前沿技术讲习班

Advanced Technology Tutorial

第四期：深度学习与自然语言处理

深度学习与知识图谱

刘知远
清华大学
liuzy@tsinghua.edu.cn

韩先培
中科院软件所
xianpei@nfs.iscas.ac.cn

前沿技术讲习班
Advanced Technology Tutorial

面向知识图谱构建的语义关系抽取

韩先培
xianpei@nfs.iscas.ac.cn
中国科学院软件研究所

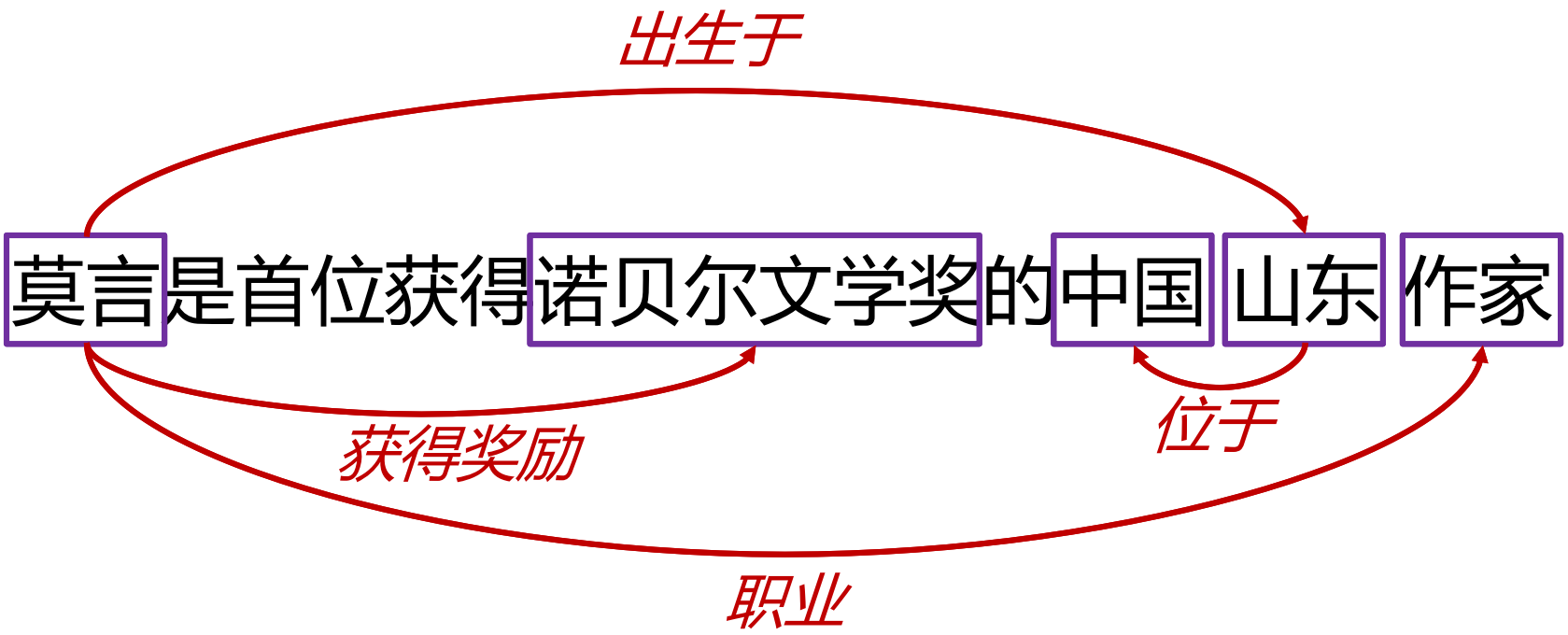
语义关系

- 语义关系**描述实体及概念之间的关联与交互**，是人类知识和知识图谱的核心组成部分
  - 首都（北京，中国）
  - ISA（中书省，唐代中央行政机构）
  - 获得奖励（莫言，诺贝尔文学奖）...

- 语义关系来自于
  - 日常生活中积累的对世界的认识和经验
    - ✓ *医生救治病人，老师教育学生...*
  - 来自于特定信息（文本、视频...）的表述
    - ✓ *欧冠-马竞2-0巴萨进四强 拜仁客场2-2晋级*

语义关系抽取

- 识别文本中实体和实体之间的语义关系并将其分类
  - 输入：自然语言文本语料库，Web，...
  - 输出：文本中包含的特定语义关系



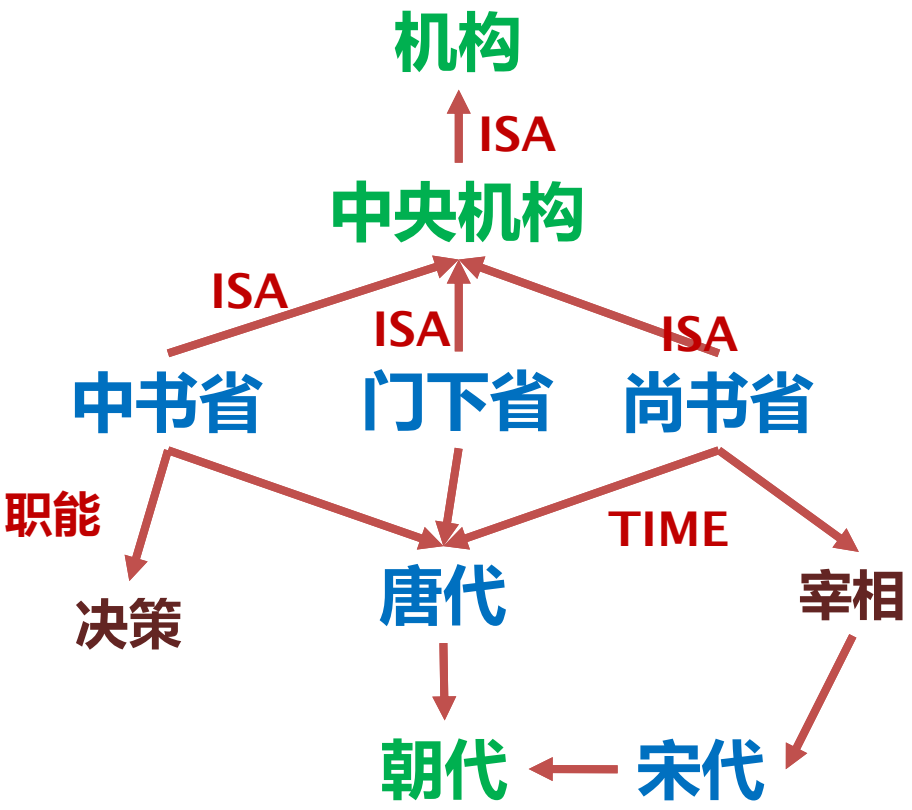
为什么需要语义关系抽取？

- 文本语义理解和智能信息服务的基础技术
- 知识图谱的核心知识获取技术

海量文本

唐朝中央的三省中书、门下和尚书，分别负责决策、审议和执行。三省的长官都是宰相，相权分散。

结构化知识网络



关系抽取方法

- **有监督方法**
  - **核心：表示模型**
  - 基于特征的分类，基于Kernel的分类，基于深度学习的分类…
  - 优点：性能更好
  - 缺点：需要大量标注语料，需要预先确定抽取的目标，领域自适应性不强

- **无/弱监督方法**
  - **核心：学习算法/策略**
  - Bootstrapping, 远距离监督，无监督（聚类），自学习技术…
  - 优点：可扩展，适合开放/Web环境下的信息抽取
  - 缺点：性能比有监督方法有差距

有监督语义关系抽取

如何构建有监督关系抽取模型

- **1. 选择要抽取哪些关系**
  - CEO，创始人，总部位置，收购，...
- **2. 标注关系实例**
  - *[乔布斯]$_{arg1}$是[苹果公司]$_{arg2}$的CEO* → CEO
  - *[华为]$_{arg1}$总部位于[深圳]$_{arg2}$* → 总部位置
- **3. 选择一种关系实例表示**
  - 特征向量，句法树（依存路径），Embedding...
- **4. 找一个已有分类模型并训练**
  - SVM, MaxEnt, Deep Neural Network...
- **5. 将模型应用于测试集，调试并评估性能**

