

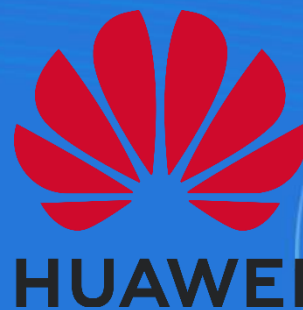
Efficient Document-level Event Extraction via Pseudo-Trigger-aware Pruned Complete Graph

Tong Zhu¹, Xiaoye Qu², Wenliang Chen¹, Zhefeng Wang²,
Baoxing Huai², Nicholas Yuan², Min Zhang¹

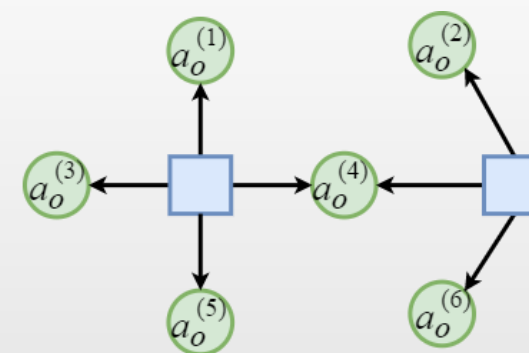
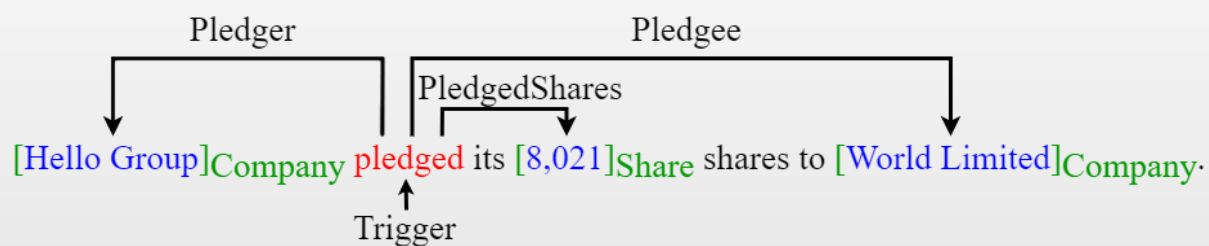
<https://github.com/Spico197/DocEE>



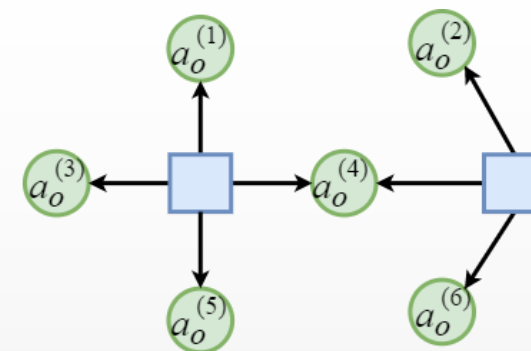
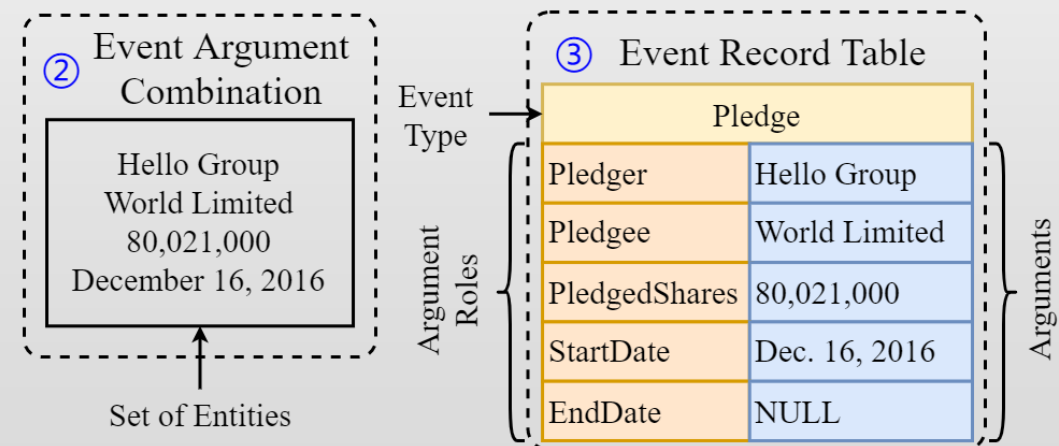
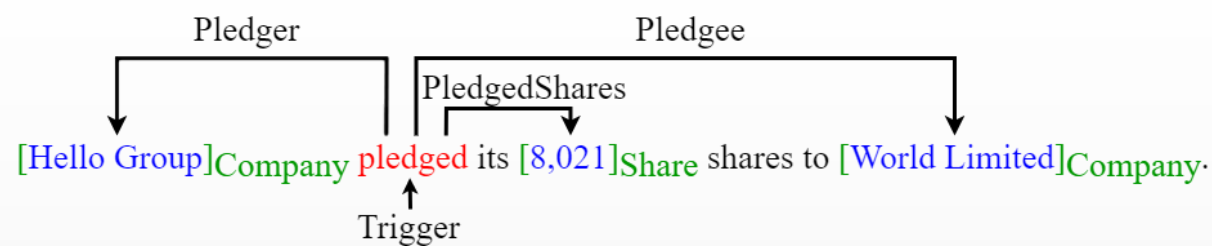
¹Soochow University



