实体消歧

赵 军（jzhao@nlpr.ia.ac.cn）

中国科学院自动化研究所
模式识别国家重点实验室

中国科学院自动化研究所
Institute of Automation, Chinese Academy of Sciences

概述

- 实体消歧概述
- 基于无监督的实体消歧
- 基于知识库链接的实体消歧

实体消歧定义

- 命名实体的歧义指的是一个实体指称项可对应到多个真实世界实体，例如，给定如下的四个实体指称项“*Michael Jordan*”

MJ1: Michael Jordan is a researcher in machine learning.

MJ2: Learning in Graphical Models: Michael Jordan

MJ3: M. Jordan wins NBA MVP.

MJ4 : Michael Jordan plays basketball in Chicago Bulls.

- 确定一个实体指称项所指向的真实世界实体，这就是命名实体消歧

网络媒体中的实体歧义

Chen Guangcheng

Morph	Target	Motivation
Blind Man (瞎子)	Chen Guangcheng(陈光诚)	Sensitive
Kimchi Country (泡菜国)	Korea (韩国)	Vivid
Rice Country (米国)	United States (美国)	Pronunciation
Miracle Brother (奇迹哥)	Wang Yongping (王勇平)	Irony

实体消歧分类

- 基于聚类的实体消歧
  - 把所有实体指称项按其指向的目标实体进行聚类
  - 每一个实体指称项对应到一个单独的类别

MJ1: Michael Jordan is a researcher in machine learning.

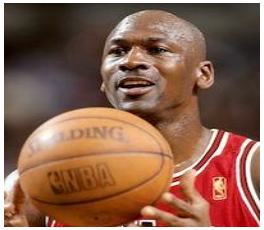
MJ2: Research in Graphical Models: Michael Jordan

MJ3: M. Jordan wins NBA MVP.

MJ4 : Michael Jordan plays basketball in Chicago Bulls

- 基于实体链接的实体消歧
  - 将实体指称项与目标实体列表中的对应实体进行链接实现消歧

MJ4 : Michael Jordan plays basketball in Chicago Bulls



概述

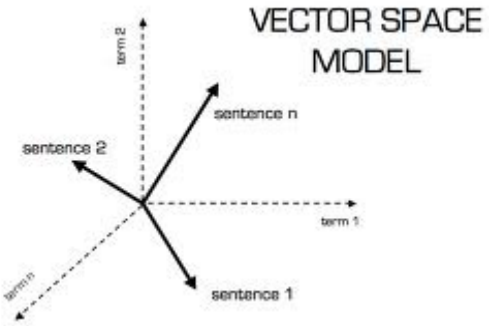
- 实体消歧概述
- 基于无监督的实体消歧
- 基于知识库链接的实体消歧

基于聚类的实体消歧

- 基本思路
  - 同一指称项具有近似的上下文
  - 利用聚类算法进行消歧
  - 核心问题：选取何种特征对于指称项进行表示
    - 词袋模型(Bagga et al., COLING, 1998)
    - 语义特征(Pederson et al., CLITP, 2005)
    - 社会化网络(Bekkerman et al., WWW, 2005)
    - 维基百科的知识(Han and Zhao, CIKM, 2009)
    - 多源异构语义知识融合(Han and Zhao, ACL, 2010)

基于聚类的实体消歧:词袋模型
(Bagga et al. COLING 1998)

- 利用待消歧实体周边的词来构造向量
- 利用向量空间模型来计算两个实体指称项的相似度，进行聚类



MJ1: Michael Jordan is a researcher in machine learning.

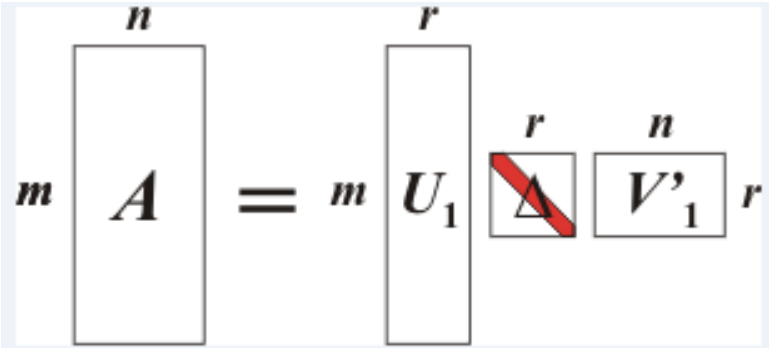
MJ1 → researcher → machine → learning

MJ1: Michael Jordan plays basketball in Chicago Bulls

MJ4 → plays → basketball → Chicago → Bulls

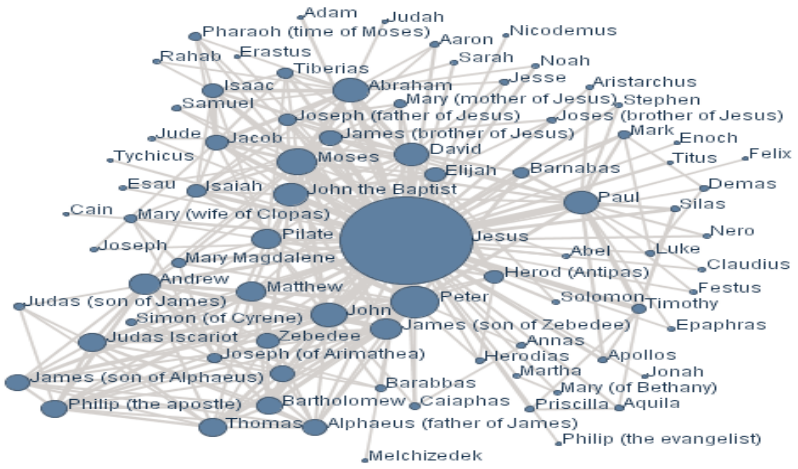
基于聚类的实体消歧: 语义特征
(Pederson et al. CLITP 2005)

