标注规范及审核规范

标注： 对chatGPT生成的数据进行人工的实体和关系标注

审核： 对现有标注数据进行纠偏和实体名称聚焦（减少代词的标注和关系指向代词）

> 我们标注的目标是导入人物知识图谱进行信息结点（对应实体）和链接（对应关系）的补充，结点需要尽可能详细，尽量减少代词

> 实体的span不要重叠 ⁽⁽(°▫°)⁾⁾! 一个实体只能有一个标签！！！

> 标注关系的 就近原则： 关系所连接的两个实体尽量在 以逗号或句号分割的 同一个句子段里；如果句子中出现多个相同的详细的实体，都需要标记出实体，但是关系连接按照就近原则和层级原则

> 标注关系的 层级原则： 实体通过关系连接到最低层级的实体 ，可用于后续的 推理，比如：中国台湾、台湾海军 和 反舰导弹 ，反舰导弹 归属到 台湾海军 即可；

> 每个事件的 arguments 中的实体，需要对应到实体类型中，如果缺少标注，可以进行补充，并且不能超过实体类型涉及的范围

实体类型

1. 人物

> 历史人物、人物；尽量以具体的人物名字为主

下面的duie2数据集就是有很多的代词或者略微模糊的群体代词，尽量少标，但是如果句子中有出现具体的名字一定要标，代词看在句子中的重要程度自行斟酌哈 ✧⁺、(ˊ▾ˋ)ノ⁺✧

```
这里的人物例子有一些是反例，里面包含了很泛的人物群体指代，我们在构建人物知识图谱结点的时候要更加具体的名称。
```

```
人物 ： ['网友', '女兵', '特种兵', '军人', 'VBSS小组成员', '海豹特战队员CharlesKeating', '指挥员', '飞行员', '杨宇军', '记者', '设计师', '网友', '战士们', '武僧', '军事爱好者', '电子爱好者', '普京', '网友', '海军飞行员张超', '飞行员', '地勤人员', '飞行员', '网友', '游人', '中国军人', '边防连官兵', '步兵', '郝静', '尹淳']
```

2. 地点

> （地域、人文政治地区、地点、城市）
> 非政治实体的、表示地点和区域的实体（偏向于非人造的）
> 2.1 地址：西源大道2006号，北纬3.6°
> 2.2 水体：长江、池塘、海峡、印度洋
> 2.3 自然地区：海岛,南沙群岛,礁,开采区域,中国西面的欧亚地震带
> 2.4 一般地区：沿海地区、中国西南地区

例句:

```
位于浙江省宁波市红星镇鄂塞村壶山西路街道金拱门巷17号的中微型探测卫星博物馆于2019年12月7日开业啦。
```

这里的 浙江省宁波市红星镇鄂塞村 是 政治实体， 壶山西路街道金拱门巷17号 是 地点，中微型探测卫星博物馆 是 设施， 2019年12月7日 是 时间； 中微型探测卫星博物馆 和 浙江省宁波市红星镇鄂塞村 是 归属 关系；这句话没有包含事件，这样就标注完成了。

3. 政治实体

> 一个或多个国家、城市、省（州）和行政区域等政治实体
> *政治实体和组织可能有点混淆，注意区分*

```
政治实体 ： ['越南', '俄罗斯', '美国', '中国南海', '印度', '中方', '伊朗', '印度', '日本', '美国', '澳大利亚', '加拿大', '美国', '中国', '共和国', '北京', '东南亚', '俄罗斯', '中国', '俄罗斯', '俄', '中国', '北京方面', '美国华盛顿', '日本', '马来西亚', '我国'...]
```

