CCKS2021中文地址任务

1. 大赛概况

人类的活动离不开位置，从空间上可以表征为坐标，从文本上表征为通讯地址。通讯地

址广泛存在于电商物流、政府登记、金融交通等领域。对通讯地址的分析、聚合服务已

经是一项重要基础服务，支撑着诸多互联网场景，比如地图搜索、电商物流分析等。实

际应用中，地址文本存在写法自由、缺省别名多、地域性强等特点，对地址的解析、归

一和匹配等都造成困难。针对这些难点， CCKS2021大会举办此次地址评测任务。该评测包含2个子任务，分别是:中文地址要素解析、地址文本相关性。参赛队伍可以任选其一参加，也可以参加两个任务。

1. 赛程安排

本次大赛分为报名组队、初赛、复赛和决赛三个阶段，具体安排和要求如下：

报名组队——————5月17日—6月25日

初赛阶段——————5月24日—7月2日

复赛阶段——————7月12日—8月13日

决赛答辩——————10月下旬

报名组队与实名认证（2021年5月17日—6月25日）

报名方式：5月17日阿里天池平台(<https://tianchi.aliyun.com/>)将开放本次比赛的组队报名、登录比赛官网，完成个人信息注册，即可报名参赛；

选手可以单人参赛，也可以组队参赛。组队参赛的每个团队2-3人，每位选手只能加入一支队伍；

选手需确保报名信息准确有效，组委会有权取消不符合条件队伍的参赛资格及奖励；

选手报名、组队变更等操作截止时间为6月25日23：59：59；

各队伍（包括队长及全体队伍成员）需要在6月25日23：59：59前完成实名认证（认证入口：天池官网-右上角个人中心-认证-支付宝实名认证），**未完成认证的参赛团队将无法进行后续的比赛；**

大赛官方钉钉群请搜索Group Number：32039787，或扫描以下二维码加入，最新通知将会第一时间在群内同步：


初赛阶段（2021年5月24日-2021年7月2日，UTC+8）

初赛的几个关键时间点： 5月24号天池平台将开放竞赛数据集和系统测评。

选手报名成功后，参赛队伍通过天池平台下载数据，本地调试算法，在线提交结果。

初赛提供训练数据集，供参赛选手训练算法模型；同时提供测试数据集，供参赛选手提交评测结果,参与排名。

初赛时间为2021年5月24日-2021年7月2日，系统每天提供2次评测机会，系统进行实时评测并返回成绩，排行榜每小时进行更新，按照评测指标从高到低排序。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示。

初赛淘汰：2021年7月2日上午9：59：59，**初赛阶段未产出成绩的队伍将被取消复赛参赛资格。**

初赛结束，初赛排名前100名的参赛队伍将进入复赛，复赛名单将在7月9日21：59：59之前公布。

复赛阶段（2021年7月12日—2021年8月13日，UTC+8）

复赛阶段测试数据不可下载，采用docker镜像的提交方式，“容器镜像”提交说明在复赛前公布。第一次接触docker可直达教程(链接https://tianchi.aliyun.com/competition/entrance/231759/tab/174?spm=5176.12281976.0.0.4d0d5330fADbWw)。

本阶段，系统每天提供2次实时评测，每小时更新排行榜，按照评测指标从高到低排序。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示，但最终入围决赛的排行榜则是根据参赛团队在复赛最后一周(8月9日-8月13日)提交的历史最优成绩而决定。复赛提交截止时间8月13日中午11：59：59。

本阶段内，选手需保证最后提交的是最优模型对应的完整端到端代码（包含数据处理和模型训练等）并能运行复现最优成绩。复赛结束后，该阶段最优成绩对应提交的镜像将直接用于代码审核，如最优成绩对应的镜像代码不是完整代码运行得出，将会直接淘汰，因此如果最后阶段出现无法复现的最优成绩可在复赛提交结束前联系组委会协助删除最优记录，复赛结束后不再受理。

