0. 写在前面

1 本课程总体结构

章节		教学内容
第一章 引言 (刘均, 2)		概念与研究背景；主要任务；挑战与研究方向；相关资源
第二章：自然语言的统计特性 (刘均, 1)		Zipf定律、Heaps定律、Benford 定律。
第三章：语言模型	词袋模型 (刘均, 3)	语言模型；词袋模型（BoW）；TF-IDF。 NLU任务：情感分析、文本聚类。
	概率语言模型 (李辰, 6)	概率语言模型；n-gram 模型；最大似然估计；平滑技术。 NLU任务：分词、语义关系抽取。
	主题模型 (刘均, 6)	生成模型；主题模型的图表示；LSA、PLSA、LDA；NMF等。 NLU任务：话题检测、推荐。
	神经网络语言模型 (李辰, 6)	分布式表示；C&W、CBOW、Skip-Gram、Glove等。 NLU任务：对话、实体消歧。
第四章：机器翻译	概述 (李辰, 1)	面临的挑战；发展历程；方法类别及特点；MT评估。
	统计机器翻译 (李辰, 3)	统计MT；Noisy Channel模型；IBM模型。
	神经网络机器翻译 与大语言模型 (刘均, 4)	RNN与LSTM简介；Encoder-Decoder框架；Attention模型；大语言模型。

- 这门课由于由两位老师授课，个人感觉结构比较混乱
- 由于时间紧任务重经费无，所以笔记还是按PPT内容和以上结构展开，即使有很多不合理的地方

2 考试有关事项

1. 词性标注

1.1. 基本概念

1️⃣词性标注概念：

1. 概念：给定一个句子，为其中的每个词分配适当的词性
2. 示例：Mr./NNP Vinken/NNP is/VBZ chairman/NN of/IN Elsevier/NNP N.V./NNP
3. 顺序：分词→词性标注→句法分析/命名实体识别/情感分析

2️⃣词性的消息来源：词语$\xrightarrow{\text{本身词性}}$预选语料库中最高频词性$\xrightarrow[\text{消除多义词歧义造成的误差}]{\text{上下文}}$精确词性

1.2. 马尔可夫模型标注器

1️⃣模型概述

1. 基本要素：

序列	含义	对应马可夫模型中
T	单词的词性集(形容词/名词/动词.....)	状态空间
O	输入的单词集	观测空间
W	对单词词性的标注序列	状态序列

2. 基本假设：有限历史(一词の词性只依赖于其前一词の词性)+时间不变性(这种依赖不随时间改变)

2️⃣模型原理：通过最大化$P(T|W)$ $\xrightarrow{\text{找到}}$最佳的词性标注序列$T=\{t_1,t_2,...,t_n\}$

1. 贝叶斯分解：$P(T|W)=\frac{P(W|T)\times P(T)}{P(W)}\propto P(W|T)\times P(T)$

Item	联合概率分解	含义
$P(W\mid T)$	$\prod_{i=1}^{n}P\left(w_i\mid t_i\right)$	每个单词由对应的词性生成
$P(T)$	$\prod_{i=1}^{n}P\left(t_i\mid t_{i-1}\right)$	词性序列之间具有转移关系

