# BERT

**B**idirectional **E**ncoder **R**epresentations from **T**ransformers

김윤아

# 목차

# BERT란?

구글에서 2018년도에 발표한 언어모델로,

Transfomer 아키텍쳐 기반으로 하고, 양방향 학습을 사용하여 좋은 성능을 보인다.

큰 양의 텍스트 데이터를 사용하여 사전학습 되며,

이후 특정 자연어 처리 작업에 fine-tuning하여 사용할 수 있다.

# BERT의 특징

**①BERT는 양방향성을 포함하여 문맥을 더욱 자연스럽게 파악할 수 있다.**

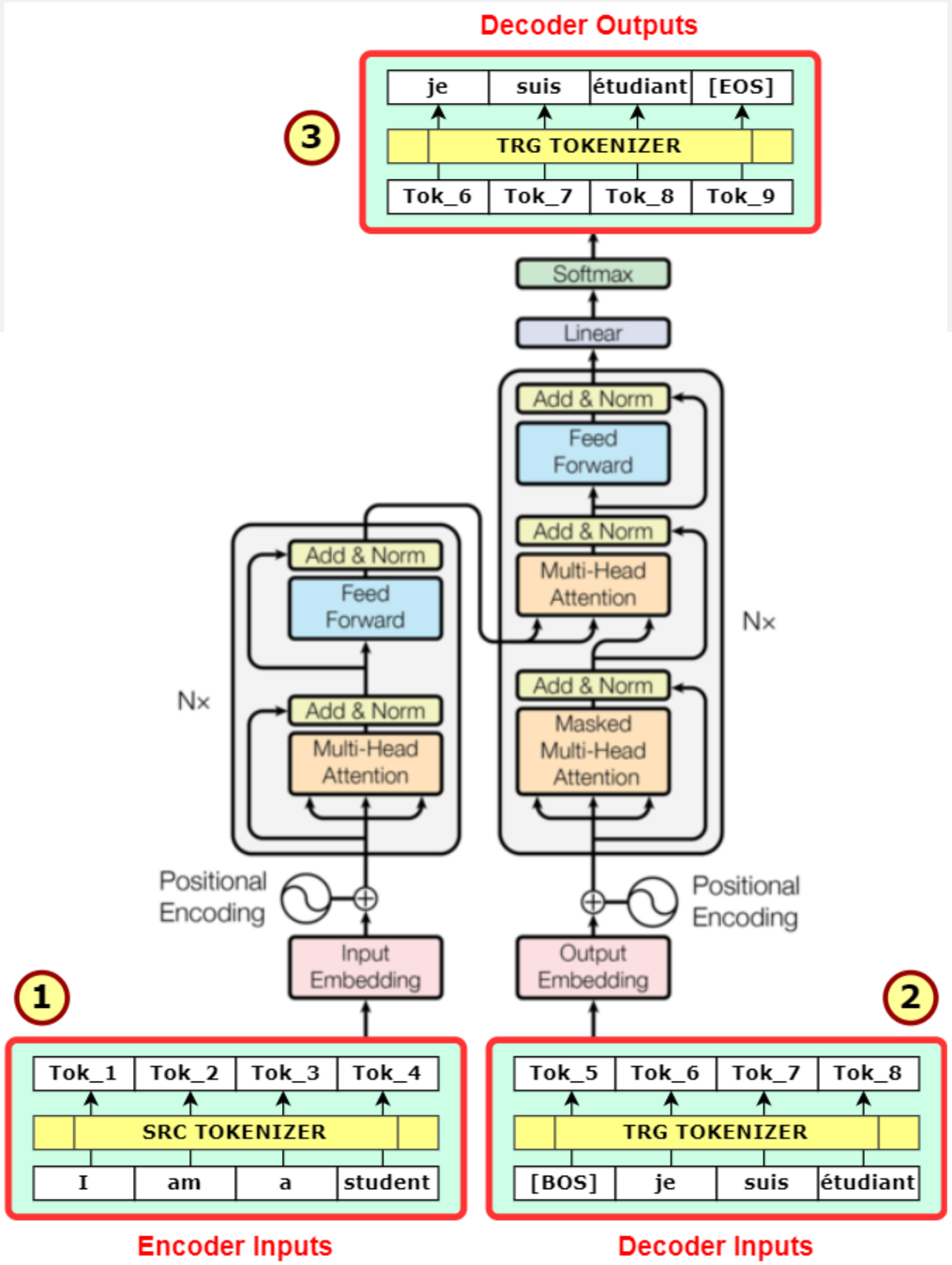
# BERT의 특징

**②BERT는 pre-training이된 모델이다.**

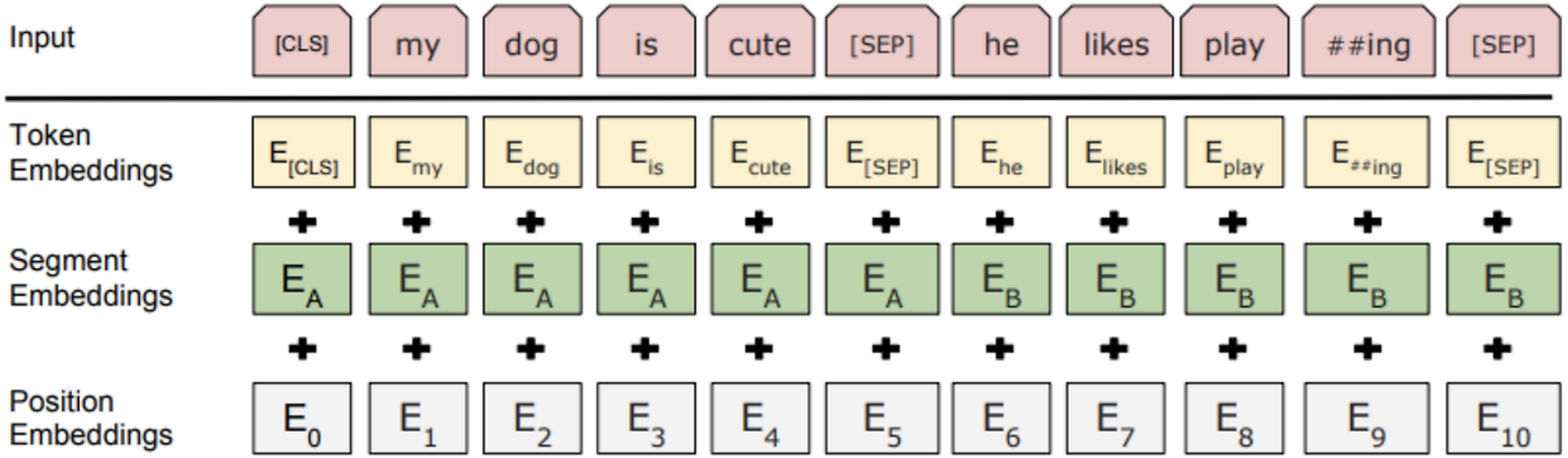
BERT 모델은 기본적으로 대량의 단어 임베딩 등에 대해 사전 학습이 되어 있는 모델
상대적으로 적은 자원만으로도 충분히 자연어 처리의 여러 일을 수행할 수 있다.

# 기존 모델 구조

기존 모델은 대부분 encoder-decoder 으로 이루어져있음
(GPT, Transfomer 구조 모두)
input에서 text의 표현을 학습하고,
decoder에서 우리가 원하는 task의 결과물을
만드는 방식으로 학습이 진행



# BERT의 Input




세 가지 임베딩(Token, Segment, Position)을 사용해서 문장을 표현

# BERT의 Input

BERT Input : Token Embedding + Segement Embedding + Position Embedding

BERT는 세 가지 임베딩을 합치고 Layer 정규화와 Dropout을 적용하여 입력으로 사용

## Token Embeddings

WordPiece 토크나이저를 사용하여 입력 텍스트를 토큰으로 분할

두 가지 특수토큰(CLS,SEP)을 사용하여 문장을 구별

## Segment Embeddings

토큰으로 나누어진 단어들을 다시 하나의 문장으로 만듬

첫번째 SEP 토큰까지는 0으로 그 이후 SEP 토큰까지는 1 값으로 마스크를 만들어 각 문장들을 구분

## Position Embeddings

토큰의 순서를 인코딩
Transformer 에서는 Sigsoid 함수를 이용한 Positional encoding을 변형하여 Position Embeddings를  사용