核心问题

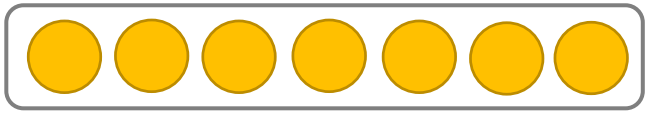
语义关系抽取的表示问题

- 将待抽取的关系实例表示为适合机器学习算法输入的形式

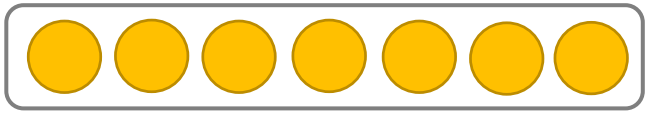
关系实例

机器学习算法输入：特征表示

<Arg1>莫言</Arg1>是首位获得<Arg2>诺贝尔文学奖</Arg2>的中国山东作家



<Arg1>莫言</Arg1>是首位获得诺贝尔文学奖的中国山东<Arg2>作家</Arg2>



语义关系实例表示的挑战

- **结构性**：关系实例中的词语必须按照特定的结构组合才能表达特定语义关系（词序列+内在语法结构）
  - 小明的哥哥是小强 V.S. 小强的哥哥是小明
  - 词袋子(BoW)模型无法使用
- **稀疏性**：语义关系通常只用很少的几个词语表达
  - 莫言，山东高密人
  - 相比之下，文档分类任务中的文档通常有几百到几千词
- **多样性**：关系表达多种多样
  - 莫言获得了诺贝尔奖，莫言领取了诺贝尔奖，诺贝尔奖得主莫言
- **外部性**：表示通常还依赖于外部知识
  - 中国山东高密：中国是国家，山东是省，高密是县

语义关系实例表示的三种策略

- **人工特征模板**：人工构建启发式特征
- **结构化核方法**：直接使用语义关系实例的原有结构化表示（词序列、依存路径、句法树），核心是构建可分类结构化对象的学习算法
- **表示学习方法**：基于深度神经网络，将结构化的关系表示转换为低维连续向量

人工特征模板

- 人工定义一系列的特征，使用0-1向量来表示一个关系实例
  - 莫言是首位获得诺贝尔文学奖的 --> [特征1， 特征2， ...]
- 实体特征：用来捕捉并表示关系论元的语义
  - *E1_Word=莫言, E1_POS=NR, E1_Type=人物, E1=Subject_of_Sentence, ...*
  - *通常会使用额外的资源：BrownCluster，WordNet，...*
- 关系特征：用来捕捉两个论元之间的关联
  - 两个论元之间的词: *within_获得，within_是*
  - 两个论元周围特定窗口内的词: *Right_E2_的*
  - 词/词性序列：*2gram，3gram，...*
    - ✓ *E1_是，获得_E2*
  - ...

人工特征模板—常用特征

- **Words**：两个提及内包含的所有词，以及两个提及之间的所有词；
- **Entity Type**：两个提及的实体类型（PERSON，ORGANIZATION，LOCATION， FACILITY，GPE等）；
- **Mention Level**：两个提及分别是命名性、代词性还是名词性；
- **Overlap**：两个提及之间的词数、提及数，它们是在一个名词短语内、动词短语内、还是介词短语内；
- **词序列**：两个提及之间的词/POS序列；
- **Dependency**：在依存树中与每个提及依存的词、词性和Chunk标记；
- **Parse Tree**：在句法树中，连接两个提及的路径上的非终极结点标记序列。

人工特征模板-分析

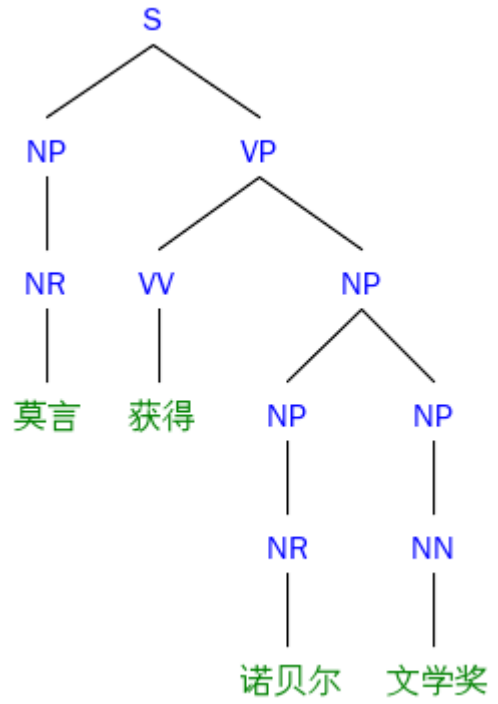
- **性能：**
  - 在ACE 2005数据集上，Jiang & Zhai (2007)获得的性能是**P=0.715, R =0.694, F=0.715**
- **优点：**
  - 系统容易构建,可以无缝使用现有的各种分类模型
  - Debug容易，可解释性强
- **缺点：**
  - 需要人工定义大量特征模板（好坏受经验影响）
  - 面临严重的特征稀疏问题：在KBP数据集中，超过四百万个单独特征，出现次数超过5的只有60万个，大部分实例只有10-30个特征
  - 很难捕捉整体的结构信息

结构化核方法

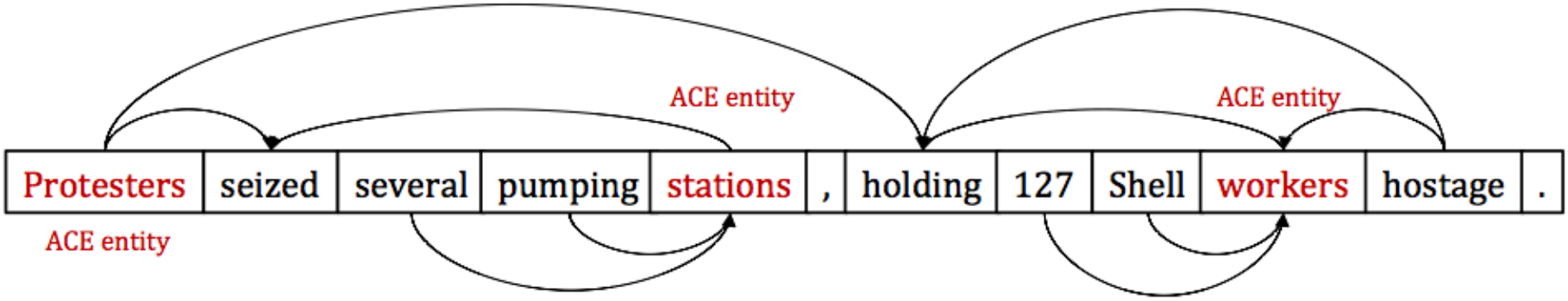
- 直接使用语法结构来表示关系实例
  - 最短依存路径，句法子树，语义增强句法子树
- 基于结构化关系表示，构建两个结构之间的相似度（Kernel），使用SVM来进行训练和预测
- 核心问题：
  - 如何表示复杂结构中结构化信息
  - 如何计算两个结构之间的现实度

结构化核方法-表示

- 最短依存路径（Bunescu and Mooney，2005）
- 句法子树（Zhang et al.，2006)
  - Minimum Complete Tree， Path-Enclosed Tree, Context-Sensitive Path Tree…
- 扩展：在句法结构中加入语义信息
  - feature-enriched/semantic tree kernel [Plank & Moschitti, 2013; Sun & Han, 2014]



