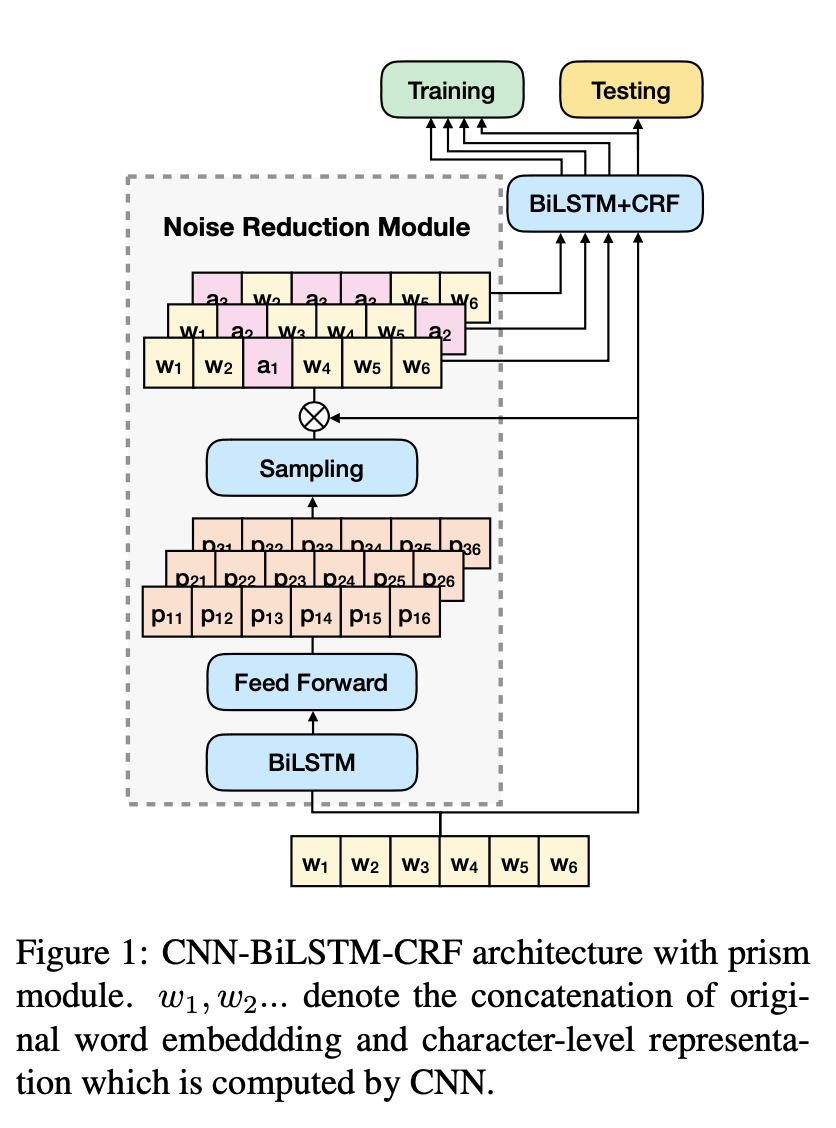
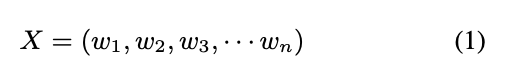
**阅读笔记一**

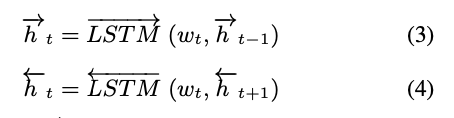
1. 贡献
2. 本文提出一种棱镜模块，用于消除单词的多余语义并减少模型输入层的噪声。
3. 也介绍了一种能够与现有模型一起训练该棱镜模块的结构
4. 方法