榜单将在复赛截止后公布。复赛结束后，组委会将对排行榜TOP 20参赛队伍进行最优提交成绩的模型和完整代码审核，该阶段最优成绩对应提交的镜像将直接用于代码审核，选手需保证提交最优模型对应完整端到端代码（包含数据处理和模型训练等）且能运行复现最优成绩，不接受随机成绩。如最优成绩对应的镜像代码不是完整代码运行得出，将会直接淘汰。对于未提交、复现未成功或审核不通过的队伍，将取消决赛资格和比赛奖励。

最终审核通过的前10名参赛队伍晋级决赛。

决赛答辩（暂定10月）

入围线下决赛答辩名单通知时间为8月31日23：59：59前。

入围团队需要在10月1日17：59：59前需要提交答辩PPT，并在现在线下决赛前一天参与决赛彩排完成设备调试。线下决赛具体时间将在复赛结束后公布。

决赛评分参考：复赛榜单、代码质量和答辩。

答辩需要准备答辩材料，包括答辩PPT(中英文均可)、参赛总结、算法核心。本次赛事决赛入围团队的最终得分将由复赛成绩、决赛答辩成绩加权得出，评分权重为：复赛成绩占80%，决赛答辩成绩占20%。

另：本次赛事若要求论文环节，则将以CCKS 2021的通知要求为准。

1. 参赛对象

大赛面向全社会开放，个人、高等院校、科研单位、企业、创客团队等人员均可报名参赛。

阿里巴巴集团、蚂蚁金服、菜鸟等BU员工均可报名参赛，规则如下：

阿里系员工报名参赛，可参与排名，不得领取奖金，但有资格领取纪念奖；

选手完成阿里云账号注册，并在天池个人中心页面内邮箱改为阿里巴巴员工邮箱，组织单位体现所在BU（“天池平台-个人中心-所在机构”）；

队伍内有一位阿里系员工（包括非正式员工），即属于阿里巴巴集团内部参赛团队。

\*注：大赛主办和技术支持单位如有机会接触赛题背景业务、产品、数据的员工，则自动退出比赛，放弃参赛资格。

1. 奖项设置

每个子任务总奖金共5万元，奖金设置如下：
第一名：两万元人民币
第二名：一万两千元人民币
第三名：八千元人民币
技术创新奖(2名)：五千元人民币

1. 大赛组织

组织机构：

英特尔

CCF自然语言处理专委会

阿里云天池平台

任务组织者：

谢朋峻

王潇斌

丁瑞雪

陈漠沙（阿里云-天池平台）

张梅山（CCF自然语言处理专委会）

任务联系人：

王潇斌（czwangxiaobin@foxmail.com）

丁瑞雪(ruixue\_ding@buaa.edu.cn)

任务沟通钉钉群：32039787

1. 任务一：中文地址要素解析
	1. 背景

地址是日常生活中一种重要的文本信息，诸多场景需要登记地址，如电商购物、外卖配送、人口普查、水电气开户等。常见的地址一般包含以下几类信息：

行政区划信息，如省、市、县、乡镇信息;

路网信息，如路名，路号，道路设施等;

详细地址信息，如POI (兴趣点)、楼栋号、户室号等;

非地址信息，如补充说明，误输入等;

地址要素解析是将地址文本拆分成独立语义的要素，并对这些要素进行类型识别的过程。地址要素解析与地址相关性共同构成了中文地址处理两大核心任务，具有很大的商业价值。目前中文地址领域缺少标准的评测和数据集，这次我们将开放较大规模的标注语料，希望和社区共同推动地址文本处理领域的发展。

* 1. 赛题描述

中文地址要素解析任务的目标即将一条地址分解为上述几个部分的详细标签，如：

**输入**：浙江省杭州市余杭区五常街道文一西路969号淘宝城5号楼，放前台

**输出**：Province=浙江省 city=杭州市 district=余杭区 town=五常街道 road=文一西路road\_number=969号 poi=淘宝城 house\_number=5号楼 other=，放前台

(详细的标签体系及标注规范将随标注数据一起公布)

* 1. 数据说明

标注数据集由训练集、验证集和测试集组成，整体标注数据大约2万条左右。地址数据通过抓取公开的地址信息（如黄页网站等）获得， 均通过众包标注生成，详细标注规范将会在数据发布时一并给出。