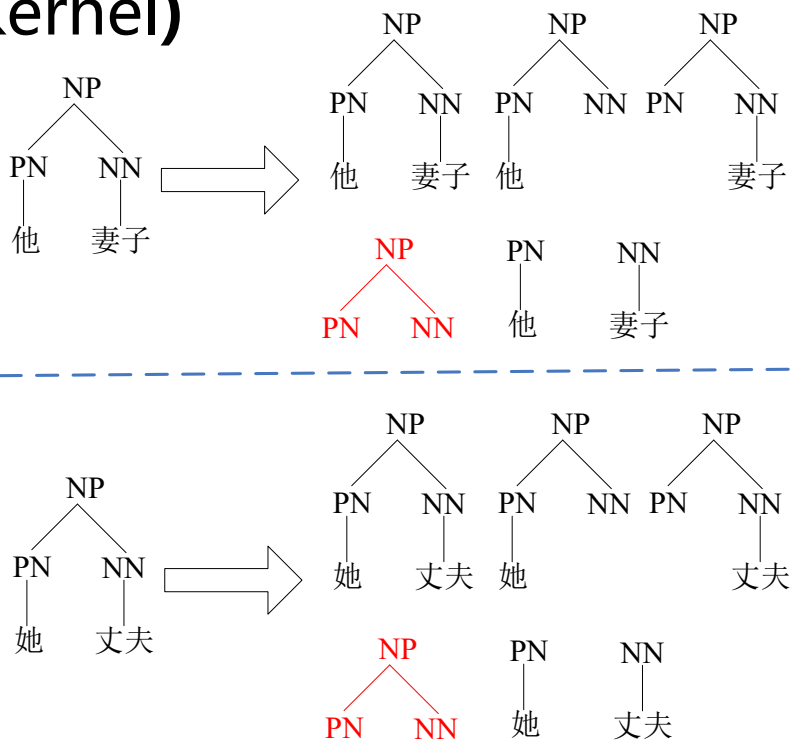
句法子树



最短依存路径

结构化核方法—Kernel

- **最短依存路径相似度**：Subsequence Kernel
- **句法子树相似度：卷积树核函数(**CTK, Convolution Tree Kernel)



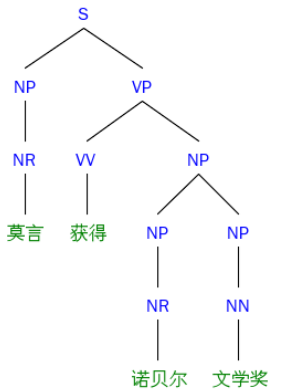
两棵句法树的相似度为相似子树的个数（CTK）：
由于两棵树在所有6个子树片段中有1个片段相同，所以两棵树的相似子树数量为1。
（Example from钱龙华）

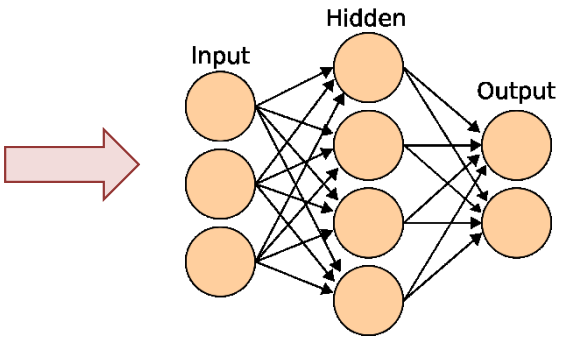
- 卷积树核ACE 2005 性能：**P=0.82, R=0.70 , F=0.76（State-of-the-Art）**

表示学习方法—基于深度神经网络

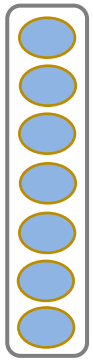
- 使用神经网络来学习关系实例的表示：
  - Recursive networks, Convolutional networks, Recurrent networks, 混合网络
- 考虑关系实例的结构（句法、词序列），同时加入语义信息（词向量）
- 上述Embedding表示可以与之前的特征组合一起用来分类