3.1 洲：亚洲、非洲
3.2 国家：中国、巴西、美国
3.3 州或省：河北、四川、宾夕法尼亚
3.4 县或地区：华中地区、新英格兰地区、花莲县

3.5 市、乡、镇、村
3.6 中东、欧美、东亚、欧盟

比如，下面的例句：

发生在布鲁塞尔的连环爆炸案再次把恐怖阴云带到欧洲,自巴黎暴恐案以来,西欧这片原本宁静的土地似乎成为了恐怖组织的新目标。

这里的 布鲁塞尔 、 欧洲 、 巴黎 、 西欧 都是 政治实体

4. 组织

> （公司、集团、企业、服务社、协会、研究所、劳工组织\军队、党派！！！）
> 表示一个或多个公司、机构、单位和组织等的实体
> 简而言之：标注 完整的全名

4.1 政府部门（人民解放军、信息产业部、国务院）
4.2 商业机构（新浪、Google、诺基亚、科大讯飞）
4.3 娱乐机构（英皇、歌舞团、环球唱片有限公司）
4.4 非政府组织机构（联合国维和部队、共产党、研究所、协会、北约）
4.5 新闻机构（人民日报、美国之音、出版社、新闻社）
4.6 宗教机构（基督教、罗马教廷、道教）
4.7 医学机构（研究中心、中科院神经所、生物医学实验室）
4.8 体育机构（体育协会、皇家马德里、费城76人）

注意：如果 组织 前面有 国家，如： 意大利依维柯公司 ，直接使用这个更加详细的实体作为 组织 。~~公司的优先级比党派低。~~**~~和军事政治相关的信息我们就拆的细一点，其他的我们可以粗一点标注~~**

例句--政治相关：

印度的人民党总理莫迪在12号宣布一项重大事件

~~标注~~ ~~印度人民党总理~~ ~~为 职务，~~ ~~莫迪~~ ~~为 人物，二者关系为 任命为~~

实体标注： 印度 为 政治实体， 人民党 为 组织， 莫迪 为 人物，扔掉总理（职务），**党派关系 的标注优先级最大**
关系标注： 人民党 隶属于 印度， 莫迪 的 党派 是 人民党，后续可以用 推理 来判断

印度人民党总理莫迪在12号宣布一项重大事件

实体标注： 印度人民党 为 政治实体， 莫迪 为 人物，扔掉总理（职务），**党派关系 的标注优先级最大**
关系标注： 印度人民党， 莫迪 的 党派 是 人民党，后续可以用 推理 来判断

谭志强，男，中共党员，出生于2001年。

实体标注： 谭志强 为 人物， 中共 为 组织，
关系标注： 谭志强 的 党派 是 中共

例句--一般组织：

塔吉克斯坦科学院是塔吉克斯坦最高科研机构，位于首都杜尚别，成立于1951年，是前苏联科学院的分院。

塔吉克斯坦科学院， 前苏联科学院 都是 组织； 塔吉克斯坦， 杜尚别 都是 政治实体；

塔吉克斯坦科学院， 杜尚别 都 隶属于 塔吉克斯坦

例子--注意：军队 的话还是带上 政治实体 比较详细比较好。

组织 ： ['越军', '解放军', '中国海军舰艇编队', '俄罗斯海军', '海军', '舰队', '里海舰队', '南部军区', '俄空军', '沈空航空兵某师', '波音公司', '美军', '美国空军', '中国海军', '我军', '我海军', '南京军区某部']

5. 教育机构

> （学校、小学、中学、大学）

注意与组织区分

教育机构 ： ['山东大学', '钟祥市实验小学', '西北农业大学', '武汉大学', '中国人民解放军艺术学院戏剧系', '山东师范大学', '四川省委党校', '同济医科大学', '白塔寺武校', '潭口中学', '华师大研究生课程班', '上海音乐学院']

6. 时间

时间 ： ['2000年代', '早期', '近期', '近日', '2016年7月至8月', '近日', '5月11日', '2017年11月15日', '深秋时节', '今年3月30日', '5月31日', '30日', '本月3日', '缅怀日', '新年度', '初春']

