北京大学

基于知识的文本语义理解与文本生成

赵东岩
北京大学王选计算机研究所
2020年9月6日

内容提纲

- 背景介绍
- 文本语义理解与生成
- 研究进展

内容提纲

- 背景介绍
- 文本语义理解与生成
- 研究进展

类人智能研究动态

- IBM
  - Jeopardy!：知识问答机器人 — 子问题理解、多线索汇总、置信度评估
  - Debater：辩论机器人 — 首次呈现的人机辩论
- 考试机器人
  - Aristo、Plato计划：AI2面向基础教育问题理解的解题系统
  - Todai Robot：2021年考上东京大学 — 深层语义分析 + 文本推理
- Google Answer Engine
- Bert，GPT-3

机器如何做到语义理解——可理解、可解释

◆人是如何理解语义和知识的?

已知:四川的国民生产总值远超渝、滇、贵、藏。
试问: 四川是西南地区最富庶的么?

◆高考Robot的基础技术研究

- ➢ 构建基础语义资源库及深度语义分析技术平台
- ➢ 研制大规模知识库构建技术，构建学科知识库
- ➢ 提出语义与知识表示方法、研制深度语义理解技术
- ➢ 实现面向初等教育问题求解的知识推理

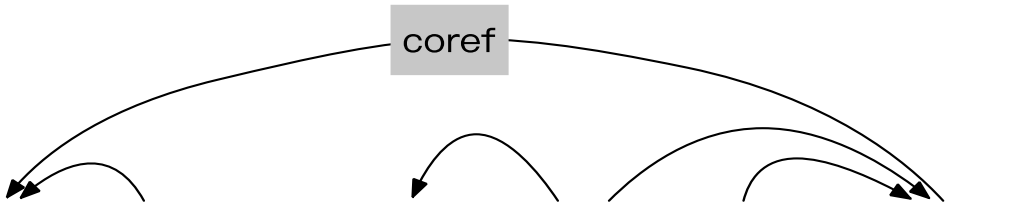
内容提纲

- 背景介绍
- 基于知识的文本语义理解与生成
- 研究进展

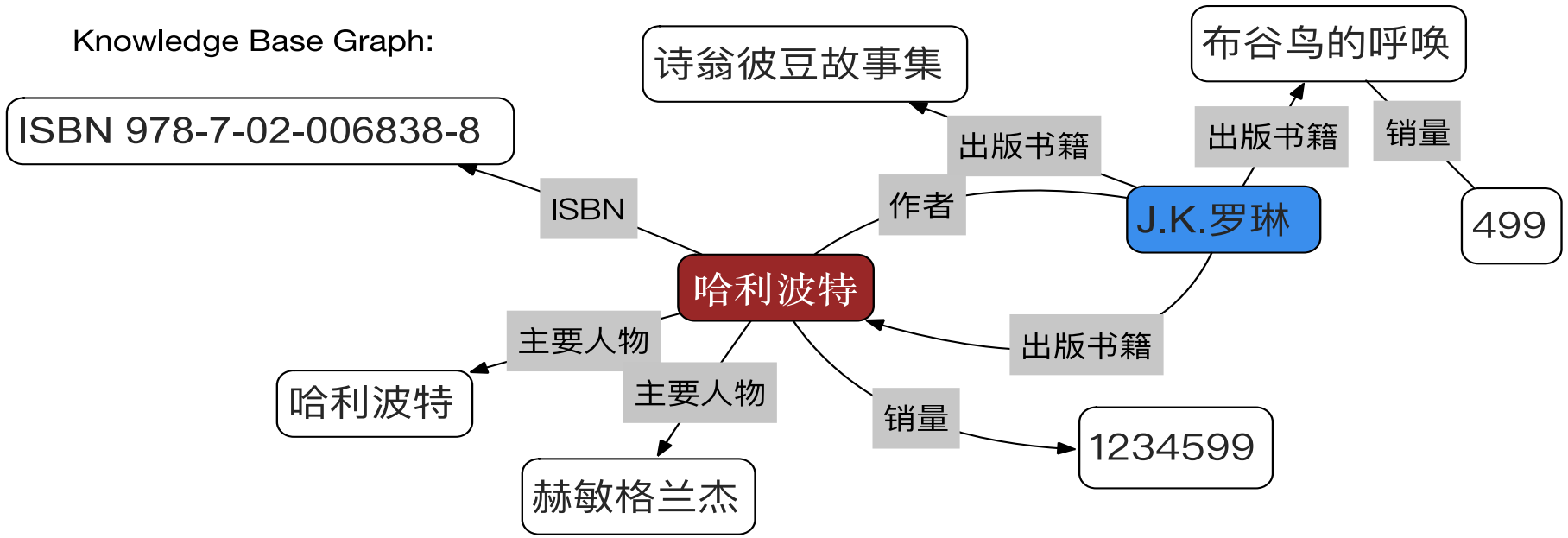
基础：基于知识图谱的语义理解

◆ 将自然语言问题解析为与知识库关联的结构化查询语句

C: 类别型短语
R: 关系型短语
E: 实体型短语
V: 变量型短语
Agg: 函数式短语



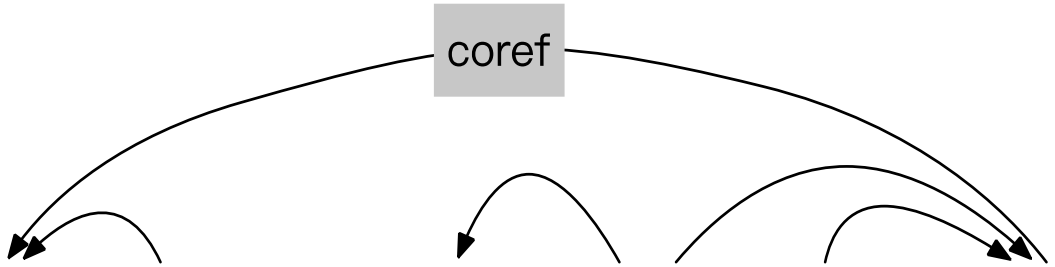
Knowledge Base Graph:



基于知识图谱的语义理解

结构化查询语句SPARQL

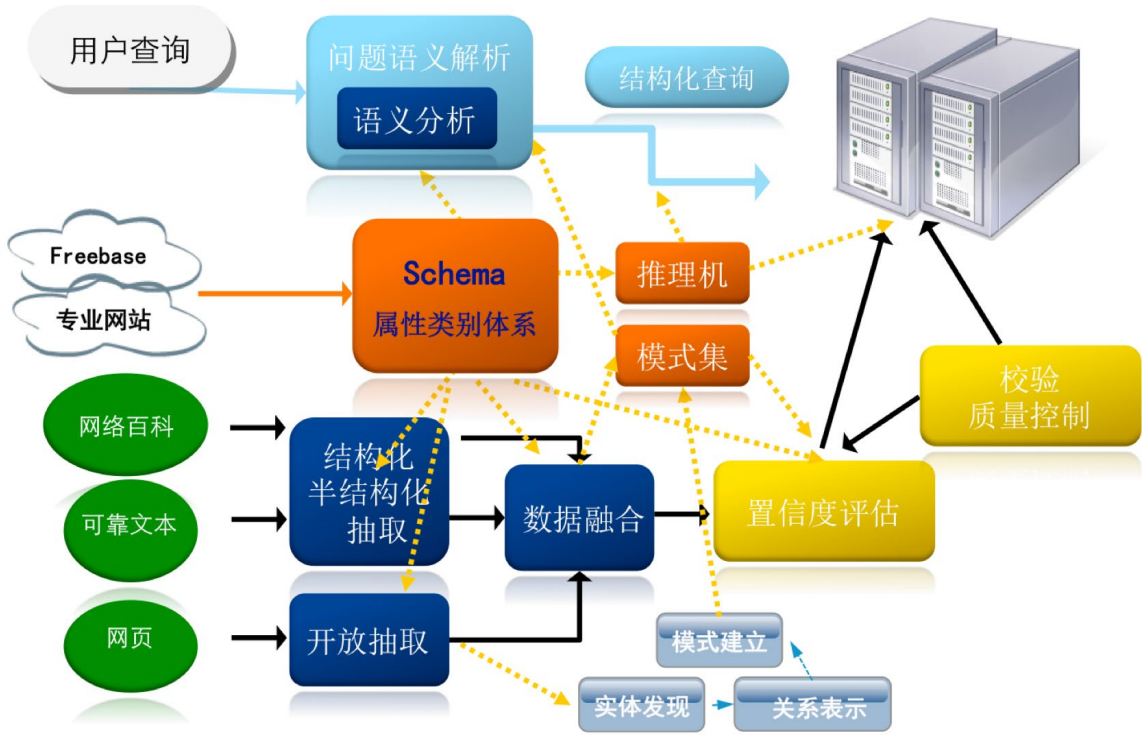
C: 类别型短语
R: 关系型短语
E: 实体型短语
V: 变量型短语
Agg: 函数式短语




```
select ?y where { 哈利波特 作者 ?x
                  ?x 出版书籍 ?y
                  ?y 销量 ?z } orderBy ?z
```

研究工作之一： 知识图谱构建

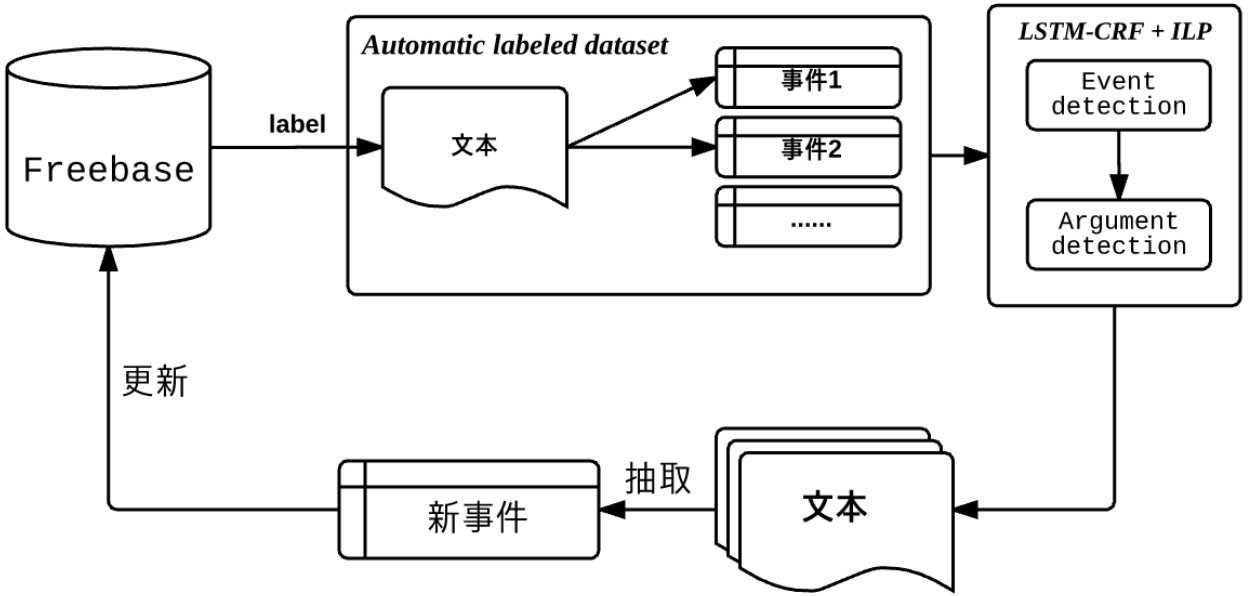
- **研究内容和研究框架**
  - **研究内容：**如何从开放域网络信息资源中萃取以实体及实体间关系形式存在的知识条目；并据此构建以图模式存储的结构化语义知识库
  - **研究方法：**
    - **开放域半结构化知识抽取**
    - **非结构化文本的实体关系抽取**
    - **知识库补全**



[Chen et al, AI Journal 2018; Ze

如何扩大优质样本?

