自然语言处理与大模型
笔记

整理人：Your Name

2026 年 1 月 8 日

目录

1 绪论 (2 学时宗成庆)

1.1 基本概念

- **语言与自然语言：**语言是用于表达意思、交流思想的工具，也是一个抽象的数学系统。自然语言是指人类社会发展中自然产生的语言。
- **学科术语辨析：**

NLU (自然语言理解) 侧重探索人类语言认知过程，模仿人类思维，属于 AI 早期核心问题 (1956s)。

CL (计算语言学) 侧重语言学角度，通过建立形式化计算模型来分析语言，偏基础理论 (1960s)。

NLP (自然语言处理) 侧重工程与技术实现，利用计算机对文本进行加工处理 (1970-80s)。

HLT (人类语言技术) 范围更广，涵盖语音、文本等多模态技术 (1980s)。

- **重要历史节点：**
  - **1954 年：**Georgetown 大学与 IBM 用 IBM-701 实现了世界上第一个机器翻译系统。
  - **1956 年：**达特茅斯会议，人工智能（AI）诞生，NLU 成为 AI 核心问题之一。
  - **1966 年：**美国科学院发布 **ALPAC 报告**，认为机器翻译昂贵且质量低，直接导致 NLP 研究进入长达十年的低谷期。

1.2 问题挑战

自然语言处理的核心困难在于歧义性 (Ambiguity) 和不确定性。

1. **歧义现象：**
   - 分词/结构歧义：例如“门把手弄坏了”可切分为“门/把手/弄坏了”或“门/把/手/弄坏了”。
   - 语义歧义：例如“喜欢乡下的孩子”（是喜欢“乡下”这个地方，还是喜欢那里的“孩子”?）。
2. **未知语言现象：**新词（如“内卷”、“给力”）、新含义（“潜水”）、新用法的不规范性。
3. **知识获取困难：**常识往往是隐蔽的，且不同语言间存在概念差异（Culture Gap）。

1.3 技术方法

NLP 技术发展主要经历了三个范式的演变：

1.3.1 1. 理性主义 (Rationalism) - 符号逻辑

- **核心思想：**基于规则 (Rule-based)，通过人类专家编写规则和词典，进行“分析-转换-生成”。
- **代表人物：**Chomsky (句法结构)。
- **优缺点：**可解释性强，但覆盖率低，鲁棒性差，难以处理大规模真实文本，跨语言移植难。

1.3.2 2. 经验主义 (Empiricism) - 统计学习

- **核心思想**：基于统计 (Statistical)，利用大规模真实语料，统计语言规律的可能性（概率）。
- **代表模型**：HMM, SVM, CRF, n-Gram。
- **核心公式 (SMT 噪声信道模型)**：

$$\hat{T} = \arg\max_T P(T|S) = \arg\max_T \frac{P(T) \times P(S|T)}{P(S)} \approx \arg\max_T P(T) \times P(S|T)$$

  其中 $P(T)$ 为语言模型 (Language Model)，保证译文通顺；$P(S|T)$ 为翻译模型 (Translation Model)，保证语义对应。

- **优缺点**：不需要深层句法分析，开发周期短；但难以处理长距离依赖，对语料规模和质量依赖大。

1.3.3 3. 连接主义 (Connectionism) - 深度学习

- **核心思想**：基于神经网络，使用连续向量 (Embedding) 表示语言单位，端到端 (End-to-End) 训练。
- **代表模型**：CNN, RNN, LSTM, Transformer, BERT, GPT 系列。
- **演变路线**：SMT (1989) → NMT (2013) → ChatGPT (2022)。
- **优缺点**：性能卓越，无需繁琐的人工特征工程；但模型如同“黑盒”可解释性差，且对算力和数据要求极高。

1.4 课程内容与考核

- **课程结构**：基本概念 → 基础理论 (统计/N-gram) → 传统技术 (分词/句法/语义) → 深度学习与 LLMs (Transformer/BERT/GPT/多模态) → 应用系统。
- **考核方式**：闭卷考试 (60%) + 课程实践 (40%)。