- 词袋模型，没有考虑词的语义信息
- 利用SVD分解挖掘词的语义信息
- 利用词袋和浅层语义特征，共同表示指称项，利用余弦相似度来计算两个指称项的相似度



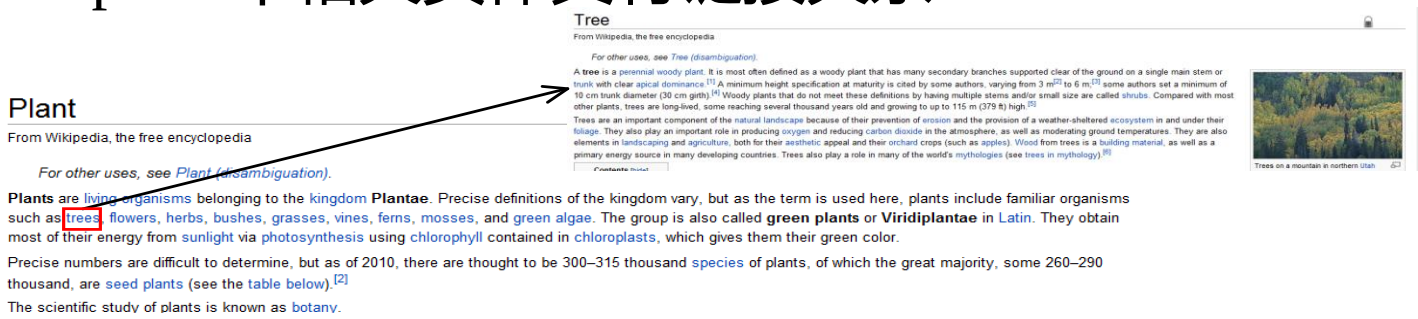
基于聚类的实体消歧：社会化网络
(Bekkerman et al. WWW 2005)

- 不同的人具有不同的社会关系
  - MJ（BasketBall）: Pippen, Buckley, Ewing, Kobe…
  - MJ（Machine Learning）: Liang, Mackey, Wauthier…
- MJ, Pippen, Buckley, Ewing, Kobe等的社会化关联信息所表现出来的网页链接特征，对网页进行聚类，从而实现网页内的人名聚类消歧。

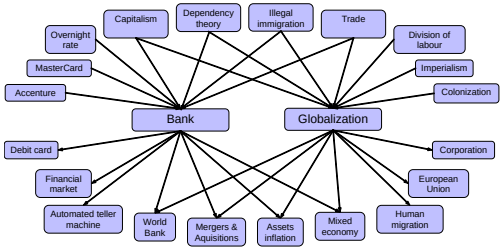


基于聚类的实体消歧: Wikipedia (Han CIKM 2009) (1/3)

- Wikipedia中相关实体具有链接关系



- 这种链接关系反映条目之间的语义相关度
  - D. Milne and Ian H. Witten 2008: The higher semantic related Wikipedia concepts will share more semantic related concepts.



The concepts has links with a and b

$$sr(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

The Whole Wikipedia concepts

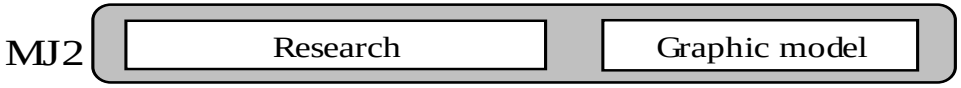
The concepts has links with a and b

基于聚类的实体消歧: Wikipedia (Han CIKM 2009) (2/3)

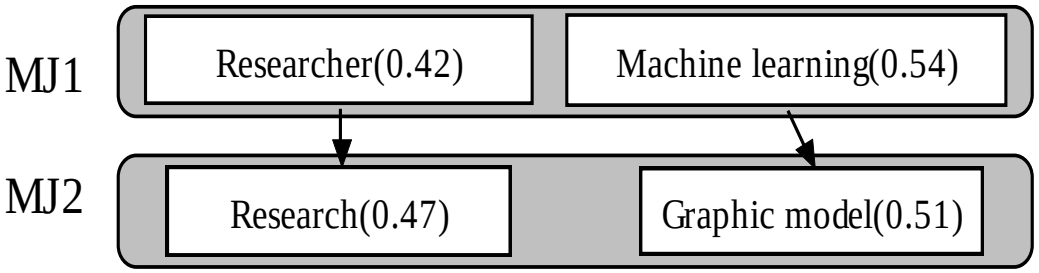
- 用实体上下文的维基条目对于实体进行向量表示

MJ1: Michael Jordan is a Researcher in machine learning.

