

自然语言处理

信息抽取（关系抽取部分）

孙承杰 杨沐昀

sunchengjie@hit.edu.cn

哈尔滨工业大学计算学部
语言技术研究中心

主要内容

- 信息抽取的定义、任务及发展
- 命名实体识别
- 实体链接
- 关系抽取

主要内容

- 信息抽取的定义、任务及发展
- 命名实体识别
- 实体链接
- 关系抽取

实体链接的定义

□ 将“实体提及”链接到知识库中对应的实体

昨天下午，<ORG>市政协</ORG>、<ORG>市委统战部</ORG>联合举办<LOC>北京市</LOC><ORG>全国政协</ORG>委员视察考察活动，围绕历史文化街区改造和疏解整治促提升专项行动进展等视察并座谈。<ORG>市政协</ORG>主席<PER>吉林</PER>参加。

<PER>吉林</PER> → 吉林（北京市政协党组书记、主席）

实体链接的定义

吴京（某学者）

```
{ 'alias': [],  
  'subject_id': '86620',  
  'data': [{ 'predicate': '出生地', 'object': '浙江东阳' },  
    { 'predicate': '摘要',  
      'object': '吴京，男，1971年5月出生，浙江东阳人，  
      中共党员，东南大学土木工程学院获博士，东南大学土木工  
      程学院教授、博士生导师，东南大学土木工程学院建筑工程  
      系主任。' },  
    { 'predicate': '性别', 'object': '男' },  
    { 'predicate': '毕业院校', 'object': '东南大学' },  
    { 'predicate': '中文名', 'object': '吴京' },  
    { 'predicate': '义项描述', 'object': '东南大学教授  
' },  
    { 'predicate': '国籍', 'object': '中国' },  
    { 'predicate': '出生日期', 'object': '1971年5月' },  
    { 'predicate': '职业', 'object': '东南大学教授，博  
      导' },  
    { 'predicate': '学位', 'object': '博士' },  
    { 'predicate': '标签', 'object': '行业人物、人物、  
      教师' } ],  
  'type': 'Person',  
  'subject': '吴京' }
```

吴京（某演员）

```
{ 'alias': ['오 경', 'Jason Wu'],  
  'subject_id': '159056',  
  'data': [{ 'predicate': '出生地', 'object': '北京' },  
    { 'predicate': '外文名', 'object': 'Jason Wu、오 경（韩  
      语）' },  
    { 'predicate': '摘要', 'object': '吴京，1974年4月3日出生  
      于北京，毕业于北京体育大学。' },  
    { 'predicate': '代表作品', 'object': '流浪地球、战狼Ⅱ、  
      战狼、狼牙、杀破狼2、男儿本色、少林武王、小李飞刀、太极宗  
      师、' },  
    { 'predicate': '毕业院校', 'object': '北京体育大学' },  
    { 'predicate': '星座', 'object': '白羊座' },  
    { 'predicate': '配偶', 'object': '谢楠' },  
    { 'predicate': '义项描述', 'object': '中国内地男演员、导  
      演' },  
    { 'predicate': '出生日期', 'object': '1974年4月3日' },  
    { 'predicate': '身高', 'object': '175cm' },  
    { 'predicate': '职业', 'object': '演员、导演' },  
    { 'predicate': '标签', 'object': '娱乐人物、人物、导演、  
      演员' } ],  
  'type': 'Person',  
  'subject': '吴京' }
```

实体链接的定义

```
{
  "text_id": "1",
  "text": "《琅琊榜》海宴_【原创小说|权谋小说】",
  "mention_data": [
    {
      "mention": "琅琊榜",
      "offset": "1"
    },
    {
      "mention": "海宴",
      "offset": "5"
    },
    {
      "mention": "原创小说",
      "offset": "9"
    },
    {
      "mention": "权谋小说",
      "offset": "14"
    }
  ]
}
```



```
{
  "text_id": "1",
  "text": "《琅琊榜》海宴_【原创小说|权谋小说】",
  "mention_data": [
    {
      "kb_id": "2135131",
      "mention": "琅琊榜",
      "offset": "1"
    },
    {
      "kb_id": "10572965",
      "mention": "海宴",
      "offset": "5"
    },
    {
      "kb_id": "215143",
      "mention": "原创小说",
      "offset": "9"
    },
    {
      "kb_id": "NIL_Work",
      "mention": "权谋小说",
      "offset": "14"
    }
  ]
}
```

实体链接定义的延伸

- Linking free text to entities
 - ▣ Any piece of text
 - news documents
 - blog posts
 - tweets
 - queries
 - ...
 - ▣ Entities (typically) taken from a knowledge base
 - Wikipedia
 - Freebase
 - ...

