**CCKS 2018 面向中文电子病历的命名实体识别**

 本任务由清华大学知识工程实验室及医渡云（北京）技术有限公司联合主办，是CCKS 2017 CNER评测的改进和完善。在专业医生团队的指导和参与下，从临床应用的具体实际出发，在全新的电子病历数据上构建标注质量更高，更具有挑战性的医疗实体识别任务。它是围绕中文电子病历语义化开展的一系列评测竞赛中的第一步，为后续即将推出的医疗实体结构化、医疗实体标准化以及医疗实体关系抽取等任务打下坚实基础。

1. **任务定义及描述**

本评测任务为面向中文电子病历的命名实体识别，即对于给定的一组电子病历纯文本文档，任务的目标是识别并抽取出与医学临床相关的实体提及（entity mention），并将它们归类到预定义类别（pre-defined categories），比如症状、药品、手术等。

* 1. **形式化定义**

**输入**：

 1.电子病历的自然语言文本

 2.预定义类别：

**输出**：

 实体提及和所属类别对的集合：

 其中是出现在中的医疗实体提及（mention），和分别表示在中的起止位置，表示所属的预定义类别。要求实体提及之间不重叠，即。

* 1. **实体类别定义**

 结合数据源“现病史记录”的内容及特点，本任务实体类型聚焦在**症状、药物、手术**三大类。由于症状类型实体多表现为结构化形式，本次任务将症状类型进一步细化，划分为三类：解剖部位 （复合症状的主体）、症状描述（复合症状的描述）及独立症状。最终将本任务预定义类别限定为以下五类：

1. **解剖部位**：由多种组织构成能行使功能的结构单位，例如“腹部”
2. **症状描述**：指患者患病后对机体生理功能异常的自身体验和感觉，同时需与解剖部位联合输出，例如“不适”，需与腹部组合输出“腹部不适”
3. **独立症状**：指患者患病后对机体生理功能异的自身体验和感觉，可独立输出，例如“眩晕”
4. **药物**：用来治疗、预防或促进健康的一种化学物质
5. **手术**：指医生用医疗器械对病人身体进行的切除、缝合等治疗
6. **数据及标注说明**

本任务采用的数据集由医渡云（北京）技术有限公司提供，并组织专业的医生团队对数据进行整理和标注，仅限CCKS 2018 竞赛评测用。

电子病历主要有两类,即门诊病历和住院病历。门诊病历通常较短，包含信息较少，也缺乏对患者治疗情况的跟踪，因此本任务仅考虑住院病历。住院病历内容包括住院病案首页、**入院记录**、病程记录、医嘱单、辅助检查报告单、病理资料等。入院记录记录了患者入院后经诊治医师通过问诊、查体、辅助检查等获得的相关资料，具体内容包括一般项目、主诉、**现病史**、既往史、个人史、家族史、月经婚育史、体格检查、辅助检查、诊断等，其中，**现病史**是住院病历的重点内容，着重于了解患者本次疾病的发生、演变、诊疗经过， 蕴含了更丰富的医疗信息，因此，本次评测主要面向**现病史**部分进行医疗命名实体的识别和抽取。

数据集分为训练集和测试集，由于医疗数据尤其是电子病历数据的特殊性及标注工作的艰巨性，我们将分批逐步发布标注数据，初步计划发布600个现病史文档作为训练集，200-400份作为测试集。

有关数据集的详细描述及标注规范，将随数据的发布一并加以说明，任务书中不再赘述。

1. **评测指标**

本任务采用精确率（Precision）、召回率（Recall）以及F1-Measure作为评测指标。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。集合元素为一个实体提及，表示为四元组，表示文档，和分别对应实体提及在文档中的起止下标，表示实体提及所属预定义类别。分别从两个层面进行评价。

* 1. **严格指标**

 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。由此得到严格评测指标：

, ,

* 1. **松弛指标**

 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。由此得到松弛评测指标：

, ,

最后，按照预定义类别的5个不同类别，对每个子类进行分开评测，共得到12个评测结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 解剖部位 | 症状描述 | 独立症状 | 药品 | 手术 | 综合 |
| 严格指标 |  |  |  |  |  |  |
| 松弛指标 |  |  |  |  |  |  |

1. **任务提交指南**

每一个参赛队需提交的材料如下：

1. 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 2018官网）

结果文件用“参赛队名\_result.txt”命名，文件格式另行通知。代码及其文档需打包成一个zip文件，用code.zip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，且所得结果文件相一致。方法描述文档用CNER.pdf命名，包含算法描述及参数设置，pdf格式存储，页数不超过5页。

1. **任务交流平台**

本任务的交流平台为：

ccks2018-cner@googlegroups.com

所有报名参赛的参赛队至少要有一名成员加入到该讨论组中，后续所有有关本任务评测及数据的相关说明和通知将只在讨论组中发布和交流，不再另行发邮件或更新官网通知，请所有参赛队务必在报名成功之后加入该讨论组。

1. **时间安排**

与官网的时间表不一致，以本任务书中的时间表为准，后续如果有调整，也将在谷歌讨论组中更新说明。

* 2018年4月3日，发布任务描述
* 2018年4月3日~7月15日，参赛队伍报名
* 2018年5月11日~5月15日 签署数据使用承诺书
* 2018年5月16日，发布训练数据初版，提交讨论组讨论
* 2018年6月11日，修定训练标注数据，提交最终版
* 2018年7月15日，发布测试数据
* 2018年7月20日，提交测试结果
* 2018年8月5日，提交评测论文

任务负责人：

张江涛 清华大学知识工程实验室 zhang-jt13@mails.tsinghua.edu.cn

焦增涛 医渡云（北京）技术有限公司zengtao.jiao@yiducloud.cn