BART: Denoising sequence-to-sequence pre-training for natural language generation

23.05.26

이정민

BART

Paper

❖ **BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension**

- [2019, arXiv] 23/05/22 기준 5125회 인용

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer
Facebook AI
{mikelewis,yinhanliu,naman}@fb.com

Abstract

We present BART, a denoising autoencoder for pretraining sequence-to-sequence models. BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and many other more recent pretraining schemes. We evaluate a number of noising approaches, finding the best performance by both randomly shuffling the order of the original sentences and using a novel in-filling scheme, where spans of text are replaced with a single mask token. BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks. It matches the performance of RoBERTa with comparable training resources on GLUE and SQuAD, achieves new state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks, with gains of up to 6 ROUGE. BART also provides a 1.1 BLEU increase over a back-translation system for machine translation, with only target language pretraining. We also report ablation experiments that replicate other pretraining schemes within the BART framework, to better measure which factors most influence end-task performance.

masked tokens are predicted (Yang et al., 2019), and the available context for replacing masked tokens (Dong et al., 2019). However, these methods typically focus on particular types of end tasks (e.g. span prediction, generation, etc.), limiting their applicability.

In this paper, we present BART, which pre-trains a model combining Bidirectional and Auto-Regressive Transformers. BART is a denoising autoencoder built with a sequence-to-sequence model that is applicable to a very wide range of end tasks. Pretraining has two stages (1) text is corrupted with an arbitrary noising function, and (2) a sequence-to-sequence model is learned to reconstruct the original text. BART uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and many other more recent pretraining schemes (see Figure 1).

A key advantage of this setup is the noising flexibility; arbitrary transformations can be applied to the original text, including changing its length. We evaluate a number of noising approaches, finding the best performance by both randomly shuffling the order of the original sentences and using a novel in-filling scheme, where arbitrary length spans of text (including zero length) are replaced with a single mask token. This approach generalizes the original word masking and next sentence prediction objectives in BERT by forcing the model to reason more about overall sentence length and make longer range transformations to the input.

BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks. It matches the performance of RoBERTa

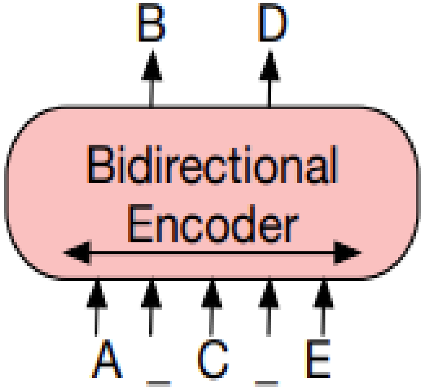
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

BART

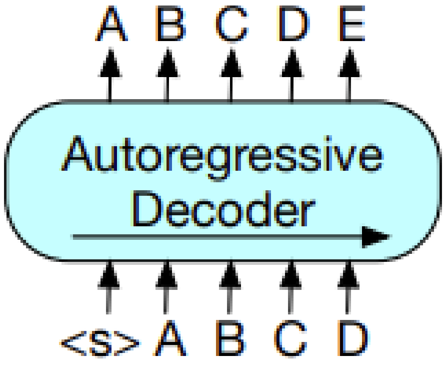
Denoising sequence-to-sequence pre-training for natural language generation

❖ 등장 배경

- Maksed token이 무엇인지와 그것의 위치를 예측하는 연구들이 활발하게 진행
- 해당 연구들은 특정 type과 task에만 국한됨



BERT



GPT

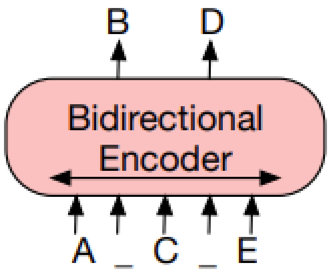
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

BART

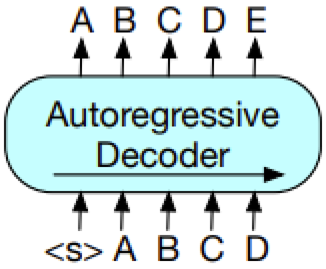
Denoising sequence-to-sequence pre-training for natural language generation

❖ BERT vs GPT vs BART

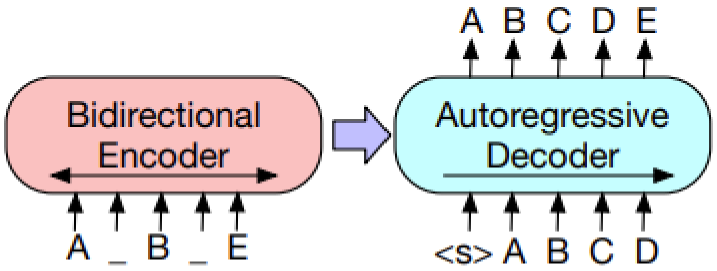
- BERT : Masked token을 활용한 **Auto Encoding** 방식(Transformer Encoder 사용)
- GPT : **Autoregressive** 방식(Transformer Decoder 사용)
- BART: **Bidirectional Encoder + Autoregressive Decoder**



