基于多知识库的表格实体链接研究

严晟嘉 09013119
指导老师：李慧颖教授
顾问老师：漆桂林教授
特别感谢：吴天星博士

提纲

* 动机
* 方法一
* 方法二
* 实验与分析
* 总结与展望

提纲

* 动机
* 方法一
* 方法二
* 实验与分析
* 总结与展望

动机

表格实体链接任务

* 建立表格单元格中的字符串指称与其在给定知识库中的参考实体的链接

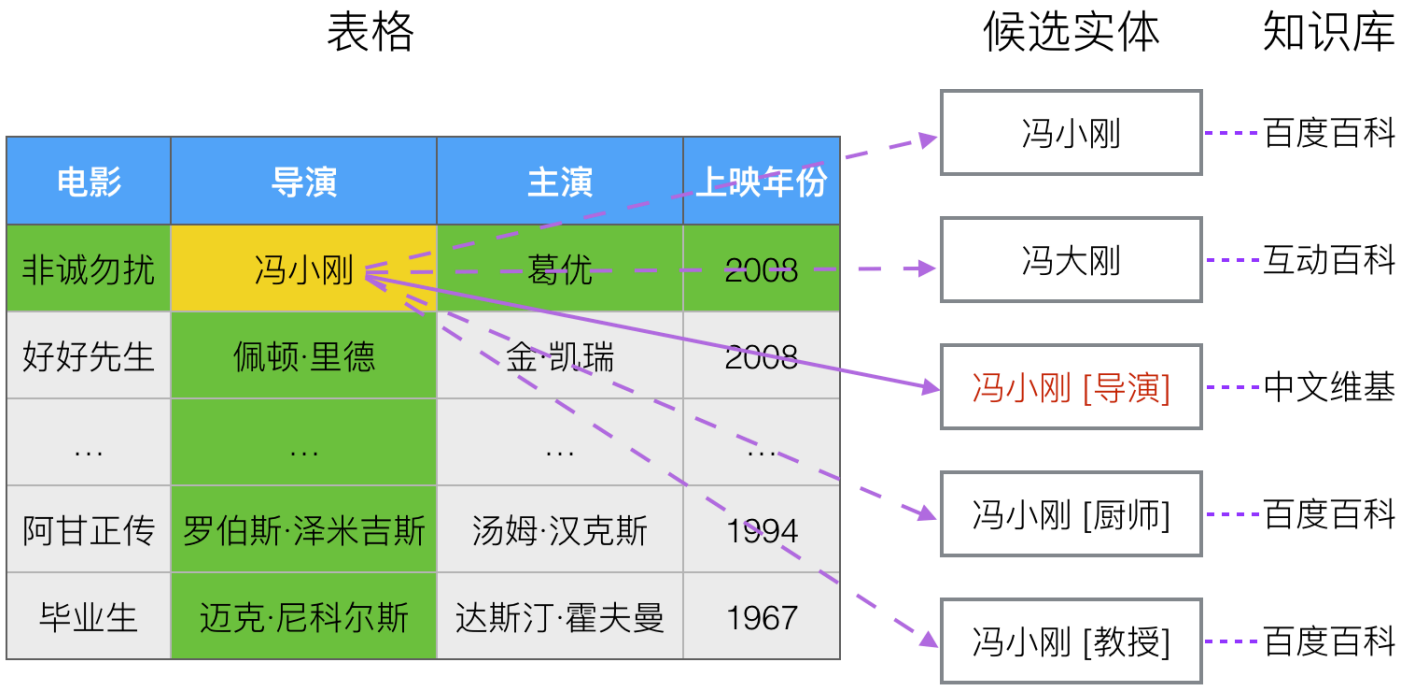

图 1 一个表格实体链接的例子

问题与解决方案

当前研究工作中存在的主要问题：

* 强烈依赖于特定信息，比如表格的表头和目标知识库中的实体类别，然而这些特定信息并不总是存在

* 只考虑用一个单知识库进行实体链接，并不能总是保证很好的实体覆盖度 (Coverage)

