

# CCL 2019自然语言处理国际前沿动态综述 句法分析方向

屠可伟（上海科技大学）

<http://faculty.sist.shanghaitech.edu.cn/faculty/tukw/>




上海科技大学  
ShanghaiTech University

# 趋势统计

---

## ▶ “Tagging, Chunking, Syntax and Parsing” Area

	投稿量	总投稿量占比	本领域接收率	总体接收率
ACL 2017	78	6.0%	26.9%	23.3%
ACL 2018	61	3.9%	?	24.9%
ACL 2019	99	3.7%	27.3%	22.7%



# 概要

---

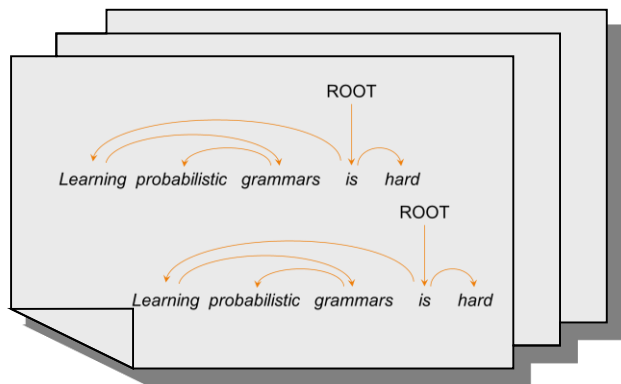
- ▶ 句法分析前沿动态综述（2019年论文为主）
  - ▶ 有监督句法分析
  - ▶ 无监督句法分析
  - ▶ 跨领域、跨语言句法分析



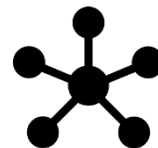
# 有监督句法分析

---

## 训练语料



## 句法分析器



每个句子都有句法分析标注（树库）

### ▶ 近几年趋势

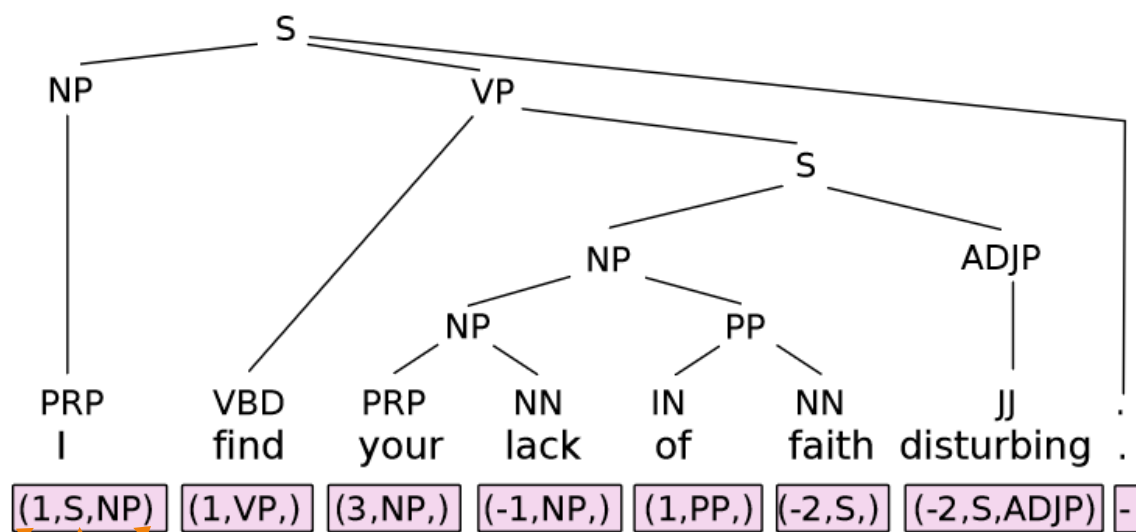
- ▶ 神经网络结合传统方法
- ▶ 准确率不断提升，上升空间不大
- ▶ 论文：深入分析 + 新方法



