

## 1. Bootstrap

seeds/seed tuple

bootstrap

Seeds/seed tuple

作者	书名
A	123
B	456
C	7890

作者 → 书名  
→ 已知

step 1: 生成规则      文本 → 规则

1. 李航马3统计全析 → 规则:  $X \text{ 马 } Y$
2. 机器学习是由周志华写的 → 规则:  $Y \text{ 是由 } X \text{ 写的}$
3. ....

↓ 由此得到规则库

step 2: 生成 tuple

→ 规则 → tuple

→ 新的文本去循环是否能够匹配规则库

将新的 tuple 将入到 seed 中

缺点:

规则的准确性低; 人力成本;

存在一些修饰语 (解决: 近似匹配);

Error accumulation;

Bootstrap: 1. 生成规则; 2. 生成 tuple; 如此循环

Snowball: 1. 生成规则; 2. 生成 tuple; 3. 评估规则; 4. 评估 tuple; 循环

## 2. snowball

SNOW BALL (Bootstrap)

seed tuple	
ORG	LOC
Microsoft	Redmond
IBM	Armonk
Borg	Seattle
Intel	Santa Clara

Fact:  $length = 2$

1. Computer Serves at Microsoft's headquarters in Redmond.
2. In mid-afternoon, Redmond-based Microsoft fell.
3. The Armonk-based IBM Introduces...
4. Borg's headquarters in Seattle.
5. Intel's Santa Clara, cut price of --

⇒ 近似算法  
相似度  
(tag → tuple)  
相似 tuple

Step1: 生成规则

Pattern representation →

1 ⇒ (Serves at) Microsoft's headquarters in Redmond ( )

Left Middle Right

$|f_{\text{Left}}| = |V|$   $|f_{\text{Middle}}| = |V|$   $|f_{\text{Right}}| = |V|$

$\langle \text{Left} \rangle \langle \text{ORG} \rangle \langle \text{Middle} \rangle \langle \text{Loc} \rangle \langle \text{Right} \rangle \leftarrow \text{规则}$

From 海上漂来 to Everyone:  
只不过这个实体再出现的概率比较低  
错误放大!

From 好嗨哟 to Everyone:  
这个标记是不是要用到实体识别呀?

From 好嗨哟 to Everyone:  
left去多少个词合适呀?

From 海上漂来 to Everyone:  
所以在近似算法之前还有一步是对新规则进行一个量化表示

From 好嗨哟 to Everyone:  
实体类型是指? 例如: 上海! = 北京, 上海! = 书本  
城市 = 城市, 城市! = 书本?

To: Everyone More

Type message here...

Fact:  $length = 2$

1. Computer Serves at Microsoft's headquarters in Redmond.
2. In mid-afternoon, Redmond-based Microsoft fell.
3. The Armonk-based IBM Introduces...
4. Borg's headquarters in Seattle.
5. Intel's Santa Clara, cut price of --

⇒ 近似算法  
相似度  
(tag → tuple)  
相似 tuple

Step1: 生成规则

Pattern representation →

1 ⇒ (Serves at) Microsoft's headquarters in Redmond ( )

Left Middle Right

$|f_{\text{Left}}| = |V|$   $|f_{\text{Middle}}| = |V|$   $|f_{\text{Right}}| = |V|$

$\langle \text{Left} \rangle \langle \text{ORG} \rangle \langle \text{Middle} \rangle \langle \text{Loc} \rangle \langle \text{Right} \rangle \leftarrow \text{规则}$

怎么设计算相似度?

