



课程重要知识点回顾

邵研

2023年11月20日



- 自然语言处理定义：
 - NLP, 计算语言学, 语言技术;
- NLP典型应用：
 - 机器翻译, 自动摘要, 文本分类与信息过滤, 信息检索, 自动问答, 文本挖掘, 情感分析;
- NLP的难点：
 - 歧义, 形态, 句法, 语义, 语用;
- NLP发展历程：
 - 大模型时代



- 基于规则/语言学理论的建模方法为什么无法解决大部分自然语言处理的问题？



- HMM原理介绍：
 - 生成式序列模型，生成概率、转移概率；
- Viterbi算法：
 - 动态规划算法，时间复杂度；
- 最大似然法：
 - 统计方法，拉普拉斯平滑；
- NLP应用举例
 - 词性标注任务。



- HMM为什么叫隐马尔可夫模型？它与马尔可夫链有什么关系？
- 一阶HMM为什么要做独立性假设？高阶HMM会有什么问题？
- HMM有哪些局限性？
- 如果没有已知的标注数据，如何用无监督的方式得到HMM模型的参数？
- 哪些任务适合用HMM来做建模？哪些不适合？



- 线性模型介绍：
 - 特征, 权重, 偏置, 线性回归, 超平面;
- 线性感知器：
 - 在线学习, 学习速率, 损失函数, 多分类线性感知器;
- 逻辑回归：
 - 交叉熵, 梯度下降;
- SVM：
 - 支持向量, 最优决策边界, 线性不可分问题, 核函数, 特征工程
- 条件随机场：
 - 生成模型与判别模型



- 不同的线性模型之间有什么区别？各有什么优劣？
- 逻辑回归需要解决什么问题？引入逻辑函数的目的是什么？
- 特征工程应该怎么去做？
- HMM和CRF的区别和联系是什么？HMM有什么优势？
- 线性模型的优势是什么？



- 词嵌入：
 - 自监督学习，预训练，向量化；
- 卷积神经网络：
 - 卷积核，通道，池化，一维CNN，二维CNN；
- 循环神经网络：
 - 梯度爆炸，梯度消失，门结构，RNN-CRF；
- 注意力机制和Transformer：
 - KQV，自注意力机制，多头注意力机制；
- 深度学习框架：
 - Numpy，动态计算图，静态计算图，PyTorch，TensorFlow



- 词嵌入的目的是什么？有没有什么局限性？
- CNN和RNN的区别与联系是什么？各有什么优劣？
- 引入注意力机制要解决什么问题？为什么要使用多头注意力机制？有没有什么潜在问题？
- Transformer的主要优势是什么？最大的局限性是什么？如果没有位置编码会怎样？
- Numpy与TensorFlow和PyTorch的最大区别是什么？



- 图模型介绍：
 - 常见的图结构，图，节点，边；
- 随机游走：
 - PageRank, TextRank, 阻尼系数；
- 标签传播算法：
 - 种子样本，半监督算法；
- 图神经网络：
 - 稀疏邻接矩阵，表示学习，图卷积，图模型中的池化和注意力机制
- NLP应用实例：
 - 图构建方法，文本分类，知识图谱问答



- NLP中什么样的问题适合用图模型来建模?
- 利用循环神经网络作为图神经网络的信息传递函数会有什么问题?
- 图神经网络与其他面向序列的神经网络相比有什么优点?
- 图神经网络存在哪些问题?



- 强化学习介绍：
 - 环境状态, 智能体, 奖励, 策略, 打折率, 价值函数;
- 马尔可夫决策过程：
 - MRP, MDP, 蒙特卡洛方法, 贝尔曼方程 (等式), 贝尔曼期望 (最优) 方程, 最优策略, 策略迭代, 价值迭代;
- 时序差分算法：
 - 无模型强化学习, 时序差分误差, SARSA算法, Q-learning算法;
- DQN算法和策略梯度算法：
 - 连续状态空间, 经验回收, 目标网络, REINFORCE, Actor-Critic, TRPO, PPO
- NLP应用实例：
 - 实时机器翻译, 任务对话系统



- 为什么要引入折扣率？折扣率取0和取1时代表什么？
- 策略迭代与价值迭代的区别与联系是什么？
- SARSA算法与Q-learning算法的主要区别是什么？各有什么优劣？
- 在DQN中引入经验回收和目标网络的目的是什么？
- 价值优化和策略优化的联系和区别是什么？
- NLP任务中使用强化学习最大的限制是什么？
- 强化学习与隐马尔可夫模型有什么联系和区别？



- 不同的分类任务：
 - 二分类, 多分类, 层次分类, 多标签分类;
- 分类任务的评价方式：
 - 准确率, 召回率, F值, 宏平均, 微平均;
- 分类方法：
 - N-gram, TF-IDF, 统计方法, 深度学习方法;
- 聚类任务：
 - 簇
- 聚类方法：
 - K-MEANS, AGNES, DBSCAN, OPTICS, LDA, VAE