我的解决方案：

两个基于多知识库的 Web 表格实体链接的通用方法

* 多知识库 -> 替代单知识库

* 通用 -> 不使用任何特定信息

提纲

* 动机
* 方法一
* 方法二
* 实验与分析
* 总结与展望

方法一：两步走

该方法包含**两个主要的步骤**

(1) 使用一个基于图的随机游走算法

进行单知识库实体链接

(2) 通过多知识库间的 sameAs 关系和

三条启发式规则提升第一步的链接质量

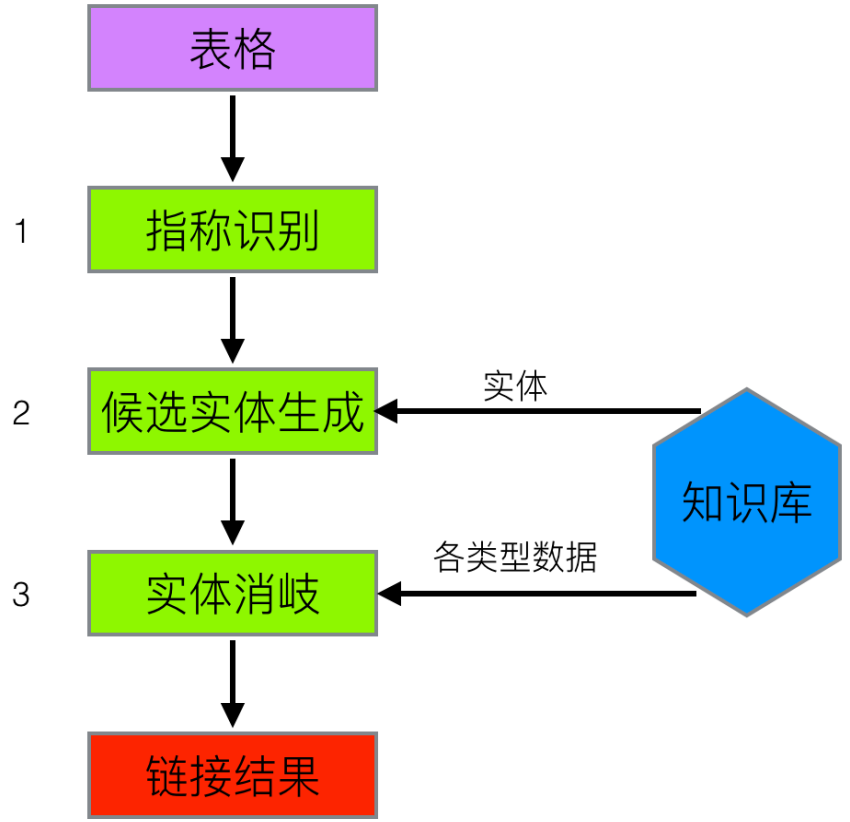

图 2 实体链接一般流程

单库链接

1) **指称识别**
   识别出表格单元格中的每个指称

2) **候选实体生成**
   为每个指称生成来自给定知识库的候选实体集合

3) **实体消岐**
   从指称的候选实体集合中挑选一个实体作为该指称对应的参考实体
   分为三小步:
   a) 构建实体消岐图
   b) 计算实体链接影响因子
   c) 迭代概率传播

候选实体生成

1) 如果给定知识库中的一个实体 e 包含某个指称 m 的分词集合中的至少一个单词，那么实体 e 可能是指称 m 的一个候选参考实体；然后给定一个阈值，如果指称 m 和实体 e 的字符串相似度高于阈值，则将实体 e 加入候选实体集合。

2) 如果给定知识库中的一个实体 e 在 BabelNet 中的一个同义词 s 包含某个指称 m 的分词集合中的至少一个单词，那么实体 e 可能是指称 m 的一个候选参考实体；然后检测指称 m 和同义词 s 的字符串相似度，如果高于阈值，则将实体 e 加入候选实体集合。

