**CCKS 2020 技术评测任务书**

# 面向中文电子病历的医疗实体及事件抽取

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, 2019相关评测任务的基础上进行了延伸和拓展。

本任务包括两个子任务：

1）**医疗命名实体识别**

2）**医疗事件抽取**

参赛队可同时选择两个子任务参赛，也可选择任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**

本任务为中文病历医疗实体识别任务，即对于给定的一组电子病历纯文本文档，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。

相比2019命名实体识别任务我们做了如下调整：

1. 标注了新数据对训练数据进行扩充
2. 提供实体词表及大量非标注数据供参赛者使用
   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**
   1. **数据标注说明**

词表及电子病历数据由医渡云（北京）技术有限公司编写，标注数据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. **数据示例**

**{**

**"originalText": "患者3月前因“直肠癌”于在我院于全麻上行直肠癌根治术（DIXON术），手术过程顺利，术后给予抗感染及营养支持治疗，患者恢复好，切口愈合良好。，术后病理示：直肠腺癌（中低度分化），浸润溃疡型，面积3.5\*2CM，侵达外膜。双端切线另送“近端”、“远端”及环周底部切除面未查见癌。肠壁一站（10个）、中间组（8个）淋巴结未查见癌。，免疫组化染色示：ERCC1弥漫（+）、TS少部分弱（+）、SYN（-）、CGA（-）。术后查无化疗禁忌后给予3周期化疗，，方案为：奥沙利铂150MG D1，亚叶酸钙0.3G+替加氟1.0G D2-D6，同时给与升白细胞、护肝、止吐、免疫增强治疗，患者副反应轻。院外期间患者一般情况好，无恶心，无腹痛腹胀胀不适，无现患者为行复查及化疗再次来院就诊，门诊以“直肠癌术后”收入院。"**

**"entities": [**

**{**

**"label\_type": "疾病和诊断",**

**"overlap": 0,**

**"start\_pos": 8,**

**"end\_pos": 11**

**},**

**{**

**"label\_type": "手术",**

**"overlap": 0,**

**"start\_pos": 21,**

**"end\_pos": 35**

**},**

**{**

**"label\_type": "疾病和诊断",**

**"overlap": 0,**

**"start\_pos": 78,**

**"end\_pos": 95**

**}**

**]**

**}**

* 1. **数据集描述**

本次评测的训练数据有：

1. 1500条标注数据
2. 1000条非标注数据。
3. 6个类别的6292个实体词词表

标注数据集统计如下表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 文本 | 疾病和诊断 | 检查 | 检验 | 手术 | 药物 | 解剖部位 | 总数 |
| 训练集 | 1500 | 6211 | 1490 | 1885 | 1327 | 2841 | 12660 | 26414 |

1. **评价指标**

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。

3.1 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：

, ,

3.2 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6个不同类别，对每个子类进行分开评测，共得到14个评测结果：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 疾病和诊断 | 检查 | 检验 | 手术 | 药物 | 解剖部位 | 综合 |
| 严格指标 |  |  |  |  |  |  |  |
| 松弛指标 |  |  |  |  |  |  |  |

**二、****医疗事件抽取**

1. **任务定义及描述**

本任务为中文病历医疗事件抽取任务，即给定主实体为肿瘤的电子病历文本数据，定义肿瘤事件的若干属性，如肿瘤大小，肿瘤原发部位等，识别并抽取事件及属性，进行文本结构化。

本任务提供少量标注数据、大量非标注数据集及词表，旨在训练数据有限的情况下，利用非标注文本和半监督等方法提升模型性能。更接近真实世界的场景。

* 1. **事件模板定义**

**事件主实体：肿瘤**

**属性1：原发部位 【某种疾病最先发生于的组织或者器官】**

**属性2：病灶大小 【原发部位的大小】**

**属性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】**

每个文本的一个属性可能出现0个或多个属性实体，比如多个原发部位。

* 1. **认识示例**



1. **数据集描述**
   1. **数据标注说明**

电子病历数据由医渡云（北京）技术有限公司编写，标注数据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. **数据示例**

