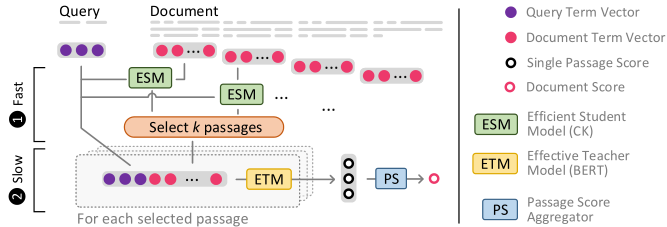
动机：

1）目的：用大型预训练语言模型（例如BERT）对文档各段落的查询延迟很高，设计新模型提高计算效率。

2）思路：采用内联级排序模型，先用蒸馏过的学生模型来对文档中的所有段落进行排序，得到分数最高的top k个段落，再将这k个段落放进教师模型再次进行排序，得到最终的查询结果。

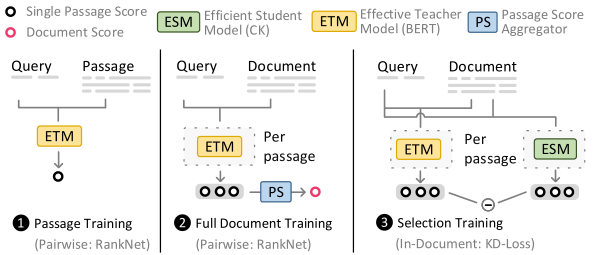
解决思路：

1）框架：包含两个级联阶段，第一阶段是轻量级且快速的选择模型，第二阶段是高消耗且高精度的BERT-based模型。



2）细节：第一阶段使用轻量级且快速的选择模型ESM，对正片文档中的每一个段落进行评分，并从中选出评分最高的前k个段落，将其送至第二阶段的BERT-based模型，记为ETM模型。ETM对这k个段落再次进行评分，输出最终的文档评分。

3）模型优化：



4）细节：

第一步，在段落对集合上对ETM进行训练，以优化其段落排名能力。段落集是正负段落对组成的，使用二元交叉熵损失函数进行优化。

第二步：将ETM优化扩展到全文档排名，即训练对全文档所有段落进行排名的能力。

第三步：使用知识蒸馏将ETM中的训练到ESM上，以强化ESM的段落排序和选择能力。

5） ESM：CK模型，Conv-KNRM的一种变体，CK 模型使用 CNN 层转换查询和段落表示，并使用余弦函数计算它们的相似性，然后由具有不同分布参数的高斯核激活

6） ETM：BERT排名模型。

模型下载地址：https://github.com/sebastian-hofstaetter/intra-document-cascade

实验：

1）数据集：

a. TREC DL 2019：2019 TREC Deep Learning Track中用于文档排名任务的 43 个查询

b. MS MARCO：从包含在MS MARCO开发集中的Bing 查询日志中采样的5,193个查询。

2）baseline：

a. BM25：

b. TKL：一种非 BERT 局部自注意力排序模型，具有内核池化。Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and AllanHanbury. 2020. Local Self-Attention over Long Text for Efficient DocumentRetrieval. In Proc. of SIGIR.

c. PARADEMax-Pool：非常接近我们的 All-BERT 基线，因为它使用 BERT ranker 对每个段落进行评分，并在轻量级层中聚合段落表示Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PA-RADE: Passage Representation Aggregation for Document Reranking. arXivpreprint arXiv:2008.09093 (2020).

d. PARADETF：使用额外的 Transformer 块来聚合段落表示

总结：IDCM是一种文档内级联排序模型，它提供了最先进的有效性，同时与非级联完整 BERT 排序模型相比，将中值查询延迟提高了四倍以上。实现与完整 BERT 模型相同的有效性的关键步骤是使用 BERT 段落分数来训练更有效的选择模块的知识提炼训练。模型应用的知识蒸馏为所有段落提供了自我监督的教师信号，无需手动注释，这不仅改善了模型在部署场景中的查询延迟，而且还提供了用逐步训练的教师模型替代人工注释劳动和成本的效率。

其他：