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

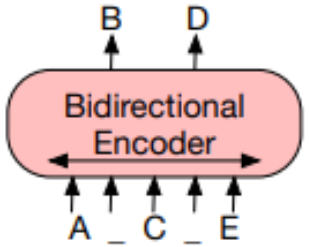
ACL 2020

발제자 : 안제준

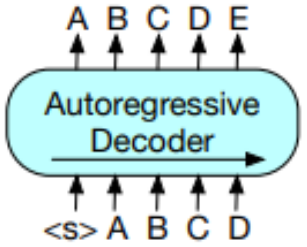
목차

- Introduction
  - BART Motivate
  - BART Outline
- Model
  - Architecture
  - Pre-training BART
- Fine-tuning BART
- Comparing Pre-training Objectives
  - Comparison Objectives
  - Tasks
  - Result
- Qualitative Analysis
- Conclusions

BART Motivate



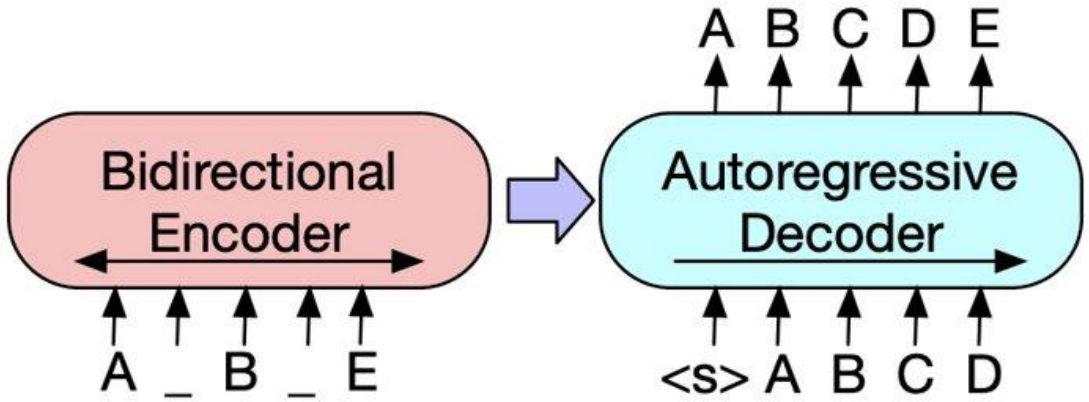
(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.

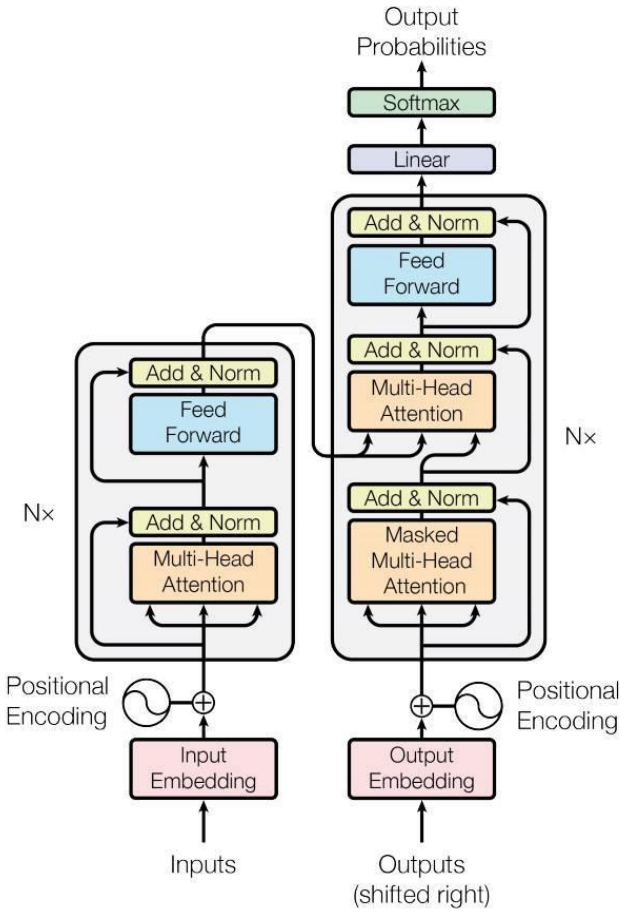
- BERT :
  - 랜덤 토큰은 마스크로 대체
  - 문서는 양방향 인코딩
  - 누락된 토큰은 독립적으로 예측
  - 따라서, BERT를 생성에 쉽게 사용할 수 없음.
  - BERT는 엔코더이기 때문에 Generation task에 대응할 수 없음.
- GPT :
  - 토큰은 자동 회귀로 예측
  - GPT가 생성에 사용될 수 있음을 의미
  - 그러나, 단어는 왼쪽 문맥에서만 조건화할 수 있으므로 양방향 상호 작용을 학습할 수 없다.
  - 디코더만 존재하기 때문에 양방향 문맥정보를 반영하지 못함

