FDDC 2018

2018全球金融数据探索与发现大赛

FINANCIAL DATA DISCOVERY COMPETITION

A股上市公司公告信息抽取
从具体到抽象

队伍名称：gogogo
当前方向：医疗大数据、医疗AI

赛题（Competition question）

公告类型	主键	第1列	第2列	第3列	第4列	第5列	第6列	第7列	第8列
股东增减持	1-2-4	公告id	股东全称	股东简称	变动截止日期	变动价格	变动数量	变动后持股数	变动后持股比例
重大合同	1-2-3	公告id	甲方	乙方	项目名称	合同名称	合同金额上限	合同金额下限	联合体成员
资产重组	1-2-3	公告id	交易标的	标的公司	交易对方	交易标的作价	评估方法		
定向增发	1-2	公告id	增发对象	增发数量	增发金额	锁定期	认购方式		

4 kinds of Listed Company Announcements and 28 fields to be extracted.

评估方法： $F1 = \frac{2 \times Recall \times Precision}{Recall + precision}$

Evaluation method

赛题（Competition question）

杭州泰格医药科技股份有限公司（以下简称“公司”）于2014年6月5日接到股东QM8 LIMITED（以下简称“QM8”）的告知函，QM8于2014年6月4日通过深圳证券交易所大宗交易系统累计减持公司无限售流通股2,680,000股，减持数量占公司总股本比例为1.255%，具体情况如下：

一、股东减持

1、股东减持股份情况

股东名称	减持方式	减持时间	减持均价（元）	减持股数（股）	减持比例（%）
QM8	大宗交易	2014-06-04	29.76	2,680,000	1.255

2、股东本次减持前后持股情况

股东名称	股份性质	减持前持有股份		减持后持有股份	
		股数（股）	占总股本比例（%）	股数（股）	占总股本比例（%）
QM8	全部为流通股	19,825,920	9.282	17,145,920	8.027

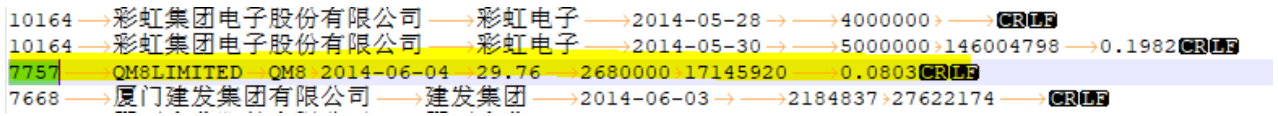
PDF

杭州泰格医药科技股份有限公司（以下简称“公司”）于2014年6月5日接到股东QM8 LIMITED（以下简称“QM8”）的告知函，QM8于2014年6月4日通过深圳证券交易所大宗交易系统累计减持公司无限售流通股2,680,000股，减持数量占公司总股本比例为1.255%，具体情况如下：

股东名称	减持方式	减持时间	减持均价（元）	减持股数（股）	减持比例（%）
QM8	大宗交易	2014-06-04	29.76	2,680,000	1.255

股东名称	股份性质	减持前持有股份		减持后持有股份	
		股数（股）	占总股本比例（%）	股数（股）	占总股本比例（%）
QM8	全部为流通股	19,825,920	9.282	17,145,920	8.027

HTML

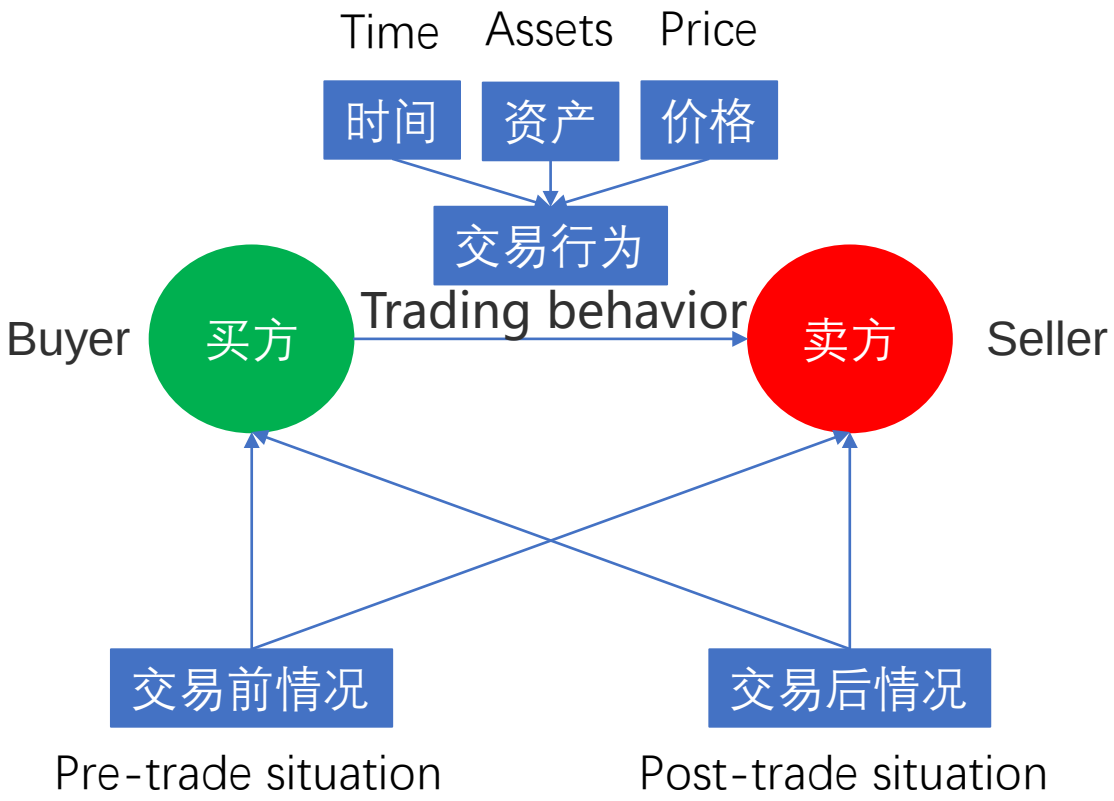


结果 Results

赛题理解(Understanding question)

大部分**资产交易类**公告内容都可以抽象为如下的模型

Most of the contents of asset trading announcements can be abstracted into the following model



