Paper Review

[NAACL 2019] BERT: Pre-training of Deep Bidirectional Transformers For Language Understanding

Dept. of Computer Science & Engineering
Artificial Intelligence and Data Mining Lab(AIDML)
202122029 Meeyun Kim

Contents

1. Introduction

2. About BERT

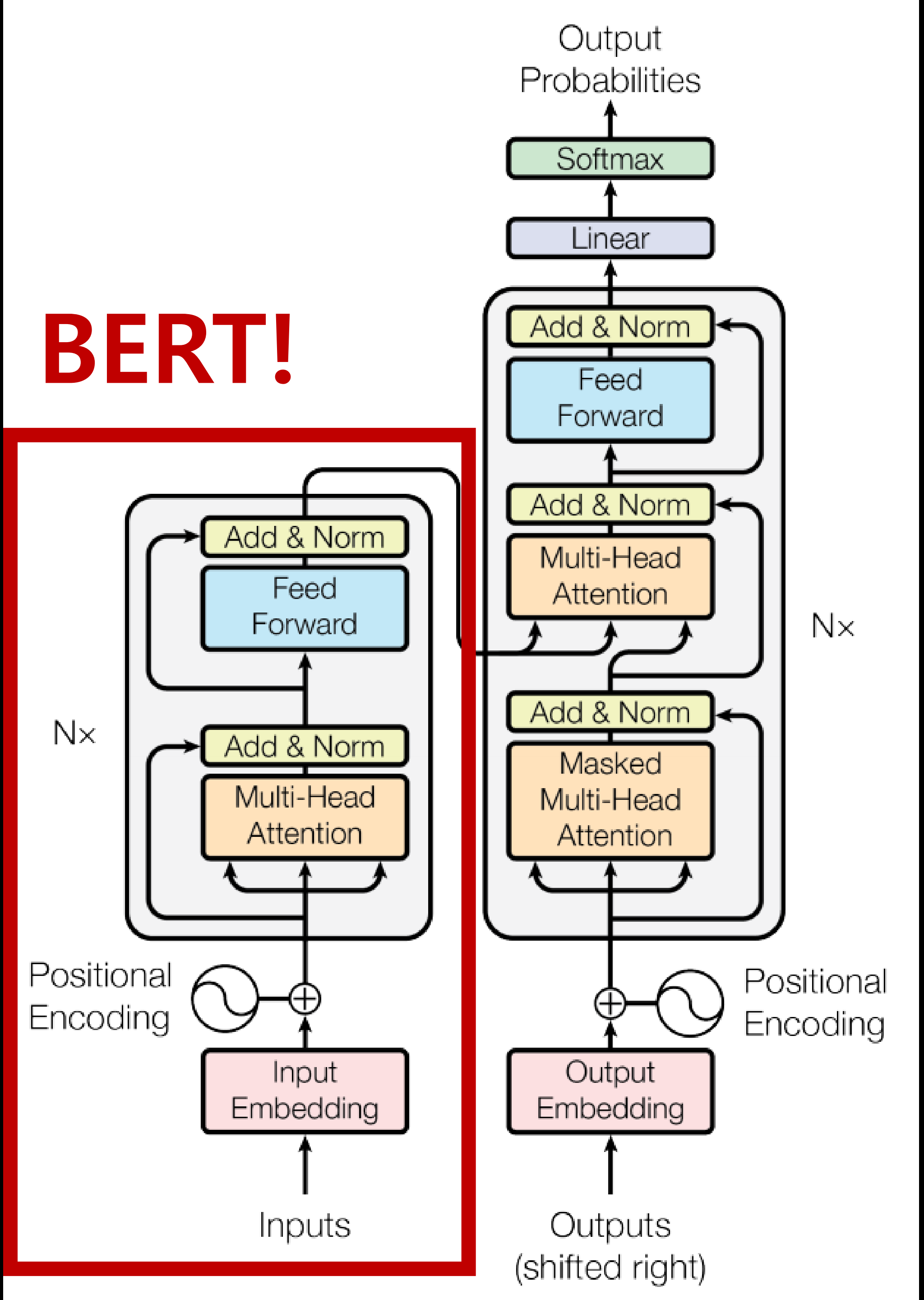
3. Result

1. Introduction

0) Transformer

- Encoder + Decoder 구조

- **Transformer의 Encoder 부분을 발전시킨 것이 BERT!**



1) What is BERT?

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- **Pre-Training**, **Fine-Tuning** 방식을 적용, 개선하여 transfer learning을 용이하게!
- **Masked Language Model(MLM)**을 이용한 성능 향상.

* Language Modeling

- 문장이 있을 때, 다음에 올 단어를 예측하는 task.

