**CCKS 2020 技术评测任务书**

**面向中文短文本的实体链指任务**

# 任务描述

本评测任务围绕实体链指技术，结合其对应的AI智能应用需求，在CCKS 2019面向中文短文本的实体链指任务的基础上进行了拓展与改进，主要改进包括以下几部分：

（1）去掉实体识别，专注于中文短文本场景下的多歧义实体消歧技术；

（2）增加对新实体（NIL实体）的上位概念类型判断；

（3）对标注文本数据调整，增加多模任务场景下的文本源，同时调整了多歧义实体比例。

面向中文短文本的实体链指，简称EL（Entity Linking）。即对于给定的一个中文短文本（如搜索Query、微博、对话内容、文章/视频/图片的标题等），EL将其中的实体与给定知识库中对应的实体进行关联。

传统的实体链指任务主要针对长文本，长文本拥有丰富的上下文信息，能辅助实体进行歧义消解并完成实体链指，相比之下，针对中文短文本的实体链指存在很大的挑战，主要原因如下：

（1）口语化严重，导致实体歧义消解困难；

（2）短文本上下文语境不丰富，须对上下文语境进行精准理解；

（3）相比英文，中文由于语言自身的特点，在短文本的链指问题上更有挑战。

此次任务的输入输出定义如下：

**输入：**

 中文短文本以及该短文本中的实体集合。

**输出：**

 输出文本此中文短文本的实体链指结果。每个结果包含：实体mention、在中文短文本中的位置偏移、其在给定知识库中的id，如果为NIL情况，需要再给出实体的上位概念类型（封闭体系的概念详见附件）

**示例输入：**

{

 "text\_id":"1",

 "text":"《琅琊榜》海宴\_【原创小说|权谋小说】",

 "mention\_data":[

 {

 "mention":"琅琊榜",

 "offset":"1"

 },

 {

 "mention":"海宴",

 "offset":"5"

 },

 {

 "mention":"原创小说",

 "offset":"9"

 },

 {

 "mention":"权谋小说",

 "offset":"14"

 }

 ]

}

**示例输出：**

{

 "text\_id":"1",

 "text":"《琅琊榜》海宴\_【原创小说|权谋小说】",

 "mention\_data":[

 {

"kb\_id":"2135131",

 "mention":"琅琊榜",

 "offset":"1"

 },

 {

"kb\_id":"10572965",

 "mention":"海宴",

 "offset":"5"

 },

 {

"kb\_id":"215143",

 "mention":"原创小说",

 "offset":"9"

 },

 {

"kb\_id":" NIL\_Work ",

 "mention":"权谋小说",

 "offset":"14"

 }

 ]

}

说明：对于实体有歧义的查询，系统应该有能力来区分知识库中链接的候选实体中哪个实体为正确链指的实体结果。例如，知识库中有8个不同的实体都可能是『琅琊榜』的正确链指结果，因为知识库中的这8个实体都可以通过『琅琊榜』的字面表达查找到，但是我们在给定的上下文中（『海燕』、『原创小说』、『权谋小说』），有足够的信息去区分这些候选实体中，哪个才是应该被关联上的结果。

# 数据描述

## 2.1. 知识库

该任务知识库来自百度百科知识库。知识库中的每个实体都包含一个subject\_id(知识库id)，一个subject名称，实体的别名，对应的概念类型，以及与此实体相关的一系列二元组< predicate，object>（<属性，属性值>）信息形式。知识库中每行代表知识库的一条记录（一个实体信息），每条记录为json数据格式。

示例如下所示：

{

 "subject\_id":"1000131",

 "subject":"小王子",

 "alias":[

 "Le Petit Prince",

 "The Little Prince",

 "リトルプリンス 星の王子さまと私",

 "小王子"

 ],

 "type":[

 "Work"

 ],

 "data":[

 {

 "predicate":"外文名",

 "object":"Le Petit Prince"

 },

 {

 "predicate":"发行公司",

 "object":"派拉蒙影业"

 },

 {

 "predicate":"类型",

 "object":"奇幻"

