

对从 Web 中抽取出来的列表进行实体链接

第一步：生成候选实体集合

构建 Dictionary (<key, key.value>集合)。4 种资源用于构建 Dictionary:

- (1) Entity page: key 为 page 的 title, key.value 为该页面所描述的实体;
- (2) Redirect page: key 为 page 的 title, key.value 为该页面中 key 所指向的具体页面的 entity;
- (3) Disambiguation page: key 为 page 的 title, key.value 为该页面中所列出来的 entities;
- (4) Hyperlink in Wikipedia article: key 为超链接的锚文本, key.value 为该超链接所指向的 entity。

有了 Dictionary 就可以直接查询 Dictionary 以获得某一字符串对应的候选实体。

第二步：实体消歧

链接质量评估：两方面知识

- (1) Prior probability: 实体的流行度，通过指向该实体的 link 数来表示。计算公式如下：

$$P_{pr}(r_{i,j}) = \frac{\text{count}(r_{i,j})}{\sum_{u=1}^{|R_i|} \text{count}(r_{i,u})}$$

其中 $|R_i|$ 表示列表中第 i 项的候选实体个数。对于不同的候选实体，其流行度是不一样的。

- (2) Coherence: 候选实体间的语义一致性。表示同一列中候选实体中语义相似性 (semantically similar)。计算公式如下：

$$Coh(r_{i,j}) = \frac{1}{|L| - 1} \sum_{u=1, u \neq i}^{|L|} Sim(r_{i,j}, m_u)$$

其中 $|L|$ 表示列表中的 item 数目 (行数)。

$Sim(a,b)$ 的值由两部分组成：

$$Sim_{hr}(r_{i,j}, m_u)$$

和

$$Sim_{ds}(r_{i,j}, m_u)$$

其中 Sim_{hr} 表示 type hierarchy based similarity, Sim_{ds} 表示 distributional context similarity。

1) Sim_{hr} 的计算

计算两个类别 t_1 和 t_2 的 hierarchy based similarity 的公式为：

$$Sim_{hr}(t_1, t_2) = \frac{2 \times \log(P(t_0))}{\log(P(t_1)) + \log(P(t_2))}$$

其中 $P(t)$ 为随机选择一个实体 e ，该实体 e 属于类别 t 的概率值。 $P(t)$ 的计算在文献 “An information-theoretic definition of similarity” 中进行了介绍。

因此，两个候选实体的 type 集合的相似度计算如下所示：

$$Sim_{hr}(T(e_1) \rightarrow T(e_2)) = \frac{\sum_{t_1 \in T(e_1)} Sim_{hr}(t_1, \varepsilon(t_1))}{|T(e_1)|}$$

其中 $\epsilon(t1)$ 表示在实体 $e2$ 的 type 集合中, 使得值 $Sim_{hr}(t1, t2)$ 最大的 type $t2$ 。
最终, 两个候选实体 $e1$ 和 $e2$ 的 hierarchy based similarity 的计算公式为:

$$Sim_{hr}(e_1, e_2) = \frac{Sim_{hr}(T(e_1) \rightarrow T(e_2)) + Sim_{hr}(T(e_2) \rightarrow T(e_1))}{2}$$

2) Sim_{ds} 的计算

Sim_{ds} 表示的是两个候选实体的上下文语义一致性。

候选实体的上下文的获得: 收集维基百科页面中实体 e 的上下文。 e 的上下文为 e 出现的位置的前后若干个词 (停用词除外), 综合实体 e 的所有上下文 (实体 e 可能出现在不同的页面中, 或者同一页面出现多次), 并记录其上下文单词出现的次数。

Sim_{ds} 的计算公式为:

$$Sim_{ds}(e_1, e_2) = \frac{\sum_{i=1}^g a_i * b_i}{\sqrt{\sum_{i=1}^g a_i^2} * \sqrt{\sum_{i=1}^g b_i^2}}$$

其中, a_i, b_i 表示实体 $e1$ 和 $e2$ 的上下文中某一个词出现的次数。 g 表示两个候选实体 $e1$ 和 $e2$ 的上下文的交集的词个数。