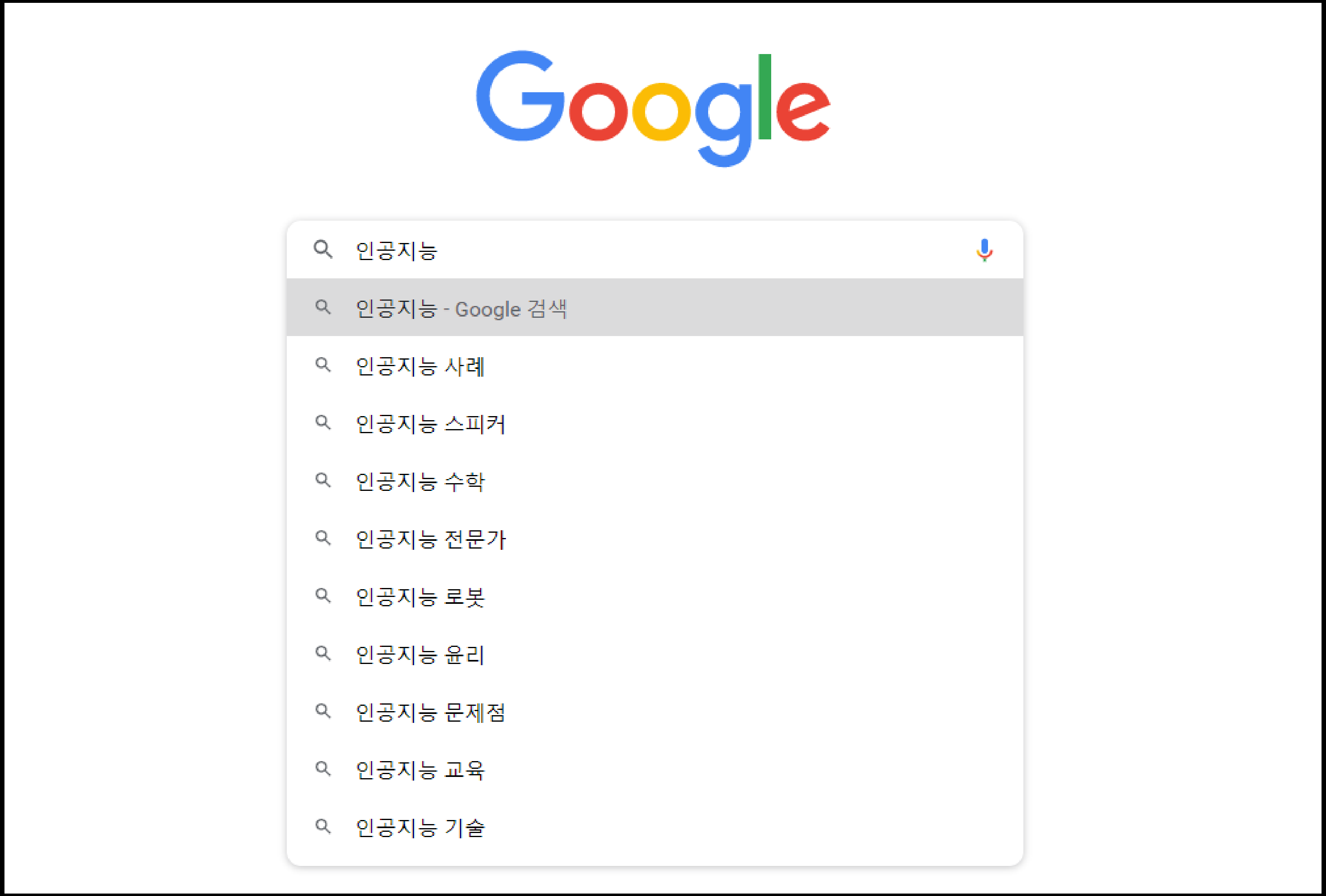
$$\textit{the students opened their } \underline{\qquad\qquad} \rightarrow \begin{cases} \textit{books} \\ \textit{laptops} \\ \textit{minds} \end{cases}$$

- 조건부 확률을 이용해 n번째 나올 단어의 확률 계산.

$$\prod_{i=1}^{n} P(w_i | w_1, \ldots, w_{i-1})$$

- 확률값 $P(books), P(laptops), P(minds)$를 비교.

* Language Modeling (Cont'd)



2) Why Bidirectional?

- **Unidirectional**한 구조는 성능을 제한할 수 있음.
- **left-to-right** 진행 구조 → **양쪽 문맥 이해 힘듦.**
- ex) **OpenAI GPT** (next token만을 맞출 수 있는 language model 방식.)

2. About BERT

1) Model Architecture

- **BERT$_{BASE}$: L=12, H=768, A=12, Total Parameters=110M**

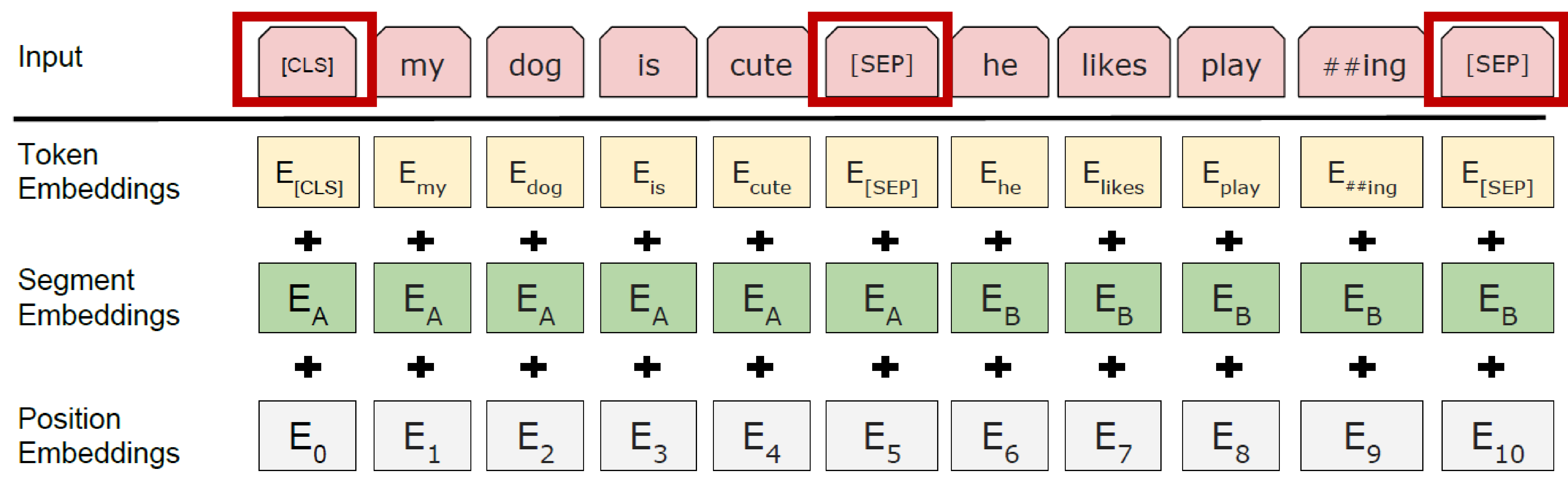
(OpenAI GPT와 hyper parameter가 동일!)

