实体、链接、共指标注说明书

# 标注工具

使用inception进行实体标注，依照《inception用户手册》（https://inception-project.github.io/releases/0.18.0/docs/user-guide.html）使用。

# 数据源

中时新闻网新闻数据，[www.chinatimes.com](http://www.chinatimes.com)。

文件路径：apple：/home/disk2/nuclear/news\_data/PCube/CT

# 人工标注规范

## 实体识别标注规范

标注的实体包括 人物（PER）、组织机构（ORG）、地点（LOC）、国家/政权（REG）、其它（OTH）

### 人物实体

人物分人名和职位，当一个词语是人名时，标注为人物，当一个职位做主语且能够指向唯一人物实体时标注为实体。若职位与任职人同时出现，职位不能指向唯一人物时，不标注职位为人名。

### 组织机构实体

组织机构主要指政府部门、公司、研究所、学校等，经济报道中常出现的 道琼斯指数、标普指数、香港恒生指数等股指，严格意义上不算组织机构，但由于这些实体比较重要，故也按照组织进行标注。

### 地点实体

地点一般指地名，需要是一个有着明确地理范围的地方，如洲、国、省、州、市县、区、村、街道、商场等地，一般作为句子的定语、状语部分出现。特别的，如果一个国家以句子的主语、宾语出现，则将国家视为特殊的 国家/政权实体，而非地点。

### 国家/政权实体

国家/政权实体指政治上被广泛承认的国家或者存在但没有被承认的政权，当一个国家、地区在句子中做主语宾语时，该国家、地区则应该被视为 国家/政权 实体，而非简单的地点实体

### 其它实体

其它实体为为无法被归类为上述实体的其它重要实体，当这类实体与人物、组织机构、地点、国家/政权实体产生关系时考虑进行标注。但该类实体数目较少。

### 特殊情况

重叠实体出现时，应该以最长匹配为准。若最长实体无法进行实体链接，可以将其中的定语部分作为重叠实体标记并链接，以表明无法消歧的实体的一定所属关系。

## 实体链接标注规范

实体链接以不完整的wikidata为知识库。凡是能够在inception提供的wikidata知识库中搜索到的实体均被链接到wikidata中，并以wikidata的实体编号为唯一编号，编号是一串以Q开头且跟有1位以上数字的字符串组成。但需要注意的是，有部分实体在wikidata网站本身可以检索到，但由于未知原因无法在inception的消歧中被检索，在标注中遇到的这类实体均被收录在：附录1.无法消歧的实体列表中。

在标注实体链接时，遵从同一个文件内实体的表述相同且实体类型相同的必定指的是同一个实体的假设，故仅标注每一个一个实体表述第一次出现的位置，再次出现实体的表述相同且实体类型相同时一般不进行标注。如果同样的实体表述在同一篇文档中不是第一次出现，且标注了与第一次出现时不同的链接，则说明此次实体表述所指代的实体与之前的不同，但之后若要再次出现没有标注的相同实体表述，仍以第一次出现时的标注为准。例如：【美】（标注LOC Q30）股……欧【美】（标注LOC Q49）……【美】企（标注LOC）则欧【美】应该链接到Q49， 美企连接到Q30。如果实体的表述相同但实体的类型不同，不应该当做指代同一个实体。例如：中【美】（标注REG Q30）贸易战……欧【美】（标注LOC Q30）市场，二者不能被当作指代同一个实体。

## 实体共指标注规范

实体共指以实体间的特殊关系（Corefernce）为标记，共指消解同样遵从相同文章内实体的表述相同且实体类型相同的必定指的是同一个实体的假设。

若共指的实体为可以链接的实体，则在标注时仅连接不同表述第一次出现的时候（即有链接标注的位置）。如果同一个实体有三种及以上的表述，则Corefernce的标记是从第一个出现连接到第二个，之后再从第二个连接到第三个，以此类推。

