

任务四：面向金融领域的小样本跨类迁移事件抽取 评测第一名报告

队伍：SaltyFishes

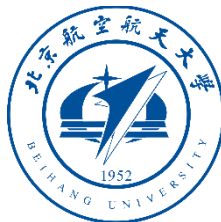
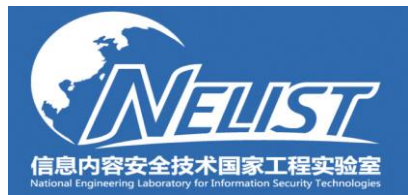
成员：盛傢伟¹、李倩²、黑一鸣²、郁博文¹

指导老师：王丽宏³、郭舒³

1 中国科学院信息工程研究所

2 北京航空航天大学

3 国家互联网应急中心



Outline

1. 赛题介绍
2. 方案流程
3. 实验验证
4. 方案总结

Outline

1. 赛题介绍
2. 方案流程
3. 实验验证
4. 方案总结

赛题介绍

➤ 任务：面向金融领域的小样本跨类迁移事件抽取

- ✓ **事件抽取**：给定金融领域新闻资讯句子，识别事件的**触发词**（Event Trigger）和对应**事件类别**（Event Type），并抽取对应的**事件论元**（Event Arguments）。
- ✓ **跨类迁移**：给定两种事件类别，包括**样本较多的原始大类别**和**样本较少的目标小类别**。任务主要评测训练样本较少的**目标事件类别**的抽取能力。

示例数据：

输入：

刚刚，A公司发布情报通告，称已于2019年10月28日向广州知识产权法院就B公司涉嫌滥用市场支配地位等相关事宜提起诉讼，并于2019年11月4日得到受理。

输出：

- 事件类型：起诉
- 触发词：诉讼
- 原告（公司）：A公司
- 被告（公司）：B公司
- 起诉日期：2019年10月28日

- **原始大事件类别：**
质押、股份股权转让、投资、起诉和减持
- **目标小事件类别：**
收购、担保、中标、签署合同和判决

任务难点

➤ 问题1：跨类迁移问题

- 原始事件类别的平均训练数据量： **936**
- 目标事件类别的平均训练数据量： **179**
- ✓ 原始事件类别的数据量明显高于目标事件类别

➤ 问题2：事件重叠问题

- ✓ 单个触发词可能触发多个事件
- ✓ 单个论元可能在多个事件中作为不同角色
- 以触发词为标志，事件重叠的比例占到**33.9%**

source Types	质押 pledge	投资 investment	股份股权转让 share transfer	高管减持 reduction	起诉 prosecution
Data Size	815	1083	1581	670	533
target Types	收购 acquisition	判决 judgment	中标 win bid	签署合同 sign contract	担保 guarantee
Data Size	200	200	200	132	163

Sentence

世纪华通/ 作价/ 298.03亿元/ 收购/ 盛跃网络/ 100%/ 股权。
Shijihuatong/set a price of/ 29.803 billion yuan/ to acquire/ Shengyue Network's/ 100% equity.

Event 1

Event Type: 投资/ investment

Trigger: 收购/ acquire

Arguments

Sub-company: 世纪华通/ Shijihuatong

Obj-company: 盛跃网络/ Shengyue Network

Money: 298.03亿元/ 29.803 billion yuan

Event 2

Event Type: 股份股权转让/ share transfer

Trigger: 收购/ acquire

Arguments

Sub-company: 盛跃网络/ Shengyue Network

Obj-company: 世纪华通/ Shijihuatong

Money: 298.03亿元/ 29.803 billion yuan

Proportion: 100%/ 100%

Collateral: 股权/ equity

Outline

1. 赛题介绍
2. 方案流程
3. 实验验证
4. 方案总结

任务分解

➤从概率的角度上，本方案提出以下事件抽取任务的分解方式：

$$P(C,T,A|s;\Theta) \propto P(C|s;\Theta_1)P(T|s,C;\Theta_2,\Theta_3)P(A|s,C,T;\Theta_2,\Theta_4)$$

