Natural Language Processing I

4. 中文分词

WU Xiaokun 吴晓堃

xkun.wu [at] gmail

2022/02/28

N元语法

中文分词语料库：统计

- PKU：1998年《人民日报》
- MSR：微软亚洲研究
- 繁体：CITYU （香港城市大学）、AS （台湾中央研究院）

语料	字符	词目	词频	词长
PKU	183万	6万	111万	1.6
MSR	405万	9万	237万	1.7
AS	837万	14万	545万	1.5
CITYU	240万	7万	146万	1.7

语料	字符	词目	词频	词长	OOV
PKU	17万	1万	10万	1.7	5.75%
MSR	18万	1万	11万	1.7	2.65%
AS	20万	2万	12万	1.6	4.33%
CITYU	7万	1万	4万	1.7	7.40%

中文分词语料库：比较

- 规模：AS > MSR > CITYU > PKU
- 从 OOV 看，难度：CITYU > PKU > AS > MSR
- 汉语平均词长：1.7
  - 长词都是低频词
- 汉语常用词汇量：10万级

语料	字符	词目	词频	词长
PKU	183万	6万	111万	1.6
MSR	405万	9万	237万	1.7
AS	837万	14万	545万	1.5
CITYU	240万	7万	146万	1.7

语料	字符	词目	词频	词长	OOV
PKU	17万	1万	10万	1.7	5.75%
MSR	18万	1万	11万	1.7	2.65%
AS	20万	2万	12万	1.6	4.33%

训练语法模型

训练：统计二元、一元语法词频

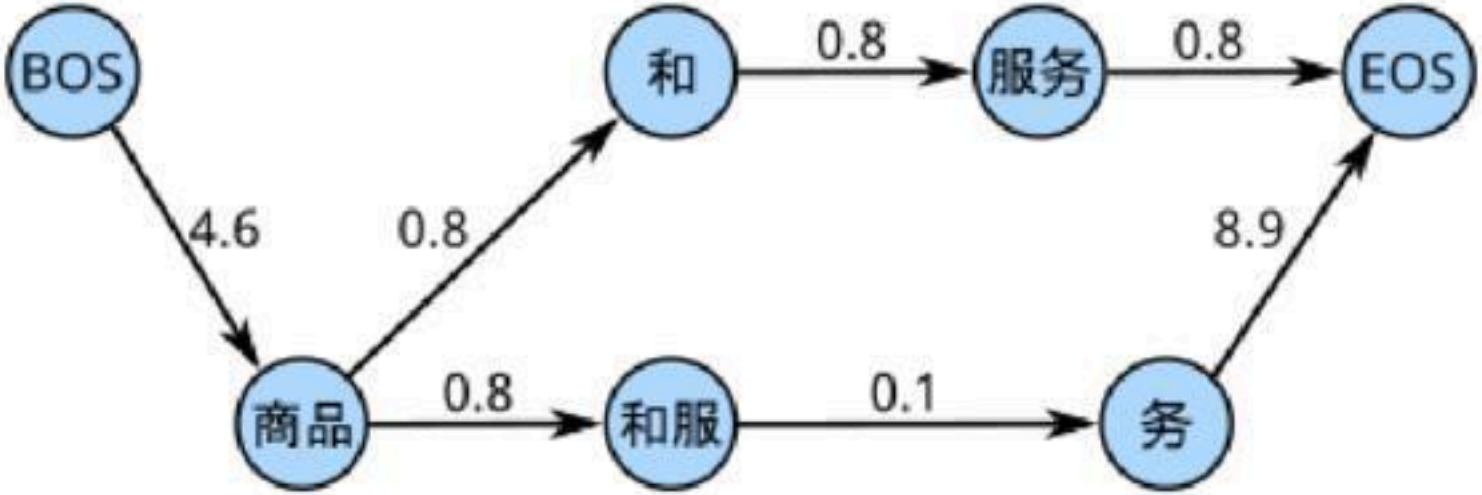
- **词频词典**：单词+词频[+词类]
- 计算概率分布：极大似然估计+平滑策略

词典可以看成模型的**存储形式**

词图

起点到终点每条路径代表一种分词

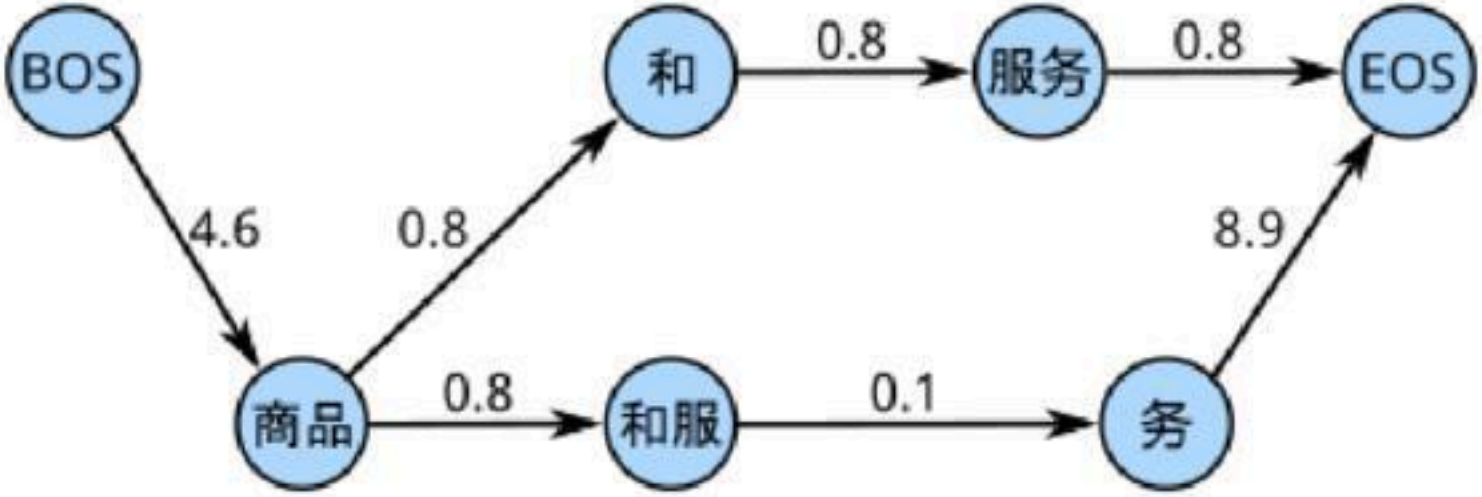
- 目标：找出**最合理**的路径
  - 以二元语法的概率作为距离：最长路径，或负对数的最短路径
- 例如：商品和服务