2 统计学习基础 (4 学时宗成庆)

复习重点: 本章复习重点

1. **概率论与随机过程**：掌握语言的稳态遍历性假设。
2. **齐夫定律 (Zipf's Law)**：理解长尾效应及其对 NLP 数据稀疏的影响。
3. **信息论核心指标**：熵、联合熵、条件熵、互信息、KL 散度、交叉熵、困惑度。
4. **统计学习分类**：生成式模型 (HMM) vs 判别式模型 (CRF, SVM)。
5. **最大熵模型 (MaxEnt)**：原理、约束条件、参数估计算法 (GIS, IIS, L-BFGS)。
6. **条件随机场 (CRF)**：解决标注偏置问题、全局归一化、Viterbi 解码。

2.1 概率论略览

2.1.1 基本概念回顾

- **贝叶斯法则 (Bayes' Theorem)**：NLP 中常用于求解逆向概率（如拼写纠错、机器翻译）。

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \propto P(X|Y)P(Y)$$

  其中 $P(Y)$ 为先验概率（Language Model），$P(X|Y)$ 为似然概率（Observation Model）。

- **最大似然估计 (MLE)**：

$$\hat{\theta} = \arg\max_{\theta} P(D|\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \ln P(x_i|\theta)$$

2.1.2 随机过程 (Stochastic Process)

语言被视为一个随机过程 $\{\xi_t, t \in T\}$。为了使语言的统计规律可被研究，NLP 引入了两个重要假设：

1. **稳态性 (Stationarity)**：语言的统计特性（如词频、共现概率）不随时间推移而改变。即 $P(w_t) \approx P(w_{t+k})$。
2. **遍历性 (Ergodicity)**：单个样本在长时间内的统计特性等于所有样本在同一时刻的统计特性。这意味着我们可以通过在大规模语料库（空间）上的统计来近似语言（时间）的概率分布。

2.2 齐夫定律 (Zipf's Law)

- **定义**：在自然语言语料库中，一个单词出现的频率 f 与其排名 r 成反比。

$$f \times r = C \quad (\text{常数})$$

  或者写成对数形式，$\log(f)$ 与 $\log(r)$ 呈线性关系：$\log f = \log C - \log r$。

- **长尾效应 (Long Tail Effect)：**
  - **高频词：**极少数词（如"the", " 的", " 是"）占据了极高的频率。
  - **低频词（长尾）：**绝大多数词出现的频率非常低。
  - **NLP 挑战：**无论语料库多大，总会遇到未登录词 (Unknown Words/OOV) 和数据稀疏 (Data Sparsity) 问题。

2.3 信息论基础

2.3.1 熵 (Entropy) 与语言熵

衡量随机变量的不确定性。

$$H(X) = -\sum_{x} P(x) \log_2 P(x)$$

- **常识数据：**英文字母的熵约为 4.03 bits，汉字的熵约为 9.71 bits（冯志伟, 1989）。

2.3.2 互信息 (Mutual Information)

- **定义：**$I(X;Y) = H(X) - H(X|Y)$。表示知道 Y 后 X 不确定性的减少量。
- **点式互信息 (PMI)：**用于衡量两个具体事件（如两个词）的相关性。

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **应用：**在汉语分词中，如果两个字 x, y 的 PMI 值很高，说明它们结合紧密，倾向于成词；反之则可能需要断开。

2.3.3 相对熵与交叉熵 (重点辨析)

- **相对熵 (KL Divergence)：**衡量两个分布 P (真实) 和 Q (模型) 的距离。

$$D_{KL}(P||Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

- **交叉熵 (Cross Entropy)：**

$$H(P,Q) = H(P) + D_{KL}(P||Q) = -\sum_{x} P(x) \log Q(x)$$

- **关系与作用：**在机器学习中，最小化交叉熵等价于最小化相对熵。因为真实分布 $P(x)$ 是固定的，其熵 $H(P)$ 是常数。我们通过最小化交叉熵损失函数，使模型分布 $Q(x)$ 逼近真实分布 $P(x)$。

