信息抽取前沿动态

陈玉博

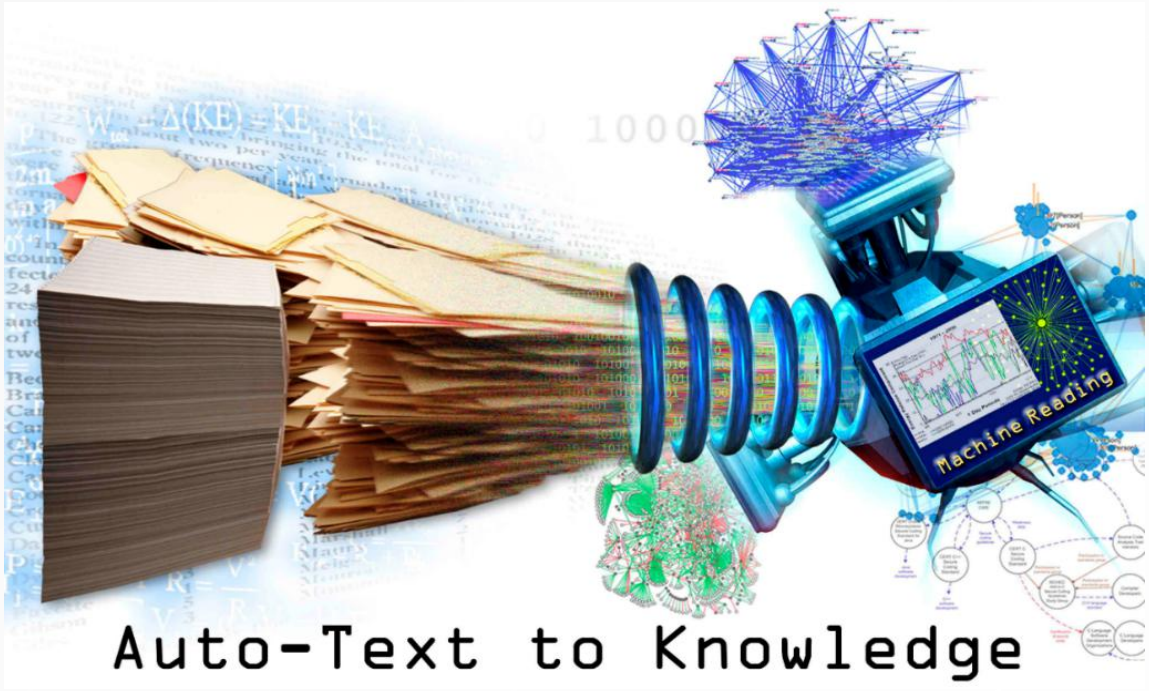
中国科学院自动化研究所

模式识别国家重点实验室

信息抽取

- 定义: 从自然语言文本中抽取指定类型的实体、关系、事件等事实信息，并形成结构化数据输出的文本处理技术。

(Grishman, 1997)



主要任务

识别出待处理文本中指定类型的命名实体。

例：我想买个最新款的苹果，大家都建议我买华为。

实体识别

实体消歧

消除实体的歧义，将实体按语义进行聚类或链接到知识图谱上

例：

知 识 抽 取

获取实体的属性或实体之间的语义关系

例： 新款苹果1200万像素，后置三摄。

关系抽取

事件抽取

抽取用户感兴趣的事件信息，如什么人、在什么时间做了什么事

例：隔壁实验室的学霸昨天花9999元给女朋友买了一个最新款的苹果。

任务难点

- 知识层面
  - 知识类型丰富多样：语言知识，世界知识，常识知识，领域知识等
  - 知识结构多元复杂：概念，实体，属性，关系，事件等
- 语言层面
  - 歧义性：他刚**离开**公司 VS 他永远**离开**了我们
  - 多样性：我 VS 爷 VS 娘 VS 朕 VS 哀家 VS 鄙人 VS 老夫 VS 小生 VS 贫僧 VS 贫道 VS 洒家 VS 小可 VS 在下 VS 本座 VS ……
  - 非规范性:火星文、表情符号、错误……

相关评测

- MUC（Message Understanding Conferences, 1987-1997）
  - 由美国国防高级研究计划委员会DARPA资助
  - 主要是英文，后两届扩展到中文
  - 任务：命名实体识别，模板关系抽取等等

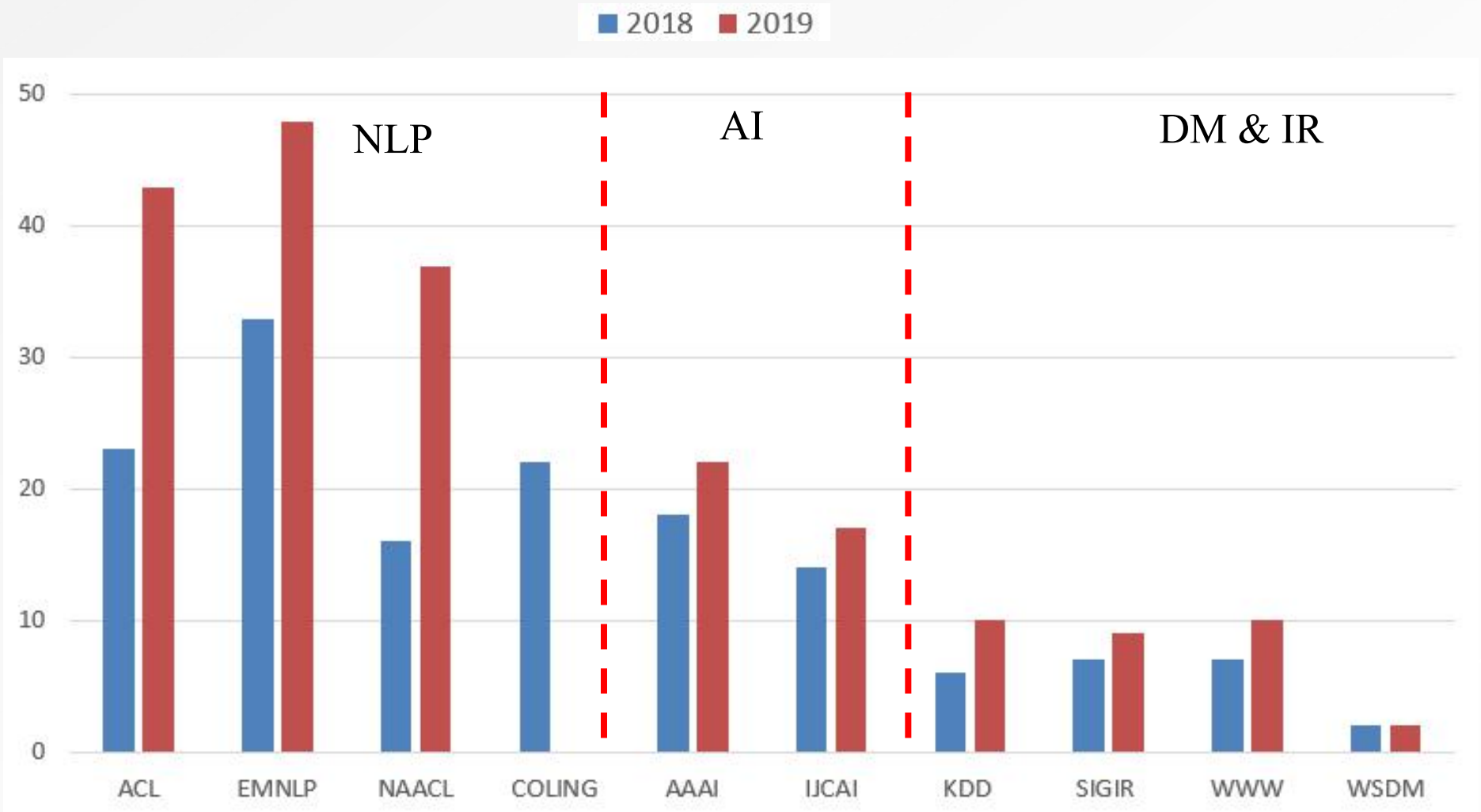
- ACE（Automatic Content Extraction, 1999-2008 ）
  - 由美国国家标准与技术研究所NIST主办
  - 2009起，ACE变成了TAC（Text Analysis Conference）的一项子任务
  - 英文、中文、阿拉伯文等
  - 任务：命名实体识别，实体消歧，关系抽取，事件抽取等等

