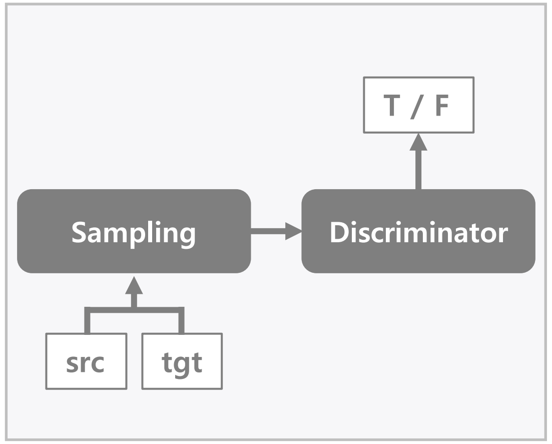
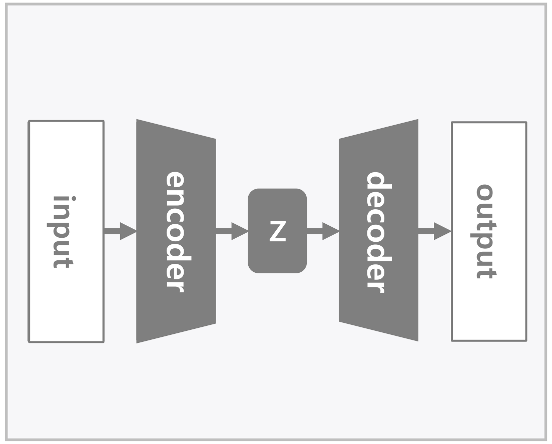
Unsupervised Machine Translation Using Monolingual Corpora Only