2. 联合概率分解：$P(T|W)=\prod_{i=1}^{n}P\left(w_i|t_i\right)P\left(t_i|t_{i-1}\right)$

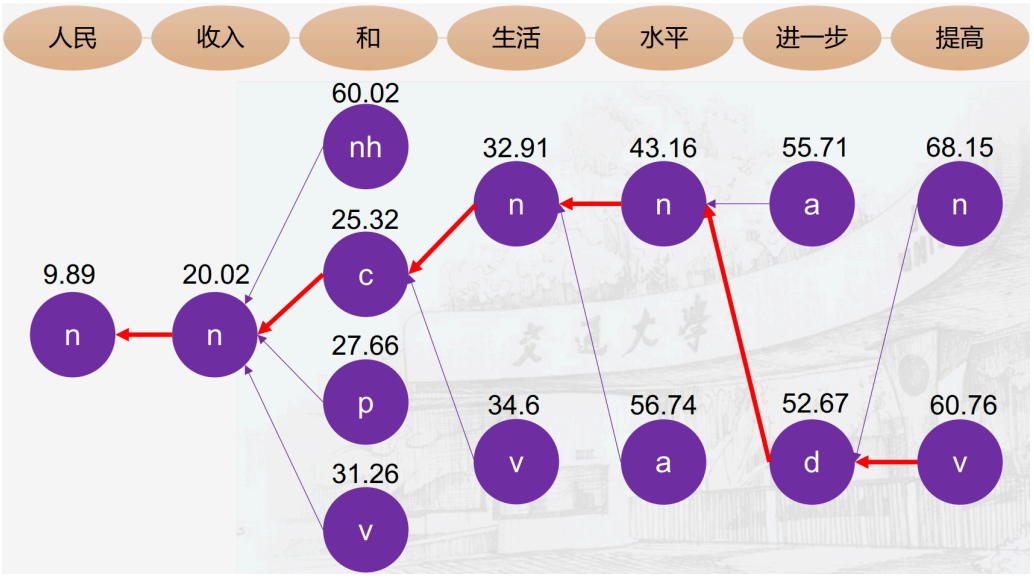
3️⃣模型训练与预测

1. 模型训练：

概率类型	公式	含义
标注转移概率	$P\left(t^k\mid t^j\right)=\frac{C\left(t^j,t^k\right)}{C\left(t^j\right)}$	词性t^j转移到t^k的次数/词性t^j的总出现次数

概率类型	公式	含义
词生成概率	$P\left(w^l \mid t^j\right)=\frac{C\left(w^l, t^j\right)}{C\left(t^j\right)}$	词w^l被标注为词性t^j的次数/词性t^j总出现次数

2. 模型预测：示例



4️⃣其它事项

1. 未登录词：即训练语料库中从未出现过的词，可认为$P(w|t)=\frac{1}{|可能词性|}$/假设其词性可任选.....
2. 平滑问题：采用Laplace/Good-Turing平滑技术，或者收集更数据