- TAC-KBP(Knowledge Base Population) (2009-2019)
  - 任务：实体识别和消歧、属性抽取、事件抽取、情感分析等等

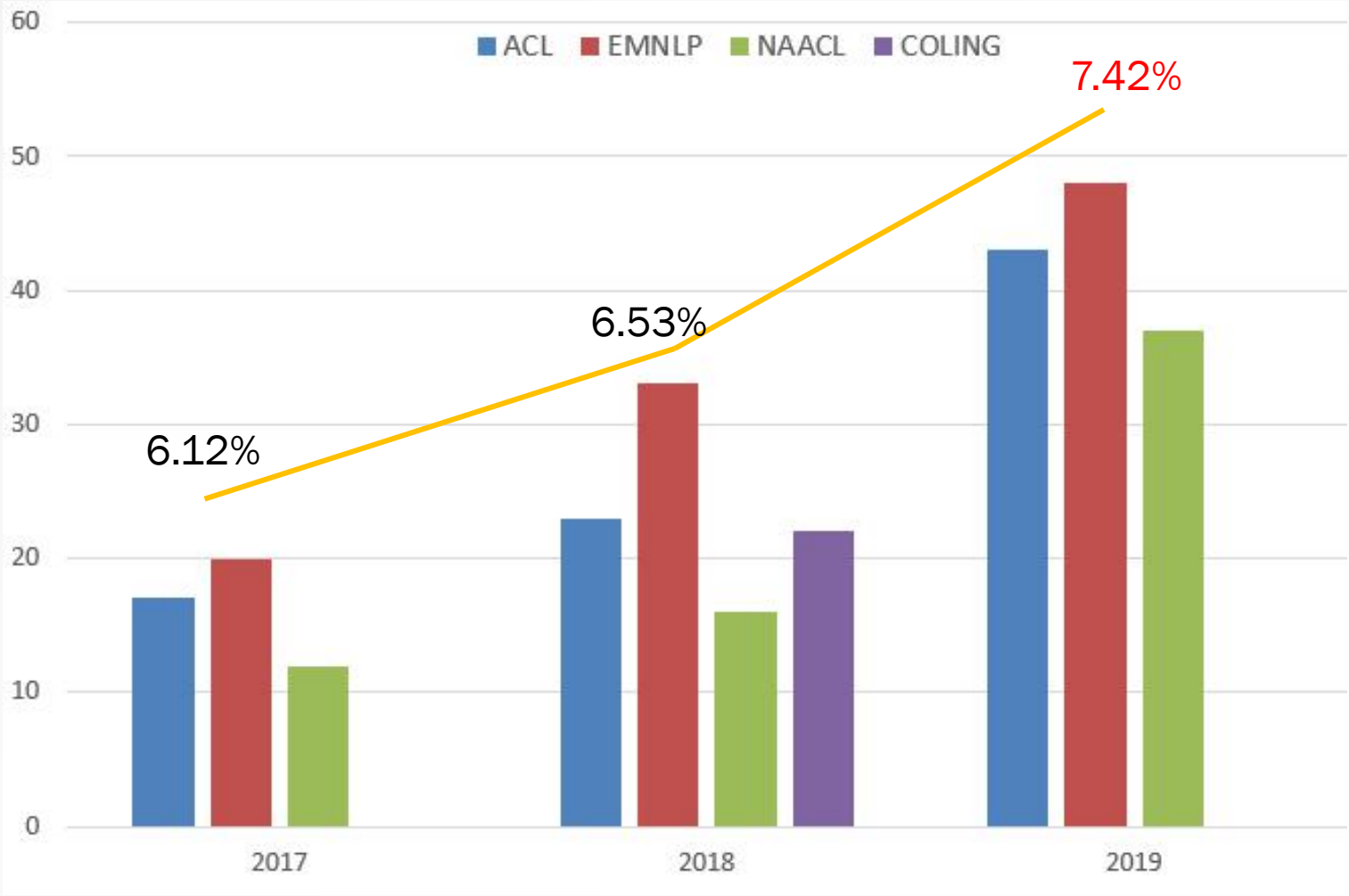
相关评测

任务	相关评测
实体识别	MUC, ACE, TAC-KBP, SigHAN, CoNLL
实体消歧	TAC-KBP, ACE WePS
关系抽取	MUC, ACE, TAC-KBP, SemEval，FewRel，DocRED
事件抽取	MUC, ACE, TAC-KBP, TDT, BioNLP

