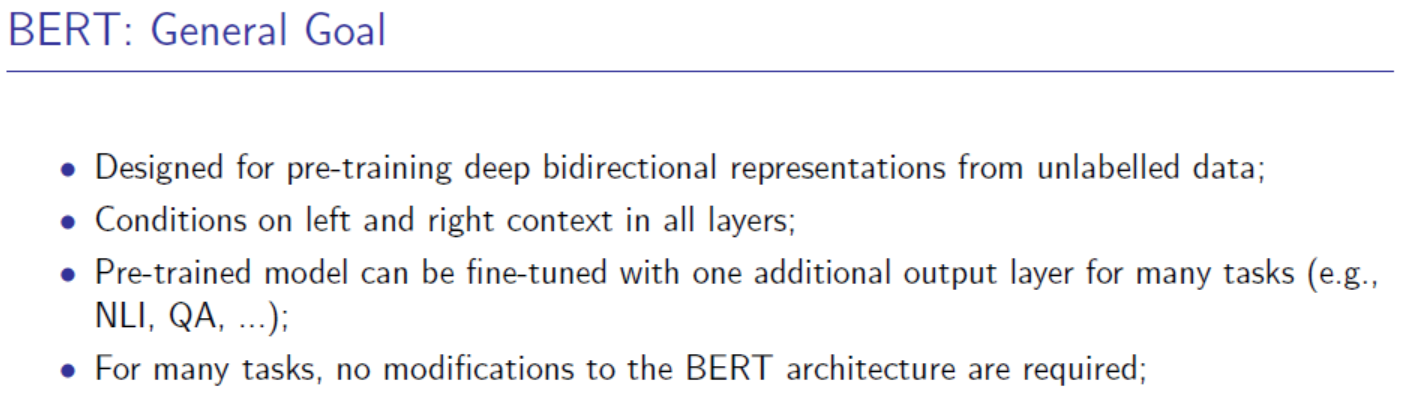
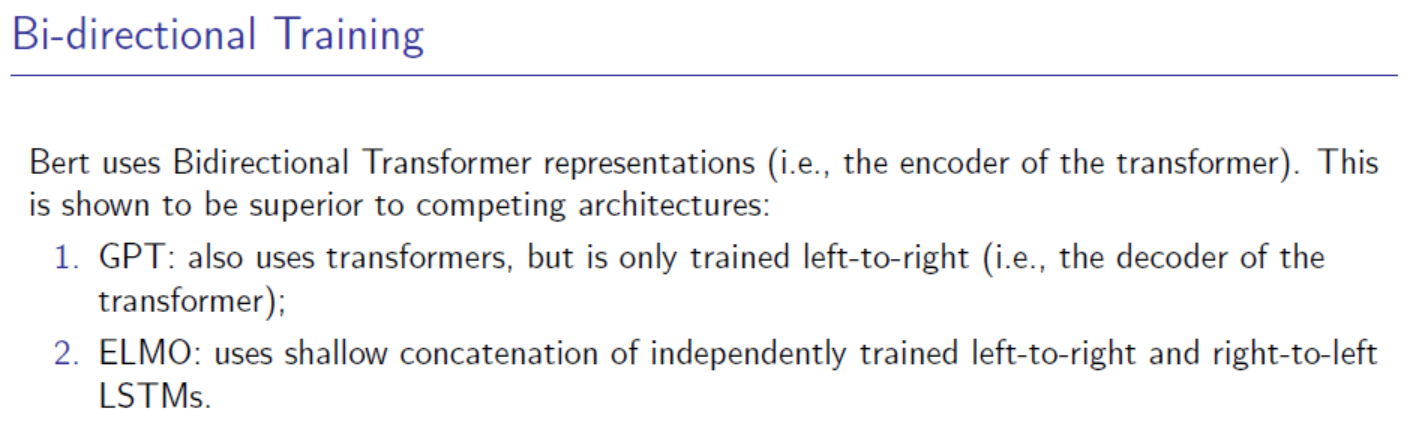
## 引入