实体链接的应用

□ Enable

- semantic search
- advanced User Interface(UI)/User Experience(UX)
- automatic document enrichment
- inline annotations
- ontology learning, KB population

□ “Use as feature”

- to improve classification; retrieval; word sense disambiguation; semantic similarity;...
- dimensionality reduction (e.g., term vectors)

进行实体链接的一般步骤

1. Determine “linkable” phrases
 - ▣ mention detection – **MD**
2. Rank/Select candidate entity links
 - ▣ link generation – **LG**
 - ▣ may include NILs (null values, i.e., no target in KB)
3. Use “context” to disambiguate/filter/improve
 - ▣ disambiguation – **DA**

Open Evaluation

□ Language-Independent Named Entity Recognition

- <http://www.cnts.ua.ac.be/conll2003/ner/>
- <http://www.cnts.ua.ac.be/conll2002/ner/>

EU NNP B-NP B-ORG
rejects VBZ B-VP O
German JJ B-NP B-MISC
call NN I-NP O
to TO B-VP O
boycott VB I-VP O
British JJ B-NP B-MISC
lamb NN I-NP O
.. O O

Peter NNP B-NP B-PER
Blackburn NNP I-NP I-PER

Open Evaluation

□ Text Analysis Conference (TAC) 2017

▣ Adverse Drug Reaction Extraction from Drug Labels (ADR)

- test various natural language approaches for their information extraction performance on adverse drug reactions (ADR).

▣ Knowledge Base Population (KBP)

- KBP tracks develop technologies for building and populating knowledge bases (KBs) from unstructured text.
 - Cold Start KB (CSKB), Entity Discovery and Linking (EDL), Slot Filling (SF), Event, Belief and Sentiment (BeSt)

Open Evaluation

- ❑ The BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenge evaluation
 - ▣ http://biocreative.sourceforge.net/biocreative_2_gn.html
 - ▣ Systems will be required to return the EntrezGene (formerly Locus Link) identifiers corresponding to the human genes and direct gene products appearing in a given MEDLINE abstract.
 - ▣ This has relevance to improving document indexing and retrieval, and to linking text mentions to database identifiers in support of more sophisticated information extraction tasks. .

Open Evaluation

□ CCKS2017

□ 问题命名实体识别和链接

■ 输入:

- 李娜是在哪一年拿的澳网冠军?

■ 输出:

- 李娜|||澳网\t李娜 (中国女子网球名将) |||澳大利亚网球公开赛

□ 电子病历命名实体识别

- 女性, 88岁, 农民, 双滦区应营子村人, 主因<症状和体征><身体部位>右髌部</身体部位>摔伤后疼痛肿胀, 活动受限5小时</症状和体征>于2016-10-29; 11: 12入院。

- <http://www.ccks2017.com/index.php/eval/>

Open Evaluation

□ CCKS 2019 中文短文本的实体链指

▣ https://biendata.com/competition/ccks_2019_el/

□ 输入:

```
{  
  "text_id": "1",  
  "text": "比特币吸粉无数，但央行的心另有所属 | 界面新闻 · jmedia"  
}
```

□ 输出:

```
{  
  "text_id": "1",  
  "text": "比特币吸粉无数，但央行的心另有所属 | 界面新闻 · jmedia"  
  "mention_data": [  
    {  
      "kb_id": "278410",  
      "mention": "比特币",  
      "offset": "0"  
    },  
    {  
      "kb_id": "199602",  
      "mention": "央行"
```

