哈爾濱工業大學
社会计算与信息检索研究中心

基于增强学习的股市涨跌预测技术

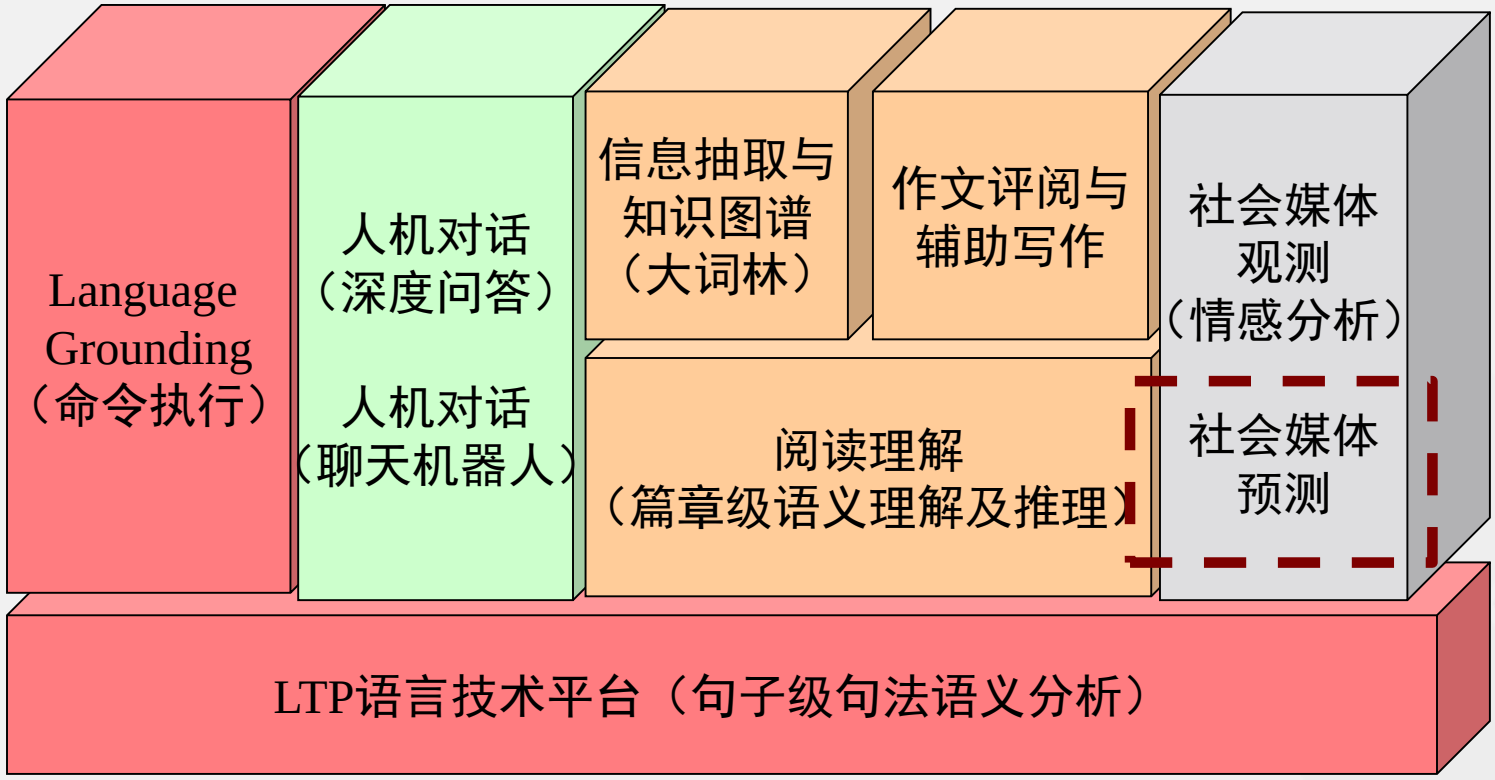
丁 效
哈工大社会计算与信息检索研究中心

2018年8月3日，中国 · 哈尔滨

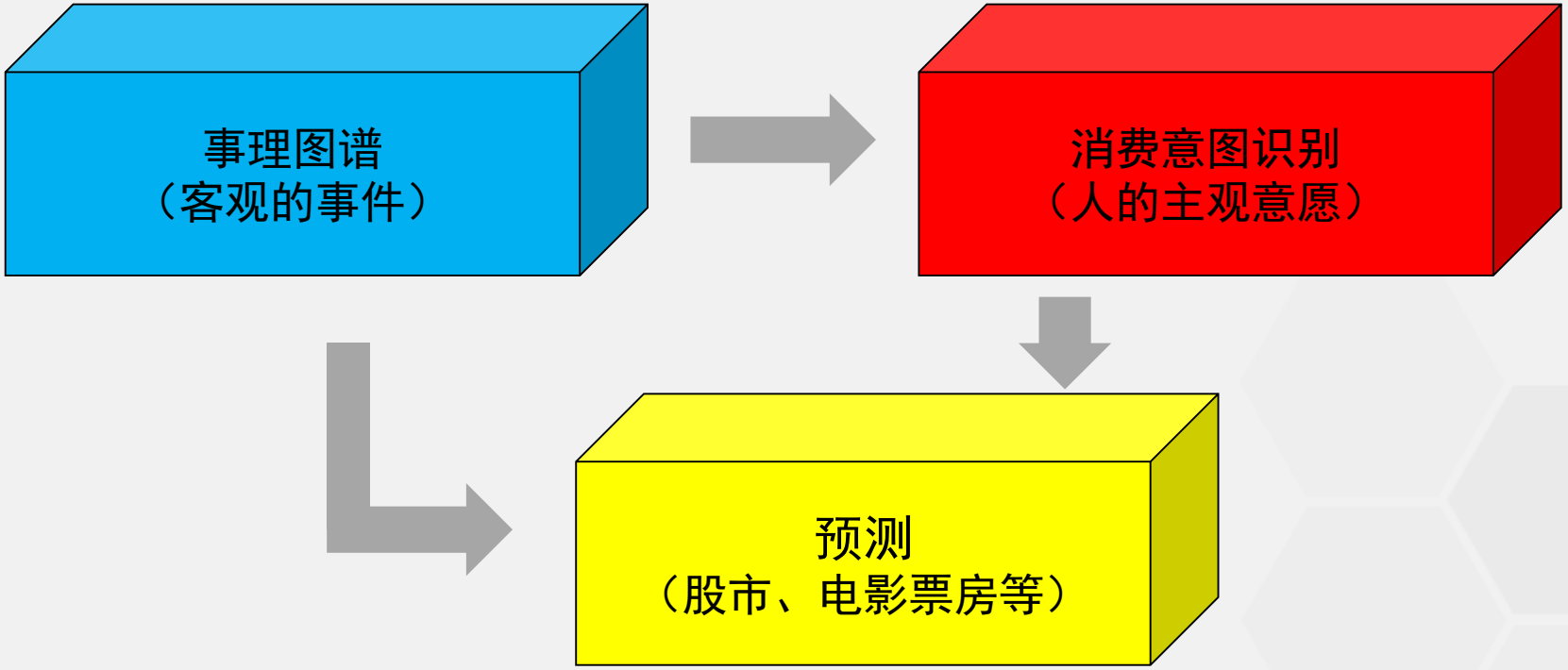
哈工大社会计算与信息检索研究中心

- 2000年9月1日，成立“信息检索研究室HIT-IRLab”
- 2011年5月18日，更名为“社会计算与信息检索研究中心HIT-SCIR”
  - SP组正式成立
- 技术理想
  - “以中文技术，助民族复兴”
- 实验室口号
  - “理解语言，**认知社会**”
- 实验室文化
  - “有爱 力行 乐学 日新”