BERT



GPT



BART

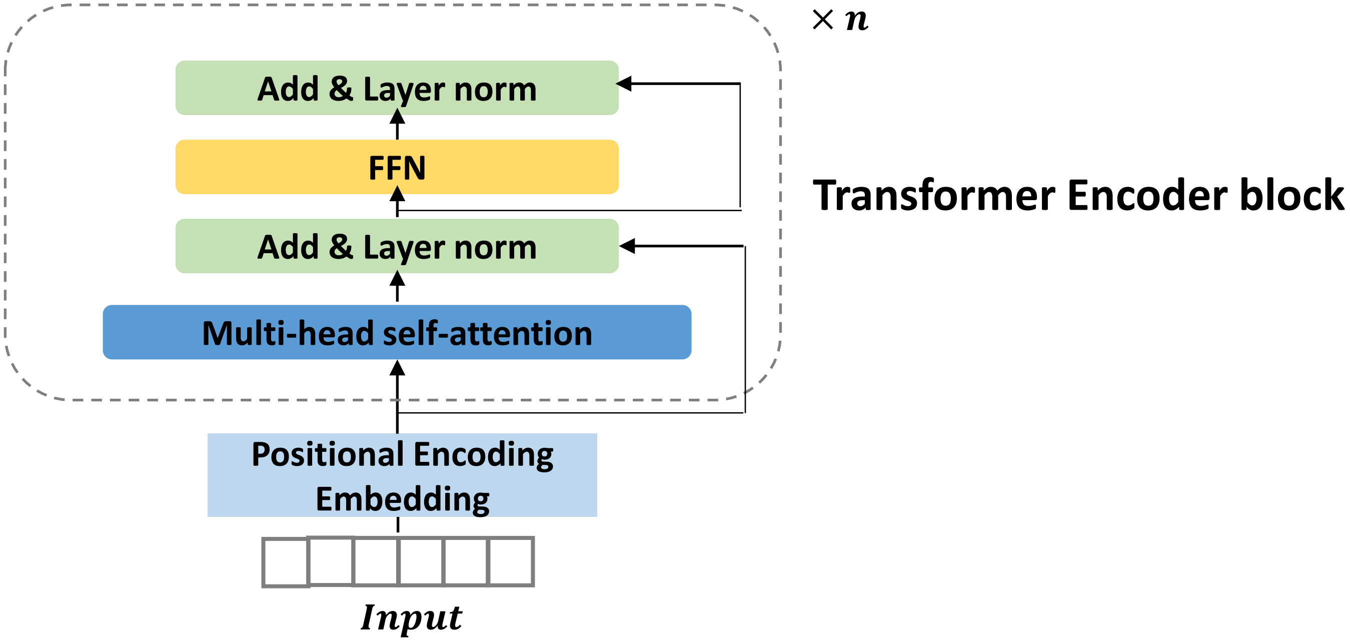
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

BART

Denoising sequence-to-sequence pre-training for natural language generation

❖ BERT

- Transformer Encoder를 활용

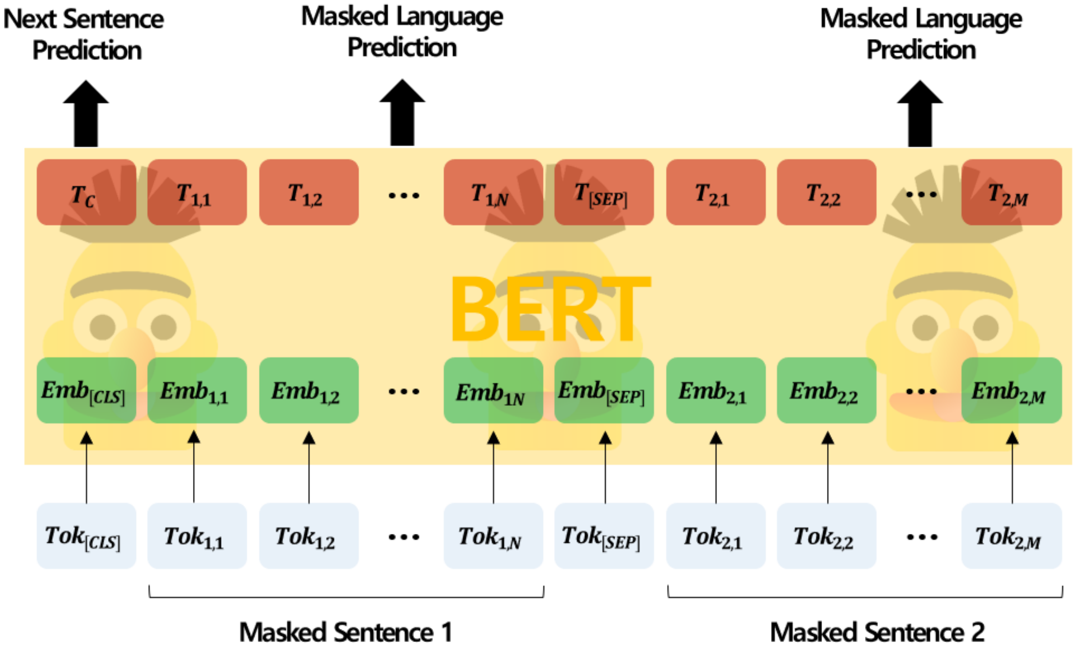


BART

Denoising sequence-to-sequence pre-training for natural language generation

❖ BERT

- Transformer Encoder를 활용
- Masked Language Prediction 및 Next Sentence Prediction을 수행하는 구조



