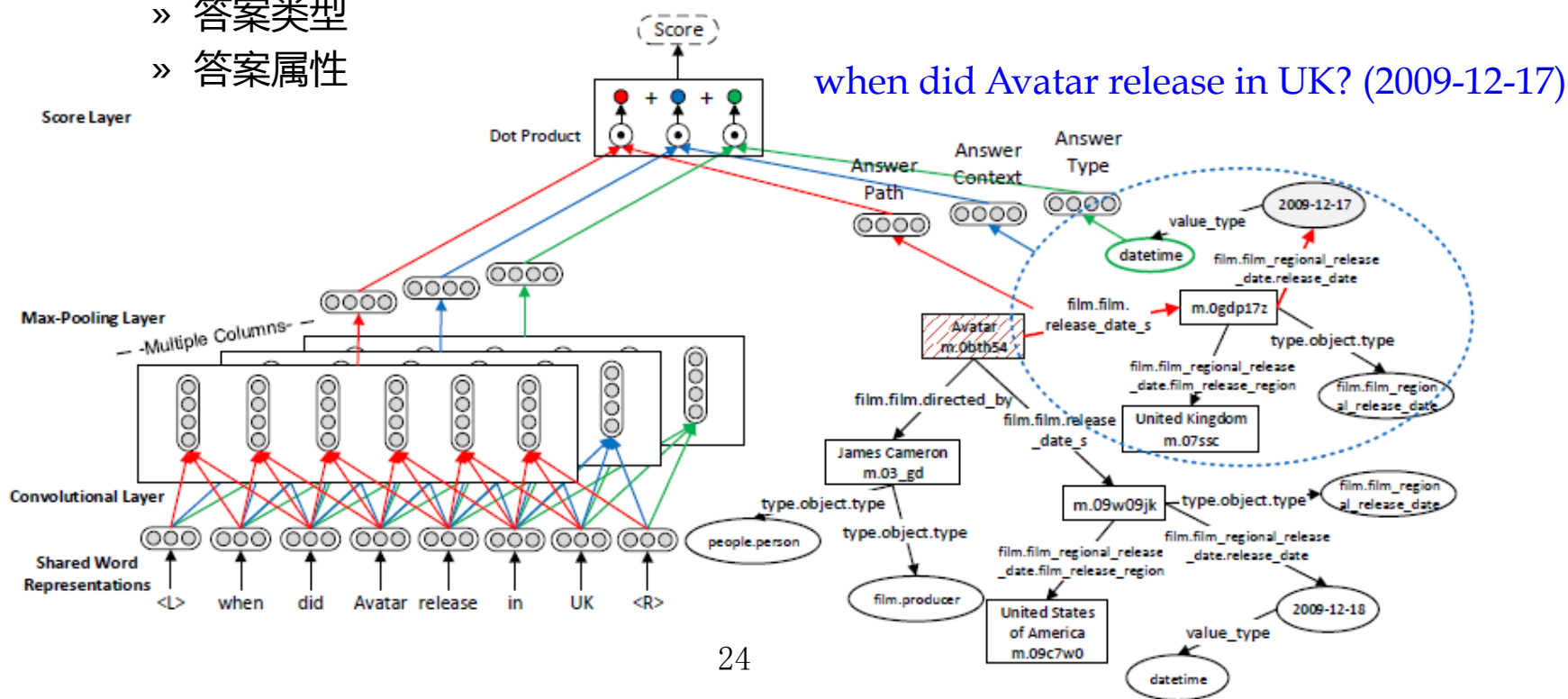
» $a' = \sum_{a \in A(q)} s(q, a)$

使用CNN组合问句表示 [Dong et al. ACL 2015]

●核心思想

- 使用CNN组合问句语义
- 使用答案的不同上下文重要度
 - » 查询实体-答案路径
 - » 答案类型
 - » 答案属性

MCCNN
多列CNN



符号表示方法 vs 分布表示方法

● 符号表示的方法

- 将自然语言问句转化为知识符号的逻辑语义表达式，在知识图谱上进行基于语义的查询
- 准确度好，覆盖度低，泛化能力弱，不方便扩展，推理效率低
- 用户可理解，流水线方式

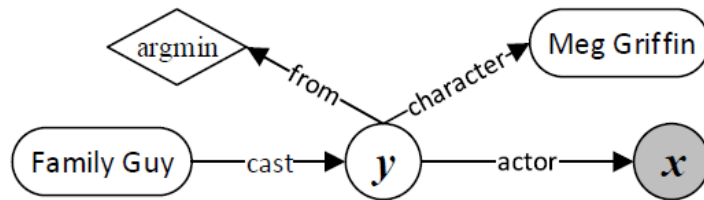
● 分布表示的方法

- 将自然语言问句和知识图谱都转化为低维空间中的数值表示，在低维数值空间中计算知识图谱中的知识元和问句语义的匹配程度
- 准确度弱，覆盖度好，泛化能力强，容易扩展和融合到其他系统，效率相对高
- 不可解释，端到端方式

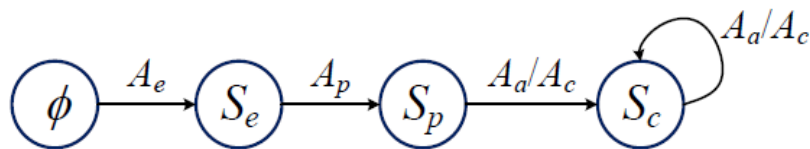
融合的方法 [Yih et al, ACL 2015 最佳论文提名]

- 大框架是符号的方法，把问句转换成查询图
- 在转换过程中，比如关键实体提取、查询路径的抽取以及其他限制条件和约束的抽取等，都用到了基于表示学习的匹配方式。
- 综合两方面优势，取得目前最好效果

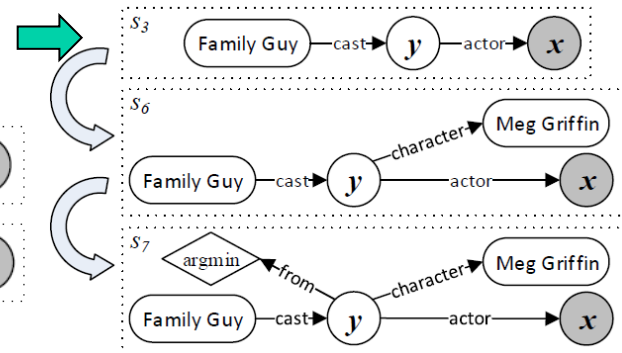
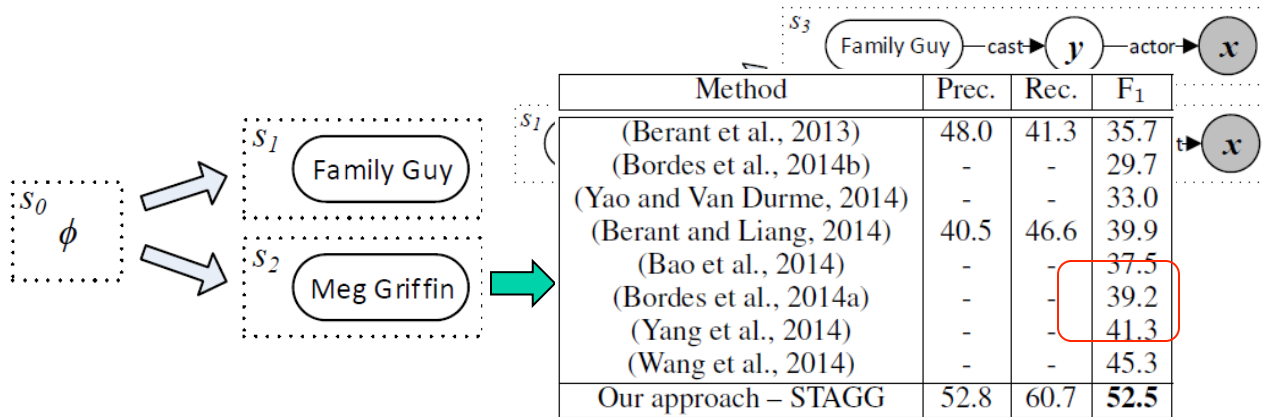
“Who first voiced Meg on Family Guy?”