# 新方法

## ▶ Sequence Labeling Parsing

- ▶ 把成分句法分析变成一个序列标注任务
- ▶ 优点：更快！



与下一个词的公共祖先数（相对前一个值）

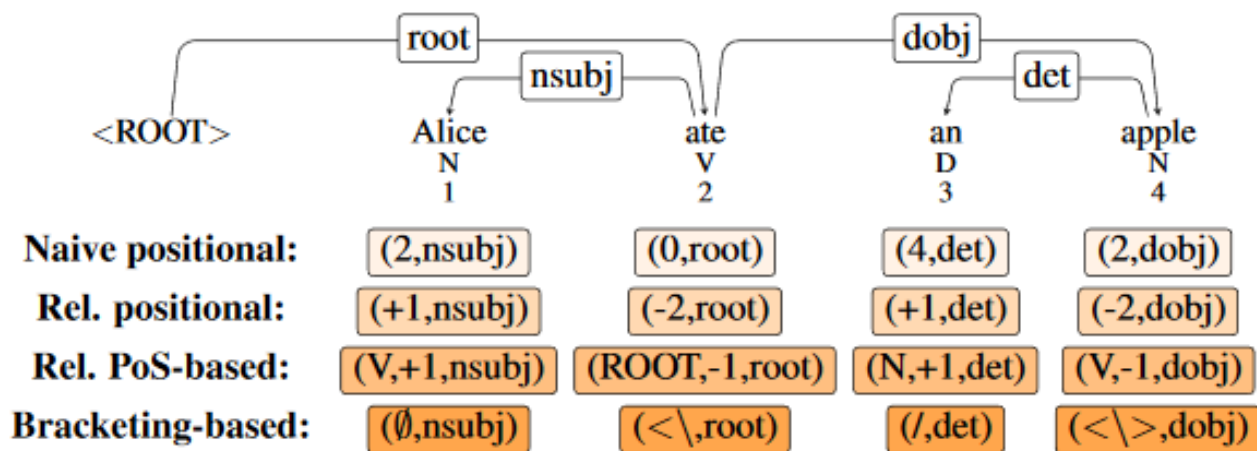
与下一个词的最低公共祖先

最低公共祖先到该词的 unary chain

# 新方法

## ▶ Sequence Labeling Parsing的新工作

- ▶ 改进编码方式、三部分分开预测
- ▶ 用于依存分析



## ▶ 同时进行成分和依存分析

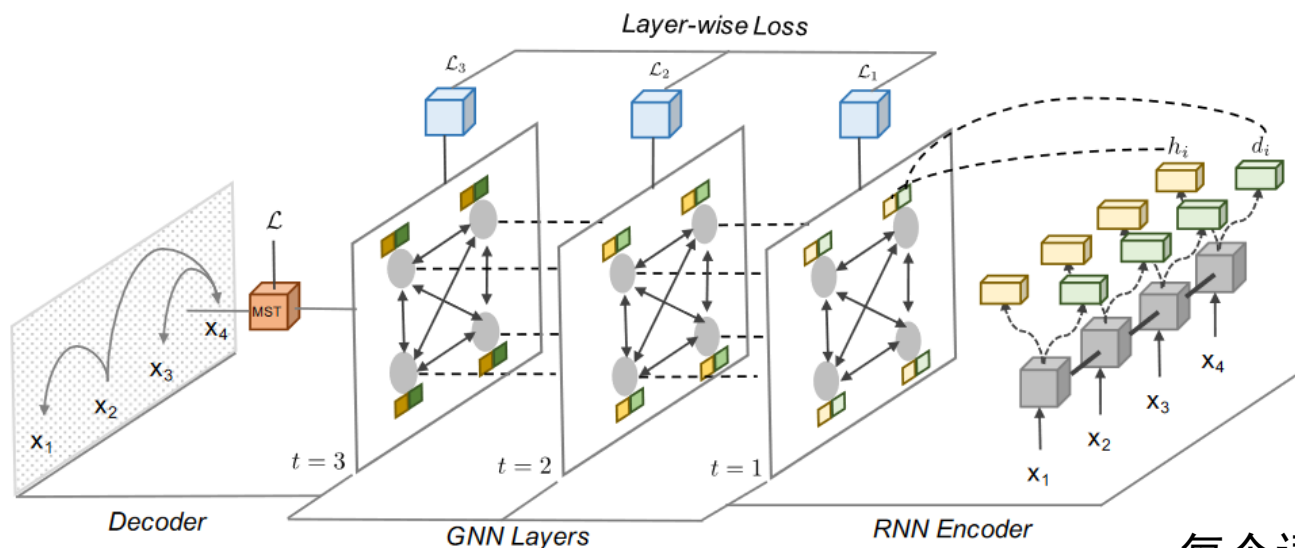
Vilares, et al., Better, Faster, Stronger Sequence Tagging Constituent Parsers (NAACL 2019)

Strzyz, et al., Viable Dependency Parsing as Sequence Labeling (NAACL 2019)

Strzyz, et al., Sequence Labeling Parsing by Learning across Representations (ACL 2019)

