

# 实体链接

## (Entity Linking)

---

汪 鹏 , 潘 喆



`pwang@seu.edu.cn`

东南大学 计算机科学与工程学院

---

课程主页: <https://github.com/npubird/KnowledgeGraphCourse>

# 提纲

- 一、实体链接背景场景
- 二、实体链接方法简介
  - ✓ 基于概率生成模型的方法
  - ✓ 基于主题模型的方法
  - ✓ 基于图的方法
  - ✓ 基于深度学习的方法
  - ✓ 无监督方法

# 实体链接的问题背景

**实体链接**是指将文档中出现的文本片段，即实体指称 (entity mention) 链向其在特定知识库 (Knowledge Base) 中相应条目 (entry) 的过程。



# 实体链接的问题背景

- 例如：

86年的电视剧西游记是对小说西游记最经典的改编。

- 链接结果：

86年的电视剧西游记（1986年杨洁执导央视版电视剧）是对小说西游记（中国古典长篇小说）最经典的改编。

【对其中的专有名词进行正确的标注】

# 实体链接的应用场景

- 文本分类和聚类
- 信息检索
- 知识库构建
- 智能问答
- ...

# 实体链接的步骤

- 命名实体识别（之前已经讨论过了）
- 词义消歧

# 实体链接的测评

会议	组织者	评测	年份	文本语种	文本风格	知识库	需识别提及	新实体聚类
TAC	NIST	KBP Entity Linking	2009-2010	英	新闻、博客	英文维基百科	否	否
		KBP Entity Linking	2011	英、 <b>中</b>	新闻、博客	英文维基百科	否	<b>是</b>
		KBP Entity Linking	2012-2014	英、 <b>中</b> 、西	新闻、博客	英文维基百科	否	<b>是</b>
		KBP Cold Start	2012-2014	英	新闻、博客	英文维基百科	否	<b>是</b>
		KBP Entity Discovery & Linking	2015	英	新闻、博客		<b>是</b>	<b>是</b>
NLPCC	CCF	Entity Linking	2013	<b>中</b>	微博	百度百科	否	否
		Entity Linking	2014	<b>中</b>	新闻、微博	中文维基百科	否	否
		Entity Linking	2015	<b>中</b>	新闻、微博	中文维基百科	<b>是</b>	否
SIGIR	Microsoft Google Yahoo!	ERD	2014	英	搜索引擎查询 新闻、网页	Freebase中有维基百 科链接的条目	<b>是</b>	否
WWW		Microposts NEEL	2014	英	微博	英文DBPedia	<b>是</b>	否

表 实体链接相关评测

# 实体链接的方法

简单的分类：

- 基于概率生成模型的方法
- 基于主题模型的方法
- 基于图的方法
- 基于深度学习的方法
- 无监督方法
- ...



## 基于概率生成模型的方法

- 人们在进行链接工作时，使用了大量关于实体的知识：
  1. 实体的知名度
  2. 实体的名字分布
  3. 实体的上下文分布
- 提出了实体-提及模型来融合上述异构知识

