**面向中文电子病历的医疗实体识别及属性抽取**

本任务是CCKS围绕中文电子病历语义化开展的系列评测的一个延续，在CCKS 2017，2018医疗命名实体识别评测任务的基础上进行了延伸和拓展。

包括两个子任务：1）**医疗命名实体识别**：由于国内没有公开可获得的面向中文电子病历医疗实体识别数据集，本年度保留了医疗命名实体识别任务，对2017年度数据集做了修订，并随任务一同发布。2）**医疗实体及属性抽取（跨院迁移）**：在医疗实体识别的基础上，对预定义实体属性进行抽取。本任务为迁移学习任务，即在只提供目标场景少量标注数据的情况下，通过其他场景的标注数据及非标注数据进行目标场景的识别任务。

参赛队可同时选择两个子任务参赛，也可选择任一一个子任务单独参赛。

**一、医疗命名实体识别**

1. **任务定义及描述**

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），比如疾病、治疗、检查检验等。

* 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. **预定义类别**

预定义类别定义如下：

**1) 疾病和诊断**：医学上定义的疾病和医生在临床工作中对病因、病生理、分型分期等所作的判断。

**2) 检查：** 影像检查（X线、CT、MR、PETCT等）+造影+超声+心电图，未避免检查操作与手术操作过多冲突，不包含此外其它的诊断性操作，如胃镜、肠镜等。

**3）检验**： 在实验室进行的物理或化学检查，本期特指临床工作中检验科进行的化验，不含免疫组化等广义实验室检查

**4) 手术**： 医生在患者身体局部进行的切除、缝合等治疗，是外科的主要治疗方法。

**5) 药物**： 用于疾病治疗的具体化学物质。

**6) 解剖部位**： 指疾病、症状和体征发生的人体解剖学部位。

1. **数据集描述**

数据集由医渡云（北京）技术有限公司编写，并由医渡云公司组织专业的医学团队进行人工标注，仅限CCKS竞赛评测用。

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. **评价指标**

采用精确率（Precision）、召回率（Recall）以及F1-Measure作为评测指标。并按照预定义类别的5个不同类别，对每个子类进行分开评测。

**二、****医疗实体及属性抽取（跨院迁移）**

1. **任务定义及描述**

给定的一组电子病历纯文本文档，定义若干与医疗相关的目标字段，如肿瘤大小，肿瘤原发部位等，任务的目标是识别并抽取目标字段的答案实体，比如左肺，肺叶等。

本任务为实体及属性抽取问题的跨院迁移任务，即在只提供需要识别的场景的少量标注数据的情况下，提供大量来自其他场景的已知数据分布不同的标注数据，以及大量非标注数据，进行目标场景的识别任务。

* 1. **问题实例**



* 1. **实体类别定义**

结合数据源“癌症医疗影像检查与结论”的内容及特点，本任务字段类型聚焦在**癌症原发部位，病灶大小，**以及**癌症转移部位**三大类。每个文本的一个字段可能出现多个或0个目标实体，比如多个原发部位。

**1) 原发部位：**某种疾病最先发生于的组织或者器官，如肺癌原发与左肺上叶；

**2) 病灶大小：**原发部位的大小，通常以最大直径或者大小直径表示；

**3) 转移部位：**某种疾病从最先发生的组织或者器官转移到的其他组织或器官；

1. **数据集描述**

本次评测的训练数据将分为三部分(具体数量待定)：

1. 900条非目标场景的标注数据；
2. 100条目标场景的标注数据；
3. 1000条各个场景的非标注数据。

此外，本评测将使用400条目标场景的标注数据作为最终评测的测试集。有关数据集的详细描述也将随数据的发布一并加以说明。

1. **评价指标**

由于每个文本的一个目标字段可能出现多个实体，评测指标使用实体而非字段来计算准召率，最终使用实体的f1值作为评测指标。

**三、任务提交指南**

每一个参赛队需提交的材料如下（分子任务提交）：

1. **结果文件**
2. **代码**
3. **方法描述文档**（非评测论文，评测论文撰写要求见CCKS 2019官网）

结果文件用“参赛队名\_result.txt”命名，文件具体格式另行通知。代码需打包成zip文件，用“参赛队名\_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，且运行结果与结果文件相一致。方法描述文档用“参赛队名\_method.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。

**四、任务交流平台**

本任务的交流平台为：

[CCKS2019-clinic@googlegroups.com](mailto:CCKS2019-clinic@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。

加入讨论组时请务必注明所属参赛队队名及所属单位，加入讨论组有问题的参赛队请联系：[zhang-jt13@tsinghua.org.cn](mailto:zhang-jt13@tsinghua.org.cn)。

**五、时间安排**

时间安排初定如下，后续如果有调整，将在讨论组中更新说明。

* 评测任务发布： 4月1日
* 报名时间：4月1日—7月10日
* 训练数据发布(第一批)：5月6日
* 训练数据发布(第二批)：5月31日
* 测试数据发布：7月20日
* 提交测试结果：7月30日
* 评测论文提交：8月15日
* CCKS会议日期：8月24日—27日

任务组织者：

张江涛 清华大学知识工程实验室

李文琪，医渡云（北京）技术有限公司

汤步洲 哈尔滨工业大学（深圳）

任务联系人：

张江涛，zhang-jt13@tsinghua.org.cn

陈泽曦，zexi.chen@yiducloud.cn