

基于中文维基百科构建的 知识库的谓词归一化

韩喆

iampkuhz@gmail.com

2015 年 4 月 2 日

- 背景
 - 基于维基百科的知识库
- Motivation
- 实验方法
- 特征选取
- 实验效果
 - 分析和改进

Background

基于维基百科的知识库

张家辉



2010年8月24日参加电影《线人》江展首映礼。

男演员

羅馬拼音 Cheung Ka Fai
英文名 Nick Cheung
國籍  中国 (香港)
籍貫 廣東番禺
出生 1967年12月2日 (47歲)
 英屬香港
語言 粵語、英語、普通話
配偶 關詠荷 (2003年至今)
兒女 張童 (Brittany Cheung)
- 2006年01月24日 (9歲)
活躍年代 1987年至今
經紀公司 錢瑋[1]
獎項 最佳男演員 - 上海國際電影節
2013年《线人》- 狂輝
最佳男演員 - 香港電影評論學會

Subject	Predicate	Object
张家辉	罗马拼音	Cheung Ka Fai
张家辉	英文名	Nick Cheung
张家辉	国籍	中国(香港)
张家辉	籍贯	广东番禺
张家辉	出生	1967年12月2日 (46岁)
张家辉	语言	粤语
张家辉	语言	英语
张家辉	语言	普通话
张家辉	配偶	关咏荷(2003年至今)
张家辉	儿女	张童(Brittany Cheung)
张家辉	活跃年代	1989年至今
张家辉	经纪公司	钟珍

- 知识库的谓词数量多
 - 1.59w, 手工排查后变成 1.4w
- 谓词冗余
 - 含有“邮政”的谓词 (17):
INSEE/邮政编码、ISO 3166-2 邮政简写、美国邮政编号、美国邮政编码、邮政、邮政代码、邮政信箱、邮政分区、邮政区号、邮政号码、邮政简称、邮政编号、邮政编号字母、邮政编码、邮政编码 FSA、邮政编码首字母、邮政缩写
- 进行谓词归一

- 假设/前提

- 我们在 1.4w 个候选谓词内部进行实验，提供任意两个谓词的相似性，进而判断是否是相同谓词
- 假设所有字符相同的谓词都是同一谓词，所有字符不同的谓词都非同谓词
 - 姚明：出生：上海 vs 刘翔：出生地：上海
 - 姚明：出生：上海 vs 刘翔：出生：1983 年 7 月 13 日
- 转换问题为二分类：给定任意两个谓词对，判断其是否是相同谓词
 - 训练数据格式：[*true/false*, *PredicateId*₁, *PredicateId*₂]
 - 测试数据格式：[*PredicateId*₁, *PredicateId*₂]

- 实验环境/数据
 - 标注了 1700 多个谓词对
 - 谓词对本身根据规则（有一定拼音、字符串等特征的相似性）抽取，非随机抽取两个谓词加入训练/测试数据中
 - 785 个相同谓词对（47.3），873 个不同谓词对
 - 测试集 1000 个单词对，训练集 500 个单词对

● 实验步骤

- ① 对于每个谓词 (1/14000), 统计其信息 (不同类别的特征)

出生 : pinyin={chusheng}, Content={出生}, SubjectCategory={(篮球运动员,10),(足球运动员,100),(政治人物,50)}...

- ② 对于任意两个谓词, 比较其每类特征的相似性, 转化为数值, 生成特征向量

出生, 出生地 : pinyinSim=0.67, ContentSim=0.67, SubjectCategorySim=0.38,...

- ③ 对于训练数据, 提取特征向量, 训练模型
- ④ 对于测试数据, 提取特征向量, 根据模型预测是否为同一谓词

- 已选特征
 - 文本相似度
 - 拼音相似度
 - 词频相似度
 - wikitext 相似度
 - 主体的二级类别相似度

- 文本相似度

- ① 相同单词个数/总长度 (2 维)
- ② $\min(\text{编辑距离}/\text{总长度}, 1)$ (2 维)
- ③ 61.8% correct on SVM

- 拼音相似度

- ① 同文本相似度计算方式, 比较字符相同时改用拼音判段是否相同
- ② 53.3% correct on SVM

- 词频相似度















- ① 初衷是希望出现频率差别越大的谓词越应当合并 (判重), 实际基本没有效果

Method

● wikitext 相似度: 期望的重点

張家輝	
男演員	
罗马拼音	Cheung Ka Fai
英文名	Nick Cheung
国籍	 中国（香港）
籍贯	广东番禺
出生	1967年12月2日（47岁） <div> 英属香港</div>
语言	粤语、英语、普通话
配偶	关咏荷（2003年至今）
儿女	张童（Brittany Cheung） <div>- 2006年01月24日（9岁）</div>
活跃年代	1987年至今
经纪公司	锺珍 ^[1]