2.3.4 困惑度 (Perplexity)

语言模型评价指标。$PP(S) = 2^{H(L)}$。困惑度越小，模型越好。

2.4 统计学习模型分类

2.4.1 生成式模型 (Generative Model)

- **建模对象**：联合概率分布 $P(X, Y)$。
- **原理**：先对 $P(X, Y)$ 建模，再利用贝叶斯公式求 $P(Y|X)$。
- **典型模型**：n-gram, HMM, Naive Bayes。
- **特点**：收敛速度快，由于学习了数据的生成机制，可以处理隐变量；但很难融合复杂的重叠特征。

2.4.2 判别式模型 (Discriminative Model)

- **建模对象**：条件概率分布 $P(Y|X)$ 或直接学习决策函数 $f(X)$。
- **原理**：直接寻找不同类别之间的最优分类面。
- **典型模型**：SVM, MaxEnt, CRF, Neural Networks。
- **特点**：准确率通常更高，可以灵活利用各种上下文特征（Arbitrary overlapping features）；但训练开销通常较大。

3 统计学习基础 (4 学时宗成庆)

复习重点: 本章核心考点导航

1. **基础理论：**语言的随机过程假设、齐夫定律（长尾效应）。
2. **信息论公式：**必须背诵熵、联合熵、条件熵、互信息的定义式及其推导关系。
3. **模型分类：**生成式 (HMM) vs 判别式 (MaxEnt, CRF, SVM) 的区别。
4. **最大熵模型：**基于约束的熵最大化原理、GIS 算法。
5. **条件随机场 (CRF)：**克服标注偏置问题、特征函数设计、Viterbi 解码算法（重点）。

3.1 概率论与随机过程

3.1.1 语言的随机过程假设

自然语言被视为一个随机过程 $\{\xi_t, t \in T\}$。为了使统计方法可行，NLP 通常引入两个强假设：

1. **稳态性 (Stationarity)：**语言的统计特性（如词频、共现概率）不随时间推移而改变。即 $P(w_t) \approx P(w_{t+k})$。
2. **遍历性 (Ergodicity)：**单个样本在长时间内的统计特性等于所有样本在同一时刻的统计特性。
   - *物理意义*：我们可以通过在大规模语料库（空间）上的统计频率，来近似语言（时间）的真实概率分布。

3.1.2 齐夫定律 (Zipf's Law)

- **定义：**单词在语料库中的频率 f 与其排名 r 成反比。

$$f \times r = C \quad \Rightarrow \quad \log f = \log C - \log r \tag{1}$$

  在双对数坐标系下，表现为一条斜率为负的直线。

- **长尾效应 (Long Tail)：**
  - **高频词：**极少数词（如"the", " 的"）覆盖了绝大部分文本。
  - **低频词：**绝大多数词汇处于“长尾”部分，频率极低。
  - **NLP 困境：数据稀疏 (Data Sparsity)** 是统计 NLP 的核心难题。无论语料库多大，总会有未登录词 (OOV)。

3.2 信息论基础

3.2.1 熵 (Entropy)

衡量随机变量 X 的平均不确定性。

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad \text{(单位: bits)} \tag{2}$$

注：常识数据——汉字的熵约为 *9.71 bits*，英文字母约为 *4.03 bits*。

3.2.2 联合熵 (Joint Entropy)

描述一对随机变量 (X, Y) 平均所需要的信息量。

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y) \tag{3}$$

3.2.3 条件熵 (Conditional Entropy) [补全]

在已知随机变量 X 的情况下，随机变量 Y 的不确定性。

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y|x) \tag{4}$$

链式法则 (Chain Rule)：联合熵等于边缘熵加上条件熵。

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{5}$$

3.2.4 互信息 (Mutual Information)

衡量两个变量的相关性，即知道 Y 后 X 不确定性的减少量。

$$I(X; Y) = H(X) - H(X|Y) \tag{6}$$

$$= \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \tag{7}$$

- **性质**：$I(X; Y) \geq 0$。当且仅当 X, Y 独立时取 0。
- **应用**：在汉语分词中，点式互信息 $PMI(x, y)$ 常用于判断两个字是否成词（结合紧密则 PMI 高）。