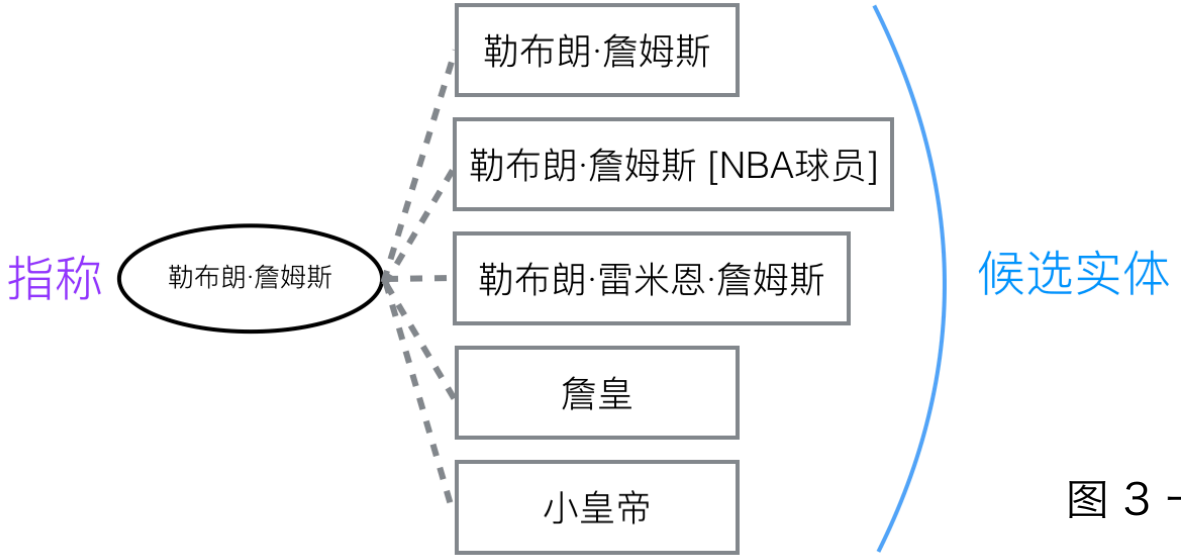

图 3 一个指称与它的候选实体集合

实体消岐

表格中同行同列的指称可能是语义相关的。

一个指称与其候选实体可能是语义相关的。

同一张表格中指称的候选实体之间可能是语义相关的。

利用这些语义相关性可以提升实体链接的质量。

* 马尔科夫链可以用于捕捉这些关系。

* 消岐的过程就是给候选实体排名的过程，与 PageRank 思想类似。

电影	导演	主演	上映年份
非诚勿扰	冯小刚	葛优	2008
好好先生	佩顿·里德	金·凯瑞	2008
…	…	…	…
阿甘正传	罗伯斯·泽米吉斯	汤姆·汉克斯	1994
毕业生	迈克·尼科尔斯	达斯汀·霍夫曼	1967