- 为什么需要用F值来衡量分类效果？只用准确率或者召回率会有什么问题？
- 分类模型与序列标记模型有什么区别和联系？
- 不同文本聚类算法的优势和局限性是什么？



- 机器翻译：
 - Seq2Seq, 源语言, 目标语言, 统计机器翻译, 神经网络机器翻译, Transformer;
- 机器翻译挑战：
 - 文化差异, 谚语习语, 长距离依赖;
- 文本摘要类型：
 - 抽取式, PageRank, 压缩式, 生成式, Seq2Seq, 有监督, 无监督
- 评价指标：
 - BLEU, ROUGE;



- 序列标记任务和Seq2Seq的任务主要区别是什么？机器翻译和文本摘要可否用序列标记任务实现？
- 利用BLEU或者ROUGE来衡量机器翻译和文本摘要的效果会有什么问题？



- 问答系统：
 - 知识问答，知识库检索，符号推理，向量检索；
- 问答难点：
 - 语义歧义，结构歧义；
- 对话系统基本架构：
 - 语音识别，自然语言理解，对话管理，自然语言生成，语音合成；
- 对话系统实现方式：
 - 基于检索，基于生成
- 对话管理：
 - 闲聊机器人，面向任务的聊天机器人，有限状态机，填槽



- 问答系统和对话系统有什么区别和联系？
- 对话和问答系统中基于检索和基于生成的方法各有什么优劣？
- 在构建闲聊机器人和任务型聊天机器人时，优化目标有什么不同？
- 完整的对话系统分为哪些主要模块？能否用端到端的方式实现？



- 知识图谱概念：
 - 有向图，实体，关系，三元组，本体；
- 知识图谱规范：
 - W3C, RDF;
- 知识抽取：
 - 本体构建，信息抽取，实体识别，关系抽取，实体链接，事件抽取；
- 知识图谱应用：
 - 检索，KBQA，图谱表示学习



- 以知识图谱为代表的结构化数据，相比非结构化数据有什么优势和不足？
- 知识图谱的本体有什么作用？如果不定义本体会怎么样？
- 实体识别中需要解决哪两个子问题？为什么还需要实体链接（对齐）？
- 大模型时代知识图谱的价值是什么？知识图谱与大模型如何实现相互促进与融合？



- 大语言模型介绍：
 - 词嵌入, 语言模型, 迁移学习;
- Elmo模型：
 - 双向LSTM, CNN字符编码, 多层特征融合;
- BERT模型：
 - Transformer, WordPiece Tokenizer, 随机词掩码, 是否是下一句;
- BERT模型的变体：
 - BERT模型的问题, 更多预训练数据, 改进预训练目标, 获取更多监督信息, roBERTa, ALBERT, Electra, BART, T5, ERNIE
- NLP应用实例：
 - BERT-(LSTM)-CRF



- 预训练模型的效果与哪些因素相关？
- BERT模型的主要问题是什么？其他模型都是怎么做优化的？
- 大规模预训练模型在实际应用中有哪些挑战？
- 为什么小模型难以使用更多的预训练数据来提升效果？



- GPT基础模型：
 - GPT的预训练目标，下游任务，GPT与BERT的区别；
- GPT模型演进历程：
 - 零样本学习，上下文学习，模型规模与模型能力，Scaling Law，GPT3存在的问题，开源平替；
- ChatGPT：
 - RLHF，对齐，“对齐税”
- GPT4：
 - 多模态能力，人工通用智能，存在的问题
- NLP应用实例：
 - 传统文本生成-NLP任务，代码能力，辅助规划和决策，科研辅助，创意生成（AIGC）



- GPT基础模型与BERT相比有哪些优势和劣势？
- 什么是零样本学习？零样本有什么优势？为什么零样本学习很有挑战性？
- 零样本学习，单样本学习和少样本学习有什么区别和联系？
- Scaling Law对于大模型的训练有什么重要意义？
- 大语言模型的快速发展给人类社会带来了哪些挑战？



- 提示词工程：
 - 参数设定, 提示词要素, 提示词设计技巧, 思维链;
- 检索增强：
 - 大语言模型的时效性, 检索模块, 外部知识源;
- 开源生态：
 - LLaMA系列, GLM系列, Baichuan系列
- 大模型微调：
 - Prefix Tuning, P-Tuning, Adaptor, Delta-Tuning, LoRA, FireFly, Vicuna,
- NLP应用实例：
 - 展厅导览机器人



- 提示词工程能解决哪些问题？不能解决哪些问题？
- 为什么思维链能够提升大模型的效果？使用思维链有哪些缺点？
- 相比全参数微调，使用Prefix Tuning, LoRA等方法有什么优势和劣势？
- 我们为什么需要本地微调大语言模型？直接使用商业模型的接口（如OPENAI GPT API）会有哪些问题？



- 模型压缩和加速
- 预训练语料的挖掘和构建
- 模型的可解释性
- 跨语言研究
- 多模态大模型
- 大模型与AI for Science
- 大模型的评估
- 模型的自适应和个性化
- 具身智能



- 你对哪一个方向最感兴趣？请基于你阅读相关文献，简述该方向的发展现状和未来展望。



问题?