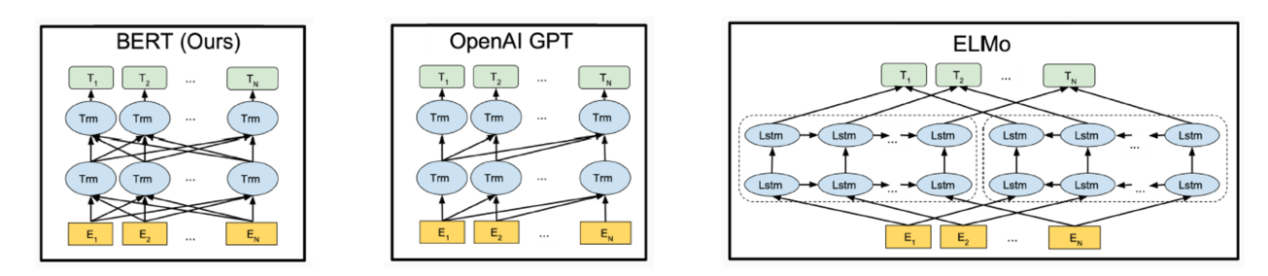


Elmo：如果可以通过预训练的方法初始化一个语言模型，用预训练好的语言模型的参数用

GPT：有mark attention机制（防偷窥）

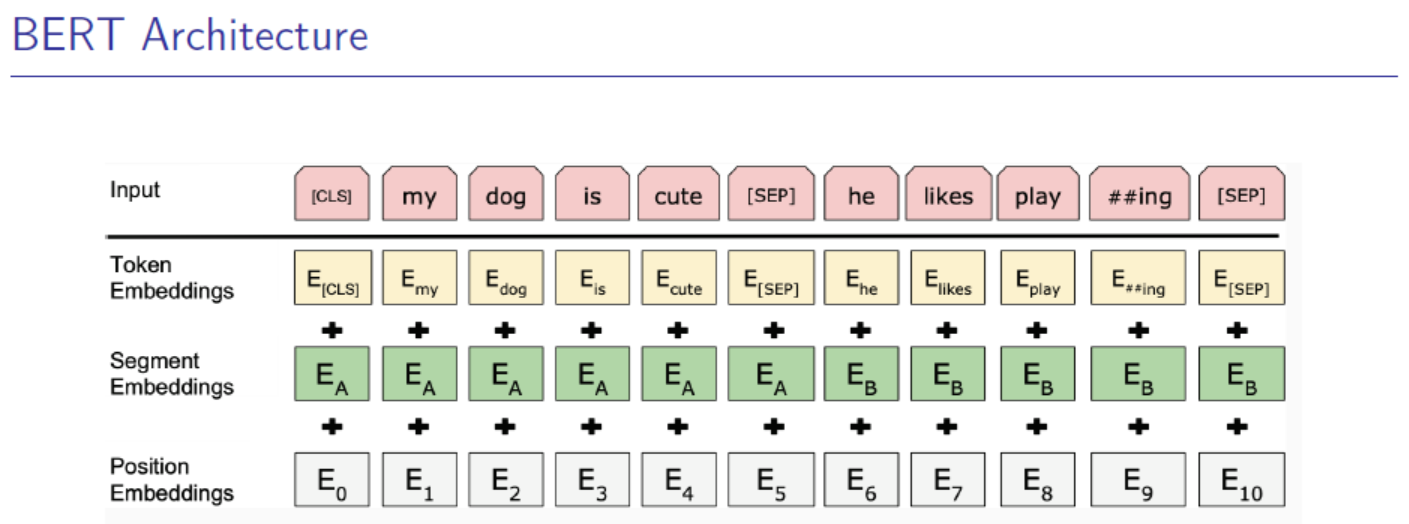
Bert-->发扬光大预训练





## General Architecture

Bert input：在预训练的时候输入是两句话拼接起来

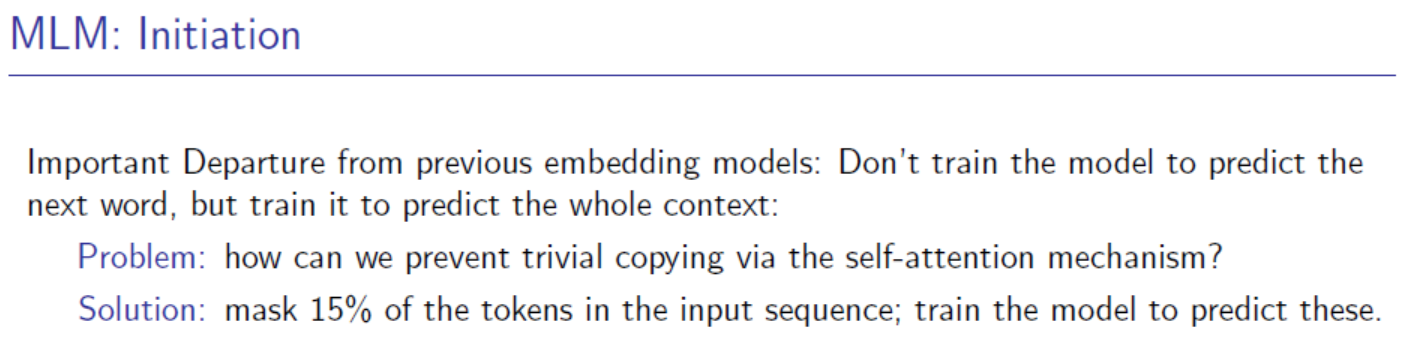