链接质量可由如下公式表示:

$$LQ(r_{i,j}) = \alpha * P_{pr}(r_{i,j}) + \beta * \frac{1}{|L| - 1} \sum_{u=1, u \neq i}^{|L|} Sim_{hr}(r_{i,j}, m_u) + \gamma * \frac{1}{|L| - 1} \sum_{u=1, u \neq i}^{|L|} Sim_{ds}(r_{i,j}, m_u)$$

其中三个和项前面的参数为权重, 三个权重之和为 1。实体消歧的最终结果就是使得链接质量取得最大值。

权重的学习: 通过 max-margin technique 来得到。具体形式如下:

$$\forall l_i, \forall r_{i,j} \neq m_i \in R_i : LQ(m_i) - LQ(r_{i,j}) \geq \xi_{l_i}$$

Dump 文件的下载地址: <https://dumps.wikimedia.org/zhwiki/>

Index of /zhwiki/		
../		
20150301/	05-Mar-2015 15:28	-
20150325/	30-Mar-2015 01:47	-
20150417/	21-Apr-2015 09:46	-
20150515/	29-May-2015 17:18	-
20150602/	15-Jun-2015 23:40	-
20150703/	19-Jul-2015 19:05	-
20150807/	15-Aug-2015 09:32	-
20150826/	29-Aug-2015 10:10	-
20150901/	10-Sep-2015 19:30	-
20151002/	14-Oct-2015 00:42	-
20151020/	01-Nov-2015 10:43	-
20151102/	12-Nov-2015 18:16	-
20151123/	29-Nov-2015 06:23	-
latest/	29-Nov-2015 06:23	-

Index of https://dumps.wikimedia.org/zhwiki/20151123/		
zhwiki-20151123-abstract-zh-cn.xml	973.2 MB	
zhwiki-20151123-abstract-zh-tw.xml	973.2 MB	
2015-11-26 22:09:47	done	List of all page titles
zhwiki-20151123-all-titles.gz	21.6 MB	
2015-11-26 22:09:24	done	List of page titles in main namespace
zhwiki-20151123-all-titles-in-ns0.gz	8.5 MB	
2015-11-23 15:50:44	done	List of annotations (tags) for revisions and
zhwiki-20151123-change tag.sql.gz	9.8 MB	

Entity page 信息

Index of https://dumps.wikimedia.org/zhwiki/20151123/		
zhwiki-20151123-iwlinks.sql.gz	24.0 MB	
2015-11-23 15:50:10	done	Redirect list
zhwiki-20151123-redirect.sql.gz	7.9 MB	
2015-11-23 15:50:06	done	Nonexistent pages that have been protected.
zhwiki-20151123-protected titles.sql.gz	48 KB	
2015-11-23 15:50:03	done	Name/value pairs for pages.
zhwiki-20151123-page props.sql.gz	17.1 MB	

Redirect page 信息

← → ↻ <https://dumps.wikimedia.org/zhwiki/20151123/>

应用 谷歌搜索引擎_香港... 在线文档-jdk_7u4 Scholar--Google f... Glgoo 学术搜索 正则基础之——贡...

Last dumped on 2015-11-02

Dump complete

Verify downloaded files against the [\(md5\)](#), [\(sha1\)](#) checksums to check for corrupted files.

2015-11-24 16:26:30	done	Articles, templates, media/file descriptions, and primary meta-
		zhwiki-20151123-pages-articles-multistream.xml.bz2 1.2 GB
		zhwiki-20151123-pages-articles-multistream-index.txt.bz2 19.8 MB
2015-11-26 22:08:54	skipped	All pages with complete edit history (.7z)
2015-11-26 22:08:54	skipped	All pages with complete page edit history (.bz2)

Hyperlink 信息以及实体上下文信息

2015-11-23 15:34:11	done	Wiki category membership link records.
		zhwiki-20151123-categorylinks.sql.gz 144.7 MB
2015-11-23 15:31:19	done	Wiki page-to-page link records.
		zhwiki-20151123-pagelinks.sql.gz 524.5 MB
2015-11-23 14:57:36	done	Metadata on current versions of uploaded media/files.
		zhwiki-20151123-image.sql.gz 6.0 MB
2015-11-23 14:57:33	done	A few statistics such as the page count.
		zhwiki-20151123-site_stats.sql.gz 812 bytes