Constraints & Aggregations



构造查询图，查询图扩展到正确形式



提纲

●深度问答中的知识表示

- 符号表示
- 分布表示

●关键技术

- 知识抽取
- 问句语义解析
- 推理规则学习

●资源融合

- 文本、抽取知识库与知识图谱

●展望

信息抽取

- ✓ 实体识别
- ✓ 关系识别
- ✓ 实体链接
- ✓ 事件识别

实体关系抽取

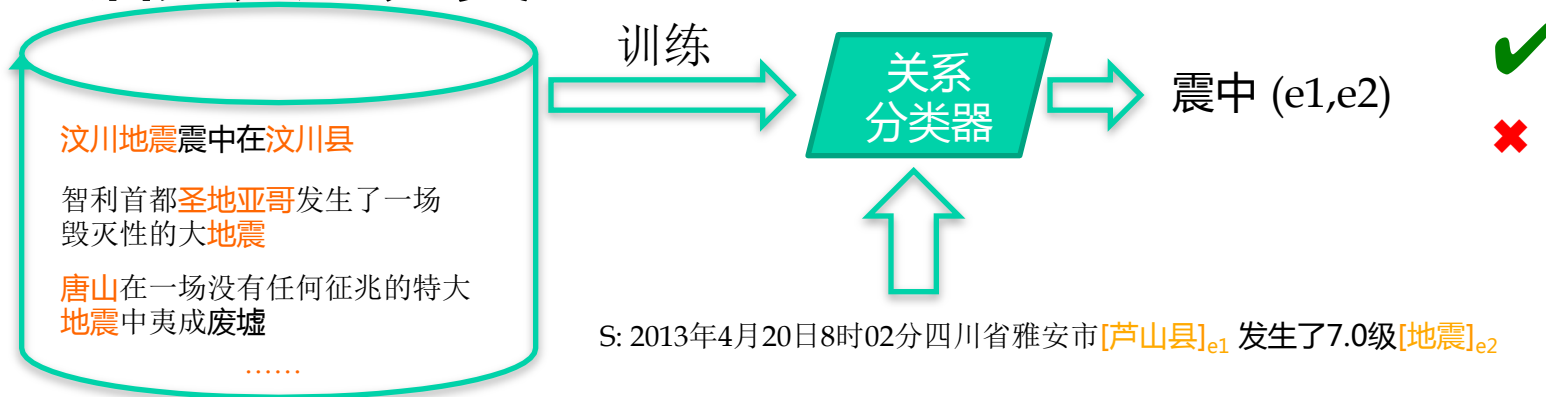
[Zeng et al, COLING 2014 Best Paper Award]
[Zeng et al, EMNLP2015]

●任务：判别句子中实体之间的语义关系

S: 2013年4月20日8时02分四川省雅安市[芦山县]_{e1} 发生了7.0级[地震]_{e2}

震中 (e1,e2)

●语义关系分类



难点：抽取实体间语义关系的文本特征表示

● 传统特征提取需要NLP预处理+人工设计的特征

2013年4月20日8时02分四川省雅安市[芦山县]_{e1}发生了7.0级[地震]_{e2}

震中 (e1,e2)

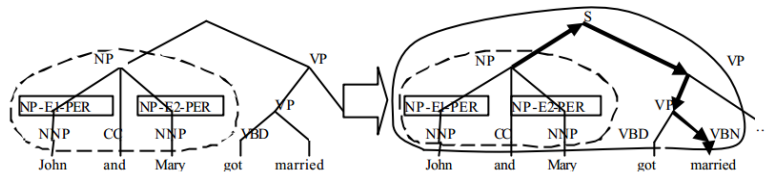
传统特征：

Words： 芦山县_{m11}, 地震_{b1}, 发生_{b2}, 在₂₁

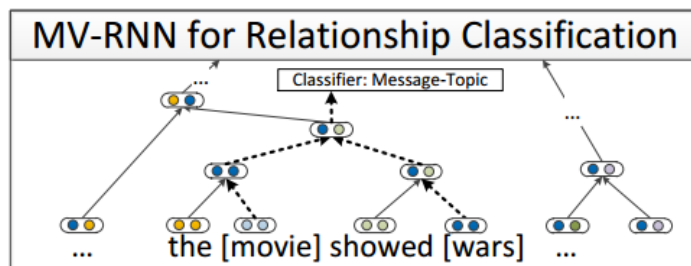
Entity Type： Noun_{m1}, Location_{m2}

Parse Tree： Location-VP-PP-Noun

Kernel Feature:



基于神经网络的特征抽取：



- 问题1：对于缺少NLP处理工具和资源的语言，无法提取文本特征
- 问题2：NLP处理工具引入的“错误累积”
- 问题3：人工设计的特征不一定适合当前任务

解决方法

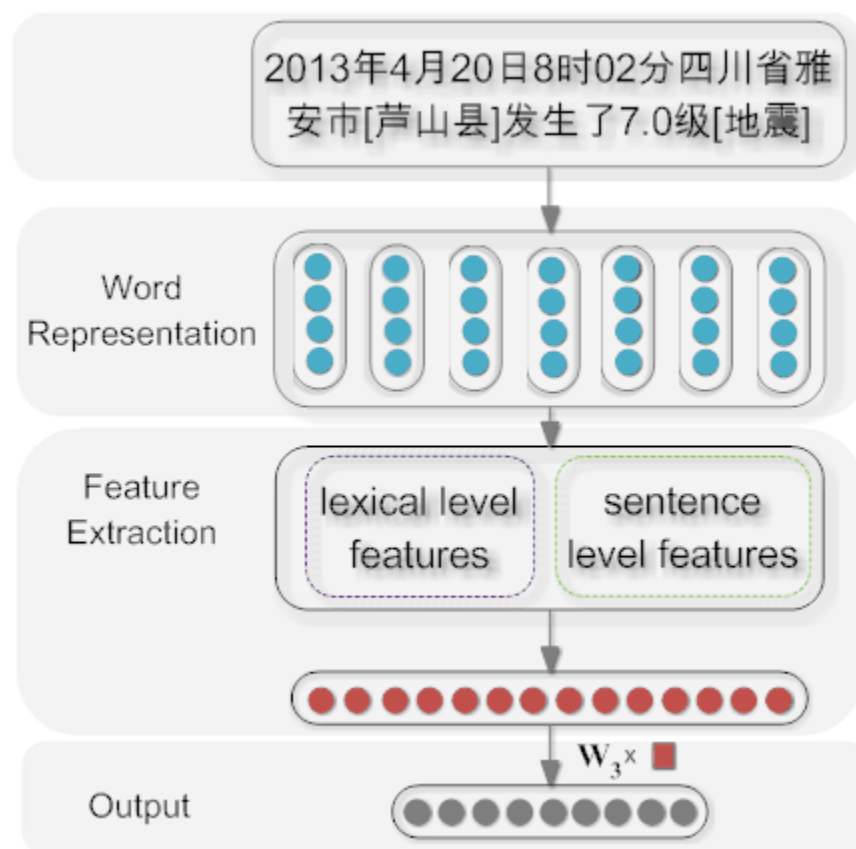
● 基于卷积神经网络的实体关系分类方法

- 通过CNN学习文本语义特征
- 不需要人工设计特征

通过Word Embeddings挖掘词汇的语义表示

Lexical Level Features: 实体本身的语义特征

Sentence Level Features: 通过CNN网络挖掘句子级别的文本特征



实验结果

Classifier	Feature Sets	F1
SVM	POS, stemming, syntactic patterns	60.1
	+ WordNet	74.8
	+ TextRunner, FrameNet,...	82.2
MVRNN	-	79.1
Ours	-	80.6

	特征集合	F1
state-of-the-art	POS, Prefixes, Levin classed morphological, WordNet, FrameNet, dependency parse, NomLex-Plus, TextRunner, ProBank, Google n-gram	82.2
深度卷积网络	上下文	80.6
深度卷积网络 + 专家知识	上下文+WordNet	82.7

SemEval-2010 Task 8 英文数据集

面向大规模知识库的关系抽取

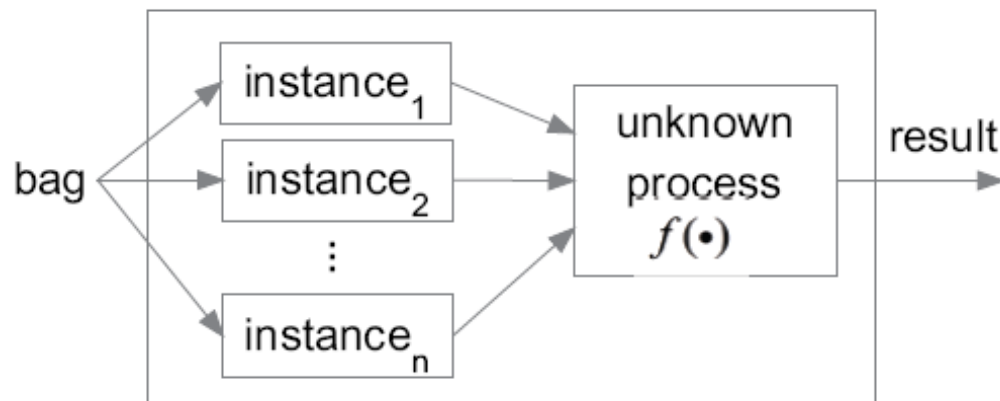
● 实体类别和关系类别巨大

- Freebase : 4000多万实体 , 上万个属性关系 , 24多亿个事实三元组
- DBpedia : 400多万实体 , 48,293种属性关系 , 10亿个事实三元组
- NELL : 519万实体 , 306种关系 , 5亿候选三元组
- Knowledge Vault: 4500万实体 , 4469种关系 , 2.7亿三元组

