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

知识链接：从文本到概念

韩先培
xianpei@nfs.iscas.ac.cn

中文信息处理研究室，中国科学院软件研究所

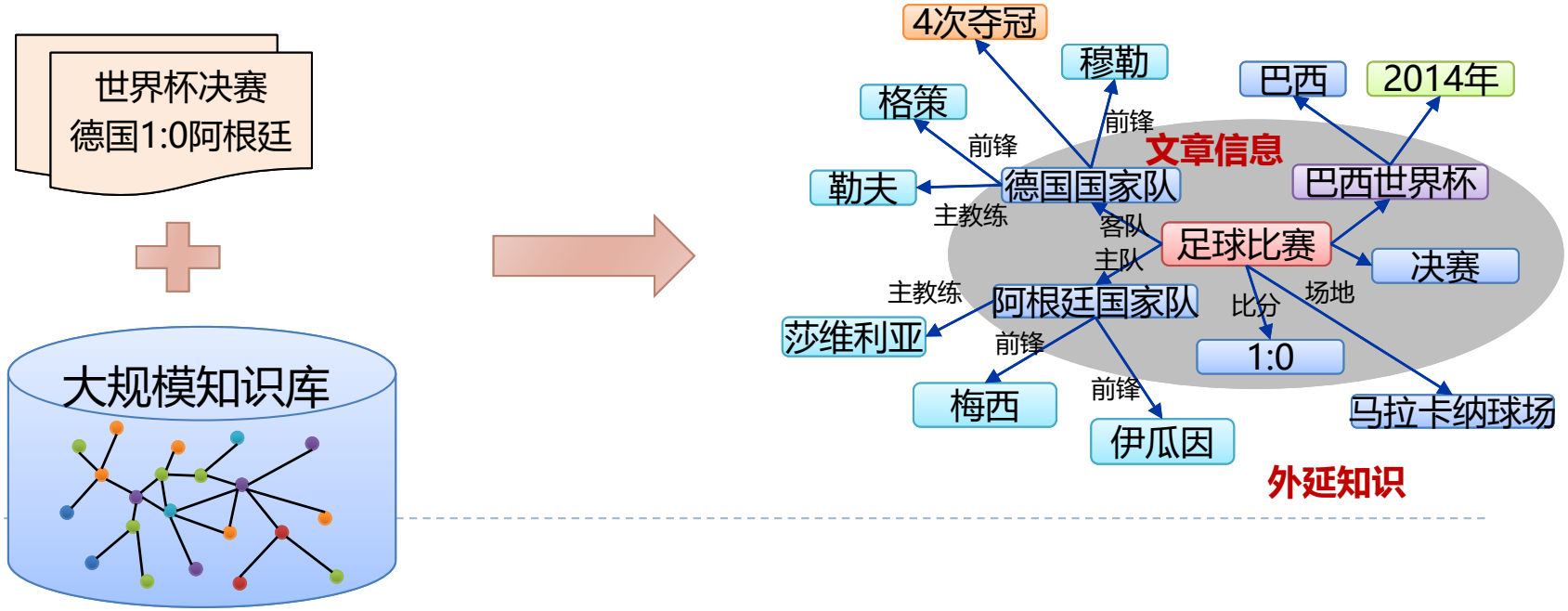
Part3: 2016-10-14 烟台 CCL 2016

提纲

- 任务
- 关键技术
  - 引用表构建
  - 实体知识挖掘与表示
  - 链接推理算法
- 总结及展望

背景：基于知识的文本语义理解

- 知识图谱包含了大量的世界知识，实体链接是实现了文本到知识的连接，是基于知识的文本语义理解的基础技术

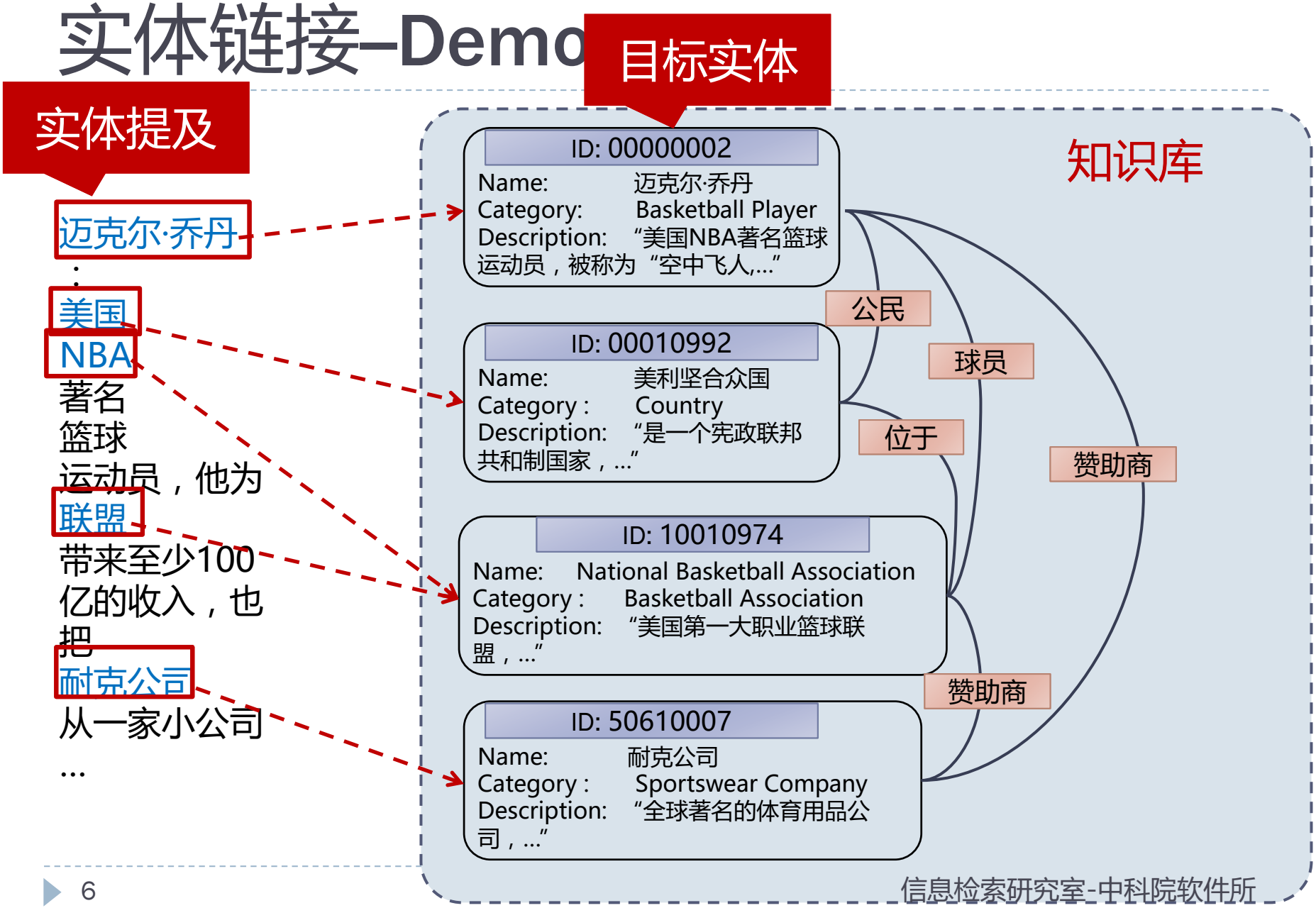


实体链接任务

1. 给定一系列需要链接的**文本提及（mentions）**
2. 和需要链接到的**目标知识库 (KB)** ，如Wikipedia, Freebase,Yago,…
3. 确定文本提及与目标知识库中实体的**一一对应关系**