事件核心为买卖双方和资产，优先提取

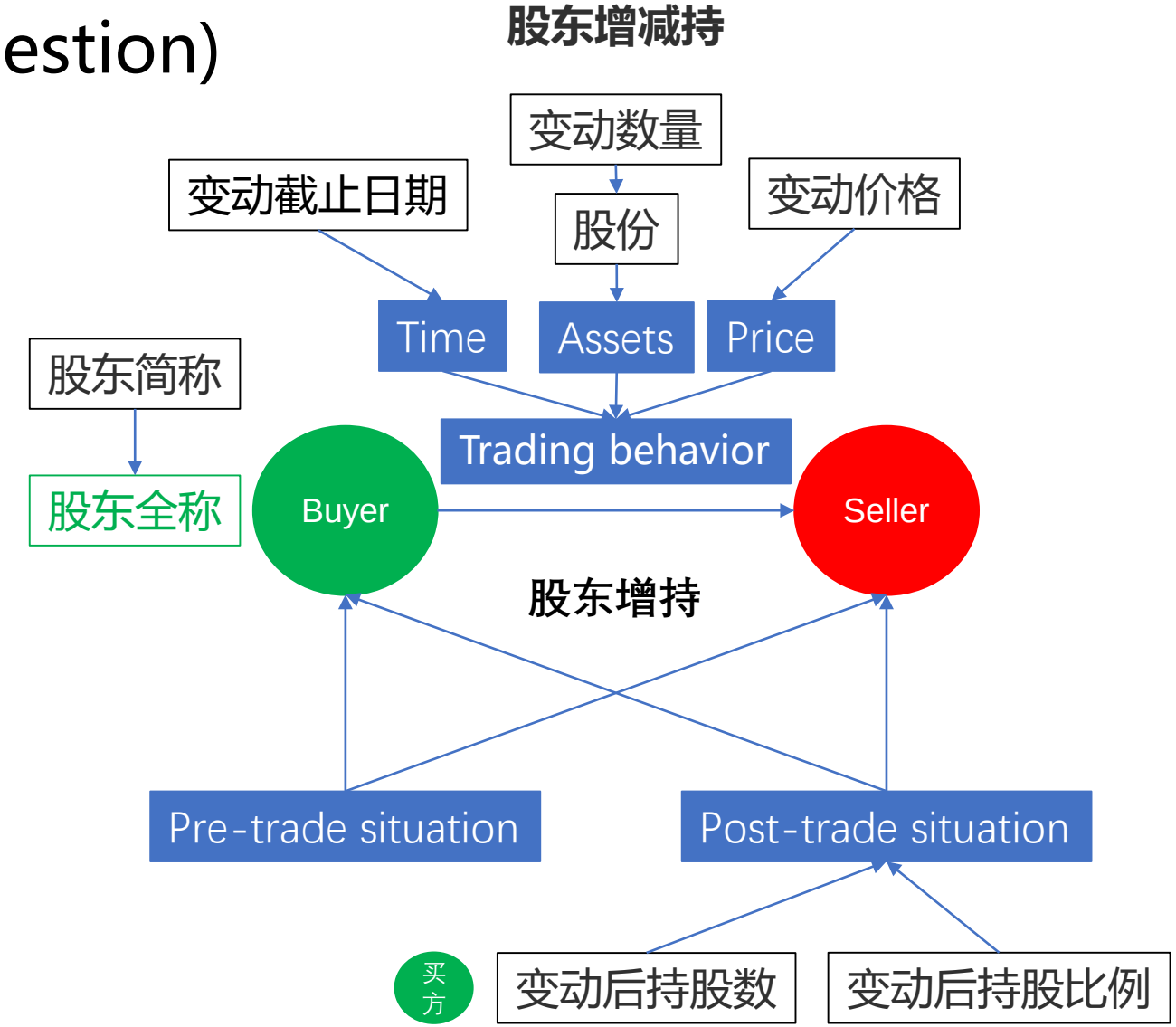
The key of the event is the buyer and seller and the assets, priority extraction.

赛题理解(Understanding question)

核心行为拆解(Core behavior dismantling)

股东买股份(Shareholders buy shares)

股东卖股份(Shareholders sell shares)

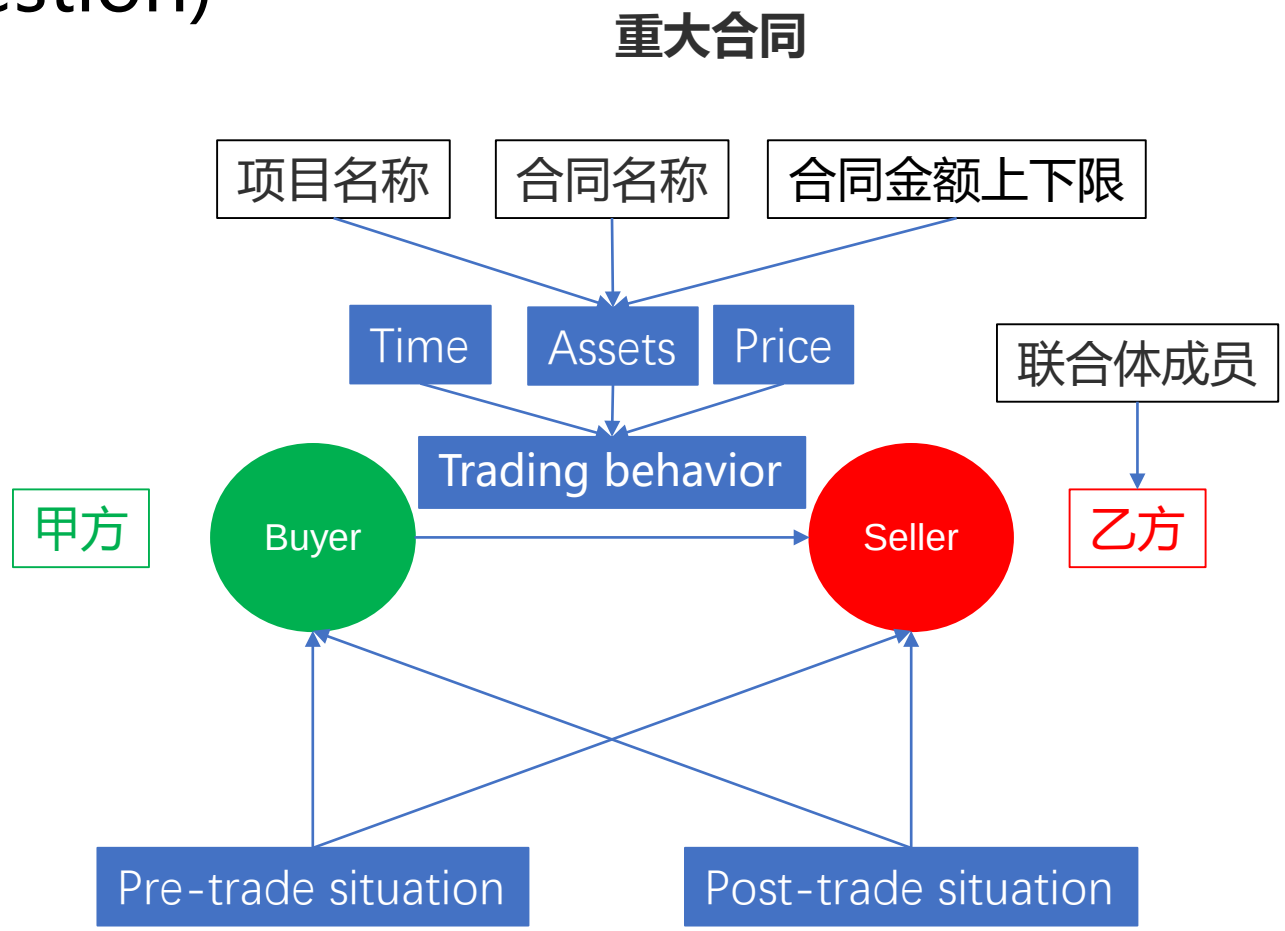


赛题理解(Understanding question)

核心行为拆解(Core behavior dismantling)

甲方买项目（合同）
Party A buys the project(contract)

乙方卖项目（合同）
Party B sells the project(contract)