与“冯小刚”语义相关的指称

图 4 一个表格实例

因此，方法一使用了一个基于图的随机游走算法来对一张表格中的所有指称进行联合消岐。

马尔科夫链　　Personalized PageRank

构建实体消岐图

为每张表格构建独特的实体消岐图：无向带权图，两种结点，两种边

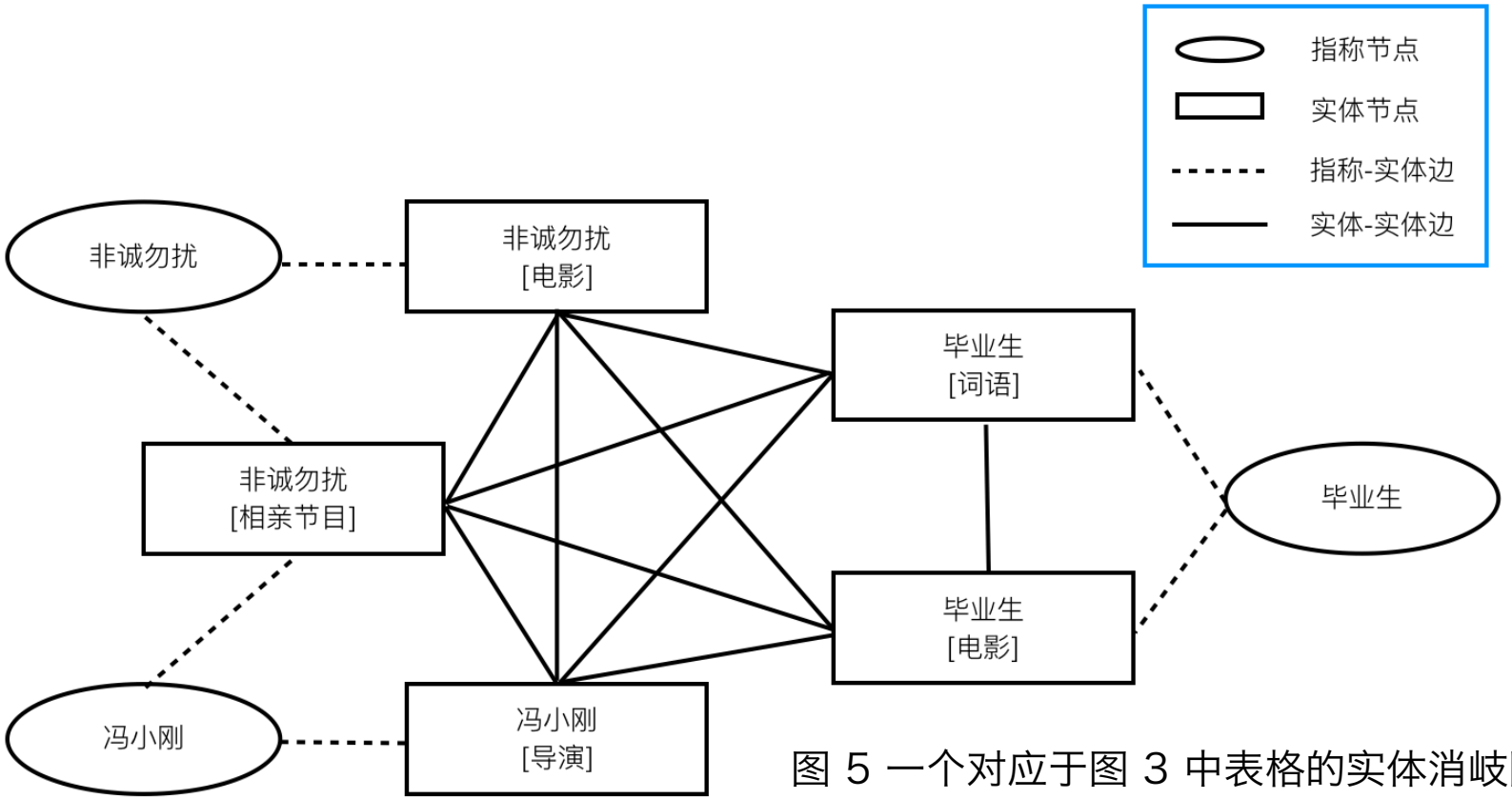

图 5 一个对应于图 3 中表格的实体消岐图的一部分

计算EL影响因子

实体链接影响因子分为两类：

(1) 指称结点上的概率：指称的初始权重值

每个指称都认为是平等的，当表格中有 k 个指称，那么每个指称的权重值初始化为 1/k 。

(2) 边上的概率：结点间的语义相关度

* 指称与实体之间的语义相关度：
    * 字符串相似度特征
    * 指称-实体上下文相似度特征
* 实体与实体之间的语义相关度：
    * 三元组关系特征
    * 实体-实体上下文相似度特征