BART Outline



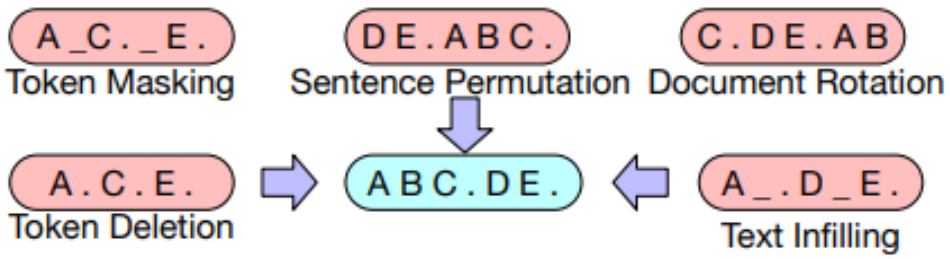
- BART :
  - Bidirectional과 Auto-Regressive Transformer 를 합친 모델
  - seq2seq 모델로 만들어진 denoising autoencoder
  - 많은 종류의 downstream 태스크에서 잘 동작

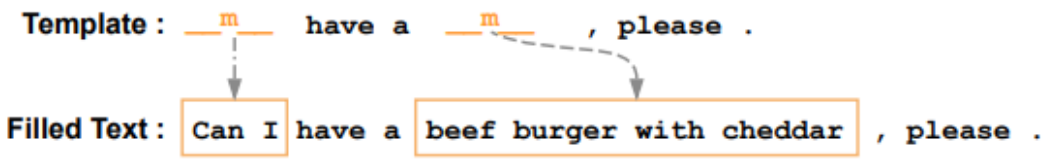
Model Architecture



- Standard Sequence-to-Sequence Transformer 구조
- 단, 디코더에서는 GPT에서 사용하는 ReLU 활성화 함수를 GeLU로 바꿈
- 파라미터 초기화를 N(0,0.2) (N(0,0.2)는 표준 정규분포를 따르는데 평균은 0, std는 0.2의 분포를 갖게 하였다는 뜻)
- base 모델은 엔코더와 디코더에 각각 6개의 레이어를 사용하였고 large 모델은 12개의 레이어를 사용
- 디코더의 각 레이어가 엔코더의 최종 hidden 레이어와 cross-attention을 수행
- BERT(엔코더)는 단어를 유추해내기 위해 추가적인 feed-forward 네트워크를 사용하지만, BART는 그렇지 않다.(엔코더가 바로 masking된 단어를 유추하지 않기 때문)

