# 1：显式关系识别

## 1.1 识别结果如下（仅仅使用显示连词词典识别）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **关系编号** | **标注** | **识别** | **识别正确** | **P** | **R** | **F** |
| **1时序** | 674 | 602 | 563 | 0.94 | 0.84 | 0.88 |
| **2因果** | 2097 | 2046 | 1998 | 0.97 | 0.95 | 0.96 |
| **3比较** | 1948 | 1829 | 1731 | 0.95 | 0.89 | 0.92 |
| **4扩展** | 3180 | 3025 | 2826 | 0.93 | 0.89 | 0.91 |

## 1.2 旧版本体系下的识别结果为

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **关系编号** | **标注** | **识别** | **识别正确** | **P** | **R** | **F** |
| **1时序** | 502 | 525 | 490 | 0.93 | 0.98 | 0.95 |
| **2因果** | 1571 | 1614 | 1554 | 0.96 | 0.99 | 0.98 |
| **3条件** | 685 | 682 | 664 | 0.97 | 0.97 | 0.97 |
| **4比较** | 2260 | 2112 | 2045 | 0.97 | 0.90 | 0.94 |
| **5扩展** | 2208 | 2571 | 1996 | 0.78 | 0.90 | 0.84 |
| **6并列** | 765 | 509 | 397 | 0.78 | 0.52 | 0.62 |

## 1.3 去年师姐的数据

其中： “3条件” 合并到了”2因果”；”5 扩展” 和 “6并列” 合并到了 “4扩展”

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 关系 | 1时序 | 2因果 | 3条件 | 4比较 | 5扩展 | 6并列 |
| P | 95.12% | 96.86% | 98.9% | 99.23% | 86.83% | 63.64% |
| R | 71.56% | 88.76% | 90.45% | 84.31% | 82.24% | 50.6% |
| F | 81.67% | 92.63% | 94.49% | 91.16% | 84.47% | 56.38% |

# 2：连词识别

## 2.1 常见词性的出现次数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **词性** | **个数** | **词性** | **个数** | **词性** | **个数** |
| a | 6660 | m | 5551 | p | 6413 |
| b | 1088 | n | **34472** | q | 3084 |
| c | 6494 | nd | 2706 | r | 6869 |
| d | **11808** | nh | 1700 | u | **13313** |
| e | 1 | ni | 113 | v | **39039** |
| g | 0 | nl | 668 | wp | **13835** |
| h | 6 | ns | 3967 | ws | 466 |
| i | 1107 | nt | 1929 | x | 0 |
| j | 917 | nz | 430 |  |  |
| k | 115 | o | 4 |  |  |

## 2.2 连词识别数量(所有词均考虑的情况下)

|  |  |  |  |
| --- | --- | --- | --- |
| **过滤条件** | **正样例** | **负样例** | **总数** |
| **不过滤** | 3772 (2.3%) | 159146 (97.7%) | 162918 |
| **使用词性过滤** | 3710 (2.7%) | 131590 (97.3%) | 135300 |
| **停用词+词性过滤** | 3284 (3.4%) | 92200 (96.6%) | 95484 |

## 2.3 连词歧义性分析

所有出现的连词个数为：7848

|  |  |  |
| --- | --- | --- |
| 连词指向关系种类 | 次数 | 比例 |
| 指向1个关系 | 5225 | 0.665775 |
| 指向2个关系 | 1997 | 0.25446 |
| 指向3个关系 | 594 | 0.075688 |
| 指向4个关系 | 32 | 0.004077 |
| 指向5个关系 | 0 | 0 |
| 指向6个关系 | 0 | 0 |

# 3：隐式关系识别

## 3.1 数据分布

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 关系  数量 | 0木有 | 1时序 | 2因果 | 3比较 | 4扩展 |
| 训练+测试 | 16+0 | 141+22 | 826+275 | 328+69 | 3849+927 |
| 比例 |  |  |  |  |  |

## 3.2 BaseLine

Svm训练参数："-c", "2.0", "-g", "0.125","-w1","15","-w2","4","-w3","10","-w4","1"

