Knowledge Base Unification via Sense Embeddings and Disambiguation

Claudio Delli Bovi @uniroma1 罗马大学

Luis Espinosa-Anke @upf 庞培法布拉大学

Roberto Navigli @uniroma1

icst-wip

2015 年 10 月 22 日

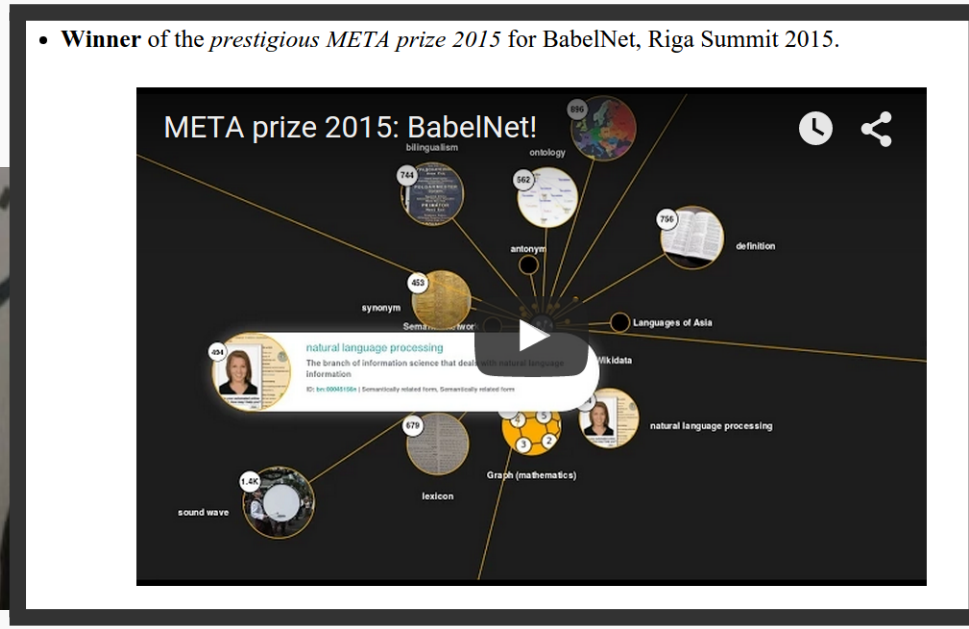
Outline

- summary
  - 问题、方法概述
  - 相关工作
  - 实验使用的工具
- 归一化（Unification）的方法
  - 实体消歧（entity disambiguation）
  - 关系对应（relation alignment）
- 实验

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
- Roberto Navigli
  - Claudio 的导师，07 年博士毕业，BabelNet/Babelfy

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
- Roberto Navigli
  - Claudio 的导师，07 年博士毕业，BabelNet/Babelfy

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

...

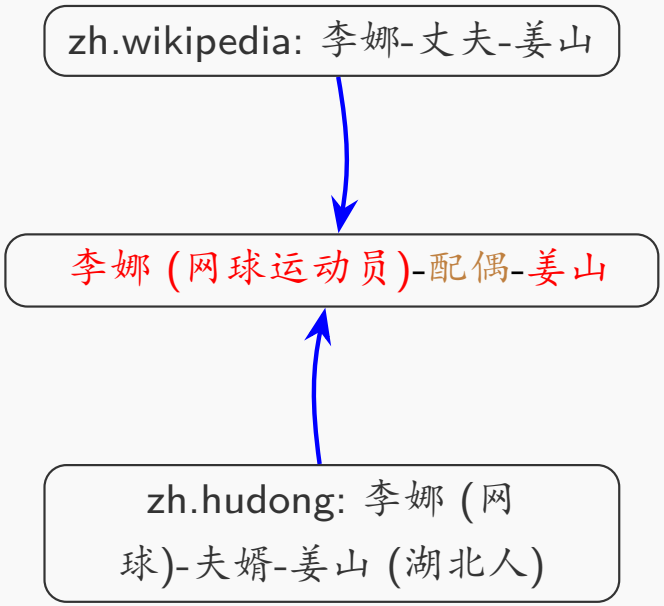
关系

配偶

出生日期

子女

...



如何整合多个知识库?

- 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）
- 给定一个标准的数据库（左上的实体 + 关系）
- 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

...

