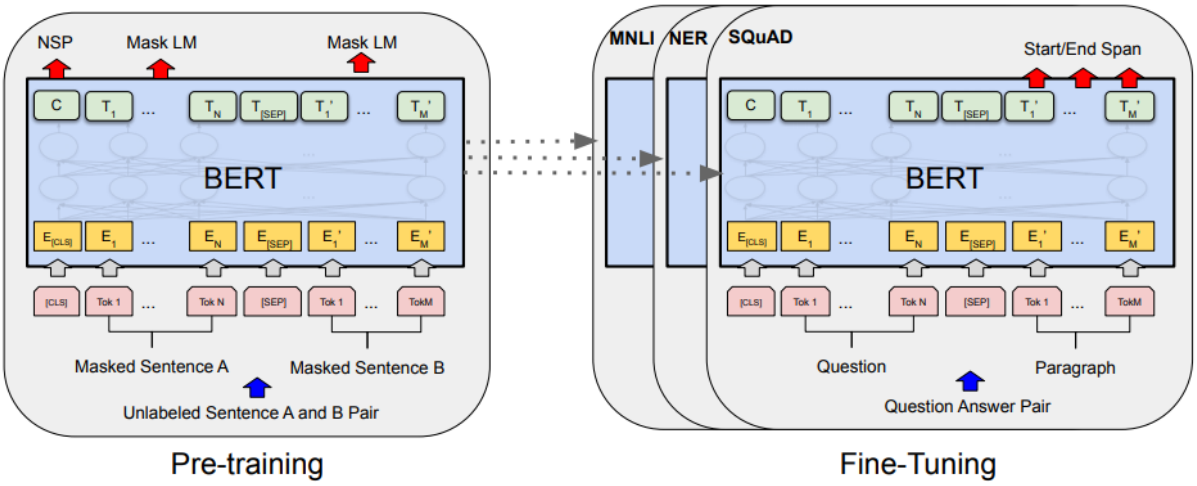
BERT 내용정리



- BERT는 절대 위치 임베딩을 사용하는 모델이므로 일반적으로 왼쪽보다는 오른쪽의 입력을 채우는 것이 좋습니다.
- BERT는 마스크 언어 모델링(MLM) 및 다음 문장 예측(NSP) 목표로 훈련되었습니다. 마스크 토큰을 예측하고 일반적으로 NLU에서 효율적이지만 텍스트 생성에는 최적이 아닙니다.
- 무작위 마스킹을 사용하여 입력을 손상시킵니다. 더 정확히 말하면 사전 학습 중에 지정된 비율의 토큰(일반적으로 15%)이 다음과 같이 마스킹됩니다.
    - 확률 0.8의 특수 마스크 토큰
    - 확률 0.1로 마스크된 것과 다른 랜덤 토큰
    - 확률 0.1의 동일한 토큰
- 모델은 원래 문장을 예측해야 하지만 두 번째 목적이 있습니다. 입력은 두 문장 A와 B(그 사이에 분리 토큰 있음)입니다. 확률 50%로 문장은 코퍼스에서 연속적이고 나머지 50%에서는 관련이 없습니다. 모델은 문장이 연속적인지 여부를 예측해야 합니다.

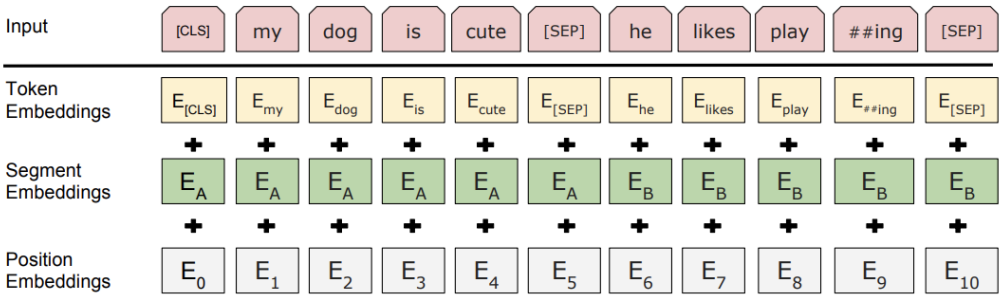


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

<Input>

- Token Embedding: Word Piece 임베딩 방식 사용, 각 Char(문자) 단위로 임베딩을 하고, 자주 등장하면서 가장 긴 길이의 sub-word를 하나의 단위로 만듭니다. 자주 등장하지 않는 단어는 다시 sub-word로 만듭니다. 이는 이전에 자주 등장하지 않았던 단어를 모조리 'OOV'처리하여 모델링의 성능을 저하했던 'OOV'문제도 해결 할 수 있습니다.
- Segment Embedding: Sentence Embedding, 토큰 시킨 단어들을 다시 하나의 문장으로 만드는 작업입니다. BERT에서는 두개의 문장을 구분자([SEP])를 넣어 구분하고 그 두 문장을 하나의 Segment로 지정하여 입력합니다. BERT에서는 이 한 세그먼트를 512 sub-word 길이로 제한하는데, 한국어는 보통 20 sub-word가 한 문장을 이룬다고 하며 대부분의 문장은 60 sub-word가 넘지 않는다고 하니 BERT를 사용할 때, 하나의 세그먼트에 128로 제한하여도 충분히 학습이 가능하다고 합니다.
- Position Embedding: BERT의 저자는 이전에 Transformer 모델을 발표하였는데, Transformer란 CNN, RNN 과 같은 모델 대신 Self-Attention 이라는 모델을 사용하는 모델입니다. BERT는 Transformer의 인코더, 디코더 중 인코더만 사용합니다. Transformer Self Attention은 입력의 위치를 고려하지 않고 입력 토큰의 위치 정보를 고려합니다. 그래서 Transformer모델에서는 Sinusoid 함수를 이용하여 Positional encoding을 사용하고 BERT는 이를 따서 Position Encoding을 사용합니다.

데이터들을 임베딩하여 훈련시킬 데이터를 모두 인코딩 하였으면, pre_trainning에 들어가게 됩니다. 기존의 방법들은 보통 문장을 왼쪽에서 오른쪽으로 학습하여 다음 단어를 예측하는 방식이거나, 예측할 단어의 좌우 문맥을 고려하여 예측하는 방식을 사용합니다.
특히나 BERT는 언어의 특성을 잘 학습하도록

- MLM(Masked Language Model)
- NSP(Next Sentence Prediction)

위 두가지 방식을 사용합니다!