**任务：面向中文短文本的实体链指任务**

# 任务描述

本评测任务为面向中文短文本的实体识别与链指，简称ERL（Entity Recognition and Linking）。即对于给定的一个中文短文本（如搜索Query、微博、用户对话内容、文章标题等），ERL系统会识别其中的实体，并与给定知识库中的对应实体进行关联。ERL任务过程中需要进行实体识别等子任务，这些子任务的训练可以使用额外的资源。

传统的实体链指任务主要是针对长文档，长文档拥有在写的上下文信息能辅助实体的歧义消解并完成链指。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。

**输入：**

 输入文件包括若干行中文短文本。

**输出：**

 输出文本每一行包括此中文短文本的实体识别与链指结果。每个结果（结果用\2分隔）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，多个结果间用\1分隔开。

**示例：**

 输入：

李娜温网冠军

 输出：

 李娜温网冠军 \1 李娜 \2 1 \2 5285 \1 温网 \2 3 \2 336310 \1 冠军 \2 5 \2 32330

 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48个不同的实体都可能是『李娜』的正确链指结果，因为知识库中的这48个实体都可以通过『李娜』的字面表达查找到，但是我们在给定的上下文中（『温网』、『冠军』），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

该任务的参考知识库包括来自百度百科知识库的约39万个实体。知识库中的每个实体都包含一个KB-ID，一个字符串名称，及与此实体相关的一系列三元组<subject，predicate，object>信息形式。知识库中每行代表知识库的一条记录，每条记录的格式为一个json格式。predicate-id和object-id的值都为0，subject-id的值为一个正整数，说明subject总是对应知识库中的一个实体。

例如，

{

 "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": ["Movie"],

 "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"

 }

 ]

}

知识库数据分布如下:

|  |  |
| --- | --- |
| **类型** | **数量** |
| 实体数量 | 398,082 |
| SPO数量 | 3,564,565 |
| 实体描述数据 | 361,778 |
| 实体平均属性数量 | 9 |
| 实体描述数据平均长度 | 103 |

## 2.2 标准数据集

标注数据集由训练集、开发集和评估集组成。其中训练集中包括9万条短文本标注数据，开发集和评估集分别包括1万条短文本数据，3份数据均通过百度众包标注生成（人工进行评估，其平均准确率约为93%，召回率94%）。

标注数据集中每条数据的格式为：

{

 "text\_id": "1",

 "text": "李娜温网冠军",

 "mention\_data": [

 {

 "mention": "李娜",

 "offset": 0,

 "kb\_id": "5285"

 },

 {

 "mention": "温网",

 "offset": 2,

 "kb\_id": "336310"

 },

 {

 "mention": "冠军",

 "offset": 4,

 "kb\_id": "32330"

 }

 ]

}

标注数据集主要来自于：真实的互联网网页标题数据，是用户检索Query对应的有展现及点击的网页，短文本平均长度为21.73中文字符，覆盖了不同领域的实体（包括各垂类的实例、概念），如人物、电影、电视、小说、软件、组织机构、事件等垂类，以及通用概念。

<br>

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 |
| 年轻的母亲1最初完整版 |
| 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准集合S进行比较来计算准确率(Precision)，召回率(Recall)和F-1分值(F-1 score)。

评估标准如下：

详细说明如下，给定短文本输入（用Query表示，其属于golden标注集），此Query中有N个实体mention：，每个实体mention链接到知识库的实体id为：，实体标注系统输出标注结果如下：，则实体标注的准确率定义如下：



实体标注的召回率定义如下：



实体标注的F1值定义如下：



# 任务提交

本次任务将采取刷榜的方式，验证集发布后，允许参赛队伍多次向平台提交结果，文件命名为“参赛队名称\_valid\_result.txt”，格式与任务描述中的示例输出相同（注意保留<>或""），排名随时更新。参赛队伍可在评测集发布之前随时上传验证集的计算结果，管理系统会及时更新各队伍的最新排名情况；

测试集发布后，允许参赛队伍多次提交测试集结果文件。

**最终提交文件要求：**每一个参赛队需提交的材料如下。

1. 测试集结果文件，用result.txt命名（UTF-8格式），格式如下：

例如给定查询：”李娜温网冠军”，查询id为1234。一个由Team ABC提交的第一个评估结果中的一组解释为：

{

 "team-tag": "ABC",

 "submit-tag": "S1",

 "text\_id": "1234",

 "text": "李娜温网冠军",

 "mention\_data": [

 {

 "mention": "李娜",

 "offset": 0,

 "kb\_id": "5285",

 "confidence": 0.806850

 },

 {

 "mention": "温网",

 "offset": 2,

 "kb\_id": "336310",

 "confidence": 0.783856

 },

 {

 "mention": "冠军",

 "offset": 4,

 "kb\_id": "32330",

 "confidence": 0.579471

 }

 ]

}

offset偏移计算：基于utf8数据格式计算文本offset，中文、英文及符号均计算为1个字符，offset初始值为0。

1. 相关代码及说明
2. 方法描述文档（非评测论文，评测论文撰写要求见CCKS 2019官网）

以上三个文件需在任务提交截止日期前发送至邮箱ccks2019\_erl@163.com。邮件的标题为：“CCKS-ERL-参赛队名称”，例如“CCKS-ERL-火箭队”。

代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），用code.xxx命名，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（<https://biendata.com/competition/ccks_2019_el>）展开，请有意向的参赛队伍关注平台上的竞赛列表。

# 时间安排

评测任务发布：4月1日

报名时间：4月1日 - 7月10日

训练及验证数据发布：4月20日

测试数据发布：7月20日

提交测试结果：7月25日

评测论文提交：8月15日

# 组织者

汪琦： ccks2019\_erl@163.com

冯知凡：fengzhifan@baidu.com

张扬： zhangyang08@baidu.com