Model Pre-training BART

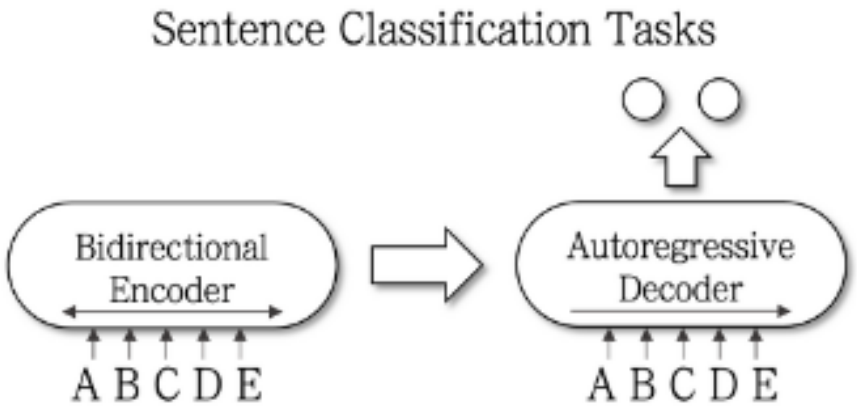




Text Infilling 예시

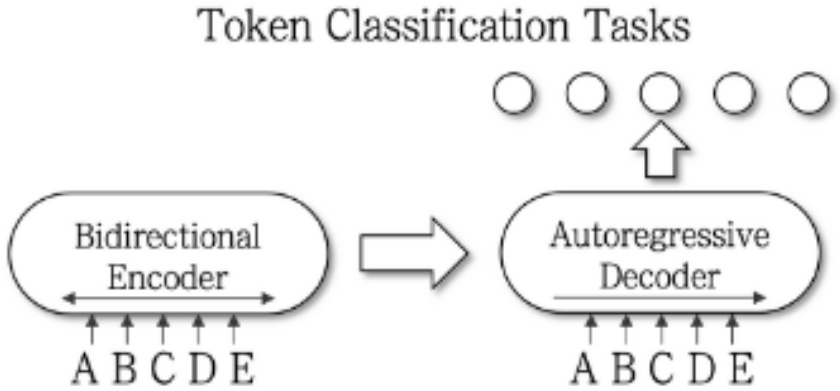
- **Token Masking** :
  - 랜덤한 토큰이 샘플링되어 [MASK]토큰으로 치환
  - 모델은 MASK 토큰이 어떤 토큰이었는지 예측
  - 무작위로 token을 mask하고 복구
- **Token Deletion** :
  - 랜덤한 토큰들이 input에서 치환되지 않고 제거
  - 기존 token masking과는 다르게, 모델은 인풋의 어느 위치가 없어졌는지에 대한 정보도 함께 결정
  - Random으로 token을 삭제하고 복구
- **Text Infilling** :
  - 포아송 분포에 따르는 길이의 text span을 생성
  - 이를 하나의 mask token으로 masking
  - Text span을 생성 mask하고 복구
- **Sentence Permutation** :
  - 하나의 문서가 마침표를 기준으로 문장별로 모두 분리
  - 분리된 문장들은 순서가 랜덤으로 섞임
  - 모델은 섞인 토큰들을 원래의 순서로 배열
  - Document를 문장 단위로 섞고 복구
- **Document Rotation** :
  - 하나의 토큰이 랜덤으로 동일한 확률로 선택
  - 문서가 섞여 해당 토큰이 문서의 시작지점
  - 모델이 그 문서의 시작점을 찾는 것을 통해 훈련