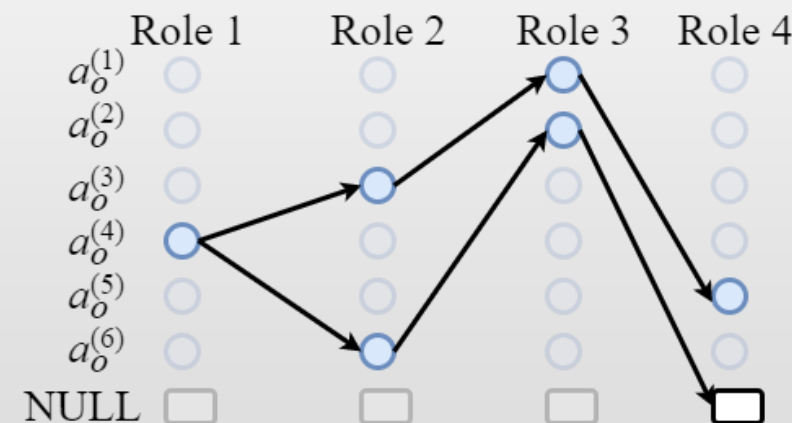
²Huawei Cloud



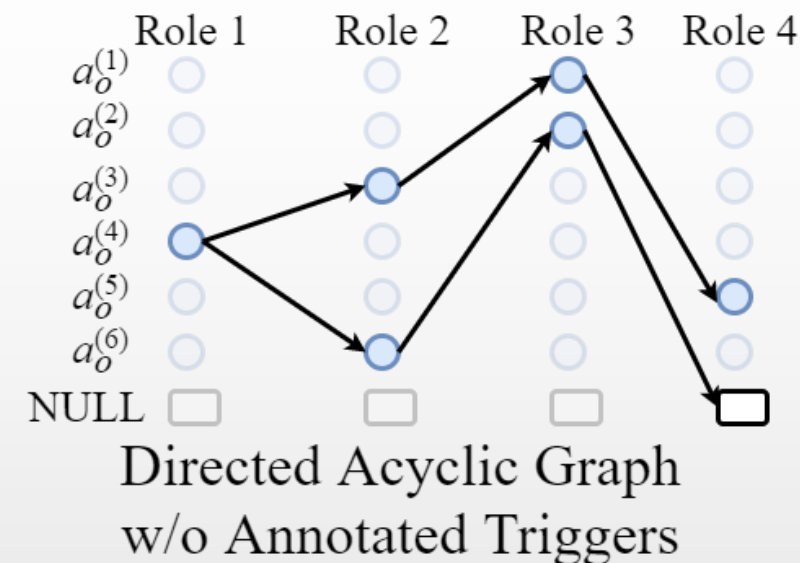
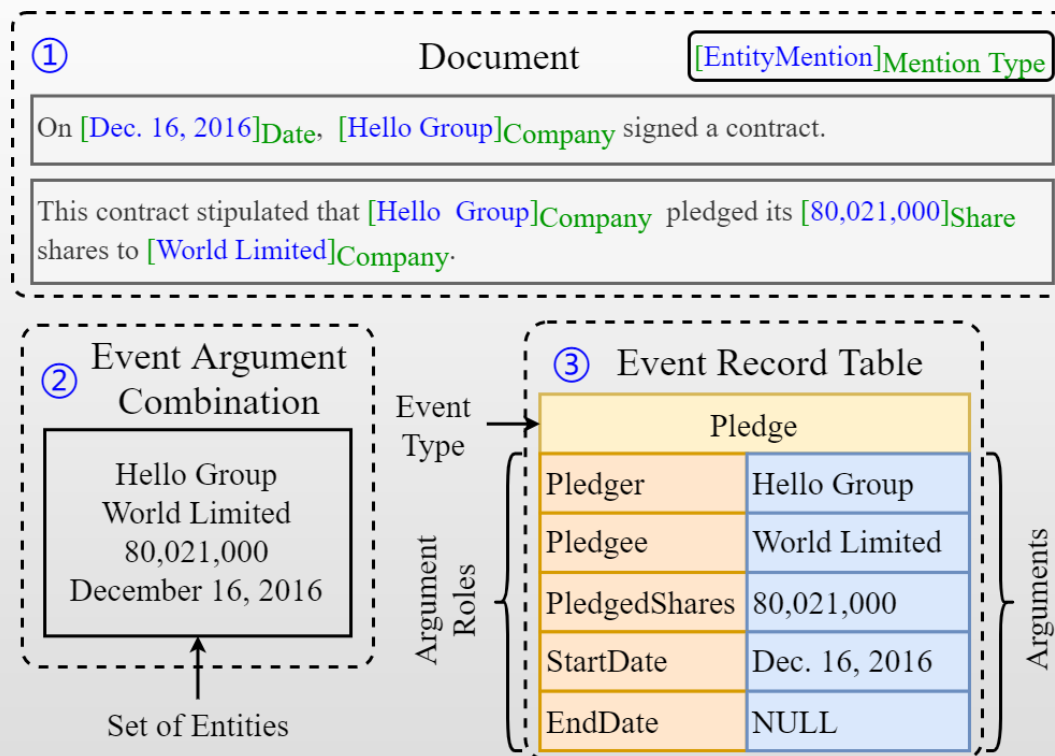
Annotated-Trigger-centered Trees



