


# 科技文献标注手册v1.2


17 2021-04-22 v1.1

17 2021-05-18

## 实体标注

- 任务TASK：应用程序，要解决的问题，要构建的系统。
  - 例如，信息提取，机器读取系统，图像分割等
- 方法Method：xx方法，xx模型，要使用的系统或工具，系统的组成部分，框架。
  - 例如，语言模型，CORENLP，POS解析器，内核方法等。
  -  例如：图像识别模型，电力负荷预测模型
- 评估指标 Metric：可以表达系统/方法质量的指标，度量或实体。
  - 例如，F1，BLEU，Precision，Recall，ROC曲线，均值倒数，均方误差，鲁棒性，时间复杂度等
- 材料Material：数据，数据集，资源，语料库，知识库。
  - 例如，图像数据，语音数据，立体图像，双语词典，释义问题，CoNLL，Panntreebank，WordNet，Wikipedia等
- 其他科学术语Other Scientific Terms：属于科学术语但不属于上述任何类别的短语
  - 例如，物理或几何约束，定性先验知识，话语结构，句法规则，话语结构，树，节点，树核，特征，噪声，准则
- 通用Generic：可能指实体但本身并不提供信息的一般术语或代词，通常用作连接词。
  - 例如，该模型，该方法，先验知识


## 关系标注


 注意：除了[简称Hyponym-of]这种关系可以跨相邻两句，对于其他关系我们只标注句内关系，只标注句内实体之间的关系，只标注句内实体的关系。

- 用于Used-for：B用于A，B模型A，A受B训练，B漏洞利用A，A基于B。例如：
  - **TISPER**系统旨在支持许多文本应用程序。
  - **Xxxx**方法模拟了用户的熟练程度。
  - **Xxxx**算法利用了局部平稳性。
- 简称Hyponym-of：B是A的下位词，B是A的缩写在后文出现，B是A的一种类型，<下位词B，hyponym of，A>。例如：
  - **TUIT**是一个软件库
  - **NLP**应用程序，例如**机器翻译和语言生成**
- 比较Compare：对称关系（用蓝色表示实体）。相反，比较两个模型/方法，或列出两个相对的实体。例如。
  - 与**定量先验**不同，**定性先验**常常被忽略。
- 连接Conjunction：对称关系（使用蓝色表示实体）。充当类似角色或使用/合并和。例如。
  - 从**人类专家**或**知识库**中获得
  - **NLP**应用程序，例如**机器翻译和语言生成**

# FAQ

---

- 实体标注的编著错误的常见的错误有？
  - 除了标签错误之外，实体在选择的时候，边界一定不要手抖，不要误选首尾多字的现象。
- 文本中所有实体都要标注吗？还是说只标注存在关系的实体？
  - 所有的实体都标注，可以不存在依赖关系的实体
  - **标题和摘要**都需要标注
- 同一个实体能否多标？面对多个潜在标签的时候，如何结合上下文语境标注？
  - 同一个实体只能有一个标签，请仔细斟酌
  - 同一个实体在前后文不同的语境里面，标签请尽可能的保持前后一致，但不做强制要求，请结合语境做出合适的判断。
- 何为语义信息丰富的片段和语句，如何标注？
  - 是指实体分布十分密集的语句，例如标题，结论句之类，标注此类语句之时，可以允许适当带一些修饰词，标注的实体应该尽可能的保持语义完整
- 是否允许嵌套标注？
  - 无须嵌套标注，请根据句子的语义丰富程度选择标注的粒度
- 那些是一定要标注出来的，应当重点关注的部分？
  - 同一种方法或者模型的表述一定要标注出来
  - 出现了某某明显特征的短语或者句式，“xxx方法”，“xxx指标”，‘xxx之上实验得出’，‘基于xxx’
- 标记更可能多的信息还是标记部分但是能够泛化的信息？
  - 在语义不丰富的语句标注中，应进行最小粒度进行标注
  - 例如：“收敛成功率”，在语义丰富的句子中标注出一个实体，反之仅将“成功率”和“收敛”当作两个实体
- 关系如何标注，能否标注多元关系？
  -  除了简称Hyponym-of之外，关系只标注句内的关系
  - 不允许标注多元关系，一对实体只能标注一对二元关系
  - 一个实体可以和多个其他实体标注多种类型的二元关系
- 什么是对称关系和非对称关系？对称关系之外，其他关系的方向是怎么样的？
  - 一对关系之中的头实体和尾实体可以互换位置，不影响语义，一般而言对称关系的两个实体都是同种类型的。
  - 对称关系之外，比如说B用于A，关系名是used for，标注方向只能是 B（头实体） + used for + A（尾实体）
- 如何标注other scientific term类别的实体？
  - 属于科技论文之中常用语，但是又不好归类的实体，一般此类实体同<method,task ,material , metric>有较强的联系
  - 标注此类实体之前请先考虑是否能够归为上述四类(method,task , material , metric)的实体，如若不能再考虑标注成other scientific term
- 通用实体与其他科学术语有什么区别？通用实体标注是否有条件，比如只有在其有指代意义或者涉及到关系的时候才标注？
  - 两者有一定的相似性，其他科学术语othersciterm相比较于通用generic应该更具有表达意义，标注的时候的思考优先级应该是：othersci > generic
  - 为了区分这两者，othersci实体应该通常和其他实体<method , task , metric , material>存在关系，若不存在关系则标注成generic。

- 对于代词/下位词统一标注成generic：例如：我们的方法，我们的系统，过去的方法，存在的研究etc..
-  对于包含了英文解释/缩写的标注, 如何标注?
  - 例如：许瓦兹-克里斯托(Schwarz-Christoffel Mapping,SCM)映射方法.
  - 需要完全标注，本例之中则为【许瓦兹-克里斯托(Schwarz-Christoffel Mapping,SCM)映射方法】

## 常见标注问题

- 标签错误
  - 尽可能避免，请标注同学和质检同学仔细核对
- 关系漏标
  - 因为关系和实体不是以句为单位的标注，可能会存在漏标的情况比较严重？
  - 若是以句为单位是否为好一些？况且除了简称关系之外不存在跨句关系。
  - 最后再次全部标注所有的简称关系
- 边界问题
  - 对于【图像识别模型】，【电力负荷预测模型】这中标注成【方法】
  - 若是出现‘xxx是用于图像识别的模型’，‘电力负荷预测的模型’这样子，标注成【电力负荷预测】，【图像识别】，作为任务
- 一致性问题
  - 在同一个篇章之内，同一实体，存在前后标签不一致问题
  - 标准不统一，例如：【图像识别模型】(方法),【电力负荷预测模型】(通用)

## 返回格式json

```
{
  Id :{xxx }

  Tokens : {
    title : xxxxx,
    abstract : xxxxx
  }
  Entity: {
    [entity : xxx , entity_type : xxx , start : xxx , end:xxx , position=title/abstract]
    ,
    [entity : xxx , entity_type : xxx , start :xxx , end :xxx , position =
title/abstract]
  }
  # start : xxx , end:xxx 之中的 xxx 为分别为实体在篇章之中的起始位置
  Relationship:
  {
    [relation_type : xxx , head : xxx , tail:xxx]
    [relation_type : xxx , head : xxx , tail:xxx]
    [relation_type : xxx , head : xxx , tail:xxx]
  }
  # head : xxx , tail:xxx 之中的xxx 分别为head 和 tail 实体的在entity字典之中的index
}
```