- **基于新闻事件触发的事件抽取**
  - 可靠表格信息作为远程监督



property / id	company_acquired	acquiring_company	date	divisions_formed
m.07bh4j7	Remedy Corp	BMC Software	2004	Service Management Business Unit
m.05nb3y7	aQuantive	Microsoft	2007	NONE

[Zeng et al., AAAI 2018]

研究工作— 事件抽取

- **可与传统手工标注相兼容**
  - **自动发现相关事件的触发词**
  - **可覆盖64.7%的ACE 2005 数据集**

Event types	Trigger candidates	Percentage
film_performance	play, appear, star, cast, portray	0.72
award_honor	win, receive, award, share, earn	0.91
education	graduate, receive, attend, obtain, study	0.83
acquisition	acquire, purchase, buy, merge, sell	0.81
employ._tenure	be, join, become, serve, appoint	0.79

研究工作— 事件抽取

- 利用wiki的 business acquisition, winning of the Olympics games, 和 awards winning in entertainment 三个大型表格
- 没有任何人工标注

Event type	Entries	Positive	EC	KAD	AAD
Acquisition	690	414	87.0%	72.0%	69.6%
Olympics	2503	1460	77.2%	64.5%	38.6%
Awards	3039	2217	95.0%	82.8%	58.6%

知识图谱的动态构建

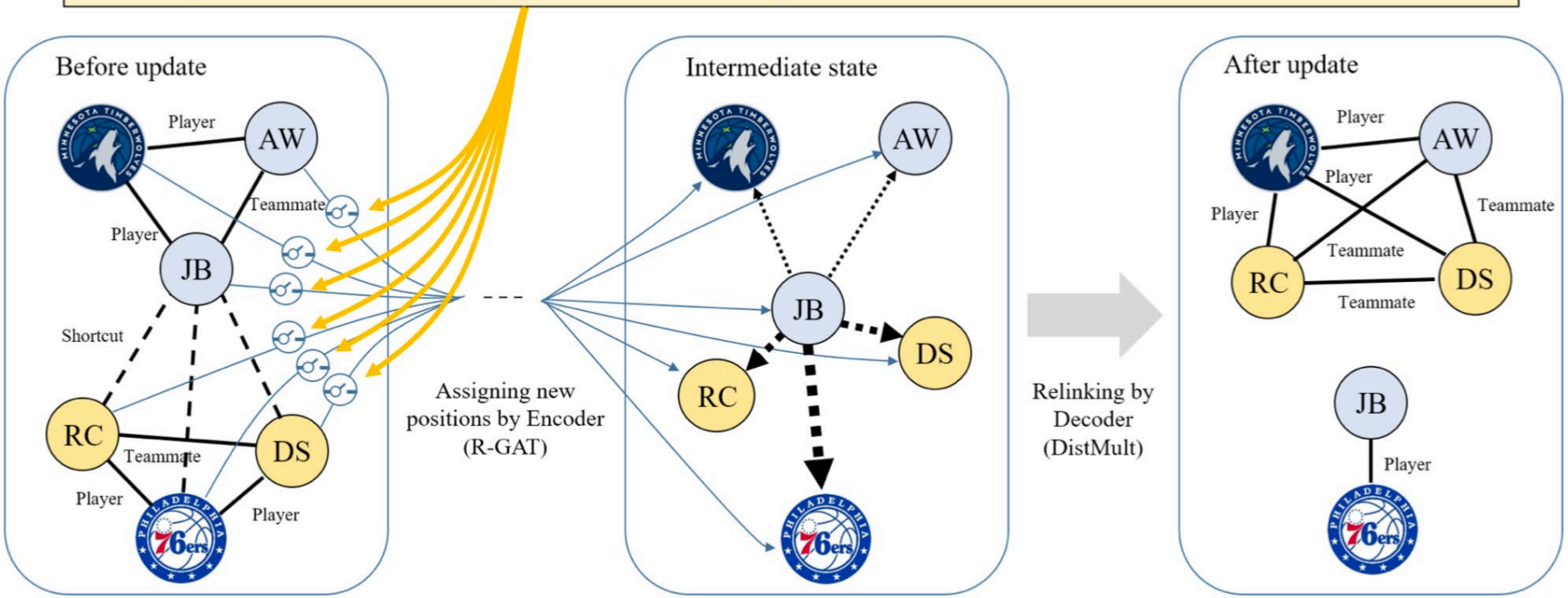
• 新闻事件触发的知识更新

- **问题：新闻事件与知识图谱更新机制间的gap**
  - 新闻文档中同时包括显式知识与隐式知识的更新
  - 知识库的更新机制依赖于具体新闻事件类型
  - 在NLP及KG研究领域中上属于空白

- **方法：利用图神经网络刻画新闻事件与知识更新机制间的关系**
  - 新闻事件解析与知识库更新间的协同学习
  - 图神经网络模型学习控制知识图谱上的消息传递机制
    - 学习特定事件语义
    - 控制消息传递路径
  - 可以从有限的标注样本当中自发的学习到特定事件类型的更新规则
  - 可扩展应用至许多人机结合的场景 （EMNLP 2019）

研究工作—新闻事件触发的知识更新

The *Minnesota Timberwolves* announced the team has acquired forward **Robert Covington**, forward **Dario Šarić**, guard **Jerryd Bayless** and a 2022 second-round draft pick from the *Philadelphia 76ers* in exchange for forward **Jimmy Butler** and center **Justin Patton**.



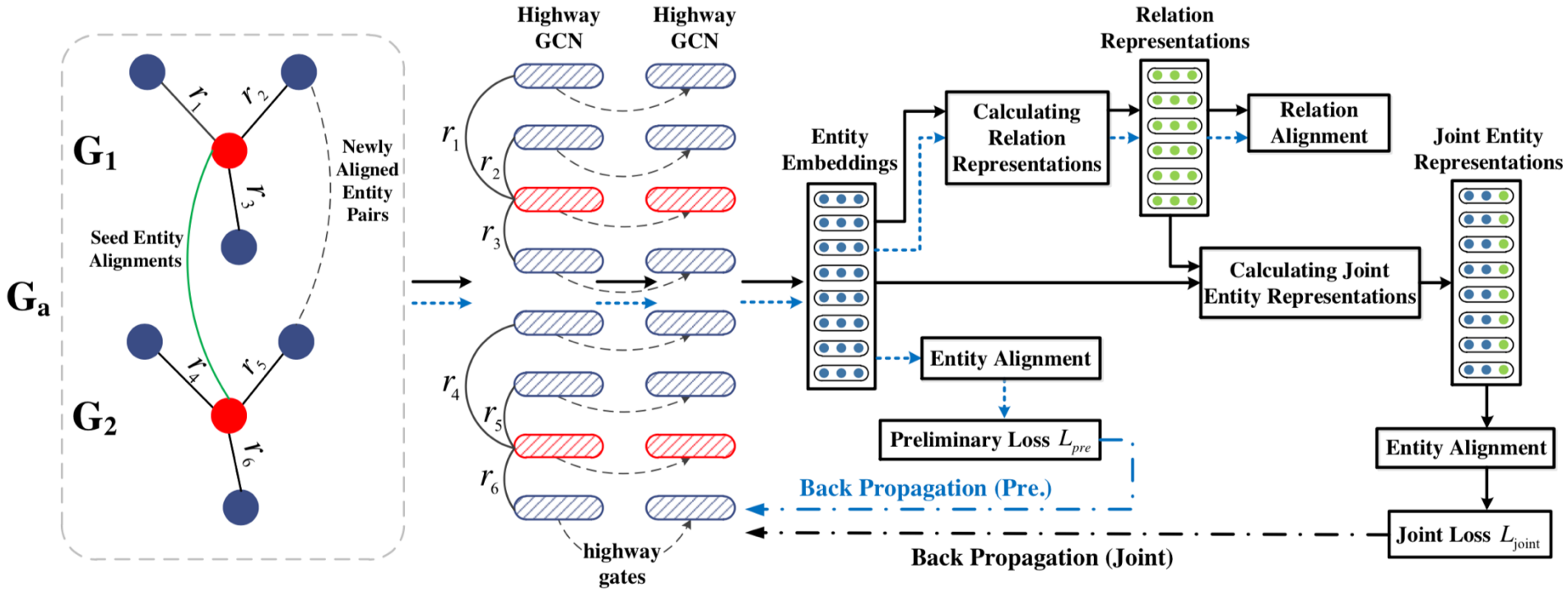
知识图谱的融合

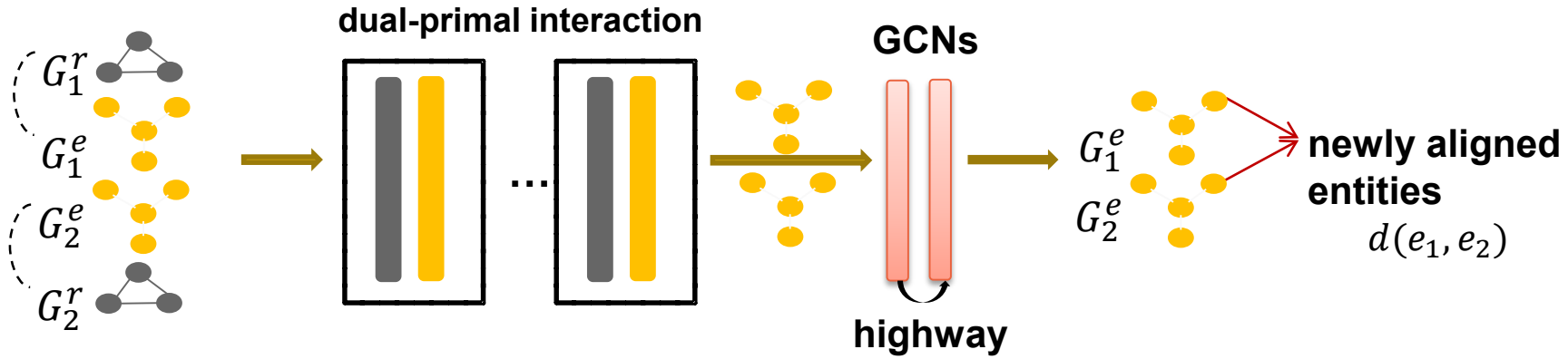
• 多源异构知识图谱的实体对齐

- **问题：多源、异构、跨语言知识图谱的差异性**
  - 不同来源实体的可用信息差别很大
    - 结构完全不可比
    - 名称表述差别大
    - 机器翻译等外部工具结果不稳定
  - 可用训练数据很少

- **方法：基于图神经网络的实体结构语义表示及匹配**
  - 利用对偶图等设计弥补实体和关系数目上的巨大差异
  - 利于图匹配技术引入全局信息促进实体对齐
  - 利用节点邻域信息引入模拟的实体关系信息，提升实体语义表示
  - 已发表在ACL2020/2019、EMNLP2019、IJCAI2019上