总目标

事件分类

触发词抽取

事件论元抽取

其中，s为给定的句子；C为s中的事件类型集合；T为s中的事件触发词集合；A为s中的事件论元集合

➤ 方案思路：

- ✓ 给定事件类型，抽取对应触发词
- ✓ 给定事件类型和触发词，抽取对应事件论元
- ✓ 任务参数共享，跨类参数共享

# Sentence 世纪华通/ 作价/ 298.03亿元/ 收购/ 盛跃网络/ 100%/ 股权。		
# 事件分类 ->给定: Sentence	返回: <投资>, <股份股权转让>	
# 触发词抽取 ->给定: <投资> ->给定: <股份股权转让>	返回: 收购 返回: 收购	
# 事件论元抽取 ->给定: <投资>; 收购 ->给定: <股份股权转让>; 收购	返回: <Sub-company> <Obj-company> <Money> 返回: <Sub-company> <Obj-company> <Money>	世纪华通 盛跃网络 298.03亿元 盛跃网络 世纪华通 298.03亿元

方案框架

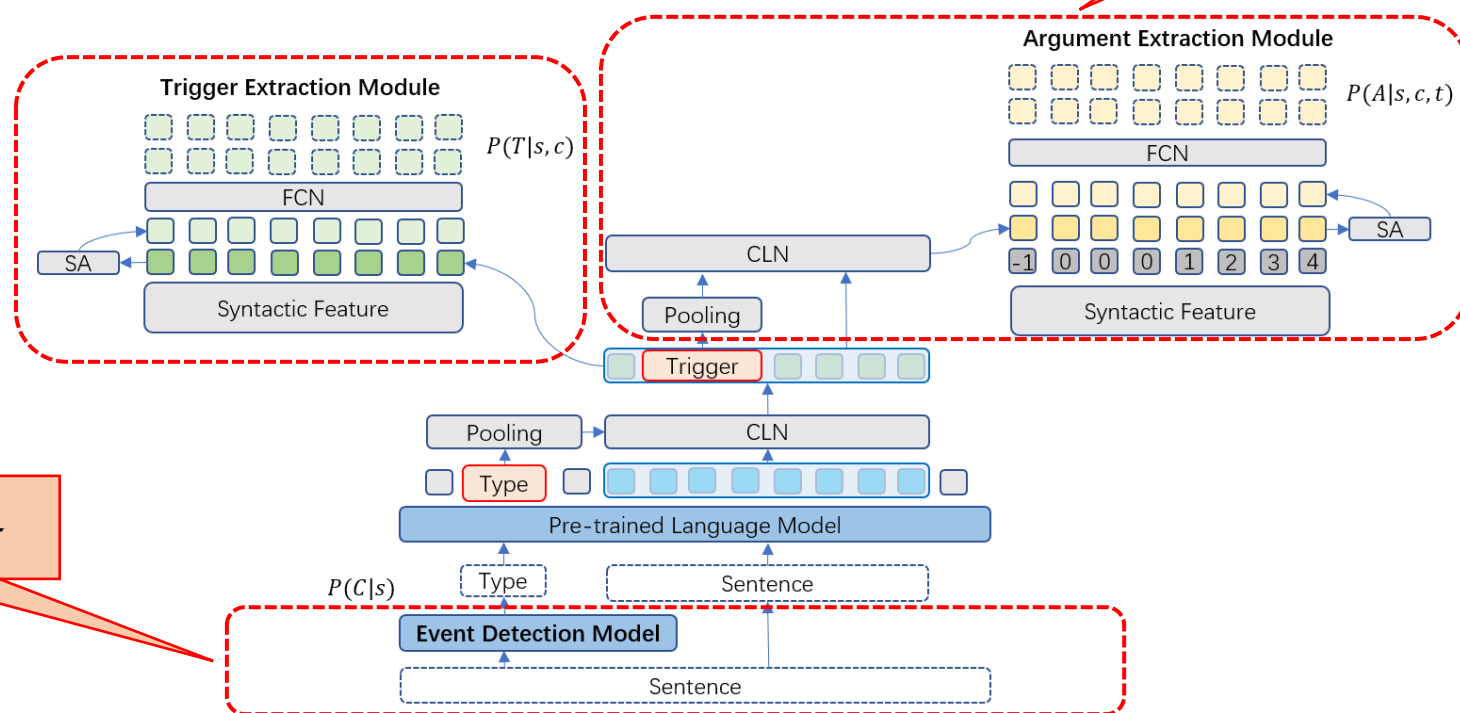
➤ 本方案包括两个模型：

- 事件检测模型：事件检测任务 $P(C|s; \Theta_1)$
- 事件抽取模型：触发词抽取任务 $P(T|s, C; \Theta_2, \Theta_3)$ ；论元抽取任务 $P(A|s, C, T; \Theta_2, \Theta_4)$

论元抽取任务

触发词抽取任务

事件检测任务



EDM:事件检测模型

➤ 事件检测模型 (Event Detection Model, EDM)

- ✓ 多标签多类型文本分类：单个句子中可能存在多个事件类型 $P(C|s; \Theta_1)$
- ✓ 预训练语言模型 (Pre-trained Language Model, PLM)：提升特定任务的抽取能力

➤ 方案：

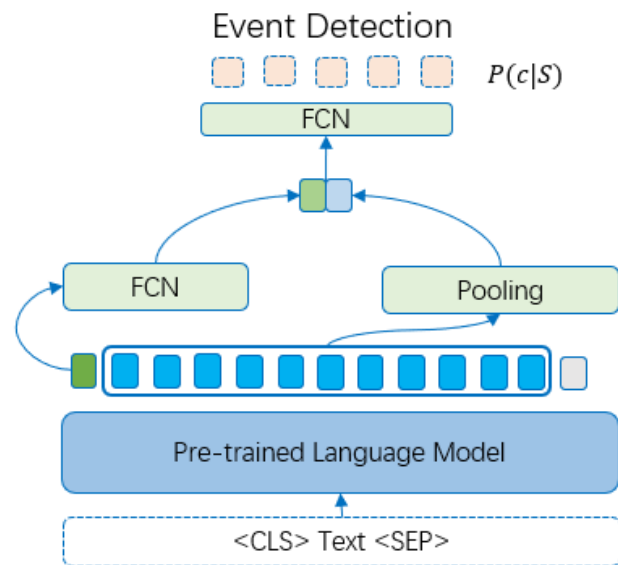
- 通过PLM，得到句子的表示向量 z_{sent}
- 计算句子 s 中发生事件类型 c 的概率：

