NLP 基础算法梳理(One-hot, Word2vec, Elmo, Bert, GPT)

Friday, May 28, 2021 2:13 AM

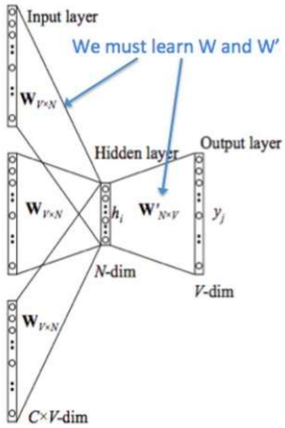
教程来自Hongyi Li的教程和Shusen Wang的教程以及Jay Alammar的博客

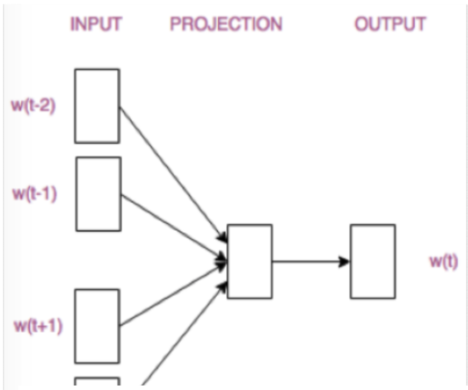
自从seq2seq模型引入到NLP领域之后，后续的模型基本上都延续了这encoder-decoder的结构。

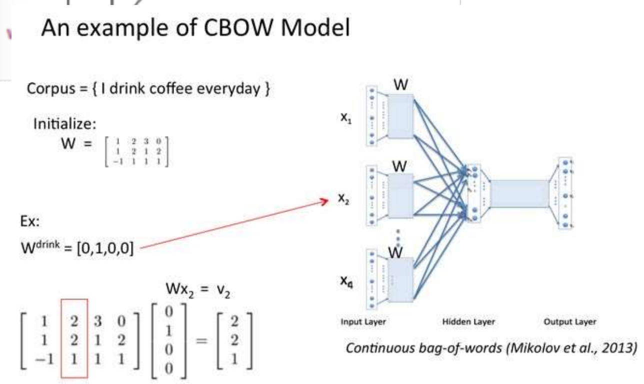
其中 encoder 的目的是将字词进行编码，得到字词在多维向量空间representation，这个过程叫word embedding

1. Embedding的方法更迭

- One-hot encoding：0,1对词库里的所有词进行编码。
  - 缺点是词的数量增多的时候会出现维度的爆炸；此外还有矩阵稀疏的问题。
- Word2vec，推进了NLP领域的发展
  - CBOW





