**Subword**

与tokenization的区别？

1. 传统词表示方法无法很好的处理未知或罕见的词汇（OOV问题）
2. 传统词tokenization方法不利于模型学习词缀之间的关系
3. Character embedding作为OOV的解决方法粒度太细
4. Subword粒度在词与字符之间，能够较好的平衡OOV问题

BPE（字节对编码）

答：原理：统计每一个连续字节对的出现频率，选择最高频者合并成新的subword

Wordpiece

答：WordPiece算法可以看作是BPE的变种。不同点在于，WordPiece基于概率生成新的subword而不是下一最高频字节对。