MASS : Masked Sequence to Sequence Pre-training for Language Generation

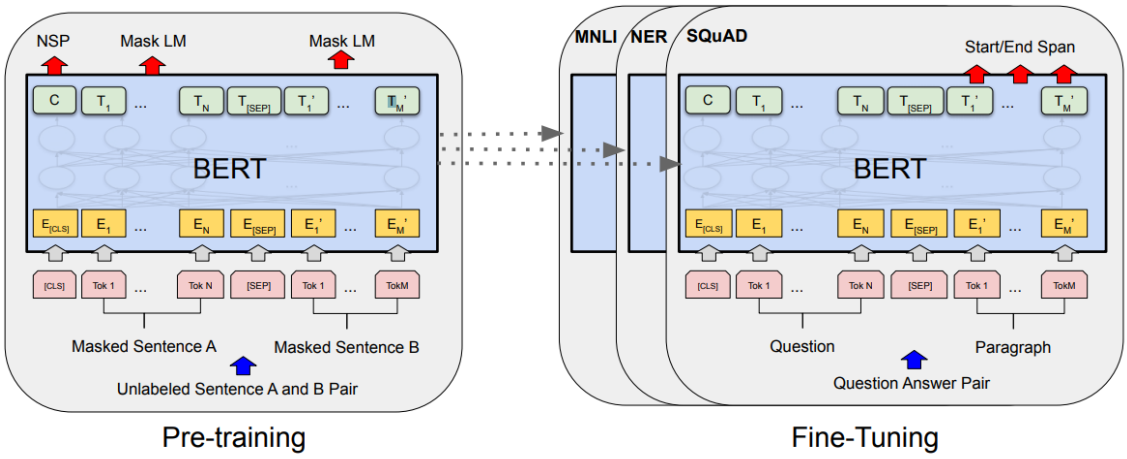
Motivation

natural language generation에 적합한 pre-training 방식을 만들자!

Introduction

▼ NLP의 최근 트렌드 : transfer learning

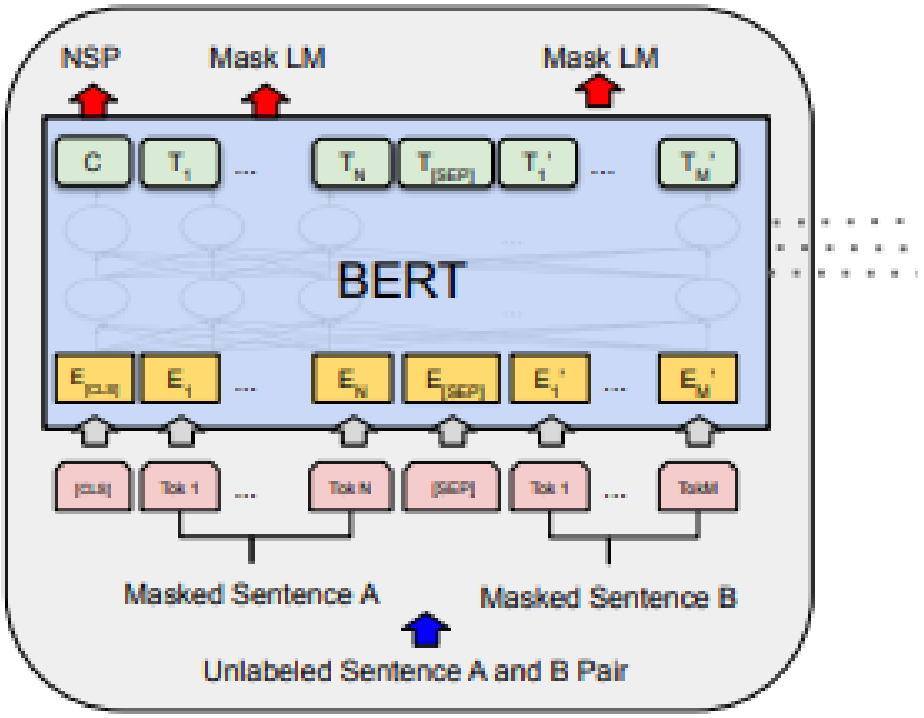
- Network를 large corpus를 통해 학습한 언어 지식을 downstream task에 transfer 하는 방법은 NLP research area에서 필수로 자리 잡음



- Pre-trained 된 모델은 downstream task에 비교적 적은 데이터로도 효과적으로 fine-tuning 가능
- Transfer learning의 등장 이후 많은 NLP task들은 human level을 능가하는 성능을 보이기 시작함

▼ Language generation 에서도 이러한 transfer learning을 효과적으로 적용할 수 있을까?