信息抽取会议论文分布


2018
2019
50
40
30
20
10
0
NLP
AI
DM & IR
ACL
EMNLP
NAACL
COLING
AAAI
IJCAI
KDD
SIGIR
WWW
WSDM

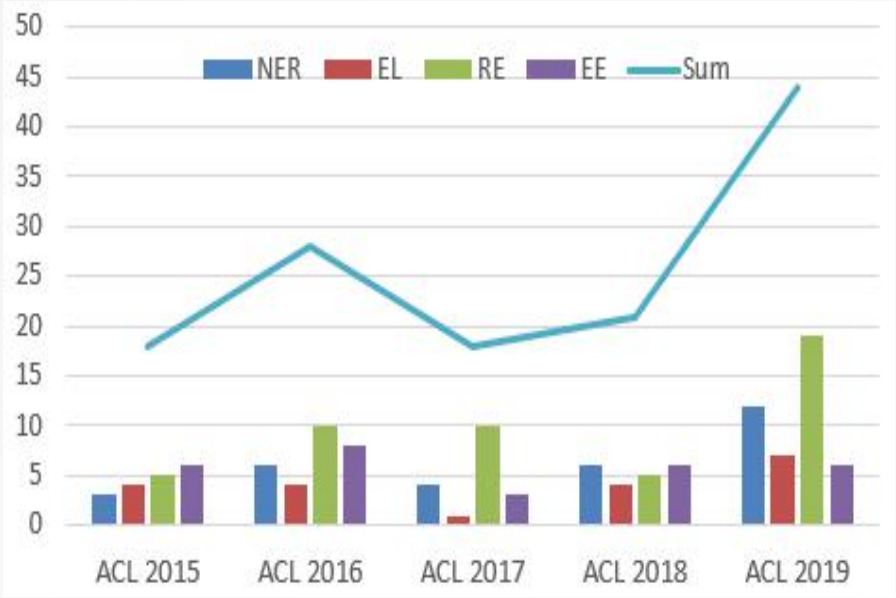
信息抽取会议论文数量及比例

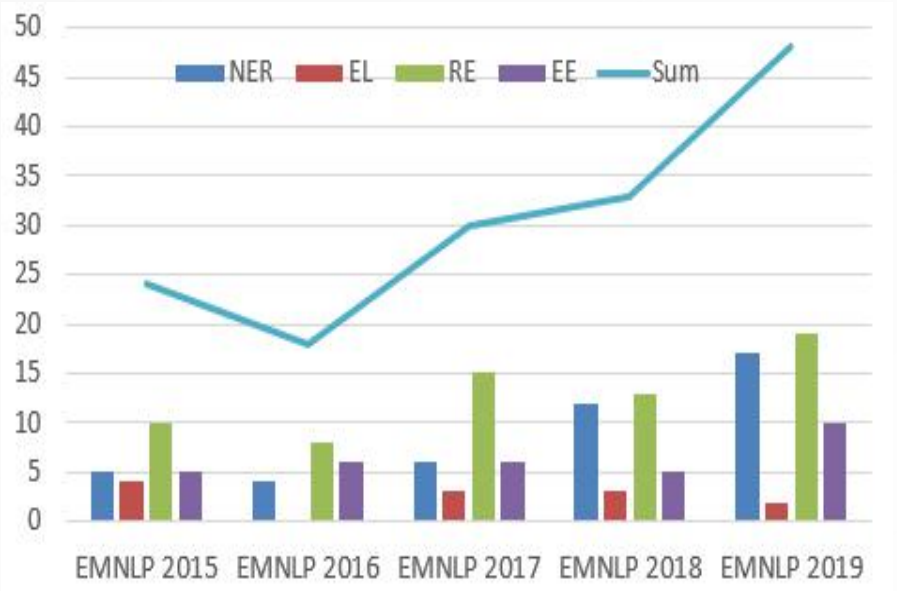


相关信息统计自ACL Anthology: https://www.aclweb.org/anthology/

NAACL17年未召开利用16年代替，COLING两年一次

信息抽取子任务会议论文数量





关系抽取相关的研究最多，事件抽取逐渐成为研究的热点

相关信息统计自ACL Anthology: https://www.aclweb.org/anthology/

研究热点

ACL 2019

EMNLP 2019

任务挑战

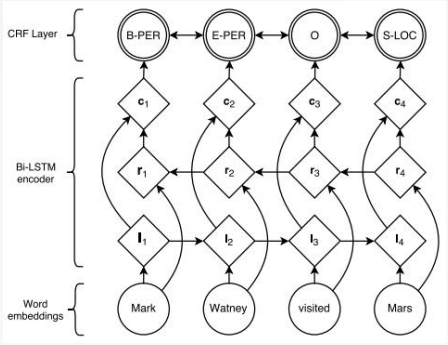
特征 如何（自动）抽取到有效、鲁棒的特征

语料 如何（自动）构建大规模、高质量语料

任务 如何实现多任务联合抽取

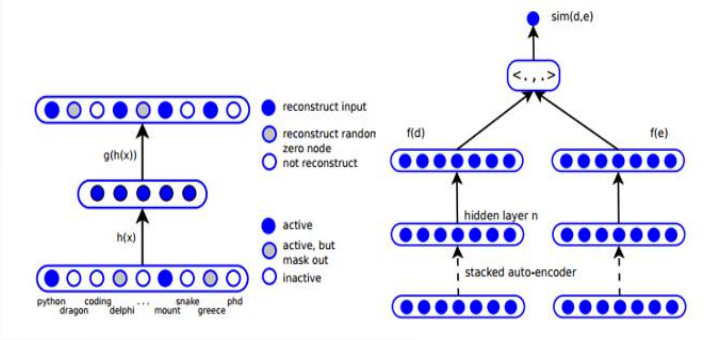
研究趋势

- **特征：**特征工程 → 特征学习
- **解决的问题：**人工设计特征，可扩展性差，依赖复杂NLP工具



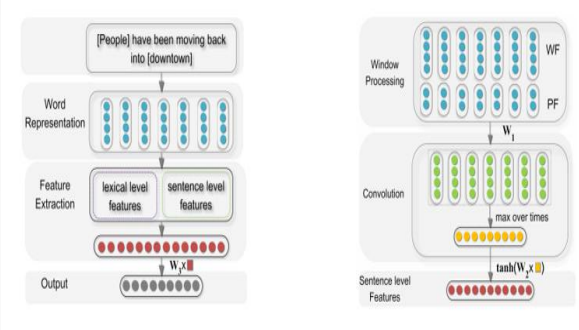
实体识别

实体消岐



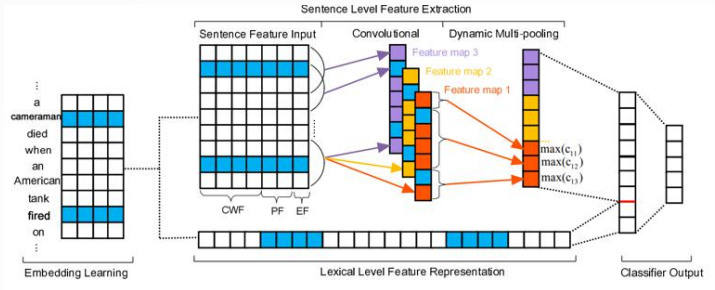
LSTM： Neural Architectures for Named Entity Recognition（Lample et al. NAACL 2016）

Auto-encoder： Learning Entity Representation for Entity Disambiguation（He et al. ACL 2013）



