Knowledge Base Unification via Sense Embeddings and Disambiguation

EMNLP2015

Claudio Delli Bovi @uniroma1 罗马大学

Luis Espinosa-Anke @upf 庞培法布拉大学

Roberto Navigli @uniroma1

hanzhe@icst-wip

2015 年 10 月 23 日

Outline

- summary
  - 作者简介
  - 相关工作
  - 论文问题、方法概述
  - 实验使用的工具 (穿插说)
- 归一化（Unification）的方法
  - 实体消歧（entity disambiguation）
  - 关系对应（relation alignment）
- 实验/总结

summary/author



Claudio Delli Bovi

罗马大学博士，计算语言学方向，主要研究语法、句法结构，目前在做 WSD

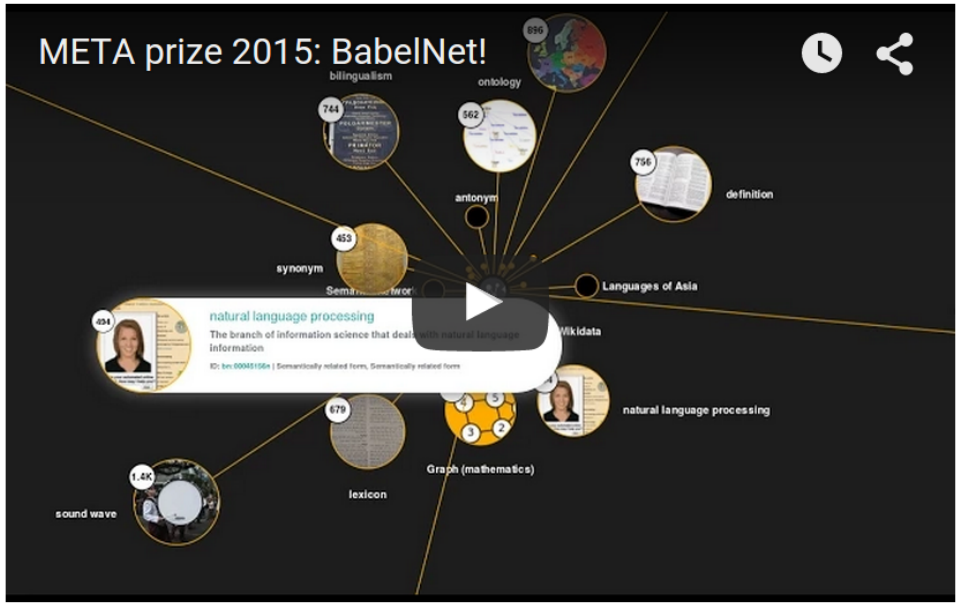
Luis Espinosa-Anke

庞培法布拉大学博士，在做一些信息抽取，语义网有关的东西

Roberto Navigli

Claudio 的导师，07 年博士毕业，**BabelNet/Babelfy**

- **Winner** of the *prestigious META prize 2015* for BabelNet, Riga Summit 2015.



Claudio Delli Bovi

罗马大学博士，计算语言学方向，主要研究语法、句法结构，目前在做 WSD

Luis Espinosa-Anke

庞培法布拉大学博士，在做一些信息抽取，语义网有关的东西

Roberto Navigli

Claudio 的导师，07 年博士毕业，**BabelNet/Babelfy**

概述

— 动机/方法简述

summary/motivation

实体

李娜 (网球运动员)

李娜 (歌手)

李娜 (游泳运动员)

姜山 (网球运动员)

张家辉

...

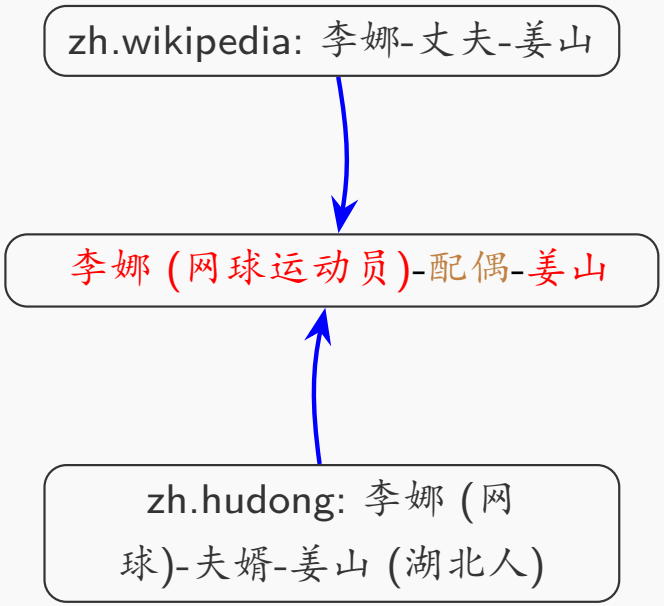
关系

配偶

出生日期

子女

...



如何整合多个知识库?

假设

- 全部在 subject-predicate-object 数据库上
- 知识库可能是完全非结构化的（主语只是普通字符串）
- 给定一个完备的语义集合（左上的实体集合）
- 其他的数据库可能存在歧义，结构不标准

summary/motivation

实体

李娜 (网球运动员)

李娜 (歌手)

李娜 (游泳运动员)

姜山 (网球运动员)

张家辉

...