Accuracy = 65.89% (852/1293) P: 0.66、R: 1.0 、F: 0.79 宏平均和微平均如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **关系类型** | **P** | **R** | **F** |  |
| **1** | 0.045 | 0.017 | 0.025 |  |
| **2** | 0.327 | 0.213 | 0.258 |  |
| **3** | 0.087 | 0.053 | 0.066 |  |
| **4** | 0.814 | 0.717 | 0.763 |  |

## 3.3 基于序列标注的数据统计

5821个句间关系(包括相邻和不相邻)，其中相邻的句间关系共：3608(62.0%)对。

### 3.3.1 显式关系和隐式关系的转移矩阵

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **显隐式** | Exp-Next | Imp-Next | 总和 | **显隐式** | Exp-current | Imp-Current |
| Exp | 495 | 617 | 1112 | Exp-Next | 44.51% | 24.76% |
| Imp | 666 | 1830 | 2496 | Imp-Next | 55.73% | 73.32% |

### 3.3.2各个关系之间的转移矩阵(包括了显式和隐式)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **关系** | **1** | **2** | **3** | **4** | 总和 |
| **1时序** | 53 | 14 | 28 | 96 | 191 |
| **2因果** | 9 | 69 | 68 | 271 | 417 |
| **3比较** | 22 | 79 | 73 | 301 | 475 |
| **4扩展** | 112 | 275 | 309 | 1829 | 2525 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **关系** | 1 | 2 | 3 | 4 |
| **1时序** | 27.75% | 7.33% | 14.66% | 50.26% |
| **2因果** | 2.16% | 16.55% | 16.31% | 64.99% |
| **3比较** | 4.63% | 16.63% | 15.37% | 63.37% |
| **4扩展** | 4.44% | 10.89% | 12.24% | 72.44% |

**示例：**

万客隆开启了台湾自助、开放式的仓储批发购物模式，法商家乐福也在一九八九年进入台湾。一九九五年，美国好市多到台湾筹备开店；一九九九年法国零售品牌吉安进入量贩市场。九八年为了扩大市场占有率，属于润泰集团的大润发购并了大买家、亚太等量贩店，使得店数达到十六家。

3.4 旧版本6类关系的分布

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 关系数量 | 1 | 2 | 3 | 4 | 5 | 6 |
| 训练+测试 | 142 + 37 | 936 + 246 | 30 + 10 | 379 + 103 | 3184 + 779 | 1068 + 265 |
| 比例 | 2.47% | 16.31% | 0.52% | 6.60% | 55.48% | 18.61% |
| 总和(7169) | 179 | 1182 | 40 | 472 | 3963 | 1333 |
|  |  |  |  |  |  |  |

## 3.4 基于Word Vector的隐式句间关系识别

使用停用词过滤掉其中的无意义词，使用的普通加和

"-c", "2.0", "-g", "0.001953125"

Accuracy = 67.19085060757683% (940/1399) (classification)

SenseID: 1 P: NaN R: NaN F: NaN

SenseID: 2 P: 0.432 R: 0.147 F: 0.219

SenseID: 3 P: NaN R: NaN F: NaN

SenseID: 4 P: 1.000 R: 0.011 F: 0.022

SenseID: 5 P: 0.674 R: 0.918 F: 0.777

SenseID: 6 P: 0.747 R: 0.623 F: 0.679

使用加权平均，对不同的语义角色的词设置不同的权值

if( pos.startsWith("a") ) weight = 2; //形容词

if( pos.startsWith("u") ) weight = 2; //语气词

if( pos.startsWith("n") ) weight = 3; //名词

if( pos.startsWith("v") ) weight = 4; //动词

"-c", "8.0", "-g", "0.0078125","-w2","3","-w4","9","-w5","1","w6","3",

# 4：EDU切分情况

All EDU:438

采用完全相同的(词数少于3个的)：arg1Correct:203 arg2Correct:202 argAllCorrect:186

相同部分超过7成的：

All EDU:432

arg1Correct:38 arg2Correct:34 argAllCorrect:13

Same Percentnge 70 percent arg1CorrectIn:31 arg2Correct:27

Same Percentnge 80 percent arg1CorrectIn:27 arg2Correct:27

Same Percentnge 90 percent arg1CorrectIn:23 arg2Correct:26

正确率为：54.6%

# 1：连词识别错误的情况：

(1) 标注错误

“突然提出离婚就是因为人口普查有机可趁 ，想趁机为非婚生子女报户口 ，”