# 新方法

## 基于图神经网络的依存分析



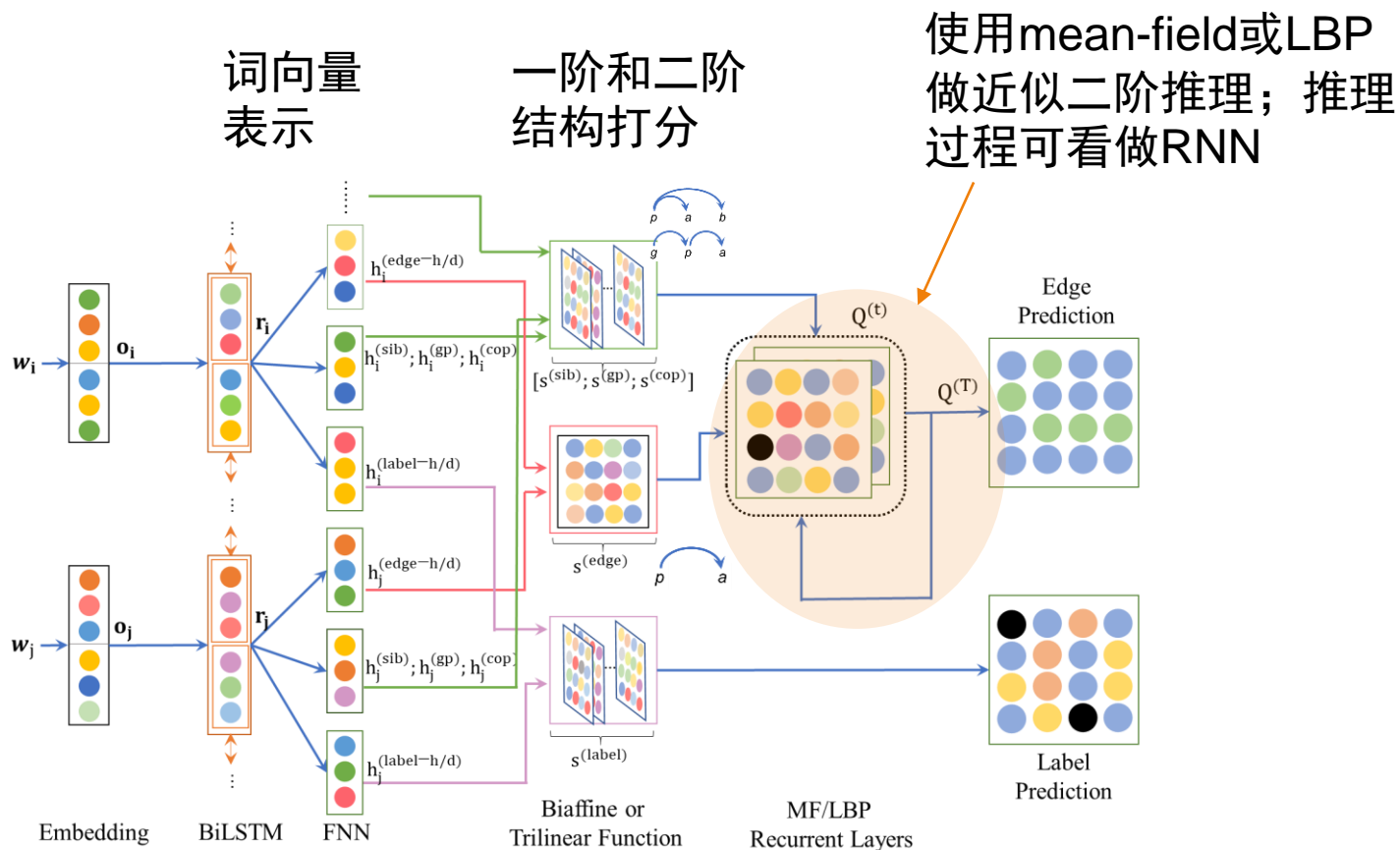
使用新的词向量表示计算边的分数，跑MST

使用一阶信息构建图，基于二阶结构不断更新向量表示

每个词有作为head和dependent两个向量表示

# 新方法

## 二阶依存分析 → 端到端神经网络



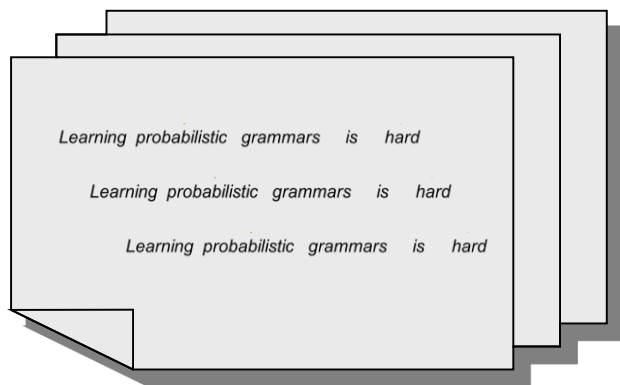
Wang, et al., Second-Order Semantic Dependency Parsing with End-To-End Neural Networks (ACL 2019)



# 无监督句法分析

---

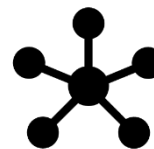
## 训练语料



训练句子没有句法分析标注



## 句法分析器



(部分工作假设学习时有下游任务反馈)

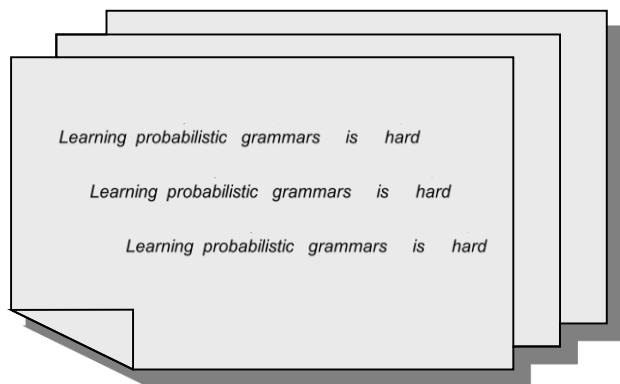
- ▶ 近两年**突然升温**
  - ▶ ACL 2019有7篇相关论文，NAACL/EMNLP也有多篇论文
  - ▶ ICLR 2019 best paper