实体链接–Demo


实体提及
迈克尔·乔丹
.
.
美国
NBA
著名
篮球
运动员，他为
联盟
带来至少100
亿的收入，也
把
耐克公司
从一家小公司
…
目标实体
知识库
ID: 00000002
Name: 迈克尔·乔丹
Category: Basketball Player
Description: “美国NBA著名篮球运动员，被称为“空中飞人,…”
ID: 00010992
Name: 美利坚合众国
Category : Country
Description: “是一个宪政联邦共和制国家，…”
ID: 10010974
Name: National Basketball Association
Category : Basketball Association
Description: “美国第一大职业篮球联盟，…”
ID: 50610007
Name: 耐克公司
Category : Sportswear Company
Description: “全球著名的体育用品公司，…”
公民
球员
位于
赞助商
赞助商

实体链接子任务

- 实体提及识别
  - *S: 在旧金山的发布会上，苹果为开发者推出新编程语言Swift*
    → {“旧金山”，“苹果”，“Swift”}
- 候选目标实体选取
  - “苹果” → {*苹果(水果)，苹果公司，苹果(电影)，苹果(银行)*, …}

核心研究问题

- 实体知识挖掘及表示
  - 苹果公司上下文：{开发，iOS，旧金山，编程…}
  - 苹果水果上下文：{维生素，价格，美味，甜，…}
- 链接推理
  - Sim(“苹果”, S, *苹果公司*) = 0.7 **正确**
  - Sim(“苹果”, S, *苹果(水果)*) = 0.1 **错误**

提纲

- 任务
- 关键技术
  - 引用表构建
  - 实体知识挖掘与表示
  - 链接推理算法
- 总结及展望

引用表

- 引用表存储：
  - 所有可用于链接的名字
  - 名字 → 实体 的映射关系
- 构建方法：锚文本挖掘，重定向页面，消歧义页面，缩写扩展

名字	目标实体	次数
苹果	水果苹果	10000
	苹果公司	3000
	苹果电脑	2030
	苹果《电影》	200
	苹果银行	10
AI	Artificial intelligence	581
	Game artificial intelligence	48
	Ai (singer)	10
	Angel Investigations	9
	Strong AI	3
	…	…

引用表

- 基于引用表的提及识别
  - 识别文本中所有的ngram，如果一个ngram是引用表中的名字，则识别为一个提及
  - 通常会加一些启发式规则过滤
    - 必须是一个名词短语
    - Keyphraseness
- 基于引用表的候选实体选取
  - 直接查表
- 趋势：Joint NER and EL，降低错误传递

苹果在开始重点进军AI方向

名字	目标实体	次数
苹果	水果苹果	10000
	苹果公司	3000
	苹果电脑	2030
	苹果《电影》	200
	苹果银行	10
AI	Artificial intelligence	581
	Game artificial intelligence	48
	Ai (singer)	10
	Angel Investigations	9
	Strong AI	3
	…	…

提纲

- 任务
- 关键技术
  - 引用表构建
  - 实体知识挖掘与表示
  - 链接推理算法
- 总结及展望

实体链接的核心科学问题

- 实体名的歧义性：
  - 中关村的苹果不错 → 苹果电脑
  - 新发地的苹果不错 → 水果苹果

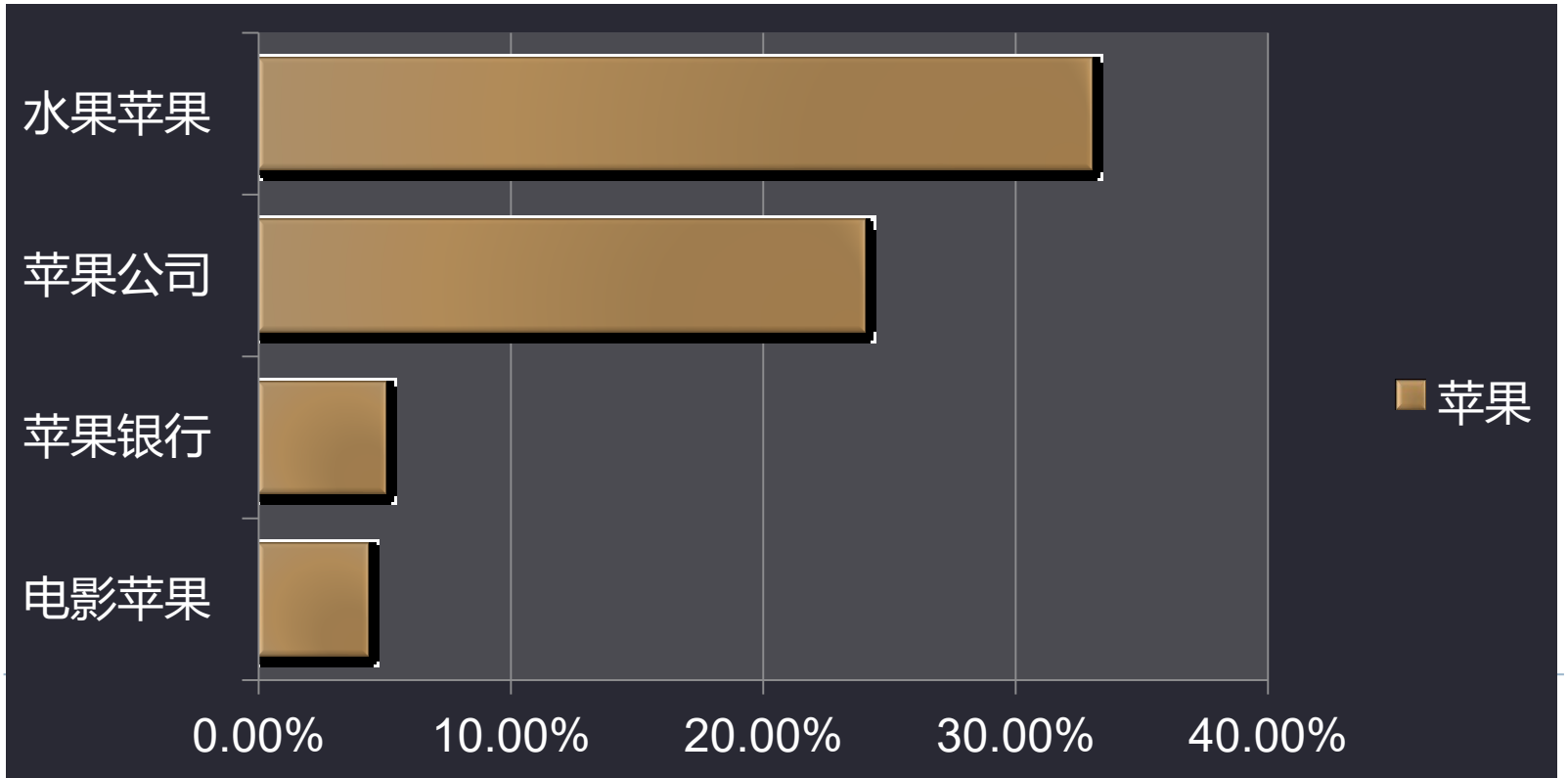
- 需要有实体相关的知识来帮助我们消除实体的歧义
  - 中关村→ IT → 苹果电脑
  - 新发地 → 菜市场批发 → 水果苹果