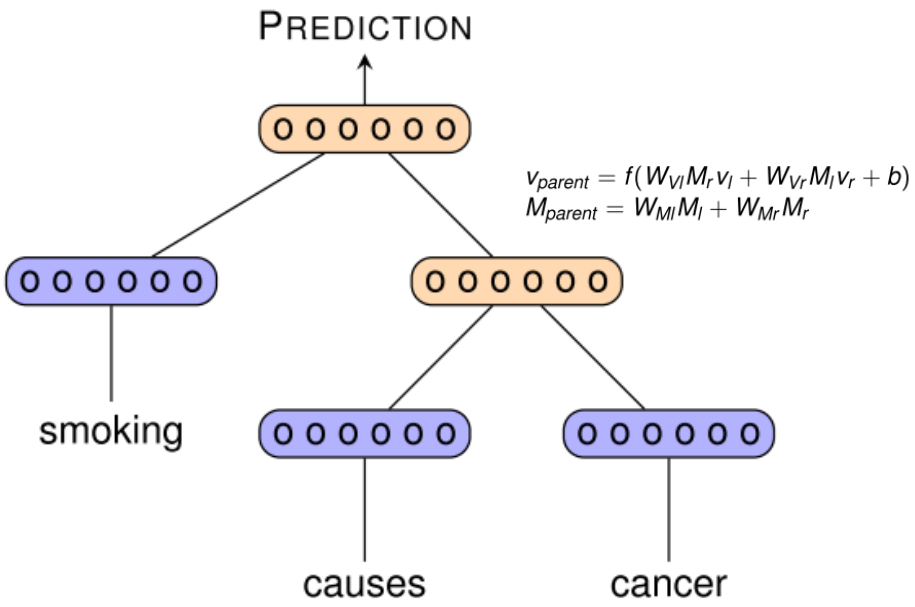
结构化关系实例

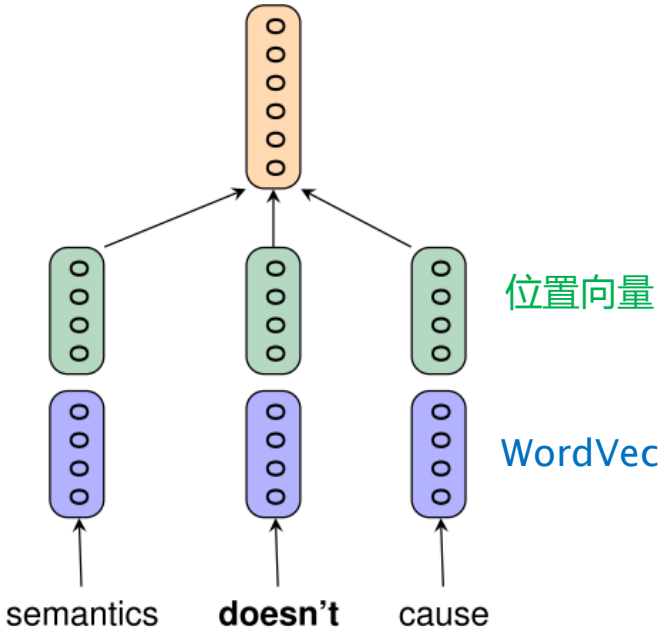




低维连续空间向量

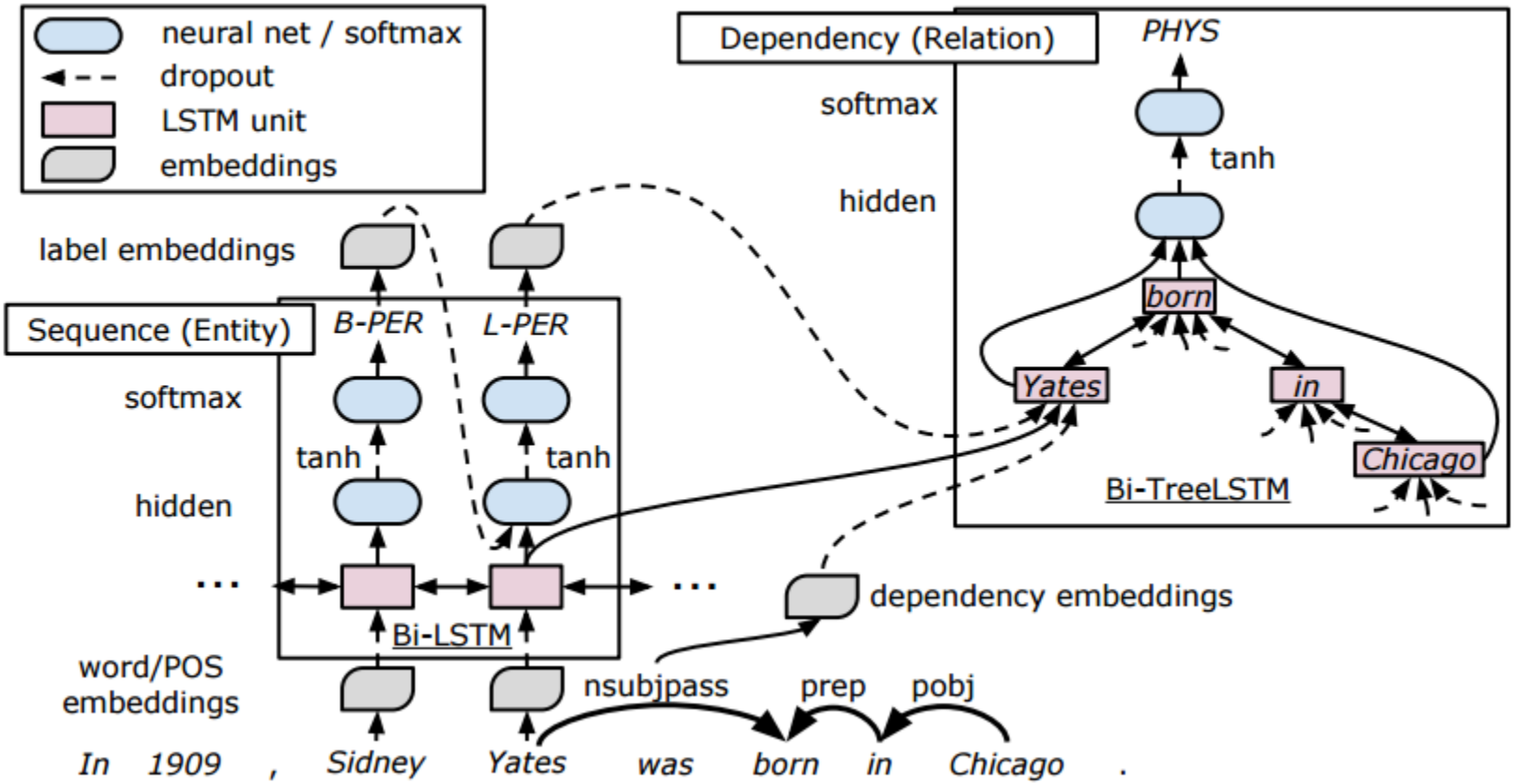






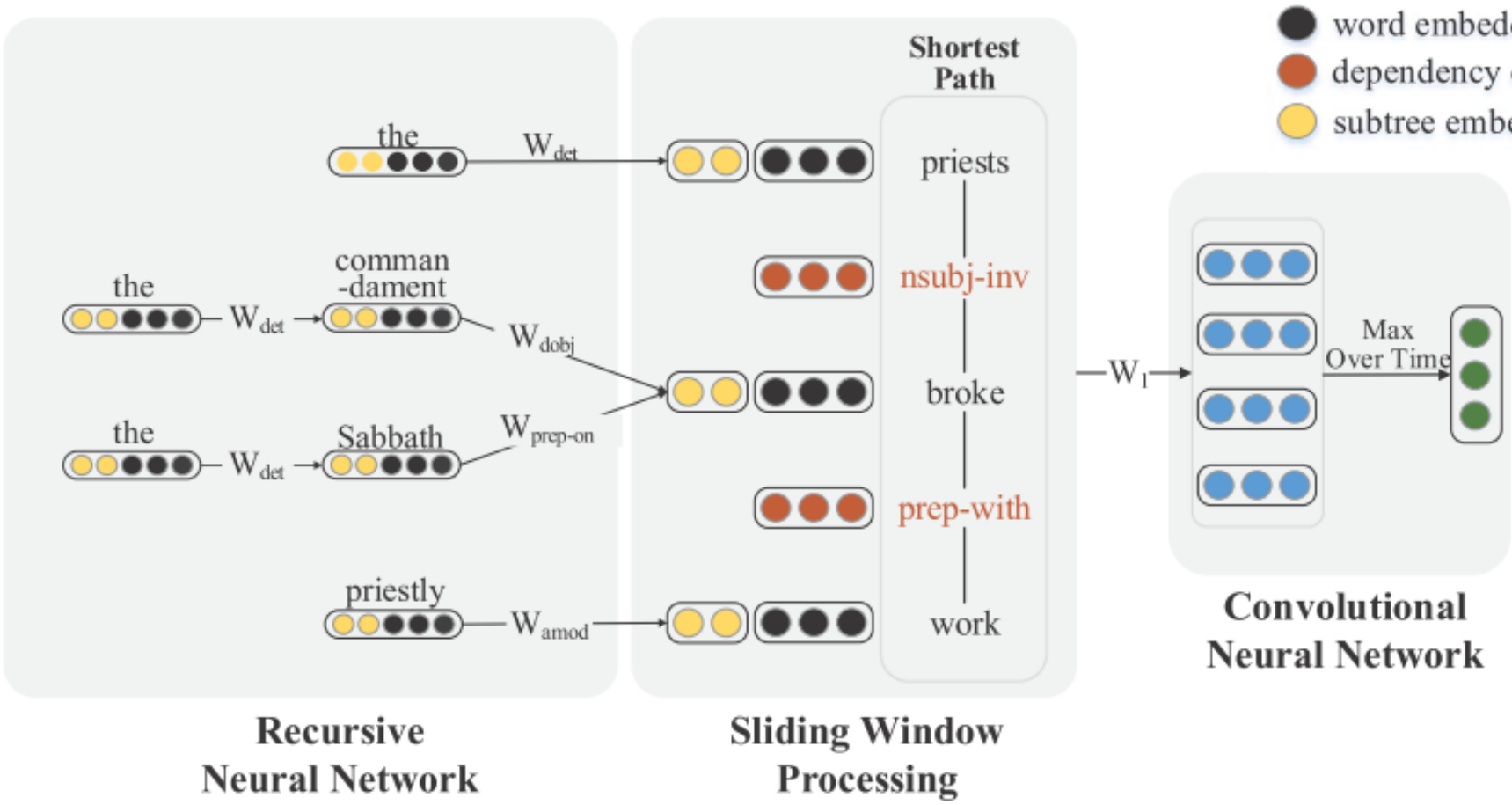
Recursive networks：基于句子的句法结构，自底向上递归的组合不同部分的表示 [Socher & al., 2012]

CNN：按特定窗口大小，依次组合窗口内词的信息，并通过max-pooling来得到最后的关系实例表示 [Zeng & al., 2014]



LSTM：

按照关系实例里面词的顺序，依次处理每一个词并更新关系实例的表示 [Li & al., 2015; Miwa & Bansal, 2016]



混合网络：

基于增强最短路径的深度学习表示（Liu et al., 2015）：

- 使用RNN来学习依存路径上每个词的增强表示
- 使用CNN来学习依存路径的表示

表示学习方法—关系分类性能

- 在关系分类任务（SemEval-2010 Task 8 dataset）是当前的State-of-the-art表示方法

Classifier	Feature set	F_1
SVM	POS, WordNet, prefixes and other morphological features, depdency parse, Levin classes, PropBank, FanmeNet, NomLex-Plus, Google n-gram, paraphrases, TextRunner	82.2
RNN	Word embeddings	74.8
	Word embeddings, POS, NER, WordNet	77.6
MVRNN	Word embeddings	79.1
	Word embeddings, POS, NER, WordNet	82.4
CNN	Word embeddings	69.7
	Word embeddings, word position embeddings, WordNet	82.7
Chain CNN	Word embeddings, POS, NER, WordNet	82.7
FCM	Word embeddings	80.6
	Word embeddings, depedency parsing, NER	83.0
CR-CNN	Word embeddings	82.8†
	Word embeddings, position embeddings	82.7
	Word embeddings, position embeddings	**84.1**†
SDP-LSTM	Word embeddings	82.4
	Word embeddings, POS embeddings, WordNet embeddings, grammar relation embeddings	**83.7**