若共指的实体为无法链接的实体则连接遵从就近原则，连接与第二个表述最近的上一个表述位置即可。

# 导出格式

导出格式采用Inception自创的WebAnno TSV 3.3 File format格式，输出的文件类型为.tsv格式，编码类型为UTF-8。具体的格式以<https://inception-project.github.io/releases/0.18.0/docs/user-guide.html#sect_webannotsv>的说明为准，这里仅说明与实体、链接、共指相关的部分。

对于一篇文章，导出文件以句子为单位进行划分。开头的5行为无效信息（inception导入信息），从第6行开始为第一句话的原句，原句行以#Text开头，分词与分词之间有空格，紧随原句行的是各个词语的标注行。最后一个词语行与下一个句子的原句行之间有一个空行。

对于词语行，每行用7列描述一个词语，每列的含义分别为：词语编号、文中位置、词语本身、实体链接页面、实体类型、实体与另一实体的封闭式关系类型、另一实体的编号。每列的描述具体如下：

## 词语编号

词语编号的格式形如“\d+-\d+(.\d+)\*”，即由一串数字，短线、第二串数字组成，可能还在最后有“.”和第三串数字。第一串数字的含义为句子序号，从1开始；第二串数字的含义为句子内词语序号，从1开始；当出现.和第三串数字时，表示上一行词语内部有实体，该行即为上一行词语的子字符串，第三串数字为实体编号，该行的后续列与正常行完全一致。

## 文中位置

文中位置的格式形如“\d+-\d+”，即由一串数字，短线、第二串数字组成。第一串数字的含义为词语的起始字符在整个文章中的索引号，从0开始编号；第二串数字的含义为词语的终止字符在整个文章中的索引号，从0开始编号。

## 词语本身

词语本身的格式为一个字符串，长度不小于1，是该词语的表述。

## 实体链接页面

使用inception进行实体链接时所链接到的知识库页面，不同的知识库应该有所不同，wikidata知识库为链接到其页面，格式为下划线或星号或URL，形如[http://www.wikidata.org/entity/Q\d+(\[\d+\])\*](http://www.wikidata.org/entity/Q\d+(\%5b\d+\%5d)*)。当该列为下划线时，表示该词语不属于实体，当该列为星号时，表示该词语属于实体但没有找到对应的链接页面（或已在前面的共指实体标记出）。当该列为URL时，表示该实体链接到了该页面，URL的末尾可能跟有一个中括号包围数字的部分，该部分存在时表示该实体包含不止一个词语，括号中的数字表示该跨词语实体的编号，与实体类型的编号对应。

## 实体类型

实体类型为标注时定义的实体类型，本次标注中有 PER、ORG、LOC、REG、OTH 5种，类型后面可能跟有一个中括号包围数字的部分，该部分存在时表示该实体包含不止一个词语，括号中的数字表示该跨词语实体的编号。若该行为下划线，则表示该词语不属于实体

## 封闭式关系类型

封闭式关系类型为标注时连接另一实体与本实体的关系类型，在本文档中仅有Coreference一种类型，表示共指关系

## 另一实体的编号

封闭式关系的源实体编号，分四种情况说明。

情况1：源实体、本实体均为单词语实体，编号即为该词语的编号（第一列）；

情况2：源实体为多词语实体，本实体为单词语实体，编号为源实体的第一个词语的编号+[源实体编号\_0]，实体编号指其在实体类型列结尾中括号内的数字，下同；

情况3：源实体为单词语实体，本词语为多词语实体，编号为源实体的第一个词语的编号+[0\_本实体编号]；

情况4：源实体、本实体均为多词语实体，编号为源实体的第一个词语的编号+[源词语编号\_本实体编号]；

情况5：源实体为词语内部实体，编号即为该词语内部实体的编号（第一列）。

# 最终格式转换

实体、链接、共指部分的导出数据可以使用PCube项目代码目录下PCube/preporecess/ann2train.py的parseNERData函数进行转换，转换时需要将导出的数据解压到/home/disk2/nuclear/PCubeAnn/AnnExport/目录下，总目录名重命名为NER，即/home/disk2/nuclear/PCubeAnn/AnnExport/NER目录下为无数篇文章的标注信息目录的状态。执行完成后所有标注数据以文章为单位，保持原文件名，以json的文件格式保存至/home/disk2/nuclear/PCubeAnn/Traindata/NER目录下。

