# 自然语言处理

从形式语言理论到大规模预训练模型

MarkZZZ WeChat: MarkZZZ20XX

## 课程简介

自然语言处理（Natural Language Processing, NLP）是人工智能与计算语言学的核心交叉领域，研究如何使计算机理解、生成和分析人类语言。本课程以博士研究生为目标受众，系统讲授 NLP 的数学理论基础、核心算法与前沿模型。

课程从形式语言理论出发，建立正则语言、上下文无关文法与 CYK 算法的严格数学框架；深入讲授语言模型的概率论基础（n-gram、Kneser-Ney 平滑、香农熵与困惑度的信息论解释）；在词表示部分，完整推导 Word2Vec 的负采样目标函数与 GloVe 的矩阵分解视角。

序列建模部分涵盖 HMM 的前向-后向算法与 Viterbi 解码的完整推导、条件随机场（CRF）的对数线性模型理论；神经序列模型部分从 RNN 梯度消失的理论分析出发，深入 LSTM 遗忘门的设计动机与 GRU 的等价性。

Attention 机制与 Transformer 架构是课程的核心章节，涵盖自注意力的复杂度分析、位置编码的设计理论与多头注意力的表达能力。预训练语言模型部分系统讲授 BERT（MLM/NSP 目标）、GPT（自回归建模）、RoBERTa 与 T5 的统一框架。机器翻译章节提供 IBM Model 1 的完整 EM 推导与 BLEU 评价指标的统计基础。

课程后半部分涵盖信息抽取（BiLSTM-CRF、关系抽取、事件检测）、问答与阅读理解（SQuAD、DPR 密集检索）、情感分析与数据偏见量化，以及可解释 NLP 与对抗鲁棒性。所有核心定理均给出完整证明，算法分析包含复杂度论证，为从事 NLP 研究的学生提供坚实的理论基础。

## 适合人群

- 计算机科学、人工智能、语言学等专业的博士研究生
- 从事自然语言处理、计算语言学和大模型研究的科研人员
- 希望系统掌握 NLP 理论基础与前沿方法的高级工程师
- 对语言模型、信息抽取、机器翻译等方向有深入研究需求的研究者

## 前置知识

- **概率论与数理统计**：条件概率、贝叶斯定理、最大似然估计、EM 算法基础
- **线性代数**：矩阵运算、奇异值分解（SVD）、矩阵微积分
- **信息论**：熵、互信息、KL 散度、困惑度
- **机器学习基础**：逻辑回归、梯度下降、正则化、神经网络基础
- **深度学习**：反向传播、PyTorch/TensorFlow 编程能力
- **离散数学**：图论基础、形式语言与自动机（有益但非强制）
- **编程能力**：Python、NumPy/Pandas，能够实现并调试 NLP 模型

# 1 课程内容

| 讲次 | 主题 | 内容概要 |
|---|---|---|
| 1 | 语言学基础与形式语言 | 乔姆斯基层级 (Chomsky Hierarchy)、正则语言与有限自动机、上下文无关文法（CFG）与下推自动机、CFG 的 Chomsky 范式（CNF）转换算法、CYK 动态规划算法（$O(n^3\|G\|)$ 推导）、歧义性理论、句法分析树与成分结构、依存文法基础 |
| 2 | 语言模型基础 | 语言模型的概率定义、n-gram 模型与马尔可夫假设、最大似然估计、数据稀疏与平滑策略、Laplace 平滑、Jelinek-Mercer 插值、Kneser-Ney 平滑的完整推导、困惑度 (Perplexity) 的信息论基础、香农熵与交叉熵、语言模型的极限定理 |
| 3 | 词表示学习 | 词袋模型 (Bag-of-Words)、TF-IDF 权重的信息论解释、词-文档矩阵的 SVD 分解（LSA）、Word2Vec 的 Skip-gram 与 CBOW 目标函数推导、负采样 (Negative Sampling) 的完整梯度推导、层次 Softmax、GloVe 的加权最小二乘目标与矩阵分解等价性证明、词向量的代数性质（语义偏移）、fastText 与子词表示 |
| 4 | 序列标注：HMM 与 CRF | 隐马尔可夫模型（HMM）的形式定义与三个基本问题、前向算法（Forward Algorithm）完整推导、后向算法（Backward Algorithm）、前向-后向算法用于参数估计（Baum-Welch）、Viterbi 算法（完整动态规划推导与正确性证明）、MEMM 与标注偏置问题、条件随机场（CRF）的对数线性模型定义、全局归一化与 Viterbi 解码 |
| 5 | 神经序列模型 | 循环神经网络（RNN）的形式定义与展开图、BPTT（Backpropagation Through Time）推导、梯度消失/爆炸的理论分析（Jacobian 矩阵谱半径定理）、LSTM 的遗忘门/输入门/输出门完整推导、细胞状态的梯度流分析、GRU 的简化结构与 LSTM 等价性讨论、双向 RNN 与堆叠 RNN |

