Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

开放式信息抽取技术

门元昊

导师：罗森林

2023年05月03日

内容提要

- 背景简介
- 基本概念
  - 信息抽取
  - Tagging-based Model
  - Generative Model
  - 序列标记问题
- 算法原理
  - Transformer-Based OIE
  - Meta-Learning OIE
- 应用总结
- 参考文献

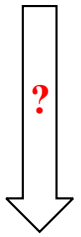
内容提要

- 预期收获
  - 1. 了解开放式信息抽取的基本概念
  - 2. 理解开放式信息抽取的算法原理
  - 3. 了解开放式信息抽取的应用和发展方向

背景简介

- 如何从文本中提取结构化信息?
- 如何根据大量文本数据构建知识图谱?

Deep learning is a class of ML algorithms that uses multiple layers to extract features from the raw input



(Deep learning; is a class of; ML algorithms)
(Deep learning; uses; multiple layers)
(Deep learning; extract; features; from the raw input)

竞赛简介

信息安全与对抗技术竞赛(Information Security and Countermeasures Contest，简称ISCC)，由罗森林教授创建，自2004年起每年举办一届，面向全国大学生组织个人挑战赛和分组对抗赛，第1届-第16届累积参加人数超过35000人，全国参加院校数700所以上。大赛宗旨是"**提升信息安全意识，普及信息安全知识，实践信息安全技术，共创信息安全环境，发现信息安全人才**"。同时探索信息对抗技术及其相关专业工程教育的新途径。

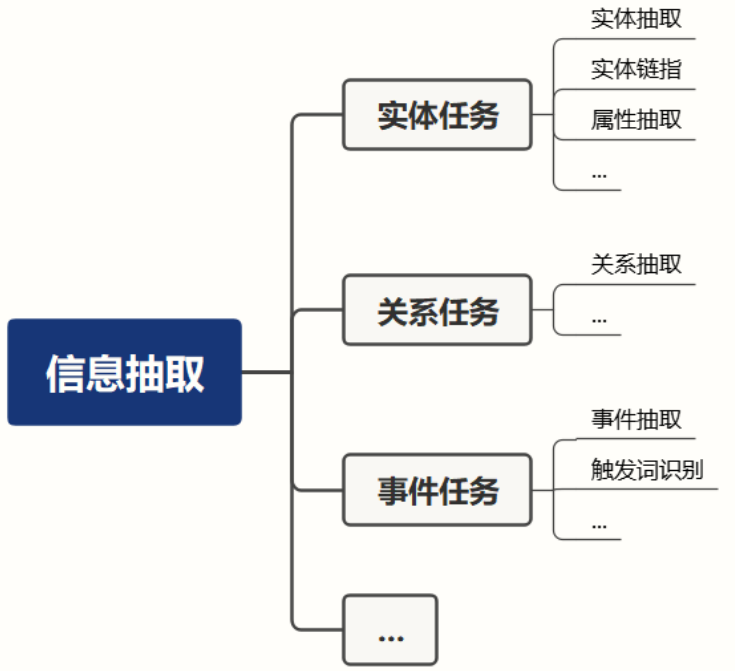
2007年8月，竞赛活动进一步得到了教育部高教司、工业和信息化部人事司的肯定。经批准，在全国大学生电子设计竞赛中增设一项信息安全技术专题邀请赛，即增设《全国大学生电子设计竞赛信息安全技术专题邀请赛》，且于2008年起每两年举办一次，为全国大学生提供了更多的机会，对向全国范围普及和推动信息安全技术具有十分重要的作用。

ISCC经过多年的发展，竞赛平台日渐完善、知识范围不断拓展、攻防方式逐步丰富。现在，每年一度的ISCC已成为全国各地信息安全人才思想、知识、技术交流的大好机会。参赛者组成正越来越丰富，早已经不再局限于在校的本科生和硕博研究生，越来越多在校的教职员工、已毕业的校友以及兄弟学校的学生甚至校外安全组织均有人注册参赛，可见竞赛的影响力仍在逐步扩大。

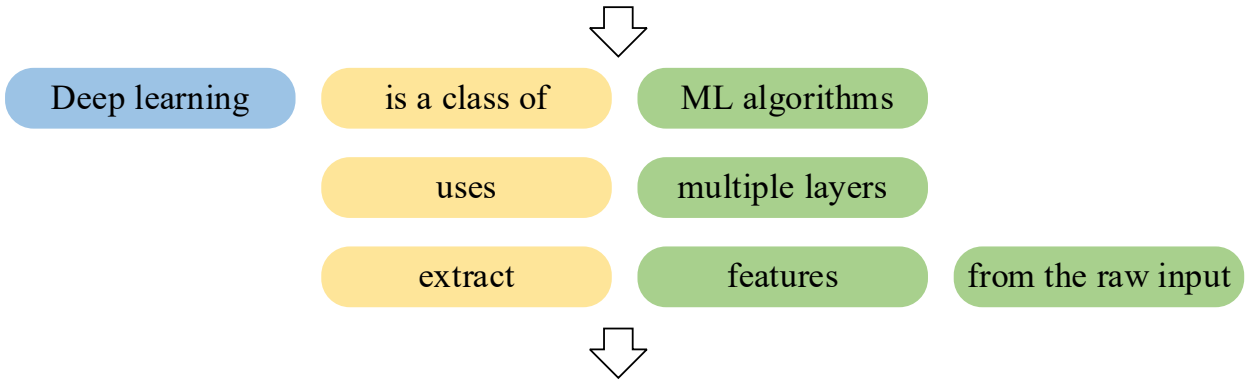
基本概念

基本概念

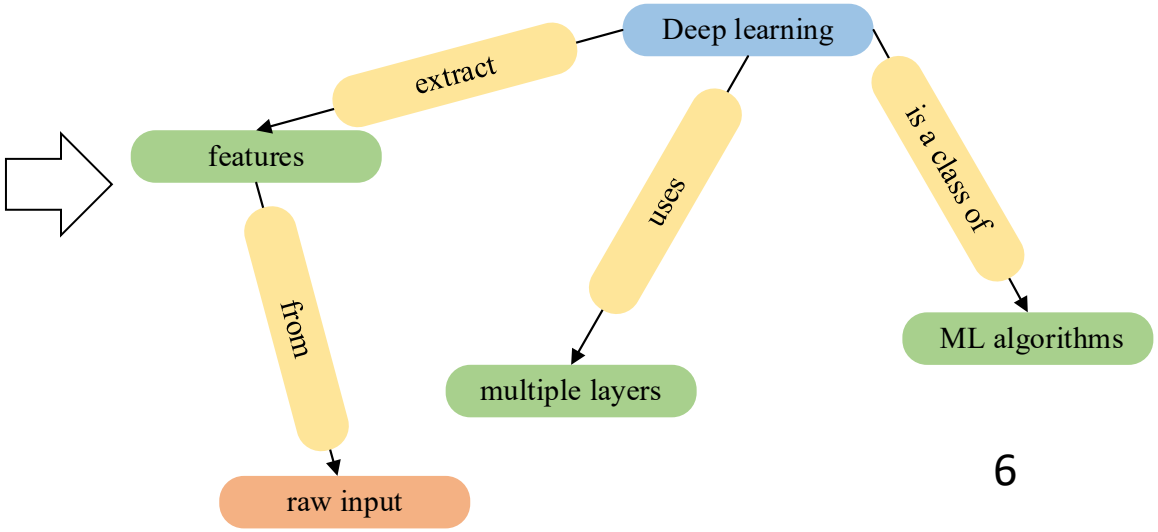
- 信息抽取
  - 对文本内包含的信息进行结构化处理
  - 输入原始文本，输出固定格式的信息
- 开放式信息抽取
  - 不依赖于预先定义的模式（schema）
  - 以n元关系组的形式提取事实



