CCKS 2020 南昌

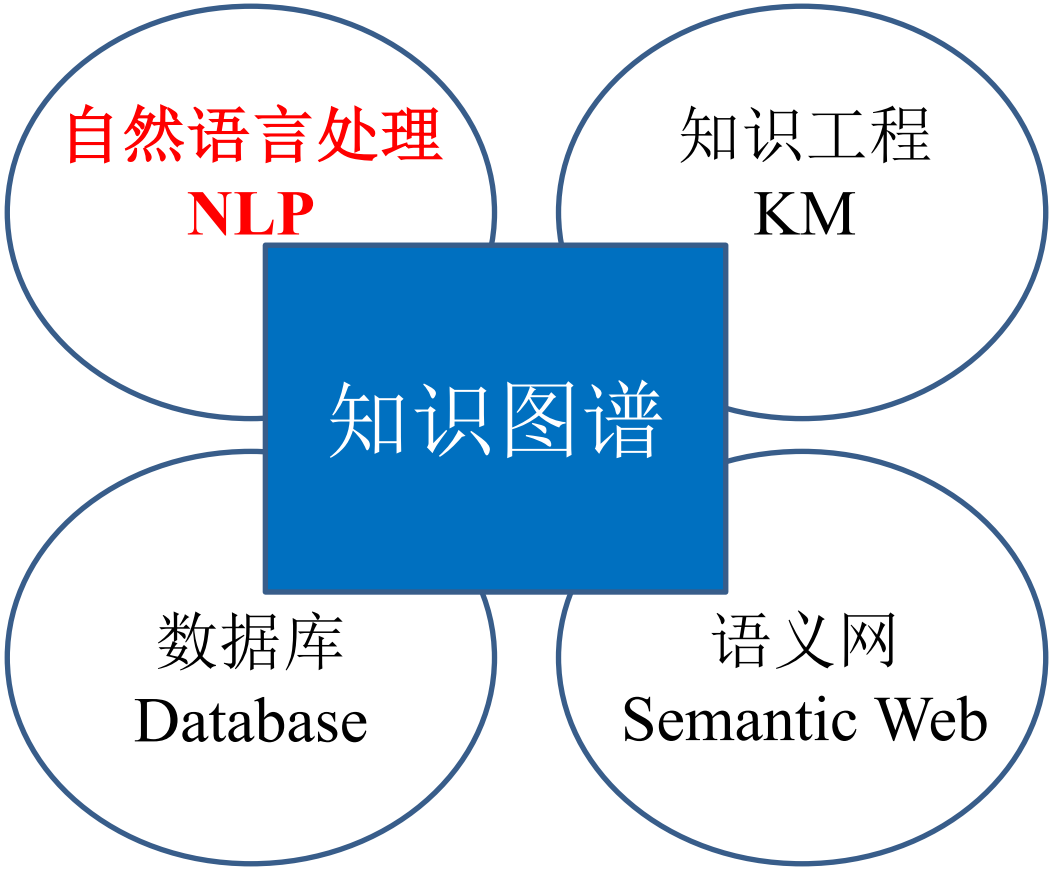
知识图谱研究进展

- 自然语言处理视角

毛先领

北京理工大学计算机学院

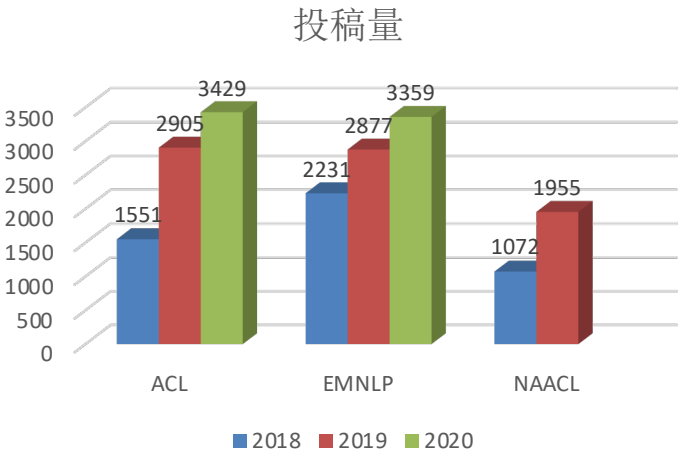
知识图谱渊源

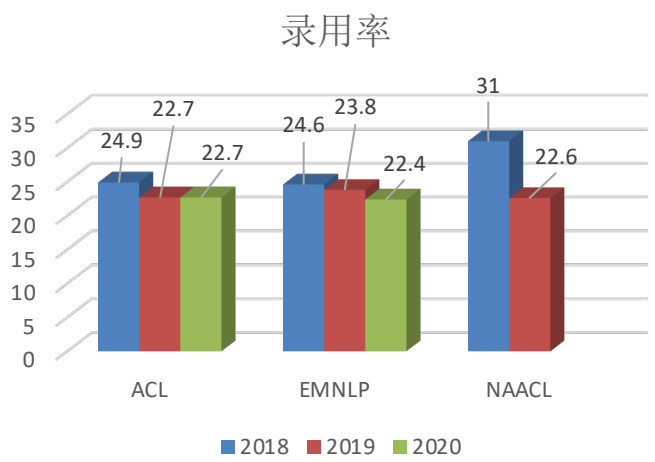


NLP 三大会议投稿量/录用率趋势统计

❊投稿量继续猛增

❊录用率略降





		提交量	接收量	录用率
ACL	All	3429	779	22.7
	Long	2244	571	25.4
	Short	1185	208	17.6
EMNLP	All	3359	754	22.4

*ACL2020由ACL在北美举办，故没有NAACL2020数据

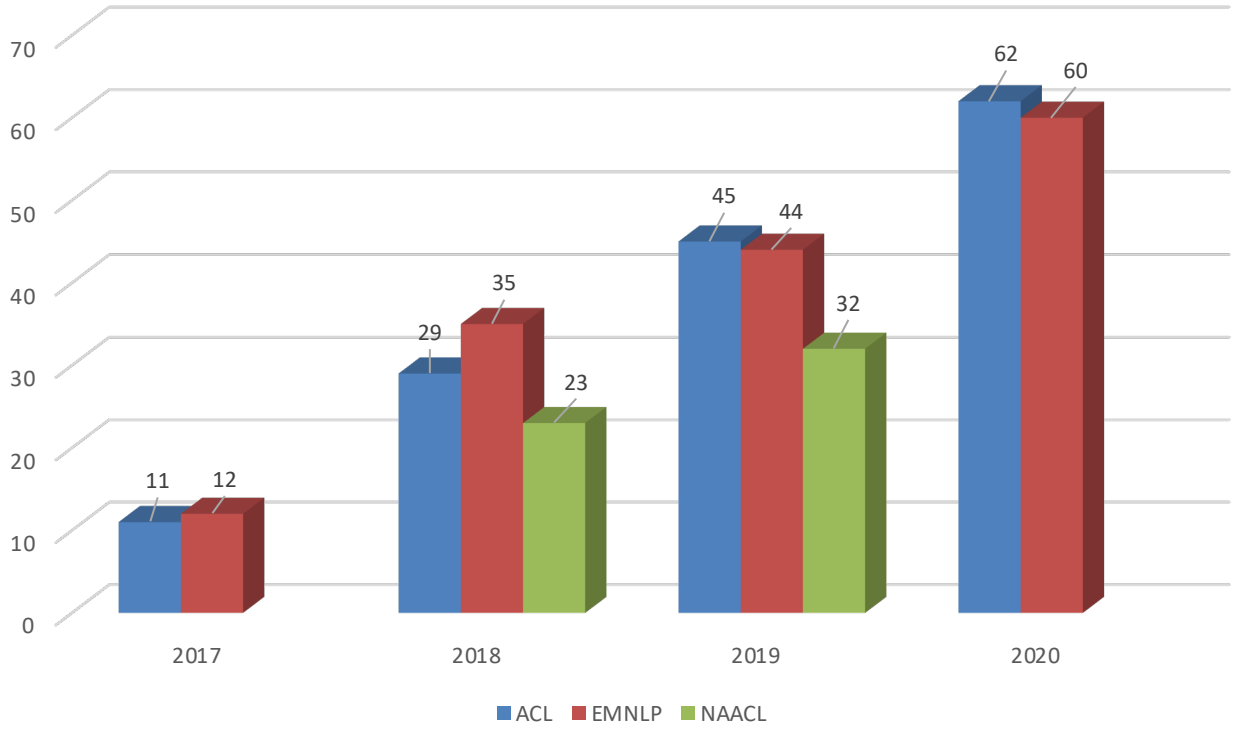
NLP 2020 三大会议领域统计

- 以ACL 2020为例
- 热门领域
  - Information Extraction & Text Mining（313）
  - Machine Learning for NLP（296）
  - Dialogue and Interactive System（250）
  - Machine Translation （245）
- 领域录用率
  - Information Extraction & Text Mining（23.0%）
  - Machine Learning for NLP（22.6%）
  - Dialogue and Interactive System（24.8%）
  - Machine Translation （27.8%）

知识图谱论文统计

❊知识图谱论文粗略统计

❄检索关键词“knowledge”