Annotated-Trigger-centered Trees



Directed Acyclic Graph
w/o Annotated Triggers

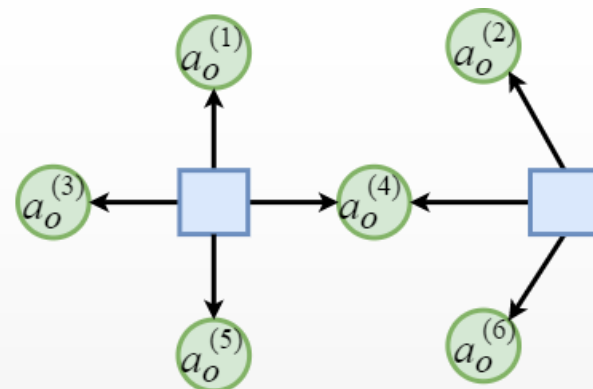
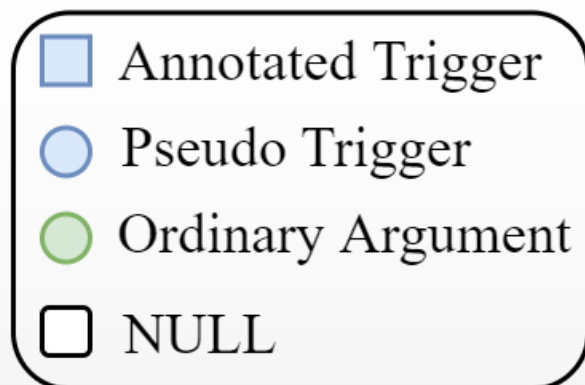