MJ2: Research in Graphical Models: Michael Jordan



- 利用维基条目之间的相关度计算指称项之间的相似度（解决数据稀疏问题）



实验比较
(Han CIKM 2009) (3/3)

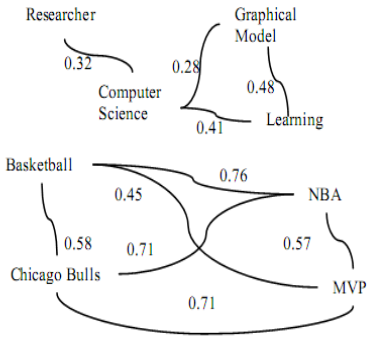
- 使用WePS数据集测试
- 使用结构化关联语义核的实体相似度能够提升10.7%的消歧性能

Method	WePS1_training		
	Pur	Inv_Pur	F
BOW	0.71	0.88	**0.78**
SocialNetwork	0.66	0.98	**0.76**
WikipediaConcept	0.80	0.88	**0.82**
WS-SameWeight	0.84	0.89	**0.85**
WS	0.88	0.89	**0.87**

基于聚类的实体消歧: 多源异构知识
(Han ACL 2010) (1/3)

- 仅仅考虑Wikipedia一种知识源，覆盖度有限
- 多源异构知识的挖掘与集成
  - 知识源中存在大量的多源异构知识
  - 挖掘和集成多源异构知识可以提高实体消歧的性能
  - Wikipedia
    - 用于捕捉概念之间的语义关联
  - WordNet
    - 用于捕捉词语之间的语言学关联
  - Web网页库
    - 用于捕捉命名实体之间的社会化关联

基于聚类的实体消歧：多源异构知识
(Han ACL 2010) (2/3)



- 多源异构知识的表示框架：语义图
  - 等同概念识别
  - 概念连接
    - 同时捕捉显式语义知识和结构化语义知识
      - **语义图的边（显式语义知识）**——建模了所有从知识源中直接抽取出的概念之间的显式语义关联
      - **语义图的结构（结构化语义知识）**——建模了概念之间的隐藏语义关联
- 语义图中语义知识的挖掘和融合算法
  - 计算原则：“如果一个概念的邻居概念与另一个概念存在语义关联，则这个概念也与另一个概念存在语义关联”
  - 语义关联在图中的传递性

$$S_{ij} = \lambda \sum_{l \in N_i} \frac{A_{il}}{d_i} S_{lj} + \mu A_{ij}$$

邻居节点传递　　显式语义关联

基于聚类的实体消歧:多源异构知识
(Han ACL 2010) (3/3)

- 实验比较
  - 使用WePS数据集测试
  - 使用多源知识能够有效提高消歧的准确度

	WePS2_test		
	Pur	**Inv_Pur**	**F**
BOW	0.80	0.80	**0.77**
SocialNetwork	0.62	0.93	**0.70**
SSR-NoKnowledge	0.84	0.80	**0.80**
SSR-NoStructure	0.84	0.83	**0.81**
SSR-NE	0.78	0.88	**0.80**
SSR-WordNet	0.85	0.82	**0.83**
SSR-Wikipedia	0.84	0.81	**0.82**
SSR	0.90	0.86	**0.88**

基于聚类的实体消歧：评测(1/2)

- WePS： Web People Search Evaluation
  - WePS1是SEMEVAL2007的子任务
  - WePS2是WWW的一个workshop
  - 任务：Web环境中的人名消歧，即给定一个包含某个歧义人名的网页集合，按照网页中人名指称项所指向的人物概念来对网页进行聚类，以及抽取一个网页中关于某个人的特定属性来辅助进行人名消歧
  - 评测方法

$$\text{Purity} = \sum_i \frac{|C_i|}{n} \max \text{Precision}(C_i, L_j)$$

$$\text{Inverse Purity} = \sum_i \frac{|L_i|}{n} \max \text{Precision}(L_i, C_j)$$

$$\text{F} = \frac{1}{\alpha \frac{1}{\text{Purity}} + (1-\alpha)\frac{1}{\text{Inverse Purity}}}$$

基于聚类的实体消歧：评测(2/2)

		Macro-averaged Scores F-measures			
rank	team-id	$\alpha = ,5$	$\alpha = ,2$	Pur	Inv_Pur
1	CU_COMSEM	,78	,83	,72	,88
2	IRST-BP	,75	,77	,75	,80
3	PSNUS	,75	,78	,73	,82
4	UVA	,67	,62	,81	,60
5	SHEF	,66	,73	,60	,82
6	FICO	,64	,76	,53	,90
7	UNN	,62	,67	,60	,73
8	*ONE-IN-ONE*	*,61*	*,52*	*1,00*	*,47*
9	AUG	,60	,73	,50	,88
10	SWAT-IV	,58	,64	,55	,71
11	UA-ZSA	,58	,60	,58	,64
12	TITPI	,57	,71	,45	,89
13	JHU1-13	,53	,65	,45	,82
14	DFKI2	,50	,63	,39	,83
15	WIT	,49	,66	,36	,93
16	UC3M_13	,48	,66	,35	,95
17	UBC-AS	,40	,55	,30	,91
18	*ALL-IN-ONE*	*,40*	*,58*	*,29*	*1,00*