Deep learning is a class of ML algorithms that uses multiple layers to extract features from the raw input

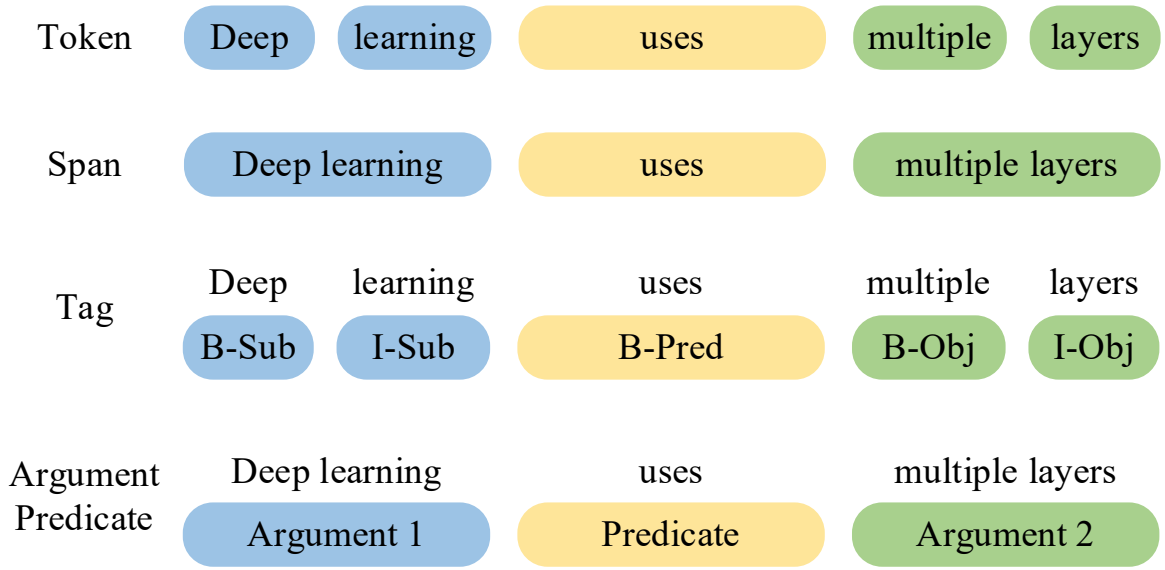


(Deep learning; is a class of; ML algorithms)
(Deep learning; uses; multiple layers)
(Deep learning; extract; features; from the raw input)



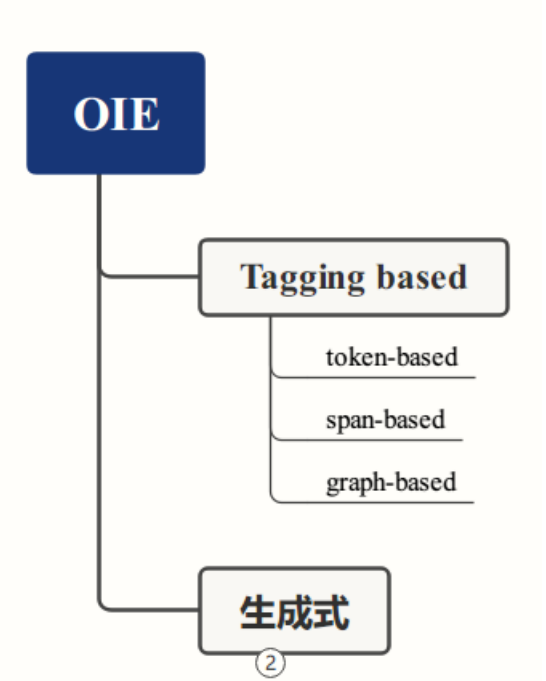
基本概念

- 概念辨析
  - token
    - 标记，指语句中一个独立的词或符号
  - span
    - 跨度，通常由一个或多个标记组成
  - tag
    - 标签，标记或跨度对应的类别
  - argument和predicate
    - 关系元组的组成部分
    - 单个元组通常由一个predicate和一个以上的argument组成
    - $(arg1, pred, arg2, \dots)$



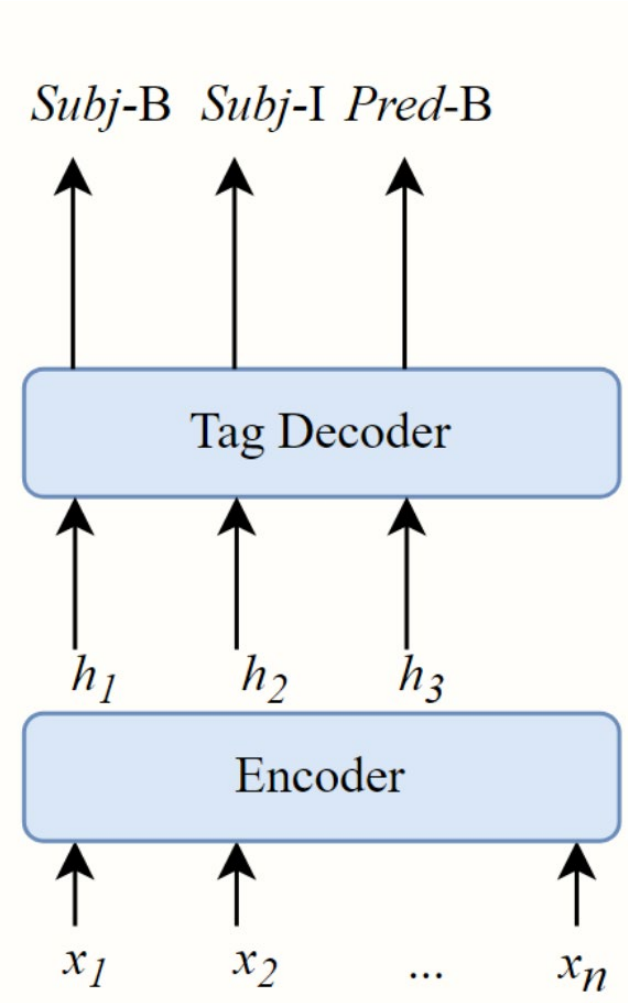
基本概念

- Tagging based Model
  - 以序列标记任务的角度处理OIE问题
  - 给定序列和标签，学习单个标记或跨度的概率分布
  - 常见类型有token、span、graph三种
- token-based（基于标记）
  - 依次判断每个token是否属于arg或pred
- span-based（基于跨度）
  - 将token划分为span，直接判断span是否为arg或pred
- graph-based（基于图）
  - 基于token span构建图，直接识别关系三元组