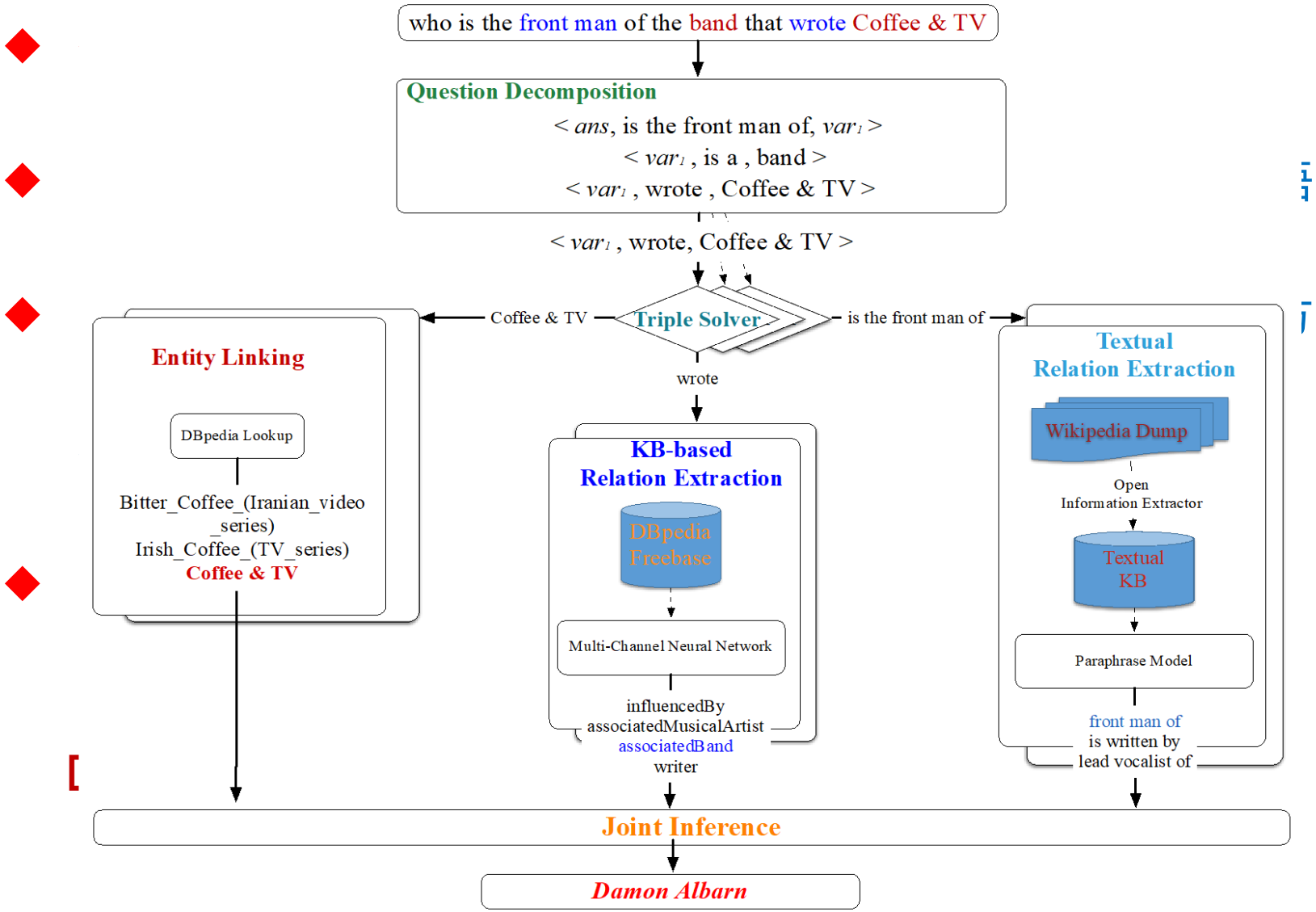
研究工作—多源异构图谱的实体对齐





研究工作之二— 语义理解与问题求解

- 基于大规模结构化知识资源的语义分析与理解



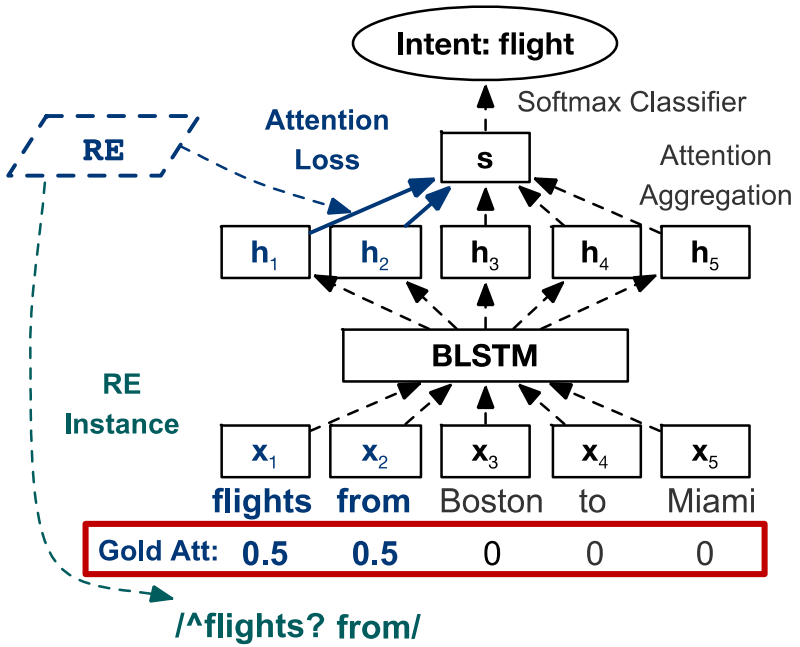
小样本学习问题

- **专家知识与神经网络模型的多方位结合**
  - **问题：将离散专家知识应用在神经网络模型**
    - 典型专家知识：正则表达式
    - 形式多样、复杂度可控
    - 与大规模标注数据之间存在互补性
  - **方法：将正则表达式以不同形式融入神经网络**
    - 以口语理解为典型应用场景
      - 意图识别：句子分类
      - 槽位解析：序列标注
    - 利用正则表达式对生数据进行初标注
      - 分别在输入端、特征构造、输出端、注意力模块等不同层次融合
      - 非常适宜小样本学习（few-shot learning）

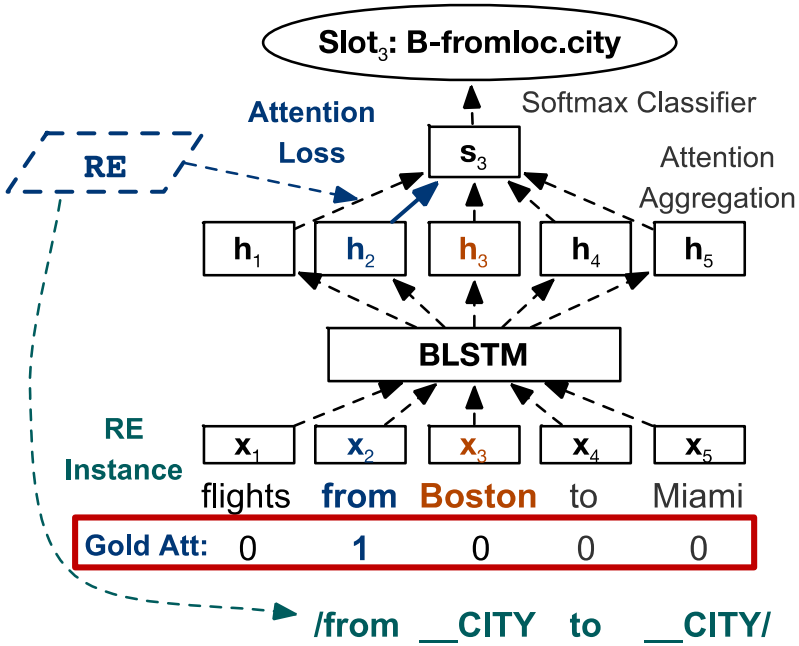
  已发表在ACL 2018 上

研究工作—基于正则表达式学习模型

- **Positive Attention** & **Negative Attention**



Intent Detection



Slot Filling

小样本的学习效果

- **Intent Detection——Macro F1/Accuracy**

	5-shot	10-shot	20-shot
base	45.28 / 60.02	60.62 / 64.61	63.60 / 80.52
feat	49.40 / 63.72	64.34 / 73.46	65.16 / 83.20
ouput	46.01 / 58.68	63.51 / 77.83	69.22 / **89.25**
att	**54.86 / 75.36**	**71.23 / 85.44**	**75.58** / 88.80
RE		70.31 / 68.98	

- **Slot Filling——Macro/Micro F1**

	5-shot	10-shot	20-shot
base	60.78 / 83.91	74.28 / 90.19	80.57 / 93.08
feat	**66.84 / 88.96**	**79.67 / 93.64**	**84.95 / 95.00**
ouput	63.68 / 86.18	76.12 / 91.64	83.71 / 94.43
att	59.47 / 83.35	73.55 / 89.54	79.02 / 92.22
RE		42.33 / 70.79	

研究工作—复杂问题的分步求解

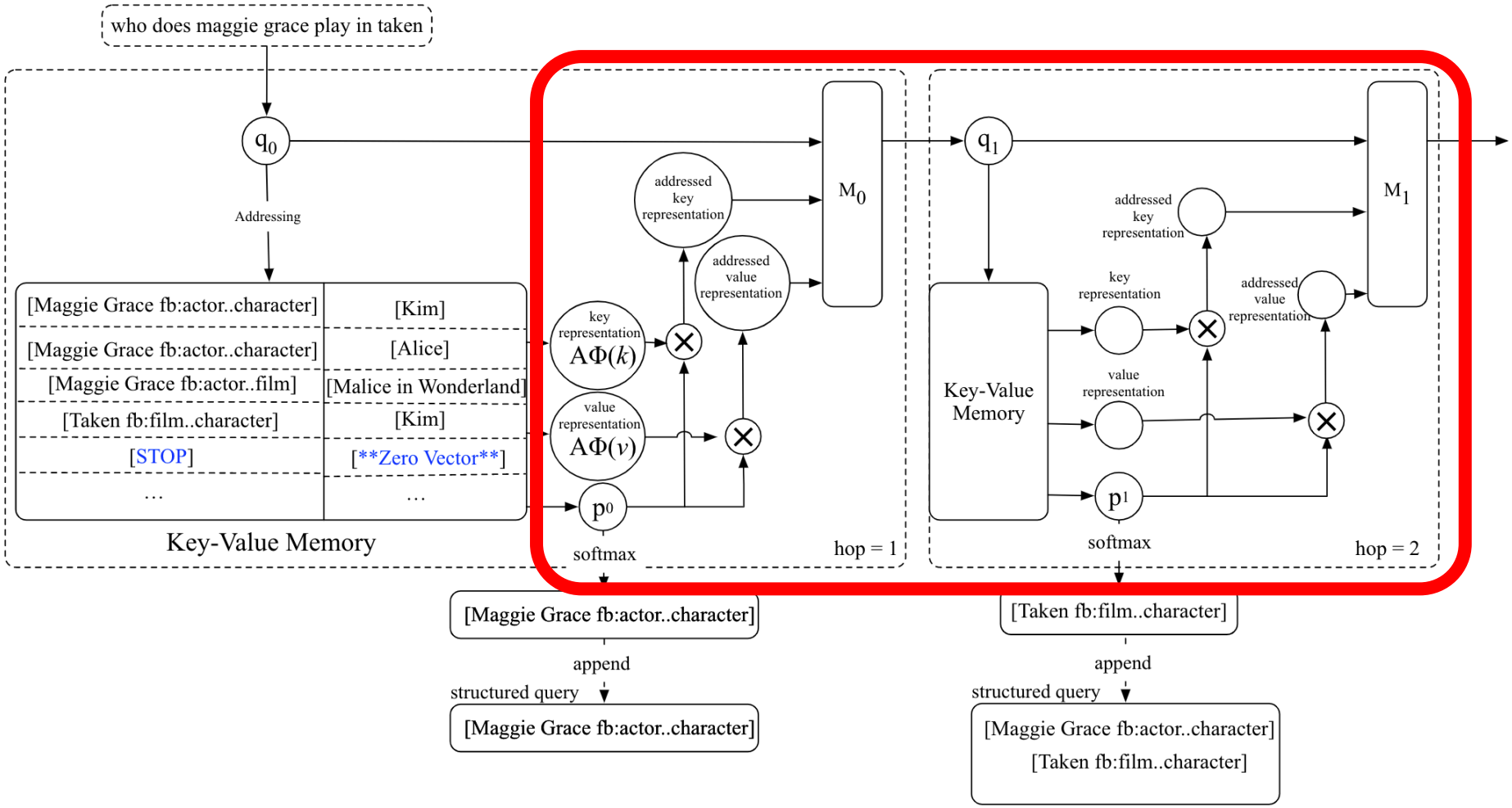
– 基于Key-Value Memory的自然语言问答技术

- **问题：**
  - 对复杂问题的准确解析依赖于较多的人工规则
  - 对复杂问题缺乏有效的问题拆分与组合技术
  - 对复杂问题缺乏有效的知识推理机制

- **方法：**
  - 利用 Key-Value Memory 存储结构化知识资源
  - 利用多次读取记忆模块来分步解析复杂问题
  - 利用记忆模块更新原始问题，分离已读取的部分
  - 学习 STOP 机制，从而避免无效的反复读取记忆模块