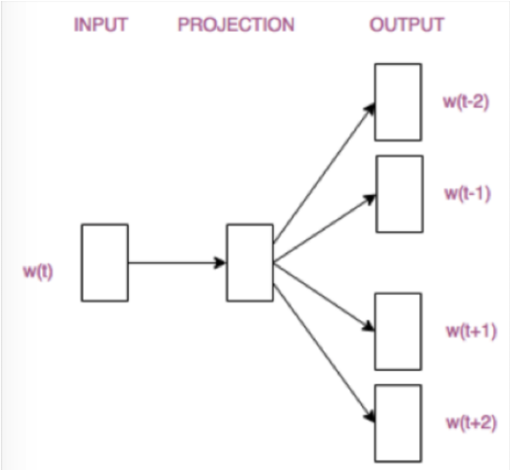


优点：向量运算有实际的意义，如King-Man+woman = queen

缺点：每个单词只有一个向量来表征，没法分辨同义词

为了减少训练的时间，有同义词之间只是微调参数，引入了霍夫曼树的方法

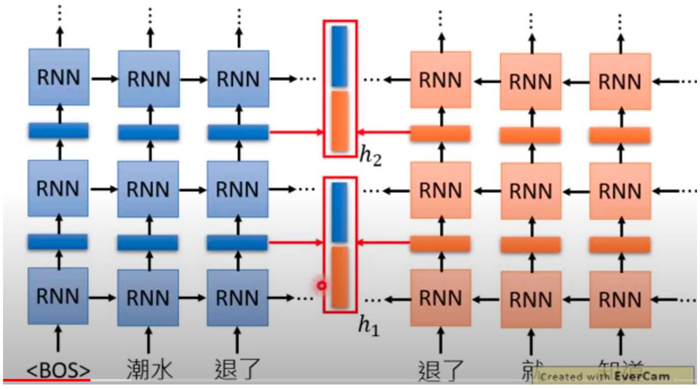
  - Skip-gram



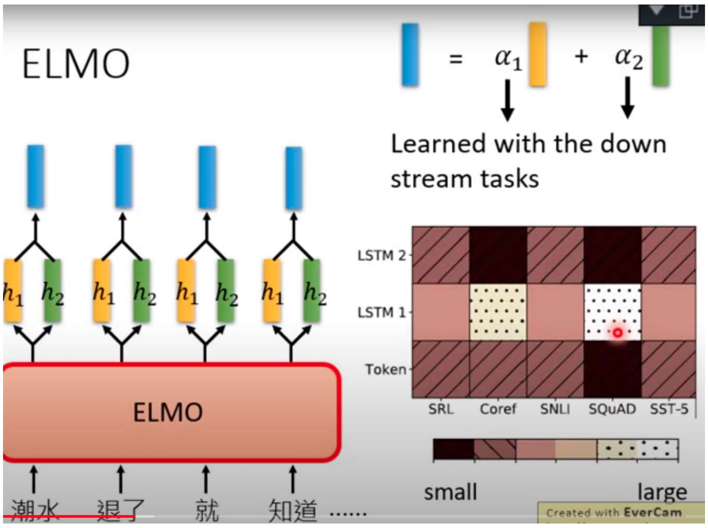
优点：解决了contextual的问题，根据上下文意思的不同，每个词的Embedding是不一样的

缺点：LSTM是seq2seq的，速度慢

- ELMo（Embedding from language models）



每一层的LSTM能够生成一个latent representation

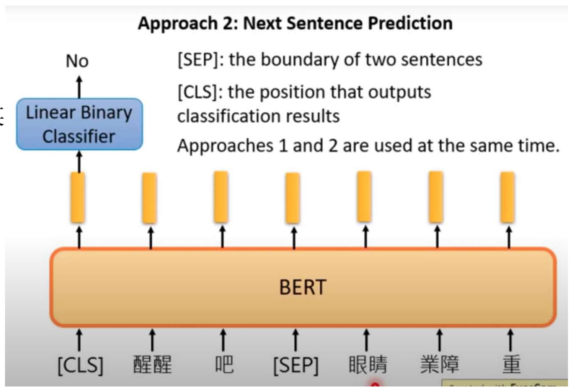


不同任务的a1和a2是不一样的，这两个参数根据不同的任务来训练

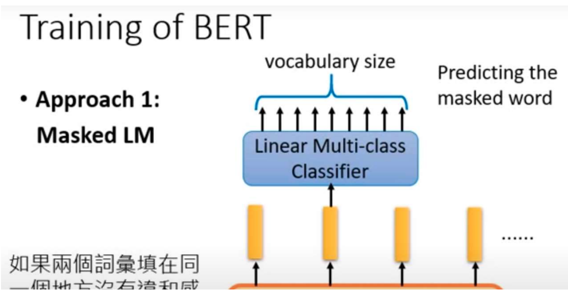
右下角的图表示的是哪一层的参数对任务的影响最大

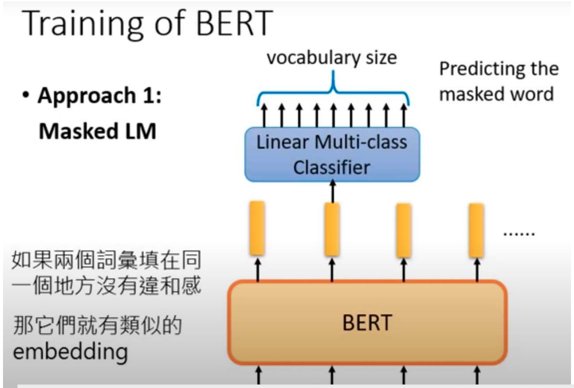
- BERT（Bidirectional Encoder representation from transformers）