关系抽取

事件抽取



CNN： Relation Classification via Convolutional Deep Neural Network （Zeng et al. COLING 2014）

DMCNN： Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks（Chen et al. ACL 2015）

研究趋势

- **特征**：词→句子→篇章
- **解决的问题**：仅利用词或者句子信息，忽略篇章信息

他离开了公司。

Transport

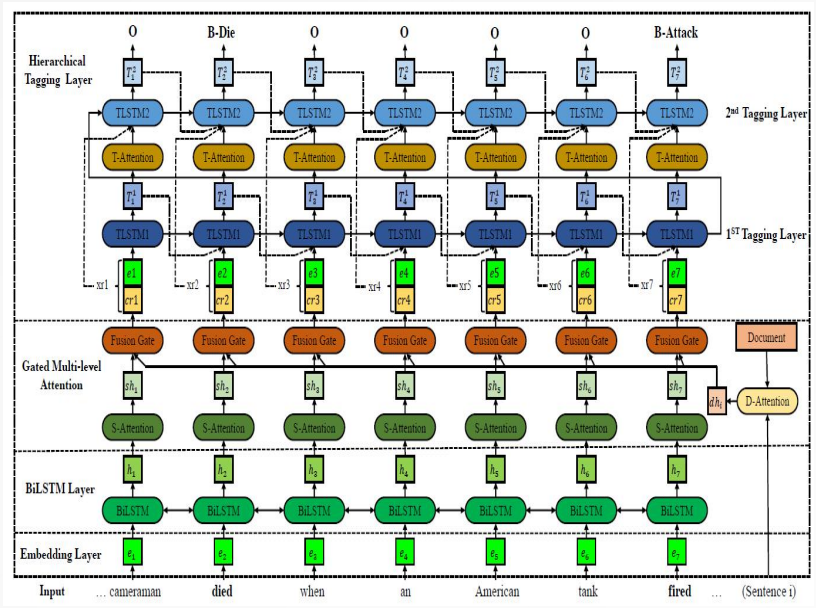
他打算先去超市买点东西再回家，因为最近总加班，很久没有这么早下班了。

End-position

我们为他开了个一个离职趴。

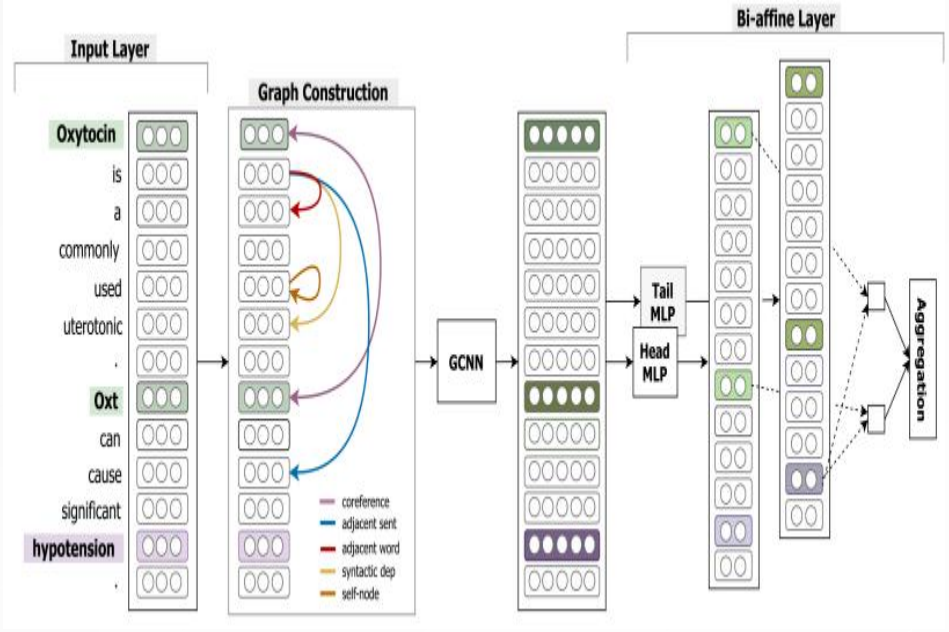
研究趋势

- **特征**：词→句子→篇章
- **解决的问题**：仅利用词或者句子信息，忽略篇章信息



事件抽取

Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention Mechanisms (EMNLP 2018)



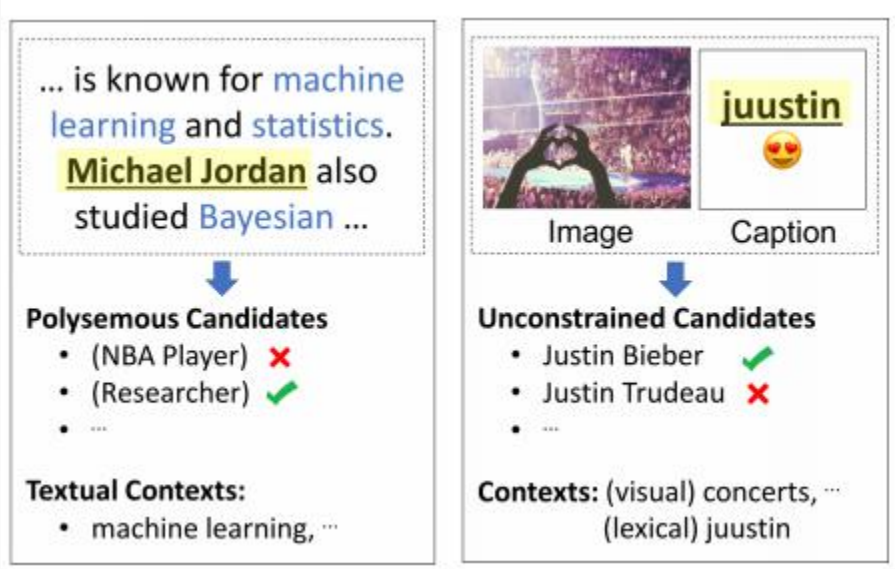
关系抽取

Inter-sentence Relation Extraction with Document-level Graph Convolutional Neural Network (ACL 2019)