可用于实体链接的实体知识

- 实体知名度
- 实体上下文
- 实体语义关联度
- 文章主题
- 知识网络

实体知名度

- 表示一个实体被人们知道了解的程度
- 高知名度的实体更有可能在文章中被提起
- 通常被建模为一个先验概率P(e)



实体上下文

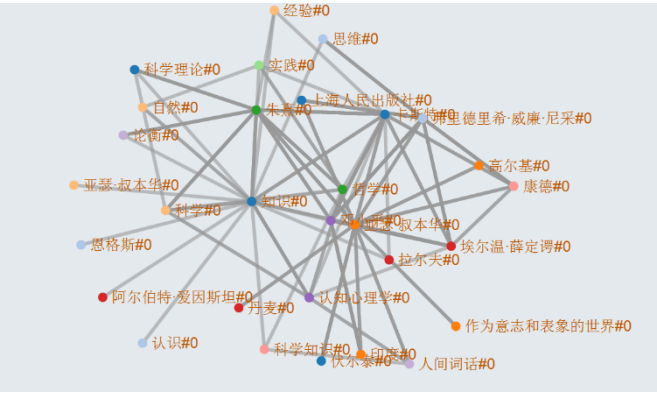
- 特定实体的上下文规律性
  - 周围出现iPad，酷，视网膜屏的苹果更可能是苹果公司
  - 周围出现好吃、甜、一斤的苹果更可能是水果苹果
- 可建模为特征向量，词分布语言模型

苹果公司

苹果银行

实体语义关联度

- 捕捉实体和实体之间的语义关系
- 相关实体更容易同时出现在一篇文章中
  - 苹果电脑与乔布斯、iPad、iPhone、库克
  - 苹果与葡萄、桃子、苹果汁、酒
  - 苹果电影与范冰冰、华星、华谊
- 如何衡量两个实体之间的相关度
  - 在知识网络中的距离
  - 在文章中共现的次数



	贝叶斯网络	芝加哥公牛
机器学习	0.74	0.00
NBA	0.00	0.71

文章主题

- 一篇文章中的实体应当与其主题相关
  - 苹果公司更容易出现在IT相关主题的文档中
  - 水果苹果更容易出现在吃或农业相关的文档中
  - 电影《苹果》倾向于出现在娱乐相关的新闻中
- 通常被建模为文章的Topic分布

苹果公司

计算机	媒体	软件
Topic(**Computer**)	Topic(**Video**)	Topic(**Software**)
Computer	*Video*	*Computer software*
CPU	*Mobile phone*	*Microsoft Windows*
Hardware	*Mass media*	*Linux*
Personal computer	*Music*	*Web browser*
Computer memory	*Television*	*Operating system*

Table 2. The 3 topics where the *Apple Inc.* has the

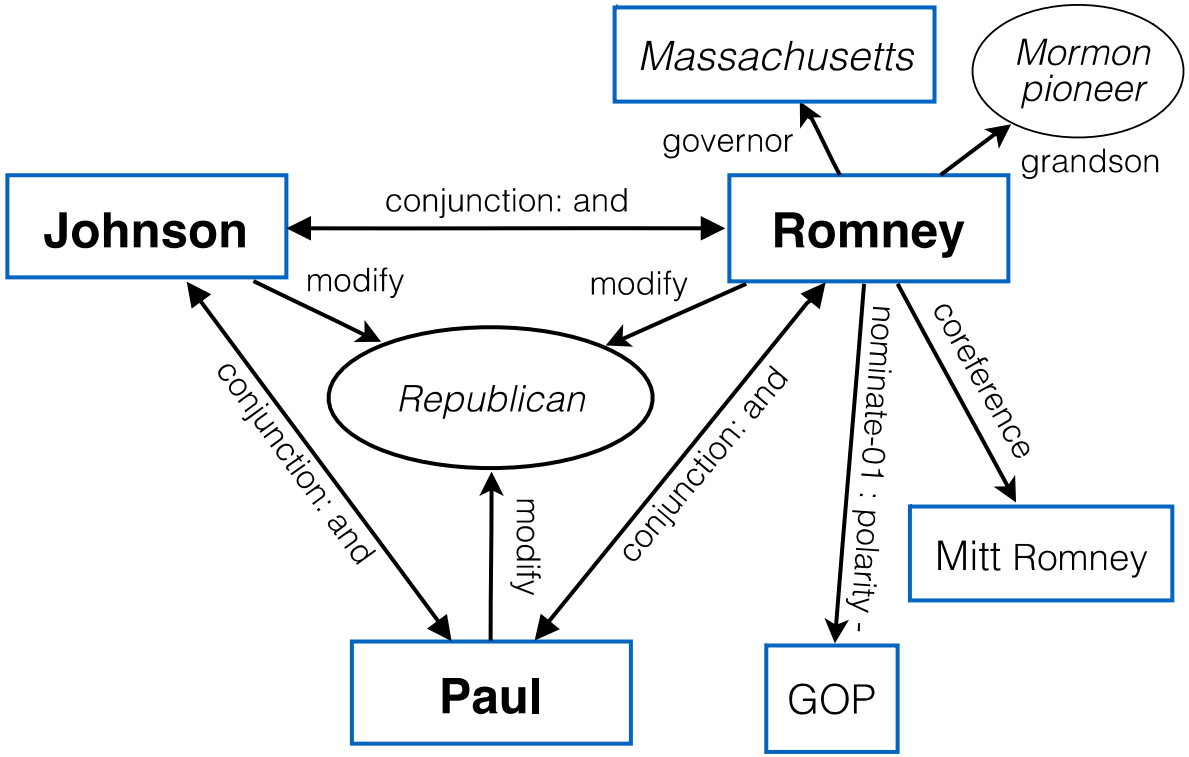
水果苹果

酒	食物	植物
Topic(**Wine**)	Topic(**Food**)	Topic(**Plant**)
Wine	*Food*	*Plant*
Grape	*Restaurant*	*Flower*
Vineyard	*Meat*	*Leaf*
Winery	*Cheese*	*Tree*
Apple	*Vegetable*	*Fruit*

