**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。
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 |

数据集分为training dataset、test dataset以及unlabeled dataset。具体描述如下：

* 1. **Training dataset**

1. 电子病历数据组织成四个文件夹：一般项目、病史特征、诊疗过程、出院情况;

2. 每个目录下存储文本数据（\*.txtoriginal.txt）及对应的标注结果数据(\*.txt)，通过文件名进行一一对应。如“一般项目-1.txt”对应“一般项目-1.txtoriginal.txt”;

3. 标注结果文件包含以下4个字段, 字段与字段之间用tab键隔开。

* （实体名称）
* （实体名称在原文中出现的起始位置）
* （实体名称在原文中出现的终止位置）
* （实体所属类别）
	1. **Test dataset**

数据格式与training dataset完全相同。要求参赛者输出的标注结果与training dataset的标注结果完全相同（因格式问题导致无法正确评价，将不计成绩）。

* 1. **Unlabeled dataset**

本任务将提供800-2000条未标注的数据集，作为参赛队进行unsupervised 或者semi-supervised learning的补充数据集。

1. **评价**

本任务采用精确率（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。将从两个层面进行评价。

* 1. **Strict Metrics**

 我们定义与严格等价，当且仅当：

基于以上等价关系，我们定义集合与的严格交集为。由此得到严格评价指标：

, ,

* 1. **Relaxed Metrics**

 我们定义与松弛等价，当且仅当：

基于以上等价关系，我们定义集合与的松弛交集为。由此得到松弛评价指标：

, ,

最后，按照category的5个不同类别，对每个子类进行分开评价，也就是每一个子类得到两个evaluation（strong vs weak），共得到12个评价结果，如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Symptom | disease | Exam | Treatment | Body | overall |
| Strict |  |  |  |  |  |  |
| relaxed |  |  |  |  |  |  |

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。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~6月15日，陆续发布训练数据
* 2017年6月15日，发布unlabeled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

**另：本任务创建了一个交流互动的google group**

**ccks2017-cner@googlegroups.com**

**所有参赛队务必加入这个group，所有有关本任务的最新动态以及调整，都会在group中发布，若group中发布的信息与本任务书有冲突，皆以group发布的信息为准，切记切记！**