实验室研究方向定位

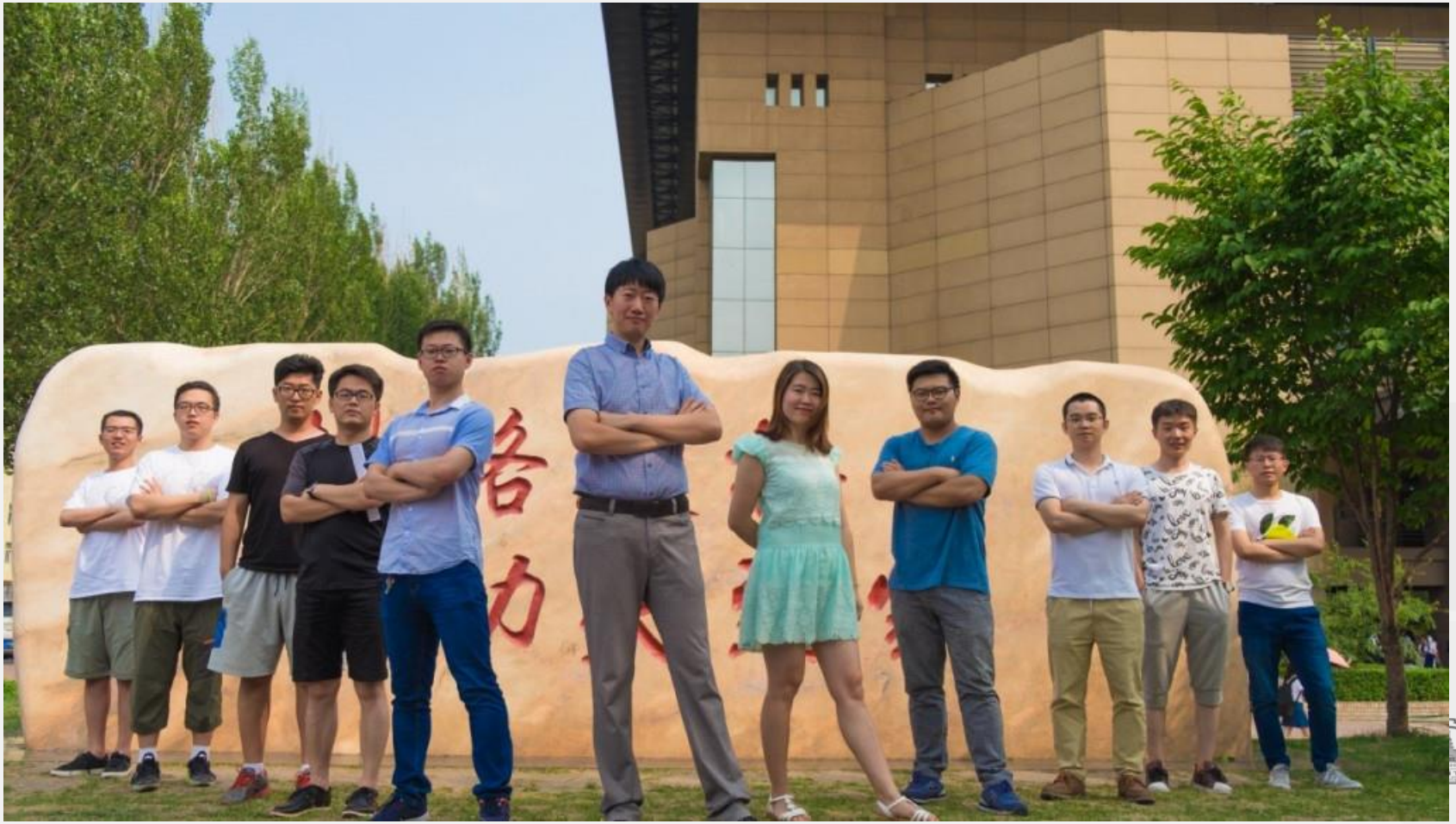

Language Grounding（命令执行）
人机对话（深度问答）
人机对话（聊天机器人）
信息抽取与知识图谱（大词林）
作文评阅与辅助写作
社会媒体观测（情感分析）
阅读理解（篇章级语义理解及推理）
社会媒体预测
LTP语言技术平台（句子级句法语义分析）

社会预测（SP）组研究方向



社会预测（SP）组人员

- 组长：刘挺 教授/博导；副组长：丁效 博士/助理研究员
- 三名博士生、六名硕士生、两名本科生



目录

AI+金融成为热点

- 金融科技（FINTECH）在2011年被正式提出之前，2016年开始成为整个金融行业的主旋律
- 2017年7月20日，国务院印发《新一代人工智能发展规划》，明确提出智能金融产业建设，“提升金融多媒体数据处理与理解能力，创新智能金融产品和服务，发展金融新业态”
- 2017年度各种AI会议上FINTECH也成为大家关注的热点

国内外金融科技发展现状

- **文艺复兴公司（大奖章基金）**
  - **通过模型对股票、期货、货币等主要投资标的的价格进行监控，作出买入或卖出指令**
  - **94-14年，平均年化收益率71.8%**
  - **02年底至05年底，规模50亿美元大奖章基金已为投资者支付了60多亿美元回报**
- **Kensho公司**
  - **利用自然处理技术从经济报告、货币政策、时政新闻等多方位研究市场动态**
  - **2013年成立，高盛投资1500万美元，2018年被S&P GLOBAL5.5亿美元收购**

量化交易方式已成为华尔街上的主流，也将成为国内股票市场的发展趋势。

- **国内IT及金融公司布局智能金融业务**

公司名称	金融业务
腾讯	微众银行；和泰人寿；众安保险；理财通；财富通（微信支付）；腾讯征信
阿里	蚂蚁金服（网商银行、国泰财险、众安保险、天弘基金、蚂蚁财富、支付宝、花呗、借呗、芝麻信用、网金社）
百度	百度金融（百信银行、百度钱包、百度理财、有钱花、白金交）
京东	京东金融（京东支付、京东白条、京东金条、京保贝、京东众筹、京东财富）
网易	网易理财、网易支付、立马理财、网易有钱、网易小贷
新浪	新浪支付、微财富、房金所、操盘宝
小米	小米支付、新网银行、小米理财、小米小贷
滴滴	融资租赁、汽车分期、滴滴保险
奇虎360	360金融（你财富、私银家、360淘金、360借条）
58同城	58钱柜、58贷款、长银五八消费金融公司
美团	美团小贷、美团支付、吉林亿银银行
搜狐	搜狐金服（搜易贷、狐狸慧赚、小狐分期）

金融领域的发展变革

- **人工处理->智能计算分析**
- **海量的金融信息，超过了人类处理的极限**

“20世纪90年代，一个基金经理把市场当天产生的研报、舆情、新闻、交易数据看完，大概需要10个小时；2010年，大概需要10个月的时间；2016年，把当天市场信息看完，需要20年的时间，相当于整个职业生涯”

——《智能革命》

研究框架


股票融合预测模型
金融事件分析
金融事理图谱
股票数据与形态技术分析
金融文本处理与分析
数据实时获取（文本数据，股票数据）
择时预测
选股预测
知识图谱
深度学习
自然语言处理
数据挖掘
数据采集

多源金融文本数据实时采集

- 金融数据分析需求：
  - 信息充分，多种来源的金融数据
  - 分秒必争，实时采集与分析数据

- 实时采集新闻、论坛、博客、公告、研报等多种金融相关信息

金融数据采集与抽取结果

事理图谱的研究动机

- 现有的知识库普遍是以“概念及概念间的关系”为核心的，缺乏对“事理逻辑”知识的挖掘

- 事理逻辑（事件之间的演化规律与模式）是一种非常有价值的常识知识，挖掘这种知识对我们认识人类行为和社会发展变化规律非常有意义

什么是事理图谱?

