

# Constructing a Comprehensive Knowledge Base from DBpedia

Xuying@Monash

# Introduction to DBpedia

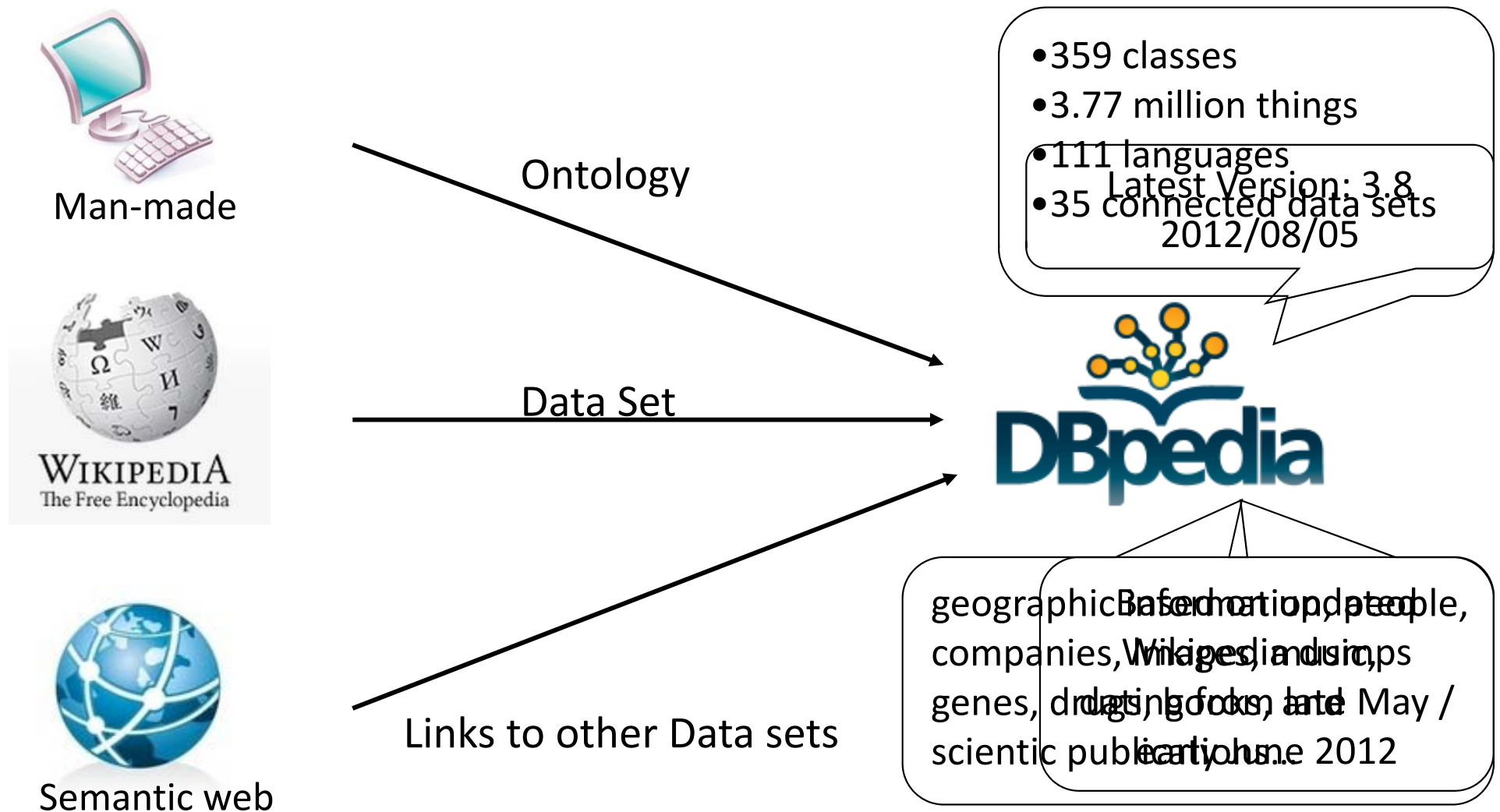


Fig. 1 Introduction to DBpedia

[illegible]

2

# Problem Statement

- DBpedia is less formally structured;
- Data quality is lower and there are inconsistencies within DBpedia;
- Localized DBpedia is version based rather than up-to-date;
- Italians know more about Italian villages.