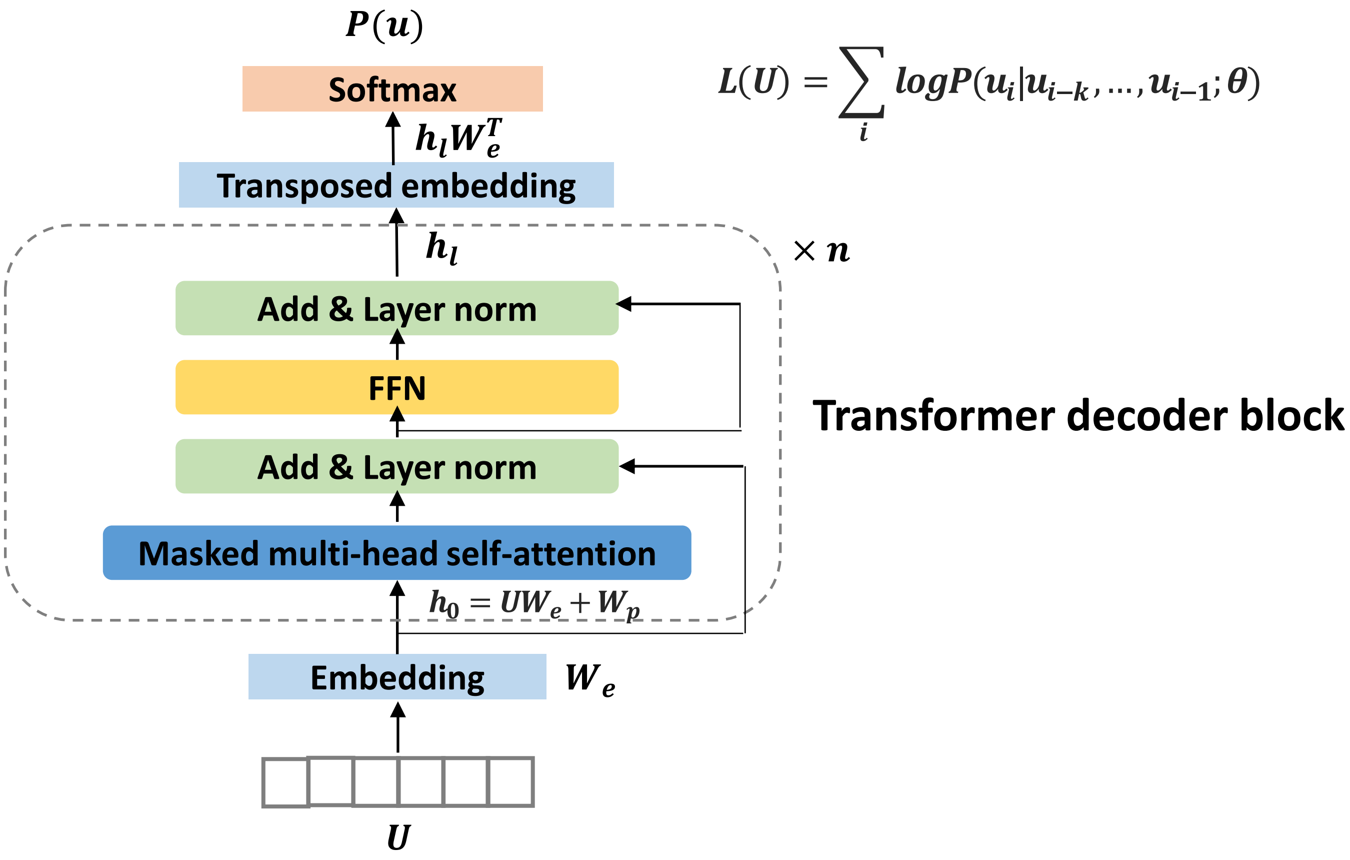
Next Sentence Prediction
Masked Language Prediction
Masked Language Prediction
T_C $T_{1,1}$ $T_{1,2}$ ... $T_{1,N}$ $T_{[SEP]}$ $T_{2,1}$ $T_{2,2}$ ... $T_{2,M}$
BERT
$Emb_{[CLS]}$ $Emb_{1,1}$ $Emb_{1,2}$ ... $Emb_{1,N}$ $Emb_{[SEP]}$ $Emb_{2,1}$ $Emb_{2,2}$ ... $Emb_{2,M}$
$Tok_{[CLS]}$ $Tok_{1,1}$ $Tok_{1,2}$ ... $Tok_{1,N}$ $Tok_{[SEP]}$ $Tok_{2,1}$ $Tok_{2,2}$ ... $Tok_{2,M}$
Masked Sentence 1
Masked Sentence 2

BART

Denoising sequence-to-sequence pre-training for natural language generation

❖ GPT

- Transformer Decoder를 활용



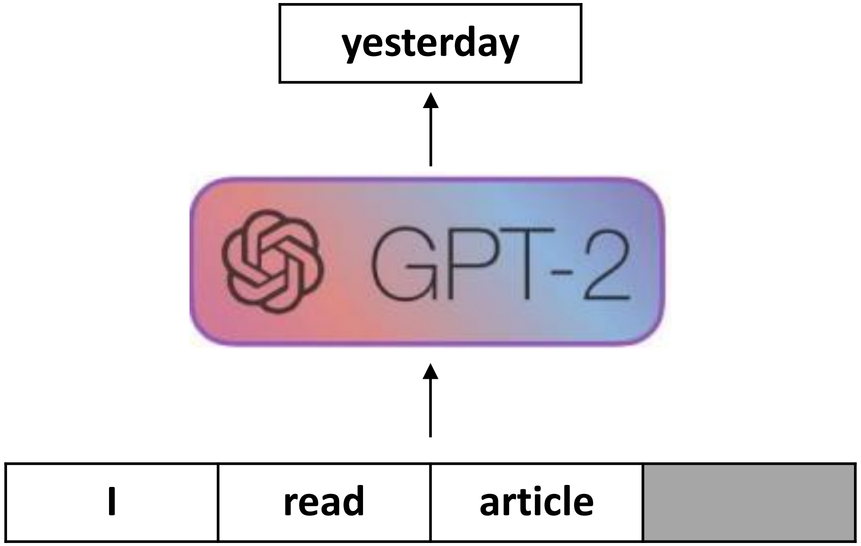
$$L(U) = \sum_i logP(u_i|u_{i-k}, \dots, u_{i-1}; \theta)$$