Fine-tuning BART



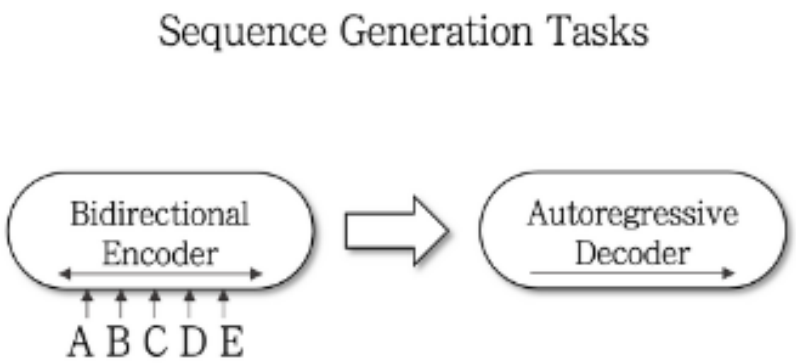
- **Sequence Classification Tasks**
  - Sequence Classification Task는 어떠한 시퀀스를 분류하는 Task
  - 대표적으로 주어진 문장이 문법적으로나 영어적으로 합당한지 분류하는 GLUE의 CoLA가 있다.
  - 해당 태스크에서는 동일한 입력이 엔코더와 디코더에 들어가고,
  - 마지막 decoder token의 마지막 hidden state이 새로운 multi-class 선형 분류기에 들어가게 된다.
  - 이 방법론은 BERT가 CLS 토큰을 분류하는 것에 영감을 얻었다.
  - 그렇지만, 여기에선 추가적인 토큰을 마지막에 추가하여 디코더에 있는 마지막 토큰의 representation이 전체 입력과 attention을 수행할 수 있도록 하여
  - 마지막 output은 모든 입력을 반영할 수 있도록 하였습니다.

Fine-tuning BART



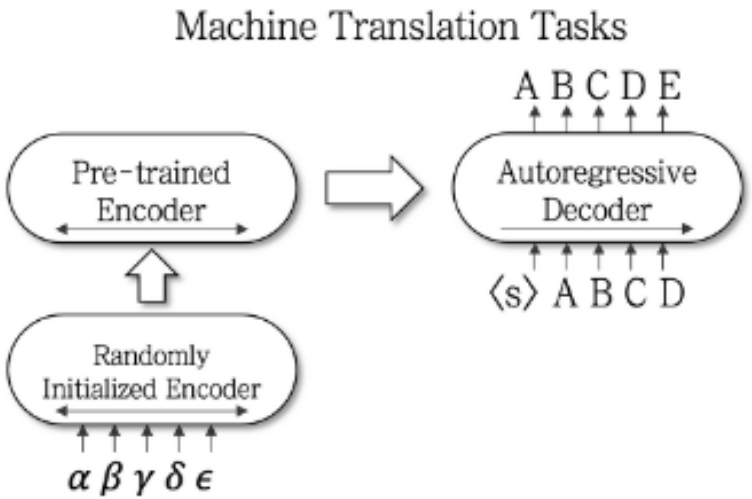
- **Token Classification Tasks**
  - Token classification Task는 토큰단위로 분류를 수행하는 태스크입니다.
  - 대표적으로 주어진 본문 내에서 정답을 찾아야 하는 SQuAD가있습니다.
  - SQuAD는 정답에 해당되는 Start Point와 End Point의 토큰을 찾아야 합니다.
  - BART에서는 모든 문서를 엔코더와 디코더를 입력으로 하고, 디코더의 가장 위의 hidden state를 각 토큰의 representation으로 사용하였습니다.
  - 각 토큰들의 representation을 분류하는데 사용합니다.

Fine-tuning BART



- **Sequence Generation Tasks**
  - 기존 BERT는 할 수 없었던 태스크 중 하나인 Generation Task
  - 그 이유는 BART가 autoregressive 디코더를 가지고 있기 때문
  - abstractive question answering과 summarization과 같은 sequence generation 태스크에 바로 적용 할 수 있다
  - 이러한 태스크들은, 입력 시퀀스가 복사되고 조절되는 특징이 있는데, 이는 denoising pre-training objective랑 밀접하게 연관되어 있다.
  - 엔코더의 입력은 입력 시퀀스가 되고, 디코더는 autoregressive하게 출력을 생성합니다.

