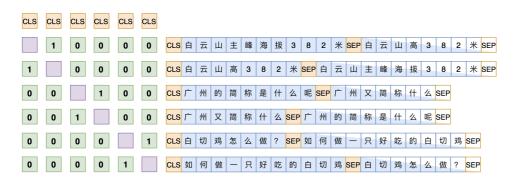
## SimBert & RoFormer-Sim

### **SimBert**

参考: https://spaces.ac.cn/archives/7427

# 模型训练图



#### 模型创新点

- 1. 以UniLM模型为基础训练的有监督模型,因此具有检索相似句和生成相似句的能力
- 2. 爬取百度知道2200万相似句组进行训练,因此对问句的检索能力和生成能力好

#### **RoFormer-Sim**

参考: https://spaces.ac.cn/archives/8454

## 模型创新点

与SimBert相比, RoFormer-Sim创新点主要体现在以下两个方面:

- 1、SimBert训练语料是从百度知道爬取到的相似句组,而RoFoemer-Sim还收集了通用类相似句,通过以下两个方案收集
  - 第一个方案是基于"同一个问题的答案是相似的"思想,假如我们有现成的问答语料,该语料对于同一个问题有多个答案,那么我们可以将每个答案分句,然后用一个现成的相似度函数来比较答案之间的相似度,挑出相似度超过某个阈值的句对作为相似句对使用;
  - 第二个方案则是基于"同一篇章的句子是相似的"思想,它更加简单直接一点,就是将每个篇章分句,然后用一个现成的相似度函数两两计算相似度,挑出相似度超过某个阈值的句对作为相似句对使用,显然该方案的合理性更弱,所以它的阈值也更高。
- 2、训练时还随机地将输入句子的部分token替换为[MASK],训练目标是输入一个带噪声的句子,输出原句子的一个相似句