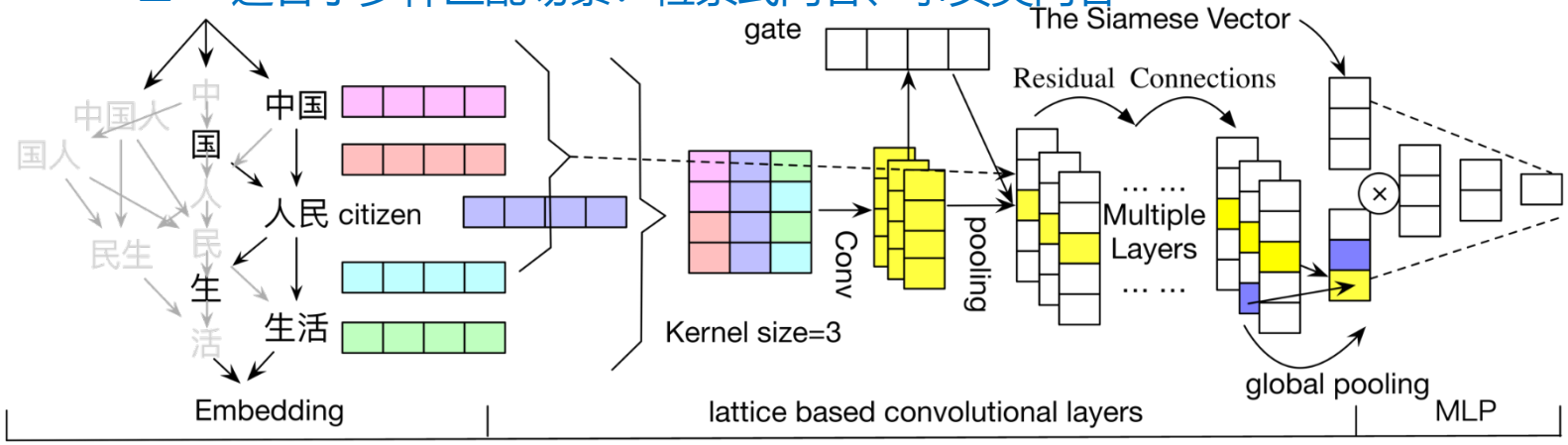
已发表在 NAACL 2019

复杂问题的分步求解


who does maggie grace play in taken
q_0
Addressing
[Maggie Grace fb:actor..character] [Kim]
[Maggie Grace fb:actor..character] [Alice]
[Maggie Grace fb:actor..film] [Malice in Wonderland]
[Taken fb:film..character] [Kim]
[STOP] [**Zero Vector**]
…
Key-Value Memory
key representation $A\Phi(k)$
value representation $A\Phi(v)$
p^0
softmax
addressed key representation
addressed value representation
M_0
hop = 1
q_1
Key-Value Memory
key representation
value representation
p^1
softmax
addressed key representation
addressed value representation
M_1
hop = 2
[Maggie Grace fb:actor..character]
append
structured query
[Maggie Grace fb:actor..character]
[Taken fb:film..character]
append
structured query
[Maggie Grace fb:actor..character]
[Taken fb:film..character]

中文知识库问答

- **基于多粒度匹配的中文自然语言问答技术**
  - **问题：**
    - 用词等细微的表述差异导致无法匹配
    - 常见匹配方法无法利用局部失配信息
    - 在中文处理中尤其严重
  - **方法：**
    - 利用Lattice-CNN进行多粒度：词、短语、不连续短语串的混合匹配
    - 缓解中文分词不一致、同义异形词等带来的问题
    - 适合于多种匹配场景：检索式问答、事实类问答



已发表在 AAAI 2019

研究工作之三—答案生成

- **基于异构知识的自然语言答案生成**
  - 希望答案句回答准确、表述自然、减少冗余

◆ **问题：**

- ◆ 通常只返回一个实体或数值，不能以自然语言形式进行回答
- ◆ 多数工作仅依赖一种知识来源，无法利用多种知识资源
- ◆ 多数知识资源之间可能存在重复、冲突矛盾

◆ **方法：**

- □ 利用 Key-Value Memory 建模异构知识资源
- □ 分离 Decoding state 挑选精确的候选信息
- □ 利用 Cumulative Attention 机制处理冗余信息
- □ 利用异构知识资源提高答案丰富度

[Fu et al, NAACL 2018]

研究工作—自然语言答案生成

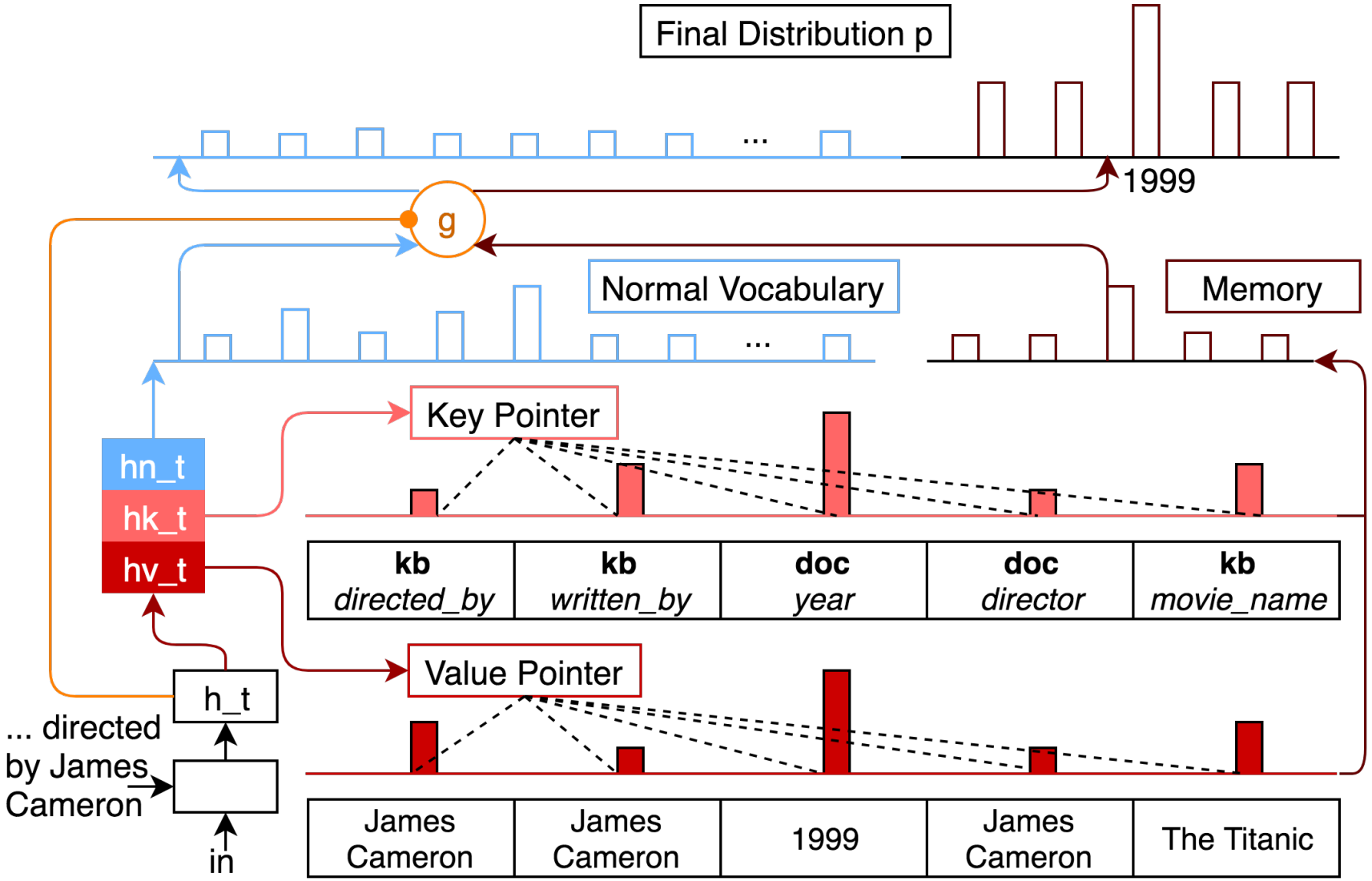
kb *directed_by*	Mark Haggard
kb *written_by*	Bruce Kimmel
doc *year*	1976
doc *directed*	Bruce Kimmel
kb *movie_name*	The First Nudie Musical
doc *director*	Mark Haggard

Q: the film The First Nudie Musical was directed by who?