# 无监督句法分析

---

## 训练语料

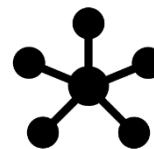


训练句子没有句法分析标注



(部分工作假设学习时有下游任务反馈)

## 句法分析器



### ▶ 新趋势

- ▶ 过去十几年：无监督依存分析为主
- ▶ 这两年：无监督**成分**分析强势回归
  - ▶ NAACL 2019：成分2、依存0
  - ▶ ACL 2019：成分5、依存2



# 生成式模型

---

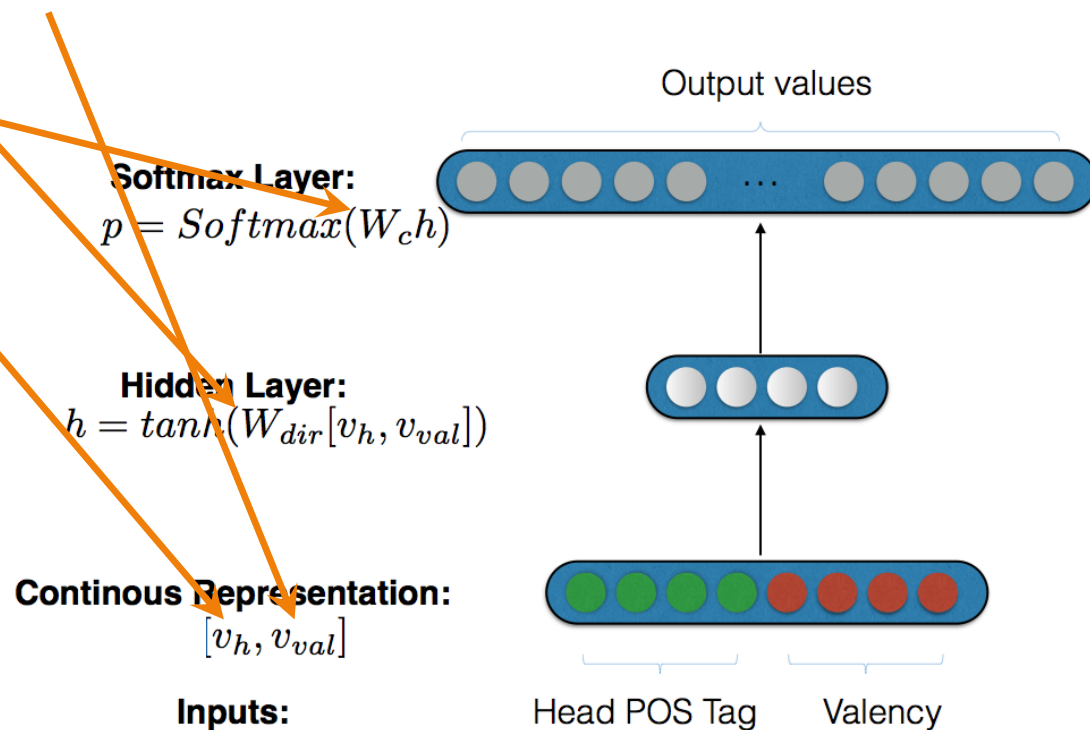
- ▶ 建模  $P(\text{sentence}, \text{parse})$ ，一般分解成多个规则概率的乘积
  - ▶ 概率上下文无关文法 (PCFG)
  - ▶ Dependency Model with Valence (DMV)
- ▶ 训练目标:  $P(\text{sentence})$
- ▶ 训练算法: Expectation-maximization