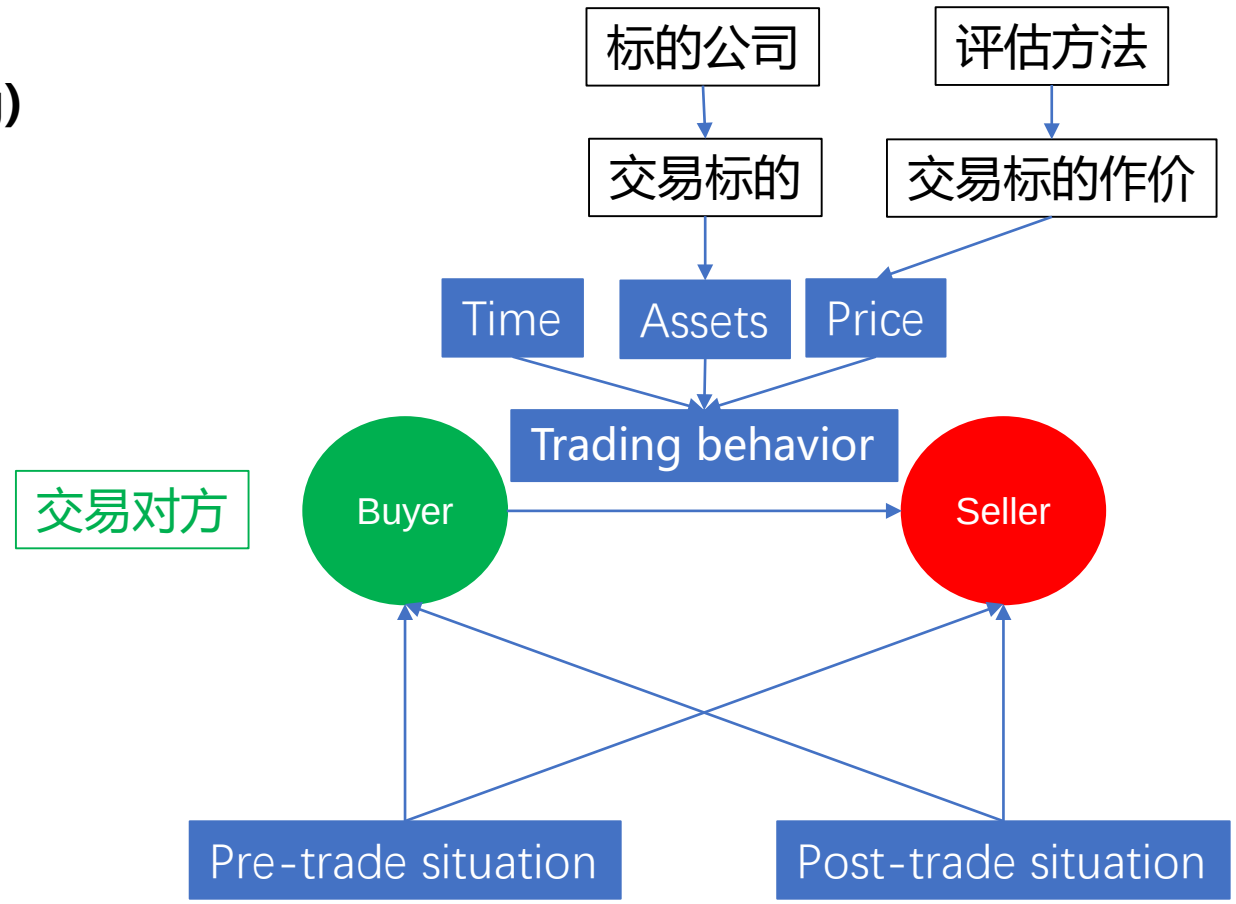
赛题理解(Understanding question)

核心行为拆解(Core behavior dismantling)

公司买标的(Company buys target)

公司卖标的(Company sells target)

公司拿自己的资产换交易对方资产
Asset replacement

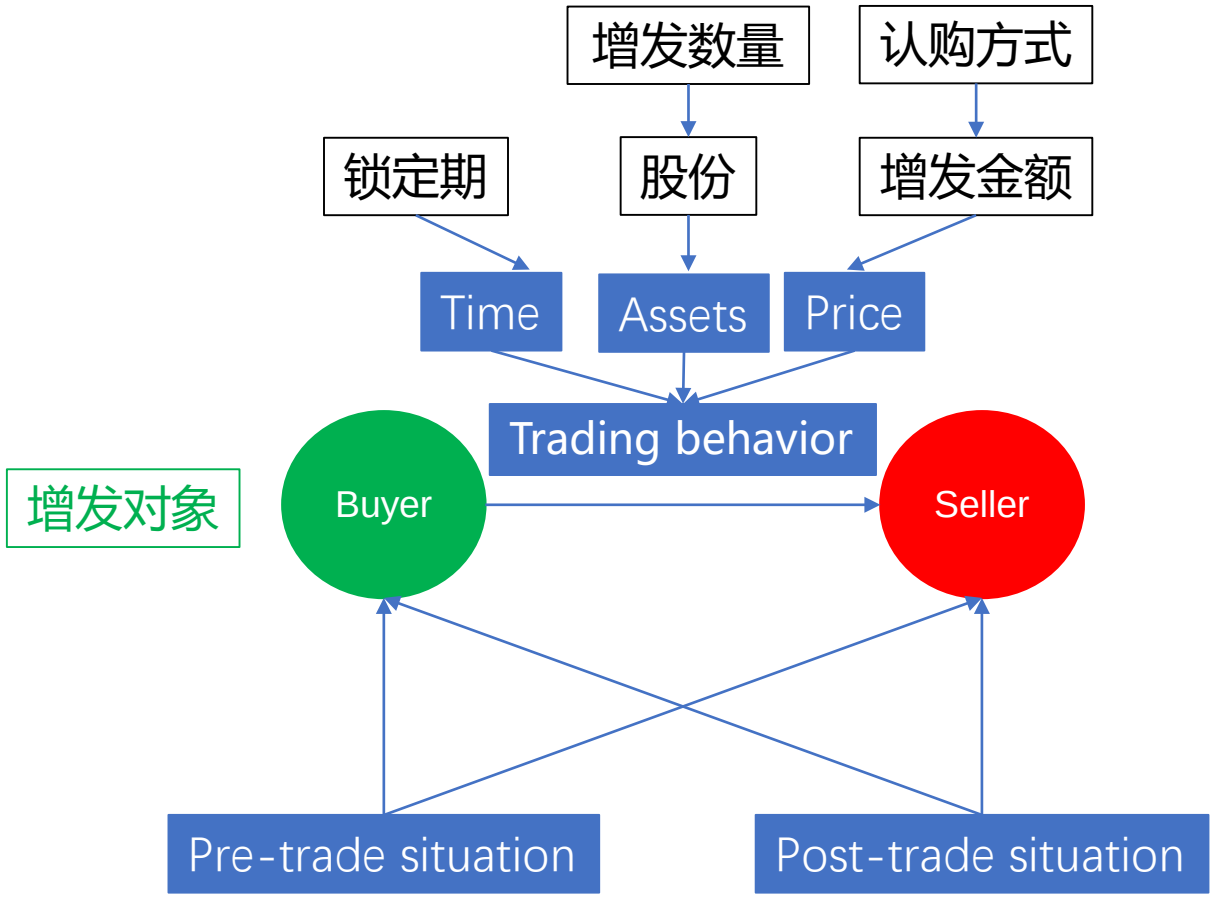


赛题理解(Understanding question)

定向增发

核心行为拆解(Core behavior dismantling)

股东买股份(Shareholders buy shares)



数据集情况(DataSets)

训练集(Training set)