主要内容

- 信息抽取的定义、任务及发展
- 命名实体识别
- 实体链接
- 关系抽取

关系抽取

- 预定义关系抽取
- 开放域关系抽取
- 关系推理

关系抽取的定义

- 自动识别由一对实体和联系这对实体的关系构成的相关三元组
 - ▣ 二元关系
 - 比尔盖茨是微软的CEO
 - CEO(比尔盖茨, 微软)
 - ▣ 多元关系一般转化为二元关系处理
 - Michael Jordan获得1997/98赛季的MVP
 - Award(Michael Jordan, 1997/98赛季, MVP)

预定义关系抽取

□ 任务

- ▣ 给定实体关系类别，给定语料，抽取目标关系对

□ 评测语料 (MUC, ACE, KBP, SemEval)

- ▣ 专家标注语料，语料质量高
- ▣ 抽取的目标类别已经定义好

ACE实体关系类别

	PER	ORG	GPE	LOC	FAC	WEA	VEH
P	Per_Social.Bus Per_Social.Family, Per_Social.Lasting, Gen_Aff.Ideology, Gen_Aff.CRRE	Org_Aff.Employment, Org_Aff.Ownership, Org_Aff.Student/Alum, Org_Aff.Sports_Affiliation, Org_Aff.Investor/Shareholder, Org_Aff.Membership, Org_Aff.Founder, Gen_Aff.CRRE	Physical.Located, Physical.Near, Org_Aff.Employment, Org_Aff.Investor/Shareholder, Org_Aff.Founder, Gen_Aff.CRRE	Physical.Located, Physical.Near, Gen_Aff.CRRE	Physical.Located Physical.Near, Agent/Artifact.UOIM	Agent/Artifact.UOIM	Agent/Artifact.UOIM
O		Part_Whole.Subsidiary, Org_Aff.Investor/Shareholder, Org_Aff.Membership	Part_Whole.Subsidiary, Org_Aff.Investor/Shareholder, Gen_Aff.Loc/Origin	Gen_Aff.Loc/Origin	Agent/Artifact.UOIM	Agent/Artifact.UOIM	Agent/Artifact.UOIM
G		Org_Aff.Investor/Shareholder,Org_Aff.Membership,	Physical.Near, Part_Whole.Geographical Org_Aff.Investor/Shareholder	Physical.Near, Part_Whole.Geographical	Agent/Artifact.UOIM	Agent/Artifact.UOIM	Agent/Artifact.UOIM
L			Physical.Near, Part_Whole.Geographical	Physical.Near, Part_Whole.Geographical	Physical.Near, Part_Whole.Geographical		
O			Physical.Near, Part_Whole.Geographical	Physical.Near, Part_Whole.Geographical	Physical.Near, Part_Whole.Geographical		
C						Part_Whole.Artifact	
F							Part_Whole.Artifact
A							
C							
W							
E							
A							
V							
E							
H							

TAC-KBP实体关系类别

Person	Organization	Geo-Political Entity
alternate names	alternate names	alternate names
age	political/religious affiliation	capital
birth: date, place	top members/employees	subsidiary orgs
death: date, place, cause	number of employees	top employees
national origin	members	political parties
residences	member of	established
spouse	subsidiaries	population
children	parents	currency
parents	founded by	
siblings	founded	
other family	dissolved	
schools attended	headquarters	
job title	shareholders	
employee-of	website	
member-of		
religion		
criminal charges		

预定义关系抽取方法

- 采用机器学习的方法，将关系实例转换成特征表示，在标注语料库上训练生成分类模型，再识别实体间关系。
 - ▣ **基于特征向量方法**：最大熵模型(Kambhatla 2004)和支持向量机(Zhao et al., 2005; Zhou et al., 2005; Jiang et al., 2007)等
 - ▣ **基于核函数的方法**：浅层树核 (Zelenko et al., 2003)、依存树核 (Culotta et al., 2004)、最短依存树核 (Bunescu et al., 2005)、卷积树核 (Zhang et al., 2006; Zhou et al., 2007)
 - ▣ **基于神经网络的方法**：递归神经网络 (Socher et al., 2012)、基于矩阵空间的递归神经网络 (Socher et al., 2012)、卷积神经网络 (Zeng et al., 2014)

基于特征向量的关系抽取方法

- 常用的句法特征 (Syntactic features)
 - ▣ the entities themselves
 - ▣ the types of the two entities
 - ▣ word sequence between the entities
 - ▣ number of words between the entities
 - ▣ path in the parse tree containing the two entities
- 常用的语义特征 (Semantic features)
 - ▣ path between the two entities in the dependency parse
- Both the semantic and syntactic features extracted are presented to the classifier in the form of a feature vector, for training or classification.