实体、链接、共指部分的数据最终转换格式如下：

[{"token": 句子原句，无分词空格, "wikiLink": [{"form": 实体在文本中的表述, "type": 实体类型, "wiki\_id": 实体的wikidataID, "idx":[实体在句子中的起始字符索引，实体在句子中的结束字符索引]}, ……], "ner": [字符0类型,字符1类型, 字符2类型, ……]}, ……]

注意，每个句子的排列顺序以原文顺序为准，每个wikiLink按照实体的结束字符索引由小到大排列，每个ner与token的每个字符（单个汉字）对应，ner采用BIO标注格式，即实体的开头标注“B-类型”，其余部分标注“I-类型”，非实体标注“O“。共指消解的关系目前尚无使用的需求，本文不予处理。该最终格式同样是人立方3.0系统实体、链接算法的输出格式。

# 附录

附录1.无法消歧的实体列表

|  |  |  |
| --- | --- | --- |
| 实体名 | wiki实体名 | wiki编号 |
| 行政院 | Executive Yuan | Q715055 |
| 政院 | Executive Yuan | Q715055 |
| 南非 | South Africa | Q258 |
| 美国国务卿 | United States Secretary of State | Q14213 |
| 大陆政府 | Government of the People's Republic of China | Q936190 |
| 中国政府 | Government of the People's Republic of China | Q936190 |
| 高雄市政府 | Kaohsiung City Government | Q6366077 |
| 联合国 | United Nations | Q1065 |
| 辉瑞 | Pfizer | Q206921 |
| 瑞典 | Sweden | Q34 |
| 美国国务院 | United States Department of State | Q789915 |
| 美国务院 | United States Department of State | Q789915 |
| 白俄罗斯 | Belarus | Q184 |
| 塞尔维亚 | Serbia | Q403 |
| 伊朗 | Iran | Q794 |
| 三倍券 | Triple Stimulus Vouchers | Q97345665 |
| 麦当劳 | McDonald’s | Q38076 |
| 开元食品 | Creation Food | Q30947484 |
| 伦敦 | London | Q84 |
| 瑞士 | Switzerland | Q39 |
| 以色列 | Israel | Q801 |
| 加拿大 | Canada | Q16 |
| 芬兰 | Finland | Q33 |
| 卢森堡 | Luxembourg | Q32 |
| 西班牙 | Spain | Q29 |
| 葡萄牙 | Portugal | Q45 |
| 巴西 | Brazile | Q155 |
| 波兰 | Poland | Q36 |
| 东北 | Northeast China | Q1057925 |
| 美国政府 | Federal Government of the United States of America | Q48525 |
| 联邦政府 | Federal Government of the United States of America | Q48525 |
| 川普政府 | Federal Government of the United States of America | Q48525 |
| 川普当局 | Federal Government of the United States of America | Q48525 |
| 拜登政府 | Federal Government of the United States of America | Q48525 |
| 俄罗斯 | Russia | Q159 |
| 俄国 | Russia | Q159 |
| 俄 | Russia | Q159 |
| 大立光 | Largan Precision | Q6488962 |
| 大立光电 | Largan Precision | Q6488962 |
| 敦睦舰队 | ROCN Midshipmen Cruising & Training Squadron | Q15910206 |
| 墨西哥 | Mexico | Q96 |
| 纽约 | New York City | Q60 |
| 加州大学柏克莱分校 | University of California, Berkeley | Q168756 |
| 纽约市 | New York City | Q60 |
| 实体名 | wiki实体名 | wiki编号 |

该名单的json格式副本保存在/home/nuclear/PCube3.0/preprocess/CannotLink.json