规则1  $p = \langle L_1, T_1, M_1, T_2, R_1 \rangle$

规则2  $s = \langle L_2, T_2, M_2, T_3, R_2 \rangle$

$Sim(p, s) = 0$  if  $T_1 \neq T_2$  or  $T_2 \neq T_3$


⇒  $\frac{\mu_1 \cdot L_1 + \mu_2 \cdot M_1 + \mu_3 \cdot R_1}{\mu_1 + \mu_2 + \mu_3}$  otherwise

$\mu_1 + \mu_2 + \mu_3 = 1$   $\mu_1 = \mu_2 = \mu_3 = \frac{1}{3}$

为什么可以用内积表示相似度：

因为在把规则表示成向量的时候，已经做了归一化。所以不用除以模长。

### 2.1 step1 生成模版 pattern



① 生成模板 (pattern)

1.  $\begin{pmatrix} \text{Serves} & \text{at} \end{pmatrix} \xrightarrow{\text{ORG}} \begin{pmatrix} \text{'s} & \text{headquarters} & \text{in} \end{pmatrix} \xrightarrow{\text{LOC}} \begin{pmatrix} \end{pmatrix}$
2.  $\begin{pmatrix} \text{In} & \text{mid-afternoon} \end{pmatrix} \xrightarrow{\text{LOC}} \begin{pmatrix} - & \text{based} \end{pmatrix} \xrightarrow{\text{ORG}} \begin{pmatrix} \text{fell} \end{pmatrix}$
3.  $\begin{pmatrix} \text{The} \end{pmatrix} \xrightarrow{\text{LOC}} \begin{pmatrix} - & \text{based} \end{pmatrix} \xrightarrow{\text{ORG}} \begin{pmatrix} \text{Introduced} & \text{a} \end{pmatrix}$
4.  $\begin{pmatrix} \text{operate} & \text{from} \end{pmatrix} \xrightarrow{\text{ORG}} \begin{pmatrix} \text{is} & \text{headquarters} & \text{in} \end{pmatrix} \xrightarrow{\text{LOC}} \begin{pmatrix} \end{pmatrix}$

\*  $\begin{pmatrix} \text{Serves} & \text{at} \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 0 & 0 & \dots & 1 & \dots & 0 & 1 & \dots \end{pmatrix}$   
 $\rightarrow \begin{pmatrix} 0 & 0 & 0 & \dots & 0.75 & \dots & 0 & 0.75 & \dots \end{pmatrix}$   
 $0.75^2 + 0.75^2 \approx 1$

① 生成模板 (pattern)

1.  $\begin{pmatrix} \text{Serves} & \text{at} \end{pmatrix} \xrightarrow{\text{ORG}} \begin{pmatrix} \text{'s} & \text{headquarters} & \text{in} \end{pmatrix} \xrightarrow{\text{LOC}} \begin{pmatrix} \end{pmatrix} \rightarrow P_1$
2.  $\begin{pmatrix} \text{In} & \text{mid-afternoon} \end{pmatrix} \xrightarrow{\text{LOC}} \begin{pmatrix} - & \text{based} \end{pmatrix} \xrightarrow{\text{ORG}} \begin{pmatrix} \text{fell} \end{pmatrix} \rightarrow P_2$
3.  $\begin{pmatrix} \text{The} \end{pmatrix} \xrightarrow{\text{LOC}} \begin{pmatrix} - & \text{based} \end{pmatrix} \xrightarrow{\text{ORG}} \begin{pmatrix} \text{Introduced} & \text{a} \end{pmatrix} \rightarrow P_3$
4.  $\begin{pmatrix} \text{operate} & \text{from} \end{pmatrix} \xrightarrow{\text{ORG}} \begin{pmatrix} \text{is} & \text{headquarters} & \text{in} \end{pmatrix} \xrightarrow{\text{LOC}} \begin{pmatrix} \end{pmatrix} \rightarrow P_4$

\*  $\begin{pmatrix} \text{Serves} & \text{at} \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 0 & 0 & \dots & 1 & \dots & 0 & 1 & \dots \end{pmatrix}$   
 $\rightarrow \begin{pmatrix} 0 & 0 & 0 & \dots & 0.75 & \dots & 0 & 0.75 & \dots \end{pmatrix}$   
 $0.75^2 + 0.75^2 \approx 1$

假设: 模板之间的相似度  $\Rightarrow$  clustering

① clustering (模板)