- 事理图谱：Event Evolutionary Graph

- 结构上：事理图谱是一个有向有环图，节点代表事件，有向边代表事件之间的顺承、因果关系
- 本质上：事理图谱是一个事理逻辑知识库，描述了事件之间的演化规律和模式

事理图谱的应用

- 事理图谱可应用于事件预测、常识推理、消费意图挖掘、对话生成、问答系统、辅助决策等任务中
- 大规模事理图谱将和传统知识图谱一样，具有非常巨大的应用价值

事理图谱与知识图谱的区别与联系

	事理图谱	知识图谱
研究对象	谓词性事件及其关系	名词性实体及其关系
组织形式	有向图	有向图
主要知识形式	事理逻辑关系，以及概率转移信息	实体属性和关系
知识的属性	事件间的演化关系多数是不确定的	多数实体关系是确定性的

事理图谱中的事件定义

- 前人工作: 事件是特定时间、地点下的一个状态变化
  - 经典的事件抽取和分类任务, ACE 2005
  - 话题检测与跟踪

- 事理图谱中的事件:
  - 不是具体事件，而是抽象事件
  - 表示为泛化、语义完备的谓词短语或片段
  - “吃火锅”，“看电影”，“去机场”是合理的事件表达
  - “去地方”，“做事情”，“吃”是不合理或不完整的事件表达

事件间顺承关系

- 顺承关系(suquential)是指两个事件在时间上相继发生的偏序关系

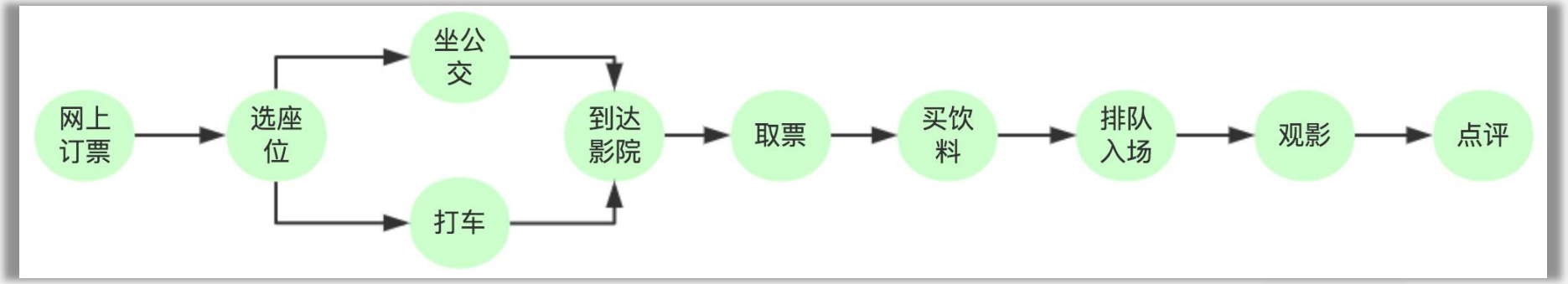
吃过午饭后，小明到前台买单，然后离开了餐馆。

事件间因果关系

- 因果关系(causal)是指两个事件之间，前一事件（原因）导致后一事件（结果）的发生
- 因果关系是顺承关系的子集
  - 满足发生时间上的偏序约束

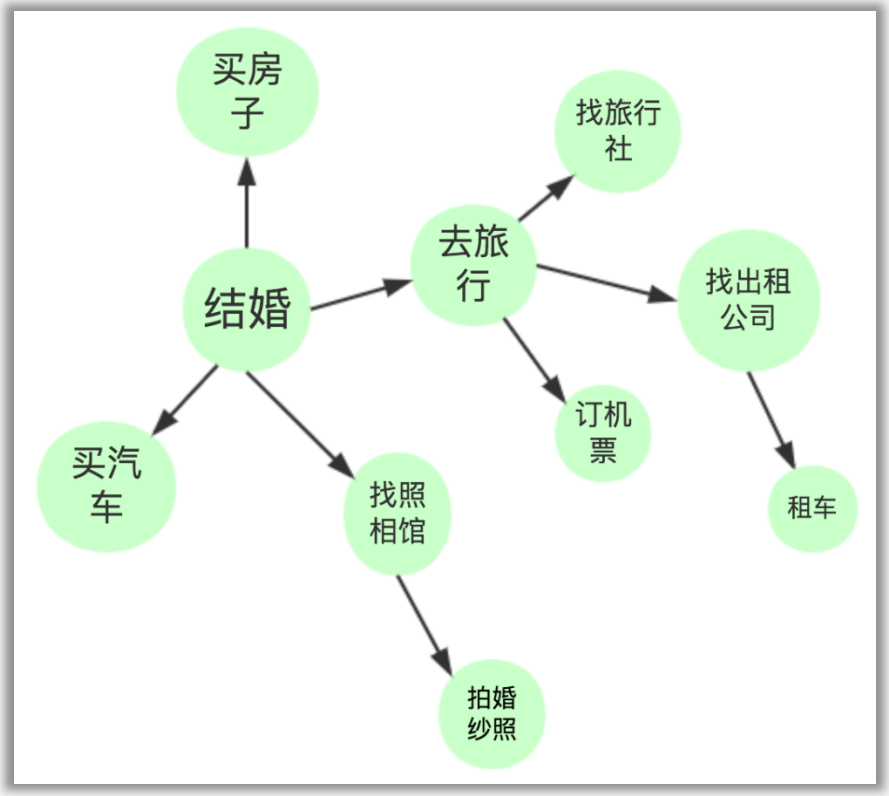
核泄漏引起了严重的海洋污染。

事理图谱的三种拓扑结构



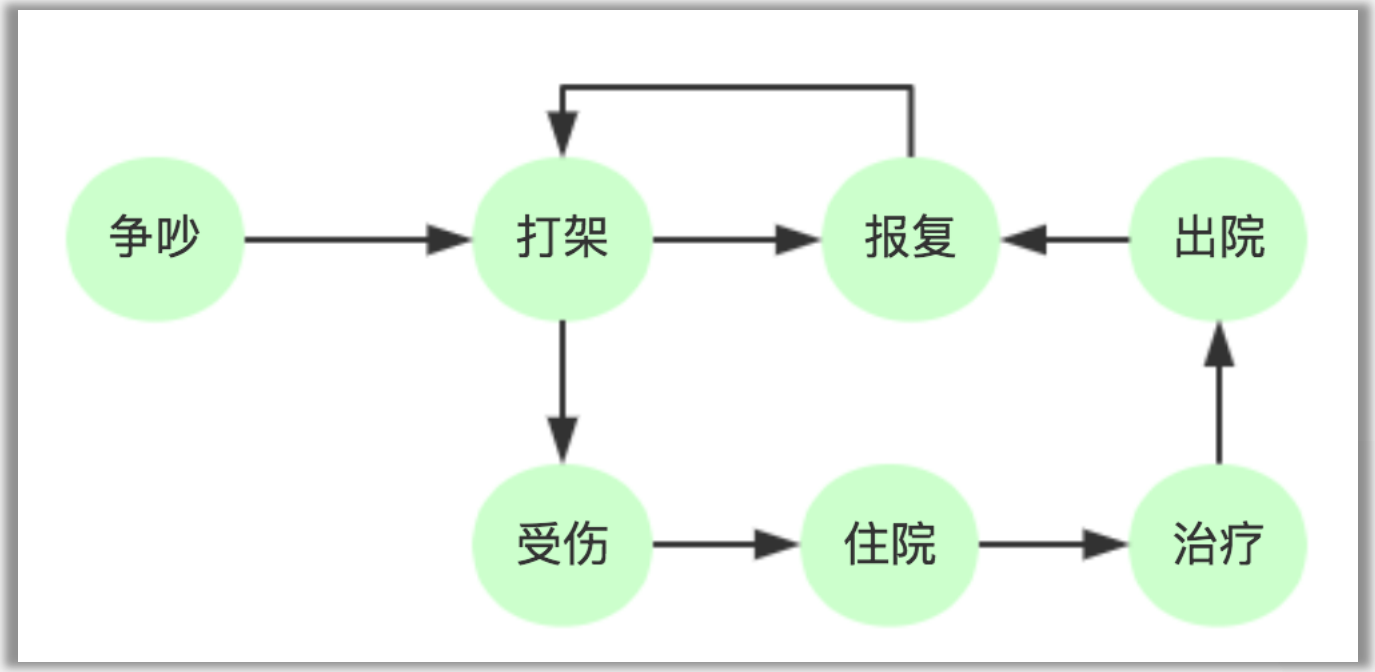
‘看电影’场景下的**链状**事理图谱

事理图谱的三种拓扑结构



‘结婚’场景下的**树状**事理图谱

事理图谱的三种拓扑结构



‘打架’场景下的**环状**事理图谱

哈工大在事理图谱方面的探索

金融领域事理图谱的构建与应用

哈爾濱工業大學
社会计算与信息检索研究中心

样本分析

- 财经新闻中包含大量事件之间的**因果关系**
  - 显式因果（带关联词）：“塑化剂事件导致白酒股大跌。”
  - 隐式因果（不带关联词）：“百度Q2财报：净利同比增82.9%，股价盘后上涨7%”