WePS 1

		Macro-averaged Scores F-measures			
rank	run	$\alpha = ,5$	$\alpha = ,2$	Pur	Inv_Pur
	BEST-HAC-TOKENS	,90	,89	,93	,88
	BEST-HAC-BIGRAMS	,90	,87	,94	,86
1	PolyUHK	,88	,87	,91	,86
2	UVA_1	,87	,87	,89	,87
3	ITC-UT_1	,87	,83	,95	,81
	CHEAT_SYS	,87	,94	,78	1,00
4	UMD_4	,81	,76	,95	,72
5	XMEDIA_3	,80	,76	,91	,73
6	UCI_2	,80	,84	,75	,89
7	LANZHOU_1	,80	,78	,85	,77
8	FICO_3	,80	,76	,90	,73
	HAC-BIGRAMS	,78	,64	,96	,67
9	UGUELPH_1	,74	,84	,64	,95
10	CASIANED_4	,73	,77	,72	,83
	HAC-TOKENS	,71	,64	,96	,60
11	AUG_4	,69	,68	,79	,68
12	UPM-SINT_4	,67	,70	,69	,74
	ALL_IN_ONE	,67	,79	,56	1,00
13	UNN_2	,64	,59	,80	,57
14	ECNU_1	,53	,56	,60	,63
15	PRIYAVEN	,53	,49	,71	,48
16	UNED_3	,51	,48	,71	,48
17	BUAP_1	,37	,30	,89	,27
	ONE_IN_ONE	,34	,27	1,00	,24

WePS 2

小结

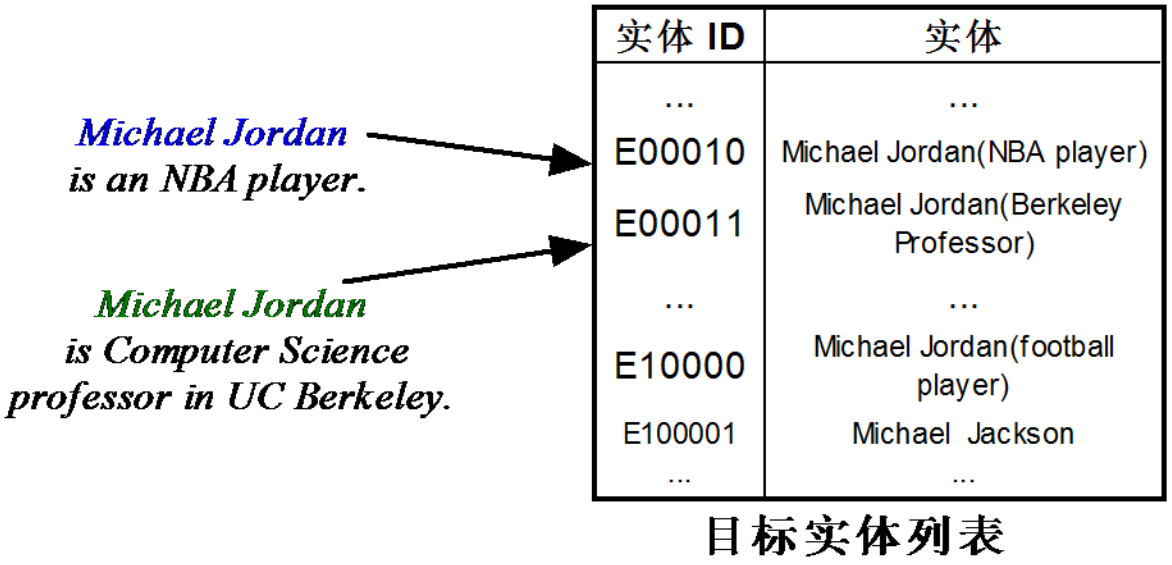
- 主要研究集中在实体指称项的语义表示
- 已有工作大多是通过扩展特征，增加更多的知识来提高消歧精度
- 挑战
  - 消歧目标难以确定
  - 缺乏实体的显式表示

概述

- 实体消歧概述
- 基于无监督的实体消歧
- 基于知识库链接的实体消歧

实体链接的任务

- 任务
  - 给定实体指称项和它所在的文本，将其链接到给定知识库中的相应实体上



实体链接主要步骤

- 主要步骤
  - 候选实体的发现
    - 给定实体指称项，链接系统根据知识、规则等信息找到实体指称项的候选实体
  - 候选实体的链接
    - 系统根据指称项和候选实体之间的相似度等特征，选择实体指称项的目标实体

实体指称项文本：Michael Jordan is a former NBA player, active businessman and majority owner of the Charlotte Bobcats.

候选实体：
Michael Jordan (basketball player)
Michael Jordan (mycologist)
Michael Jordan (footballer)
Michael B. Jordan
Michael H. Jordan
Michael-Hakim Jordan
Michael Jordan (Irish polotician)
…

候选实体发现

- 如何根据实体指称项找出候选实体
  - 利用Wikipedia的信息
  - 利用上下文信息

实体指称项	候选实体
Michael Jordan	Michael Jordan (basketball) Michael Jordan (mycologist) Michael Jordan (football) Michael B. Jordan (American actor) …
AI	Artificial intelligence Ai (singer) …
…	…

利用Wikipedia信息获取候选实体

- 利用Wikipedia中锚文本的超级链接关系
  - Michal Jordan is a former NBA player

- 利用Wikipedia中的消歧页面

Michael Jordan is an American basketball player.