Table 3. The 3 topics where the fruit *Apple* has the

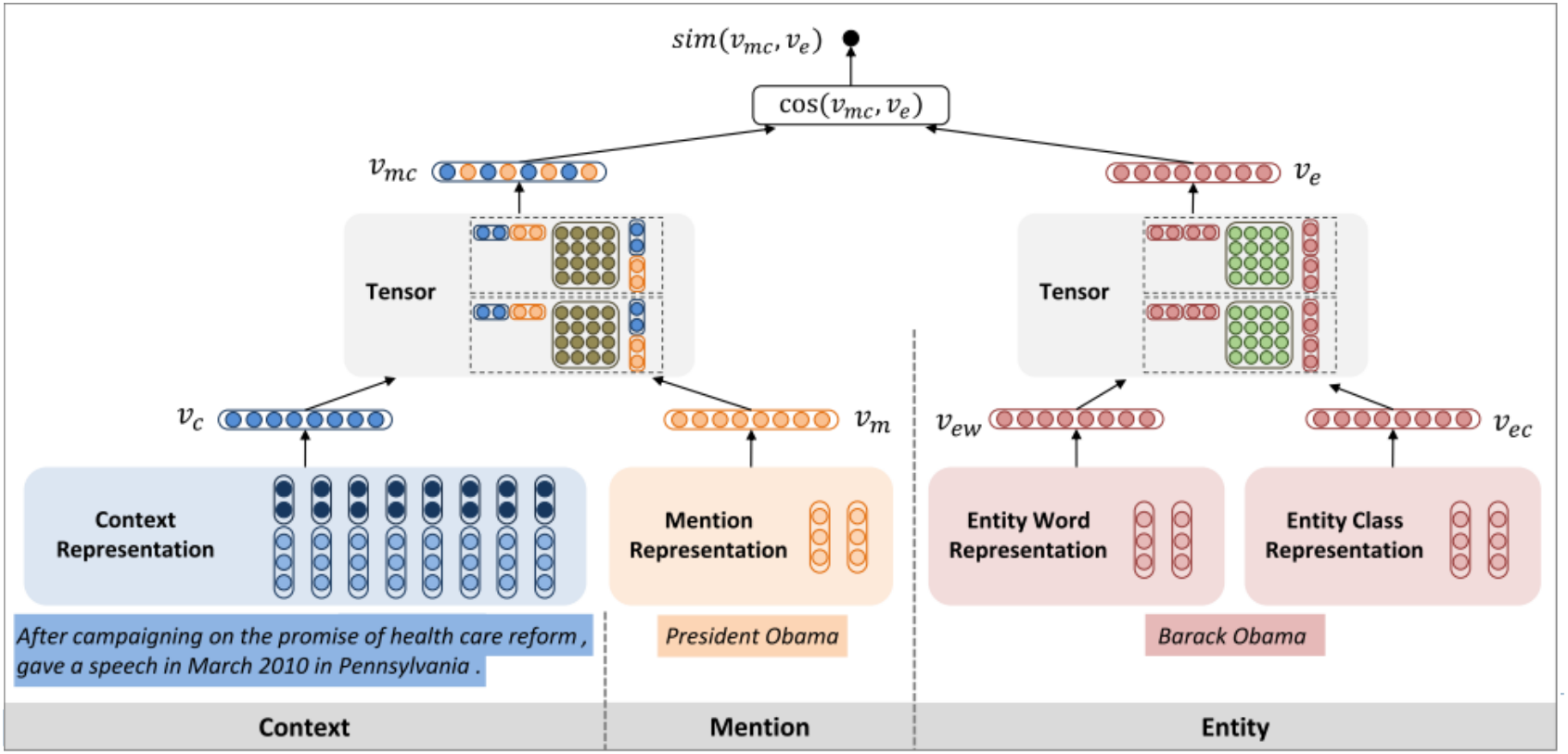
知识网络

- 利用实体与文章中其它实体之间的语义关系来表示实体(Pan et al., 2015)
  - 使用AMR parsing来获取实体间关系



基于深度学习的统一表示学习

- 将实体、上下文、提及等文本信息利用NN映射到连续低维空间中（He et al., ACL 13, Sun et al., IJCAI’15,…）
- 使用NN来同时考虑多方面信息之间的组合、转换和交互



一些有用但是难以构建的知识

- 作者知识领域
- 倾向性偏好
  - 科黑 vs 科密
- 媒体偏好
  - 新华社 vs 微博
- 地点偏好
  - 北京 vs 广州
  - 宿舍 vs 图书馆
- 精准度与构建成本之间的权衡

提纲

- 任务
- 关键技术
  - 引用表构建
  - 实体知识挖掘与表示
  - 链接推理算法
- 总结及展望

链接推理算法

- 链接推理算法就是综合实体知识进行决策的过程
- 中关村的苹果不错→ 水果苹果？苹果电脑？

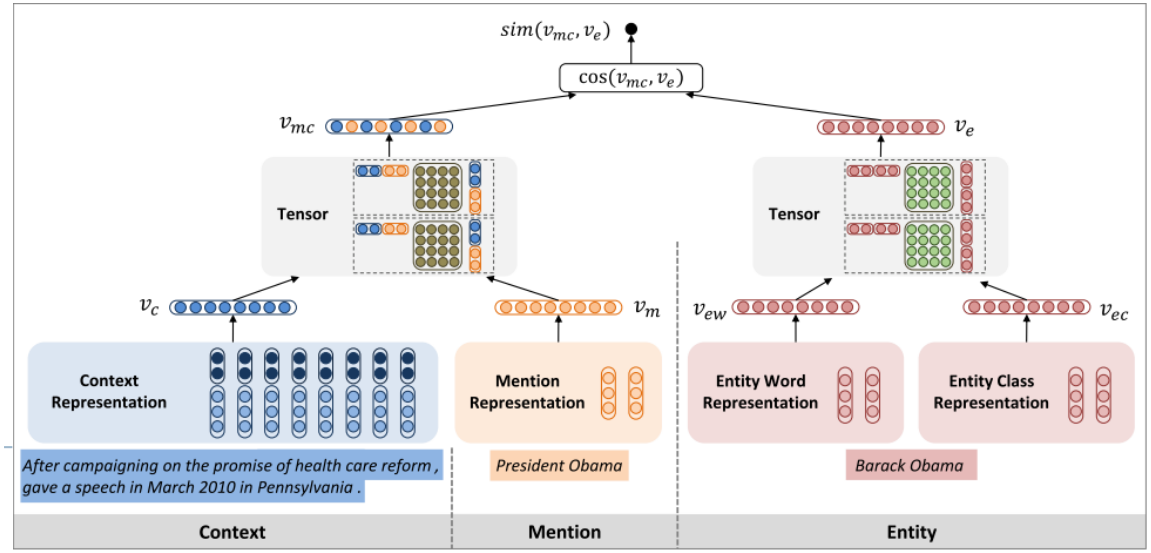
- 推理算法
  - 局部推理
  - 全局推理

局部推理

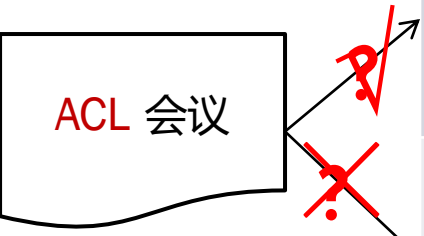
- 考虑单个实体的上下文，不考虑文章中其它实体对该实体的影响
  - 文本相似度
  - 统计模型
  - Learning To Rank

