自然语言处理

2022年秋季

黄河燕、鉴萍

北京理工大学 计算机学院

hhy63, pjian@bit.edu.cn

NLP中的序列评估与序列标注

黄河燕，鉴萍
北京理工大学 计算机学院
hhy63, pjian@bit.edu.cn

大纲

- 序列评估
  - 语言模型
  - 汉语分词
- 序列标注

汉语分词

- 汉语分词(Chinese word segmentation,CWS)
  - 曾经是中文信息处理最重要的任务之一
  - 典型的序列评估问题，直接用LM即可解
  - 可以转化为序列标注问题
  - 存在有效的无监督方法
  - 核心内容：消除分词歧义、OOV(out of vocabulary, 未登录词)

汉语分词

- 介绍以下方法
  - 基于词典的机械分词法
  - 基于n-gram的方法
  - 基于字标注的方法
  - 基于理解的方法
  - 深度神经网络方法

汉语分词

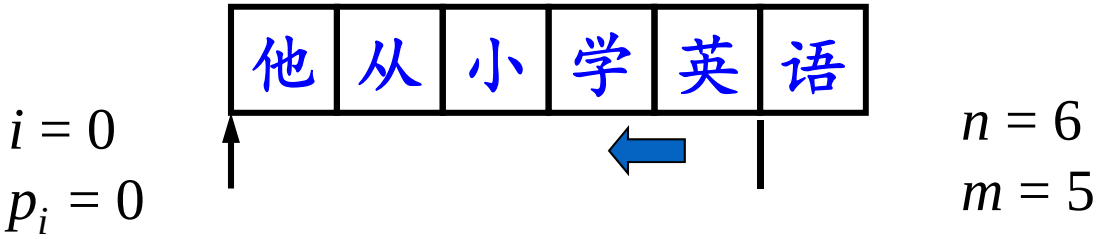
- 基于词典的机械分词方法
  - 最大匹配(maximum matching)
  - 最短路径(shortest-paths)
  - 全切分(omni-segmentation)

基于词典的机械分词

最大匹配

- 举例
  - 输入：他从小学英语
  - 词典：他，从小，从，小学，小，英语

(假设词典中最长词的长度为5)

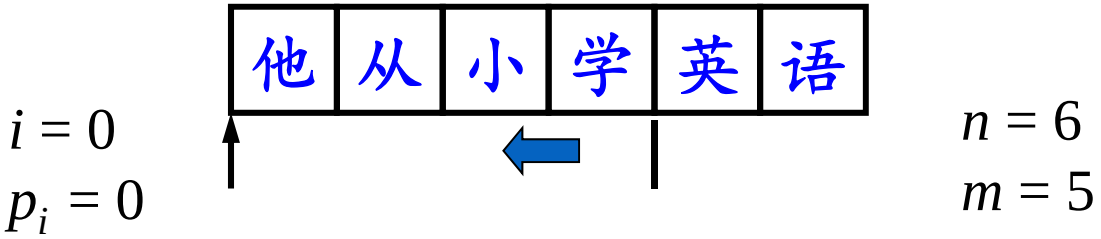


基于词典的机械分词

最大匹配

- 举例
  - 输入：他从小学英语
  - 词典：他，从小，从，小学，小，英语

(假设词典中最长词的长度为5)

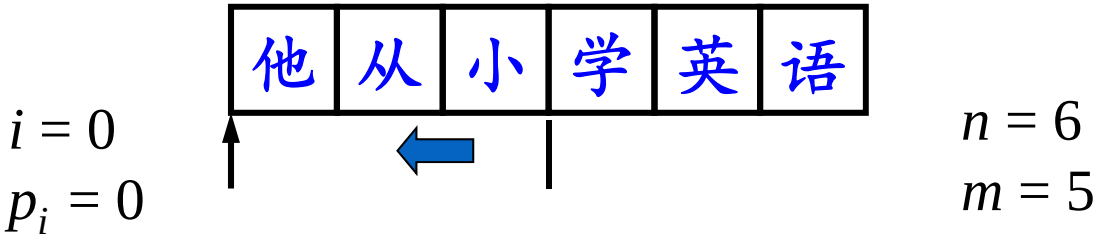


基于词典的机械分词

最大匹配

- 举例
  - 输入：他从小学英语
  - 词典：他，从小，从，小学，小，英语

(假设词典中最长词的长度为5)

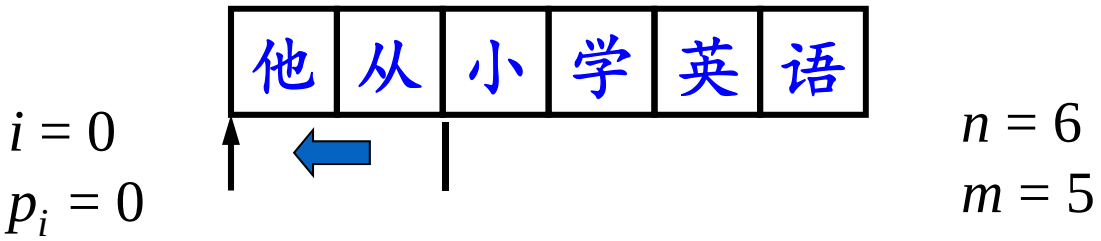


基于词典的机械分词

最大匹配

- 举例
  - 输入：他从小学英语
  - 词典：他，从小，从，小学，小，英语

(假设词典中最长词的长度为5)



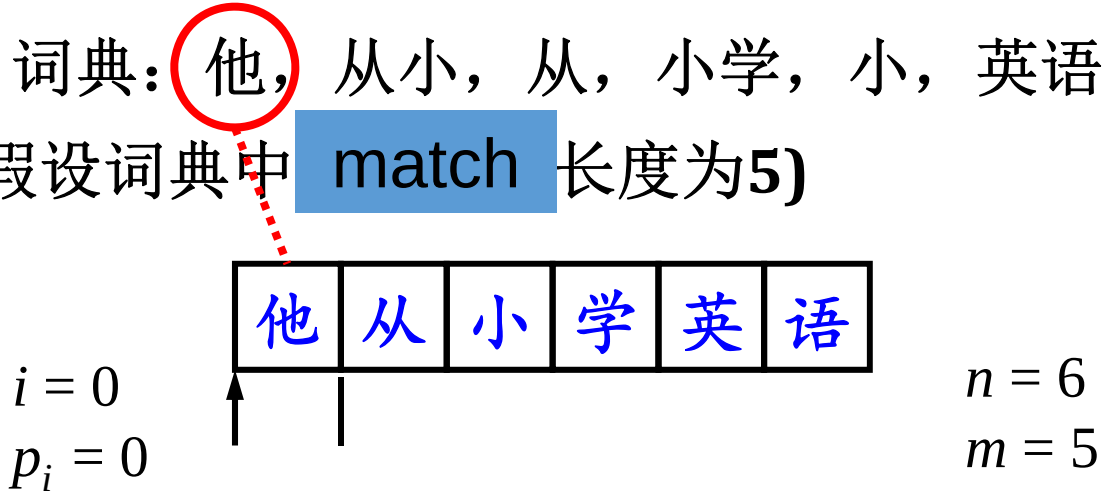
基于词典的机械分词

最大匹配

举例

- 输入：他从小学英语
- 词典：他，从小，从，小学，小，英语

(假设词典中 match 长度为5)

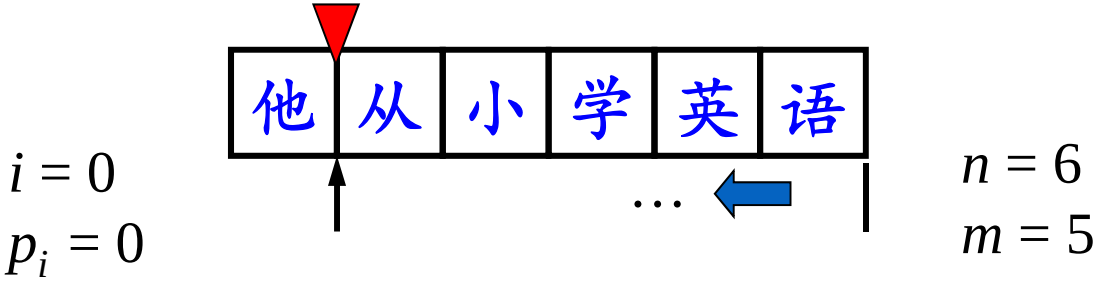


基于词典的机械分词

最大匹配

- 举例
  - 输入：他从小学英语
  - 词典：他，从小，从，小学，小，英语

(假设词典中最长词的长度为5)

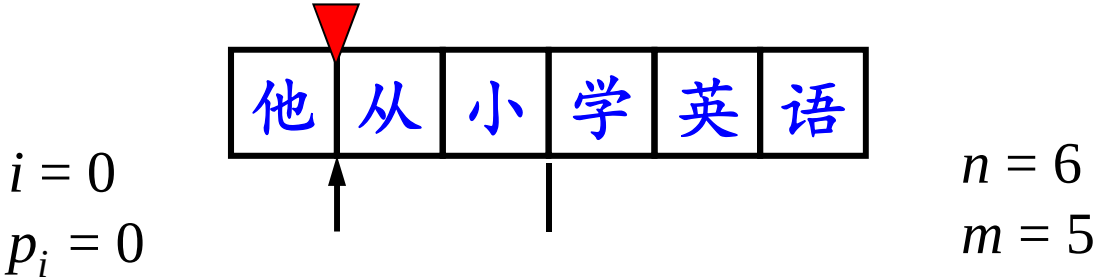


基于词典的机械分词

最大匹配

- 举例
  - 输入：他从小学英语
  - 词典：他，从小，从，小学，小，英语

(假设词典中最长词的长度为5)



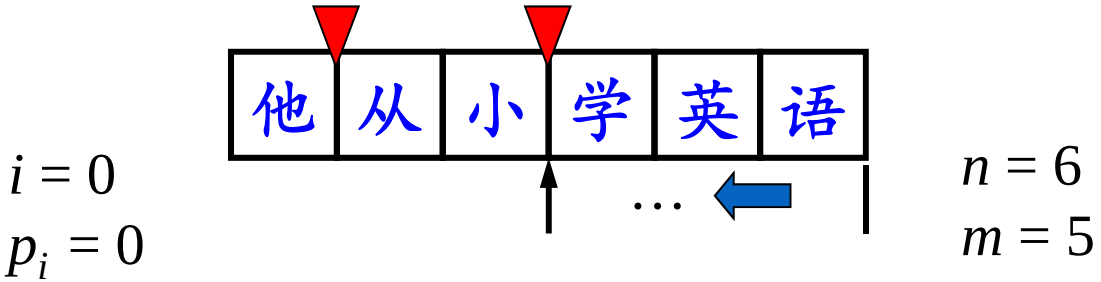
基于词典的机械分词

最大匹配

举例

- 输入：他从小学英语
- 词典：他，从小，从，小学，小，英语

(假设词典中最长词的长度为5)

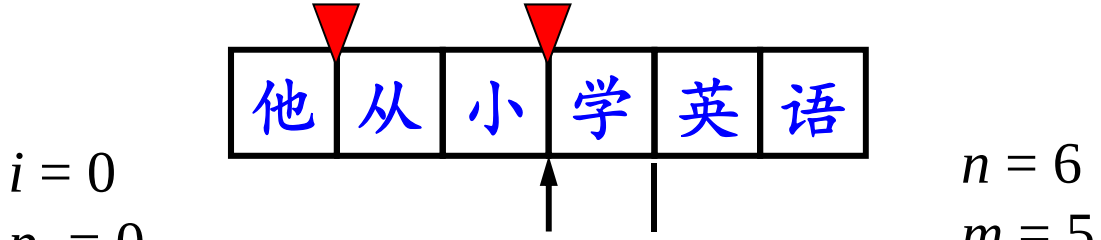


基于词典的机械分词

最大匹配

- 举例
  - 输入：他从小学英语
  - 词典：他，从小，从，小学，小，英语

(假设词典中最长词的长度为5)



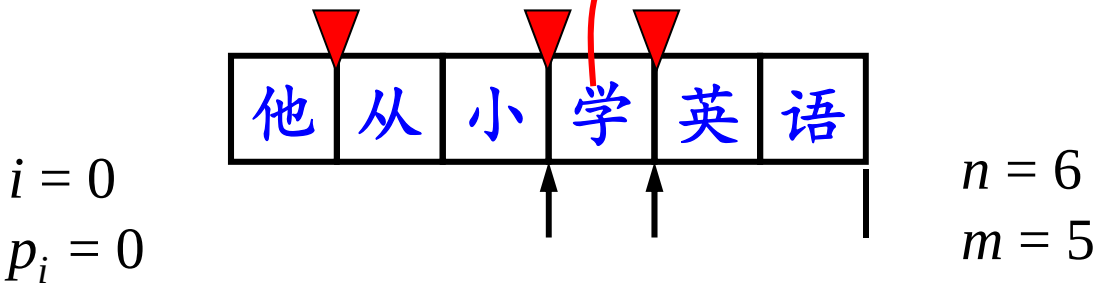
基于词典的机械分词

最大匹配

举例

- 输入：他从小学英语
- 词典：他，从小，从，小学，小，英语，学

(假设词典中最长词的长度为5)



发现缩减到最小单位——一个字，依然没有和词典里匹配的词，则将该单字词作为新发现的词加入词典

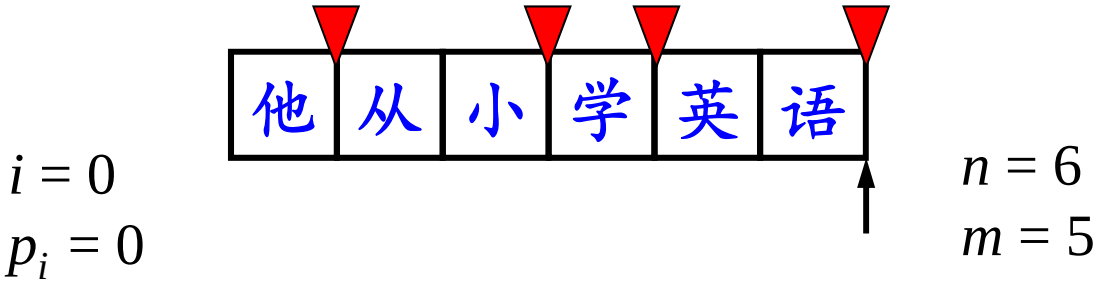
基于词典的机械分词

最大匹配

举例

- 输入：他从小学英语
- 词典：他，从小，从，小学，小，英语，学

(假设词典中最长词的长度为5)



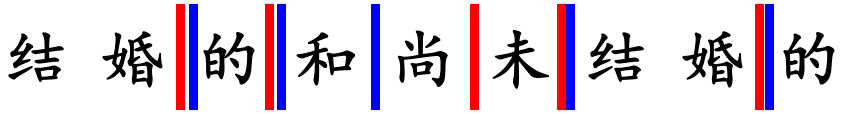
FMM results： 他 从小 学 英语

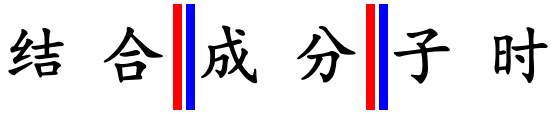
BMM results：他 从 小学 英语

基于词典的机械分词

- 单向最大匹配忽略交叉和组合歧义，双向最大匹配好一些

- 双向最大匹配无法排除歧义链长为偶数的交叉歧义和组合歧义





基于词典的机械分词

MM方法的歧义检测能力

- 汉语文本中90.0%左右的句子，FMM和BMM的切分完全重合且正确；
- 9.0%左右的句子FMM和BMM的切分不同，但其中必有一个是正确的(歧义检测成功)；
- 只有不到1.0%的句子，
  - 或者**FMM**和**BMM**的切分虽重合但却是错的
  - 或者**FMM**和**BMM**切分不同但两个都不对**(**歧义检测失败**)**

