任务：面向金融领域的因果事件要素抽取及事件相似度计算

任务描述

金融领域知识图谱和事件图谱可以支持智能投顾、风险管控等众多业务场景，而事件抽取及事件对齐是知识图谱、事件图谱构建的重要环节。

本任务旨在从海量财经新闻中抽取因果事件并补齐相关事件要素，并将相同的事件聚合到一起。具体分为两个子任务：（1）子任务一为因果事件要素的抽取，该子任务目标是继因果关系抽取后，从因果事件所在篇章抽取原因事件要素和结果事件要素；（2）子任务二为事件相似度计算，旨在根据事件描述和事件要素评估事件相似度，并将相同事件进行合并。

**子任务一：因果事件要素抽取**

**该任务目标是继因果抽取后，从因果事件所在篇章抽取原因事件要素和结果事件要素。**即给定文本T，及其包含的原因事件C和结果事件E，分别抽取C和E中的事件要素，如下：

|  |  |  |
| --- | --- | --- |
|  | 事件要素 | 事件要素描述 |
| Core-argument | actor | 施事，指事件动作主体，一个或多个实体构成 |
| action | 动作，描述事件行为或状态，多为动词 |
| object | 受事，指事件动作客体，一个或多个实体构成 |
| Financial-Adjuncts | region | 区域，指事件涉及地点、区域等 |
| industry | 行业，指事件涉及的行业/产业/市场，指生产同类产品或多类产品集合划分的经济活动类别 |
| organization | 机构，指事件涉及机构/公司/企业名称 |
| product | 产品，指事件涉及产品，通常指经营活动中的具体产品 |

**输入**：一段文本，其包含的原因事件C和结果事件E

**输出**：原因事件和结果事件的事件要素

**示例：**

样例一：

输入：{“text\_id”:”123456”, ”text”: “恒大地产2012年拟裁员30%：1月4日，**恒大地产**（3333.HK）的一位副总裁向记者确认，公司内部发文，2012年将裁员30%。知情人士透露，按步骤初期只裁10%，因为2009年裁员的时候，矛盾冲突比较激烈。该副总称**裁员**的**原因**主要是**项目销售问题**。”, “cause\_effect\_list”: [{“id”:”123456\_0”, “cause\_mention”:”项目销售问题”,”effect\_mention”:”裁员”}]}

输出：{“text\_id”:”123456”,” cause\_effect\_list”:[{“id”:”123456\_0”, “cause\_mention”:”项目销售问题”, ”cause\_actor”:”项目”, “cause\_action”:”销售问题”, “cause\_object”:””, “cause\_region”:[],”cause\_industry”:[”地产”], “cause\_organization”:[“恒大地产”], “cause\_product”:[], “effect\_mention”:[“裁员”], ”effect\_actor”:[“恒大地产”], “effect\_action”:[“裁员”], “effect\_object”:””, “effect\_region”:[], ”effect\_industry”:[“地产”], “effect\_organization”:[“恒大地产”], “effect\_product”:[]}]}

样例二：

输入：{“text\_id”:”123457”, “text”: “雷诺拟在华统一延质保周期：**雷诺**中国执行总裁陈国章近日在接受记者采访时表示，雷诺正在考虑统一延长质保周期，但**延保**会**导致车价上升**，具体的定价策略正在商讨中。” , “cause\_effect\_list”: [{“id”:”123457\_0”, “cause\_mention”:” 延保”,”effect\_mention”:”车价上升”}]}}

输出：{“text\_id”:”123457”, cause\_effect\_list”:[{“id”:”123457\_0”, “cause\_mention”:” 延保”, ”cause\_actor”:” 雷诺”, “cause\_action”:” 延保”, “cause\_object”:””, “cause\_region”:[],”cause\_industry”:[], “cause\_organization”:[“雷诺”], “cause\_product”:[], “effect\_mention”:[“车价上升”], ”effect\_actor”:[“车价”], “effect\_action”:[“上升”], “effect\_object”:””, “effect\_region”:[], ”effect\_industry”:[], “effect\_organization”:[“雷诺”], “effect\_product”:[“车”]}]}

数据描述

本次数据主要来自金融领域的公开新闻报道，训练集、验证集及测试集说明如下：

**子任务一：**

**训练集&验证集：**

在训练及验证数据发布阶段，我们会发布5k条左右的事件对和1k条验证集。训练集每行为一个json格式，“text\_id”为样本编号，“text”为文本，“cause\_mention”为原因事件文本，“cause\_actor”为原因事件施事要素，“cause\_action”为原因事件动作，“cause\_object”为原因事件受事，“cause\_region”为原因事件涉及区域，“cause\_industry”为原因事件涉及行业，“cause\_organization”为原因事件涉及机构，“cause\_product”为原因事件提及产品，effect结果事件诸如上包含effect\_mention, effect\_actor, effect\_action, effect\_object, effect\_region, effect\_industry, effect\_organization, effect\_product。其中，事件要素取值可能重复与嵌套，如cause\_organization和effect\_organization取值一致等。

