# 实体标注说明书

## 1 标注工具

使用BRAT进行实体标注，依照《BRAT使用说明书》使用。

## 2 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

## 3 标注数据结构

每篇文章对应一个json文件，包含字段entities、raw。

* raw字段：文章原文，使用list存放，其结构为[句子1, 句子2, 句子3, …]，每个元素是文章中的一句话。句子也使用list存放，其结构为[‘早上’, ‘好’, ‘！’]，每个元素是句子中的一个词。
* entities字段：命名实体识别的正确结果，使用list存放，其结构为[实体1, 实体2, 实体3, …]，每个元素是一个命名实体。实体使用dict存放，其包含字段type、name、sent\_id、pos。
  + - type字段：实体类型，使用string存放，包含‘PER’、‘ORG’、‘LOC’三种情况。
    - name字段：实体名，使用list存放，其结构为[‘北京’, ‘邮电’, ‘大学’]，每一个元素是实体名中包含的一个词。
    - sent\_id字段：实体所在的句子索引，使用int存放。
    - pos字段：实体在所在句子中的位置，使用list存放，其结构为[0, 2]，第一个元素是起始位置的索引，第二个元素是结束位置的索引+1。

示例：

{

‘raw’: [[‘北京’, ‘邮电’, ‘大学’, ‘开学’, ‘了’, ‘！’], [], []],

‘entities’: [

{

‘type’: ‘PER’,

‘name’: [‘北京’, ‘邮电’, ‘大学’]

‘sent\_id’: 0

‘pos’: [0, 3]

},

{…},

{…},

…

]

}

## 4 工具标注方法

现有命名实体识别工具已经可以达到98%左右的准确率，这给本次数据标注带来了极大的帮助。使用哈工大开源NLP工具ltp，可以轻松获得较高准确率的实体识别结果。再使用python编码，将标注结果转为ann格式文件，BRAT可以对此进行读取，方便进行下一步人工筛查。

## 5 人工标注方法

依照《BRAT使用说明书》进行标注配置，标注类型为PER、ORG、LOC，并修改合适的标签颜色。进入BRAT工具web端后可以直接进行人工标注，以及对工具标注结果进行筛查。

## 6 格式转换

完成工具+人工的标注后，可以从BRAT工作目录下获得txt文件和ann文件，使用python编码完成txt+ann文件到目标数据结构json文件的转换，完成标注工作。