부스트캠프 AI Tech 2기

boostcamp ai tech

BART

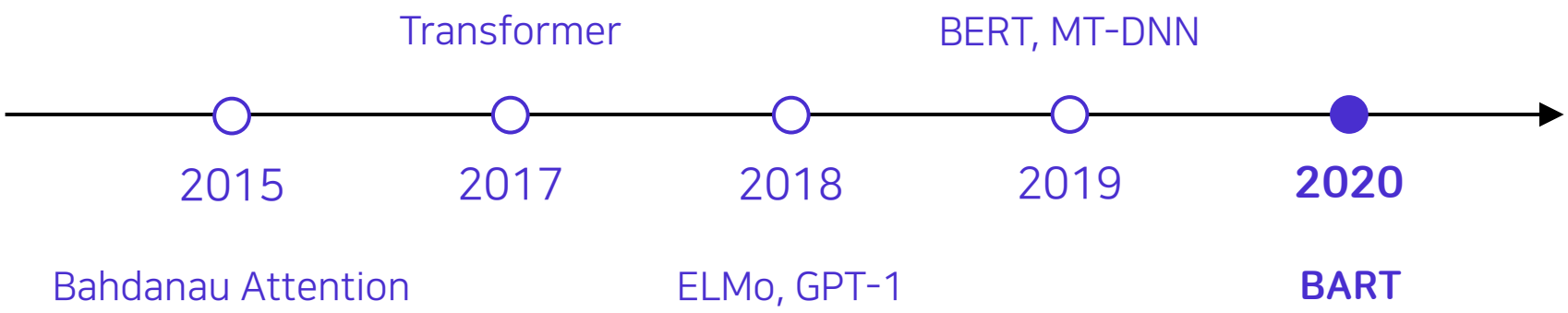
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, ACL 2020

Email : jinmang2@gmail.com

GitHub : github.com/jinmang2

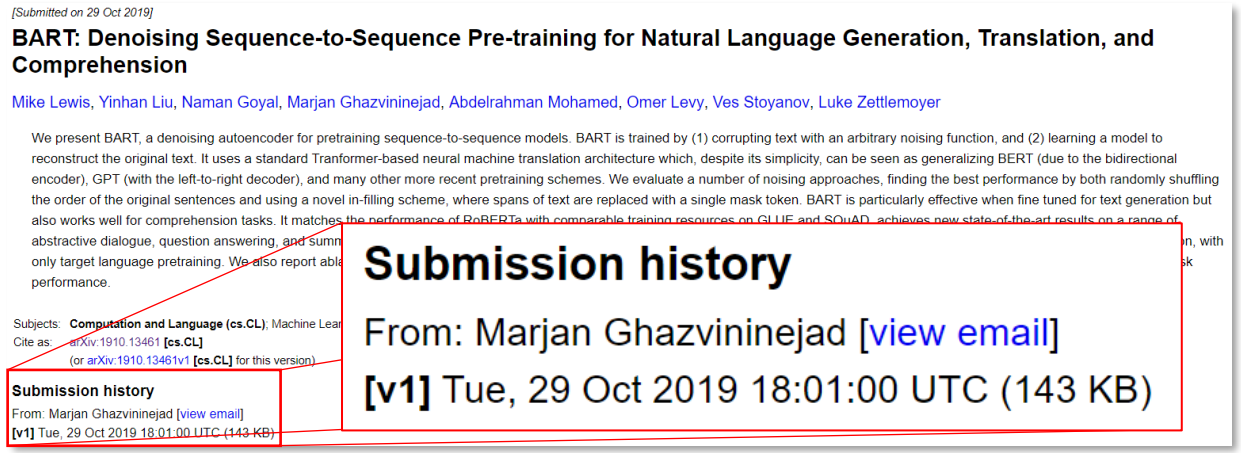
Huggingface Hub: huggingface.co/jinmang2

Boostcamp AI Tech 2 NLP 논문 모임에 오신 여러분 환영합니다!



BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

Lifelog of BART



[Submitted on 29 Oct 2019]

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

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 summ…n, with only target language pretraining. We also report abla…sk performance.

Subjects: **Computation and Language (cs.CL)**; Machine Lear…
Cite as: arXiv:1910.13461 **[cs.CL]**
(or arXiv:1910.13461v1 **[cs.CL]** for this version)

Submission history
From: Marjan Ghazvininejad [view email]
[v1] Tue, 29 Oct 2019 18:01:00 UTC (143 KB)

Submission history

From: Marjan Ghazvininejad [view email]

[v1] Tue, 29 Oct 2019 18:01:00 UTC (143 KB)

ACL Anthology FAQ Corrections Submissions Search...

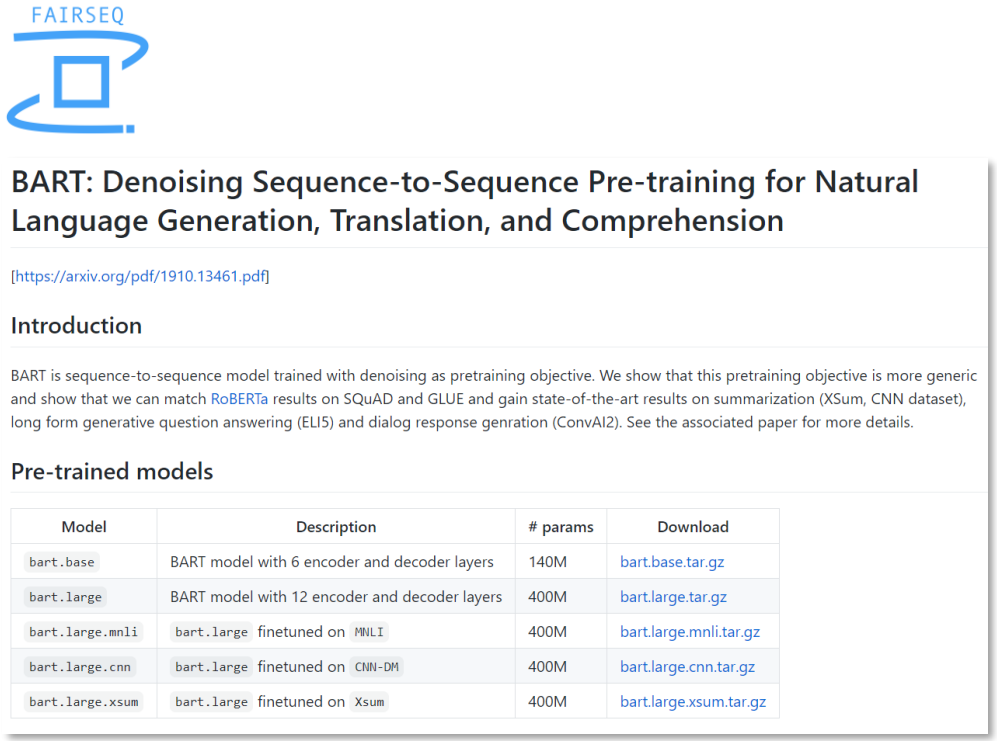
Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

Dan Jurafsky, Joyce Chai, Natalie Schluter, Joel Tetreault (Editors)

Anthology ID: 2020.acl-main
Month: July
Year: 2020
Address: Online
Venue: ACL
SIG: –
Publisher: Association for Computational Linguistics
URL: https://www.aclweb.org/anthology/2020.acl-main
DOI: –
Bib Export formats: BibTeX MODS XML EndNote
PDF: https://www.aclweb.org/anthology/2020.acl-main.pdf

PDF (full)
BibTeX
Search

https://arxiv.org/abs/1910.13461
https://www.aclweb.org/anthology/2020.acl-main.703/



FAIRSEQ

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

[https://arxiv.org/pdf/1910.13461.pdf]

Introduction

BART is sequence-to-sequence model trained with denoising as pretraining objective. We show that this pretraining objective is more generic and show that we can match RoBERTa results on SQuAD and GLUE and gain state-of-the-art results on summarization (XSum, CNN dataset), long form generative question answering (ELI5) and dialog response genration (ConvAI2). See the associated paper for more details.

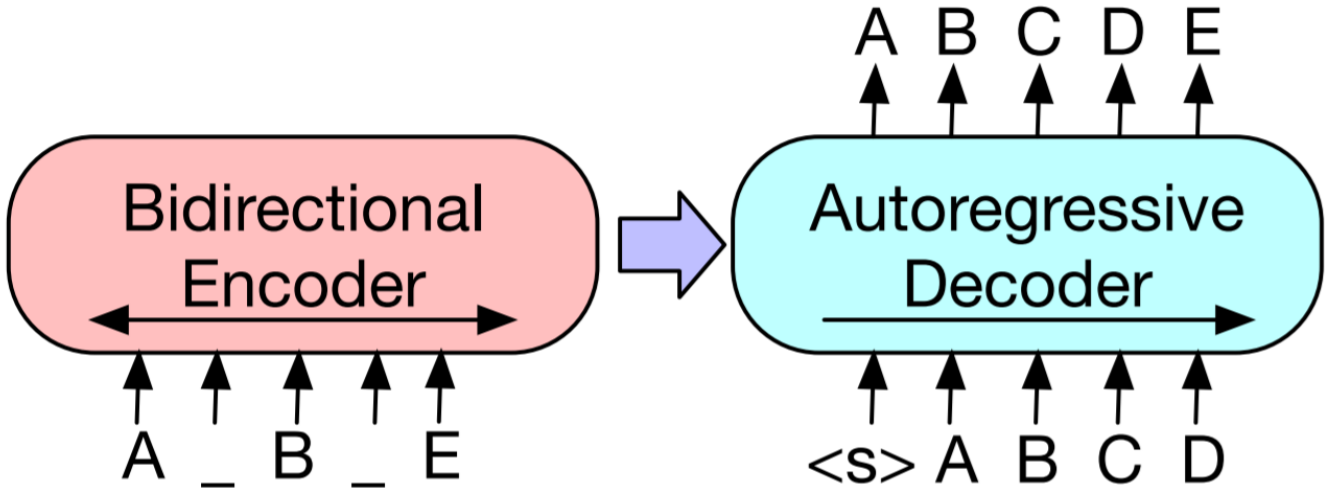
Pre-trained models

Model	Description	# params	Download
`bart.base`	BART model with 6 encoder and decoder layers	140M	bart.base.tar.gz
`bart.large`	BART model with 12 encoder and decoder layers	400M	bart.large.tar.gz
`bart.large.mnli`	`bart.large` finetuned on `MNLI`	400M	bart.large.mnli.tar.gz
`bart.large.cnn`	`bart.large` finetuned on `CNN-DM`	400M	bart.large.cnn.tar.gz
`bart.large.xsum`	`bart.large` finetuned on `Xsum`	400M	bart.large.xsum.tar.gz

