简答猜测

1. 简述最大匹配和最大概率分词法的原理和区别

最大匹配分词算法：

原理：从左到右，按照设定的最大长度，在词典中匹配最长的词语。

最大概率分词算法：

原理：基于统计模型和概率计算，选择使得整体概率最大的分词方式。

区别和比较：

1. **原理不同：**
   - 最大匹配分词法基于固定长度的匹配，而最大概率分词法基于概率计算和上下文信息。
2. **处理歧义的能力：**
   - 最大匹配分词法在处理歧义时可能出现问题，因为它只依赖于固定长度的匹配。而最大概率分词法能够考虑上下文信息，对歧义性词汇有更好的处理能力。
3. **算法复杂性：**
   - 最大匹配分词法相对简单，容易实现。最大概率分词法需要建立统计语言模型，相对复杂。
4. **依赖词典：**
   - 两者都依赖于一个词典，但最大匹配分词法更注重匹配长度，而最大概率分词法更灵活，能够更好地适应不断变化的语境。
5. **适用场景：**
   - 最大匹配分词法适用于简单场景，对分词速度要求高的情况。最大概率分词法适用于需要更高分词准确性的场景，例如自然语言处理任务。

2. 简述一下基于HMM的词性标注算法

初始化：为每个词初始化一个隐含状态（词性标签）。

状态转移：根据状态转移概率，对每个词的隐含状态进行转移。

发射概率：根据发射概率，为每个隐含状态生成一个观测（词性标签）。

Viterbi解码：利用动态规划算法，找到最可能的词性标注序列。

模型训练：通过训练语料库，估计模型参数。

评估：使用评价指标对模型性能进行评估。

应用：对新的句子进行词性标注。

3. 简述生成模型的beam-search、top-k采样、top-p采样解码算法，以及区别

Beam Search:

原理：

- Beam Search 是一种贪心搜索算法，它在每个生成步骤选择最有可能的前 k 个候选序列。

步骤：

1. 在初始时，选择模型生成的前 k 个候选序列。
2. 对于每个候选序列，生成下一个词，得到新的 k 个候选序列。
3. 重复此过程，直到生成的序列达到指定的长度或者满足终止条件。

特点：

- 每个步骤保留了 k 个最有可能的序列，因此具有一定的宽度。

Top-K Sampling:

原理：

- Top-K Sampling 在每个生成步骤从模型生成的概率分布中选择概率最大的前 K 个词，然后在这 K 个词中按概率进行随机采样。

步骤：

1. 对于每个生成步骤，从模型输出的概率分布中选择概率最大的前 K 个词。
2. 对这 K 个词按照概率进行随机采样，选择最终生成的词。

特点：

- 具有一定的随机性，使得生成的序列更加多样化。

Top-P Sampling ：

原理：

- Top-P Sampling 在每个生成步骤从模型生成的概率分布中选择概率累积达到一定百分比 P 的词，然后在这个累积概率内按照概率进行随机采样。

步骤：

1. 对于每个生成步骤，从模型输出的概率分布中选择概率累积达到一定百分比 P 的词。
2. 对这个累积概率内的词按照概率进行随机采样，选择最终生成的词。

特点：

- 可以控制生成过程的多样性，当 P 接近 1 时，采样更加多样化；当 P 接近 0 时，趋向于贪心采样。

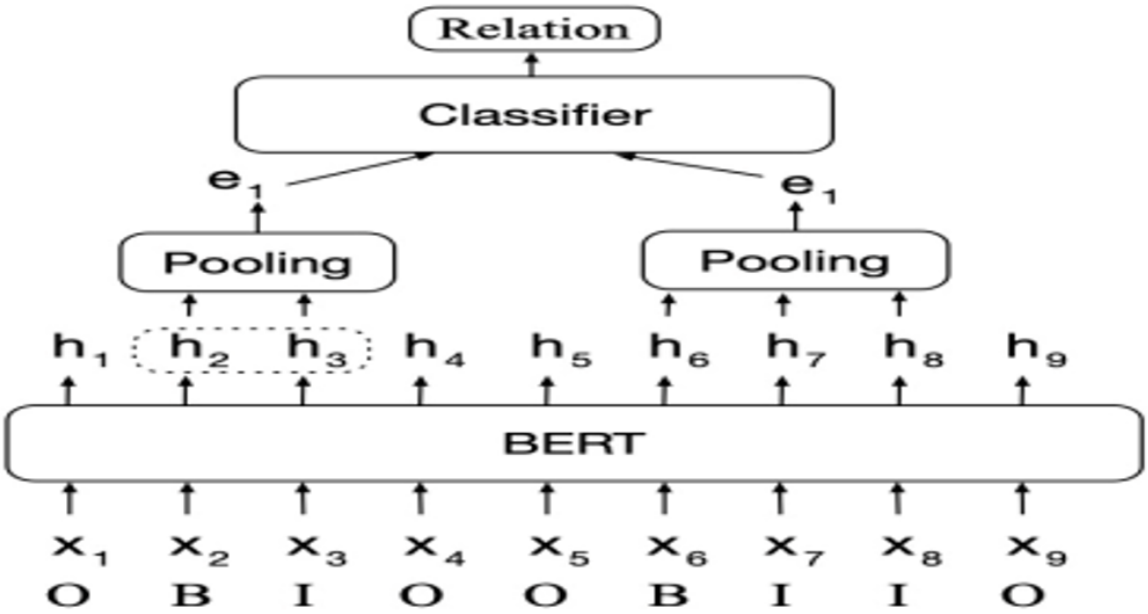
区别：

1. **搜索策略：**
   - Beam Search 是一种宽度优先的搜索策略，保留每个步骤的 top-k 个候选序列。
   - Top-K Sampling 和 Top-P Sampling 是一种随机采样策略，根据概率选择词语，使得生成更具有随机性。
2. **宽度和深度：**
   - Beam Search 在每个步骤保留多个序列，具有一定的宽度。
   - Top-K Sampling 和 Top-P Sampling 是基于概率分布进行采样，具有一定的深度。
3. **多样性控制：**

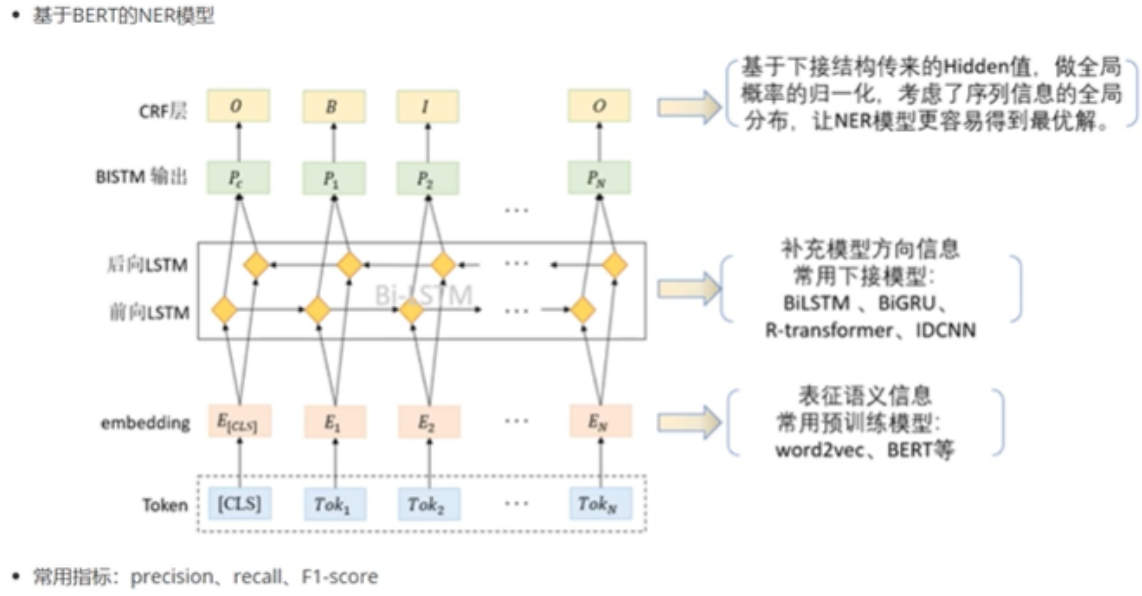
- Beam Search 可能导致生成的序列相对单一。
- Top-K Sampling 和 Top-P Sampling 具有一定的随机性，可以更好地控制生成序列的多样性。