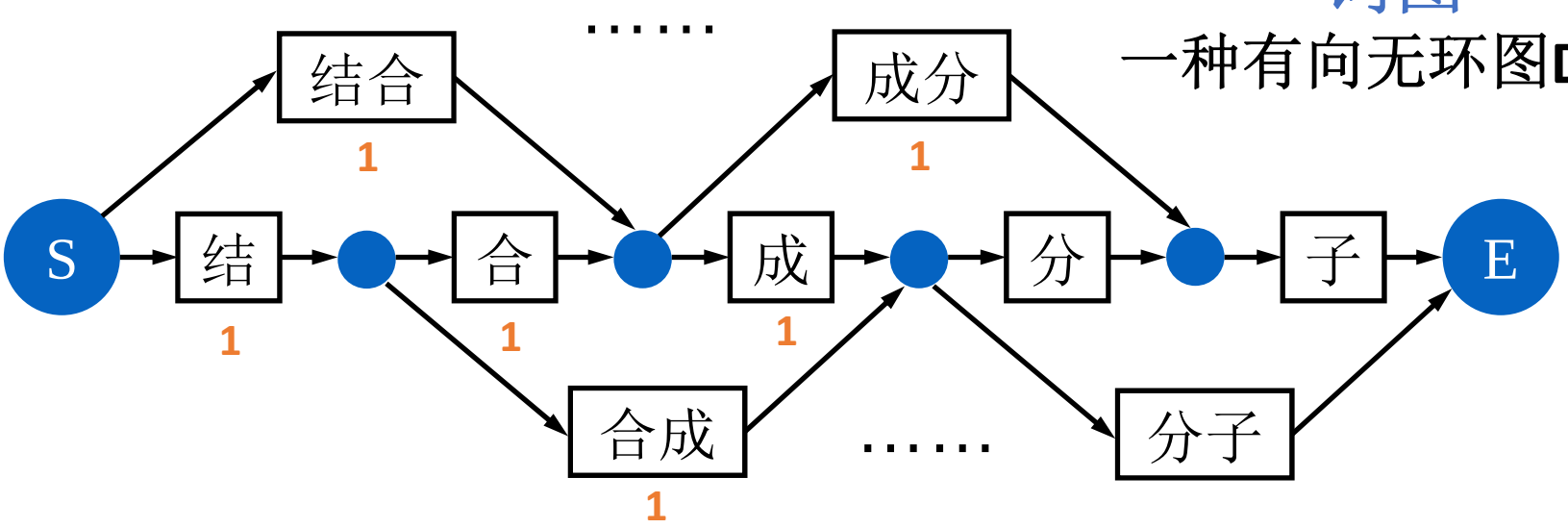
基于词典的机械分词

最短路径法

- 选择包含词最少的切词路径

词图

一种有向无环图**DAG**

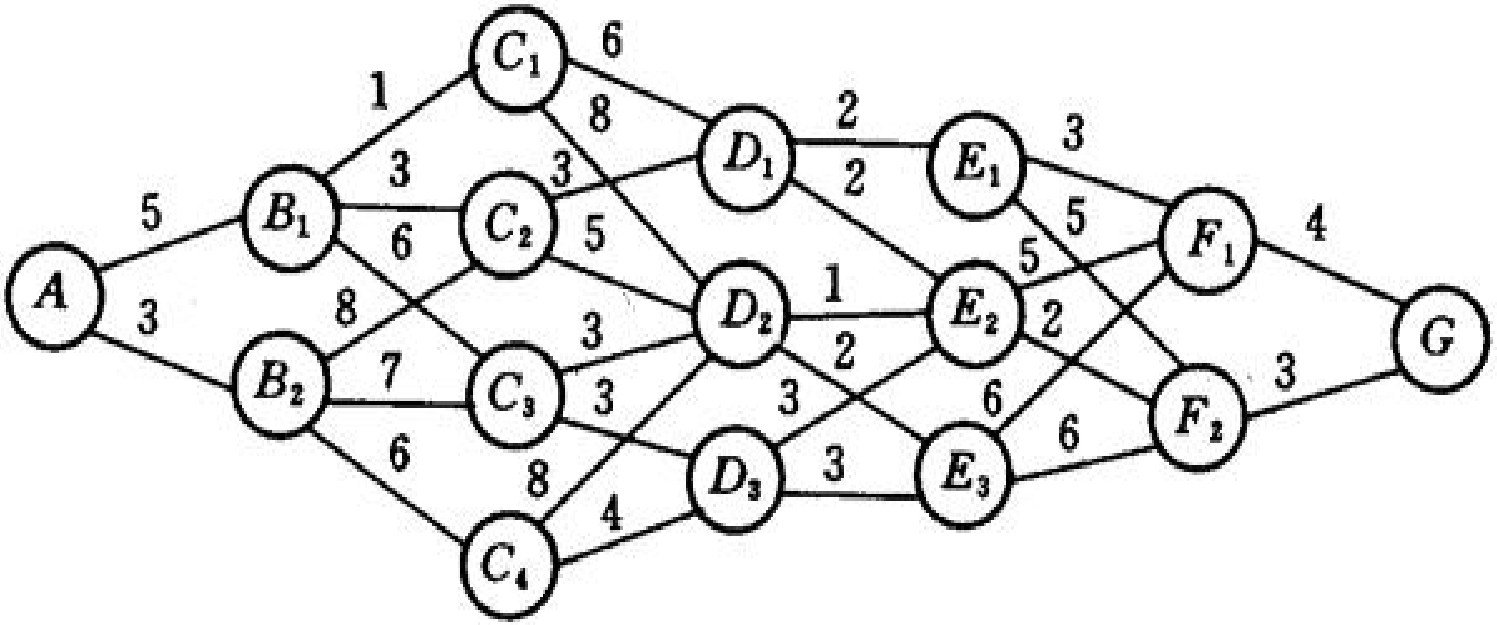


- 算法：Dijkstra算法

  动态规划(dynamic programming)

基于词典的机械分词

- ✓ 补充：动态规划搜索最短路径



$$S_{AE_2} = Shortest\{(S_{AD_1} + D_{D_1E_2}), (S_{AD_2} + D_{D_2E_2}), (S_{AD_3} + D_{D_3E_2})\}$$

基于词典的机械分词

最短路径法

- 切分原则符合汉语自身规律(节约原则)
- 不能发现所有的组合歧义
- 最短路径可能有多条

  结　合成　分子

  结合　成　分子

  结合　成分　子
- 低效

基于词典的机械分词

全切分

- 切出所有的组合，再用n-gram求概率最大
- 没有盲区，但会切出很多垃圾

汉语分词

- 介绍以下方法
  - 基于词典的机械分词法
  - 基于n-gram的方法
  - 基于字标注的方法
  - 基于理解的方法
  - 深度神经网络方法

基于n-gram模型

基于n-gram

- 输入：$C = c_1 c_2 \ldots c_l$
- 输出：$W = w_1 w_2 \ldots w_m$
- 根据Bayes法则：

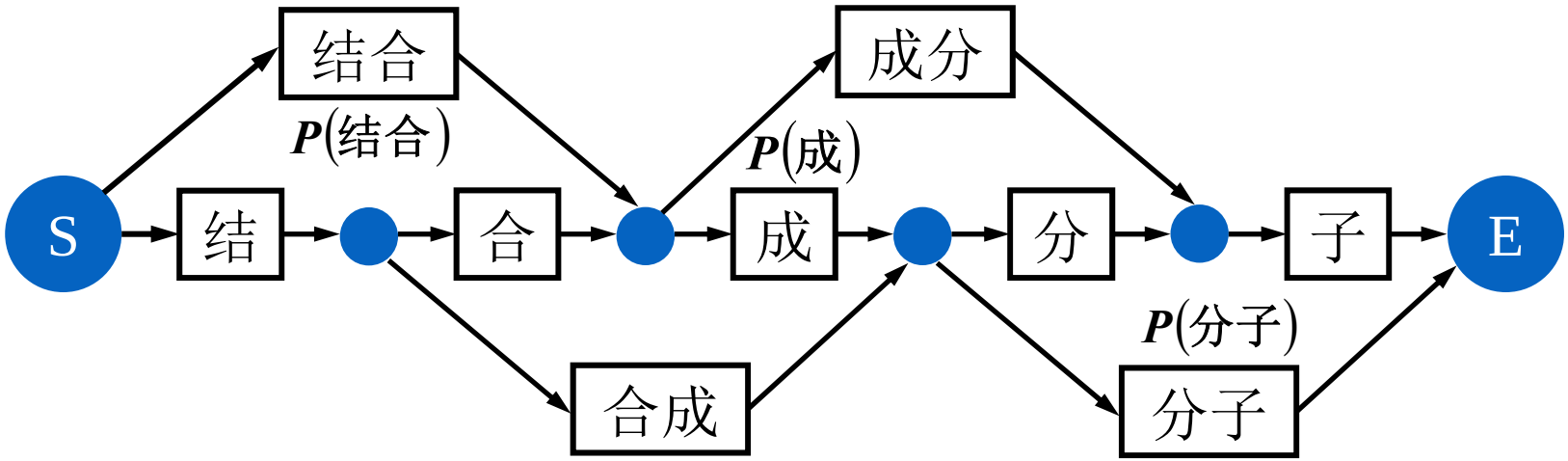
$$W^* = \arg\max_W P(W|C)$$

$$= \arg\max_W \frac{P(C|W)P(W)}{P(C)}$$

$$= \arg\max_W P(W)$$

基于n-gram模型

- 1-gram切分词图

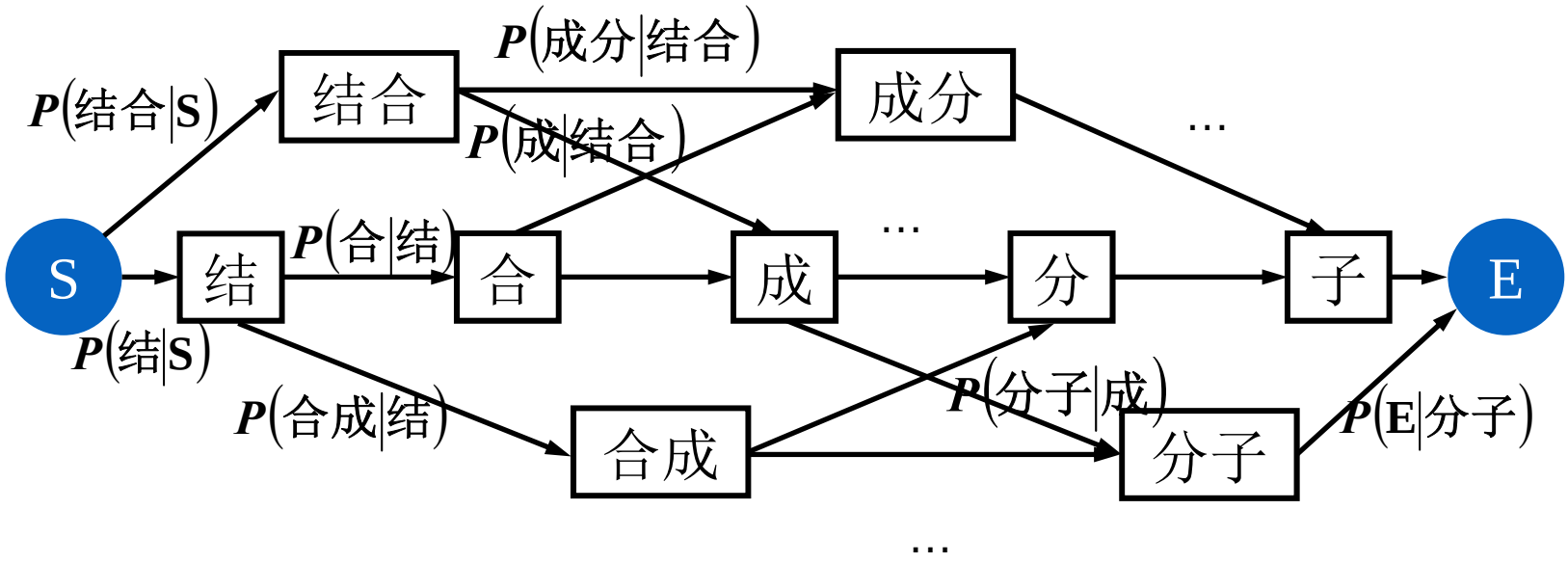


$$P(\text{结合 成 分子}) = P(\text{结合})P(\text{成})P(\text{分子})$$

概率统计方法，需要标注了词边界的语料

基于n-gram模型

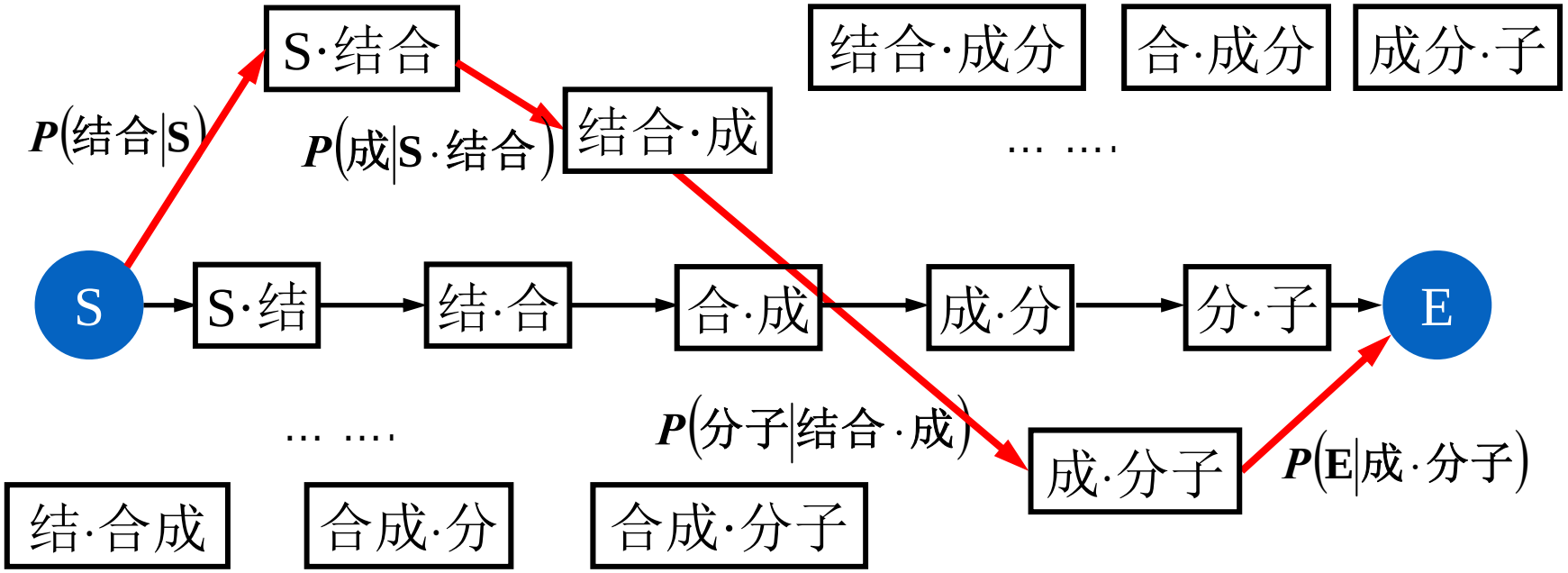
- 2-gram切分词图



$$P(\text{结合 成 分子}) = P(\text{结合}|\text{S})P(\text{成}|\text{结合})P(\text{分子}|\text{成})P(\text{E}|\text{分子})$$

基于n-gram模型

- 3-gram切分词图



$$P(\text{结合 成 分子}) = P(\text{结合}|\text{S})P(\text{成}|\text{S}\cdot\text{结合})P(\text{分子}|\text{结合}\cdot\text{成})P(\text{E}|\text{成}\cdot\text{分子})$$

基于n-gram模型

基于n-gram

- 基于n-gram模型，分词可以看作是分词图上的最佳路径搜索
- 算法可以用动态规划
- 但不能识别OOV

汉语分词

- 介绍以下方法
  - 基于词典的机械分词法
  - 基于n-gram的方法
  - 基于字标注的方法　如何转化为一个序列标注问题？
  - 基于理解的方法
  - 深度神经网络方法

基于字标注

将一个NLP问题转化为序列标注问题——一种重要的思路

基于字标注

- 将分词看作字的分类，转化为序列标注问题

  结 合 成 分 子

  B E S B E

- 平衡看待词表词和OOV
- 可以采用典型序列标注模型
  - **Cascade models：SVM, ME, Perceptron (**每个位置上确定性地分类**)** 会在下一节序列标注中进一步介绍
  - **Integrated models： HMM, MEMM, CRF**

汉语分词

- 介绍以下方法
  - 基于词典的机械分词法
  - 基于n-gram的方法
  - 基于字标注的方法
  - 基于理解的方法
  - 深度神经网络方法

基于理解的方法

- 切分歧义的排除有时需要更深层的语言知识的支撑
  - POS

    他俩儿谈恋爱是从头年元月开始的

    segmentation a: **... 是 从头 年 元月 ...**

    VC AD M NT

    segmentation b: **... 是 从 头年 元月 ...**

    VC P NT NT

基于理解的方法

- 切分歧义的排除有时需要更深层的语言知识的支撑
  - 句法结构

    什么时候我才能克服这个困难？

    Segmentation a:

    什么　时候　我　才　能　克服　这个　困难

    Segmentation b:

    什么　时候　我　才能　克服　这个　困难

切分**b**得不到可信的句法树，因此被拒绝

基于理解的方法

- 切分歧义的排除有时需要更深层的语言知识的支撑
  - 语义搭配

    他学会了解数学难题

    Segmentation a:

    他　学会　了　解　数学　难题

    Segmentation b:

    他　学会　了解　数学　难题

“解”需要“题目”或“钮扣”作为宾语，所以选**a**

基于理解的方法

- 歧义的排除有时需要更深层语言知识的支撑
  - 语用

    今天做核酸的队长死了。

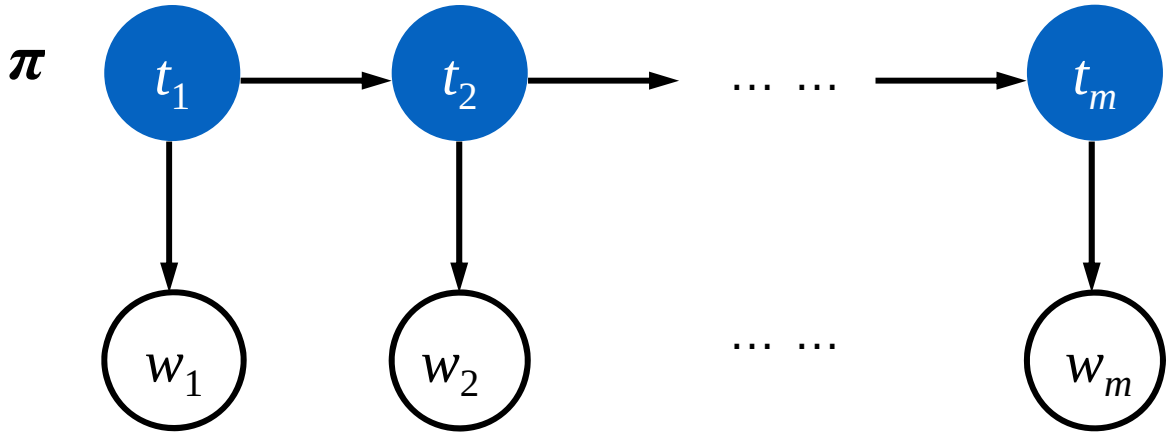
“95%的左右的切分歧义可以通过句法或句法以下的知识来解决，只有5%必须依仗语义和语用知识。”