- **BERT$_{LARGE}$: L=24, H=1024, A=16, Total Parameters=340M**

(number of layers - **L**, the hidden size - **H**, the number of self-attention heads - **A**)

- **BookCorpus**(800M words) + **English Wikipedia**(2,500M words)

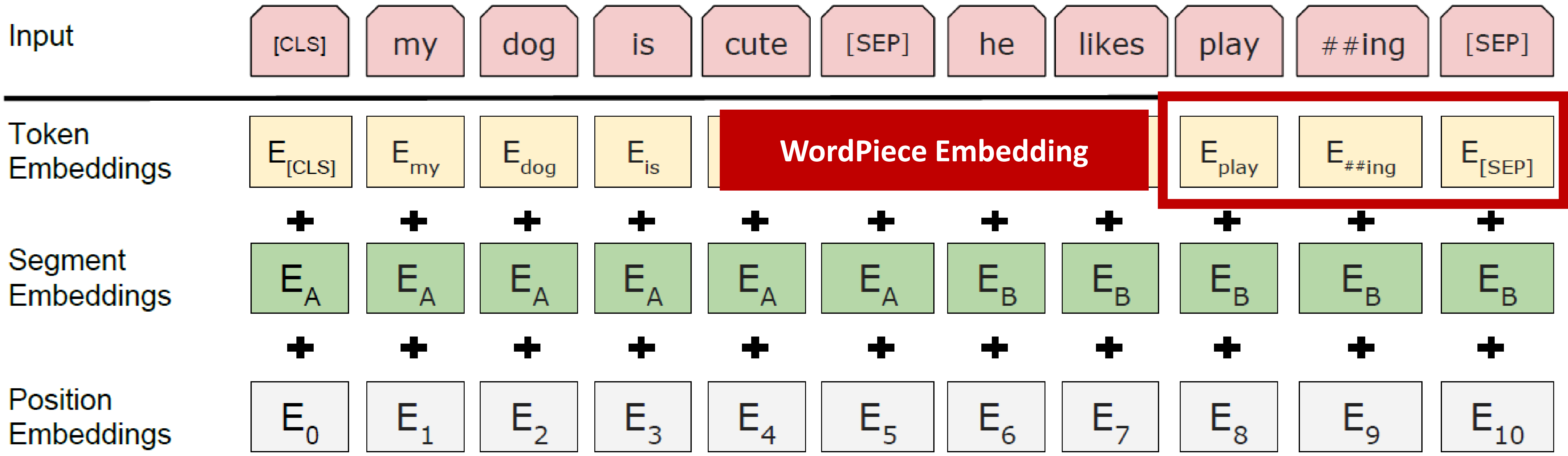
2) Input Representation



[CLS] : Classification task에 사용되기 위한 vector. 문장 전체가 하나의 vector로 표현된 special token.

[SEP] : 두 문장이 input으로 들어왔을 때, 이를 구분하기 위한 token.

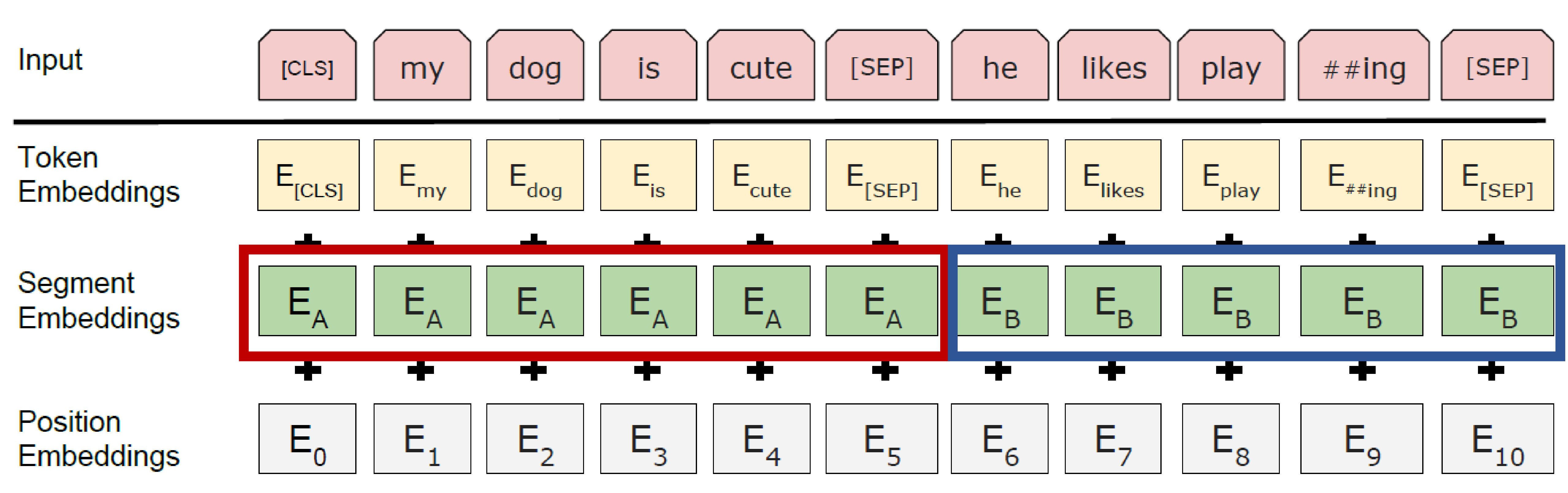
2) Input Representation (Cont'd)