基于核函数的关系抽取方法 (1)

- String-kernels (Lodhi et al., 2002)
- Given two strings x and y , the string-kernel computes their similarity based on the number of subsequences that are common to both of them.

$$\begin{aligned}\phi(x = cat) &= [\phi_a(x) \dots \phi_c(x) \dots \phi_t(x) \dots \phi_{at}(x) \dots \phi_{ca}(x) \dots \phi_{ct}(x) \dots \phi_{cat}(x) \dots] \\ &= [\lambda \quad \dots \lambda \quad \dots \lambda \quad \dots \lambda^2 \quad \dots \lambda^2 \quad \dots \lambda^2 \quad \dots \lambda^3 \quad \dots] \\ &\qquad \qquad \qquad \lambda \in (0, 1]\end{aligned}$$

$$i = i_1, i_2, \dots, i_{|u|} \quad [u = x[i] \quad l(i) = i_{|u|} - i_1 + 1$$

$$\begin{aligned}\phi_u(x) &= \sum_{i: u=x[i]} \lambda^{l(i)} & K(x, y) &= \phi(x)^T \phi(y) \\ & & &= \sum_{u \in U} \phi_u(x)^T \phi_u(y)\end{aligned}$$

U is the set of all possible ordered subsequences present in strings x and y

基于核函数的关系抽取方法 (2)

- In relation extraction, if x^+ and x^- are objects representing positive and negative entity-relation examples respectively and if y is the test example, $K(x^+, y) > K(x^-, y)$ implies that y contains a relation or otherwise.
- In practice $K(x, y)$ is the similarity function used in classifiers like SVMs, Voted Perceptron etc.
- For the task of relation extraction, objects x^+ , x^- and y can be represented as
 - ▣ word sequences around the entities under question
 - ▣ parse trees containing the entities

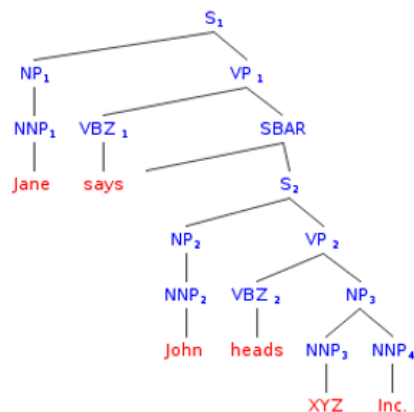
基于核函数的关系抽取方法 (3)

□ Bag of features Kernel

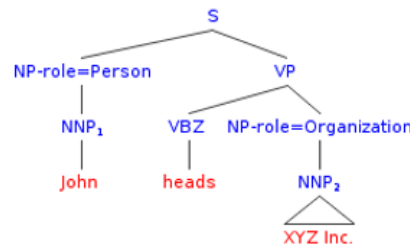
- ▣ A sentence $s = w_1, \dots, e_1, \dots, w_i, \dots, e_2, \dots, w_n$ containing related entities e_1 and e_2 can be described as
 - $s = s_b e_1 s_m e_2 s_a$.

□ Tree Kernels

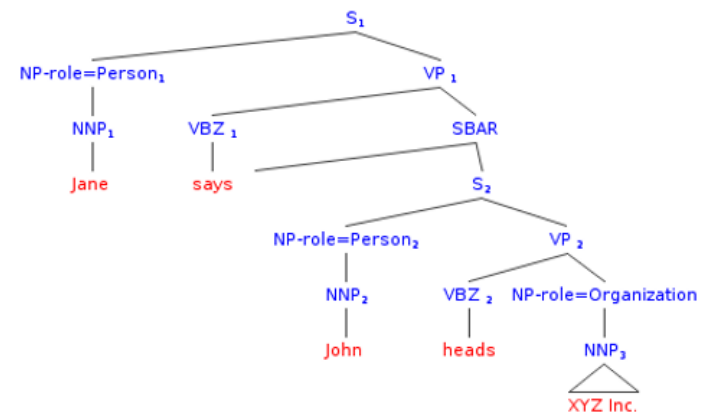
- ▣ The tree kernel computes a weighted sum of the number of subtrees that are common between the two shallow-parse trees.