 }

 ]

}

## 2.2 标准数据集

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10万条左右，数据均通过百度众包标注生成，详细标注质量将会在数据发布时一并给出。

标注数据集中每条数据的格式为：

{

 "text\_id":"1",

 "text":"《琅琊榜》海宴\_【原创小说|权谋小说】",

 "mention\_data":[

 {

"kb\_id":"2135131",

 "mention":"琅琊榜",

 "offset":"1"

 },

 {

"kb\_id":"10572965",

 "mention":"海宴",

 "offset":"5"

 },

 {

"kb\_id":"215143",

 "mention":"原创小说",

 "offset":"9"

 },

 {

"kb\_id":" NIL\_Work ",

 "mention":"权谋小说",

 "offset":"14"

 }

 ]

}

标注数据集主要来自于：真实的互联网网页标题数据、视频标题数据、搜索Query

标注文本对象的示例数据如下：

|  |
| --- |
| **标注文本对象** |
| 刘德华演过无间首第几部 |
| 小米枪战游戏不可以玩跳伞 |
| 圣杯传奇，「探宝者」圣杯传说，圣杯传奇的起源 |
| 佛歌100首经典歌李娜 |
| 杨洋郑爽再度携手《爱上你是我做过最好的事》 |

# 评价指标

我们可以看到对于给定的中文短文本查询，EL系统输出的结果中包含给定中文短文本查询中出现的所有命名实体的链接结果。我们通过将输出结果与人工标注的集合进行比较来计算准确率(Precision)，召回率(Recall)和F-1分值(F-1 score)。

具体计算过程如下所示：

给定短文本输入（用Text表示，其属于golden标注集），此Text中有N个mention：$M\_{n}=\{m\_{1},m\_{2},m\_{3}…m\_{n}\}$ ，每个mention链接到知识库的实体id为：$E\_{n}=\{e\_{1},e\_{2},e\_{3}…e\_{n}\}$ ，实体标注系统输出标注结果如下：$E'\_{n}=\{e'\_{1},e'\_{2},e'\_{3}…e'\_{n}\}$ ，则实体标注的准确率定义如下：

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

实体标注的召回率定义如下：

$$R=\frac{\sum\_{nϵN}^{}|E\_{n}∩E'\_{n}|}{\sum\_{nϵN}^{}|E\_{n}|}$$

实体标注的F1值定义如下：

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

特别注意，在计算评价指标时，对NIL实体的上位概念类型判断结果NIL\_Type与实体的关联id等价处理。

# 任务提交方式

本次任务将采取刷榜的方式，将会发布A榜和B榜两个测试集，A榜测试集用于日常刷榜和模型迭代，不计入最终比赛成绩，测试集A发布后，允许参赛队伍每天至多提交一次结果，文件命名为“\*.txt”，格式与任务描述中的示例输出相同，排名每十分钟更新一次。

测试集B为最终评估比赛成绩的有效评测集，待测试集B发布后，允许参赛队伍每天至多提交一次，总共至多提交七次结果。

**所有获奖团队需额外提交相关材料以供资格审查，要求如下**：

1. 测试集B结果文件，用result.txt命名（UTF-8格式）
2. 相关代码及说明
3. 方法描述文档（非评测论文，评测论文撰写要求见CCKS 2020官网）

将以上三个文件在任务提交截止日期前发送至邮箱ccks2020\_el@126.com。邮件的标题为：“CCKS2020-EL-参赛队名称”，例如“CCKS2020-EL-TeamKG队”。

代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），作为邮件附件传送，要求提交所有的程序代码及相关的配置说明，**程序应当可以运行且所得结果与result.txt相符**。如果方法使用了额外资源，要求说明并提供资源文件或地址。

本次评测将依托Biendata平台，请有意向的参赛队伍关注平台上的比赛列表，同时创建了微信群作为官方信息发布出口和答疑渠道，您可以扫描下方二维码添加小助手，由小助手邀请进入官方微信群。