A: The First Nudie Musical is a 1976 American motion picture directed by Mark Haggard and Bruce Kimmel.

memory from topic word | memory from knowledge base | memory from documents

自然语言答案生成模型


Final Distribution p
...
1999
g
Normal Vocabulary
...
Memory
Key Pointer
hn_t
hk_t
hv_t
kb directed_by
kb written_by
doc year
doc director
kb movie_name
Value Pointer
h_t
... directed by James Cameron
in
James Cameron
James Cameron
1999
James Cameron
The Titanic

内容提纲

- 背景介绍
- 文本语义理解技术
- 研究进展

研究进展—语义知识库构建与问答

- **大规模异构知识资源的语义网络构建技术**
  - **构建[illegible]5500万、[illegible]0**
    - [illegible]含类别概念10万个，类别信息300万条
    - [illegible]自然语言[illegible]
    - [illegible]项NSFC、4项863项目和重点研发计划支[illegible]

研究进展—高考地理因果类问题解答

- **考题示例**

题干:上海设立我国第一个自由贸易区的原因有？
选项: 1对外开放程度高，外向型经济特色显著 2矿产资源丰富，工业比较发达 3海上交通便利，便于货物运输 4区位条件优越，经济腹地广阔 A 123　　B 124　　C234　　D 134

- **挑战**
  - 直接/间接因果关系
  - 覆盖面广、时事性、需额外时间、地点、数字、事实等背景判断
  - 指代消解、篇章理解等困难

高考地理因果类问题解答

- **考题示例**

题干:马达加斯加地广人稀，水稻种植有着得天独厚的优越性，全国各地都有栽培，但仍不能完全自给。马达加斯加稻米不能自给的原因是？
选项: A粮食需求量增长过快 B农业技术落后产量低 C岛国 种植面积有限 D国际市场稻米价格低廉

- **挑战**
  - 需额外地点、数字、事实等背景判断
  - 篇章理解
  - 矛盾说法

因果类问题解答框架

核心因果指示词
词向量扩展
因果指示词

维基百科
基于标签的话题筛选
百度百科
因果指示词筛选
基于关键词的话题筛选
教材关键词统计
地理教材

地理因果语料库
因果指示词筛选
因果关系初步提取
地理因果关系库

地理因果关系知识库
基于谓词的因果要素抽取
基于引导词的因果要素抽取

断言组排名
逻辑打分

(可比的)因果断言组
拆分筛选
选项断言
题目背景
题干
选项
地理问题

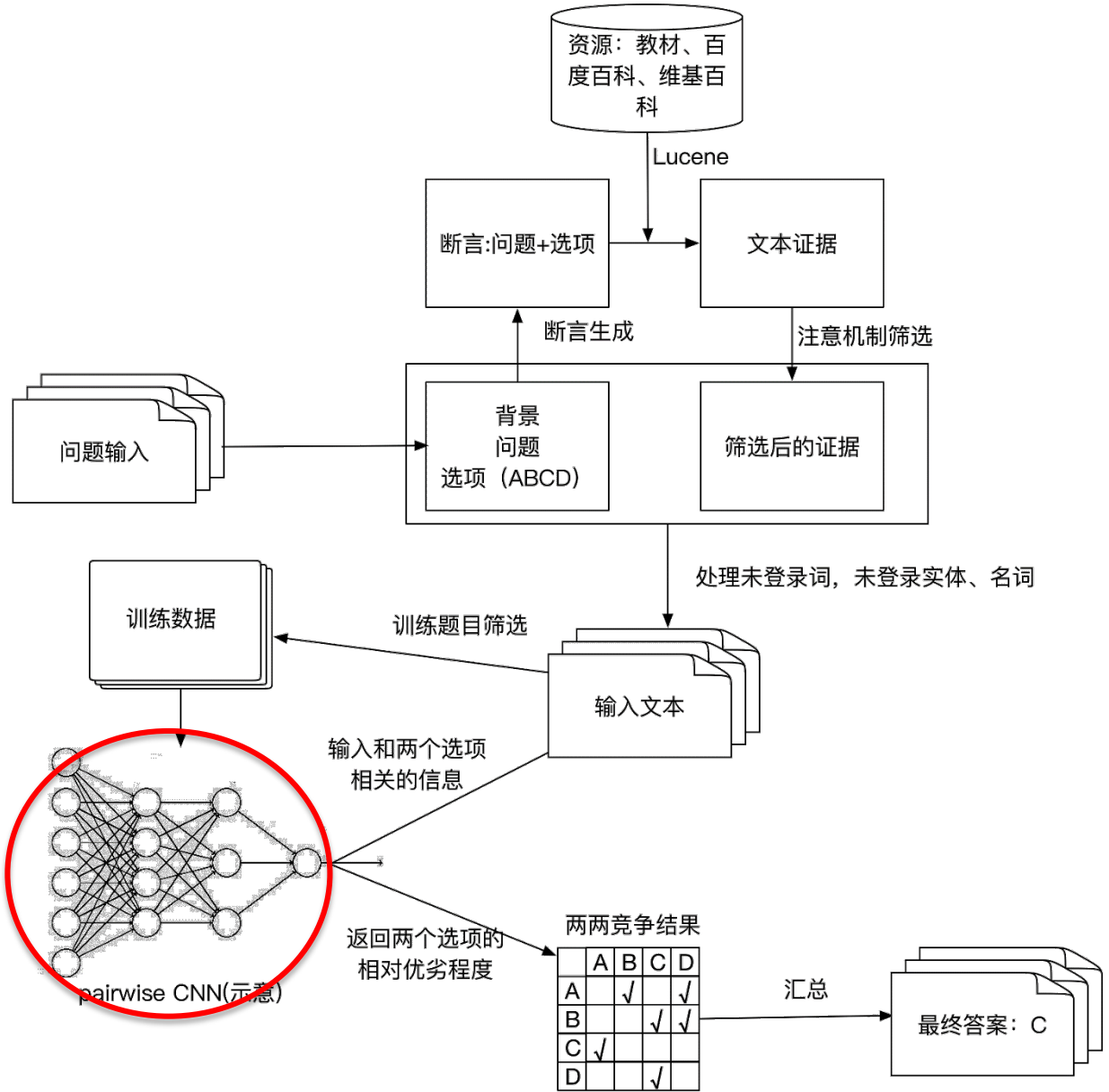
地理因果逻辑知识库构建

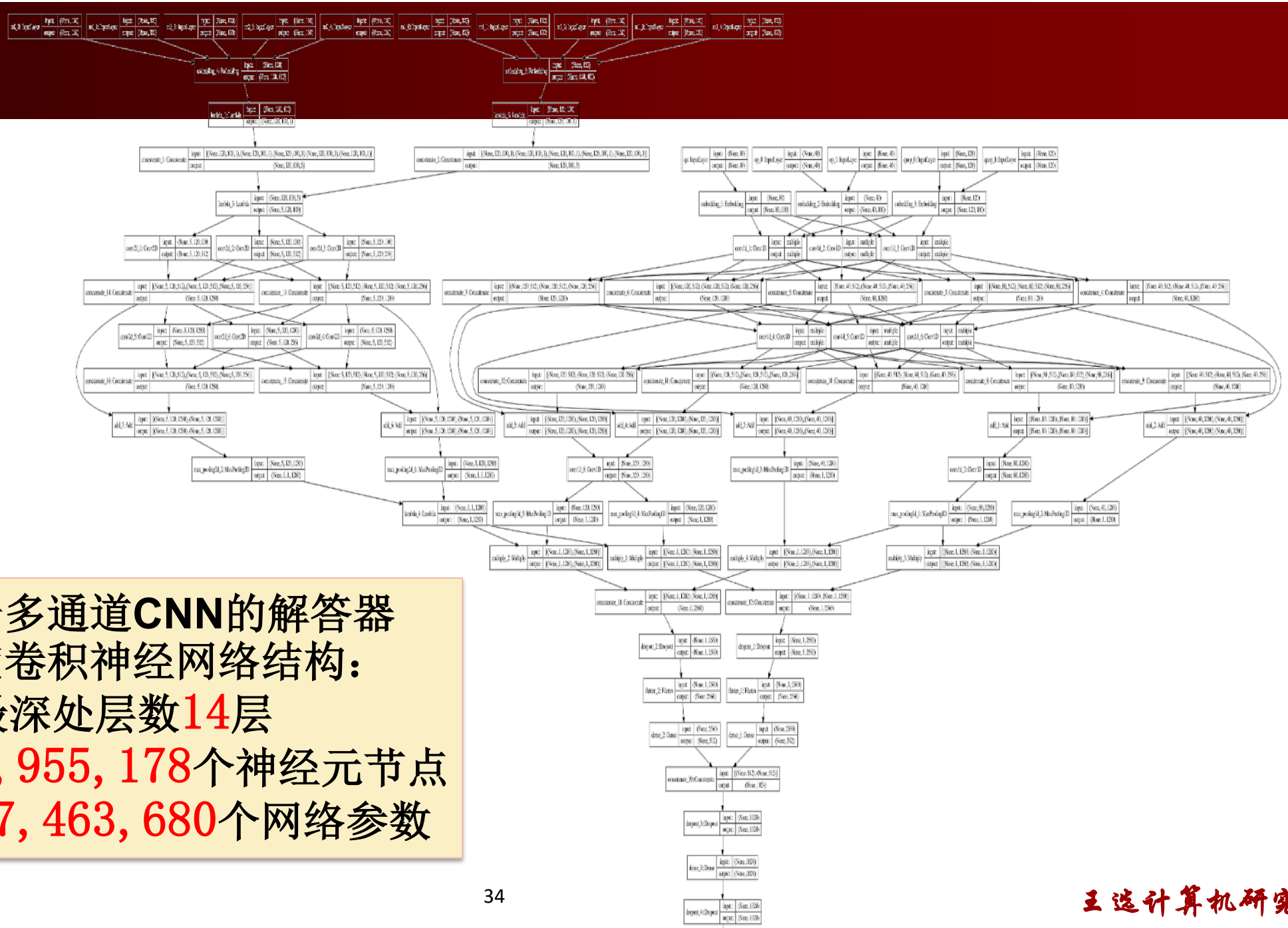
- 从地理教材、百度百科和中文维基百科得到13.49万地理因相关的因果逻辑关系条目
  - 组成**地理因果逻辑关系知识库**

地理教材	维基百科	百度百科	总计
928	13010	120962	**134900**

- XML格式存储，总大小92.8MB

融合多种知识资源的事实类问题求解





基于多通道CNN的解答器
加宽卷积神经网络结构：

- 最深处层数14层
- 3, 955, 178个神经元节点
- 27, 463, 680个网络参数

2017北京高考地理第7题
内蒙古：
A. 水域面积大，水能资源丰富
B. 受降水影响，森林覆盖率东部大于西部
C. 地势平坦，宜大幅度提高城市建设用地比例
D. 其他及未利用地面积比黔少，后备土地资源不足

断言补全

A. 内蒙古水域面积大，水能资源丰富
B. 内蒙古受降水影响，森林覆盖率东部大于西部
C. 内蒙古地势平台，宜大幅度提高城市建设用地比例
D. 内蒙古其他及未利用地面积比黔少，后备土地资源不足

神经网络多通道输入

[选项断言] 内蒙古 受 降水 影响 __<0>__ 森林 __NN_34__ 东部 大于 西部

[证据] 内蒙古 自治区 __NN_25__ 广袤__<0>__所处 纬度 较高__<0>__高原 面积 大__<0>__距离 海洋 较远，降水量 由 东北 向 西南 递减 __<0>__

未登录词处理

检索知识库 & 排序 & 过滤

【证据】内蒙古自治区域势广袤，所处纬度较高，高原面积大，距离海洋较远，降水量由东北向西南递减。[来源：百度百科]

根据教材
百度百科
维基百科

地理学科知识库

神经网络比较器

选项两两打分结果：

净角分 \ 视角分	A	B	C	D
A	A:B→.090	.910	A:C→.765	.235
B	B:A→.876	B:C→.930	B:D→.743	.257
C	C:A→.235	C:B→.070	C:D→.433	.567
D	D:A→.124	D:B→.257	D:C→.567	D:0.954分