$$p(c|s; \Theta_1) = \text{sigmoid}(\mathbf{w}_c \cdot \mathbf{z}_{sent}),$$

- 采用二元交叉熵损失训练模型

$$Loss(\Theta_1) = \sum_{m=1}^M y_m * \log(p_m) + (1 - y_m) * \log(1 - p_m)$$

其中， M 是训练集大小， y 为真实标签， p 为预测标签

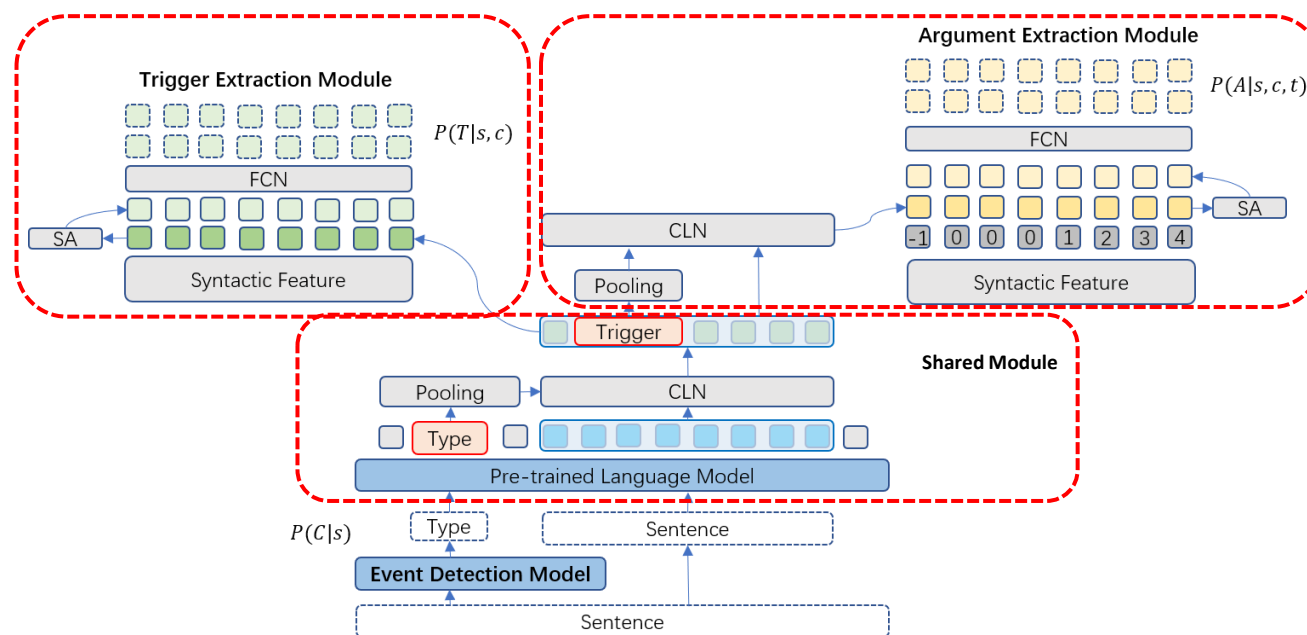


EEM:事件抽取模型

➤ 为了完成触发词抽取和论元抽取任务，设计事件抽取模型（Event Extraction Model, EEM）

➤ 模型包括：

- 共享模块：共享的文本表示 (Θ_2)
- 私有模块：
 1. 触发词抽取模块TEM (Θ_3)
 2. 论元抽取模块AEM (Θ_4)



共享模块：文本编码

➤ 根据指定的事件类别，编码文本语义表示

- ✓ 共享文本编码，增强触发词抽取任务和论元任务的关联

➤ 方案：

- 将事件类别 c 拆解成token，作为PLM的输入 X

$$X : \langle \text{CLS} \rangle + \text{type tokens} + \langle \text{SEP} \rangle + \text{word tokens} + \langle \text{SEP} \rangle.$$

- 得到token的语义表示： $H = \text{PLM}(X)$
- 借助CLN模块，将类型编码 H_c 注入到文本编码 H_s 中

$$H_{s-ty} = \text{CLN}(H_s, \text{MeanPooling}(H_c)),$$

- H_{s-ty} 为事件类别条件下的文本表示，用于后续抽取模块

Conditional Layer Normalization, CLN:

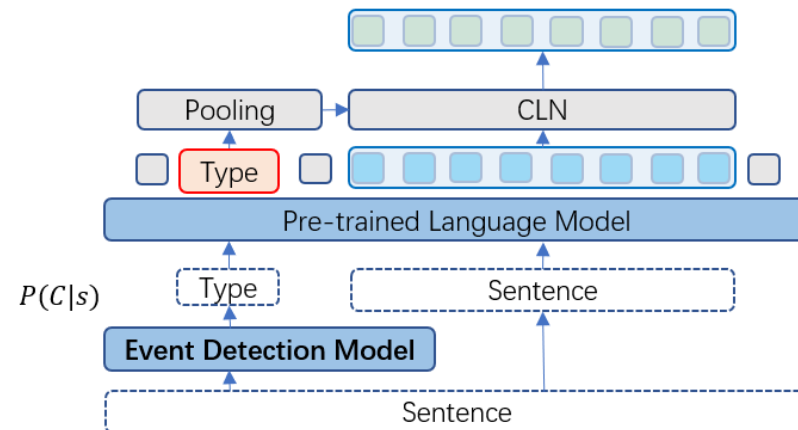
给定条件信息向量 \mathbf{c} 和单词向量 \mathbf{x} ,

CLN将条件信息融合到向量的均值和方差中

$$\text{CLN}(\mathbf{x}, \mathbf{c}) = \gamma_c \odot \left(\frac{\mathbf{x} - \mu}{\sigma} \right) + \beta_c,$$

$$\mu = \frac{1}{d} \sum_{i=1}^d x_i, \sigma = \sqrt{\frac{1}{d} \sum_{i=1}^d (x_i - \mu)^2},$$

$$\gamma_c = W_\gamma \mathbf{c} + \mathbf{b}_\gamma, \beta_c = W_\beta \mathbf{c} + \mathbf{b}_\beta,$$