研究趋势

- **特征：**文本特征→多模态特征
- **解决的问题：**仅利用文本特征，忽略其它模态特征

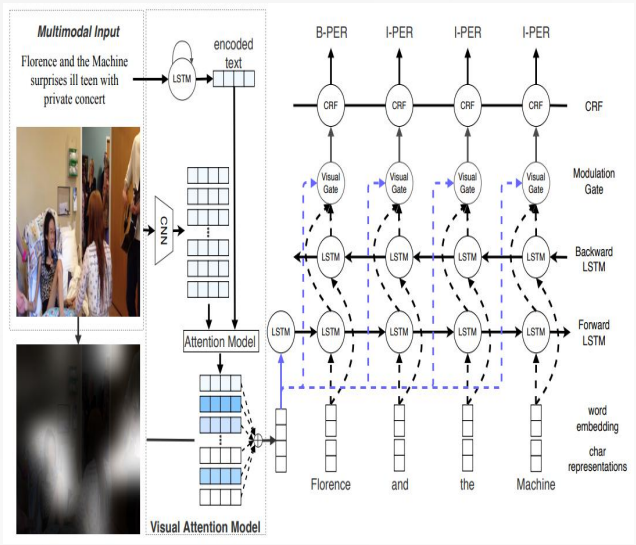
乐队 or 棒球队？



歌手 or 加拿大总理？

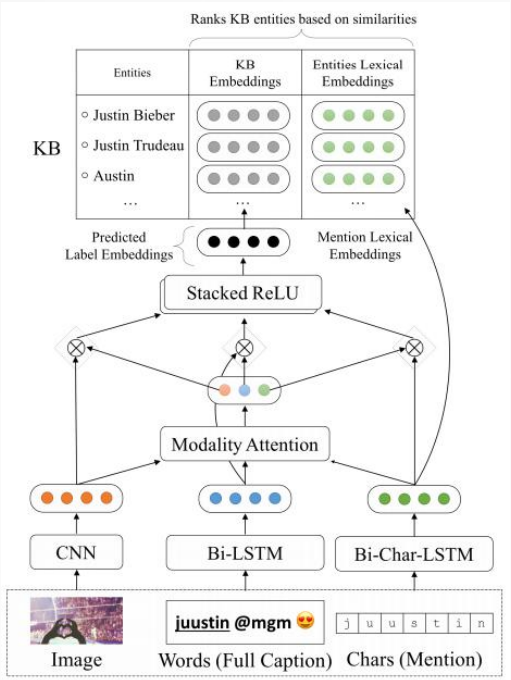
研究趋势

- **特征**：文本特征→多模态特征
- **解决的问题**：仅利用文本特征，忽略其它模态特征



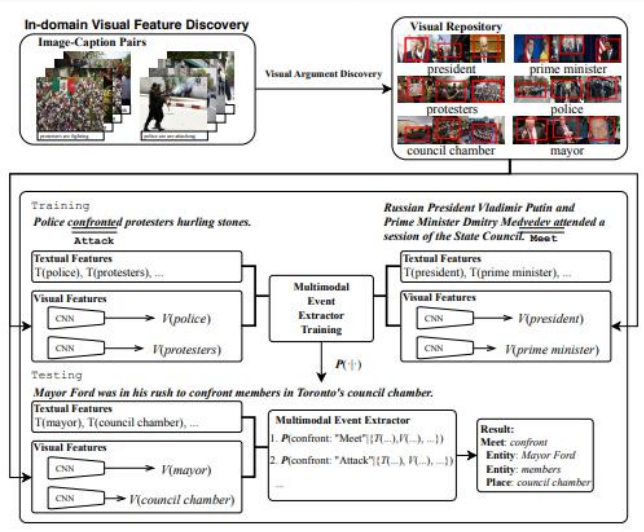
实体识别

Visual Attention Model for Name Tagging in Multimodal Social Media (ACL 2018)



实体消歧

Zeroshot Multimodal Named Entity Disambiguation for Noisy Social Media Posts (ACL 2018)

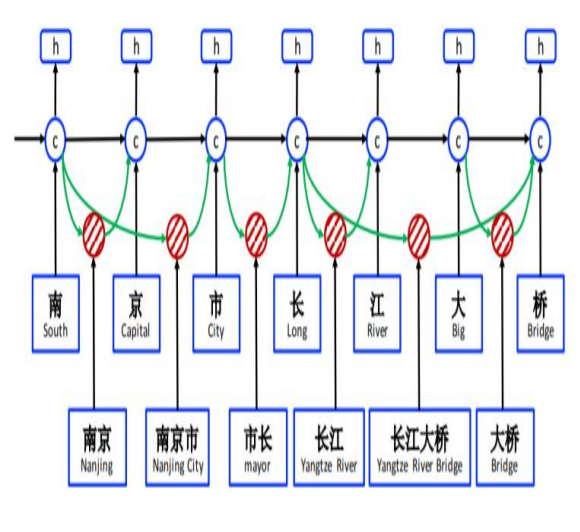


事件抽取

Multi-media Multi-lingual Knowledge Extraction and Inference System （TAC 2019）

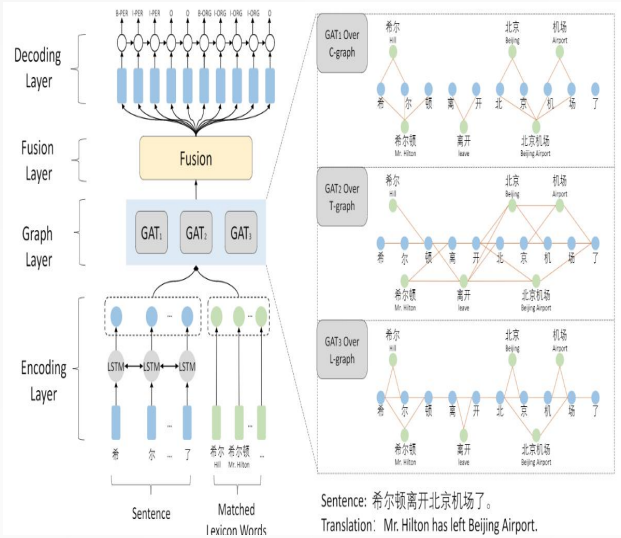
研究趋势

- **特征**：融入知识
- **解决的问题**：仅利用文本，忽略知识



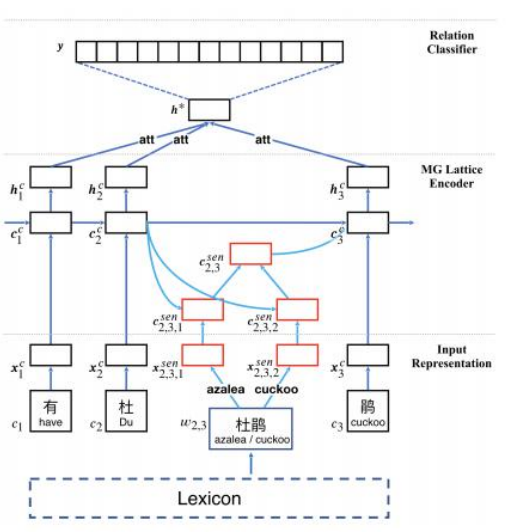
实体识别

Chinese NER Using Lattice LSTM (ACL 2018)



实体识别

Leverage Lexical Knowledge for Chinese Named Entity Recognition via Collaborative Graph Network (EMNLP 2019)

