25自然语言处理-试题回忆

编者	Egopposer（line2345）
日期	2025/5/29

0、前言及试题概览

试题均为个人回回忆，存在大量描述误差，示例错误等；

莫得答案(ゝ∀･)

*斜体*为记忆模糊的题目，不保证可信度，**黑体**为我认为需要关注的点。

试题概览：（主观评价较多，仅供娱乐）

题量	大
难易度（软院专业课中）	10/10
送分题占比	10%
背诵记忆占比	10%
数理难度	7/10
21-23均分	83.966
21-23平均满绩率	34%

最难的一集，部分题目极偏；画的重点也不是重点，题目大量深度学习内容（尤其是Transformer）建议重读；

本课程是软院内少数只看课件不能完全涵盖考试内容的课程，建议看完课件再看看书，最好拓展性学习内容；

1、选择题（15*2=30）

- 句子`乒乓球拍卖完了`属于
    - 交集型歧义
    - 组合型歧义
- 语料库语言学的研究目的不包括
    - 语料库的编撰与建立
    - 语料库的使用
    - 语料库的加工和统计
    - 语料库的管理
- bigram认为目标词的出现概率依赖于
    - 当前词的频率
    - 上一个词的频率

    - 上两个词的频率
    - 历史所有词的频率
- 当句子上下句独立时，其分布趋近于
    - 高斯分布
    - 二项分布
    - 泊松分布
    - 正态分布
- BERT的预训练任务中，不包含的任务是
    - 下一句预测
    - 句子分类
    - 掩码自编码
    - 自回归编码
- Skip-Gram与CBOW相比，最大的区别在于其对（）进行预测
    - 目标词向量
    - 上下文词向量
    - 一个错误的答案
    - 另一个错误的答案
- Glove相较于Elove，还额外提供了
    - 上下文语义
    - 全局共现统计频率
    - .....
    - .....
- 最早在深度神经网络中引入非线性激活函数的目的是
    - 加快梯度收敛速度
    - 防止网络过拟合
    - 赋予模型拟合复杂函数的能力
    - 好玩
- 在Transformer中，Layer-Normalization的位置在
    - FFN之前
    - 残差连接之后
    - 位置编码之后
    - Attention计算之后
- BART的训练目标是
    - 预测被MASK的词语
    - 恢复含有噪声的输入文本
    - 预测目标词的上下文向量
    - 捕获文本中丰富的语义信息
- 有以下文法: $G = (\{S, A, B\}, \{a, b\}, P, S)$,其中:
  P: $\mathrm{S} \rightarrow \mathrm{aAB}$;
  $\mathrm{A} \rightarrow \mathrm{a}|\mathrm{aA}\ \mathrm{B} \rightarrow \mathrm{b}|\mathrm{bB}$
  求 $\mathrm{L}\,(\mathrm{G}) =$?
    - $L = \{a^n b^m \mid n \geq 0, m \geq 1\}$
    - $L = \{a^n b^m \mid n \geq 1, m \geq 1\}$
    - $L = \{a^n b^m \mid n \geq 2, m \geq 0\}$
    - $L = \{a^n b^m \mid n \geq 2, m \geq 1\}$

2、填空题（10*1=10）

- CRF的特征函数一般包含（）和（）
- GPT仅使用了Transformer的（）层，而BERT使用了（）层
- *命名实体标注的经典方法是（）*
- *一种常用的基于统计方法的词性标注的方法是（）*
- 语义角色分析是对谓词结构的（）进行分析，以`the boy next the door`这句话为例，`the boy`是（），`the door`是（）
- HMM的参数估计需要使用（）算法

3、简答题（5*4=20）

> 邪门，完全不往重点考

- 简述RNN（例如LSTM）与Transformer各自的优劣，并阐述Transformer为什么特别适用于现代自然语言处理
- 简述Attention机制的工作原理，并阐述Attention机制在NLP任务中的使用和优势
- 比较生成式模型和判别式模型的区别
- 简述BPE的机制，并阐述如何使用BPE解决OOV问题

4、综合题（8+10+12+10）

1、给定句子

```
<s>The dog ate the cat</s>
<s>The cat sat on the mat</s>
<s>The bird sat on the mat</s>
```

- 使用2-gram模型，计算句子`The dog sat on the mat`的概率。
- 使用加一法进行数据平滑后，再次计算第一问句子出现的概率，同时描述数据平滑对词组`sat on`出现概率的影响

2、给定HMM参数（原题给出的是概率），使用Viterbi算法求解观测序列`{walk,shop,clean}`出现时最可能得状态序列。

状态转移概率	**Rainy**	**Sunny**
Rainy	0.7	0.3
Sunny	0.5	0.5

发射概率	**walk**	**shop**	**clean**
Rainy	0.3	0.2	0.5
Sunny	0.5	0.4	0.1

初始概率	**Rainy**	**Sunny**
	0.5	0.5

3、给定句子`这是伊朗核问题的解决方案（忘了）`

- 使用`arc-eager`算法对该句子进行依存语法分析，给出每一步的`action`；
- 下面是给出的正确计算得到的依存句法树和通过模型计算得到的依存语法树，计算UA，LA，DA；

4、现在有一个微博文本（小于140字）情感二分类任务，样本存在噪声和标签稀疏的问题；请你设计方案，分别给出数据预处理、特征表示、模型选取、训练方法和测试指标的具体方案；

> 疑似考察篇章分布式表示