TEM:触发词抽取模块

➤ $P(T|s, C; \Theta_2, \Theta_3)$: 抽取指定事件类别条件下的事件触发词

- ✓ 为了增强跨事件类别的触发词抽取能力，各个事件类别共享触发词解码器

➤ 方案 (Trigger Extraction Module, TEM)

- 私有编码层：基于自注意力和语法特征

$$H_{sa-ty p} = SA(H_{s-ty p}).$$

$$H_{tri} = H_{s-tyr} \oplus H_{sa-tyr} \oplus H_{syn}.$$

其中, H_{sa-tyr} 为自注意力层输出, H_{syn} 为语法特征,

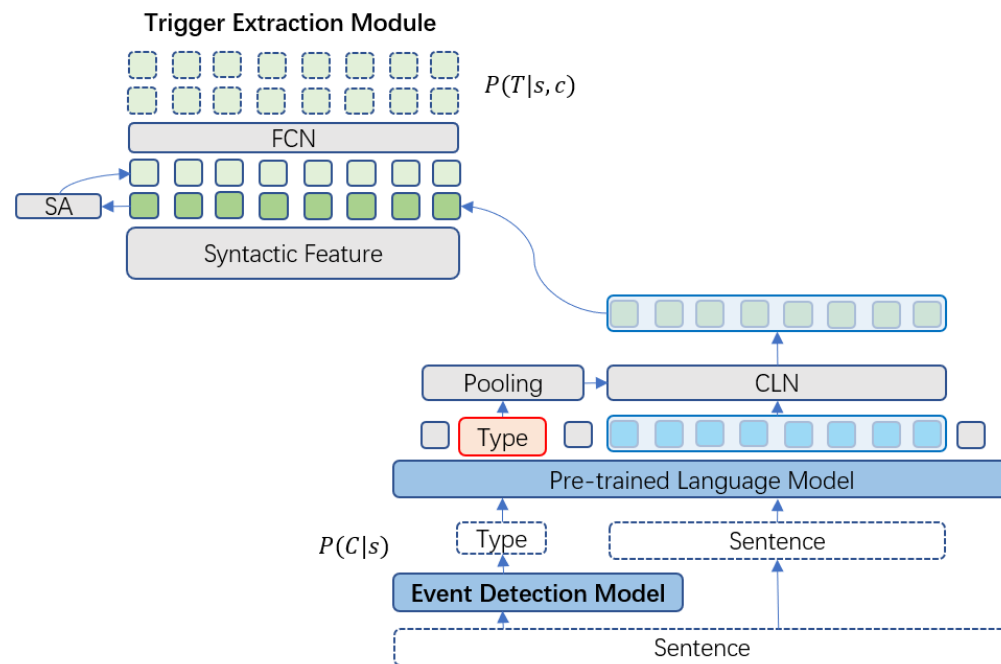
- 解码层：预测触发词的开始位置和结束位置

$$p(t^{(b)}|x_i, c; \Theta_2, \Theta_3) = \text{sigmoid}(w_{t^{(b)}} * h_{tri,i}),$$

$$p(t^{(e)}|x_i, c; \Theta_2, \Theta_3) = \text{sigmoid}(w_{t^{(e)}} * h_{tri,i})$$

- 训练损失: 解码层两部分二元交叉熵损失之和

$$Loss_{tri}(\Theta_2, \Theta_3) = w_t * Loss_{t(b)}(\Theta_2, \Theta_3) + (1 - w_t) * Loss_{t(e)}(\Theta_2, \Theta_3),$$