- 상식적으로 생각해 보면 안 될 이유가 없다. 그렇다면 기존의 pre-trained 된 모델을 바로 사용하면 손쉽게 해결되지 않을까?
- 그렇지 않다. 기존의 pre-training 기법은 language generation model architecture과 차이가 있기 때문
- 기존의 pre-training 기법은 encoder only 모델. 즉, input sentence를 embedding하고, 그 결과를 바탕으로 바로 classification하는 방식으로 모델이 학습된다.

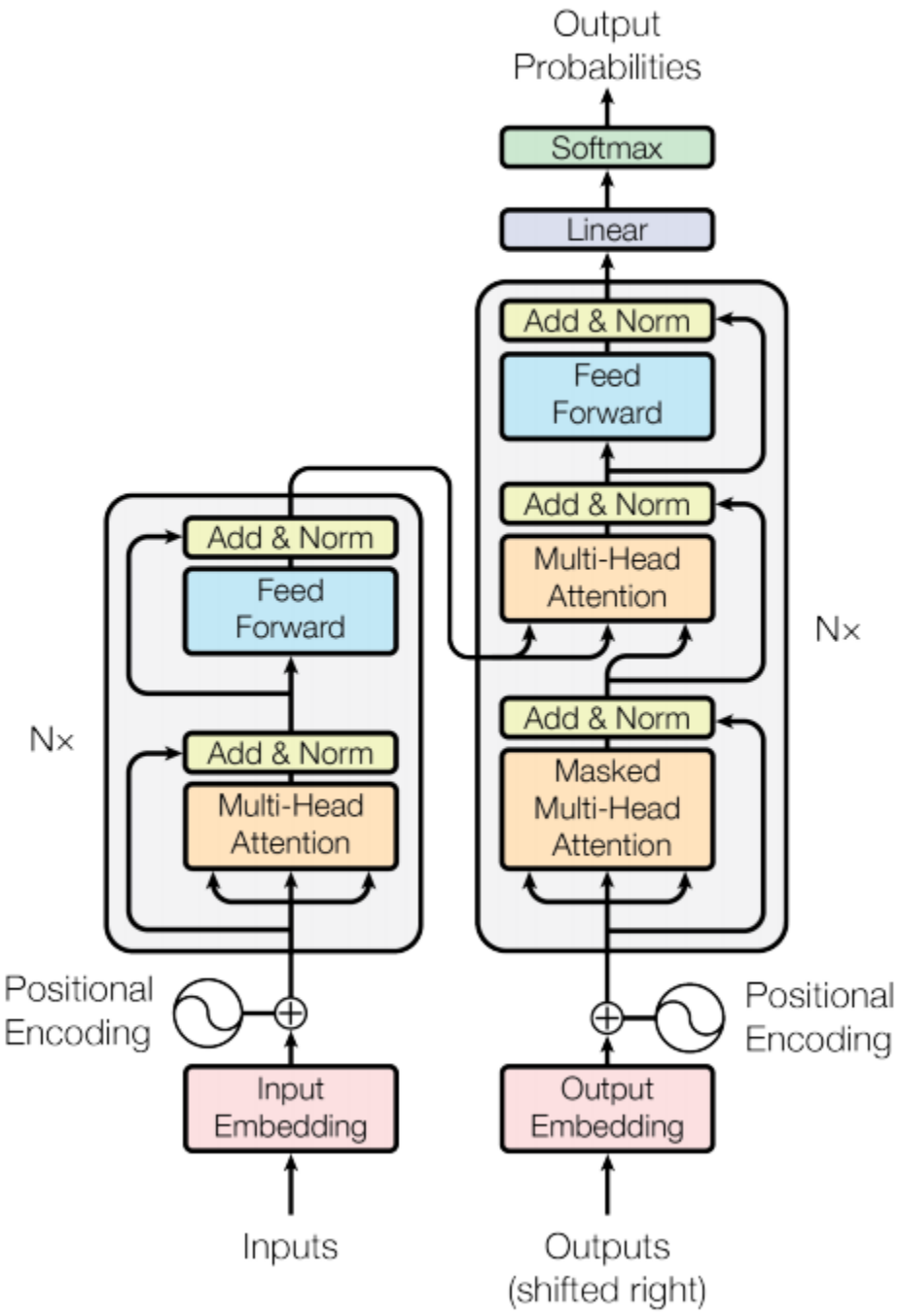


Pre-training

- 따라서, 이러한 기법은 sentence embedding 기반의 classification에서 최적의 성능 향상을 기대할 수 있다. 실제로 이들 논문이 성능을 평가한 tasks들은 전부 classification tasks들임

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
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**

- 그런데 language generation (사실 보다 정확하게 표현하면 seq2seq 저자들이 말하는 language generation은 sequence to sequence를 의미하는듯 image captioning, language model은 해당사항 없음)은? Encoder와 decoder 모두가 존재한다. 즉, source sentence와 target sentence를 모두 modeling할 수 있어야 한다.



- 기존의 pre-training 기법을 이러한 encoder-decoder 구조를 가진 모델에 적용하려면?

1. pre-trained 된 encoder를 seq2seq model의 encoder와 decoder에 동시에 적용하는 방법 → 딱 봐도 이상함