(a) Parse Tree



(b) Positive Relation Example



(c) Negative Relation Example

基于神经网络的关系抽取方法

- **主要问题**：如何设计合理的网络结构，从而捕捉更多的信息，进而更准确的完成关系的抽取
- **网络结构**：不同的网络结构捕捉文本中不同的信息
 - ▣ 递归神经网络（Recursive Neural Network, RNN）
 - 网络的构建过程更多的考虑到句子的句法结构，但是需要依赖复杂的句法分析工具
 - ▣ 卷积神经网络（Convolutional neural network, CNN）
 - 通过卷积操作完成句子级信息的捕获，不需要复杂的NLP工具

基于卷积网络的预定义关系抽取

□ 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)

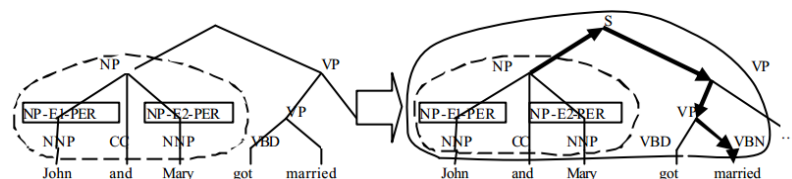
传统特征:

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:



□ 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

基于卷积网络的预定义关系抽取

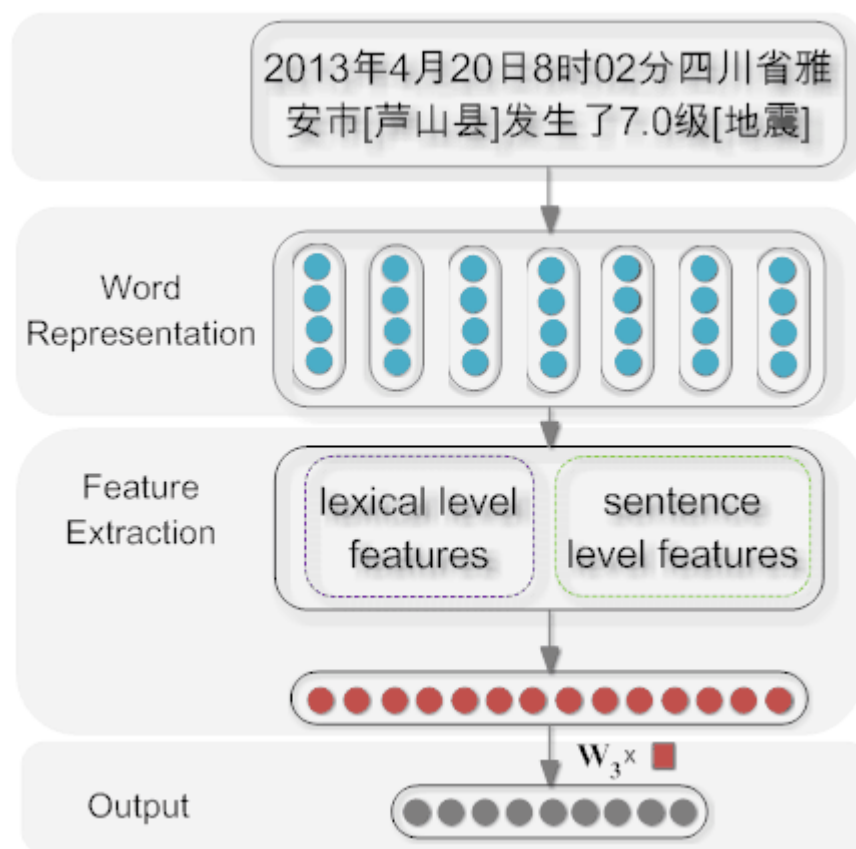
基于卷积神经网络的实体关系分类方法

- 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



基于卷积网络的预定义关系抽取

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 英文数据集上的实验结果

开放域关系抽取

□ 实体类别和关系类别不固定、数量大

- ▣ 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

□ 难点问题

- ▣ 如何获取训练语料
- ▣ 如何获取实体关系类别
- ▣ 如何针对不同类型目标文本抽取关系

□ 需要研究新的抽取方法

- ▣ 基于句法的方法
- ▣ 基于知识监督的方法

开放域关系抽取：基于句法的方法

- 通过识别表达语义关系的短语来抽取实体之间的关系
 - (华为, 总部位于, 深圳), (华为, 总部设置于, 深圳), (华为, 将其总部建于, 深圳)
- 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
- 优点：无需预先定义关系类别
- 缺点：语义没有归一化，同一关系有不同表示