# Wikipedia v.s. Baidu Baike


女歌手	
国籍	 中华人民共和国
籍贯	沈阳市
出生	1967年11月27日（44岁）
职业	女歌手、演员
语言	普通话（国语）
配偶	高峰（？——？） <sup>[1]</sup> 孟桐（2006年——） 长子：小名高兴（父高峰，2004年出生） 长女：小名苹果（父孟桐，2007年出生）
儿女	
音乐类型	华语流行音乐
活跃年代	1990年代至今
唱片公司	福茂唱片（1994—1996） 科艺百代（1997—2000） 华纳唱片（2001—2004） 亚神音乐（2011—）
经纪公司	银鱼音乐（1998—2004） 银河树文化（2011—）

Fig. 4 Wikipedia Inforbox

Similarities: both have the most basic and significant information in the title card, which consists of a description paragraph and an inforbox.

Interesting discovery:

1, Baidu Baike always contains information which are more specific for Chinese.

中文名:	那英	主要成就:	东方风云榜最佳女歌手
别名:	格格		MTV亚洲大奖最受欢迎歌手
国籍:	中国		台湾金曲奖最佳国语女歌手
民族:	满族		香港十大中文金曲全国最佳女歌手
出生地:	中国辽宁		北京流行音乐典礼杰出成就大奖
出生日期:	1967-11-27（农历：十月廿六）	身高:	168厘米
职业:	演员，歌手		
代表作品:	《征服》《心酸的浪漫》《一笑而过》 《那又怎样》《长镜头》		
新浪微博:	<a href="http://t.sina.com.cn/naying">http://t.sina.com.cn/naying</a>		

Fig. 5 Baidu Baike Inforbox

# Wikipedia v.s. Baidu Baike

2, Baidu Baike does better in catching up with the latest events happened in China.

《中国好声音》四大导师团队成员				
刘欢团队	▪ 李代沫 (第一期)	▪ 徐海星 (第一期)	▪ 王乃恩 (第二期)	▪ 刘
	▪ 袁娅维 (第三期)	▪ 佳宁组合 (第三期)	▪ 权振东 (第四期)	▪ 刘
	▪ 郑虹 (第四期)	▪ 陈斌 (第五期)	▪ 刘昊霖 (第五期)	▪ 李
	▪ 吉克隽逸 (第五期)	▪ 山野 (复活赛)		
那英团队	▪ 张玉霞 (第一期)	▪ 黄鹤 (第一期)	▪ 赵露 (第一期)	▪ 黄勇 (第一期)
	▪ 张玮 (第一期)	▪ 多亮 (第二期)	▪ 卓义峰 (第四期)	▪ 梁博 (第四期)
	▪ 李敏 (第五期)	▪ 汪妤凌 (第五期)	▪ 歌浴森 (第五期)	▪ 侯祖辛 (第六期)
	▪ 王琪玮 (第六期)	▪ 张赫宣 (复活赛)		
庾澄庆团队	▪ 李维真 (第一期)	▪ 王韵壹 (第二期)	▪ 葛林 (第二期)	▪ 吴莫愁 (第三期)
	▪ 褚乔 (第三期)	▪ 陈俊彤 (第四期)	▪ 谢丹 (第四期)	▪ 阿蜜丝女孩 (第四期)
	▪ 金池 (第四期)	▪ 王克 (第五期)	▪ 魏语诺 (第五期)	▪ 陶虹旭 (第六期)
	▪ 赵可 (第六期)	▪ 大山 (复活赛)		
杨坤团队	▪ 邹宏宇 (第一期)	▪ 丁丁 (第二期)	▪ 平安 (第二期)	▪ 伍佳丽 (第三期)
	▪ 张玮琪 (第三期)	▪ 倪雅丰 (第三期)	▪ 丁少华 (第三期)	▪ 汪洋 (第三期)
	▪ 金志文 (第四期)	▪ 关喆 (第四期)	▪ 周礼虎 (第五期)	▪ 黄一 (第五期)
	▪ 尼克 (第六期)	▪ 曹寅 (第六期)		