词图

起点到终点每条路径代表一种分词

- 目标：找出**最合理**的路径
  - 以二元语法的概率作为距离：最长路径，或负对数的最短路径
- 例如：商品和服务



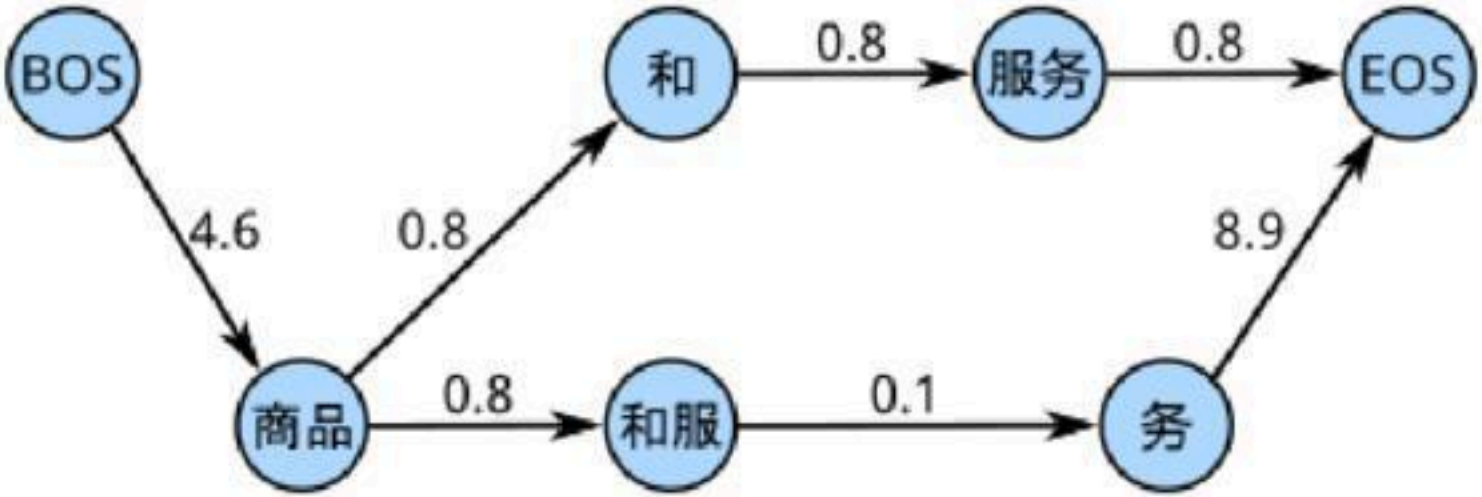
观察：决定路径的关键点在路径最后

- 全局算法，如动态规划

Viterbi 算法

Markov 链构成的网状图上的最短路径

1. 前向：更新最小花费、前驱指针
2. 后向：回溯前驱指针



评测：N元语法

算法	P	R	F_1	R_{OOV}	R_{IV}
最长匹配	91.80	95.69	93.71	2.58	98.22
二元语法	**92.62**	**96.85**	**94.69**	2.58	**99.41**

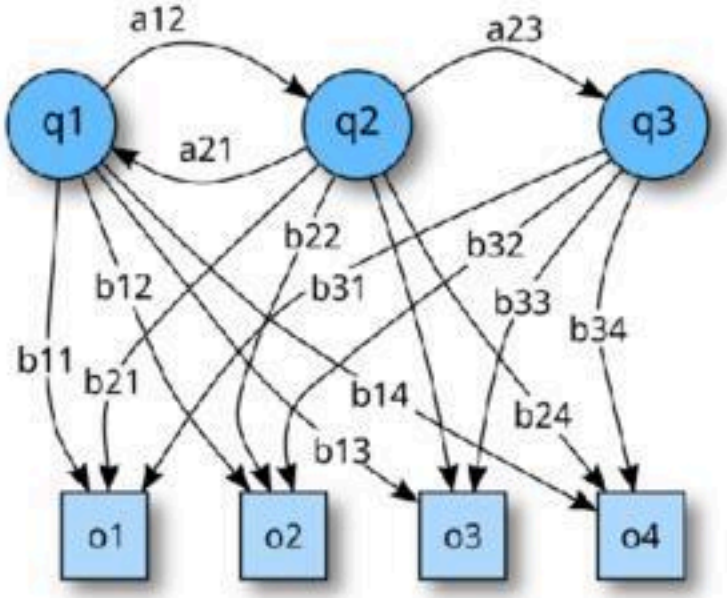
- 二元语法全面胜出
- 召回率没有提升：词图的构建不变

隱式Markov模型

隐式Markov模型

隐式Markov模型 hidden Markov model (HMM)

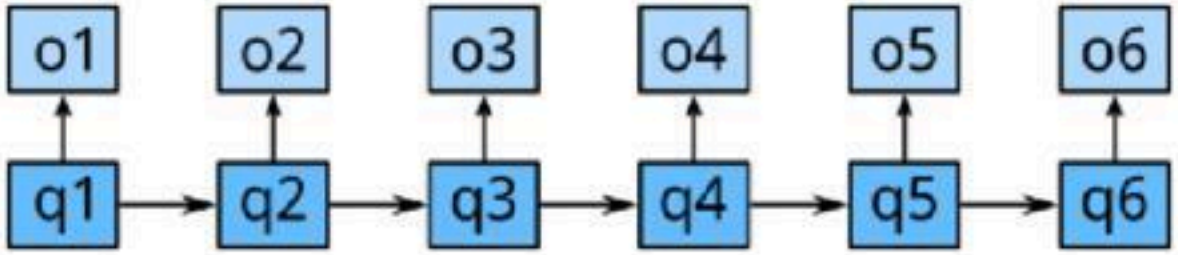
- 隐状态：Markov链，相互转化的**内因**
- 输出变量：由隐状态直接生成的**表象**



问题：依赖链长，计算复杂。

HMM：一阶模型

一阶HMM的两个假设



- 一阶Markov假设：状态概率只取决于**前一个**状态
  - 称为**状态转移概率**

$$P(q_i|q_1..q_{i-1}) = P(q_i|q_{i-1})$$

- 独立性假设：输出变量的概率只取决于**直接关联**的隐变量
  - 称为**观测似然**，或**"发射概率"**

$$P(o_i|q_1..q_T, o_1..o_T) = P(o_i|q_i)$$

评测：隐式Markov模型

算法	P	R	F_1	R_{OOV}	R_{IV}
最长匹配	91.80	95.69	93.71	2.58	98.22
二元语法	**92.62**	**96.85**	**94.69**	2.58	**99.41**
一阶HMM	78.49	80.38	79.42	41.11	81.44
二阶HMM	78.34	80.01	79.16	**42.06**	81.04