BART

Denoising sequence-to-sequence pre-training for natural language generation

❖ GPT

- Transformer Decoder를 활용
- 다음 순서의 token을 예측하도록 사전 학습 수행(autoregressive)



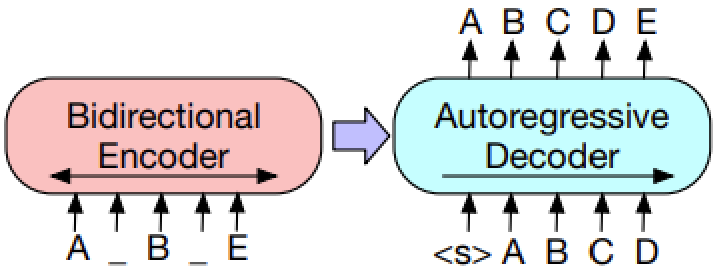
$$L(U) = \sum_{i} logP(u_i|u_{i-k}, \dots, u_{i-1}; \theta)$$

BART

Denoising sequence-to-sequence pre-training for natural language generation

❖ Architecture

- Base model: 6 layers in encoder and decoder
- Large model: 12 layers in encoder and decoder
- BERT와의 차이점
  - ➢ Decoder에서 Encoder의 마지막 layer와 cross-attention 수행
  - ➢ Feed-Forward Network 사용하지않음



BART

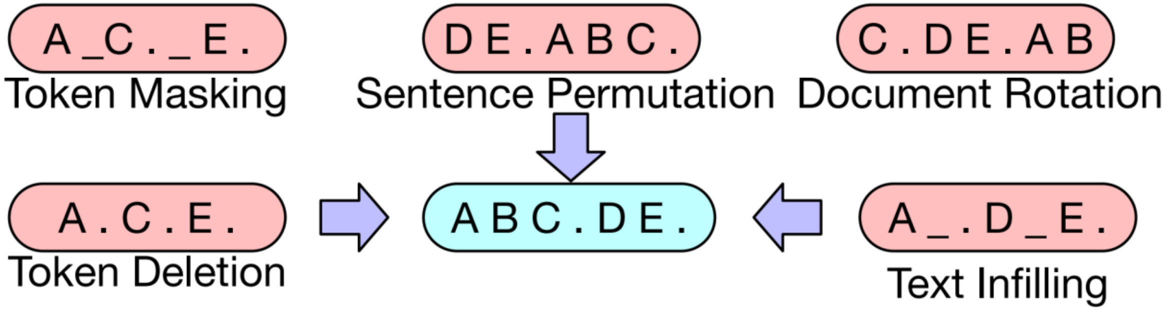
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

BART

Denoising sequence-to-sequence pre-training for natural language generation

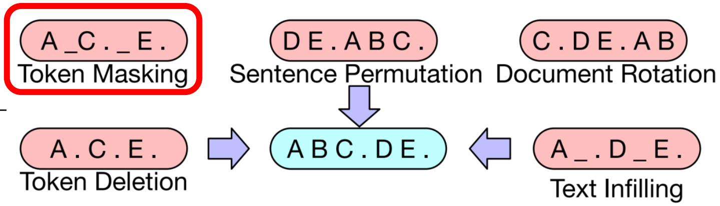
❖ Token Corrupt

- Token Masking
- Token Deletion
- Text Infilling
- Sentence Permutation
- Document Rotation



Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

BART



A_C._E.
Token Masking

DE.ABC.
Sentence Permutation

C.DE.AB
Document Rotation

A.C.E.
Token Deletion

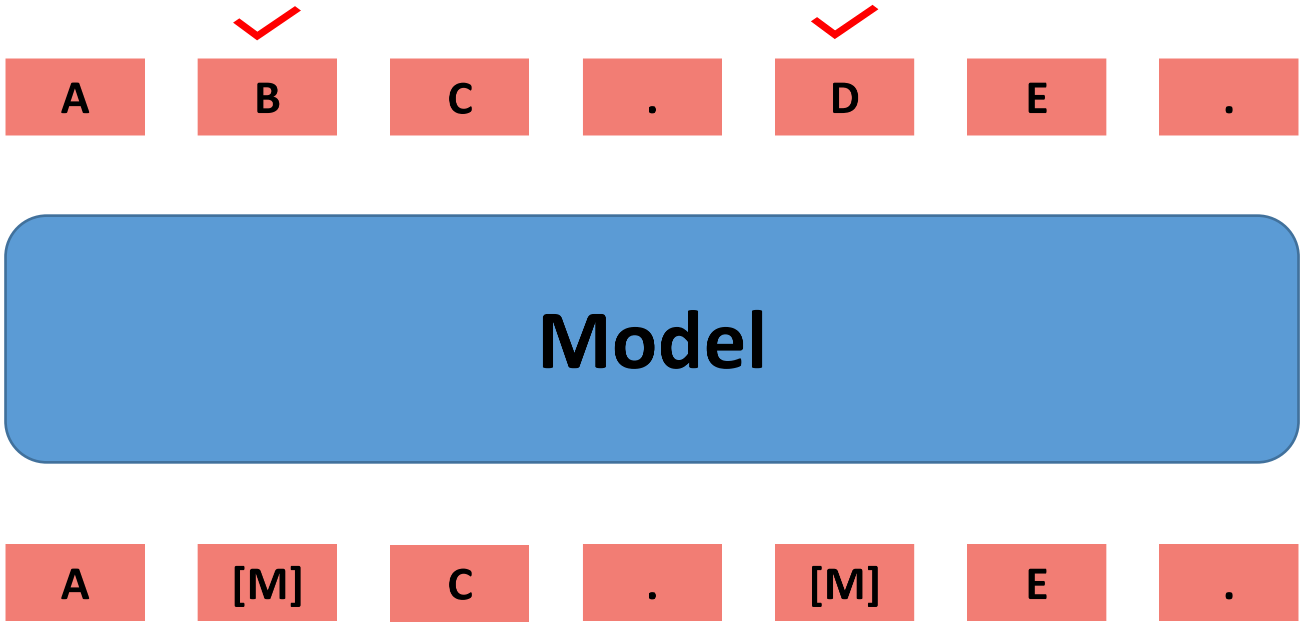
ABC.DE.

A_.D_E.
Text Infilling

Token Corrupt

❖ Token Masking

- BERT의 방식과 동일
- Masked token 예측



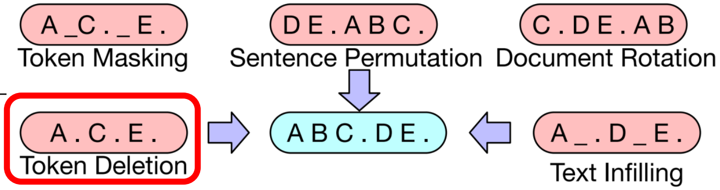
A B C . D E .

Model

A [M] C . [M] E .

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

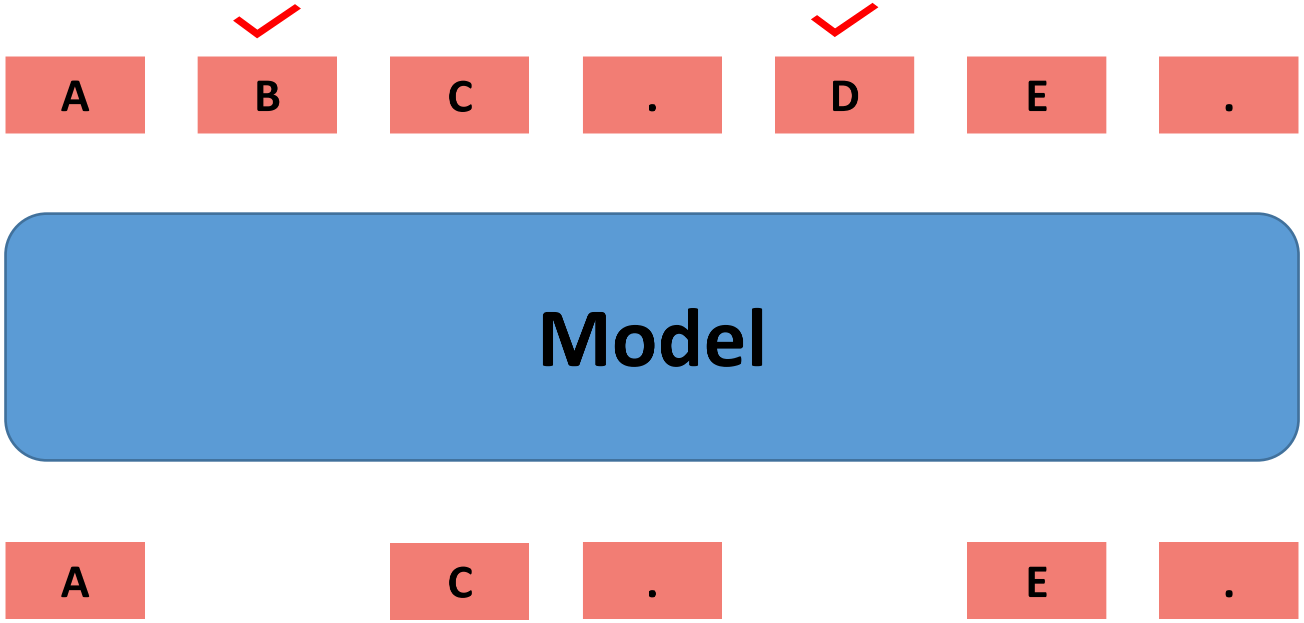
BART



Token Corrupt

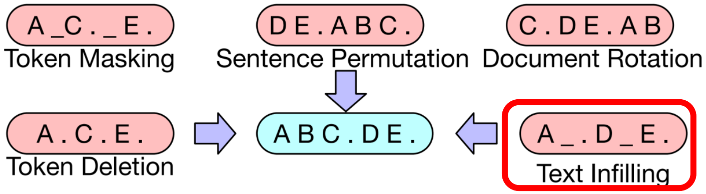
❖ Token Deletion

- 임의의 token 삭제
- 삭제한 token의 위치와 삭제된 token이 어떤 token이였는지를 예측



Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

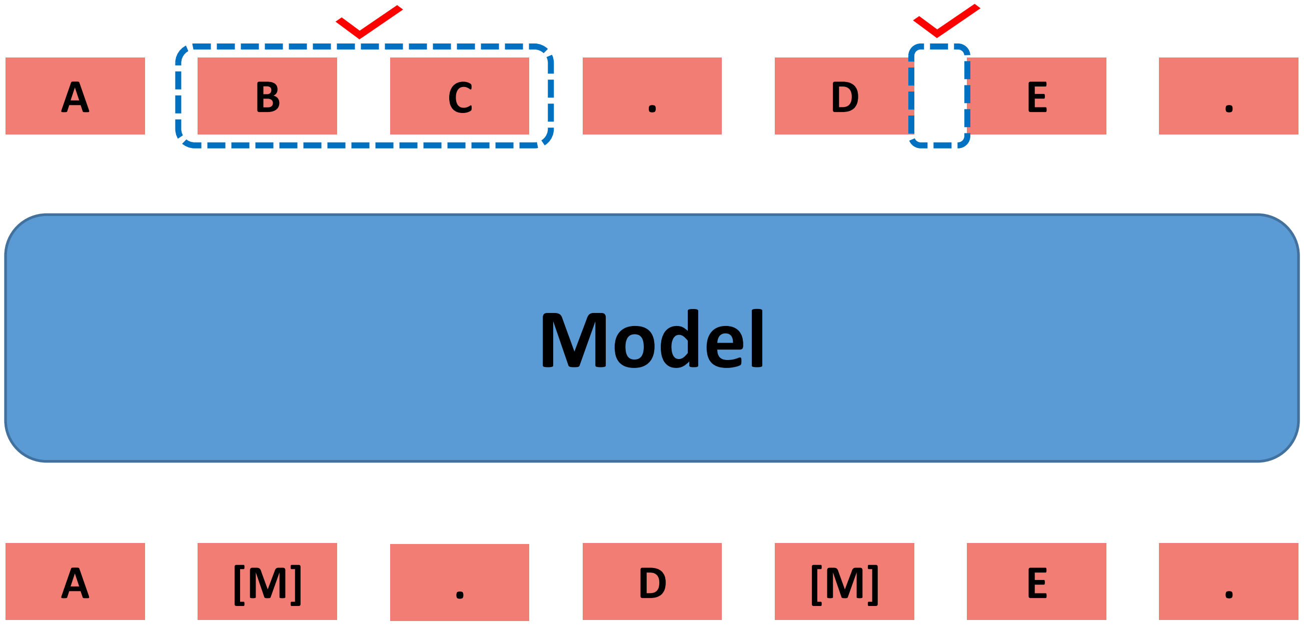
BART



Token Corrupt

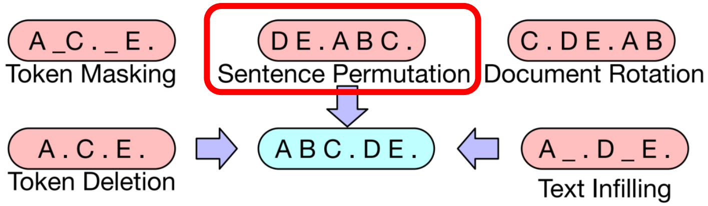
❖ Text Infilling

- 포아송 분포($\lambda = 3$)로 부터 span length sampling 후 해당 길이 만큼 masked token으로 대체
- Span length=0인 경우 해당 위치에 masked token 추가
- Masked token 위치에 어떤 token이, 몇 개의 token이 있었는지 예측



Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

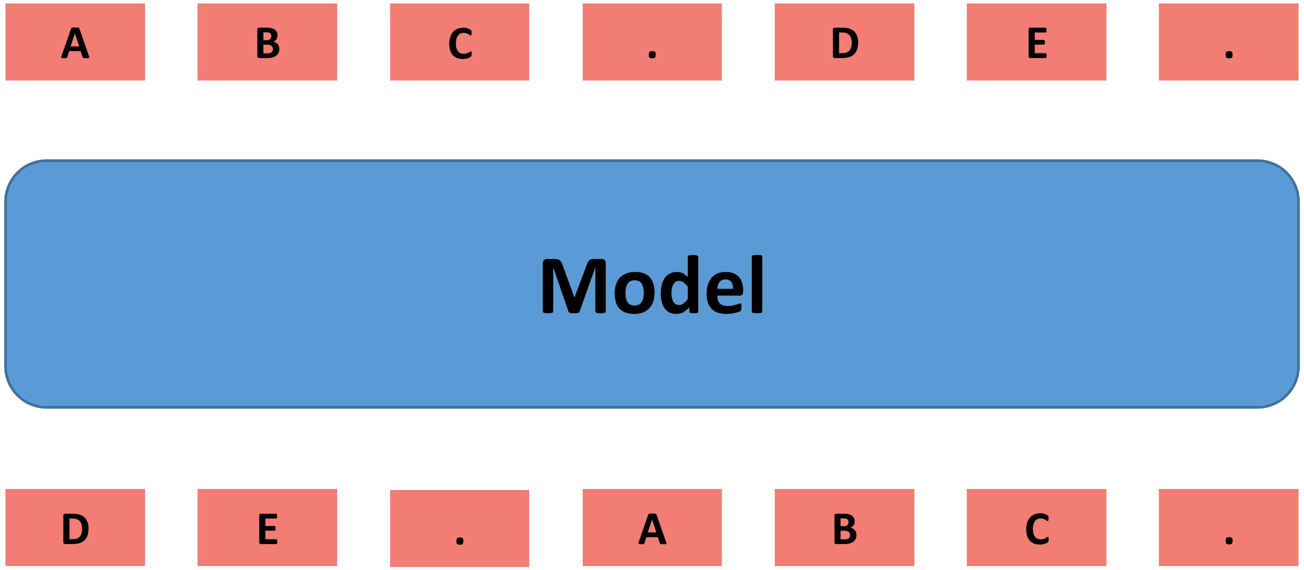
BART



Token Corrupt

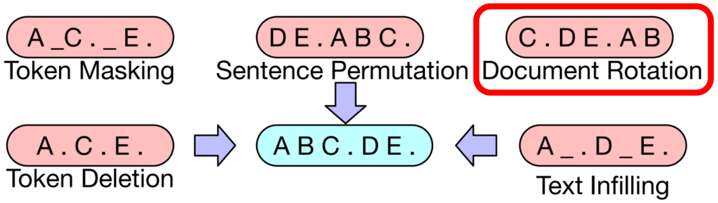
❖ Sentence Permutation

- 문장의 순서를 임의로 섞은 후 원래 문장 순서 예측



Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

BART



A _C . _E .
Token Masking

D E . A B C .
Sentence Permutation

C . D E . A B
Document Rotation

A . C . E .
Token Deletion

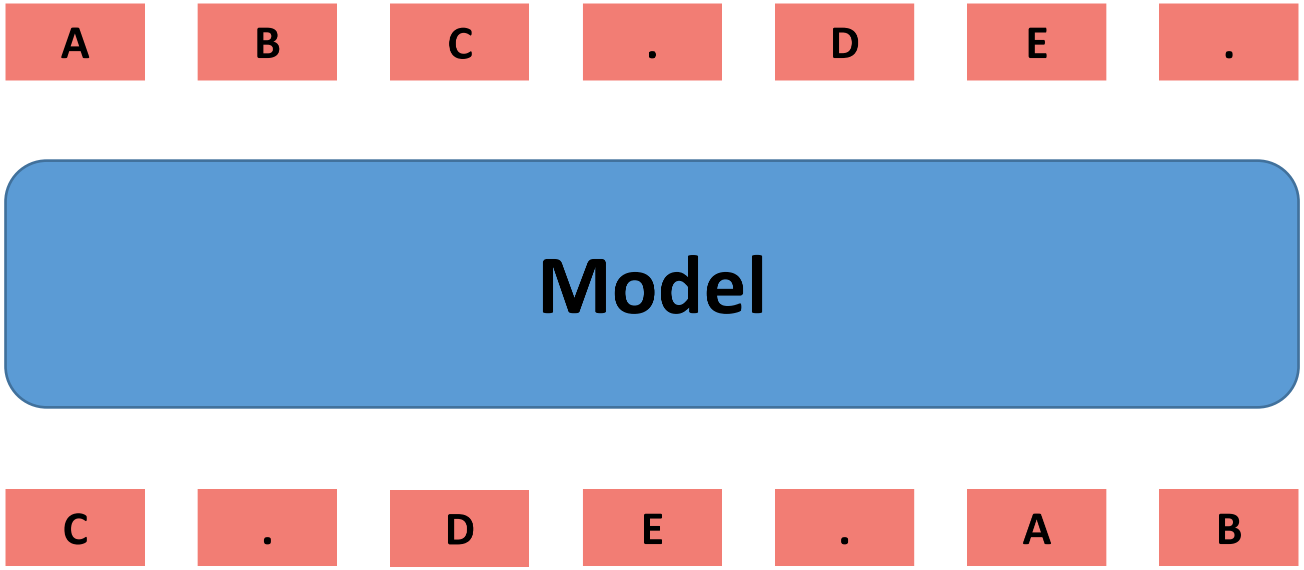
A B C . D E .

A _ . D _E .
Text Infilling

Token Corrupt

❖ Document Rotation

- Uniform 분포로부터 임의로 한 token 선택
- 선택된 token이 시작점이 되도록 rotation 수행
- 문서의 시작점을 예측할 수 있도록 학습



A B C . D E .

Model

C . D E . A B

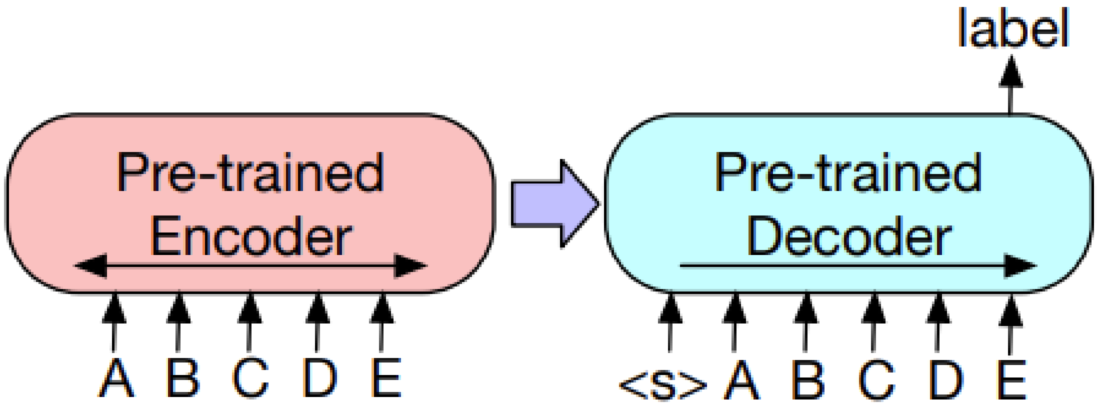
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

BART

Fine-tuning BART

❖ **Sequence Classification Tasks**

- Encoder와 Decoder에 같은 입력값이 들어감
- Final decoder layer의 hidden state 값이 multi-class linear classifier의 입력으로 들어감



Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

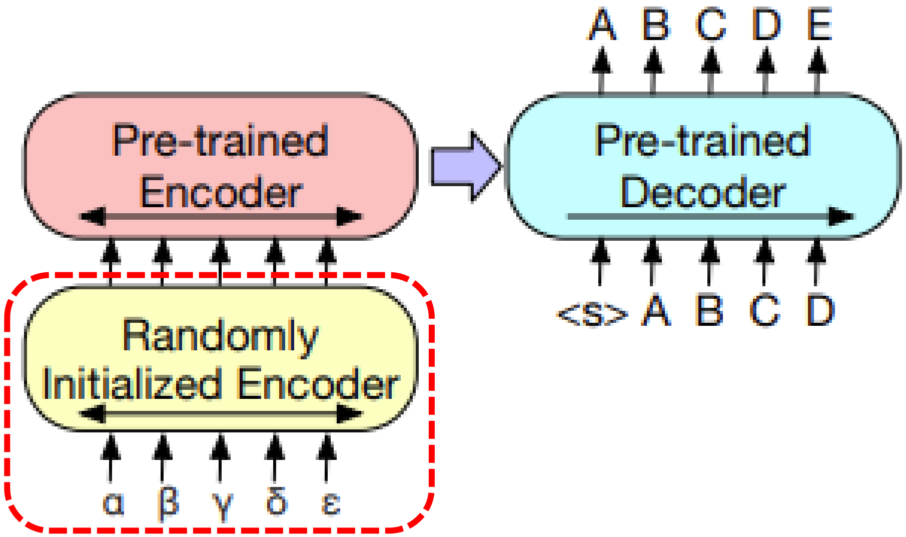
BART

Fine-tuning BART

❖ Machine Translation

- BART 전체 모델을 single pretrained decoder로 사용
- BART의 encoder embedding layer를 randomly initialized encoder로 대체

BART's encoder embedding layer



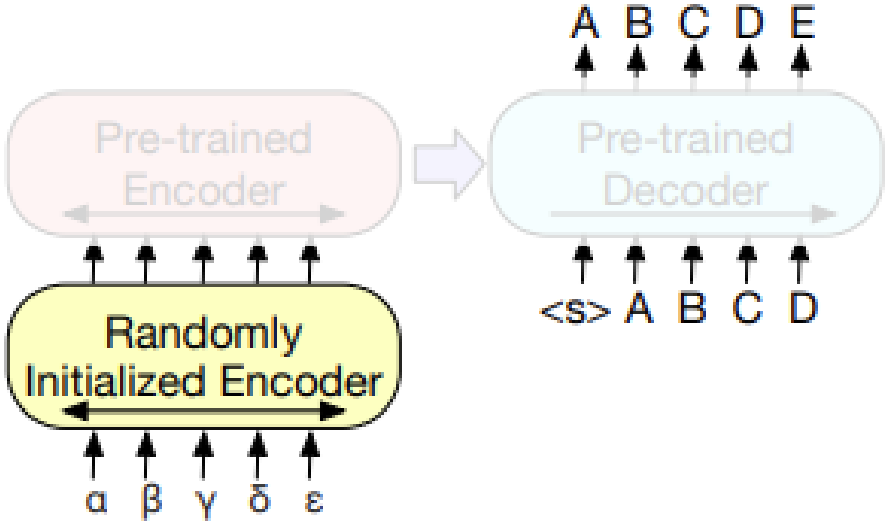
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

BART

Fine-tuning BART

❖ Machine Translation

- Step1: 대부분의 BART parameter들을 freeze 시키고 해당 parameter들만 update
  - ➢ **Randomly initialized encoder / BART positional embeddings / BART encoder의 첫번째 layer의 self-attention input projection matrix**
- Step2: 적은 수의 iteration만큼 전체 parameter 학습



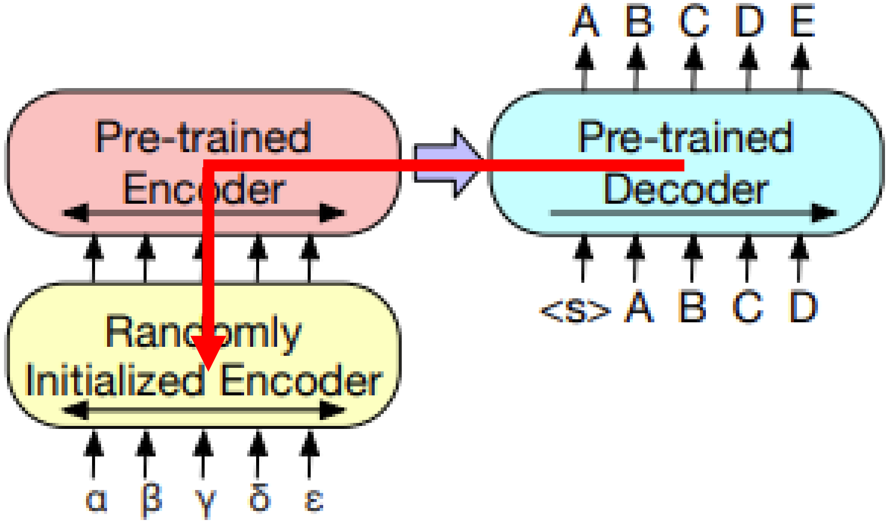
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

BART

Fine-tuning BART

❖ Machine Translation

- Step1: 대부분의 BART parameter들을 freeze 시키고 해당 parameter들만 update
    - ➢ **Randomly initialized encoder / BART positional embeddings / BART encoder의 첫번째 layer의 self-attention input projection matrix**
- Step2: 적은 수의 iteration만큼 전체 parameter 학습



Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.