AEM:论元抽取模块

➤ $P(A|s, C, T; \Theta_2, \Theta_4)$: 抽取指定事件类型和指定触发词的事件论元

✓ 为了增强跨事件类别的论元抽取能力, 各个事件类别下的相同论元类型共享解码器

➤ 方案(Argument Extraction Module, AEM) :

- 私有编码层: 类似于TEM

$$H_{arg} = H_{s-tri} \oplus H_{sa-tri} \oplus R \oplus H_{syn},$$

其中, H_{s-tri} 为触发词条件下的文本表示; H_{sa-tri} 为自注意力层输出;

R 为当前位置到触发词的相对位置编码; H_{syn} 为语法特征

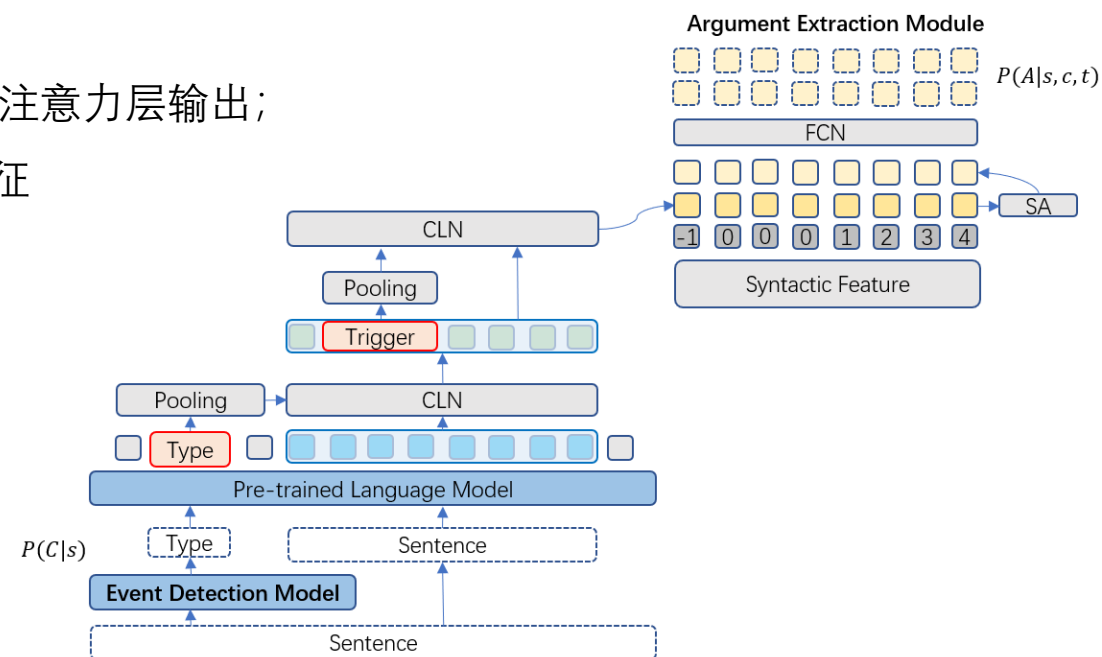
- 解码层: 预测论元类型 a_k 的开始位置和结束位置

$$p(a_k^{(b)}|x_i, c, t; \Theta_2, \Theta_4) = \text{sigmoid}(w_{a_k^{(b)}} * h_{arg,i}),$$

$$p(a_k^{(e)}|x_i, c, t; \Theta_2, \Theta_4) = \text{sigmoid}(w_{a_k^{(e)}} * h_{arg,i}),$$

- 训练损失: 解码层两部分二元交叉熵损失之和

$$Loss_{arg}(\Theta_2, \Theta_4) = \sum_{k=1}^K w_a * Loss_{a_k^{(b)}}(\Theta_2, \Theta_4) + (1 - w_a) * Loss_{a_k^{(e)}}(\Theta_2, \Theta_4), \quad (16)$$



模型训练

➤ EDM: 独立学习 $Loss(\Theta_1)$

➤ EEM: 联合学习TEM和AEM, 合并两部分损失

$$Loss(\Theta_2, \Theta_3, \Theta_4) = w_j * Loss_{tri}(\Theta_2, \Theta_3) + (1 - w_j) * Loss_{arg}(\Theta_2, \Theta_4)$$