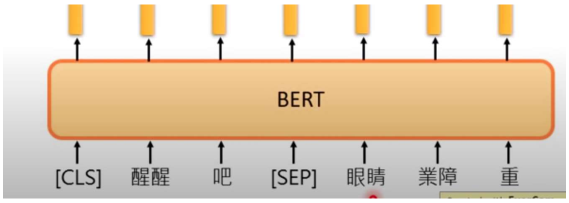
两种训练方法，一种是Masked LM，预测句子中被遮盖住的某

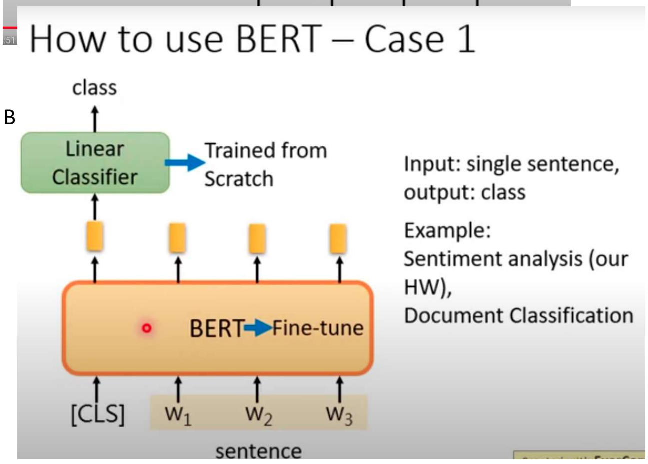


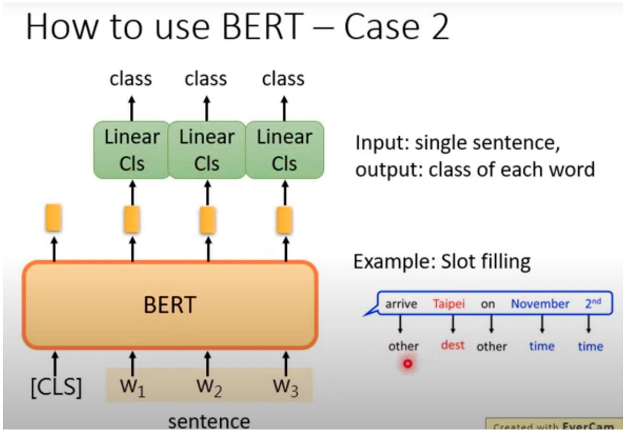
一般情况下两种都是一起训练的

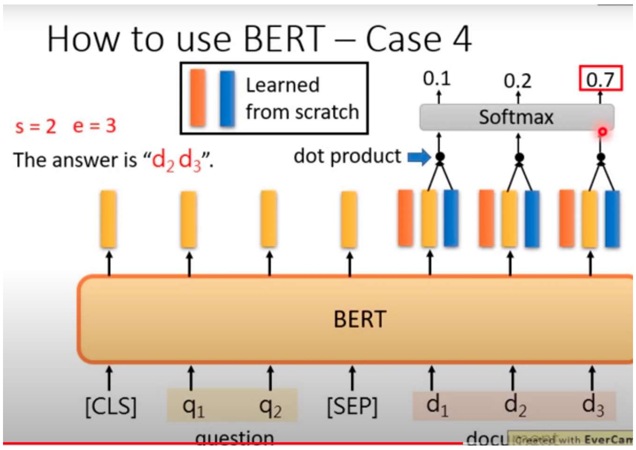


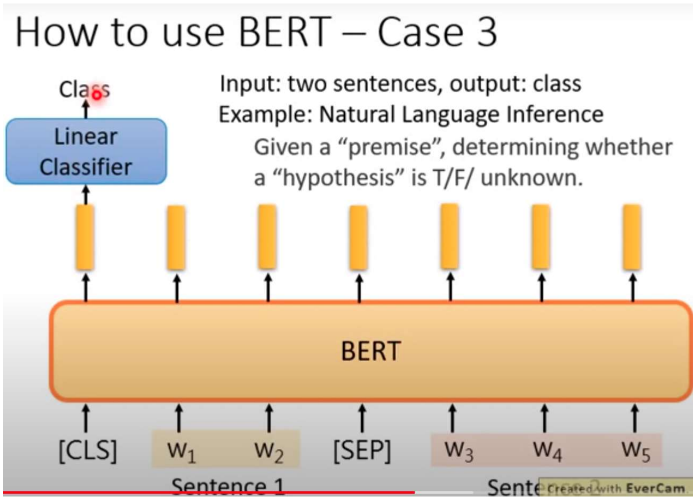






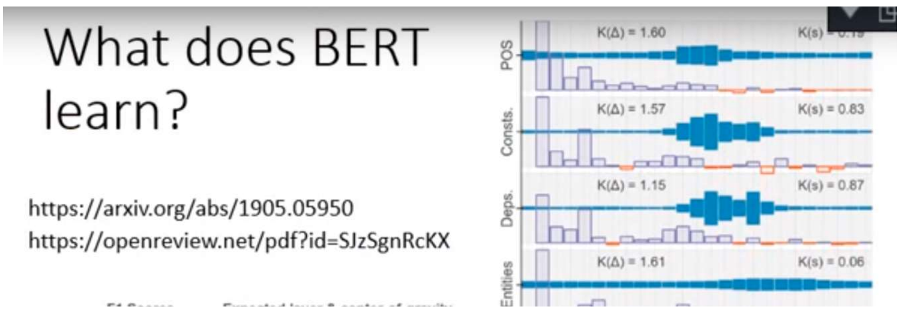






这个分层，就是把每一层encoder的hide state提取出来

其实各个层级的hiden state加起来显示的F1是不一样的