*ACL2020由ACL在北美举办，故没有NAACL2020数据

知识图谱论文统计

- KG 相关论文（人工看摘要）
  - ACL 2019：38
  - ACL 2020：51
- NLP研究中几大KG任务

	ACL2019	ACL2020
其他(KG应用)	12	3
关系(关系分类、关系抽取)	8	13
实体(实体识别、实体链接)	7	25
问答(知识库问答)	6	1
表示(KG嵌入表示、图表示)	5	9

NLP中KG“构建与应用”的几大趋势

- 知识图谱构建
  - 实体相关研究
    - 多模态、低资源
  - 关系抽取相关研究
    - 联合抽取、开放抽取、文档级抽取、低资源
  - 模型的可解释性
- 知识图谱应用
  - 融合知识的预训练语言模型
  - 知识与推理
  - 融合知识的NLP任务

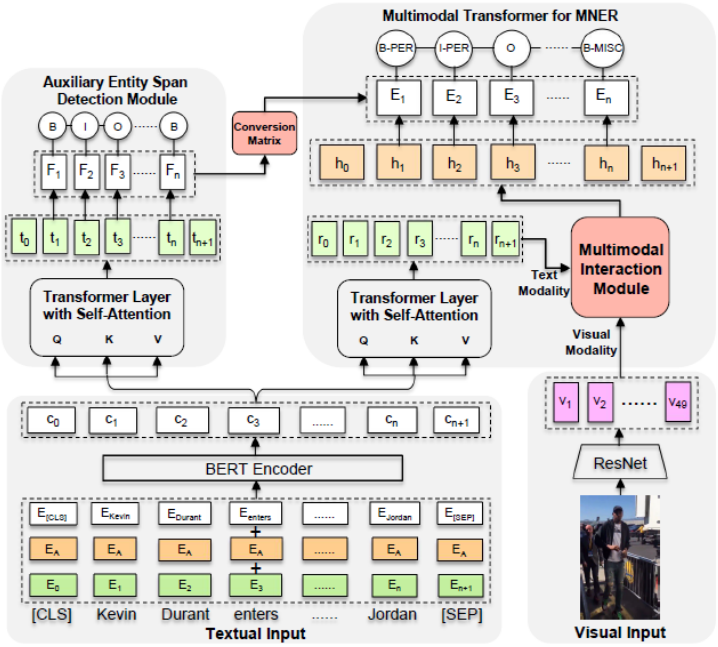
实体相关研究

- 多模态NER
  - Code and Named Entity Recognition in StackOverflow (ACL2020)
  - Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer (ACL2020)

实体相关研究

- 多模态NER
  - 社交媒体中的帖子经常含有图片
  - 视觉语境带来了丰富的信息，也带来了视觉语境偏差
  - Jianfei Yu et al. 提出了统一的多模态Transformer框架

(a). **[Kevin Durant PER]** enters **[Oracle Arena LOC]** wearing off — White x **[Jordan MISC]**



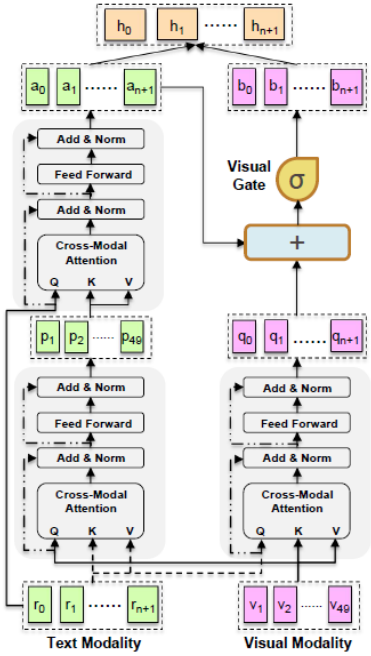
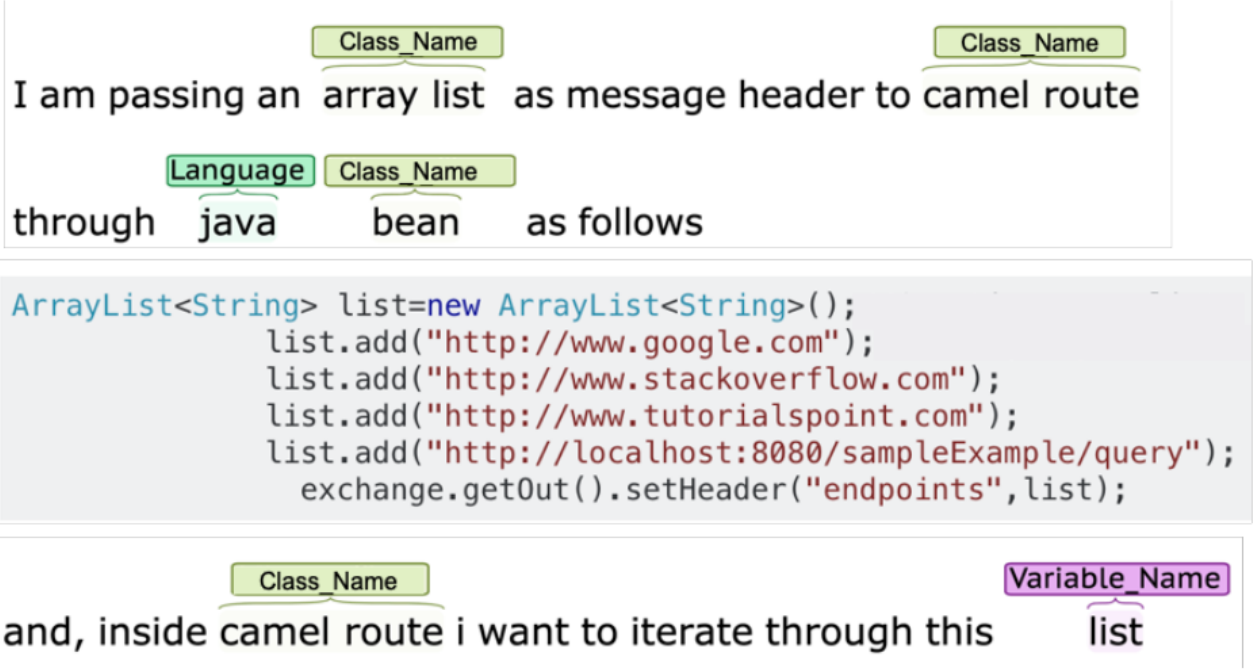


Figure 2: (a). Overall Architecture of Our Unified Multimodal Transformer. (b). Multimodal Interaction (MMI) Module.

Jianfei Yu et al. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer (ACL2020)

实体相关研究

- 多模态NER
  - Jeniya Tabassum et al. 利用代码中含有的丰富信息来加强StackOverflow 上帖子文本的命名实体识别



Jeniya Tabassum et al. Code and Named Entity Recognition in StackOverflow (ACL2020)

实体相关研究

❋低资源NER（1）

- From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains (ACL2020)
- Improving Low-Resource Named Entity Recognition using Joint Sentence and Token Labeling (ACL2020)
- Named Entity Recognition without Labelled Data: A Weak Supervision Approach (ACL2020)
- Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language (ACL2020)
- Design Challenges in Low-resource Cross-lingual Entity Linking (EMNLP2020)
- Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference (ACL2020)
- Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language (ACL2020)
- Multi-Cell Compositional LSTM for NER Domain Adaptation (ACL2020)

实体相关研究

❋低资源NER（2）

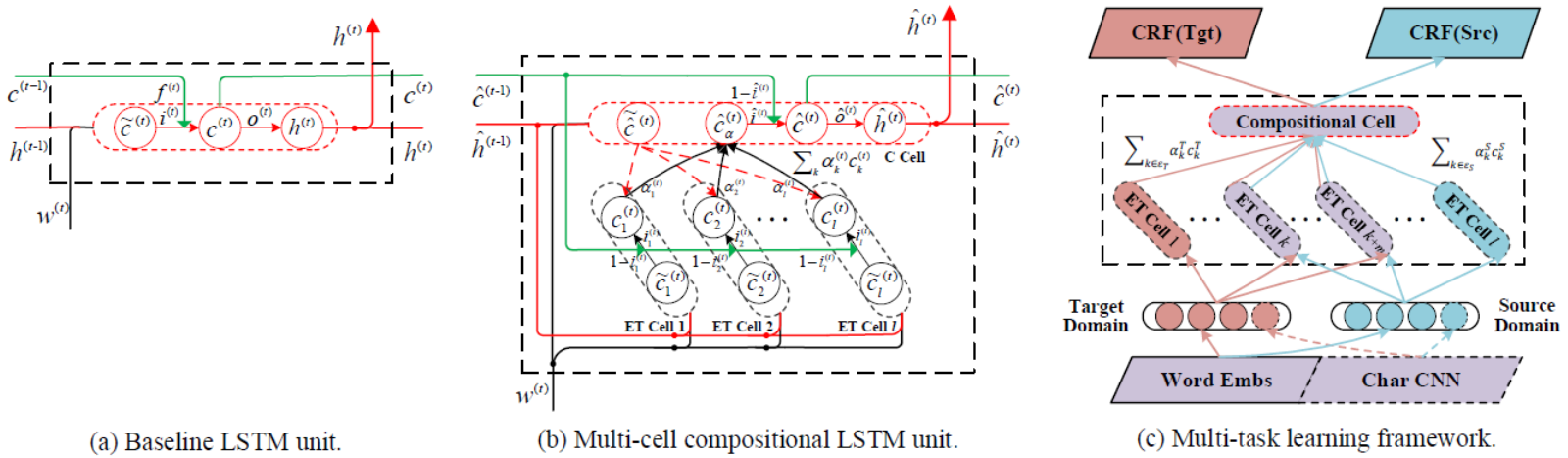
- Soft Gazetteers for Low-Resource Named Entity Recognition (ACL2020)
- Multi-Cell Compositional LSTM for NER Domain Adaptation (ACL2020)
- TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition (ACL2020)
- Empower Entity Set Expansion via Language Model Probing (ACL2020)