# 生成式模型

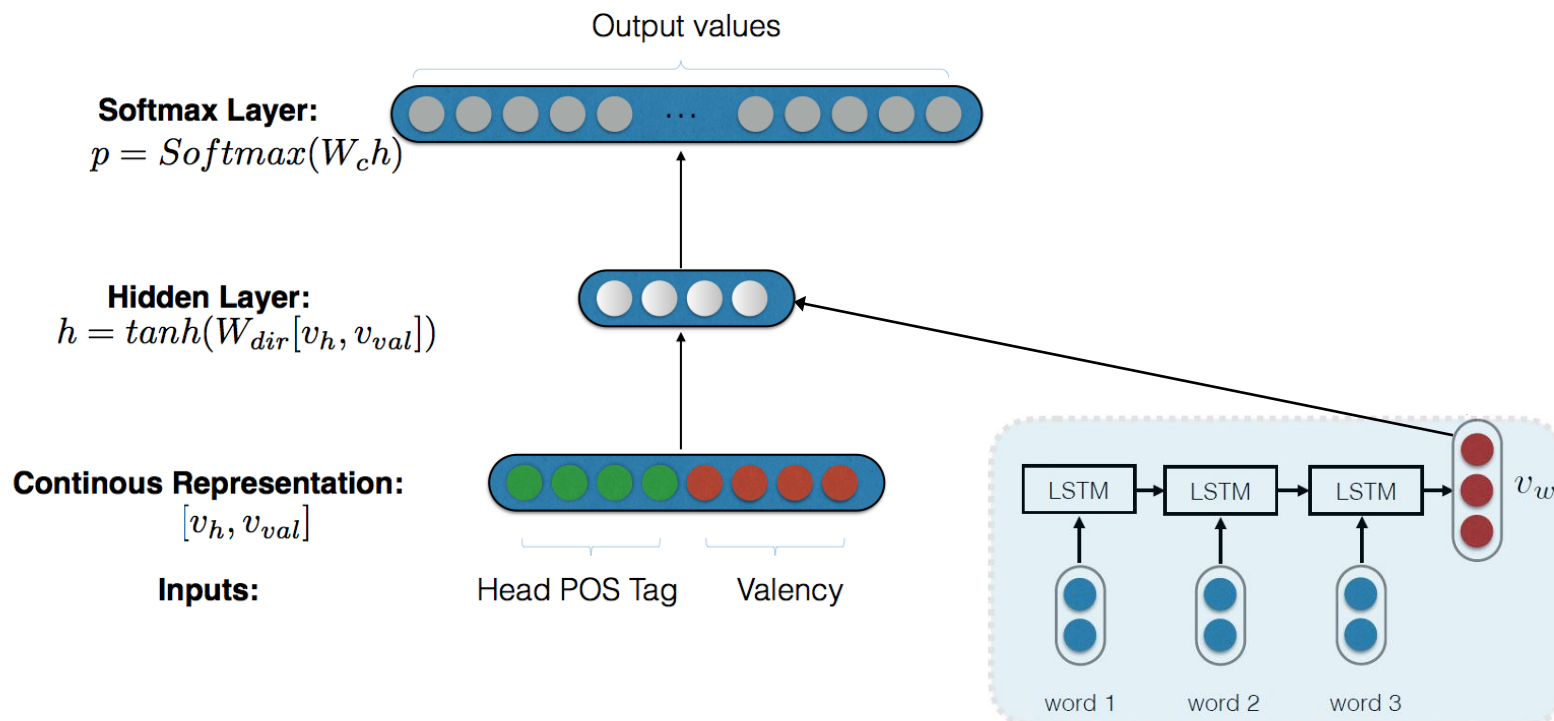
## ► Neural DMV

$$P(\text{child} \mid \text{head}, \text{direction}, \text{valency})$$



# 生成式模型

## ► Discriminative Neural DMV

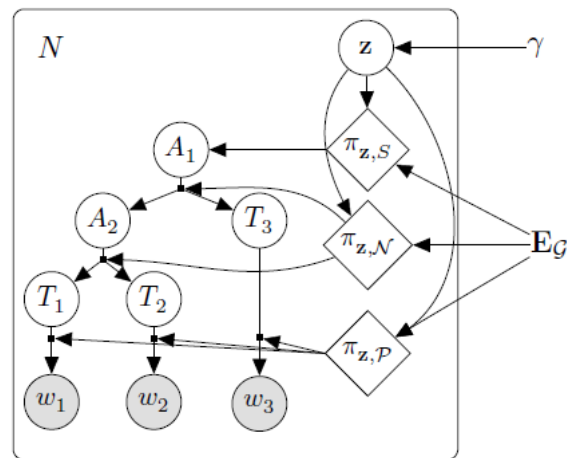
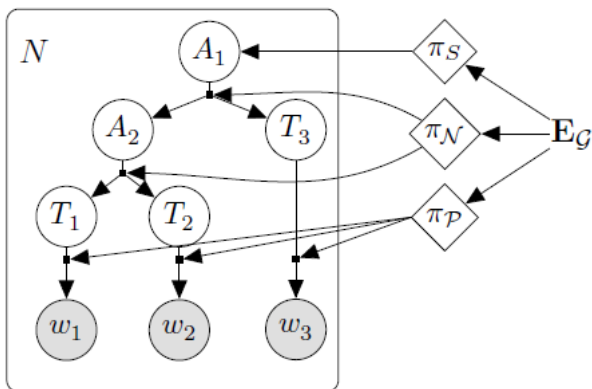


D-N-DMV objective:  $J(\theta) = \sum_y P_\theta(x, y | v_w)$  or ELBO

Han, et al., Enhancing Unsupervised Generative Dependency Parser with Contextual Information (ACL 2019)

# 生成式模型

- ▶ Compound PCFG ← 相似的方法，用于PCFG
  - ▶ 使用terminal/nonterminal embedding输入神经网络计算PCFG规则概率
  - ▶ 使用LSTM+Gaussian采样出句子embedding，输入到神经网络，从而影响规则概率的计算
  - ▶ 训练目标：ELBO



