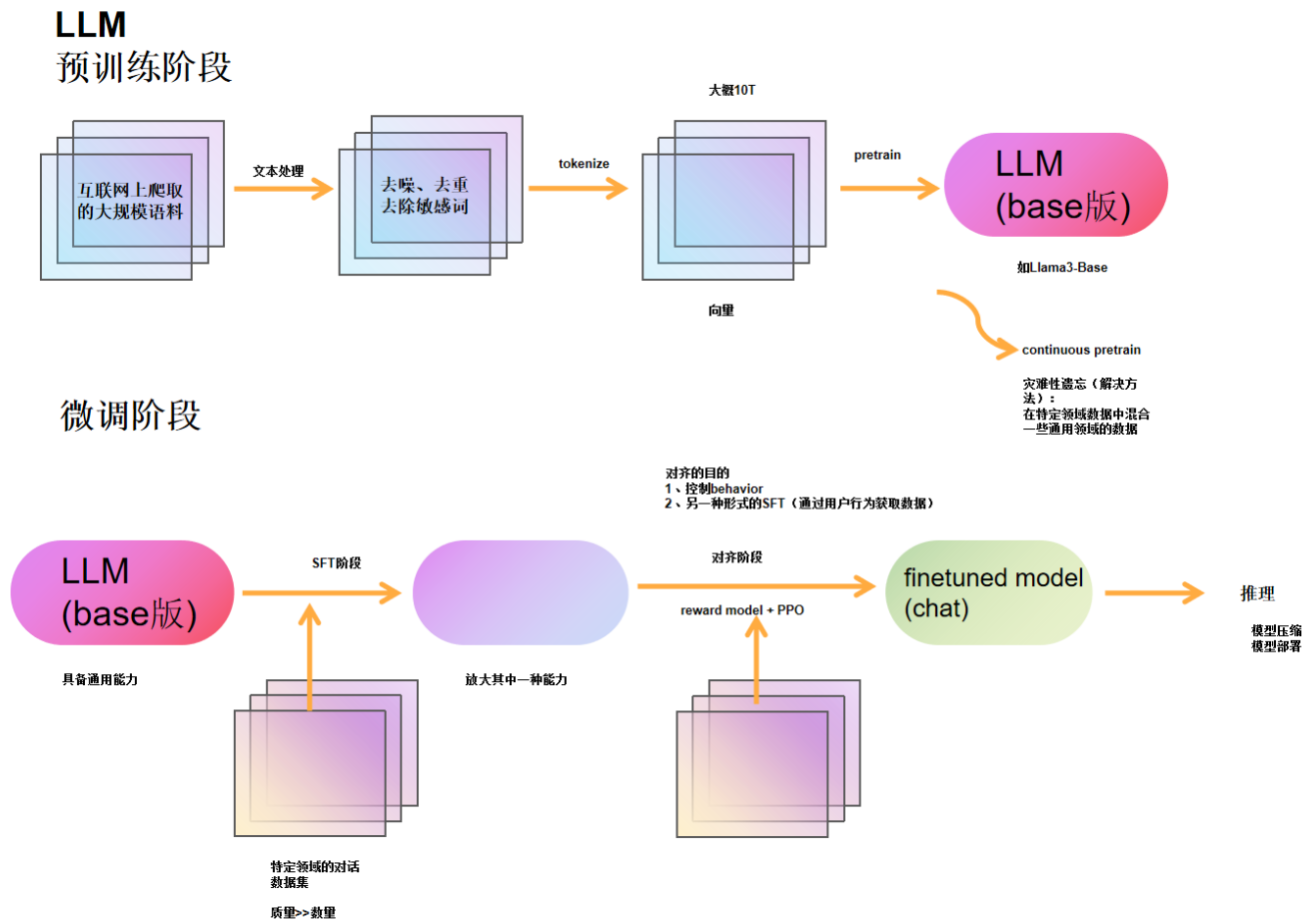
LLM训练流程



【大语言模型LLM基础之Tokenizer完全介绍-哔哩哔哩】 https://b23.tv/2kdTKxf

LLM中的tokenizers

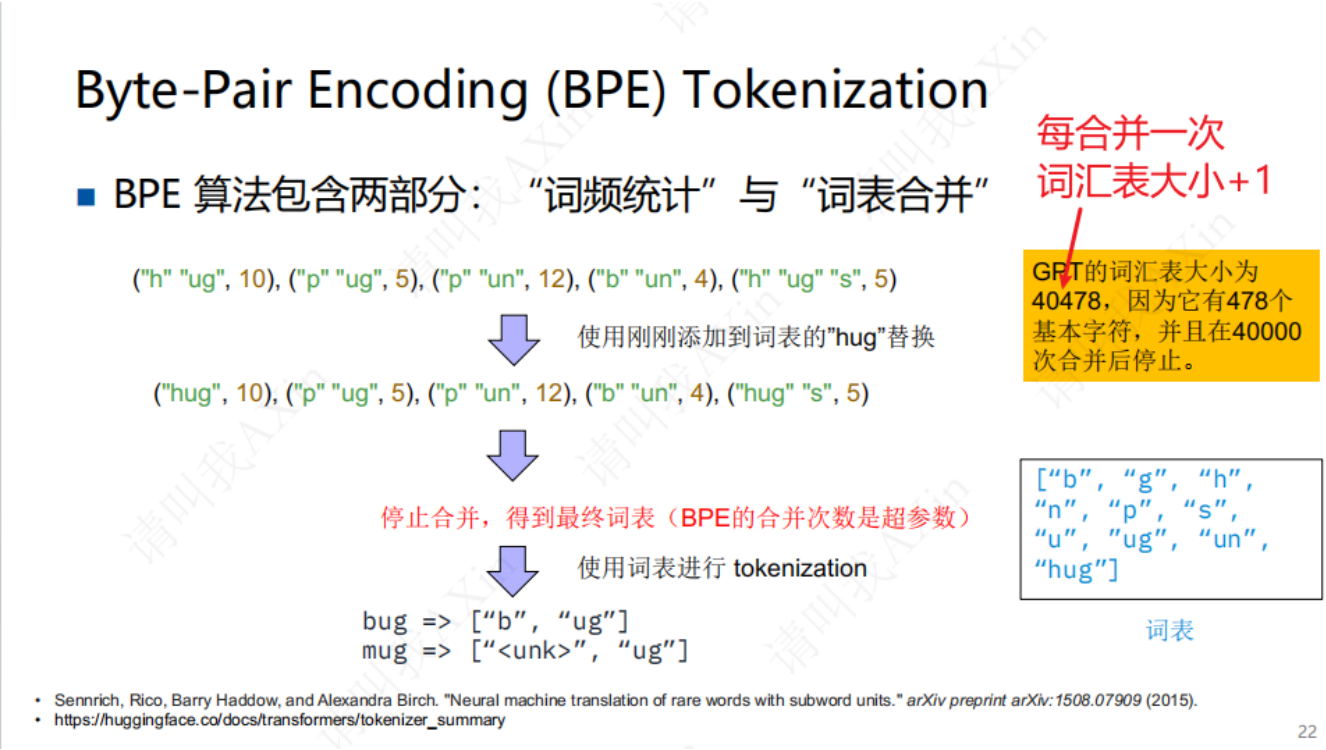
三种不同分词粒度的Tokenizers

- word-based
- character-based
- subword-based
  - WordPiece：BERT、DistilBERT
  - Unigram：XLNet、ALBERT
  - BPE（Byte-Pair Encoding）：GPT-2、RoBERTa
  - SentencePiece

BPE

词频统计->词表合并

设置：BPE的合并次数



改进：BBPE

WordPiece

WordPiece Tokenization

https://static.googleusercontent.com/media/research.google.com/ja//pubs/archive/37842.pdf

- 大体和BPE类似
  - 构建基本词表时，除第一个字母，会添加##作为前缀（BERT）

```
word => [w, ##o, ##r, ##d]
```

  - 使用类似联合概率的大小而不是次数对 token 进行合并，公式如下：

$$\text{pair得分} = \frac{\text{pair出现的次数}}{\text{token1出现的次数} \times \text{token2出现的次数}}$$