任务本身的难点：

- 无触发词
- 长文本处理

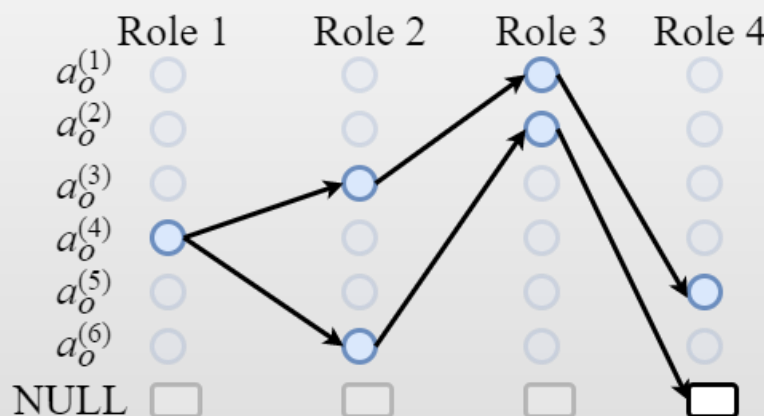
当前方法的痛点：

- 训练慢：4-8卡跑将近一周
- 推理时消耗资源多

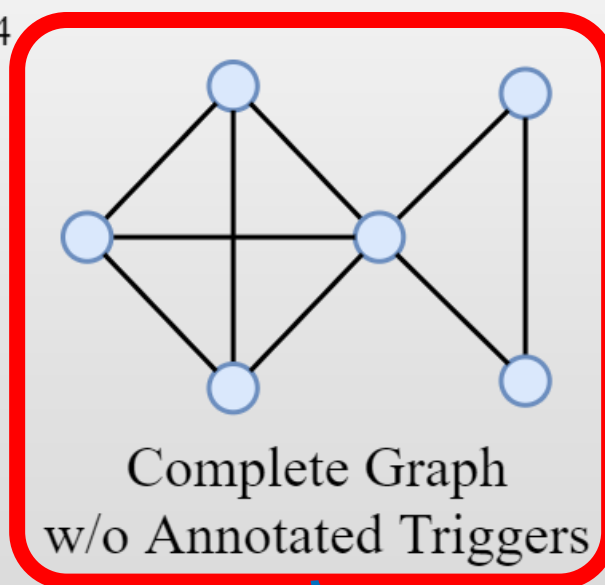


Annotated-Trigger-centered Trees

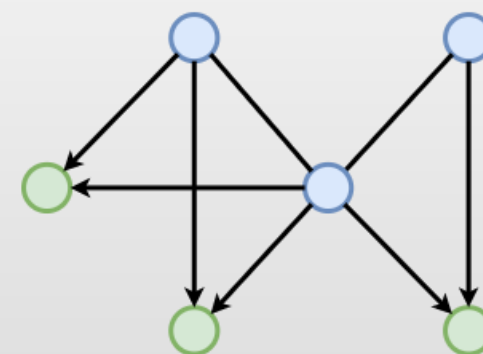
如果每个事件实例只选择一个核心论元，则退化为句级事件抽取中的常用组合方法



Directed Acyclic Graph
w/o Annotated Triggers



Complete Graph
w/o Annotated Triggers

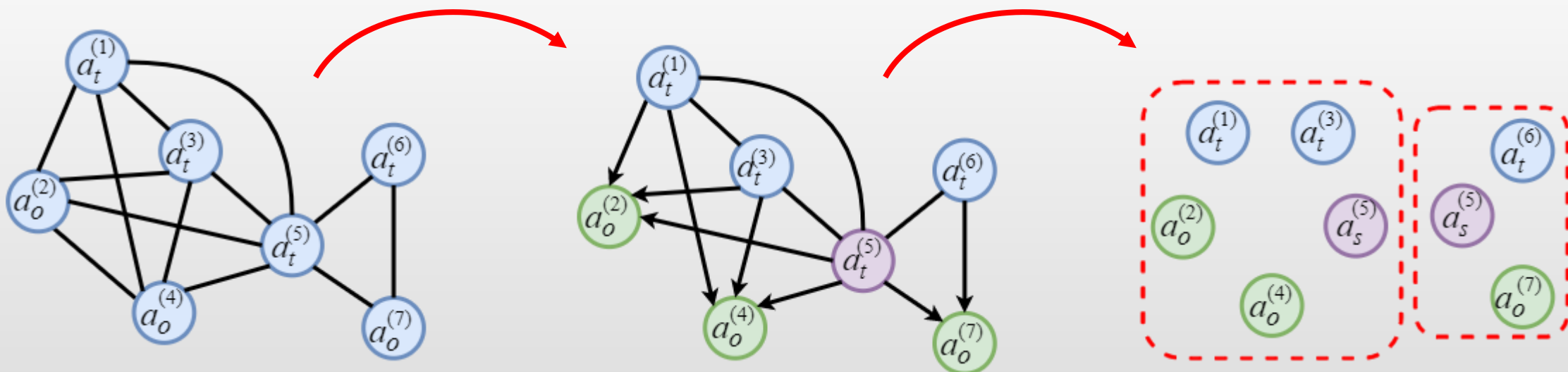


Pruned Complete Graph
w/o Annotated Triggers

通过选择重要的论元，可以进行剪枝操作

如何选择最“重要”的论元，
从而构建为剪枝完全图？

如何根据构建的图解码出一个
事件组合？



触发词在事件抽取中究竟承担什么角色？或者有什么特征？

- 存在性 (Existence)：触发词在实例中必须存在，从而指示 (identify) 事件实例
- 区分性 (Distinguishability)：触发词不被共享，可以区分不同的事件实例

可以使用 一组论元角色 R 对应的 论元 作为 事件实例的 “伪触发词”

$$\text{一组论元角色 } R \text{ 的存在性} = \frac{R \text{ 对应的论元中, 至少有一个论元在实例中存在的数量}}{\text{整个事件实例的数量}}$$

$$\text{一组论元角色 } R \text{ 的区分性} = \frac{R \text{ 对应的论元中, 至少有一个论元在实例中存在的数量}}{\text{整个事件实例的数量}}$$

$$\text{重要性} = \text{存在性} \times \text{区分性}$$

Document \mathcal{D}_1

	