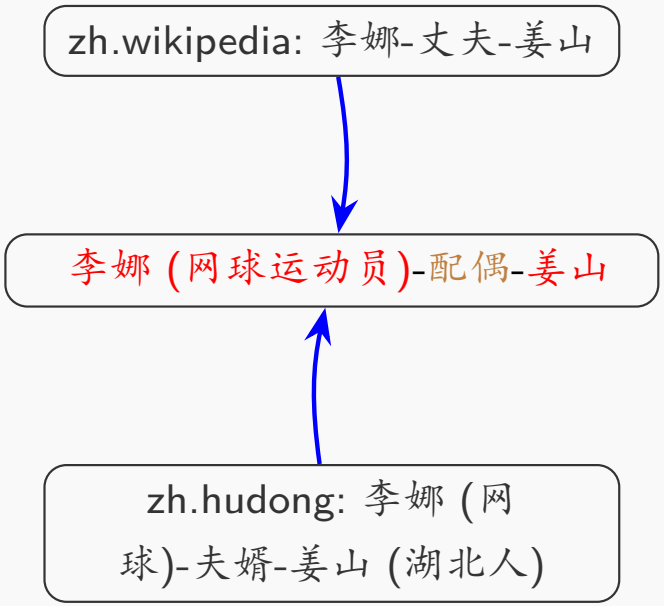
关系

配偶

出生日期

子女

...



如何整合多个知识库?

过程

- **整合多个知识库**（消歧 + 谓词统一）
- 怎么把李娜-丈夫-姜山转换成标准的结构?
  - 消除其他数据库中主体、客体的歧义
  - 将不同数据库中含义相同的谓词合并

summary/method

实体

李娜 (网球运动员)

李娜 (歌手)

李娜 (游泳运动员)

姜山 (网球运动员)

张家辉

...

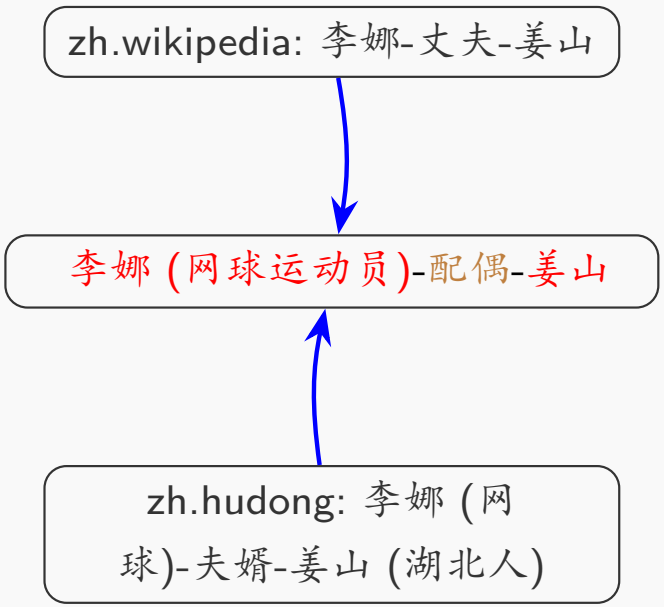
关系

配偶

出生日期

子女

...



- 方法
- 先对**实体消歧**，再对应不同数据库的**relation 归一**
- **实体消歧**【1】
  - 对一条三元组的主体、客体的语义候选的所有组合，如果存在一组已经非常好了，选为种子
- **实体消歧**【2】
  - 找出好的 relation （主体的向量集中，比如都表示人; 客体的向量集中，比如都表示地点）

summary/method

实体

李娜 (网球运动员)

李娜 (歌手)

李娜 (游泳运动员)

姜山 (网球运动员)

张家辉

...

关系

配偶

出生日期

子女

...

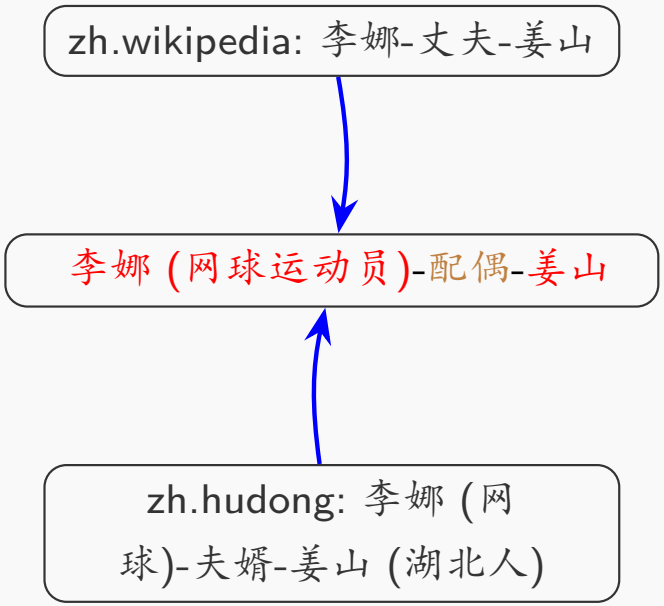
【一】

微软-CEO-纳德拉

搜狐-CEO-张朝阳

苹果 (?)-CEO-库克

...



- 方法
- **实体消歧**【3】
  - 【一】对于好的 relation，该关系内的所有三元组的主体、客体相互间相似，可以作为其中一个主体做消歧时的文本

summary/method

实体

李娜 (网球运动员)

李娜 (歌手)

李娜 (游泳运动员)

姜山 (网球运动员)

张家辉

...

关系

配偶

出生日期

子女

...