https://github.com/pytorch/fairseq/tree/master/examples/bart

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

What is BART? 🤔

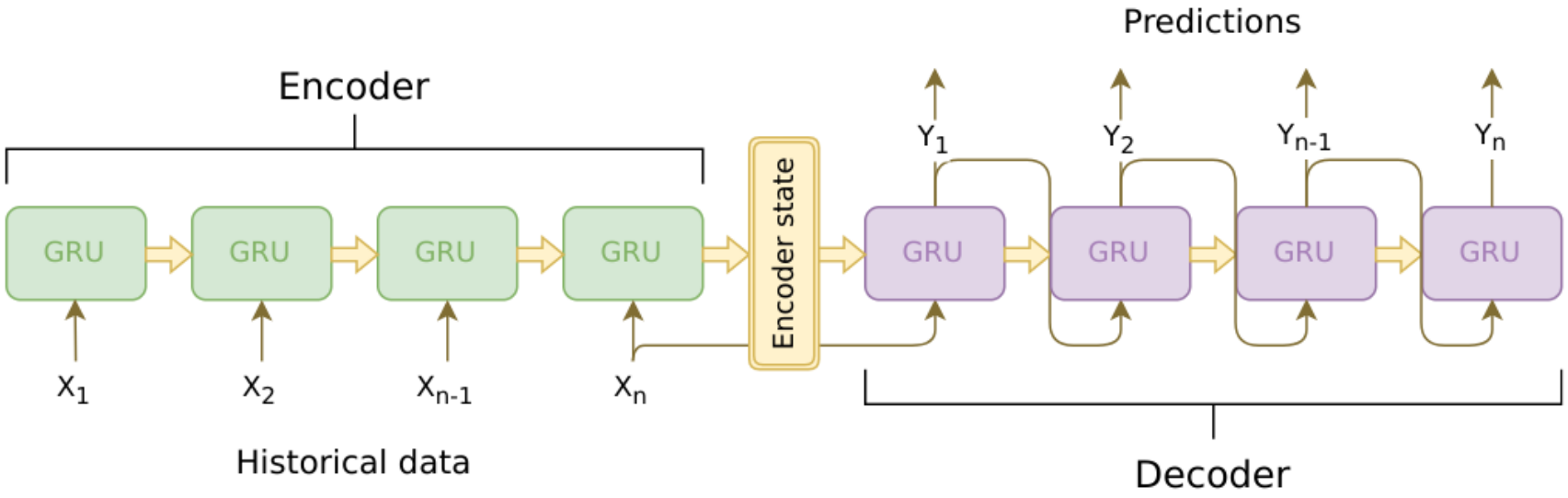


Sequence-to-Sequence model trained with denoising as pretraining objective

https://www.aclweb.org/anthology/2020.acl-main.703/

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

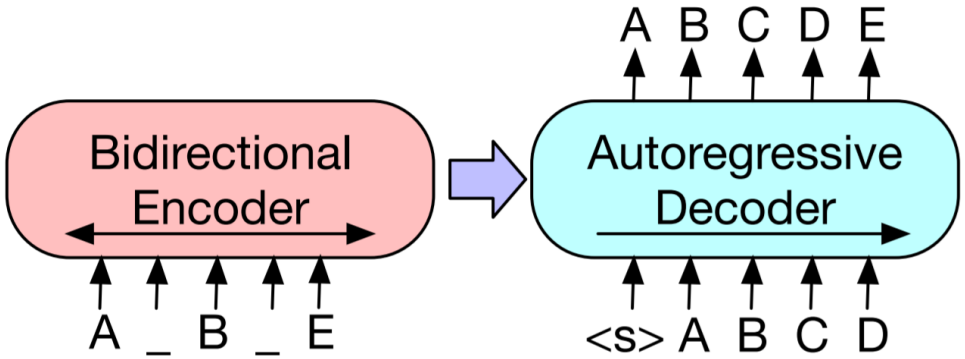


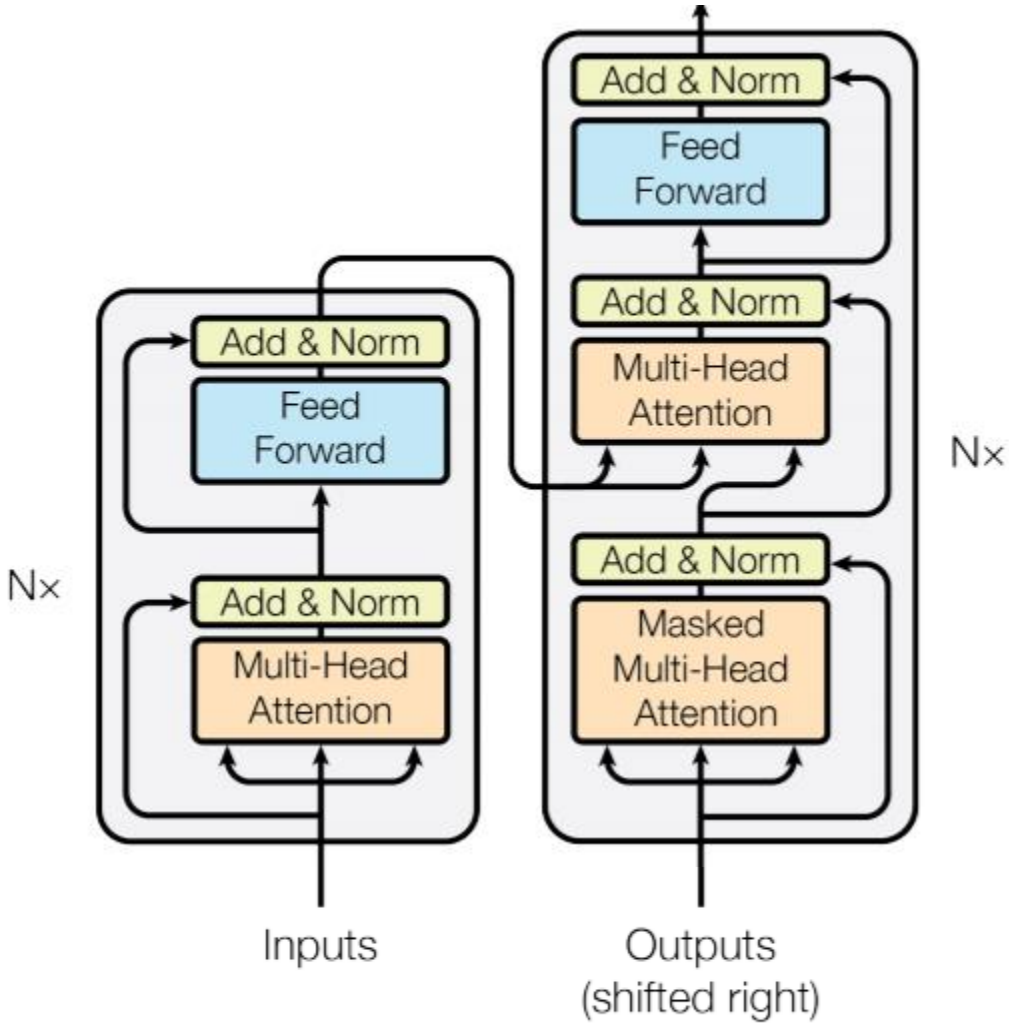
Seq2Seq: Input Sequence를 입력으로 받고 Output Sequence를 출력하는 모델 !! (AutoEncoder)