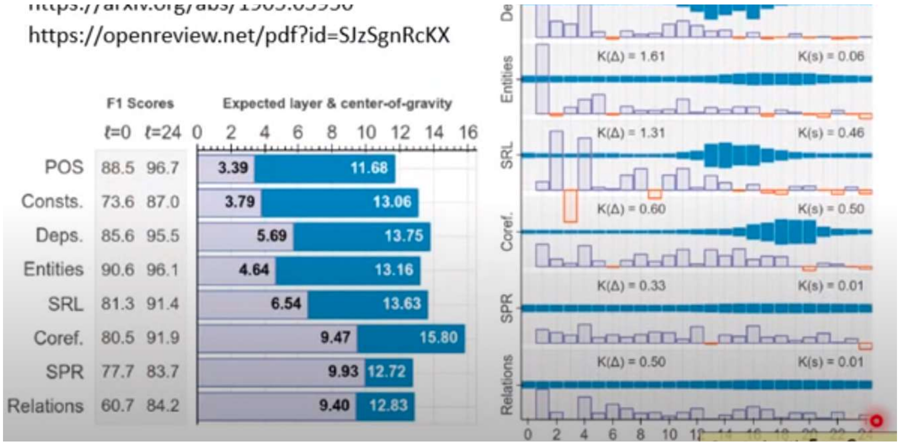


FROM PAPER:
Dissecting Contextual Word Embeddings: Architecture and Representation

WHAT DO YOU LEARN FROM CONTEXT, PROBING FOR

BERT Rediscovers the Classical NLP Pipeline

Evaluation as word representations
- MultiNLI dataset
- Semantic Role Labeling, Ontonotes dataset
- Constituency parsing, Penn Treebank