# 奖励设置

除CCKS组委会设置的1,5000元奖金外，百度额外提供3,0000元奖金给获奖团队，奖金设置如下：

第一名：20,000元

第二名：10,000元

第三名：5,000元

创新奖：10,000元

另外，我们还准备了丰富的周级奖励礼品和精美的参赛T恤。

# 时间安排

评测任务发布：3月20日

报名时间：3月20日-9月10日 23:59

测试数据A提交：4月7日-9月14日 23:59

测试数据B（最终评测集）发布：9月15日 00:00

提交最终测试结果：9月15日-9月21日 23:59

评测排名及结果复审：9月22日-9月30日

评测论文提交截止日期：10月25日

CCKS会议日期(评测报告及颁奖)：11月12日-15日

# 组织者信息

实体链指组委会邮箱：ccks2020\_el@126.com

汪琦：wangqi31@baidu.com

冯知凡：fengzhifan@baidu.com

柴春光：chaichunguang@baidu.com

# 附录

**NIL实体的上位类型集合**

|  |  |  |
| --- | --- | --- |
| 概念体系 | 类型定义 | 说明 |
| Event | 事件活动 | 突发事件,政治事件,历史事件,网络事件,社会事件,颁奖活动,案件,事故,仪式,慈善活动,军事活动,体育赛事,娱乐活动,展出活动,学术活动,宗教活动,考试等... |
| Person | 人物 | 娱乐人物,政治人物,经济人物,军事人物,医护人物,专家学者,体育人物,文化人物,宗教人物,艺术人物,历史人物,教育人物,新闻人物,法律人物,虚拟人物... |
| Work | 作品 | 著作,动漫作品,栏目,文学作品,音乐作品,美术作品,舞蹈作品,摄影作品,戏剧作品,工艺品,剧场表演,节目,系列作品,诗词古文,电影,电视剧，电视节目（综艺等）... |
| Location | 区域场所 | 行政区划,景点,建筑物,POI,地名等... |
| Time&Calendar | 时间历法 | 节日,日期,时代和历法等... |
| Brand | 品牌 | 服饰品牌,电脑品牌,橱柜品牌,护肤品品牌,汽车品牌,饮料品牌,皮具品牌,家具品牌,白酒品牌,手表品牌,厨具品牌,珠宝品牌,运动品牌,建材品牌,零食品牌,金融品牌等... |
| Natural&Geography | 自然地理 | 生命现象,化学现象,物理现象,天文现象,天气现象,山川,河流,湖泊等... |
| Game | 游戏 | 手机游戏,网络游戏,线下游戏等各种游戏 |
| Biological | 生物 | 动物,微生物,真菌,植物 |
| Medicine | 药物 | 药物,药品 |
| Food | 食物 | 饮品,食材,菜肴,乳类,甜点,小吃,零食 |
| Software | 软件 | 手机软件,游戏软件,应用软件等 |
| Vehicle | 车辆 | 汽车,卡车,救护车,摩托车,以及各种车辆和车辆相关的... |
| Website | 网站平台 | 平台网站,社区网站,行业门户等 |
| Disease&Symptom | 疾病症状 | 疾病,症状 |
| Organization | 组织机构 | 组织团体,企业,政府机构,医疗机构,金融机构,传媒机构,教育机构,法律服务机构，科研机构，医疗机构等... |
| Awards | 奖项 | 科技奖项,电影奖项,音乐奖项,文学奖等... |
| Education | 教育 | k12教育,职业教育,成人教育,考研,留学,在线教育,职业教育等... |
| Culture | 文化 | 专业术语,成语,习俗,民俗,文艺,艺术 |
| Constellation | 星座 | 天文学和占星学以及其他星座相关的... |
| Law&Regulation | 法律法规 | 法律,法规,规章,制度,协议,条约等... |
| VirtualThings | 虚拟事物 | 虚拟场景,虚拟场所,虚拟物品 |
| Diagnosis&Treatment | 诊断治疗方法 | 手术,疾病的诊断等... |
| Other | 其他 | 无法分类的其他类别 |