https://www.aclweb.org/anthology/2020.acl-main.703/

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives



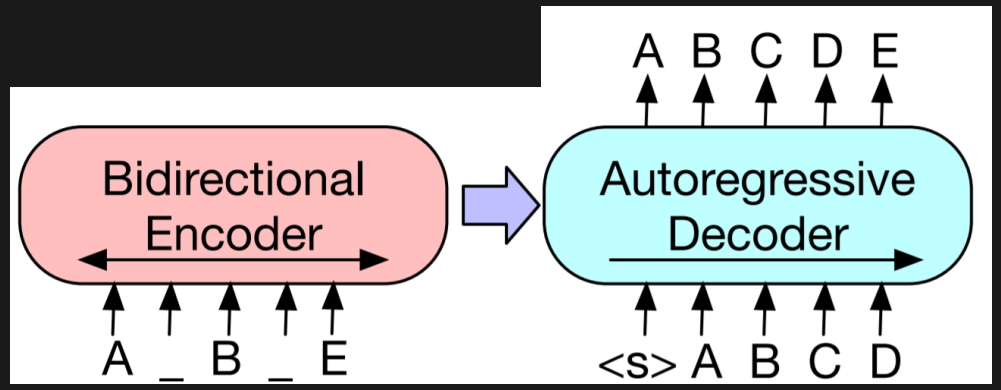


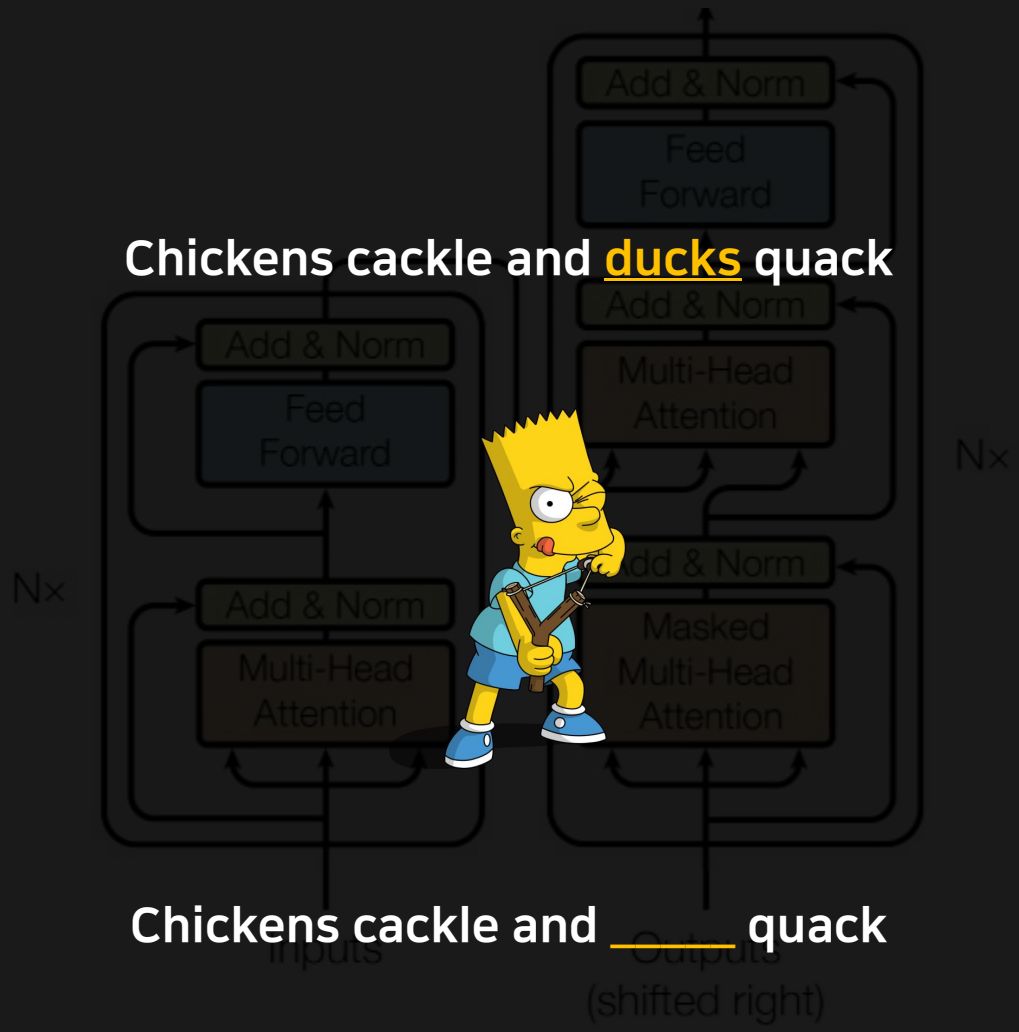
BART의 Encoder-Decoder를 Transformers로! (BERT-GPT)

https://arxiv.org/pdf/1706.03762.pdf

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives





Sequence-to-Sequence model trained with denoising as pretraining objective

https://www.deviantart.com/mighty355/art/Bart-Simpson-01-483298269
https://www.gwern.net/docs/psychology/writing/1953-taylor.pdf

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

Self-supervised 방법론들은 광범위한 NLP task에서 괄목할 성과를 거둠

- Word2Vec, ELMo, BERT, SpanBERT, XLNet, Roberta

이 중 가장 성공적인 접근은 Cloze Tasks에서 영감을 받은 MLM(Masked Language Models)의 변형

- Wilson L Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Bulletin*
- In *XLNet*, Masked Token이 예측될 순서를 개선 (Permutation Operation)
- In *SpanBERT*, Masked Token의 분포를 개선 (많은 NLP Task에서 text span 간 관계 추론이 필요)
- In *UniLM*, Masked Token을 대체할 Context를 개선

그러나 기존의 방법론들을 특정 End Task 형태에 집중하여 활용성이 떨어짐

- XLNet, SpanBERT, UniLM

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

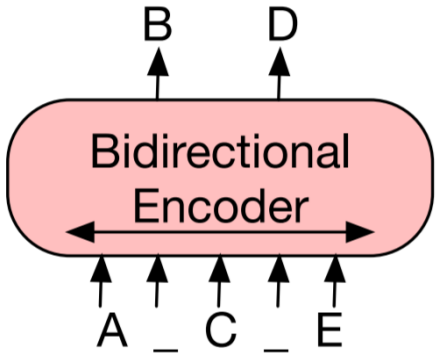
BERT

BART

https://pngio.com/PNG/a127850-bert-png.html
https://www.pngwing.com/en/free-png-dkxeal

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives



BERT

- ✓ Transformers의 Encoder로 모델 구축
- ✓ Cloze Task의 영감을 받은 MLM Pre-Train 목적함수와
- ✓ Next Sentence Prediction Pre-Train 목적함수로 학습
- ✓ Albert, Roberta 등 후속 연구에 의하면 MLM이 진또배기!
- ✓ By XLNet Paper, BERT is based on denoising auto-encoding

$$\log p_\theta(\bar{x}|\hat{x}) \approx \sum_{t=1}^{T} m_t \log p_\theta(x_t|\hat{x})$$

- ✓ Independence Assumption

  모든 masked token들은 독립적으로 재구축됨
- ✓ Input Noise

  Pretrain-Finetune discrependancy
- ✓ 위의 단점을 XLNet은 Autoregressive한 LM으로 해결코자 함

  PLM(Permutation Language Model)

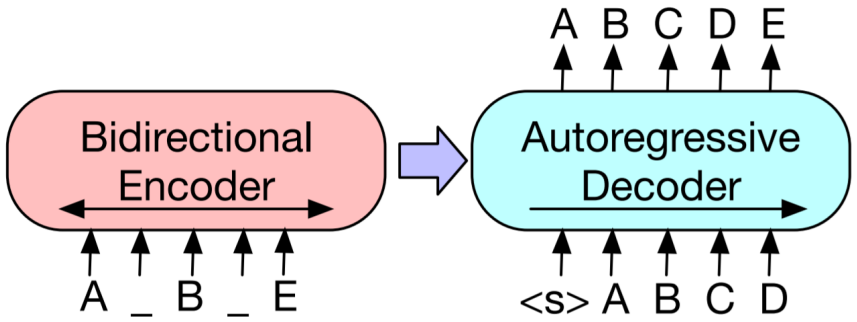
https://arxiv.org/pdf/1906.08237.pdf

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

Base: 6 layers
Large: 12 layers

BART Architecture & Denoising Objectives

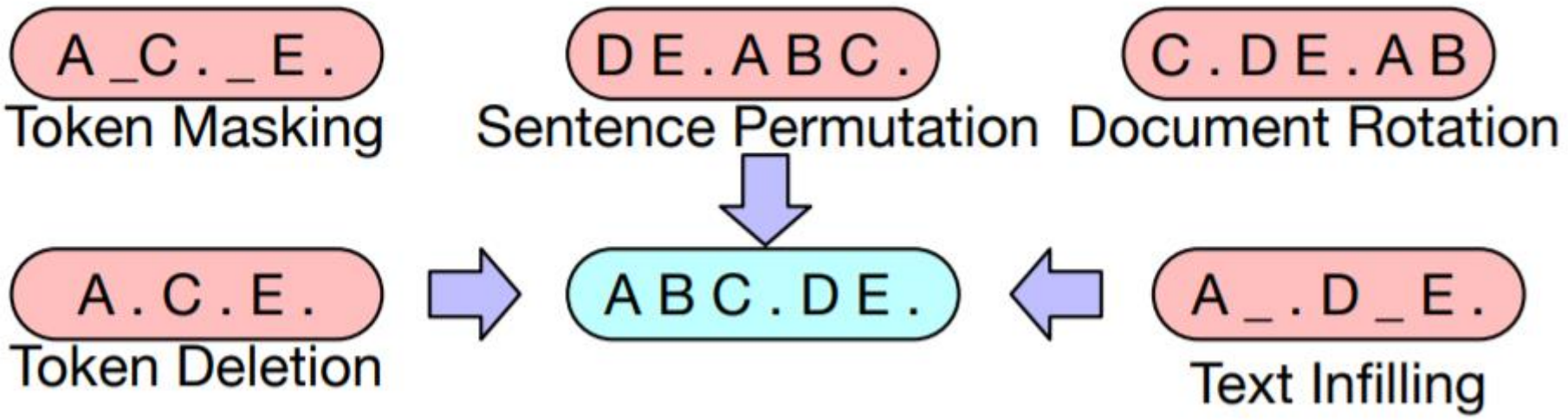
- ✓ BART는 Encoder-Decoder 모델! (Transformers)
- ✓ Bidirectional Transformer's Encoder (as like BERT)
  - Decoder의 각 Layer별 Cross-Attention을 수행
  - Word Prediction 전 추가적인 FFN 필요 X
- ✓ Autoregressive Transformers' Decoder (as like GPT)
  - ReLU대신 GeLUs 사용
  - Parameter를 $N(0, 0.02)$로 초기화
- ✓ BERT의 단점들을 개선 가능
- ✓ XLNet, SpanBERT, UniLM 등 기존의 MLM 변형체들 채용
- ✓ Noise Flexibility; origin text에 임의의 변환을 적용할 수 있음



BART

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

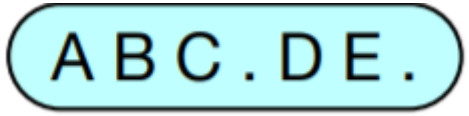


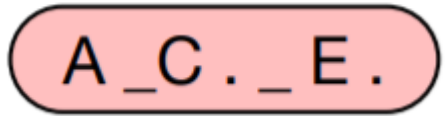
어떤 Noise를 줬을까??

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

Token Masking





BERT (Devlin et al., 2019)와 동일하게,

Random Token들을 sampling하고 이를 [MASK] token으로 대체

어제 그녀와 만났다.

그녀와의 데이트는 너무나도 달콤했다.