表示学习方法—关系抽取性能

- 在关系抽取任务上（ACE 2005）上性能仍然离之前的方法有差距
- 根据（Nguyen & Grishman, 2015）的实验结果，CNN Embedding在ACE 2005上的性能为：
  - **P=0.71， R =0.54, F=0.61(State-of-the-Art性能：P=0.82, R=0.70 , F=0.76)**
- 原因分析：
  - 关系抽取任务与关系分类任务的差异
  - 在ACE 2005数据集中，90%的实体对之间都不存在语义关系 – 关系检测是核心
  - 相比之下， SemEval-2010下只有17%的实体对之间是不存在语义关系的 – 关系分类是核心

关系实例表示总结

- **三种方法**
  - 人工特征向量
  - 结构化核方法
  - 基于神经网络的表示学习（RNN，CNN,...）

- **性能**
  - 在关系分类任务上，基于NN的表示学习效果最好
  - 在关系抽取任务上，结构化核方法表示最好
  - 人工特征向量都能取得具备相当竞争力的性能
    - ✓在定义良好特征集上，性能一般比state-of-the-art系统低2~5个点左右

无/弱监督关系抽取

无/弱监督语义关系抽取

- **标注训练语料需要耗费大量的时间和精力**
  - **关系类别数目巨大**
  - **语料跨领域**
  - **…**

- **Bootstrapping**：不需要标语料，只需要大文档集

- **Distant Supervision**：不需要标语料，需要知识库

- **Open Information Extraction**：不需要标语料，不需要知道要抽哪些关系

BOOTSTRAPPING

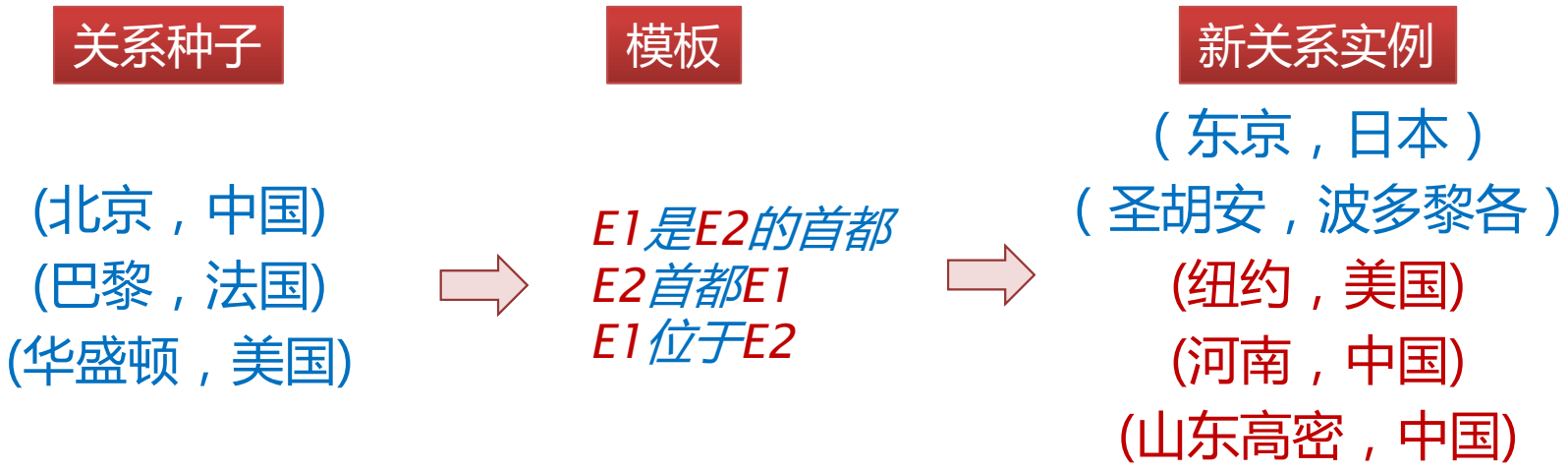
Bootstrapping

- Bootstrapping：一种经典（***古老/不好发论文***）、有不足之处、非常实用的方法

- 初始化：
  - 一些种子实体对
  - (北京，中国)，(巴黎，法国)，(华盛顿，美国)
- 扩展Expansion
  - 新模板： *E1是E2的首都，E2首都E1, …*
  - 新实例：（东京，日本），（圣胡安，波多黎各）
- 迭代数次，输出结果
- 主要的困难
  - 语义漂移（Semantic drift）