7. 数字

> 除了时间以外的数字，主要和金钱相关，注意与时间区分

数字 ： ['1607人', '1105亩', '1亿元', '8.06亿人民币', '11万余平方米', '1390.5米', '72.10万平方公里', '1180公顷', '4288米', '100万元', '500万日元', '人民币12886.30万元', '560人']

8. 设施

> 建筑、机场、铁路、桥、XX水库等表示设施本身而非地点的实体
> 偏向于人造建筑
> *和地点会混淆*

设施 ： ['卫星', '塔台', '塔台', '大桥', '大桥', '跨刻赤海峡大桥公路', '跨刻赤海峡大桥', '河南省共产主义渠获嘉段', '四座新的导弹掩体', '军事设施', '诺福克海军基地', '三菱下关江浦工场', '华东某机场', '乌姆罗伊军营', '乌姆罗伊军营停机坪', '尼鲁姆–杰卢姆水电站']

例句：

庆营嘉和苑占地10453平方米，总建筑面积约3.1万平方米，其中住宅面积2.7万平方米，商业面积4000平方米。

庆营嘉和苑 标记为 设施， 注意 这个不是标注为 地点、组织 ⊙_⊙
10453平方米， 3.1万平方米， 2.7万平方米， 4000平方米 标记为 数字

9. 装备

> 表示一个或多个武器装备的实体
> 补充： XX技术 也可算作 装备

注意： 中国T-55坦克 需要拆分成 中国 （政治实体）， T-55坦克 （装备）

装备 ： ['99A坦克', '国产T-55M3型主战坦克', '该型坦克', 'T-M3主战坦克', '125毫米滑膛炮', 'L7型105毫米坦克炮', 'T-55/54坦克', '轰-6K', '歼-10', '直-20', '直-8G', '导弹驱逐舰长沙舰', '鞍鞯斯坦号护卫舰', '导弹自动拦截技术']

10. JS活动（军事活动）

> 对于军事行动的概括性称呼，如军事代号、行动代号、军事活动名称等
> （后期生成或爬取数据）

注意：装备试验 如 飞机起降试验 也算作 军事活动实体

军事行动 ：['“巴巴罗萨”行动', '诺曼底登陆', '闪电行动', '反潜作战', '反潜演练', '登陆战演习', '飞机起降试验', '国庆阅兵式'...]

职务

> 好利来公司的董事长， 总理， 美国总统 这些都标为职务，
> 民进党候选人 注意，要拆出

例句：

时任日本美术家协会主席、著名美学家宇智波佐助先生在生病期间说：听我说谢谢你，因为有你，温暖了四季~~

日本美术家协会主席， 著名美学家 都是 职务， 宇智波佐助 是 人物

时任羊村大学教务处主任、校长的宇智波佐助先生在生病期间说：听我说谢谢你，因为有你，温暖了四季~~

羊村大学教务处主任、校长 整体是 职务，因为 实体的连续性， 校长 是这个大学的职务，不要拆开来好吧(⚭-⚭)ว， 宇智波佐助 是 人物

关系类型

1. 隶属于

> 组织、政治实体、教育机构 --> 组织、政治实体、教育机构

> 注意！！！ 没有 人物

组织隶属于组织或政治实体、政治实体隶属于政治实体的关系

```
隶属于  ：[['俄罗斯海军', '组织', '-->', '俄罗斯', '政治实体'], ['里海舰队', '组织', '-->', '俄罗斯海军', '组织'], ['海军', '组织', '-->', '南部军区', '组织'], ['南部军区', '组织', '-->', '俄罗斯', '政治实体'], ['海豹特战队员CharlesKeating', '人物', '-->', '美军', '组织'], , ['国防部', '组织', '-->', '中方', '政治实体'], ['杨宇军', '人物', '-->', '国防部', '组织'], ['日本自卫队', '组织', '-->', '日本', '政治实体']]
```