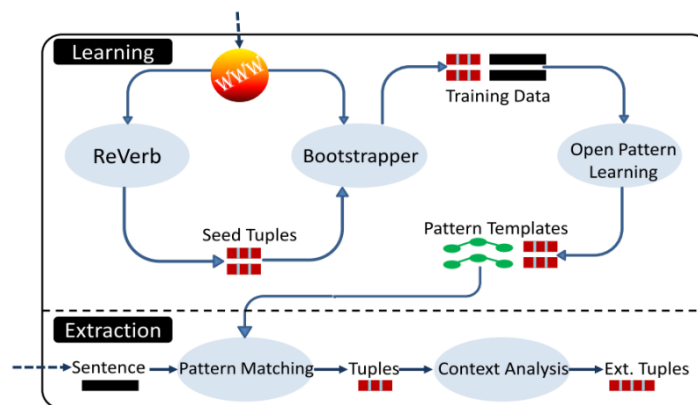
$$V \mid VP \mid VW^*P$$

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



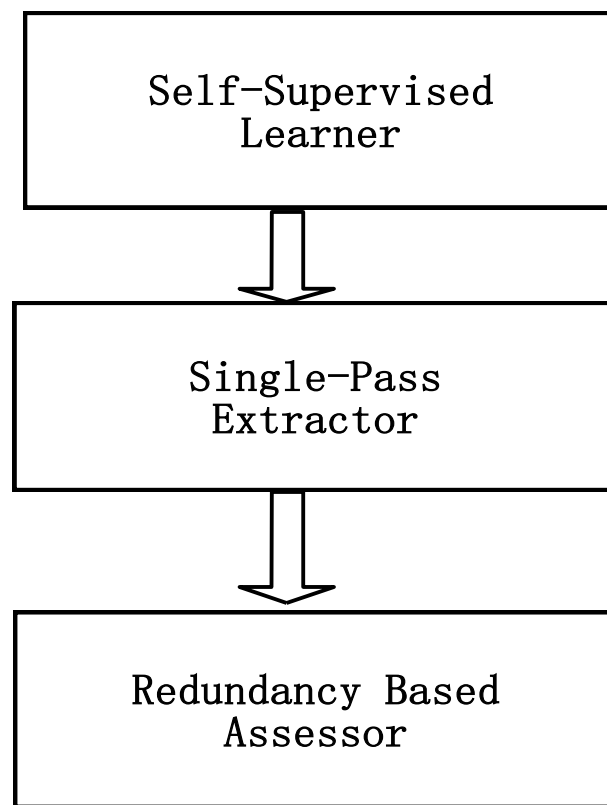
开放域关系抽取：基于句法的方法

□ 步骤：

- 离线的训练集产生：利用简单的启发式规则，在宾州树库上产生训练语料
- 离线的分类器训练：提取一些浅层句法特征，训练分类器，用来判断一个元组是否构成关系
- 在线关系抽取：在网络语料上，找到候选句子，提取浅层句法特征，利用分类器，判断抽取的关系对是否“可信”
- 在线的关系可信度评估：利用网络海量语料的冗余信息，对可信的关系对，进行评估

