**一、简答题（共 20 分，每题约 1.3 分，共 15 小题）**

1. **自然语言处理的难点与“歧义”现象**

语言具有多层次结构，组合爆炸，需要上下文信息（同一词/句子在不同上下文有不同意思），容易有歧义等

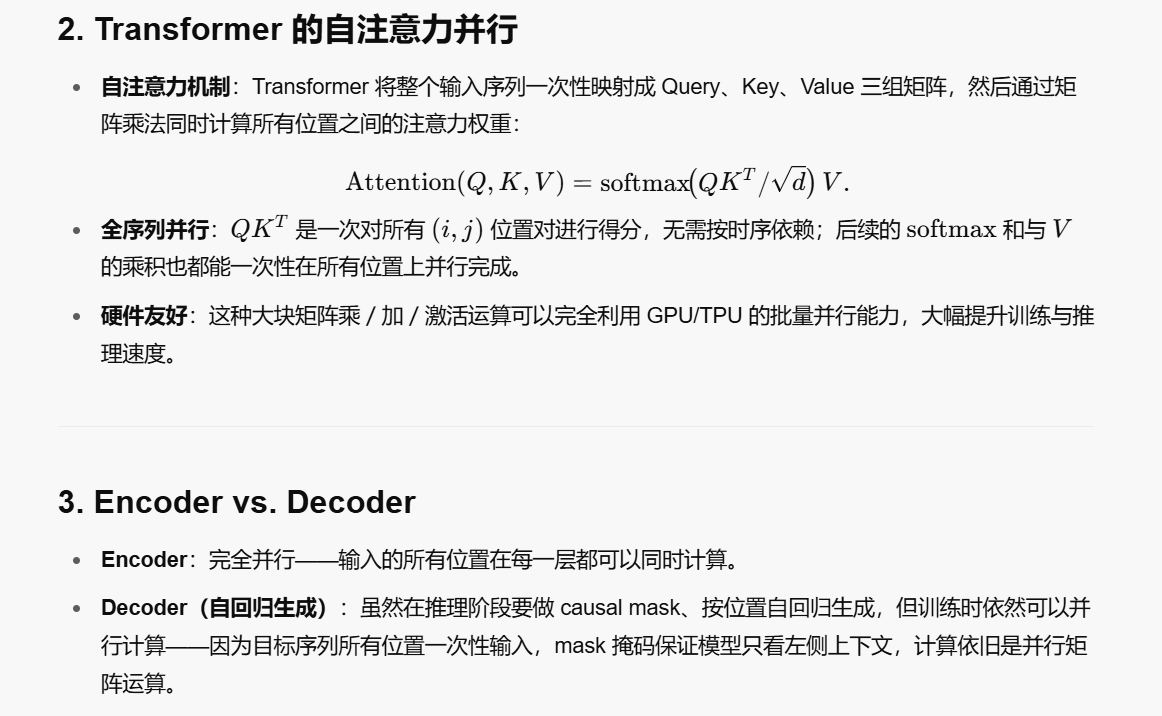
1. **自然语言处理的发展历史**

发展早期（50-60）注：1950 图灵测试 符号学派，随机学派 布朗语料库

理性主义（70-80） 基于逻辑，规则，随机

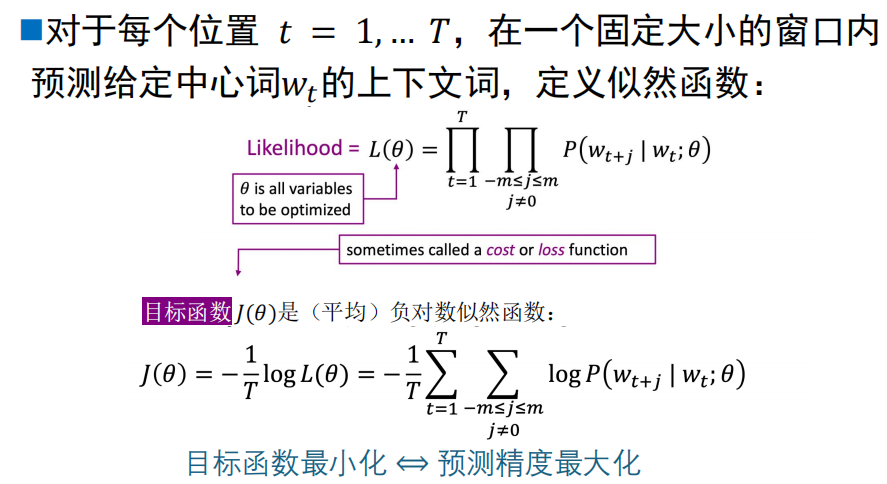
经验主义（90-21世纪）机器学习和数据驱动

深度学习（2006-2018）



大模型（2019-至今）

1. **自然语言处理的三大主要流派**
   * **符号主义、统计方法、联结主义**
2. **分布式表示的本质**
   * 将文本表示为**低维的稠密向量**，并在低维表示空间中利用表示向量之间的计算关系体现**文本之间的语义关联**。 用词在文本中的邻居表示它。
   * 优势在于**平滑泛化、参数共享、缓解稀疏**。
3. **列举两种经典的词嵌入算法**
   * **Word2Vec**（CBOW 与 Skip-gram），Glove