## 优化任务：Masked language model and next sentence predicate

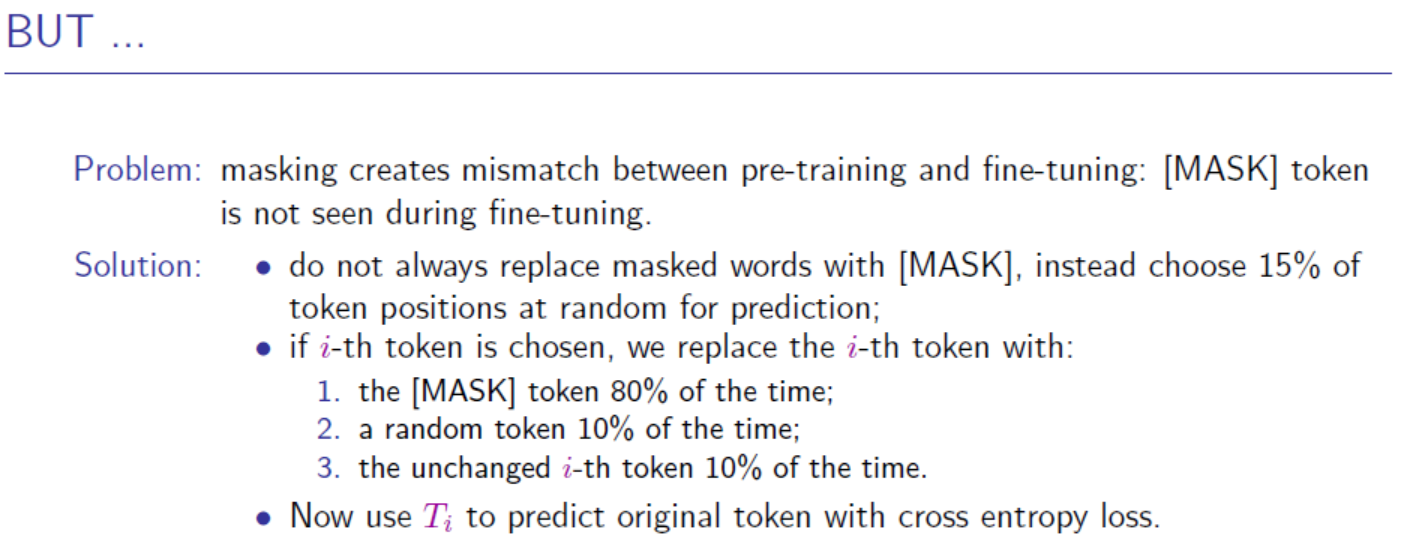
#### 3.1 Masked language model

Bert之前：从前面的单词预测

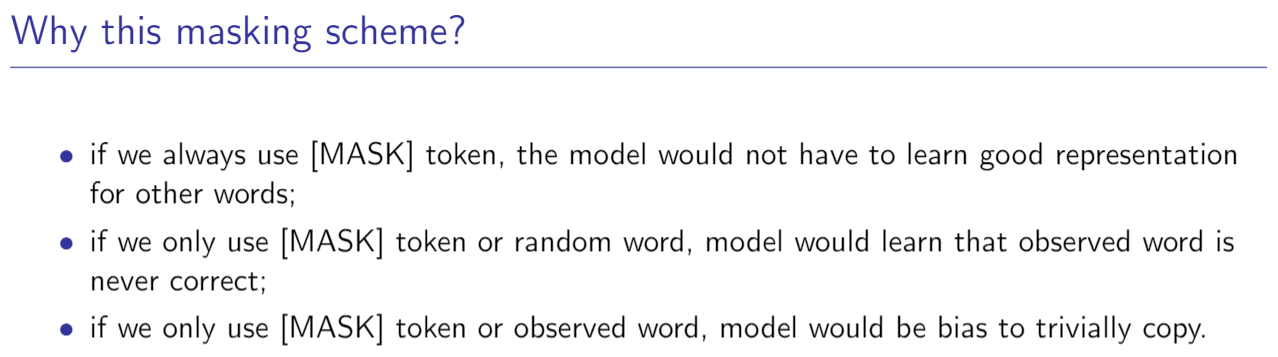
Bert ：从两边预测中间（mark：就是中间要预测的词，bert中把这个扣掉）