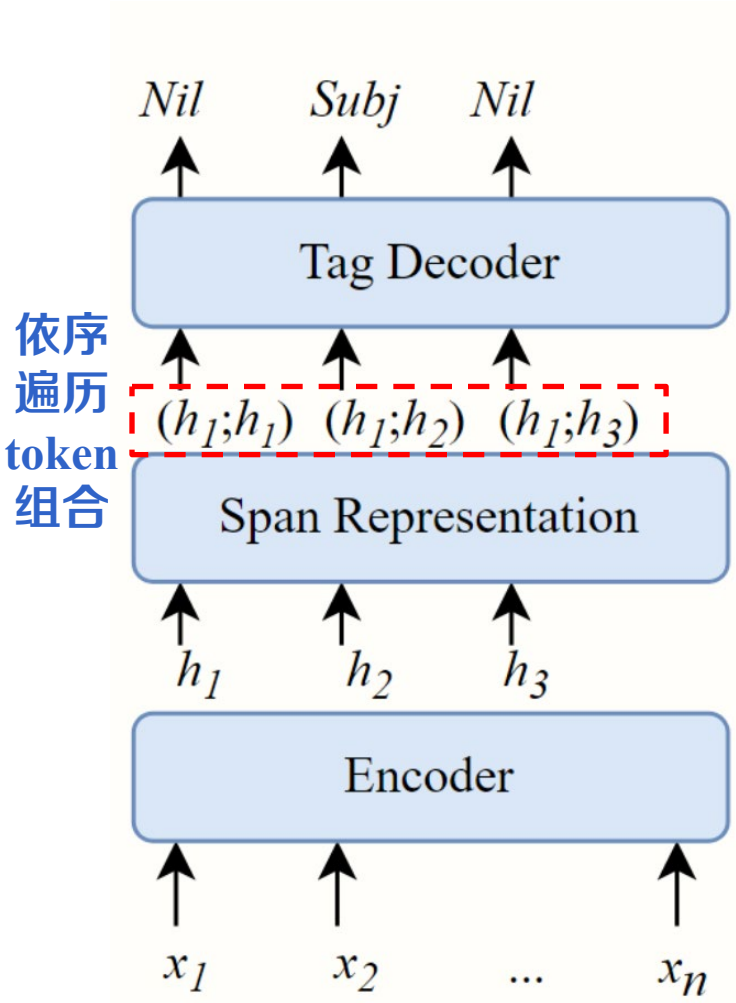


基本概念

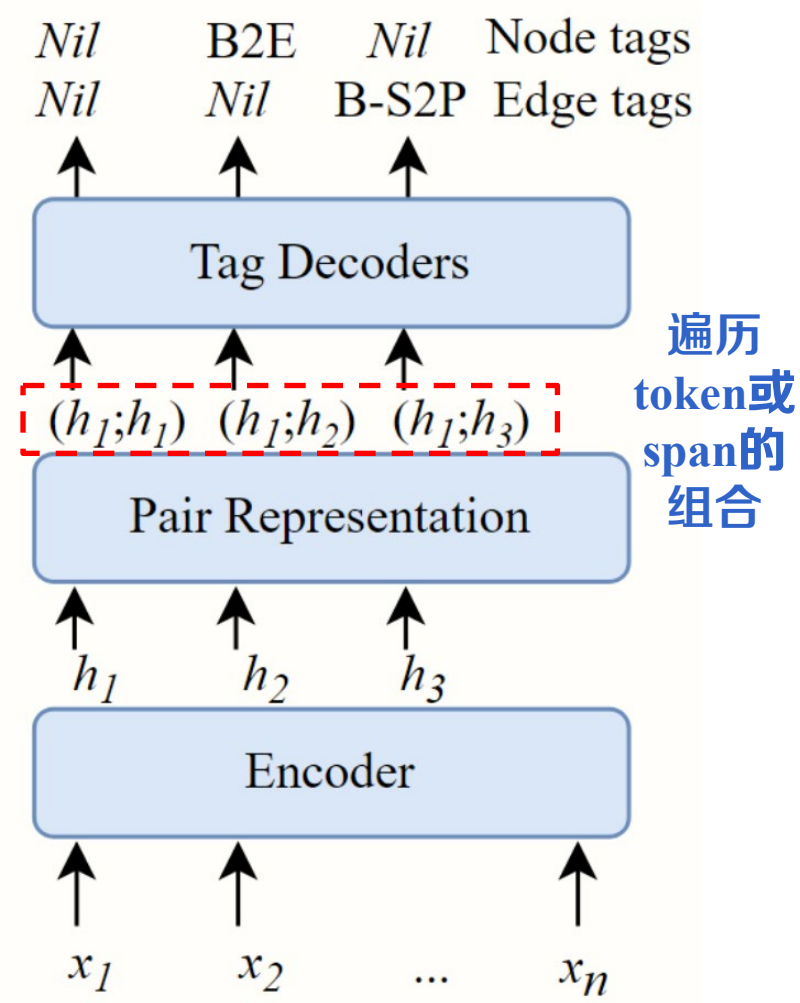
- **Tagging-based Model**



token-based



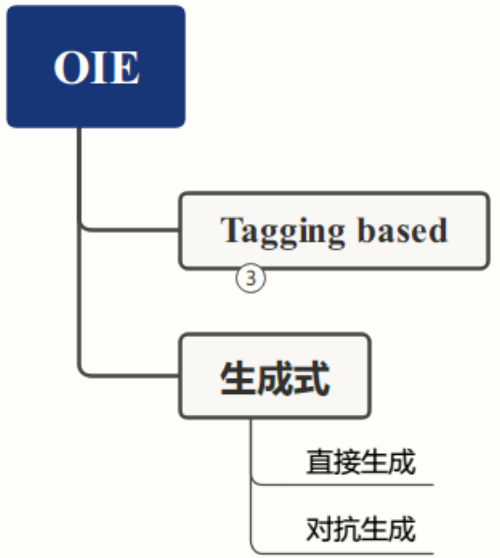
span-based



graph-based

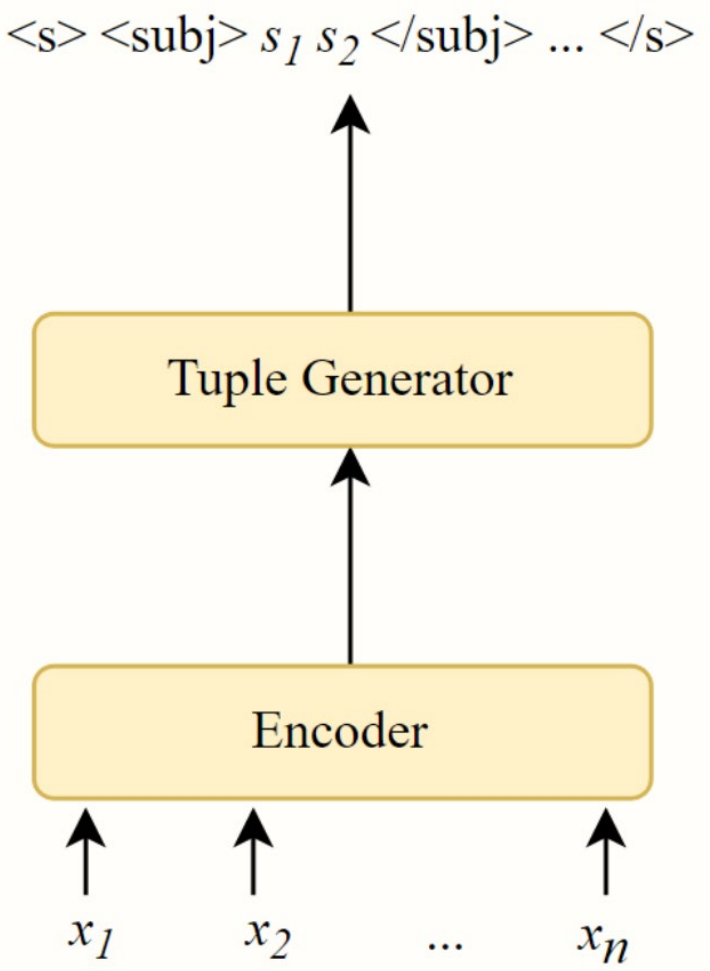
基本概念

- Generative Model（生成式模型）
  - 以序列生成任务的角度处理OIE问题
  - 给定语句，输出生成的抽取序列
  - 常见类型有直接生成和对抗生成
- 直接生成
  - 编码器：依据给定语句，提取语句上下文表示