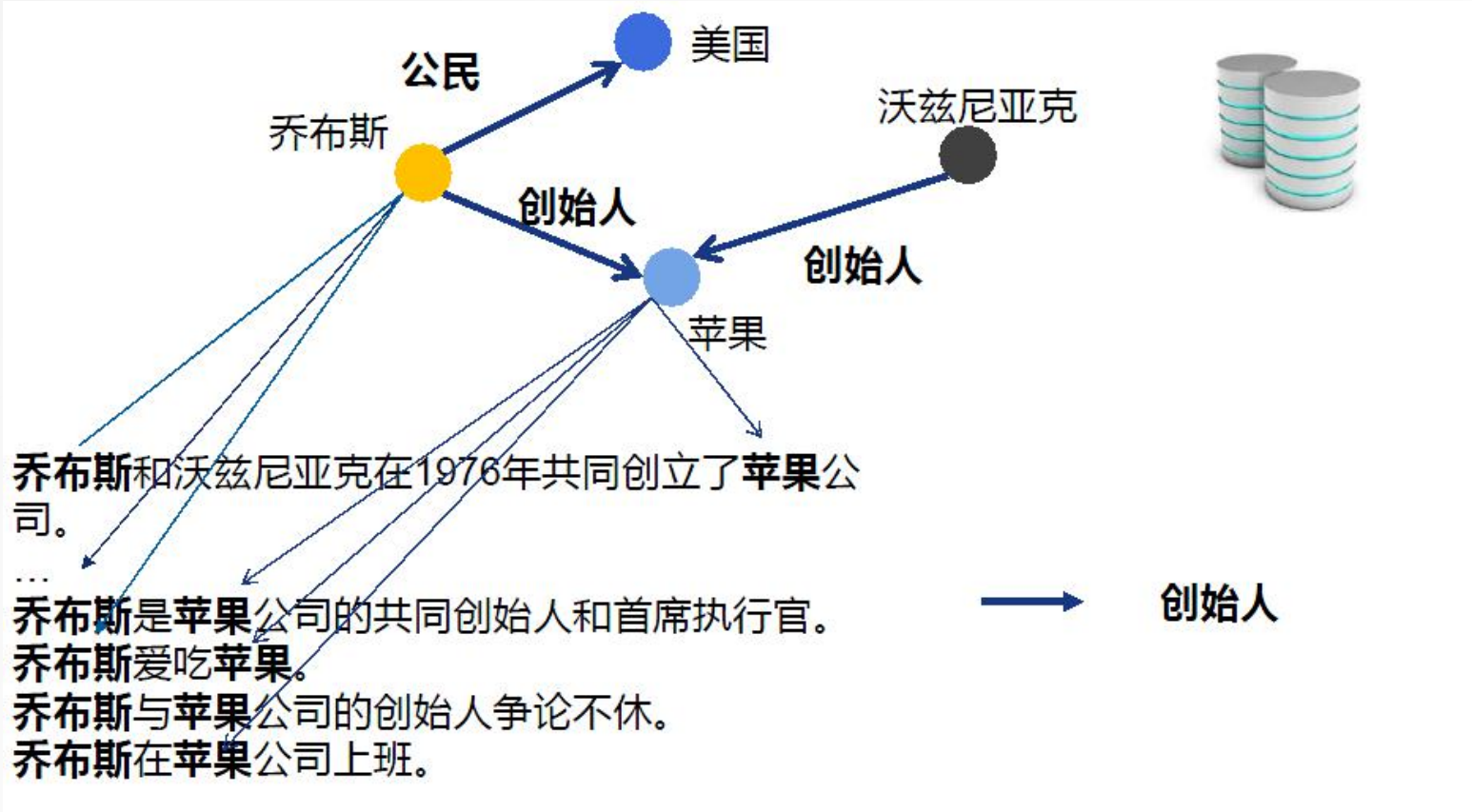


关系抽取

Chinese Relation Extraction with Multi-Grained Information and External Linguistic Knowledge (ACL 2019)

研究趋势

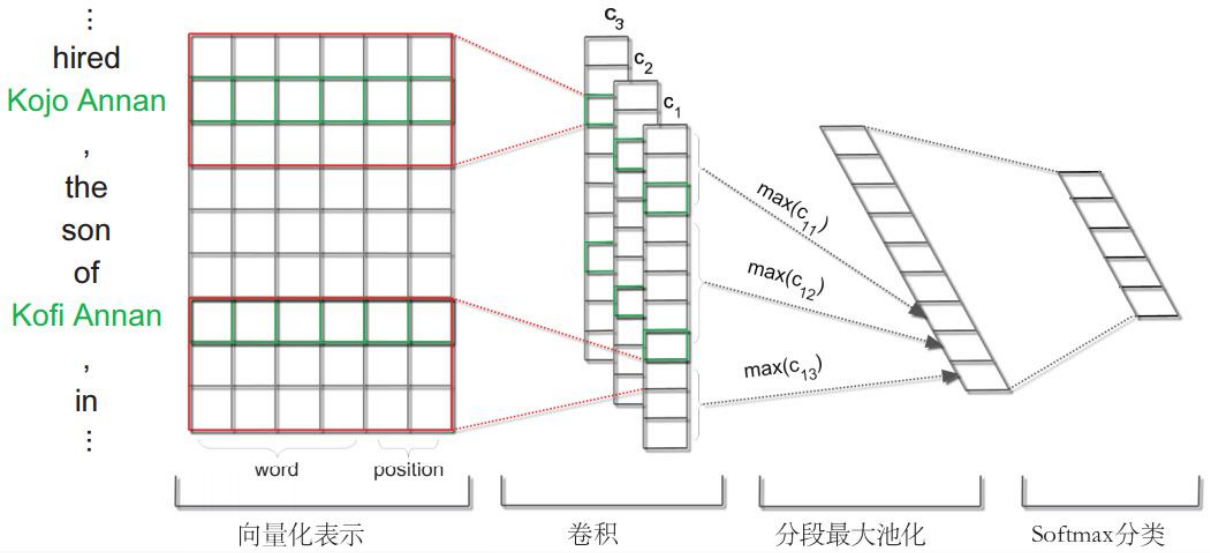
- **语料：**人工标注→ 自动回标



Distant supervision for relation extraction without labeled data（ Mintz et al. ACL 2009）