Fine-tuning BART



- **Machine Translation**
  - 영어를 다른 언어로 번역하는 Machine Traslation Task 입니다.
  - 해당 태스크에서는 엔코더를 2가지 스텝으로 훈련시킵니다.
  - 두 방식 모두 BART 모델의 출력으로 cross-entropy 로스로 역전파를 적용해 수행합니다.
  - 첫번째 step에서 BART의 대부분의 파라미터를 freeze하고
  - 랜덤으로 초기화된 (embedding layer를 대체하는) source 엔코더, BART의 positional embedding, 그리고 첫번째 엔코더 레이어의 self-attention input projection matrix만 학습시킵니다.
  - 두번째 step은 모든 모델 파라미터를 작은 수의 iteration으로 학습

Comparing Objectives

- Language Model
  - 훈련은 GPT모델이 훈련하는 방식으로, left-to-right 트랜스포머 언어 모델
  - 해당 모델은 BART의 디코더와 동일하지만, cross-attention을 수행하지 않습니다.
- Permuted Language Model
  - 해당 모델은 XLNet 기반 모델이며, ⅙ 토큰만큼 샘플링하여 이를 랜덤한 순서로 autoregressive하게 생성하는 모델입니다.
  - 다른 모델과의 동일한 비교를 위해, 기존 XLNet에서 수행하였던 relative positional embedding이나 segment간의 attention을 수행하지 않았다고 합니다.
- Masked Language Model
  - BERT와 같은 모델링 방법이며, 15%의 토큰을 [MASK] 토큰으로 치환하고 모델을 각 토큰마다 기존 토큰을 예측하도록 훈련합니다.
- Multitask Masked Language Model
  - UniLM에서 제안한 방법으로, Masked Language Model을 추가적인 self-attention mask를 통해 훈련
  - Self attention의 mask는 다음과 같은 비율로 랜덤하게 선택됩니다. : ⅙ left-to-right, ⅙ right-to-left, ⅓ unmasked,
  - 그리고 ⅓의 토큰이 unmask
- Masked Seq-to-Seq
  - MASS에 영감을 받은 모델로, 50%의 토큰을 포함하는 span을 마스크
  - seq2seq 모델로 마스크된 토큰을 예측하도록 훈련합니다.

Task

- SQuAD
  - 위키피디아에서 따온 본문과 질문이 주어지면 주어진 본문로부터 정답에 해당되는 text span을 찾는 문제
- MNLI
  - 두개의 문장에 대한 classification 태스크로 하나의 문장이 다른 문장을 entail, 즉 이전 문장과 이후 문장의 관계가 성립하는지 예측하는 태스크
- ELI5
  - 긴 형식의 abstractive question answering 태스크 입니다. BART 모델은 문제와 추가적인 문서를 붙인 것으로 컨디션을 주어 답을 생성합니다.
- XSum
  - 뉴스 요약 태스크로 많이 함축된 요약을 생성해야합니다.
- ConvAI2
  - 대화의 답변에 대한 generation 태스크로, context와 persona(화자)를 컨디션으로 줍니다.
- CNN/DM
  - 뉴스 요약 데이터셋입니다. 이 데이터셋의 요약본은 입력 문서와 밀접하게 연관되어 있음

Result

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**

예측 성능

	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
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

생성 성능

Result

	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**

요약 성능

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

번역 성능

Qualitative Analysis

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.

Conclusions

- 손상된 문서를 원본에 매핑하는 방법을 배우는 사전 훈련 방식 인 BART를 소개했습니다.

- 향후 연구는 사전 훈련을 위해 문서를 손상 시켜서 특정 최종 작업에 맞게 조정할 수 있는 새로운 방법을 모색해야합니다.

포아송 분포

정의 [편집]