只在预训练的时候用到marks，BUT（只在预训练的时候用到mark，训练的时候就没了mark）

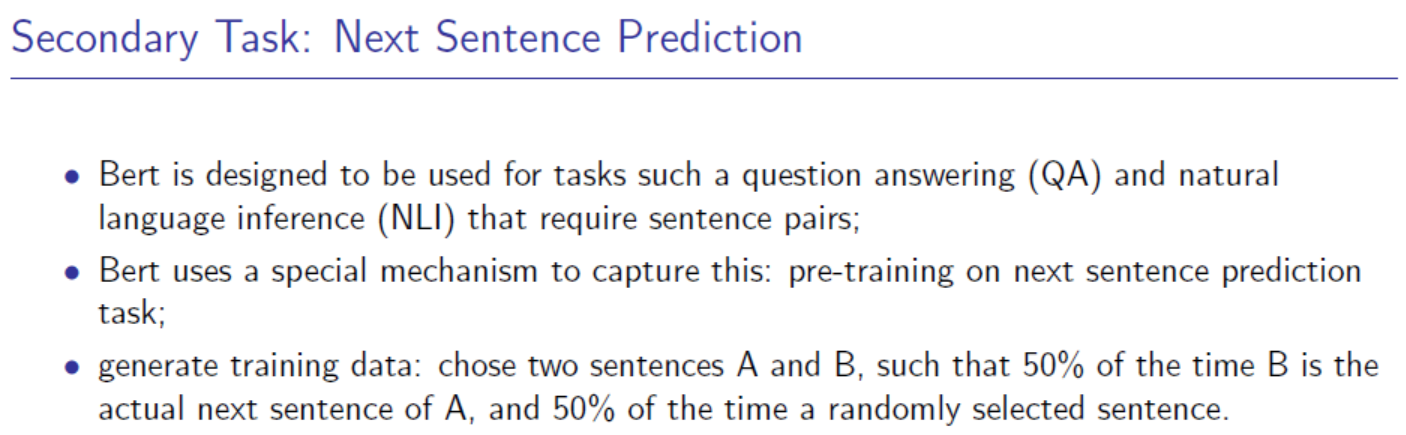


如果不这么做：



