Project 3
Neural Machine Translation

Team 5
박석훈, 고명준, 손기훈, 서경원

목차

1. 프로젝트 개요

[프로젝트 목표]
1. 각 영어문장을 한글로 번역

[성과달성 지표]
1. 두 테스트 셋에 대한 조화평균 점수

[개발 환경]
Google Colab Pro + (언어: Python)

2. 프로젝트 팀 구성 및 역할

훈련생	담당 업무
박석훈(팀장)	- 프로젝트 총괄, 결과 보고서 작성
고명준	- 코드 작성, 디버깅
서경원	- 논문 리뷰, 자료 조사
손기훈	- 디버깅, 자료 조사

* 담당 업무는 주요 담당 업무이며, 전원 각 업무에 참여

3. 프로젝트 진행 프로세스

구분	기간	활동	비고
팀 빌딩, 업무 분담	2/28(월)	착수 회의, 역할 분담	
베이스라인 코드 리뷰	3/1(화)~3/4(금)	베이스라인 코드 확인	각 역할 전원 참여
디버깅	3/1(화)~3/14(월)	파라미터 조정, 모델 변경	각 역할 전원 참여
결과 정리, 보고서 작성	3/14(월)~3/15(화)	결과 정리, 보고서 작성	각 역할 전원 참여

mBART(Multilingual Denoising Pre-training for Neural Machine Translation)

- 다국어 모델을 사용한 단일언어 Bart 객체 seq2seq 모델
- 인코더, 디코더, 텍스트 재구성 등에 집중한 다른 모델과는 달리, 전체 텍스트를 25개 국어로 denoising한 seq2seq 모델
- 한 번에 모든 언어에 대해 훈련하고, 언어 군에 대해 fine-tuning에 필요한 파라미터를 제공
- 인코더 12 layer, 디코더 12 layer. layer-normalization layer를 인코더와 디코더 위에 하나씩 포함

mBART(Multilingual Denoising Pre-training for Neural Machine Translation)

- 데이터로는 CC25 Corpus를 사용
- BART가 오직 영어로만 사전학습된 것에 비해, mBART는 다국어 학습을 진행
- 언어별 데이터 크기가 달라, 리밸런싱 과정을 거침

Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

Multilingual Denoising Pre-training for Neural Machine Translation

https://arxiv.org/abs/2001.08210

mBART(Multilingual Denoising Pre-training for Neural Machine Translation)

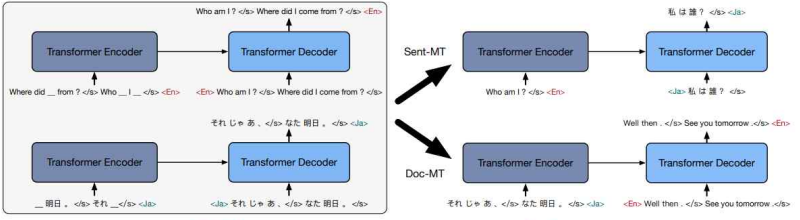

Figure 1: Framework for our Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained model is used for all tasks.

mBART(Multilingual Denoising Pre-training for Neural Machine Translation)

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

Table 4: **Comparison with Other Pre-training Approaches** on WMT16 Ro-En.

+ Back Translation Back-translation (BT, Sennrich et al., 2016b) is a standard approach to augment bi-text with target side monolingual data. We combine our pre-training with BT and test it on low resource language pairs – En-Si and En-Ne – using the FLoRes dataset (Guzmán et al., 2019). For a fair comparison, we use the same monolingual data as (Guzmán et al., 2019) to generate BT data. Figure 2 shows that initializing the model with our mBART25 pre-trained parameters improves BLEU scores at each iteration of back translation, resulting in new state-of-the-art results in all four translation directions.

4. 수행 결과

- epoch = 1
- batch_size = 4
- gradient_accumulation = 4
- learning_rate = 3e-5
- Score:0.19822