Fig. 7 Name List of The Voice of China in Baidu Baike

季度	教练/选手			
	杨坤	那英	刘欢	庾澄庆
1	邹宏宇 丁丁 平安 伍佳丽 张玮琪 倪雅丰 丁少华 汪洋 金志文 关喆 周礼虎 黄一 尼克 曹寅	黄鹤 黄勇 张玮 赵露 张玉霞 多亮 卓义峰 梁博 李敏 汪妤凌 歌浴森 侯祖辛 王琪玮 张赫宣	徐海星 李代沫 刘悦 王乃恩 袁娅维 佳宁组合 权振东 刘振宇 郑虹 吉克隽逸 李行亮 陈斌 刘昊霖 李昊瀚	李维真 王韵壹 葛林 吴莫愁 褚乔 陈俊彤 谢丹 阿蜜丝女孩 金池 王克 魏语诺 陶虹旭 赵可 大山

Fig. 6 Name List of The Voice of China in Wikipedia

# DBpedia Extraction Framework

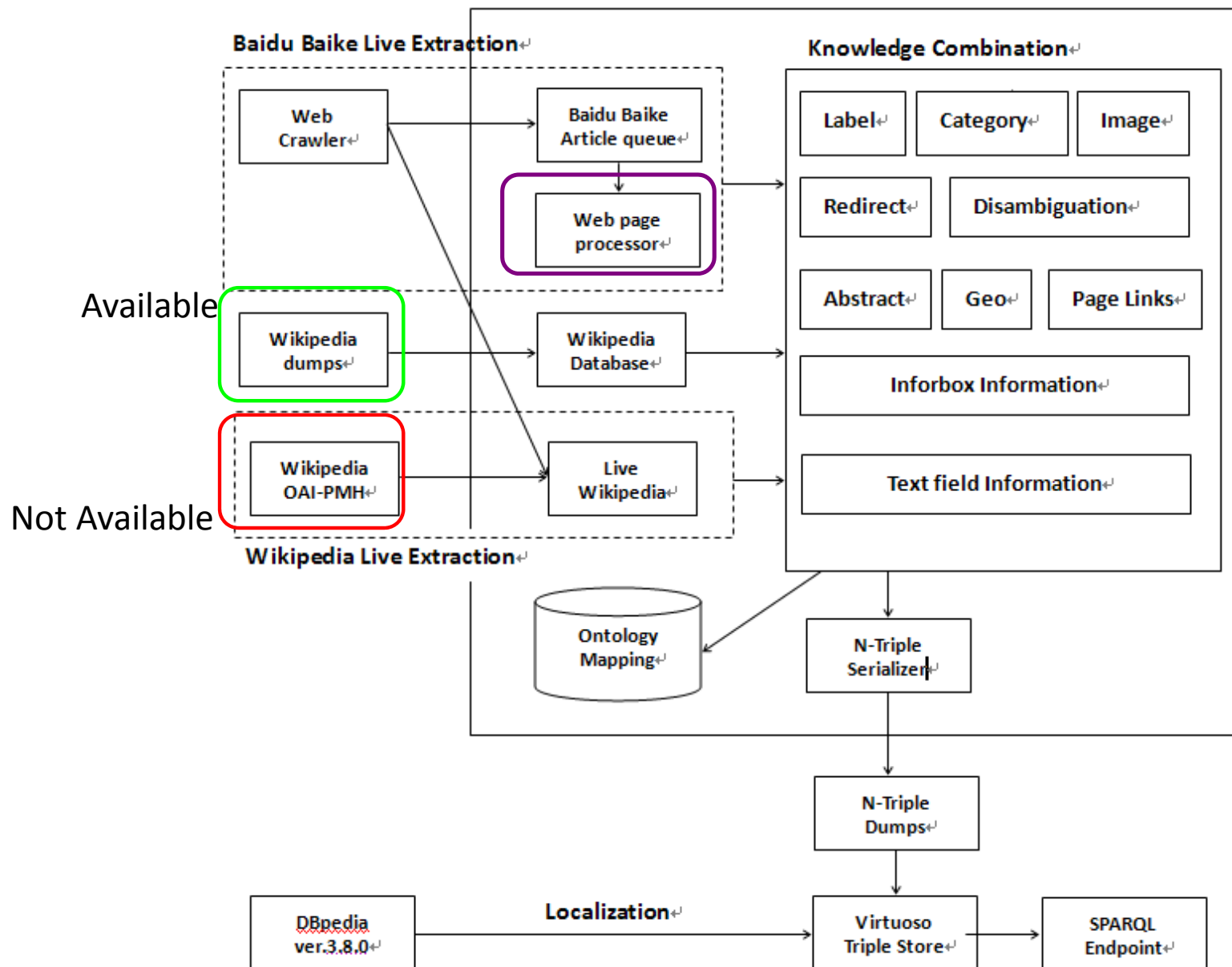


Fig. 3 Framework of knowledge base construction



# Source code comparison

```

<tr>
  <td style="white-space:nowrap;">
    <b> 国籍 </b>
  </td>
  <td class="adr" colspan="2">
    <span class="country-name">
      <span class="flagicon">
        <a title="中华人民共和国" href="/wiki/%E4%B8%
          国 </a>
      </span>
      <br>
    </td>
  </tr>

```

Triple: <http://dbpedia.org/resource/Naying> <http://dbpedia.org/property/countryName> “中国” @zh

```

  <td style="white-space:nowrap;">
    <b> 唱片公司 </b>
  </td>
  <td class="org" colspan="2">
    <a title="福茂唱片" href="/wiki/%E7%A6%81
      <small> (1994 - 1996) </small>
    <br>
    <a title="科艺百代" href="/wiki/%E7%A7%9:
      <small> (1997 - 2000) </small>
    <br>