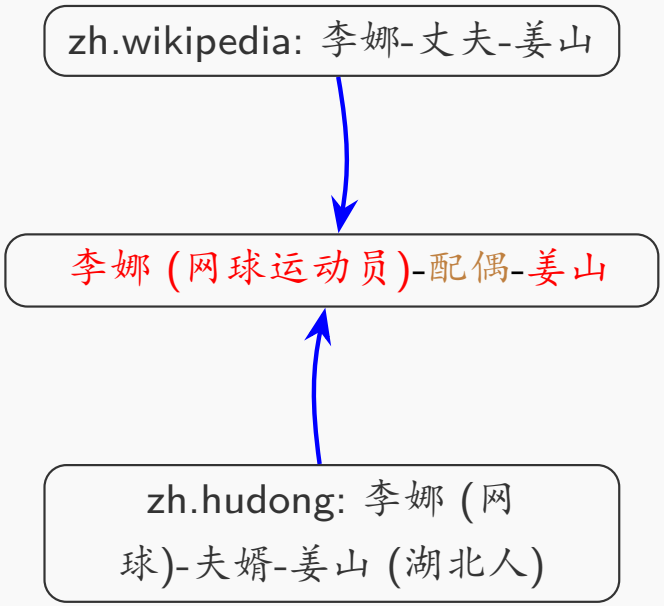
【二】

花果山-所在-连云港

全国政协-所在-北京

大众-所在-黑龙江

(这里指大众乡) ...



- 方法
- **实体消歧**【3】
  - 【一】对于好的 relation，该关系内的所有三元组的主体、客体相互间相似，可以作为其中一个主体做消歧时的文本
  - 【二】对于不好的 relation 对应的一条三元组的主语或客体，只能通过 relation 名字来消歧

summary/method

实体

李娜 (网球运动员)

李娜 (歌手)

李娜 (游泳运动员)

姜山 (网球运动员)

张家辉

...

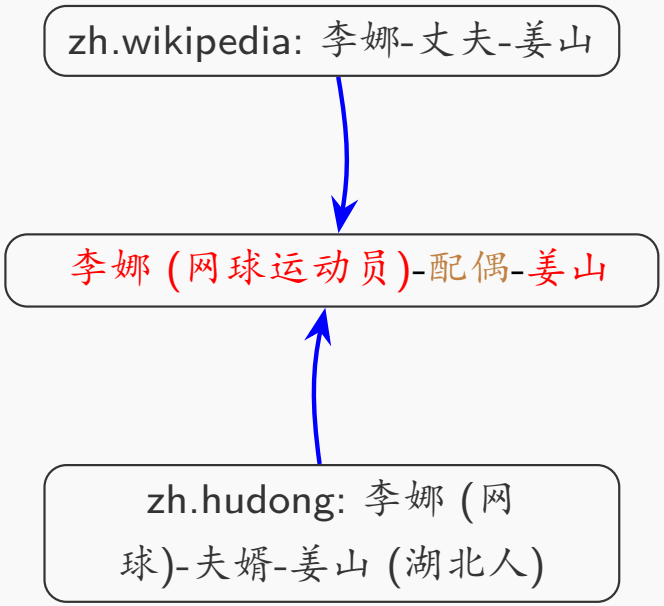
关系

配偶

出生日期

子女

...



- 方法
- **实体消歧**【3】
  - 【一】对于好的 relation，该关系内的所有三元组的主体、客体相互间相似，可以作为其中一个主体做消歧时的文本
  - 【二】对于不好的 relation 对应的一条三元组的主语或客体，只能通过 relation 名字来消歧

summary/method

实体

李娜 (网球运动员)

李娜 (歌手)

李娜 (游泳运动员)

姜山 (网球运动员)

张家辉

...

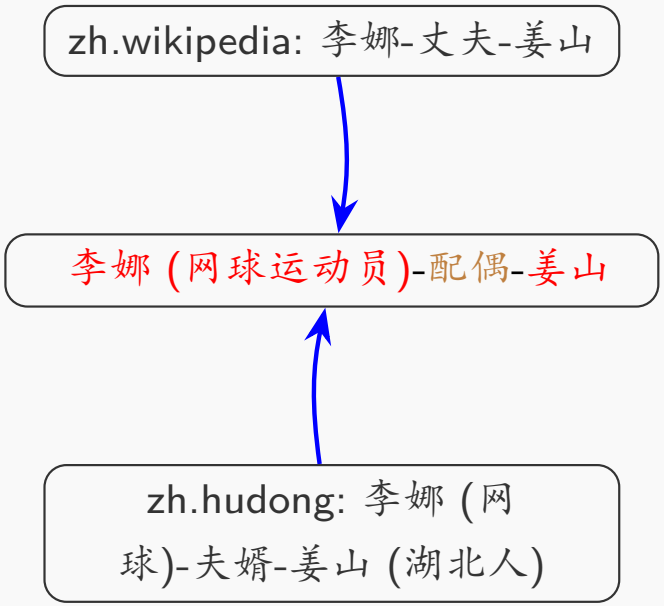
关系

配偶

出生日期

子女

...