整合最 A:1.73分 C:0.74分

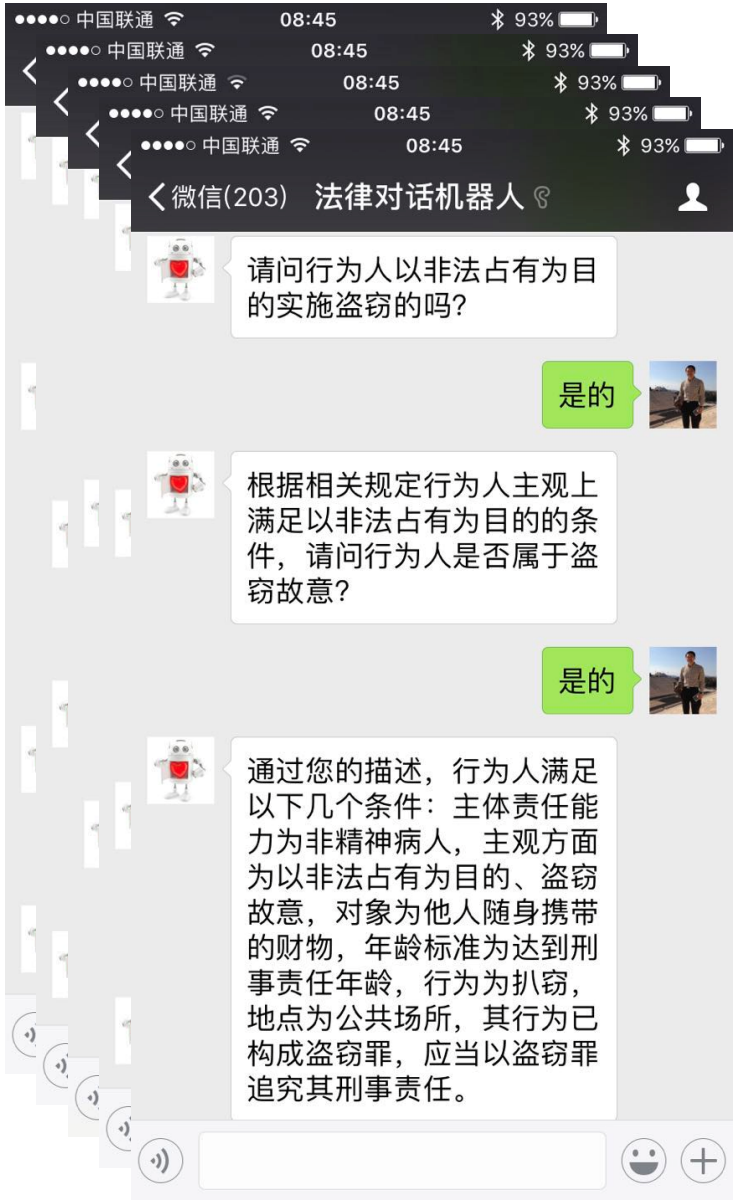
答案:B

应用进展—基于知识理解与生成的智能应用

- ◆语义搜索技术
  - ◆智能搜索引擎
  - ◆智库系统、情报分析系统
- ◆+问答技术
  - ◆知识问答系统——Answer Engine
  - ◆辅助诊疗系统；智能审判系统
- ◆+对话技术
  - ◆理财顾问机器人；法律助理机器人；智能客服机器人

法律对话机器人——Lawbot

- 多轮主动式对话技术
  - 问：偷二百元算不算犯罪?
  - ……
  - Rot: 当事人是否精神正常?
  - ……
  - R: 在哪里偷的钱
  - H: 在百货商场
  - ……
  - R: 偷了谁的钱?
  - ……
  - R: 其行为构成/尚未构成犯罪

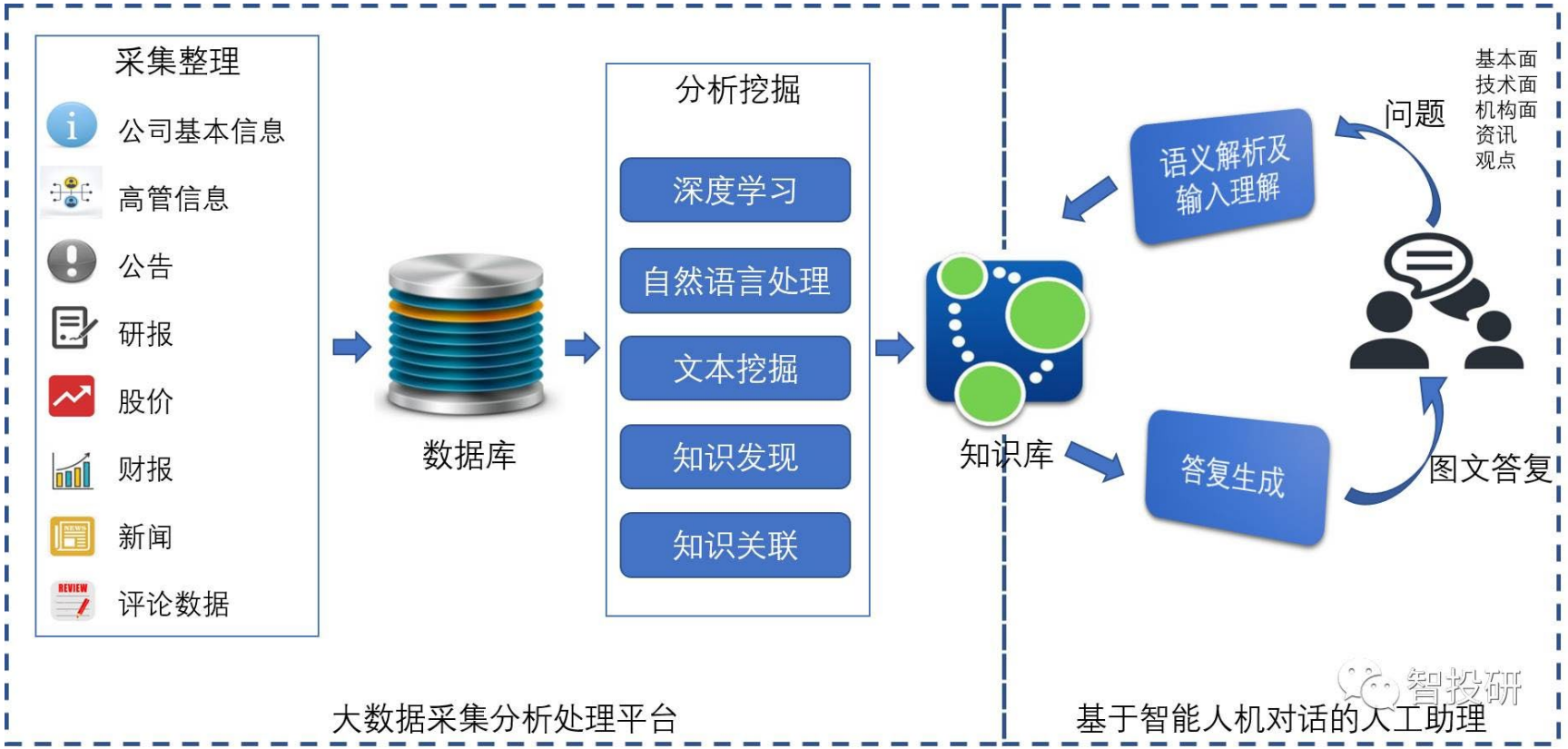


研究工作进展— 知识服务应用

- **保险条款解析**
  - **构建保险产品知识图谱**
  - **保险产品的量化评估**

智能投资助手

基于知识图谱和对话技术的智能证券投资助手



智能投资助手-整合多源数据

数据收集、整理

- 公司、人员、行业、产品、知识产权
- 基本信息、财务数据、股价信息，及周边数据（大宗商品价格等）
- 事件及关系（如投融资、并购重组、产品及项目、诉讼）
- 个股、板块、行业、上下游
- 包括文本类、数值类的信息

分析挖掘

- 挖掘行业、板块、个股之间的关联关系
  - 传导效应
  - 轮动效应
- 数据模型与时间序列
- 事件模型，跟踪及因果关系分析
- 预测模型

数据情况-上市公司

- 3300多家基本信息及9万多名高管
- 200多万公告（含年报、季报），回溯至IPO，其中包含了历年（季）的财务数据
- 50万研究报告（公司研究、行业研究、宏观研究、投资策略）
- 二级市场交易数据（实时股价、近7年的历史量价信息500万条），行业指数
- 130万会议议案，挖掘7类共3万多条事件（提供担保、申请额度、签订协议、使用资金、设立公司、收购股权、开展业务）
- 公司行业、主营产品、主营地区，行业上下游关系、板块信息
- 财经资讯：30多万，含多年的公司资讯18万
- 可扩展：工商登记，股东，涉诉，人事、知识产权，行业专业数据等

智能投资助手-分析与模型

产业链分析

- 通过构造上下游产业链知识图谱，基于经济基本面建立传导模型
- 当产业链中重要节点的状态发生变化时，将启动沿产业链传导推理引擎，自动给出影响范围、对象和程度，为事件引发的基本面分析做支持