3.2.5 交叉熵 (Cross Entropy) 与相对熵 (KL Divergence)

- **相对熵 (KL 散度)**：衡量两个分布 P (真实) 和 Q (模型) 的距离（非对称）。

$$D_{KL}(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

- **交叉熵**：NLP 模型训练常用的损失函数。

$$H(p, q) = H(p) + D_{KL}(p||q) = -\sum_{x} p(x) \log q(x)$$

- **困惑度 (Perplexity)**：$PP = 2^{H(L,q)}$。语言模型评价指标，越小越好。

3.3 统计学习模型分类

- **生成式模型 (Generative Model)：**
  - 建模联合概率 $P(X,Y)$，然后由贝叶斯公式求 $P(Y|X)$。
  - 代表：**HMM (隐马尔可夫)**, Naive Bayes, n-gram。
  - 优点：收敛快；缺点：难以处理复杂的重叠特征。
- **判别式模型 (Discriminative Model)：**
  - 直接建模条件概率 $P(Y|X)$ 或决策函数 $f(X)$。
  - 代表：**CRF (条件随机场)**, MaxEnt (最大熵), SVM, 神经网络。
  - 优点：准确率通常更高，可灵活使用任意全局特征（如“当前词的词性取决于整句话的句法结构”）。

3.4 最大熵模型 (MaxEnt)

3.4.1 核心原理

“承认无知，保留最大不确定性”：在满足所有已知约束条件（特征期望一致）的模型集合中，选择熵最大的那个模型。

3.4.2 数学定义

- **特征函数** $f_i(x,y)$：二值函数，描述 (x,y) 是否满足某条件。
- **约束条件：**模型对特征 f_i 的期望 $E_P(f_i)$ 等于训练数据的经验期望 $\tilde{E}(f_i)$。

$$\sum_{x,y} \tilde{P}(x)P(y|x)f_i(x,y) = \sum_{x,y} \tilde{P}(x,y)f_i(x,y) \tag{8}$$

- **最终形式** (对数线性模型)：利用拉格朗日乘子法求解，得到：

$$P_w(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(x,y)\right) \tag{9}$$

3.4.3 参数估计 (GIS 算法)

通用迭代尺度法 (GIS) 是一种专门用于训练 MaxEnt 参数 λ_i 的迭代算法。

- **核心思路：**通过不断修正参数 λ，使得模型特征期望 E_P 逼近经验期望 $\tilde{E}$。
- **更新公式：**

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \frac{1}{C} \ln \frac{\tilde{E}[f_i]}{E^{(t)}[f_i]}$$

其中 C 是特征总数常数（若不满足常数需引入松弛特征）。

- **局限：**收敛极慢。现代通常使用 L-BFGS (拟牛顿法) 代替 GIS。

3.5 条件随机场 (CRF)

3.5.1 为什么需要 CRF？

1. **对比 HMM**：HMM 假设观察值独立（Output Independence），限制了特征的使用。CRF 没有此假设，可以利用长距离上下文特征。

2. **对比 MEMM (最大熵马尔可夫)**：MEMM 存在标注偏置问题 (Label Bias Problem)。因为 MEMM 进行的是局部归一化（每个状态单独归一），倾向于选择出度较少的状态路径。CRF 进行全局归一化，彻底解决了此问题。

3.5.2 线性链 CRF 模型

最常用于序列标注（如分词、词性标注、NER）。

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{i=1}^{n}\sum_{k} \lambda_k f_k(y_{i-1}, y_i, X, i)\right) \tag{10}$$

- $t_k(y_{i-1}, y_i, X, i)$: 转移特征 (依赖于前后标签关联)。
- $s_l(y_i, X, i)$: 状态特征 (依赖于当前标签与观测值的关联)。

3.5.3 CRF 的解码：Viterbi 算法

问题：给定模型参数 λ 和观测序列 X，找到概率最大的标记序列 $Y^* = \arg\max_Y P(Y|X)$。这是一个动态规划问题。暴力搜索复杂度为 $O(N^T)$，Viterbi 复杂度为 $O(T \cdot N^2)$。