感知机

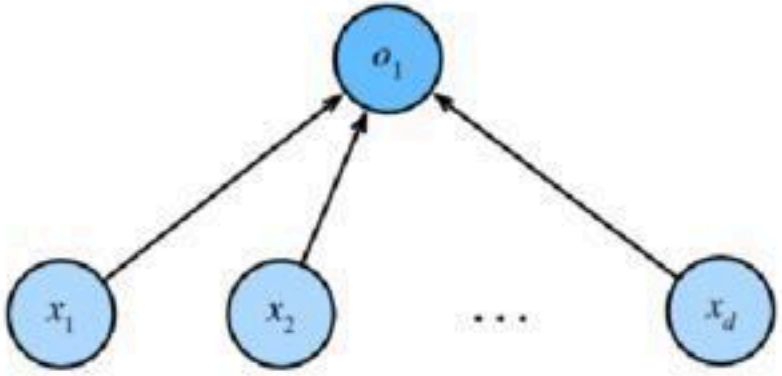
感知机二分类模型

输入：$\mathbf{x}$；参数：$\mathbf{w}, b$；输出：

$$o = \sigma\left(\langle \mathbf{w}, \mathbf{x} \rangle + b\right), \sigma(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{otherwise} \end{cases}$$

二分类问题，激活函数通常输出：$\{-1, 1\}$

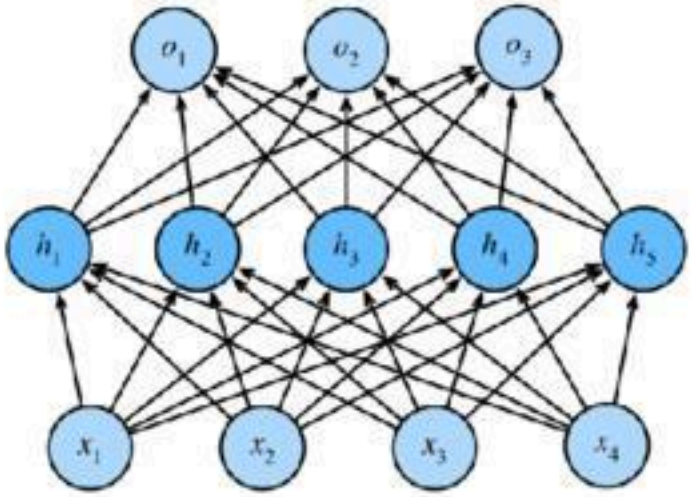
- 回归：实数；softmax回归：概率
- 线性部分不变，**只改变激活函数**


o_1
x_1
x_2
...
x_d

单隐藏层：小批量计算

$$\mathbf{H} = \sigma(\mathbf{X}\mathbf{W}^h + \mathbf{b}^h)$$
$$\mathbf{O} = \mathbf{H}\mathbf{W}^o + \mathbf{b}^o$$

- $\mathbf{X} \in \mathbb{R}^{B\times d}, \mathbf{W}^h \in \mathbb{R}^{d\times l}, \mathbf{b}^h \in \mathbb{R}^{1\times l}$
- $\mathbf{H} \in \mathbb{R}^{B\times l}, \mathbf{W}^o \in \mathbb{R}^{l\times C}, \mathbf{b}^o \in \mathbb{R}^{1\times C}$