经济事件影响

- 基于金融知识图谱和推理逻辑，找到未来趋势的变化或者解释已经发生过的事情
- 如：行业中发生某一事件时，可沿产业链向上游进行传导推理，并生成分析影响报告

研报评价

- 研报分析回溯及结果跟踪
- 对研报（按人、机构）的正确率进行第三方评价

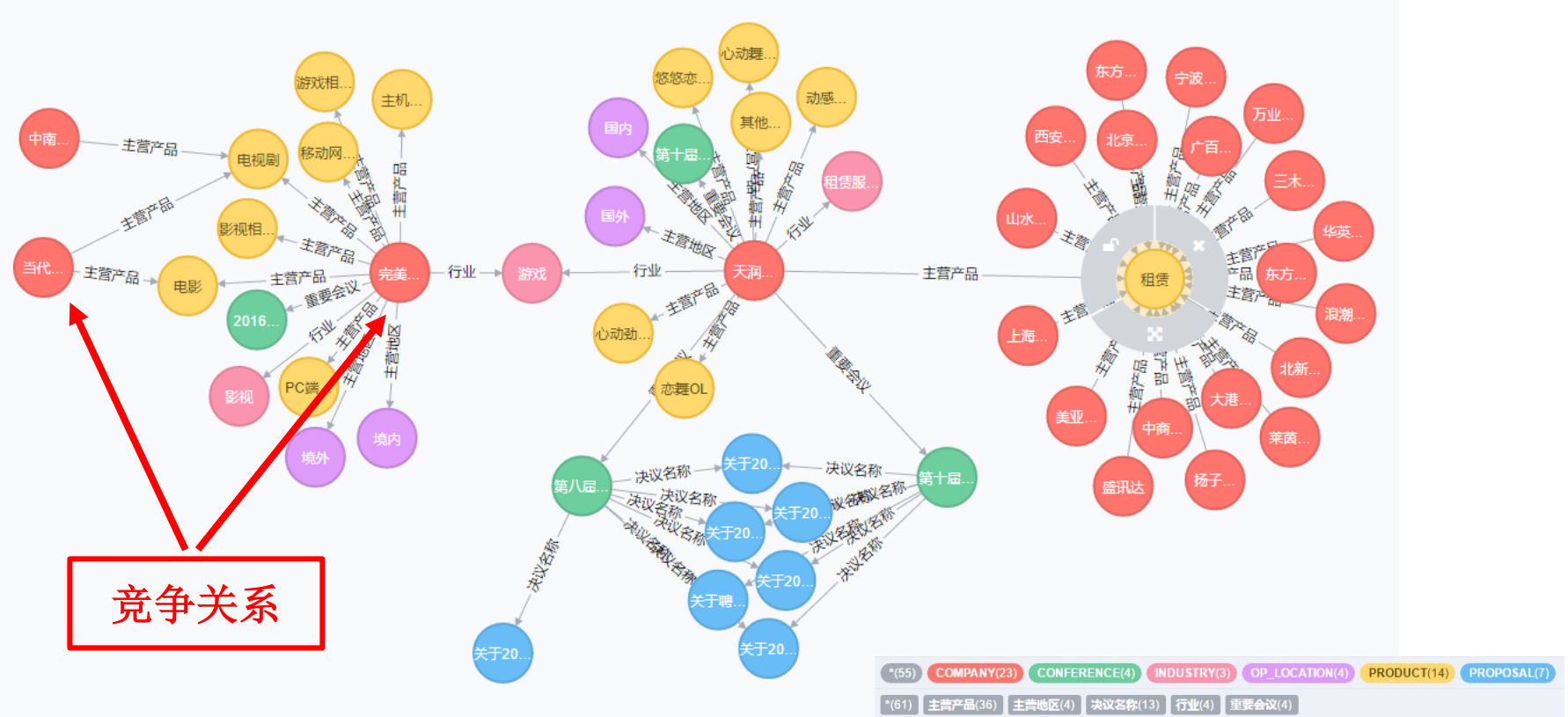
财报解读

- 分析提取财报的核心要素
- 针对不同的关注点给出相应的解读

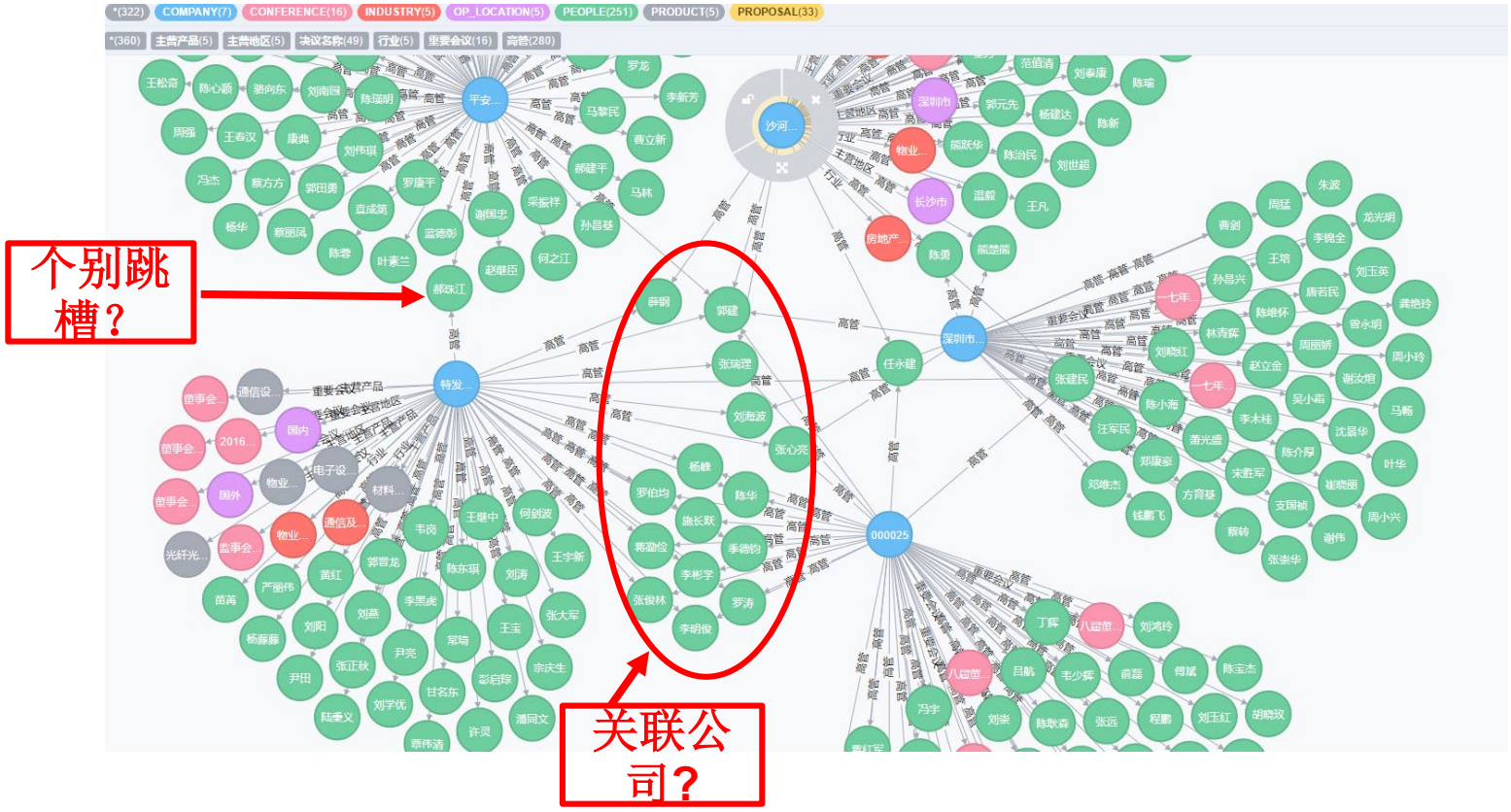
文本量化因子

- 基于文本的量化因子发现及因子有效性评价

知识图谱示例：公司、产品、行业、地区，会议及决议

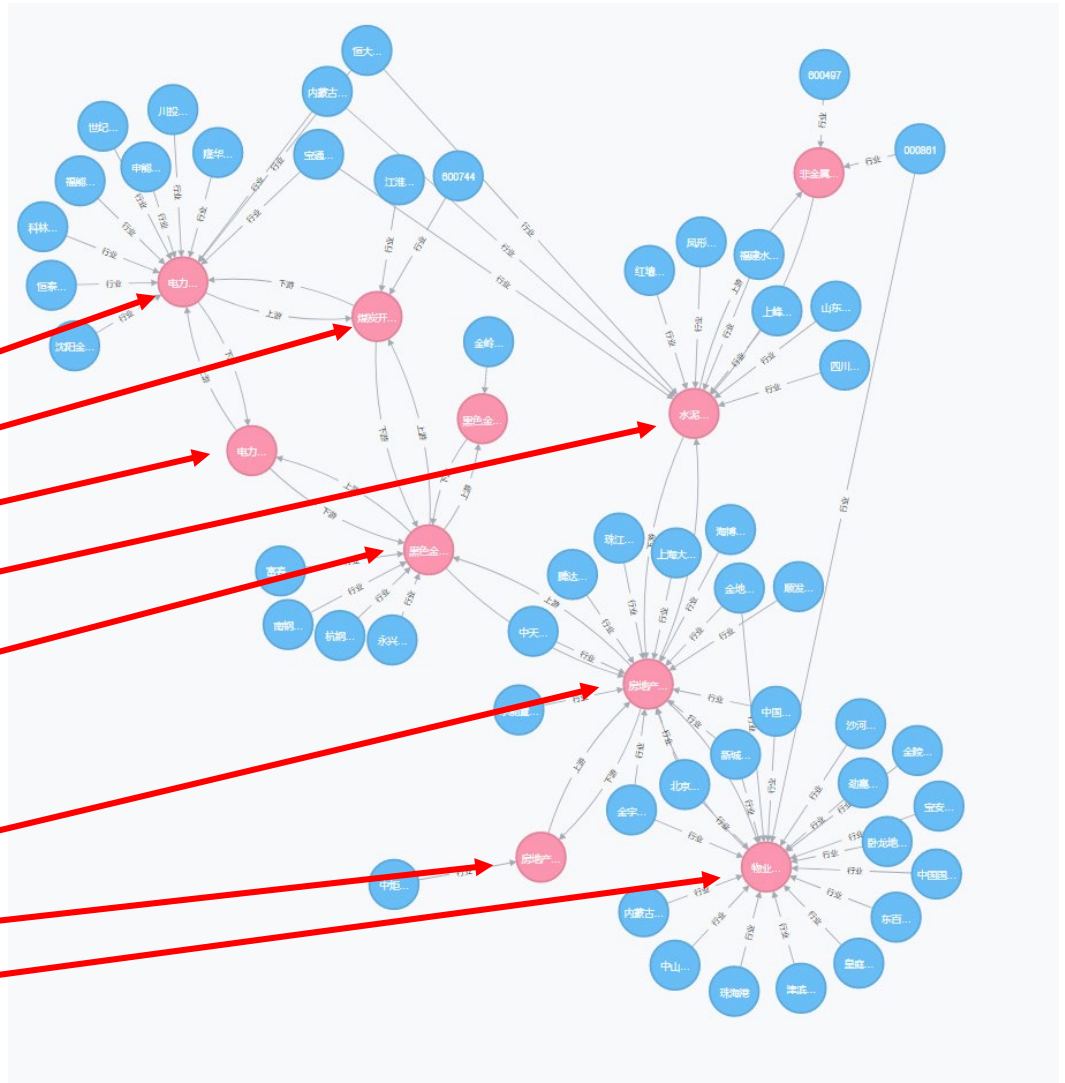


知识图谱示例：高管及公司的关联关系



公司、行业上下游关系

可分析重大经济事件在产业链的传播和影响情况

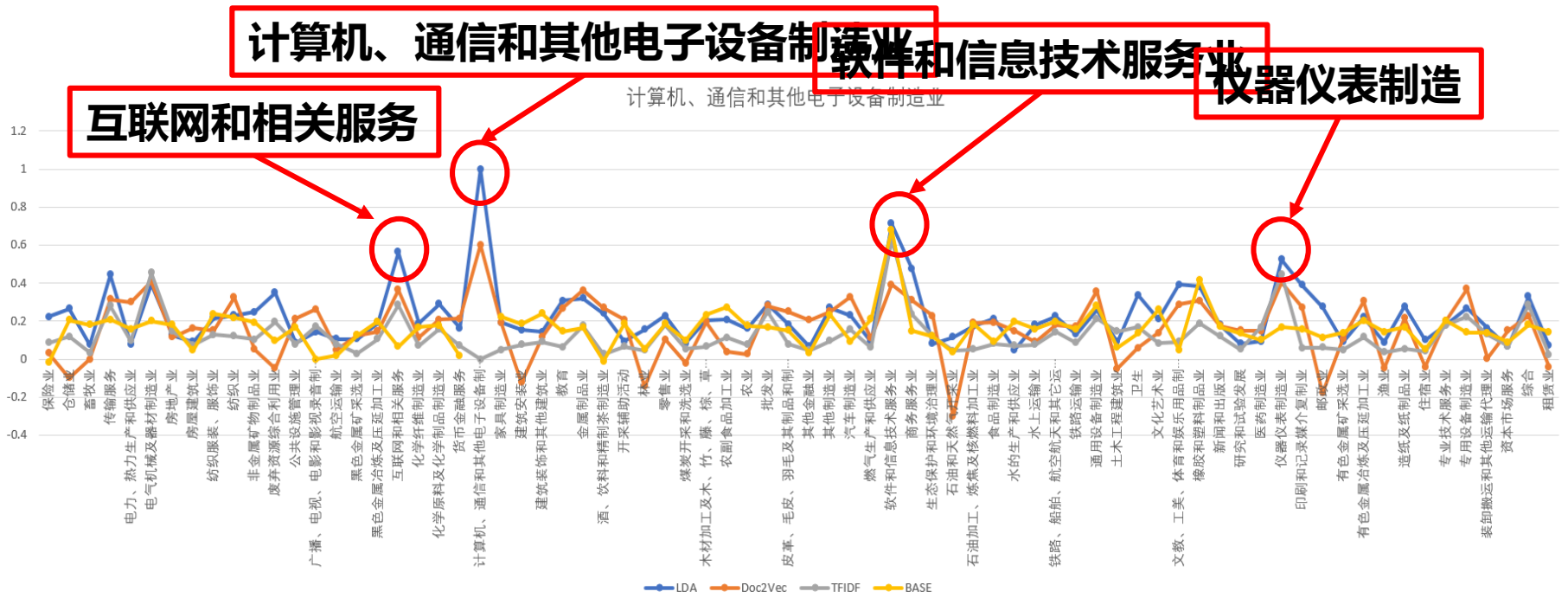


电力生产
煤炭开采
电力供应
水泥制造
黑色金属冶炼和压延加工业

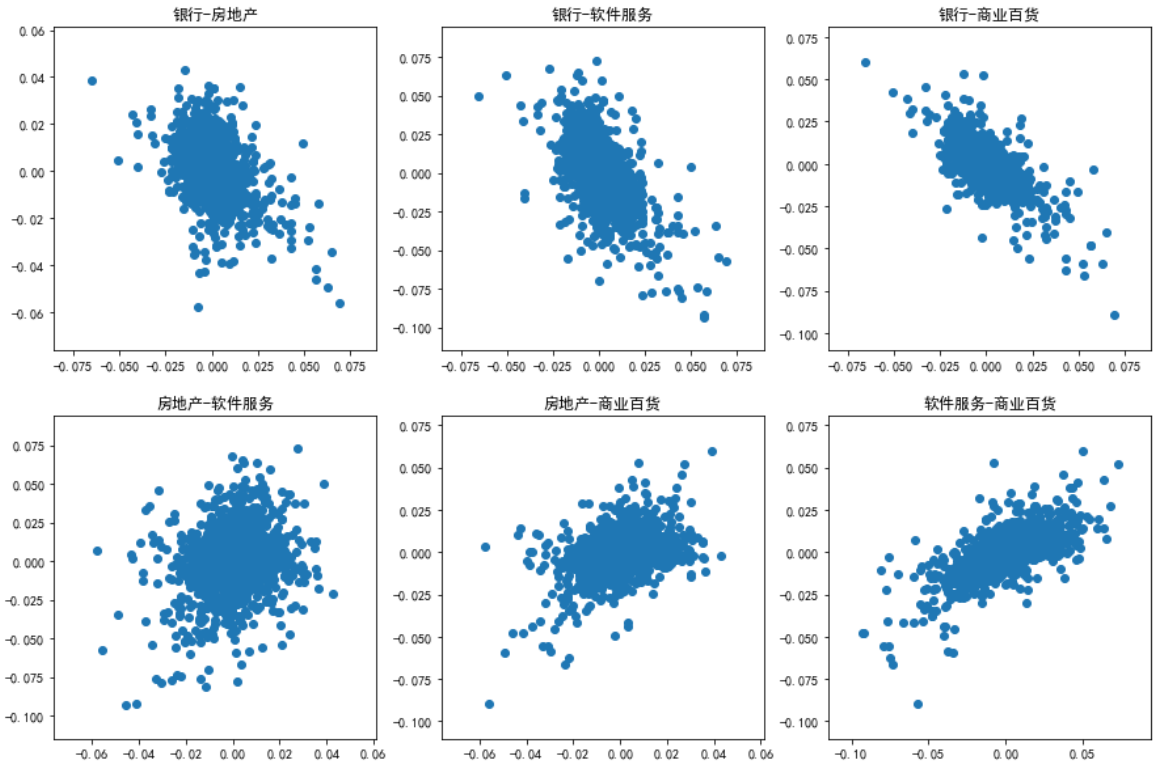
房地产开发经营
房地产中介服务
物业管理

行业相关性分析

使用多种文本相似性分析及数值分析方法



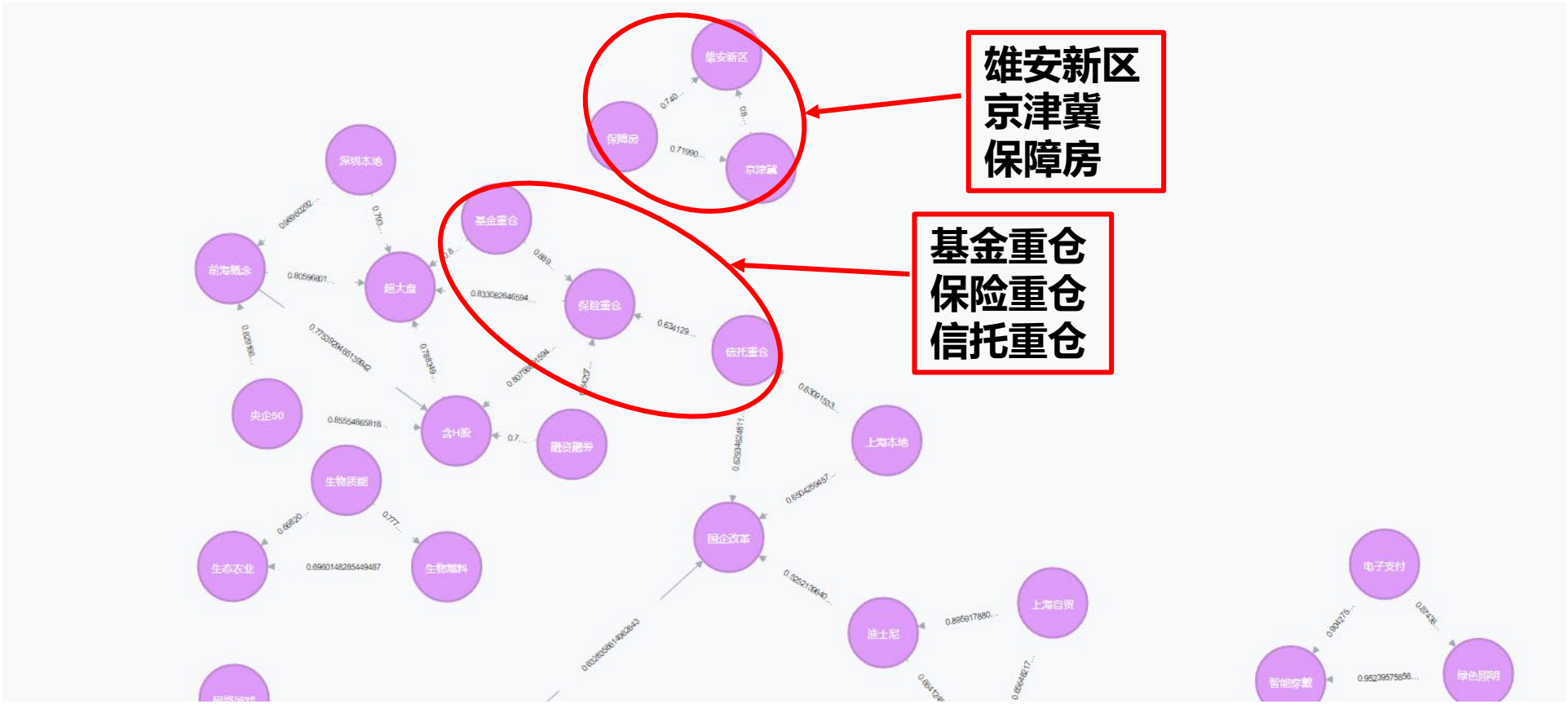
行业相关性分析（部分行业指数）



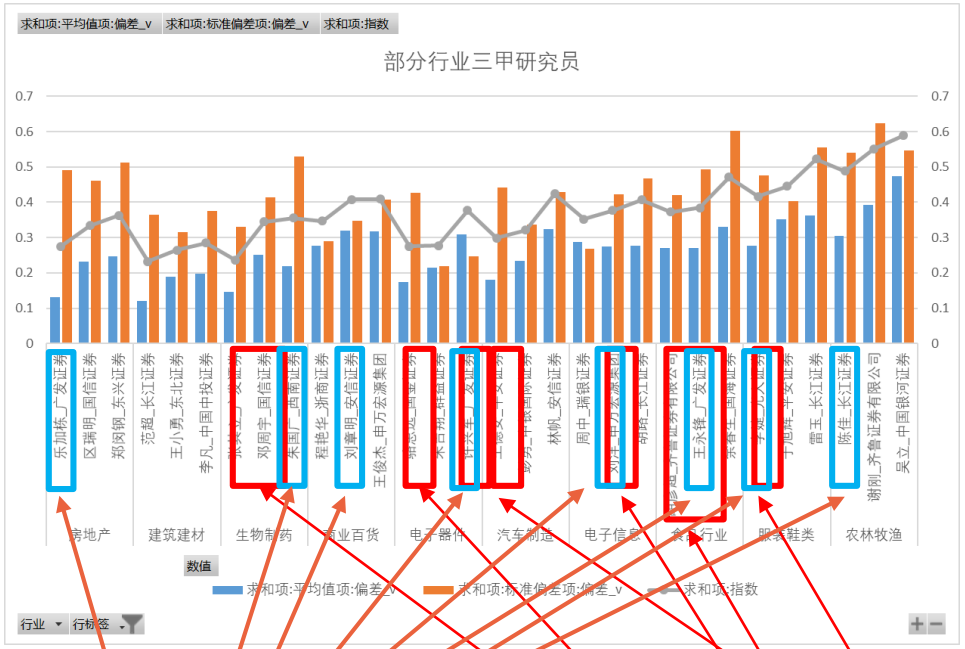
- 指数涨跌，扣除大盘影响
- 银行与软件存在较弱的负相关
- 银行与商业百货的负相关较为明显
- 房地产与软件服务独立性强
- 房地产商业百货存弱正相关
- 软件服务与商业百货较强正相关
- 轮动效应

板块相关性分析

使用数值分析方法（历史股价信息）



证券分析师第三方评价