□ 出发点：

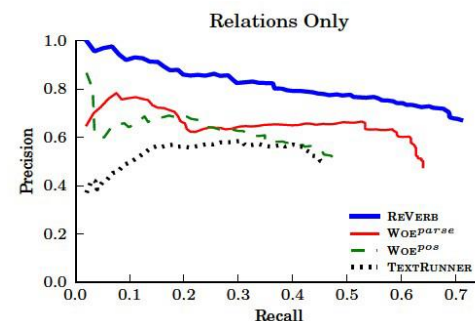
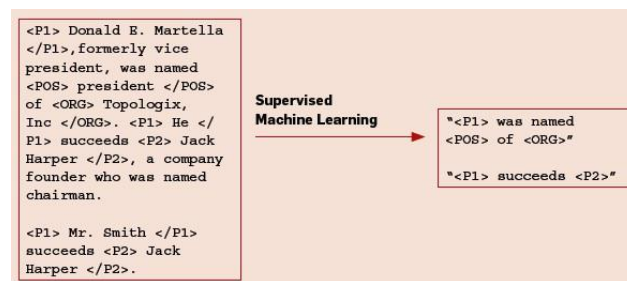
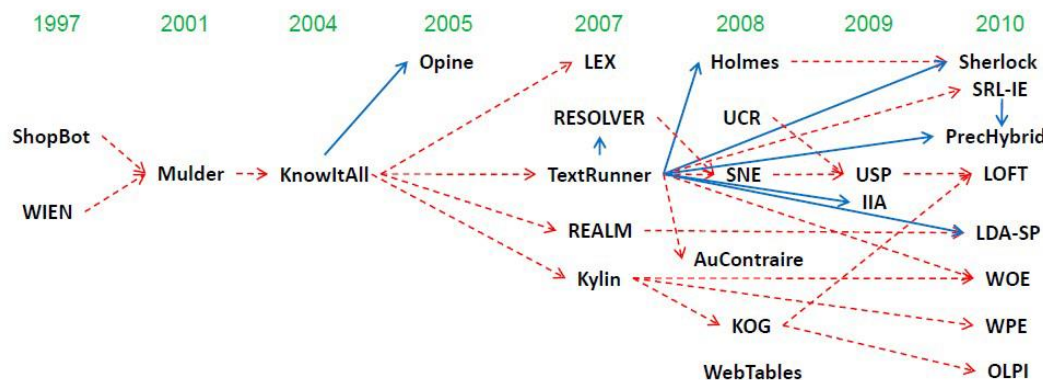
- 关系类别的产生：动词作为关系类别
- 训练语料的产生：通过句法关系引出语义关系



开放域关系抽取代表系统

TextRunner、ReVerb、WOE、OLLIE

- 从Wikipedia Infobox获得关系名
- 通过在句法树上回标获得句法关系模板



开放域关系抽取: 基于知识监督的方法

□ 任务: 在Wikipedia文本中抽取关系 (属性) 信息

□ 难点

- 无法确定关系类别

- 无法获取训练语料

□ 方法

- 在Infobox抽取关系信息

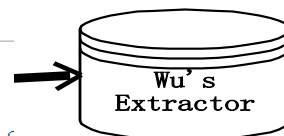
- 在Wikipedia条目文本中进行回标, 产生训练语料

Input

Tsinghua University
From Wikipedia, the free encyclopedia
(Redirected from [Qinghua](#))

For the university in Taiwan, see [National Tsing Hua University](#).

Tsinghua University (THU; simplified Chinese: 清华大学; traditional Chinese as [Qinghua](#), is a university in Beijing, China. The school is o: 1911 under the name "Tsinghua Xuetang" or "Tsinghua College" and was re: was founded in 1925 and the name "National Tsinghua University" was ado: *Commitment*, Tsinghua University describes itself as being dedicated to global development.^[1] Tsinghua is often ranked as the first or second t rankings.^{[2][3][4]}



Output

Tsinghua University 清华大学	
	
Motto	自强不息, 厚德载物
Motto in English	Self-discipline and Social Commitment
Established	1911
Type	Public
President	Gu Binglin
Academic staff	2, 857
Undergraduates	13, 915
Postgraduates	12, 831
Location	Beijing, People's Republic of China
Campus	Urban, 395 ha (3.95 km ²)
Flower	Redbud and Lilac
Colors	Purple and White
Affiliations	ABARU, APRU, C9
Website	www.tsinghua.edu.cn

开放域关系抽取:基于知识监督的方法

Clearfield County, Pennsylvania	
Statistics	
Founded	March 26, 1804
Seat	Clearfield
Area	
- Total	2,988 km ² (1,154 mi ²)
- Land	sq mi (km ²)
- Water	17 km ² (6 mi ²), 0.56%
Population	
- (2000)	83,382
- Density	28/km ²

Clearfield County was created on 1804 from parts of Huntingdon and Lycoming Counties but was administered as part of Centre County until 1812.