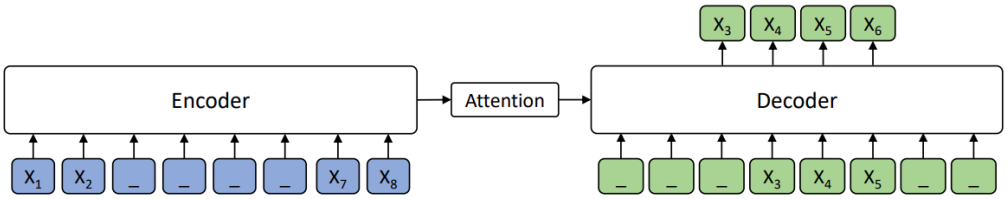
2. encoder를 따로 pre-train 시키고, decoder를 따로 pre-train 시켜서 각각 적용 → encoder와 decoder 사이의 connection이 고려가 안 된 상태로 pre-train됨

- 따라서 natural language generation을 위해서는 encoder-decoder의 형태를 가진 아키텍쳐에 프리트레인 된 모델이 필요
- 이 논문의 contribution : encoder와 decoder를 가진 구조를 large-corpus에 효과적으로 pre-train 하는 방법을 제안

Methodology

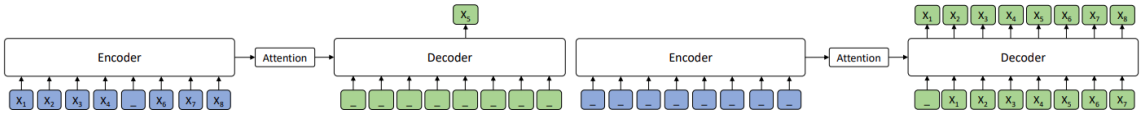
▼ Masked sequence to sequence pre-training (MASS)

- 문장의 일부(span)를 masking하고 그 masking된 부분을 예측



- e.x. 나는 배가 고프다. 그래서 밥을 먹는다. 그러면 배가 부르다.
  1. 나는 배가 고프다. ___ __ ___. 그러면 배가 부르다. (마스킹)
  2. (예측)
     1. encoder → 나는 배가 고프다. ___ __ ___. 그러면 배가 부르다.
        decoder → 그래서
     2. encoder → 나는 배가 고프다. ___ __ ___. 그러면 배가 부르다.
        decoder → 그래서 밥을
     3. encoder → 나는 배가 고프다. ___ __ ___. 그러면 배가 부르다.
        decoder → 그래서 밥을 먹는다
- span bert와 상당히 유사. 그러나 span bert는 masked된 단어를 encoder에서 바로 classification하고 mass는 decoder가 예측. 따라서 mass는 masked된 단어끼리의 dependency를 capture할 수 있다는 점에서 차이가 있다.

