LLMs Tokenizer 篇

来自： AiGC面试宝典

宁静致远

2023年09月29日 12:20

Byte-Pair Encoding(BPE)篇

1 Byte-Pair Encoding(BPE) 如何构建词典?

1. 准备足够的训练语料;以及期望的词表大小;
2. 将单词拆分为字符粒度(字粒度)，并在末尾添加后缀“”，统计单词频率
3. 合并方式:统计每一个连续/相邻字节对的出现频率，将最高频的连续字节对合并为新的子词;
4. 重复第3步，直到词表达到设定的词表大小;或下一个最高频字节对出现频率为1。

> 注：GPT2、BART和LLaMA就采用了BPE。

WordPiece 篇

1 WordPiece 与 BPE 异同点是什么?

本质上还是BPE的思想。与BPE最大区别在于:如何选择两个子词进行合并

- BPE是选择频次最大的相邻子词合并;
- WordPiece算法选择 能够提升语言模型概率最大的相邻子词进行合并，来加入词表

> 注：BERT采用了WordPiece。

SentencePiece 篇

简单介绍一下 SentencePiece 思路?

把空格也当作一种特殊字符来处理，再用BPE或者来构造词汇表。

> 注：ChatGLM、BLOOM、PaLM采用了SentencePiece。

对比篇

1 举例 介绍一下 不同 大模型LLMs 的分词方式?

模型	词表大小	分词结果	长度
LLaMA	32000	['男', '<0xE5>', '<0x84>', '<0xBF>', '何', '不', '<0xE5>', '<0xB8>', '<0xA6>', '<0xE5>', '<0x90>', '<0xB4>', '<0xE9>', '<0x92>', '<0xA9>', ',', '', '收', '取', '关', '山', '五', '十', '州', '。']	24
Chinese LLaMA	49953	['男', '儿', '何', '不', '带', '吴', '钩', ',', '', '收取', '关', '山', '五十', '州', '。']	14
ChatGLM-6B	130528	['男儿', '何不', '带', '吴', '钩', ',', '收取', '关山', '五十', '州', '。']	11
ChatGLM2-6B	65024	['男', '儿', '何', '不', '带', '吴', '钩', ',', '', '收取', '关', '山', '五十', '州', '。']	14
Bloom	250880	['男', '儿', '何不', '带', '吴', '钩', ',', '', '收取', '关', '山', '五十', '州', '。']	13
Falcon	65024	['男', '儿', '', '', '', '不', '带', '吴', '', '', '', '', '', '收', '取', '', '', '山', '五', '十', '州', '。']	22

2 介绍一下 不同 大模型LLMs 的分词方式 的区别?

模型	词表大小	中文平均token数	英文平均token数	中文处理时间(s)	英文处理时间(s)
LLaMA	32000	1.45	0.25	12.6	19.4
Falcon	65024	1.18	0.235	21.395	24.73
Chinese LLaMA	49953	0.62	0.249	8.65	19.12
ChatGLM-6B	130528	0.55	0.19	15.91	20.84
ChatGLM2-6B	65024	0.58	0.23	8.899	18.63
Bloom	250880	0.53	0.22	9.87	15.6

1. LLaMA的**词表是最小的**，LLaMA**在中英文上的平均token数都是最多的**，这意味着LLaMA对中英文分词都会比较碎，比较细粒度。尤其在中文上平均token数高达1.45，这意味着LLaMA大概率会将中文字符切分为2个以上的token。
2. Chinese LLaMA扩展词表后，**中文平均token数显著降低**，会将一个汉字或两个汉字切分为一个token，提高了中文编码效率。
3. ChatGLM-6B是平衡中英文分词效果最好的tokenizer。由于**词表比较大**，**中文处理时间也有增加**
4. BLOOM虽然是**词表最大的**，但由于是多语种的，在中英文上分词效率与ChatGLM-6B基本相当。