● 人工标注训练集不可行 , 需要寻求无指导或弱指导的关系抽取方法

基于分段卷积神经网络的多示例弱监督关系抽取

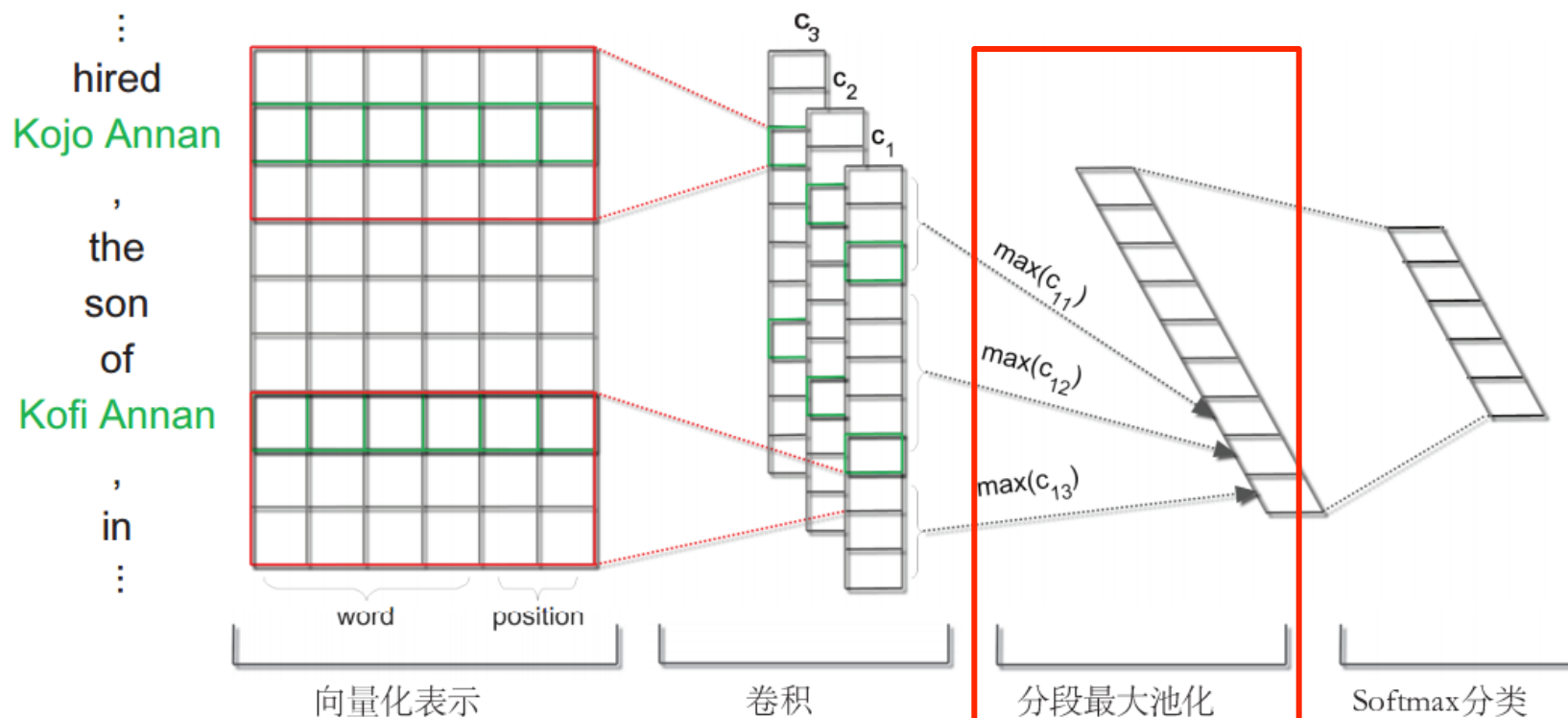
- 将弱监督关系抽取看作是多示例问题：
 - Steve Jobs was the **co-founder** and CEO of Apple and formerly Pixar.
 - Steve Jobs **passed away** the day before Apple unveiled iPhone 4S in late 2011.
 - The history of Microsoft began on April 4, 1975, when it was **founded** by Bill Gates and Paul Allen in Albuquerque.



基于分段卷积神经网络的多示例弱监督关系抽取

- 利用分段卷积网络自动学习特征

- 设计分段最大池化层，保留句子中结构化信息



实验

●数据集：Riedel等开发

- 知识库：Freebase
- 文本：纽约时报 (NYT)

●评价方法

- Held-out 评价：以Freebase 中存在关系的三元组作为标准
- 人工评价：去掉已经在Freebase 中存在的实体对，人工标注top N 结果

实验 (1)

- Word2vec训练参数

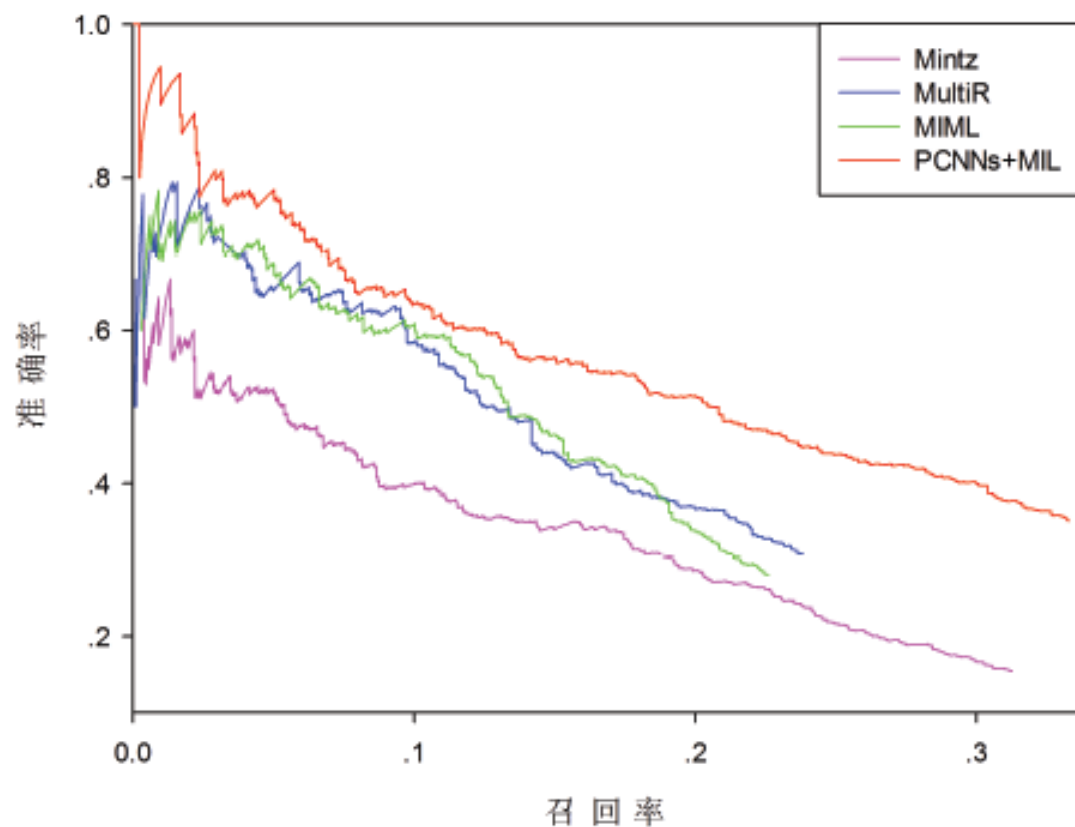
参数	hs	iter	size	binary	window	sample	negative	min-count
值	0	3	50	0	5	1e-4	5	0

- 模型参数确定

卷积窗长度	$w = 3$	} 网格搜索
滤波器个数	$n = 230$	
位置向量维数	$d_p = 5$	
词向量维数	$d_w = 50$	
Mini-batch大小	$b_s = 50$	
Adadelta参数	$\rho = 0.95, \varepsilon = 1e^{-6}$	
Dropout概率	$p = 0.5$	

实验 (2)

- Held out实验结果



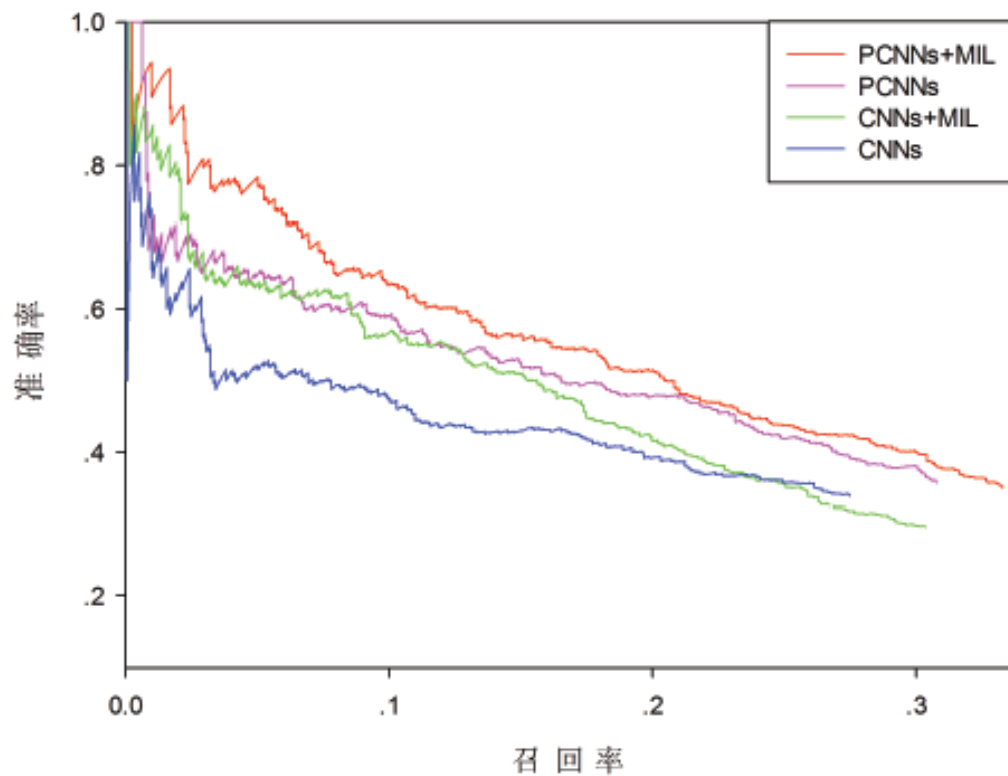
实验 (2)