实体消岐图的每个结点和每条边上都被赋予了一个概率，实体结点上的概率代表该实体成为某指称的参考实体的概率（初始化为0）。

迭代概率传播

迭代概率传播将不同的EI影响因子结合起来得到实体链接结果。

给定一个有 n 个结点的实体消岐图 G，我将 G 表示为一个 $n \times n$ 的邻接矩阵 A，A_{ij} 指的是结点 i 到结点 j 的转移概率。

定义了一个 $n \times 1$ 的矩阵 r 来记录所有结点上的概率，并且使用以下公式迭代地计算 r 直到收敛。

$$\boldsymbol{r}^{t+1} = ((1-d) \times \frac{\boldsymbol{E}}{n} + d \times \boldsymbol{A}) \times \boldsymbol{r}^{t}$$

t：迭代轮数

d：衰减系数

E：$n \times n$ 单位阵

Zhishi.me 上的一个测试

1) 将方法一中的单库链接算法应用到超过70000张 Web 表格上，在 Zhishi.me 中的三个知识库上都跑一遍。

2) 对于每个指称，如果最终链接到的来自不同知识库的任意两个实体间没有 sameAs 关系，那么认为此时 EL 出现了冲突。

3) 根据统计，有 38.94% 的指称的链接结果存在这种冲突。

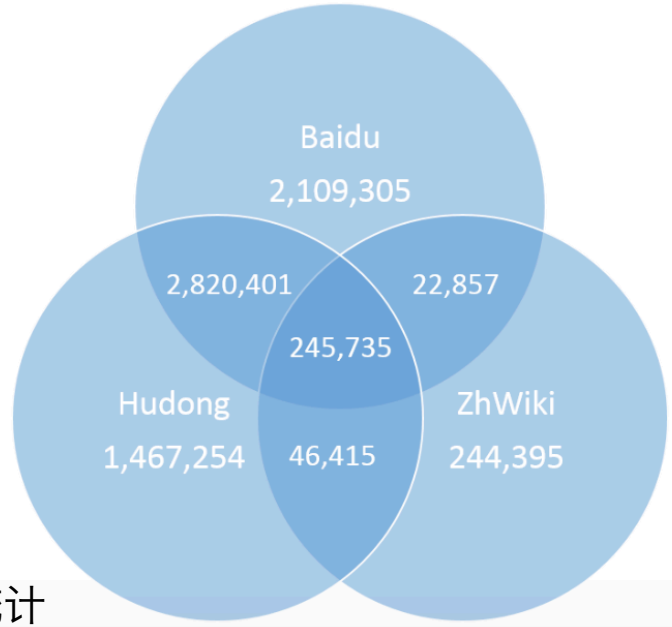

图 6 Zhishi.me 数据统计

多库优化

为了解决实体链接结果冲突的问题，方法一中利用了多知识库间的 sameAs 关系和三条启发式规则来优化单库链接的结果。

三条启发式规则：

* **规则一**: 如果一个实体分组的*平均排名*、*最高排名*在所有分组中排名最高，并且该组的*实体数量*不少于知识库数量的一半，那么选择这一组作为给定指称的最终实体链接结果。

* **规则二**: 如果有两个或多个实体分组的*平均排名*、*最高排名*相同并在所有分组中排名最高，并且这些分组的*实体数量*不少于知识库数量的一半，那么从这些分组中随机挑选一组作为给定指称的最终实体链接结果。