## 基于概率生成模型的方法

- 一个实体的名字通常是固定的，且以一定的概率出现。
- 指称的上下文与实体越匹配，则越可能链接到对应实体：

苹果 上下文包含 性能、续航等，则有可能指科技公司。

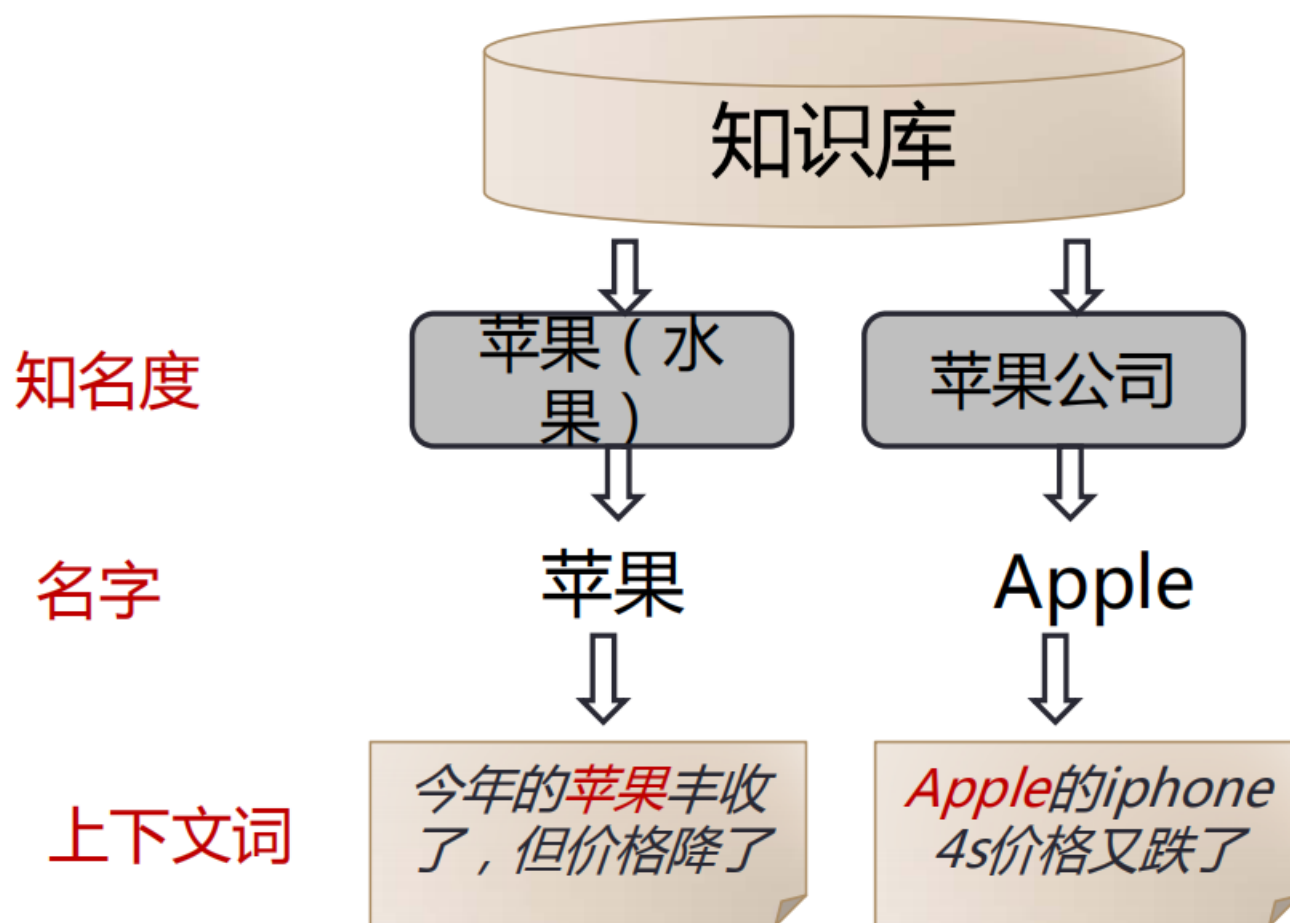
苹果 上下文包含 口感、色泽等，则有可能指水果。

## 基于概率生成模型的方法

- 利用 M&W 相似度可以计算出候选实体与上下文中其他实体的相关性。

$$\rho^{\text{MW}}(a,b) = 1 - \frac{\log(\max(|in(a)|, |in(b)|)) - \log(|in(a) \cap in(b)|)}{\log(|W|) - \log(\min(|in(a)|, |in(b)|))}$$