### 3.2 next sentence prediction

词Word ，词的变种subword = sub + word

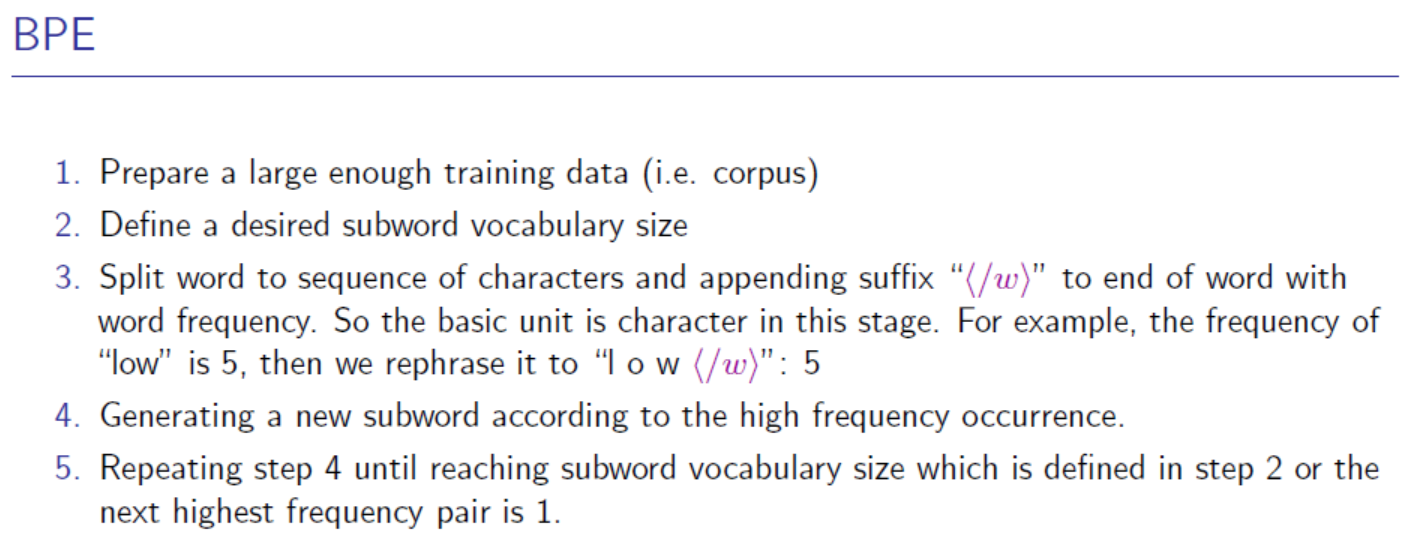


## Subword Units 