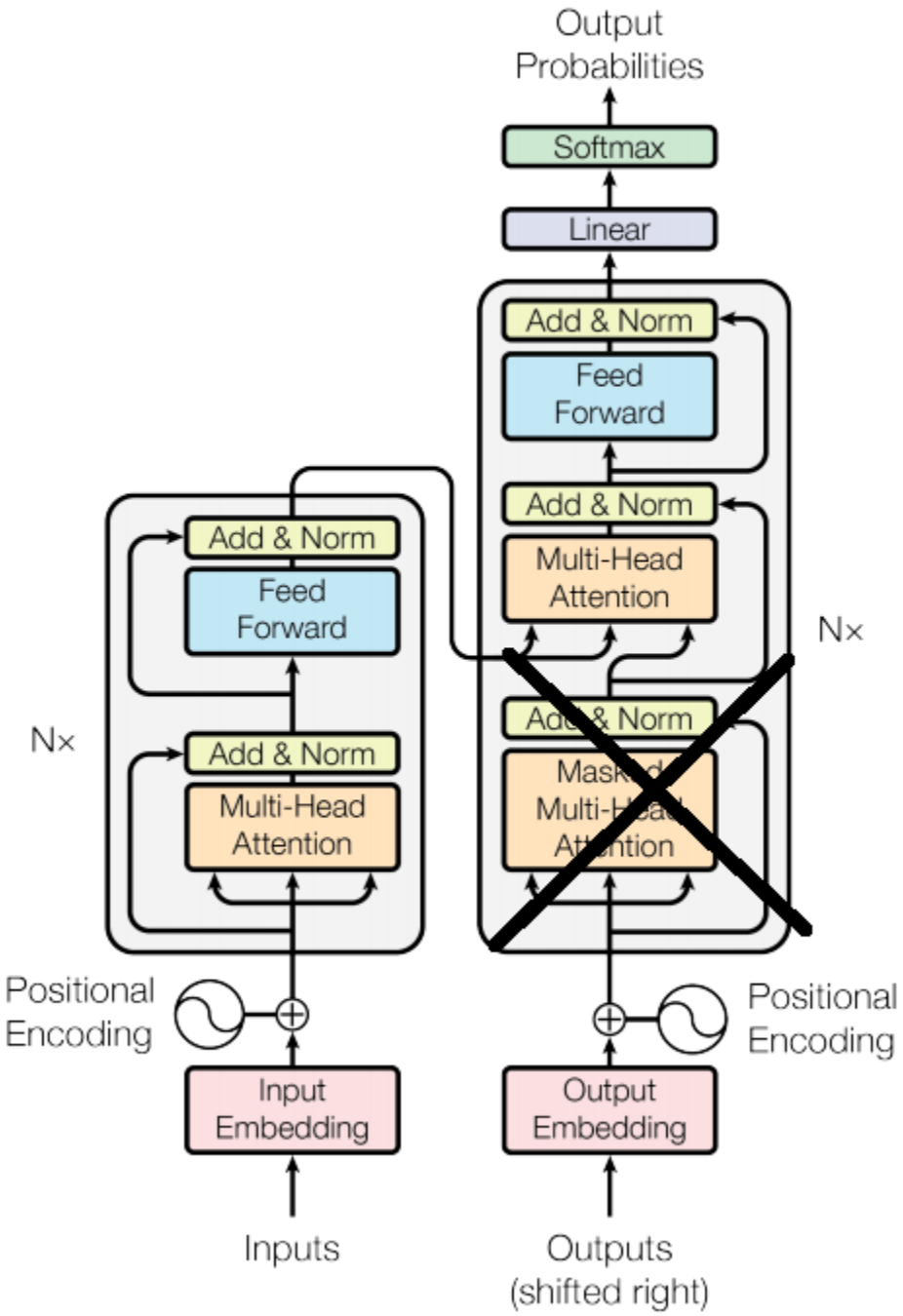
▼ MASS는 standard language model 및 MLM (bert로 대표되는)의 generalized version이다