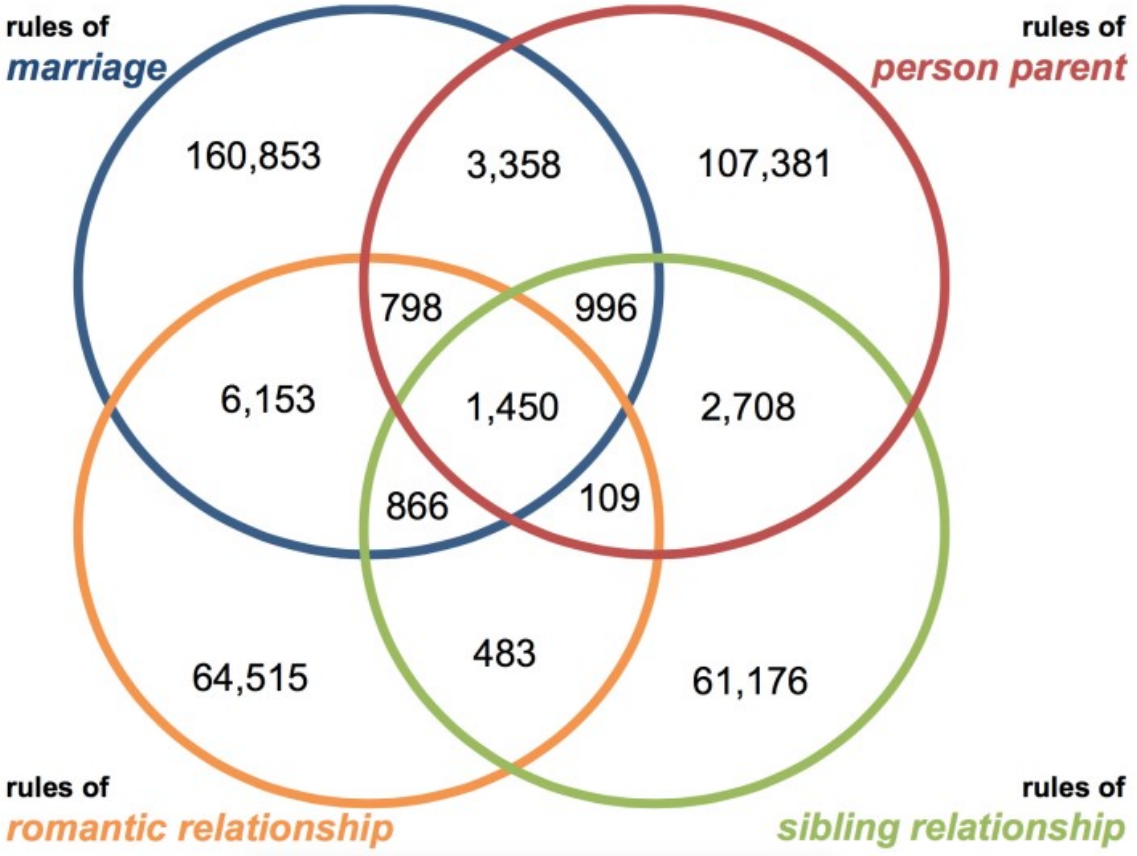
Bootstrapping-语义漂移

- 语义漂移问题：迭代会引入噪音实例和噪音模板



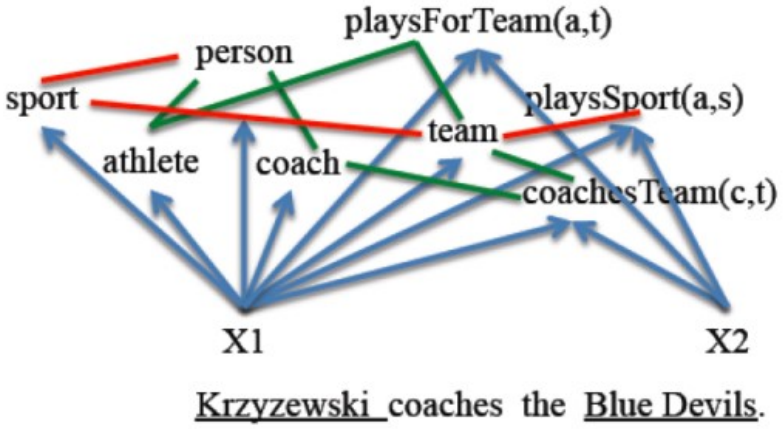
Bootstrapping-语义漂移

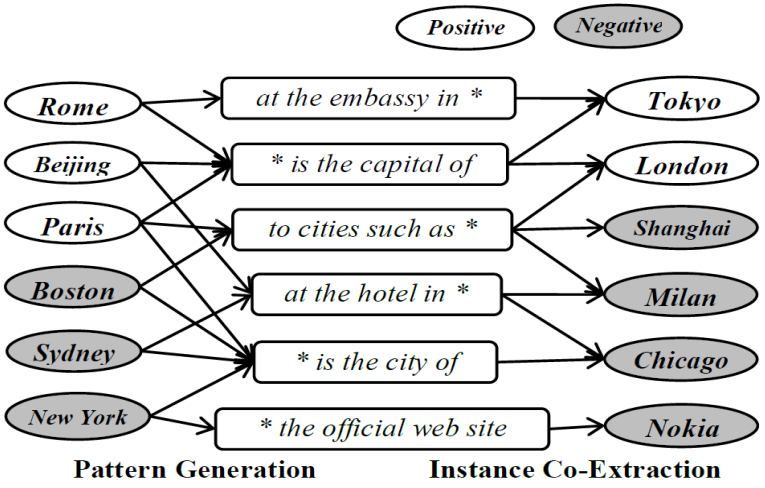
- 根据[Krause&al.,2012]，人物之间四种关系Pattern的交叉程度



Bootstrapping-语义漂移解决方案

- Mutual exclusive Bootstrapping (McIntosh et al., 09)：同时扩展多个互斥类别，一个实体对只能属于一个类别
- Coupled training（Carlson & al., 10）：建模不同抽取关系之间的约束，寻找最大化满足这些约束的抽取结果
- Co-Bootstrapping(Shi et al. 14)：引入负实例来限制语义漂移



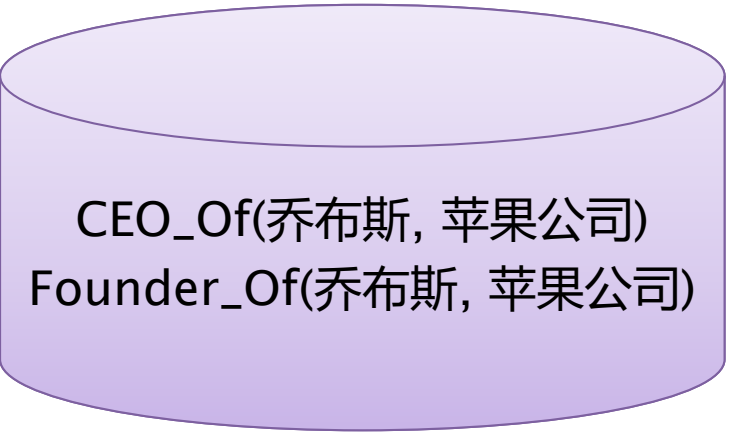


DISTANT SUPERVISION

知识监督开放抽取-Distant Supervision

- **Distant Supervision**: 使用知识库中的关系启发式的标注训练语料
  - WordNet, Freebase,Yago,DBPedia, WikiData…
- 启发式标注所有句子作为训练语料
- 使用最分类器来构建系统

知识库



标注训练语料

Relation Instance	Label
S1: 乔布斯是苹果公司的创始人之一	Founder-of, CEO-of
S2: 乔布斯回到了苹果公司	Founder-of, CEO-of

远距离监督方法—主要难点

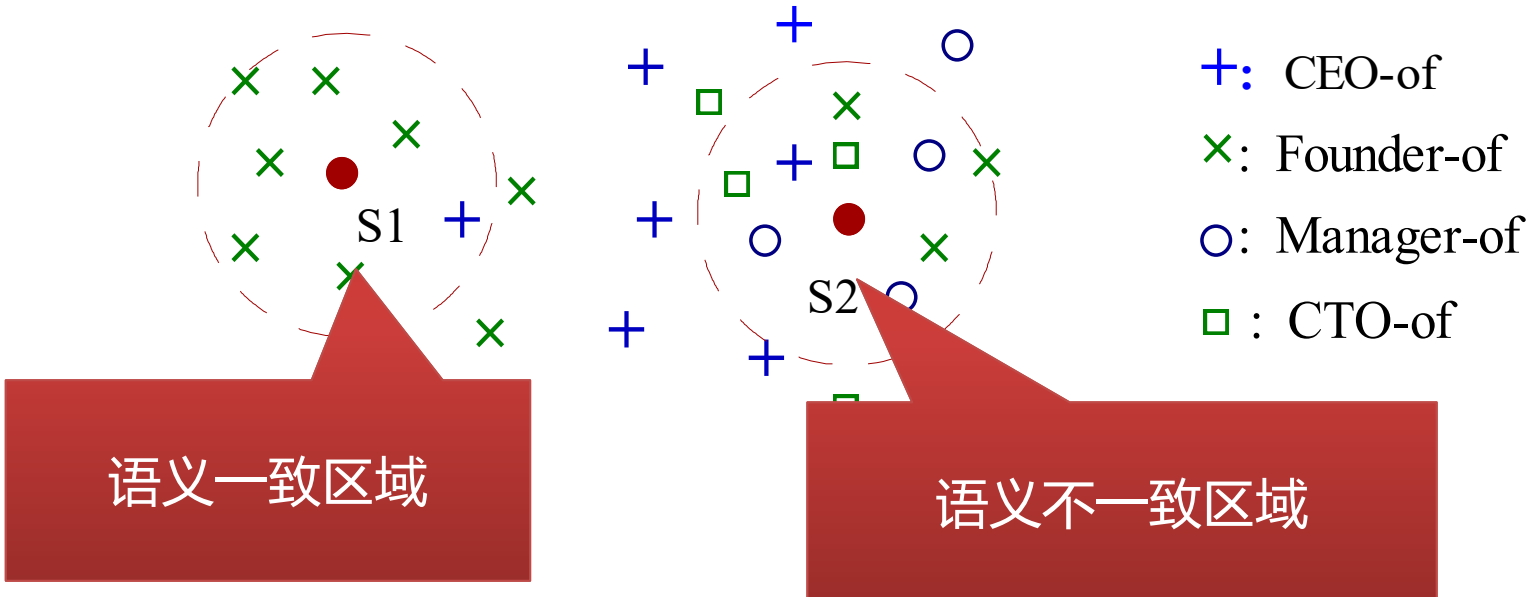
- **DS假设**: 每一个同时包含两个实体的句子都会表述这两个实体在知识库中的对应关系
- **主要难点**：带来大量噪音训练实例，严重影响抽取性能