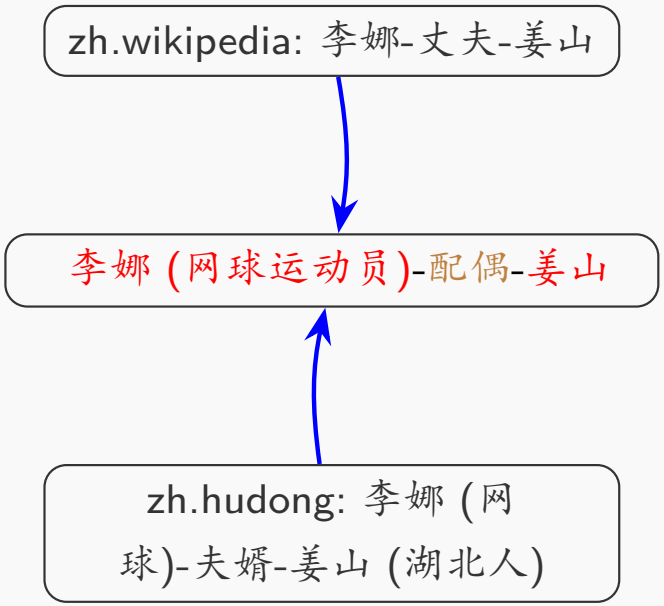
关系

配偶

出生日期

子女

...



如何整合多个知识库?

- **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

...

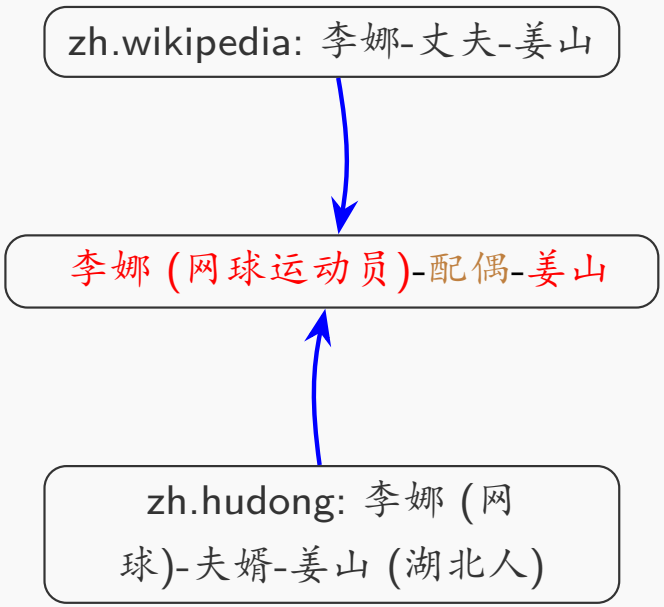
关系

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**
- 预处理：对于标准数据库的实体得到一个语义向量
- **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，如果存在一组组合起来已经非常好了，选为种子三元组
  - 计算每种 relation 的种子三元组中主体、客体的语义向量的均值作为该关系的特征主语向量、特征客体向量

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

...

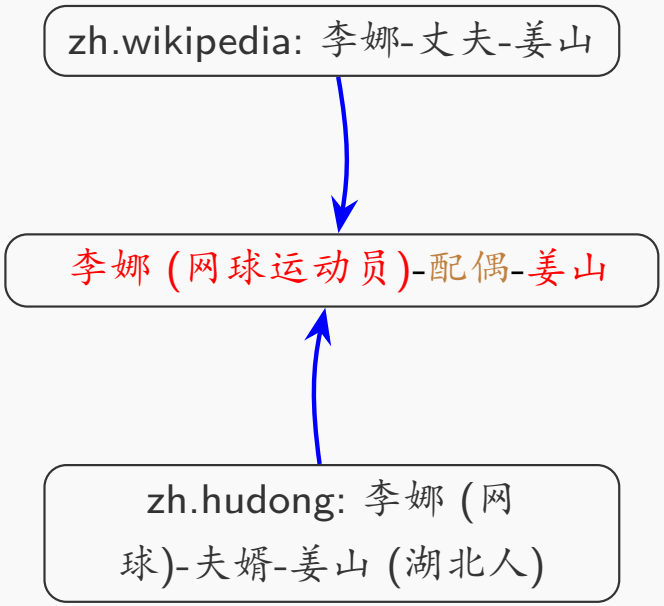
关系

配偶

出生日期

子女

...



- 方法
- **实体消歧**【2】
  - 【一】对于主客体特征向量比较好的 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

...

