A Prior Information Enhanced Extraction Framework for Document-level Financial Event Extraction

SUDA-HUAWEI

王海涛　朱桐　瞿晓晔　王铭涛　张国梁

陈文亮　王喆锋

苏州大学　华为技术有限公司

CONTENT

目　录

1

任　务　定　义

Task Definition

任务定义
Task Definition

- 任务：从文本中抽取事件类型和对应的事件要素。

- 输入：
- 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。

- 输出：

股权冻结	被冻结股东	冻结金额	冻结开始日期	冻结结束日期
	天津顺航海运有限公司	181，015，974	2017年1月12日	—

2

数 据 分 析

Data Analysis

数据分析

Data Analysis

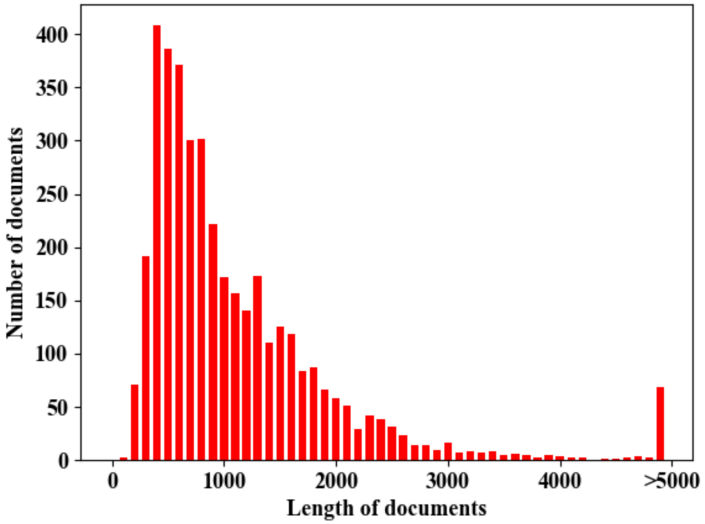
- 来源：金融领域的新闻、公告文本

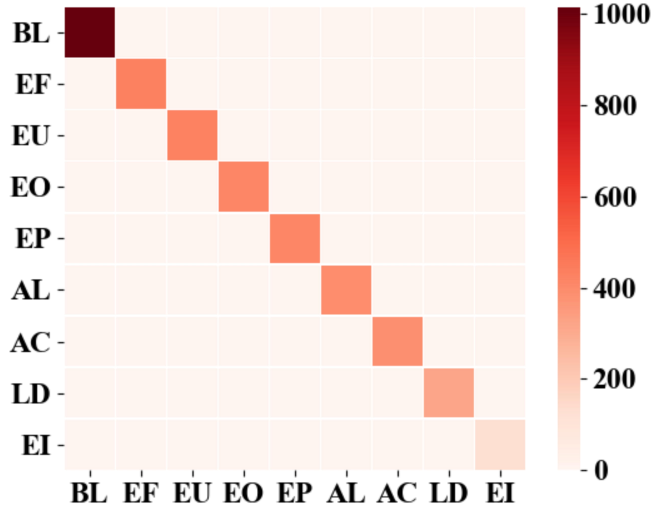
- 训练集：3956条数据，5521个事件，9种事件类型
- 验证集：750条数据
- 测试集：28096条数据，其中1000条左右参与结果评测

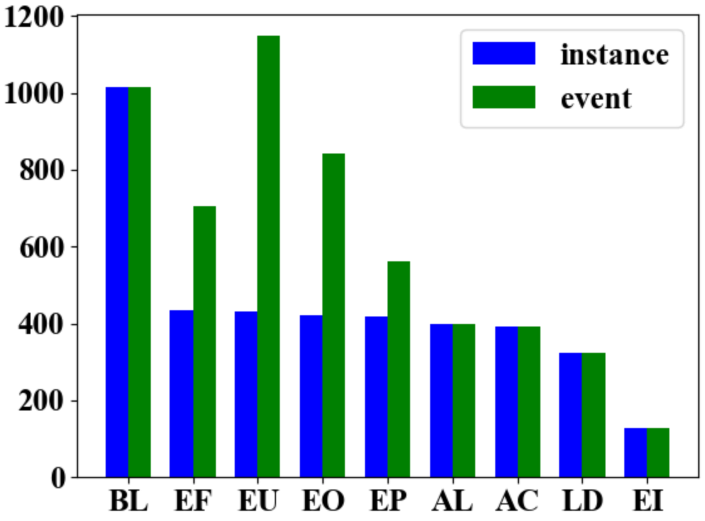
数据分析

Data Analysis

- 训练集中81.7%的文本长度大于512
- 训练集中每个文本中只存在一种事件类型
- 训练集中股权质押、股权冻结、股东增持、股东减持的文本中一般存在多个同类型事件，其余事件类型文本中只存在一个事件