**原文：右肺癌化疗后，对比2016-11-29CT： 右上肺病变较前范围稍缩小，周边少许炎症较前稍减少。 两肺散在小结节，大致同前。 左侧锁骨下区、纵隔多发淋巴结，考虑转移，较前稍缩小。 肝囊肿。 左肾小囊肿。右肺癌化疗后，对比2016-11-29CT： 右肺上叶见不规则结节状、片状病灶，边界不清，最大层面大小约12mm×8mm，边缘呈分叶状，增强扫描不均匀强化，紧贴斜裂胸膜，部分范围较前略缩小，右上肺见少许斑片状稍高密度影，边界不清，较前明显减少。左下肺（se8，im96）、左上肺（se8，im221）及右下肺（se8，im104）散在数个小类结节，边界清，大者直径3mm，大致同前。 右肺上叶前内基底段支气管变窄，基底段支气管分支管壁增厚，气管及其余支气管分支通畅。 左侧锁骨下区、两下上气管旁、血管前、主动脉弓旁、主肺动脉窗、隆突上见多发淋巴结，部分相互融合，大者短径约5mm，轻度强化，较前稍缩小。 两侧胸腔未见积液，左侧胸膜未见明显增厚。心包未见明显积液。 肝脏形态正常，各叶比例在正常范围以内，其外形轮廓光整，肝内见多个低密度灶，边界清，大者直径约9mm，未见强化。 肝内胆管正常，其内未见结石影，胆囊大小正常，其内未见结石影，胆总管未见扩张，其内未见结石影。肝门区正常。门静脉未见异常。 脾大小正常，密度均匀。胰腺大小、形态正常，密度均匀。 左肾见低密度灶，边界清，大小约4mm×3mm，未见强化。 右肾及两肾下腺未见异常。 膈脚后、腹主动脉旁未见肿大淋巴结。 扫描范围未见骨质破坏征象。**

**肿瘤原发部位**：**右肺上叶**

**原发病灶大小： 12mm×8mm**

**转移部位**：**左侧锁骨下区、纵隔多发淋巴结**

* 1. **数据集描述**

本次评测的训练数据有：

1. 1400条标注数据
2. 1300条非标注数据。
3. 863个实体词表

标注数据集统计如下表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 文本 | 肿瘤原发部位 | 原发病灶大小 | 转移部位 | 总数 |
| 训练集 | 1400 | 1209 | 590 | 1013 | 2812 |

此外，本评测将使用300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事件属性可能出现多个属性实体，评测指标使用属性实体而非属性来计算准召率，最终使用属性实体的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 2020官网）

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：

[ccks2020-clinic@googlegroups.com](http://ccks2020-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、评测平台**

与往年一致，组委会依托Biendata（<https://www.biendata.com/>）作为在线评测平台，所有参赛队最终结果均需在Biendata上指定的时间窗内进行提交。本年度评测组织者还会在Biendata上发布一个baseline方案和实现。提供一个基准的性能，供参赛队伍参考。有关baseline的发布事宜，敬请关注CCKS 2020官网通知。

**六、奖金设置**

本任务奖金设置如下。

每个子任务第一名5k，第二名3k，第三名2k。

另设技术创新奖5k。

**七、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。

* 任务准备时间：3月1日—3月20日
* 评测任务发布：3月20日
* 报名时间：3月20日—7月10日
* 签署数据使用承诺书：3月20日-7月10日
* 训练及验证数据发布：3月20日
* 测试数据发布：7月10日
* 提交测试结果：7月20日
* 评测论文提交：8月5日
* CCKS会议日期(评测报告及颁奖)：8月15日—18日

**任务组织者**：

**张江涛 清华大学知识工程实验室**

**焦增涛，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**

**梁世浩，shihao.liang@yiducloud.cn**

**温清华，wtsinghua1@163.com**