**CCKS 2020 技术评测任务书**

# 面向金融领域的篇章级事件主体与要素抽取

# 任务描述

“事件抽取”是舆情监控领域和金融领域的重要任务之一，“事件”在金融领域是投资分析，资产管理的重要决策参考；事件也是知识图谱的重要组成部分，事件抽取是进行图谱推理、事件分析的必要过程。“事件抽取”的挑战体现在文本的复杂和任务的复杂。文本的复杂体现在事件抽取的输入文本可能是句子、段落或者篇章，不定长度的文本使得限制文本长度的模型无法使用；任务的复杂体现在事件识别的任务包括：事件类型识别，事件主体抽取和事件要素抽取。

本次评测任务的文本范围包括互联网上的新闻文本，上市公司发布的公告文本（PDF文档已转成无结构化的文本内容）。 本次评测任务的事件类型包括：财务造假、偿付能力不足、高层失联/去世、企业破产、重大资损、重大赔付、重大事故、股权冻结、股权质押、增持、减持等。本次评测任务把发生特定事件类型的主体称为事件主体，如 “公司A向公司B赔付”对于事件类型“重大赔付”的事件主体是“公司A”。事件要素为该事件类型的所有属性信息，如事件类型“破产清算”的事件要素包括“发布公告时间”、“破产清算的公司”、“受理法院”、 “公司所属行业”、 “裁定时间”。

本次评测包括两个子任务：篇章事件主体抽取和篇章事件要素抽取。具体内容如下：

**子任务一：事件主体抽取**

**该任务旨在从文本中抽取事件类型和对应的事件主体**。即给定文本T，抽取T中所有的事件类型集合S，对于S中的每个事件类型s，从文本T中抽取s的事件主体。其中各事件类型的主体实体类型为公司名称或人名或机构名称

**输入**：一段文本T

**输出**：事件类型和事件主体

**示例**：

***样例1***

*输入：*“*法院裁定公司A需向公司B一次性赔付500万*”

*输出：事件类型：“重大赔付”。 事件主体：*“*公司A*”

***样例2***

*输入：“公司A高管将100万股股权质押给银行，质押开始日期为2018年6月1日，期限为5年*”

*输出： 事件类型：*“*股权质押*”*。事件主体；*“*公司A*”

**子任务二：篇章事件要素抽取**

**该任务旨在从文本中抽取事件类型和对应的事件要素。**即给定文本T，抽取T中所有的事件类型集合S，对于S中的每个事件类型s，从文本T中抽取s的事件要素。

其中各事件类型的要素实体类型如下表：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 高层失联去世 | 事件要素 | 公司名称 | 高层人名 | 高层职务 | 失联/去世时间 |  |
| 事件要素类型 | 公司 | 人名 | 职称 | 时间 |  |
| 企业破产 | 事件要素 | 公告时间 | 公司名称 | 债权债务总额 | 债权清偿率 |  |
| 事件要素类型 | 时间 | 公司 | 数字和单位 | 数字 |  |
| 重大资损 | 事件要素 | 公告时间 | 公司名称 | 资损金额 |  | 　 |
| 事件要素类型 | 时间 | 公司 | 数字和单位 |  | 　 |
| 重大赔付 | 事件要素 | 公告时间 | 公司名称 | 赔付金额 | 赔付对象 | 赔付截止时间 |
| 事件要素类型 | 时间 | 公司 | 数字和单位 | 公司 | 时间 |
| 重大事故 | 事件要素 | 公告时间 | 公司名称 | 事故地址 |  |  |
| 事件要素类型 | 时间 | 公司 | 地址 |  |  |
| 股权冻结 | 事件要素 | 冻结股数 | 被冻结股东名称 | 起始日期 | 截止时间 |  |
| 事件要素类型 | 数字和单位 | 公司/人名 | 时间 | 时间 |  |
| 股权质押 | 事件要素 | 质押股数 | 质押股东名称 | 起始时间 | 截止时间 |  |
| 事件要素类型 | 数字和单位 | 公司/人名 | 时间 | 时间 |  |
| 增持 | 事件要素 | 增持股数 | 增持方股东名称 | 起始时间 | 截止时间 |  |
| 事件要素类型 | 数字和单位 | 公司/人名 | 时间 | 时间 |  |
| 减持 | 事件要素 | 减持股数 | 减持方股东名称 | 起始时间 | 截止时间 |  |
| 事件要素类型 | 数字和单位 | 公司/人名 | 时间 | 时间 |  |