例句：

```
美国的非洲司令部位于。。。
```

中间有 的 进行切割，所以 标注 美国 （政治实体）、 非洲司令部 （组织）
二者是隶属于

2. 搭载

> 搭乘着、装载了（以装载为主）

装备 --> 装备

一个装备搭载了人物 或 搭载了另一个装备的关系

```
搭载  ：  [['卡玛斯卡车', '装备', '-->', '普京', '人物'], ['T-M3主战坦克', '装备', '-->', 'L7型105毫米坦克炮', '装备'], ['MP7冲锋枪', '装备', '-->', '安保人员', '人物'], ['辽宁舰', '装备', '-->', '李克强', '人物'], ['F-4EJ战机', '装备', '-->', '两名飞行员', '人物'], ['该车', '装备', '-->', 'ZTM-130mm机炮', '装备']]
```

3. 归属

> 装备、设施 --> 人物、组织、政治实体
> 装备 --> 设施

装备归属于组织或政治实体的关系

如果句子中存在 组织 和 政治实体 如：台湾海军 、 中国台湾 ，可以自主选择把对应的 装备、设施 如： XX型无人机 归属到 就近的 最低层 组织 和 政治实体 ， 归属 议按照层级来 连接关系，比如 连接到 台湾海军

```
归属  ：  [['国产T-55M3型主战坦克', '装备', '-->', '越南陆军', '组织'], ['T-M3主战坦克', '装备', '-->', '越军', '组织'], ['常规潜艇', '装备', '-->', '解放军', '组织'], ['AH-64D阿帕奇攻击直升机', '装备', '-->', '美国', '政治实体'], ['新一代航母003号', '装备', '-->', '我海军', '组织']]
```

4. 丈夫

> 头实体 的 丈夫 是 尾实体

```
丈夫  ：  [['张玛莉', '人物', '-->', '李忠琛', '人物'], ['郭蔼明', '人物', '-->', '刘青云', '人物'], ['蓝为洁', '人物', '-->', '汤晓丹', '人物']]
```

5. 妻子

> 头实体 的 妻子 是 尾实体

```
妻子  ：  [['李忠琛', '人物', '-->', '张玛莉', '人物'], ['刘青云', '人物', '-->', '郭蔼明', '人物'], ['汤晓丹', '人物', '-->', '蓝为洁', '人物']]
```

6. 子女

> 头实体 的 子女 是 尾实体

子女 ： [['王桂荃', '人物', '-->', '梁思礼', '人物'], ['杨丽华', '人物', '-->', '宇文娥英', '人物'], ['周宣帝', '人物', '-->', '宇文娥英', '人物']]

7. 党派 （高优先级）（后期生成或爬取数据）

> 人物 的 党派 是 组织
> 注意关系区分： 工作于

[['希拉里', '人物', '-->', '民主党', '组织']]

美国 和 民主党 是要拆分的

注意下面的例句，遇见能标记出党派的文本段，记得切分 职务：

特朗普，男，中共党员，曾潜入CIA进行长达20年的工作生活。

句子中标注实体： 特朗普 为 人物， 中共 为 组织，CIA 为 组织，
句子中标注关系： 特朗普 的 党派 为 中共， 特朗普 工作于 CIA
(ง ˙ω˙)✧

8. 母校

> 人物 的 母校 是 教育机构

母校 ： [['陈维涵', '人物', '-->', '中央戏剧学院', '教育机构'], ['朴信惠', '人物', '-->', '中央大学', '教育机构'], ['奥尔洛娃', '人物', '-->', '俄罗斯下诺夫哥罗德州大学', '教育机构']