- 人工评价实验结果

Top N	Mintz	MultiR	MIML	PCNNs+MIL
Top 100	0.77	0.83	0.85	0.86
Top 200	0.71	0.74	0.75	0.80
Top 500	0.55	0.59	0.61	0.69
Average	0.676	0.720	0.737	0.783

实验 (3)

- 分段最大池化和多示例学习的效果



问句语义解析

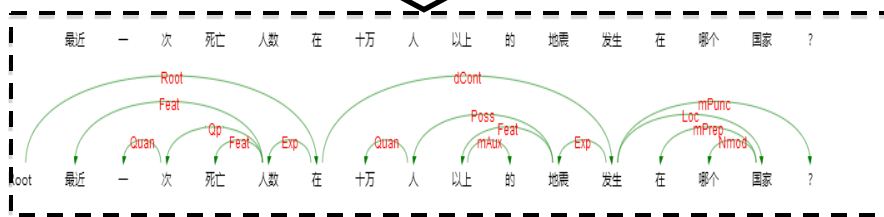
Semantic Parsing

自然语言问句的语义解析 [He et al, EMNLP 2014]

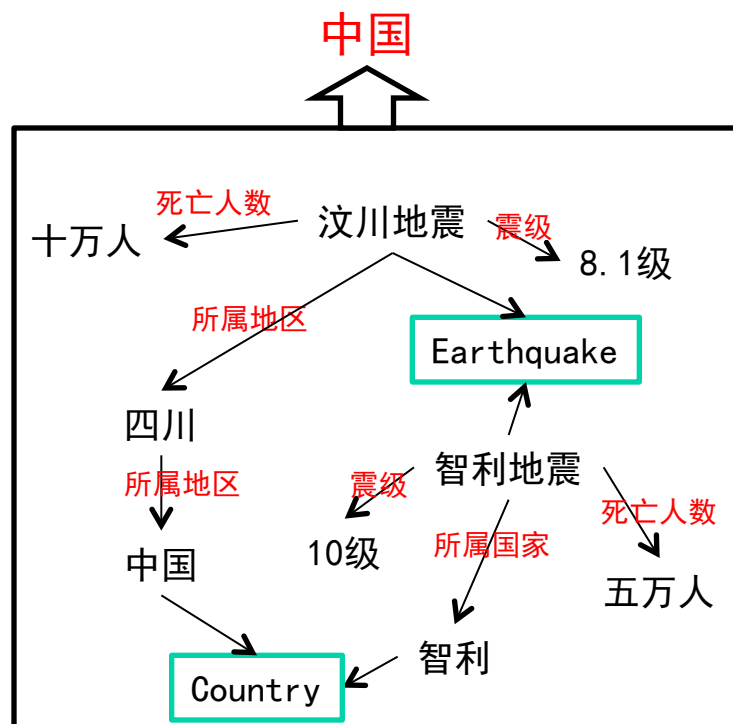
- 将自然语言问句转换为结构化查询语句（如SPARQL语句）

最近一次死亡人数在十万人以上的地震发生在哪个国家？

语义解析



```
SELECT DISTINCT ? X
WHERE {
  ?y 所属国家 ?x; ?y 死亡人数 ?p;
  ?y 时间 ?t; argmin(?t-now);
  max(?t-now, 0); max(?p, 十万);
}
```



自然语言问句到结构化查询语句的映射

- 基本步骤

最近一次死亡人数在十万人以上的地震发生在哪个国家？

短语切分：

十万人

地震

死亡人数

发生

国家

资源映射：

dbo:十万人

dbr:死亡人数

dbo:Earthquake

dbr:所属地区

dbo:Country

语义组合：

$\langle ?y, \text{dbr:死亡人数}, \text{dbo:十万人} \rangle$

$\langle ?y, \text{dbr:所属地区}, \text{dbo:?x} \rangle$

$\langle ?y, \text{dbr:isa}, \text{dbo:地震} \rangle$

$\langle ?x, \text{dbr:isa}, \text{dbo:国家} \rangle$

```
SELECT DISTINCT ? X
WHERE {
  ?y 所属国家 ?x; ?y 死亡人数 ?p;
  ?y 时间 ?t; argmin(?t-now);
  max(?t-now, 0); max(?p, 十万);
}
```

传统方法：规则和模板驱动的方法

- 语义词典

Words	Category
	Syntax : Semantics
Texas	NP : texas
borders	(S\NP)/NP : $\lambda x. \lambda y. \text{borders}(y, x)$
Kansas	NP : kansas
Kansas city	NP : kansas

- 组合规则

$$X/Y : f \quad Y : g \Rightarrow X : f(g) \quad (>)$$

$$Y : g \quad X \backslash Y : f \Rightarrow X : f(g) \quad (<)$$

$$X/Y : f \quad Y/Z : g \Rightarrow X/Z : \lambda x. f(g(x)) \quad (> \mathbf{B})$$

$$Y \backslash Z : g \quad X \backslash Y : f \Rightarrow X \backslash Z : \lambda x. f(g(x)) \quad (< \mathbf{B})$$

$$\begin{array}{c}
 \text{New York} \quad \text{borders} \quad \text{Vermont} \\
 \hline
 \text{NP} \quad (S \backslash NP) / NP \quad \text{NP} \\
 \text{ny} \quad \lambda x \lambda y. \text{next_to}(y, x) \quad \text{vt} \\
 \hline
 (S \backslash NP) \\
 \lambda y. \text{next_to}(y, \text{vt}) \\
 \hline
 S \\
 \text{next_to}(\text{ny}, \text{vt})
 \end{array}$$

$$\begin{array}{c}
 \text{what} \quad \text{states} \quad \text{border} \quad \text{texas} \\
 \hline
 S / (S | NP) \quad S | NP / (S | NP) \quad S \backslash NP / NP \quad \text{NP} \\
 \lambda f \lambda x. f(x) \quad \lambda f \lambda x. \text{state}(x) \wedge f(x) \quad \lambda y \lambda x. \text{next_to}(x, y) \quad \text{tex} \\
 \hline
 S | NP / NP \\
 \lambda y \lambda x. \text{state}(x) \wedge \text{next_to}(x, y) \\
 \hline
 S | NP \\
 \lambda x. \text{state}(x) \wedge \text{next_to}(x, \text{tex}) \\
 \hline
 S \\
 \lambda x. \text{state}(x) \wedge \text{next_to}(x, \text{tex})
 \end{array}$$

存在的问题

- 传统方法

- 无法处理大规模链接数据上的真实问句

- 自动方法

- 模板自学习
 - 需要集成实体链接、关系链接、语义资源构造、逻辑表达式生成等多个技术

难点：歧义严重

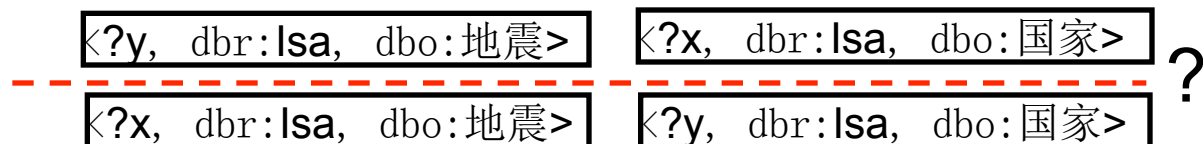
- 短语切分

- 死亡？死亡人数？

- 资源映射

- 地震→歌曲名
- 地震→事件类别
- 地震→人名

- 语义组合



- 传统Pipeline策略，逐步解决每一步的歧义问题
- 缺点：错误积累与传递
- 是否可以利用各步骤之间的相互影响进行联合消歧

基于马尔科夫逻辑网络的联合消歧

● 基于马尔科夫逻辑网络的多任务学习

- 针对每一个子任务的规则写成一阶逻辑公式
- 利用马尔科夫逻辑网络学习各个子任务间的相互影响

求解目标：

短语切分



资源映射

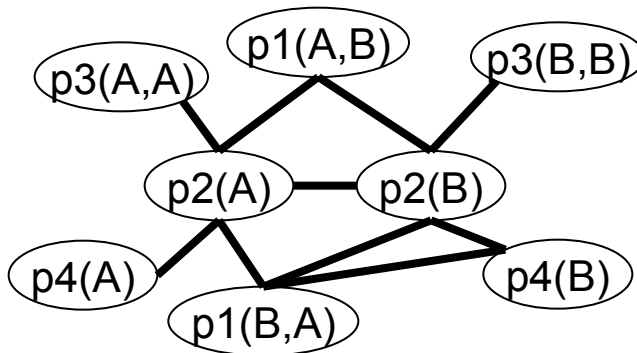


语义组合



<i>hasPhrase(i)</i>	The <i>i</i> th candidate phrase has been chosen
<i>hasResource(i, j)</i>	The <i>i</i> th phrase is mapped to the <i>j</i> th semantic item
<i>hasRelation (ri, rj, rr)</i>	The semantic item <i>ri</i> and <i>rj</i> can be grouped together with the relation type <i>rr</i>

$$p(y) = \frac{1}{Z} \exp\left(\sum_{(\phi_i, w_i) \in L} w_i \sum_{c \in C^{\phi_i}} f_c^{\phi_i}(y)\right)$$