- 除因果关系外，也存在大量事件**顺承关系**
  - “停牌了近半个月的科大讯飞（002230.SZ）复牌，股价开盘即涨停。”

金融事理图谱构建目标

- 目标
  - 针对金融领域，从财经新闻中挖掘、构建与经济变化情况（尤其是导致股票涨跌变化）相关的顺承、因果事理关系，形成金融领域的事理图谱
- 方案
  - 因果、顺承关系抽取
  - 将“因果事件对”形成图谱形态

“因果对”抽取方法

- 利用因果触发词构造填充模板，进行正则匹配，得到cause和effect子句
  - 例如：(.+)(导致|引起|造成)(.+)(下跌|上涨)
- 对cause和effect子句进行分词、词性标注，进行词性过滤，得到动词、名词和形容词构成的句子主干作为最终的cause和effect

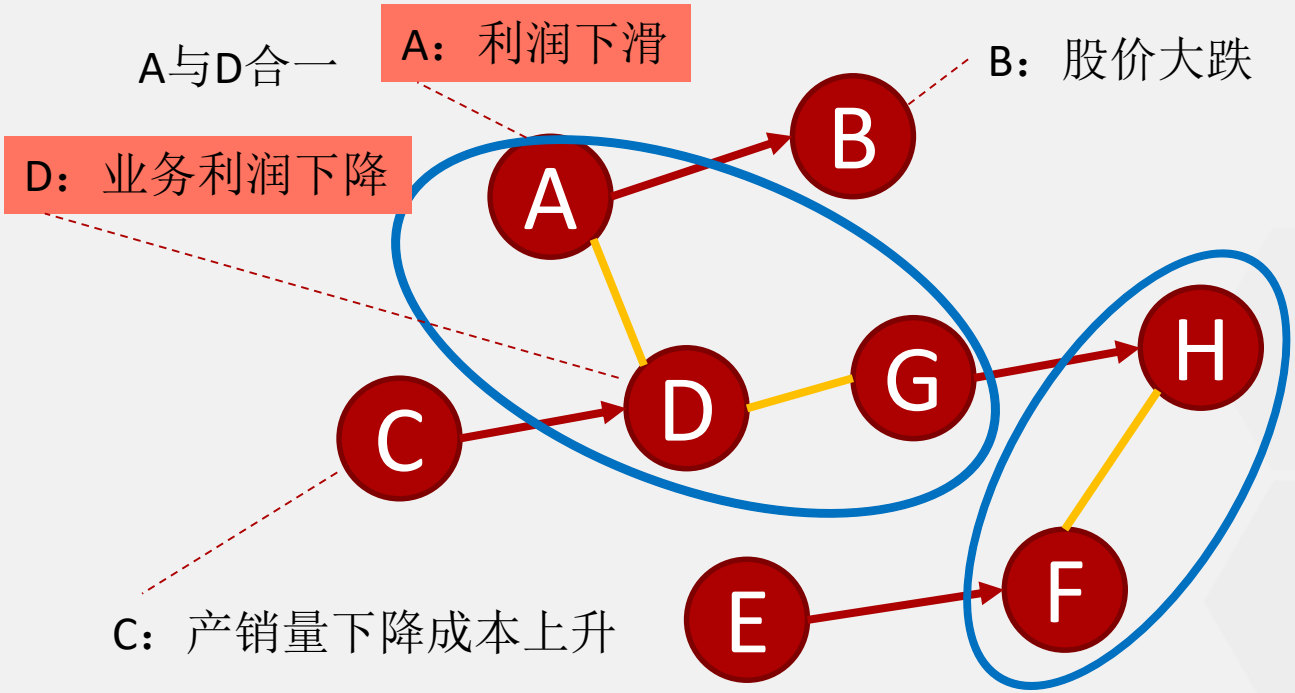
Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Chengxiang Zhai. Constructing and embedding abstract event causality networks from text snippets. WSDM 2017.

典型抽取结果实例

- “中兴通讯利润下滑**引发**股价大跌”
- “中兴通讯天价赔偿传闻**导致**股票大跌”
- “双汇并购史密斯菲尔德消息**使得**双汇发展股价大涨”
- “肉制品产销量下降成本上升**造成**肉制品业务利润下降”
- “相关竞争对手进入**导致**产品毛利率市场份额下降”

因果事件对到图谱

- 通过对某些事件的合一，将因果对形成图谱形态



Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan. Deep Learning for Event-Driven Stock Prediction. IJCAI, 2015.

通过相似度计算，实现事件的合一

- 尝试多种事件相似度度量方法
- 事件表示
  - 动词、名词词袋
  - 动词、名词、形容词词袋
  - 所有词的average embedding
  - 动词、名词、形容词的average embedding
  - Event embedding
  - Phrase embedding
- 相似度计算
  - 杰卡德相似度
  - cosine相似度

金融事理图谱构建

- 数据准备：

	文件名	文件体积
财经新闻 (6.7G)	163_news.txt	1.1G
	by_news.txt	1.4G
	guba_extract_addtime.txt	419M
	hexun47w_time.txt	1.8G
	Resset_News_addtime.txt	16M
	tencentNetease_addtime.txt	2.1G
开放域新闻 (19G)	news_paper.txt(2005~2014)	19G

金融事理图谱构建

- 因果对抽取:
- 给每一条因果边都加上所在的**上下文**;

  给每一条因果边都加上该篇新闻发布的**时间戳**;

  因果事件对抽取的精细化，提升抽取效果;

  1000条人工标注句子上的抽取结果评估如下:

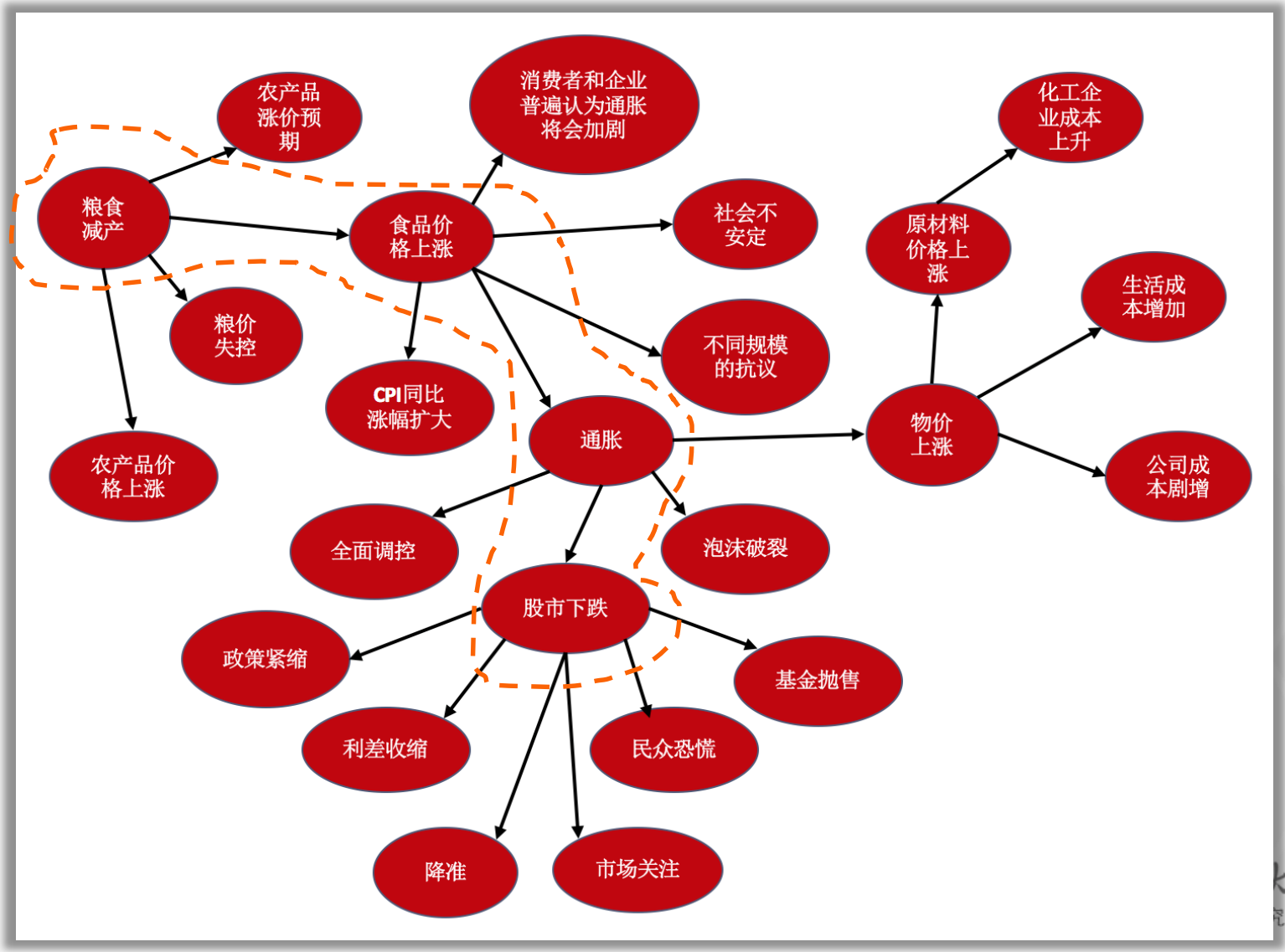
	Accuracy	Precision	Recall	F1
因果对事件抽取	62.1%	93.7%	59.3%	72.6%