	Plankton
	Krabs
	NULL
	Dec. 16, 2016

	
	Plankton
	Sandy
	NULL
	NULL

	
	Squidward
	Pearl
	3,456,000
	Nov. 16, 2016

Document \mathcal{D}_2

	
	Patrick
	NULL
	6,800
	Jan. 16, 1993

	
	Gary
	Patrick
	NULL
	NULL

	
	NULL
	SpongeBob
	6,800
	Dec. 16, 2016

对每一组可能的角色都进行计算

For $|\mathcal{R}| = 2$
 $|\text{snail}| = 6$ records

Roles	Existence	Distinguishability	Importance	Sort ▼
{🍌, ⭐}	$\frac{1+1+1+1+1+1}{6}$	$\frac{1+1+1+1+1+1}{6}$	$1 \times 1 = 1 \rightarrow \{\text{🍌}, \text{⭐}\}$ ✓	
{🍌, 🦀}	$\frac{1+1+1+1+1+1}{6}$	$\frac{0+0+1+1+1+1}{6}$	$1 \times \frac{4}{6} = \frac{2}{3}$	
{🍌, 🦋}	$\frac{1+1+1+1+1+1}{6}$	$\frac{1+1+1+1+1+1}{6}$	$1 \times 1 = 1$	
{⭐, 🦀}	$\frac{1+1+1+1+1+1}{6}$	$\frac{1+1+1+1+1+1}{6}$	$1 \times 1 = 1$	
{⭐, 🦋}	$\frac{1+1+1+1+1+1}{6}$	$\frac{1+1+1+1+1+1}{6}$	$1 \times 1 = 1$	
{🦀, 🦋}	$\frac{1+0+1+1+0+1}{6}$	$\frac{1+1+1+1+1+1}{6}$	$\frac{4}{6} \times 1 = \frac{2}{3} \rightarrow \{\text{🦀}, \text{🦋}\}$	