Michael Jordan may also refer to:

- Michael Jordan (mycologist), English mycologist
- Michael Jordan (footballer) (born 1986), English goalkeeper (A
- Michael B. Jordan (born 1987), American actor
- Michael I. Jordan (born 1957), American researcher in machine
- Michael H. Jordan (d. 2010), American executive for CBS, Pep
- Michael-Hakim Jordan (born 1977), American professional bas
- Michael Jordan (Irish politician), Irish Farmers' Party TD from W

- 利用Wikipedia中的重定向页面

Michael jordan

From Wikipedia, the free encyclopedia

Redirect page

↳ Michael Jordan

利用上下文获取缩略语候选实体
(Zhang IJCAI 2011)

- 问题
  - 缩略语在实体指称项中十分常见，据统计，在KBP2009的测试数据，在3904个实体指称项中有827个为缩略语
- 动机
  - 缩略语指称项具有很强的歧义性，但它的全称往往是没有歧义的
  - ABC和American Broadcasting Company, AI和Artificial Intelligence等
  - 在实体指称项文本中，缩略语的全称出现过
- 解决方法
  - 利用人工规则抽取实体候选

候选实体链接

- 如何进行实体链接
  - 基本方法：计算实体指称项和候选实体的相似度，选择相似度最大的候选实体
  - 单一实体链接
    - BOW模型 (Honnibal TAC 2009, Bikel TAC 2009)
    - 加入候选实体的类别特征（Bunescu et al., EACL 2006）
    - 加入候选实体的流行度等特征（Han et al., ACL 2011）
  - 协同实体链接
    - 利用实体之间类别的共现特征（Cucerzan, EMNLP 2007）
    - 利用实体之间链接关系（Kulkarni et al., KDD 2009）
    - 利用同一篇文档中不同实体之间存在的语义关联特征（Han et al., SIGIR 2011）

实体链接的基本方法
(Honnibal TAC 2009, Bikel TAC 2009)

- 基于词袋子模型计算相似度
  - 将实体指称项上下文文本与候选实体上下文文本表示成词袋子向量形式，通过计算向量间的夹角确定指称项与候选实体相似度，系统选择相似度最大的候选实体进行链接

$$score(q, e_k) = cos(q.T, e_k.T) = \frac{q.T}{\|q.T\|} \frac{e_k.T}{\|e_k.T\|}$$

$$\hat{e} = arg \max_{e_k} score(q, e_k)$$

类别特征
(Bunescu EACL 2006)

- 动机
  - 候选实体的文本内容可能太短，会导致相似度计算的不准确
  - 加入指称项文本中的词与候选实体类别的共现特征
    - 例：除了计算待消歧文本和实体Wikipedia文本John Williams (composer)的相似度外，还考虑当前文本中的词语与**Music, Art** 等类别的共现信息
- 方法
  - 训练SVM分类器对候选实体进行选择
  - 训练数据由Wikipedia中的超级链接获得
  - 所采用的特征
    - 文本相似度
    - 指称项文本中词与候选实体类别的共现信息

John Williams (composer): Category={**Music, Art**…}
John Williams (wrestler): Category={Sport,…}
John Williams (VC): Category={Bank,…}

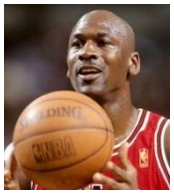
Williams has also composed numerous classical concerti, and he served as the principal conductor of the Boston Pops Orchestra from 1980 to 1993

类别：music

实体流行度等特征
(Han ACL 2011)

- 动机
  - 传统的方法仅仅是计算实体指称项与候选实体的相似度，忽略了候选实体的背景知识与先验信息，如实体本身的流行度、实体与指称项的关系等
- 方法
  - 考虑实体的背景知识，将实体的背景知识融入到实体链接的过程，实体的背景知识和先验信息主要有
    - 实体流行度：实体e在知识库中的概率P(e)
    - 名称的知识：指称项s指向实体e的概率P(s|e)
    - 上下文知识：实体e出现在特定上下文环境c的概率P(c|e)

协同实体链接



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

- 实体指称项与目标实体的语义相似度
- 目标实体之间的语义相似度

协同学习策略

- 动机
  - 同一篇文档中实体之间具有语义相关性
  - 利用Pairwise优化策略

$$\frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in S_0} r(y_s, y_{s'}) + \frac{1}{|S_0|} \sum_{s \in S_0} w^\top f_s(y_s).$$

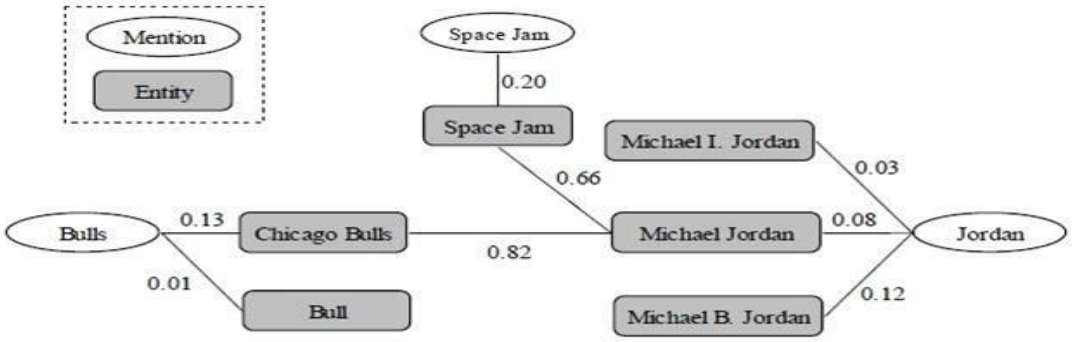
任意两个目标实体之间的语义相关度

实体指称项到目标实体的语义相似度

- 目标实体的语义相关度计算方法：
  - 利用实体类别重合度计算目标实体语义相似度（Cucerzan, EMNLP 2007）
  - 利用实体之间链接关系计算目标实体语义相似度（Kulkarni , KDD 2009 ）