[WordPiece Embedding]

- 명확한 의미 전달이 가능해져서 model에 명확성 부여. (ex. play+ing)
- 신조어가 있는 입력에 대한 성능 향상. (ex. googling)

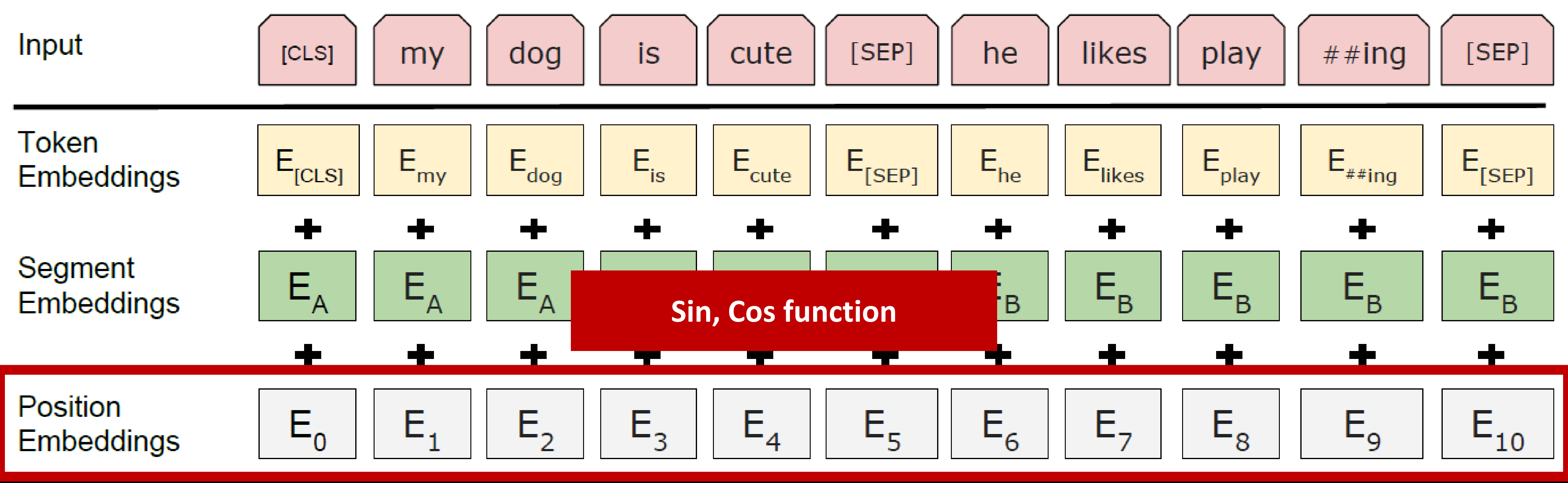
2) Input Representation (Cont'd)



[Segment Embeddings]

- 두 개의 문장이 input일 때, 각 문장에 서로 다른 값을 더해주어 문장 구분.
- E_A와 E_B는 고정된 값, 문장이 하나일 경우엔 E_A만 더해줌.

2) Input Representation (Cont'd)



[Position Embeddings]

- sin과 cos의 출력 값은 입력에 따라 달라짐.
- 무한대 길이의 입력 값도 **상대적인 위치**로 출력이 가능.