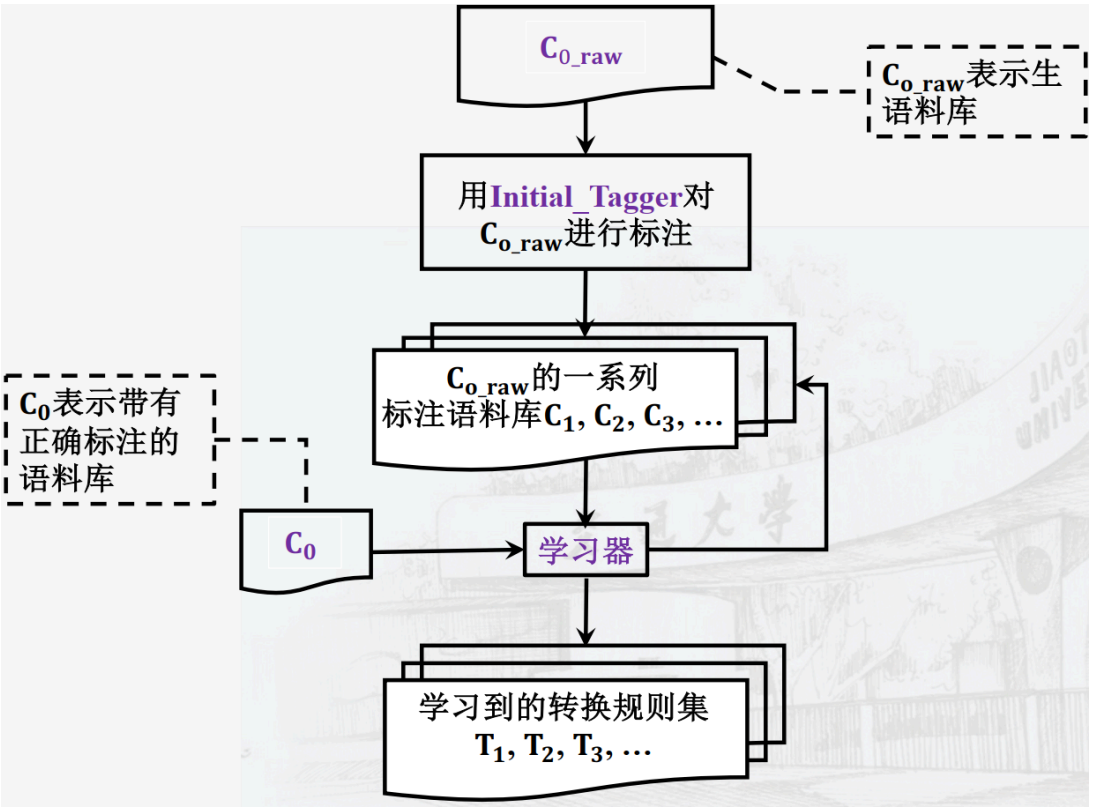
1.3. 基于转换的词性标注

1️⃣概述

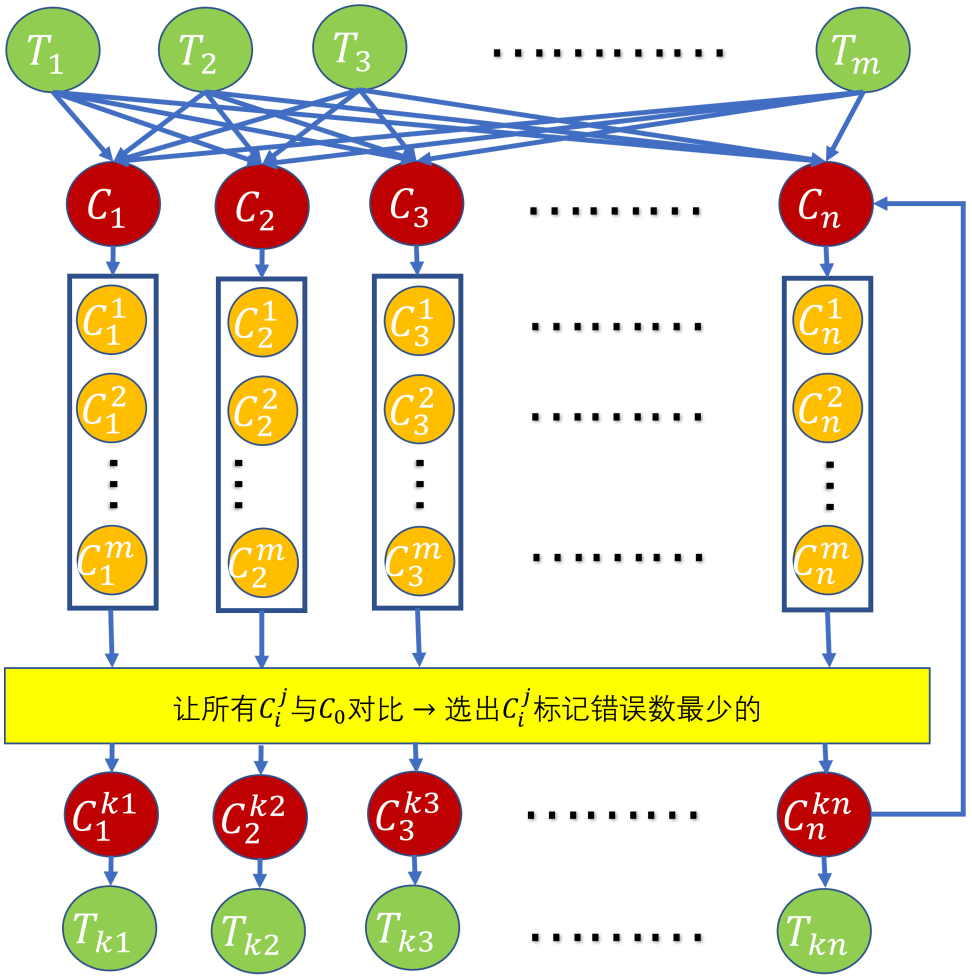
1. 基本思想：
   - 修正规则：正确结果要不断修正得到，修正方式有迹可循
   - 转换规则：计算机可以学习修正过程(记录为转换规则)，并用学得转换规则进行词性标注
2. 转换规则：

组成	含义	示例
改写规则	将一个词性转换成另一个词性	将一个词性从动词改为名词
激活环境	激发改写规则的条件	该词左边第一个词的词性是量词

2️⃣转换规则学习器算法



1. 生成：生成语料库$\mathbb{C}_{0\text{-raw}}$ $\xrightarrow[\text{标注}]{\text{初始标注器}}$有词性标记的语料$\mathbb{C}_1/\mathbb{C}_2/\mathbb{C}_3$……
2. 找错：（语料库$\mathbb{C}_1/\mathbb{C}_2/\mathbb{C}_3$…… $\xleftrightarrow{\text{比较}}$正确标记预料库$\mathbb{C}_0$）$\xrightarrow{\text{得到}}$ $\mathbb{C}_1/\mathbb{C}_2/\mathbb{C}_3$……中词性标注错误数
3. 循环：更新原有$\{C_1, C_2, ..., C_n\}$序列$\xrightarrow{\text{循环}}$得到最终的规则序列$\{T_a, T_b, ..., T_x\}$



- 生成：将所有的规则T_j应用到C_i上$\xrightarrow{\text{得到}}$ mn个C_i^j
- 选取：在每列对比所有

$$(C_i^j \xleftrightarrow{\text{错误数}} C_0) \xrightarrow[(\text{假设是}C_i^k)]{\text{选取错误数最少的}} \begin{cases} \text{认定}T_k\text{为这次学习得到的转换规则} \\ \text{将}C_i^k\text{作为新的待修改语料库} \end{cases}$$

- 迭代：让$\{C_1^{k_1}, C_2^{k_2}, ..., C_n^{k_n}\}$代替$\{C_1, C_2, ..., C_n\}$，重复以上步骤
- 终止：迭代到错误率小于阈值，输出最终的规则序列$\{T_a, T_b, ..., T_x\}$

2. 句法解析

2.1. 基本概念与概述

1 模型描述：

1. 条件：给定一个句子s及其语法G，以$P(t|(s,G))$概率生成分析树t，并且$\sum_t P(t|(s,G))=1$
2. 目的：找出最大化$P(t|(s,G))$的t，即最有可能的句法树

2 与语言模型：

1. 句子概率：语言模型中句子以$P(s)$概率生成，若考虑句法结构则有$P(s)=\sum_t P(s,t)$

2. 最优分析：句法分词旨在最大化$P(t|s)$ $\xrightarrow[P(s)是关于t的常数]{P(t|s)=\frac{P(t,s)}{P(s)}}$变为直接最大化$P(t,s)$