逻辑规则

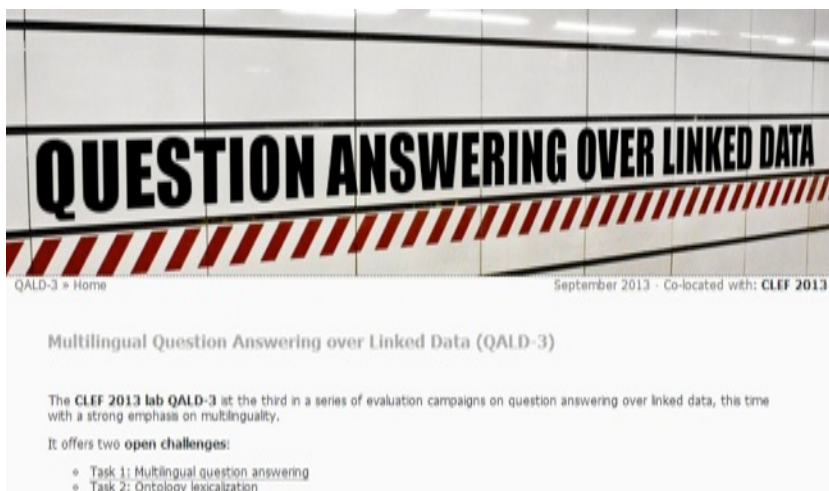
sf1	$\text{priorMatchScore}(p, r, s) \Rightarrow \text{hasPhrase}(p)$
sf2	$\text{priorMatchScore}(p, r, s) \Rightarrow \text{hasResource}(p)$
sf3	$\text{phrasePosTag}(p, \text{pt}+) \wedge \text{resourceType}(r, \text{rt}+) \Rightarrow \text{hasResource}(p, r)$
sf4	$\text{phraseDepTag}(p1, p2, \text{dp}+) \wedge \text{hasResource}(p1, r1) \wedge \text{hasResource}(p2, r2) \Rightarrow \text{hasRelation}(r1, r2, \text{rr}+)$
sf5	$\text{phraseDepTag}(p1, p2, \text{dp}+) \wedge \text{hasResource}(p1, r1) \wedge \text{hasResource}(p2, r2) \wedge \neg \text{hasMeanWord}(p1, p2) \Rightarrow \text{hasRelation}(r1, r2, \text{rr}+)$
sf6	$\text{phraseDepTag}(p1, p2, \text{dp}+) \wedge \text{hasResource}(p1, r1) \wedge \text{hasResource}(p2, r2) \wedge \text{phraseDepOne}(p1, p2) \Rightarrow \text{hasRelation}(r1, r2, \text{rr}+)$
sf7	$\text{hasRelated}(r1, r2) \wedge \text{hasResource}(_, r1) \wedge \text{hasResource}(_, r2) \Rightarrow \text{hasRelation}(r1, r2, _)$
sf8	$\text{hasQueryRelated}(r1, r2, r3) \Rightarrow \text{hasRelation}(r1, r2, \text{rr}1) \wedge \text{hasRelation}(r2, r3, \text{rr}2)$

短语p1与短语p2之间存在dp的句法关系，短语p1映射为资源r1，短语p2映射为资源r2，则：在本句子中，r1和r2之间存在rr的语义关系

- 利用Markov Network将解决多个任务的规则放在统一的学习框架下，有效地利用了多个规则间的相互影响进行联合消歧
- 利用Markov Network上的学习，解决了规则冲突问题

实验

● 国际问答技术评测QALD



Test set	System	#T	#Q	#A	P	R	F1
QALD-3	CASIA (He et al., 2013)	99	52	29	0.56	0.3	0.38
	Scalewelis (Joris and Ferré, 2013)	99	70	32	0.46	0.32	0.38
	RTV (Cristina et al., 2013)	99	55	30	0.55	0.3	0.39
	Intui2 (Corina, 2013)	99	99	28	0.28	28	0.28
	SWIP (Pradel et al., 2013)	99	21	15	0.71	0.15	0.25
	Ours	99	75	45	0.6	0.46	0.52
QALD-4 ²⁰	gAnswer	50	25	16	0.64	0.32	0.43
	Intui3	50	33	10	0.30	0.2	0.24
	ISOFT	50	50	10	0.2	0.2	0.2
	RO FII	50	50	6	0.12	0.12	0.12
	Ours	50	26	15	0.58	0.3	0.4

QALD3, QALD4

基于马尔科夫逻辑网的联合寻优，能够有效地克服自然语言问句到结构化查询语句转换过程中的错误累积问题

知识推理

- ✓ 链接预测（知识库补全）
 - ✓ 推理规则挖掘
 - ✓ 文本蕴含推理

知识推理

- 事实推理

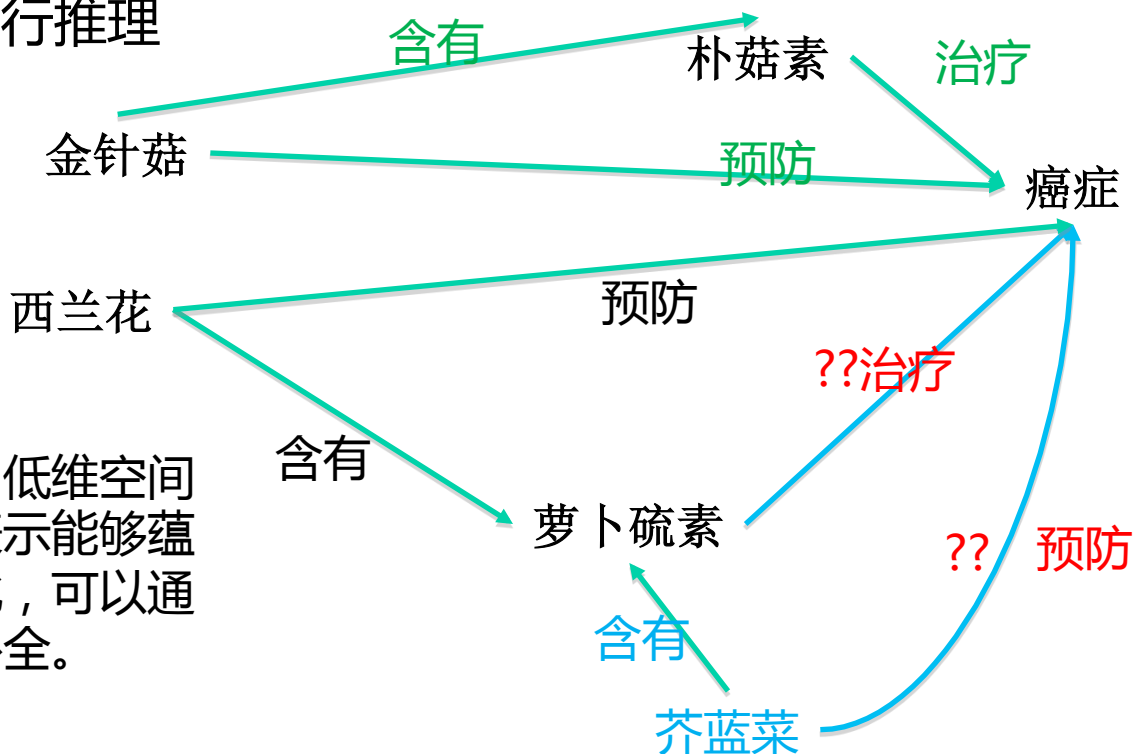
- 给定实体，判断它们间是否存在某关系
- 给定两个实体和关系，判断该事实是否成立

- 推理规则学习

- 问答中的查询推理需要判断关系或关系组合之间是否具有逻辑推理关系

链接预测

- 知识图谱本身就蕴涵着推理模式
- 利用现有知识对未知知识进行推理



把知识库中的实体和关系表示为低维空间的对象及其它们的操作，这种表示能够蕴涵它们在知识库中的特性，因此，可以通过推理对未知知识进行补全。

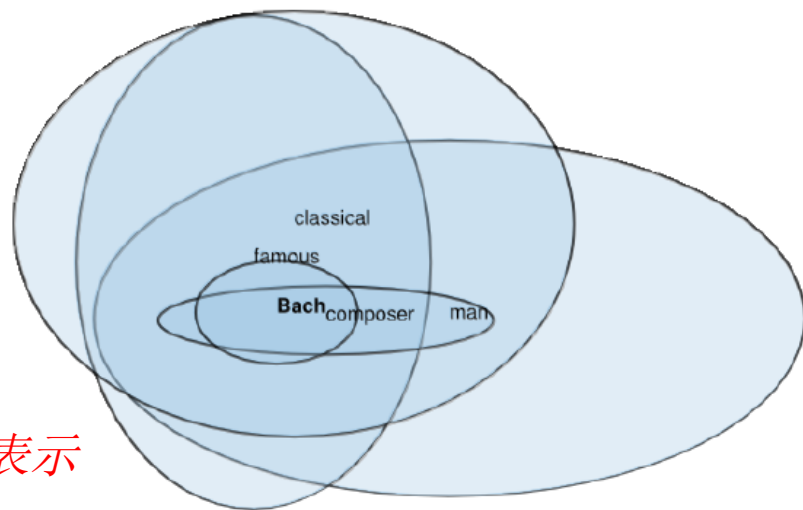
吃哪些蔬菜预防癌症？

方法1：基于高斯分布表示学习的补全方法

[He et al, CIKM 2015]