其中, $w_j \in (0,1)$ 为折中系数

➤ 训练&测试

- ✓ 在训练过程中, 采用groudtruth训练所有模块, 加速模型的训练
- ✓ 在测试过程中, 依次完成事件检测, 触发词抽取和论元抽取任务

提升策略

1. 继续预训练

- PLM在通用语料上训练，往往与特定领域存在语义偏置
- 为了充分利用**所有比赛数据**，适应金融领域语料，在预训练的RoBERTa上按原预训练损失继续训练

2. 模型集成

- 为了充分利用所有**有标签的训练数据**，采用K-fold交叉检验的方式划分本地训练集和本地验证集
- 在K个数据集划分上训练模型，将结果按投票法集成

3. 利用数据伪标签

- 为了充分利用所有**无标签的测试数据**，我们用中间模型在测试集上预测伪标签
- 将伪标签数据和真实标签数据合并，训练新的模型

4. 后处理规则

- 利用论元长度约束、类别约束等，去除冗余结果

Outline

1. 赛题介绍
2. 方案流程
3. 实验验证
4. 方案总结

数据集

➤ 数据集类别信息和数据集划分信息

- ✓ 测试集为官方给出，训练集和验证集为本地划分

source Types	质押 pledge	投资 investment	股份股权转让 share transfer	高管减持 reduction	起诉 prosecution
Data Size	815	1083	1581	670	533
target Types	收购 acquisition	判决 judgment	中标 win bid	签署合同 sign contract	担保 guarantee
Data Size	200	200	200	132	163

Table 1. Statistics of each event type in the dataset.

	training	validation	testing
source Types	2,459	273	163,763
target Types	738	82	93,610

Table 2. Data partition for training, validation and testing.

消融实验

➤ 为了验证EEM中各子模块的有效性，进行了消融实验

	Trigger Extraction			Argument Extraction			F1-mean
	P	R	F1	P	R	F1	
1. complete model	.969	.979	.970	.844	.969	.889	.930
2. w/o pseudo-label data	.940	.952	.939	.845	.863	.838	.888
3. w/o source data	.941	.945	.934	.818	.865	.823	.878
4. repl PLM: BERT	.901	.924	.904	.789	.825	.789	.846
5. repl PLM: RoBERTa	.931	.938	.929	.837	.886	.828	.879
6. repl CLN: concat	.931	.931	.926	.807	.860	.814	.870
7. w/o layer lr	.946	.945	.940	.799	.869	.816	.878
8. w/o syntactic feature	.921	.924	.917	.863	.874	.856	.887

Table 3. Results on validation data.

✓ 经过实验验证，EEM中各个子部分都能够起到相应的作用

提交结果


CCKS & 招商银行 • ¥ 44,500 • 404 支单人队伍 • 13 支多人队伍 • 440 名参赛者

CCKS 2020：面向金融领域的小样本跨类迁移事件抽取

组队截止时间 2020-09-10

开始时间 2020-03-20

结束时间 2020-09-20

#	队伍名	分数	最终提交次数
1	SaltyFishes 	0.87813	18
2	ianma	0.87664	19
3	HelloWorld_HW	0.85071	18
4	=★baseline★=	0.84943	21
5	aj47	0.84397	20
6	inspur	0.84339	11
7	juzibot 	0.83781	13
8	xiashurang	0.83516	9
9	renweixuan	0.83157	21
10	Aster 	0.82921	17
11	天大智算队 	0.56975	1

Outline

1. 赛题介绍
2. 方案流程
3. 实验验证
4. 方案总结

结论

- 针对比赛中的跨类事件抽取问题和事件要素重叠问题，本文提出
 - ✓ 通过跨类共享抽取模块参数，提升跨类别的事件抽取能力
 - ✓ 通过分解抽取任务，抽取指定条件下的事件要素，解决事件抽取中的要素重叠问题
- 通过实验验证了方案模块的有效性，并在比赛中取得了第一名的成绩

Thanks!
Questions and Advices?

Email: shengjiawei@iie.ac.cn