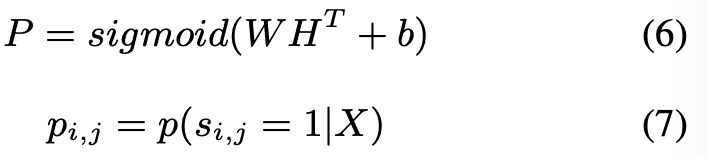
**棱镜模块：通过取代句子的某些单词来获取低噪声的句子**

给定的由n个单词组成的句子X：，其中是句子中第i个单词的词向量。每个单词有m个不同方面的意义，用A来表示，其中，是单词的第i层意思的词向量。

用双向LSTM来捕获输入的单词间的特征

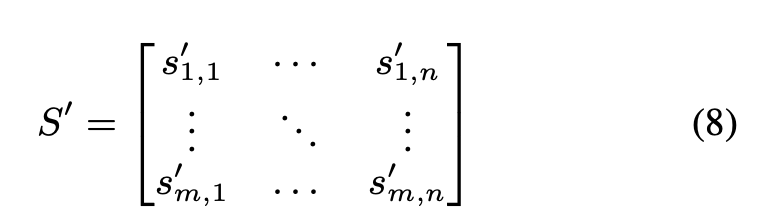
，，。

定义变量来表示第j个单词s是否被第i层意思取代，概率矩阵P(大小为m×n)由构成，是的概率:



其中W（m×2h）是隐层向量H(2h×n)的权重，b为偏置。

是由参数化的具有多项式分布的随机变量，用概率对进行采样：

其中，第i行表示用第i个方面的意思替换。

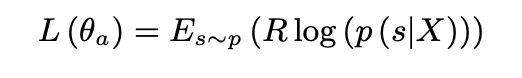
句子用替代完毕之后，得到m个句子，每个句子都从不同方面进行了降噪处理，然后这些去噪的句子和原始语句组成的并行语句被用作下游模型的输入。

**模型训练：棱镜模块与下游模型一起训练，模型中的参数可分为两个部分，下游模型的和。**

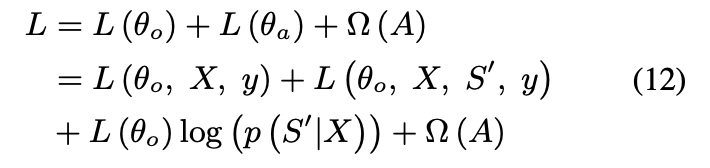
优化的目标在于提高模型的精度，由于模型的输入包含词向量和抽象单词的向量，故参数的损失函数为：



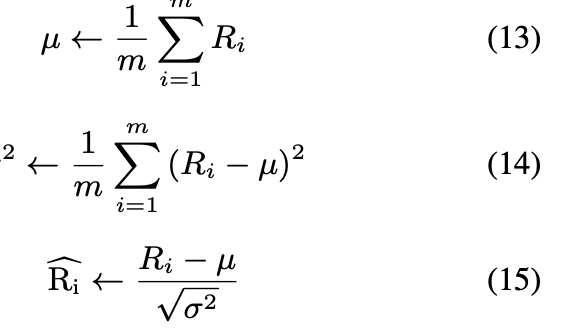
而参数的优化目标是用适当的意思替代适当的词，由于离散的变量，故损失函数是不可微的，本文用梯度/增强的方法优化，损失函数为：