3) Pre-training Tasks

BERT = Pre-training + Fine-tuning

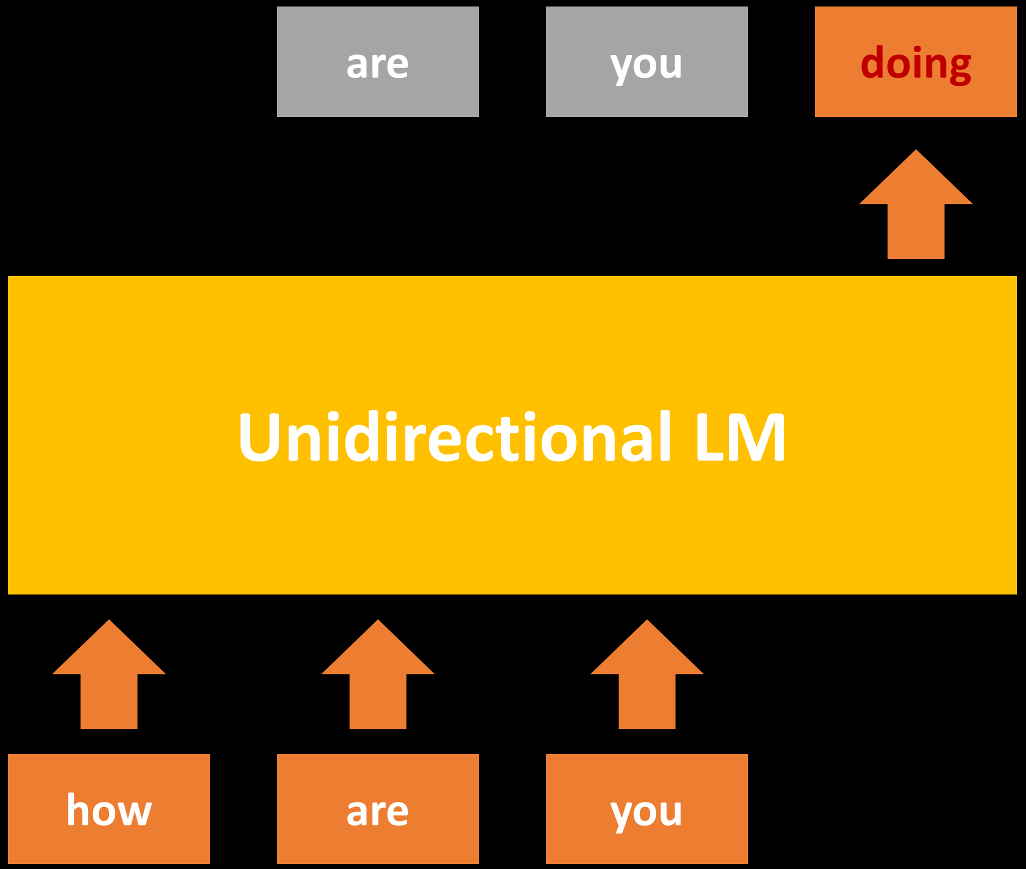
(1) Masked Language Model(MLM)

(2) Next sentence prediction

3) Pre-training Tasks

(1) Masked Language Model(MLM)

Unidirectional(Traditional) LM

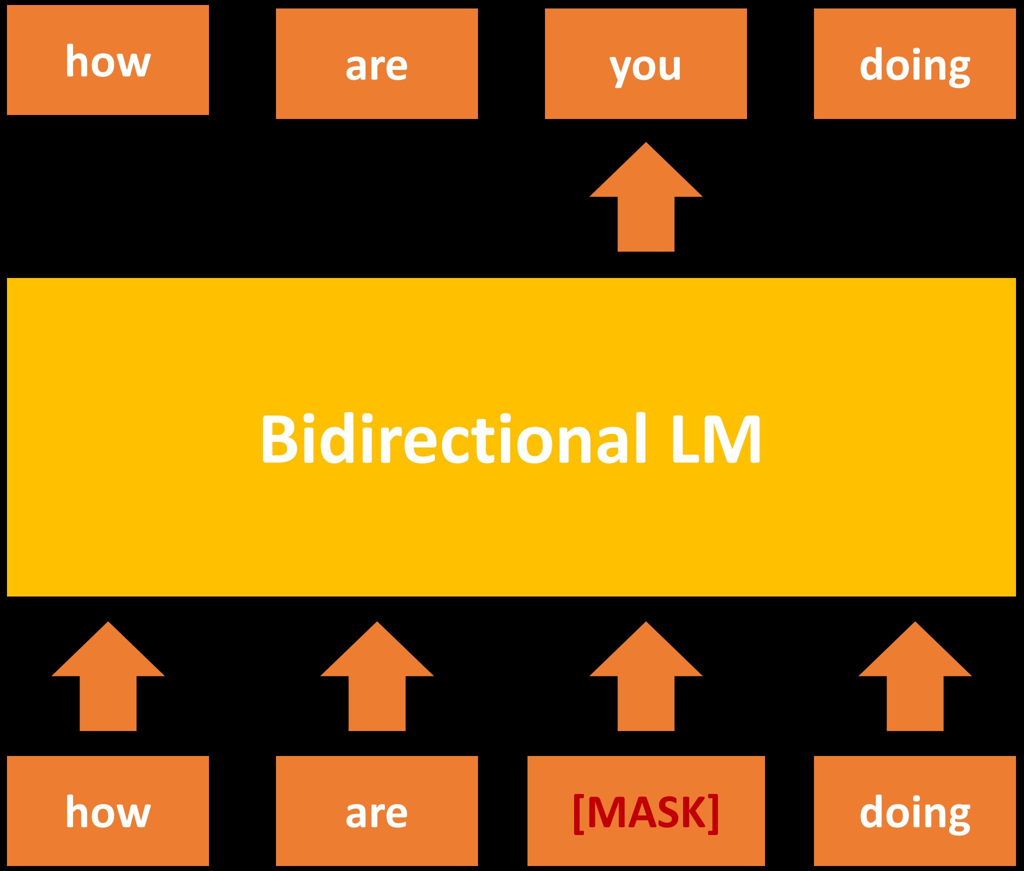


- Traditional한 Language model
- 단방향으로 학습됨. (ex. GPT)

3) Pre-training Tasks

(1) Masked Language Model(MLM) (Cont'd)

Bidirectional(Masked) LM(BERT)



- **BERT**에서 보여주는 MLM의 형태.

- 문장 전체를 학습하되, **[MASK]**로 가려진 단어를 예측하도록 학습됨.

3) Pre-training Tasks

(1) Masked Language Model(MLM) (Cont'd)

- **Generator**가 input 단어 중 random한 **15%**를 **[MASK]** token으로 바꿔줌.

  → **80% : [MASK]** token

  → **10% : Random** token

  → **10% : Unchanged** token

- 이 과정을 통해 BERT는 **context를 파악하는 능력**을 기르게 됨.

