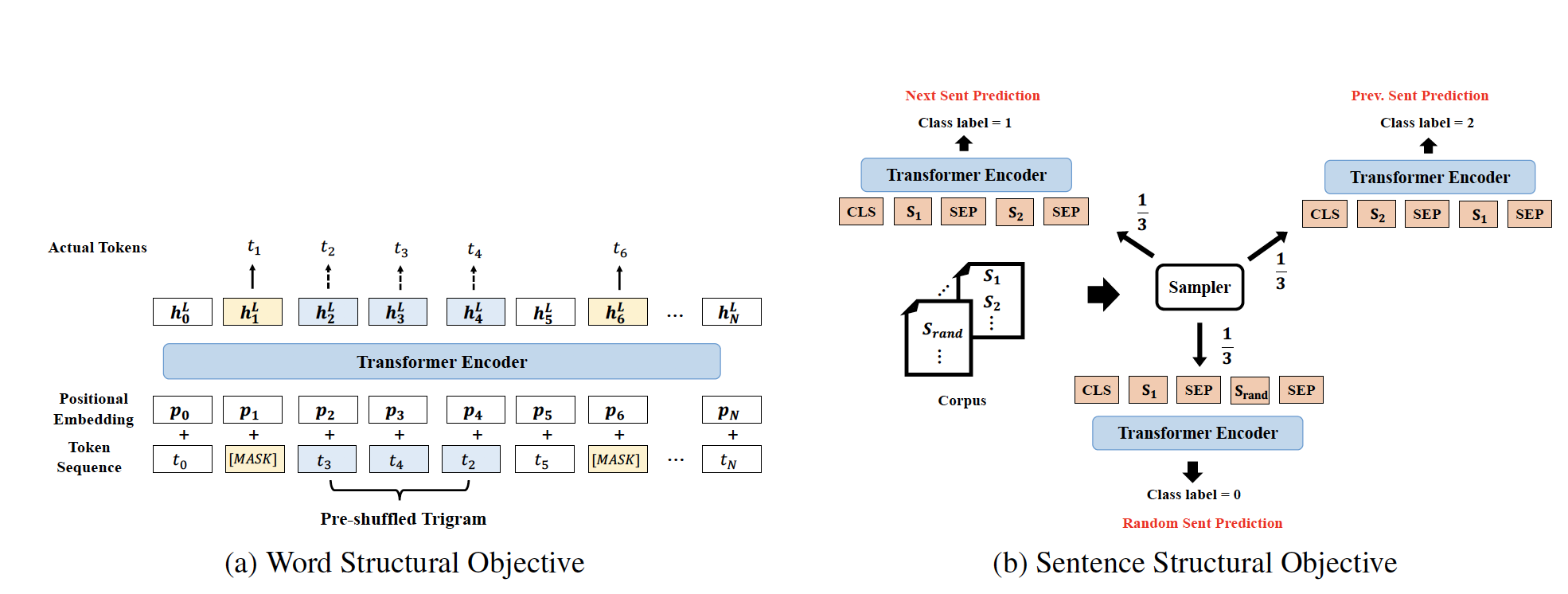
本课程介绍了StructBert和T5，重点介绍Roberta和AIBERT的改进方式以及使用方式，然后介绍上海交大Retro-Reader的算法。

第一部分介绍了StructBert的模型，这是阿里巴巴2019年放出的模型，是Incorporating Language Structures Into Pre-Training for Deep Learning Understanding的简称，属于基于bert的预训练模型，是从预训练任务的角度修改bert模型，并没有改变bert的基本结构，基本保留了wordpiece，absolute positional encoding的相关特性。其贡献集中于将两个新型的预训练任务（如下图所示，词级别的预训练任务(a)和句子级别的预训练任务(b)），将其加入bert原始预训练任务并且去掉了NSP任务，去掉NSP任务的原因是NSP会带来整体结果性能的下降。



StructBert在基于词级别的任务上是输入一个句子中打乱的单词，引入了pre-shuffled trigram进行预测。

T5的模型较大，一般认为是一个较大的实验报告来看待，含有丰富的参考资料。T5是Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer(Google 2019)的简称，它属于一个通用的框架，基于C4数据集（Colossal Clean Crawled Corpus，超大型干净爬取数据，取结尾有标点的句子、去掉包含脏话的网页、）进行了大量的实验。其主要任务是将所有NLP任务都转化成Text-to-Text任务，采用相同的模型、损失函数、训练过程和解码过程。其按照编码方式分成三种transformer架构为Language Model ，Prefix LM，Encoder-Decoder类型。对比之后还是任务Encoder-Decoder结果较好。其实验结论是模型越大越好，但是为了考虑实际环境而提供5个大小不一的模型级别，进行混合多模型任务的训练，使得Text-to-Text框架统一，编码器解码器共享参数不会使性能下降太多，对阈内未标记的数据训练可以提高下游性能。少模型训练模型的效果好于小模型用更多的数据去训练，但是对于下一步的展望是不能一味追求模型大，怀疑mask作为目标状态的效果，需要用更有效的方法 去学到更通用的知识，T5天然有上下游的gap，因此该翻译任务效果并不好。

RoBERTa是facebook提出的BERT优化版本，是A Robustly Optimizer BERT Pretraining Approach的简称，是保持结构不变，整体改动并不大。通过让模型接受更多的数据来进行预测效果的提升，其主要集中在更长的训练时间、更大的batch size，以及更多的训练数据进行着手。在训练方法上同样去掉了NSP预测任务，并且进行动态MASK，引入了新型的文本编码方式。NSP的局限性体现在输入空间的局限性，由于句子长短不一，容易出现补padding的时候造成预测的准确性下降。BERT级别的文本编码是字符级别的文本编码BPE词典达到30K，而RoBERTa是基于byte(字节)级别的进行文本编码从而增加数据量到50K。

ALBERT主要集中在进行轻量级的BERT设计。该方法是通过矩阵分解和参数共享减少参数，通过SOP替换NSP，并且采用n-gram MASK进行改进。这种矩阵分解的思想是通过因子嵌入参数化来实现，在两个大维度之间加入一个小维度起到降维的作用。SOP的本质是逆序，通过补偿一部分因为嵌入和FFN共享所损失的性能，将负样本换成了同一篇文章中的两个逆序的句子，SOP主要是进行主题预测和通顺度预测。其中的N-GRAM MASK是预测包含更完整语义信息的n-gram片段。AIBERT将窄而深的模型转变成了宽而浅的模型。

RetroReader是Retrospective Reader for Machine Reading Comprehension的简称，属于一种回顾式阅读器，可以更好的进行答案鉴定，该阅读器首次提出了显著性检验方法。采用了两阶段阅读和一个验证进行模型架构设计。