Algorithm 1 Viterbi 算法 (CRF 解码)

Input: 观测序列 X，特征函数集 $\{f_k\}$，权重 $\{\lambda_k\}$，状态集 S (大小 m)

Output: 最优路径 Y^*

定义 $\delta_t(j)$：时刻 t 到达状态 j 的最大非归一化概率（分数）。

定义 $\psi_t(j)$：时刻 t 到达状态 j 的最优路径中，前一个时刻的状态（回溯指针）。

1. 初始化 (t=1)

for $j = 1 \to m$ **do**

  $\delta_1(j) = \sum_k \lambda_k f_k(y_0 = \text{start}, y_1 = j, X, 1)$

  $\psi_1(j) = 0$

end for

2. 递推 (t=2 to T)

for $t = 2 \to T$ **do**

  for $j = 1 \to m$ (当前状态) **do**

    核心公式：寻找前一时刻哪个状态 i 转移到 j 得分最高

    $\delta_t(j) = \max_{1 \le i \le m} [\delta_{t-1}(i) + \sum_k \lambda_k f_k(y_i, y_j, X, t)]$

    $\psi_t(j) = \arg\max_{1 \le i \le m} [\delta_{t-1}(i) + \dots]$

  end for

end for

3. 终止与回溯

最大分数 $P_{max} = \max_j \delta_T(j)$

最优路径终点 $y_T^* = \arg\max_j \delta_T(j)$

for $t = T - 1 \to 1$ **do**

  $y_t^* = \psi_{t+1}(y_{t+1}^*)$

end for

return 序列 $Y^* = (y_1^*, \dots, y_T^*)$

复习重点: Viterbi 总结

Viterbi 本质就是在一个有向图（T 列 x N 行）中寻找一条边权重之和最大的路径。

- δ (Delta): 记录走到当前的“最高分”。
- ψ (Psi): 记录“从哪儿来的”，用于最后倒着找回去。

4 隐马尔可夫模型与条件随机场 (3 学时宗成庆)

4.1 马尔科夫模型

4.2 隐马尔可夫模型（HMM）

4.3 前向算法

4.4 后向算法

4.5 维特比算法

4.6 参数学习

4.7 HMM 在 NLP 中的应用

4.8 条件随机场及其应用

5 语言模型 (6 学时张家俊)

复习重点: 本章核心考点

1. **马尔可夫假设**：理解 N-gram 模型如何利用有限历史近似全概率。
2. **参数估计**：掌握最大似然估计 (MLE) 的计算公式及缺陷。
3. **数据平滑**：能够解释为什么需要平滑（数据稀疏），并掌握加 1 法、Good-Turing、Katz 后退、线性插值的基本原理。
4. **评价指标**：困惑度 (Perplexity) 的定义与计算。

5.1 n 元文法 (N-gram Model)

5.1.1 模型定义

语言模型 (Language Model, LM) 的核心任务是计算一个句子（词序列）的概率 $P(S)$，或者根据上下文预测下一个词 $P(w_i|w_1 \dots w_{i-1})$。

- **链式法则 (Chain Rule)**：

$$P(S) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\dots P(w_m|w_1 \dots w_{m-1}) = \prod_{i=1}^{m} P(w_i|w_1 \dots w_{i-1})$$

- **问题**：随着句子变长，参数空间呈指数级爆炸（L^m），无法直接估计。

5.1.2 马尔可夫假设 (Markov Assumption)

假设当前词出现的概率只与前面 $n-1$ 个词（历史）相关。

$$P(w_i|w_1 \dots w_{i-1}) \approx P(w_i|w_{i-n+1} \dots w_{i-1})$$

5.1.3 常见模型

1. **一元文法 (Unigram, n=1)**：词与词相互独立。

$$P(S) \approx \prod_{i=1}^{m} P(w_i)$$

2. **二元文法 (Bigram, n=2)**：只看前 1 个词（一阶马尔可夫链）。

$$P(S) \approx \prod_{i=1}^{m} P(w_i|w_{i-1})$$

3. **三元文法 (Trigram, n=3)**：只看前 2 个词（二阶马尔可夫链）。

$$P(S) \approx \prod_{i=1}^{m} P(w_i|w_{i-2}w_{i-1})$$

注：为了处理句首句尾，通常引入标记 $\langle BOS \rangle$ *(Begin)* 和 $\langle EOS \rangle$ *(End)*。