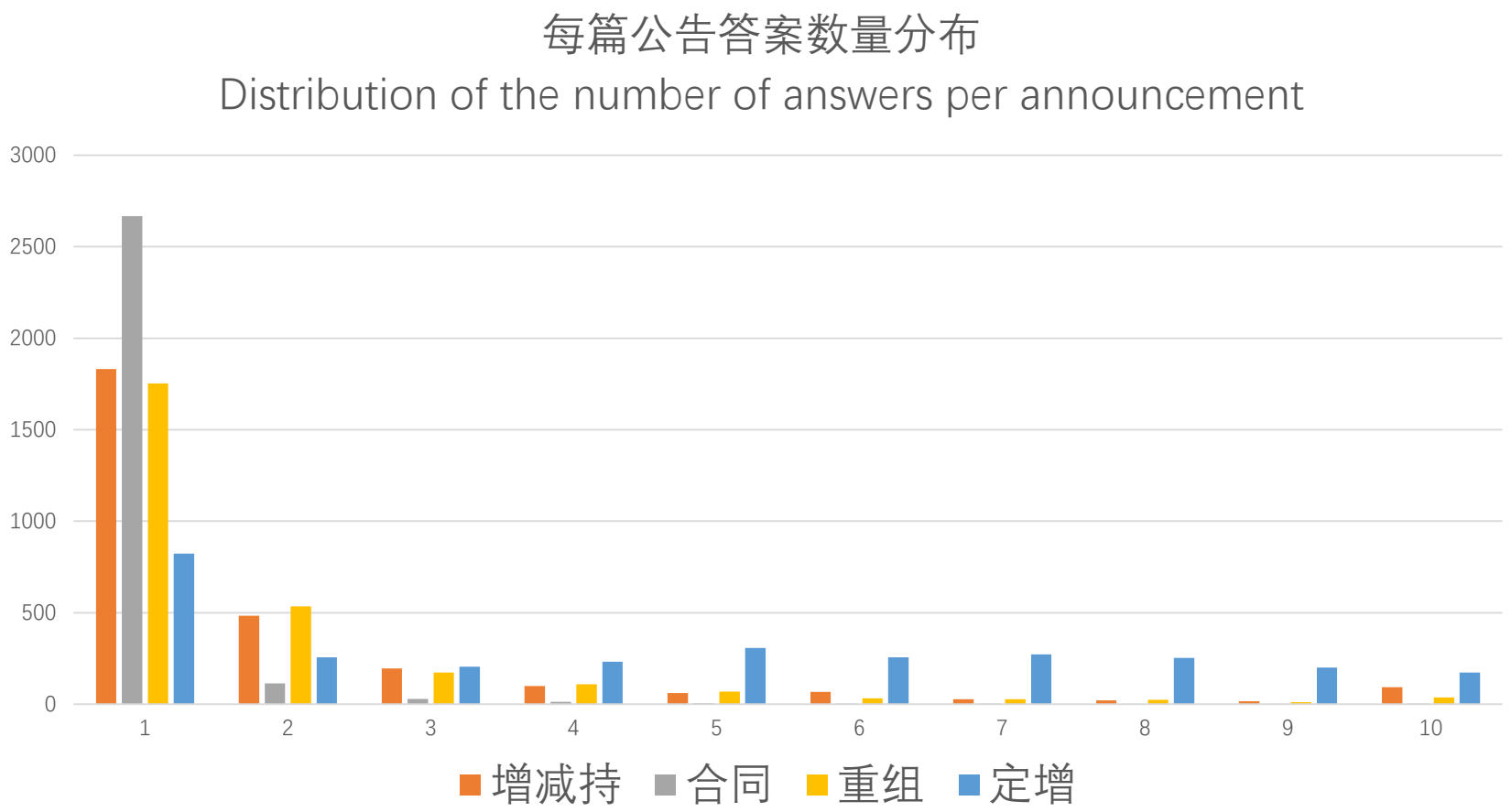
公告类型 Types	文档数量 Doc Num	平均PDF大小(KB) Avg PDF Size(KB)	平均HTML大小（KB） Avg HTML Size(KB)	答案数量 Answer Num
股东增减持	3000	130	6	6452
重大合同	3000	139	6.1	3186
资产重组	2772	3322	2182	5471
定向增发	3000	716	370	13499

结果(Results)

测试集结果(Testing set result)

公告类型 Types	测试数据集 Testing Set	F1	训练集答案数量 Training set Num
股东增减持	FDDC_announcements_round1_test_b_20180708	0.74	6452
重大合同	FDDC_announcements_round1_test_b_20180708	0.53	3186
资产重组	FDDC_announcements_round2_test	0.55	5471
定向增发	FDDC_announcements_round1_test_b_20180708	0.81	13499

训练集答案分布情况（Training set answer distribution）



文档撰写逻辑分析（Document writing logic analysis）

一个事件怎么表述?
How to describe an event?

两三个事件怎么表述?
How to describe the two or three events?

很多事件怎么表述?
How to describe many events?

文档撰写逻辑分析（Document writing logic analysis）

一个事件: 在一个或少数段落中直接陈述
An event: direct describe in one or a few paragraphs.

近日,公司收到 山东电力集团公司 甲方 发出的中标通知书,通知公司中标 山东中心智能化仓储系统项目 项目名称 ,公司为该项目提供自动化立体仓库、自动输送及分拣系统等,中标金额 4245.380001 万 合同金额 元(肆仟贰佰肆拾伍万叁仟捌佰圆零壹分),占2011 年度经审计营业总收入5.42%。

两三个事件怎么表述：在一个或少数段落中使用并列关系的连词（及、和、分别、顿号等）
How to describe two or three events: use conjunctions (and, respectively, and pauses) in one or a few paragraphs.

万马电缆分别向 电气电缆集团 NAME2 、 张德生 NAME2 、 金临达实业 NAME2 发行股份购买其持有的 万马高分子 NAME1 100%股权 NAME0 ;向 电气电缆集团 NAME2 、 潘玉泉 NAME2 、 张云 NAME2 发行股份购买其持有的 天屹通信 NAME1 100%股权 NAME0 ;向 电气电缆集团 NAME2 、 王一群 NAME2 、 普特实业 NAME2 发行股份购买其持有的 万马特缆 NAME1 100%股权 NAME0 。

文档撰写逻辑分析（Document writing logic analysis）

很多事件怎么表述：采用表格或列举形式
How to describe many events: table or list form

按照本次发行股票数量上限 3,600 万股测算，本次非公开发行前后公司股本变动情况具体如下：

股东名称	本次发行前		本次发行后	
	持股数量（万股）	持股比例	持股数量（万股）	持股比例
巴玛投资	-	-	3,600.00	20.08%
施延军	3,970.20	27.70%	3,970.20	22.14%
施延助	1,462.50	10.20%	1,462.50	8.16%
施雄飚	975.00	6.80%	975.00	5.44%
施文	975.00	6.80%	975.00	5.44%
薛长煌	760.00	5.30%	760.00	4.24%
其他股东	6,189.80	43.19%	6,189.80	34.52%
合计	14,332.50	100%	17,932.50	100.00%

近日，本公司中标以下重大工程：

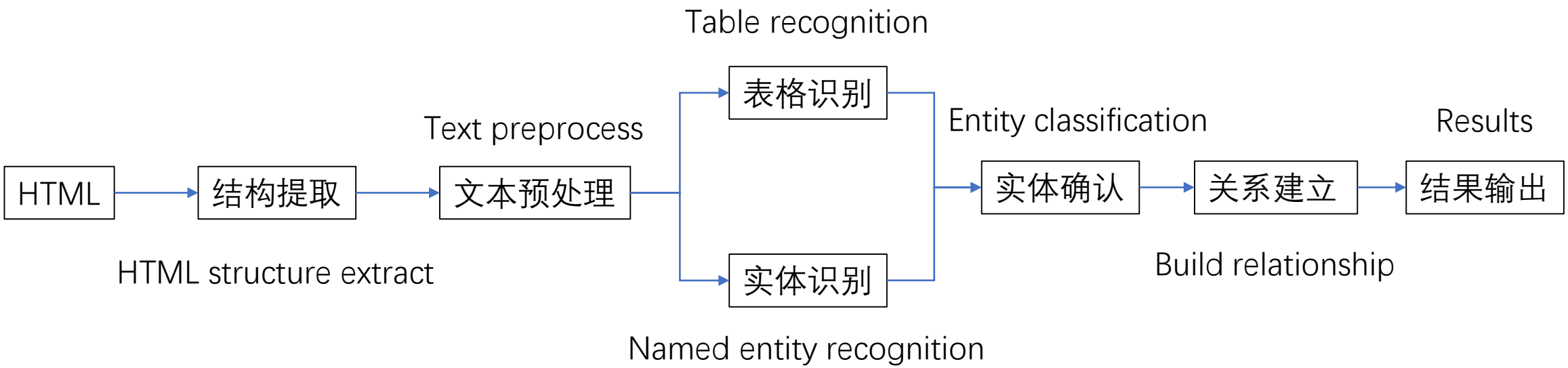
一、本公司收到杭甬铁路客运专线建设筹备组发出的中标通知书，通知本公司所属子公司中铁十七局集团有限公司、中铁二十四局集团有限公司分别中标新建杭州至宁波铁路客运专线工程第 HYZQ-2、HYZQ-3 标段，中标价合计人民币 624,917.84 万元，约占本公司中国会计准则下 2007 年营业收入的 3.52%。