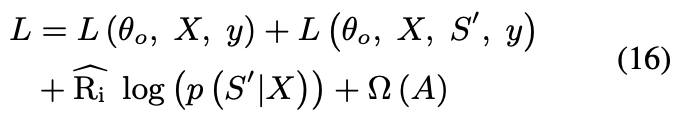


引入一个惩罚项：其中F表示范数，I表示恒等矩阵，通过将等式(2)的A矩阵归一化而来，考虑通过该类矩阵采样而来，对于所有参数，损失函数为：



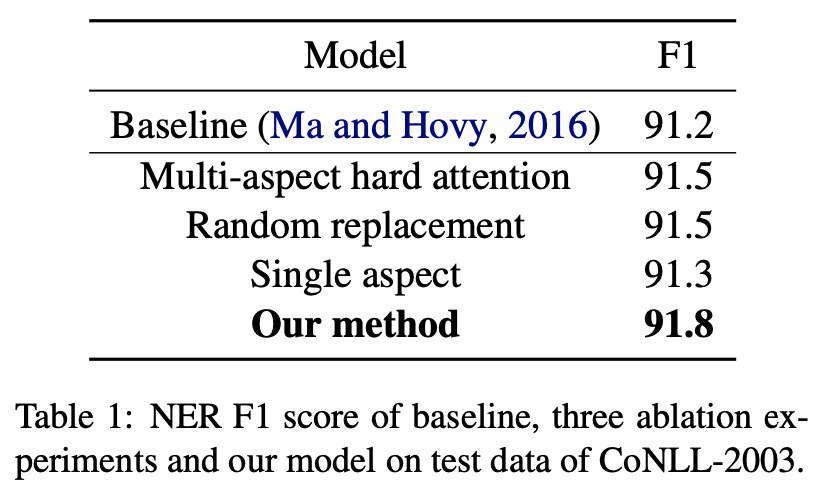
奖励归一化：一种减少方差并使训练过程稳定的新方法，通过使奖励的平均值为0且方差为1来标准化奖励。