请理解：并不是说必须用XXX信息才能分词，你用n-gram依然可以——“知识”来自于你的语料，但你可能做不对，结果可能是无法解释的

基于理解的方法并不常用

汉语分词

- 举例：一个分词、POS标注联合模型(joint model)



词序列：$W = w_1 w_2 \ldots w_m \quad m \geq 1$

POS序列：$W = t_1 t_2 \ldots t_m \quad m \geq 1$

汉语分词

1. 先固定一个词串(一个分词候选)
2. 这时，这是一个HMM模型
3. 基于观测值(固定的这个词串)，推导最优状态序列(最优POS串)
4. 根据这个最优POS串的概率来计算候选分词路径的概率，选最大的那个

$$T_k^* = \underset{T}{\operatorname{argmax}}\, P(T|W_k)$$

$$= \underset{T}{\operatorname{argmax}}\, \frac{P(W_k|T)P(T)}{P(W_k)}$$

$$= \underset{T}{\operatorname{argmax}} \prod_i P(w_i|t_i)P(t_i|t_{i-1})$$

$$P(w_i|t_i) = \frac{f(w_i, t_i)}{f(t_i)}$$

$$P(t_i|t_{i-1}) = \frac{f(t_{i-1}, t_i)}{f(t_{i-1})}$$

$$P(W_k) = \frac{P(W_k|T_k^*)P(T_k^*)}{P(T_k^*|W_k)} = \frac{P(W_k|T_k^*)P(T_k^*)}{\prod_i P(t_i|w_i)}$$

$$W^* = \underset{W_k}{\operatorname{argmax}}\, P(W_k)$$

汉语分词

- 介绍以下方法
  - 基于词典的机械分词法
  - 基于n-gram的方法
  - 基于字标注的方法
  - 基于理解的方法
  - 深度神经网络方法

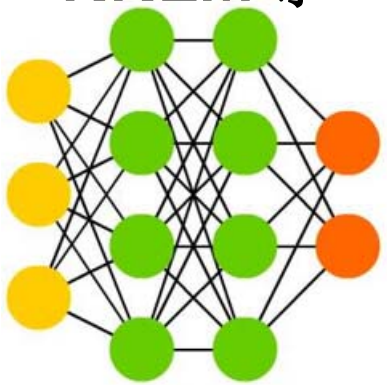
如何将已有概率统计方法转化为**NN**方法？

1. 将**n-gram**用**NN**模型代替；
2. 用基于**NN**的序列标注模型…

NN汉语分词

- 一个NN分词模型

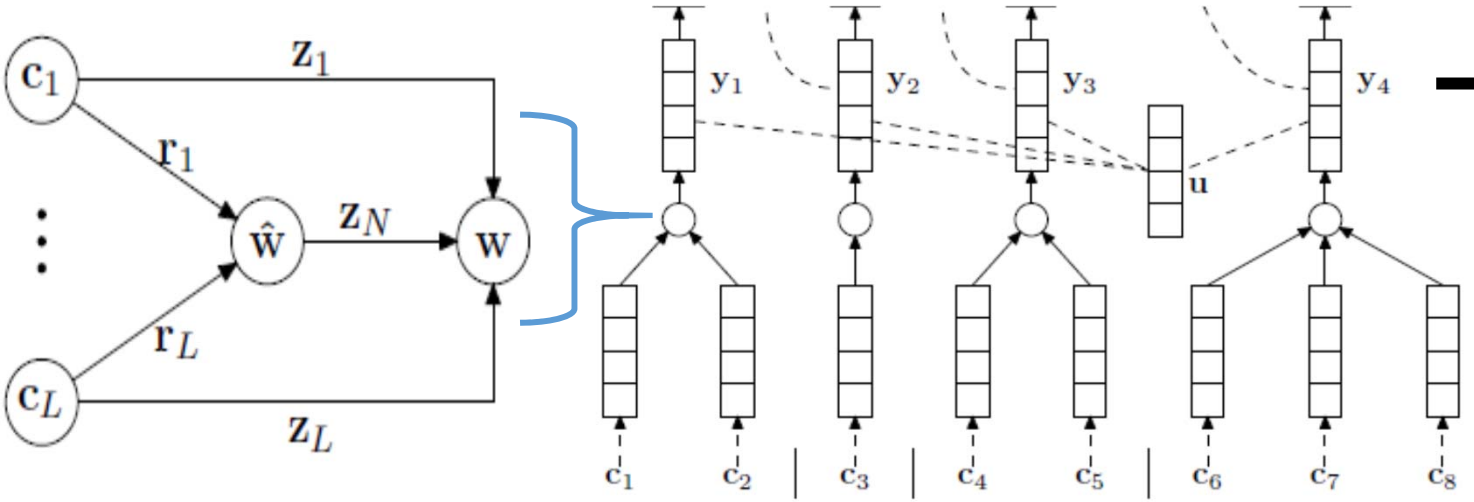
Deng Cai & Hai Zhao (ACL2016)



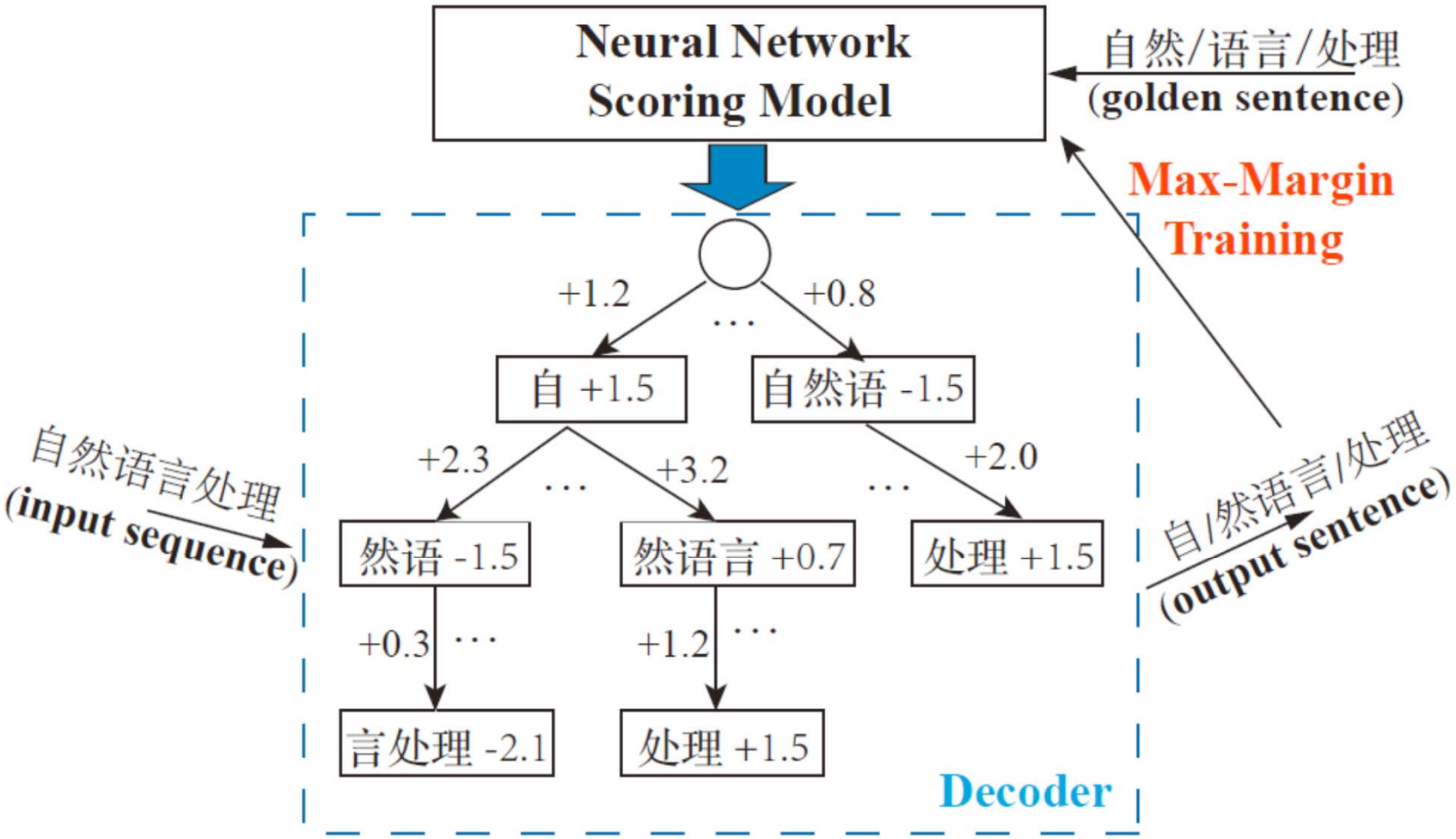
基于**LM**获得外部打分**(linking score)**

Decoding (Beam search)

字组合获得内部打分**(word score)**

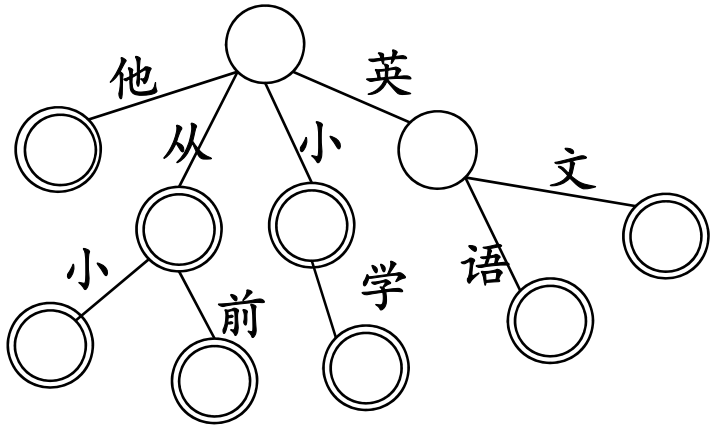


NN汉语分词


Neural Network
Scoring Model
自然/语言/处理
(golden sentence)
Max-Margin
Training
+1.2
+0.8
自 +1.5
自然语 -1.5
+2.3
+3.2
+2.0
然语 -1.5
然语言 +0.7
处理 +1.5
+0.3
+1.2
言处理 -2.1
处理 +1.5
Decoder
自然语言处理
(input sequence)
自/然语言/处理
(output sentence)

✓ Jieba分词的主要技术

1. 构建Trie树，扫描句子构成词图



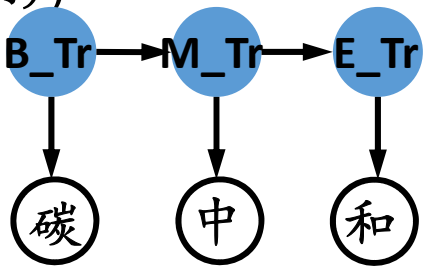
输入：他从小学英语
词图：
{0:[0]; 1:[1,2]; 2:[2,3]; 3:[3]; 4:[5]}

2. 采用动态规划查找最大概率路径, 找出基于词频(1-gram)的最大切分组合

$$P(\text{Node}_n) = 1.0,$$

$$P(\text{Node}_{n-1}) = P(\text{Node}_n) \times max\big(P(\text{最后一个词})\big) \ldots$$

3. 对于OOV，采用基于汉字成词能力的HMM，使用Viterbi算法求最优标注序列

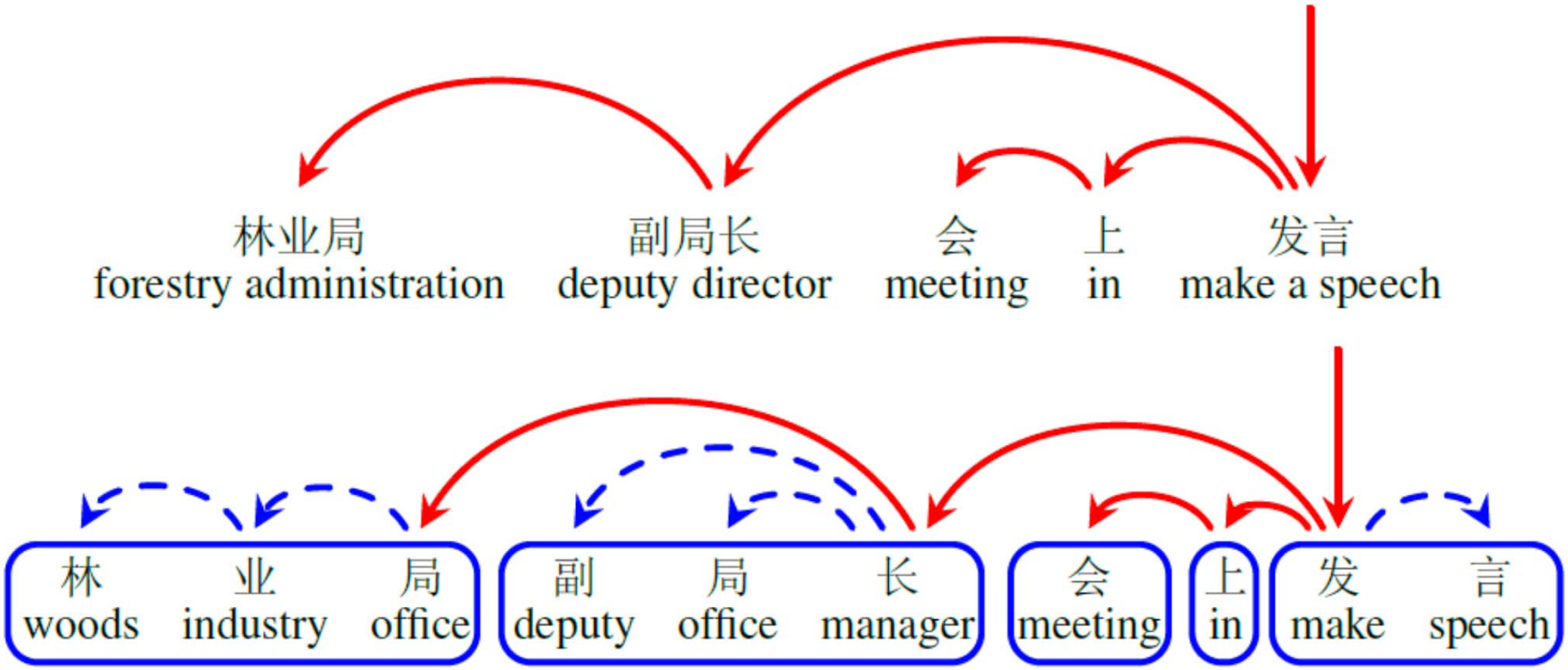


汉语分词

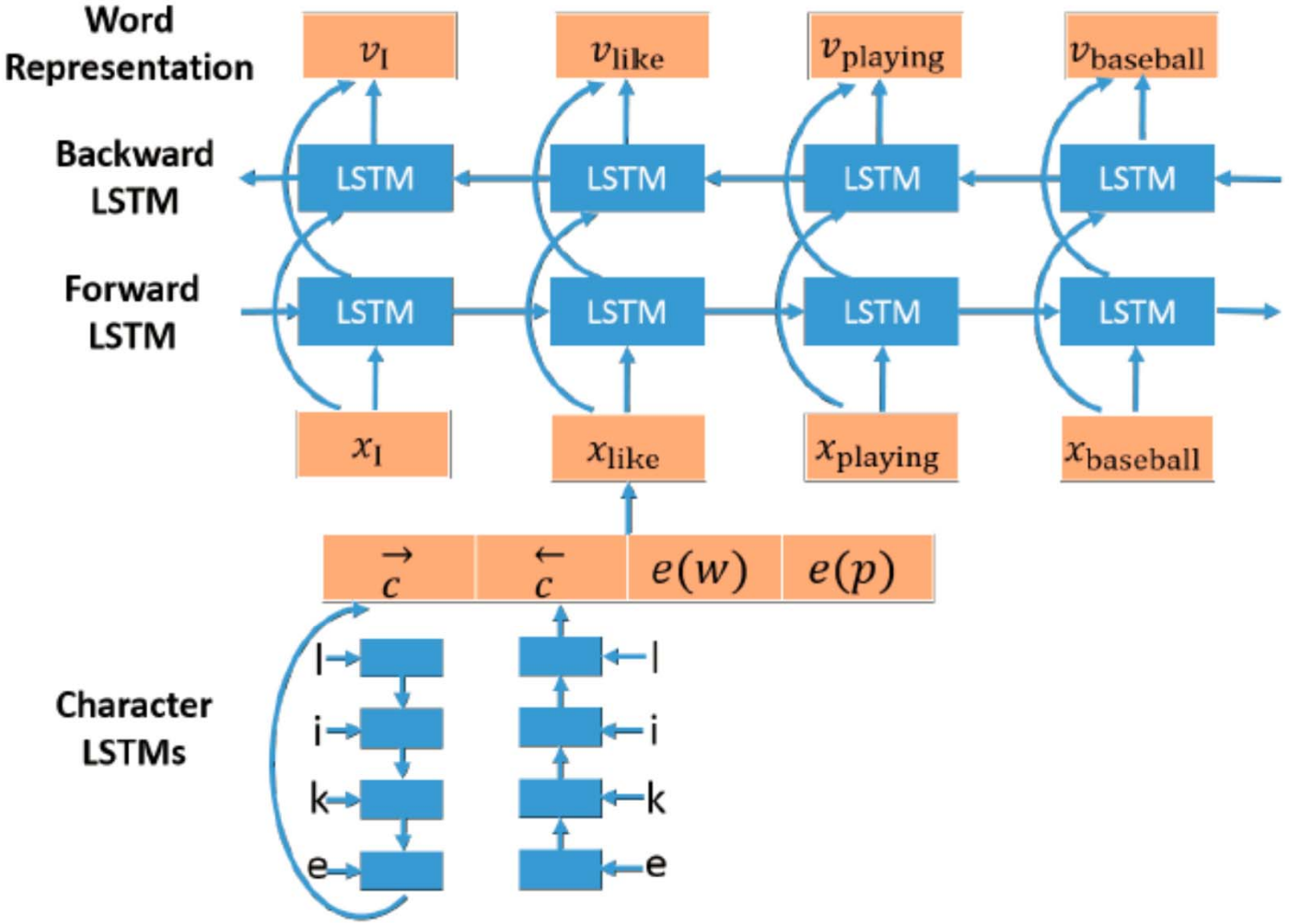
- 分词在现阶段特别是预训练范式下用得少了
- 词法分析的核心工作形态分析(分析词的组成)也只多见于特殊语言或特殊任务
- 有一些工作对词以下(under-word)成分建模