标注结果为：289 297 298 356 1 0 300 304 1 2-1-2结果在先

标注连词：是因为

识别结果：因为

-------识别出了连词但是与标注的不一致--------

支持 电影 降低 票价 理由 是 脱离 群众 实际 生活 水平 的 高 价位 违背 了 电影 行业 服务 于 人民 大众 的 宗旨

427 438 441 495 1 0 441 445 1 2-2-2推论在先 标注连词：理由是

(2) 连词定义错误

“大家 的 设想 以及 大家 的 关切”

标注结果：257 264 265 275 1 0 265 268 1 4-5平行

标注连词：以及

识别结果：无连词

“复杂的雷达系统以及反对抗装置 “

标注结果：199 210 210 220 1 0 210 213 1 4-5平行

标注连词：以及

识别连词：无连词

## 2：连词识别和标注的连词不一致

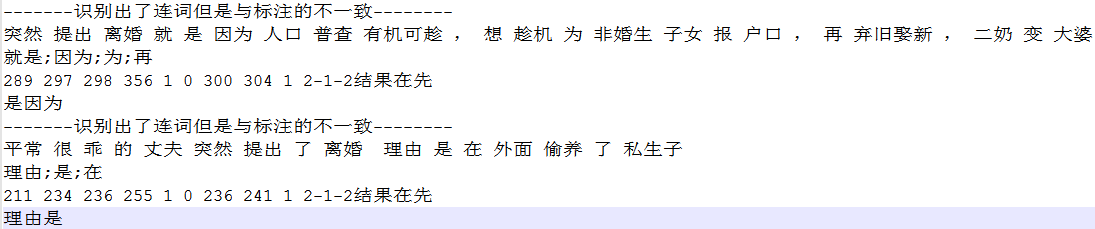
-------识别出了连词但是与标注的不一致--------

他 将 邀请 南国 总统 米洛舍维奇 和 南国 反对党 总统 候选人 科什图尼察 到 莫斯科 举行 谈判

将

183 201 202 236 1 0 202 203 1 4-5平行 和

原始标注就错误的连词有：“和”，部分“以及”