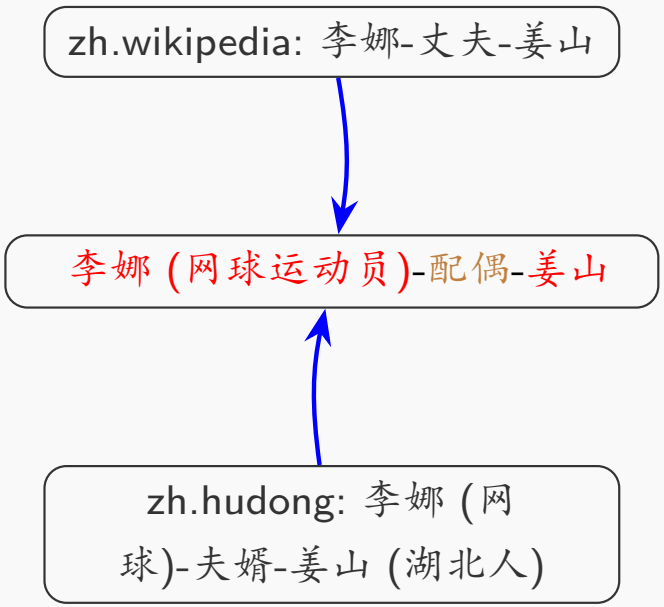
【一】

微软-CEO-纳德拉

搜狐-CEO-张朝阳

苹果 (?)-CEO-库克

...



- 方法
- **实体消歧**【2】
  - 【一】对于主客体特征向量比较好的 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】对于主客体特征向量不好的 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

...

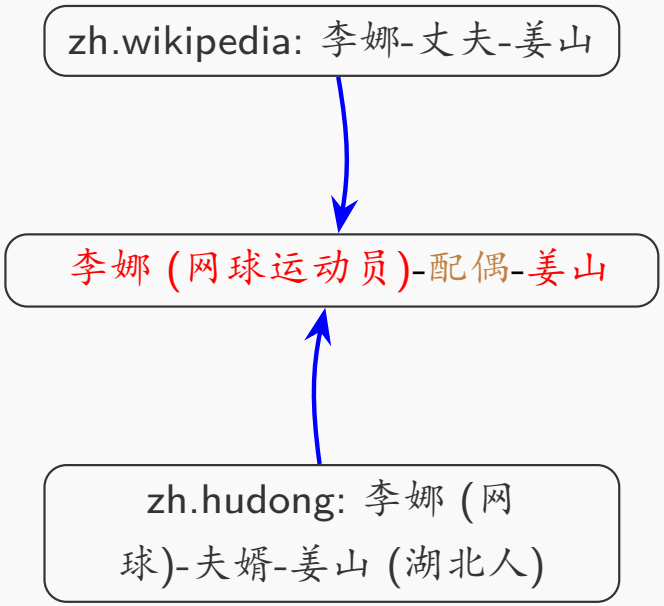
【二】

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**【2】
  - 【一】对于主客体特征向量比较好的 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

...

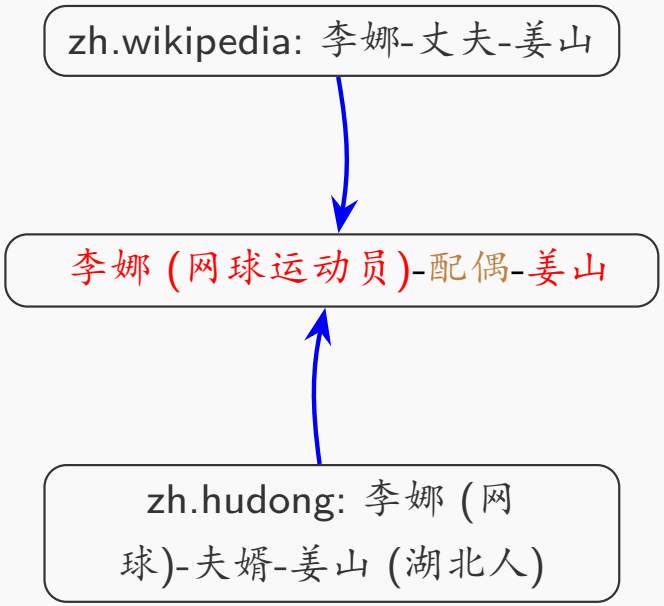
关系

配偶

出生日期

子女

...



- 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