  - 解码器（生成器）：基于上下文表示，顺序生成序列
- 对抗生成
  - 鉴别器：与已有标记比较，判别生成标记真实性

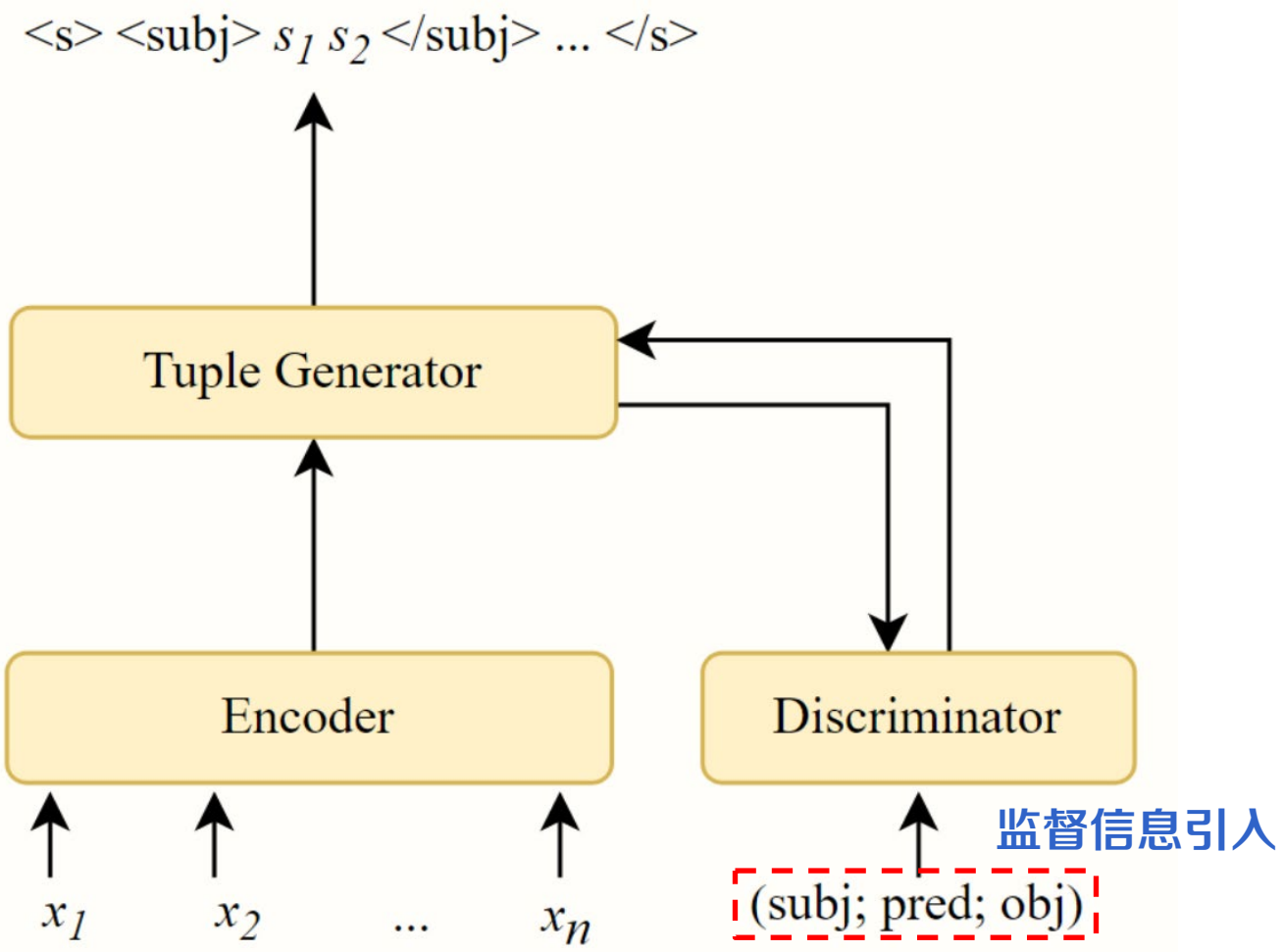


基本概念

- **Generative Model**



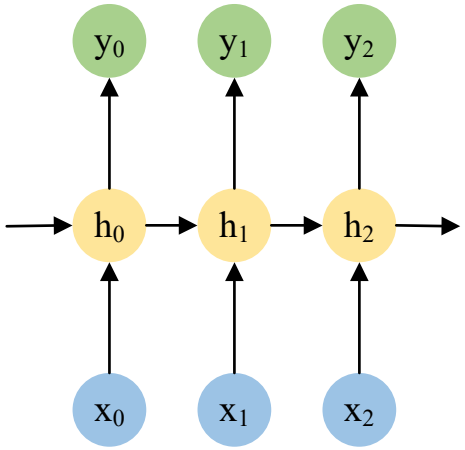
直接生成



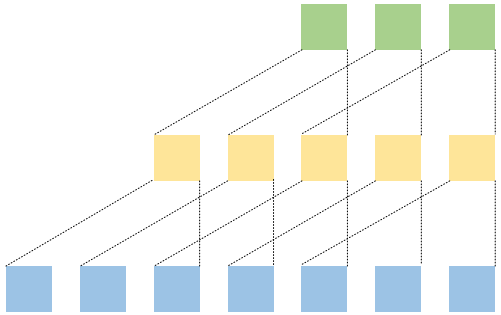
对抗生成

基本概念

- 序列编码问题
  - RNN
    - 递归式处理，$y_t = f(y_{t-1}, x_t)$
    - 优势：结构简单，模型逻辑贴合序列的形式
    - 劣势：缺乏并行能力，速度慢；长期依赖问题，难以学习全局信息