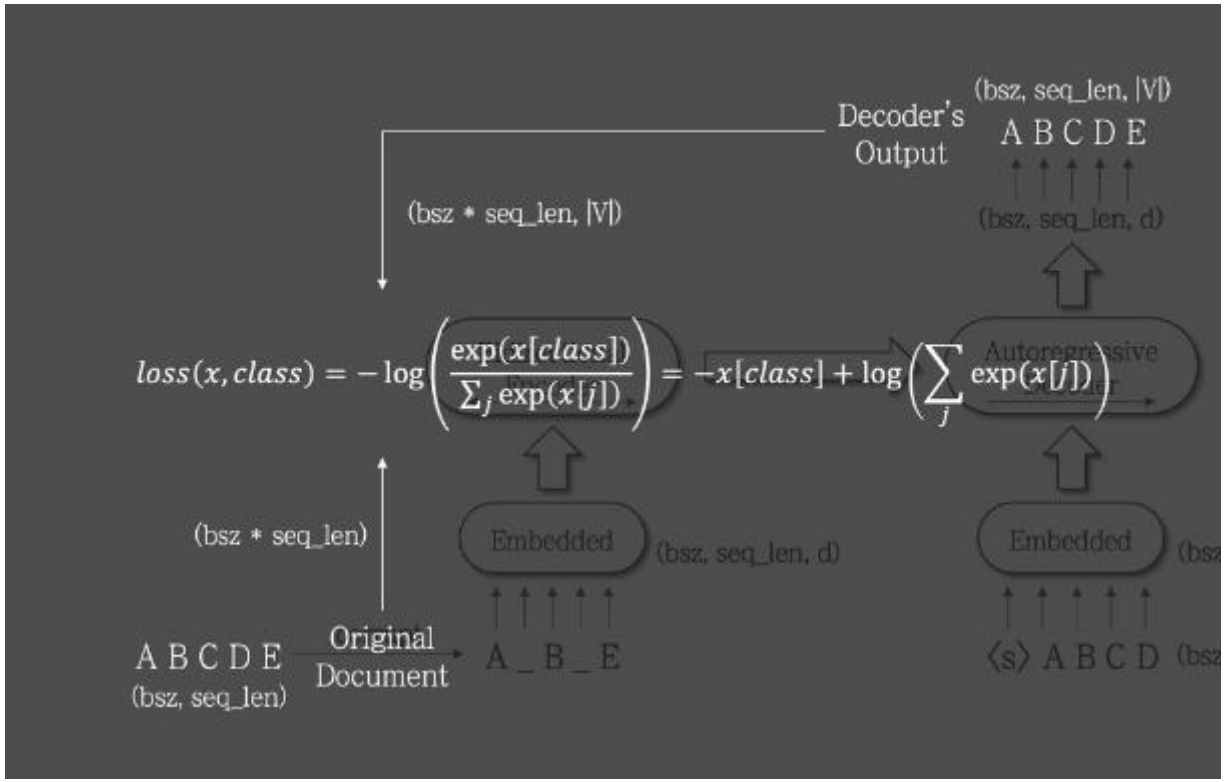
정해진 시간 안에 어떤 사건이 일어날 횟수에 대한 기댓값을 λ라고 했을 때, 그 사건이 k회 일어날 확률은 다음과 같다.

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

여기서 e는 자연상수이다.

BART 모델에서는 람다 초깃값을 3으로 설정

Task	설명
Sequence Classification	같은 input을 encoder와 decoder로 넣어주고, decoder의 최종 output을 multi-class linear classifier에 활용
Token Classification	완전한 Document를 Encoder와 Decoder에넣어주고, Decoder의 Output을 각 Token의 Representation으로 사용 (Classifier에사용됨)
Sequence Generation	BART는 Autoregressive decoder이므로 모델 그대로 fine-tuning 가능 (Question and Answering, Summarization)
Machine Translation	Encoder의 embedding layer를 new initialized source encoder로 대체하고 pre-trained BART 모델을 이용함 (1) BART 대부분 parameter는 freeze, 일부만 update (2) 적은 수의 iteration 동안 전체 parameter를 update


(bsz, seq_len, |V|)
Decoder's Output
A B C D E
(bsz * seq_len, |V|)
(bsz, seq_len, d)
$loss(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right) = -x[class] + \log\left(\sum_j \exp(x[j])\right)$
Autoregressive
(bsz * seq_len)
Embedded
(bsz, seq_len, d)
Embedded
A B C D E
Original Document
A _ B _ E
⟨s⟩ A B C D
(bsz, seq_len)