- ✓ k-means
- ✓ Spectral Clustering:  $\begin{pmatrix} \text{matrix} \end{pmatrix} \rightarrow$
- ✓ Hierarchical Clustering

问题:  $\tilde{P} = \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 \end{pmatrix}$

1.  $[P_1]$
2.  $[P_1, P_2]$   $\text{sim}(P_1, P_2) > 0.7$
3.  $[P_1, P_2, P_3]$   $\text{sim}(P_1, P_3) = 0.1, \text{sim}(P_2, P_3) = 0.2$
4.  $[P_1, P_2, P_3, P_4]$   $\text{sim}(P_1, P_4) = 0.15, \text{sim}(P_2, P_4) = 0.2$
5.  $[P_1, P_2, P_3, P_4, P_5]$   $\text{sim}(P_1, P_5) = 0.75$

②  $[P_1, P_2], [P_2, P_3]$   
 $\downarrow$  平均  $\downarrow$  平均  
 Centroid Centroid

## 2.2 step2 生成 tuple

② 生成 tuple  $> 0.7$

1.  $\begin{pmatrix} \text{ORG} & \text{LOC} \end{pmatrix}$
2.  $\begin{pmatrix} \text{ORG} & \text{LOC} \end{pmatrix}$
3.  $\begin{pmatrix} \text{ORG} & \text{LOC} \end{pmatrix}$
4.  $\begin{pmatrix} \text{ORG} & \text{LOC} \end{pmatrix}$

规则

Text

ORG LOC

ORG LOC

ORG LOC

ORG LOC

for ORG for LOC for

$\begin{pmatrix} \text{ORG} & \text{LOC} \end{pmatrix}$

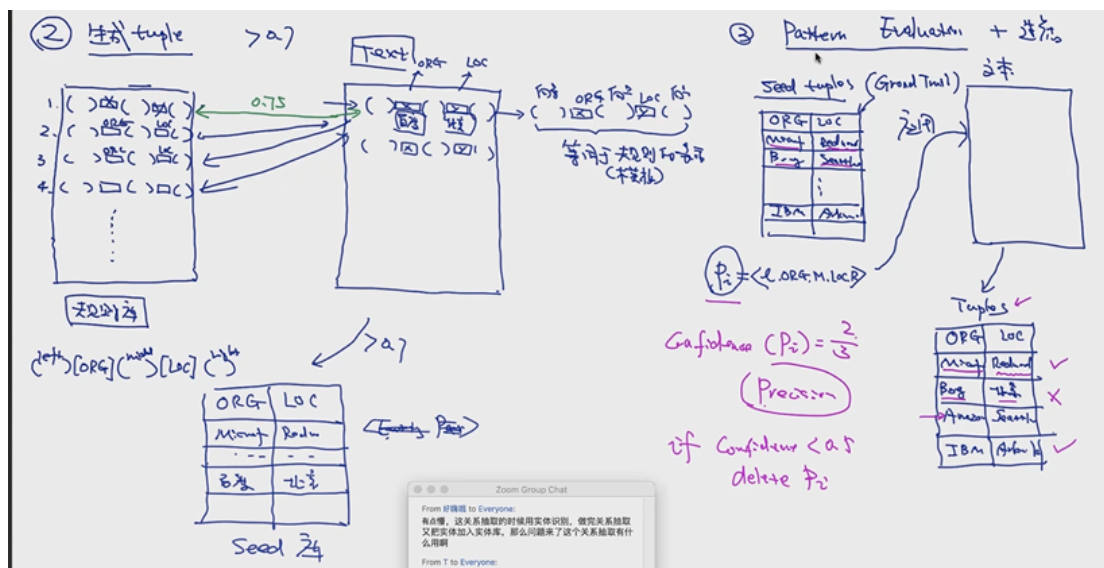
等同于规则 (模板)

$> 0.7$

ORG	LOC
Mount	Radu
...	...
Base	北

$\langle \text{Entity Pair} \rangle$

## 2.3 step3 模版评估 pattern evaluation + 过滤



## 2.4 step4 tuple evaluation + 过滤