# 自编码器模型

---

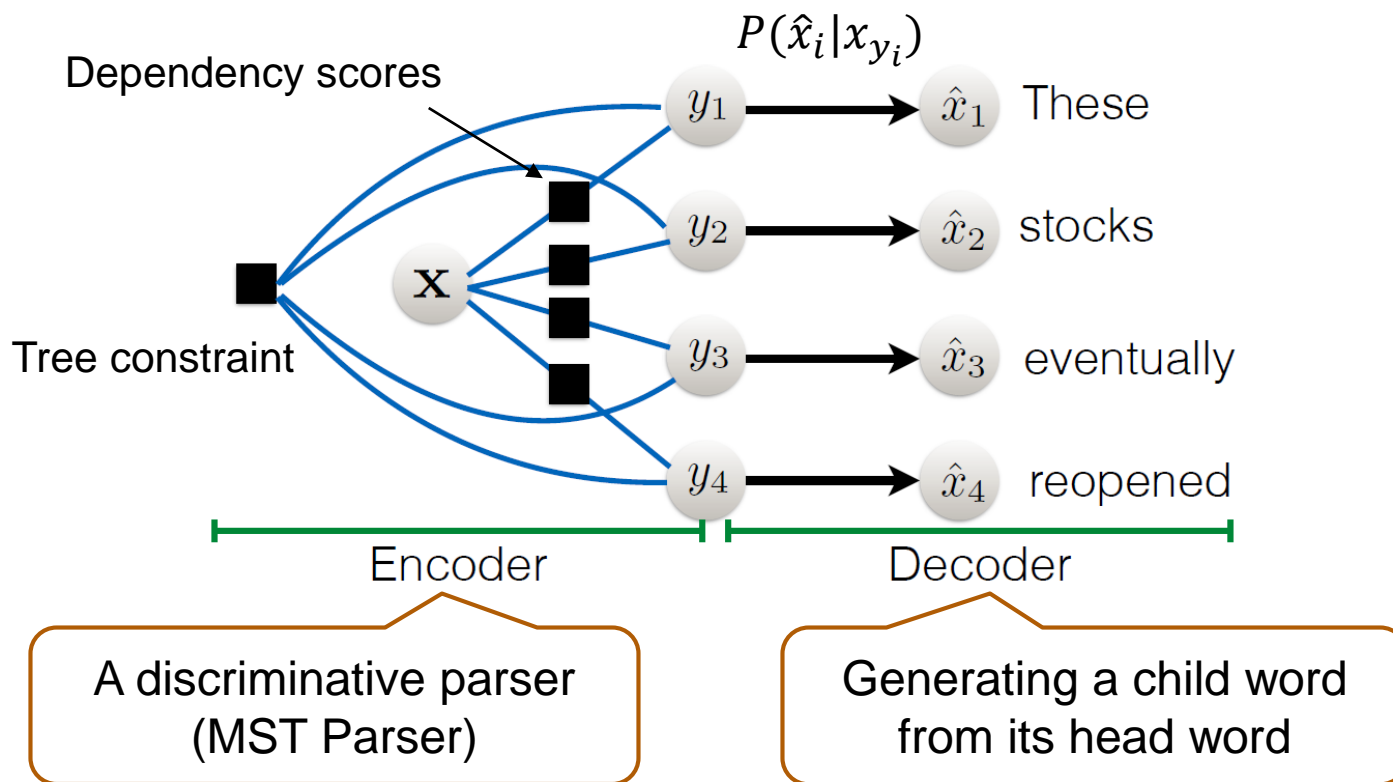
- ▶ 编码器：句子  $\rightarrow$  句法树
- ▶ 解码器：句法树  $\rightarrow$  句子
- ▶ 训练目标：
  - ▶ 重构输入
  - ▶ ELBO  $\leftarrow$  更常见