    <a title="华纳唱片 (台湾)" href="/wiki/%E8%
      <small> (2001 - 2004) </small>
    <br>
    <a title="亚神音乐" href="/wiki/%E4%BA%9
      <small> (2011 - ) </small>
    </td>
</tr>

```

```

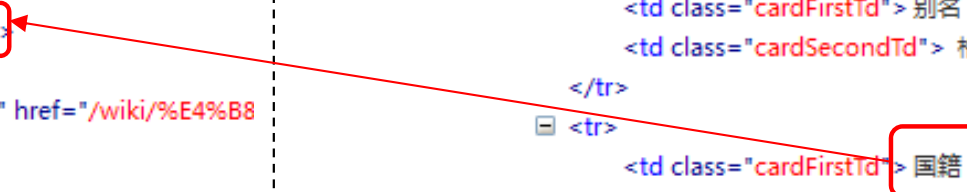
<tr>
  <td class="cardFirstTd"> 中文名 : </td>
  <td class="cardSecondTd"> 那英 </td>
</tr>
<tr>
  <td class="cardFirstTd"> 别名 : </td>
  <td class="cardSecondTd"> 格格 </td>
</tr>
<tr>
  <td class="cardFirstTd"> 国籍 : </td>
  <td class="cardSecondTd">
    <a href="/view/61891.htm" target="_blank"> 中国 </a>
  </td>
</tr>

```

```

<div class="card-summary-content">
  <p>
    那 ( Nǎ ) 英，中国著名女歌手，籍贯沈阳市。
    <a href="/view/2314.htm" target="_blank"> 满族 </a>
    人，有说为
    <a href="/view/143387.htm" target="_blank"> 叶赫那拉 </a>
    氏后人。她是华语乐坛90年代首屈一指的实力派天后，多次在央视
    <a href="/view/629988.htm" target="_blank"> 春晚 </a>
    演唱歌曲，出演过多部影视剧。2001年，凭借专辑《
    <a href="/view/866727.htm" target="_blank"> 心酸的浪漫 </a>
    》成为迄今唯一获得
    <a href="/view/1717215.htm" target="_blank"> 台湾金曲奖 </a>
    最佳国语女歌手奖的内地歌手。2011年，时隔九年，那英重归乐坛，于9月30日发行专辑《
    <a href="/view/6087894.htm" target="_blank"> 那又怎样 </a>
    》。
  </p>
</div>