- 方法
- 不同知识库的**relation 归一**
- 通过每两个知识库的每任两个 relation 的主客体特征向量计算相似性

related work/introduction

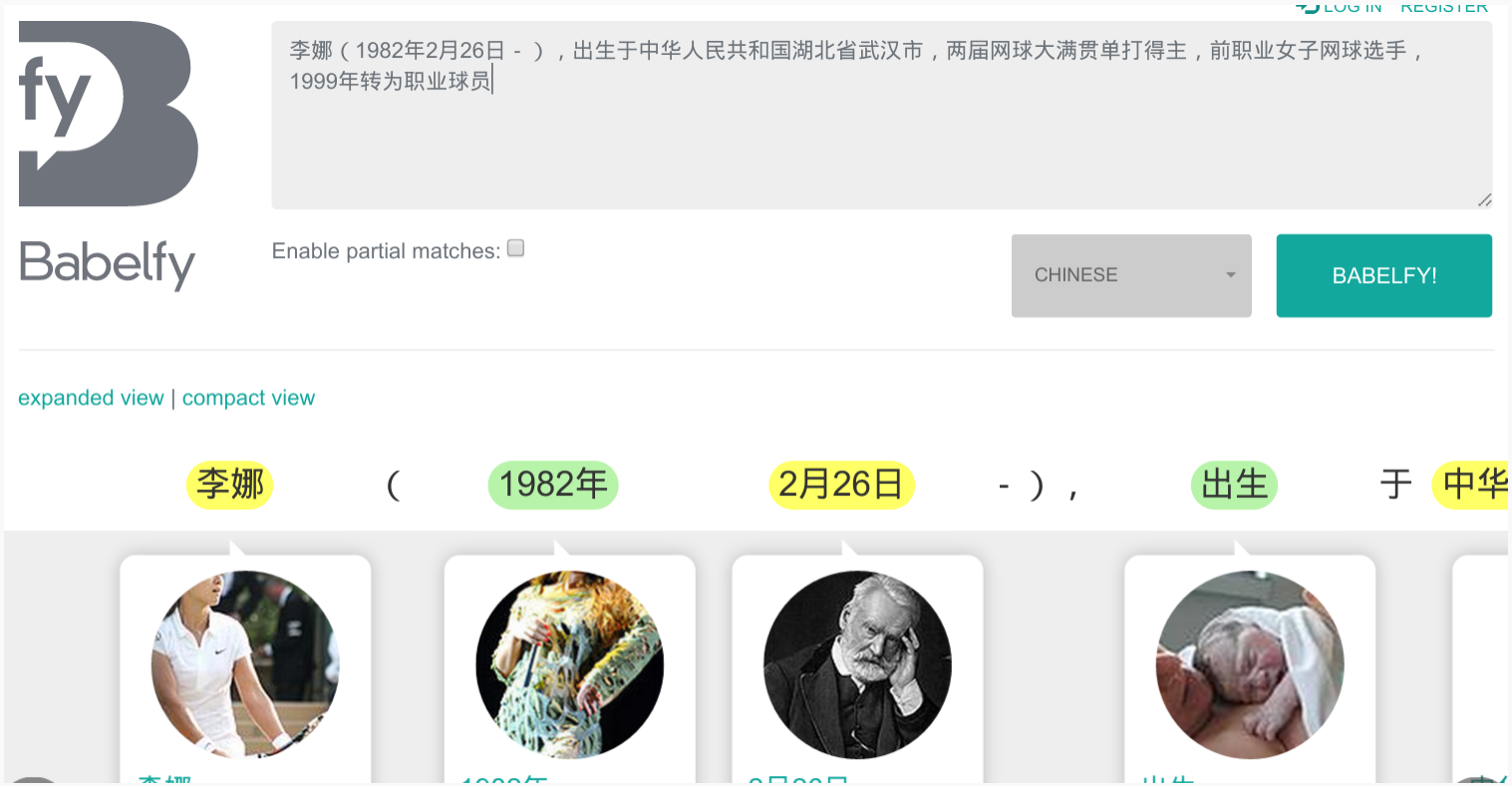
- Open Information Extraction
  - 从 Web-scale 级别的自然语言信息中抽取结构化/格式化的数据
  - e.g. DBpedia, Freebase, YAGO,...
  - 提升效果/去除噪声数据的方法
    - matrix factorization, distant supervision, multi-instance,...
  - 知识库补全（Knowledge Base completion）
    - 少量结构化数据和大量的半结构化数据相互提升准确率

related work

- BabelNet
  - 罗马大学开发的跨语言语义网，以“姚明”为例
    - “entity”有中文标签，可以显示中文，“relation”纯英文
    - “relation”有重复，无意义的

related work

- Babelfy
  - 罗马大学开发的多语言的消歧和实体链接集成的工具
    - 链接到 BabelNet
    - 分词效果不好;
    - 有一定的消歧能力：能找出网球运动员李娜，找不到歌手李娜



related work

两个数据库间的归一/对应

- Dutta et al. (2014): NELL 数据集的 augment 链接到 DBpedia 的实体
  - 使用一阶逻辑和马尔可夫网络
- Grycner and Weikum (2014) :PATTY 的 pattern 和 wordNet 的谓词对应
- Lin et al. (2012) : 使用 freebase 的类别来链接 reverb 中的潜在实体
  - reverb record:[2047542 Bilberry also_contains vitamin_C bilberry also_contain vitamin_c 1 0.94124 http:...]