实体相关研究

❊低资源NER

❊跨领域

- ❁不同领域信息的共通性，实现知识迁移
- ❁基于不同领域的实体类型相对稳定(例如人名地名)的现象， Chen Jia et al. 提出了multi-cell compositional LSTM的结构



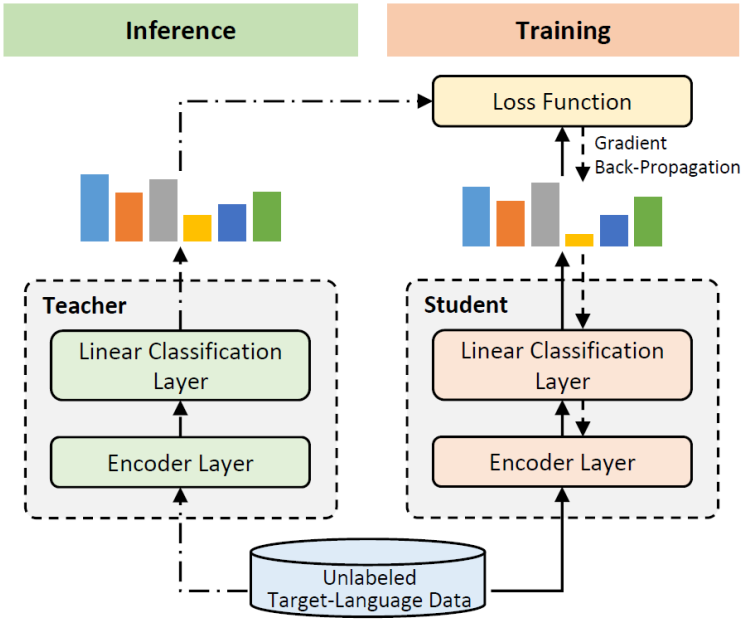
(a) Baseline LSTM unit. (b) Multi-cell compositional LSTM unit. (c) Multi-task learning framework.

- ❁当应用于跨领域的数据时，现有NER方法普遍面临鲁棒性问题，Wang et al. 提出了一个评估NER模型跨领域鲁棒性的框架

[1] Chen Jia et al. Multi-Cell Compositional LSTM for NER Domain Adaptation (ACL2020)
[2] Jing Wang et al. Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic (ACL2020)

实体相关研究

- 低资源NER
  - 跨语言
    - Teacher-Student Model
    - Unlabeled data in target language



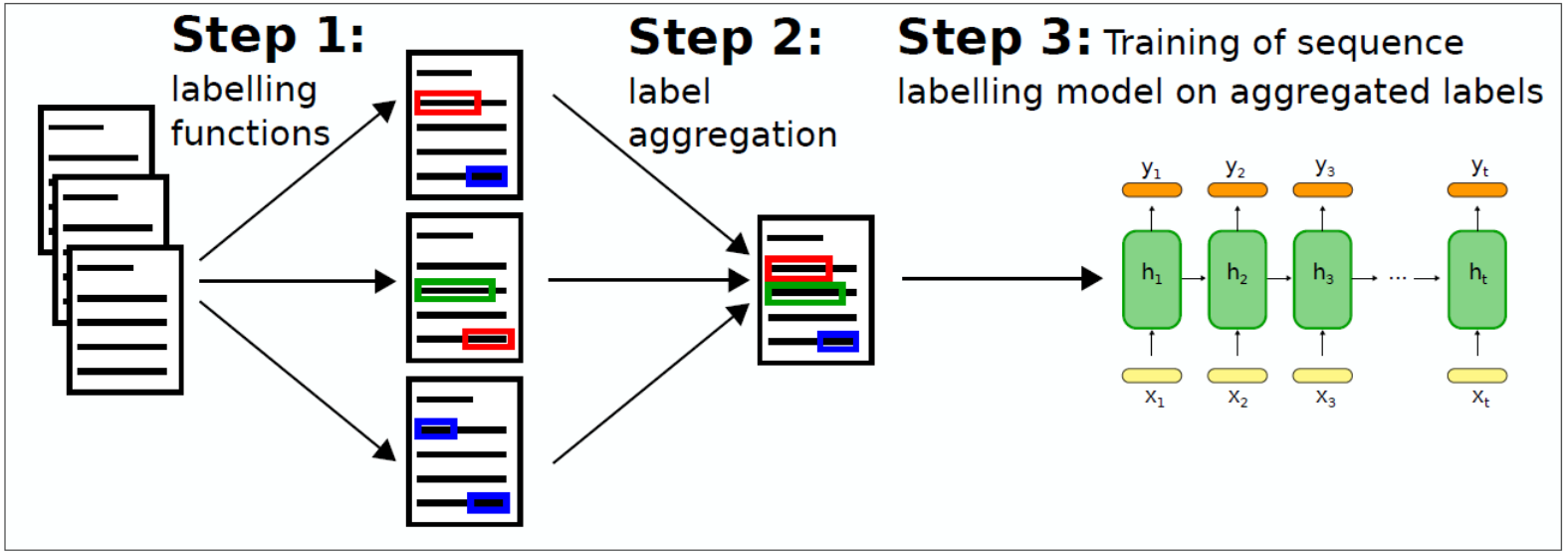
	es	nl	de
Täckström (2012)	61.90	59.90	36.40
Rahimi et al. (2019)	71.80	67.60	59.10
Chen et al. (2019)	73.50	72.40	56.00
Moon et al. (2019)†	76.53	**83.35**	72.44
Ours-avg	77.75	80.70	74.97
Ours-sim	**78.00**	81.33	**75.33**

Table 3: Performance comparisons of **multi-source** cross-lingual NER. **Ours-avg**: averaging teacher models (Eq. 7) . **Ours-sim**: weighting teacher models with learned language similarities (Eq. 11). † denotes the reported results *w.r.t.* freezing the bottom three layers of $\text{BERT}_{\text{BASE}}$.

Qianhui Wu et al. Single-Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language (ACL2020)

实体相关研究

- 低资源NER
  - Weak Supervision
    - No labelled Data
    - Labelling functions
    - Label aggregation



Pierre Lison et al. Named Entity Recognition without Labelled Data: A Weak Supervision Approach (ACL2020)

实体相关研究

- 低资源NER
  - 数据集的自动构建
    - Entity Trigger: A group of words in a sentence that helps to explain why humans would recognize an entity in the sentence.

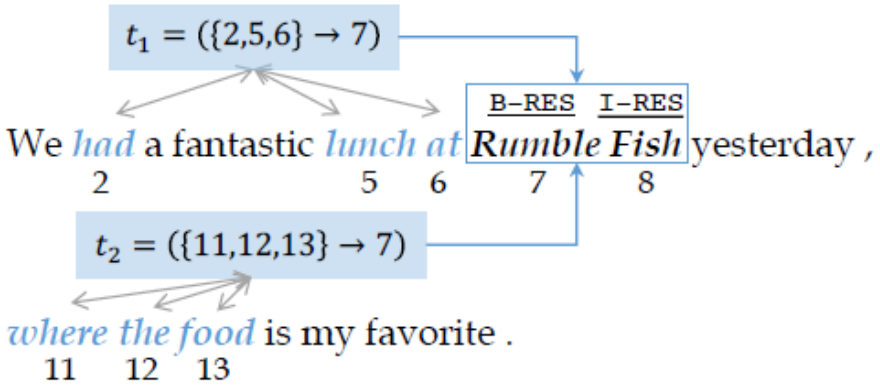


Figure 1: We show two individual **entity triggers**: t_1 ("*had ... lunch at*") and t_2 ("*where the food*"). Both are associated to the same entity mention "*Rumble Fish*" (starting from 7th token) typed as restaurant (RES).

Bill Yuchen Lin et al. TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition; (ACL 2020)

NLP中KG“构建与应用”的几大趋势

- 知识图谱构建
  - 实体相关研究
    - 多模态、低资源
  - **关系抽取相关研究**
    - 联合抽取、开放抽取、文档级抽取、低资源
  - 模型的可解释性
- 知识图谱应用
  - 融合知识的预训练语言模型
  - 知识与推理
  - 融合知识的NLP任务

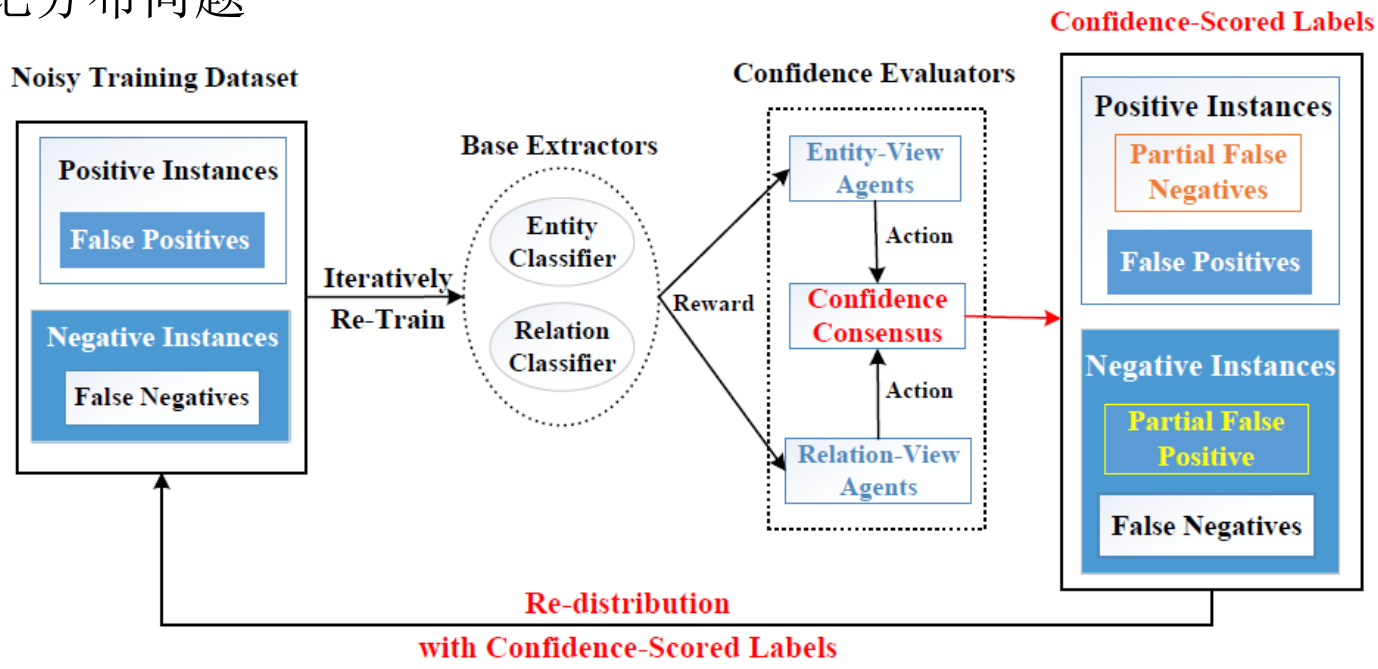
关系抽取相关研究

- 联合抽取、开放抽取、文档级抽取、低资源
  - In Laymans Terms： Semi-Open Relation Extraction from Scientific Texts (ACL2020)
  - Reasoning with Latent Structure Refinement for Document-Level Relation Extraction (ACL2020)
  - Relabel the Noise： Joint Extraction of Entities and Relations via Cooperative Multiagents (ACL2020)
  - Dialogue-Based Relation Extraction (ACL2020)
  - ZeroShotCeres： Zero-Shot Relation Extraction from Semi-Structured Webpages (ACL2020)
  - Relation Extraction with Explanation (ACL2020)
  - Revisiting Unsupervised Relation Extraction (ACL2020)

关系抽取相关研究

❊联合抽取

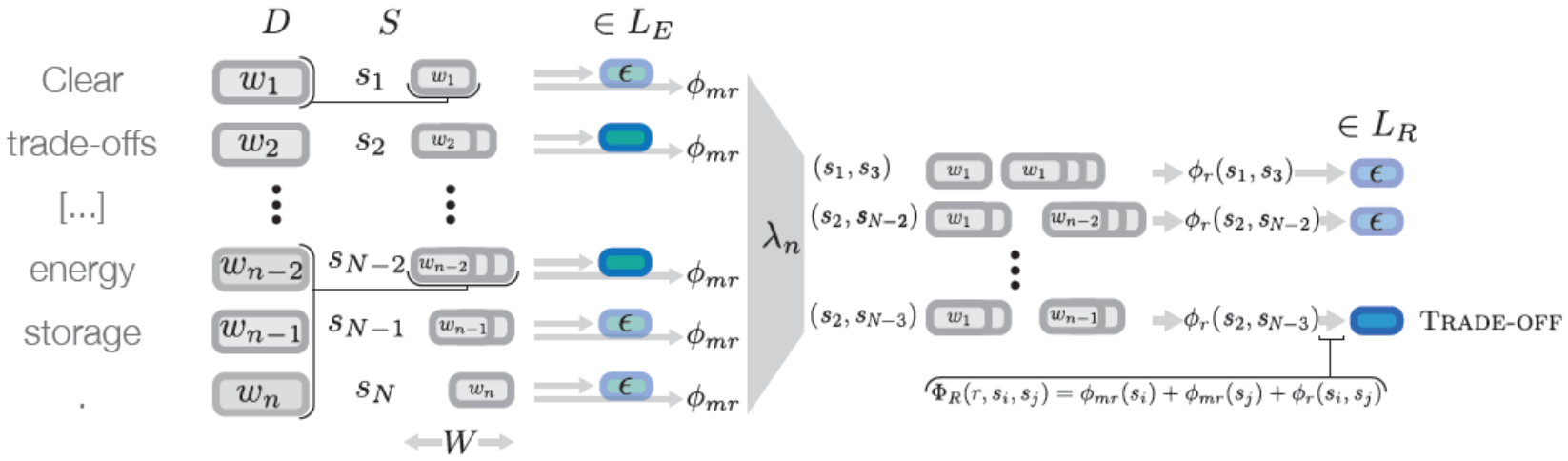
- ❊联合抽取共享NER和RE的信息，信息利用更充分
- ❊Chen et al. 采用联合抽取解决远程监督存在的噪声标记和移位标记分布问题



Daoyuan Chen et al. Relabel the Noise: Joint Extraction of Entities and Relations via Cooperative Multiagents (ACL2020)

关系抽取相关研究

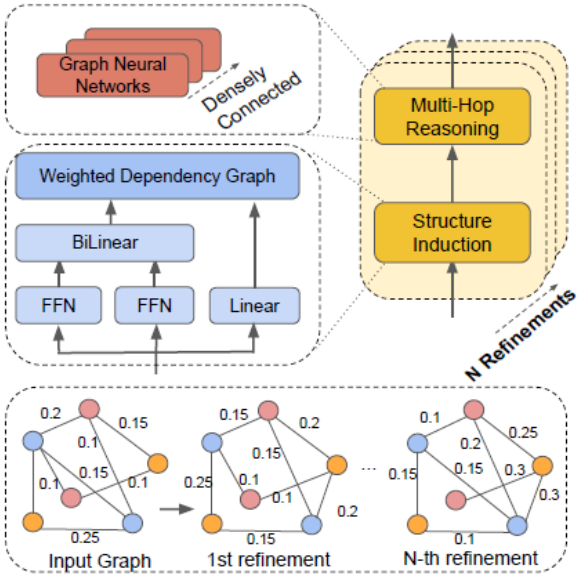
- 开放抽取
  - 关系抽取只能抽取文本的一小部分信息
  - Open IE在处理科学文本时遇到长句子和复杂句子表现不佳
  - Ruben Kruiper et al. 结合两种系统的输出，提出半开放的关系抽取

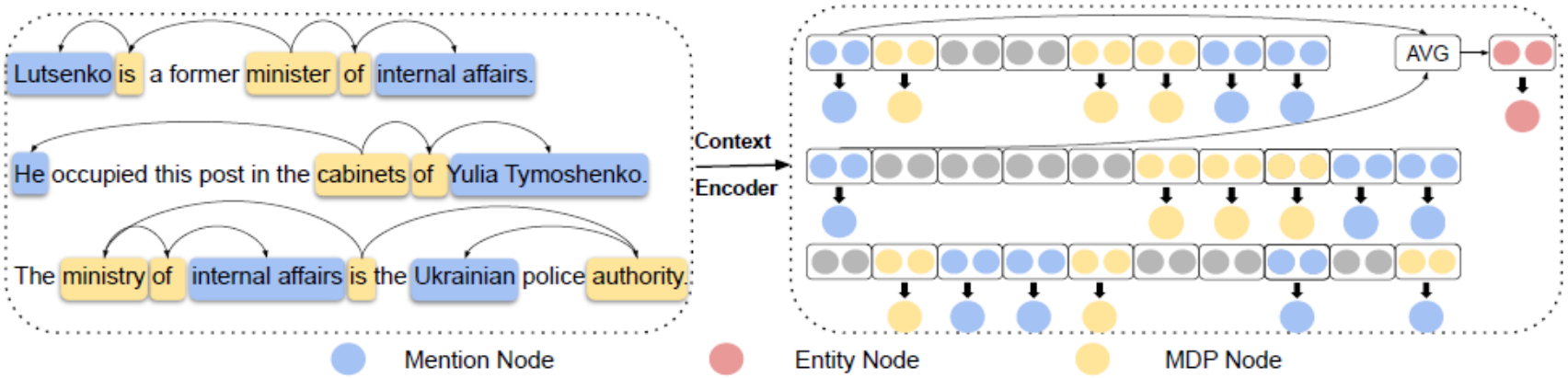


Ruben Kruiper et al. In Laymans Terms：Semi-Open Relation Extraction from Scientific Texts (ACL2020)

关系抽取相关研究

- 文档级抽取
  - 利用整篇文档中的上下文信息加强关系抽取
  - 需要捕获句子间实体的复杂交互
  - Guoshun Nan et al. 提出Latent Structure Refinement (LSR)模型动态地学习文档级结构，并以端到端方式进行预测



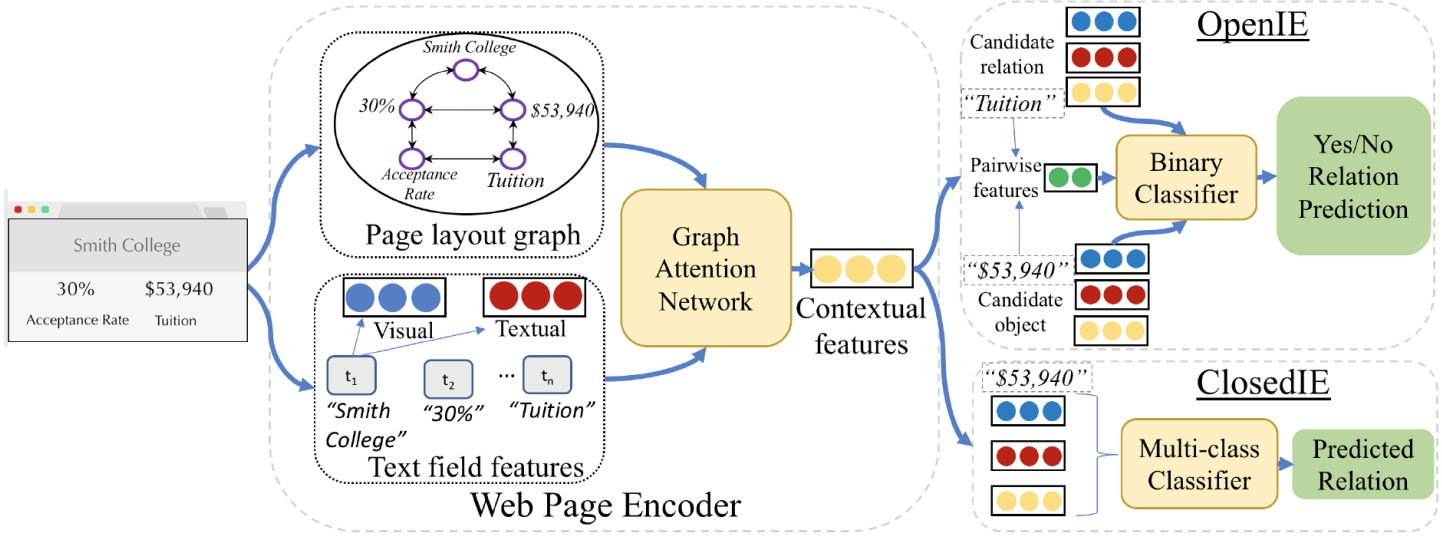


Guoshun Nan et al. Reasoning with Latent Structure Refinement for Document-Level Relation Extraction (ACL2020)

关系抽取相关研究

❊低资源抽取

- 从半结构化文本提取信息的工作需要依赖给定模板，模板需要通过手工标记或远程监督获得
- Colin Lockard et al. 提出了一种“Zero-Shot”开放域关系抽取的解决方案，使用基于图神经网络的方法在网页上构建文本的丰富表示以及它们之间的关系，使其泛化成为新的模板



Colin Lockard et al. ZeroShotCeres: Zero-Shot Relation Extraction from Semi-Structured Webpages (ACL2020)

NLP中KG“构建与应用”的几大趋势

- 知识图谱构建
  - 实体相关研究
    - 多模态、低资源
  - 关系抽取相关研究
    - 联合抽取、开放抽取、文档级抽取、低资源
  - **模型的可解释性**
- 知识图谱应用
  - 融合知识的预训练语言模型
  - 知识与推理
  - 融合知识的NLP任务

模型的可解释性

❊相关工作

- Learning Collaborative Agents with Rule Guidance for Knowledge Graph Reasoning (EMNLP 2020)
- Relation Extraction with Explanation (ACL2020)
- TACRED Revisited： A Thorough Evaluation of the TACRED Relation Extraction Task (ACL2020)
- Rationalizing Medical Relation Prediction from Corpus-level Statistics (ACL2020)
- Instance-Based Learning of Span Representations： A Case Study through Named Entity Recognition (ACL2020)

模型的可解释性

❊代表性工作

- ❊ 前人的工作主要集中在提高性能上，对可解释性的研究很少
- ❊ Hamed Shahbazi et al. 提供了一个标注的测试集，以评估远程监督下的关系抽取模型的可解释性好坏，并做了相关的实验分析
- ❊ Wang et al.提出了一种新的可解释的医学关系预测框架，以期为医学关系预测的合理化提供参考

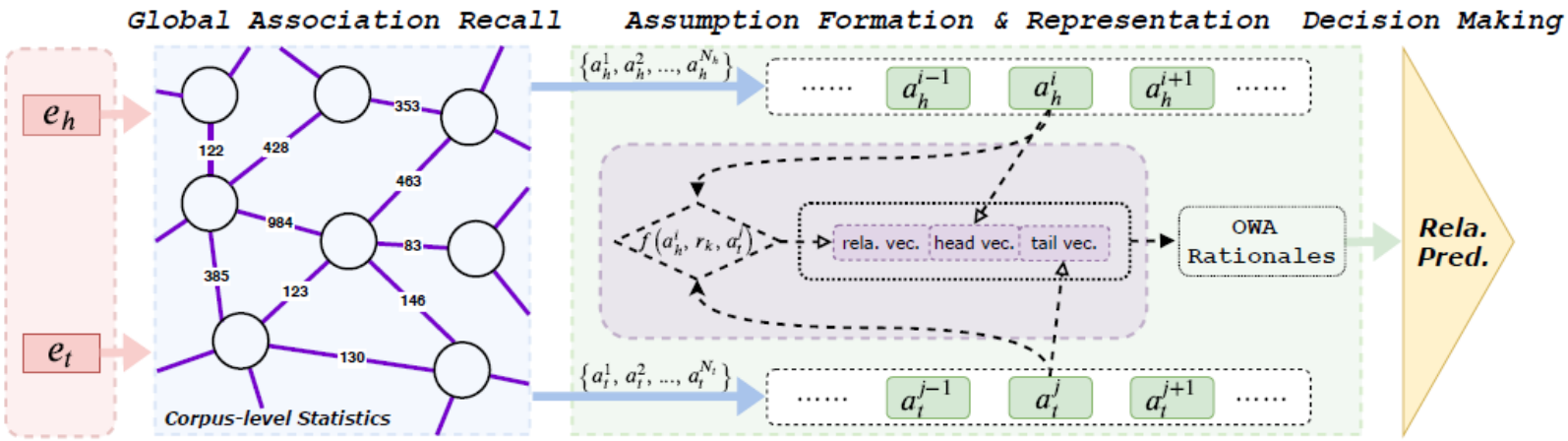


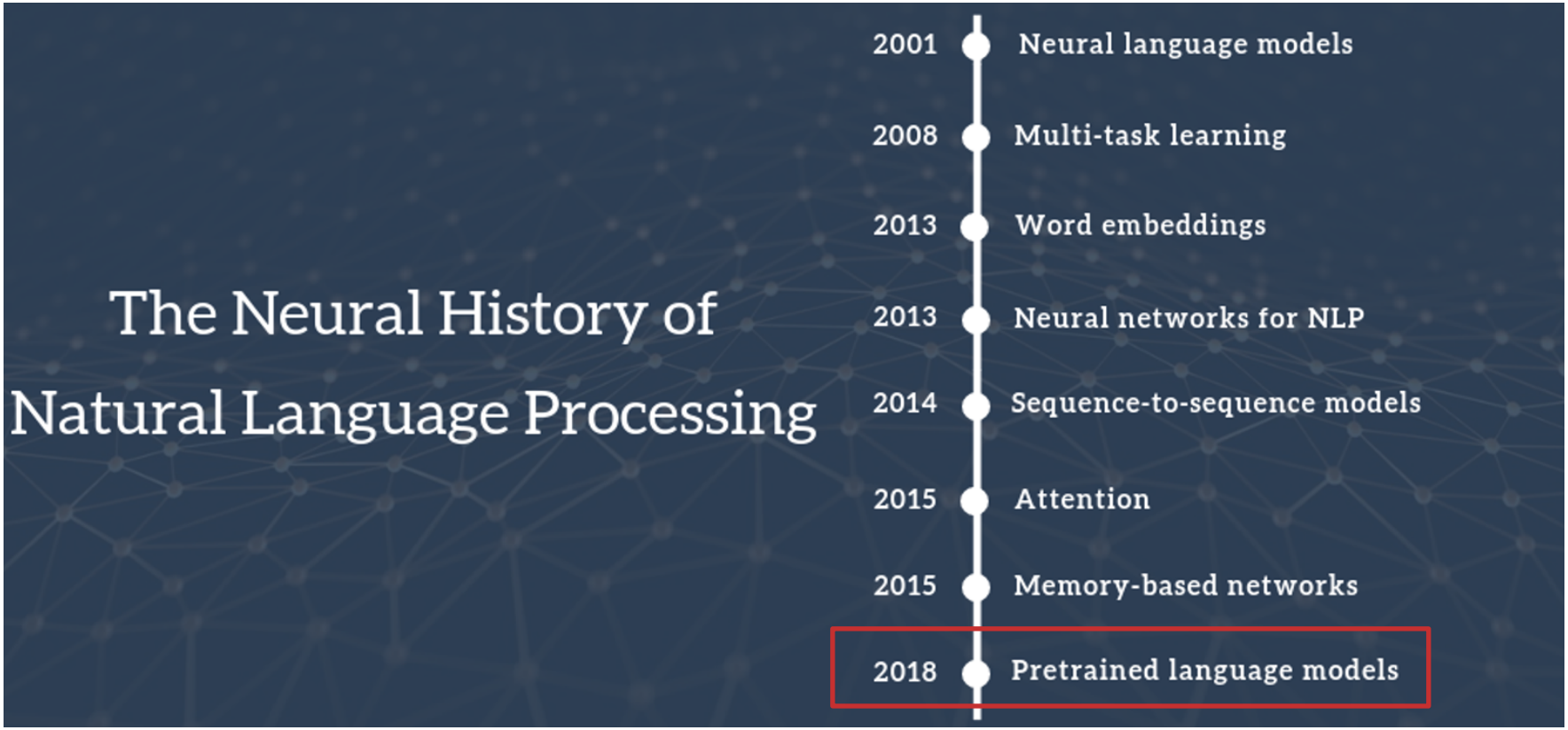
Figure 3: Framework Overview.

[1] Hamed Shahbazi et al. Relation Extraction with Explanation (ACL2020)
[2] Zhen Wang et al. Rationalizing Medical Relation Prediction from Corpus-level Statistics (ACL2020)

NLP中KG“构建与应用”的几大趋势

- 知识图谱构建
  - 实体相关研究
    - 多模态、低资源
  - 关系抽取相关研究
    - 联合抽取、开放抽取、文档级抽取、低资源
  - 模型的可解释性
- 知识图谱应用
  - **融合知识的预训练语言模型**
  - 知识与推理
  - 融合知识的NLP任务

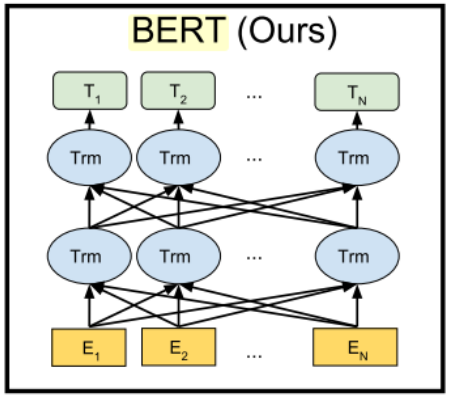
融合知识的预训练语言模型

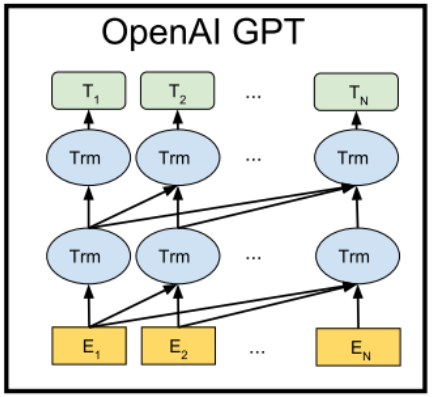


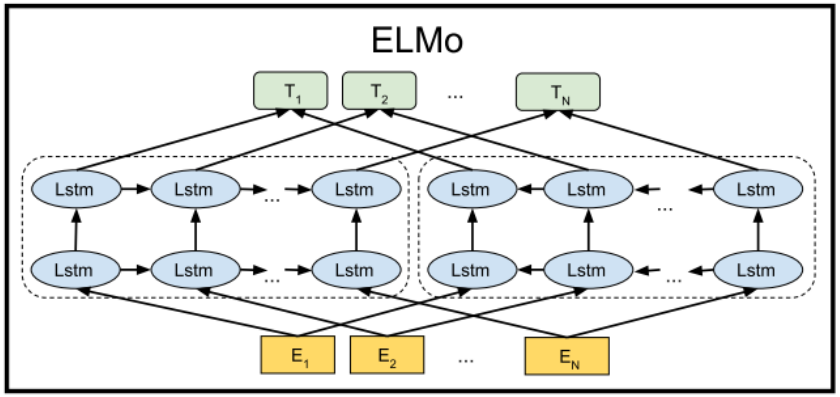
http://ruder.io/a-review-of-the-recent-history-of-nlp/

融合知识的预训练语言模型

- ELMo: 双向LSTM
- GPT: 单向Transformer
- BERT: 双向Transformer (NAACL2019 Best paper)
- XLNet，GPT3





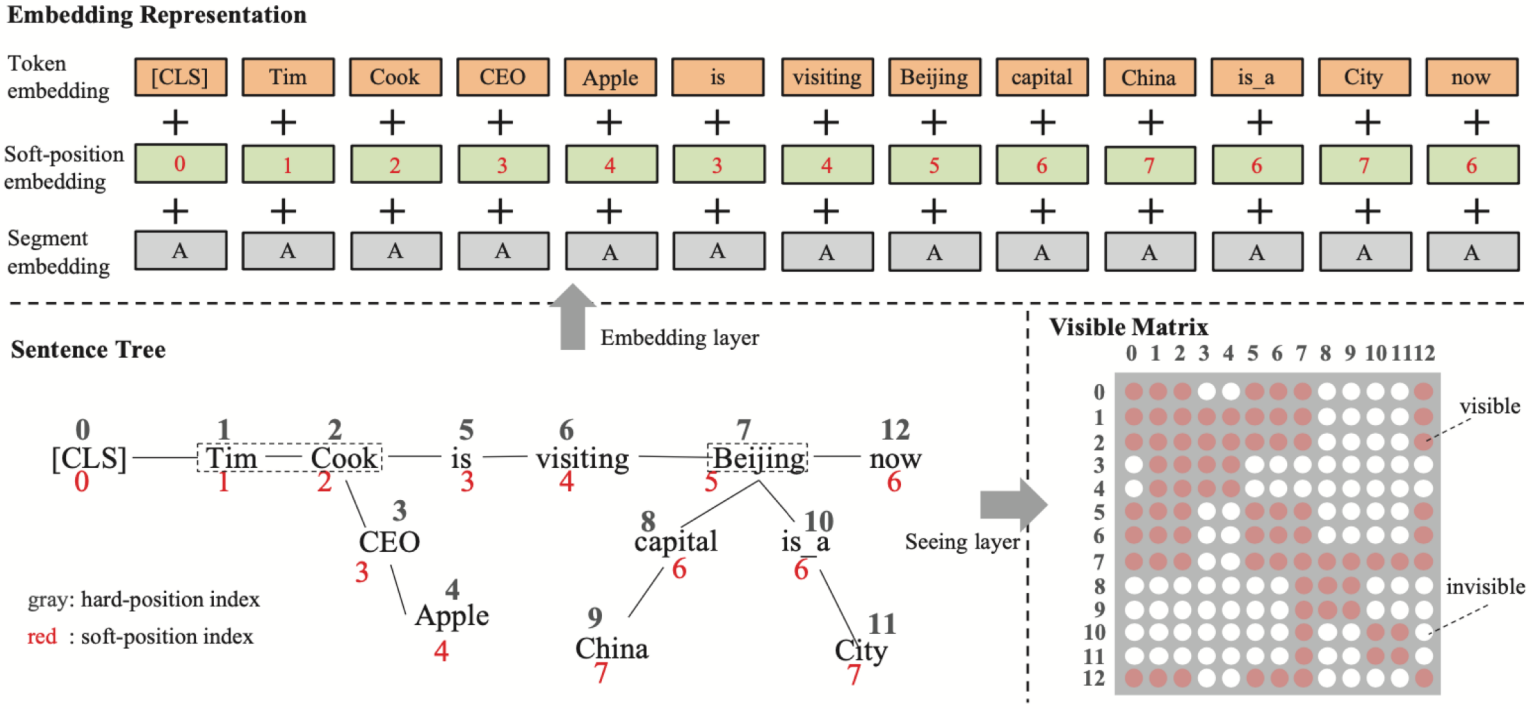


[1] Peters et al. Deep contextualized word representations (NAACL2019).
[2] Alec Radford et al. Improving language understanding with unsupervised learning.
[3] Jacob Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (NAACL2019).
[4] Tom B. Brown et al. Language Models are Few-Shot Learners. (2020)

融合知识的预训练语言模型

❊知识增强的预训练语言模型

- ❊ K-BERT：知识支持的语言表示模型，可以合并特殊领域知识
- ❊ 通过**结合知识图谱**，K-BERT不仅在特定领域上明显胜过BERT，而且在开放域的任务上也胜过BERT



Wejie Liu et al. K-BERT: Enabling Language Representation with Knowledge Graph.(AAAI2020)

NLP中KG“构建与应用”的几大趋势

- 知识图谱构建
  - 实体相关研究
    - 多模态、低资源
  - 关系抽取相关研究
    - 联合抽取、开放抽取、文档级抽取、低资源
  - 模型的可解释性
- 知识图谱应用
  - 融合知识的预训练语言模型
  - **知识与推理**
  - 融合知识的NLP任务

知识与推理

- 问题：现有神经网络模型缺乏解释性，难以实现推理
- 可能的解决思路
  - 资源
    - 考虑知识
    - 引入常识
  - 模型
    - Memory Network
    - Graph Neural Network

知识与推理

❊引入知识的多跳推理

- ❊对话生成模型ConceptFlow，利用常识KG来显式地对对话流进行建模

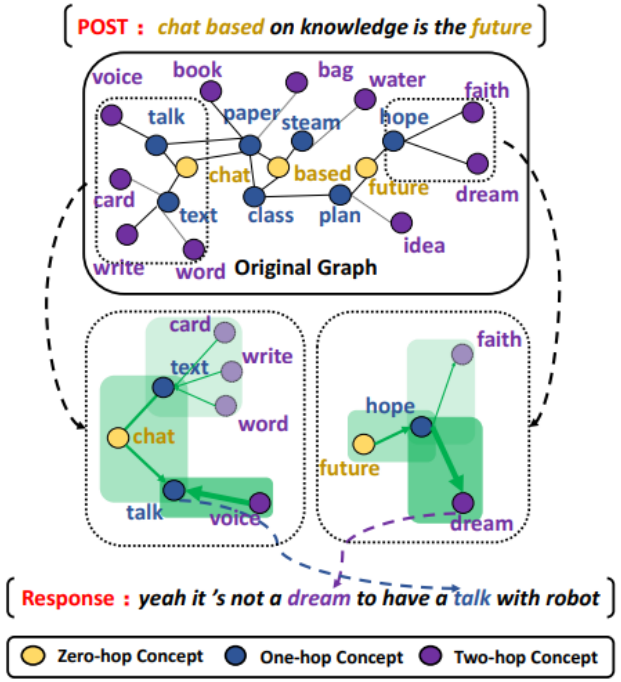


Figure 1: An Example of Concept Shifts in a Conversa-

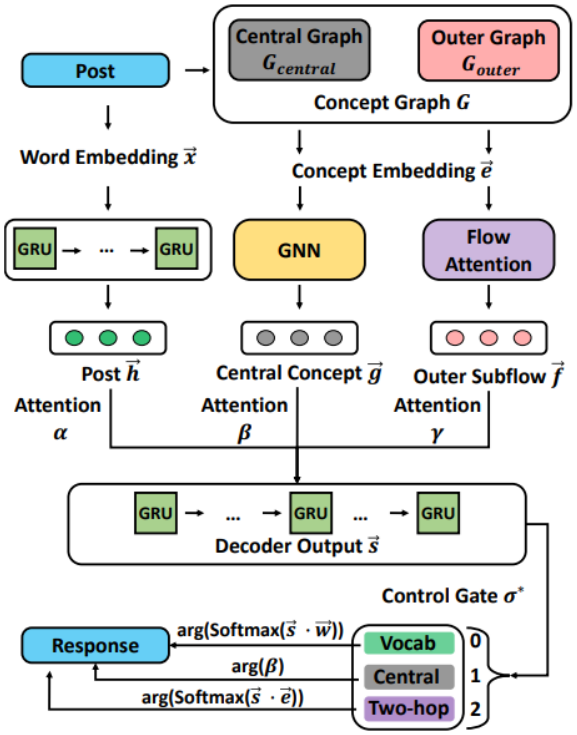
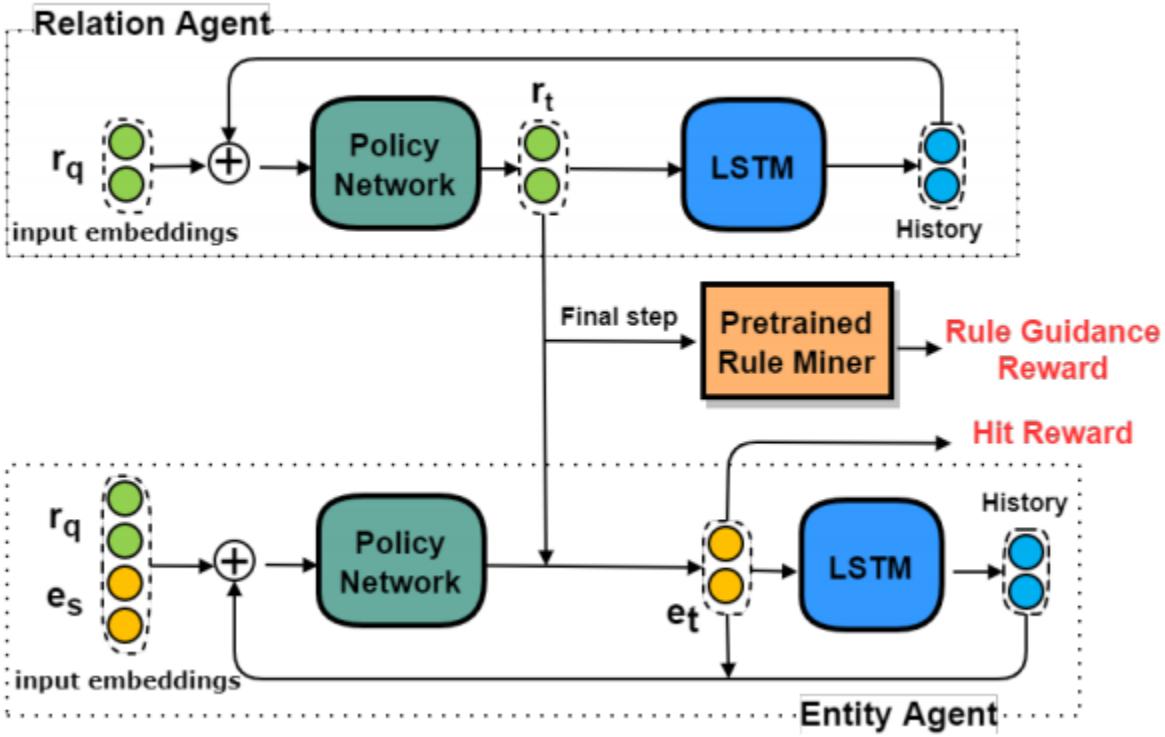


Figure 2: The Architecture of ConceptFlow.

Houyu Zhang et al. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs (ACL2020)

知识与推理

※知识推理，提高可解释性



Deren Lei et al. Learning Collaborative Agents with Rule Guidance for Knowledge Graph Reasoning (EMNLP2020)

NLP中KG“构建与应用”的几大趋势

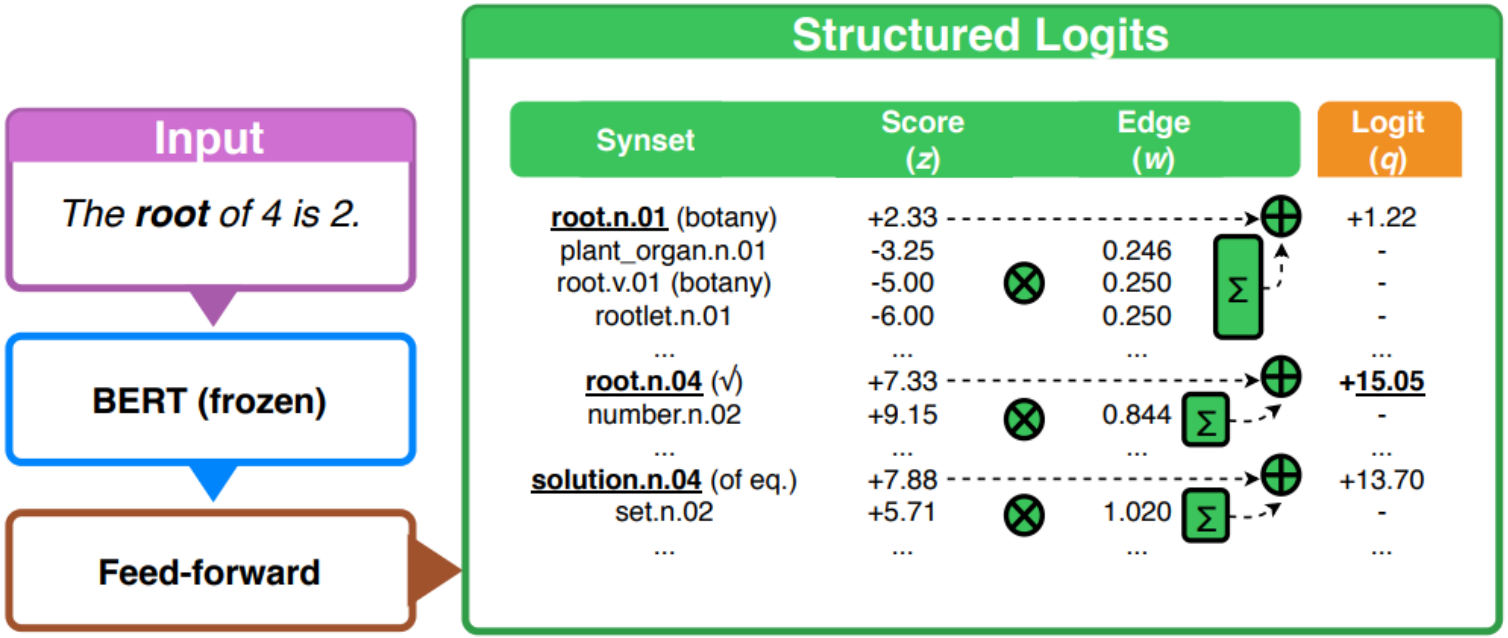
- 知识图谱构建
  - 实体相关研究
    - 多模态、低资源
  - 关系抽取相关研究
    - 联合抽取、开放抽取、文档级抽取、低资源
  - 模型的可解释性
- 知识图谱应用
  - 融合知识的预训练语言模型
  - 知识与推理
  - **融合知识的NLP任务**

融合知识的NLP任务

- ❊词义消歧
- ❊文本生成
- ❊事件因果分析
- ❊对话生成
- ❊阅读理解
- ❊.....

融合知识的NLP任务

※基于知识的词义消歧



Michele Bevilacqua et al. Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information (ACL2020)

融合知识的NLP任务

※基于知识的对话生成

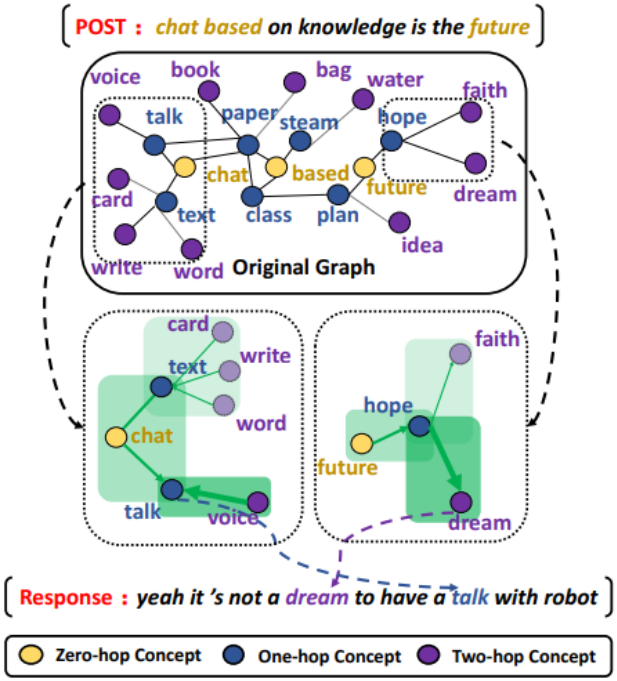


Figure 1: An Example of Concept Shifts in a Conversa-

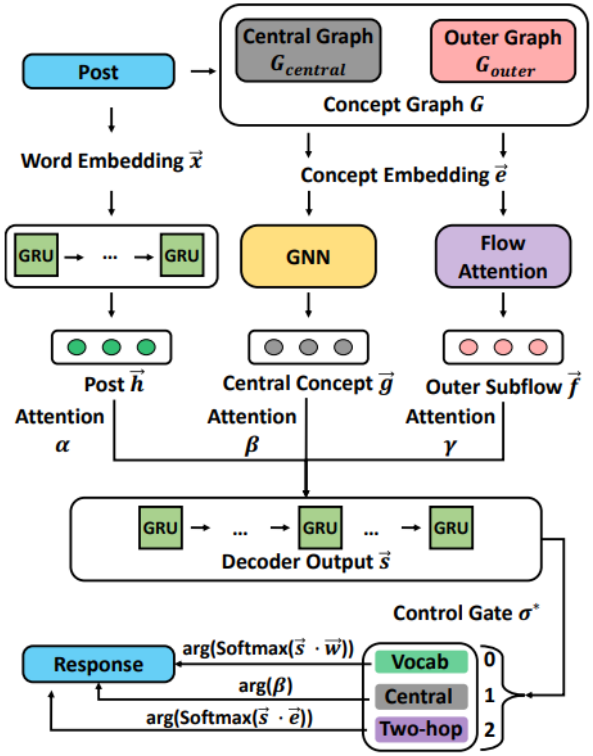


Figure 2: The Architecture of ConceptFlow.

Houyu Zhang et al. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs (ACL2020)

融合知识的NLP任务

❊基于知识的文本生成

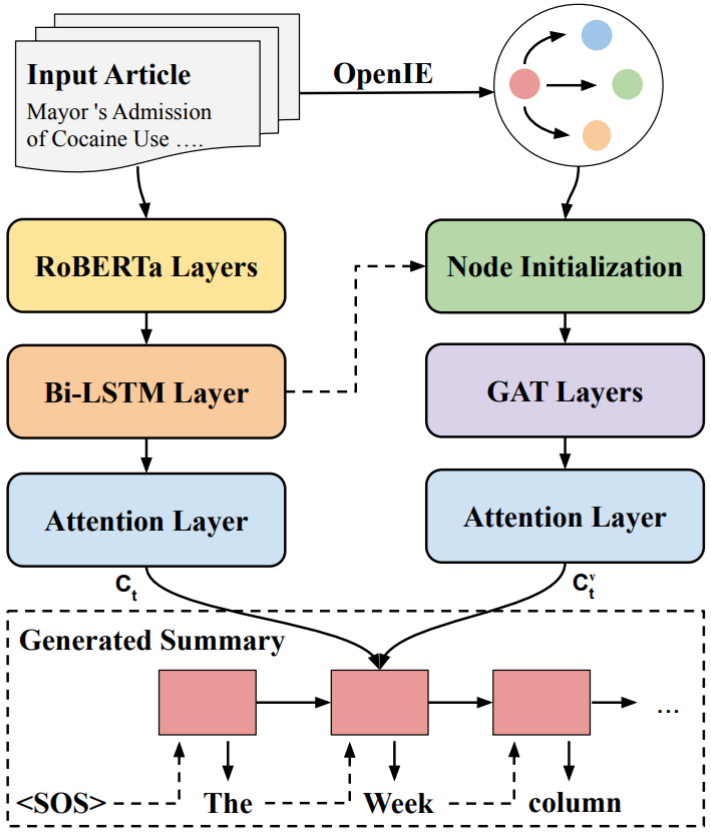
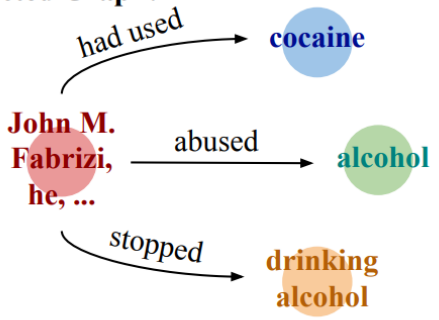


Figure 2: Our ASGARD framework with document-

Input Article of New York Times:
John M. Fabrizi, the mayor of Bridgeport, admitted on Tuesday that **he** had used **cocaine** and abused **alcohol** while in office.
Mr. Fabrizi, who was appointed mayor in 2003 after the former mayor, Joseph P. Ganim, went to prison on corruption charges, said **he** had sought help for his drug problem about 18 months ago and that **he** had not used drugs since.
About four months ago, **he** added, **he** stopped **drinking alcohol**.

Constructed Graph:



Summary by Human:
The Week column. **Mayor John Fabrizi** of Brigeport, Conn, publicly admits **he** used **cocaine** and abused **alcohol** while in office; says **he** stopped **drinking alcohol** and sought help for his drug problem about 18 months ago.

[1] Luyang Huang et al. Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward (ACL2020)
[2] Martin Schmitt et al. An Unsupervised Joint System for Text Generation from Knowledge Graphs and Semantic Parsing (EMNLP2020)

融合知识的NLP任务

- 基于知识的阅读理解
  - Synonym Knowledge Enhanced Reader for Chinese Idiom Reading Comprehension (COLING 2020)

- 基于知识的事件因果分析
  - KnowDis: Knowledge Enhanced Data Augmentation for Event Causality Detection via Distant Supervision (COLING 2020)

总结

- 知识图谱构建
  - 实体相关研究
    - 多模态、低资源
  - 关系抽取相关研究
    - 联合抽取、开放抽取、文档级抽取、低资源
  - 模型的可解释性
- 知识图谱应用
  - 融合知识的预训练语言模型
  - 知识与推理
  - 融合知识的NLP任务

换个角度

- 知识图谱正在走向实用
  - 研究条件逐渐放宽
    - 低资源
  - 研究条件逐渐考虑实际情况
    - 考虑的信息越来越多，如联合抽取、开放抽取、文档级抽取、多模态抽取等
- 知识图谱正在“增效降本”的道路上奔跑
  - 标注语料耗时费力
  - 低资源
- 知识图谱的应用将越来越广泛
  - 对话生成、阅读理解、词义消歧.....

Q&A

谢谢！！

感谢辛成鑫博士协助整理