文本相似度推理

- 将提及m和候选实体e分别表示为特征向量
- 选择最大化sim(m, c, e)的实体作为目标链接对象
- 中关村的苹果不错→ 水果苹果？苹果电脑？
  - 水果苹果和苹果电脑的上下文
  - 相关度(中关村，水果苹果) = 0.1
  - 相关度(中关村，苹果电脑) = 0.7



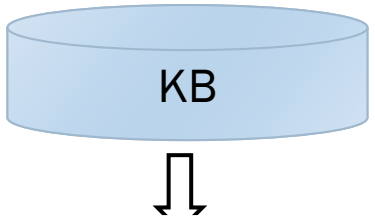
统计模型：建模人综合知识进行消歧的过程



候选链接实体	知名度	名字知识	上下文
Association for Computational Linguistics-计算语言学协会	经常说起 ACL协会	通常被叫做 ACL	ACL协会每年组织一次 会议
Atlantic City Line-亚特兰大城市快轨	几乎不谈起 亚特兰大城市快轨	只在车站代码上叫做 ACL	Atlantic City Line 从没组织过会议

实体-提及模型(EM Model)

在实体-提及模型中,每一个命名性提及m都被建模为通过下述生成过程(generative story)产生的样本：



1. EM Model根据实体的知名度 P(e)选

实体的知名度知识、名字知识和上下文知识依次被建模为概率分布P(e), P(s|e), P(c|e)

3. EM Model根据实体的上下文知识P(c|e)输出提及m的上下文c

乔丹在1984年加入NBA

基于实体-提及模型的实体链接

- 基于上述模型, 实体e是提及m目标实体的概率:

$$P(m,e) = P(s,c,e) = P(e)P(s \mid e)P(c \mid e)$$

- 模型选择能最大化条件概率P(e|m)的实体e作为其提及m的目标实体

$$\mathrm{e} = \arg\max_{e} \frac{P(m,e)}{P(m)} = \arg\max_{e} P(e)P(s \mid e)P(c \mid e)$$

Learning to Rank(Zheng et al., 2010)

- 将所有不同匹配evidence都建模为不同的特征
- 使用Learning to Rank算法来综合不同的特征对最终排序的影响

Set	Features in Set
Set1	Surface Features
Set2	Set1+TF-IDF Features
Set3	Set2+AllWordsInSource
Set4	Set3+NENumMatch
Set5	Set4+CountryInTitle Features
Set6	Set5+CountryInText Features
Set7	Set6+CityInTitleMatch
Set8	Set7+MatchType

Table 2: Feature Sets

Algorithm	Accuracy	Improvement over SVM
ListNet	**0.9045**	+18.5%
Ranking Perceptron	0.8842	+15.8%
SVM	0.7636	-
Perceptron	0.7546	-1.2%