能适应多个数据库归一的方法

- Riedel et al. (2013): 训练含有隐式特征（latent feature）的实体、关系向量
- Dong et al. (2014): 使用 freebase 数据训练一个概率模型来判断抽取结果的准确度

related work

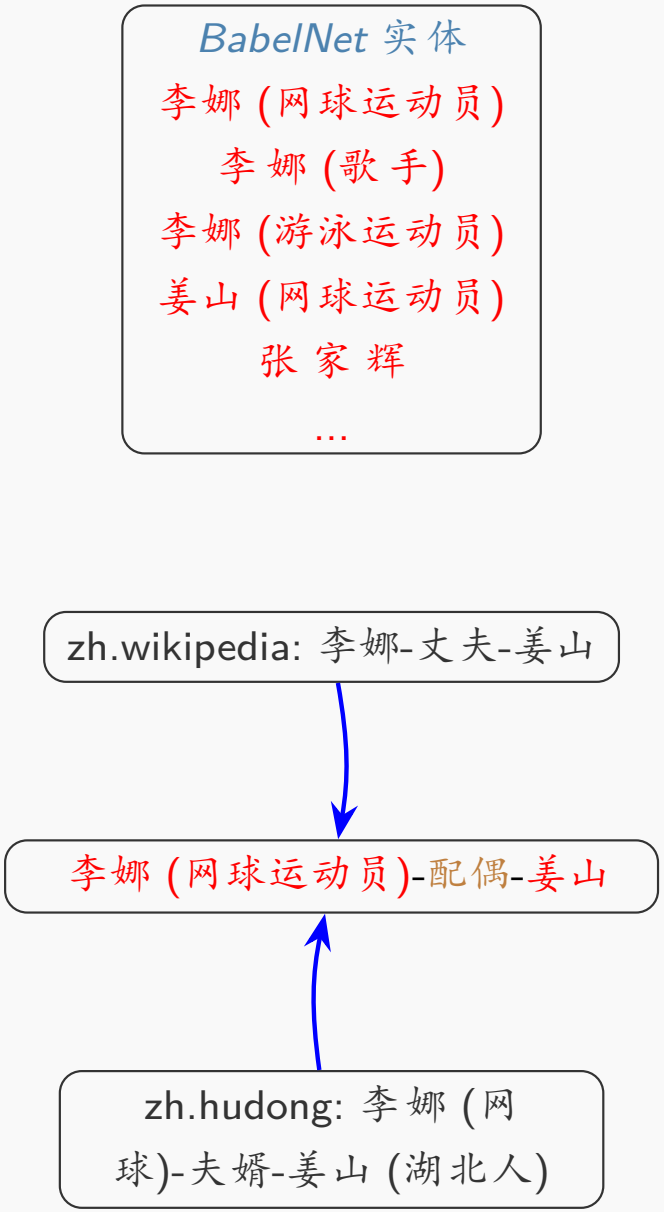
在关系抽取和知识补全上使用 Embedding models

- Socher et al., 2013; Weston et al.,2013; Bordes et al., 2013
- 作者认为这些模型停留在（surface level），不能表达普世的语义（common semantic）

归一化方法

— 实体消歧

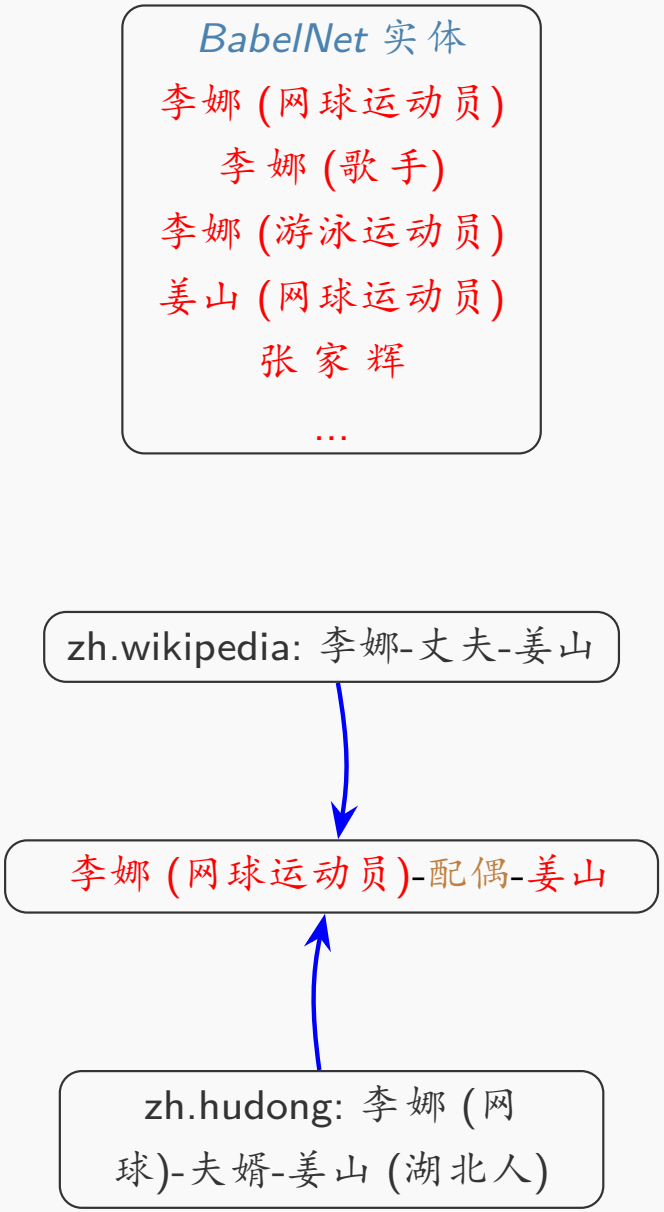
method/disambiguation



【实体消歧】准备

- BabelNet 是一个语义集合（sense inventory），与常见知识库的实体对应
- 消歧结束
  - 知识库的主语/客体对应到 BabelNet 上
- 实体/语义向量
  - 利用一个庞大的标注语料训练词向量
  - 采用 SENSEMBED（罗马大学 2015ACL）训练而非 word2vec，因为后者不能区分多义词