# 基于概率生成模型的方法



# 基于概率生成模型的方法

基于上述模型, 实体 $e$ 是提及 $m$ 目标实体的概率

$$P(m, e) = P(s, c, e) = P(e)P(s | e)P(c | e)$$

知名度

名字概率

上下文概率

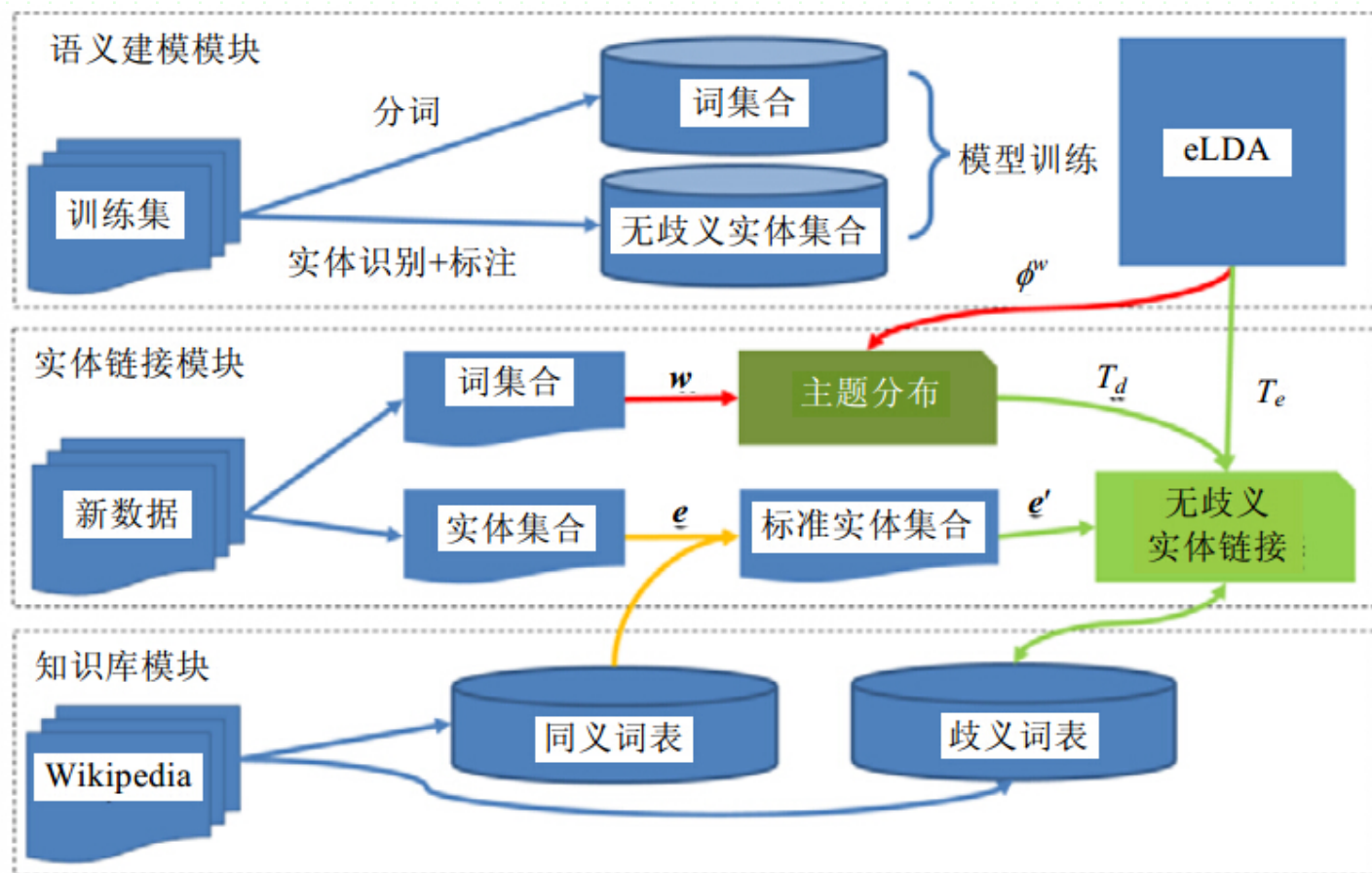
# 基于主题模型的方法

强假设：

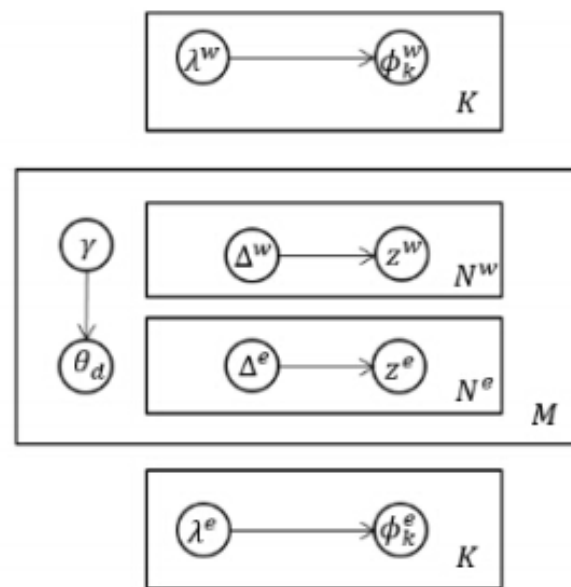
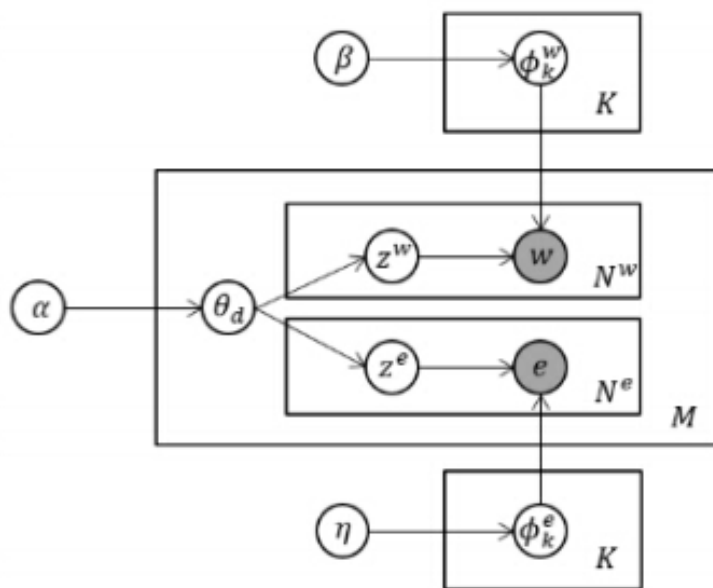
- 同一篇文本中的实体应当与文本的主题相关：

和科技、手机等有关文章也更有可能出现**苹果公司**，而不是**水果**。

# 基于主题的知识推理



# 基于主题的知识推理





# 基于主题的知识推理

概率主题模型属于生成模型,对于由  $M$  个文档组成的文档集中的一个包含  $N_d^w$  个词、 $N_d^e$  个实体的文档  $d$ , eLDA 认为,该文档的生成过程如下:

- 1) 对于每个主题  $k$ ,根据狄利柯雷(Dirichlet)分布,分别生成词和实体的在主题上的分布:

$$\phi_k^w \sim \text{Dir}(\beta), \phi_k^e \sim \text{Dir}(\eta)$$

- 2) 对每个文档集中的文档  $d \in \{1, \dots, M\}$ , 执行步骤 3)~步骤 9)
- 3) 为当前文档  $d$  生成主题分布  $\theta_d \sim \text{Dir}(\alpha)$
- 4) 对  $d$  中的每个词  $n \in \{1, \dots, N_d^w\}$ , 执行步骤 5)、步骤 6)
- 5) 为当前词生成主题  $z_{d,n} \sim \text{Mult}(\theta_d)$
- 6) 生成当前的词  $w_{d,n} \sim \text{Mult}(\phi_{z_{d,n}}^w)$
- 7) 对  $d$  中的每个实体  $n' \in \{1, \dots, N_d^e\}$ , 执行步骤 8)、步骤 9)
- 8) 为当前实体生成主题  $z_{d,n'} \sim \text{Mult}(\theta_d)$
- 9) 生成当前的实体  $e_{d,n'} \sim \text{Mult}(\phi_{z_{d,n'}}^e)$