Its county seat is Clearfield.

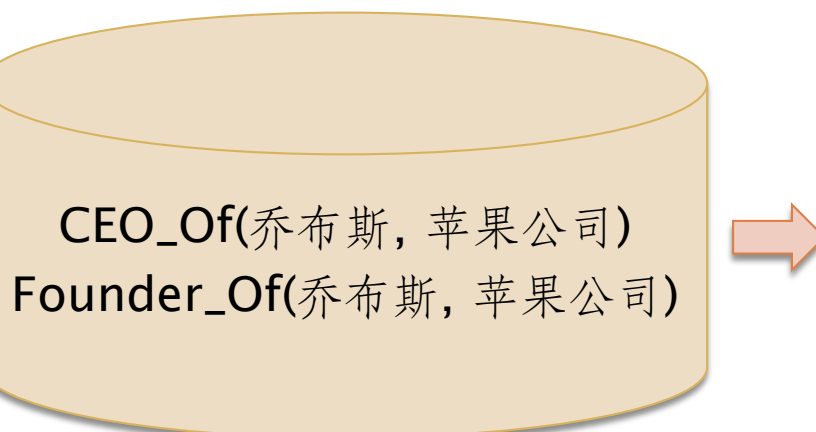
2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water.

As of 2005, the population density was 28.2/km².

开放域关系抽取: 远距离监督 (Distant Supervision) 方法

- 开放域信息抽取的一个主要问题是缺乏标注语料
- (Mintz et al., ACL09) 首次提出了 Distant Supervision 的思想
- Distant Supervision: 使用知识库中的关系启发式的标注训练语料

知识库



CEO_Of(乔布斯, 苹果公司)
Founder_Of(乔布斯, 苹果公司)

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

开放域关系抽取: 远距离监督 (Distant Supervision) 方法

- **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
- 基于上述假设标注所有句子作为训练语料
- 使用最大熵分类器来构建IE系统
- 知识库: Freebase, 文本: Wikipedia
- 最大的问题: 噪音训练实例

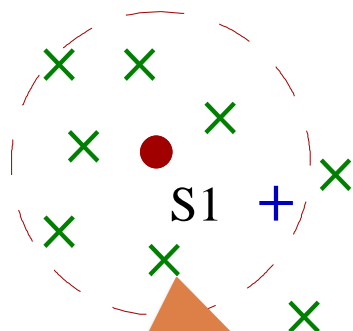
噪音训练实例

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	✗
S2: 乔布斯回到了苹果公司	Founder-of	✗
S2: 乔布斯回到了苹果公司	CEO-of	✗

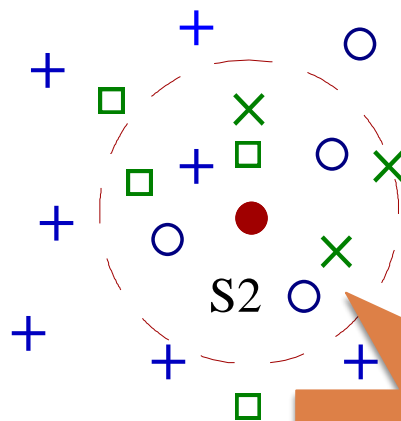
开放域关系抽取: 远距离监督 (Distant Supervision) 方法

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



语义一致区域



语义不一致区域

+ : CEO-of
x : Founder-of
○ : Manager-of
□ : CTO-of

Data Set

□ 关系分类

▣ SemEval-2010 Task 8

■ <https://github.com/sahitya0000/Relation-Classification>

□ 关系抽取

▣ ACE 2005

■ <https://catalog.ldc.upenn.edu/LDC2006T06>

▣ OpenNRE

■ <https://github.com/thunlp/OpenNRE>

□ 关系推理

▣ FB15k-237

■ <https://github.com/deepakn97/relationPrediction/tree/master/data>

Useful links

- 关系抽取（分类）总结（最后更新2019-12-31）
 - ▣ <http://shomy.top/2018/02/28/relation-extraction/>