金融事理图谱构建

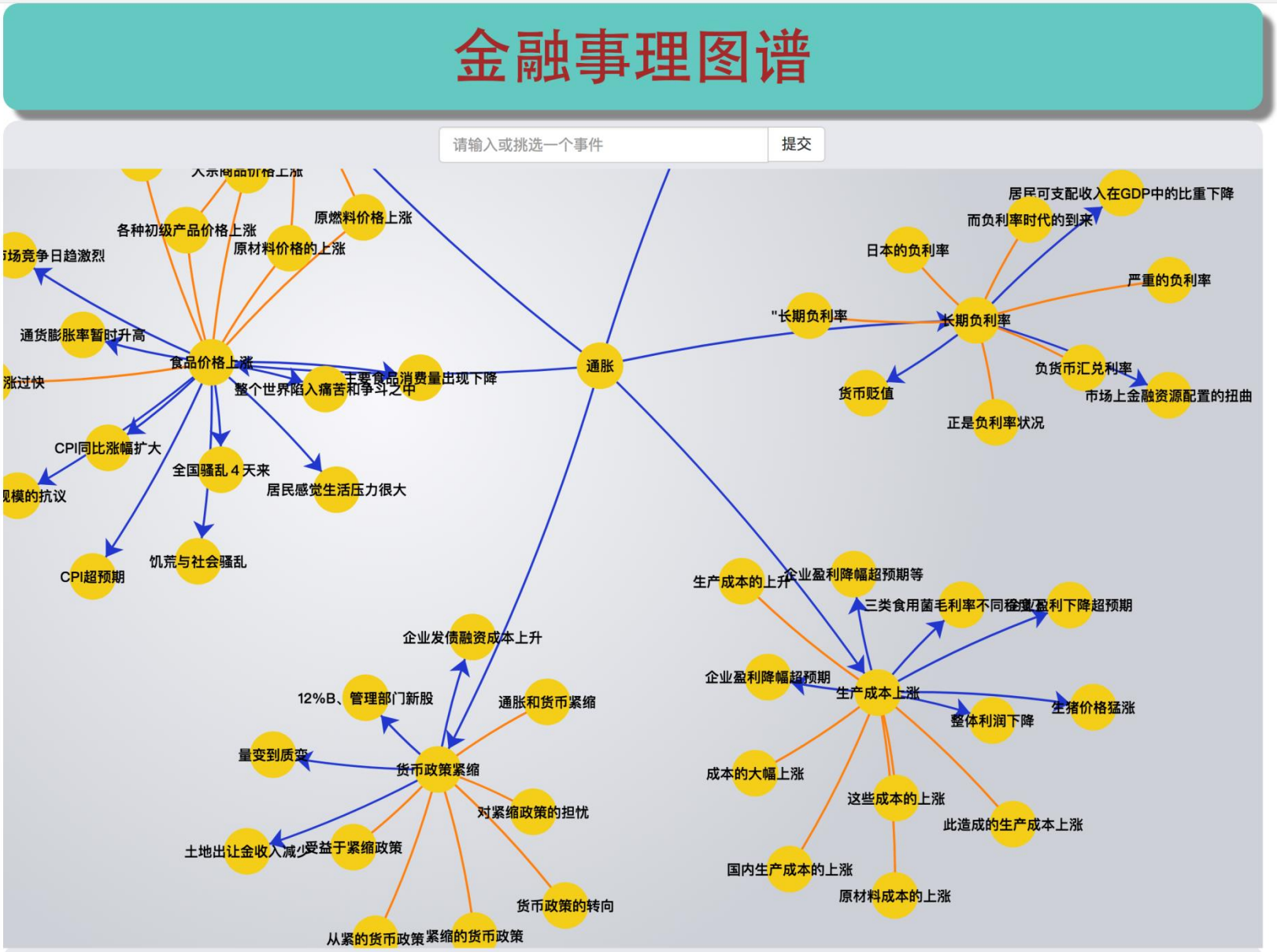
- 基于上述方法，构建了第二版本的金融事理图谱
- 主要改进：扩大数据源；提升因果事件对抽取效果

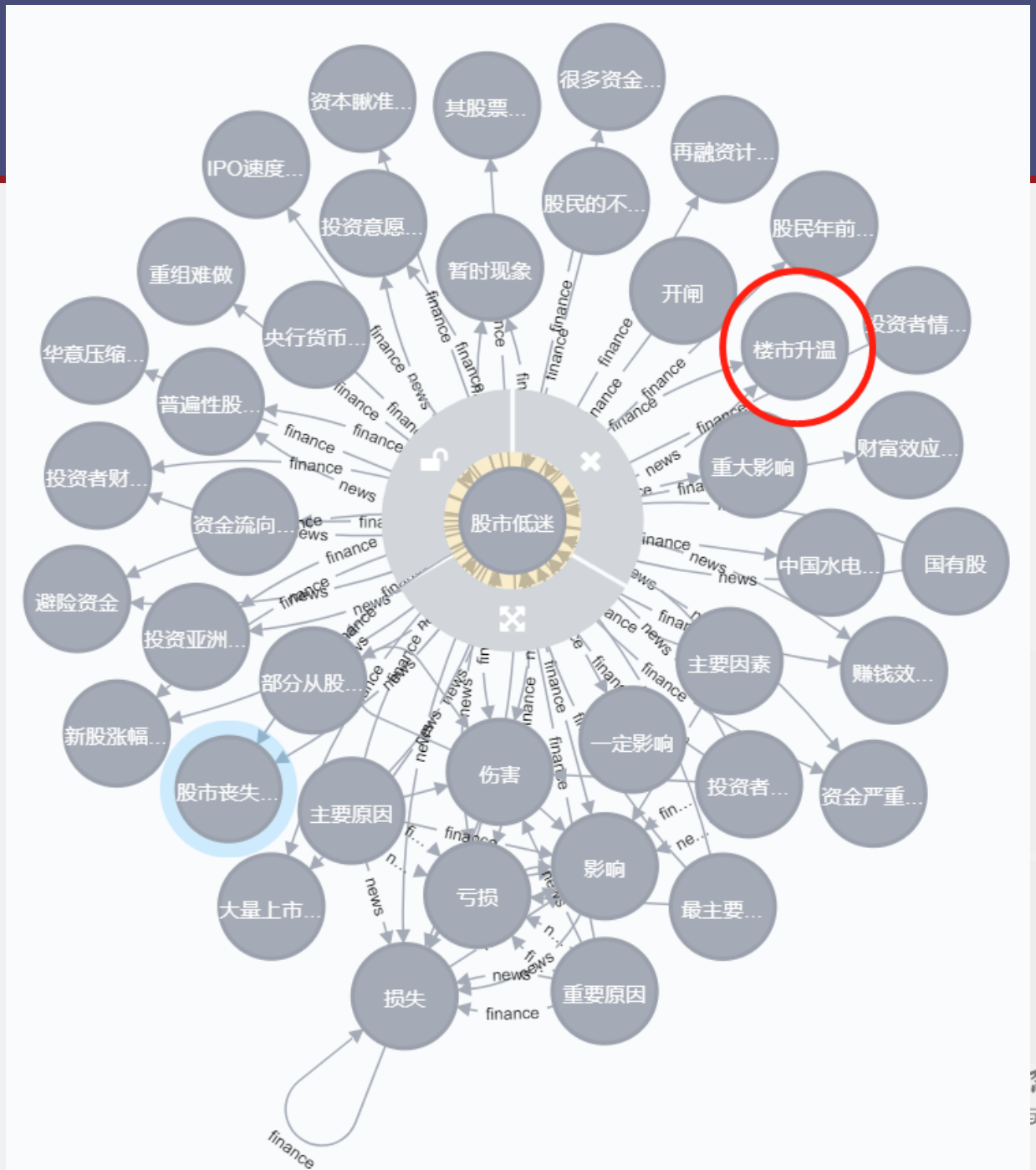
	因果对数目	事件数目
第一版本金融图谱规模	154,233	247,926
第二版本金融图谱规模	1,873,140	1,542,516