# 自编码器模型

## ► CRF-AE

- 训练目标：重构输入的概率

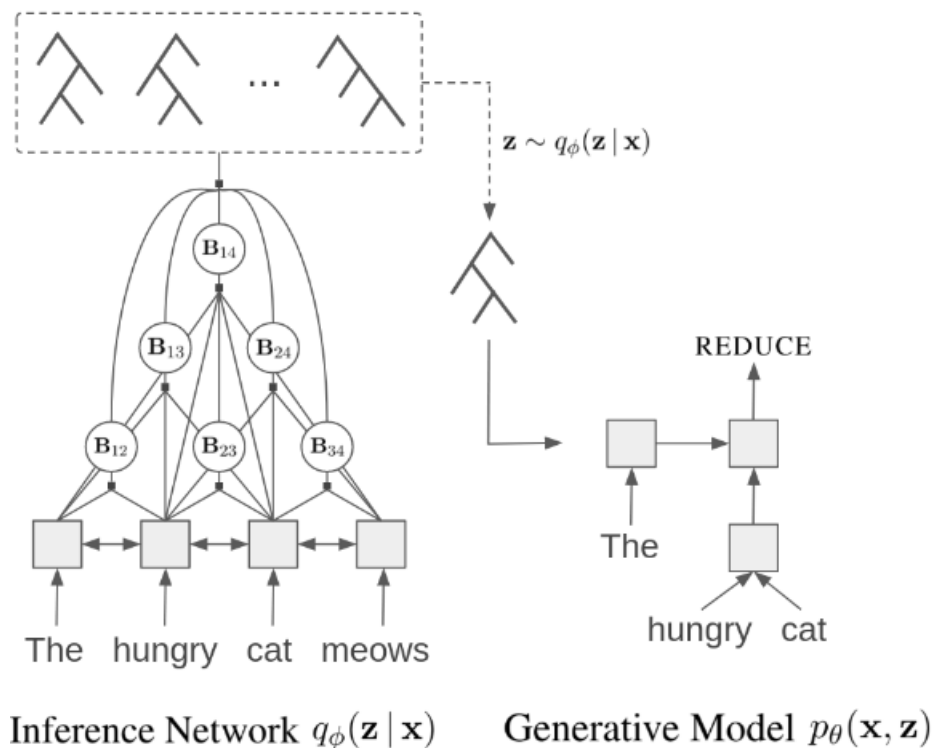




# 自编码器模型

## 基于recurrent neural network grammar (RNNG)的方法

- ▶ 编码器：  
discriminative RNNG  
或 CRF-parser
- ▶ 解码器：  
generative RNNG
- ▶ 训练目标：ELBO

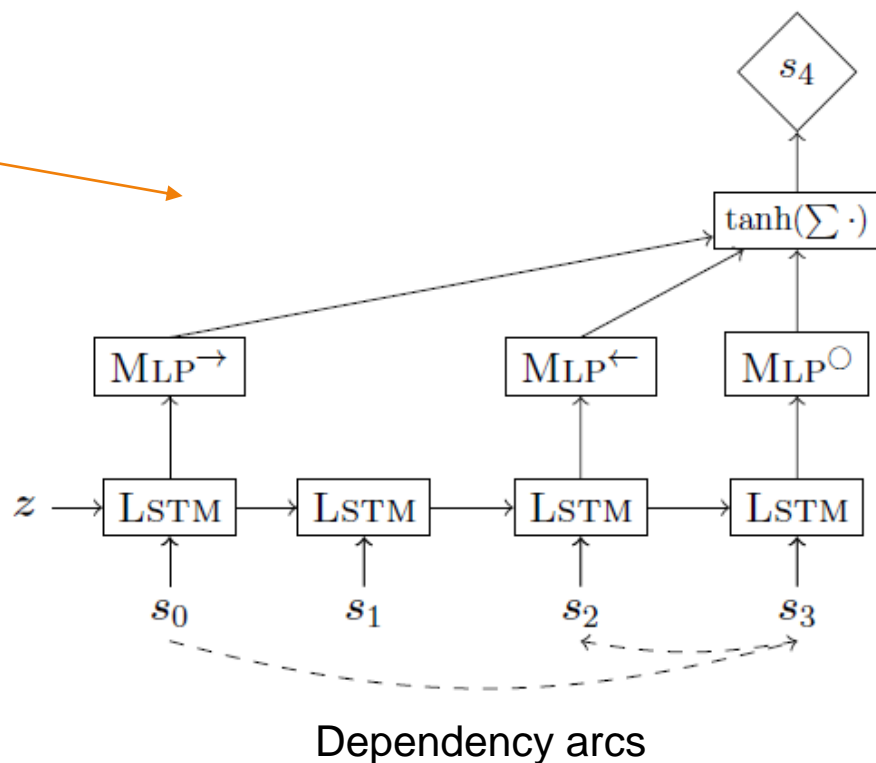


Li, et al., Dependency Grammar Induction with a Neural Variational Transition-based Parser (AAAI 2019)

Kim, et al., Unsupervised Recurrent Neural Network Grammars (NAACL 2019)