  - CNN
    - 窗口式遍历，$y_t = f(x_{t-1}, x_t, x_{t+1})$
    - 优势：并行计算，速度快；容易捕捉一部分全局信息
    - 劣势：需要通过堆叠增大感受野，面对长序列时堆叠过深



  - Attention
    - $y_t = f(x_t, A, B)$，A, B为序列
    - Transformer

推荐阅读：BFS学术报告-Transformer中的Multi-Head Attention-王睿怡
BFS学术报告-预训练语言模型GPT3-高依萌

算法原理

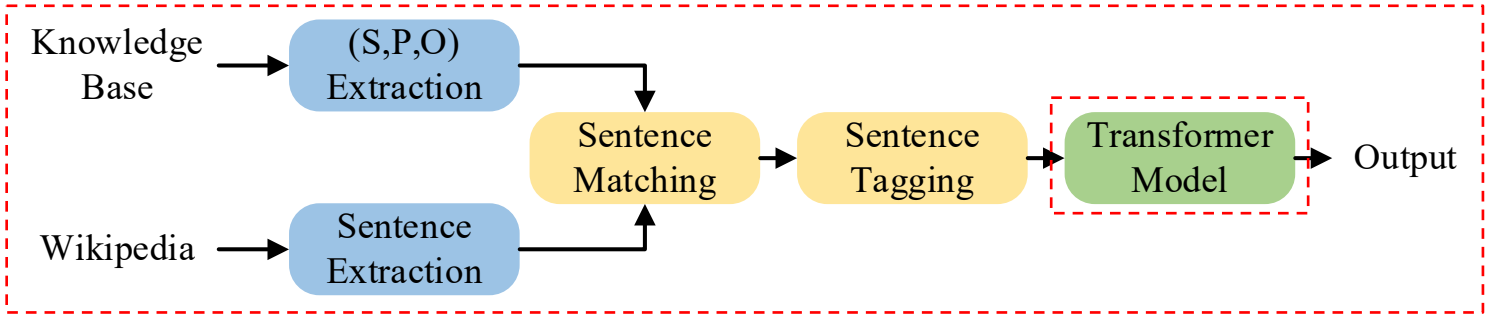
TIPO—Transformer

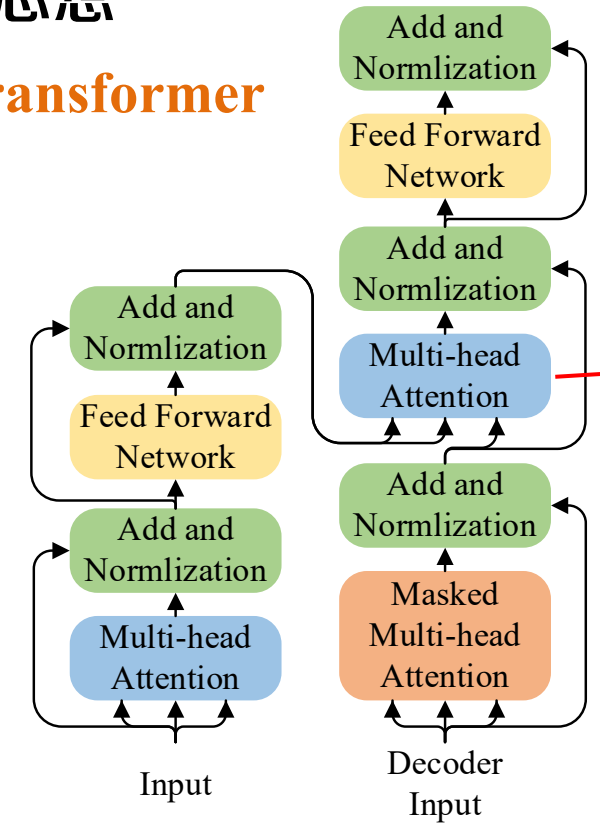
T	目标	从文本中提取信息三元组（S, P, O）
I	输入	文本数据（Wikidata、Wikipedia dump）
P	处理	使用Transformer直接进行端到端处理
O	输出	信息三元组（S, P, O）

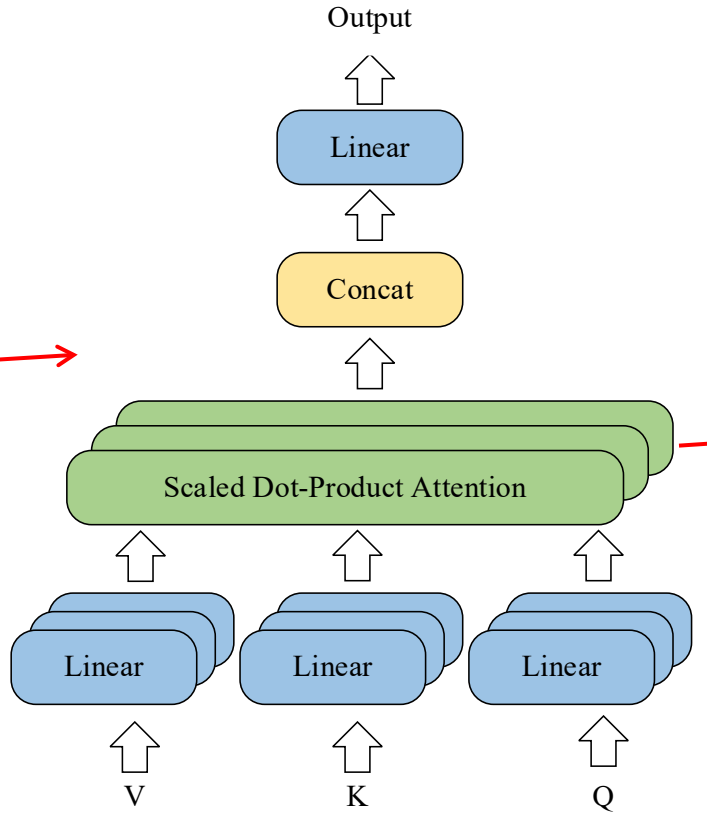
P	问题	手工构建提取模式所需人力和时间成本高昂
C	条件	不使用预先定义的提取模式
D	难点	有效表征语句上下文并抽取关系元组
L	水平	Engineering Applications of Artificial Intelligence, 2021