二、本公司收到宁杭铁路有限责任公司发出的中标通知书，通知本公司所属子公司中铁十七局集团有限公司、中铁二十四局集团有限公司分别中标新建南京至杭州铁路客运专线站前及相关工程第 NHZQ-3、NHZQ-4 标段，中标价合计人民币 871,250.47 万元，约占本公司中国会计准则下 2007 年营业收入的 4.91%。

三、本公司收到天平铁路有限公司发出的中标通知书，通知本公司所属子公司中铁二十一局集团有限公司、中铁十九局集团有限公司、中铁建电气化局集团有限公司分别中标新建铁路天水至平凉线工程第 TP-TJ1、TP-TJ2、TP-ZH 标段，中标价合计人民币 380,919.39

四、本公司收到南宁铁路局发出的中标通知书，通知本公司所属子公司中铁十二局集团有限公司、中铁二十三局集团有限公司、中铁十一局集团有限公司、中铁二十五局集团有限公司分别中标湘桂铁路永州至柳州段扩能改造工程第 XG-3、XG-5、XG-6、XG-7 标段，中标价合计人民币 935,092.49 万元，约占本公司中国会计准则下 2007 年营业收入的 5.27%。

抽取模型建立(Extraction model)



抽取模型建立——结构提取(Extraction model-HTML structure extract)

标签层级	标签名称	释义	子属性	子属性释义
0		公告标签	title	公告标题
1		公告段落	title	段落标题
2		公告正文		
3	<hidden>	页码	name	页码计数，从0开始
3	<table>	表格		
4	<tbody>	表格		
5	<tr>	表格行		
6	<td>	表格列(单元格)	rowspan	单元格跨度
7	<image>	图片		

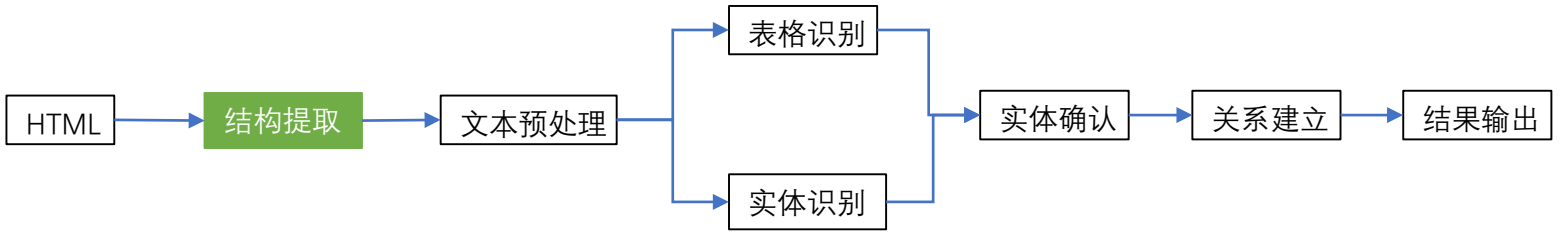
- 标签层级序号大的标签中不可以包含层级低的标签
- 段落(paragraph)中可以包含子段落
- 正文(content)中不会再包含content
- 表格(table)，页码(hidden) 必须在content标签中
- 表格(table)中只有一个tbody标签，并且不会包含子表格
- 一个表格(tbody)会有多行(tr)，一行(tr)会有多列(td)
- 一个表格行(tr)或表格列(td)中不会再有其他的tr或td
- 图片(image)必须在content标签或td标签中，image在本次比赛中无需使用，可以直接过滤

工具(Tools)：Beautifulsoup

表格：表头识别、表格跨页合并、rowspan处理
Table: Header、Merge cross-page table、rowspan recognition

段落标题：数字开头、内容矫正
Paragraphs title：Starting with numbers、Content correction

图片：删除
Images：Delete



抽取模型建立——文本预处理(Extraction model-Text preprocess)

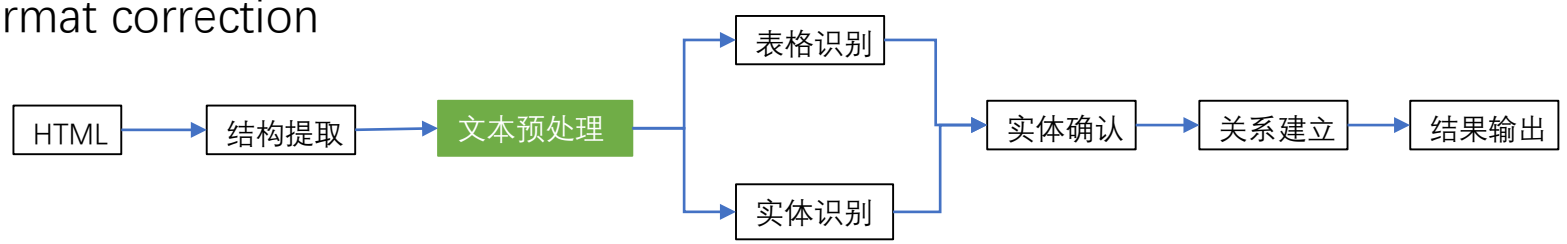
全角字符转换：％ -> %
Full-width character conversion

空行、空格删除： 2007年 12月4 日，本公司接第一大股东
Blank line, space deletion

数字格式转换：68,059,079股 -> 68059079
Number format conversion

数字单位转换：合计金额不低于4亿元人民币 ->合计金额不低于 400000000元人民币
unit conversion

格式纠错：表格转换错误，语句非正常断句等（HTML转换问题）。
Format correction



抽取模型建立——表格识别(Extraction model-Table recognition)

表格单位
Unit

减持期间	减持均价（元）	减持股数（万股）	减持比例
13 年 7 月 19 日	10.85	200.00	0.31%

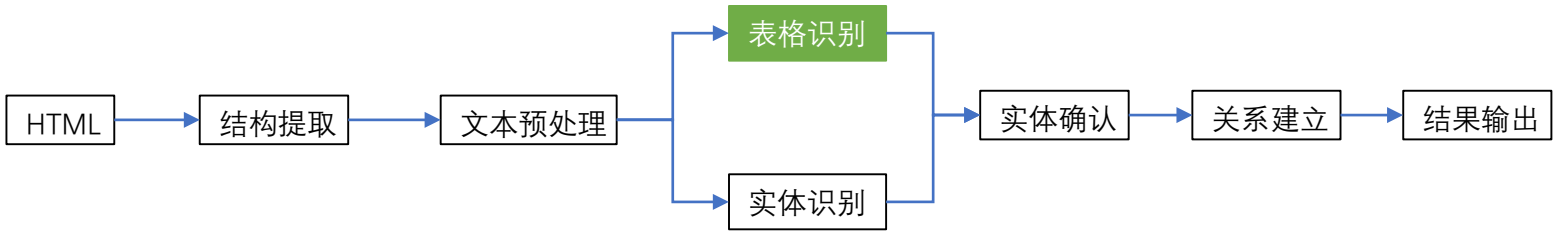
表格上下文语义
Table context semantics

1、本次减持股份情况

股东名称	减持方式	减持期间	减持均价(元)*	减持股数（股）	减持比例(%)
复星医药产业	集中竞价	2014 年 1 月 9 日 -2014 年 5 月 26 日	22.82	1,210,061	1.00
	合　计	/	22.82	1,210,061	1.00