噪音训练实例

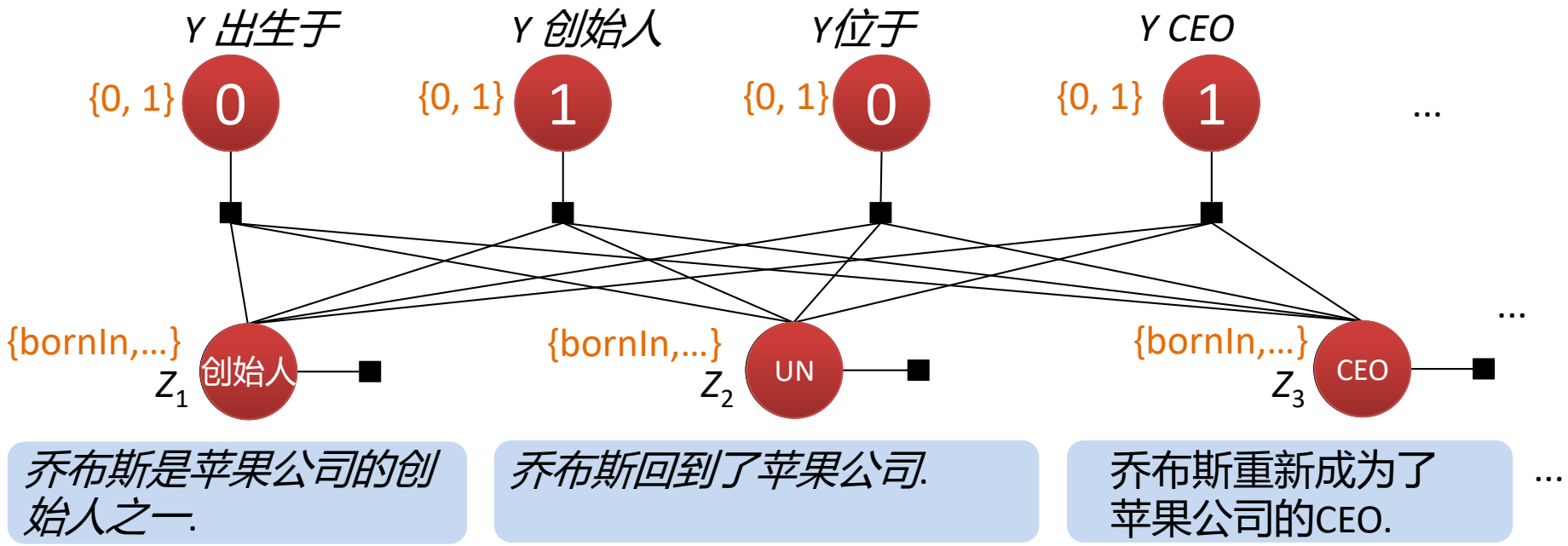
Relation Instance	Label	
S1:乔布斯是苹果公司的创始人之一	Founder-of	√
S1:乔布斯是苹果公司的创始人之一	CEO-of	X
S2:乔布斯回到了苹果公司	Founder-of	X
S2:乔布斯回到了苹果公司	CEO-of	X

基于噪音实例去除的DS方法

- 通过去除噪音实例来提升远距离监督方法的性能
- 假设：一个正确的训练实例会位于语义一致的区域，也就是其周边的实例应当都有相同一致的Label
  - 基于生成式模型的方法（Takamatsu et al. ACL 12）
  - 基于稀疏表示的方法（Han et al. ACL 14）



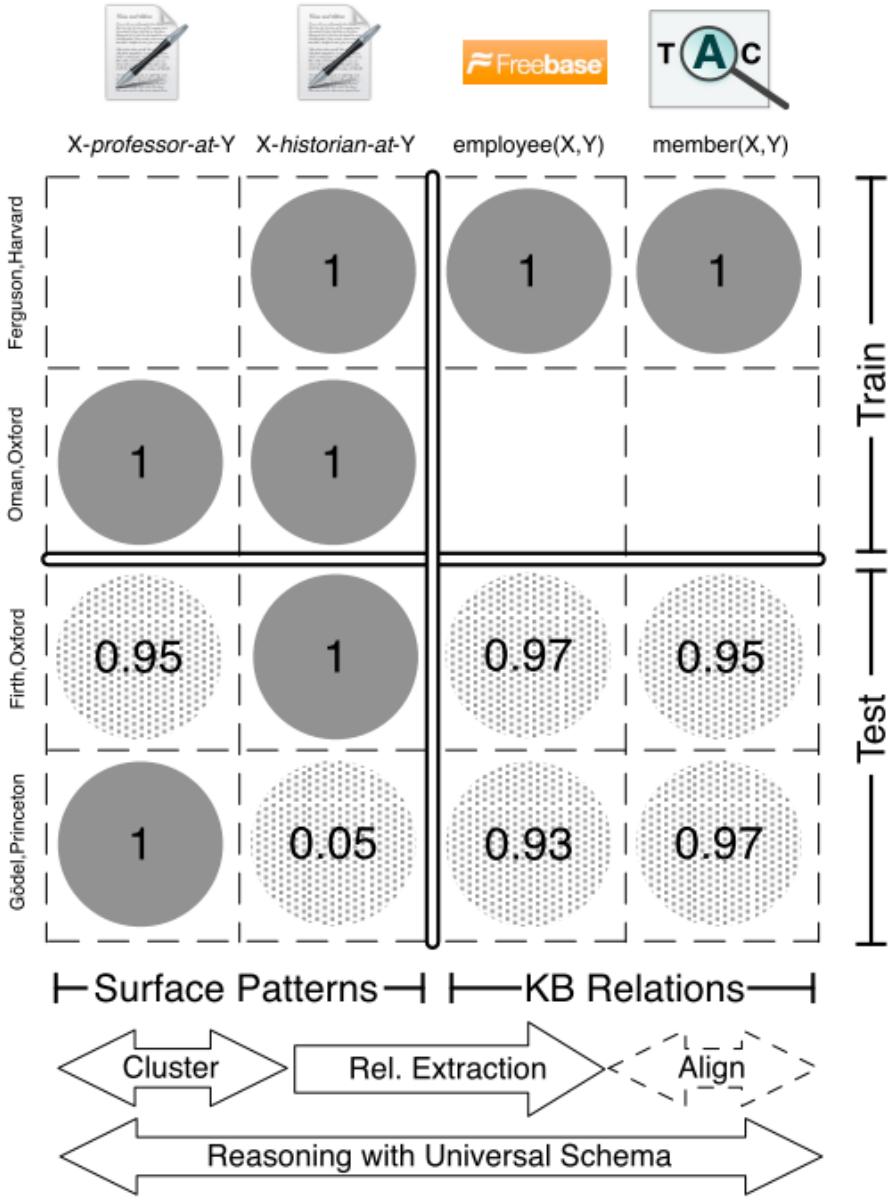
基于多实例学习的DS方法



- 一个实体对由一个句子集合表示
- 关系实例label被建模为hidden variable，使用Factor Graph来表示多个变量之间的关系（Surdeanu et al. EMNLP 12, ...）
- AtLeastOne假设：只要实体对的一个句子具有特定关系，那么该实体对也就具有该关系

基于协同推荐的DS方法

- 使用矩阵来表示实体对与Pattern，实体对与语义关系，Pattern与语义关系之间的关联
- 关系抽取任务被建模为矩阵填空问题
- 基于协同过滤推荐的方法(Riedel et al. NAACL 13)
- 基于Low-Rank矩阵分解的方法(Fan et al. ACL 14)



