**CCKS 2017 电子病历命名实体识别任务说明**

1. **任务定义及描述**

本评测任务为面向电子病历的命名实体识别，简称CNER （Clinical Named Entity Recognition）。即对于给定的一组电子病历文档（纯文本文件），任务的目标是识别并抽取出与医学临床相关的实体名字（entity mention），并将它们归类到预先定义好的类别（pre-defined categories），比如疾病、症状、检查等。

本年度的评测任务是CCKS 2016影视领域实体发现与实体链接评测的一个延续，目的旨在促进限定领域实体识别与实体链接的研究发展。本次评测由清华大学知识工程实验室、微软亚洲研究院以及北京极目云健康科技有限公司联合主办。

* 1. **输入和输出**
* **输入**

任务的输入为一组电子病历数据。所谓病历，是指病人在医院诊断治疗全过程的原始记录，它包含有首页、病程记录、检查检验结果、医嘱、手术记录、护理记录等等。根据CNER任务的特点，我们将原始的电子病历数据进行了裁减和整理，重点将病程记录及检查检验结果抽取出来，以纯文本的格式进行存储。单个病人的单次就诊记录为一个条目。

* **输出**

任务的输出包含：

1) 给定文档中出现的与医学相关的实体名字（字符串边界）;

2) 每一个实体名字对应的类别。

* 1. **实体类别定义**

本次评测将命名实体的类别限定在以下5类：

* **症状和体征**：症状是患者描述的主观感受、体征则是外部观察到的客观事实。举例来说，流鼻涕、头昏及体温超过摄氏三十八度，这三者都是感冒的元素：但流鼻水及头昏是感冒的症状；而体温超过三十八度，则是感冒的体征。
* **检查和检验**：指通过实验室技术、医疗仪器设备为临床诊断、治疗所提供的依据。
* **疾病和诊断**：疾病是机体在一定的条件下，受病因损害作用后，因自稳调节紊乱而发生的异常生命活动过程；诊断，是指根据症状来识别病人所患何病。
* **治疗**：通常是指干预或改变特定健康状态的过程。为解除病痛所进行的活动，如药物、手术等。
* **身体部位**：指疾病、症状和体征发生的人体解剖学部位。

标注例子如下（注意：实体与实体之间存在嵌套关系）：

女性，88岁，农民，双滦区应营子村人，主因<症状和体征><身体部位>右髋部</身体部位>摔伤后疼痛肿胀，活动受限5小时</症状和体征>于2016-10-29；11：12入院。

1.患者老年女性，88岁；

2.既往<症状和体征>体健</症状和体征>，<症状和体征>否认药物过敏</症状和体征>史。

3.患者缘于<症状和体征>5小时前不慎摔伤，伤及<身体部位>右髋部</身体部位></症状和体征>。伤后患者<症状和体征>自感伤处疼痛</症状和体征>，呼我院120接来我院，查<检查和检验><身体部位>左髋部</身体部位>部X光片</检查和检验>示：<检查和检验><身体部位>左侧粗隆间</身体部位>骨折</检查和检验>。给予<治疗>补液</治疗>等<治疗>对症治疗</治疗>。患者<症状和体征>病情平稳</症状和体征>，以<疾病和诊断>左侧粗隆间骨折</疾病和诊断>介绍入院。患者自入院以来，<症状和体征>无发热</症状和体征>，<症状和体征>无头晕头痛</症状和体征>，<症状和体征>无恶心呕吐</症状和体征>，<症状和体征>无胸闷心悸</症状和体征>，<症状和体征>饮食可</症状和体征>，<症状和体征><身体部位>小便</身体部位>正常<症状和体征>，<症状和体征>未排<身体部位>大便</身体部位></症状和体征>。

4.<检查和检验>查体</检查和检验>：<检查和检验>T36.1 C <检查和检验>，<检查和检验>P87次/分<检查和检验>，<检查和检验>R18次/分<检查和检验>，<检查和检验>BP 150/93mmHg <检查和检验>,<症状和体征><身体部位>心肺</身体部位><检查和检验>查体</检查和检验>未见明显异常</症状和体征>，专科情况：<症状和体征><身体部位>右下肢</身体部位>短缩畸形约 2cm </症状和体征>，<症状和体征><身体部位>右髋部</身体部位>外旋内收畸形</症状和体征>，<症状和体征><身体部位>右髋部</身体部位>压痛明显，叩击痛阳性</症状和体征>,<症状和体征><身体部位>右髋关节</身体部位>活动受限</症状和体征>。<症状和体征><身体部位>右足背动脉</身体部位>波动好</症状和体征>，<症状和体征><身体部位>足趾</身体部位>感觉运动正常</症状和体征>。

5.辅助检查：本院<检查和检验><身体部位>右髋关节</身体部位>正位片</检查和检验>：<检查和检验><身体部位>右侧股骨粗隆间</身体部位>骨折</检查和检验>