3) Pre-training Tasks

(2) Next Sentence prediction (NSP)

- **두 문장** 사이의 관계를 이해하기 위한 task를 위함. (like **QA**)

- **Binarized next sentence prediction task**

Sentence A $\xrightarrow{\text{Is Next?}}$ Sentence B

→ 50% : **Sentence B**가 실제 input의 다음 문장(**IsNext**로 labeled)

→ 나머지 50% : **B**가 **Random sentence**로 치환(**NotNext**로 labeled)

3) Pre-training Tasks

(3) Pre-training Procedure

- Pre-training의 기본적인 절차는 LM에서 수행하는 것과 동일.

- **BERT_English** : BookCorpus(800M words) **+** English Wikipedia(2,500M words)

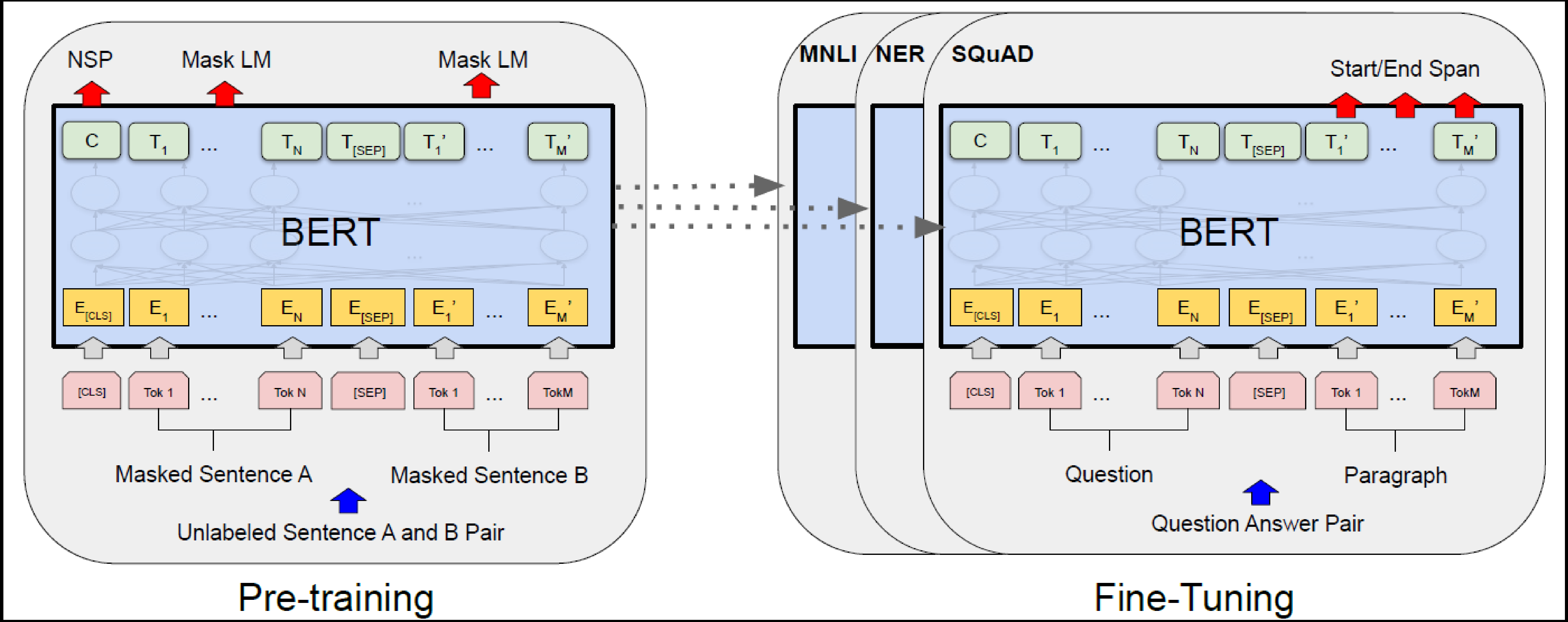
    * Wikipedia data - list, table, header등은 모두 제거. **Only text passage.**

    →**연속적인 시퀀스**만을 추출, 학습시키기 위함.