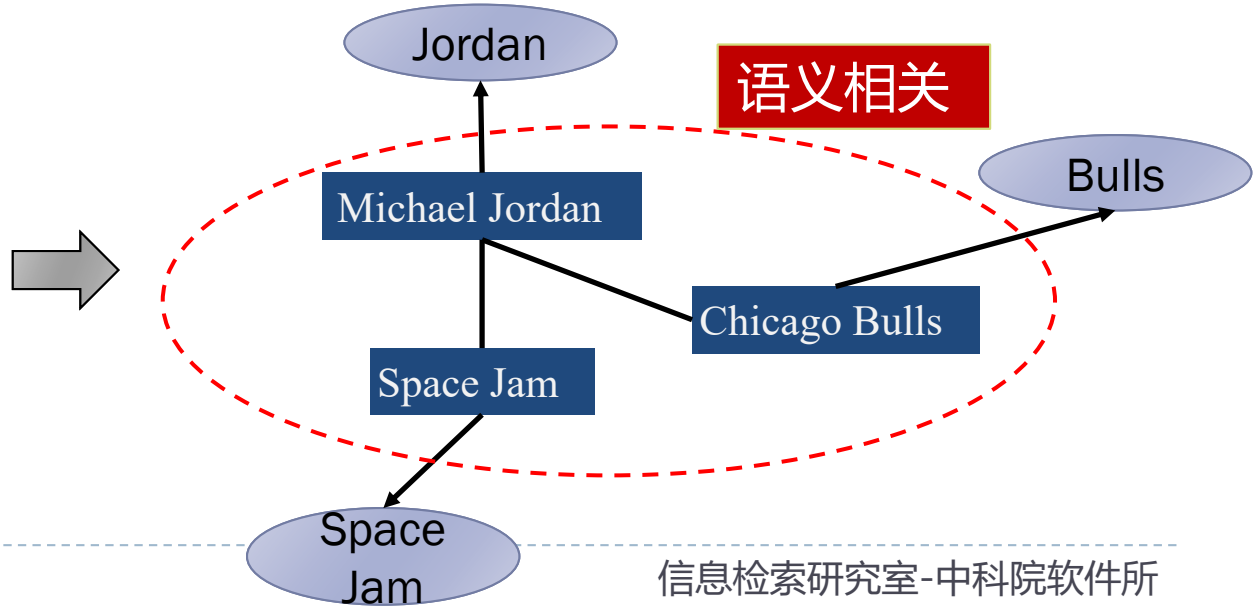
全局推理

- 单篇文本中的实体相互关联，全局推理算法进一步考虑不同实体链接决策之间的相互关联，从而提升性能
  - 基于图的全局推理算法
  - 基于统计的全局推理算法

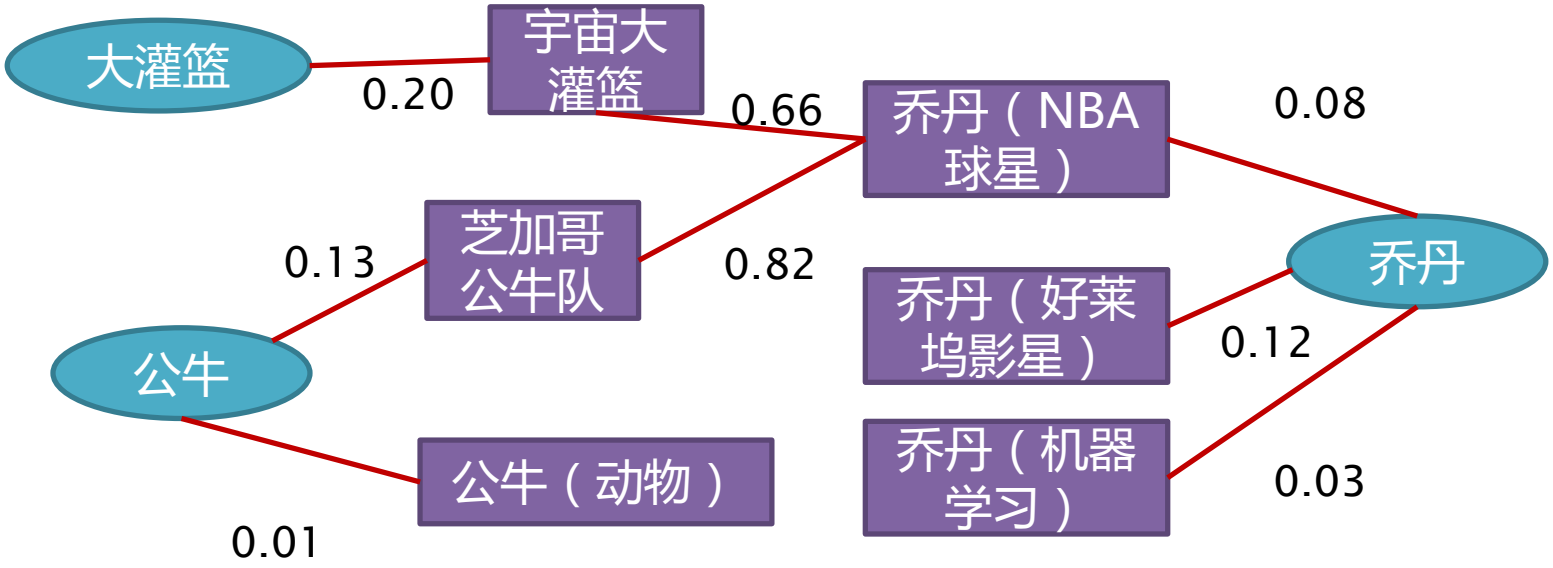
全局链接

利用目标实体之间的语义关联，协同链接单篇文本内的所有提及能有效提升实体链接性能

During his standout career at Bulls, Jordan also acts in the movie Space Jam.

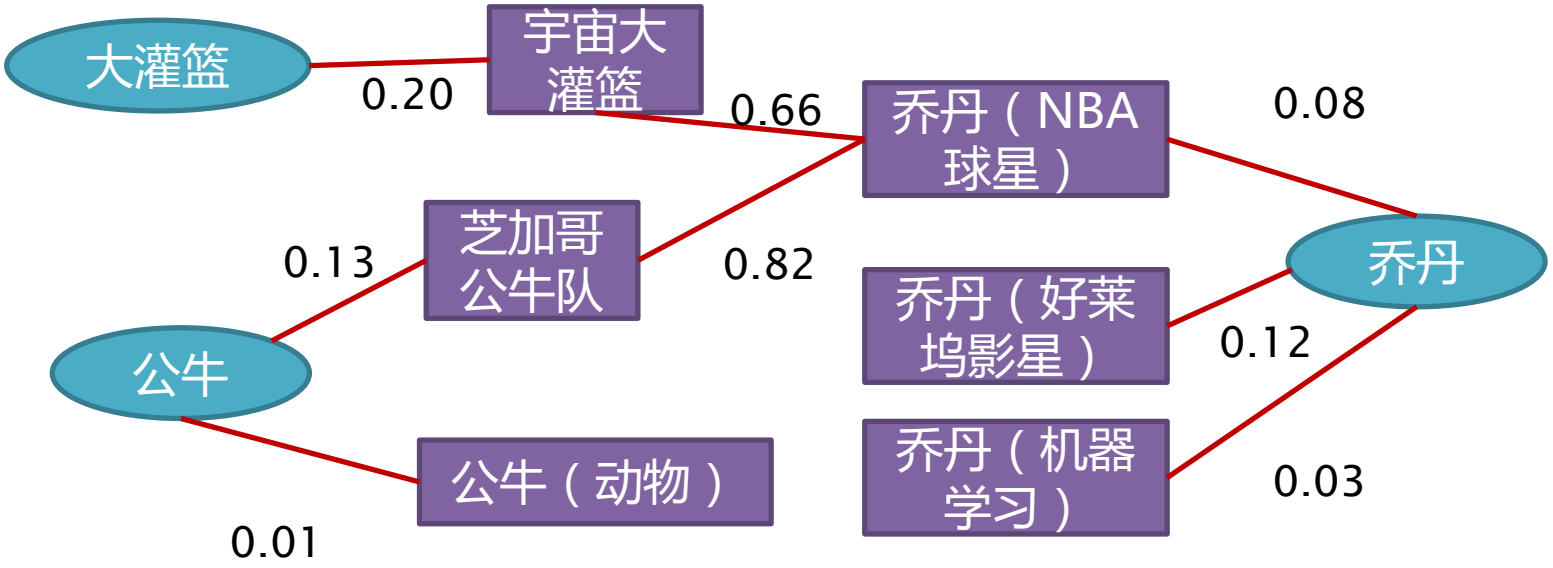


实体链接代表方法—图方法



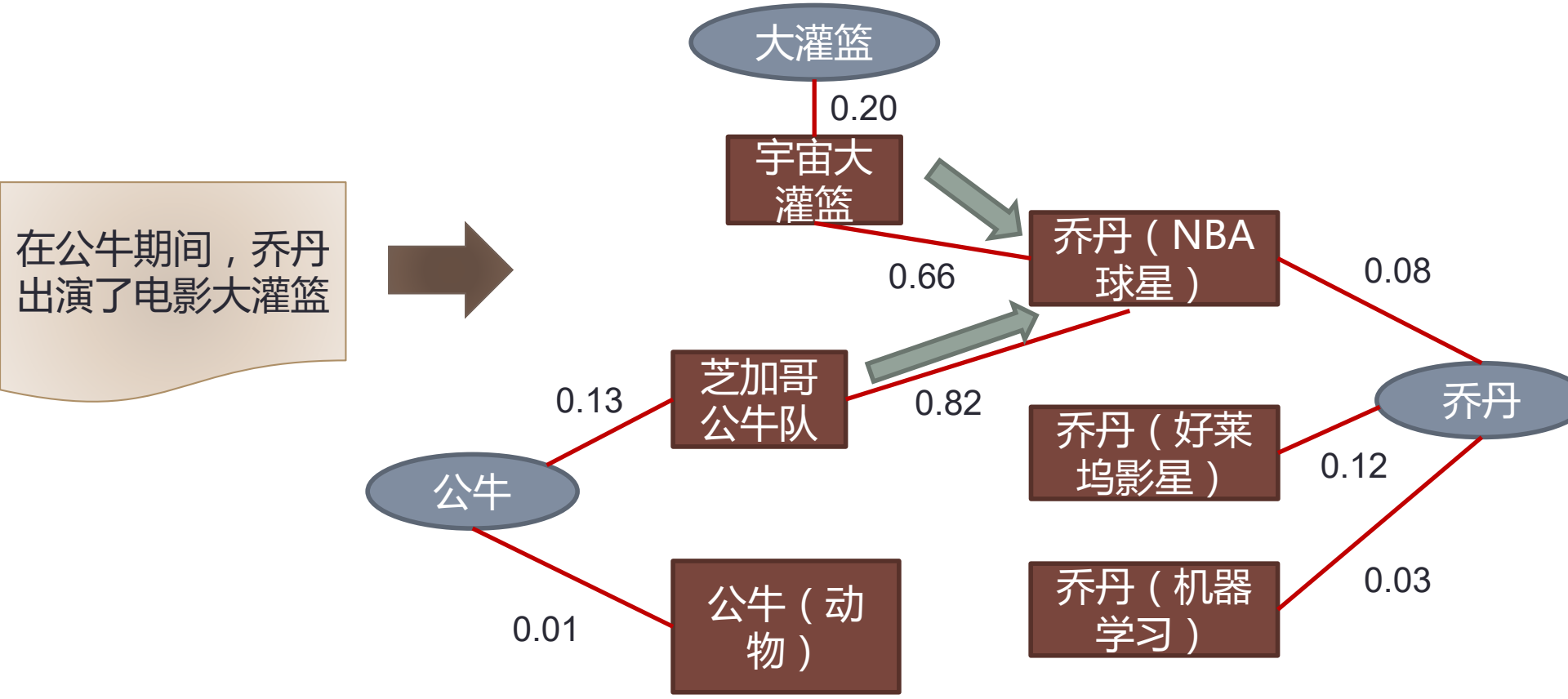
- 使用知识库中的知识来构建mention-entity graph
- 构建算法来计算最大似然链接结构
  - 同时考虑mention-entity的一致性和entity-entity之间的语义关联
  - 保证每一个mention指向且只指向一个目标实体