어제 [MASK] 만났다.

[MASK] 데이트는 너무나도 달콤했다.

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

Token Deletion

입력에서 임의로 token들을 제거

Token Masking과 다르게, Model은 어느 자리의 Token이 유실됐는지 결정해야 함.

어제 그녀와 만났다.

그녀와의 데이트는 너무나도 달콤했다.

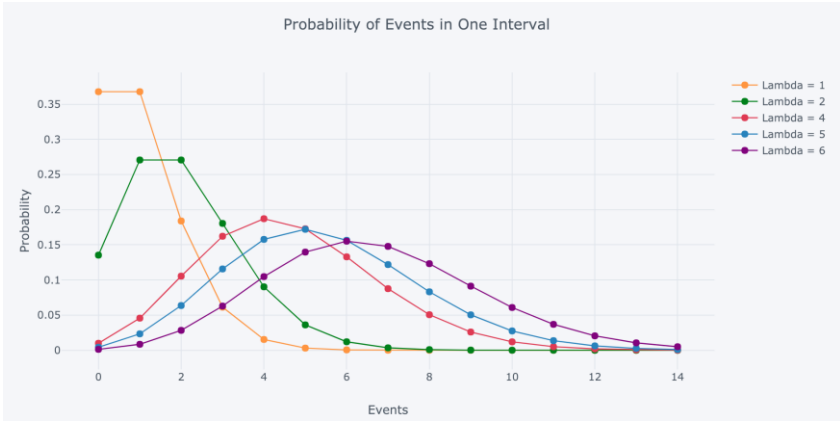
어제 만났다.

데이트는 너무나도 달콤했다.

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

Text Infilling



Poisson($\lambda = 3$) 분포에서 Span Length를 추출한 길이 만큼의

Text Span을 Sampling하고 각 Token을 단일 [MASK] token으로 대체한다.

길이가 0일 경우에는 [MASK] token을 삽입하는 것과 동일하다.

이는 모델에게 Span에서부터 얼마나 많은 Token이 유실됐는지 예측하도록 학습시킨다.

SpanBERT (Joshi et al., 2019)에서 영감을 받았다고 한다.

$$\lambda = \mathbb{E}(X) = \mathrm{Var}(X)$$

어제 그녀와 만났다.

그녀와의 데이트는 너무나도 달콤했다.

어제 그녀와 [MASK] 만났다.

[MASK] 너무나도 달콤했다.

https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

Sentence Permutation

Full Stop들을 기준으로 Document를 Sentences로 나누고

이 문장들을 임의의 순서로 섞어준다.

어제 그녀와 만났다.

그녀와의 데이트는 너무나도 달콤했다.

그녀와의 데이트는 너무나도 달콤했다.

어제 그녀와 만났다.

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

Document Rotation

임의로 Token을 선택하고 Document를 해당 Token으로 시작하도록 Rotate

이 Task는 Model이 문장의 시작이 어딘지 학습할 수 있게 도와준다.

어제 그녀와 만났다.

그녀와의 데이트는 너무나도 달콤했다.

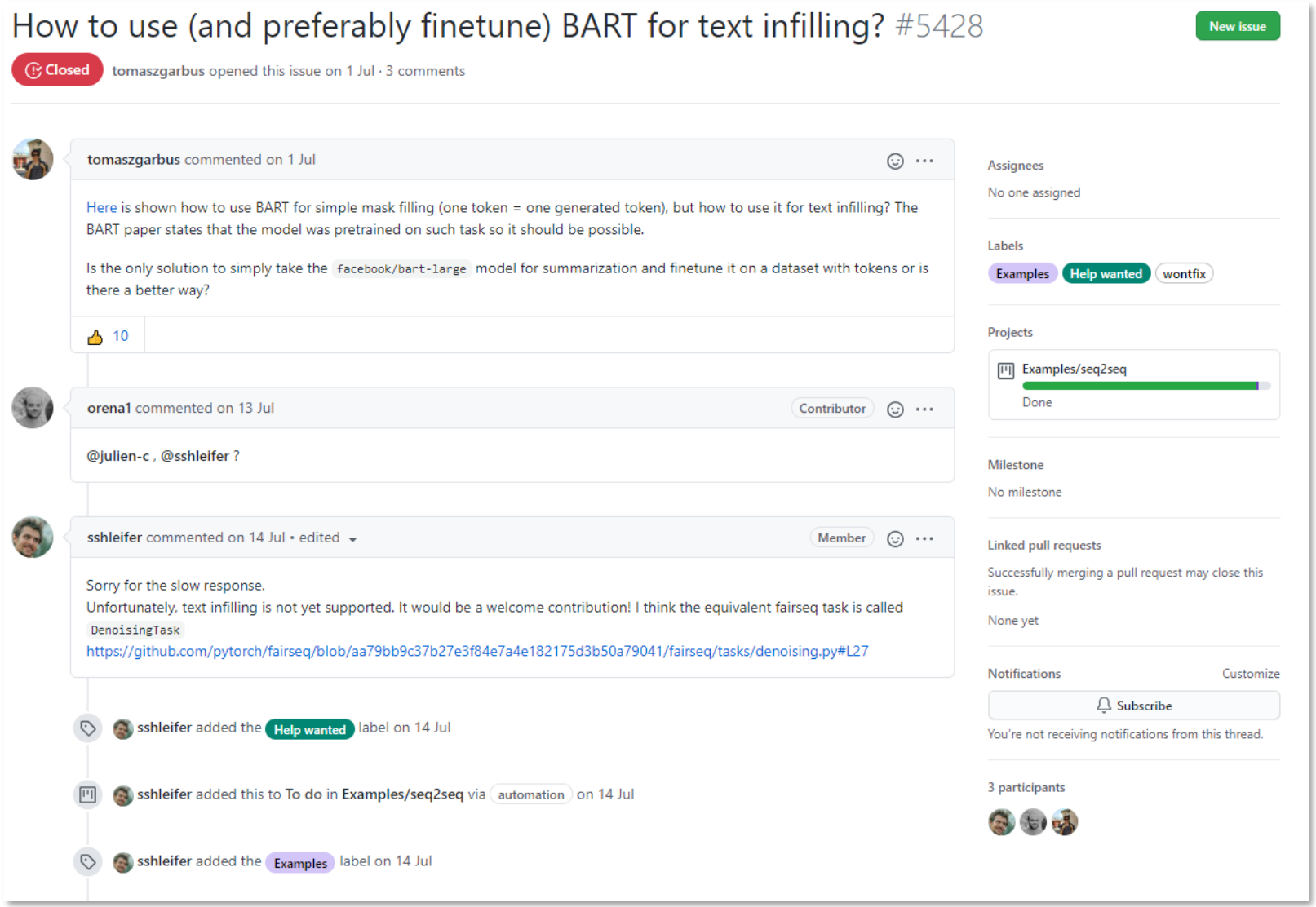
만났다.

그녀와의 데이트는 너무나도 달콤했다.

어제 그녀와

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives



How to use (and preferably finetune) BART for text infilling? #5428

New issue

Closed tomaszgarbus opened this issue on 1 Jul · 3 comments

tomaszgarbus commented on 1 Jul

Here is shown how to use BART for simple mask filling (one token = one generated token), but how to use it for text infilling? The BART paper states that the model was pretrained on such task so it should be possible.

Is the only solution to simply take the `facebook/bart-large` model for summarization and finetune it on a dataset with tokens or is there a better way?

10

orena1 commented on 13 Jul Contributor

@julien-c , @sshleifer ?

sshleifer commented on 14 Jul • edited Member

Sorry for the slow response.
Unfortunately, text infilling is not yet supported. It would be a welcome contribution! I think the equivalent fairseq task is called `DenoisingTask`
https://github.com/pytorch/fairseq/blob/aa79bb9c37b27e3f84e7a4e182175d3b50a79041/fairseq/tasks/denoising.py#L27

sshleifer added the Help wanted label on 14 Jul

sshleifer added this to To do in Examples/seq2seq via automation on 14 Jul

sshleifer added the Examples label on 14 Jul

Assignees
No one assigned

Labels
Examples Help wanted wontfix

Projects
Examples/seq2seq
Done

Milestone
No milestone

Linked pull requests
Successfully merging a pull request may close this issue.
None yet

Notifications Customize
Subscribe
You're not receiving notifications from this thread.

3 participants

https://github.com/huggingface/transformers/issues/5428

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives

```python
# Copyright (c) Facebook, Inc. and its affiliates.
#
# This source code is licensed under the MIT license found in the
# LICENSE file in the root directory of this source tree.

import logging
import os

from fairseq.data import (
    data_utils,
    Dictionary,
    AppendTokenDataset,
    DenoisingDataset,
    PrependTokenDataset,
    StripTokenDataset,
    TokenBlockDataset,
)
from fairseq.data.encoders.utils import get_whole_word_mask
from fairseq.tasks import FairseqTask, register_task
from fairseq import utils


logger = logging.getLogger(__name__)


@register_task('denoising')
class DenoisingTask(FairseqTask):
    """
    Denoising task for applying sequence to sequence denoising. (ie. BART)
    """
```

```python
class DenoisingDataset(FairseqDataset):
    """
    A wrapper around TokenBlockDataset for BART dataset.

    Args:
        dataset (TokenBlo
        sizes (List[int])
        vocab (~fairseq.d
        mask_idx (int): d
        mask_whole_words:
            over vocab in
            word. We will
        shuffle (bool, op
          Default: ``True
        seed: Seed for ra
        args: argparse ar
    """
```