4. **计算开销：**
   - Beam Search 的计算开销较大，随着 k 的增加而增加。
   - Top-K Sampling 和 Top-P Sampling 的计算开销相对较小，但仍然需要对概率分布进行计算。

关系抽取

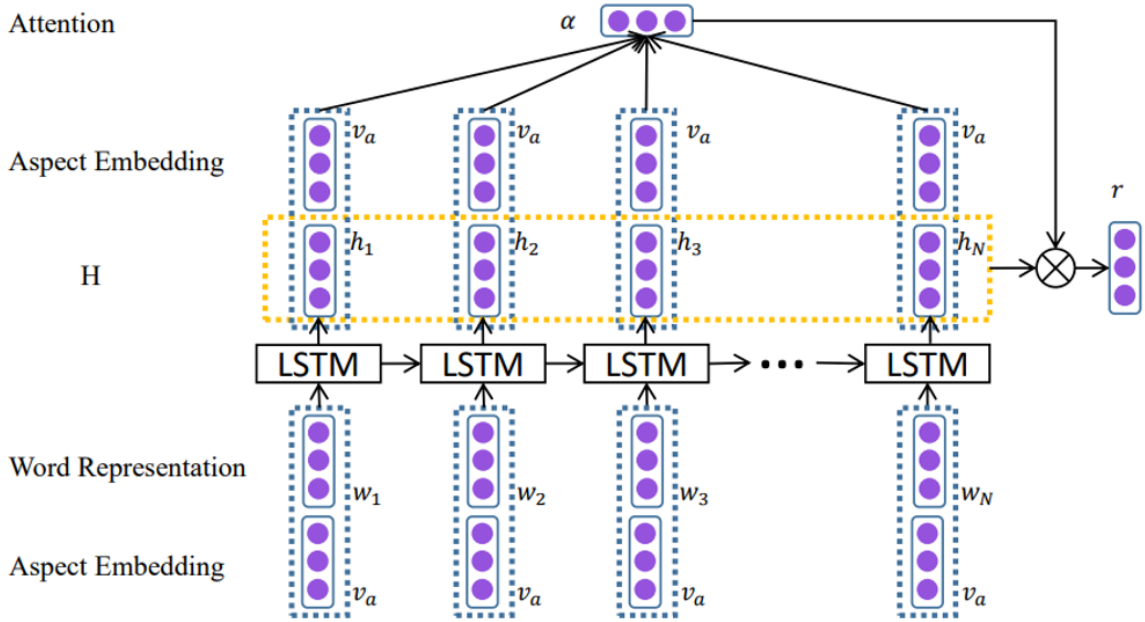


命名实体识别



情感分类

ATAE-LSTM （Attention-based LSTM with Aspect Embedding）



摘要系统

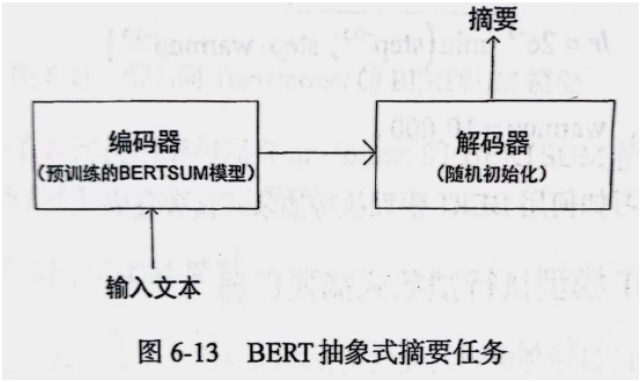


图 6-13　BERT 抽象式摘要任务