3

数 据 处 理

Data Preprocessing

数据处理
Data Preprocessing

- 还原转义符号和html标记
- 去除重复标点、冗余的空格、网页链接
- 全角转半角
- 简繁转换
- 断句，最大句长为500

	&quot	&apos	&amp	&gt	&lt	
\s	"	'	&	>	<	\n

*
替换成空格

数据处理
Data Preprocessing

- 回标数据（BIO）
- B-{事件类型}-{事件元素角色}
  - B-股权冻结-被冻结的股东、B-股权冻结-冻结金额
- I -{事件类型}-{事件元素角色}
  - I-股权冻结-被冻结的股东、I-股权冻结-冻结金额
- O

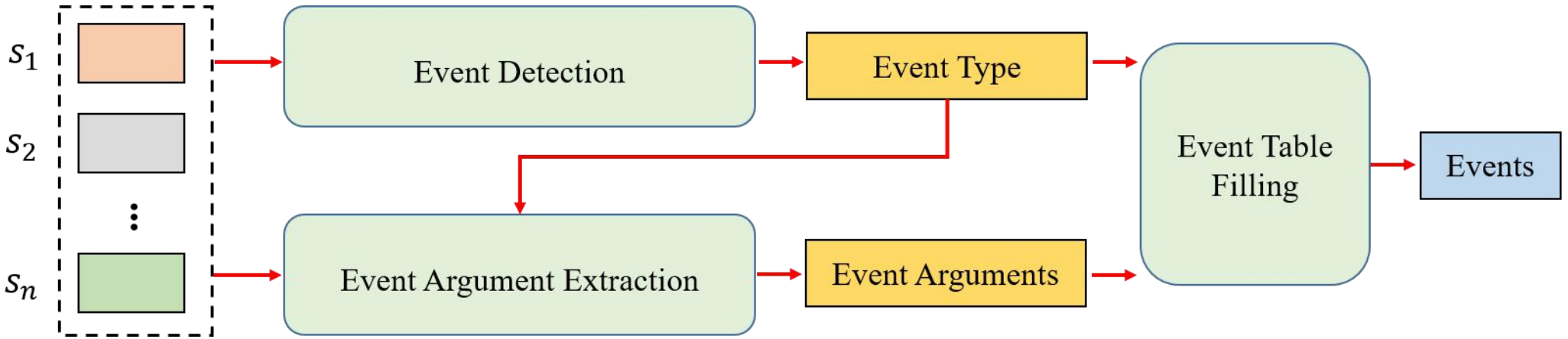
顺	航	海	运	被	冻	结	股	票	一	亿	股	。
B	I	I	I	O	O	O	O	O	B	I	O	O