2.2. 概率上下文无关文法

2.2.1. 一些基本概念与假设

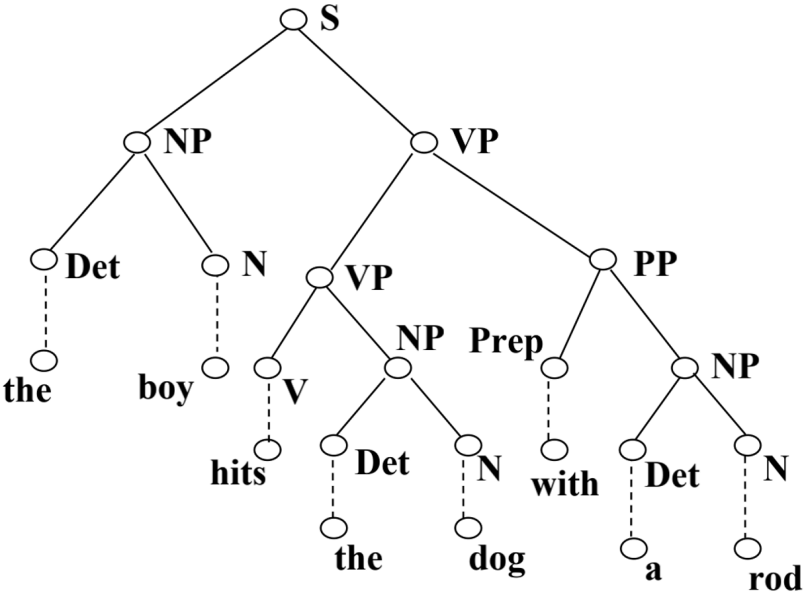
1 句子结构：

结构	含义	示例
非终止符	抽象语法成分，不直接出现在句子中	S(句子)/NP(名词短语)/VP(动词短语)
终结符	实际出现的单词或符号	cat, eats, fish...
规则	非终止符如何进一步被分为符序列/短语	$\mathrm{NP} \xrightarrow{\text{规则}} \mathrm{Det(限定词)+N}$
层次结构	规则逐步展开形成的树状结构	句法树

2 上下文无关文法(CFG)：

Item	含义	例子
CFG	细分非终止符的语法规则集	$\mathrm{NP}\rightarrow\mathrm{Det+N}/\mathrm{VP}\rightarrow\mathrm{V+NP}$
PCFG	为每条规则赋予一个概率	$P(\mathrm{NP}\rightarrow\mathrm{Det+N})=0.9/$ $P(\mathrm{VP}\rightarrow\mathrm{V+NP})=0.1$

3 句法树：用树状结构表示句子内部语法层次



结构	内容
根结点	整个句子

结构	内容
中间结点	包括非终结结点(如NP/VP等语法成分)+终结结点(如N/V等具体单词词性)
叶结点	实际的单词，与终结结点1-1对应

4️⃣模型假设

假设	含义	示例
位置不变	子树概率与在句子中位置无关	名词短语NP在句首/尾时，其结构概率相同
上下文无关	子树概率不依赖不属于该子树词	动词短语VP生成概率不依赖于句中主语NP
祖先无关	子树概率与其父/祖先节点无关	嵌套从句CP生成概率与更高层句法树无关

2.2.2. 概率上下文无关文法基本问题

2.2.2.1. 问题1: 句子概率$P(w_{1:m}|G)$计算

1️⃣Chomsky范式语法

1. 两种规则：

规则	含义	规则概率
二元规则	N^i(一个非终结符)$\xrightarrow{\text{生成}}N^iN^k$(一个非终结符)	$P(N^i\rightarrow N^jN^k\mid G)$
一元规则	N^i(一个非终结符)$\xrightarrow{\text{生成}}w^j$(一个终结符)	$P(N^i\rightarrow w^j\mid G)$

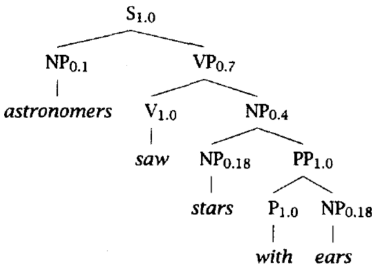
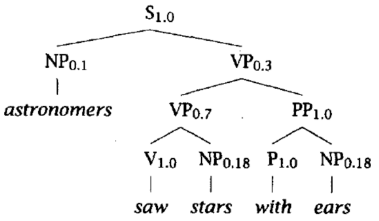
2. 参数空间：对于空间$\{N^1,N^2,...,N^n,w^1,w^2,...,w^V\}$
   - 规则数量：二元规则共n^3个，一元规则共nV个
   - 规则概率：需满足$\sum_{r,s}P\left(N^j\rightarrow N^rN^s\right)+\sum_{k}P\left(N^j\rightarrow w^k\right)=1$