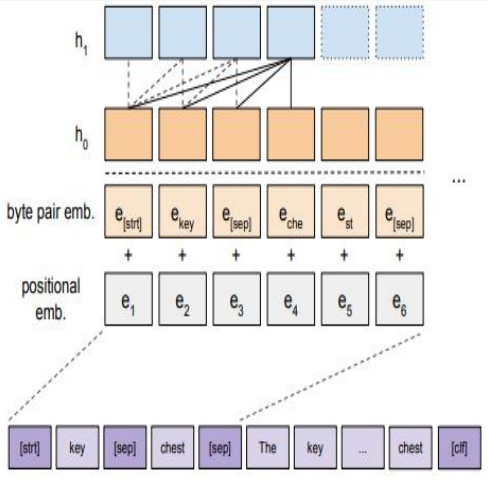
研究趋势

- **语料**：含噪数据下的知识抽取
  - 多示例学习（EMNLP 2015，ACL 2017）
  - 包级别表示（关注机制，强化学习等）（ACL 2016，AAAI 2018）
  - 数据降噪（强化学习，生成对抗网络）（ACL 2018）



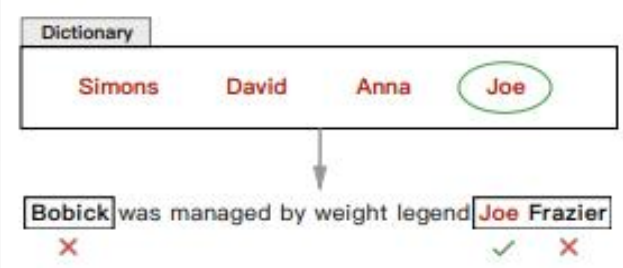
研究趋势

- **语料**：含噪数据下的知识抽取



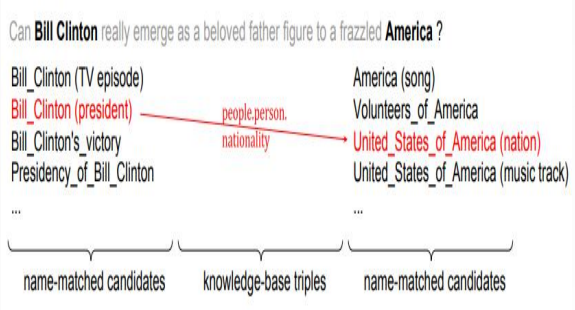
关系抽取

Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction (ACL 2019)



实体识别

Distantly Supervised Named Entity Recognition using Positive-Unlabeled Learning (ACL 2019)



实体消歧

Distant Learning for Entity Linking with Automatic Noise Detection (ACL 2019)

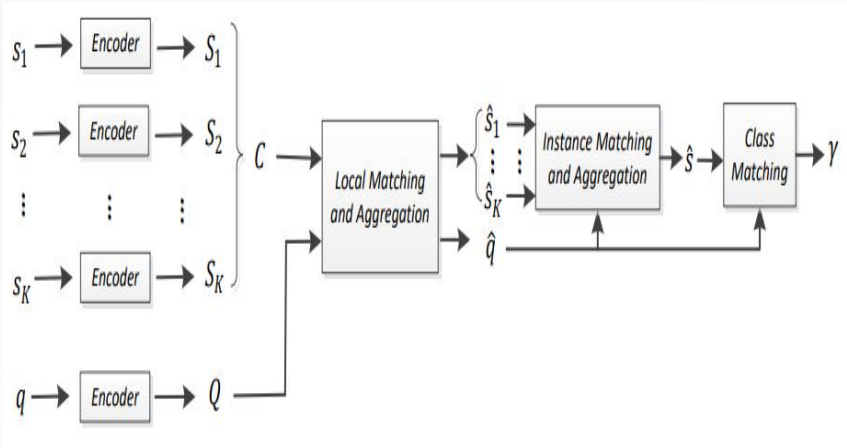
DIAG-NRE: A Neural Pattern Diagnosis Framework for Distantly Supervised Neural Relation Extraction (ACL 2019)
Self-Attention Enhanced CNNs and Collaborative Curriculum Learning for Distantly Supervised Relation Extraction (EMNLP 2019)
Improving Distantly-Supervised Relation Extraction with Joint Label Embedding (EMNLP 2019)
Looking Beyond Label Noise: Shifted Label Distribution Matters in Distantly Supervised Relation Extraction (EMNLP 2019)

研究趋势

- **语料：**小样本学习

Supporting Set	
(A) capital_of	(1) *London* is the capital of *the U.K.* (2) *Washington* is the capital of *the U.S.A.*
(B) member_of	(1) *Newton* served as the president of *the Royal Society*. (2) *Leibniz* was a member of *the Prussian Academy of Sciences*.
(C) birth_name	(1) *Samuel Langhorne Clemens*, better known by his pen name *Mark Twain*, was an American writer. (2) *Alexei Maximovich Peshkov*, primarily known as *Maxim Gorky*, was a Russian and Soviet writer.
Test Instance	
(A) or (B) or (C)	*Euler* was elected a foreign member of *the Royal Swedish Academy of Sciences*.

FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation (EMNLP 2018)



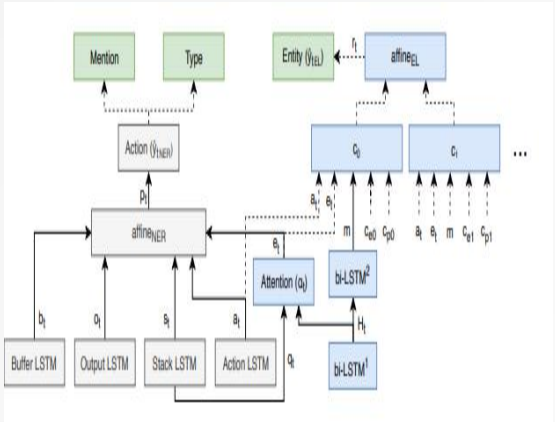
Multi-Level Matching and Aggregation Network for Few-Shot Relation Classification （ACL 2019）

FewRel 2.0: Towards More Challenging Few-Shot Relation Classification（EMNLP 2019）

Adapting Meta Knowledge Graph Information for Multi-Hop Reasoning over Few-Shot Relations （EMNLP 2019）

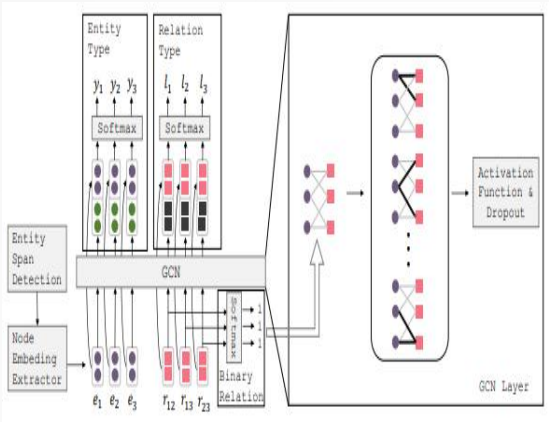
研究趋势

- **任务**：多任务联合学习
- **解决的问题**：各任务独立学习，忽略依存关系



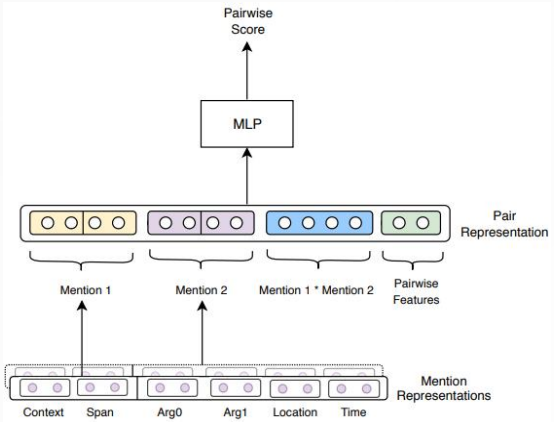
实体识别+实体链接

Joint Learning of Named Entity Recognition and Entity Linking (ACL 2019)