上述过程中的  $\text{Dir}$  表示狄利克雷分布,  $\text{Mult}$  表示多项式分布.当给定参数  $\alpha, \beta, \eta$  时,所有观察变量(即观察到的文档中的词和实体)和隐变量(如每个主题上的实体分布)的联合概率公式为

$$P(d, z^w, z^e, \Theta, \Phi^w, \Phi^e, \alpha, \beta, \eta) = P(\Theta | \alpha) P(\Phi^w | \beta) P(\Phi^e | \eta) \times \left( \prod_{n=1}^{N_d^w} P(w_n | z^w, \Phi^w) \right) \left( \prod_{n=1}^{N_d^e} P(e_n | z^e, \Phi^e) \right) \quad (1)$$

## 基于图的方法

- 重启随机游走
- 实体相似度计算：根据实体属性值的数据类型使用不同相似度计算方法来度量它们之间的相似性再使用聚合函数初始化实体间的相似度矩阵。
- 图模型构建：根据实体类型，基于中计算得到的相似度确定候选链单元，将所有候选单元作为关联图中的顶点，再基于各实体间的语义关系，确定候选链单元间的关联（即生成关联图中的边）。

# 基于图的方法

## ● 重启随机游走

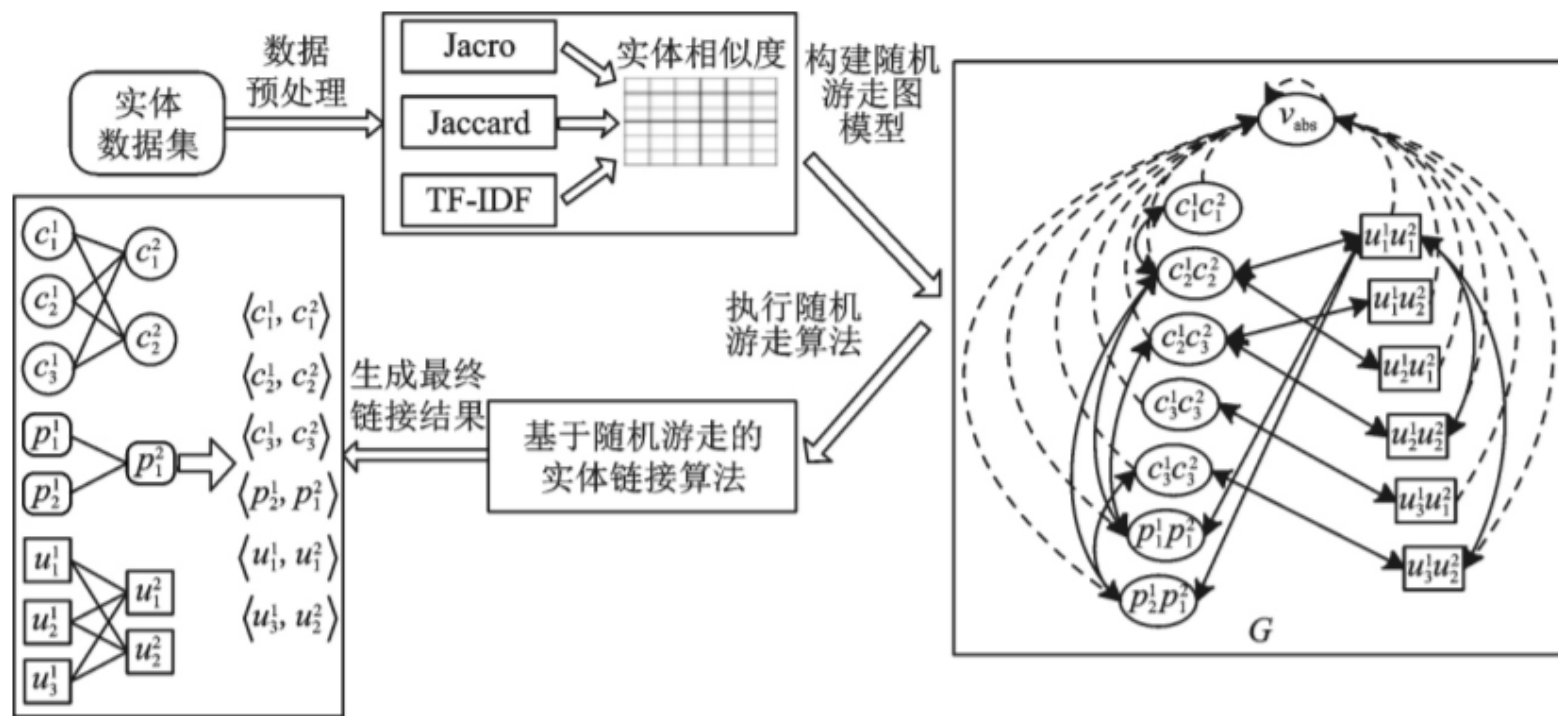


图 1 基于随机游走的实体链接模型

## 基于图的方法

- 重启随机游走:

基本思想是给定一个图，游走者从某个顶点或一系列顶点开始遍历该图。在任意一个顶点，游走者对于下一步行动有 2 种选择：

1. 以概率  $1 - c$  随机选择一条关联到当前顶点的边以游走到某个邻居顶点
  2. 以  $c$  的概率随机跳转到图中任意一个顶点
- 每次游走后，均将得到一个概率分布，将该概率分布作为下一次游走的输入，反复迭代。当满足一定前提条件时，该概率分布将会收敛到一个稳定值。