2️⃣句子概率$P(w_{1:m})=\sum_{t:\text{yield}(t)=w_{1m}}P(t)$

项	含义
$P\left(w_{1:m}\right)$	生成句子(词序列)$w_{1:m}=\{w_1,w_2,...,w_m\}$的概率
$t:\text{yield}(t)=w_{1:m}$	句法树的叶节点序列是$\{w_1,w_2,...,w_m\}$
$\sum P(t)$	所有叶节点序列是$\{w_1,w_2,\ldots,w_m\}$的句法树生成的概率总和
$P(t)$	某一句法树生成的概率，为生成句法树所有规则概率的乘积

3️⃣ 示例：考虑句子astronomers saw stars with ears

1. 句法树：

	t_1: with ears修饰stars	t_1: with ears修饰saw
句法树		

2. 规则概率：

S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → *astronomers*	0.1
VP → V NP	0.7	NP → *ears*	0.18
VP → VP PP	0.3	NP → *saw*	0.04
P → *with*	1.0	NP → *stars*	0.18
V → *saw*	1.0	NP → *telescopes*	0.1

3. 生成概率

概率	计算
$P(t_1)$	$1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18$
$P(t_2)$	$1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18$
$P(w)$	$P(t_1) + P(t_2)$

2.2.2.2. 问题2: 最佳句法分析

1️⃣ 问题描述

1. 目的：找到使句子概率最大的句法树，即最优句法树
2. 形式化：
   - 定义$\delta_i(p,q)$：即以非终结符N^i且覆盖字句$w_{p:q}$情况下，最佳解析树的概率
   - 求解方法：动态规划

2️⃣ 类Viterbi风格的动态规划求解

1. 二元规则：

$$\delta_i(p,q) \xleftarrow{N^i_{p:q}\text{子树由}N^j_{p:r}/N^k_{r+1:q}\text{构成}} \max_{j,k,r} \left(P\left(N^i \to N^j N^k\right) \times \delta_j(p,r) \times \delta_k(r+1,q)\right)$$

2. 一元规则：$\delta_i(p,p) \xleftarrow{\text{由叶节点}N^i_{p:p}\text{直接生成}w_j} P\left(N^i \to w_p\right)$

2.2.2.3. 问题2: 文法训练

1. Inside-Outside算法

   1. 内部概率&外部概率

P	公式	含义
内	$\beta_j(p,q)=P\left(w_{p:q} \mid \left(N_{p:q}^j, G\right)\right)$	由$N_{p:q}^j \xrightarrow[\text{语法}G]{\text{生成}} w_{p:q}$的概率
外	$\alpha_j(p,q)=P\left(\left(w_{1:(p-1)}, N_{p:q}^j, w_{(q+1):m}\right) \mid G\right)$	句子pq以外(基于G)的生成概率

   2. 算法公式:

$$P\left(\text{规则}N\to\alpha\text{在}w_{p:q}\right)=\alpha_i(p,q)\times P(N\to\alpha)\times\prod\beta(\text{子结构})\times\frac{1}{P\left(w_{1:m}\right)}$$

2. EM算法：优化规则的概率$P(N\to\alpha)$

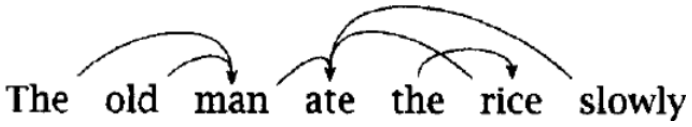
   1. E步：使用Inside-Outside算法，算出规则在未标注语料中出现次数的期望
   2. M步：更新每条规则中的概率为

$$P(N\to\alpha)=\frac{\text{规则}N\to\alpha\text{的期望值}}{\text{所有以}N\text{为左部规则的期望值总和}}$$

2.3. 其它有关内容

1. 依存句法

   1. 含义：在一句话中选一个词为中心，然后用词之间的依存关系来描述语言结构
   2. 表示：带箭头的曲线表示



2. 句法消歧分析：统计方法用概率代替语义规则，自动完成句法树的消歧和选择过程
3. 树库：包含已正确句法分析的句子及其对应解析树的集合，用于构建和训练统计句法分析器