~~**9. 研制 （为了减少和 归属于 的重叠，把 研制 去掉，并定义为 研制事件）**~~

> (ง ᵒ̌ω ᵒ̌)ง⁼³₌₃
> 人物、组织、政治实体 研制 装备

组织或政治实体研制装备的关系

10. 工作于

> 人物 工作于 组织、政治实体
> 人物 工作于 设施
> 注意关系区分：党派

2023年5月30日，<印度尼西亚><北风营区>的军官<阿里亚·穆罕默德>组织了一场<山地作战训练>。<阿里亚·穆军默德>是<印度尼西亚><民主党>的成员，他注重士兵在复杂地形条件下的应对能力，提高北风营区的战斗素质。

第一句话：
这里的 北风营区 是 设施， 阿里亚·穆罕默德 是 人物，二者是 工作于 的关系；
山地作战训练 是 军事活动；
第二句话：
阿里亚·穆罕默德 是 人物， 印度尼西亚 是 政治实体， 民主党 是 组织，有两个关系： 隶属于 和 党派

11. 任命为

> 人物 任命为 职务

事件类型

> 每个事件的 arguments 中的实体，需要对应到实体类型中，如果缺少标注，可以进行补充，并且不能超过实体类型涉及的范围
> trigger表示该事件类型的触发词，可以不和实体重合
> 由于没有专门的事件标注工具，我们用实体类型的标签来曲线救国的标记事件的trigger和对应的实体（放置在arguments中），这些实体必须和之前的实体类型重合，如何前面漏标注了，请先补充实体类型，在标记事件的实体类型，thanks (•ᴗ•)

---

1. 访问事件

trigger

句子中出现诸如： 出席 、 到访 、 暗访 、 考察 等关键词作为访问事件的触发词，标记为 访问-trigger

arguments(role)

参与人物

地点

时间

例句1:

> 2019年4月1日，希拉里一行到访台湾，蔡英文接见到访人员并举行隆重的会议。

到访 标注为 访问-trigger，
希拉里 和 蔡英文 都标注为 访问-参与人物，
台湾 标注为 访问-地点 ，
2019年4月1日 标注为 访问-时间

例句2:

> 江苏师范大学校长周汝光一行考察泰兴高新区

实体： 江苏师范大学 （职务）， 周汝光 （人物）， 泰兴高新区 （政治实体）
事件trigger： 考察 （访问-trigger）
事件实体： 周汝光 （访问-参与人物）， 泰兴高新区 （访问-地点）

例句3:
mixed_data test -> 第1008条句子

---

2. 交易事件

trigger

arguments(role)

买方

卖方

时间

交易物

金额

场所（交易地点）

---

3. 试验事件

> 武器装备（技术） 的 测试 试验 验证

trigger

arguments(role)

主体

（人、组织、政治实体；尽量以事件的发起组织为主）

场所（试验地点）

时间

试验物

（ 装备 实体 前的数量还是尽量标注一下更好！！！）

试验活动（军事活动）（补？）

---

4. 演练事件

> 部队、军队组织 的 演练

trigger

arguments(role)

主体

（人、组织、政治实体；尽量以事件的发起组织为主）

军事活动

代号、演习内容等

时间

地点

---

5. 部署事件

trigger

arguments(role)

主体

（组织、政治实体）

军事力量

组织、装备

时间

地点

部署活动（军事活动）（补？）

---

6. 游行事件

trigger

示威、抗议，静坐，罢工和暴乱

arguments(role)

主体

（人群、组织、政治实体）

时间

地点

---

7. 选举任免事件

trigger

选举、任命

arguments(role)

主体

（组织、政治实体）

党派组织、政府。。。

人物

参选人、任命人

时间

地点

职务

---

8. 研制事件

trigger

研发、创造、突破技术

arguments(role)

主体

人物、组织、政治实体

客体-研制物

装备、设施

时间

地点

意见建议（Comments）：

试验事件 和 研制 事件 有部分重叠，需要切割
还需要补充一些 和 突发事件 （如地震后派遣军队抢险救灾、维和部队对地区暴动的应急治安（不紧急的应该为部署）和地下救援任务、军事冲突等）