# 基于深度学习的方法

- 使用BiLSTM、CNN等

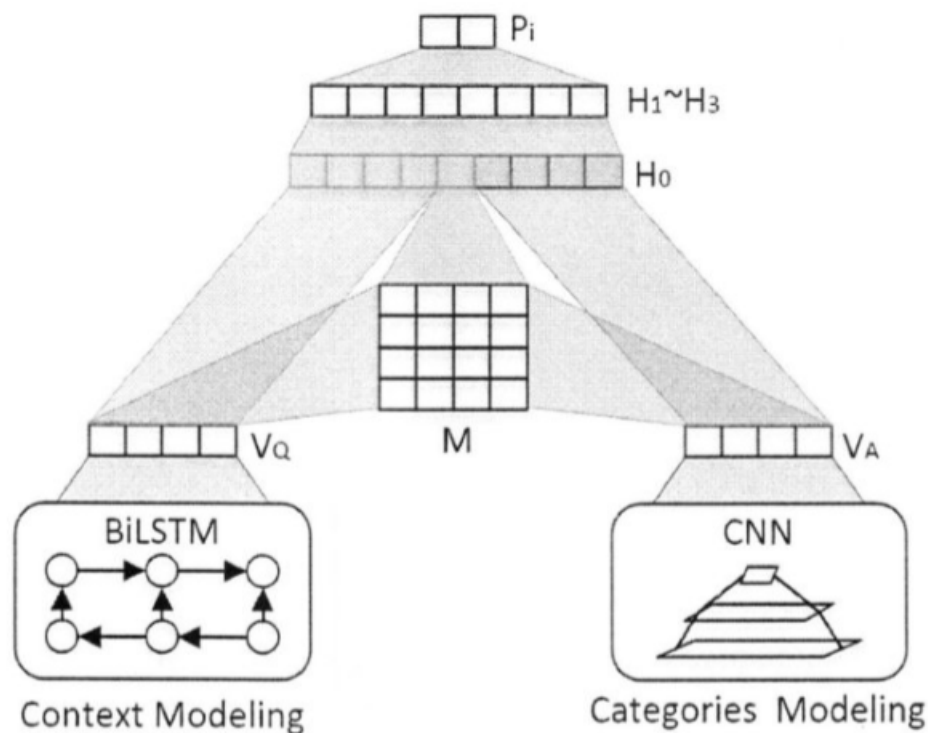


图 4.6 文本实体链接神经网络整体架构

# 基于深度学习的方法

- 借助BiLSTM、CNN等计算相似度后再基于图：

表 4.5 中文数据集上的实体链接对比试验结果

数据集		用例总数		文档总数		候选集平均大小	
训练集		743, 978		232, 493		6.032	
测试集		55, 716		15, 058		5.966	
LIEL		DSRM		Ours			
Micro	Macro	Micro	Macro	Micro	Macro		
0.7063	0.7189	0.7434	0.7296	<u>0.8107</u>	<u>0.8211</u>		
DSRM + Ours		Prior		Prior + RWR			
Micro	Macro	Micro	Macro	Micro	Macro		
0.7776	0.7821	0.6983	0.6844	0.7191	0.7123		

# 基于无监督的方法

## AAAI 18 阿里 Colink

- 协同训练算法:
- 在该框架中定义两个不同的模型：一个基于属性的模型 **fatt** 和一个基于关系的模型 **frel**。
- 这两个模型会进行二元分类预测，将一组给定实体对分类为正例（链接的）或负例（非链接的）。
- 该协同训练算法以迭代的方式不断增强这两个模型。




# 基于无监督的方法

## AAAI 18 阿里 Colink

**Input:** a source social network  $G^s$ , a target social network  $G^t$

**Output:** a set of user pairs  $S$

```
1  $S \leftarrow$  the set of seed pairs generated with seed rules;
2 repeat
3   /* generate pairs from attribute-based model */
4    $D_{att} \leftarrow f_{att}(S, G^s, G^t)$ ;
5   /* generate pairs from relationship-based model */