$\mathcal{D}_1 =$

```

    graph TD
      Krabs --> Dec16_2016[Dec. 16, 2016]
      Plankton --> Dec16_2016
      Plankton --> Sandy
      Squidward --> Nov16_2016[Nov. 16, 2016]
      Squidward --> Pearl
      Pearl --> 3456000[3,456,000]
  
```

$\mathcal{D}_2 =$

```

    graph TD
      Jan16_1993[Jan. 16, 1993] --> Gary
      Gary --> Patrick
      Patrick --> 6800[6,800]
      Patrick --> Dec16_2016[Dec. 16, 2016]
      SpongeBob --> Dec16_2016
  
```

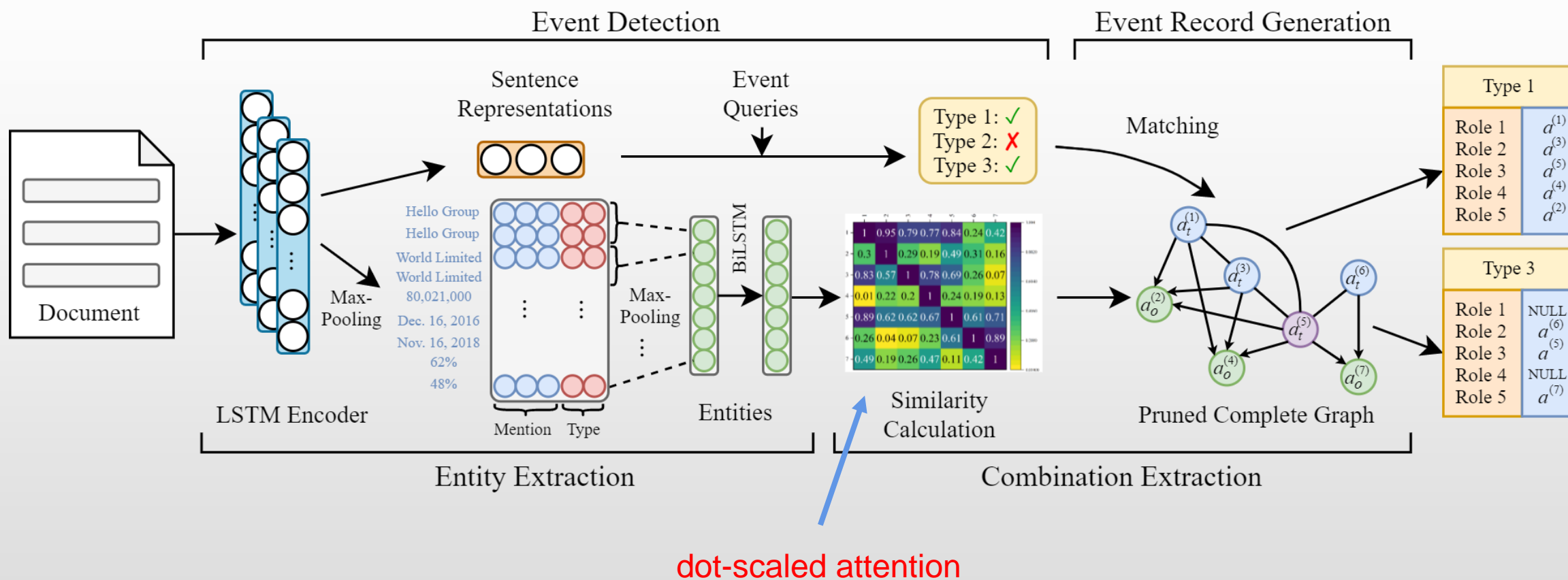
Document \mathcal{D}_1

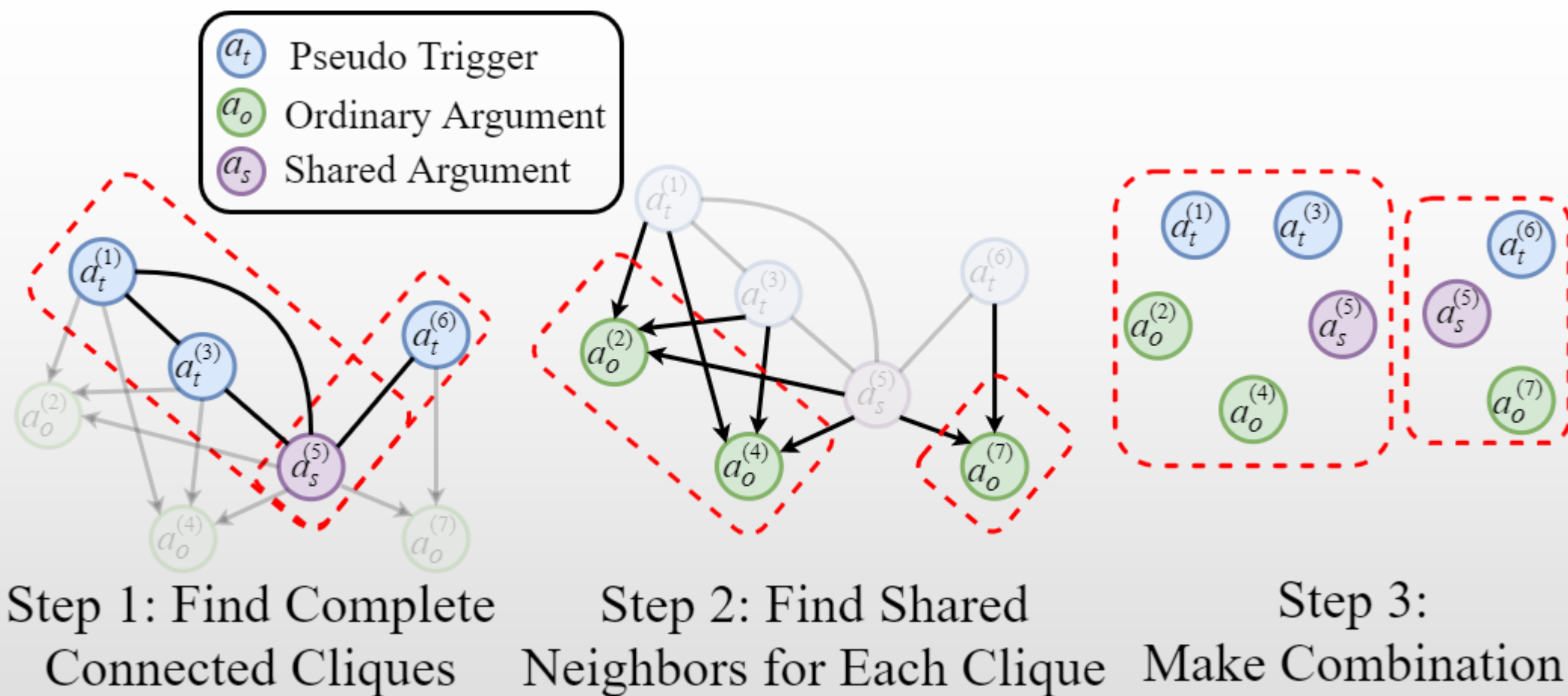
	🍌	⭐	🦀	🦋
	Plankton	Krabs	NULL	Dec. 16, 2016
	Plankton	Sandy	NULL	NULL
	Squidward	Pearl	3,456,000	Nov. 16, 2016

Document \mathcal{D}_2

	🍌	⭐	🦀	🦋
	Patrick	NULL	6,800	Jan. 16, 1993
	Gary	Patrick	NULL	NULL
	NULL	SpongeBob	6,800	Dec. 16, 2016