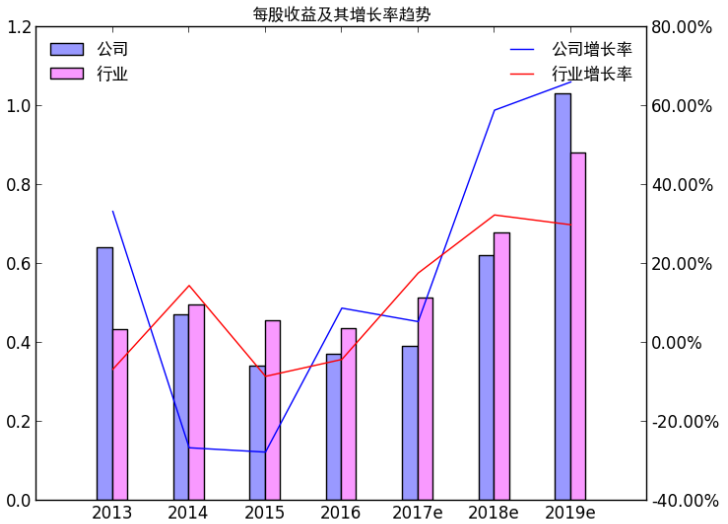


新财富2016最佳分析师

第一财经2016最佳分析师

- 分析师的公司研报中公司eps预测值与实际值偏差的回溯验证分析
- 近三年行业内报告数量为Top10的分析师中
- 以12个月内的预测偏差平均值、标准偏差，及报告数量加权形成指数
- 按行业内指数排名选出各行业的三甲分析师
  - 与部分媒体评选的2016最佳分析师的对比

智投研：基于人机对话的智能投研助手



智投研—量化投资

- **量化投资机器人**
  - 创立了两支模拟基金
  - 18年底19年初建仓
  - 跑赢对比基准和沪深300
  - 年化收益超过100%
- **19年11月6日成立实盘基金**
  - 募集350w资金
  - 依托基金公司、实盘操作
  - 跑赢对比基准中证800

日期	净值/基准净值	持股数/仓位	年化收益/基准年化
2020-09-03	1.3065 1.232	49 0.886	38.00% 28.57%
2020-09-02	1.3079 1.2396	49 0.886	38.32% 29.64%
2020-09-01	1.3034 1.239	50 0.904	37.89% 29.68%
2020-08-31	1.3003 1.2321	50 0.904	37.64% 28.91%

日期	净值/基准净值	持股数/仓位	年化收益/基准年化
2020-09-03	1.4758 1.1838	43 0.989	104.19% 36.27%
2020-09-02	1.4759 1.1912	43 0.989	104.95% 38.06%
2020-09-01	1.4632 1.1906	42 0.975	102.43% 38.16%
2020-08-31	1.4603 1.1839	42 1	102.41% 36.94%

北京大学

谢 谢!