method/disambiguation

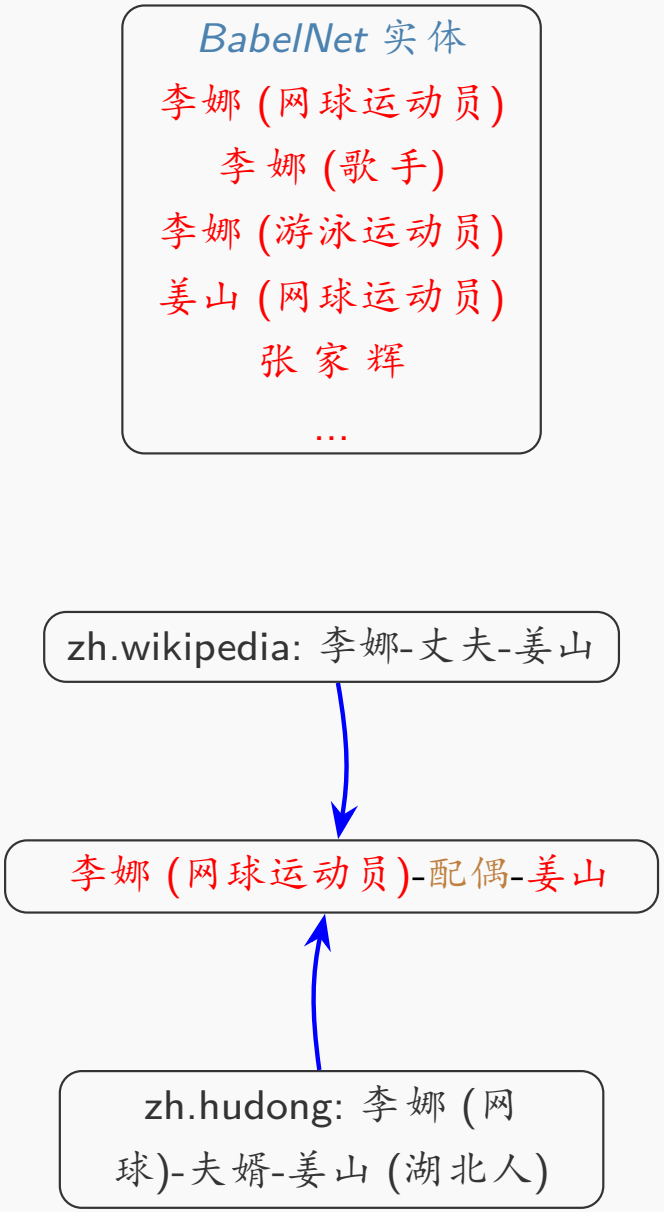


【实体消歧】准备

- BabelNet 是一个语义集合（sense inventory），与常见知识库的实体对应
- 消歧结束
  - 知识库的主语/客体对应到 BabelNet 上
- 实体/语义向量
  - 利用一个庞大的标注语料训练词向量
  - 采用 SENSEMBED（罗马大学 2015ACL）训练而非 word2vec，因为后者不能区分多义词

如果是拿维基百科训练，只要把原来正文中的实体标签"[[蘋果公司 | 苹果]]" 还原成" 蘋果公司" 其实可以解决多义的问题/训练出多个词向量【苹果、蘋果公司】?