重要性指标最大的一组触发词
 论元角色对应的论元就是实例
 中的伪触发词





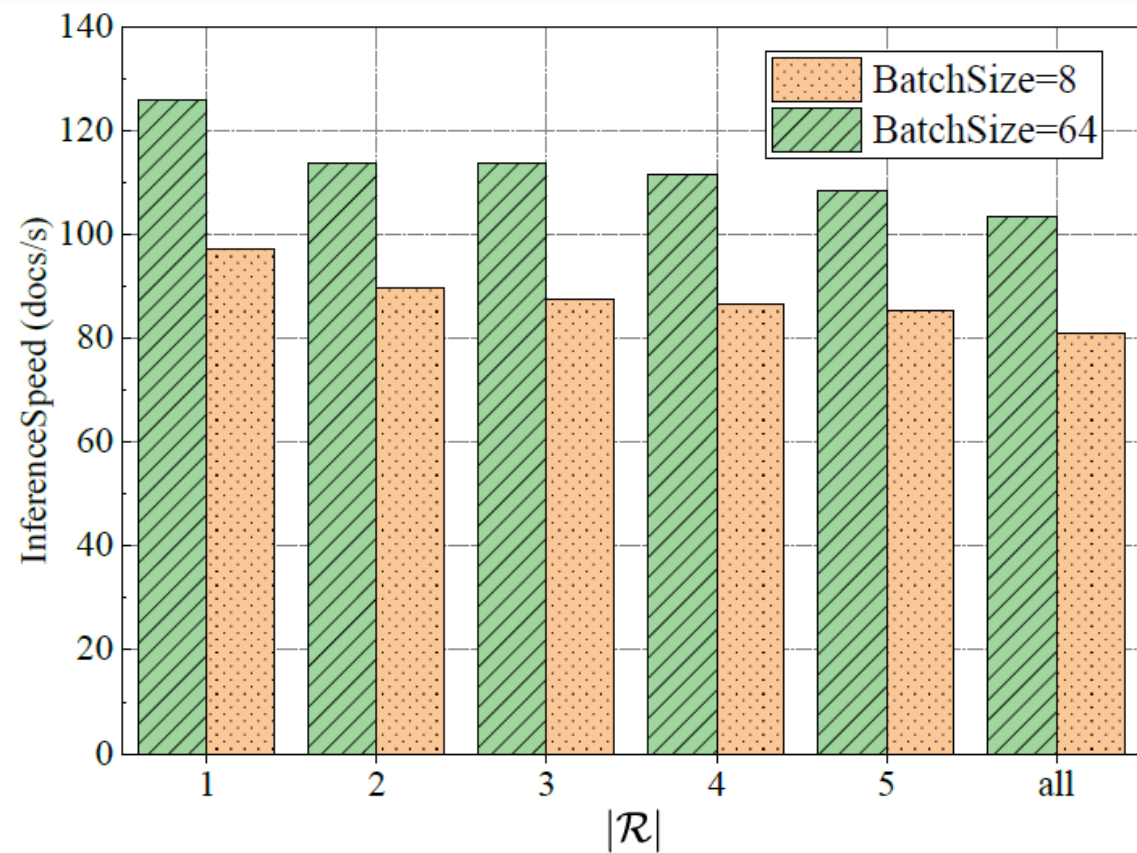
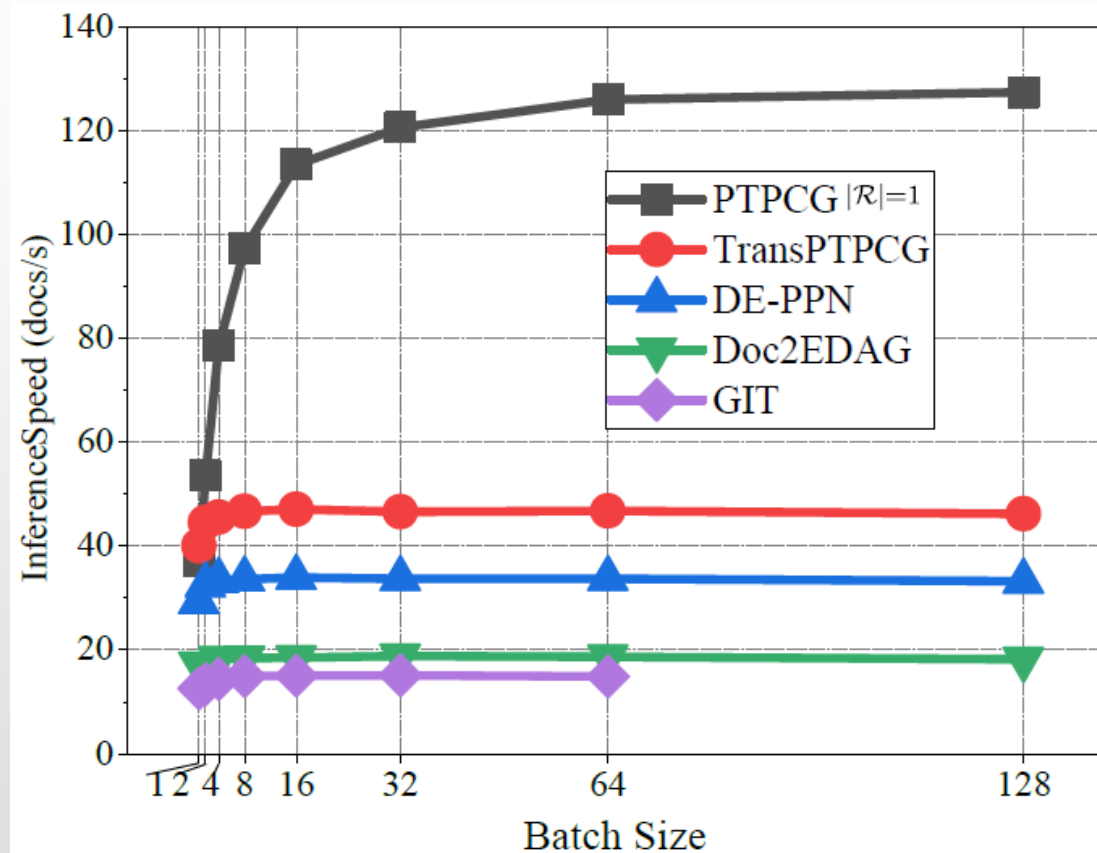
Bron-Kerbosch 算法可解

有向无环图方法

Model	#Params (w/o Emb)	GPU Hours	ChFinAnn-Single			ChFinAnn-All			DuEE-fin w/o Tgg			DuEE-fin w/ Tgg		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1
DCFEE-O*	32M (16M)	192.0	73.2	71.6	72.4	69.7	57.8	63.2	56.2	48.2	51.9	51.9	49.6	50.7
DCFEE-M*	32M (16M)	192.0	64.9	71.7	68.1	60.1	61.3	60.7	38.7	52.3	44.5	37.3	48.6	42.2
GreedyDec*	64M (48M)	604.8	83.9	77.3	80.4	81.9	51.2	63.0	59.6	41.8	49.1	59.0	42.1	49.2
Doc2EDAG*	64M (48M)	604.8	83.2	89.3	86.2	81.1	77.0	79.0	66.7	50.0	57.2	67.1	51.3	58.1
GIT*	97M (81M)	633.6	85.0	88.7	86.8	82.4	77.6	79.9	68.2	43.4	53.1	70.3	46.0	55.6
PTPCG _{\mathcal{R} =1}	32M (16M)	24.0	86.3	90.1	88.2	83.7	75.4	79.4	66.7	54.6	60.0	62.0	54.8	58.1

- 效果和DAG方法比差不多，甚至在单事件单实例上的效果比它们还好
- 模型参数非常少，只是GIT参数量的19.8%
- 训练速度非常快！单卡只要训练24小时，而GIT需要4卡训练将近1星期，GPU卡时是GIT的3.9%