数据预处理方式（sub\*\*word）

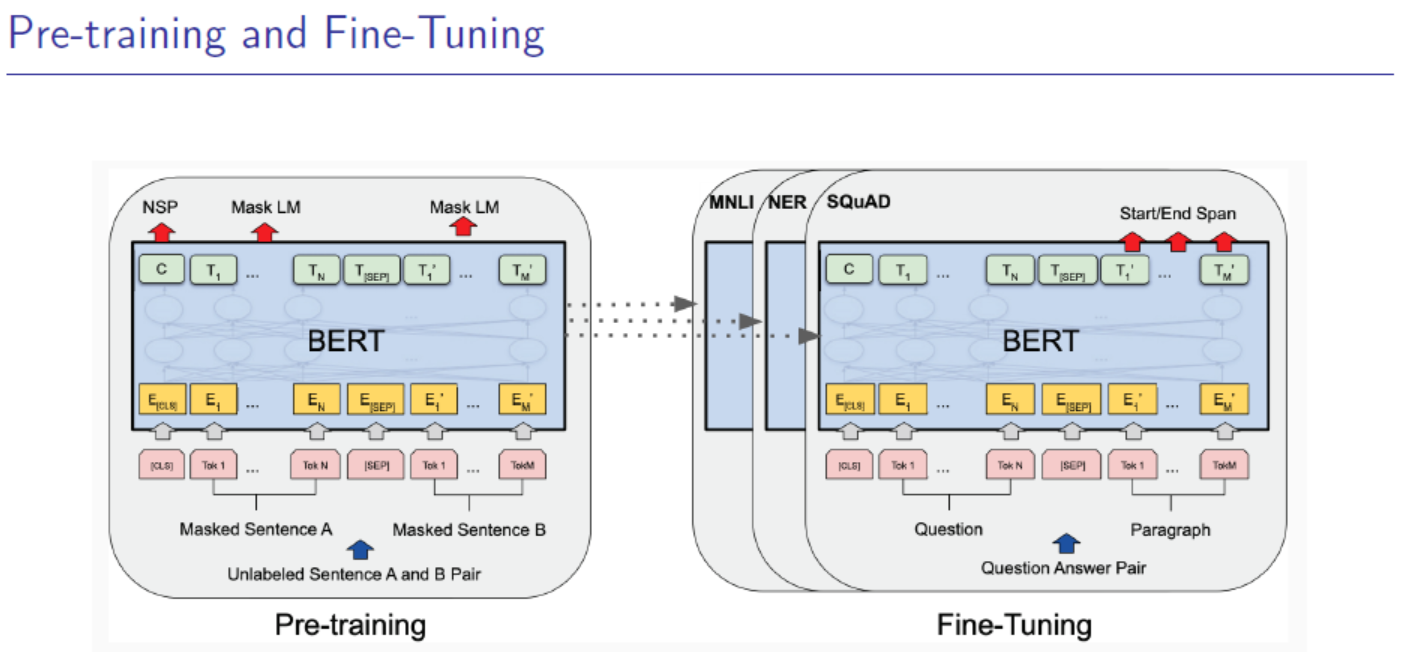
BPE

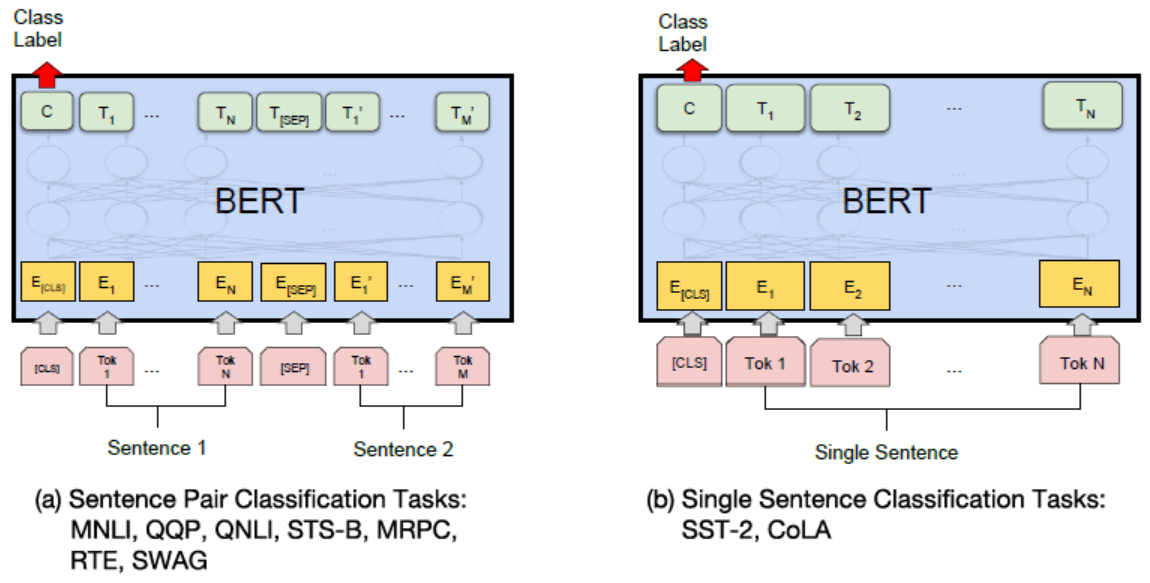
More reading:

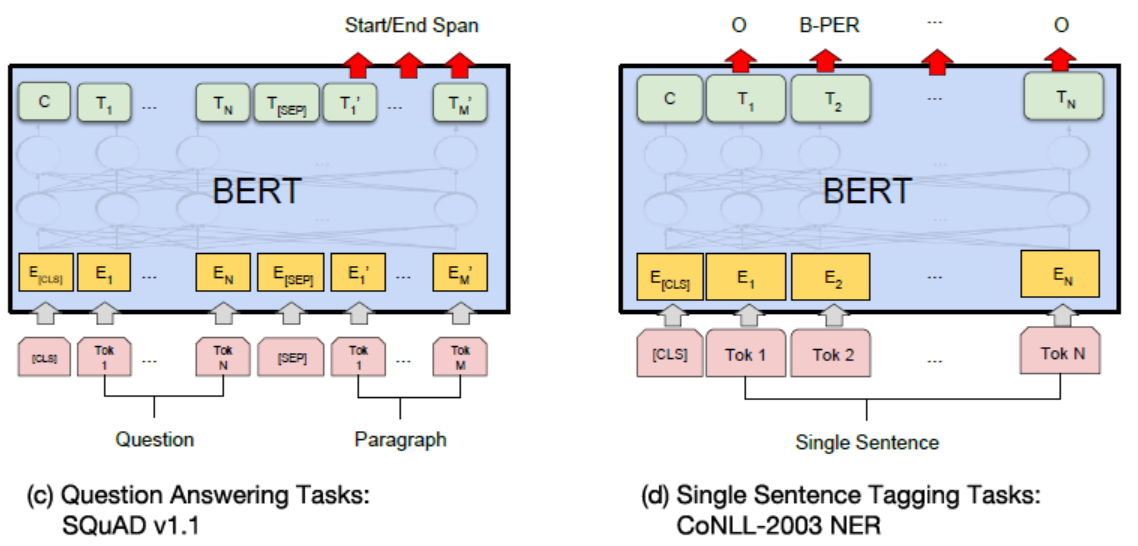
<https://blog.floydhub.com/tokenization-nlp/>

<https://zhuanlan.zhihu.com/p/86965595>  


## Pre-traning and Fine-Tuning BERT







微调的任务包括（a）基于句子对的分类任务：

MNLI：给定一个前提 (Premise) ，根据这个前提去推断假设 (Hypothesis) 与前提的关系。该任务的关系分为三种，蕴含关系 (Entailment)、矛盾关系 (Contradiction) 以及中立关系 (Neutral)。所以这个问题本质上是一个分类问题，我们需要做的是去发掘前提和假设这两个句子对之间的交互信息。

QQP：基于Quora，判断 Quora 上的两个问题句是否表示的是一样的意思。

QNLI：用于判断文本是否包含问题的答案，类似于我们做阅读理解定位问题所在的段落。

STS-B：预测两个句子的相似性，包括5个级别。

MRPC：也是判断两个句子是否是等价的。

RTE：类似于MNLI，但是只是对蕴含关系的二分类判断，而且数据集更小。

SWAG：从四个句子中选择为可能为前句下文的那个。

（b）基于单个句子的分类任务

SST-2：电影评价的情感分析。

CoLA：句子语义判断，是否是可接受的（Acceptable）

<https://zhuanlan.zhihu.com/p/48612853>