* 1. 提交说明

选手需要提交测试数据的预测结果，文件命名为：队伍名\_addr\_parsing\_runid.txt, 文件编码采用utf-8编码（无BOM头）。

结果文件分为3列，列分隔符为不可见字符，\u0001, 第一列、第二列与测试文件对应，分别为数据id, 地址原文。第三列为系统预测结果，使用BIEO标签体系，标签与类型采用“-”分隔，tag之间采用空格分隔。示例如下：

1^A浙江杭州阿里^AB-prov E-prov B-city E-city B-poi E-poi

* 1. 评估标准

我们通过将输出结果与人工标注的集合进行比较来分别计算每一种元素准确率(Precision)，召回率(Recall)和F-1分值(F-1 score)，并采用Micro-F1作为最终排名指标。

具体计算过程如下所示（以元素T为例）：

给定一条地址输入（n），此地址中有x个T元素：$G\_{n}=\{g\_{1},g\_{2},g\_{3}…g\_{x}\}$ ，系统输出标注结果中有y个T元素如下：$P\_{n}=\{p\_{1},p\_{2},p\_{3}…p\_{y}\}$ ，则实体标注的准确率定义如下：

$$P=\frac{\sum\_{nϵN}^{}|G\_{n}∩P\_{n}|}{\sum\_{nϵN}^{}|P\_{n}|}$$

实体标注的召回率定义如下：

$$R=\frac{\sum\_{nϵN}^{}|G\_{n}∩P\_{n}|}{\sum\_{nϵN}^{}|G\_{n}|}$$

实体标注的F1值定义如下：

$$F1= \frac{2\*P\*R}{P+R}$$

其中N为整个测试集。

1. 任务二：地址相关性
	1. 背景

地址文本相关性任务在现实世界中存在着广泛的应用场景，如：基于地理信息搜索的地理位置服务、对于突发事件位置信息的快速搜索定位、不同地址信息系统的对齐等等。

日常生活中输入的地址文本可以为以下几种形式：

包含四级行政区划及路名路号POI的规范地址文本；

地址要素缺省的规范地址文本，例：只有路名+路号、只有POI；

非规范的地址文本、口语化的地址信息描述，例：阿里西溪园区东门旁亲橙里；

地址文本相关性主要是衡量地址间的相似程度，地址要素解析与地址相关性共同构成了中文地址处理两大核心任务，具有很大的商业价值。目前中文地址领域缺少标准的评测和数据集，这次我们将开放较大规模的标注语料，希望和社区共同推动地址文本处理领域的发展。

* 1. 赛题描述

本评测任务为基于地址文本的相关性任务。即对于给定的一个地址query以及若干个候选地址文本，参赛系统需要对query与候选地址文本的匹配程度进行打分。

多样化的地址文本写法对地址文本的相关性任务提出的挑战如下：

同一个地址存在多种写法，没有给定的改写词表；

地址query一般存在省市区等限制条件，需要结合限制条件分析相关性；

不同地市地址规范不一，对模型泛化性提出更高要求；

* 1. 数据说明

输入：输入文件包含若干个query-地址文本对

输出：输出文本每一行包括此query-地址文本对的匹配程度，分为完全匹配、部分匹配、不匹配

示例：

 输入：

Query：江苏省南京市清水亭东路9号金域蓝湾15幢

Doc：江宁区万科金域蓝湾15栋

…………

 输出：

 完全匹配

…………

说明：

1. 完全匹配：完全匹配一般是匹配到楼栋号，若原始数据只到poi则匹配到poi

例子1：

输入：

Query：江苏省南京市清水亭东路9号金域蓝湾15幢

Doc：江宁区万科金域蓝湾15栋

输出：

完全匹配

1. 部分匹配：原始地址给到了楼栋号，备选地址中只能匹配到poi，则为部分匹配

例子2：

输入：

Query：江苏省南京市栖霞区西岗街道学森路199号保利罗兰春天13幢二单元

Doc：仙林湖学森路199号保利罗兰春天9号

输出：

部分匹配

\*\*注意： 村级别，如果是村+村组，匹配到了村，算作部分匹配；

同理，poi+subpoi，如果只匹配到了poi，算部分匹配；

村+poi，匹配到村是不匹配，匹配到poi是完全匹配；

村+门牌号+poi，匹配到村是不匹配，匹配到门牌号是部分匹配；

query为村，candidate为村+poi，匹配到村是部分匹配；

村+路+路号，匹配到村、路是部分匹配；

1. 不匹配：完全不匹配原文地址

例子2：

输入：

Query：辽宁省大连市甘井子区辛寨子街道天虹工业园区53号

Doc：新水泥路666号重工数控工业园

输出：

不匹配

标注数据集由训练集、验证集和测试集组成，整体标注数据大约5万条左右。

标注数据集中每条数据的格式为：

{

 "text\_id":"1",

 "query":"华侨村西堤1巷12栋",

 " candidate":[

 {

 "text":"华侨新村西堤一巷12号",

 "label":"部分匹配"

 },

 {

 "text":"宝安区华侨新村西堤一巷",

 "label":"部分匹配"

 },

 {

 "text":"海丰县米巷西12幢",

 "label":"不匹配"

 },

 {

 "text":"余姚市大施巷村西片12号楼",

 "label":"不匹配"

 },

 {

 "text":"中山市西堤路一巷",

 "label":"不匹配"

 }

 ]

}

* 1. 提交说明

选手需要提交测试数据的预测结果，文件命名为：队伍名\_ addr\_match\_runid.txt, 文件编码采用utf-8编码（无BOM头）。

文件中每一行为一个query对应所有候选的预测结果，以json格式输出，label键对应模型预测的匹配类型标签，其他字段与训练数据一致。具体格式示例如下:

{"text\_id": "0329e27a5a9c77b25c65d614b978492a", "query": "货站街77号院二单元", "candidate": [{"text": "货站街77-12号兴彭被装", "label": "部分匹配"}, {"text": "货站街77-10号郑交大型货运", "label": "部分匹配"}, {"text": "货站>街77-2号附近豫香居", "label": "部分匹配"}, {"text": "货站街77-2号天津狗不理包子(青年路店)", "label": "部分匹配"}, {"text": "货站街77-11号附近佳能被装", "label": "部分匹配"}]}

{"text\_id": "c80ca513f8473294f75f198e7f00235f", "query": "盛泽镇南麻工业区织庄路18号苏州佑铠科技有限公司", "candidate": [{"text": "织庄路688号附近苏州佑铠科技工业有限公司", "label": "完全匹配"}, {"text": "太湖新城镇康欣路28>号苏州佑祥精密五金科技有限公司", "label": "不匹配"}, {"text": "盛泽镇南麻开发区织庄路18号吴江利源精密整理有限公>司", "label": "部分匹配"}]}

* 1. 评估标准

我们可以看到对于给定的地址query查询，相关性系统需要对若干个候选答案指定匹配程度标签。我们通过将输出结果与人工标注的集合进行比较来计算macro F1。

给定匹配程度标签g，PREDg为被预测为g的query-地址文本对集合，GOLDg为被标注为g的query-地址文本对集合。

g标签的召回率定义为：

Rg=|intersection(PREDg,GOLDg)|/|GOLDg|

g标签的准确率定义为：

Pg=|intersection(PREDg,GOLDg)|/|PREDg|

g标签的F1定义为：

F1g=2\*Pg\*Rg/(Pg+Rg)

最终的macro F1为3个匹配程度标签F1的平均值，即：

macro F1 = (F1g1 + F1g2 + F1g3) / 3

1. 比赛规则
	1. 参赛队伍需要签署数据使用条款以获得数据使用授权。
	2. 本次比赛可以使用开源数据资源辅助训练，可以使用预训练模型优化效果，为保证比赛公平和结果可复现，使用非主办方提供的数据资源和模型资源必须确保是公开可用的资源，需要在提交文档或代码中详细声明资源出处或构建过程。
	3. 使用外部资源时确保得到资源拥有者的授权，参赛选手自己数据使用行为负责，一切因非主办方提供数据产生的法律风险由使用者自己承担。
	4. 禁止人工标注包括补充训练数据和标注测试数据。禁止使用相同功能的开源或收费软件对数据进行标注，包括用于预训练模型的训练数据。
	5. 违反以上规则将被取消参赛资格。