汉语分词

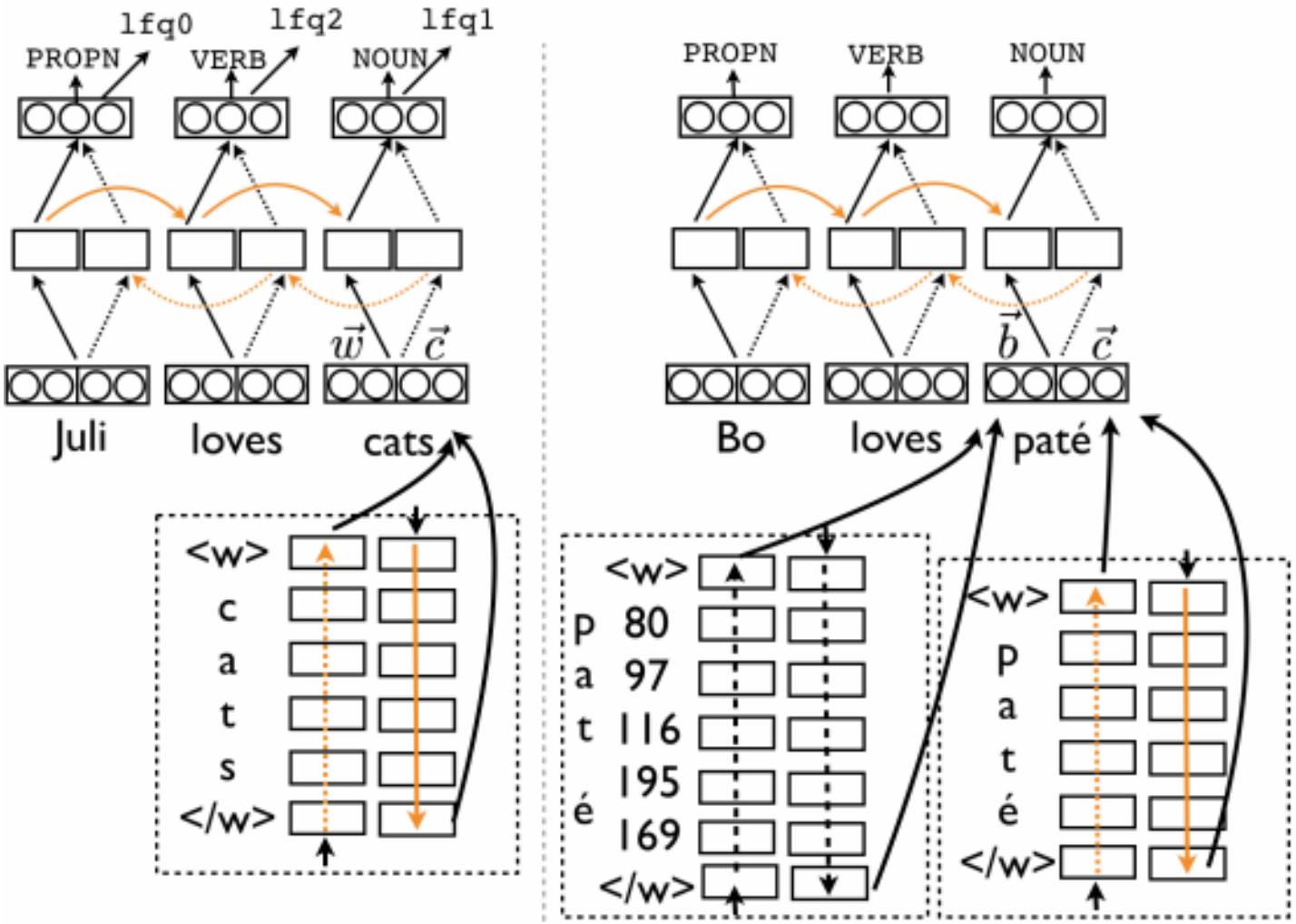
- ✓ Under-word信息在NLP中的应用



汉语分词


Word Representation
Backward LSTM
Forward LSTM
Character LSTMs
v_{I}
v_{like}
v_{playing}
v_{baseball}
LSTM
x_{I}
x_{like}
x_{playing}
x_{baseball}
$\overrightarrow{c}$
$\overleftarrow{c}$
$e(w)$
$e(p)$
I
i
k
e

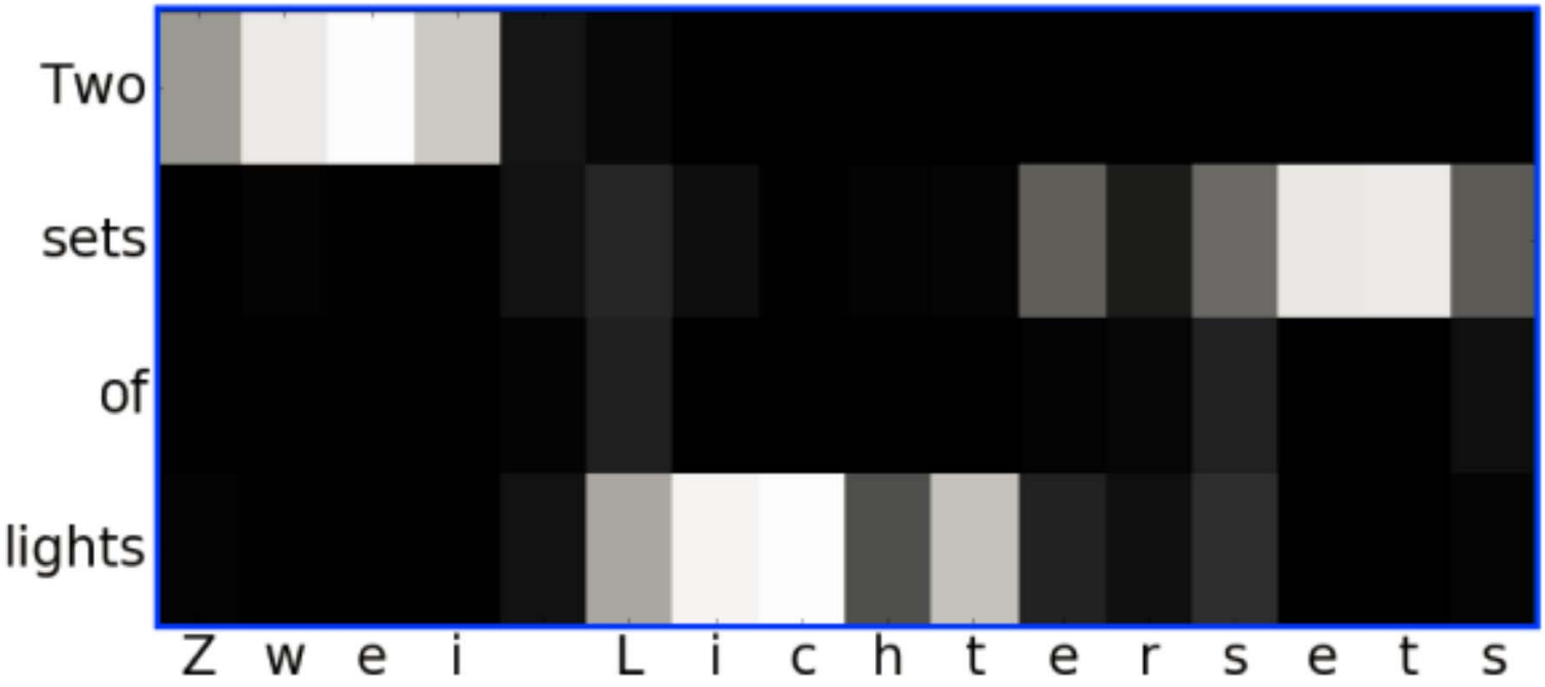
汉语分词



不仅使用了字的**RNN**，还用到了字的**brown**聚类标签的**RNN**

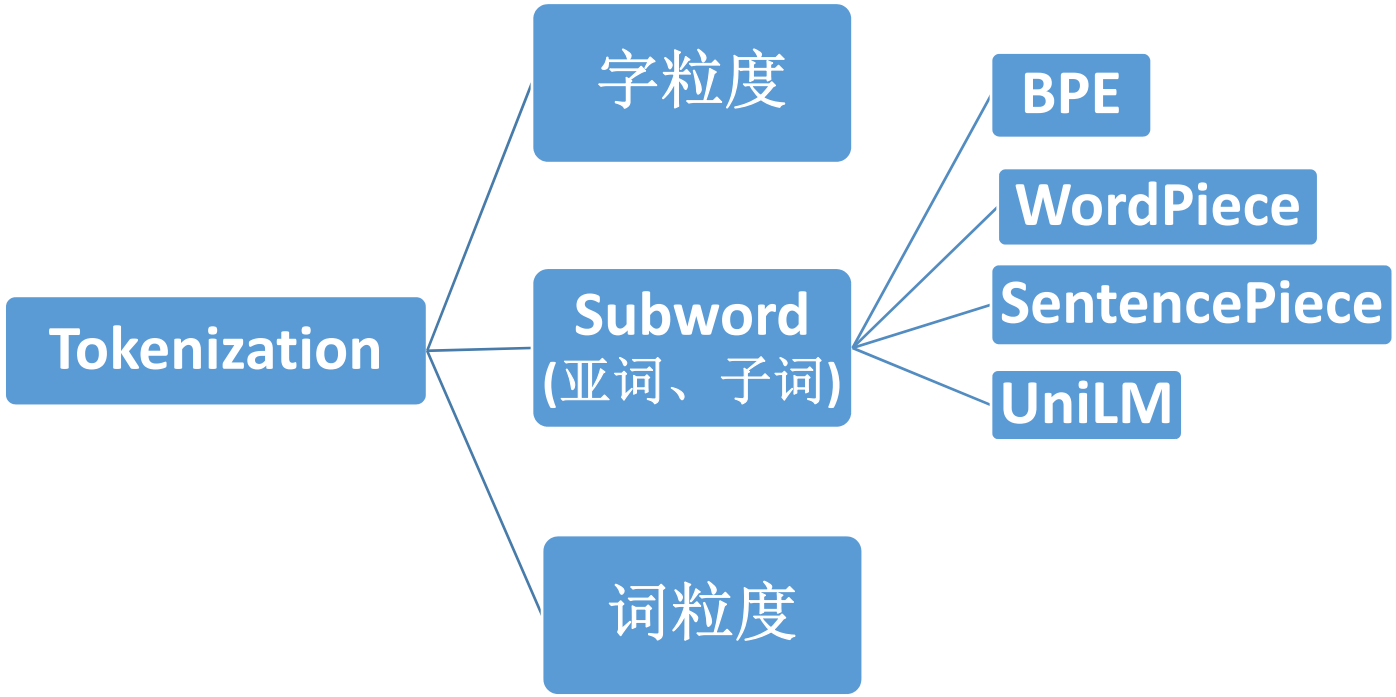
汉语分词

- Character based NMT



汉语分词

- ✓ 就目前的tokenization做总结



汉语分词

- 字粒度
  - 输入：就目前的tokenization做总结。
  - 输出：

    就 目 前 的 t o k e n i z a t i o n 做 总 结 。
- 词粒度
  - 输入：就目前的tokenization做总结。
  - 输出：

    就 目前 的 tokenization 做 总结 。

BPE (byte pair encoding，双字节编码)

- 统计字符对出现的频率，把高频的 char n-gram 当成一个整体输入单位

1. 准备训练语料，并确定期望的**Subword**词表大小；

   语料：**{'low':5, 'lower':2, 'newest':6, 'widest':3}**

2. 拆分，拆分后：**{'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w e s t </w>': 6, 'w i d e s t </w>': 3}**

3. 在语料上统计词内相邻对的频数，选取频数最高的合并成新的**Subword**单元；**{'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w es t </w>': 6, 'w i d es t </w>': 3}**

4. 重复**3**直到达到第**1**步设定的**Subword**词表大小或下一个最高频数为**1**

✓ 对于输入词，按最大匹配法匹配子词并分割

汉语分词

WordPiece

- 与BPE的差别：WordPiece选择可以最大化训练数据可能性的组合

$$\text{argmax} \log \frac{P(\text{xy})}{P(\text{x})P(\text{y})}$$

Subword x 和Subword y 合并后得到 xy

BERT的其中一个Tokenizer即WordpieceTokenizer

目前中文多用字做输入，英文等多用**WordPiece**或**BPE**等

Model	Type of Tokenizer
fast MPNet	WordPiece
PhoBERT	Byte-Pair-Encoding
T5	SentencePiece
fast T5	Unigram
fast MBART	BPE
fast PEGASUS	Unigram
PEGASUS	SentencePiece
XLM	Byte-Pair-Encoding
TAPAS	WordPiece
BertGeneration	SentencePiece
BERT	WordPiece
fast BERT	WordPiece
XLNet	SentencePiece
GPT-2	byte-level Byte-Pair-Encoding
fast XLNet	Unigram
fast GPT-2	byte-level Byte-Pair-Encoding
fast ALBERT	Unigram
ALBERT	SentencePiece
CTRL	Byte-Pair-Encoding
fast GPT	Byte-Pair-Encoding
Flaubert	Byte-Pair Encoding
FAIRSEQ	Byte-Pair Encoding
Reformer	SentencePiece
fast Reformer	Unigram
Marian	SentencePiece

一些预训练模型采用的**tokenize**算法

大纲

- 序列评估
  - 语言模型
  - 汉语分词
- 序列标注

序列标注

Train set:

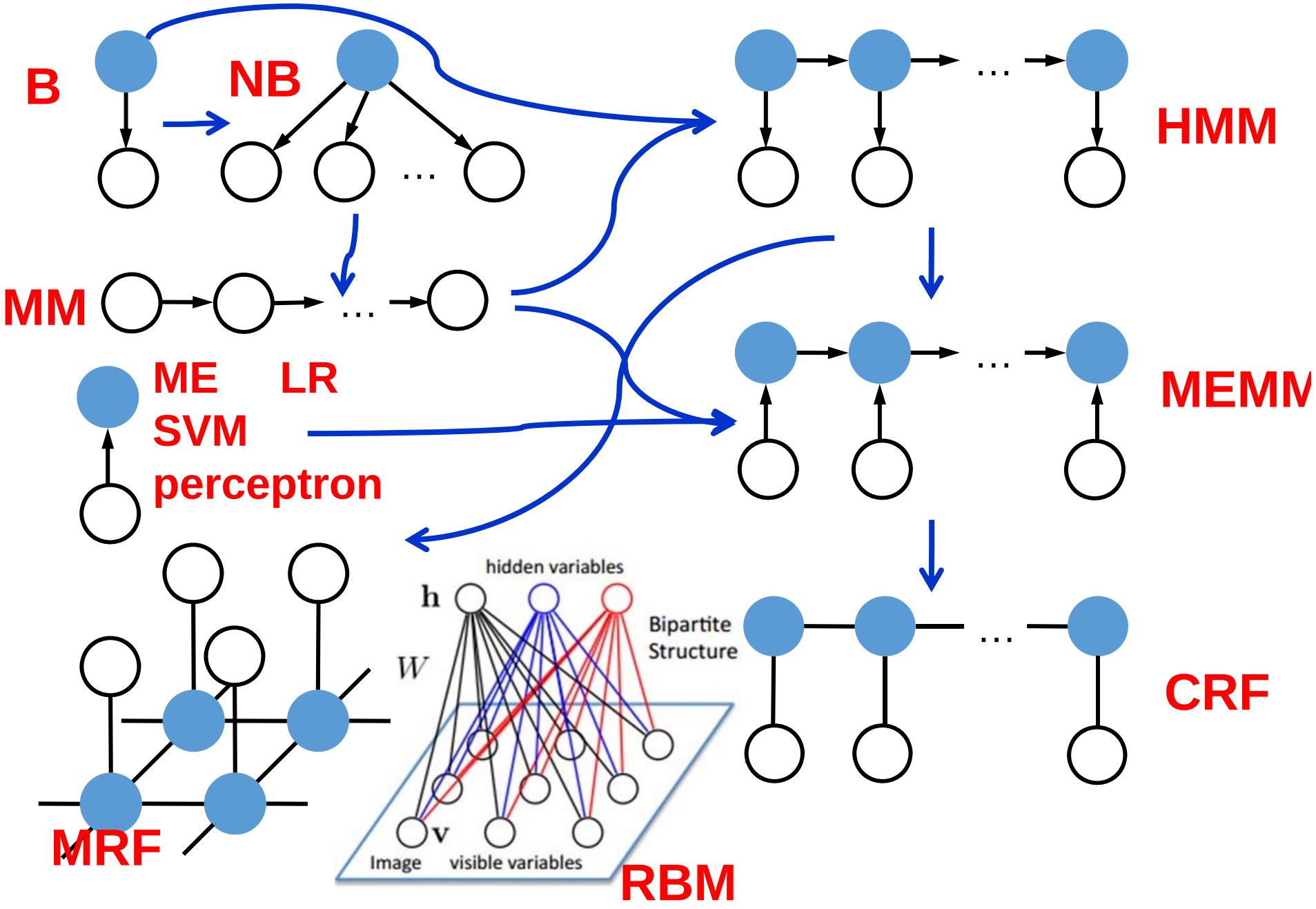
他	PN
同时	AD
希望	VV
秘鲁	NR
侨胞	NN
能	VV
继续	VV
关心	VV
和	CC
支持	VV
…	…