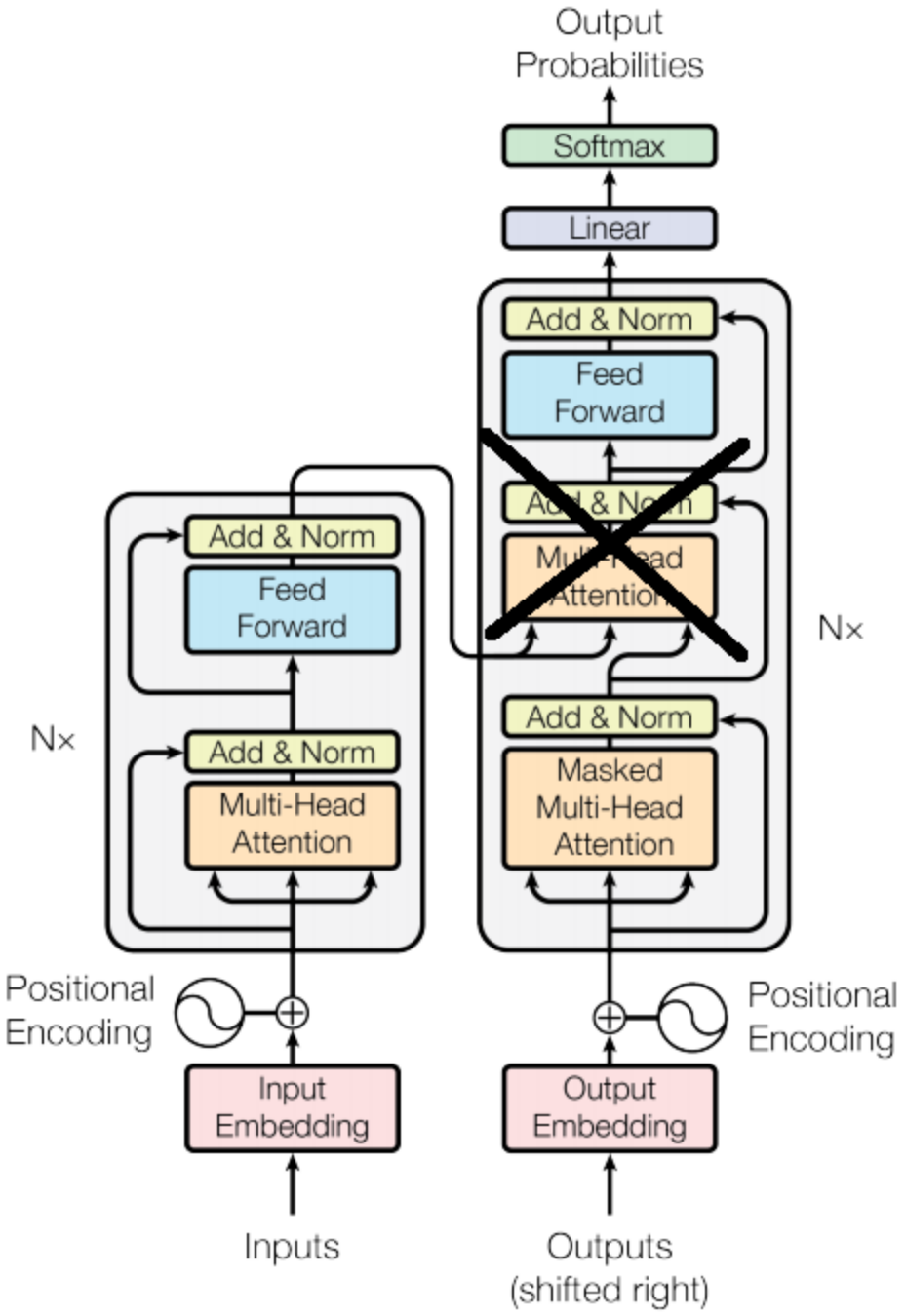
(a) Masked language modeling in BERT ($k = 1$) (b) Standard language modeling ($k = m$)