4

系　统　结　构

System Structure

系统结构
System Structure



事件类型识别
Event Detection

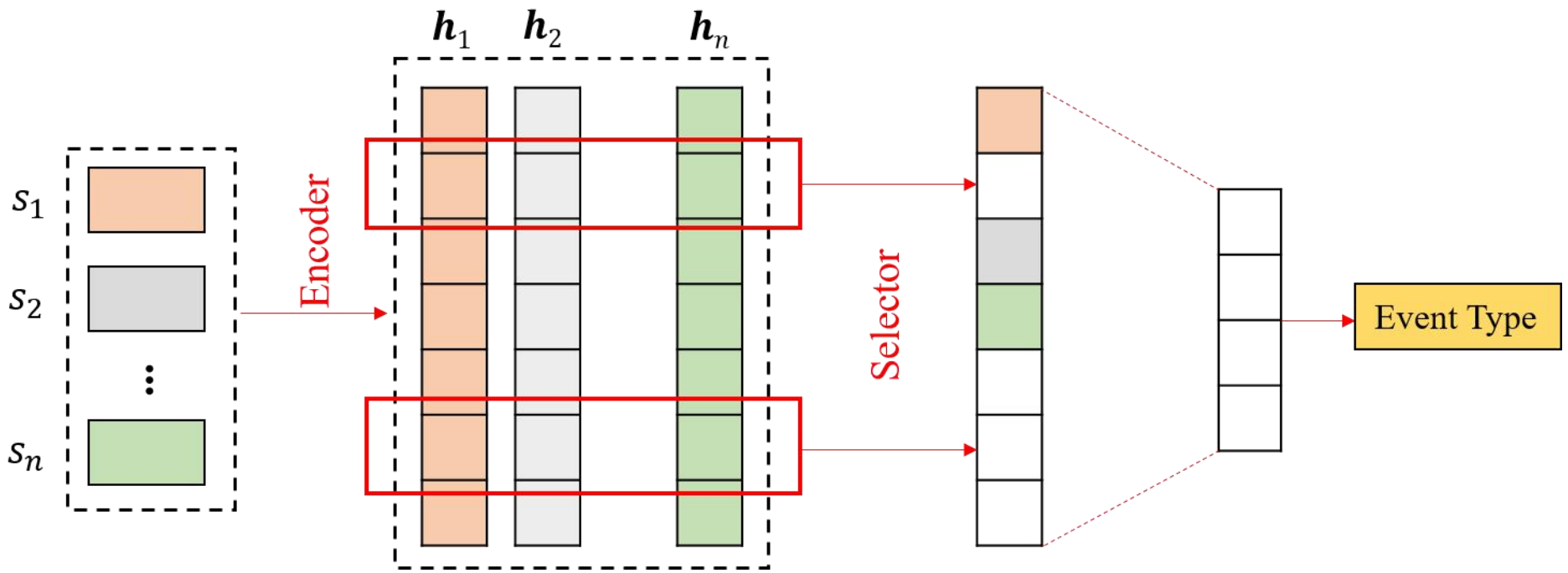
- 关系抽取：

  If two entities participate in a relation, at least one sentence that mentions these two entities might express that relation.

- 我们认为：

  If a document contains some type of event type, there is at least one sentence from this document can fully describe that event type.