| 讲次 | 主题 | 内容概要 |
| --- | --- | --- |
| 6 | Attention 机制与 Transformer | 加性注意力 (Additive Attention)、乘性注意力 (Multiplicative Attention)、缩放点积注意力（Scaled Dot-Product Attention）的温度参数分析、自注意力 (Self-Attention) 的时间/空间复杂度 $O(n^2d)$ 分析、多头注意力 (Multi-Head Attention) 的子空间投影理论、位置编码（正弦编码与可学习编码）、Transformer 完整架构（编码器/解码器）、层归一化 (Layer Normalization) 与残差连接 |
| 7 | 预训练语言模型 | 预训练-微调范式 (Pre-train & Fine-tune) 的统计学基础、BERT：掩码语言模型（MLM）目标函数、下句预测（NSP）、词片标记化（WordPiece）、GPT 系列：自回归语言建模目标、因果自注意力掩码、RoBERTa 的训练策略改进分析、ALBERT 参数共享、T5 的统一 seq2seq 框架与任务前缀设计、Scaling Law 与涌现能力 |
| 8 | 机器翻译 | 统计机器翻译（SMT）基础：噪声信道模型、IBM Model 1 的 EM 算法完整推导（E 步/M 步/收敛性）、词对齐模型（IBM Model 2-5）、神经机器翻译（NMT）：seq2seq 架构、注意力对齐机制、Transformer 翻译模型、束搜索（Beam Search）与长度归一化、BLEU 评价指标的定义、精确率 n-gram 修正与简短惩罚、BLEU 的局限性 |
| 9 | 信息抽取 | 命名实体识别（NER）：序列标注框架（BIO/BIOES 标注方案）、BiLSTM-CRF 模型推导、CRF 层的前向计算与 Viterbi 解码、关系抽取：pipeline 方法与联合学习、事件检测与论元抽取、共指消解（Coreference Resolution）：基于神经的聚类方法、零指代分析 |
| 10 | 问答与阅读理解 | 机器阅读理解（MRC）任务分类、SQuAD 数据集基线（抽取式 QA、span 预测）、BERT 微调的 QA 头设计、开放域问答（Open-Domain QA）、检索-阅读两阶段框架、密集段落检索（DPR）：双编码器训练（对比学习目标）、RAG（检索增强生成）的概率框架 |

| 讲次 | 主题 | 内容概要 |
|---|---|---|
| 11 | 情感分析与偏见 | 文档级情感分析、方面级情感分析（ABSA）、隐式情感识别、情感词典与监督学习方法、深度情感模型（TreeLSTM）、NLP 中的数据偏见：标注偏见、选择偏见、测量偏见、偏见量化指标（WEAT、StereoSet、WinoBias）、偏见缓解方法（数据增强、对抗训练、去偏投影） |
| 12 | 可解释 NLP 与鲁棒性 | 模型解释方法分类（事后解释 vs 内置解释）、基于梯度的显著图方法、LIME 局部线性近似框架、SHAP（Shapley 值）的博弈论基础、Attention 可解释性的争论（*Attention is not Explanation*）、对抗样本：文本级攻击（字符/词/句子级别）、对抗训练（PGD 变体在 NLP 中的适配）、认证鲁棒性的语言模型 |

## 2 参考文献

## 参考文献

[1] Jurafsky, D. & Martin, J. H. (2023). *Speech and Language Processing*, 3rd Edition (Draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/ ——本课程主要教材，涵盖 *NLP* 经典方法与神经方法，免费在线获取。

[2] Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. ——统计 *NLP* 的经典教材，第 *6* 章统计推断、第 *10* 章语言模型尤为重要。

[3] Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Morgan & Claypool. ——神经 *NLP* 的系统介绍，从词向量到 *RNN/CNN* 均有严格推导。

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. ——*Transformer* 架构的原始论文，*NLP* 最重要的论文之一。

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. ——*BERT* 原始论文，开启预训练-微调范式的新时代。

[6] Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. ——*GPT-3* 论文，展示大规模语言模型的涌现能力与少样本学习。

[7] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26. ——*Word2Vec* 负采样方法的原始论文，词嵌入领域的里程碑。

[8] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of EMNLP 2014*, 1532–1543. ——*GloVe* 词向量的原始论文，矩阵分解视角的词表示。

[9] Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML 2001*, 282–289. ——*CRF* 的原始论文，序列标注的理论基础。

[10] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 263–311. ——*IBM* 翻译模型 *1-5* 的原始论文，*EM* 算法在 *NLP* 中的经典应用。

[11] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27. ——*seq2seq* 模型的原始论文，神经机器翻译的基础架构。

[12] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of ICLR 2015*. ——注意力机制在 *NMT* 中的原始论文，开创注意力时代。

[13] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*. ——*RoBERTa* 的系统性消融实验，揭示 *BERT* 预训练策略的关键因素。

[14] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of KDD 2016*, 1135–1144. ——*LIME* 解释框架的原始论文。

[15] Lundberg, S. M. & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. ——*SHAP* 方法的原始论文，基于 *Shapley* 值的统一解释框架。