Paper review

Multilingual Denoising Pre-training for Neural Machine Translation

Yinhan Liu[‡*], Jiatao Gu[†*], Naman Goyal[†*], Xian Li[†], Sergey Edunov[†], Marjan Ghazvininejad[†], Mike Lewis[†], and Luke Zettlemoyer[‡]

[†]Facebook AI
[‡]Birch Technology
[†]{jgu,naman,xianl,edunov,ghazvini,mikelewis,lsz}@fb.com
[‡]yinhan@birch.ai

2021.07.15 정규헌

Index

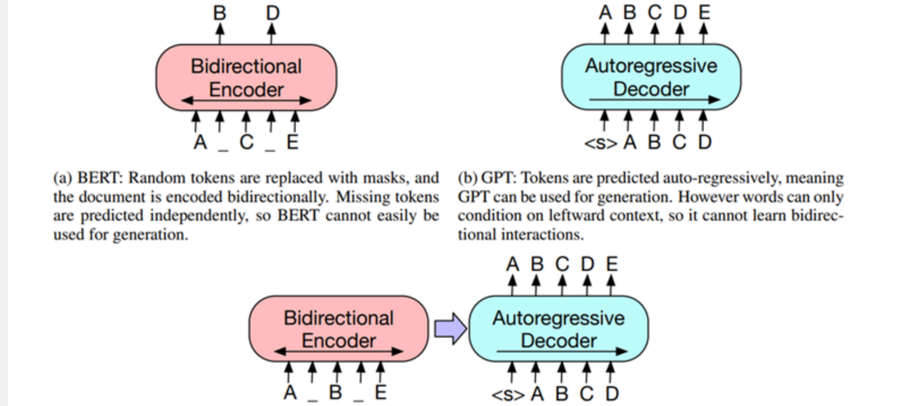
1.Introduction

BART

- masked Language Model(ex. BERT)
  - Bidirectional encoder
  - Generation task에서는 사용이 어려움
- Autoregressive decoder(ex. GPT)
  - Generation에서 사용
  - Bidirectional한 정보를 얻지 못함
- BART
  - 손상된 text를 bidirectional encoder로 인코딩
  - 정답 text에 대한 likelihood를 autoregressive decoder로 계산
  - Nosing 기법에 자유로움
  - Seq2seq Transformer 구조
  - ReLU activation 대신 GeLUs 사용

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



(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.

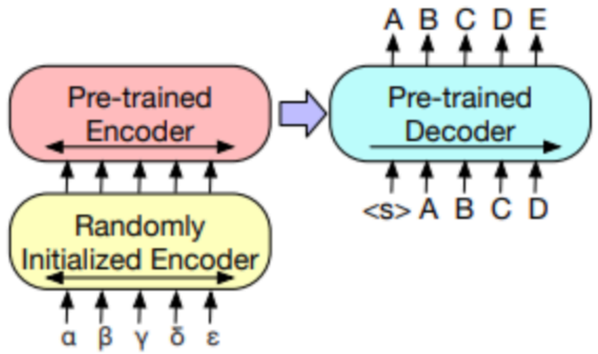
(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.

(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

1.Introduction

BART

- Pretraining BART
  - BART는 손상된 text를 복원해서 원본 text와 비교해 loss를 줄이도록 학습
  - 다양한 noising 기법들을 적용 가능 (Token masking, Permutation, Rotation, Deletion…)
- 번역을 위한 BART fine-tuning
  - 전체 BART 모델을 pretrained Decoder로 사용
  - 새로운 encoder를 추가해서 fine-tuning 함



For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

1.Introduction

mBART

- 다른 NLP task에 비해, 번역에서의 pretraining은 일반적인 관행이 아님
  - 기존 접근법은 모델을 부분적으로 pretraining하거나, 영어 말뭉치에만 집중
- 이 논문에서는 여러 언어로 전체 텍스트를 noise, 완전한 text로 재구성하는 것이 목표
  - Auto-regressive model을 pretraining
- Multilingual seq2seq denoising auto-encoder인 mBART를 소개
  - BART에 여러 언어의 대규모 단일 말뭉치를 학습시킴
  - mBART는 모든 언어에 대해 한번 훈련되어 task,language 별다른 수정이나 초기화 없이 supervised, unsupervised learning에서 모두 fine-tuning 가능
  - mBART는 언어 쌍을 통해 새로운 타입의 전송을 가능하게함
    - ex. 한국어 영어의 bi-text에 대한 fine-tuning은 다른 모든 언어에서 번역할 수 있는 모델을 만듬
    - 언어의 수와 전반적인 유사성을 포함하여 효과적인 pretrain에 가장 큰 영향을 미치는 요인에 대한 분석 제공