4) Fine-tuning Tasks

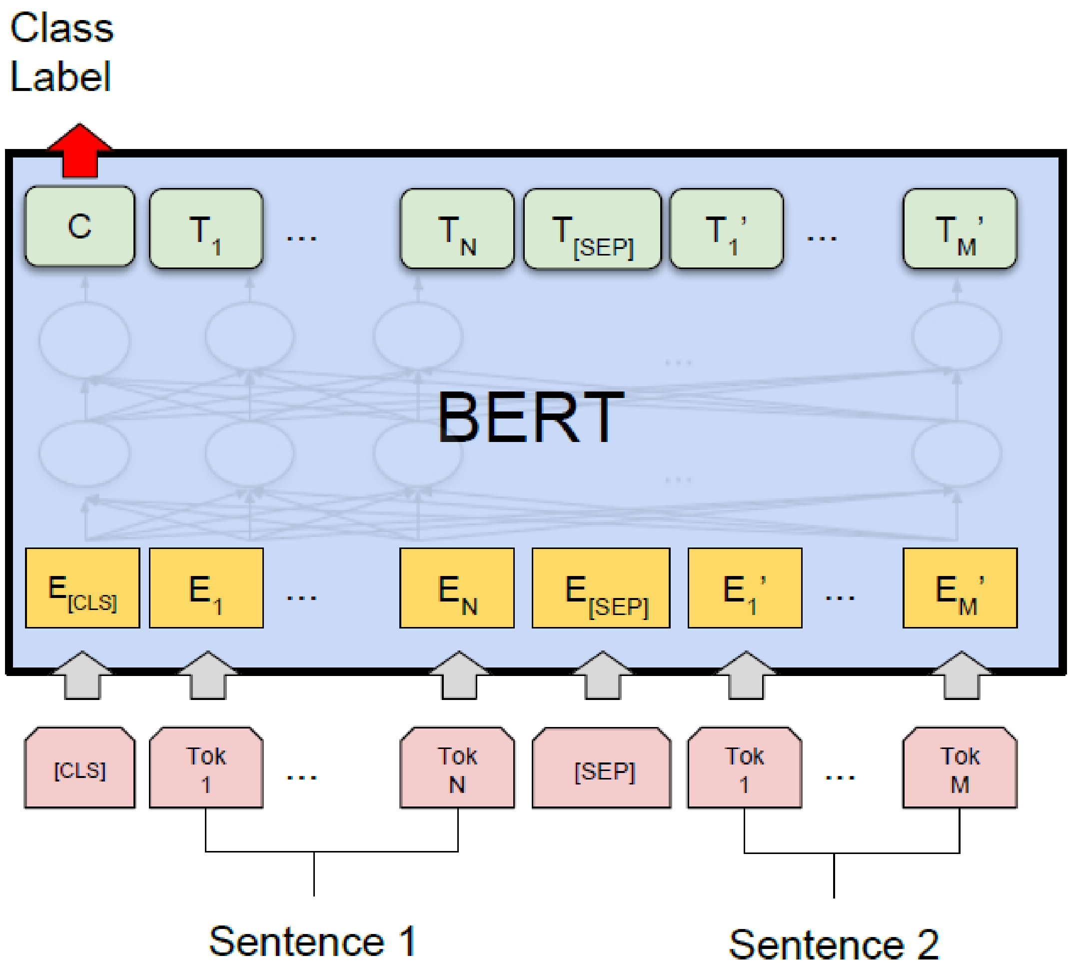
BERT = Pre-training + Fine-tuning

3) Fine-tuning Tasks

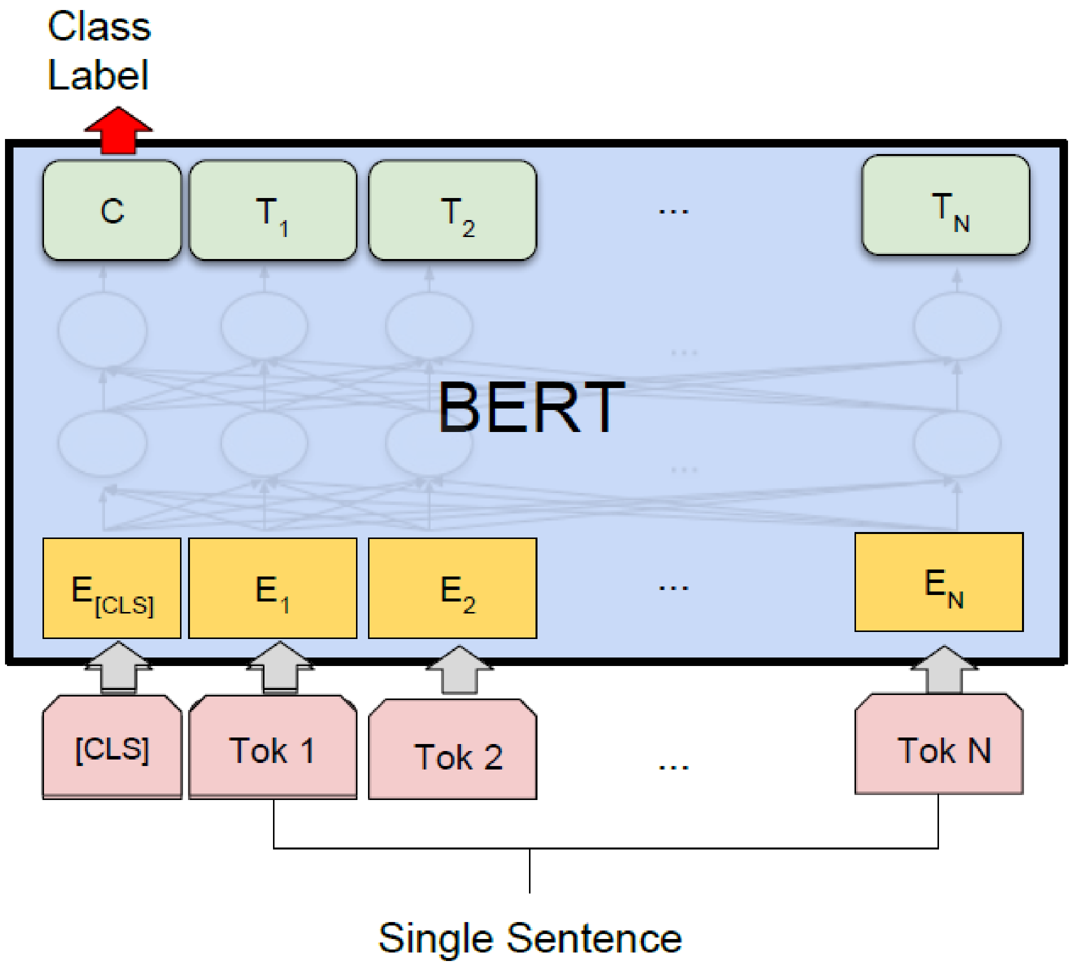


- 동일하게 pre-trained된 model들이 각자 다른 task를 수행하기 위해 모든 parameter를 **fine-tuning** 한다.