目标表格表头学习
header learning

股东全称
姓名
增持人
持有人
限售股份持有人名称
减持股东
增持方
增持人名称
股东姓名
。。。

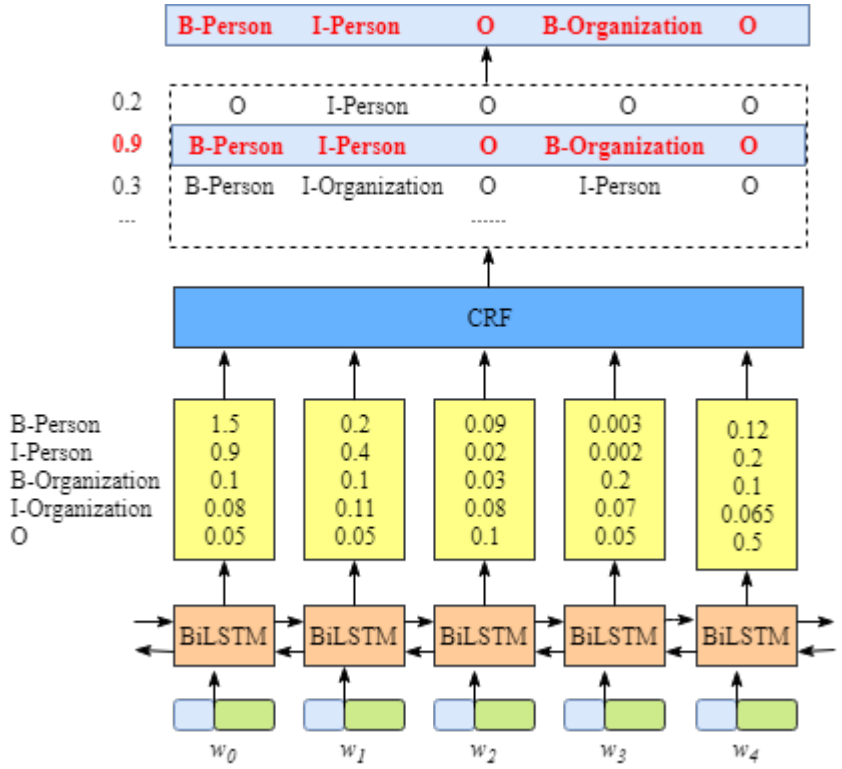


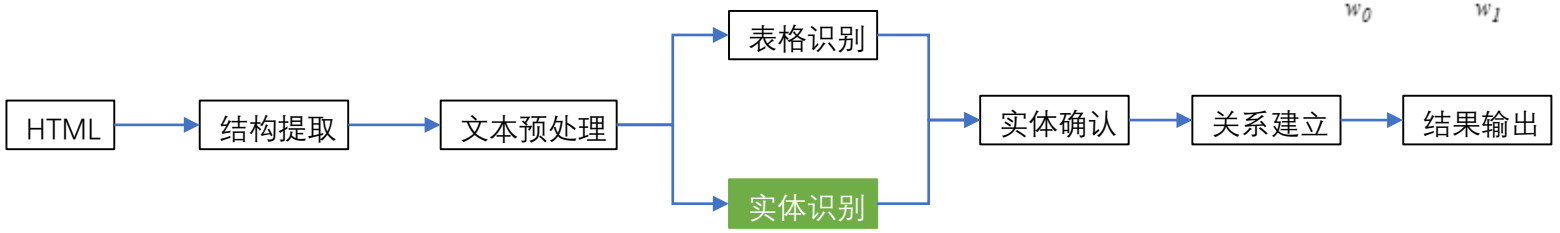
抽取模型建立——实体识别(Extraction model-Named entity recognition)

1、训练集反向标注实体建立实体识别训练集
Annotation entity create entity recognition training set

2、BiLSTM-CRF训练NER模型(Tensorflow)
Train BiLSTM-CRF NER model

3、调用模型预测实体
Predict NER

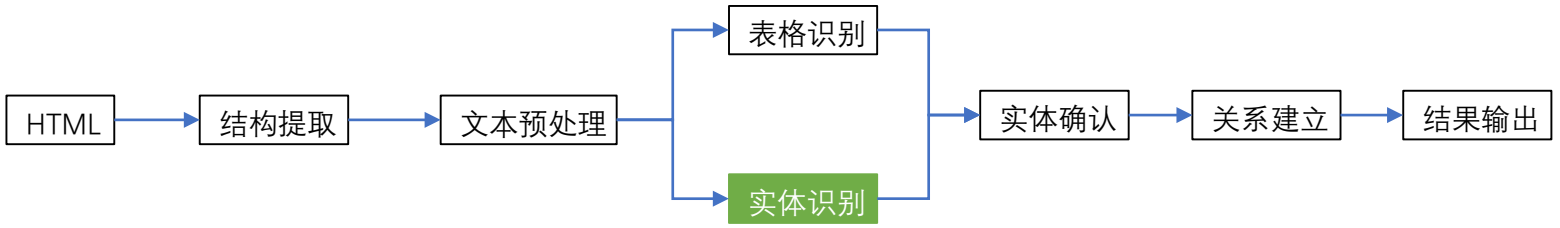




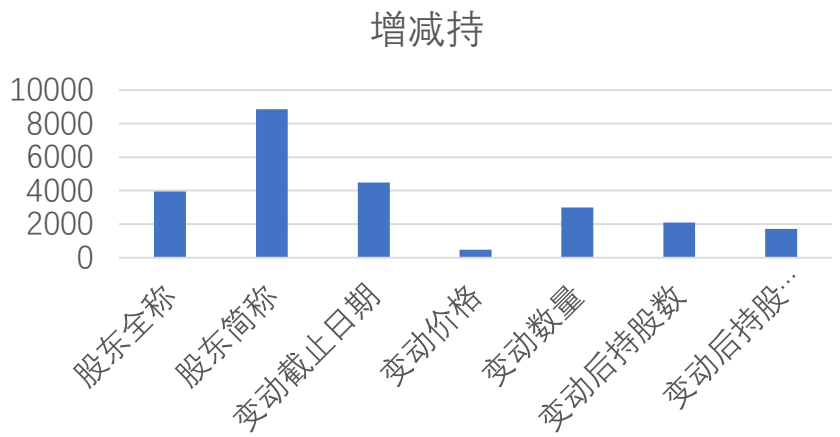
抽取模型建立——实体识别(Extraction model-Named entity recognition)