注：评估时“数字和单位”、以及“时间”会对选手答案和标准答案分别进行归一化后再评估

**输入**：一段文本T

**输出**：事件类型和事件要素

**示例**：

***样例1***

*输入：*“*本公司及董事会全体成员保证信息披露的内容真实、准确、完整，没有虚假记载、误导性陈述或重大遗漏。*

 *2017年 1月 12日，长航凤凰股份有限公司（以下简称“公司”） 通过中国登记结算有限公司系统查询获知公司第一大股东天津顺航海运有限公司（以下简称“顺航海运”）持有公司股票 181，015，974股，持股比例 17.89%被天津市第二中级人民法院（以下简称“天津二中院”）司法冻结，具体内容详见《关于公司股东股份被法院司法冻结的公告》 公告编号 2017-006号。*

 *2017年 4月 26日，顺航海运收到天津二中院（2016）津 02民初 821号《民事裁定书》，原告易涛于 2017年 4月 18日向天津二中院申请撤回对被告顺航海运的起诉，天津二中院裁定准许撤回，具体内容详见《关于公司股东股份被法院司法冻结的进展公告》 公告编号 2017-036号*”

*输出：事件类型：“股权冻结*”*。 事件要素(冻结股数)：*“*181，015，974*”*，事件要素(被冻结股东名称)：*“*天津顺航海运有限公司*”*，事件要素(执行冻结结构)：*“*天津市第二中级人民法院*”*，事件要素(起始日期) ：“2017年 1月 12日”*

# 数据描述

本次数据主要来自金融领域的新闻、公告文本，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：

**任务一：**

**训练集&验证集：**

在训练及验证数据发布阶段，我们会发布2万条左右的文本及其所标注事件类型和事件主体和1千条左右的验证文本。训练集每行4列，数据以“\t”分隔，格式为：*文本id\t文本内容\t事件类型\t事件主体，*验证集每行2列，数据以“\t”分隔，格式为：*文本id\t文本内容*

**测试集：**

在测试数据发布阶段，我们将会再发布1千条左右的文本数据集，不含标注结果，作为测试。每行2列，数据以“\t”分隔，格式为：*文本id\t文本内容*

**任务二：**

**训练集&验证集：**

在训练及验证数据发布阶段，我们会发布5千条左右的文本及其所标注事件类型和事件元素和1千条左右的验证文本。训练集每行5列，数据以“\t”分隔，格式为：*文本id\t文本内容\t事件类型\t事件要素名称\t事件要素取值，*验证集每行2列，数据以“\t”分隔，格式为：*文本id\t文本内容*

**测试集：**

在测试数据发布阶段，我们将会再发布1千条左右的文本数据集，不含标注结果，作为测试。每行2列，数据以“\t”分隔，格式为：*文本id\t文本内容*

# 评价指标

**任务1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）来评估事件主体的识别效果。采用微平均计算F值即所有样本一起计算P和R。事件类型与主体同时识别正确算该样本识别正确。

$$事件主体精确率=\frac{识别事件类型与主体和标注相同的数量}{识别出事件类型与主体总数量}$$

$$事件主体召回率=\frac{识别出事件类型与主体和标注相同的数量}{标注的事件类型与主体总数量}$$

$$事件主体F1值=\frac{2\*事件主体精确率\*事件主体召回率}{事件主体精确率+事件主体召回率}$$

**任务2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）来评估事件要素的识别效果。采用微平均计算F值即所有样本的要素一起计算P和R。一个事件类型与某一要素为一个统计项。

给定一个文档，假设答案包含两个事件实例:

A1: 事件类型1，要素一取值1，要素二取值2，要素三取值3，要素四取值4

A2: 事件类型1，要素一取值1，要素二取值6，要素三取值7，要素四取值8

假设选手给出两个预测结果：

P1: 事件类型1，要素一取值1，要素二取值2，要素三取值3，要素四取值8

P2: 事件类型1，要素一取值1，要素二取值2

P3: 事件类型2，要素一取值9，要素二取值10

评测时，会采用不放回的方式给每条答案寻找最相似的预测（即事件类型相同且相同要素最多的），例如上例中A1与P1事件类型相同且3个要素相同，A1与P2事件类型相同且2个要素相同，故与A1最相似的预测是P1，命中3个要素。由于采用不放回的方式，此时预测集合剩下P2、P3，与A2最相似的预测是P2，命中1个要素。此时两条答案均已找到最相似的预测，可以计算Precision和Recall，如下

Precision = (3+1)/(4+2+2)， Recall = (3+1)/(4+4)

注：在给每条答案寻找最相似预测时，相同事件类型的答案会按照要素个数的多少定优先级，如上例中A1和A2事件类型相同，但A1要素个数多，故优先为A1寻找最相似预测。

$$事件要素精确率=\frac{识别事件类型与要素和标注相同的数量}{识别出事件类型与要素总数量}$$

$$事件要素召回率=\frac{识别出事件类型与要素和标注相同的数量}{标注的事件类型与要素总数量}$$

$$事件要素F1值=\frac{2\*事件要素精确率\*事件要素召回率}{事件要素精确率+事件要素召回率}$$

# 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。对于任务一：每行3列，以“\t”分隔，第一列为文本id，第二列为事件类型，第三列为事件主体，即“*文本id\t事件类型\t事件主体*”。对于任务二：每行4列，以“\t”分割，第一列为文本id，第二列为事件类型，第三列为事件要素名称，第四列为事件要素取值，即“*文本id\t事件类型\t事件要素名称\t事件要素值*”

最终提交文件要求：

每一个参赛队需提交的材料如下：

1. 事件主体运行结果文件
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 2020官网）

以上三个文件需在任务提交截止日期前发送至邮箱：

**子任务1：**发送至ccks2020taskFEE\_t1@126.com。邮件的标题为：“CCKS-评测任务FEE子任务1-最终提交文件-参赛队名称”，例如“CCKS-评测任务FEE-子任务1-最终提交文件-火箭队”。

**子任务2：**发送至ccks2020taskFEE\_t2@126.com。邮件的标题为：“CCKS-评测任务FEE-子任务2-最终提交文件-参赛队名称”，例如“CCKS-评测任务FEE-子任务2-最终提交文件-火箭队”。

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(每行格式为“*文本id\t事件主体*”)。

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10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

• 评测任务发布：3月20日

• 报名时间：3月20日—7月10日

• 训练及验证数据发布：3月20日

•   验证集提交开放：3月20日23:59 - 7月17日23:59 （北京时间）

•     测试数据发布：7月27日23:59（北京时间）

•     测试集提交开放： 7月17日23:59 - 7月20日23:59（北京时间）

•     评测论文提交：8月5日

•     CCKS会议日期：8月15日—18日

# 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

第一名：人民币10000元

第二名：人民币3000元

第三名：人民币2000元

技术创新奖：人民币5000元

备注：每个子任务单独设立奖项，每个子任务最多有一个队伍获得技术创新奖，技术创新奖和前三名可以兼得，优胜者有机会进入2021年蚂蚁金服校园招聘终面或者实习生绿色通道。

#  任务组织者和联系人

**任务组织者：**

王太峰（蚂蚁金服）

王峰 （蚂蚁金服）

刘康 （中科院自动化所）

陈玉博 （中科院自动化所）

**任务联系人：**

王峰：zifan.wf@antfin.com

陈玉博：yubo.chen@nlpr.ia.ac.cn

# 任务交流平台

任务交流平台采用钉钉群的方式。所有报名参赛的参赛队要求有一名成员加入到该讨论群中，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群中发布和交流，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群。钉钉群加入二维码如下，在钉钉上扫一扫添加：



子任务1钉钉群二维码



子任务2钉钉群二维码

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：**ccks2020taskFEE\_t1@126.com或ccks2020taskFEE\_t2@126.com**。