基于图的协同链接
(Han SIGIR 2011)

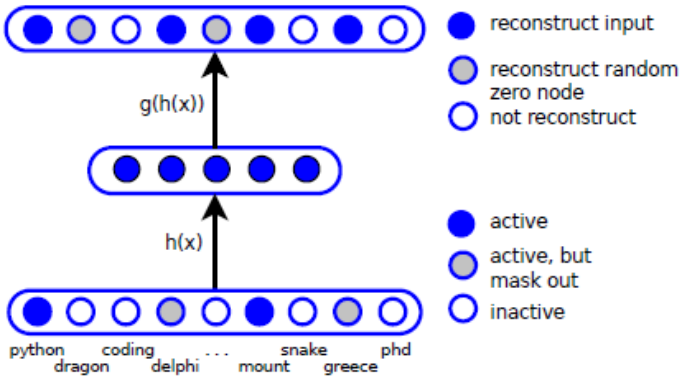
- 动机
  - Pairwise策略只考虑两两实体关系，结果不是全局最优的
  - 采用图方法，全局考虑目标实体之间的语义关联
- 方法：Referent Graph，两种关系构成
  - 指称项与实体之间的关系：该指称项文本与实体文本的相似度，由传统的VSM模型得到
  - 实体之间的语义关系：利用目标实体之间的链接关系计算实体之间的语义相关度

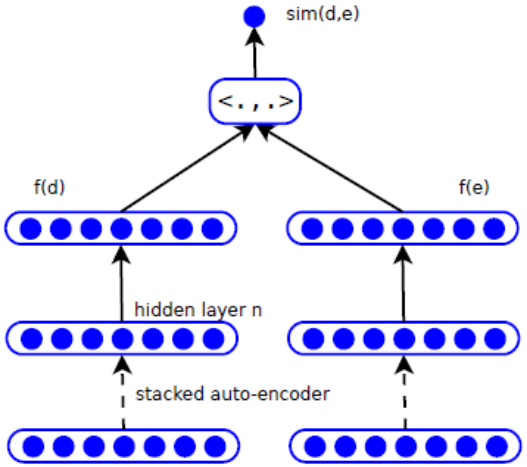


基于深度学习的方法
(He ACL 2013)

- 动机
  - 传统的方法中，计算待消歧实体上下文和目标实体语义相似度的方法（点乘，余弦相似度，KL距离等）可扩展性差，没有考虑各个概念间的内在联系
  - 在协同过滤的方法中，计算待消歧实体上下文和目标实体语义相似度也是基础工作。
- 方法