```





# Extracting information from text

公司名称：	苹果公司	公司口号：	Switch ( 变革 )
外文名称：	Apple Inc.	年营业额：	1082.5亿美元 ( 2011财年 )
总部地点：	美国加利福尼亚州库比蒂诺	员工数：	60,400 ( 2011年 )
成立时间：	1976年4月1日	现任CEO：	蒂姆·库克
经营范围：	电子科技产品		

Fig. 8 Inforbox of Apple Inc. in Baidu Baike

- Many attribute value has characteristic structures
  - Publish date: 'Digits[4]'年 Digits[2]'月 Digits[2]'日
  - Annual sales volume: 'Number' 'Monetary unit' ('Digits[4]'年)
  - Number of Employers: 'Number'('Digits[4]'年)
  - Founder: A Person List

苹果公司 ( Apple Inc. ) 是[美国](#)的一家高科技公司，2007年由苹果电脑公司 ( Apple Computer, Inc. ) 更名而来，核心业务为电子科技产品，总部位于[加利福尼亚州](#)的库比蒂诺。苹果公司由[史蒂夫·乔布斯](#)、[斯蒂夫·沃兹尼亚克](#)和Ron Wayn在1976年4月1日创立，在高科技企业中以创新而闻名，知名的产品有[Apple II](#)、[Macintosh](#)电脑、[Macbook](#)笔记本电脑、[iPod](#)音乐播放器、[iTunes](#)商店、[iMac](#)一体机、[iPhone](#)手机和[iPad](#)平板电脑等。2012年8月21日，苹果成为世界市值第一的上市公司。

Fig. 9 Introduction paragraph of Apple Inc. in Baidu Baike

# Extracting information from text

- Training Data Set: Annotation

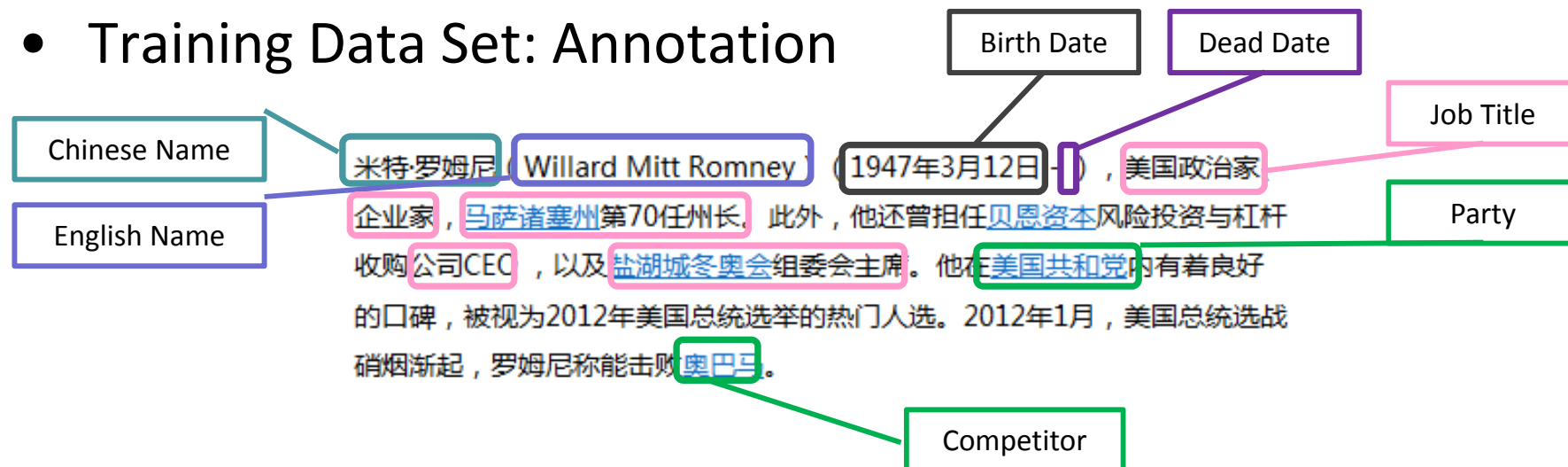


Fig. 10 Annotated paragraph

- Corresponding properties:

- <http://dbpedia.org/property/birthDate>
- <http://dbpedia.org/property/party>
- <http://dbpedia.org/property/title>
- ...