Test set:

这
个
思想
很
重要
…

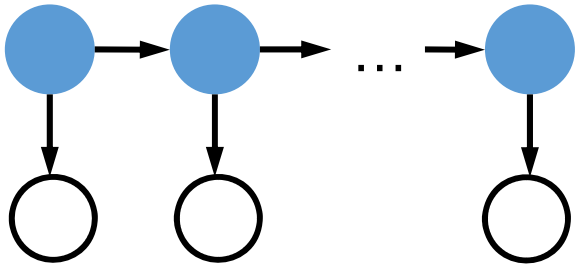
Result:

这	PN
个	M
思想	NN
很	AD
重要	AV
…	

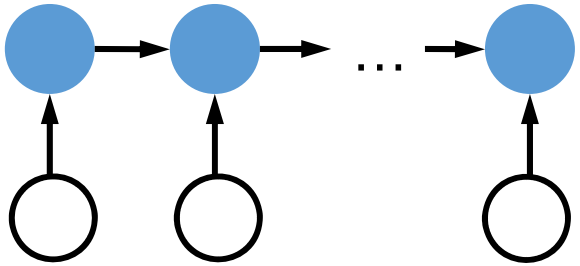

B
NB
HMM
MM
ME
LR
SVM
perceptron
MEMM
MRF
hidden variables
h
W
Bipartite
Structure
v
Image
visible variables
RBM
CRF

他	PN
同时	AD
希望	VV
秘鲁	NR
侨胞	NN
能	VV
继续	VV
关心	VV
和	CC
支持	VV
…	…

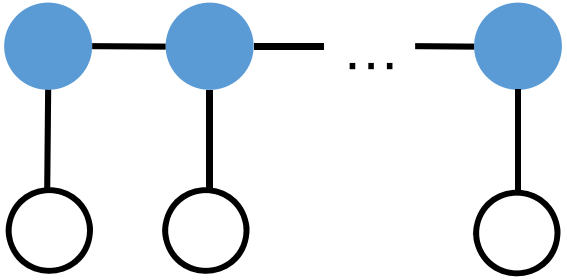
HMM, MEMM, CRF,



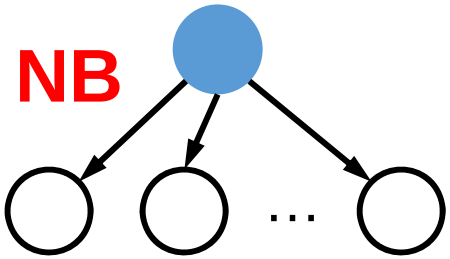
HMM

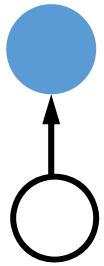


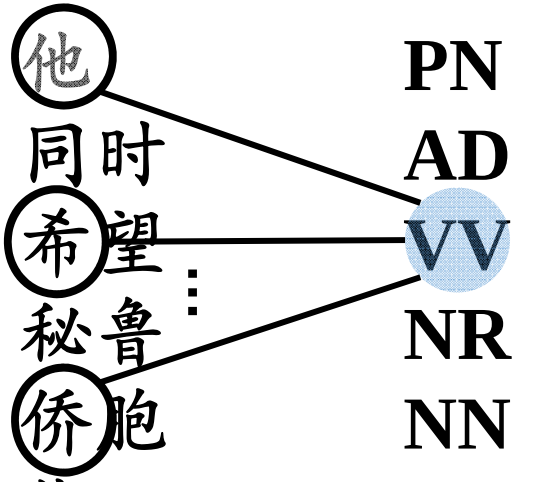
MEMM



CRF


NB


ME LR
SVM
perceptron


他 PN
同时 AD
希望 VV
秘鲁 NR
侨胞 NN
能 VV
继续 VV
关心 VV
和 CC
支持 VV
… …

NB, SVM, ME, LR

可以转化为序列标注问题的NLP任务

CWS

字	词位
他	S
同	B
时	I
希	B
望	I
秘	B
鲁	I
侨	B
胞	I
…	…

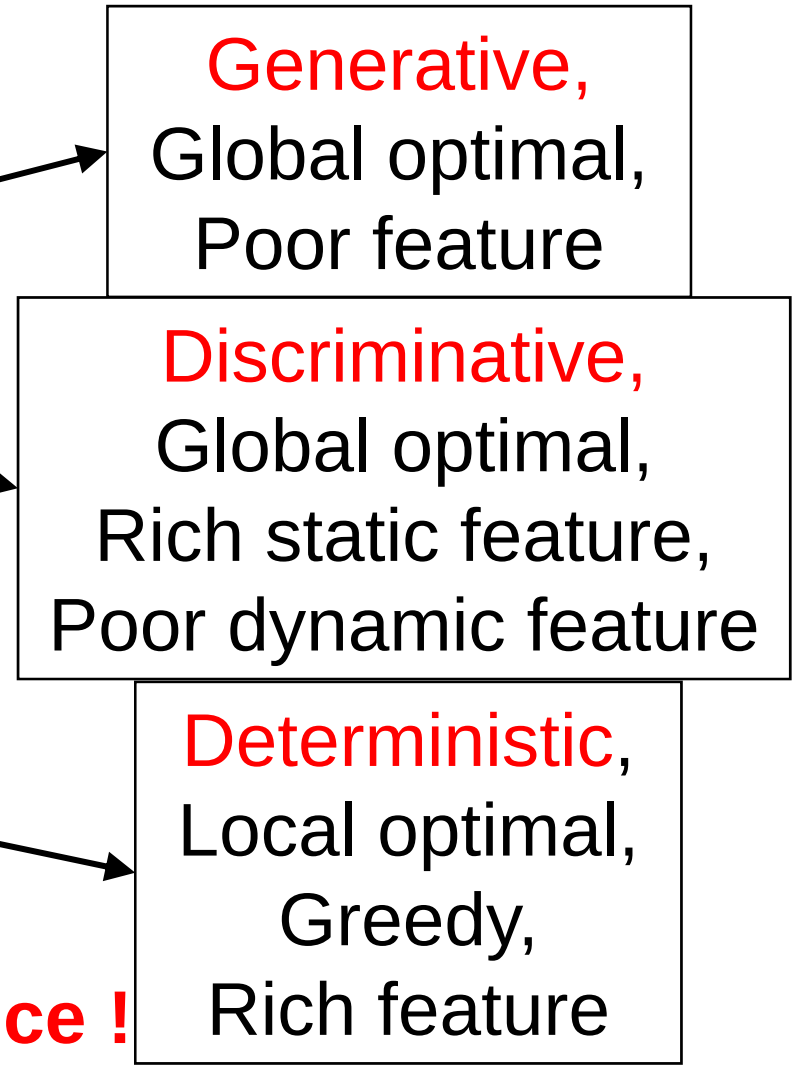
POS tagging

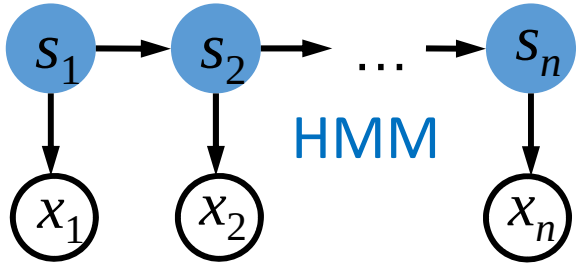
词	POS
他	PN
同时	AD
希望	VV
秘鲁	NR
侨胞	NN
能	VV
继续	VV
关心	VV
和	CC
…	…

序列标注

- 统计序列标注模型
  - Integrated models
    - **HMM**
    - **MEMM**
    - **CRF**
  - Cascade models:
    - **NB**
    - **SVM**
    - **ME**
    - **Perceptron**

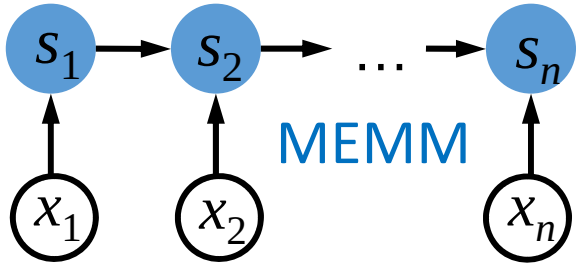
Classifier sequence !





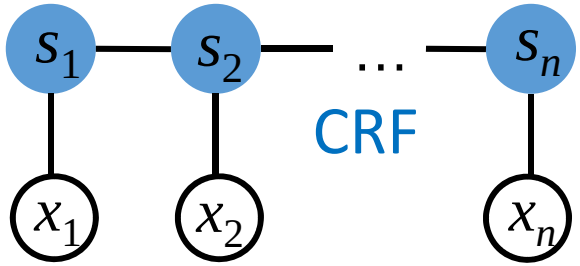
$$p(s,x) = p(s_1)\,p(x_1|s_1)\prod_{i=2}^{n} p(s_i|s_{i-1})\,p(x_i|s_i)$$

$$s^* = \arg\max_s p(s|x) = \arg\max_s \frac{p(s,x)}{p(x)}$$



$$p(s|x) = p(s_1|x_1)\prod_{i=2}^{n} p(s_i|s_{i-1},x_i)$$

$$p(s_i|s_{i-1},x_i) = \frac{1}{Z(s_{i-1},x_i)}\exp\left(\sum_k \lambda_k f_k(s_i,x_i)\right)$$



$$p(s|x) = \frac{1}{Z(s,x)}\exp\left(\sum_i\sum_k \lambda_k f_k(s_{i-1},s_i,x,i)\right)$$

$$Z(s,x) = \sum_s \exp\left(\sum_k \lambda_k f_k(s,x)\right)$$

Generative和discriminative模型——全局寻优

序列标注

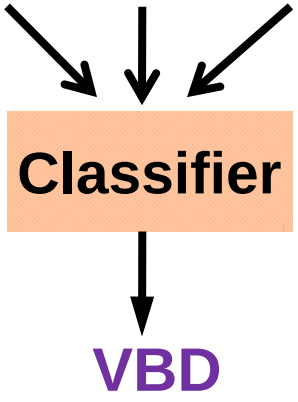
- **Sequential classifications**

Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

John saw the saw and decided to take it to the table.

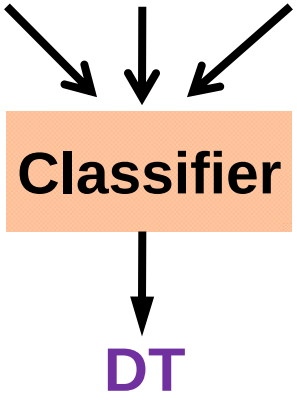


Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

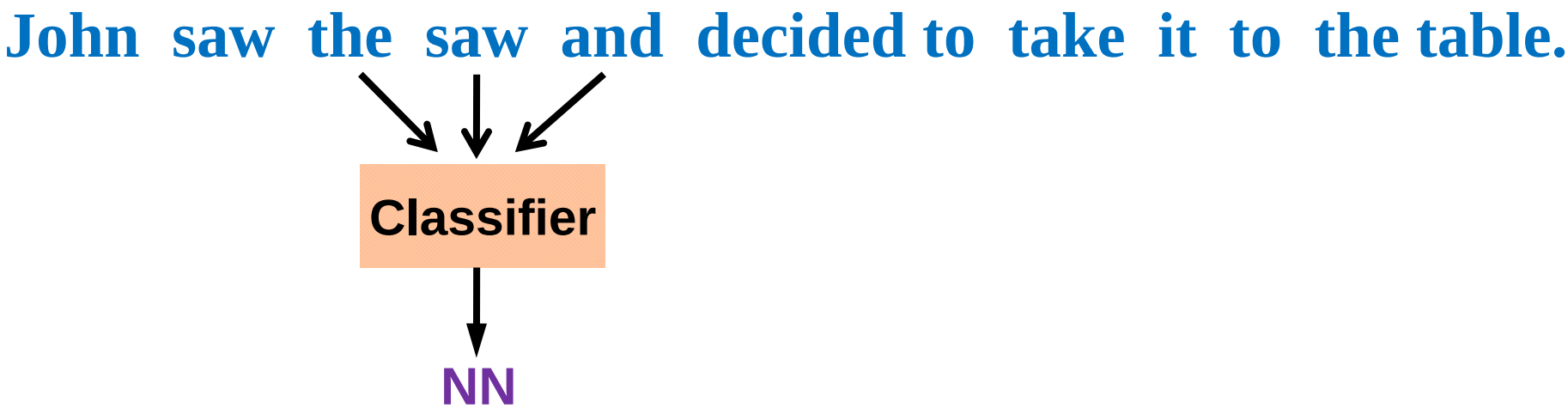
John saw the saw and decided to take it to the table.



Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

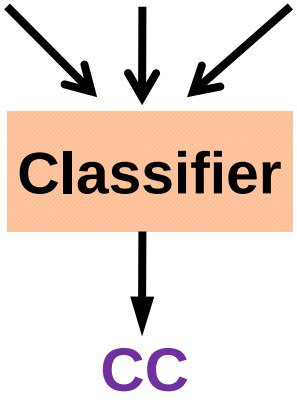


Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

John saw the saw and decided to take it to the table.

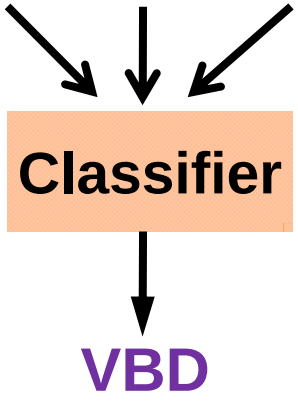


Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

John saw the saw and decided to take it to the table.

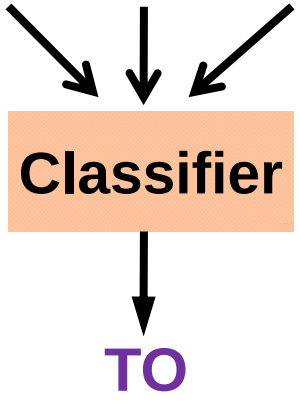


Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

John saw the saw and decided to take it to the table.

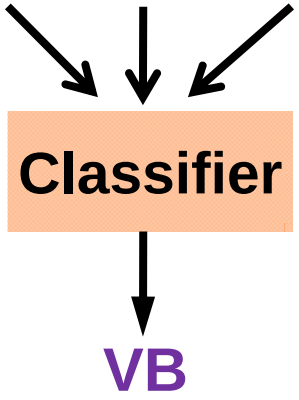


Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

John saw the saw and decided to take it to the table.

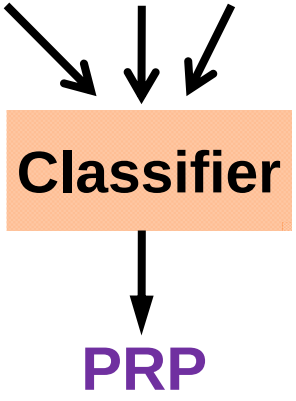


Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

John saw the saw and decided to take it to the table.

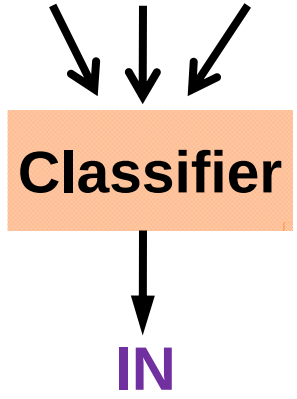


Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

John saw the saw and decided to take it to the table.

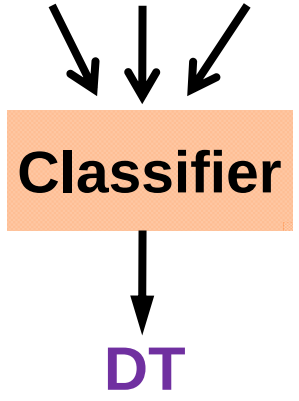


Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

John saw the saw and decided to take it to the table.