● 如何表示符号的不确定性

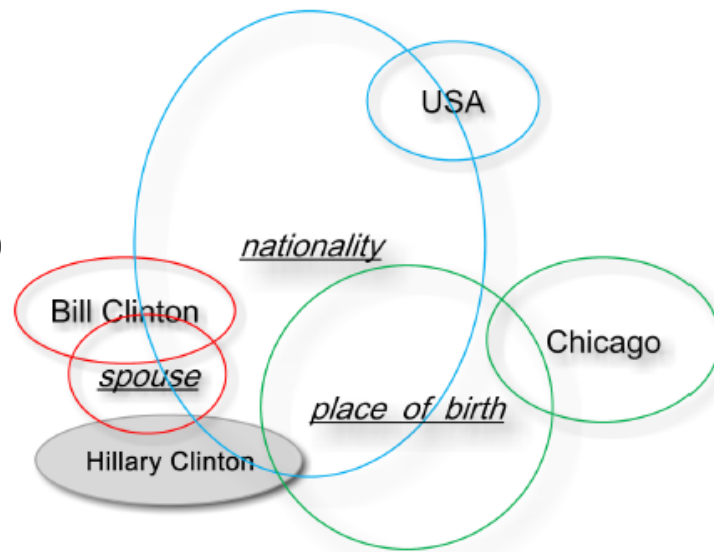
- 实体：粒度的大小（人 vs 总统）
- 关系：指示性强弱（配偶 vs 国籍）



基于密度的表示

● 多维高斯分布表示符号

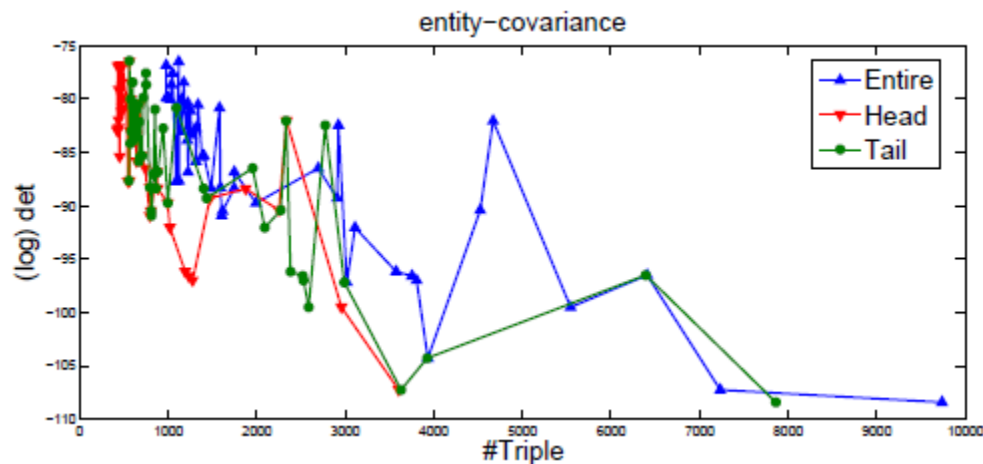
- 均值向量表示该符号的位置(含义)
- 协方差矩阵表示该符号的多样性(不确定性)



实验效果：协方差表示含义不确定性

- 实体不确定性

- 包含事实越多，
含义不确定性越小



- 关系不确定性

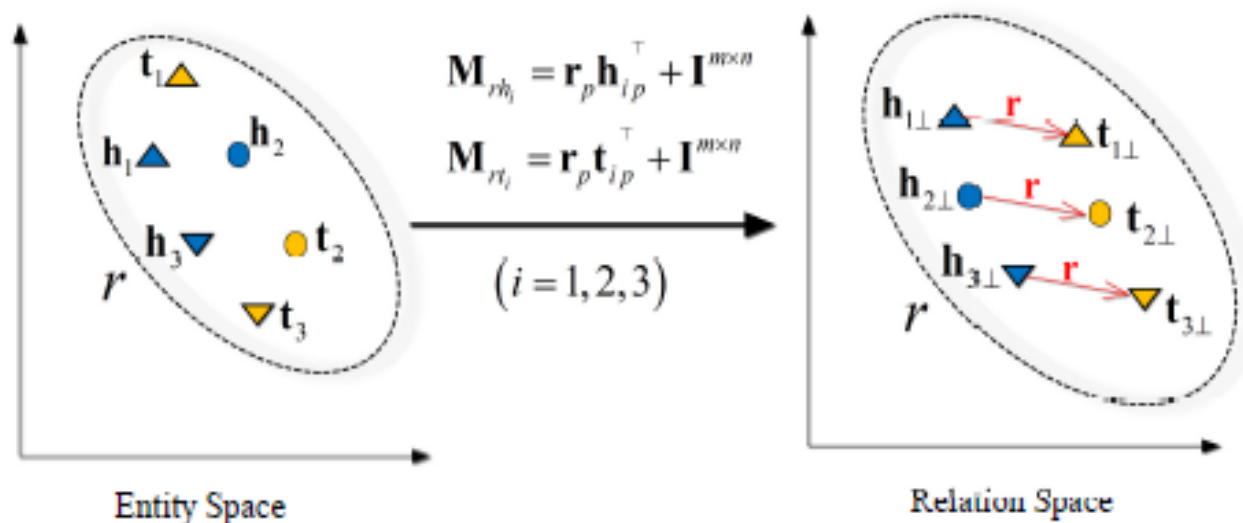
- 关系越复杂，
含义不确定性越大

Relaion	#Triplet	#Head	#Tail	Type	(log) det	trace
nationality	4198	3755	100	m-1	-4.77	67.90
place_lived	3740	2441	784	m-n	-23.02	53.24
profession	11636	4145	152	m-n	-57.45	23.10
gender	3721	3721	2	m-1	-59.53	21.35
place_of_birth	2468	2468	685	m-1	-63.42	19.41
ethnicity	2030	1610	78	m-1	-69.95	15.00
major	260	217	60	m-1	-69.59	14.62
sibling	131	111	113	1-1	-72.98	14.29
religion	1086	963	45	m-1	-75.07	12.98
spouse	427	395	385	1-1	-77.77	12.24
children	77	69	71	1-1	-76.94	12.14
parents	83	74	76	1-1	-77.20	12.03

方法2：基于动态映射矩阵的补全方法 [Ji et al, ACL 2015]

- 传统方法：实体或关系的表示与具体的事实无关，无法表示知识的多样性
- 我们的方法：通过对事实中的实体和关系构造**动态映射矩阵**，表示知识的多样性

TransD: 主要解决“一对多、多对一、多对多”类型的关系



(毛泽东, 儿子, 毛岸青)

M_1 M_2

毛泽东^{儿子} + 儿子 = 毛岸青^{儿子}

毛泽东^{儿子} + 儿子 = 毛岸英^{儿子}

M_1 M_3

(毛泽东, 儿子, 毛岸英)

链接预测

● 在WN18和FB15k上实验

Data sets	WN18				FB15K			
Metric	Mean Rank		Hits@10		Mean Rank		Hits@10	
	Raw	Filt	Raw	Filt	Raw	Filt	Raw	Filt
Unstructured (Bordes et al. 2012)	315	304	35.3	38.2	1,074	979	4.5	6.3
RESCAL (Nickel, Tresp, and Kriegel 2011)	1,180	1,163	37.2	52.8	828	683	28.4	44.1
SE (Bordes et al. 2011)	1,011	985	68.5	80.5	273	162	28.8	39.8
SME (linear) (Bordes et al.2012)	545	533	65.1	74.1	274	154	30.7	40.8
SME (Bilinear) (Bordes et al. 2012)	526	509	54.7	61.3	284	158	31.3	41.3
LFM (Jenatton et al. 2012)	469	456	71.4	81.6	283	164	26.0	33.1
TransE (Bordes et al. 2013)	263	251	75.4	89.2	243	125	34.9	47.1
TransH (unif) (Wang et al. 2014)	318	303	75.4	86.7	211	84	42.5	58.5
TransH (bern) (Wang et al. 2014)	401	388	73.0	82.3	212	87	45.7	64.4
TransR (unif) (Lin et al. 2015)	232	219	78.3	91.7	226	78	43.8	65.5
TransR (bern) (Lin et al. 2015)	238	225	79.8	92.0	198	77	48.2	68.7
CTransR (unif) (Lin et al. 2015)	243	230	78.9	92.3	233	82	44.0	66.3
CTransR (bern) (Lin et al. 2015)	231	218	79.4	92.3	199	75	48.4	70.2
TransD (unif)	242	229	79.2	92.5	211	67	49.4	74.2
TransD (bern)	224	212	79.6	92.2	194	91	53.4	77.3