实体链接代表方法—图方法



- 计算最大似然链接结构的算法
  - **寻找具有最大似然值的子图/最稠密子图**（Chakrabarti et al.: KDD’09，Hoffart et al., EMNLP’11,…）
  - **基于Graph Ranking寻找最大可能节点**（Han et al., SIGIR’11, Alhelbawy and Gaizauskas, ACL’14…）

基于图的协同推断Demo-RWR

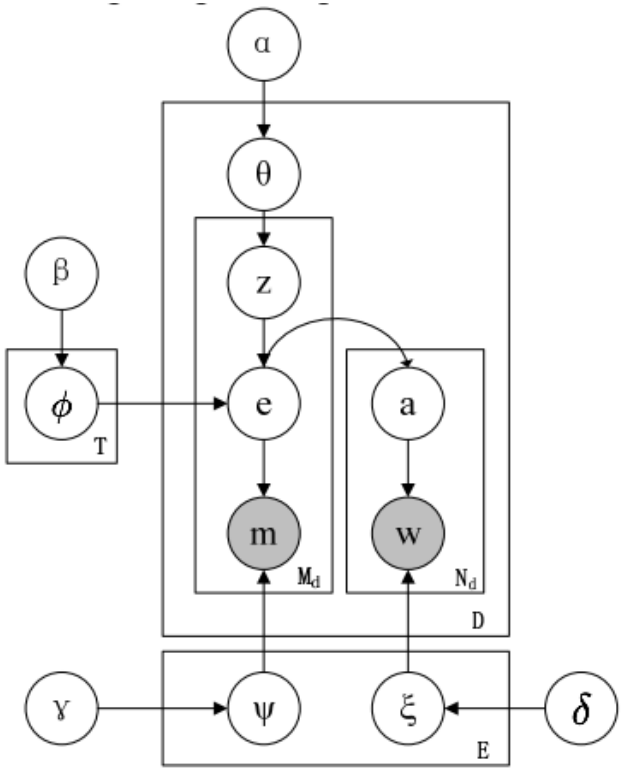


$$r^{t+1} = (1-\lambda) \times T \times r^{t} + \lambda \times s$$

- r^{t+1}：在时间 t+1的证据
- T：证据传递率矩阵(Referent Graph的归一化相邻矩阵)
- λ：证据重分配率
- s：初始证据

基于统计的全局推理算法

- 构建一个遵循如下原则的文档生成过程
  - 主题一致性假设: *一篇文章中的所有实体都围绕他的主题.*
  - 上下文一致性假设: *一个实体的上下文词都与该实体一致.*
- 实体主题模型
  - 将文档建模为一个文档-主题-实体-词的结构



The Generative Story (Topic)

- 我们通过如下文档的生成过程来展示entity-topic model:

> *At the WWDC conference, Apple introduces its new operating system release - Lion.*

Step 1: The model generates the topic distribution of the document as $\theta_d = \{Apple\ Inc.^{0.45}, Operating\ System(OS)^{0.55}\}$;

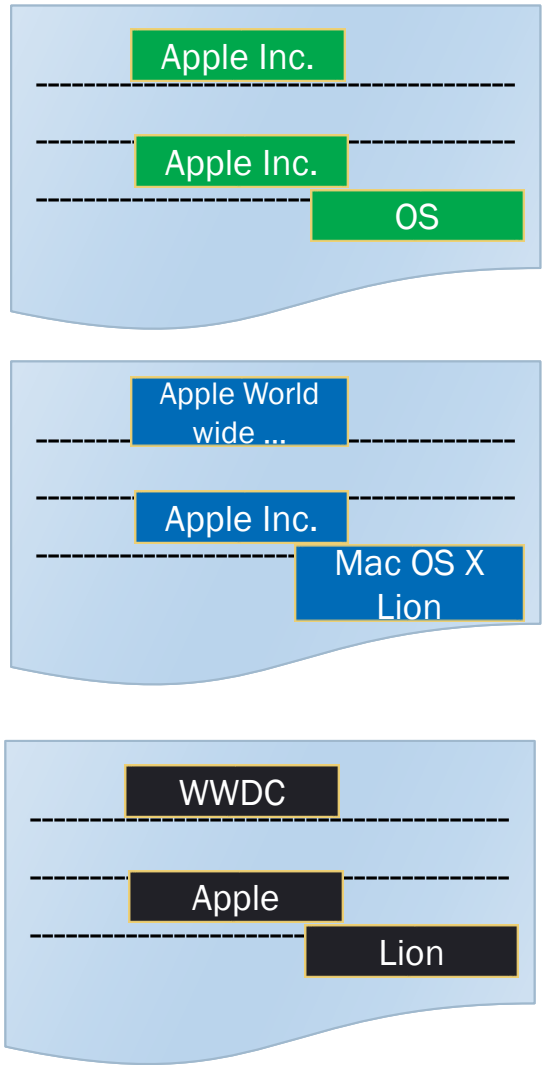
The Generative Story (Mention)

Step 2: For the three mentions in the document:

i. According to the topic distribution θ_d, the model generates their topic assignments as z_1=*Apple Inc.*, z_2 = *Apple Inc.*, z_3 = *OS*;

ii. According to the topic knowledge $\phi_{Apple\ Inc.}$, ϕ_{OS} and the topic assignments z_1, z_2, z_3, the model generates their entity assignments as e_1 = *Apple Worldwide Developers Conference*, e_2 = *Apple Inc.*, e_3 = *Mac OS X Lion*;

iii. According to the name knowledge of the entities *Apple Worldwide Developers Conference*, *Apple Inc.* and *Mac OS X Lion*, our model generates the three mentions as m_1=*WWDC*, m_2 = *Apple*, m_3 = *Lion*;

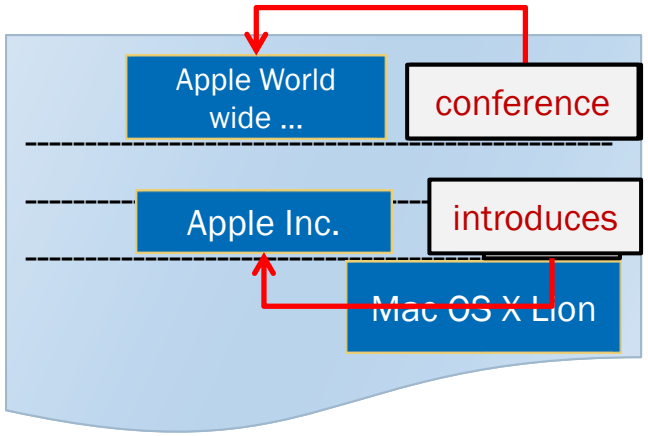


The Generative Story (Word)

Step 3: For all words in the document:

i. According to the referent entity set in document $\mathbf{e_d}$ = {*Apple Worldwide Developers Conference*, *Apple Inc.*, *Mac OS X Lion*}, the model generates the target entity they describes as a_3=*Apple Worldwide Developers Conference* and a_4=*Apple Inc.*;

ii. According to their target entity and the context knowledge of these entities, the model generates the context words in the document. For example, according to the context knowledge of the entities *Apple Worldwide Developers Conference*, the model generates its context word w_3 =*conference*, and according to the context knowledge of the entity *Apple Inc.*, the model generates its context word w_4 = *introduces*.

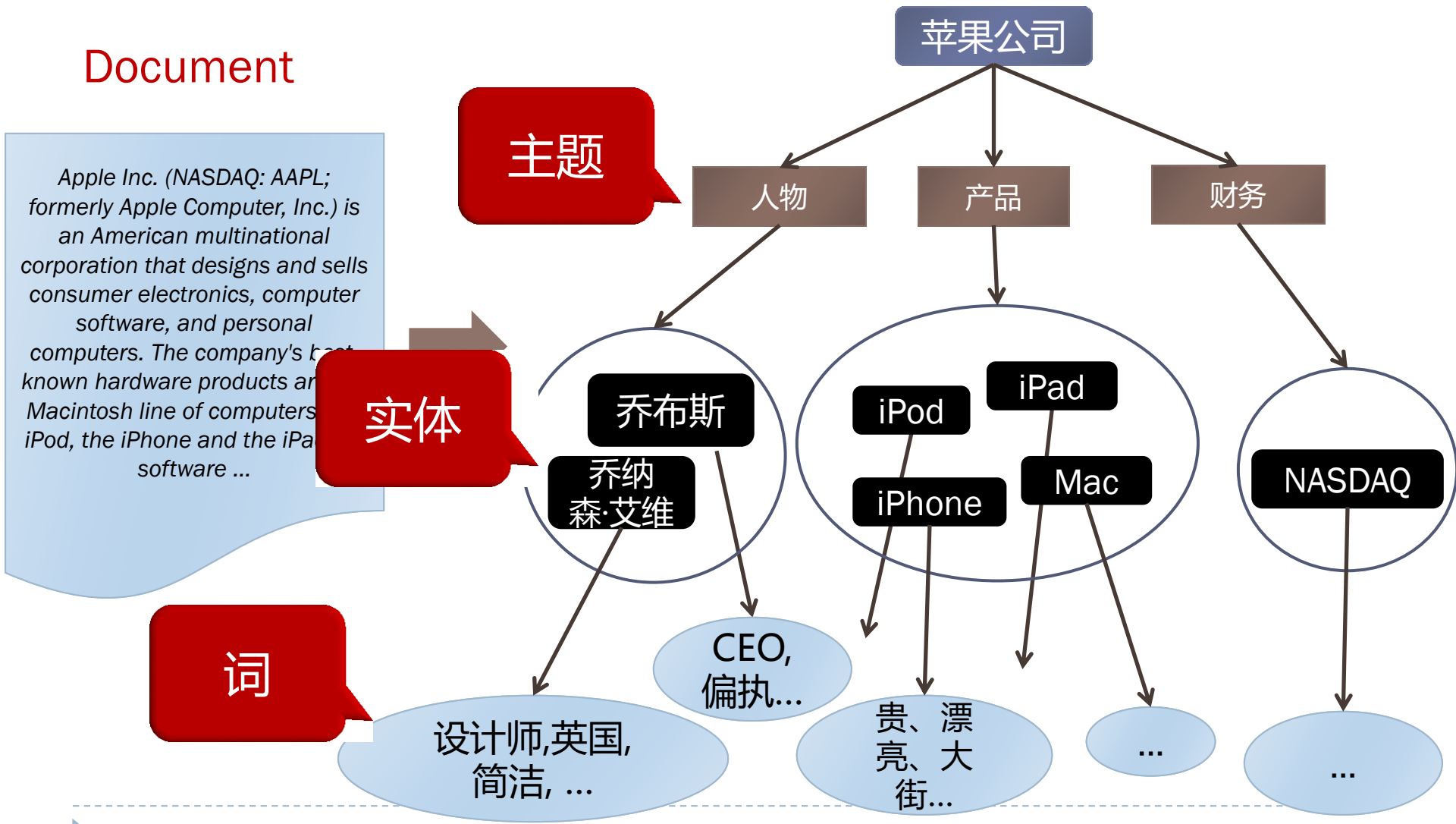


篇章主题-实体-上下文词的协同推断

内在结构

Document

Apple Inc. (NASDAQ: AAPL; formerly Apple Computer, Inc.) is an American multinational corporation that designs and sells consumer electronics, computer software, and personal computers. The company's best known hardware products are Macintosh line of computers, iPod, the iPhone and the iPad, software ...



总体性能

	Precision	Recall	F1
Wikify!	*0.55*	*0.28*	*0.37*
EM-Model	*0.82*	*0.48*	*0.61*
M&W	*0.80*	*0.38*	*0.52*
CSAW	*0.65*	*0.73*	*0.69*
EL-Graph	*0.69*	*0.76*	*0.73*
Our Method	***0.81***	***0.80***	***0.80***

Table 1. The overall results on IITB data set

局部推理

全局推理

提纲

- 任务
- 关键技术
  - 引用表构建
  - 实体知识挖掘与表示
  - 链接推理算法
- 总结及展望

总结

- 实体链接是一项解决自然语言歧义和多样性的有效技术
  - 知识图谱补全、 基于知识的文本语义理解的基础技术
- 主要研究：
  - 更多类型、更准确的异构实体知识挖掘和表示
  - 更合理、精准、快速的任务建模和推理算法
- 主要子任务
  - 实体提及识别（引用表）
  - 候选目标实体选取
  - 实体知识挖掘及表示（知名度、上下文、文本主题、知识网络、相关度…）
  - 链接推理（局部，全局）

热点及展望

- **多任务的联合**：实体识别，实体链接，关系抽取的联合…
- **实体链接系统的效率**：应用于Web环境下必须要考虑的因素（He et al., 2011；Lin & Etzioni, 2012）
- **多类型多模态上下文和知识的统一建模**：文本、图片、音频？
- **NIL实体检测和长尾实体的链接**：知识库未覆盖的实体链接仍然是开放问题
- **面向特定领域的和面向多知识库的实体链接**：目前大部分研究集中在Wikipedia，与领域知识库相关的研究较少（如电影领域的IMDb，书籍领域的豆瓣，餐馆领域的大众点评等等）Wang et al.(2012)和Zhang et al.(2016)
- **面向特定情境的实体链接**: 其它情境下（例如微博、评论、列表页面等）的实体链接研究不足：Twitter（Guo et al., 2013），Web列表的实体链接（Shen et al., 2012）

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