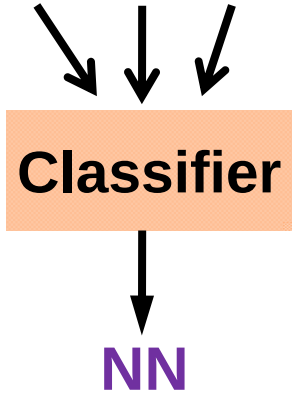


Deterministic级联模型，每个step做一次分类——greedy

序列标注

- **Sequential classifications**

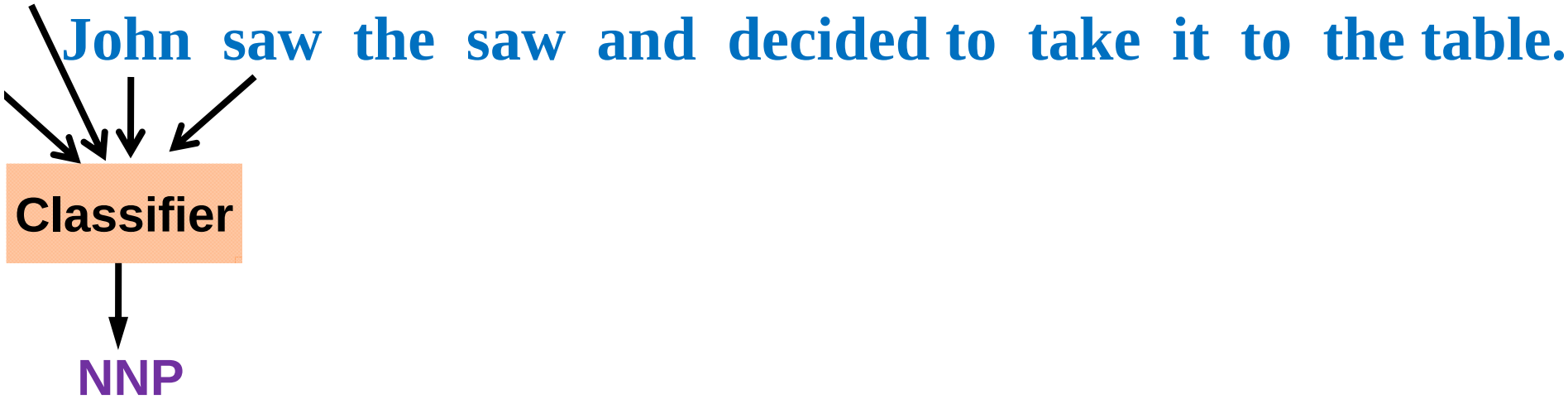
John saw the saw and decided to take it to the table.



Deterministic级联模型，每个step做一次分类——greedy

序列标注

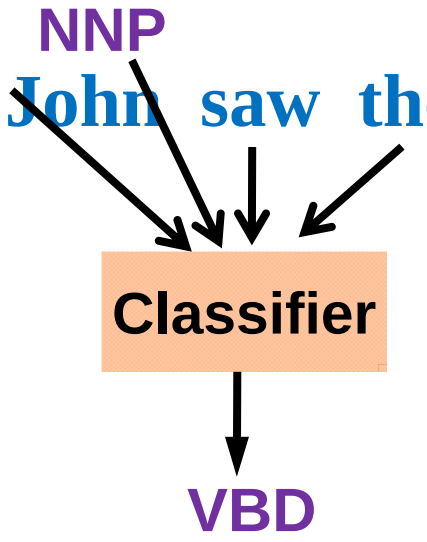
- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

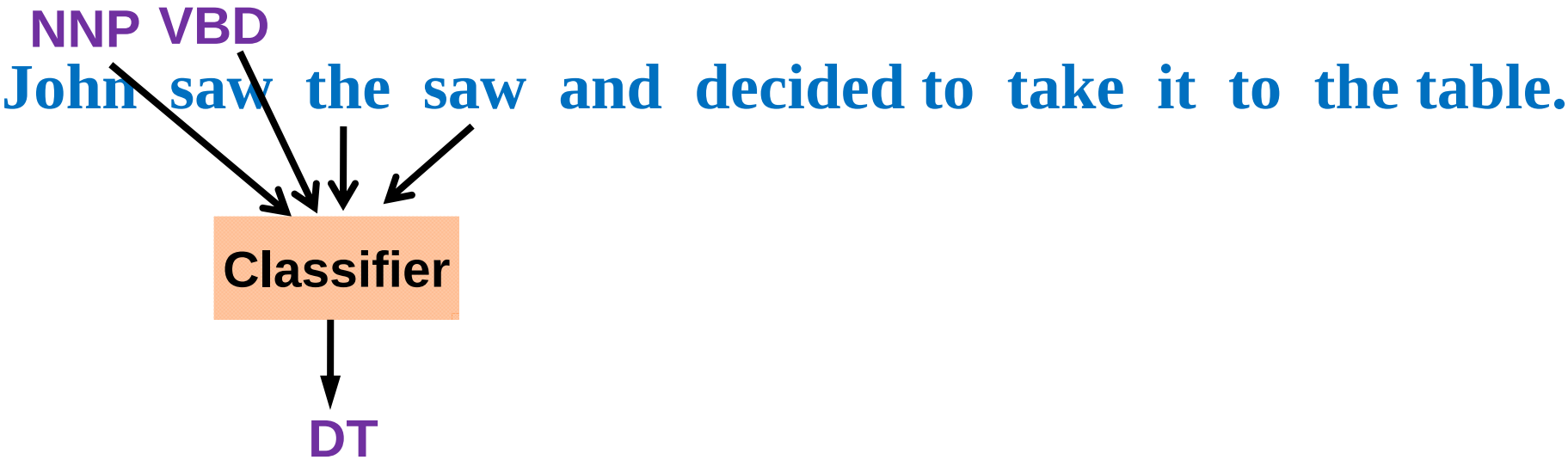
- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

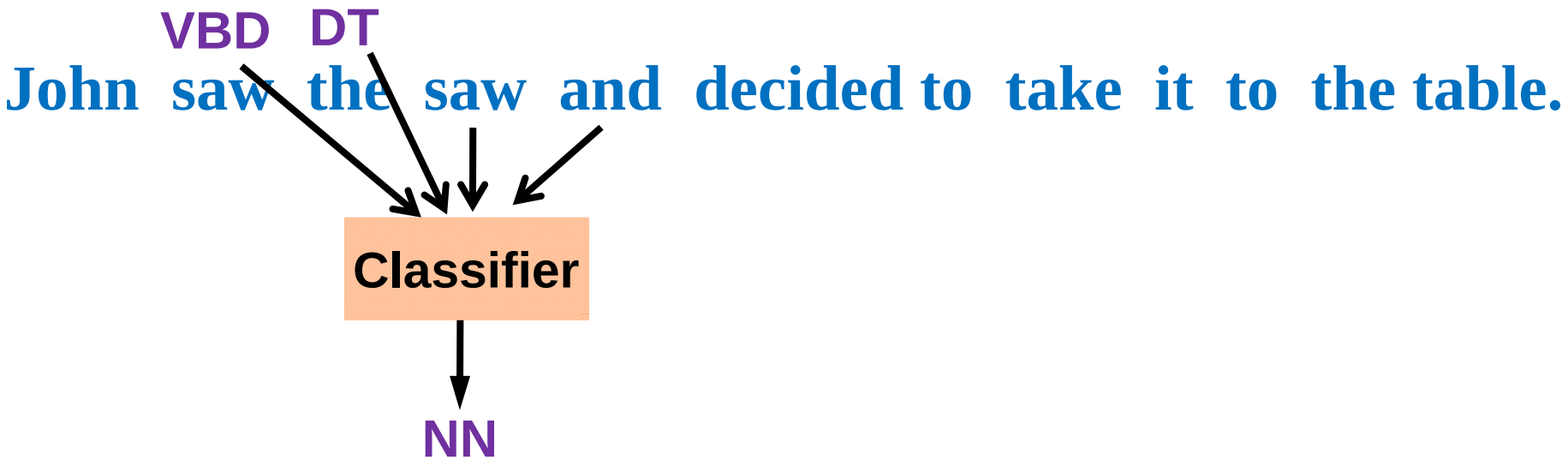
- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

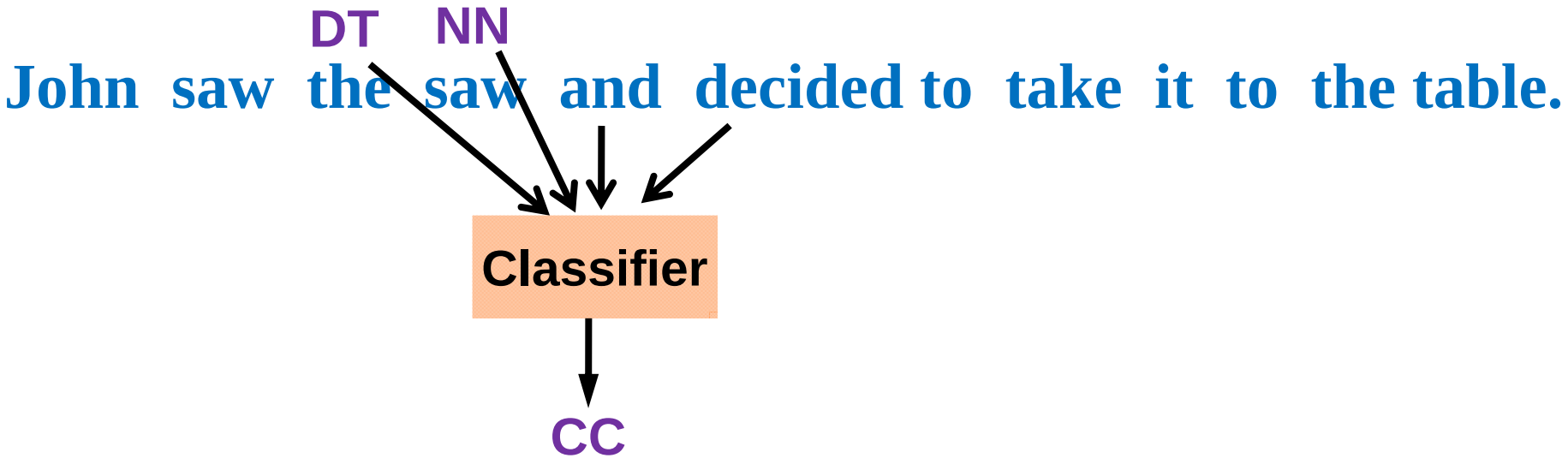
- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

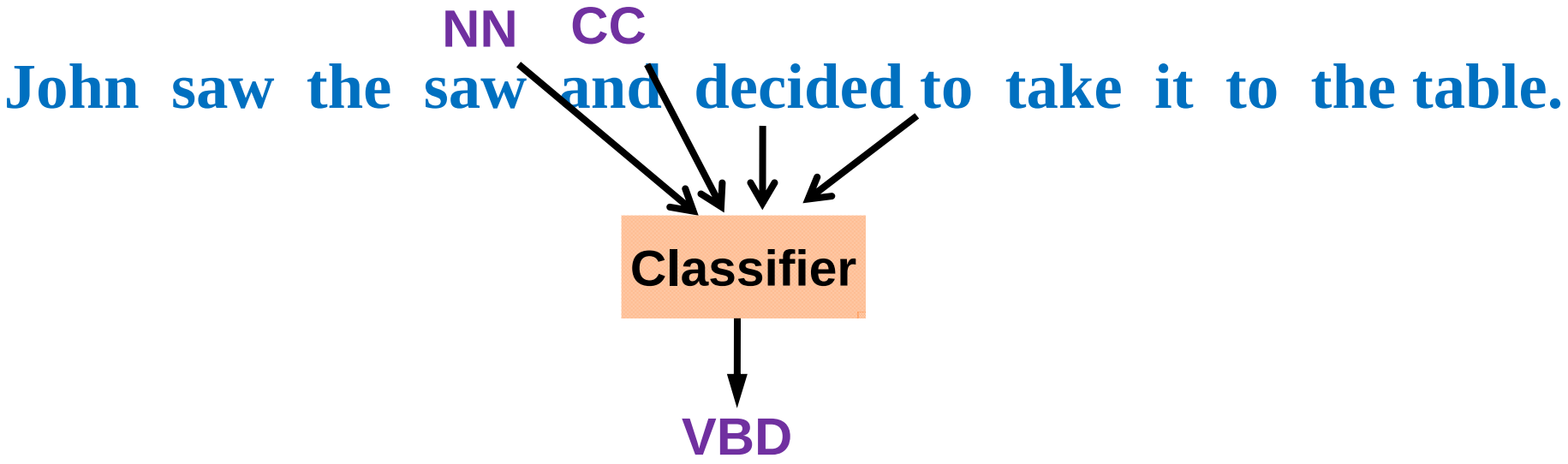
- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

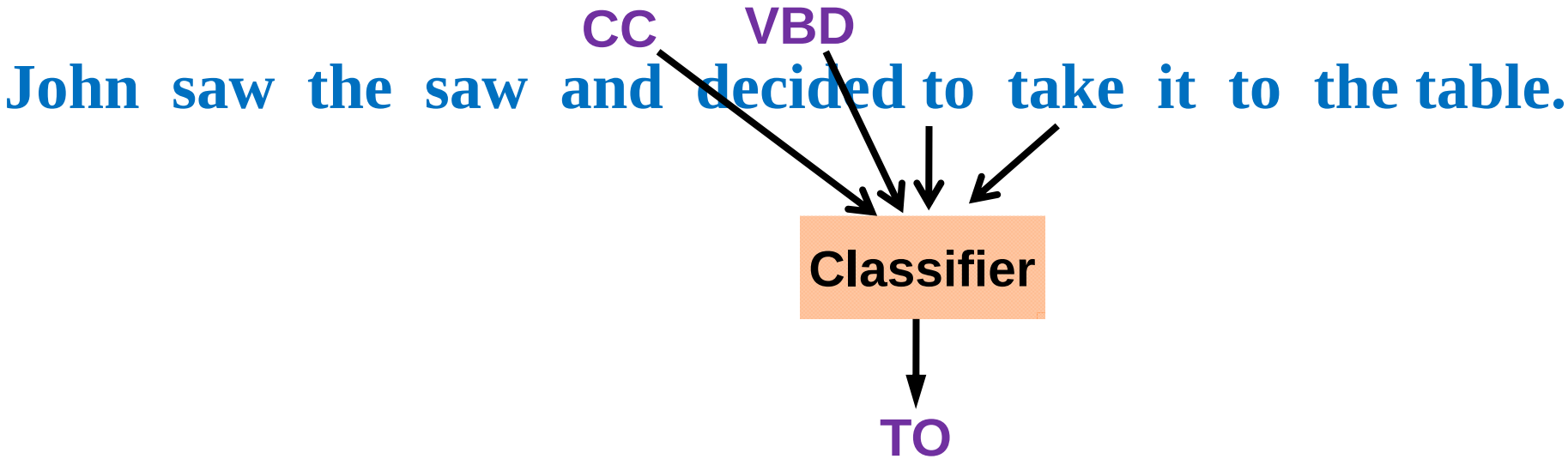
- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

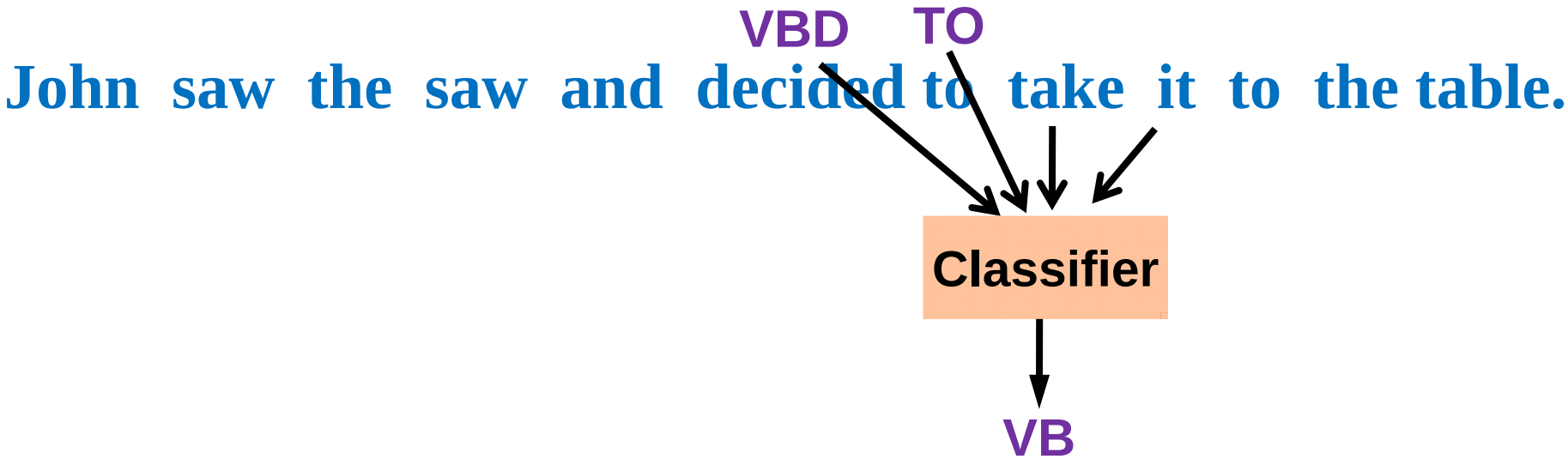
- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

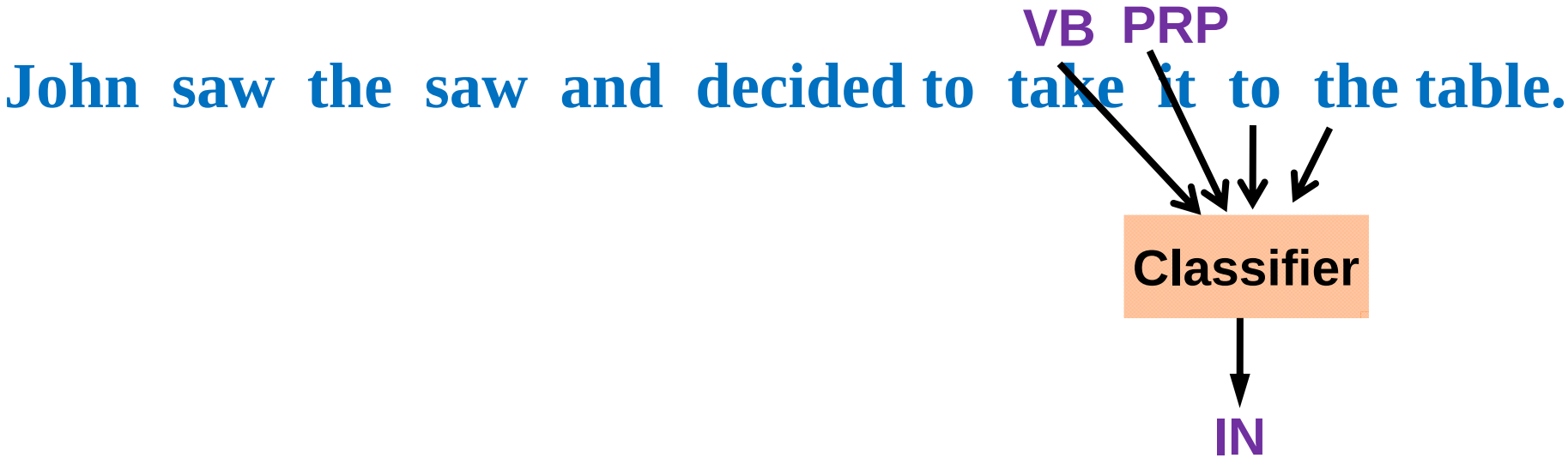
- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

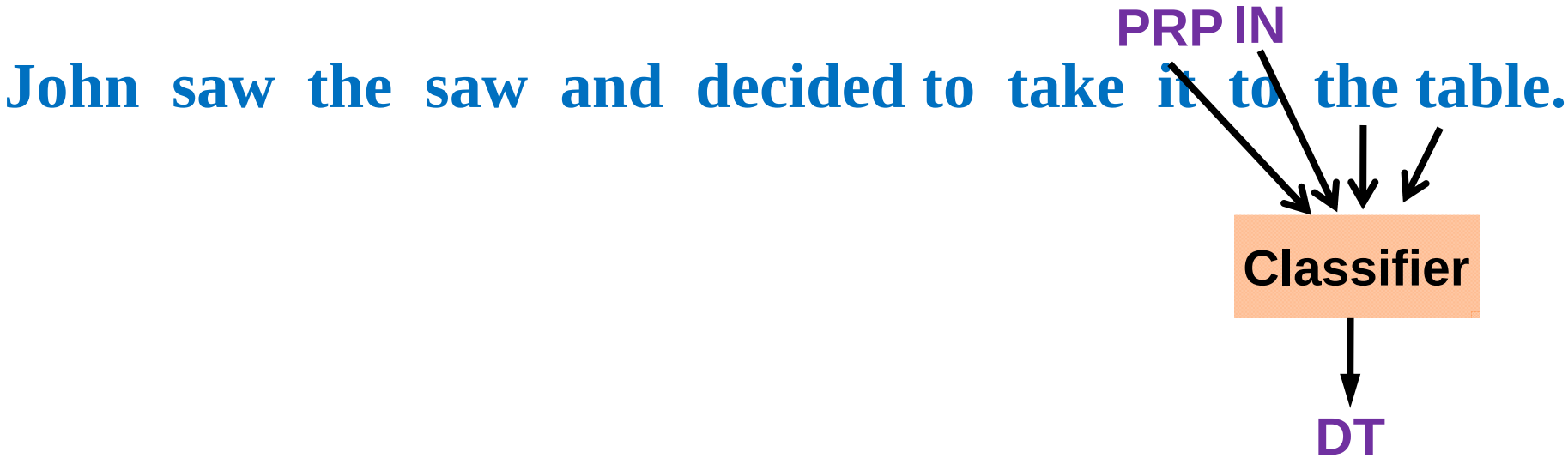
- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

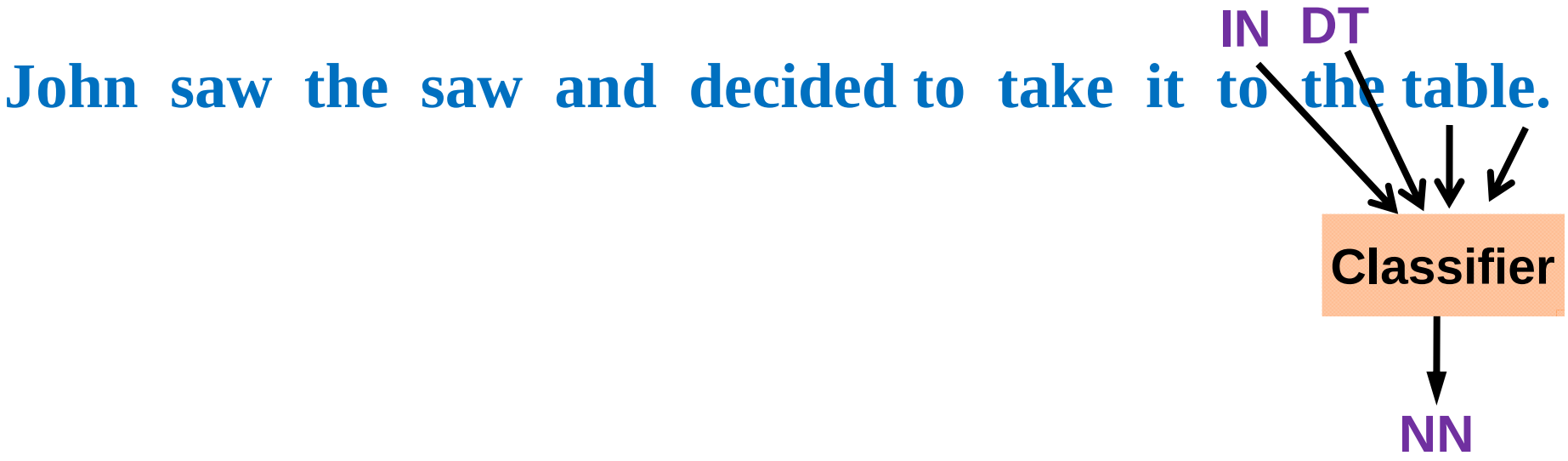
- **Sequential classifications** 使用动态特征



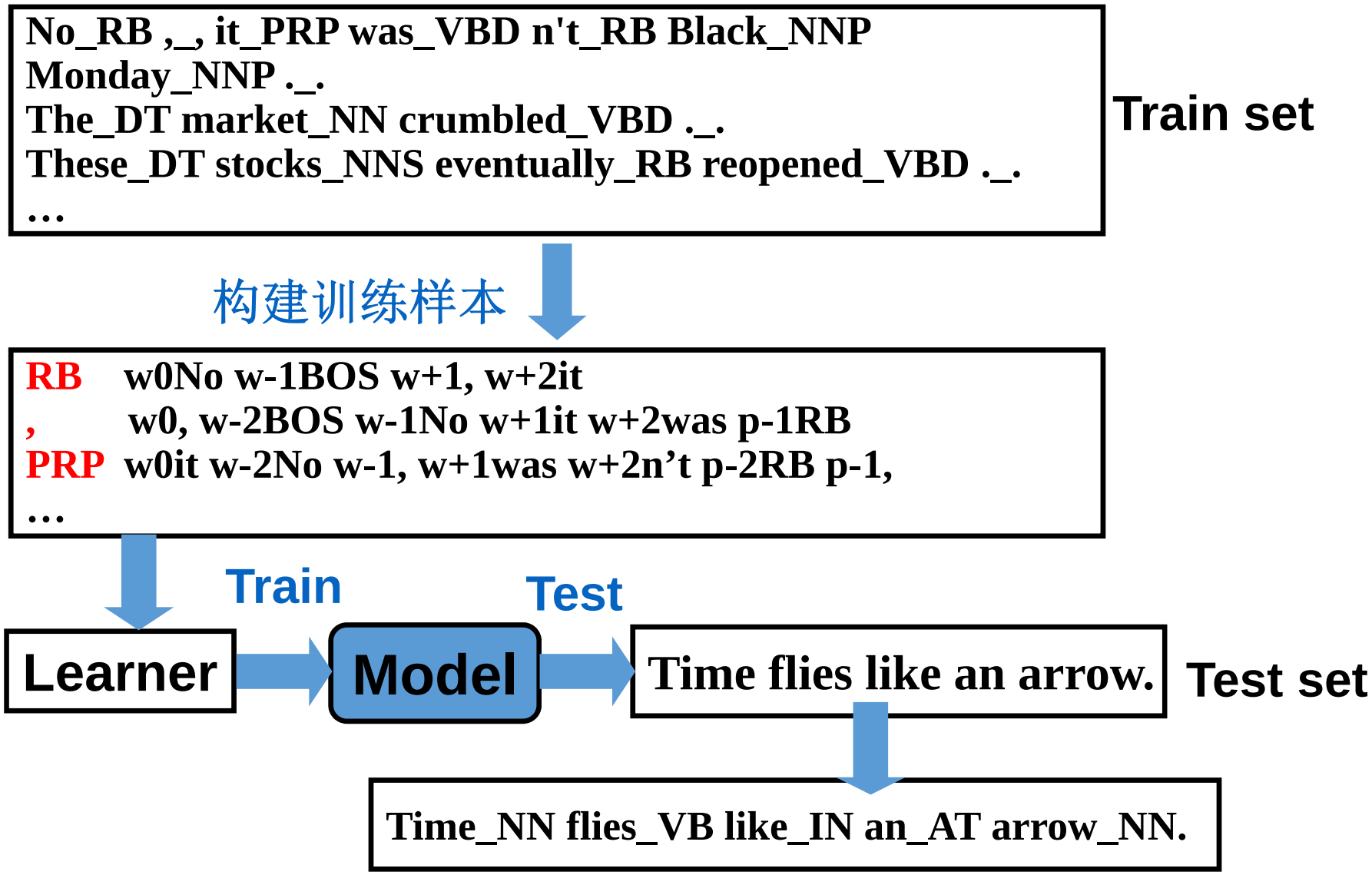
Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果

序列标注

- **Sequential classifications** 使用动态特征



Deterministic级联模型，每个step做一次分类——greedy
优点是可以使用已标注结果



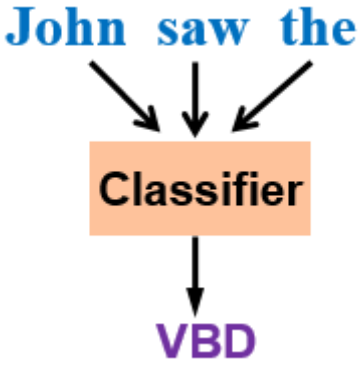
有监督的统计序列标注训练与测试

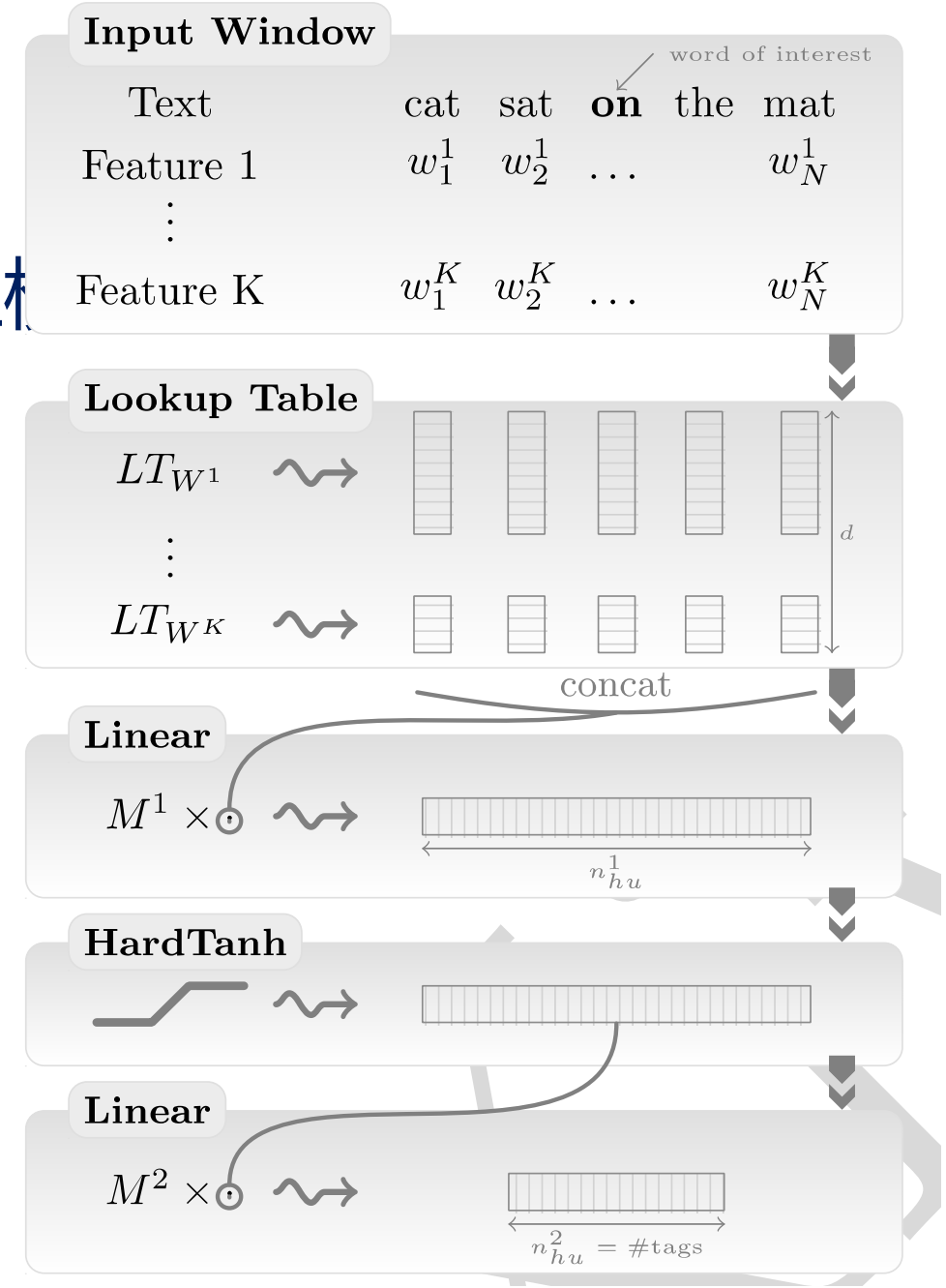
序列标注

- 最早的NN序列标注

Collobert2011: NLP (almost) from scratch

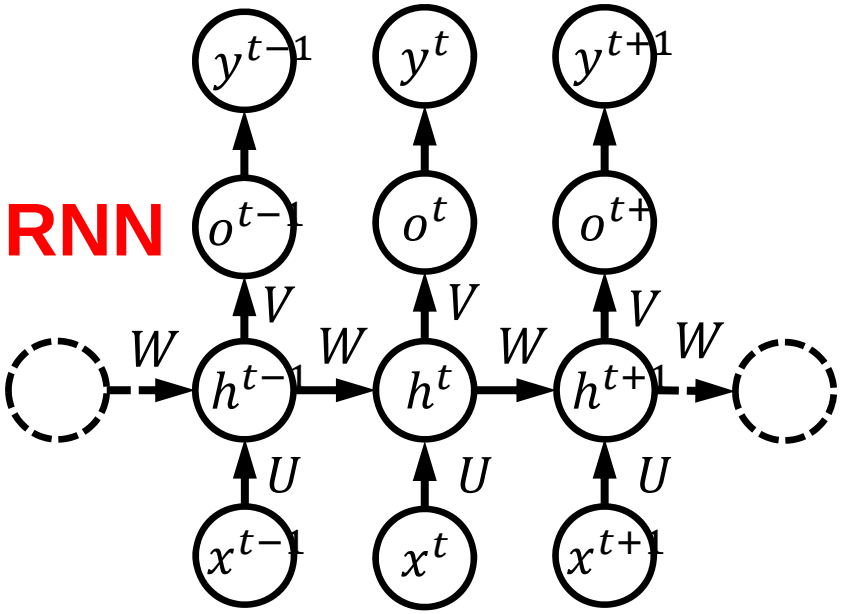
The beginning of the neural model for sequence labeling