# Extracting information from text

- Several extractors are available for training
  - Conditional Random Fields (CRFs)
  - Naïve Bayesian
  - Support Vector Machine (SVM)
  - Neural Networks

# Disambiguation

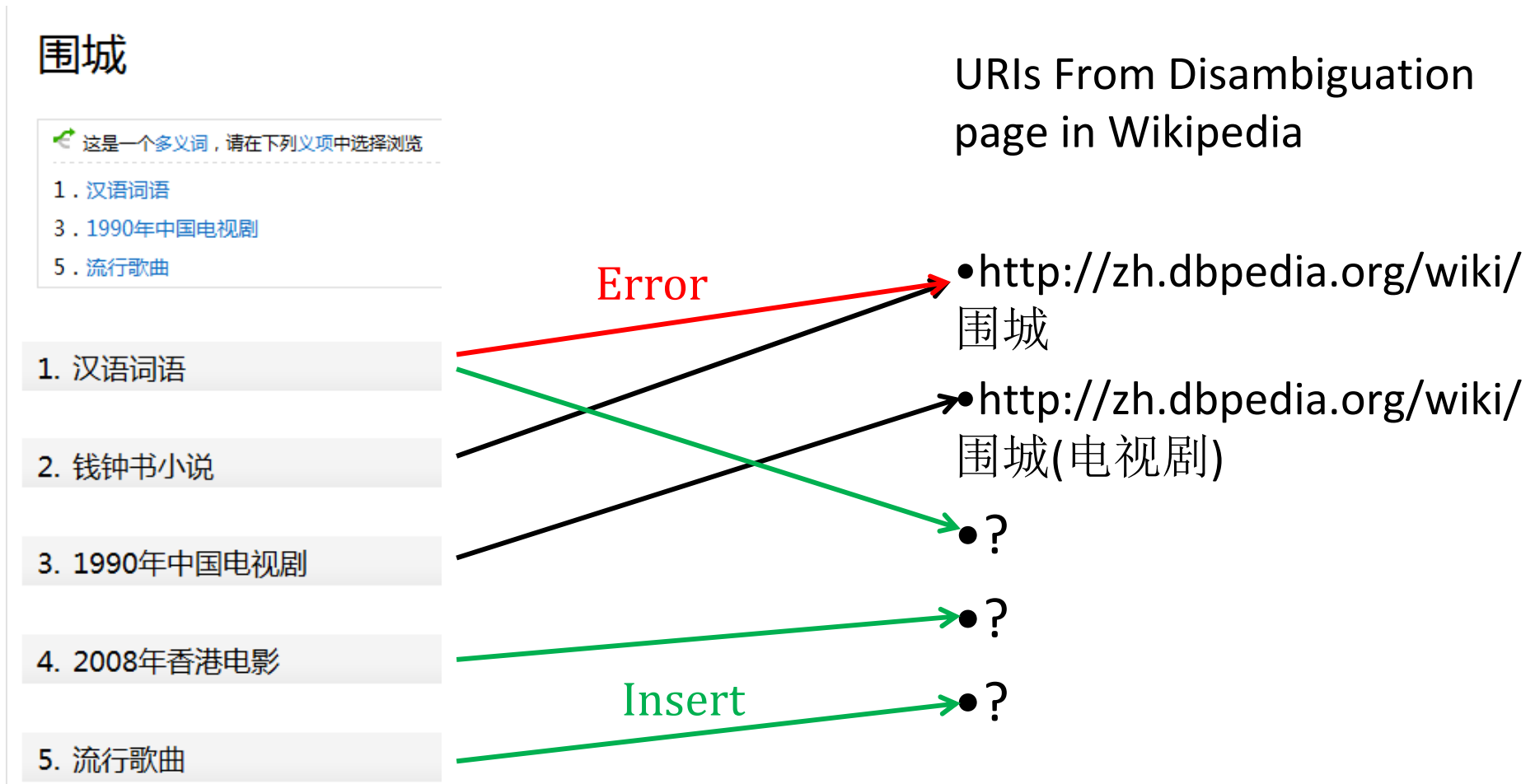


Fig. 10 Multiple results for “围城” in Baidu Baike

# Disambiguation

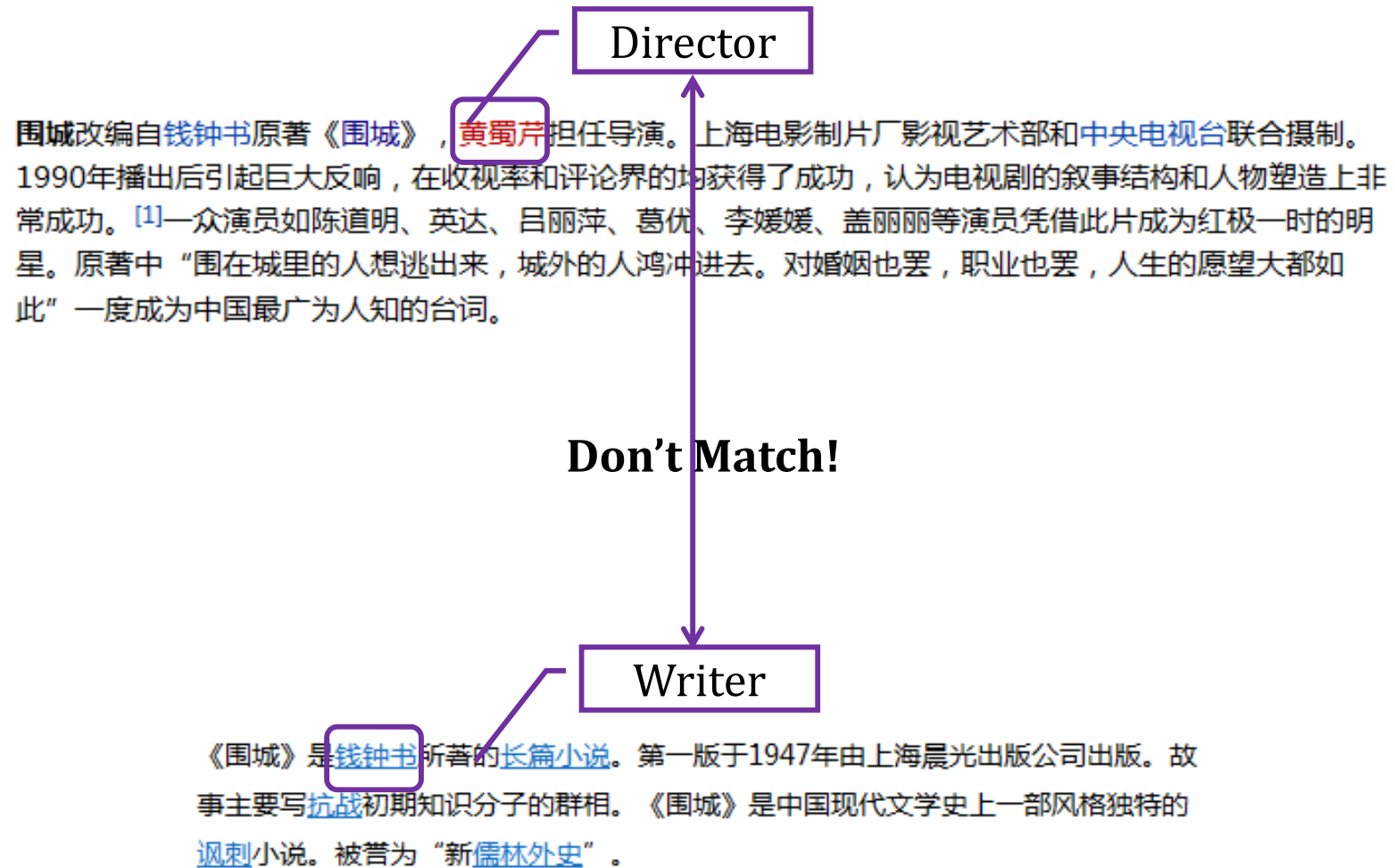


Fig. 11 Disambiguate by different attributes

# Conclusions

- This report focuses on extracting knowledge from text-based web pages.
- In fact many other work need to be done before this process
  - A web crawler that can efficiently crawl web pages both from Wikipedia and Baidu Baike
  - Tokenization, lexical analysis, gramatical analysis of the web page text;
  - Consistency checking after combination;