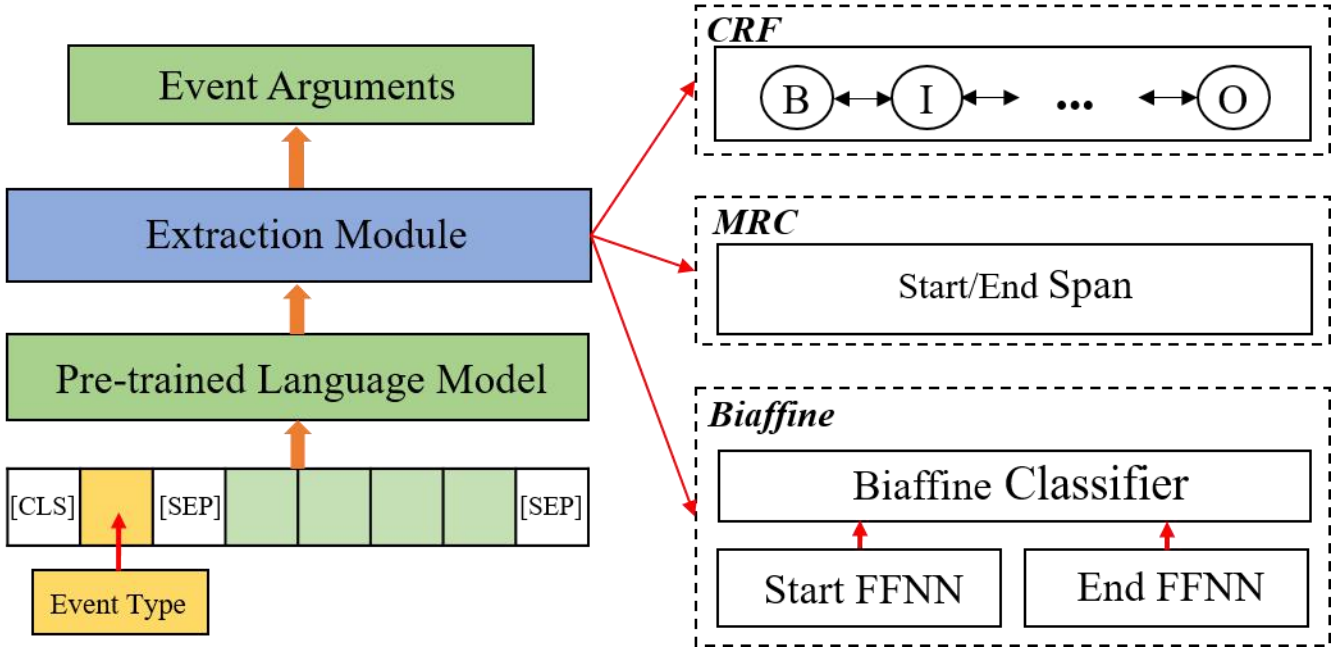
事件类型识别
Event Detection



事件类型识别
Event Detection

	CNN	BiLSTM
ONE	0.97524	0.94045
ATT	0.98233	0.97251
MAX	**0.98560**	**0.98988**

事件元素抽取
Event Argument Extraction



事件元素抽取
Event Argument Extraction

- PLM—CRF

顺	航	海	运	被	冻	结	股	票	一	亿	股	。
B	I	I	I	O	O	O	O	O	B	I	O	O

股权冻结-被冻结的股东

股权冻结-冻结金额

事件元素抽取
Event Argument Extraction

- PLM-MRC
- 构建问句
- 股权冻结-被冻结的股东：股权被冻结的股东 [1]
- 股权冻结-冻结金额：被冻结的股权金额 [2]

		顺	航	海	运	被	冻	结	股	票	一	亿	股	。
[1]	Start	1	0	0	0	0	0	0	0	0	0	0	0	0
	End	0	0	0	1	0	0	0	0	0	0	0	0	0
[2]	Start	0	0	0	0	0	0	0	0	0	1	0	0	0
	End	0	0	0	0	0	0	0	0	0	0	1	0	0

事件元素抽取
Event Argument Extraction

- PLM—Biaffine

	顺$_0$	航$_1$	海$_2$	运$_3$	被$_4$	…	票$_8$	一$_9$	亿$_{10}$	股$_{11}$	。$_{12}$
顺$_0$	0	0	0	1	0	0	0	0	0	0	0
航$_1$	-1	0	0	0	0	0	0	0	0	0	0
海$_2$	-1	-1	0	0	0	0	0	0	0	0	0
运$_3$	-1	-1	-1	0	0	0	0	0	0	0	0
被$_4$	-1	-1	-1	-1	0	0	0	0	0	0	0
…	-1	-1	-1	-1	-1	0	0	0	0	0	0
票$_8$	-1	-1	-1	-1	-1	-1	0	0	0	0	0
一$_9$	-1	-1	-1	-1	-1	-1	-1	0	2	0	0
亿$_{10}$	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0
股$_{11}$	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0
。$_{12}$	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0

事件元素抽取
Event Argument Extraction

Models	F1-score	Training Time/Epoch
PLM-CRF☆	0.82503	31min
PLM-CRF	0.84033	31min
PLM-MRC☆	0.00000	63min
PLM-MRC	**0.84777**	63min
PLM-Biaffine☆	0.82691	**18min**
PLM-Biaffine	0.84772	**18min**

☆ means no prior event type information is utilized.

事件元素抽取
Event Argument Extraction

PLM	F1-score
BERT-base	0.84615
BERT-wwm	0.84772
BERT-wwm-ext	0.84977
ERNIE	0.84298
RoBERTa-wwm-ext	0.85546
RoBERTa-wwm-ext-large	0.86533
NEZHA-large	**0.86693**