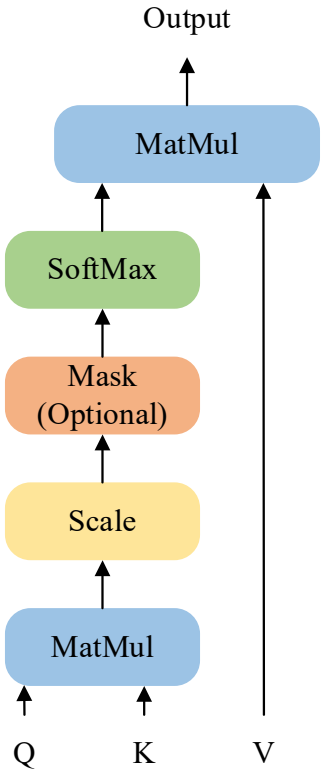
算法原理

- 整体流程
  - 远程监督构建数据集
  - 训练Transformer
- 核心思想
  - Transformer



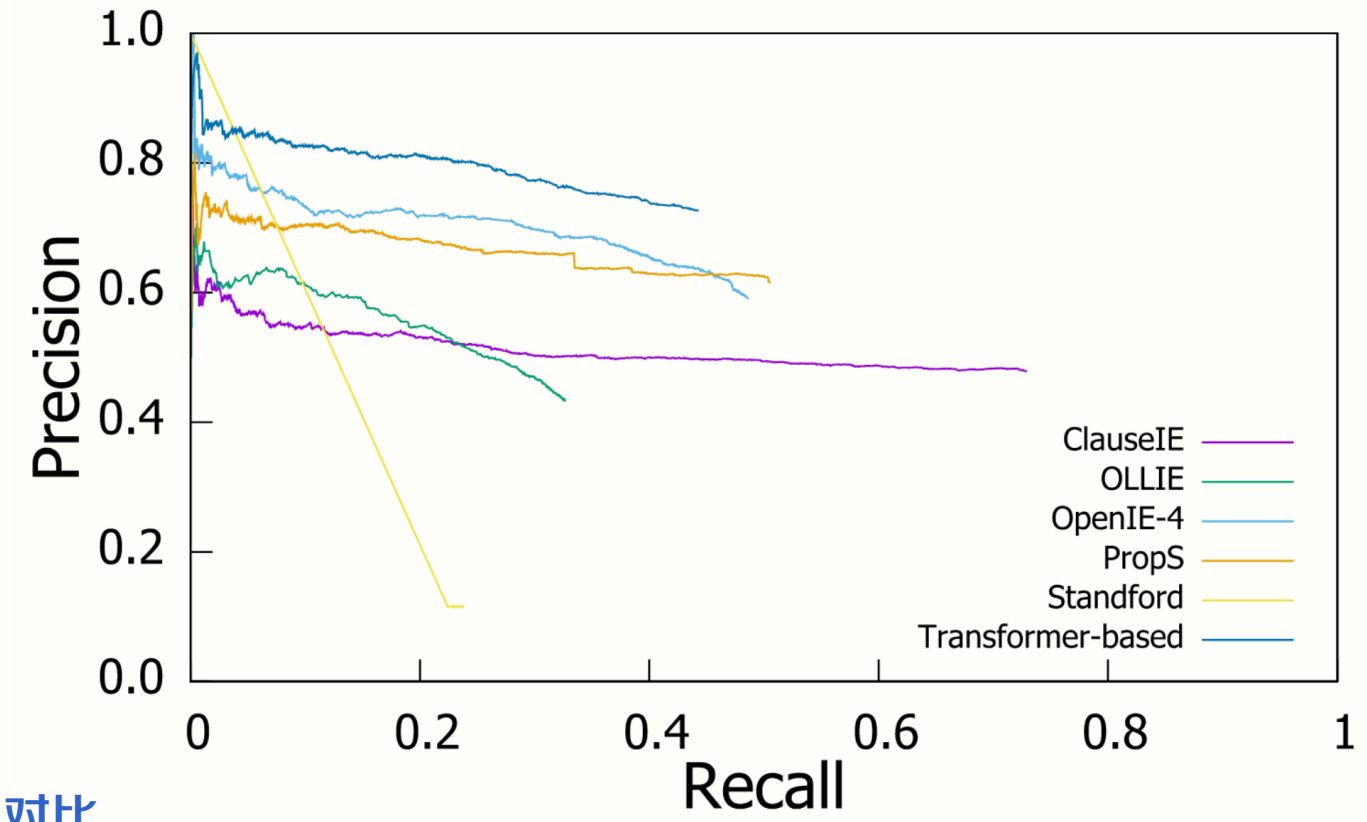






实验结果

- 对比实验结果
  - Transformer的OIE方法基本上全面优于clause-based（子句检测）方法（ClauseIE、Standford）和rule-based（人工规则构建）方法（OpenIE-4）



缺乏与其他Neural OIE方法的对比
难以验证Transformer在端到端Neural OIE方法中的先进性

TIPO——Meta

T	目标	实现对多领域文本进行信息抽取
I	输入	不同领域文本（电影、音乐、商业等）

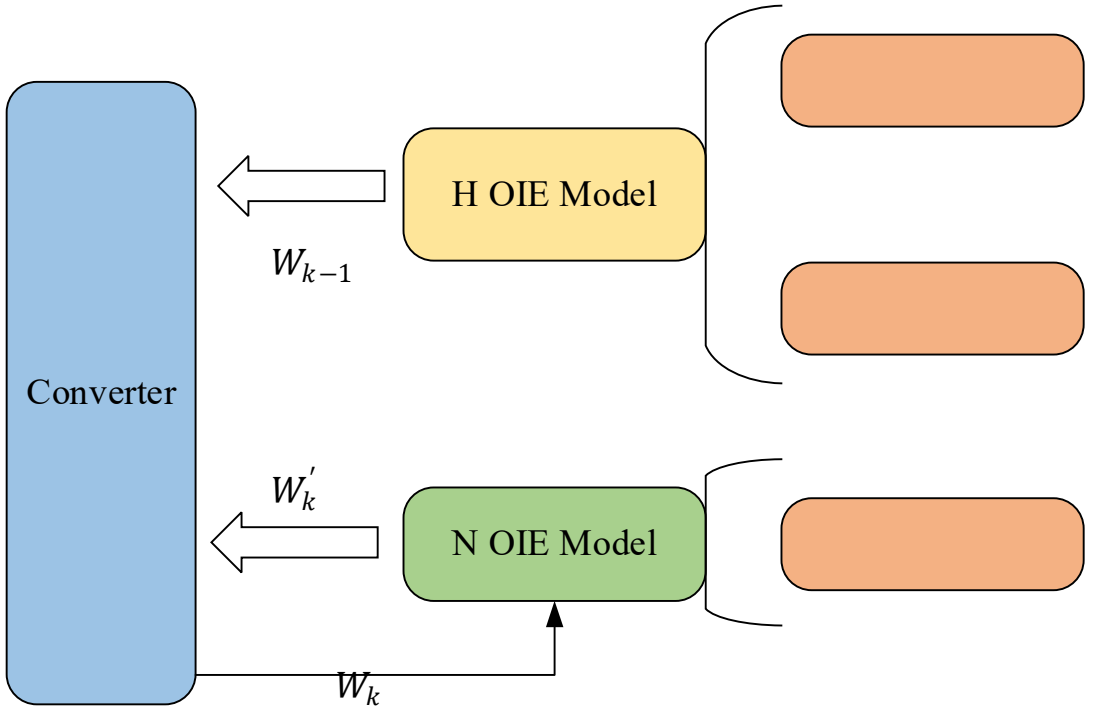
P	处理	1.初始化历史参数和转换器
		for d in domain:
		for idx in epoch:
		2.训练当前任务模型，输出任务参数
		3.转换器基于历史任务参数和当前任务参数计算更新参数
		4.当前任务模型依据转换器输出参数进行更新
		5.基于已有任务模型更新历史任务参数