# 自编码器模型

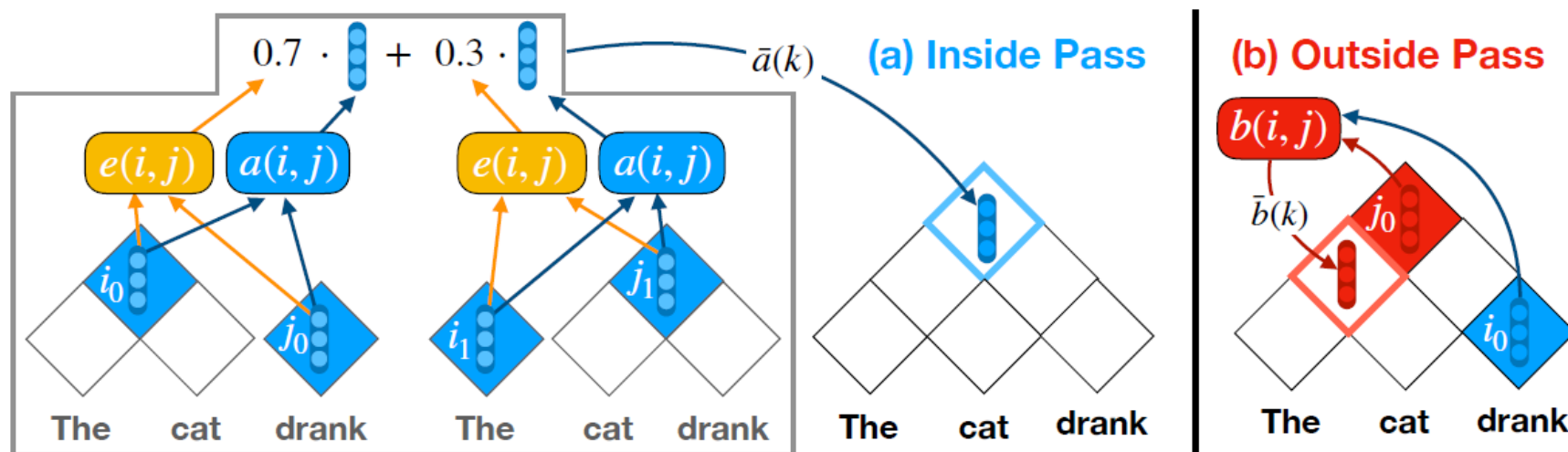
- ▶ 半监督依存分析
  - ▶ 编码器：CRF-parser
  - ▶ 解码器：GCN
  - ▶ 训练目标：ELBO



# 自编码器模型

## ▶ DIORA

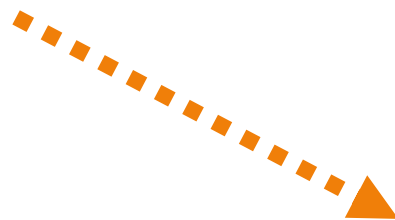
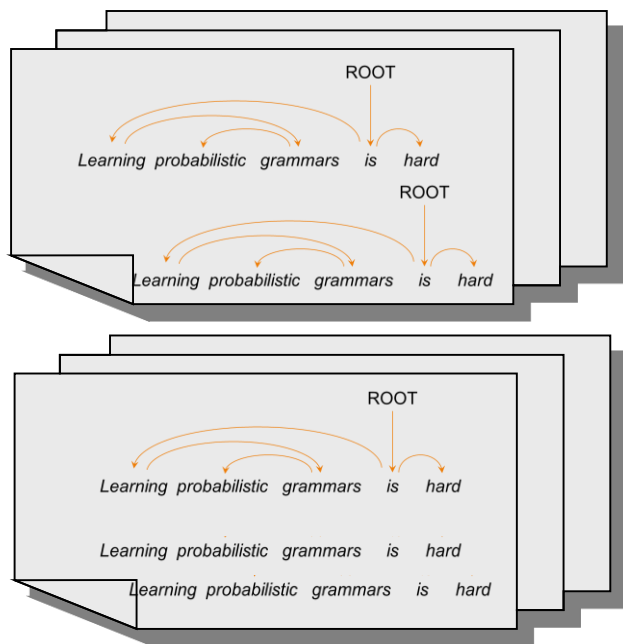
- ▶ 给定一个递归神经网络，通过类似于inside-outside的过程，计算每个词上下文的embedding，用之预测每个词
- ▶ 训练目标：最大化预测准确率



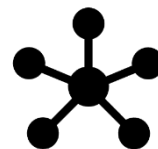
Drozdo, et al., Unsupervised Latent Tree Induction with Deep Inside-Outside Recursive Auto-Encoders (NAACL 2019)

# 跨领域、跨语言句法分析

## 多领域、多语言训练语料



句法分析器



- ▶ 实用意义巨大
- ▶ 论文数量与前两个方向平分秋色
- ▶ 评测：跨领域依存句法分析@NLPCC 2019

# 跨领域、跨语言句法分析

---

## ▶ 近期技术

- ▶ 合并训练集 (baseline)
- ▶ 多任务学习
- ▶ Domain/language embedding
- ▶ 模型参数捆绑
  - ▶ 可基于Phylogenetic Tree
- ▶ 使用大规模语料预训练的Contextual Word Embeddings (BERT, Multilingual BERT)

## ▶ 困难：各语言的词序不同

- ▶ 对于差异较大的语言，可使用对词序不敏感的模型
  - ▶ 利用typological resource构建特征
  - ▶ 可修改源语言树库，使之更匹配目标语言词序
- 



# 总结

---

- ▶ 研究现状
  - ▶ 有监督句法分析
    - ▶ 精度上升空间不大
    - ▶ 分析性研究 + 另辟蹊径
  - ▶ 无监督句法分析
    - ▶ 热度上升，成分分析 > 依存分析
    - ▶ 仍有较大探索空间
  - ▶ 跨领域、跨语言句法分析
    - ▶ 巨大的实用意义





*Thank you!*