折算为V100卡时，每个模型节省 **¥3658元**（按1卡时¥6计算）



Model	Tgg	$ \mathcal{R} $	Impt.	Dev			Online Test		
				P	R	F1	P	R	F1
Doc2EDAG	×	-	-	70.8	55.3	62.1	66.7	50.0	57.2
	✓	-	-	73.7	59.8	66.0	67.1	51.3	58.1
GIT	×	-	-	72.4	58.4	64.7	68.2	43.4	53.1
	✓	-	-	75.4	61.4	67.7	70.3	46.0	55.6
PTPCG	✓	0	62.9	73.5	59.4	65.7	67.0	50.1	57.3
	✓	1	93.7	68.8	64.2	66.4	62.0	54.8	58.1
	✓	2	97.1	64.7	64.9	64.8	59.1	56.5	57.8
	×	1	83.8	71.0	61.7	66.0	66.7	54.6	60.0
	×	2	94.3	63.8	64.8	64.3	60.2	58.4	59.3
	×	3	97.2	56.7	64.3	60.3	52.6	58.9	55.6

- DuEE-fin数据集包含了触发词标注，但存在触发词共享的情况（区分性不为100%）
- 使用伪触发词可以**辅助提升**结果
- **只使用伪触发词**的效果比只使用金标触发词的结果还要好！

$ \mathcal{R} $	Impt.	SE	ME	TotE	#links	Adj Acc.	F1
1	88.3	5.0	37.5	14.6	10,502	65.8	79.4
2	95.7	1.0	20.4	6.7	23,847	59.1	77.7
3	97.2	0.9	18.0	5.9	55,961	56.7	74.9
4	97.6	0.5	16.9	5.3	75,334	58.2	74.0
5	97.8	0.4	13.9	4.4	88,752	59.5	73.1
all	97.8	0.2	13.4	4.1	140,989	60.1	69.5

- 理论上限不为100%
 - 使用BK算法解码具有一定的理论误差，但当前模型的效果离理论误差的距离还很远
- 实体预测的结果对最终效果影响巨大
 - 当我们使用金标实体进行预测时，最终的整体F1值可以提升至少10%
- 随着伪触发词数量的增加，结果在不断下降
 - 主要原因：图中连接的数量随着伪触发词的增加而增加，预测难度也在不断加大
 - 我们需要对相似度计算和连接预测部分做进一步的优化

Visualisation on DocEE
Home
Instructions

Example
关于持股5%以上的股东部分股份质押的公告
Model
End-to-End
Tag Me
Clear

Text

恒通科技：关于持股5%以上的股东部分股份质押的公告 证券代码：300374 证券简称：恒通科技 公告编号：2018-026 北京恒通创新赛木科技股份有限公司 关于持股5%以上的股东部分股份质押的公告 北京恒通创新赛木科技股份有限公司（以下简称“公司”）近日接到公司持 股5%以上的股东 霍尔果斯恒隆德庆股权投资有限公司 质押方 （以下简称“恒隆德庆”）的函告，获悉其所持有本公司的部分股份被质押，具体事项如下：

一、股东股份质押的基本情况 1、股东股份被质押基本情况股东名称是否为第一大股东 及一致行动人质押股数质押开始日期质押到期日质权人本次质押占其所持股份比例用途恒隆德庆否 6620000 质押金额 2016年11月29日 质押开始日期 办理解除质押登记手续之日 光大证券股份有限公司 接收方 16.51% 融资 12180000 质押金额 2017年3月13日 质押开始日期 办理解除质押登记手续之日 光大证券股份有限公司 接收方 30.38% 融资 5600000 质押金额 2018年1月29日 质押开始日期 办理解除质押登记手续之日 光大证券股份有限公司 接收方 13.97% 融资 800000 质押金额 2018年2月6日 质押开始日期 办理解除质押登记手续之日 光大证券股份有限公司 接收方 2.00%补充质押 2、股东股份累计被质押的情况 截至公告披露日，恒隆德庆共持有公司股份40085760股，占公司股份总数的16.30%；其中处于质押状态的股份为25200000股，占其持有本公司股份总数的62.87%，占公司总股本的10.25%。

二、备查文件 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。1、中国证券登记结算有限责任公司股份冻结明细北京恒通创新赛木科技股份有限公司董事会 2018年2月7日 年月日股万股

Predicted Types

股权质押

Predicted Instance(s)

股权质押		股权质押		股权质押	
Role	Argument	Role	Argument	Role	Argument
质押金额	6620000	质押金额	12180000	质押金额	5600000
质押方	霍尔果斯恒隆德庆股权投资有限公司	质押方	霍尔果斯恒隆德庆股权投资有限公司	质押方	霍尔果斯恒隆德庆股权投资有限公司
质押开始日期	2016年11月29日	质押开始日期	2017年3月13日	质押开始日期	2018年1月29日
接收方	光大证券股份有限公司	接收方	光大证券股份有限公司	接收方	光大证券股份有限公司

股权质押	
Role	Argument
质押金额	800000
质押方	霍尔果斯恒隆德庆股权投资有限公司

<https://github.com/Spico197/DocEE>

<http://hlt.suda.edu.cn/docee>

Thanks Q&A

Tong Zhu
tzhu7@stu.suda.edu.cn

<https://github.com/Spico197/DocEE>