O	输出	转换器、不同领域信息抽取模型

P	问题	不同领域文本的语义信息和关系存在差异
C	条件	输入文本数据中有多个不同领域
D	难点	如何提取和应用不同领域任务中的通用知识
L	水平	Neural Computing and Applications，2022

算法原理

- 核心思想
  - 元学习
  - 不同领域同类任务具备相似性
  - 利用共同知识可提高任务效果
- 关键步骤
  - 基于领域数据，训练任务模型，获取模型参数
  - 转换器利用历史任务参数和当前任务参数，计算模型更新参数
  - 模型训练完毕后，用新模型参数更新至历史任务参数



推荐阅读：BFS学术报告-从任务划分就开始与众不同的元学习-潘珲

算法原理

- 转换器
  - 统一历史参数和当前参数的维度

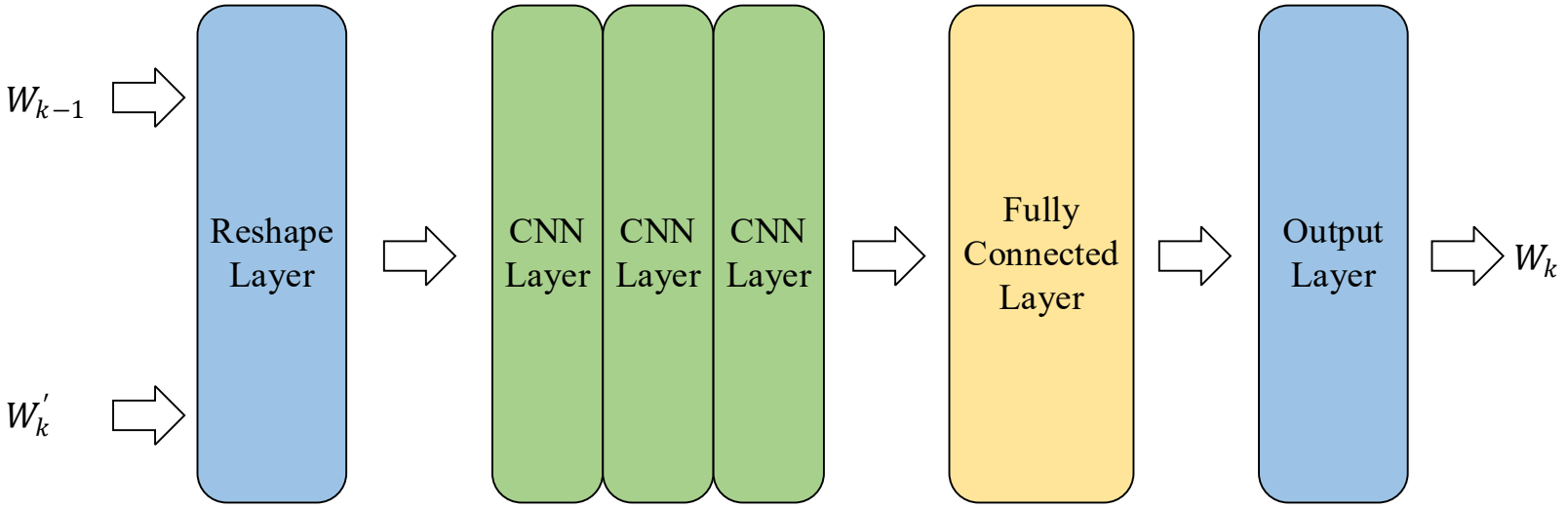
$$H_s = [W_{k-1}; W'_k; W_d],\ W_d = \frac{W_{k-1} \odot W'_k}{\|W_{k-1}\| + \zeta}$$