，其中均值μ和方差σ，表示归一化的奖励，最终损失函数为：



1. 实验结果

数据集：CoNLL2003；Baseline：Bi-directional LSTM)-CRF

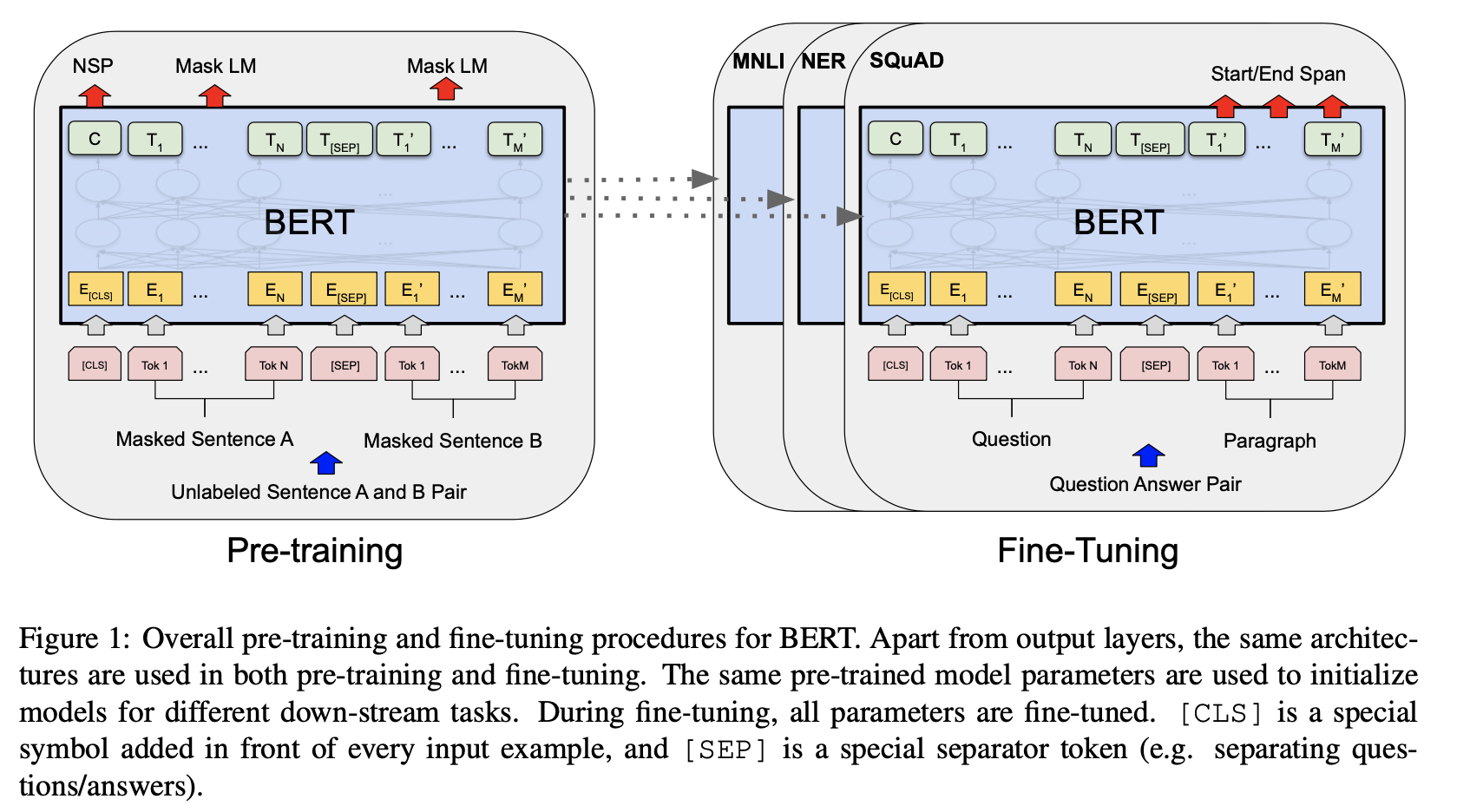


结果分析：(1) 训练每个词的抽象方面的词向量都可以捕获对任务有用的信息； (2) 模型可以学习替换由下游任务(如NER)正确引导的单词；(3) 对每个单词，一个以上的语义与任务相关。

**阅读笔记二**

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

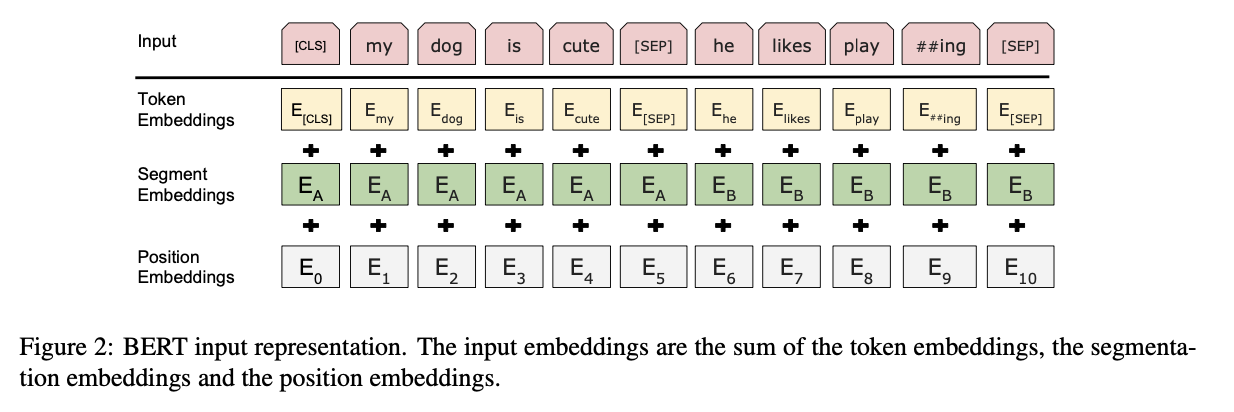
1. 贡献
2. 用Mask Language Model（MLM）来训练的深度双向语言表示向量，思路就是随机遮住15%的单词，让transformer编码器预测这个单词；
3. 两种预训练的方法：Masked模型，下一个句子的预测（NSP）
4. 基于微调的表示模型，可以在一系列句子级和单词级任务上得到比较好的性能。
5. 方法



第一步：预训练，学习输入的向量表示，包括单词、句子和位置特征信息。

第二步：微调，根据预训练得到的模型，结合具体的任务类型微调模型的参数进行训练。

1. 输入表示：此处输入表示句子或句子对



对于一个句子或句子对，BERT学习每个token、segment向量以及位置信息。

其中，token就是词向量（WordPiece），segment向量就是句子的向量表示，位置向量用transformer（注意力机制）实现。

1. 词的预训练过程：mask语言模型，遮住15%的词进行预测。

被遮住的词有80%的时间被[MASK]替代；10%的时间被一个随机词替代；10%时间保持不变。该过程的优势在于Transformer编码器会保留每个输入token的上下文表示。