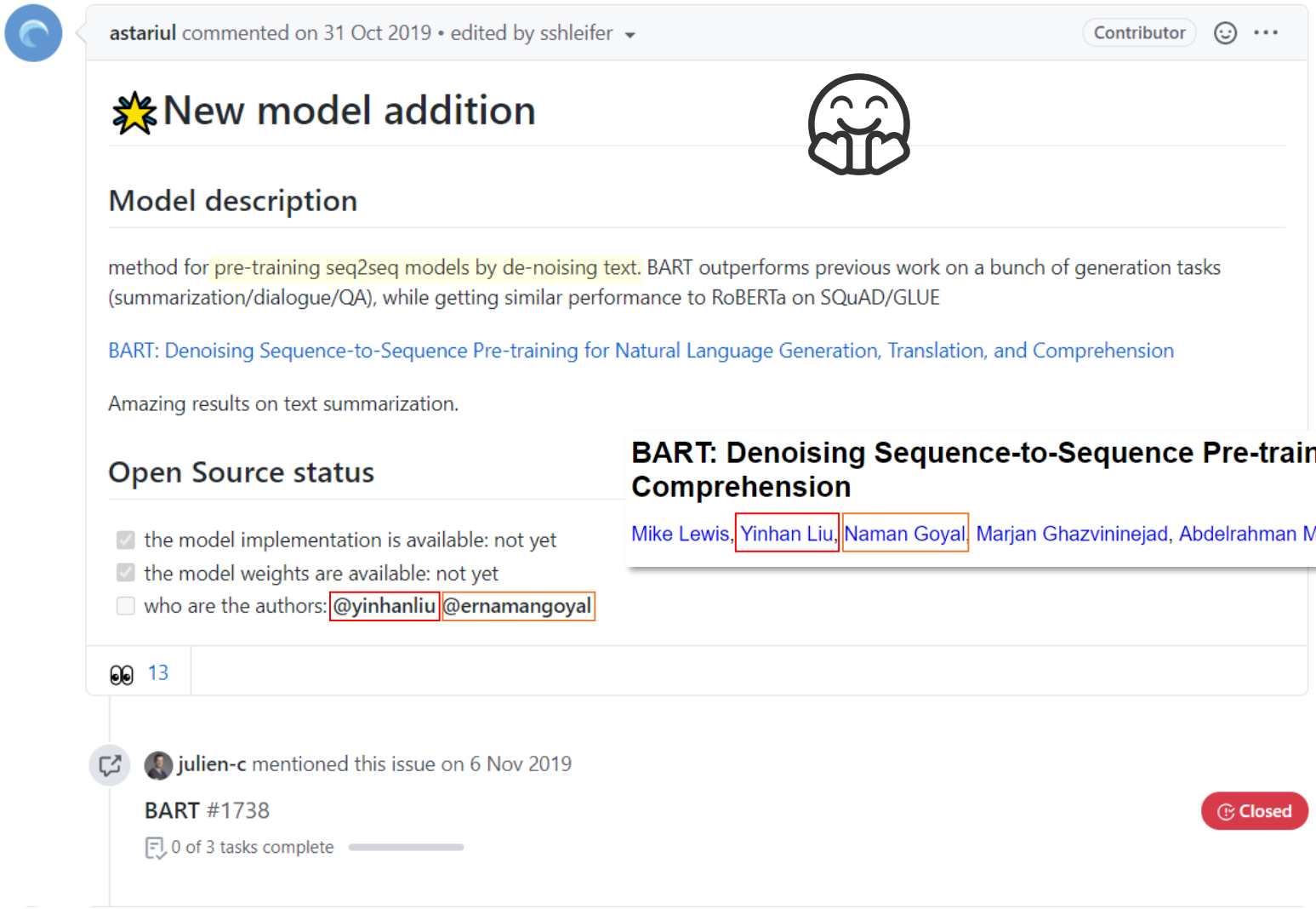
```python
    def __getitem__(self, index):
        with data_utils.numpy_seed(self.seed, self.epoch, index):
            tokens = self.dataset[index]
            assert tokens[-1] == self.eos
            source, target = tokens, tokens.clone()

            if self.permute_sentence_ratio > 0.0:
                source = self.permute_sentences(source, self.permute_sentence_ratio)

            if self.mask_ratio > 0:
                source = self.add_whole_word_mask(source, self.mask_ratio)

            if self.insert_ratio > 0:
                source = self.add_insertion_noise(source, self.insert_ratio)

            if self.rotate_ratio > 0.0 and np.random.random() < self.rotate_ratio:
                source = self.add_rolling_noise(source)
            # there can additional changes to make:
            if self.item_transform_func is not None:
                source, target = self.item_transform_func(source, target)

            assert (source >= 0).all()
            assert (source[1:-1] >= 1).all()
            assert (source <= len(self.vocab)).all()
            assert source[0] == self.vocab.bos()
            assert source[-1] == self.eos
            return {
                "id": index,
                "source": source,
                "target": target,
            }
```

Noise Flexibility

https://github.com/pytorch/fairseq/blob/master/fairseq/tasks/denoising.py
https://github.com/pytorch/fairseq/blob/master/fairseq/data/denoising_dataset.py
https://github.com/pytorch/fairseq/issues/2303

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives



astariul commented on 31 Oct 2019 • edited by sshleifer — Contributor

New model addition

Model description

method for pre-training seq2seq models by de-noising text. BART outperforms previous work on a bunch of generation tasks (summarization/dialogue/QA), while getting similar performance to RoBERTa on SQuAD/GLUE

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

Amazing results on text summarization.

Open Source status

- [x] the model implementation is available: not yet
- [x] the model weights are available: not yet
- [ ] who are the authors: @yinhanliu @ernamangoyal

13

julien-c mentioned this issue on 6 Nov 2019

BART #1738 — Closed

0 of 3 tasks complete

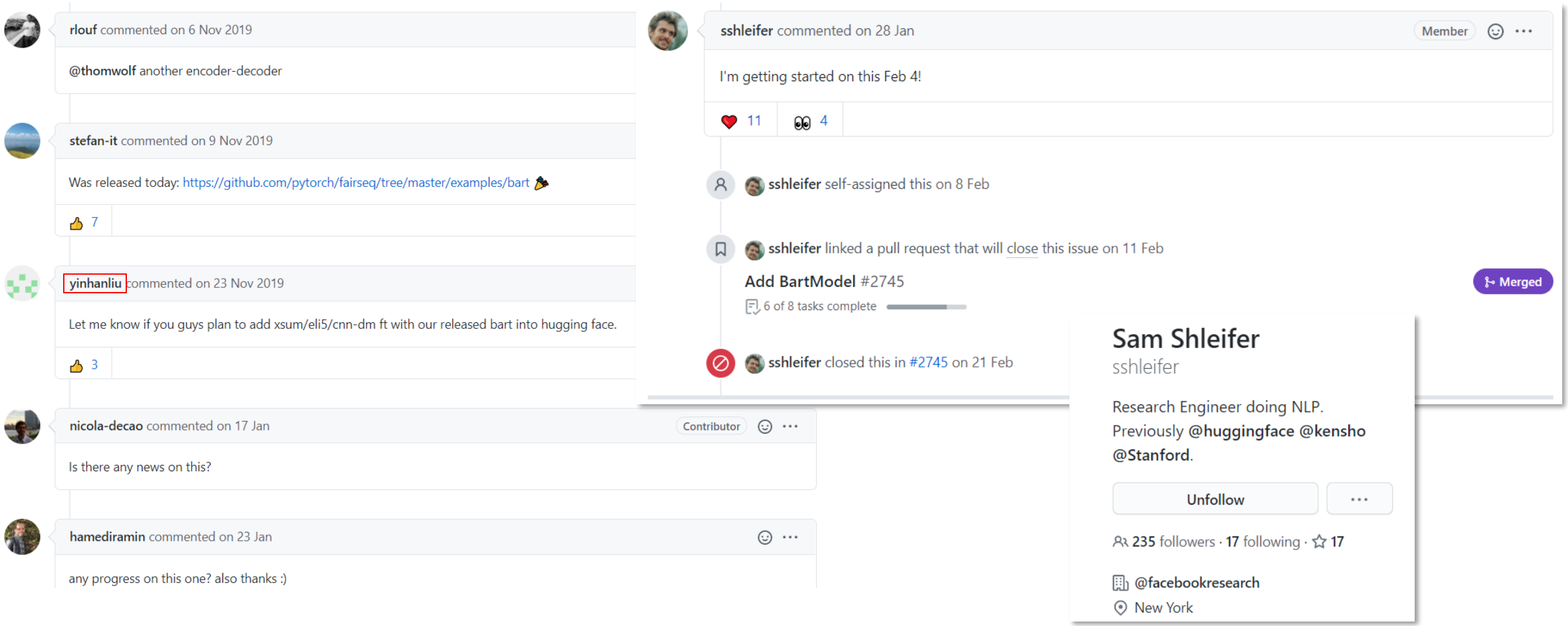
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer

https://github.com/huggingface/transformers/issues/1676

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

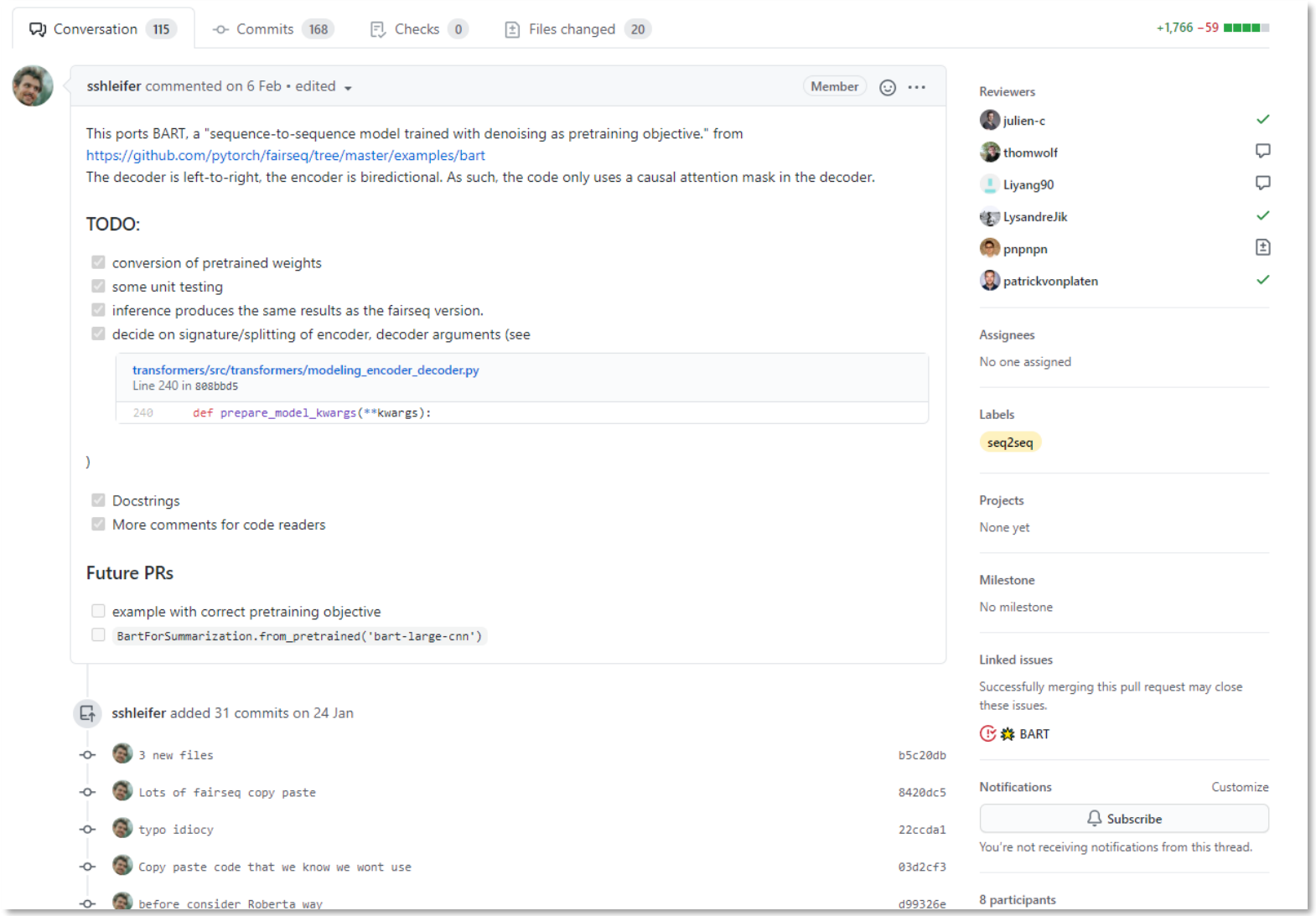
BART Architecture & Denoising Objectives



https://github.com/sshleifer

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives



TODO

- Pretrained weight conversion
- Some unit testing
- fairseq 버전과 동일한 추론 결과
- Encoder, Decoder arguments
- Docstrings
- 코드 reader들을 위한 주석 작업

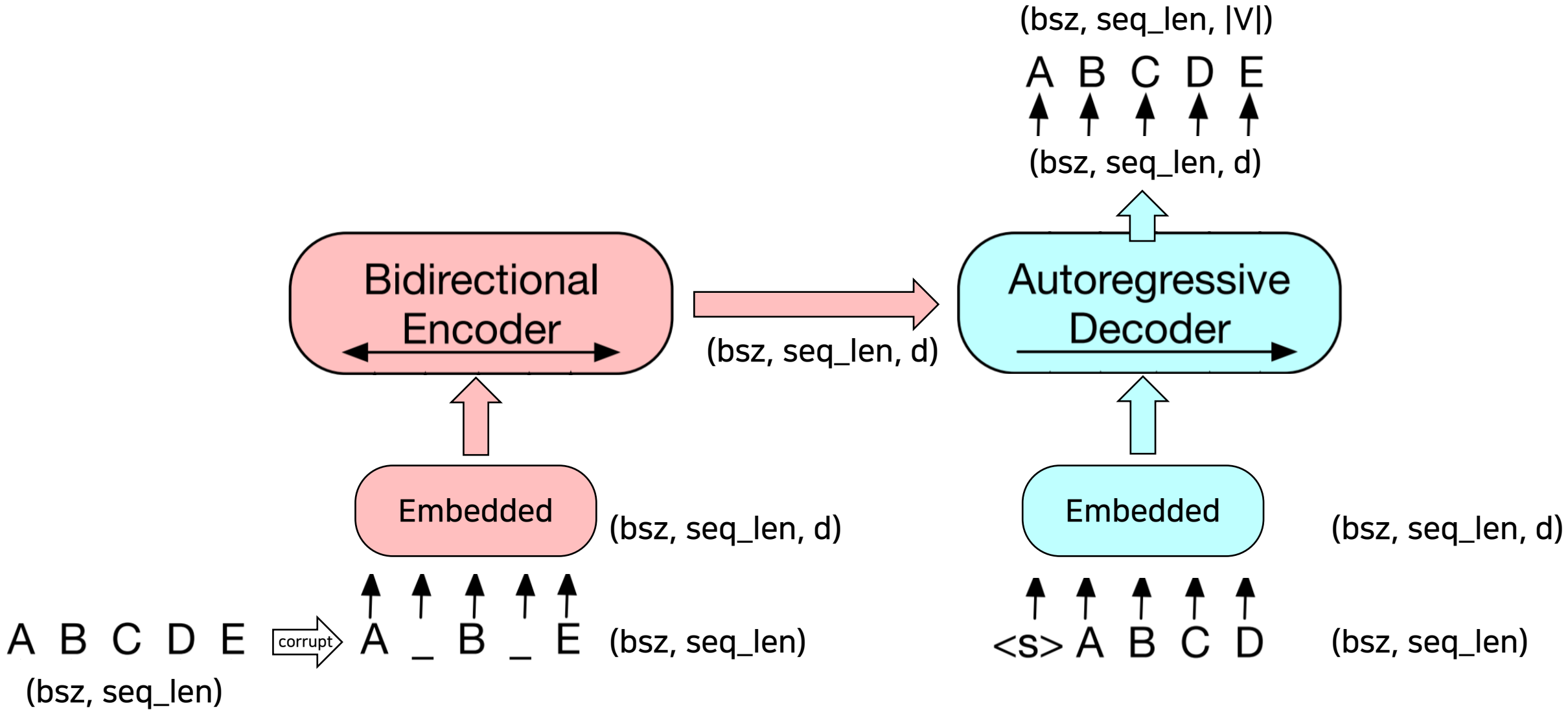
Future PRs

- Pretraining objective 예제
- BartForSummarization 구현

https://github.com/huggingface/transformers/pull/2745

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

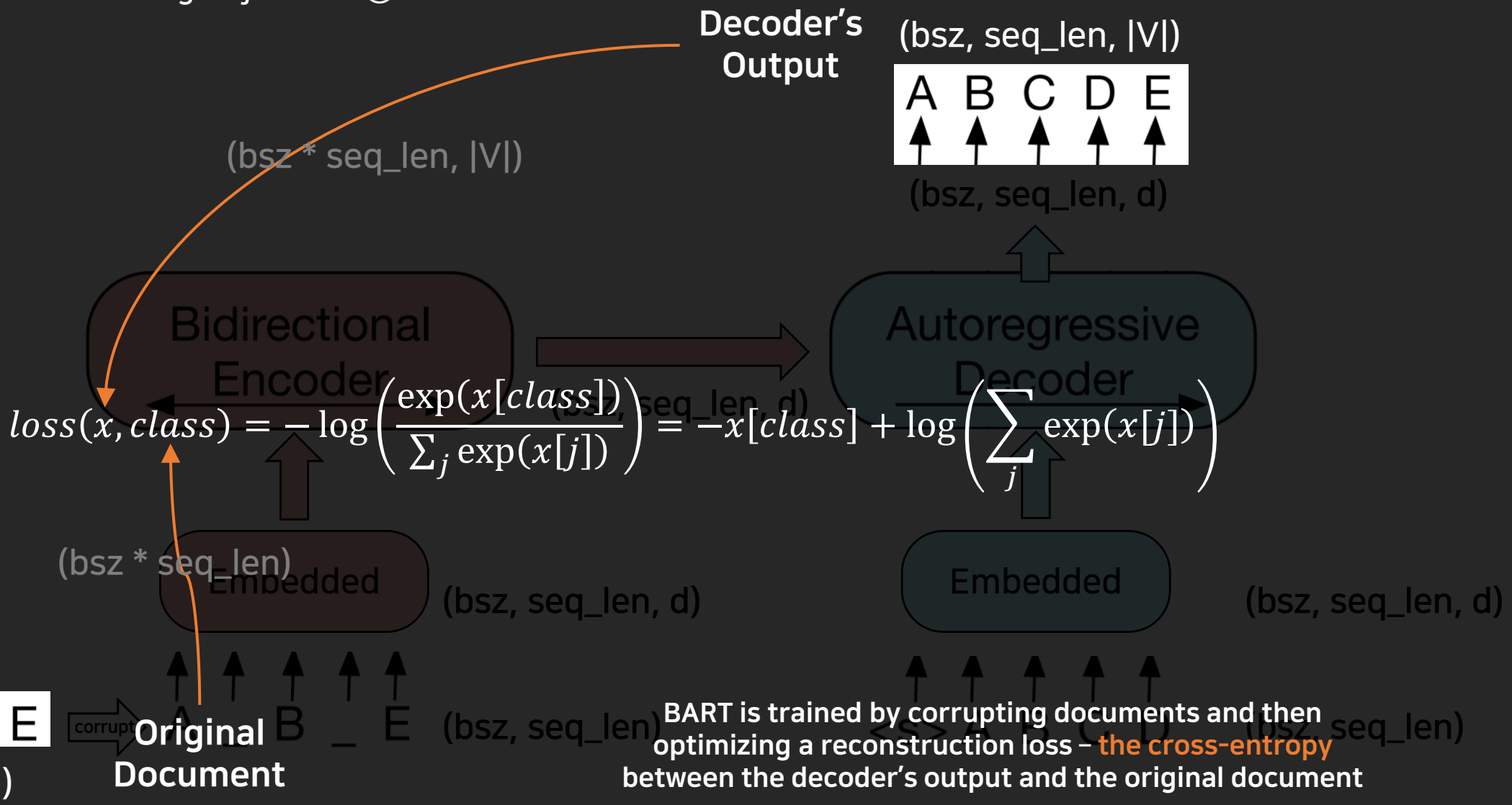
BART Architecture & Denoising Objectives



https://github.com/huggingface/transformers/blob/master/src/transformers/models/bart/modeling_bart.py