- span을 어떻게 잡는지에 따라 BERT가 될 수도 있고, 일반적인 language model이 될 수도 있다.

  span을 1로 잡을 경우 BERT가 됨. decoder 제일 아래에 있는 masked-multihead attention이 사실상 없는 것과 마찬가지가 되기 때문에 encoder만 있는 BERT와 모델이 사실상 같아짐



span을 문장 전체로 잡을 경우 standard language model이 됨. encoder에서 encoding된 게 없기 때문에 decoder의 중간 부분이 사실상 없는 것과 마찬가지 → standard한 language model이 된다.



- 따라서 MASS는 BERT와 GPT을 담는 더 큰 집합인 격이다 라는 저자들의 주장

▼ 저자들이 주장하는 MASS가 좋은 이유

- encoder와 decoder의 디펜던시를 고려할 수 있기 때문. encoder는 encoder나름대로 mask되지 않은 token들의 semantic을 효과적으로 modeling할 수 있다. 또

한, decoder는 빈칸 채우기 문제를 푸는 과정에서 중요한 단어에 대한 정보를 학습할 수 있다.

- spanbert처럼 masked 된 token들을 language modeling 방식으로 예측하기 때문에 masked token들 사이의 디펜던시를 반영할 수 있다. (spanbert는 못함)

▼ 내가 생각하는 MASS가 좋은 이유

- bidirectional한 정보를 고려하면서 모델을 학습시킬 수 있음. XLNET은 너무 무거운데 비해 MASS는 상대적으로 가벼움.

Evaluation

▼ 데이터셋

- seq2seq tasks
- machine translation, summarization, conversational response generation

▼ 학습 디테일

- 특별한한 내용은 없다. machine translation에도 적용하기 위해 여러가지 언어로 이루어진 data들을 한번에 학습시켰다는 내용 정도만 알아두면 될듯. vocabulary는 BPE로 획득. 알파벳 위주의 언어들만 묶어서 학습시켰기 때문에 가능한 세팅이었던 것 같다.

▼ 실험 결과

- 실험이 아주 깔끔하다. baseline 구성이 거의 완벽. BERT를 비롯한 publication된 모델들과의 비교는 물론 자신들 나름대로 encoder와 decoder를 초기화시킨 모델들과도 성능 평가를 진행함. encoder는 bert로 학습시키고 decoder는 language model로 학습시킨 모델과 비교 등.

Method	Setting	en - fr	fr - en	en - de	de - en	en - ro	ro - en
Artetxe et al. (2017)	2-layer RNN	15.13	15.56	6.89	10.16	-	-
Lample et al. (2017)	3-layer RNN	15.05	14.31	9.75	13.33	-	-
Yang et al. (2018)	4-layer Transformer	16.97	15.58	10.86	14.62	-	-
Lample et al. (2018)	4-layer Transformer	25.14	24.18	17.16	21.00	21.18	19.44
XLM (Lample & Conneau, 2019)	6-layer Transformer	33.40	33.30	27.00	34.30	33.30	31.80
MASS	6-layer Transformer	**37.50**	**34.90**	**28.30**	**35.20**	**35.20**	**33.10**

Method	RG-1 (F)	RG-2 (F)	RG-L (F)
BERT+LM	37.75	18.45	34.85
DAE	35.97	17.17	33.14
MASS	**38.73**	**19.71**	**35.96**

Method	Data = 10K	Data = 110K
Baseline	82.39	26.38
BERT+LM	80.11	24.84
MASS	**74.32**	**23.52**

Conclusion

- BERT로 대표되는 기존 pre-training 기법은 NLU문제를 푸는 데에 특화되어 있다. MASS는 NLG에도 효과적으로 적용될 수 있었도록 encoder-decoder 형태의 네트워크를 pre-training 하는 기법을 제시하였다