金融事理图谱的样例


农产品涨价预期
消费者和企业普遍认为通胀将会加剧
化工企业成本上升
粮食减产
食品价格上涨
社会不安定
原材料价格上涨
生活成本增加
粮价失控
不同规模的抗议
CPI同比涨幅扩大
通胀
物价上涨
农产品价格上涨
公司成本剧增
全面调控
泡沫破裂
股市下跌
政策紧缩
基金抛售
利差收缩
民众恐慌
降准
市场关注

金融事理图谱Demo



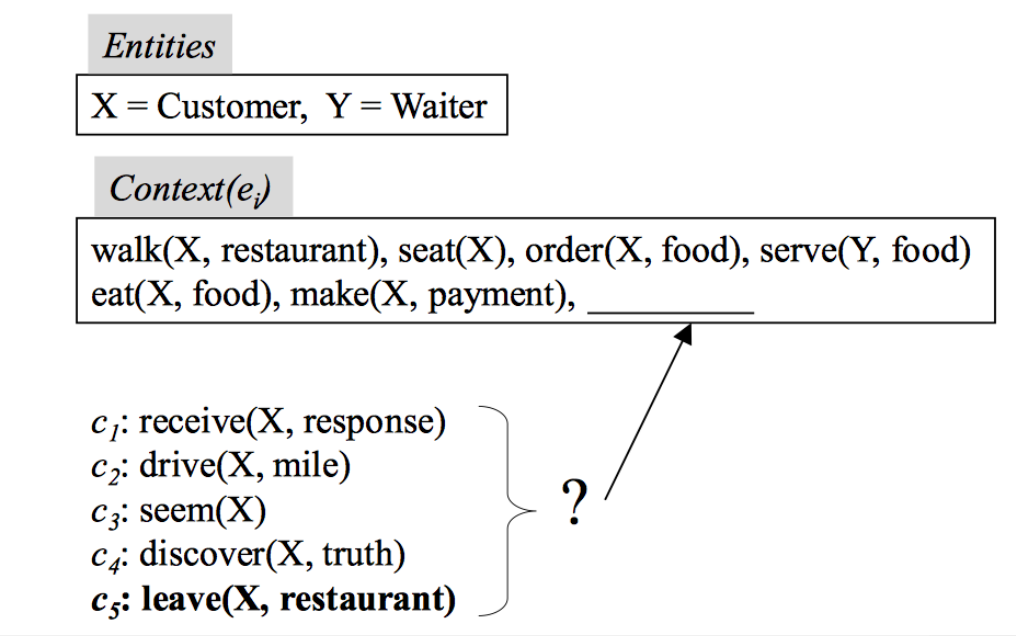


Constructing Narrative Event Evolutionary Graph for Script Event Prediction

Zhongyang Li, Xiao Ding, Ting Liu. IJCAI 2018

任务

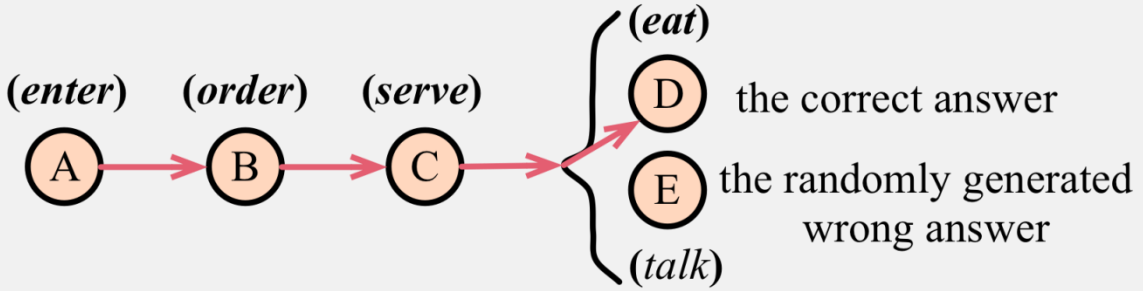
- 给定事件上文，从候选事件列表中选出接下来最有可能发生的事件



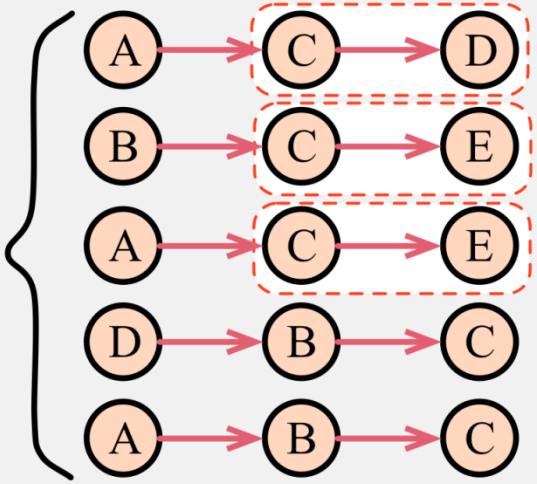
动机

1. 前人工作探索了如何利用event pair和event chain进行预测，我们的工作探索如何利用event graph的稠密连接信息来帮助事件预测
2. 图结构学习到的事件表示更加利于预测
3. 图结构能克服事件之间的不连通性，学习更好的事件关联

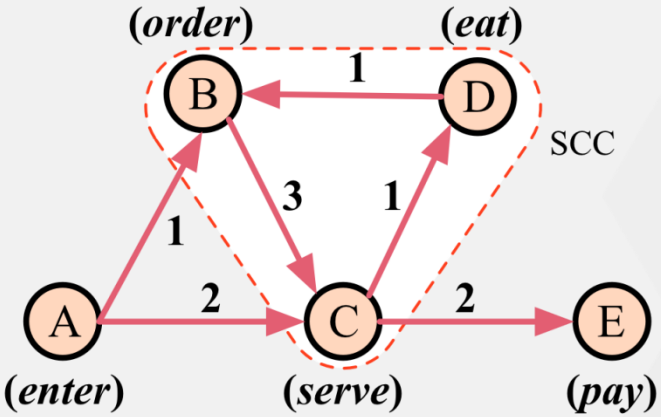
动机



(a) Given an event context (A, B, C), choose the subsequent event from D and E.