# BERT의 Pre-Training

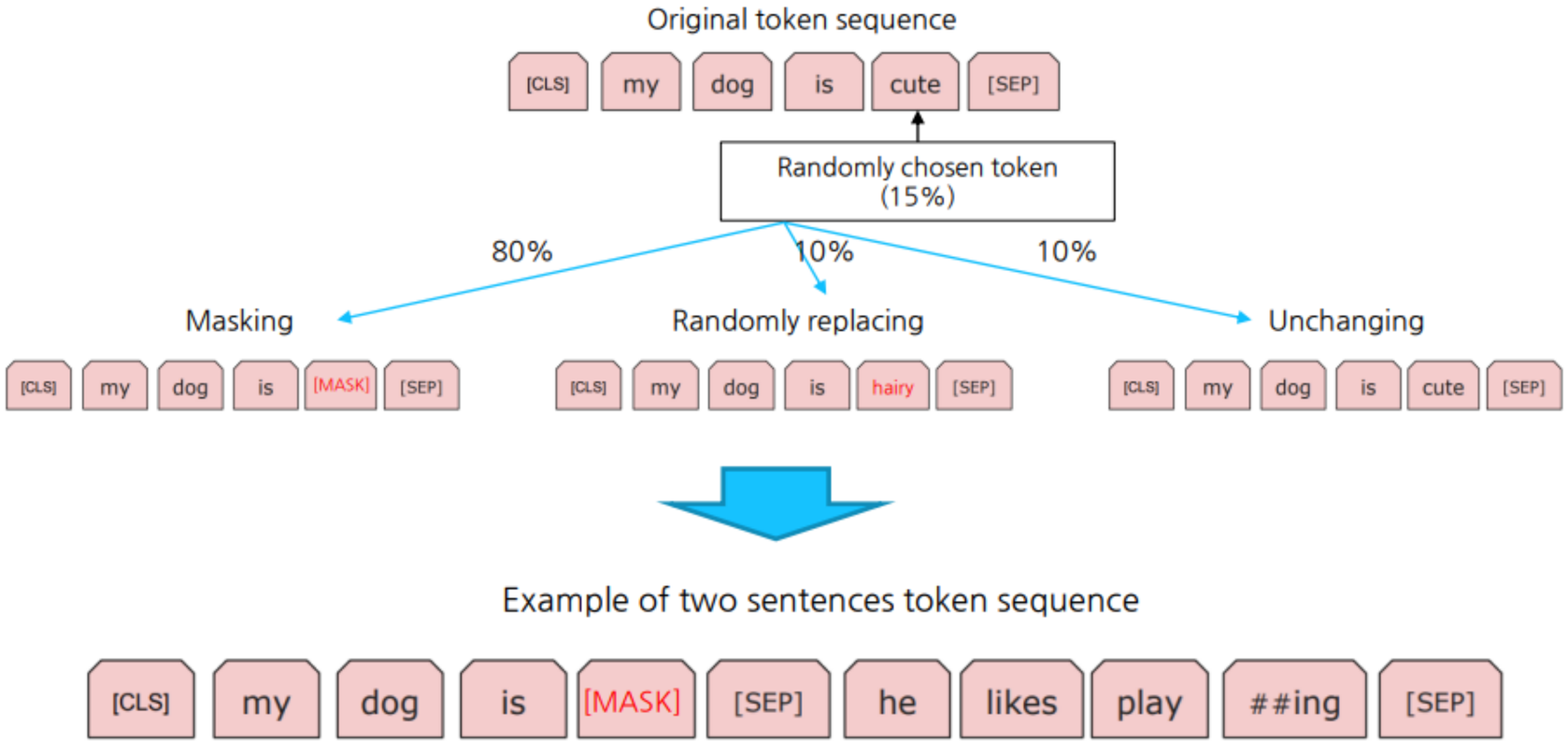
## Masked Language Model (MLM)

MLM 은 입력으로 사용하는 문장의 토큰 중 15%의 확률로 선택된 토큰을 MASK 토큰으로 변환시키고,
언어 모델을 통해 변환되기 전 MASK 토큰을 예측하는 언어 모델

ex) 나는 하늘이 예쁘다고 생각한다 -> 나는 하늘이 [Mask] 생각한다.
ex) 나는 하늘이 예쁘다고 생각한다 -> 나는 하늘이 흐리다고 생각한다.
ex) 나는 하늘이 예쁘다고 생각한다 -> 나는 하늘이 예쁘다고 생각한다.

# BERT의 Pre-Training

## Masked Language Model (MLM)



# BERT의 Pre-Training

## Next Sentence Prediction (NSP)

두 문장이 주어졌을 때, 두 번째 문장이 첫 번째 문장의 다음 문장인지 아닌지를 예측하는 것으로 학습한다.

이 방법을 통해 모델은 문장과 문맥의 의미를 파악하는 능력을 배우게된다.

〈문장유형 50%〉

첫 번째 유형: 첫 번째 문장과 두 번째 문장은 원본 문장의 이어지는 문장이다.
두 번째 유형: 첫 번째 문장과 두 번째 문장은 원본 문장에서 이어지지 않는 관계 없는 문장이다.

# 다양한 bert 모델

- **Sentence -BERT**

의미 있는 문장 임베딩을 도출할 수 있도록 modified된 BERT 네트워크

1. BERT의 [CLS] 토큰의 출력 벡터를 문장 벡터로 간주
2. BERT의 모든 단어의 출력 벡터에 대해서 평균 풀링을 수행한 벡터를 문장 벡터로 간주
3. BERT의 모든 단어의 출력 벡터에 대해서 맥스 풀링을 수행한 벡터를 문장 벡터로 간주

- **KOBERT**

KoBERT는 BERT 모델에서 한국어 데이터를 추가로 학습시킨 모델로,
한국어 위키에서 5백만개의 문장과 54백만개의 단어를 학습시킨 모델