-------识别出了连词但是与标注的不一致--------

突然 提出 离婚 就 是 因为 人口 普查 有机可趁 ， 想 趁机 为 非婚生 子女 报 户口 ， 再 弃旧娶新 ， 二奶 变 大婆

就是;因为;为;再

289 297 298 356 1 0 300 304 1 2-1-2结果在先

是因为

-------识别出了连词但是与标注的不一致--------

支持 电影 降低 票价 理由 是 脱离 群众 实际 生活 水平 的 高 价位 违背 了 电影 行业 服务 于 人民 大众 的 宗旨

理由;是

427 438 441 495 1 0 441 445 1 2-2-2推论在先

理由是

这里面的：”理由是”，”是因为”的是表示了一种方向。但是在分词之后却分开了，而在词表中此类词是出现的。建议手工处理下词表

-------识别出了连词但是与标注的不一致--------

工人 或者是 计程车 司机

或者;是

217 220 220 230 1 0 220 223 1 4-6-1相容选择

或者是

-------识别出了连词但是与标注的不一致--------

佛州 选举 的 胜利者 将 变成 为 戈尔 而 不 是 佛州 州务卿 哈里斯 先前 所 宣布 确认 的 小 布什

将;为;而;不;是;所;宣布

323 345 346 381 1 0 347 352 1 4-6-2互斥选择

而不是

-------识别出了连词但是与标注的不一致--------

梁 律师 代表会 副主席 张家荣 昨天 傍晚 在 路上 骑 车 运动 的 时候 遭到 了 两 名 歹徒 持 刀 砍伤 所 幸 没有 生命 危险

昨天;在;时候;所

0 59 61 72 1 0 61 63 1 3-2间接对比

所幸

-------识别出了连词但是与标注的不一致--------

所幸 火苗 并没有 波及 一旁 的 加油机 和 管线 否则 后果 不堪设想 。

所;并;否则

421 447 449 462 1 0 421 423 1 2-1-1原因在先

所幸

-------识别出了连词但是与标注的不一致--------

即使 穿上 绍族 传统 服式 ， 耳边 听 着 绍族 传统 乐曲 他 的 表情 还 是 一样 的 紧绷 ， 没有 一 丝 笑容

即使;还;是

333 365 368 398 3 1 333 335 333 336 375 378 378 1 3-3-1让步在先

即使;还是

-------识别出了连词但是与标注的不一致--------

不管是 华纳 还是 番刀

不管;是;还;是

395 401 402 408 2 0 395 397 402 404 1 4-5平行

不管;还是

-------识别出了连词但是与标注的不一致--------

只要 是 成年 的 勇士 至少 还 必须 拥有 三 把 番刀

只要;是;至少;还;必须

521 533 534 552 2 0 521 524 537 542 1 2-4-2-1充分条件在先

只要;还必须

仍然是因为分词的原因，导致了标注的词被分开，这里的”而不是”应该被手工修改下

-------识别出了连词但是与标注的不一致--------

昨天 晚上 １０点半 还 和 同房 的 病患 看 完 电视 休息 结果 不久 后 管理 人员 巡视 病房

昨天;还;和;结果;后

162 194 196 216 1 0 200 205 1 1-2-1先序

不久后

-------识别出了连词但是与标注的不一致--------

许朝清 毕业于 交通 大学 研究所 ， 现年 ２８ 岁 受到 轻伤 的 王贯仁 则 是 毕业于 辅仁 大学

则;是

399 426 429 454 1 0 441 444 1 4-5平行

则是

则是的原因和上面的一致

-------识别出了连词但是与标注的不一致--------

这 次 在 被 梵蒂冈 册封 的 一些 人 更 是 在 中国 土地 上 奸淫 抢掠 、 为非作歹 ， 对 中国 人民 犯下 了 不 可 饶恕 的 罪行

在;更;是;在

256 277 279 333 1 0 279 284 1 4-2递进

更是

-------识别出了连词但是与标注的不一致--------

这 是 中国 天主教徒 、 基督教徒 坚决 反对 的 也 是 中国 人民 坚决 反对 的 。

这;是;也;是

709 736 738 757 1 0 738 741 1 4-2递进

也是

-------识别出了连词但是与标注的不一致--------

如果 不 采取 有效 的 防范 措施 非洲 是 未来 ２０ 年 艾滋病 肆虐 最 严重 的 地区

如果;不;是

88 106 109 138 1 0 88 93 1 2-4-2-1充分条件在先

如果不

这里的标注明显是错误的

-------识别出了连词但是与标注的不一致--------

事发 后 有 １８ 名 乘客 及时 逃脱 存活 以至 有 １５５ 人 不幸 罹难

后;有;有

205 229 231 248 1 0 231 233 1 2-1-1原因在先

以至

词表中”以至”只标注了一次，因此被过滤掉了。

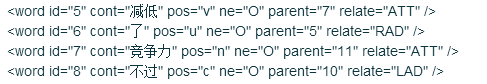