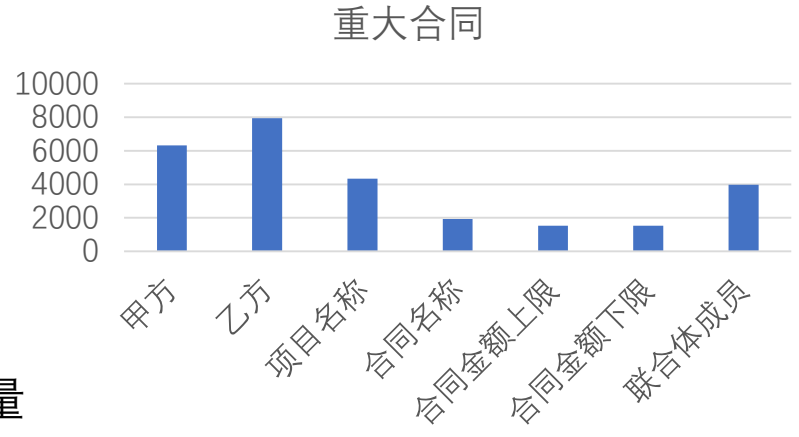
绿色为标记的实体(Green are annotation entity)

公告类型	主键	第1列	第2列	第3列	第4列	第5列	第6列	第7列	第8列	测试集F1 Test Set F1
股东增减持	1-2-4	公告id	股东全称	股东简称	变动截止日期	变动价格	变动数量	变动后持股数	变动后持股比例	90%
重大合同	1-2-3	公告id	甲方	乙方	项目名称	合同名称	合同金额上限	合同金额下限	联合体成员	80%
资产重组	1-2-3	公告id	交易标的	标的公司	交易对方	交易标的作价	评估方法			95%
定向增发	1-2	公告id	增发对象	增发数量	增发金额	锁定期	认购方式			87%

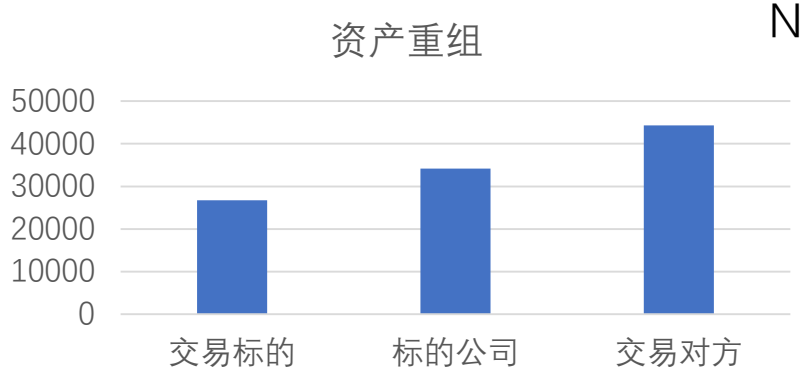


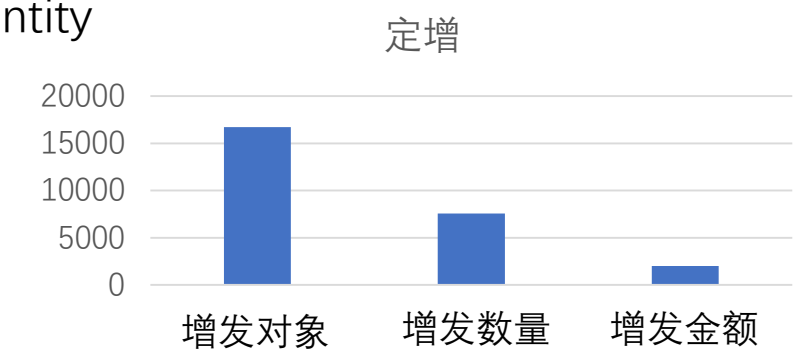
抽取模型建立——实体识别(Extraction model-Named entity recognition)

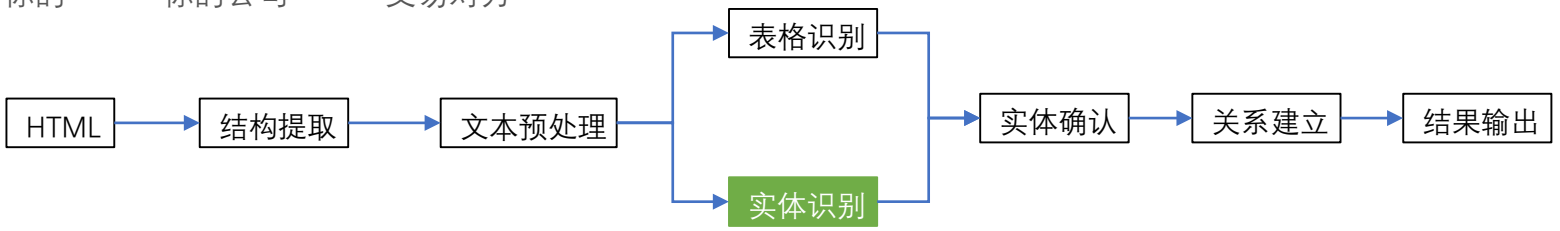




不同字段标注数量
Number of labeled entity







抽取模型建立——实体确认(Extraction model-Entity classification)

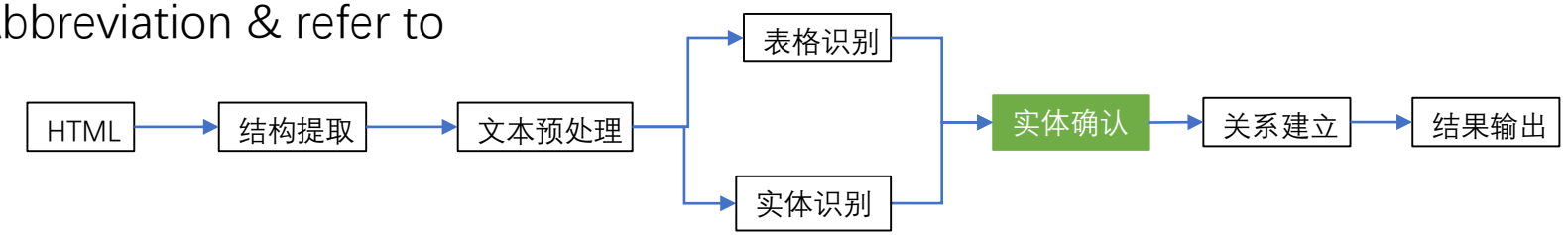
1、实体约束：明确的规则，ref [round2] FDDC_announcements_submit_notice_20180806.pdf
Rules

2、实体格式转换：日期、数量、金额，顿号分割等
Format conversion

3、实体过滤：例如全称和简称的位置关系、交易标的和标的公司的位置关系等
filter entity by location relationship

4、枚举的实体：锁定期、评估方法、认购方式
Enumeration entity

5、缩写、指代
Abbreviation & refer to



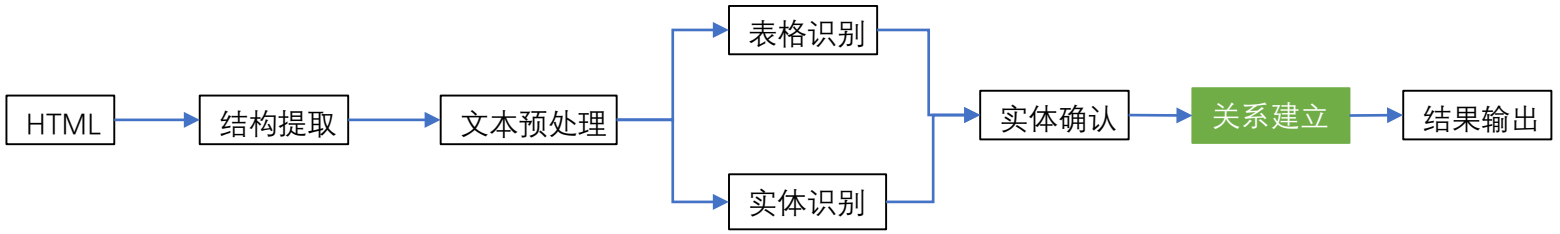
抽取模型建立——关系建立(Extraction model-Build relationship)