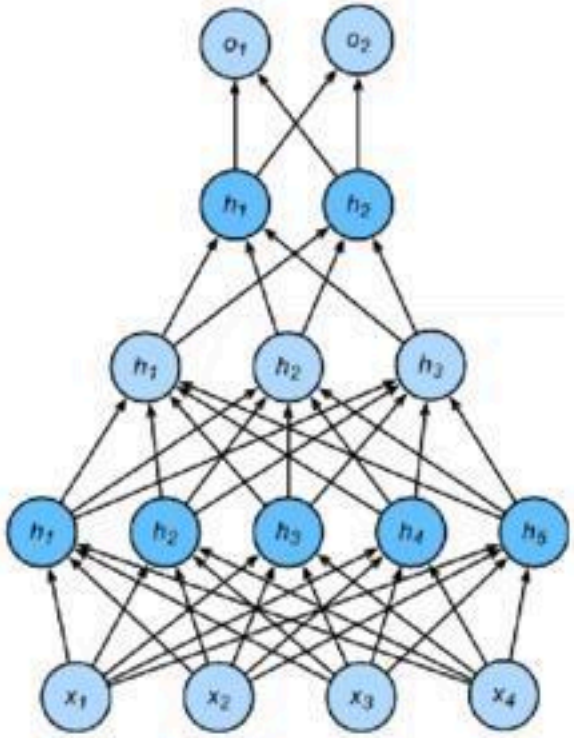


多隐藏层

$$
\begin{aligned}
\mathbf{h}_1 &= \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)\\
\mathbf{h}_2 &= \sigma(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2)\\
\mathbf{h}_3 &= \sigma(\mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3)\\
\mathbf{o} &= \mathbf{W}^o\mathbf{h}_3 + \mathbf{b}^o
\end{aligned}
$$

超参数：取决于模型的**设计**

- 隐藏层数
- 每个隐藏层的大小



评测：感知机

算法	P	R	F_1	R_{OOV}	R_{IV}
最长匹配	91.80	95.69	93.71	2.58	98.22
二元语法	92.62	**96.85**	94.69	2.58	**99.41**
一阶HMM	78.49	80.38	79.42	41.11	81.44
二阶HMM	78.34	80.01	79.16	42.06	81.04
平均感知机	**96.58**	96.34	**96.46**	70.70	97.04
结构化感知机	96.07	95.26	95.66	**72.68**	95.88

条件随机场

条件随机场 CRF

条件随机场 conditional random field (CRF)：直接计算后验，判别标签序列

$$\hat{T} = \arg\max_{T \in \mathcal{T}} P(T|S)$$

- 对比 HMM：$\hat{T} = \arg\max_T P(S|T)P(T)$

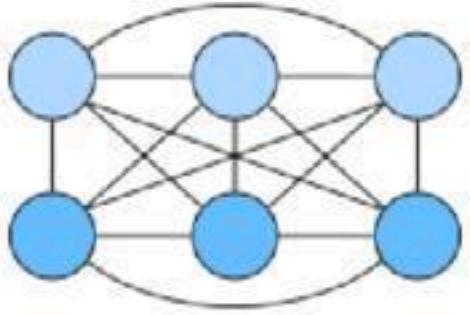
条件随机场 CRF

条件随机场 conditional random field (CRF)：直接计算后验，判别标签序列

$$\hat{T} = \arg\max_{T \in \mathcal{T}} P(T|S)$$

- 对比 HMM：$\hat{T} = \arg\max_T P(S|T)P(T)$

问题：每个时间步都计算整个标签序列T的概率



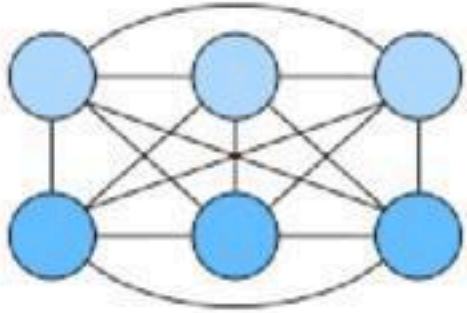
- 拆解成相关的**局部特征**，然后聚集并归一化

特征拆解

全局特征$F_k(S,T)$：每个都是整个输入序列S和输出序列T的特征

$$P(T|S)=\frac{1}{Z(S)}\exp\left(\sum_{k=1}^{K}w_kF_k(S,T)\right)$$

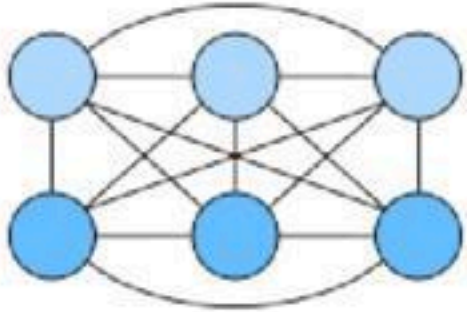
- $Z(S)=\sum_{T'\in\mathcal{T}}\exp\left(\sum_{k=1}^{K}w_kF_k(S,T')\right)$