method/disambiguation



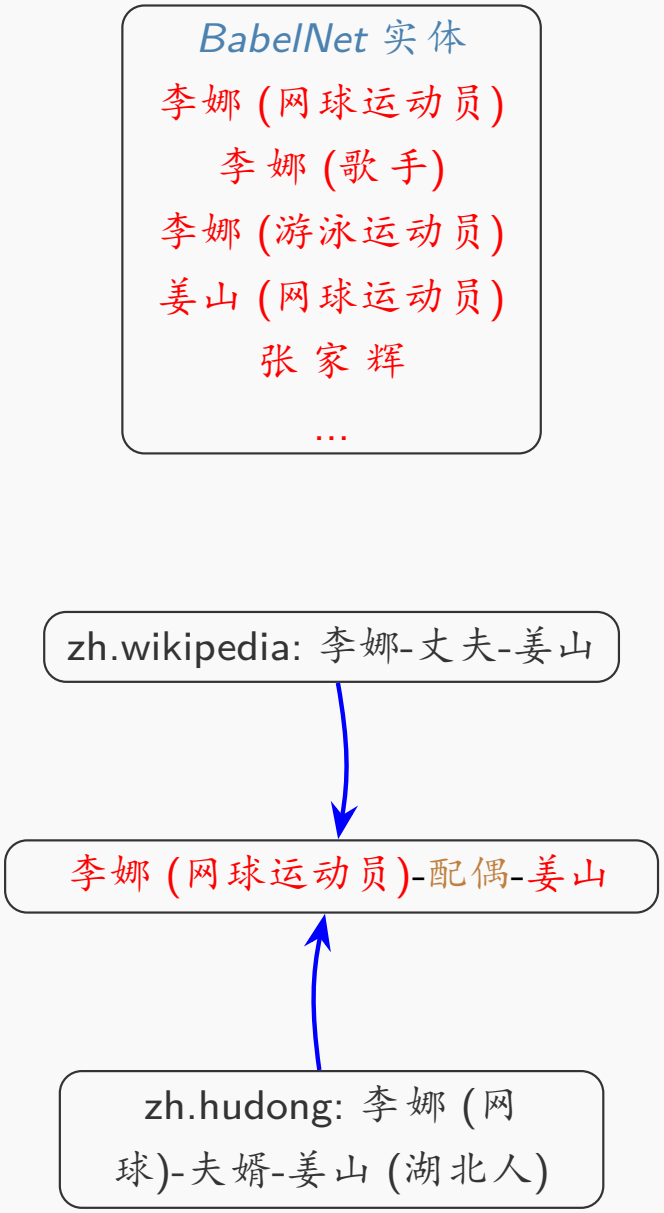
【实体消歧一】找出种子三元组

- 三元组 $<e_d, r, e_g>, e_d, e_g$ 的所有语义 $\mathbf{s}_d = \{s_d^1, ..., s_d^m\}$ and $\mathbf{s}_g = \{s_g^1, ..., s_g^{m'}\}$, 对应的词向量 $\mathbf{v}_d = \{v_d^1, ..., v_d^m\}$ and $\mathbf{v}_g = \{v_g^1, ..., v_g^{m'}\}$
- 任意两个组合中词向量余弦相似度最大的一对作为该三元组消歧的默认最优解

$$\langle v_d^*, v_g^* \rangle = \text{argmax}_{v_d \in \mathbf{v}_d,\, v_g \in \mathbf{v}_g} \frac{v_d \cdot v_g}{\|v_d\| \, \|v_g\|}$$

- 如果最优解的相似度大于一个阈值 ζ, 那么这个最有解是“完美”的，该三元组即为种子三元组
  - 这里有点问题，满足 r 的三元组的主语、客体的词向量可能是满足一个特定的角度才是标准的答案

method/disambiguation



【实体消歧二】找出完美关系

def. 完美关系

该 relation 对应的主语 (客体) 的词向量很集中，方差小

- v_D, v_G 表示关系 r 对应的所有种子三元组的主语、客体的词向量集合
- 该关系的主语、客体中心词向量为

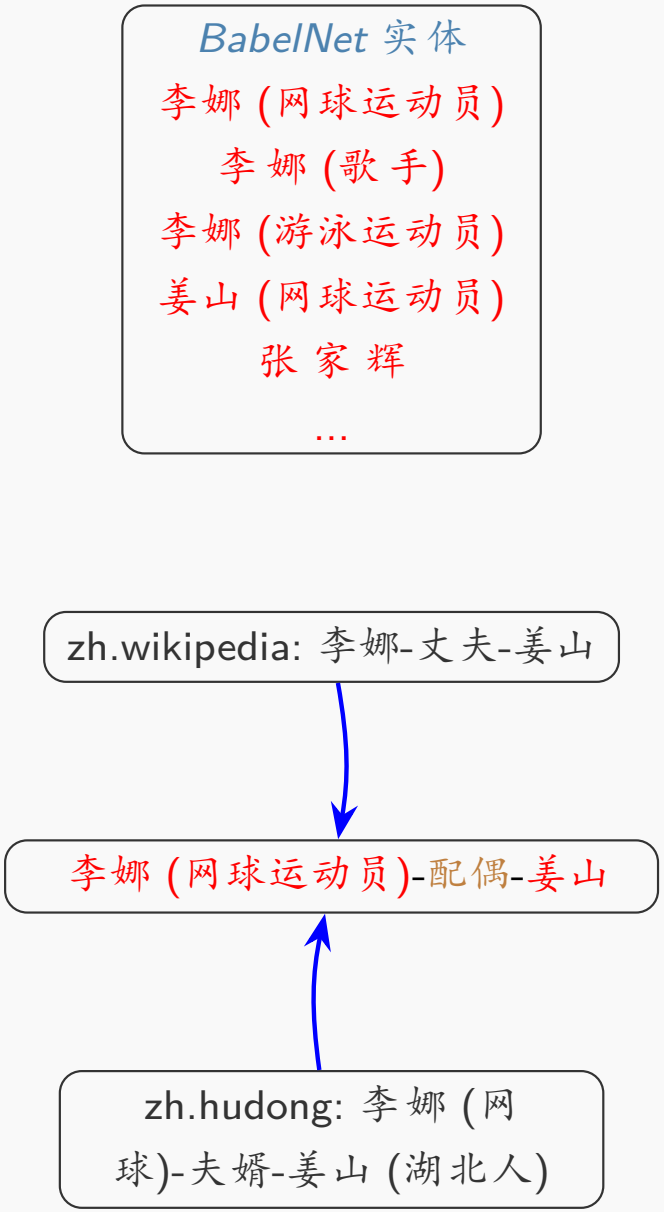
$$\mu_k = \frac{1}{|\mathbf{v}_k|} \sum_{v \in \mathbf{v}_k} \frac{v}{\|v\|}, \quad k \in \{D, G\}$$

- 该关系的主语、客体方差为

$$\sigma_k^2 = \frac{1}{|\mathbf{v}_k|} \sum_{v \in \mathbf{v}_k} (1 - cos\,(v, \mu_k))^2$$