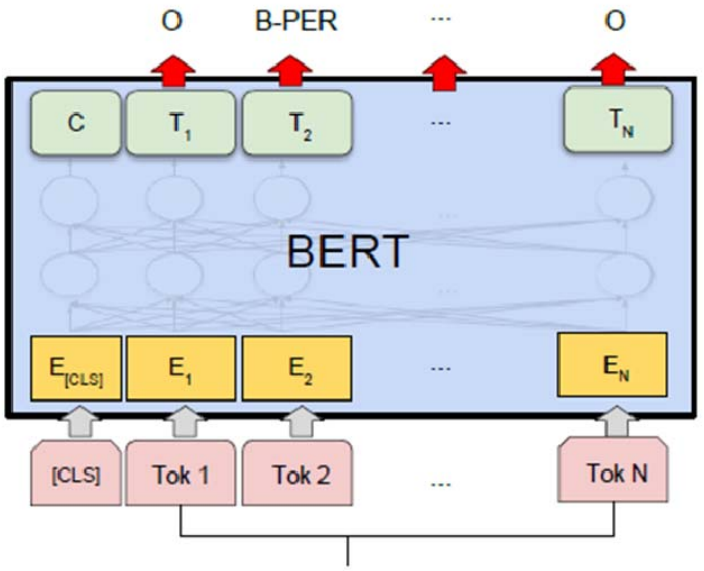




依然是特征工程，只不过使用了分布式表示

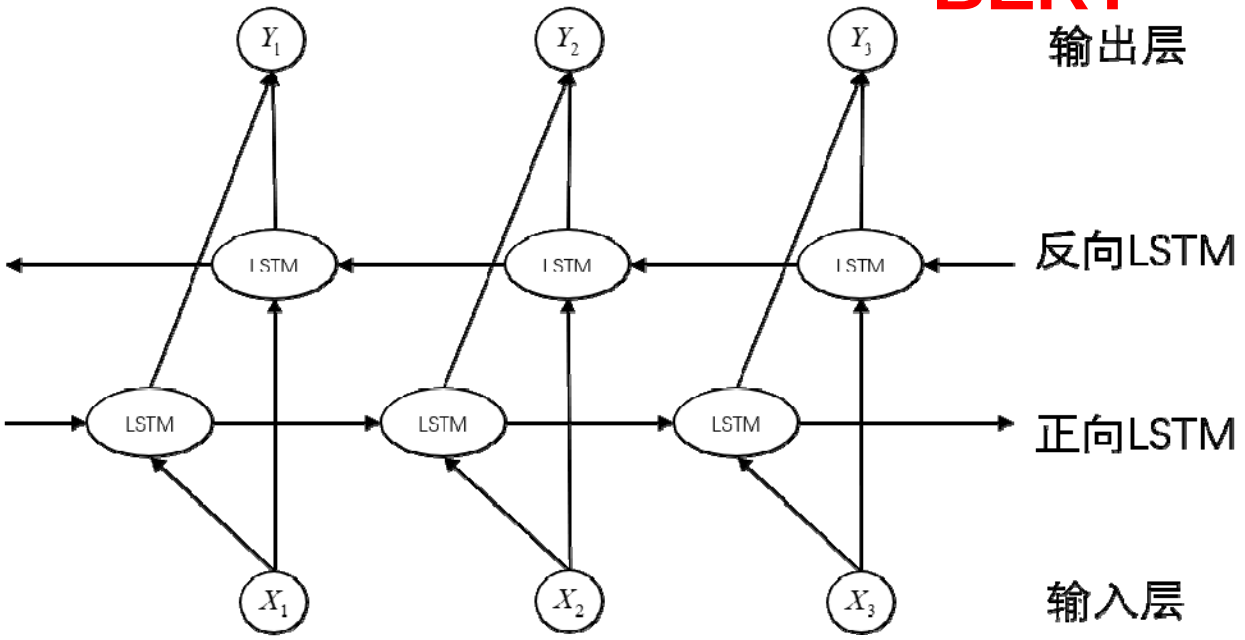
RNN





BERT

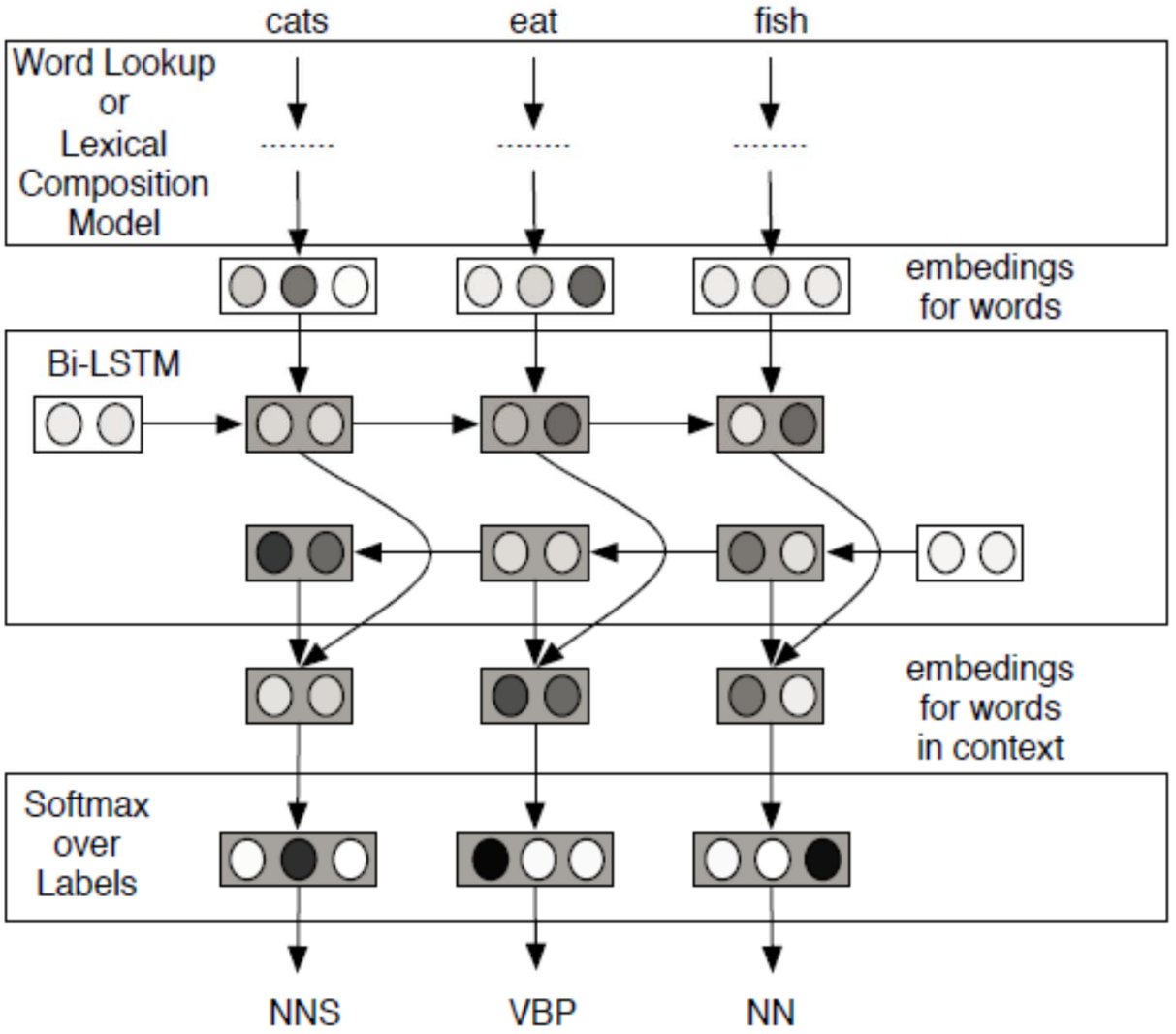
Bi-LSTM



使用更先进的模型来构建序列标注模型

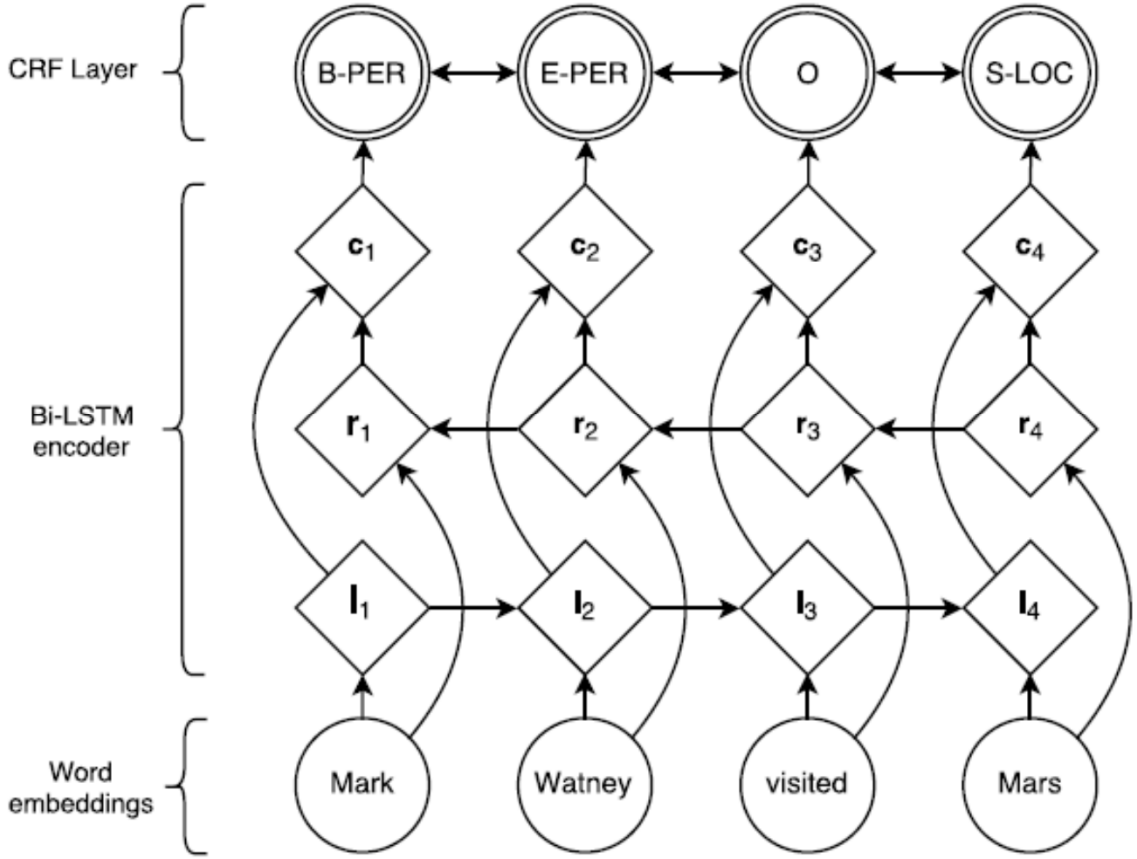
序列标注

- 一个NN的POS标注模型



序列标注

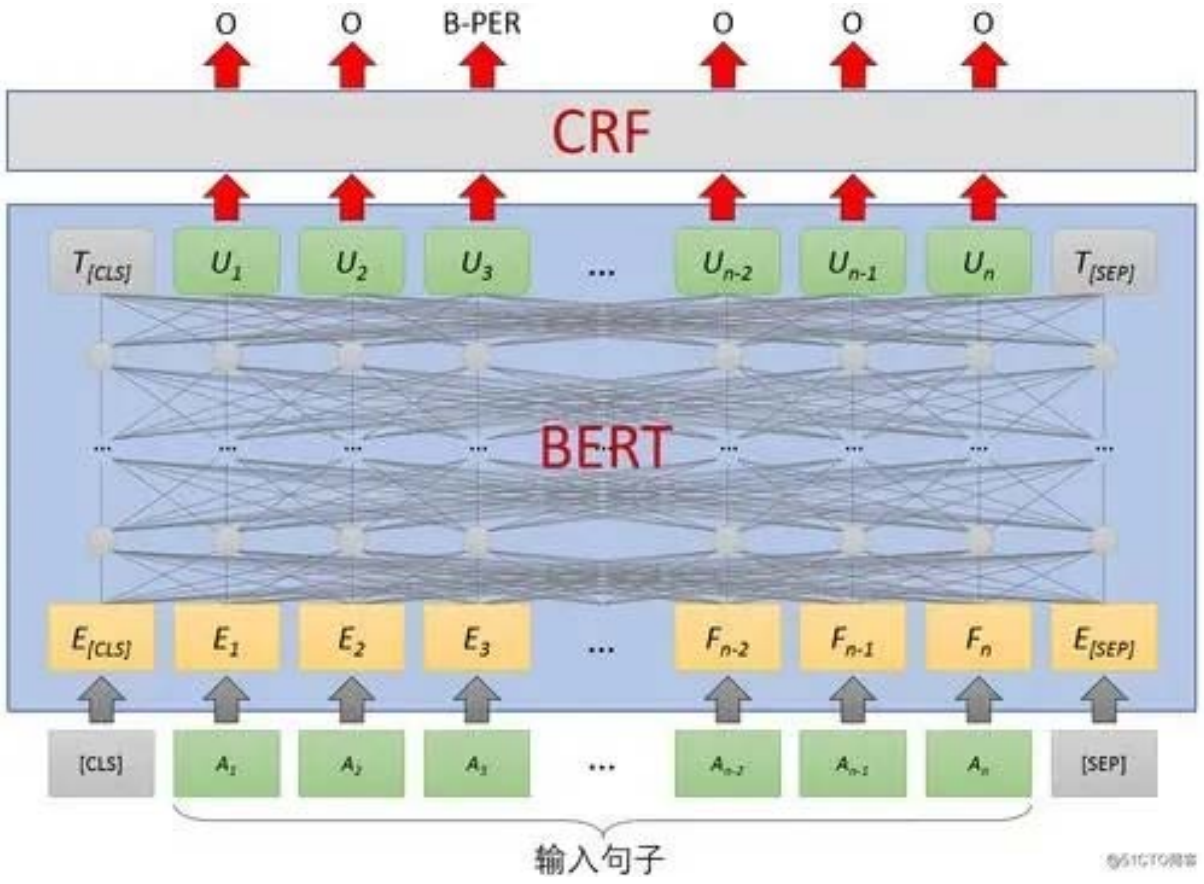
- 神经网络模型的序列标注，缺乏上下文感知

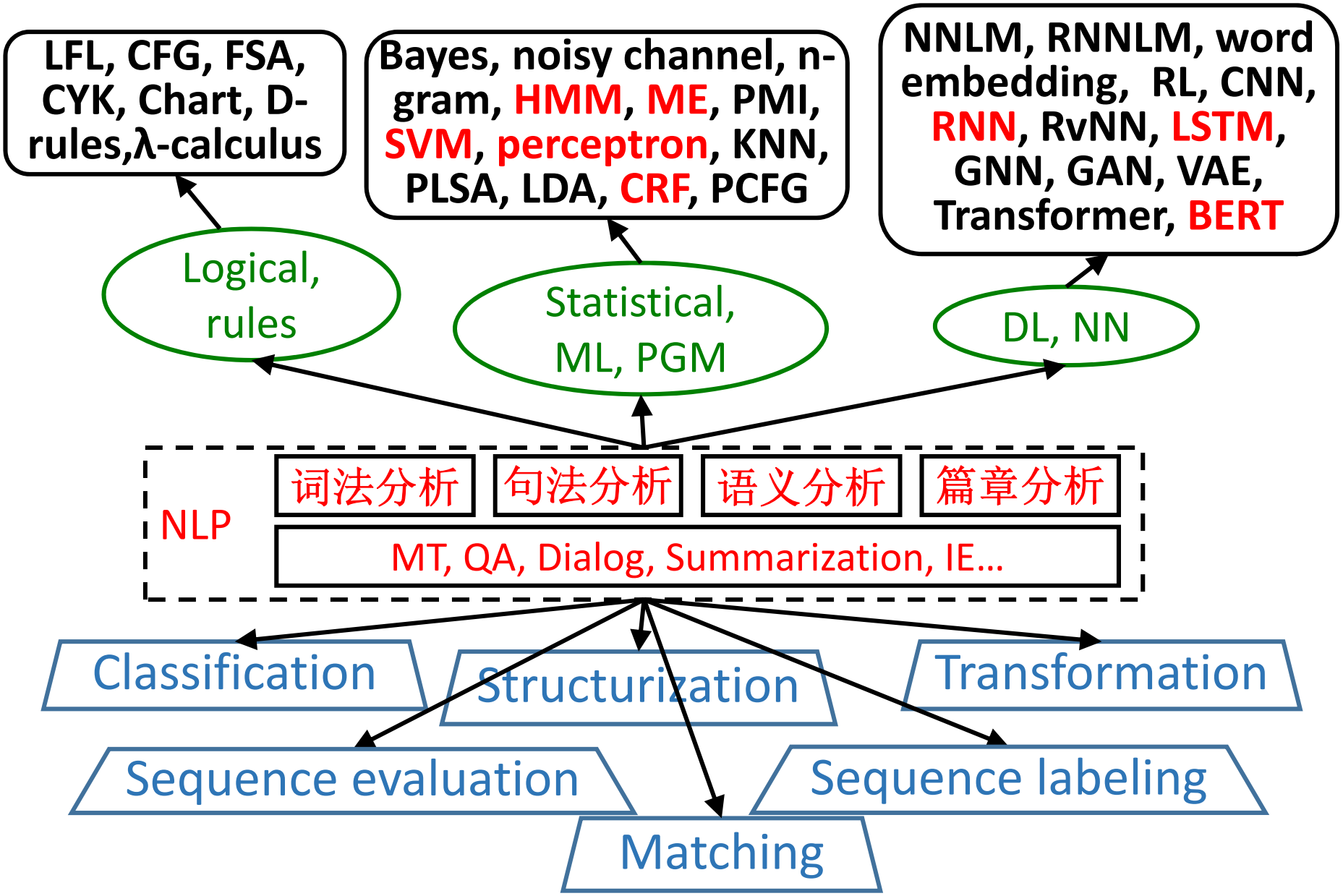


LSTM-CRF (Lample2016)

序列标注

- ✓ 目前主流的序列标注模型，常用于NER





Tasks, problems, methodologies and models in NLP