预训练：有两个相似任务A和B，A已经解决并得到了模型a，B没有解决，通过a来解决问题B

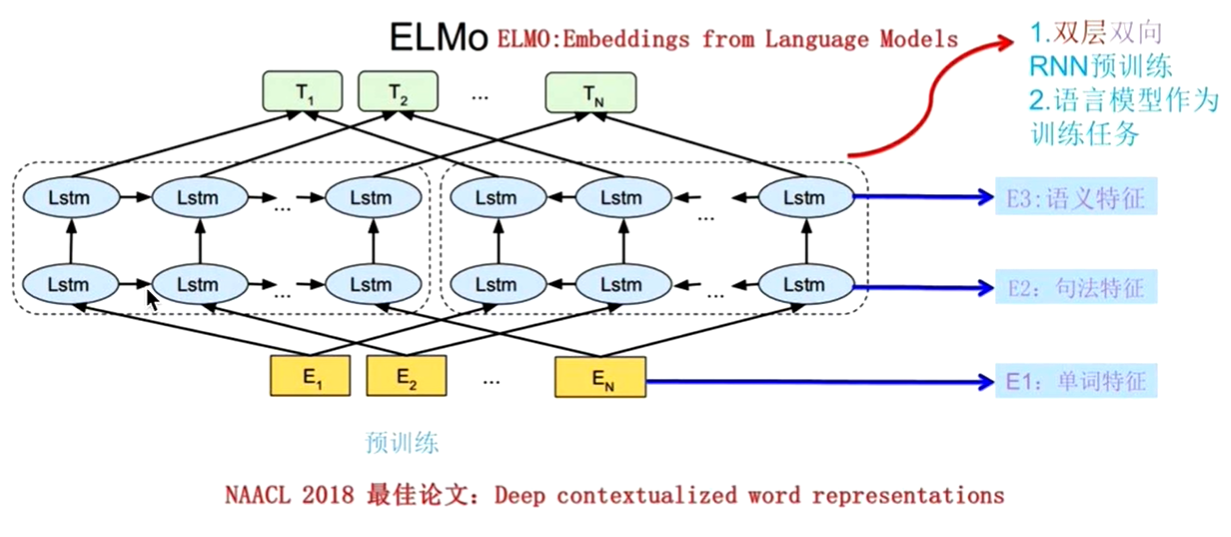
双层感知机 : softmax ( w2 ( tanh ( w1 x + b1 ) ) + b2 )

独热编码--》词向量 ( word2vec ) --》ELMO

独热编码不能够表示词间关系，并且占用空间较大

Word2vec能表示词间关系，但是不能表示单词的多意性

EMLo解决了多义词的问题，不止训练一个Q矩阵还将上下文的信息融合到了Q矩阵当中，三层信息（单词特征、语法特征、语义特征）进行叠加，E1+E2+E3=K1，得到一个新的词向量，E2和E3的不同会使得相同的E1在K1中展现不同的含义



Word2vec : 是一个神经网络语言模型其主要任务是生成词向量Q，运用了预训练的思想，细分为两种：  
1.CBOW : 一个老师教多个学生，由上下文推测某个词

2.Skip-gram：多个老师教一个学生，由某个词推测上下文

NLP里的任务：接收一个问题X（Ni+ck），给出一个回答（handsome）

预训练语言模型（给出一句话，先试用独热编码，再使用Word2vec预训练好的Q矩阵直接得到词向量，然后进行接下来的任务）两种方法1.冻结：不改变Q矩阵，2微调：随着任务的改变而改变Q矩阵

图示

描述已自动生成

NNLM模型，预测下一个词，副产品是词向量

LSTM和RNN的巨大缺点：无法并行、长期依赖。

注意力机制

背景：对于传统的CNN和LSTM等很难从大量信息鉴别出重要关键的信息，注意力机制由此提出

计算实现：存在查询对象Q和被查询对象K，计算两者点乘求相似度an=Qn \* Kn，做一层sofrmax(a1,a2,a3,...,an)进行归一化得到概率,计算(a1,a2,a3,...）\* (v1,v2,v3,...)得到新的包含注意力的V，

自注意力机制：得到的词向量不仅具有词法特征还具有语法和语义特征，解决了长期依赖、并行计算和词间语义关系问题

不仅K ~ V ~ Q三者来源于同一个X，通过X找到X中的重要信息，之后进行注意力机制的操作，计算得到Z（z1 , z2 , z3 , ... , zn）

手机屏幕的截图

描述已自动生成

图表

描述已自动生成

区别：自注意力机制qkv同源，所以一定程度可以看成qkv相等，而注意力机制kv同源或者k就是等于v，但是q和k不是同源的。并且注意力机制只是宽泛的表明了注意力机制必须得这样做，而自注意力机制是狭隘的表明了qkv必须同源，注意力机制包括自注意力机制。

总结：首先输入部分，需要做word embedding,也就是词向量化。  
然后： Q, K,V这三个矩阵是通过原始输入X做线性变换得来的。  
Q=X\* Wq, K=X\* Wx, V= X\*Wv, Wq,Wx,Wv是可训练的参数矩阵，初始值可以随机化生成。  
然后就可以通过点积运算生成注意力值了  
而模型训练的过程可以更新Wq,Wx,Wv的参数值。、

掩码自注意力机制：进行词向量的生成过程中，当词数不明确时，只能逐次生成，使得当前词向量不能包含其后词向量的内容，掩码自注意力机制分批次给出词向量信息，直到最后一个词，才将所有信息给出。

图表

描述已自动生成