2. Multilingual Denoising pretraining
pretrain mBART

- Data : CC25 Corpus (25개국 말뭉치를 up/down sampling)
- Pre-processing
  - Tokenize : Sentence-piece
  - Subword token : 250000
  - Tokenize는 fine-tuning에 도움을 줌
- Model : mBART
  - Seq2seq
  - BART에서는 오직 English , mBART는 여러 언어들을 pretraining

2. Multilingual Denoising pretraining

pretrain mBART

- Architecture
  - Standard Seq2Seq Transformer with 12 layers of encoder & decoder
  - 1024 on 16heads
  - Additional Layer Normalization layer on top of encoder&decoder
- Noise Function
  - Remove spans of text and replace them with mask token
    - mask 35% of the words in each instance
  - Permute the order of sentences within each instance

2. Multilingual Denoising pretraining
pretrain mBART

- Instance Format
  - 각 인스턴스에 Language ID 토큰<LID>를 추가
  - Pre-training 여러 문장레벨에서도 가능하게 해서 document translation도 가능
- Pretrained Models
  - mBART25 : 25 language
  - mBART06 : six European language
  - mBART02 : 4 language
  - BART-EN/RO : en/ro only

2. Multilingual Denoising pretraining

Related work

- Multi lingual Pre-training for machine translation
  - XLM / XLM-R
    - 학습된 파라미터들을 번역모델의 인코더의 초기값으로 사용
    - mBART는 XLM과 달리 seq2seq설정으로 인코더와 디코더를 동시에 pretrain : 번역에 더욱 특화
  - MASS
    - mBART와 유사하게 단어 마스킹을 사용하는 seq2seq 기반 pretraining 기법
    - MASS 디코더는 인코더에서 마스킹된 토큰만 예측, mBART는 마스킹 외에도 여러 노이즈 function을 적용 가능
  - XLM과 MASS는 pretrained model이 두 언어에 대한 번역 성능을 향상 시킨다는 증거를 보여주지 않았음

3. Sentence-level Machine Translation

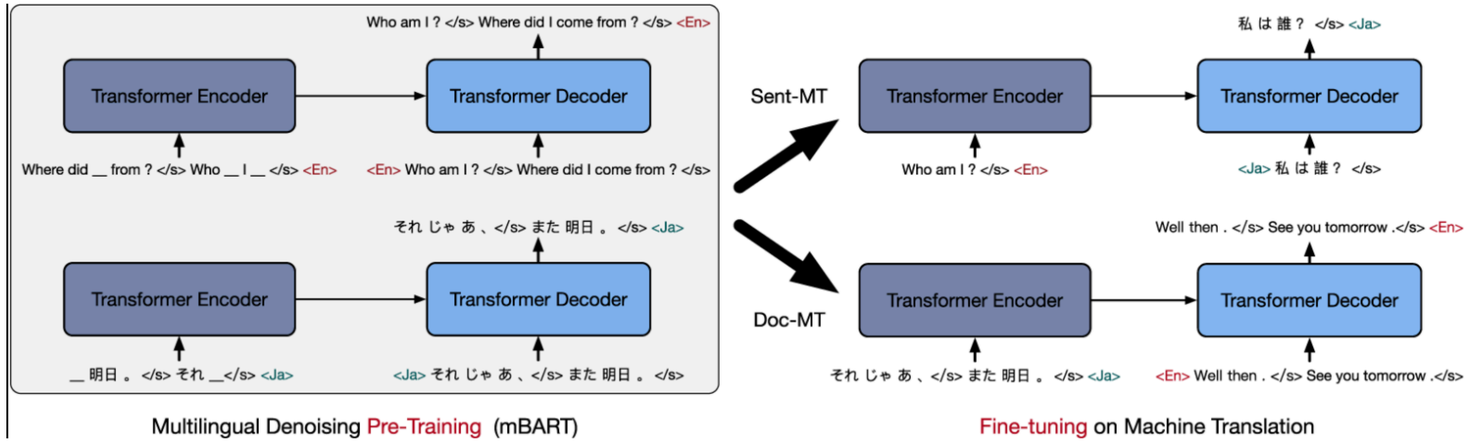
Sentence-level MT with pretrained mBART

- mBART pre-training
  - low to medium resource sentence-level MT에서 좋은 성능을 보임
  - Bi-text와 Back Translation도 함께 적용, other pre-training schemes보다 좋은 성능을 보임
  - Pre-training은 pretraining data에 존재하지 않는 언어의 번역에도 도움을 준다
- Dataset
  - 24 pairs of publicly available parallel corpora(CC25)
  - 대부분의 pairs : WMT, IWSLT에서 구함
  - 3가지로 분류 : low resource(1M), medium resource(1M< ,<10M), high resource(>10M)

