Text Summarization

Bert-based Text Summarization

1. Bert

1.1. 之前存在的问题

1.1.1. 有大量未标注的语料无法使用

1.1.2. 训练出的词向量效果并不好，获取的语义信息较浅。

1.1.3. 不能很好地利用上下文信息解决一词多义问题

1.1.4. 不能很好地应用于长文本的任务

1.1.5. 不能很好地进行迁移学习

1.2. 提出的创新点

1.2.1. word pieces

对于英文文本，可以将每个单词拆分为前缀、中缀、后缀等部分，如：playing -> play + ##ing。这样做相比 Lemmatisation 和 Stemming 可以保留单词的时态、词性等信息，并且减少了词表大小，增大了数据中的 token 数（可学习信息增多），防止 OOV。但是由于在 MLM 任务中可能 mask 的 token 是某个词的中缀，如：probability->pro + ##bali + ##lity，mask 遮住了##bali，但是根据前缀和后缀很容易猜到 mask 了什么，因此需要将整个词（pro ，##bali ，##lity）mask 掉，而不是只 mask 一个 token，这种方法叫 Whole Word Masking（WWM）。

1.2.2. Masked Language Model

每次训练时，随机固定（区别于 roberta）选择 15%的词进行 mask，为了防止模型只对 mask 的位置进行预测，降低学习难度，并且缓解和下游任务，选出的 15%里有 10%随机换成其他单词，10%不变，剩下 80%使用[mask]替换。

1.2.3. Position embedding

1.2.4. Transformer 里使用的是 Position encoding，使用公式计算好每个位置的信息，直接在 embedding 的时候使用，而 Position embedding 是训练出来的。由于 Bert 有充足的训练语料，并且注重提取完整的句子语义信息，而不是特别关注某个词的具体位置，因此使用可训练的 Position embedding 效果好。

1.2.5. Next Sentence Prediction

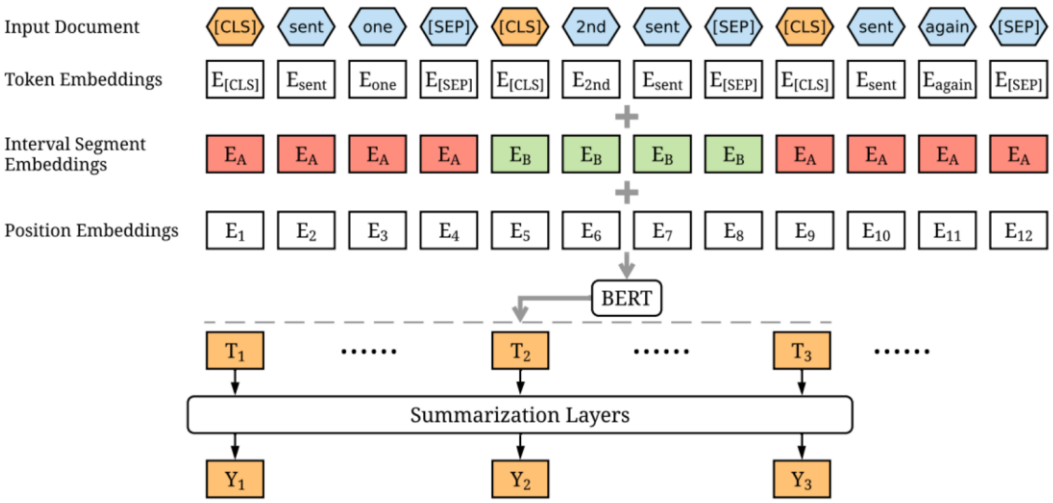
为了解决长文本的语义理解问题，Bert 加入了 NSP 任务来学习上下文的依存关系。

1.2.6. CLS 标记

Bert 在每个输入的句子开头会添加一个 CLS 标记，在训练的过程中，通过前向和后向传播，CLS 标记可以渐渐学习到整个句子的信息（其实 CLS 放在哪里都可以）

2. BertSumExt

Bert for summarization 是基于 Bert 的抽取式摘要模型，它将模型的输入格式进行了修改，原本的 Bert 输入为一个或两个句子，开头添加 CLS，句与句之间添加 SEP，而 Bert for summarizatio 是输入多个句子，每个句子开头添加 CLS，结尾添加 SEP。



将每个 CLS 标记取出，在 Bert 结构之后添加一个分类器，选择哪些 CLS 可以作为摘要抽取出来。这个分类器可以使用简单的一层或两次深度网络，也可以使用复杂一点的 transformer 的 encoder。