**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。以json格式，包含“text\_id”,“text”, “cause\_mention”, “effect\_mention”。

评价指标

**子任务一：**

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值即所有样本一起计算P和R。

假定两个样例标注为：

A1：

cause：actor：**value0** action：**value1** object：**value2**

region：**value3** industry：NULL organization：**value4**，value5，value6 product：NULL

effect：actor：value0 action：**value7** object：**value8**

region：**value3** industry：NULL organization：value4 product：NULL

A2：

cause：actor：**value10** action：**value11** object：**value12**

region：**value13** industry：NULL organization：NULL product：NULL

effect：actor：**value14** action：**value15** object：NULL

region：value13 industry：NULL organization：NULL product：value14

假设选手给出的预测结果为：

P1对应A1：

cause：actor：**value0** action：**value1** object：**value2**

region：**value3** industry：NULL organization：**value4** product：NULL

effect：actor：value9 action：**value7** object：**value8**

region：**value3** industry：NULL organization：NULL product：NULL

P2对应A2：

cause：actor：**value10** action：**value11** object：**value12**

region：**value13** industry：NULL organization：NULL product：NULL

effect：actor：**value14** action：**value15** object：NULL

region：NULL industry：NULL organization：NULL product：value16

precision\_core = (5+5)/(6+5)

recall\_core = (5+5)/(6+5)

precision\_adjuncts = (3+1)/(3+2)

recall\_adjuncts = (3+1)/(6+3)

$$事件核心要素精确率（Precission）=\frac{预测事件核心元素与标注事件核心元素重合数量}{预测事件核心元素总量}$$

$$事件核心要素召回率（Recall）=\frac{预测事件核心元素与标注事件核心元素重合数量}{标注事件核心元素总量}$$

$$事件核心要素F1值=\frac{2∗事件要素核心精确率∗事件核心要素召回率}{事件要素核心精确率+事件核心要素召回率}$$

$$事件附加要素精确率（Precission）=\frac{预测事件附加元素与标注事件附加元素重合数量}{预测事件附加元素总量}$$

$$事件附加要素召回率（Recall）=\frac{预测事件附加元素与标注事件附加元素重合数量}{标注事件附加元素总量}$$

$$事件附加要素F1值=\frac{2∗事件要素附加核心精确率∗事件附加要素召回率}{事件要素附加精确率+事件附加要素召回率}$$

**子任务二：事件相似度计算**

1. **任务介绍**

事件本体是一种以事件类为基本单元的知识表示模型，相比于传统的以“概念”为核心的本体，它可以描述包含事件动作、时间、对象、地点等要素的完整事件信息，可以保留更加丰富的语义内涵，更符合人类认知规律的知识。因此，以事件为知识表示单元的事件本体模型近年来受到学术界的广泛关注，事件本体的应用也逐步被研究人员所重视。基于事件本体的数据集成首先要解决事件本体的映射问题，即事件类与事件类之间的相似度。

1. **任务定义**

根据事件描述和事件要素的相似度衡量两个事件的相似度并将相同的事件聚合到一起。一个基本思想：如果两个事件对应的事件要素是相似的，那么认为这两个事件也是相似的。换句话说，事件要素的相似度是衡量事件相似的近似指标，对事件的相似度计算应转化为事件要素的相似度计算。这里提到的事件描述（event mention）指谓词性事件短语，事件要素可以用七元组来表示e := <actor, action, object, region, industry, organization, product>，具体定义如下：

|  |  |  |
| --- | --- | --- |
|  | 事件要素 | 事件要素描述 |
| Core-argument | actor | 施事，指事件动作主体，一个或多个实体构成 |
| action | 动作，描述事件行为或状态，多为动词 |
| object | 受事，指事件动作客体，一个或多个实体构成 |
| Financial-Adjuncts | region | 区域，指事件涉及地点、区域等 |
| industry | 行业，指事件涉及的行业/产业/市场，指生产同类产品或多类集合划分的经济活动类别 |
| organization | 机构，指事件涉及机构/公司/企业 |
| product | 产品，指事件涉及产品，通常指经营活动中的具体产品 |

输入：事件A所在文本片段、事件描述、事件要素；事件B所在文本片段、事件描述、事件要素

输出：布尔型变量，0指示事件A与事件B不相同，1指示事件A与事件B相同

1. **数据与示例**

*训练集&验证集*

在训练及验证数据发布阶段，我们会发布5k条左右的事件对和1k条验证集。训练集每行为一个json格式，“text\_id”为样本编号，“text”为文本，“mention”为事件文本，“actor”为施事要素，“action”为动作，“object”为受事，“region”为涉及区域，“industry”为涉及行业，“organization”为涉及机构，“product”为提及产品。

下例展示三个事件的示例。

|  |  |  |  |
| --- | --- | --- | --- |
| 事件要素 | $$e\_{1}$$ | $$e\_{2}$$ | $$e\_{3}$$ |
| Text | 恒大地产2012年拟裁员30%：1月4日，**恒大地产**（3333.HK）的一位副总裁向记者确认，公司内部发文，2012年将裁员30%。知情人士透露，按步骤初期只裁10%，因为2009年裁员的时候，矛盾冲突比较激烈。该副总称**裁员**的**原因**主要是**项目销售问题**。 | 恒大地产2012年拟裁员30%：1月4日，恒大地产（3333.HK）的一位副总裁向记者确认，公司内部发文，2012年将裁员30%。知情人士透露，按步骤初期只裁10%，因为2009年裁员的时候，矛盾冲突比较激烈。该副总称裁员的原因主要是项目销售问题。 | 从2021年9月份开始，恒大全国各地项目接连停工率，交楼量不足1万套导致恒大的销量大幅下降；每天都有商票逾期，债券经被停止融资融业务，裁员60%。 |
| Event mention | 项目销售问题 | 裁员 | 恒大的销量大幅下降 |
| actor | 项目 | 恒大地产 | 恒大 |
| action | 销售问题 | 裁员 | 销量下降 |
| object | -- | -- | -- |
| region | -- | -- | 全国各地 |
| industry | 地产 | 地产 | -- |
| Organ. | 恒大地产 | 恒大地产 | 恒大 |
| Product | -- | -- | -- |

*相似度判定*

|  |  |  |  |
| --- | --- | --- | --- |
| $$Sim(e\_{1},e\_{2})$$ | $$e\_{1}$$ | $$e\_{2}$$ | $$e\_{3}$$ |
| $$e\_{1}$$ | -- | 0 | 1 |
| $$e\_{2}$$ | -- | -- | 0 |
| $$e\_{3}$$ | -- | -- | -- |

0：不同

1：相同

1. **评估方法**

本次任务采用精确率（Precision, P）、召回率（Recall, R）、F1值（F1-measure, F1）来评估事件相似识别效果。

$$事件对相似精确率（Precision）=\frac{\#(预测事件对为相似==标注事件对相似)}{\#预测事件对为相似}$$

$$事件对相似召回率（Recall）=\frac{\#(预测事件对为相似==标注事件对为相似)}{\#标注事件对为相似}$$

$$事件对相似度F1值=\frac{2∗事件对相似精确率∗事件对相似召回率}{事件对相似精确率+事件对相似召回率}$$

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**子任务一：**

{” text\_id”:”xxx”,”cause\_effect\_list”:[{“id”:”xx”, “cause\_actor”: “xxx”, “cause\_action”:”xxx”, “cause\_object”:”xxx”, “cause\_region”:[xxx], “cause\_industry”:[xxx], “cause\_organization”:[xxx],”cause\_product”:[xxx], “effect\_actor”: “xxx”, “effect\_action”:”xxx”, “effect\_object”:”xxx”, “effect\_region”:[xxx], “effect\_industry”:[xxx], “effect\_organization”:[xxx],” effect\_product”:[xxx]},{…}, …]}

对于**子任务二**：

[ [event\_id, event\_id, similarity\_label]

[event\_id, event\_id, similarity\_label]

[event\_id, event\_id, similarity\_label]

…….]

最终提交文件要求：

每一个参赛队需提交的材料如下：

1. 抽取结果运行结果文件
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 2022官网）

以上三个文件需在任务提交截止日期前发送至邮箱：

CCKS22\_EE\_ECR@126.com。邮件的标题为：“CCKS-评测任务EE子任务1-最终提交文件-参赛队名称”，例如“CCKS-评测任务EE-子任务1-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致。

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。

时间安排

任务征集截止：3月24日

任务准备时间：3月25日—4月11日

评测任务发布：4月11日

报名时间：4月11日—7月25日

训练及验证数据发布：4月25日

测试数据发布：7月25日

提交测试结果：7月31日

评测论文提交：8月12日

CCKS会议日期(评测报告及颁奖)：8月25日—28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

本次任务总奖金5万元，每个子任务2.5万元，具体分配如下：

第一名（1支队伍）：人民币15000元

第二名（1支队伍）：人民币5000元

第三名（2支队伍）：人民币2500元

任务组织者

边雯皓（NEC中国研究院）

冯璐（NEC中国研究院）

丁效（哈尔滨工业大学）

高靖龙（哈尔滨工业大学）

任务联系人

边雯皓（bian\_wenhao@nec.cn）

高靖龙（jlgao@ir.hit.edu.cn）