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Architecture & Denoising Objectives



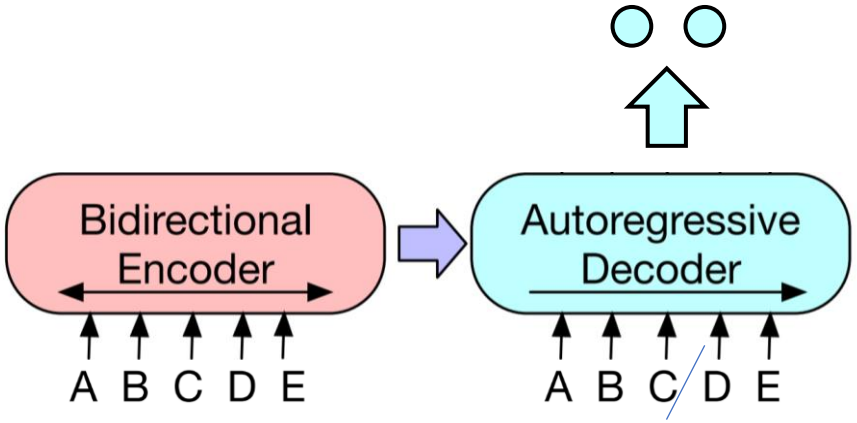
BART is trained by corrupting documents and then optimizing a reconstruction loss – the cross-entropy between the decoder's output and the original document

https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

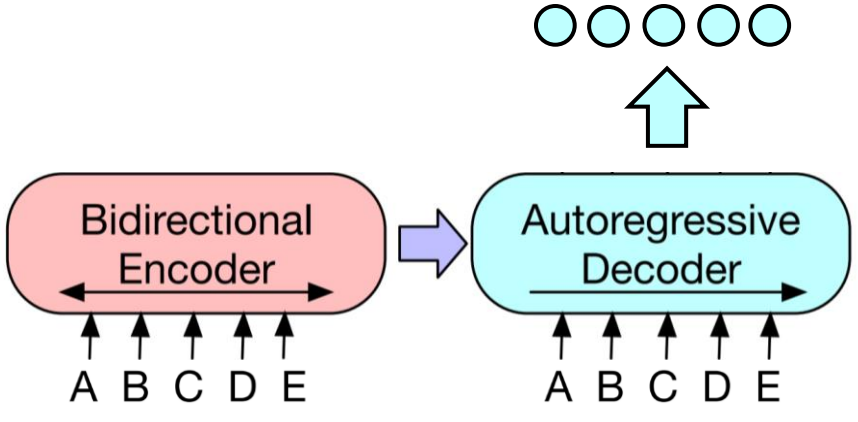
BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

Fine-tuning BART

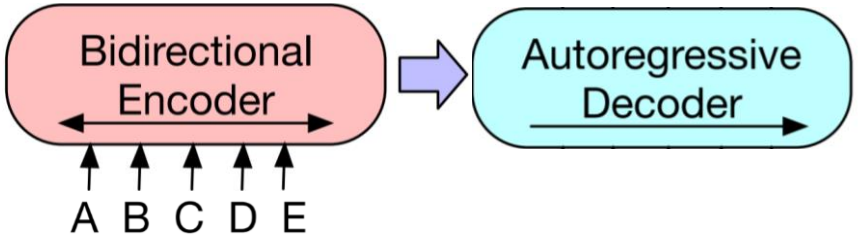
Sentence Classification Tasks



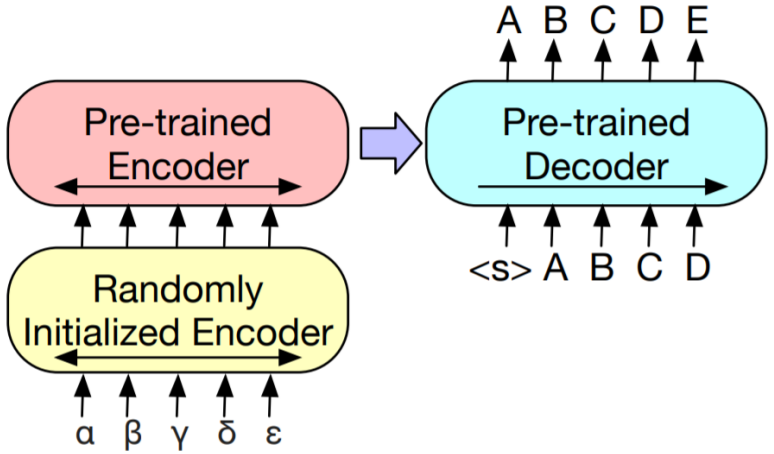
Token Classification Tasks



Sequence Generation Tasks

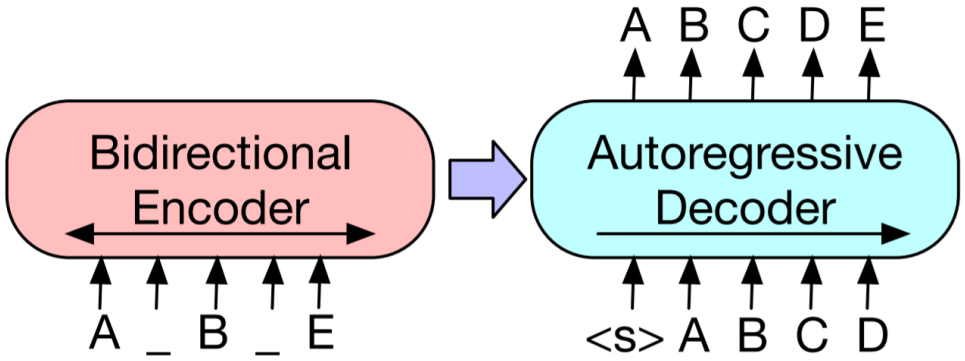


Machine Translation Tasks



BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Performance @ NLP Tasks



Task	Description
SQuAD	wikipedia 단락에 대한 대한 질문/답변, 질문과 연결된 컨텍스트를 BART의 인코더에 대한 입력으로 사용, 추가 디코더에 전달
MNLI	Fine-tuned model은 EOS 토큰이 추가된 두 문장을 연결, BART 인코더와 디코더에 전달, EOS 토큰은 문장 관계를 분류 하는데 사용
ELI5	긴 형식의 추상적인 질문/응답 데이터 질문과 supporting document의 연결을 조건으로 답변 생성
XSum	매우 추상적 요약이 포함된 뉴스 요약 데이터
ConvAI2	context와 페르소나를 조건으로 하는 대화 응답 생성
CNN/DM	뉴스 요약 데이터

Model	Description
Language Model(CLM)	GPT와 유사하게 left-to-right transformer 모델을 학습
Permuted Language Model(PLM)	XLNet을 기반으로 토큰의 1/6을 샘플링하고 auto regressively 방식으로 생성
Masked Language Model(MLM)	BERT에 따라 15% token을[MASK]로 대체하고 학습
Multitask Masked Language Model	UniLM과 같이 additional self-attention mask. Self attention mask -> 1/6 left-to-right, 1/6 right-to-left, 1/3 un-masked, 1/3의 50%는 un-masked, 나머지는 left-to-right
Masked Seq-to-Seq(MASS)	토큰의 50%를 포함하는 범위를 mask, masked된 토큰을 예측하기 위한 seq-to-seq 모델로 훈련

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Performance @ NLP Tasks

Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
Masked Language Model	90.0	83.5	24.77	7.87	12.59	7.06
Masked Seq2seq	87.0	82.1	23.40	6.80	11.43	6.19
Language Model	76.7	80.1	**21.40**	7.00	11.51	6.56
Permuted Language Model	89.1	83.7	24.03	7.69	12.23	6.96
Multitask Masked Language Model	89.2	82.4	23.73	7.50	12.39	6.74
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Performance @ NLP Tasks

encoder와 decoder의 hidden size => 12
batch size => 8,000
train steps => 500,000
tokenizing method => BPE
Text Infilling + Sentence Shuffling => masking 30% of token, permute all sentences
train step의 마지막 10% => dropout off
pre-training data => 160Gb (news + books + stories + web text)

	MNLI m/mm	SST Acc	QQP Acc	QNLI Acc	STS-B Acc	RTE Acc	MRPC Acc	CoLA Mcc
BERT	86.6/-	93.2	91.3	92.3	90.0	70.4	88.0	60.6
UniLM	87.0/85.9	94.5	-	92.7	-	70.9	-	61.1
XLNet	89.8/-	95.6	91.8	93.9	91.8	83.8	89.2	63.6
RoBERTa	**90.2/90.2**	96.4	92.2	94.7	**92.4**	86.6	**90.9**	**68.0**
BART	89.9/90.1	**96.6**	**92.5**	**94.9**	91.2	**87.0**	90.4	62.8

	SQuAD 1.1 EM/F1	SQuAD 2.0 EM/F1
BERT	84.1/90.9	79.0/81.8
UniLM	-/-	80.5/83.4
XLNet	**89.0**/94.5	86.1/88.8
RoBERTa	88.9/**94.6**	**86.5/89.4**
BART	88.8/**94.6**	86.1/89.2

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Performance @ NLP Tasks

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
ROBERTASHARE (Rothe et al., 2019)	40.31	18.91	37.62	41.45	18.79	33.90
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Performance @ NLP Tasks

	ELI5		
	R1	R2	RL
Best Extractive	23.5	3.1	17.5
Language Model	27.8	4.7	23.1
Seq2Seq	28.3	5.1	22.8
Seq2Seq Multitask	28.9	5.4	23.1
BART	**30.6**	**6.2**	**24.3**