* + （可额外： ELMo、BERT Embedding 等）

提示：One-hot不好，因为词的向量太长，容易稀疏；此外，无法捕捉到词之间的语义联系。

1. **三种词汇语义表示方式**
   * **One-hot**（高维稀疏向量）
   * **分布式词向量**（Word2Vec、GloVe）
   * **基于知识库**（WordNet 同义词集、概念图）
2. **预训练自然语言处理方法的主要步骤**
   * **无监督/自监督预训练**：大规模未标注语料上 Masked LM、Autoregressive LM、Permutation LM 等任务
   * **下游任务微调**：用少量有标注数据 Fine-tune 整个模型或部分层
3. **可以视作序列标注的任务（至少举两例）**
   * **词性标注**（POS tagging） Part of Speech 词性标注
   * **命名实体识别**（NER） name entity recognition
   * **中文分词**
4. **ChatGPT 的训练过程**
   * **预训练**：大规模网络文本上自回归 Language Model
   * **监督微调**（SFT）：用人类示例对话数据微调
   * **奖励建模**（RM）：人类标注对话优劣，训练奖励模型
   * **强化学习**（RLHF）：用 PPO 等方法最大化奖励模型分数
5. **“困惑度”（Perplexity）如何描述语言模型在测试集上的表现**

值越**低**表明模型对测试文本的预测越精准。

1. **BERT 的两个预训练任务**
   * **Masked Language Modeling (MLM)**：随机遮蔽输入词，让模型预测被遮蔽的词
   * **Next Sentence Prediction (NSP)**：下一句预测

注：下一词预测是专属自回归生成模型的，例如GPT，对于Bert，是双向的自编码结构，需要同时利用上下文的信息来预测被遮蔽的词，准确说，它需要引入掩码语言模型（随机遮蔽一些词，利用上下文信息补全），这一步是为了获得句子之内的词级别表示，同时Bert还引入了下一句预测（判断两句是否相邻）捕捉句子间的语篇/连贯信息。

1. **语言模型的本质**
   * 对自然语言序列建模，估计词序列的**联合概率**或条件概率分布
2. **n-gram 中 n 的含义**
   * 序列中**连续出现的 n 个项**（通常是**词或字**）的组合
3. **社区问答任务的分类**
   * **问答对检索**（检索最佳回答）
   * **回答生成**（自动生成回答）
   * **问句意图分类**
   * **回答质量评估**

**二、判断题（共 10 分，每题 2 分）**

判断下列说法是否正确，正确写“T”，错误写“F”。

1. **LSTM 是 RNN 的一种。**  
   **T**
2. **RNN 能很好地处理变长序列任务。**  
   **T**
3. **TextRank 是一种生成式信息抽取算法。**  
   **F** （它是一种基于图的**抽取式**算法，常用于关键词/摘要抽取）
4. **BART 是一个 Decoder-Only 模型。**  
   **F** （BART 是 Encoder-Decoder 架构）
5. **依存句法分析的输出是一棵树状结构。**  
   **T**

**三、大题（共 30 分，每题 10 分）**

1. **自然语言处理研究什么？试分别从“理解”和“生成”两个角度简要阐述。**
   * **理解**：语法分析、句法分析、语义表示、阅读理解等
   * **生成**：机器翻译、摘要生成、对话生成、文本风格化等
2. **预训练语言模型的三种主要范式**
3. **举例说明大语言模型中的 Zero-Shot、One-Shot、Few-Shot 学习。**
   * **Zero-Shot**：直接给模型一个新任务的自然语言提示（prompt），不提供任何示例，模型凭预训练知识完成任务，如 “Translate to French: Hello.”
   * **One-Shot**：提示中给出 1 个示例，如 “English→French: Hello → Bonjour. Goodbye → ?”
   * **Few-Shot**：提示中给出少量示例（通常 5~20 条），如 “Cat → 猫；Dog → 狗；Bird → ?”

**四、综合题（共 20 分）**

1. **大语言模型的主要应用场景**（8 分）
   * 文本生成（对话、写作辅助、代码生成）
   * 机器翻译
   * 问答系统与检索增强
   * 文本摘要与信息抽取
   * 风险检测与舆情分析
2. **大语言模型带来的社会伦理问题**（6 分）
   * **偏见与歧视**：模型可能放大训练数据中的性别、种族偏见
   * **错误信息/虚假信息**：生成看似真实但错误或有害内容
   * **隐私泄露**：模型可能在生成中泄露训练语料中敏感信息
   * **责任归属**：模型输出造成损害时责任难以界定
3. **从技术与应用两大层次，讨论语言模型的发展趋势**（6 分）
   * **技术层面**：更大规模的参数/数据，更高效的架构（稀疏注意力，模型剪枝，知识蒸馏），可解释性，多模态信息的融合
   * **应用层面**：行业/专业定制大模型，实时对话代理，与传统信息系统深度融合