-------识别出了连词但是与标注的不一致--------

虽然 石油 价格 大幅 上扬 减低 了 竞争力 不过 亚洲 地区 证券 和 货币 市场 的 波动 并没有 影响 到 这些 预测值

虽然;并;这些

290 313 316 357 2 0 289 293 315 319 1 3-2间接对比

虽然;不过



这里的不过被过滤掉了

-------识别出了连词但是与标注的不一致--------

只 会 搞 政治 运动 迫害 作家 迫使 作家 流亡 海外

只

860 878 880 891 1 0 879 882 1 2-1-1原因在先

迫使

迫使只在连词词典中出现了一次，因此被过滤掉了

-------识别出了连词但是与标注的不一致--------

高行健 向 瑞典 国王 以及 在场 １８００多 位 贵宾 说 他 １２ 年 前 流亡 法国 写 的 更 勤快

说;前;更

327 357 360 383 1 0 362 368 1 1-2-1先序

１２年前

１２年前，时间问题一般都会因为分词的错误导致了一段词语被分割。这个需要添加模板识别

-------识别出了连词但是与标注的不一致--------

谭盾 以 丰富 的 中国 民族 艺术 为 基础 ， 融合 了 西方 音乐 的 形式 向 世界 音乐 的 美学 出发 就 如同 香港 这个 融合 了 中 西 的 国际 城

以;为;就;这个

341 399 400 427 1 0 401 406 1 3-1-1正向对比

就如同

这里的“就如同“在词表中只出现了一次，而”如同“在词表中是没有的

-------识别出了连词但是与标注的不一致--------

过去 １００多 年 来 ， 没有 人 能够 做 成 的 ， 完成 的 他 逐渐 地 把 它 复原 出来 了

过去;来

527 561 585 604 1 0 527 536 1 1-2-1先序

过去１００多年

语义单元范围的识别有点麻烦

-------识别出了连词但是与标注的不一致--------

不止 是 古今 研究 中 所 难以 突破 的 西夏语 造字 原则 还 包括 西夏 文字 所 表现 的 音韵

是;所;还;包括;所

715 748 750 770 2 0 715 717 750 751 1 4-2递进

不止;还

属于并列连词的范围，不止…还，侧重于后面的语义。这点还要仔细考虑

-------识别出了连词但是与标注的不一致--------

或者 在 制造 上 它 的 内在 质量 不 是 很 高 ， 再有 就 是 维修 保养 、 安全 管理 这 块 呢 ， 存在 着 一些 问题

或者;在;不;是;再;有;就;是

314 343 344 384 1 0 343 346 1 4-6-1相容选择

再有

再有在分词是导致了分词错误

-------识别出了连词但是与标注的不一致--------

既 有 娱乐性 又 安全 可靠 的 游乐 设施

有;又

591 598 601 617 2 0 591 594 600 602 1 4-5平行

既有;又

-------识别出了连词但是与标注的不一致--------

济南 交警 推出 了 １＋４ 的 新 勤务 管理 模式 也 就 是 路口 安排 一 名 值勤 交警 指挥

也;就是

448 476 478 502 1 0 478 483 1 4-3-1解释说明

也就是

-------识别出了连词但是与标注的不一致--------

但 无论 是 什么 称谓 都 没 能 使 久居 三江 大地 的 康巴人 过上 真正 的 好 日子 ， ， 尤其是 １９１８年 ， 西藏 地方 政府 开始 在 昌都 进行 长 达 ３２ 年 之 久 的 黑暗 统治 ， 广大 农牧民 和 奴隶 在 三 大 领主 的 层层 盘剥 下 ， 缺吃少穿 ， 衣 不 避体 ， 过 着 牛马 不如 的 生活 。

但;都;使;尤其;是;在;在;不

228 278 277 399 1 0 279 282 1 4-2递进

尤其是

-------识别出了连词但是与标注的不一致--------

江西 ７万 钢铁 大军 并没有 因此 沾沾自喜 而是 励精图治

并;因此;而是

1239 1262 1265 1273 1 0 1265 1266 1 3-2间接对比

而

只是识别出了一个词

-------识别出了连词但是与标注的不一致--------

不 是 积极 投入 行动 而 是 先 质问 国家 药监局 官员

不;是;而是;先

492 504 506 525 2 0 492 495 507 510 1 4-6-2互斥选择

不是;而是

-------识别出了连词但是与标注的不一致--------

我 只有 更 好 的 工作 才 能 答谢 领导 对 我 的 关心

只有;更;才

685 699 701 719 2 0 687 689 701 704 1 2-4-1-1必要条件在先

只有;才能

-------识别出了连词但是与标注的不一致--------

国务院 提出 到 ２０００年 底 实现 一 控 双 达标 也 就 是 全 国 实现 主要 污染物 总量 控制 、 重点 城市 按 功能区 达标 、 工业 污染源 达标 排放