  - 通过将 pair 的频率除以其单个 token 的频率的乘积，该算法优先考虑单个 token 在词表中不太频繁的 pair 进行合并

pair得分越大，表明分子越大、分母越小

https://huggingface.co/learn/nlp-course/chapter6/6?fw=pt

24

Unigram

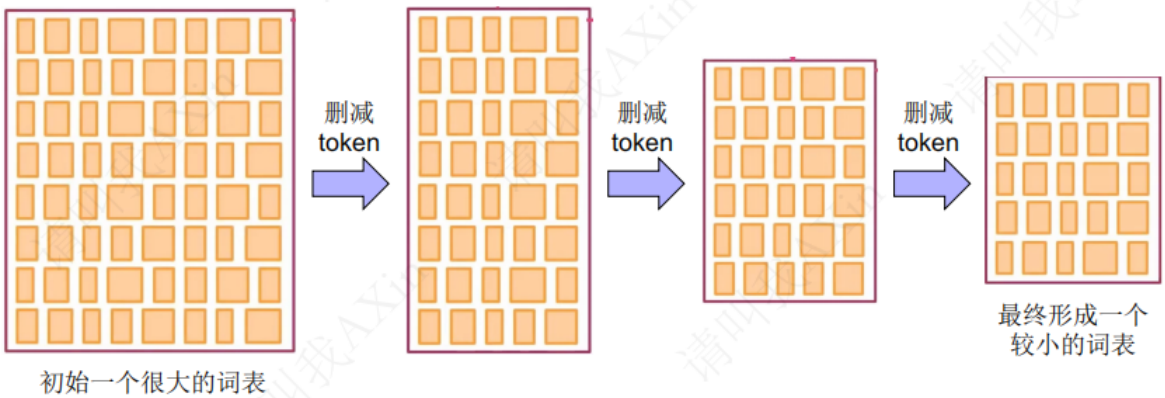
先初始化一个很大的词表（字母、单词、subword都包括）

设置：删减的次数

Unigram Tokenization

https://arxiv.org/pdf/1804.10959.pdf

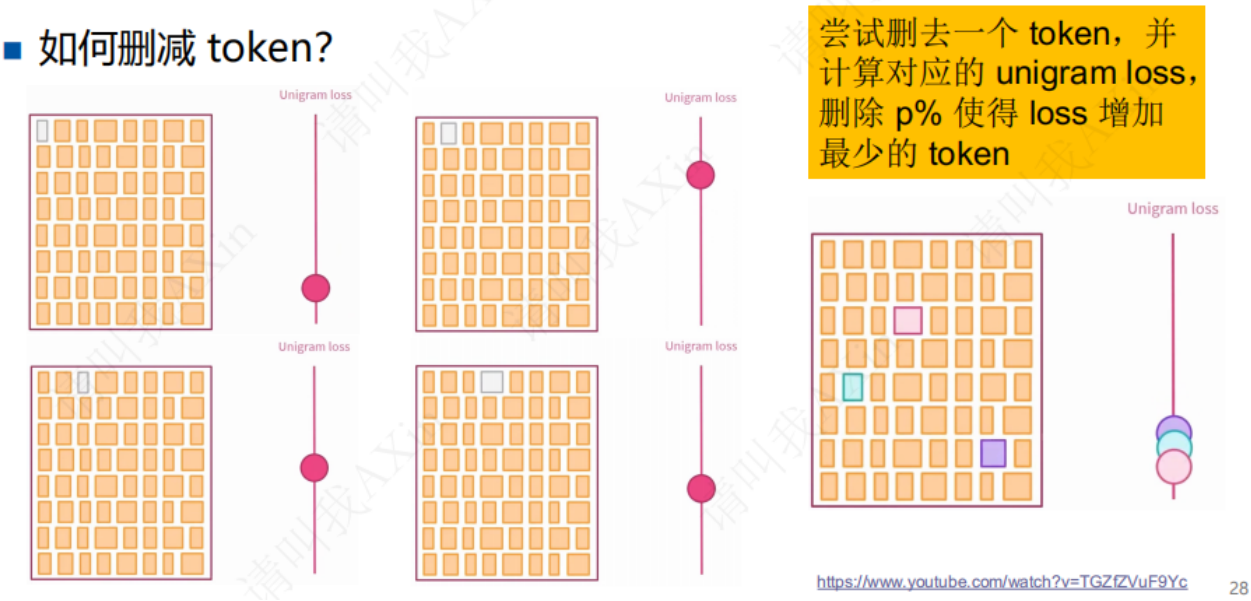
- Unigram 算法经常在 SentencePiece 中使用，是 AlBERT、T5、mBART、Big Bird 和 XLNet 等模型使用的 tokenization 算法



https://huggingface.co/learn/nlp-course/chapter6/7
https://www.youtube.com/watch?v=TGZfZVuF9Yc

27

删去对词表的表达能力影响不大的token



基于统计的划分

loss：负对数似然

SentencePiece（使用BBPE或Unigram）

解决多国语言的分词问题，输入都当做字节流（含空格）