● FB15k上根据不同属性的关系进行预测

Tasks	Prediction Head (Hits@10)				Prediction Tail (Hits@10)			
Relation Category	1-to-1	1-to-N	N-to-1	N-to-N	1-to-1	1-to-N	N-to-1	N-to-N
Unstructured (Bordes et al. 2012)	34.5	2.5	6.1	6.6	34.3	4.2	1.9	6.6
SE (Bordes et al. 2011)	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3
SME (linear) (Bordes et al.2012)	35.1	53.7	19.0	40.3	32.7	14.9	61.6	43.3
SME (Bilinear) (Bordes et al. 2012)	30.9	69.6	19.9	38.6	28.2	13.1	76.0	41.8
TransE (Bordes et al. 2013)	43.7	65.7	18.2	47.2	43.7	19.7	66.7	50.0
TransH (unif) (Wang et al. 2014)	66.7	81.7	30.2	57.4	63.7	30.1	83.2	60.8
TransH (bern) (Wang et al. 2014)	66.8	87.6	28.7	64.5	65.5	39.8	83.3	67.2
TransR (unif) (Lin et al. 2015)	76.9	77.9	38.1	66.9	76.2	38.4	76.2	69.1
TransR (bern) (Lin et al. 2015)	78.8	89.2	34.1	69.2	79.2	37.4	90.4	72.1
CTransR (unif) (Lin et al. 2015)	78.6	77.8	36.4	68.0	77.4	37.8	78.0	70.3
CTransR (bern) (Lin et al. 2015)	81.5	89.0	34.7	71.2	80.8	38.6	90.1	73.8
TransD (unif)	80.7	85.8	47.1	75.6	80.0	54.5	80.7	77.9
TransD (bern)	86.1	95.5	39.8	78.5	85.4	50.6	94.4	81.2

推理规则的学习

- 哪些蔬菜可以预防骨质疏松？
 - 西兰花含钙，钙能有效预防骨质疏松
 - 甘蓝菜含镁，镁能有效预防骨质疏松
 - 不能直接得到结果，需要知识推理

事实推理：推理出隐含的事实

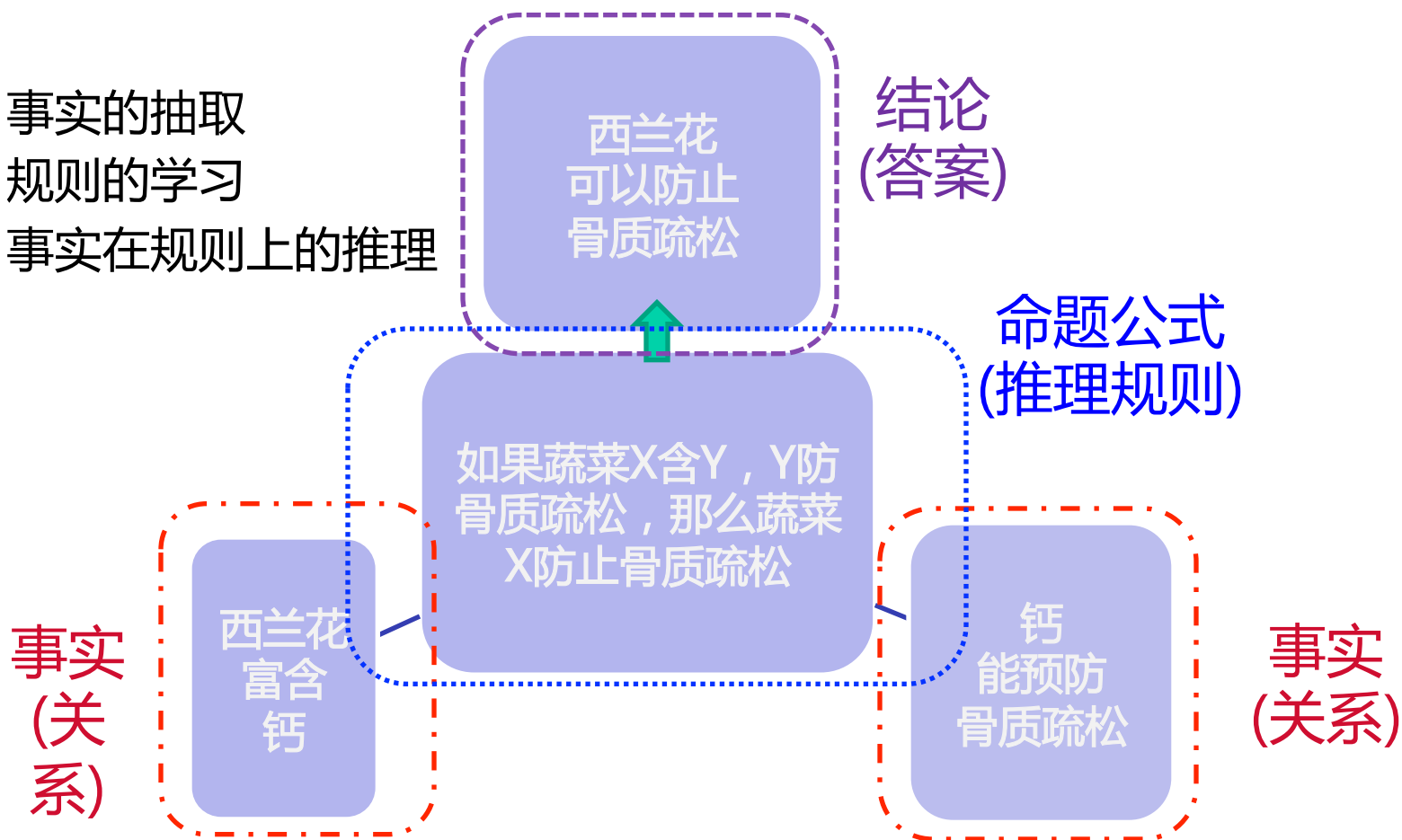
（西兰花可以防止骨质疏松）

推理规则学习：学习命题（关系）之间的逻辑推理关系

（含有能预防骨质疏松元素的蔬菜可以防止骨质疏松）

关键技术

- 事实的抽取
- 规则的学习
- 事实在规则上的推理



逻辑推理规则学习: 符号的方法

- 通过统计关系路径的共现情况学习霍恩子句表示的推理规则

[Schoenmackers et al, EMNLP 2010]

- 识别出代表性的类别和这些类别的实例(实体实例抽取)
- 发现类别间关系(关系发现)
- 发现推理规则, 计算每条推理规则的置信度 (穷举搜索所有满足条件的组合关系)
 $\text{Contains}(\text{food}, \text{chemical}) \vdash \text{IsMadeFrom}(\text{food}, \text{ingredient}) \wedge \text{Contains}(\text{ingredient}, \text{chemical});$

- 通过实体间在图中的链接特征学习关系分类器, 得到路径与关系的推理规则 [Lao et al, EMNLP 2011]

- 学习实体对属于该关系的分类器
- 提取关系的实体对在知识图谱中的链接路径作为特征
- 提取权重大的特征作为该关系的推理路径

athletePlaysForTeam

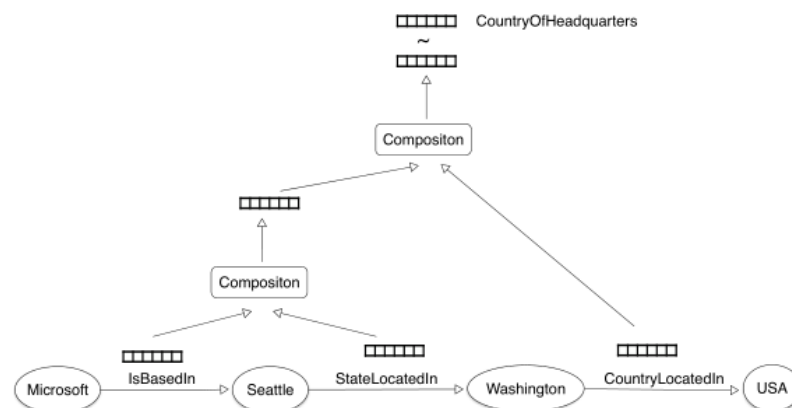
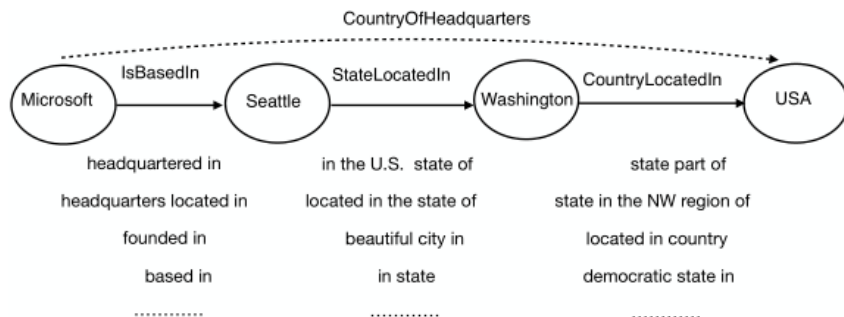
$W_1 \quad c \xrightarrow{\text{athletePlaysInLeague}} c \xrightarrow{\text{leaguePlayers}} c \xrightarrow{\text{athletePlaysForTeam}} c \text{ (teams with many players in the athlete's league)}$

$W_2 \quad c \xrightarrow{\text{athletePlaysInLeague}} c \xrightarrow{\text{leagueTeams}} c \xrightarrow{\text{teamAgainstTeam}} c \text{ (teams that play against many teams in the athlete's league)}$