联合使用多源弱监督知识

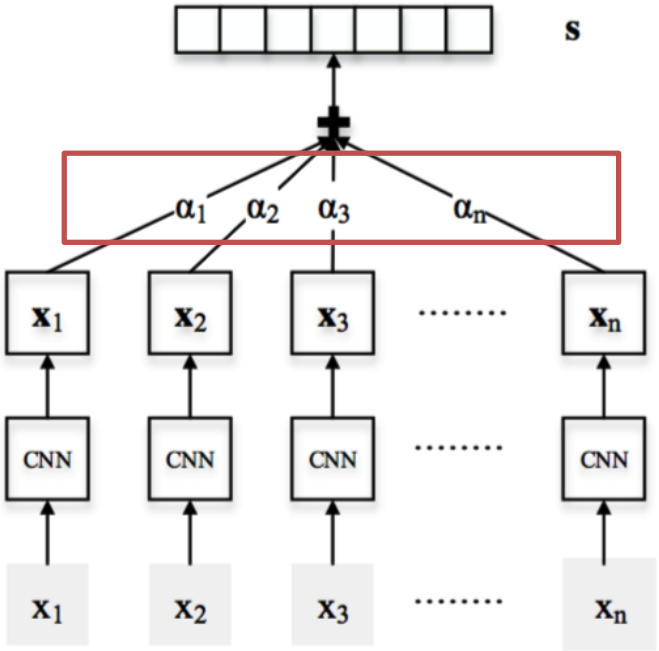
- 基于Markov Logic，同时使用多种不同的弱监督来提升关系抽取的性能（16% F1提升）（Han et al., 2016）
  - 两个实体之间关系的一致性：北京是中国的首都 => 北京是中国的政治中心
  - 具有特定关系的实体的约束：X是Y的政治制度 => Y必须是一个朝代，X必须是政治制度
  - 类似的表达具有相同的语义：X是Y的中央机构，Y的主要机构包括X，在Y的机构中X的职能是

System	P	R	F1
Mintz++	0.260	0.250	0.255
Hoffmann	0.306	0.198	0.241
Surdeanu	0.249	0.314	0.278
MLN-Base	**0.262**	**0.302**	**0.281**
MLN-Full	**0.426**	**0.259**	**0.322**

Table 1. The best F1-measures in P/R curves

基于深度神经网络的实例选择机制

- 用Attention机制来控制每个实例的表示在实体对表示中的重要性（α越大，越可能是对分类有用的实例）



Relation	employer_of
Low	When Howard Stern was preparing to take his talk show to **Sirius Satellite Radio**, following his former boss, **Mel Karmazin**, Mr. Hollander argued that ...
High	**Mel Karmazin**, the chief executive of **Sirius Satellite Radio**, made a lot of phone calls ...
Relation	place_of_birth
Low	**Ernst Haefliger**, a Swiss tenor who ... roles , died on Saturday in **Davos**, Switzerland, where he maintained a second home.
High	**Ernst Haefliger** was born in **Davos** on July 6, 1919, and studied at the Wettinger Seminary ...

$$e_i = \mathbf{x}_i \mathbf{A} \mathbf{r} \qquad \alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}$$

Lin, et al. (2016). Neural Relation Extraction with Selective Attention over Instances. ACL.

OPEN IE

Open IE

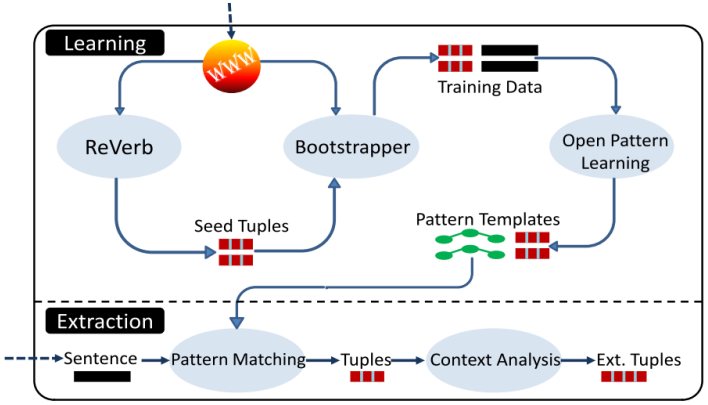
- **目标：**无需预先给定要抽取的关系类别，自动将自然语言句子转换为命题(propositions)
- **输入句子**
  - *莫言，山东高密人，是首位获得诺贝尔文学奖的中国作家*
- **输出命题**
  - (莫言, 是, 山东高密人)
  - (莫言, 是, 中国作家)
  - (莫言, 首位获得, 诺贝尔文学奖)
- **难点**
  - 复杂句子的处理
  - 关系短语语义的归一化

OpenIE：ReVerb

- 使用句法结构约束来发现关系短语
  - *动词: buy*
  - *动词 介词： buy in*
  - *动词 中间词 介词： is a city of*
- 同时使用句法和统计数据来过滤抽取出来的三元组
  - 关系短语应当是一个以动词为核心的短语
  - 关系短语应当匹配多个不同实体对
- 无需预先定义关系类别

$$V \mid VP \mid VW^*P$$
V = verb particle? adv?
W = (noun | adj | adv | pron | det)
P = (prep | particle | inf. marker)

关系短语的句法结构约束



难点1：复杂句子的处理

- 上述简单的句法模式通常不能处理复杂的句子
  - *莫言，山东高密人，是首位获得诺贝尔文学奖的中国作家*
- **解决方法：复杂句子分解**
- Clause IE[Corro & Gemulla, 2013]: 定义了7种简单句子模式和一系列句子分解规则，将复杂句子分解为简单句
- Stanford OpenIE(Angeli et. al., 2015): 使用句子的句法结构+转换规则进行简单句分解

1. SV_i
2. SV_e A
3. SV_c C
4. SV_{mt} O
5. SV_{dt} O_i O_d
6. SV_{ct} O A
7. SV_{ct} O C

难点2：语义的归一化

- X获得了Y，X领取了Y，Y得主X被认为是不同的语义关系
- 需要识别同一语义关系的不同关系短语，实现语义的归一化
- 解决方案：
  - Inference Rule Discovery：计算不同关系短语之间的相似度，识别表达相同语义的关系短语
  - 代表性模型：DIRT(Lin and Pantel, KDD01)， Soft Set inclusion (Nakashole, EMNLP 12)，Topic Model(Melamud et al., ACL 13), Jaccard (Dutta et al. WWW 15)

总部位于
(亚投行，北京)
(华为，深圳)
(IBM, 阿蒙克)
(联合国，纽约)
…

总部设置于
(亚投行，北京)
(联合国，纽约)
(红十字会，日内瓦)
(世界卫生组织, 日内瓦)
…

弱监督语义关系抽取技术总结

- **Bootstrapping**
  - 不需要标语料，只需要大文档集
  - 主要难点：语义漂移
  - 技术：Mutual Exclusive Bootstrapping, Coupled Learning，Co-Bootstrapping
- **Distant Supervision**
  - 不需要标语料，需要知识库
  - 主要难点：标注语料噪音
  - 技术：噪音实例去除，多实例学习，协同推荐，多源协同监督
- **Open Information Extraction**
  - 不需要标语料，不需要知道要抽哪些关系
  - 主要难点：复杂句子的处理，语义的归一化
  - 技术：句法模式学习，自学习技术，句子分解技术，Clustering，Inference Rule Discovery

总结

科研中的语义关系抽取

- **更好的表示**
  - 结构性、稀疏性、多样性、外部性…
  - embedding
- **更好的模型|算法**
  - 深度神经网络…
  - 统计和逻辑方法的结合…
  - 多任务、多文档、多层级joint model…
- **更少的监督/标注语料**
  - Distant supervision
  - …

现实中的语义关系抽取

- **更快的解决脏活（Dirty Work）**
  - 获得数据
  - 语料预处理
  - 获得模型训练的监督知识
- **快速的构建，快速的迭代**
  - 好的数据源胜过好的模型
  - (开始时)容易构建的系统比更高性能的系统更重要
    - ✓考虑从Bootstrapping开始
    - ✓使用现成的知识库
  - 人和机器一起监控抽取的过程和性能
  - 形成闭环：Continuous, never-ending learning
- **考虑中文的语言特点**

敬请大家批评和指导！

韩先培
xianpei@nfs.iscas.ac.cn
中文信息处理研究室，中科院软件所