提出利用深度学习的方法自动联合学习实体和文档的表示，进而完成实体链接任务（数据集：TAC-KBP 2010 和 AIDA）

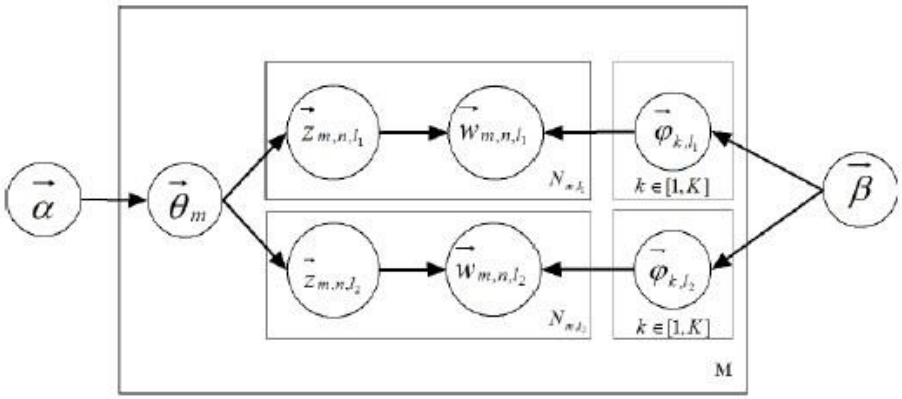




跨语言实体链接
（Zhang IJCAI 2013）

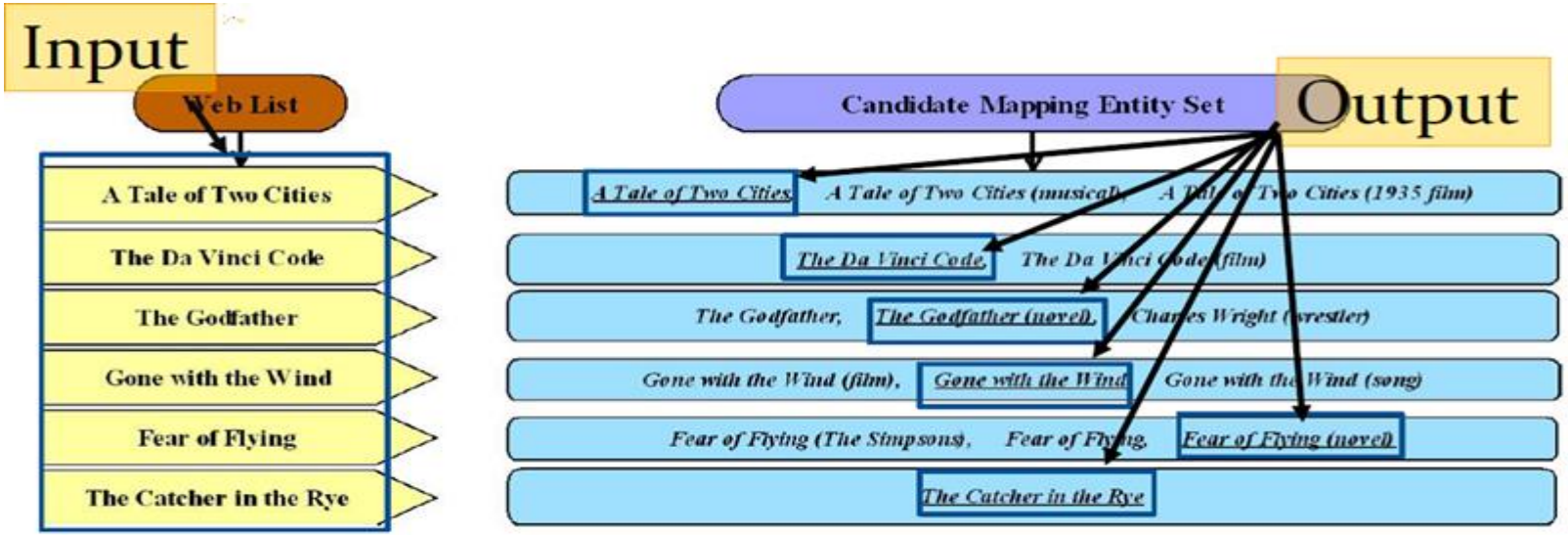
- 动机
  - 给定一种语言的实体指称项和其所在的上下文，将其链接到另外一种语言的知识库中
  - 传统方法要先翻译成目标语言，可能产生错误传递，需要大量的句子级平行的双语训练语料

- 方法：利用双语隐含主题模型将实体指称项与候选实体映射到同一个主题空间中（数据集：TAC-KBP 2011）
  - 每一个隐含主题有两种不同的分布，分别对应两种不同语言
  - 处于同一个主题分布下的两种不同语言的词的分布具有一些共性



结构化数据中的实体链接
(Shen SIGKDD 2012)

- 动机
  - 没有上下文
  - 任务与传统的实体链接不同

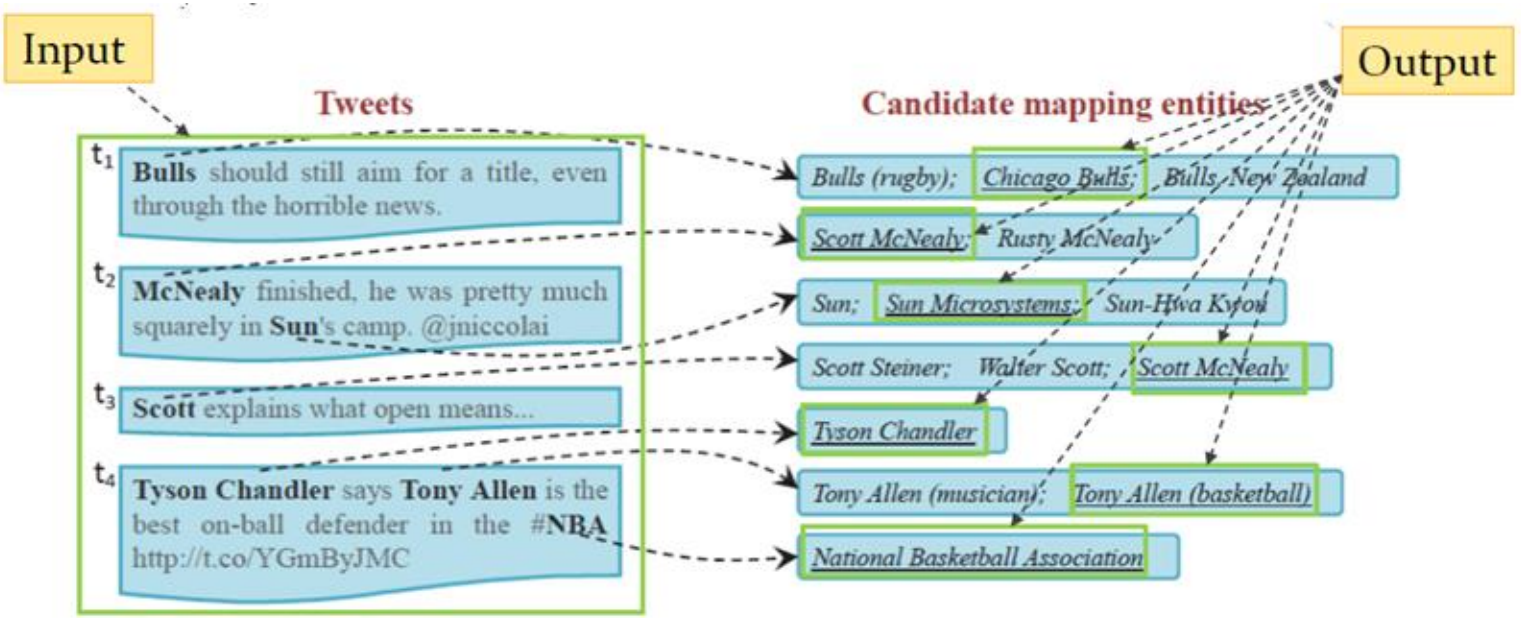


- 方法
  - 主要利用实体的流行度和实体共现类型去消歧

取自王建勇老师报告（第三届中文知识图谱研讨会）

社交数据中的实体链接
(Shen SIGKDD 2013)