逻辑推理规则学习: 分布表示方法

- 基于知识库图谱的表示学习挖掘关系路径间的蕴涵关系 [Yang et al, ICLR 2015]
 - 基于表示学习, 得到关系在低维数值空间中的表示 (向量或矩阵)
 - 通过对关系表示间的数值计算判断它们之间的推理关系
 - » $v(r1) + v(r2) \approx v(r3)$ 或 $m(r1) \cdot m(r2) \approx m(r3)$ 得到 $r1 \wedge r2 \Rightarrow r3$
- 组合路径语义学习推理规则 [Neelakantan et al, ACL 2015]

- 关系用向量表示
- 路径通过组合方式 (RNN) 进行表示



提纲

●深度问答中的知识表示

- 符号表示
- 分布表示

●关键技术

- 问句语义解析
- 知识抽取
- 推理规则学习

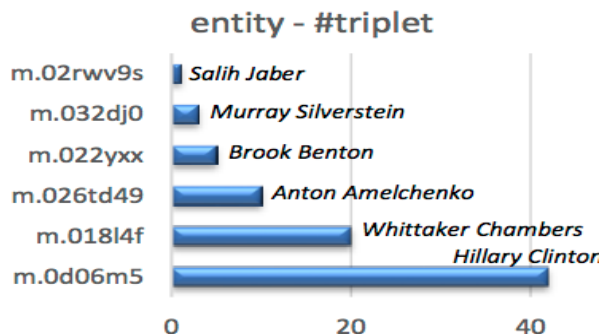
●资源融合

- 文本、抽取知识库与知识图谱

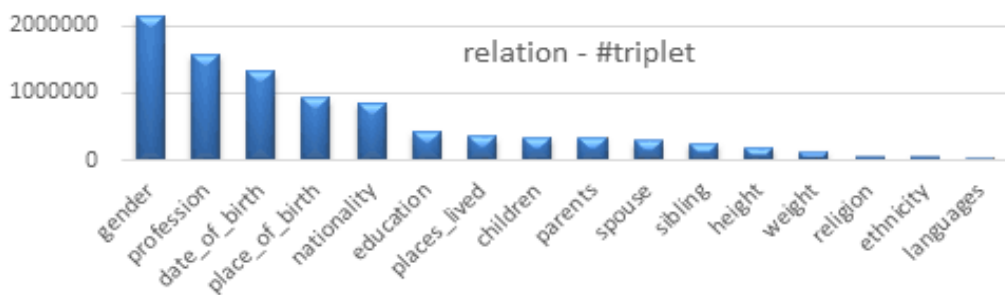
●展望

知识图谱中事实的稀疏性

- 事实的不完备



- 关系的不平衡性



- 关系、事件抽取（利用文本）
- 链接预测、知识补全（利用现有知识库）
- 多知识融合（多个知识库相互补充相互验证）

大量的知识存在于文本中（特别是最新的）

Open IE能抽取大量的事实

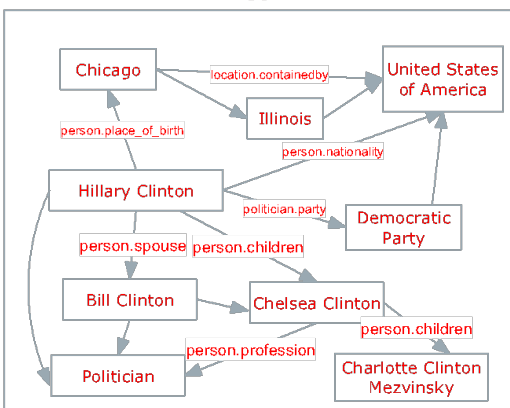
知识图谱包含规范的知识

不能回答的问题

- 条目缺失：新实体新事件，知识图谱的更新
 - “31日失事的俄罗斯客机多少人遇难？”
- 事实缺失：信息抽取，知识补全
 - “湖北有那些互联网企业？”
- 知识框架（schema）缺失：框架抽取和对齐
 - “奥巴马的爷爷、岳父是谁？”

多资源融合

KGs



Textual Triples

(Chicago, be most city in, the United States)
 (Charlotte, grandpa, Bill Clinton)
 (Chicago, locate in, Illinois)
 (Hillary, be politician in, the democratic party)
 (Edward H. Allegetti, be born in, Chicago)
 ...

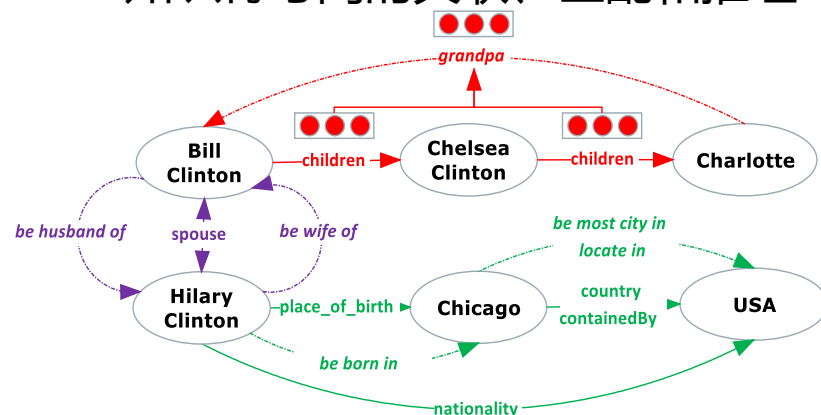
Jointly
Representations
Learning

Some of the United States' tallest towers are located in Chicago.
 Edward H. Allegetti, age 67, a resident of Naperville, IL for 26 years, passed away June 14, 2015.
 He was born July 15, 1947 in Chicago.
 Naperville is a suburb of Chicago, located in DuPage and Will counties in the State of Illinois.
 Chicago is the third most populous city in the United States.
 Bill Clinton has opened up to ... the joys of being a grandpa to Chelsea Clinton's daughter Charlotte.
 Chicago is located in northeastern Illinois on the southwestern shores of Lake Michigan.
 Hillary (born October 26, 1947) is an American politician in the Democratic Party.
 ...

Texts

1. 利用文本和抽取事实作为上下文判断实体间关系，弥补知识稀疏问题；
2. 利用知识图谱和抽取事实作为上下文判断文本相似度，弥补文本语义不明确的问题；

通过联合学习，不仅能够进行知识补充，还可以学习语言符号与知识符号间的关联、匹配和推理



基于表示学习对知识图谱 65 抽取事实和文本统一学习

表示学习框架下知识库、抽取事实和文本一体化建模

- 国家自然科学基金重点项目

- 大规模知识关联和文本语义计算方法及应用验证
(中科院自动化所, 清华, 百度)

- Google Research Award

- Representing and Reasoning Knowledge by Jointly Learning on Knowledge Graphs, Textual Triples and Unstructured Texts

提纲

●深度问答中的知识表示

- 符号表示
- 分布表示

●关键技术

- 问句语义解析
- 知识抽取
- 推理规则学习

●资源融合

- 文本、抽取知识库与知识图谱

●展望

知识图谱

- 知识抽取和补充：新实体、新框架、新事实
- 多知识的融合、文本与知识的融合
- 知识解释：实体解释、关系解释
- 知识图谱中事件的表示和应用
- 知识利用：搜索、推荐、问答...

问答系统

- 限定域领域知识图谱 → 开放域大规模知识图谱
- 单知识 → 多元异构知识
- 单关系等简单问题 → 复杂问题
- 强监督训练 → 弱监督训练
- 单回合问答 → 多回合问答（对话）
- 浅层理解(Simple Factoid QA) → 深度理解(Reading Comprehension)
- 流水线、特征工程 → 深层网络、端到端系统

Selected Publications

● 知识抽取

- Shulin Liu, Kang Liu and Jun Zhao, A Probabilistic Soft Logic Based Approach to Exploit Latent and Global Information in Event Classification, AAAI-2016
- Yubo Chen, Liheng Xu, Kang Liu and Jun Zhao, Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks, Accepted by ACL-2015.
- Daojian Zeng, Kang Liu, Yubo Chen, Jun Zhao, Distant Supervision for Relation Extraction via Sparse Representation, EMNLP-2015.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou and Jun Zhao, Relation Classification via Convolutional Deep Neural Network, COLING-2014. (Best Paper Award)

● 表示学习

- Guoliang Ji, Shizhu He, Kang Liu and Jun Zhao, Knowledge Graph Completion with Adaptive Sparse Transfer Matrix, AAAI 2016
- Shizhu He, Kang Liu, Jun Zhao, Learning to Represent Knowledge Graphs with Gaussian Embedding, CIKM 2015.
- Ji Guoliang, He Shizhu, Xu Liheng, Liu Kang and Zhao Jun, Knowledge Graph Embedding via Dynamic Mapping Matrix, ACL 2015.
- Zhuoyu Wei, Jun Zhao, Kang Liu, Zhenyu Qi, et al. Large-scale Knowledge Base Completion: Inferring via Grounding Network Sampling over Selected Instances, CIKM 2015.
- Siwei Lai, Liheng Xu, Kang Liu and Jun Zhao and Jun Zhao, Recurrent Convolutional Neural Networks for Text Classification, AAAI 2015.

● 问句解析

- Yuanzhe Zhang, Shizhu He, Kang Liu and Jun Zhao, A Joint Model for Question Answering over Multiple Knowledge Bases, AAAI-2016.
- Shizhu he, Kang Liu, Yuanzhe Zhang and Jun Zhao, Question Answering over Linked Data Using First Order Logic, EMNLP-2014.

谢谢，敬请指正

jzhao@nlpr.ia.ac.cn