* **规则三**: 如果每组的*实体数量*都小于知识库数量的一半，那么对于给定指称，原先单知识库实体链接结果保持不变。

多库优化

根据知识库间的 sameAs 关系，将给定指称的来自三个知识库的候选实体分组。

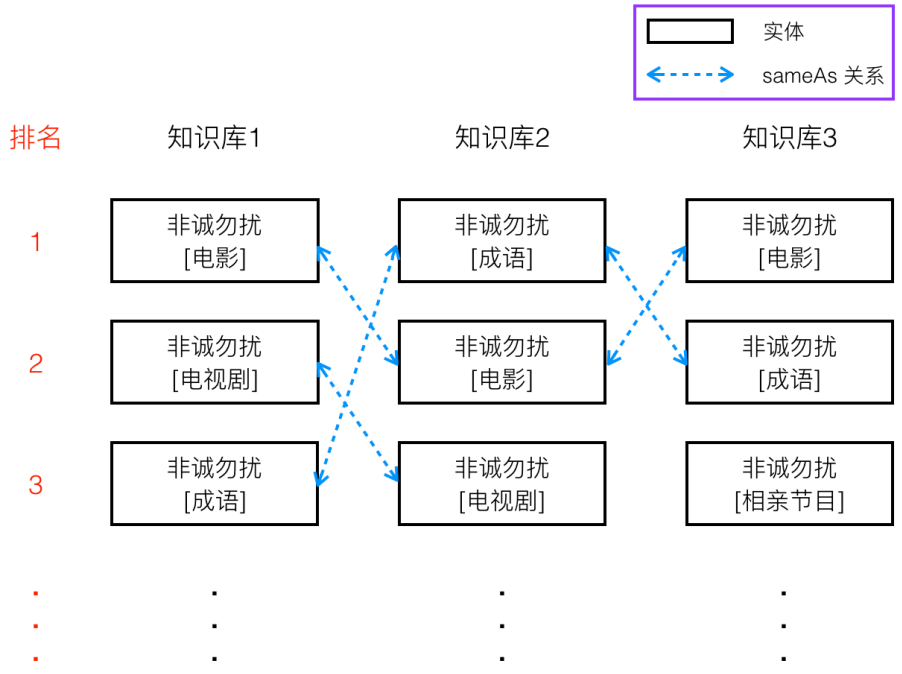

图 7 指称“非诫勿扰”在三个知识库上的候选实体排名列表

Set1={“非诫勿扰[电影]” (KB1), “非诫勿扰[电影]” (KB2), “非诫勿扰[电影]” (KB3)}
Set2={“非诫勿扰[成语]” (KB1), “非诫勿扰[成语]” (KB2), “非诫勿扰[成语]” (KB3)}
Set3={“非诫勿扰[电视剧]” (KB1), “非诫勿扰[电视剧]” (KB2)}
Set4={“非诫勿扰[相亲节目]” (KB3)}

对于 Set1 中的实体:
平均排名 = (1+2+1)/3 = 1.33
最高排名 = 1
实体数量 = 3

满足规则一!

所以 Set1 中的实体被选为指称“非诫勿扰”的最终实体链接结果

提纲

* 动机
* 方法一
* 方法二
* 实验与分析
* 总结与展望

方法二：融合

提出方法二的动机在于一步到位地得到指称在多知识库上的链接结果。

用一个统一的图模型来表示一张表格中的所有指称，其对应的候选实体以及多知识库间的 sameAs 关系。

方法二与方法一中的单库链接算法流程是大致相同的，不同的地方在于实体消岐的实现。

与方法一相比，方法二主要有**两个不同之处**:

(1) 实体消岐图中的结点定义改变了

(2) 舍弃了启发式规则

新实体消岐图

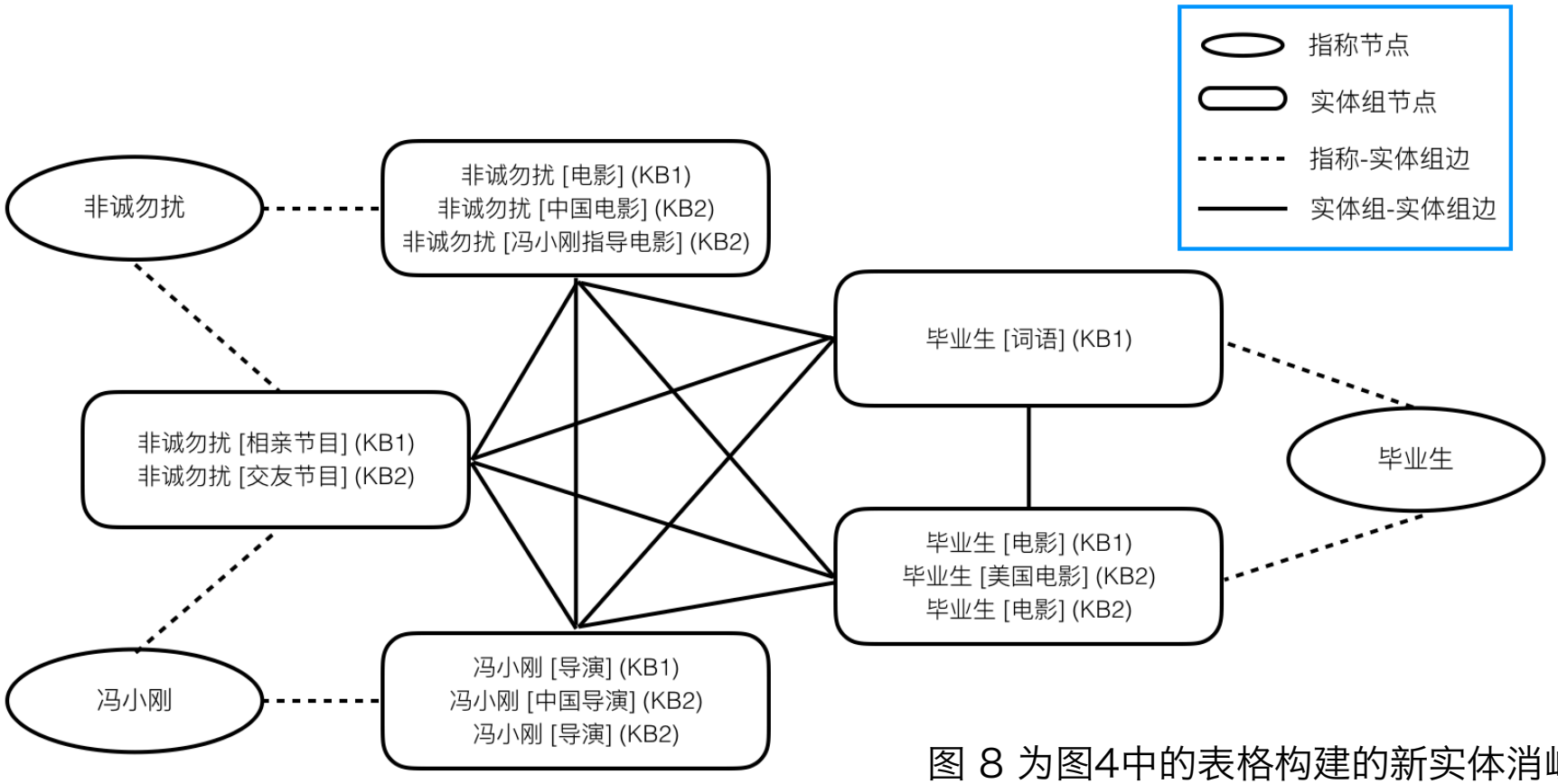

图 8 为图4中的表格构建的新实体消岐图的一部分

提纲

* 动机
* 方法一
* 方法二
* 实验与分析
* 总结与展望

数据集

因为基于多知识库的表格实体链接是一个新任务，所以缺乏基准数据集 (benchmark)。

因此需要针对我的算法构建自己的 benchmark。

表格语料库

我从70000张从 Web 上爬取的表格中随机挑选了200张。

我邀请了4位同学一起对这200张表格进行了人工标注，将表格单元格中的每个指称分别手动映射到 Zhishi.me 的三个知识库中的实体上。人工标注的结果基于服从多数原则。

实体知识库

实验以目前最大的中文百科类知识库 Zhishi.me 作为实体的来源，其包含了三个互有重叠的知识库(百度百科，互动百科，中文维基百科)。

对比实验

* *TabEL*：目前 Web 表格实体链接领域先进的系统，它使用一种使用了许多通用特征的集体分类技术来对一个给定 Web 表格中的所有指称进行联合消岐。除此之外，任何知识库都可以被应用于 *TabEL* 来执行 Web 表格上的实体链接任务。