3. Sentence-level Machine Translation

Sentence-level MT with pretrained mBART

- Fine tuning & Decoding
  - Single pair of bi-text data
  - 0.3 Dropout
  - 0.2 label smoothing
  - 2500 warm-up steps

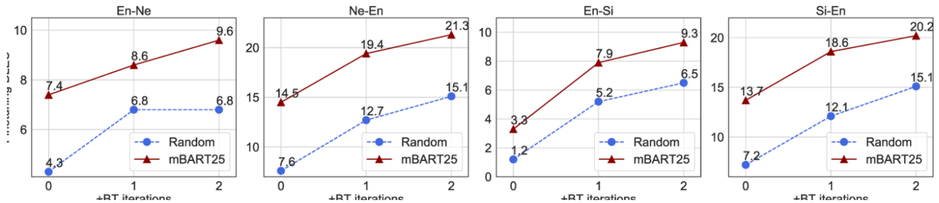


- Fine tuning 과정은 서로 다른 seed에서 안정적이 었음
- MT task에서는 sentence permutation과 word-span masking만 noising
- Special language id token은 인코더와 디코더에 넣어줌

3. Sentence-level Machine Translation

Sentence-level MT with pretrained mBART

- Results
  - Low-resource의 경우 unsupervised translation이 더 낫다
  - High resource의 경우 pretraining이 오히려 성능을 해칠 수 있다.
  - BT와 함께 pretraining하면 성능이 향상된다
  - mBART가 여러 모델들 중 가장 성능이 좋음
  - pretrain하지 않은 언어를 번역할 때도 도움이 됨



Comparison with other pre-training approaches on WMT16 Ro-En.

Pre-training		Fine-tuning		
Model	Data	En → Ro	Ro → En	+BT
RANDOM	None	34.3	34.0	36.8
XLM (2019)	En Ro	–	35.6	38.5
MASS (2019)	En Ro	–	–	39.1
BART (2019)	En	–	–	38.0
XLM-R (2019)	CC100	35.6	35.8	–
BART-EN	En	36.0	35.8	37.4
BART-RO	Ro	37.6	36.8	38.1
MBART02	En Ro	**38.5**	**38.5**	**39.9**
MBART25	CC25	37.7	37.8	38.8

4.Document-level Machine Translation

Document-level MT with mBART

- 문장이 아닌 여러문장의 doc-level로 MT 성능 측정
- Data : WMT-19, TED15
- 문장은 <\s>, 전체 인스턴스는 <LID>
- Fine-tuning & Decoding은 sentence level과 동일
- pretrain은 문서 수준의 번역 성능에서 매우 중요한 요소
  - 고품질 문서 수준 데이터를 수집하기 더 어렵기 때문
  - pretrain 했을때 모든 모델이 성능 더 우수

Table 6:

Document-level machine translation on En-De and Zh-En. (×) The randomly initialized Doc-MT model cannot produce translations aligned to the original sentences, so only document evaluation is possible.

(a) Sentence- and Document-level BLEU scores on En-De					(b) Document-level BLEU scores on Zh-En			
Model	Random		mBART25		Model	Random	mBART25	HAN (2018)
	s-BLEU	d-BLEU	s-BLEU	d-BLEU		d-BLEU	d-BLEU	d-BLEU
Sent-MT	34.5	35.9	36.4	38.0	Sent-MT	22.0	28.4	–
Doc-MT	×	7.7	**37.1**	**38.5**	Doc-MT	3.2	**29.6**	24.0

5. Conclusion

- Multilingual denoising pretraining으로 문서 및 문장 수준의 번역 task에서 성능을 크게 향상 시킴
- 감독 및 비감독 번역의 성능 또한 향상 시킴
- pretraining이 가장 효과적인 시기와 방법을 분석
- BT와 같은 다른 접근 방식과 결합 가능
- 향후 작업에서 더 많은 mBART100 모델과 같이 확장할 것