Evaluation as word representations

- MultiNLI dataset
- Semantic Role Labeling, Ontonotes dataset
- Constituency parsing, Penn Treebank
- Name entity recongnition: NER

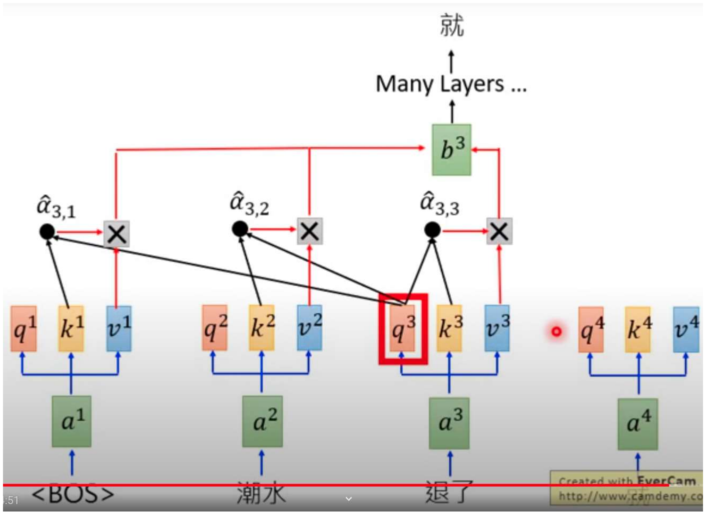
Contextual similarity

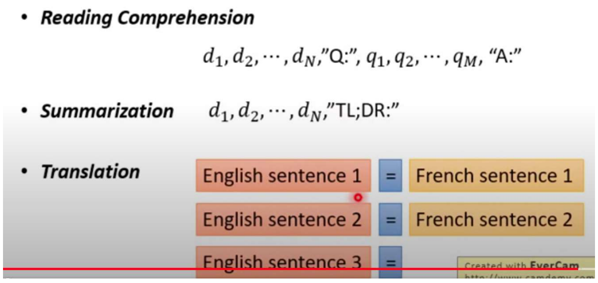
- Intra-sentence similarity
- Span representations
- Unsupervised pronominal coref

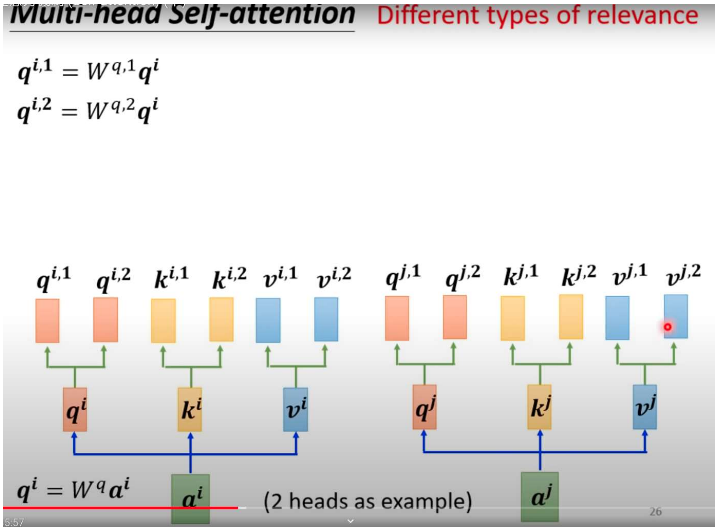
Probing contextual information

- POS：part-of-speech tagging
- Constituency parsing

- GPT（Generative Pre-training）是Transformer的decoder







2. Attention

- (NNLM) A Neural Probabilistic Language Model
- (Word2Vec) Distributed Representations of Words and Phrases and their Compositionality
- (GloVe) GloVe: Global Vectors for Word Representation
- (ELMo) Deep contextualized word representations
- (GPT) Improving Language Understanding by Generative Pre-Training
- (BERT) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding