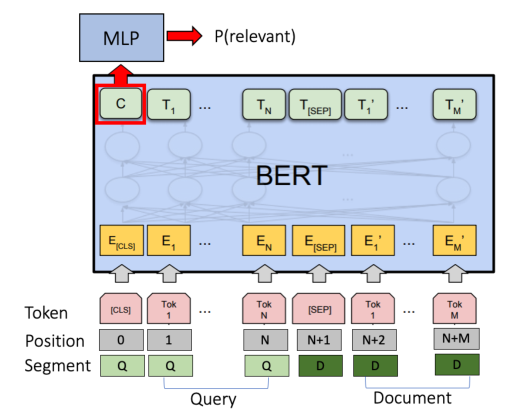
动机：

1）目的：从理解查询或文档的文本内容方面去研究IR模型。

2）思路：神经IR模型在学习查询文档相关性模式方面已取得可喜成果，现用BERT模型引入上下文文本表示为IR模型提供深入的文本理解。

解决思路：

1）框架：BERT句对分类框架

，

2）模型：模型使用现成的BERT架构。输入分三部分：一是token嵌入，将查询token和文档token的串联作为输入，使用[SEP]将其分隔为两段；二是segment嵌入，用Q表示查询的分块，D表示文档的分块；三是位置嵌入，从1到N标记每一个查询token，从N+1到M标记每一个文档token。输出部分，第一个标记[CLS]的输出嵌入被用作整个查询-文档对的表示，送入MLP模型做二分类判断。

3）其他：对于长文本，将长文本分为重叠的段落，用神经排序器独立预测每个段落的相关性，文档分数取BERT-firstP或BERT-MaxP或BERT-SumP。在大型搜索日志上精调模型来增加BERT的搜索知识。

实验：

1）数据集：

a. Robust04：0.5M个文本和249个查询组成的新闻语料库。查询包括关键字查询（标题）和较长的自然语言查询（描述）。还包括一个叙述作为相关性评估的指引。

b. ClueWeb09-B：50M个网页和200个带有标题和描述的查询。段落是使用 150 个单词的滑动窗口生成的，步长为 75 个单词。

c. Bing 搜索日志示例：0.1M 查询和 5M 查询-文档对。

2）baseline：无监督基线使用 Indri 的词袋 (BOW) 和顺序依赖模型查询 (SDM)。Learning-to-rank 基线包括具有词袋特征的 RankSVM 和 Coor-Ascent [1]。神经基线包括 DRMM [4] 和 Conv-KNRM。DRMM 使用 word2vec [5] 对单词软匹配进行建模。

基线使用标准的停用词删除和词干提取； BERT 使用原始文本。监督模型用于通过 5 折交叉验证对 BOW 检索到的前 100 个文档进行重新排序。

[1] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional

Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In WSDM.

[4] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance

Matching Model for Ad-hoc Retrieval. In CIKM.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean.

2013. Distributed Representations of Words and Phrases and their Compositional-

ity. In NIPS.

总结：自适应和微调 BERT 在两个不同的搜索任务上实现了高精度，展示了 BERT 语言建模对 IR 的有效性。传统IR模型使用关键词搜索，会去除停用词和标点符号，而停用词和标点符号构建了文本结构，所以在结合了上下文关系的BERT模型中它们扮演了重要角色。因此当系统可以对语言结构进行建模时，用自然语言编写的查询实际上可以实现更好的搜索结果。

其他：