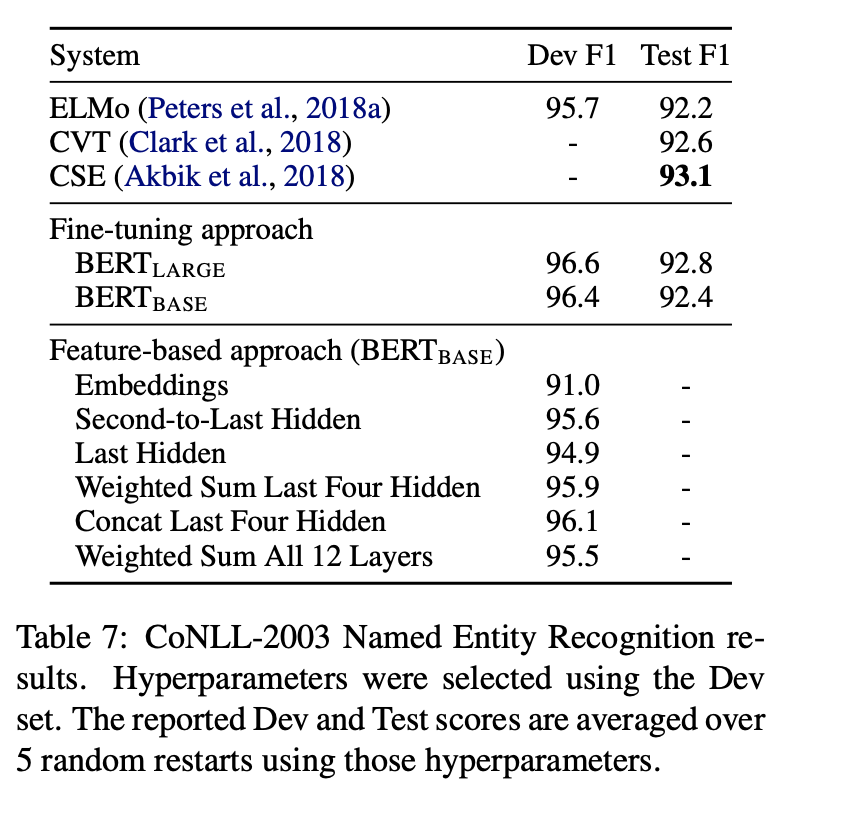
预训练过程主要用文档级别的数据，重点考虑长距离连续序列。

1. 预测下一个句子：预训练一个二分类的模型，来学习句子之间的关系

主要应用于基于两个句子的关系的任务，比如问答和推理。训练样本正负的比例为一比一，50%的时间B是A的下一句(IsNext)，50%的时间A的下一句是语料库中随机的句子(NotNext)。Figure 1中的C就被用于下一个句子的预测（NSP）任务。

1. 实验结果

该模型在11个自然语言处理任务上都取得了比较好的结果，此处仅分析命名实体识别任务的效果。



1. **其他内容**

WordPiece Embedding方法是一个将英文文本向量化的过程，学习到的是上下文无关的表示。利用 双字节编码（Byte-Pair Encoding）将单词的词义和时态分开，减少单词表的数量，提高训练速度。

具体来说，就是英文单词有不同的前后缀，以时态来讲，love, loves, loving, loved这四个词都表示love的含义，WordPiece编码单词的时候，拆分过程借鉴了数据压缩算法BPE，首先将单词划分为字符，统计字符对出现的次数，每次将出现次数最多的字符对保存下来，直到循环次数结束。

这种处理方式在遇到训练语料中未出现的新词的时候也可以通过子字符对的拼接来生成新的编码结果。