6    $D_{rel} \leftarrow f_{rel}(S, G^s, G^t)$ ;
7   /* join two sets and remove conflicting pairs */
8    $D \leftarrow merge(D_{att}, D_{rel})$ ;
9    $S \leftarrow S \cup D$ ;
10 until  $D = \emptyset$ ;
11 return  $S$ ;
```

 产业智能官



# 基于无监督的方法

- AAAI18 阿里 Colink (和无监督的方法对比)

Table 2: Performance comparison of different approaches.

Method	P	R	F1
Random-select	49.31	54.21	51.64
SiGMa	91.00	44.28	59.57
Alias-disamb	82.35	58.92	68.69
CoLink (S2S+Coarse-tuned)	86.74	83.67	85.18
CoLink (S2S+Coarse-tuned+Rev)	89.51	86.20	87.82
CoLink (S2S+Fine-tuned)	<b>91.47</b>	<b>86.70</b>	<b>89.02</b>
CoLink (S2S+Fine-tuned+Rev)	89.22	86.36	87.77
CoLink (SVM+Fine-tuned)	84.16	62.63	71.81

# 实体链接总结与展望

- 跨语言实体链接
- 利用实体链接促进自然语言处理任务
- ...

धन्यवाद

Hindi

多謝

Traditional Chinese

ขอบพระคุณ

Thai

Спасибо

Russian

Gracias

Spanish

شكراً

Arabic

*Thank You*

English

Obrigado

Brazilian Portuguese

Grazie

Italian

多谢

Simplified Chinese

Danke

German

Merci

French

நன்றி

Tamil

ありがとうございました

Japanese

감사합니다

Korean