入院后完善各项检查，给予<治疗><身体部位>右下肢</身体部位>持续皮牵引<治疗>，<治疗>应用<治疗>健骨药物</治疗>治疗</治疗>，患者<症状和体征>略发热</症状和体征>，查<检查和检验>血常规</检查和检验>：<检查和检验>白细胞数12.18\*10^9/L<检查和检验>,<检查和检验>中性粒细胞百分比92.00%</检查和检验>。<治疗>给予应用<治疗>抗生素</治疗>预防感染</治疗>。复查（<检查和检验>血常规</检查和检验>：<检查和检验>白细胞数6.55\*10^9/L<检查和检验>，<检查和检验>中性粒细胞百分比74.70%</检查和检验>，<检查和检验>红细胞数2.92\*10^12/L</检查和检验>，<检查和检验>血红蛋白94.0g/L</检查和检验>。考虑<疾病和诊断>贫血</疾病和诊断>，<治疗>指示加强营养</治疗>。<治疗>建议患者手术治疗，患者拒绝手术治疗</治疗>。<治疗>继续右下肢牵引</治疗>，患者家属要求今日出院。

患者<症状和体征>精神状况好</症状和体征>，<症状和体征>无发热</症状和体征>，诉<症状和体征><身体部位>右髋部</身体部位>疼痛</症状和体征>，<症状和体征>饮食差</症状和体征>，<症状和体征><身体部位>二便</身体部位>（大小便标为<身体部位>）正常</症状和体征>，<检查和检验>查体</检查和检验>：<症状和体征>神清</症状和体征>，<症状和体征>各项生命体征平稳</症状和体征>，<症状和体征><身体部位>心肺腹</身体部位><检查和检验>查体</检查和检验>未见异常</症状和体征>。<症状和体征><身体部位>右髋部</身体部位>压痛</症状和体征><症状和体征><身体部位>右下肢</身体部位>皮牵引固定好，无松动</症状和体征>，<症状和体征><身体部位>右足背动脉</身体部位>波动好</症状和体征>，<症状和体征><身体部位>足趾</身体部位>感觉运动正常</症状和体征>。

1. **数据集描述**

本任务采用的数据集由北京极目云健康科技有限公司提供，数据来源于其云医院平台的真实电子病历数据，共计800条（单个病人单次就诊记录），经脱敏处理，仅限CCKS 2017 竞赛评测用。数据集的初步统计信息如下：

|  |  |
| --- | --- |
| 病区 | 病例数 |
| 老年病科病房 | 120 |
| 心血管内科病房 | 115 |
| 泌尿外科病房 | 109 |
| 胃肠外科病房 | 78 |
| 儿科病房 | 66 |
| 康复科病房 | 66 |
| 骨伤科病房 | 56 |
| 神经外科病房 | 52 |
| 骨病科病房 | 47 |
| 肝胆外科病房 | 22 |
| 神经内科病房 | 21 |
| 中医一科病房 | 15 |
| 妇科组病房 | 10 |
| 心胸外科病房 | 7 |
| 消化内科病房 | 6 |
| 眼科病房 | 5 |
| 肛肠外科病房 | 3 |
| 中医科病房 | 1 |
| 口腔科病房 | 1 |

单条病例的数据内容如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 科室 | 一般项目 | 病史特征 | 入院病情 | 诊疗过程 | 出院情况 |

我们的任务主要是对后5个域中的文本进行CNER。CNER的结果输出要求如下（一条对应一个实体）：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| note\_id | Field\_id | mention | pos\_b | pos\_e | category |
| 病例编号 | 域编号 | 实体名字 | 起始下标 | 结尾下标 | 所属类别 |

本数据集分为三部分，

* training dataset：人工标注的数据集。以供参赛队进行supervised learning。
* unlabeled dataset：未标注的数据集，作为参赛队进行unsupervised 或者semi-supervised learning的补充数据集。
* test dataset：用于评价参赛队的评测结果。

本次任务发布的同时，将提供一部分样例数据供参考，完整的数据集（及相关数据格式说明）将择期发布，请关注CCKS 2017官网的“评测竞赛”：<http://www.ccks2017.com/index.php/eval/>。

1. **评价**

本评测采用精确率（Precision）、召回率（Recall）以及F1-Measure作为评价指标。参赛系统的输出结果集合记为，人工标注的结果（Gold Standard）集合记为，我们定义中的元素与中的元素等价，当且仅当：

基于以上等价关系，我们定义集合与的交集为。由此，得到实体识别的评价指标如下：

, ,

同时，按照category的5个不同类别，对每个子类进行分开评价，共得到6个评价结果，分别为：，，，，，。

1. **任务提交指南**

每一个参赛队需提交的材料如下：

1. CNER运行结果文件
2. 代码及说明
3. 方法描述文档（非评测论文，评测论文撰写要求见CCKS 2017官网）

以上三个文件需在任务提交截止日期前发送至邮箱：[zhang-jt13@mails.tsinghua.edu.cn](mailto:zhang-jt13@mails.tsinghua.edu.cn)。邮件的标题为：“参赛队名称+CNER”。

结果文件用result.json命名，采用JSON格式存储，文件格式需要与训练数据中的标注结果文件完全一样。

代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），用code.xxx命名，要求提交所有的程序代码及相关的配置说明，确保程序能够正确运行，且所得结果与result.json相符。

方法描述文档用CNER.pdf命名，包含算法描述及参数设置，需用pdf格式存储（LNCS风格的Springer 出版物格式），页数不超过5页。

1. **重要时间点**

* 2017年4月17日，发布任务描述及样例数据
* 2017年4月17日~5月20日，参赛队伍报名
* 2017年5月5日，发布训练数据
* 2017年7月5日，发布测试数据
* 2017年7月7日，提交测试结果
* 2017年7月24日，公布评测结果
* 2017年8月6日，提交评测论文