Week5

一、ERNIE(百度)

1. ERNIE1.0

百度提出的知识增强语义表示模型，全称 Enhanced Representation through knowledge Integration。

此方法与 Bert 类似，参数为 Bert Base 参数，L=12、H=768、A=12,训练数据集来自中文维基百科、百度百科、百度新闻和百度贴吧。

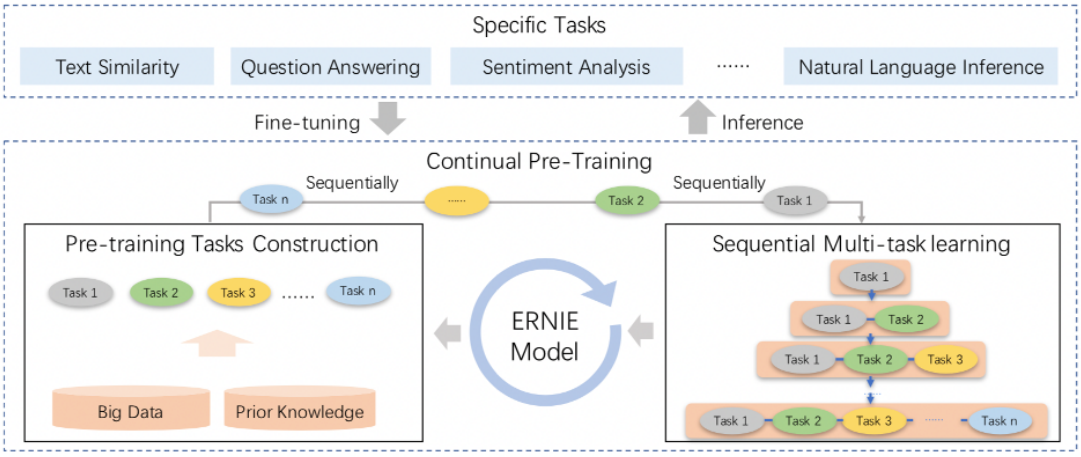
相比于 Bert 的仅仅单字 mask，ERNIE1.0 还增加了实体和短语 mask，此优化策略让模型能够学习语法和句法信息。同时，输入层使用多轮对话嵌入(Dialog embedding)，修改 NSP(Next Sentence Prediction)任务。

ERNIE1.0 在很多中文自然语言处理的任务上达到 state-of-the art，百度开放了代码和预训练模型。

2. ERNIE2.0

全称 A Continual Pre-Training Framework for Language Understanding。

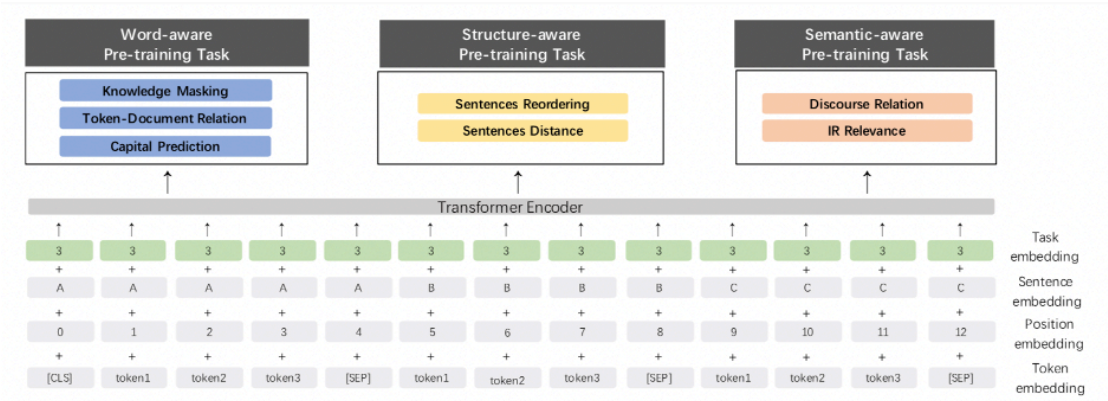
相比 1.0，ERNIE2.0 为多任务持续学习预训练框架，构建了三种类型的无监督任务,也增加了更多的语料。整体框架如下：



连续多任务学习可以不遗忘之前的训练结果，多任务高效的进行训练。使用上一任务的参数，新旧任务一起训练，将每个任务分成多次迭代，框架完成不同迭代的训练自动分配。多任务训练每个任务有独立的 loss function，句任务和词任务一起训练。

ERNIE2.0 把词法级别、语言结构级别以及语法级别的预训练任务进行连续多任务学习。词法级别识别的是 token 在段落 A 中，是否会在文档的段落 B 中出现。语言结构级别是看句子间的距离，判断句子为相连的句子，同一文档中不相连的句子还是两篇文档间的句子。语法级别任务可探寻短文本信息检索关系，识别搜索(query-title)数据，看其是否搜索并点击、搜索并展现及无关随机替换情况。

模型结构如下：



此模型目前来看是最后的中英文预训练语言模型之一，等待模型的放出或云服务。