- 该关系的总体方差为主客体方差的平均数
- 方差越小，说明该关系对应的种子三元组越好
- 如果总体方差小于某个阈值 δ，则认为其是好的关系

method/disambiguation



【实体消歧三】使用相关的文本消歧（分类）

对于任意一个三元组 $<e_d, r, e_g>$，比如想要判断 e_d 的含义

- 如果 r 是好的关系，那么 r 内的种子三元组能正确的反应 r 的含义，使用种子三元组的主语可以作为 e_d 消歧的依据
- 如果 r 不是好的关系，反之，只用当前这条三元组的 r, e_g 来作为分类/消歧的依据

归一化方法

— 关系（relation）归一

method/unification

和 disambiguation 过程类似

1. 对于每个知识库的每种关系，计算其主体、客体的平均向量
2. 对于任意两个知识库的任意关系对，计算其相似性如下

$$s_k = \frac{\mu_k^{r_i} \cdot \mu_k^{r_j}}{\|\mu_k^{r_i}\| \, \|\mu_k^{r_j}\|}$$

实验

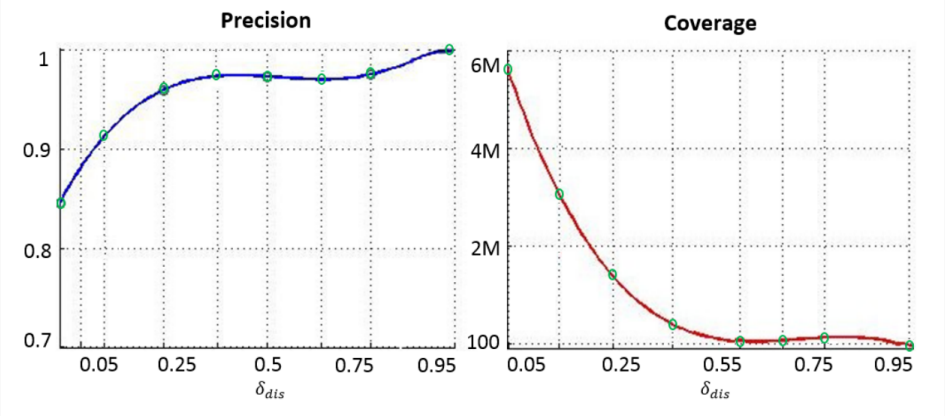
experiment

【数据集】4 个，后两个内部有链接，前两个没有

	K_U		K_D	
	NELL	REVERB	PATTY	WISENET
# relations	298	1 299 844	1 631 531	245 935
# triples	2 245 050	14 728 268	15 802 946	2 271 807
# entities	1 996 021	3 327 425	1 087 907	1 636 307

【额外实验】测试种子三元组的正确率

- 实验的时候认为种子三元组是完全正确的
- 测试是在 patty 数据集测试的，一共 15m 三元组



experiment

【主要实验一】测试消歧准确率

- 上图的 Baseline 是指不选 seed
  - 任意一个词消歧时只把他所在的那条三元组的主语、客体、谓词作为上下文消歧，和该关系的其他三元组无关
  - ζ 越大表示种子三元组越准确，准确率会更高

	SENSEMBED			Baseline		
ζ_{dis}	0.5-0.7	0.7-0.9	0.9-1.0	0.5-0.7	0.7-0.9	0.9-1.0
PATTY	.980	.980	1.000	.793	.780	1.000
WISENET	.958	.960	.973	.726	.786	.791
NELL	.955	.995	1.000	.800	.770	.885
REVERB	.930	.940	.950	.775	.725	.920

Table 2: Disambiguation precision for all KBs

【主要实验一】测试消歧准确率

- 下图的“only seed”是指只采用每个类别的种子三元组来表示给类别的主语、客体，进而预测三元组的主语、客体的类别

	$\delta_{spec} = 0.8$		$\delta_{spec} = 0.5$		$\delta_{spec} = 0.3$	
	all	only seeds	all	only seeds	all	only seeds
PATTY	62.15	26.60	52.49	24.06	40.75	21.41
WISENET	60.00	37.46	54.44	22.26	53.58	16.62
NELL	76.97	62.98	50.95	20.71	44.70	4.36
REVERB	41.20	38.57	25.14	23.70	13.37	12.75

Table 3: Coverage results (%) for all KBs

experiment

【主要实验二】测试 relation 对应的准确率

- 为了方便测试，每个数据集只取前 1w 个关系
  - 随机选取 150 对，人工判定

	PATTY-WISENET		PATTY-REVERB		NELL-REVERB	
δ_{align}	0.7	0.9	0.7	0.9	0.7	0.9
Prec.	.68	**.80**	.58	**.74**	.61	**.75**
# Align.	128k	1.2k	47k	643	2.6k	88
	PATTY-NELL		**WISENET-NELL**		**WISENET-REVERB**	
δ_{align}	0.7	0.9	0.7	0.9	0.7	0.9
Prec.	.66	**1.00**	.70	**.84**	.59	**.87**
# Align.	2.6k	57	381	34	9.9k	169

总结

conclusion

- 同一个关系的不同实例应该有相似的词向量，主语互相比较相像，客体也是
- 不同知识库公用相同的主语、客体，那么不同知识库之间的相同关系应该有相似的主语、客体形式
- 对种子三元组和完美关系这两处筛选/排序使得好的数据起到更大的作用
- 没有统一的评价体系，相互间的优劣不能对比：“我们和别的系统不同，不能照搬，无法比较效果”，很多手工筛选检查

问题？