* *LIEGE*：是一个通用方法，用于将形如列表 (List-like) 的 Web 表格 (多行一列) 中的字符串指称链接到给定知识库中的参考实体。它提出了一种使用了三个特征的迭代置换算法来执行 Web 列表中的实体链接。这个方法同样可以用于任何知识库上的 Web 表格实体链接。

* *single*：是方法一的一个退化版本。它只使用了方法一中单知识库实体链接的算法，并没有运用三条启发式规则和 sameAs 关系来执行多知识库对实体链接结果的优化算法。

* *multiple*：也是方法二的一个退化版本。在执行完单知识库实体链接算法后，它仅使用了已存在的 sameAs 关系 (不包括新学习到的 sameAs 关系) 来提升实体链接结果质量。

* *approach1*：方法一

* *approach2*：方法二

Knowledge Base	Approach	Precision	Recall	F1-score	MRR
Chinese Wikipedia	TabEL	0.823	0.809	0.816	0.858
	LIEGE	0.778	0.747	0.762	0.813
	single	0.830	0.797	0.813	0.860
	multiple	0.861	0.821	0.841	0.881
	approach1	**0.873**	**0.828**	**0.850**	**0.887**
	approach2	**0.856**	**0.830**	**0.843**	**0.814**
Baidu Baike	TabEL	0.659	0.628	0.643	0.707
	LIEGE	0.629	0.576	0.601	0.670
	single	0.696	0.652	0.673	0.725
	multiple	0.758	0.705	0.731	0.746
	approach1	**0.774**	**0.727**	**0.750**	**0.776**
	approach2	**0.769**	**0.747**	**0.758**	**0.780**
Hudong Baike	TabEL	0.681	0.649	0.665	0.780
	LIEGE	0.661	0.632	0.646	0.751
	single	0.708	0.642	0.673	0.768
	multiple	0.729	0.700	0.714	0.787
	approach1	**0.744**	**0.708**	**0.726**	**0.796**
	approach2	**0.731**	**0.712**	**0.721**	**0.788**

由三个单知识库衡量的总体实体链接结果

由整个 Zhishi.me 衡量的实体链接结果

* 方法一在整个 Zhishi.me 上的**准确率**，**召回率**和**F1值**分别为 0.831，0.903 和 0.866。
* 可以发现，召回率相比一些单库链接算法有显著的提升，说明基于多知识库的实体链接算法能够确保对实体由一个很好的覆盖率。

提纲

* 动机
* 方法一
* 方法二
* 实验与分析
* 总结与展望

总结

提出了两种新的基于多知识库的表格实体链接方法：

(1) 方法一是一个两阶段法。首先使用了一个基于图的随机游走算法进行单库链接，然后使用多库间的 sameAs 关系和三条启发式规则优化链接质量。

(2) 方法二融合了方法一中的两步。用一个统一的图模型表示表格中的指称、候选实体以及实体间的 sameAs 关系。

实验结果表明，两个方法在**准确率**、**召回率**上相比当前的一些表格实体链接算法都有所提升，这证明了基于多库的表格实体链接算法的有效性以及其能够确保一个好的实体覆盖度。

同时，这两个方法不依赖与特定信息，主要基于表格中指称与实体、实体与实体间的语义相关度来开展实体链接的工作。

除此之外，这两个表格实体链接方法可以基于任何单知识库或者相互链接的多知识库来运行。

未来工作

* 当前表格实体链接的主要挑战之一在于缺乏基准数据集，没有好的 benchmark 很难衡量算法的有效性。所以未来希望能构建更多高质量的基准数据集。

* 将上述两个方法扩展为跨语言 (cross-lingual) 的表格实体链接方法。

* 将这两个方法进行封装，做成 API 或者 Web 应用以供他人使用。

* 在实现了实体链接之后，下一步就是关系抽取。关系抽取是表格语义解释的主要任务 之一。根据表格中不同列实体链接的结果，通过两列某一行的关系即可得到表格列之间的关系，最终的结果表示为 RDF 三元组。

谢谢！

[Paper]: http://yanshengjia.com/file/seuthesis.pdf

[Code]: https://github.com/yanshengjia/link/tree/master/src