- 动机
  - 社交媒体（Twitter）是一种重要的信息来源
  - 社交媒体的上下文较短，语言表述不规范



- 方法
  - 利用tweet的用户信息和tweet的交互信息

取自王建勇老师报告（第三届中文知识图谱研讨会）

实体链接评测 (1/2)

- TAC-KBP (2009-Now)：Entity Linking
  - 任务：将文本中的目标实体链接到Wikipedia中的真实概念，达到消歧的目的
  - 评测方法：

以指称项为单位计算的准确率

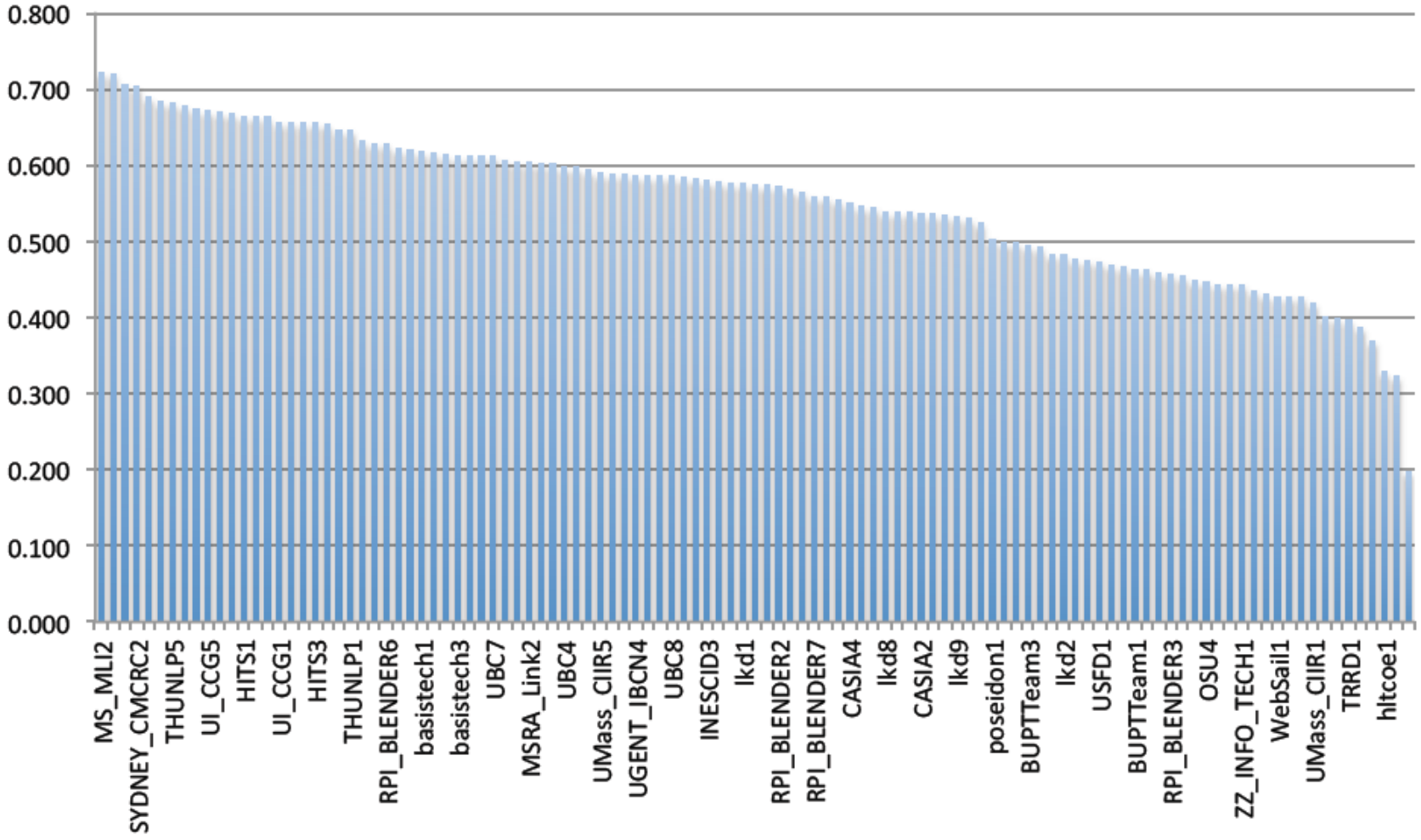
$$Accuracy_{micro} = \frac{NumCorrect}{NumQueries}$$

$$Accuracy_{macro} = \frac{\sum_{i}^{NumEntities} \frac{NumCorrect(E_i)}{NumQueries(E_i)}}{NumEntities}$$

以实体为单位计算的准确率

实体链接评测 (2/2)

2013评测结果(Micro Accuracy)



小结

- 目前实体链接方法主要是如何更有效挖掘实体指称项信息，如何更准确地计算实体指称项和实体概念之间的相似度
- 由单一实体链接向协同实体链接发展
- 难点：未登录实体的处理