3) Fine-tuning Tasks

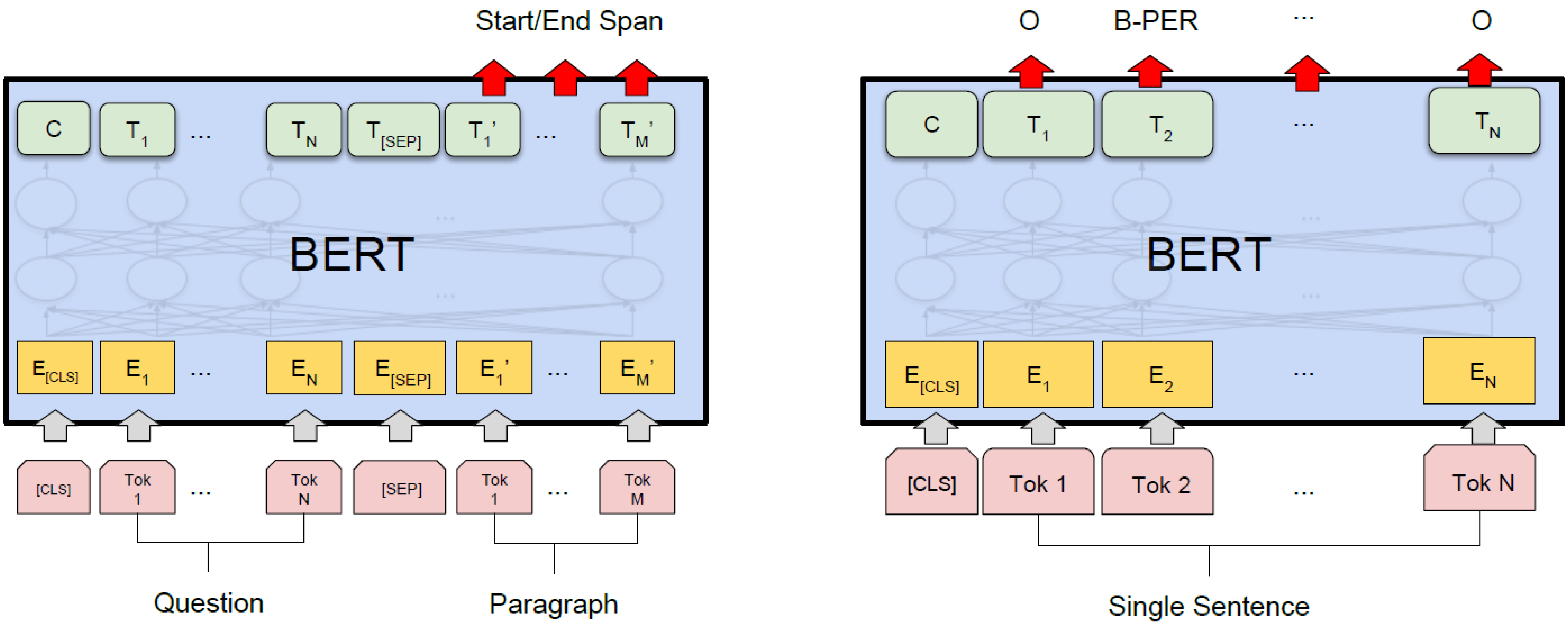


(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA

3) Fine-tuning Tasks



(c) Question Answering Tasks:
SQuAD v1.1

(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

3. Result

1) GLUE Results

— : Single Sequence task

— : Sequence Pair task

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

1) GLUE Results

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

- **GLUE → sequence classification task** (한 문장, 혹은 두 문장을 이용한 task)

- **BERT** model이 모든 task에서 **SOTA**를 달성.

2) SQuAD v1.1 Results

- **SQuAD → Question Answering task**

- **BERT$_{LARGE}$**가 **SOTA**를 달성. (with wide margin)

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT$_{BASE}$ (Single)	80.8	88.5	-	-
BERT$_{LARGE}$ (Single)	84.1	90.9	-	-
BERT$_{LARGE}$ (Ensemble)	85.8	91.8	-	-
BERT$_{LARGE}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
BERT$_{LARGE}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

3) SWAG Results

- SWAG → grounded common-sense inference task

- $\text{BERT}_{\text{LARGE}}$가 SOTA를 달성. (with wide margin)

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

4) Ablation Studies

(1) Effect of Pre-training Tasks

Tasks	Dev Set MNLI-m (Acc)	 QNLI (Acc)	 MRPC (Acc)	 SST-2 (Acc)	 SQuAD (F1)
$BERT_{BASE}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

- 논문에서 중요하다고 언급한 요소들을 제거해보며 중요성을 파악.

- **NSP**(Next Sentence Prediction)만 제거한 것, **MLM**도 제거한 것들과 결과 비교.

4) Ablation Studies

(1) Effect of Pre-training Tasks (Cont'd)

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$BERT_{BASE}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

- **NSP**를 빼면, 두 문장 간의 구조 파악을 해야 하는 **NLI** 쪽에서 성능 크게 감소.

- **MLM** 대신 **LTR**을 쓰면 성능 하락은 더욱 심각해짐.

4) Ablation Studies

(2) Effect of Model Size

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

- **모델이 커질수록**(hyperparameter 수가 늘어날수록), **정확도가 증가한다.**

QnA