  - 提取参数维度间关联关系，生成新参数

$$W_k = W_o^T GELU\left(W_f^T flatten(H^c) + b_f\right) + b_o$$

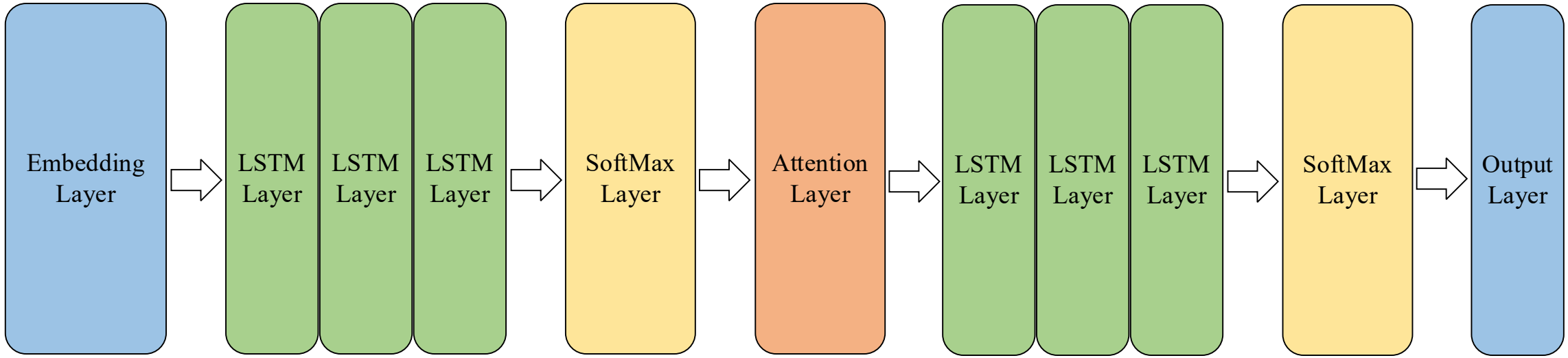
$$W_k = f(W_{k-1}, W'_k)$$

方式	缺点
降维求和	特征表达能力有限
MLP	特征表达能力有限
RNN	输出维度；长时依赖



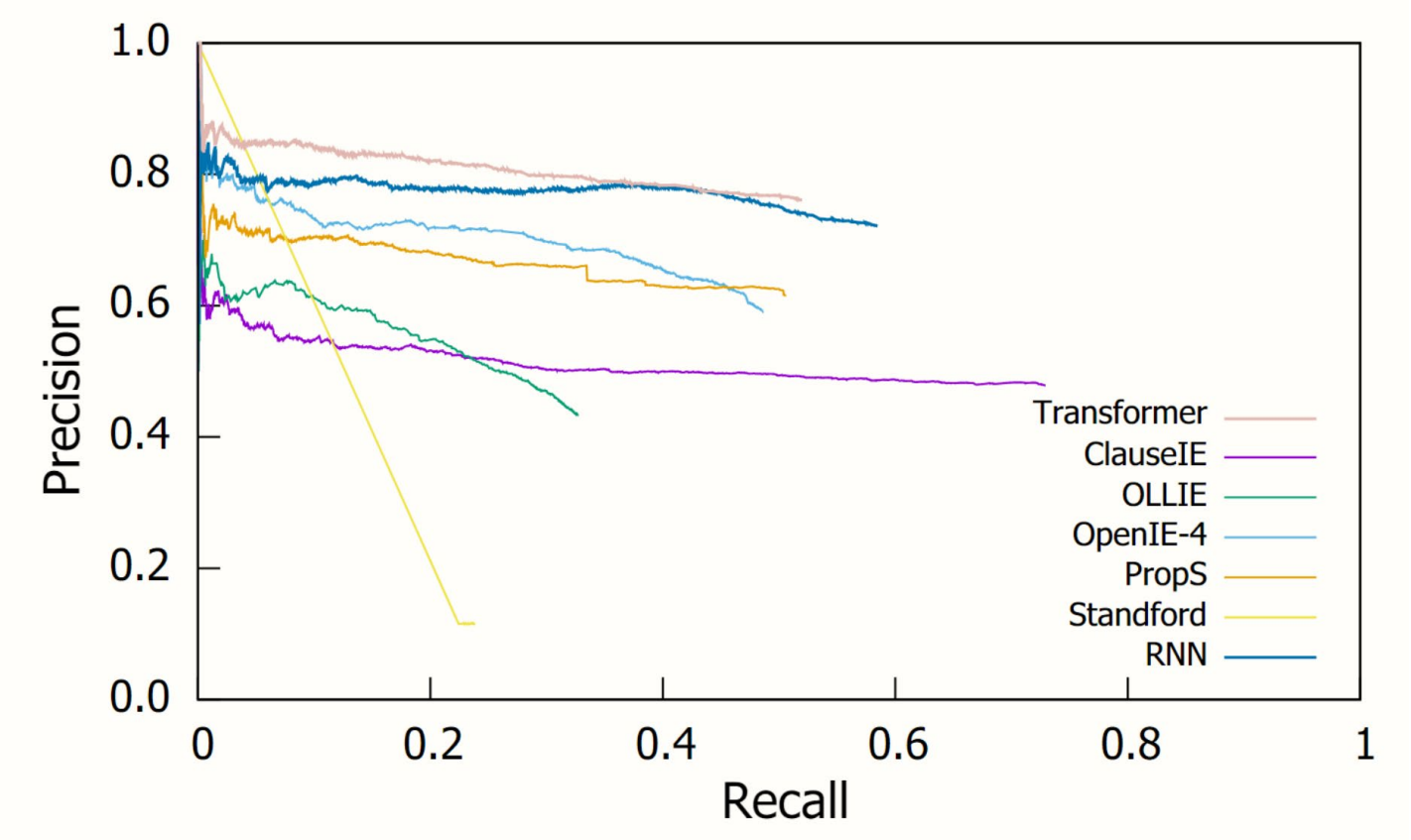
算法原理

- OIE模型
  - 常规的Tagging based Model
  - LSTM提取序列特征，注意力机制处理上下文
  - 另外也使用了一套基于Transformer搭建的OIE模型



实验结果

- OIE模型效果对比实验
  - Neural OIE方法中，Transformer整体优于RNN模型（LSTM）
  - Neural OIE方法普遍优于clause-based类方法和rule-based类方法



实验结果

- 参数更新方式效果对比实验
  - 元学习接近甚至部分优于全部重训，远好于微调和采样重训
  - 元学习训练时间略高于模型微调
  - 元学习训练时间不随领域增多而变长

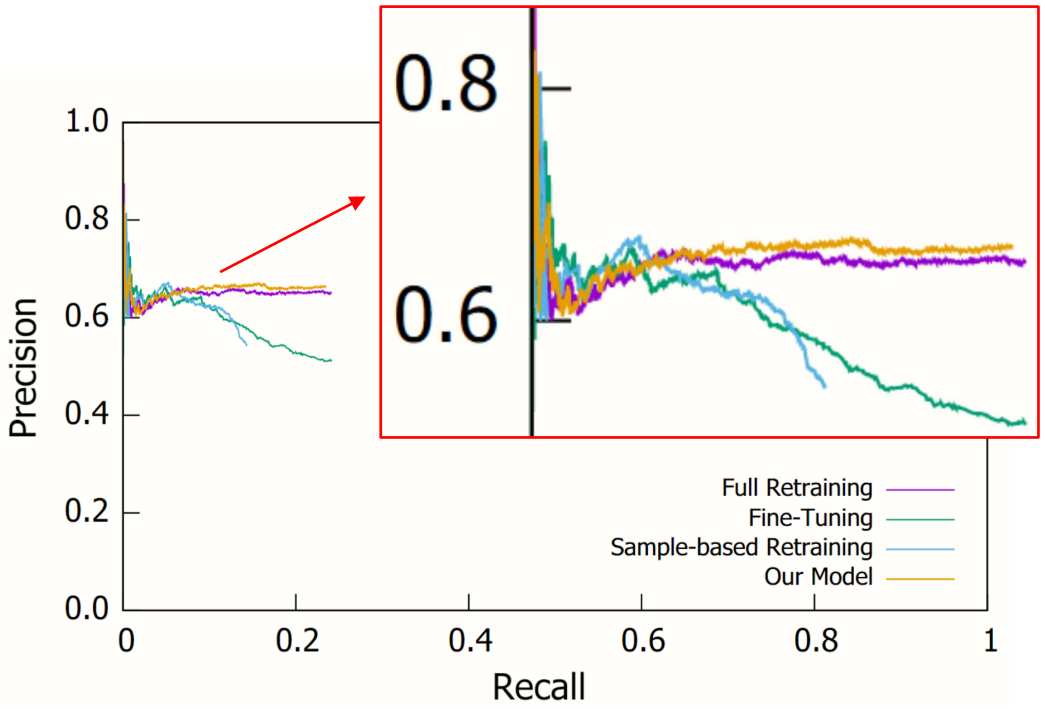


Table 8 Retraining time (min) at training (OIE:RNN) of each domain data

Domain data	0	1	2	3	4	5	6	7	8
Full-retraining	48	90	150	199	243	301	366	412	474
Fine-tuning	51	62	43	69	41	57	49	62	67
Our model	82	87	75	79	67	71	69	85	91

Table 9 Retraining time (min) at training (OIE:Transformer) of each domain data

Domain data	0	1	2	3	4	5	6	7	8
Full-retraining	32	67	102	143	181	219	267	312	352
Fine-tuning	47	59	51	49	45	39	42	37	44
Our model	67	72	59	64	58	51	52	61	55

效果与效率综合最优，性价比最高

总结

优劣分析

- Neural OIE（对比传统方法）
  - 优势
    - 精确度和召回率更高
    - 端对端方法，处理更简便
  - 劣势
    - 训练时间和计算资源开销增加
    - 数据资源需求大
    - 现有方法很大程度上依赖传统方法标注出的数据

应用总结

- 应用领域
  - 自动化知识库构建
  - 知识图谱问答
- 发展方向
  - 领域扩展，大多数研究仍局限在新闻、电影、百科等领域，需要覆盖到更多领域
  - 内容扩展，不局限于单一语言，可进一步扩展至多语言甚至多模态

参考文献

[1] Zhou S, Yu B, Sun A, et al. A survey on neural open information extraction: Current status and future directions[J]. arXiv preprint arXiv:2205.11725, 2022.

[2] Han J, Wang H. Transformer based network for open information extraction[J]. Engineering Applications of Artificial Intelligence, 2021, 102: 104262.

[3] Han J, Wang H. A meta learning approach for open information extraction[J]. Neural Computing and Applications, 2022, 34(15): 12681-12694.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。