事件表填充

Event Table Filling

- 单事件
  - 破产清算、重大资产损失、重大安全事故、高层死亡、重大对外赔付
- 多事件
  - 股东减持、股东增持、股权冻结、股权质押

事件表填充
Event Table Filling

- 公告编号:2018-025证券代码:871832证券简称:大石头主办券商:开源证券烟台大石头景观园林文化股份有限公司关于公司股东、总经理去世公告本公司及董事会全体成员保证公告内容的真实、准确、完整,没有任何虚假记载、误导性陈述或者重大遗漏,并对其内容的真实性、准确性和完整性承担个别及连带责任。
- 烟台大石头景观园林文化股份有限公司(以下简称“公司”)董事会于2018年3月25日获悉,公司股东宋国然先生因病于2018年3月24日不幸去世,享年60岁。
- 公司董事、监事、高级管理人员及全体员工,对宋国然先生的去世表示沉痛哀悼,对其家属表示深切慰问。
- 宋国然先生目前持有公司股份10,074,060股,占公司股份比例62.18%,目前为公司第一大股东。
- ……

事件表填充
Event Table Filling

- 公告编号:2018-025证券代码:871832证券简称:大石头主办券商:开源证券**烟台大石头景观园林文化股份有限公司**关于公司**股东**、**总经理去世**公告本公司及董事会全体成员保证公告内容的真实、准确、完整,没有任何虚假记载、误导性陈述或者重大遗漏,并对其内容的真实性、准确性和完整性承担个别及连带责任。
- **烟台大石头景观园林文化股份有限公司**(以下简称“公司”)董事会于2018年3月25日获悉,公司**股东宋国然**先生因病于2018年3月24日不幸**去世**,享年60岁。
- 公司董事、监事、高级管理人员及全体员工,对**宋国然**先生的去世表示沉痛哀悼,对其家属表示深切慰问。
- **宋国然**先生目前持有公司股份10,074,060股,占公司股份比例62.18%,目前为公司第一大**股东**。
- ……

事件表填充
Event Table Filling

- 一、股东股份质押的基本情况 1、股东股份被质押基本情况 陈玉忠先生分别于 2018年4月2日、2018年4月3日在中国证券登记结算有限责任公司办理完毕其所持有本公司的股票 6700万股、4600万股、4700万股质押给上海电气集团股份有限公司的质押登记手续,质押期限分别自 2018年3月30日、2018年4月2日起,具体情况如下:股东名称是否为第一大 股东及一致行动人质押股数(万股) 质押登记日期 质押到期日期质权人本次质押占其所持股份比例(%)用途 陈玉忠 是 6700 2018-03-30 2020-03-01上海电气集团股份有限公司 38.27 业务担保 陈玉忠 是 4600 2018-03-30 2020-03-01上海电气集团股份有限公司 26.28 业务担保 陈玉忠 是 4700 2018-04-02 2020-01-31上海电气集团股份有限公司 26.85 业务担保

事件表填充
Event Table Filling

- 一、股东股份质押的基本情况 1、股东股份被质押基本情况 陈玉忠先生分别于 2018年4月2日、2018年4月3日在中国证券登记结算有限责任公司办理完毕其所持有本公司的股票 6700万股、4600万股、4700万股质押给上海电气集团股份有限公司的质押登记手续,质押期限分别自 2018年3月30日、2018年4月2日起,具体情况如下：股东名称是否为第一大 股东及一致行动人质押股数(万股) 质押登记日期 质押到期日期质权人本次质押占其所持股份比例(%)用途 陈玉忠 是 6700 2018-03-30 2020-03-01上海电气集团股份有限公司 38.27 业务担保 陈玉忠 是 4600 2018-03-30 2020-03-01上海电气集团股份有限公司 26.28 业务担保 陈玉忠 是 4700 2018-04-02 2020-01-31上海电气集团股份有限公司 26.85 业务担保

评测结果
Online Results

Leaderboard A		**Leaderboard B**	
Teams	F1-score	Teams	F1-score
SUDA-HUAWEI	**0.83007**	**SUDA-HUAWEI**	**0.66996**
同花顺	0.81411	mulan	0.65043
ztjerry	0.80578	uloveqian	0.63469
mulan	0.78422	**同花顺**	0.61530
FreeWings	0.78359	LTF_	0.60464

5

总　结　展　望

Summary & Prospect

总结展望
Summary & Prospect

总结

- 数据分析和预处理很重要
- 先验信息有利于事件元素抽取
- 预训练模型

展望

- 事件表填充建模
- 事件元素抽取利用全局信息
- 考虑远程监督数据的影响

感　谢　观　看

Thanks for watching