任何侵权内容将会删除 | 百科内容须附有来源，以供查证

A A                   

- **wikitext 相似度**: 期望的重点
 - 没有固定的对应规则
 - (比方说) 编辑者在“Template: 男艺人”页面写了一个转换说明, 把“出生日期”自动转化为“出生”显示。如果没有定义, 则用模板“Template: 人物”的规则匹配。且说明页面非结构化, 不能自动抽
 - 收集了从 wikitext 抽取的三元组, 利用手写规则与从网页抽取的三元组做对应, 然后做统计

内核类别 : {(kernel type,132),(screenshot2),(logo,2) ,(name,2),(kernel,1) }

出生 : {(birth place,11470),(birth date,7598),(出生地点,7241),(出生日期,6789),(date of birth,3775),(place of birth,3690),(term start,2346),(出生地,2076),(term end,1511),(birthplace,1156)...}

Method

- wikttext 相似度: 期望的重点
 - 实验效果

- wikttext 相似度: 期望的重点
 - 实验效果
 - SVM 分类失败（全部预测为 1）

- **wikitext 相似度**: 期望的重点
 - 实验效果
 - SVM 分类失败（全部预测为 1）
 - 失败原因
 - 80% 的测试数据的相似值为 0。很多时候有一个谓词没有对应的 wikitext，尤其是出现频率少的谓词
 - 下一步修正
 - 观察没有抽到 wikitext 的谓词信息，修改代码（理论上都是可以对应有 wikitext 的）

- 主体的二级类别相似度
 - 假设前提：**相同意义的谓词，其出现的三元组主体应该是类型分布应该是一致的。**
 - “出生”作为谓词出现的三元组，主语类别分布 $\{(人物, 10000), (动物, 100)\}$
“出生日期”的主语类别分布 $\{(人物, 2000), (动物, 500)\}$
 - 实验方法
 - 利用中文维基百科的类别，“页面分类”下面的子类 (22-2) 作为类别分布的规约终点
语言, 跨學科領域, 应用科学, 文学, 艺术, 宗教, 休閒, 科技, 心理学, 人物, 地理, 人文學科, 技术, 社会, 历史, 幫助, 資訊, 科学, 總類, 自然科学, 社会科学, 哲学,

- 主体的二级类别相似度
 - 假设前提：**相同意义的谓词，其出现的三元组主体应该是类型分布应该是一致的。**
 - “出生”作为谓词出现的三元组，主语类别分布 $\{(人物, 10000), (动物, 100)\}$
“出生日期”的主语类别分布 $\{(人物, 2000), (动物, 500)\}$
 - 实验方法
 - 利用中文维基百科的类别，“页面分类”下面的子类 (22-2) 作为类别分布的规约终点
语言, 跨學科領域, 应用科学, 文学, 艺术, 宗教, 休閒, 科技, 心理学, 人物, 地理, 人文學科, 技术, 社会, 历史, 幫助, 資訊, 科学, 總類, 自然科学, 社会科学, 哲学,

- 主体的二级类别相似度实验方法一。
 - 利用维基百科的类别体系，建立所有类别到这 22 个类别的对应关系

分类:美国篮球运动员

页面分类 > 人物 > 职业 > 各职业美国人 > 美国运动员 > 美国篮球运动员

页面分类 > 人物 > 各国人物 > 各国运动员 > 美国运动员 > 美国篮球运动员

页面分类 > 人物 > 各职业人物 > 运动员 > 篮球运动员 > 美国篮球运动员

... > ... > 球类运动 > 篮球 > 篮球运动员 > 美国篮球运动员

... > ... > 篮球 > 各国篮球 > 美国篮球 > 美国篮球运动员

... > ... > 各国体育 > 美国体育 > 美国运动员 > 美国篮球运动员

... > ... > 各国体育 > 美国体育 > 美国篮球 > 美国篮球运动员

... > ... > 各国体育 > 各国运动员 > 美国运动员 > 美国篮球运动员

... > ... > 各国体育 > 各国篮球 > 美国篮球 > 美国篮球运动员

- 20 个节点宽度优先向下搜索, **深度优先失败**, 所有类别都是语言的子类
- 如果有“**雷·阿伦**: **出生**: 加利福尼亚州”

雷阿伦属于类别“美国篮球运动员”

出生: {(人物, 100), (科技, 10), ...} -> {(人物, 101), (科技, 10), ...}

- 主体的二级类别相似度
 - 实验方法二。