也;就;是;主要

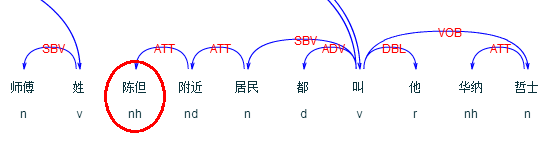
8 37 39 96 1 0 38 42 1 4-3-1解释说明

也就

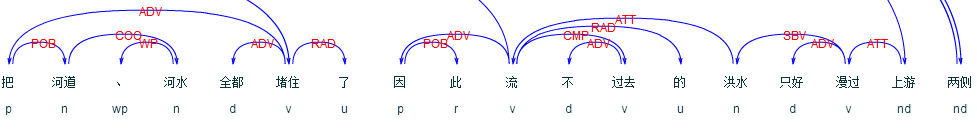
--------------------------------------------------------------------

## 1.2 LTP错误

就 是 从 日军 时代 开始 已经 三 代 单传 的 正岛 番刀 师傅 ， 师傅 姓 陈 但 附近 居民 都 叫 他 华纳 哲士



在 大雨 当时 根本 犹如 形成 一 座 漫帅 水坝 ， 把 河道 、 河水 全 都 堵住 了 因此 流 不 过去 的 洪水 只好 漫过 上游 两 侧



虽然 影响 了 交通 不过 看到 这 群 可爱 的 游行 队伍 来往 的 人 车 也 都 相当 的 体谅



**一个句子中含有多个连词**

(1) 如果 这里 的 法庭 能够 给 她 应 有 的 惩处 ， 那么 我 认为 他 应该 在 这里 接受 审判  但是 ， 我 拿 不 准 海牙 国际 法庭 会 怎么 做 ， 我 对 那个 法庭 不 怎么 了解 ， 我 并不 真正 的 信任 它 。

(2) 如果 从 纯 艺术 的 角度 来说 ， 他们 自然 每 天 都 给 过往 行人 带来 美感,但是 利文思通 不 这样 看 ， 他 看 得 更 深 一 层。

(3) 如果 我 不 走, 那 除非 我 能 陪 千千 一 辈子 。

(4)如果 你 用 每 股 一百五十 元 买 了 台积电   却 巴望 它 在 两 个 月 内 涨 到 二百二十 元

(5) 如果 台湾 劳工 也 只是 生产 帽子 ， 的确 不 划算  但 如果 台湾 劳工 可以 因为 素质 高 而 提出 好 点子 、 改善 制程 ， 那 就 值得 了

(6) 目前 的 英文 教学 用 游戏 、 活动 进行 固然 很 活泼   但 如果 只是 蹦蹦跳跳 、 大声 练习 发音 ， 小 朋友 会 产生 英文 好像 「 很 聒噪 」 的 错觉

(7) 而 如果 三 口 之 家 只 有 一 人 就业  那么 这个 家庭 的 生活 就 非常 困难 了

(8) 尽管 克林顿 即将 在 明年 元月 卸任  但是 他 的 到访 仍然 是 一 件 大 事 ， 将 有助于 双方 人员 的 接触

(9) 但是 由于 这些 声音 根本 听 不 清楚  虽然 他 象征 意义 远 大于 实质 意义 。

(10) 虽然 这边 大家 都 很 照顾  但是 她 当然 希望 自己 亲人 在 身边

(11) 由于 车速 相当 快 司机 虽然 接获 无线电 通知 ， 但是 根本 来 不 及 刹车