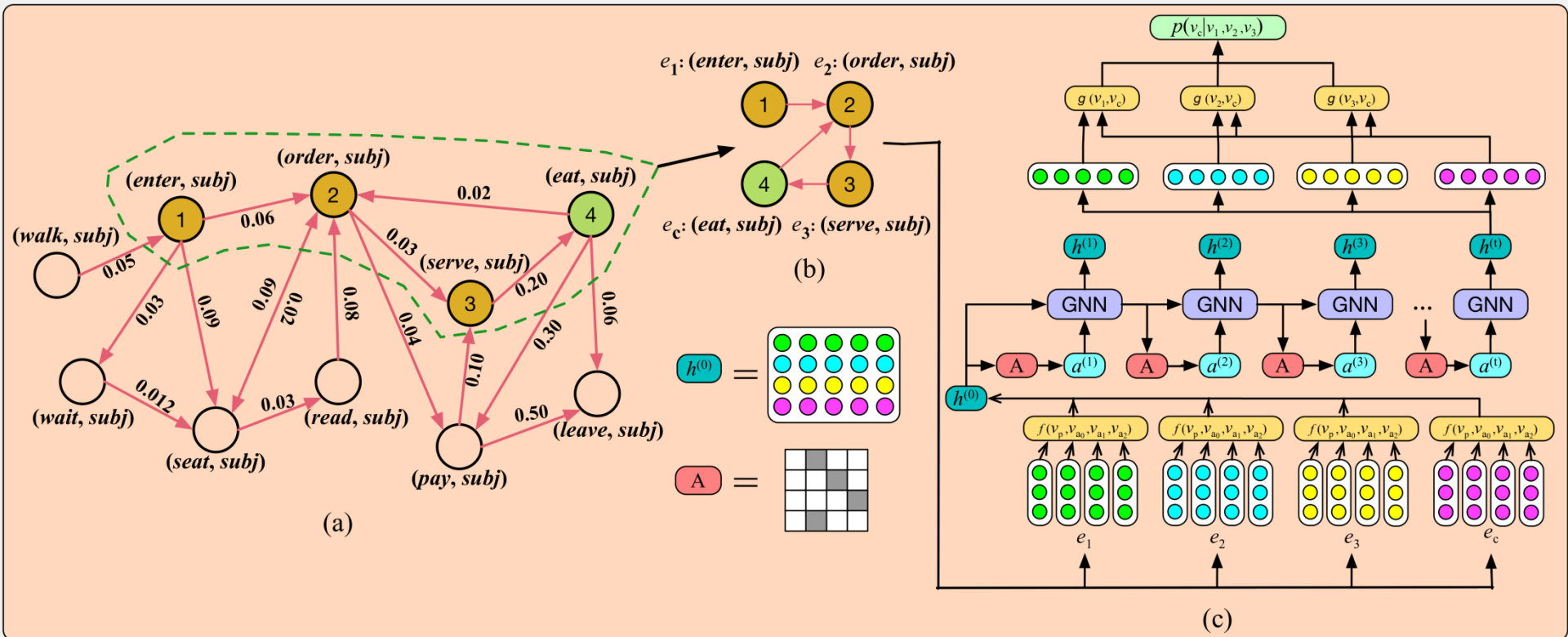
(b) Training event chains.



(c) Narrative event graph based on event chains in (b).

方法

- 我们提出了可扩展的图神经网络(Scaled Graph Neural Network-SGNN)来解决图上的预测、推理问题
  - 重点解决了GNN在大规模图结构上无法进行推理的问题



实验

Methods	Accuracy
Random	20.00
PMI [Chambers and Jurafsky, 2008]	30.52
Bigram [Jans *et al.*, 2012]	29.67
Word2vec [Mikolov *et al.*, 2013]	42.23
DeepWalk [Perozzi *et al.*, 2014]	43.01
EventComp [Granroth-Wilding and Clark, 2016]	49.57
PairLSTM [Wang *et al.*, 2017]	50.83
SGNN-attention (without attention)	51.56
SGNN (ours)	**52.45**
SGNN+PairLSTM	52.71
SGNN+EventComp	54.15
SGNN+EventComp+PairLSTM	**54.93**

Learning Sentence Representations over Tree Structures for Target-dependent Classification

Junwen Duan, Xiao Ding, Ting Liu. NAACL 2018

Motivation

- Tree structures are promising for target-dependent classification
  - Capture long-distance interactions between the target and its contexts
  - Avoid possible information loss over long sequences

- Limitation of previous work on tree structures
  - Rely on external treebank annotations or syntactic parsers
  - The tree structures are fixed

Our assumptions

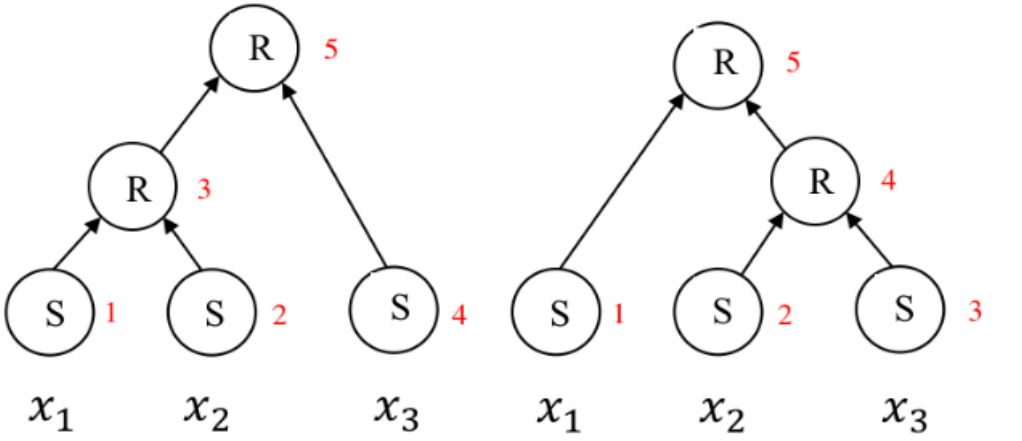


Figure 1: For input sequence $\{x_1, x_2, x_3\}$, the shift-reduce orders can be {S,S,R,S,R} and {S,S,S,R,R}, where **S** stands for SHIFT and **R** stands for REDUCE.

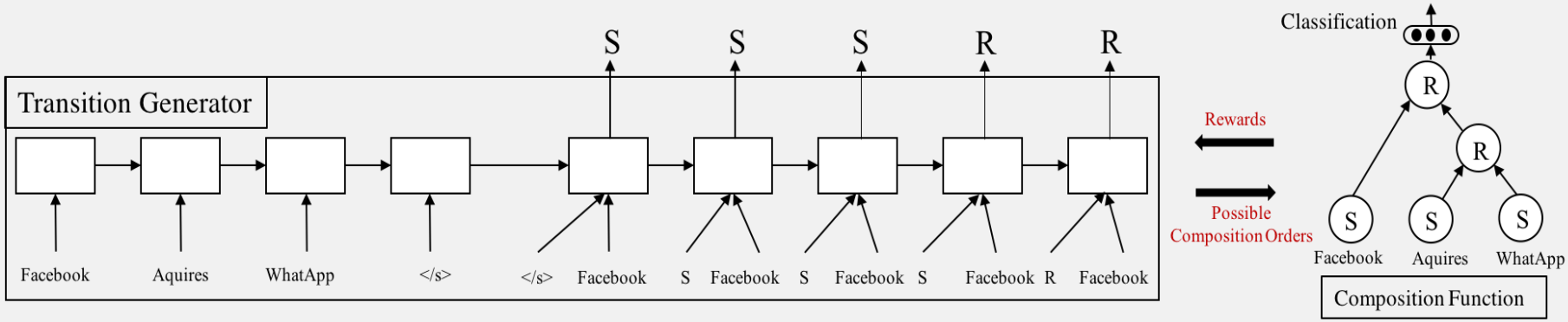
Shift-reduce parsing

We can obtain a binarized tree by using only shift and reduce transitions

Target-dependent Representation

We can obtain different vector representations for the sentence given different composition orders