Hyperlink 信息？

2015-11-23 15:49:46	done	Base per-page data (id, title, old restrictions, etc).
		zhwiki-20151123-page.sql.gz 135.2 MB
2015-11-23 15:49:07	done	Category information.
		zhwiki-20151123-category.sql.gz 5.4 MB
2015-11-23 15:49:04	done	User group assignments.
		zhwiki-20151123-user_groups.sql.gz 5 KB
2015-11-23 15:49:01	done	Wiki interlanguage link records.
		zhwiki-20151123-langlinks.sql.gz 112.0 MB
2015-11-23 15:47:30	done	Wiki external URL link records.
		zhwiki-20151123-externallinks.sql.gz 145.7 MB
2015-11-23 15:46:25	done	Wiki template inclusion link records.
		zhwiki-20151123-templatelinks.sql.gz 199.4 MB
2015-11-23 15:35:57	done	Wiki media/files usage records.
		zhwiki-20151123-imagelinks.sql.gz 40.7 MB

实体的 category (type) 信息

<http://licstar.net/archives/262>

entityandname.txt	
569358	[[黑沙环 黑沙环]]
569359	[[工业街 工业街]]
569360	[[中国大陆 中国大陆]]
569361	[[北京 北京]]
569362	[[百货公司 百货公司]]
569363	[[维兹亚州立大学 维兹亚州立大学]]
569364	[[中华人民共和国教育部 中华人民共和国教育部]]
569365	[[中华人民共和国教育部 教育部]]
569366	[[北京 北京]]
569367	[[清华大学 清华大学]]
569368	[[上海 上海]]
569369	[[上海交通大学 上海交大]]
569370	[[湖北 湖北]]
569371	[[武汉大学 武汉大学]]

entityNameDistribution.txt	
707606	2008青宁心曲:2008青宁心曲 (1)
707607	三岛站:三岛 (2)
707608	马格南:马格南 (2)
707609	李昭皇:李昭皇 (李佛金) (1) 昭圣皇后 (1) 昭圣公主 (1) 李昭皇 (23) 李佛金 (2)
707610	马桥乡:马桥乡 (4)
707611	乐华邨:乐华北邨 (1) 乐华南邨 (1) 乐华 (2) 乐华邨 (36)
707612	蘆着:蘆着 (2)
707613	横滨新都市交通:横滨新都市交通 (1)
707614	平潭镇:平潭镇 (3)
707615	乌尔班六世:乌尔班六世 (6)
707616	福特冰原岛峰:福特冰原岛峰 (1)
707617	北海舰队:北海舰队 (77)
707618	李景源:李景源 (3)
707619	加姆萨尔:加姆萨尔 (1)

wikidump0after.txt	
354802	江苏 海峡、之后在中国广东及福建沿岸登陆; 亦可能经过
354803	台风菲特 2013年 江、江苏一带沿海登陆。 这种路径的台风
354804	关岛 热带气旋在菲律宾以东或附近形成后, 先向西北偏西
354805	朝鲜半岛 带气旋在较北、较西的地方转向, 途径可以影响
354806	日本, 途径可以影响台湾、东海或中国大陆东部沿岸浙江、
354807	台风范斯高 2013年 岸浙江、江苏等地, 转向后可以影响朝鲜
354808	藤原效应 部份热带气旋因外围引导气流不明, 或受其他天
354809	季候风 部份热带气旋因外围引导气流不明, 或受其他天气
354810	台风天秤 2012年 影响如藤原效应或季候风, 路径出现打转、
354811	台风罗莎 2013年 路径出现打转、停滞等, 例如2012年的台
354812	热带气旋 台风预警是盛行地区, 于风暴可能侵袭期间, 由各
354813	热带 热带气旋是发生在、亚热带地区海面上的气旋性环流
354814	亚热带 热带气旋是发生在热带、亚热带地区海面上的气旋性环流
354815	气旋 热带气旋是发生在热带、亚热带地区海面上的气旋性环流