## 面向上市公司主营业务的实体链接任务评测

# 任务描述

实体链接(Entity Linking)，也叫实体链指，需要从非结构化数据中将预先定义好的实体类型（产品、人名、机构、地名等）识别出来，获得表示实体的词语(即所谓mention，对某个实体的指称项)，并从另一个包含若干实体的知识库(领域词库，知识图谱等)中找到mention所表示的正确实体。

本次任务主要针对上市公司的主营业务进行产品实体链接。需要获得主营业务中的产品实体，将该实体链接到产品数据库中的某一个产品实体。产品数据库将发布在竞赛平台上。比如某一公司主营业务为“主要生产日用居家小家电，生活零售用品等相关产品”，获得“日用居家小家电”这一产品实体，称为主实体；通过实体链接技术，找到产品数据库中的“生活小件家电”这一产品实体，称为链接实体。主实体与链接实体构成一个链接实体对，表示这两个实体是不同名称的相同实体。通过这些链接实体对，从而实现词语消歧以及数据源的融合。

参赛队伍需要能够准确的从公司主营业务中识别出所有产品实体，即主实体，并且需要确定每个主实体在产品数据库中是否存在链接，若存在则需要在产品数据库中找到链接实体，形成链接实体对，完成实体链接。

输入输出如下：

**输入**：一段文本T

**输出**：文本T中所有的产品实体以及每个产品实体对应的链接实体（链接实体可为空）

**样例**：

*输入：{“text\_id”:”123456”, “text”*:“主要生产日用居家小家电，生活零售用品等相关产品”}

*输出：{“text\_id”:”123456”,”EntityLink”:[(“日用居家小家电”,“生活小件电器”)),(“生活零售用品”,”日常零售用品”)]}*

# 数据描述

本次数据主要来自国内上市公司的主营业务，以及经过处理的产品数据库。样本包含正样本和负样本，训练集、验证集及测试集的说明如下：

**训练集&验证集：**

在训练及验证数据发布阶段，我们会发布？左右的文本以及每条文本的标注结果（结果中包含真实的产品实体以及每个产品实体的链接实体）。？不包含标注的文本作为验证集，选手可以将生成的验证集答案提交进行评测。在测试数据发布阶段同时发布验证集的标注结果。

**测试集：**

在测试数据发布阶段，我们将会再发布？左右的文本数据集以及一定数量的干扰数据，不含标注结果，作为测试。干扰数据指不包含真实产品实体的一段文本。

# 评价指标

**任务评价指标：**

本次任务采用精确率（Precision, P）、召回率（Recall, R）、F1值（F1-measure, F1）来评估产品实体链接的效果，使用F1作为最终评价指标。

# 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”，格式与结果文件result.txt相同，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（每日最多可上传3次），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结果文件，结果文件提交格式如下。

**结果文件格式说明：**

选手将结果保存为result.txt，以utf-8编码格式保存。使用每项结果的链接实体对和text\_id对应的标准答案作为计算评测指标。最终提交文件要求：

每一个参赛队需提交的材料如下：

1. 结果文件
2. 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 2023官网）

以上三个文件需在任务提交截止日期前发送至邮箱：

发送至[ProduceEntityAlignment@gmail.com](mailto:ProduceEntityAlignment@gmail.com)。邮件的标题为：“CCKS-评测任务EL -最终提交文件-参赛队名称”，例如“CCKS-评测任务EL-最终提交文件-飞牛队”。

结果文件用result.txt命名，采用utf-8格式存储，文件格式需要与训练数据中的标注结果文件完全一样，结果可用json解析。

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，确保程序能够正确运行，且所得结果与result.txt相符。

方法描述文档用SM.pdf命名，包含算法描述及参数设置，需用pdf格式存储，页数不超过15页。

评测平台：本次评测将依托biendata平台（https://biendata.com/）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

# 规则

1. 所有参赛选手都必须在biendata平台管理系统中注册，本次比赛的参赛对象仅限全日制在校大学生（本科、硕士、博士均可）和企业员工；

2. 参赛选手需确保注册时提交信息准确有效，所有的比赛资格及奖金支付均以提交信息为准；

3. 参赛选手在管理系统中组队，参赛队伍成员数量不得超过5个，报名截止日期之后不允许更改队员名单；

4. 每支队伍需指定一名队长，队伍名称不超过15个字符，队伍名的设定不得违反中国法律法规或公序良俗词汇，否则组织者有可能会解散队伍；

5. 每名选手只能参加一支队伍，一旦发现某选手以注册多个账号的方式参加多支队伍，将取消相关队伍的参赛资格；

6. 允许使用开源代码或工具，但不允许使用任何未公开发布或需要授权的代码或工具；

7. 除主办方提供的数据集外，参赛选手允许使用预训练数据（如词向量、字向量等），但是不能用额外的标注数据。

8. 参赛队伍可在参赛期间随时上传验证集的预测结果，一天不能超过3次 ，管理系统会实时更新各队伍的最新排名情况。

# 奖励设置

本任务设置一、二、三等奖和技术创新奖，均获得中国中文信息学会颁发的获奖证书，并有机会录用EI检索的评测论文。

# 任务组织者和联系人

**任务组织者：**

刘莹（东北大学）

**任务联系人：**

陈宇，电话：13712196871

# 任务交流平台

QQ群：496272821