5.2 参数估计

5.2.1 最大似然估计 (MLE)

利用训练语料的频率统计来近似概率。对于 n-gram 模型，参数 $P(w_i|w_{i-n+1}^{i-1})$ 计算公式为：

$$P_{MLE}(w_i|w_{i-n+1}^{i-1}) = \frac{Count(w_{i-n+1}^{i-1}w_i)}{Count(w_{i-n+1}^{i-1})} \tag{11}$$

其中：

- 分子：$w_{i-n+1}\dots w_i$ (n 个词) 在语料中共同出现的次数。
- 分母：$w_{i-n+1}\dots w_{i-1}$ (前 n-1 个词) 在语料中出现的次数。

5.2.2 实例计算 (Bigram)

语料库：

$\langle BOS\rangle$ John read Moby Dick $\langle EOS\rangle$
$\langle BOS\rangle$ Mary read a different book $\langle EOS\rangle$
$\langle BOS\rangle$ She read a book by Cher $\langle EOS\rangle$

计算 $P(read|John)$：

$$P(read|John) = \frac{Count(John, read)}{Count(John)} = \frac{1}{1} = 1$$

计算 $P(a|read)$：

$$P(a|read) = \frac{Count(read, a)}{Count(read)} = \frac{2}{3}$$

5.3 数据平滑 (Data Smoothing)

5.3.1 为什么要平滑？

数据稀疏 (Data Sparsity) 是统计 NLP 的最大挑战。

- **零概率问题**：如果测试集中出现了训练集中未见过的 N-gram，MLE 会赋予其 0 概率，导致整个句子概率为 0。
- **核心思想**：“劫富济贫”。调低高频事件的概率，将其分配给零概率或低频事件。

5.3.2 常见平滑方法

1. 加 1 法 (Add-one / Laplace Smoothing) 最简单粗暴的方法，假设每个 N-gram 至少出现一次。

$$P_{Add1}(w_i|w_{i-1}) = \frac{1 + C(w_{i-1}w_i)}{|V| + \sum_w C(w_{i-1}w)} = \frac{1 + C(w_{i-1}w_i)}{|V| + C(w_{i-1})} \tag{12}$$

- $|V|$ 是词表大小。
- **缺点**：当 $|V|$ 很大时，分配给未见词的概率过多，导致模型效果很差。

2. Good-Turing 估计法 (重点) 平滑算法的基础。利用“出现 r 次的 n-gram 的个数 n_r”来重新估计概率。

- **原理**：用出现 $r+1$ 次的事件数量来修正出现 r 次的概率。
- **修正计数**：$r^* = (r+1)\frac{n_{r+1}}{n_r}$
- **概率公式**：$P_r = \frac{r^*}{N}$ （N 为总样本数）
- 对于未见事件 ($r=0$)，分配概率 $P_0 = \frac{n_1}{N}$。

3. Katz 后退法 (Back-off)

- **思想**：如果你有可靠的高阶 n-gram 统计（次数 $> K$），就用高阶的；如果数据不够（次数 $\leq K$ 或为 0），就“后退”到低阶模型（如 trigram $\rightarrow$ bigram）。
- 需要引入归一化因子 α 以保证概率和为 1。

4. 线性插值法 (Linear Interpolation) 简单有效，广泛使用（如 Jelinek-Mercer 平滑）。

- **思想**：混合高阶和低阶模型，给它们不同的权重 λ。
- **公式 (Trigram)**：

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_3 P(w_n|w_{n-2}w_{n-1}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_1 P(w_n)$$