特征拆解

全局特征$F_k(S,T)$：每个都是整个输入序列S和输出序列T的特征

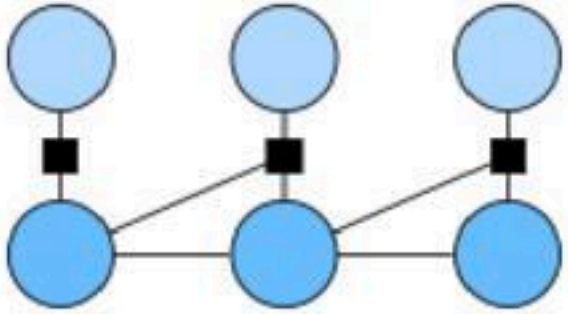
$$P(T|S) = \frac{1}{Z(S)} \exp\left(\sum_{k=1}^{K} w_k F_k(S,T)\right)$$



- $Z(S) = \sum_{T' \in \mathcal{T}} \exp\left(\sum_{k=1}^{K} w_k F_k(S,T')\right)$

简化计算：拆解成T每个位置上的**局部特征**之和

$$F_k(S,T) = \sum_{i=1}^{n} f_k(t_{i-1}, t_i, s_i, i)$$



- 称为**线性链式CRF**：特征计算只依赖于局部输出t_{i-1}, t_i

评测：条件随机场

算法	P	R	F_1	R_{OOV}	R_{IV}
最长匹配	91.80	95.69	93.71	2.58	98.22
二元语法	92.62	**96.85**	94.69	2.58	**99.41**
一阶HMM	78.49	80.38	79.42	41.11	81.44
二阶HMM	78.34	80.01	79.16	42.06	81.04
平均感知机	96.58	96.34	96.46	70.70	97.04
结构化感知机	96.07	95.26	95.66	**72.68**	95.88
条件随机场	**96.86**	96.64	**96.75**	71.54	97.33

词类标注

词类、词类标注

词类：也称词性，即单词的语法分类

- 名词、动词、形容词等

词类标注：给每个词一个词类标签

- 一词多类问题：如“她希望成为全村人的希望。”
- OOV问题：同类词义推断，如“头上戴着束发嵌宝紫金冠”

词类、词类标注

词类：也称词性，即单词的语法分类

- 名词、动词、形容词等

词类标注：给每个词一个词类标签

- 一词多类问题：如“她希望成为全村人的希望。”
- OOV问题：同类词义推断，如“头上戴着束发嵌宝紫金冠”

中文词类：属于现代研究，深受西方影响

- 标注规范存在分歧，且受版权控制

现代汉语词类表

类别	细分	举例	类别	细分	举例
名词	具体、抽象	人、意	量词	名量	把
	方位	东		动量	次
代词	人称、指示、疑问	我、这、谁	副词	程度、范围、否定	很、都、未
动词	不及物、及物	醒、看	介词	时间、原因、比较	自、因、比
	可能、必要、愿意	能、应、愿	连词	联合、偏正	和、但
	趋向	来	助词	结构、时态、语气	的、过、呢
形容词	性质、状态	大、烫	叹词	喜悦、愤怒、呼唤	哈、哼、喂

标注集

《人民日报》语料库与PKU标注集

- 1998年1月份的词性标注语料库

国家语委语料库

- 2006年国标：《信息处理用现代汉语词类标注集规范》

实验：词类标注

命名实体识别

命名实体、命名实体识别

命名实体：描述实体的词汇

- 人名 PER、组织 ORG、地点 LOC、地域 GPE

命名实体识别：专有名词的分词

- 起止范围 + 标签类型
  - 语料库颗粒较大时：分类 + 词类
  - 复合词已拆分时：规则、统计

基于规则的NER

例如：音译用字较为固定、歧义小，而名称较长

1. 粗分：确定备选词是否音译字序列
2. 从备选词向右扫描，合并音译用字

```
列宁/nh 格勒/ns 被包围，但并未被攻占。
```

基于规则的NER

例如：音译用字较为固定、歧义小，而名称较长

1. 粗分：确定备选词是否音译字序列
2. 从备选词向右扫描，合并音译用字

```
列宁/nh 格勒/ns 被包围，但并未被攻占。
```

音译名称专用词典

- 误命中问题：谨慎触发规则1

实验：命名实体识别