1、主键组合：同一个句子里面的实体组合主键
Primary key combination：both in one sentences

2、属性关联主键：主键与属性出现在同一个句子的进行组合
Attribute associated primary key

3、条件规则过滤：关键词匹配句子过滤
Sentences filter by rules

4、去重
remove duplication



模型优化(Model optimization)

1、实体标注技巧(Entity annotation skills):
按句子级别标注，同一个句子中包含所有主键则纳入训练集(primary key in one sentence)
简称、指代替换(full-short refer replacement)
信息损失，数字精度、日期简写(number precision、 date abbreviation)

2、奥卡姆剃刀：当公告结果只有一个主键时，其属性值不会产生歧义，往往属性不与主键在一个句子。
Occam's Razor: Express with minimal statements when there is no ambiguity(one result)

3、语义纠错：明显的单位错误，例如“万元”写成“元”，通过值域判断。
Semantic error correction: obviously wrong unit

4、篇章语义：复杂文档会根据目录结构选择性阅读，而非整篇通读。
Chapter semantics: read by paragraph titles

篇章语义(Chapter semantics)

以重组为例：第一次出现完整主键对的文章的章节标题，可以看到明显的概括性词语出现

Result always in featured paragraphs

模型验证策略(Model verification strategy)

先满足高覆盖率;
Recall first

再删除假阳性结果，提高准确率。
Then delete false positive results to improve precision

模型效率(Model efficiency)

普通水平的机器每篇文档控制在秒级;
Finish in seconds each doc

其中90%的时间花在实体识别
NER process cost 90% times

提升空间(How to improve)

1、规则转换为seq2seq的模型提高效果
Convert all rules to seq2seq model if there are enough data

2、全部实体改成BiLSTM+CRF建模
Predict all entity by BiLSTM+CRF model

3、手工矫正标注数据
Manually annotating data

4、关系抽取
Extract relationship by deep learning model

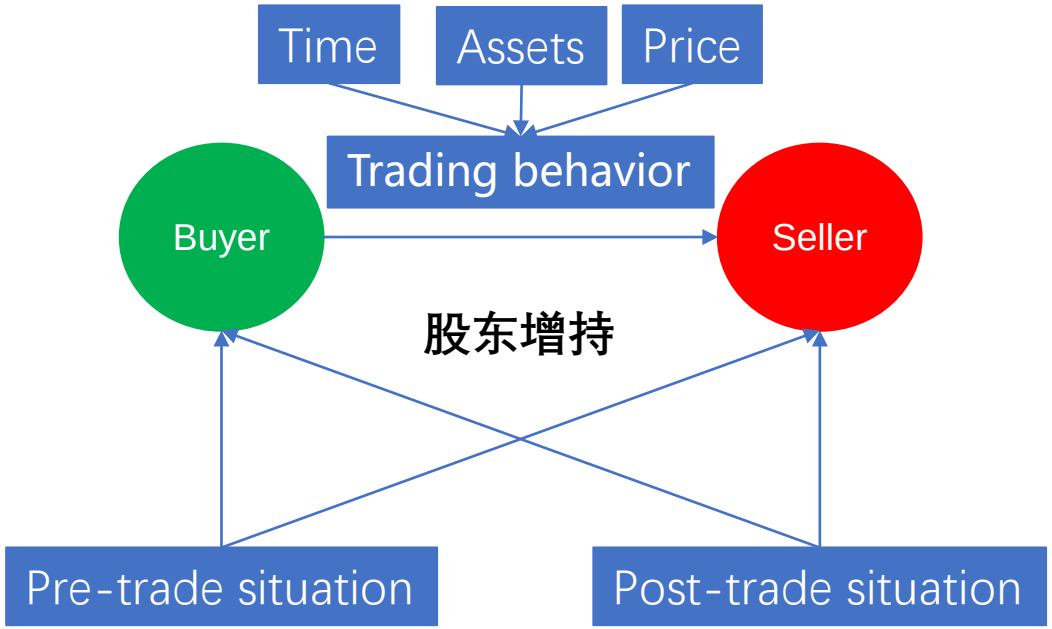
5、篇章语义建模判断结果可信度
Chapter semantics: read by paragraph titles

6、更强大的外部公司名称数据
More powerful dictionary data

展望与思考(Prospect and thinking)

大部分**资产交易类**公告内容都可以抽象为如下的模型

Most of the contents of asset trading announcements can be abstracted into the following model



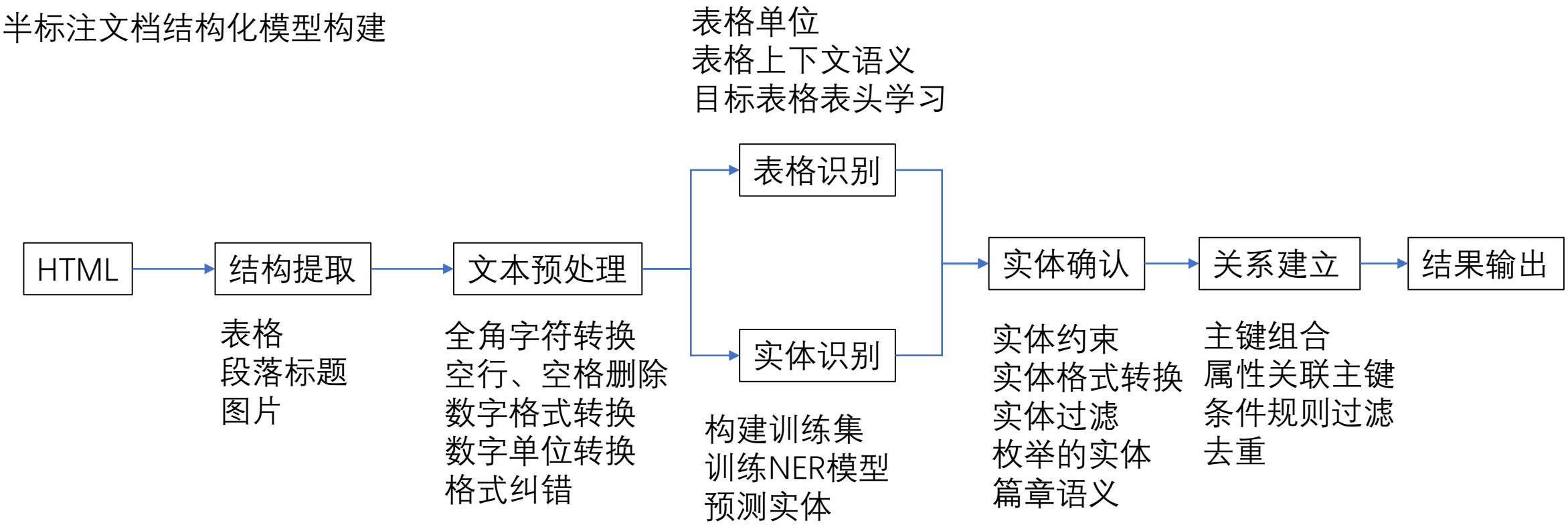
可以以事件核心为买卖双方和资产，优先提取

其他属性再与之关联

The key of the event is the buyer and seller and the assets, priority extraction.

可能的通用模型？(General model)

半标注文档结构化模型构建



致谢(Acknowledgement)

感谢主办方、评委、听众

感谢下列三方库的开发者：
Jieba
Information-Extraction-Chinese
Tensorflow
beautifulsoup


THANKS