  其中 $\sum \lambda_i = 1$。
- 参数 λ 通常通过在**留存数据 (Held-out data)** 上最大化似然度（使用 EM 算法）来获得。

5.4 神经网络概述

5.5 前馈神经网络语言模型 (FFNN-LM)

5.6 循环神经网络语言模型 (RNN-LM)

5.7 长时短时记忆网络语言模型 (LSTM-LM)

5.8 注意力机制语言模型

6 文本表示 (3 学时张家俊)

6.1 向量空间表示模型

6.2 深度学习表示模型

6.3 词语表示

6.4 句子表示

6.5 文档表示

7 词法分析与句法分析 (6 学时宗成庆)

7.1 汉语自动分词方法

7.2 分词结果评价

7.3 未登录识别

7.4 词性标注

7.5 子词切分

7.6 短语结构分析方法

7.7 依存关系分析方法

7.8 句法分析结果评价

8 篇章分析与语义分析 (3 学时宗成庆)

8.1 篇章表示理论

8.2 篇章关系分析

8.3 篇章关系应用

8.4 语义网络

8.5 语义角色标注

9 机器翻译 (4 学时张家俊)

9.1 机器翻译概论

9.2 统计机器翻译

9.3 神经机器翻译

9.4 译文质量评估

10 文本分类和聚类 (3 学时张家俊)

10.1 基于统计学习的文本分类

10.2 基于深度学习的文本分类

10.3 文本分类性能评估

10.4 文本相似度计算

10.5 文本聚类算法

10.6 文本聚类性能评估

11 信息抽取 (3 学时张家俊)

11.1 信息抽取概述

11.2 命名实体识别

11.3 实体消歧

11.4 关系抽取与知识图谱

11.5 事件抽取与事件图谱

12 预训练语言模型 (4 学时张家俊)

12.1 词向量表示回顾

12.2 ELMo 模型预训练语言模型

12.3 BERT 模型预训练语言模型

12.4 GPT 模型预训练语言模型

13 大语言模型：训练与对齐 (6 学时张家俊)

13.1 大语言模型概述

13.2 大语言模型训练数据

13.3 语言模型训练方法

13.4 大语言模型指令微调

13.5 基于人类反馈的对齐

14 多语言大模型 (3 学时张家俊)

14.1 多语言大模型方法

14.2 多语言大模型训练

14.3 多语言大模型对齐

15 提示学习 (3 学时张家俊)

15.1 基础提示学习方法

15.2 上下文学习

15.3 思维链提示

16 检索增强的大语言模型 (2 学时张家俊)

16.1 大语言模型问题分析

16.2 检索增强的大语言模型

17 课程总结与展望 (3 学时宗成庆)

17.1 课程内容回顾

17.2 学科现状分析

17.3 未来展望

17.4 关于课程考核

18 考核 (2 学时宗成庆)

18.1 考试

附录：常用 LaTeX 模板

1. 学术三线表模板

表 1: 不同模型在测试集上的性能对比（示例）

模型	准确率 (Accuracy)	精确率 (Precision)	召回率 (Recall)	F1 值
SVM	85.2%	84.1%	83.5%	83.8%
Bi-LSTM	89.5%	88.2%	89.0%	88.6%
BERT-Base	**92.3%**	**91.8%**	**92.1%**	**91.9%**

2. 图片插入模板 (已注释)

3. NLP/模式识别常用数学公式模板

常用符号： 数据集 $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$，损失函数 $\mathcal{L}(\theta)$，参数 $\boldsymbol{\theta}$。

贝叶斯公式 (Naive Bayes)：

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \propto P(c)\prod_{i=1}^{d} P(x_i|c)$$

Softmax 函数： 适用于多分类输出层：

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

自注意力机制 (Self-Attention)： Transformer 的核心公式：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

交叉熵损失 (Cross Entropy Loss)：

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

HMM 前向算法递推：

$$\alpha_t(i) = P(o_1, \ldots, o_t, q_t = i|\lambda) = \left[\sum_{j=1}^{N} \alpha_{t-1}(j)a_{ji}\right] b_i(o_t)$$