Our Approach

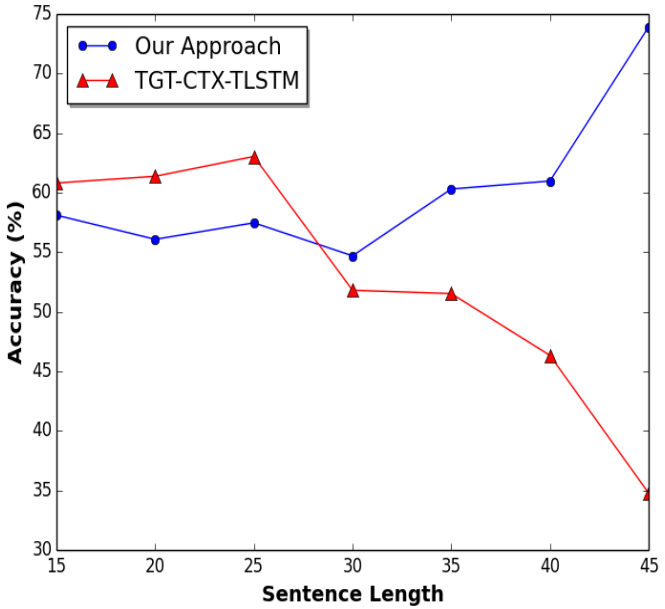


1. **Transition Generator**
   - Encoder & Attention-Decoder
   - Take the target into account at decoding
2. **Composition Function**
   - Tree-structured long short-term memory network
   - Compose words into sentence following the transition orders
3. ***Reinforce* with self-critic baseline**
   - Reward the generated structures that benefits end tasks

Experiments

Firm-specific Cumulative Abnormal Return Prediction

Method	Class	F1-score
Sentiment-based	$+CAR_3$	**0.597**
	$-CAR_3$	0.476
	Macro	0.536
Bi-LSTM	$+CAR_3$	0.557
	$-CAR_3$	0.490
	Macro	0.523
Bi-LSTM + Attention	$+CAR_3$	0.575
	$-CAR_3$	0.523
	Macro	0.549
TD-CTX-TLSTM	$+CAR_3$	0.552
	$-CAR_3$	0.570
	Macro	0.561
Our Approach	$+CAR_3$	0.572
	$-CAR_3$	**0.592**
	Macro	**0.582**



Dataset: http://ir.hit.edu.cn/~xding/

Extracting Key Evidences from Financial News Documents for Stock Market Prediction

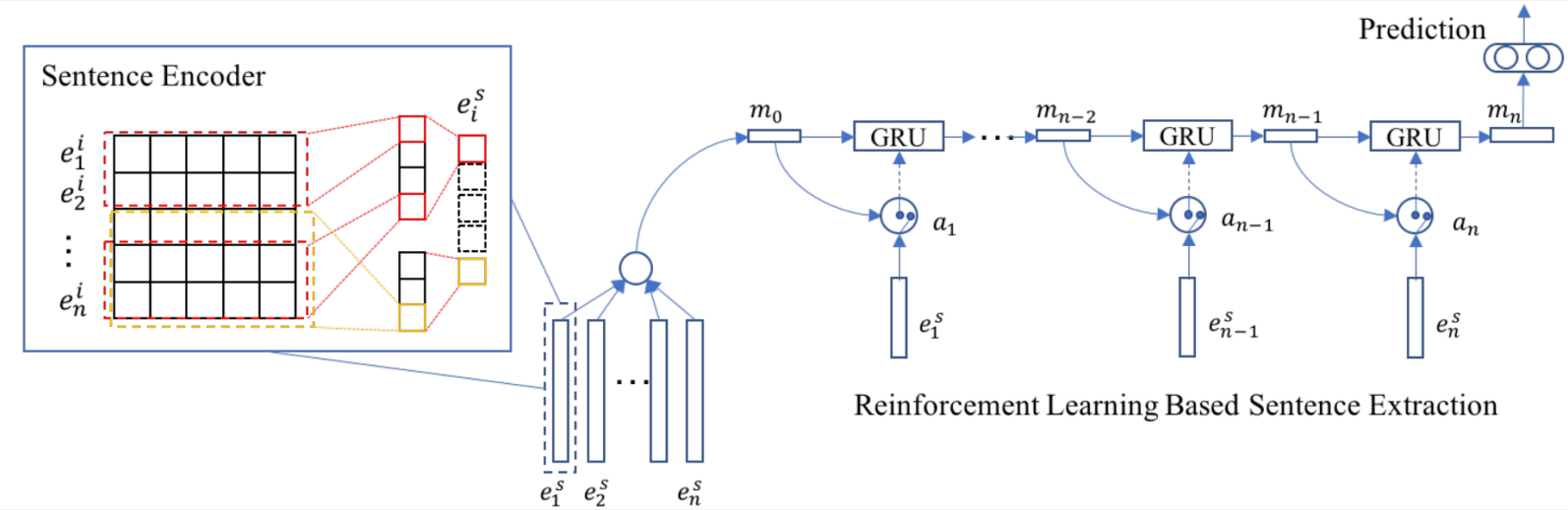
Junwen Duan, Xiao Ding, Ting Liu

Motivation

- Financial news documents are promising for stock market prediction
  - Title (Ding et al. 2014, 2015, 2016; Xie et al. 2013)
  - Leading paragraph (Chang et al. 2016; Duan et al. 2018)

- Limitation of previous work on news documents based stock prediction
  - The value of documents cannot be well explored
  - Lack of interpretability

Method



1. **Sentence Encoder**
   - Convolutional neural network based encoder
2. **Reinforcement Learning Based Sentence Extraction**
   - The extracted sentences should be beneficial for predicting the stock market movement, so that we can take advantage of the stock returns as indirect supervisions to guide the extractions
3. **Training**
   - Maximize the expected reward of the action sequence

Experiments

Firm-specific Cumulative Abnormal Return Prediction

	Training	Development	Test
$+CAR_3$	5052	253	503
$-CAR_3$	4981	259	513
#doc	10033	512	1016
#s/d	21.6	20.8	21.3

Table 2: Number of CAR_3 in the datasets

Method	Macro-F1	Accuracy
TITLE	55.64	56.88
LEAD1	57.08	57.57
LEAD3	57.56	57.56
LexRank	56.29	56.29
TextRank	57.34	57.38
CNN + AVG	57.71	57.87
HN	57.79	58.07
HAN	58.04	58.26
Our Method	**59.75**	**60.62**

Table 3: Experimental results on the test dataset, the best results are in bold.

Dataset: http://ir.hit.edu.cn/~xding/

Case Study

LexRank	the company 's shares rose 3 percent by midday on tuesday after mastercard reported $ 1.43 billion in quarterly revenue . (reporting by maria aspan . editing by derek caney , john wallace and robert macmillan)
TextRank	the company 's shares rose 3 percent by midday on tuesday after mastercard reported $ 1.43 billion in quarterly revenue . the company , which processes credit card transactions but does not lend directly to people , makes money every time someone buys something with a mastercard credit or debit card . people in asia , europe and especially latin america spent more money with maste lit and debit cards during the third quarter . the u.s. financial reform law will restrict the processing fees that mastercard and visa earn from debit card transactions . but the law is expected to have a light impact on mastercard , because of its small share of the u.s. debit market . mastercard has even said the law could help it take some market share from visa , because a provision in the law would end exclusive debit processing contracts .
Our Method	new york mastercard inc 's (ma.n) third-quarter profit rose 15 percent and beat expectations as people outside the united states bought more things and switched more of their payments from cash to plastic . the company 's shares rose 3 percent by midday on tuesday after mastercard reported $ 1.43 billion in quarterly revenue .

Noise Information

Redundant Information

总结

- 自然语言处理+金融优势互补
- 知识图谱在各个领域精耕细作，逐渐显露价值，事理图谱相关研究越来越吸引研究者关注
- 未来工作：知识图谱与事理图谱的融合，基于知识/事理图谱的金融市场分析，让事理图谱在金融领域发挥更大作用

参考文章

- Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan. Deep Learning for Event-Driven Stock Prediction. IJCAI, 2015.
- Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. EMNLP, 2014.
- Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan. Knowledge-Driven Event Embedding for Stock Prediction. COLING 2016.
- Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Chengxiang Zhai. Constructing and embedding abstract event causality networks from text snippets. WSDM 2017.
- Zhongyang Li, Xiao Ding, Ting Liu. Constructing Narrative Event Evolutionary Graph for Script Event Prediction. IJCAI 2018.
- Junwen Duan, Xiao Ding, Ting Liu. Learning Sentence Representations over Tree Structures for Target-dependent Classification. NAACL 2018.

HIT-SCIR

哈爾濱工業大學
社会计算与信息检索研究中心

请批评指正！

xding@ir.hit.eud.cn

主要合作者:

段俊文

李忠阳

赵森栋