	RO-EN
Baseline	36.80
Fixed BART	36.29
Tuned BART	**37.96**

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

BART Performance @ NLP Tasks

단순히 input 정보를 넘어서,

background knowledge를 활용하는 모습을 보여줌

- "Source Document가 journal Science의 연구임"을 말함 (1st example)
- "PG&E 가 California에 있다"고 말함 (5th example)

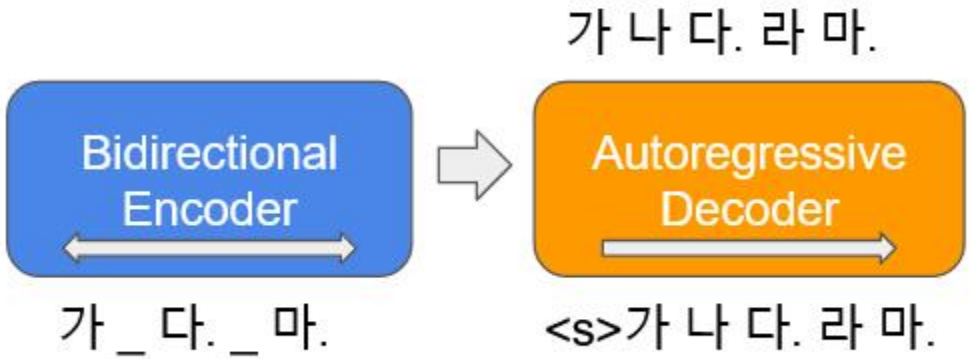
https://github.com/modulabs/beyondBERT/issues/6

Source Document (abbreviated)	BART Summary
The researchers examined three types of coral in reefs off the coast of Fiji ... The researchers found when fish were plentiful, they would eat algae and seaweed off the corals, which appeared to leave them more resistant to the bacterium Vibrio coralliilyticus, a bacterium associated with bleaching. The researchers suggested the algae, like warming temperatures, might render the corals' chemical defenses less effective, and the fish were protecting the coral by removing the algae.	Fisheries off the coast of Fiji are protecting coral reefs from the effects of global warming, according to a study in the journal Science.
Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... Prime Minister Johnson was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House."	Boris Johnson has said he will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.
According to Syrian state media, government forces began deploying into previously SDF controlled territory yesterday. ... On October 6, US President Donald Trump and Turkish President Recep Tayyip Erdoan spoke on the phone. Then both nations issued statements speaking of an imminent incursion into northeast Syria ... . On Wednesday, Turkey began a military offensive with airstrikes followed by a ground invasion.	Syrian government forces have entered territory held by the US-backed Syrian Democratic Forces (SDF) in response to Turkey's incursion into the region.
This is the first time anyone has been recorded to run a full marathon of 42.195 kilometers (approximately 26 miles) under this pursued landmark time. It was not, however, an officially sanctioned world record, as it was not an "open race" of the IAAF. His time was 1 hour 59 minutes 40.2 seconds. Kipchoge ran in Vienna, Austria. It was an event specifically designed to help Kipchoge break the two hour barrier.	Kenyan runner Eliud Kipchoge has run a marathon in less than two hours.
PG&E stated it scheduled the blackouts in response to forecasts for high winds amid dry conditions. The aim is to reduce the risk of wildfires. Nearly 800 thousand customers were scheduled to be affected by the shutoffs which were expected to last through at least midday tomorrow.	Power has been turned off to millions of customers in California as part of a power shutoff plan.

Table 7: Example summaries from the XSum-tuned BART model on WikiNews articles. For clarity, only relevant excerpts of the source are shown. Summaries combine information from across the article and prior knowledge.

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

KOBART Summarization Example



https://github.com/SKT-AI/KoBART

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

KOBART Summarization Example

Examples

- KoBART ChitChatBot
- KoBART Summarization
- NSMC Classification
- KoBART Translation

KoBART를 사용한 흥미로운 예제가 있다면 PR 주세요!

https://github.com/haven-jeon/KoBART-chatbot
https://github.com/seujung/KoBART-summarization
https://github.com/SKT-AI/KoBART/tree/main/examples
https://github.com/seujung/KoBART-translation

Data

Data	# of Sentences
Korean Wiki	5M
Other corpus	0.27B

한국어 위키 백과 이외, 뉴스, 책, 모두의 말뭉치 v1.0(대화, 뉴스, ...), 청와대 국민청원 등의 다양한 데이터가 모델 학습에 사용되었습니다.

Tokenizer

`tokenizers` 패키지의 `Character BPE tokenizer`로 학습되었습니다.

`vocab` 사이즈는 30,000 이며 대화에 자주 쓰이는 아래와 같은 이모티콘, 이모지 등을 추가하여 해당 토큰의 인식 능력을 올렸습니다.

> 😀, 😁, 😆, 😅, 🤣, .., `:-)`, `:)`, `-)`, `(-:`...

또한 `<unused0>` ~ `<unused99>`등의 미사용 토큰을 정의해 필요한 `subtasks`에 따라 자유롭게 정의해 사용할 수 있게 했습니다.

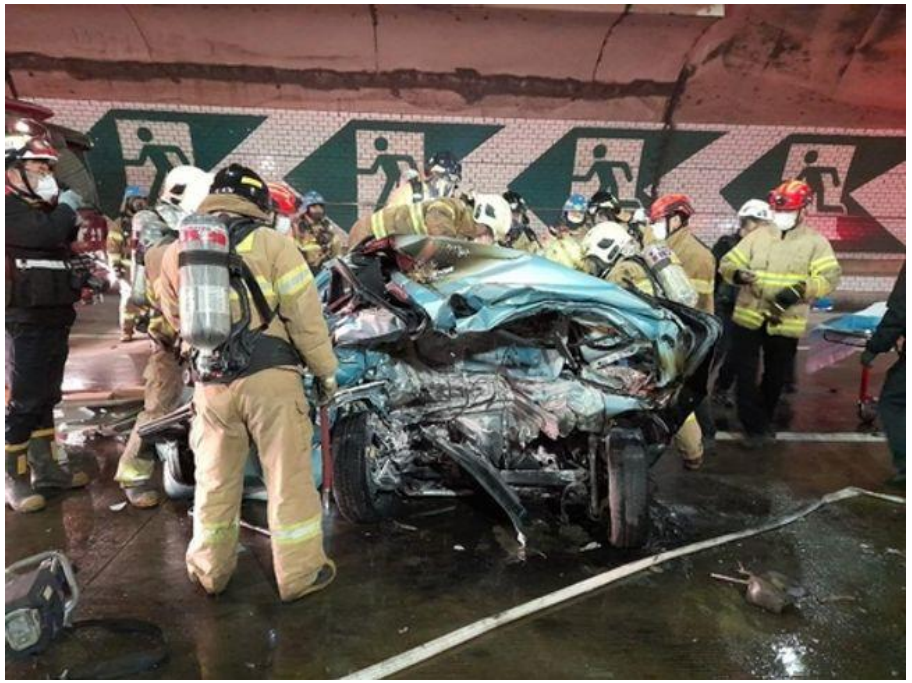
```
>>> from kobart import get_kobart_tokenizer
>>> kobart_tokenizer = get_kobart_tokenizer()
>>> kobart_tokenizer.tokenize("안녕하세요. 한국어 BART 입니다.🤣:)l^o")
['_안녕하', '세요.', '_한국어', '_B', 'A', 'R', 'T', '_입', '니다.', '🤣', ':)', 'l^o']
```

Model

Model	# of params	Type	# of layers	# of heads	ffn_dim	hidden_dims
`KoBART-base`	124M	Encoder	6	16	3072	768
		Decoder	6	16	3072	768

BART: Denoising Sequence-to-Sequence Pre-training for NLG, Translation, and Comprehension

KOBART Summarization Example



원문 뉴스

회식 후 만취 상태에서 벤츠 차량을 몰다 추돌사고를 내 앞차 운전자를 숨지게 한 40대 남성이 경찰에 붙잡혔다.

인천 중부경찰서는 17일 특정범죄가중처벌법상 위험운전치사(윤창호법), 도로교통법상 음주운전 혐의 등을 적용해 A(44)씨를 붙잡아 조사하고 있다고 밝혔다.

A씨는 앞서가던 마티즈 차량을 들이받아 운전자 B(41)씨를 숨지게 한 혐의를 받고 있다.

A씨는 전날 오후 9시 10분쯤 인천시 중구 수도권 제2외곽순환고속도로 김포방향 북항터널 내 편도 3차로 가운데 2차로에서 차량을 몰던 중 마티즈 차량 후미를 들이받은 것으로 조사됐다.

마티즈 차량은 차선을 벗어났고 차량에 불이 났다. 이 사고로 운전자 B씨는 불붙은 차에서 미처 빠져나오지 못하고 숨졌다. 불은 출동한 소방당국에 의해 19분만에 진화됐으나 마티즈 차량은 전소됐다.

당시 A씨의 혈중알코올농도는 면허취소 수치인 0.08% 이상으로 확인됐다.

A씨는 경찰 조사에서 "미추홀구에서 지인들과 회식을 했다"며 "사고 당시 기억이 잘 나지 않고 졸음운전을 한 것 같다"고 진술했다.

경찰은 A씨에게 음주운전 중 사망사고를 내면 처벌을 강화하는 '윤창호법'을 적용할 방침이다.

경찰 관계자는 "A씨를 상대로 윤창호법을 적용해 구속영장을 신청할 방침이다"고 말했다.

요약된 뉴스

인천 중부경찰서는 회식 후 만취 상태에서 벤츠 차량을 몰다 추돌사고를 내 앞차 운전자를 숨지게 한 40대 남성을 특정범죄가중처벌법상 위험운전치사, 도로교통법상 음주운전 혐의 등을 적용해 조사하고 있다.

http://20.194.43.11:7874/
https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=102&oid=119&aid=0002453221

부스트캠프 AI Tech 2기

boostcamp ai tech

Discussion

Email : jinmang2@gmail.com

GitHub : github.com/jinmang2

Huggingface Hub: huggingface.co/jinmang2

진명훈_T2216