实体识别+关系抽取

Joint Type Inference on Entities and Relations via Graph Convolutional Networks （ACL 2019）



实体共指+事件共指

Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution （ACL 2019）

GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction （ACL 2019）

Multi-Task Learning for Chemical Named Entity Recognition with Chemical Compound Paraphrase （EMNLP 2019）

Learning the Extraction Order of Multiple Relational Facts in a Sentence with Reinforcement Learning （EMNLP 2019）

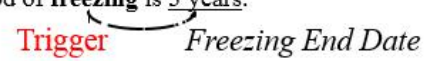
研究趋势

- **任务**：篇章级信息抽取
- **解决的问题**：句子级信信息描述不全

证券代码：600747股票简称：大连控股编号：临2017-04大连大福控股股份有限公司关于大股东股份冻结的公告本公司董事会及全体董事保证本公告内容不存在任何虚假记载、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。
公司于近日收到通知，公司第一大股东长富瑞华持有的上市公司520,000股被大连市人民法院于2017年5月5日冻结。冻结期限为3年。自转为正式冻结之日起计算。
本次轮候冻结包括孳息（包括派发的送股、转增股及现金红利），其效力从登记在先的冻结证券解除冻结且本次轮候冻结部分或全部生效之日起产生。
此次轮候冻结股数占公司总股本的35.51%,截止本公告日，长富瑞华持有本公司股份520,000,000股，占公司总股本35.51%；此次股份冻结后累计股份冻结的数量520,000,000股，占公司总股本的35.51%，经公司向大股东长富瑞华了解，此次股份冻结事项不会对公司的控制权造成影响，也不影响公司正常经营,长富瑞华将与相关方积极协商妥善处理解决相关事宜,公司将密切关注该事项的进展并及时履行信息披露义务,特此公告。
大连大福控股股份有限公司董事会二〇一七年一月十三日2

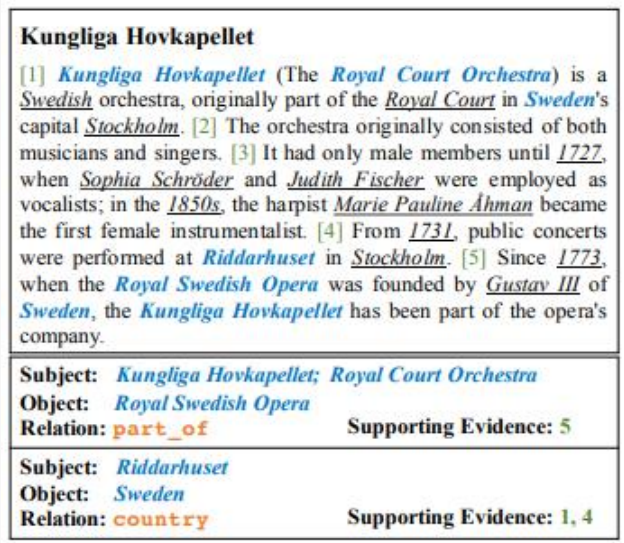
C1:长富瑞华持有的上市公司520,000股被大连市人民法院于2017年5月5日**冻结**。
C2:**冻结**期限为3年。

E1:The 520,000 shares held by Nagafu Ruihua were **frozen** by the people's Court of Dalian city in May 5, 2017.
E2:The period of **freezing** is 3 years.



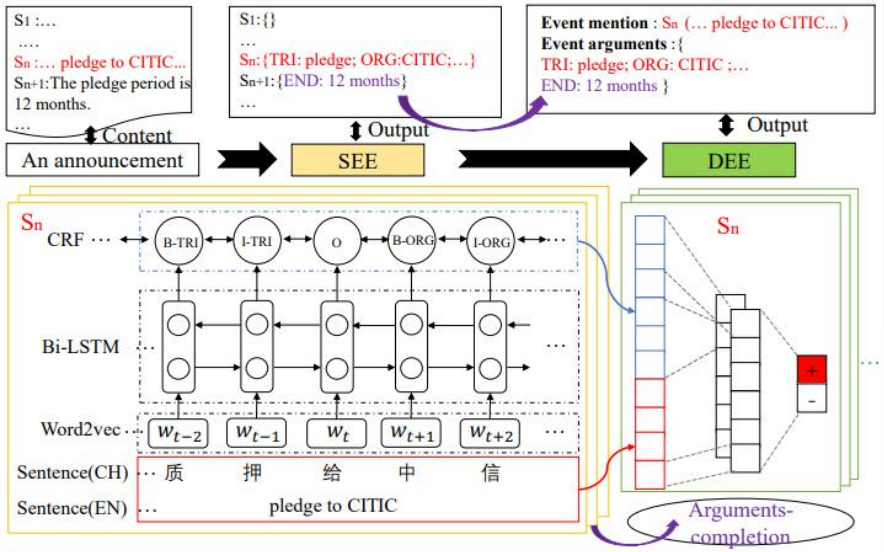
研究趋势

- **任务**：篇章级信息抽取
- **解决的问题**：句子级信信息描述不全



篇章级关系抽取

DocRED: A Large-Scale Document-Level Relation Extraction Dataset （ACL 2019）



篇章级事件抽取

DCFEE: A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data （ACL 2018）

Document-Level N-ary Relation Extraction with Multiscale Representation Learning （EMNLP 2019）
Connecting the Dots: Document-level Neural Relation Extraction with Edge-oriented Graphs （EMNLP 2019）
Doc2EDAG: An End-to-End Document-level Framework for Chinese Financial Event Extraction （EMNLP 2019）

总结

- **特征多元化**
  - 词→句子→篇章（特别是超长文本）
  - 多模态特征
  - 融入知识
- **语料构建（半）自动化**
  - 含噪样本：噪音过滤、冗余信息利用
  - 小样本：迁移学习、元学习
- **任务联合学习**
  - 多任务联合学习
  - 篇章级抽取

CASIA

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

谢谢！Q&A！