집현전 중급 14조 이원호 김택현 양수영

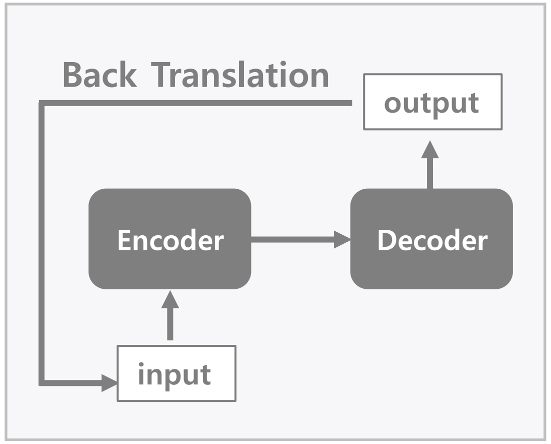
Before we start



Adversarial Learning

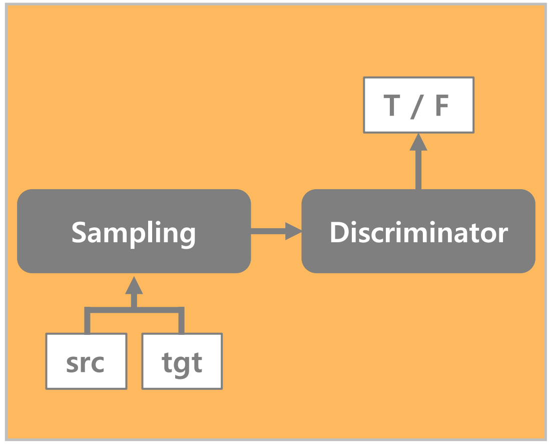


Auto Encoding

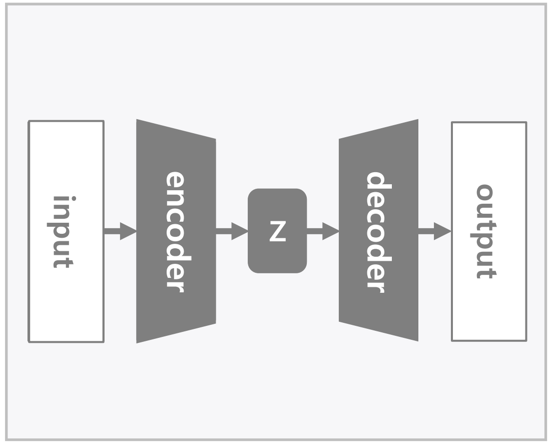


Back Translation

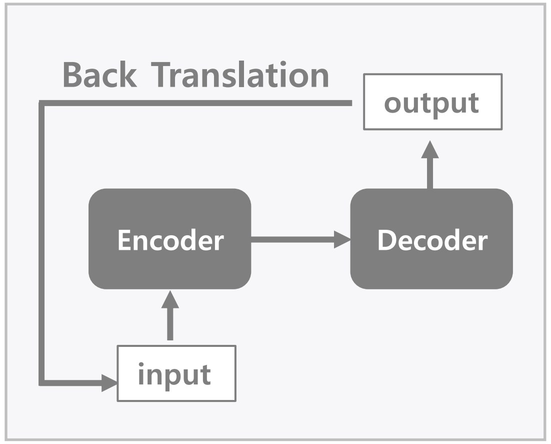
Before we start



Adversarial Learning



Auto Encoding



Back Translation

Before we start

Adversarial Attack

이미 훈련된 모델에 대해 입력 데이터를 조작해 잘못 예측하도록 함

방어책

- Adversarial Training

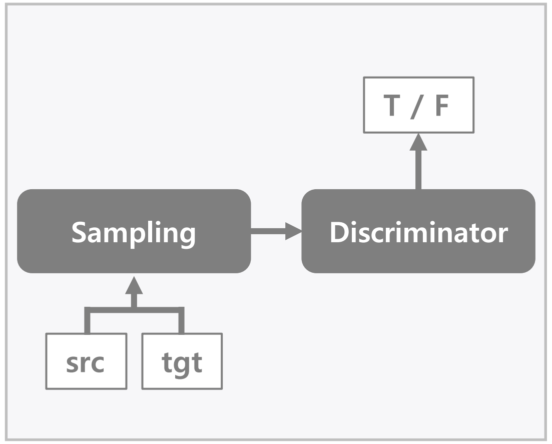
  입력 데이터와 조작된 데이터 간의 구분을 위한 새로운 신경망을 학습시켜

  모델이 임의로 조작된 데이터를 받지 않아도 이를 학습할 수 있는 방식

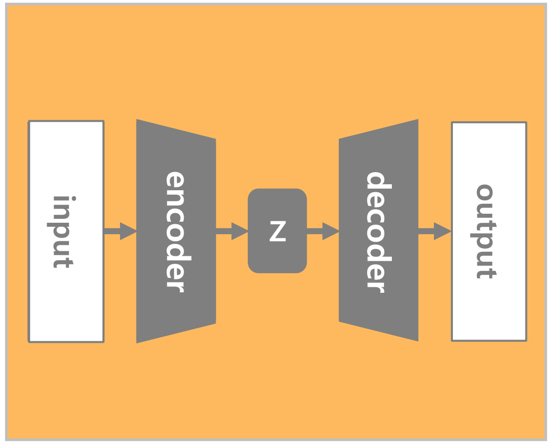
- Defensive Distillation

  조작된 데이터까지 모델에 학습시키는 방식

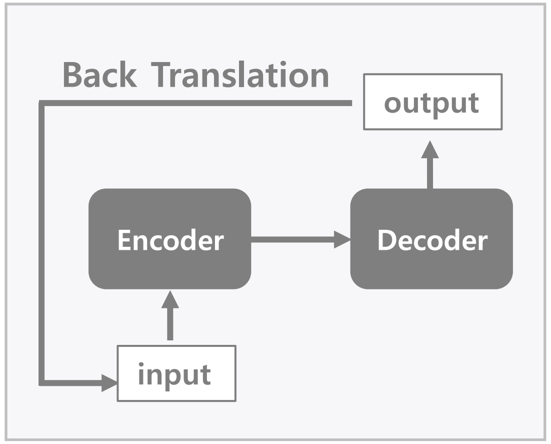
Before we start



Adversarial Learning



Auto Encoding

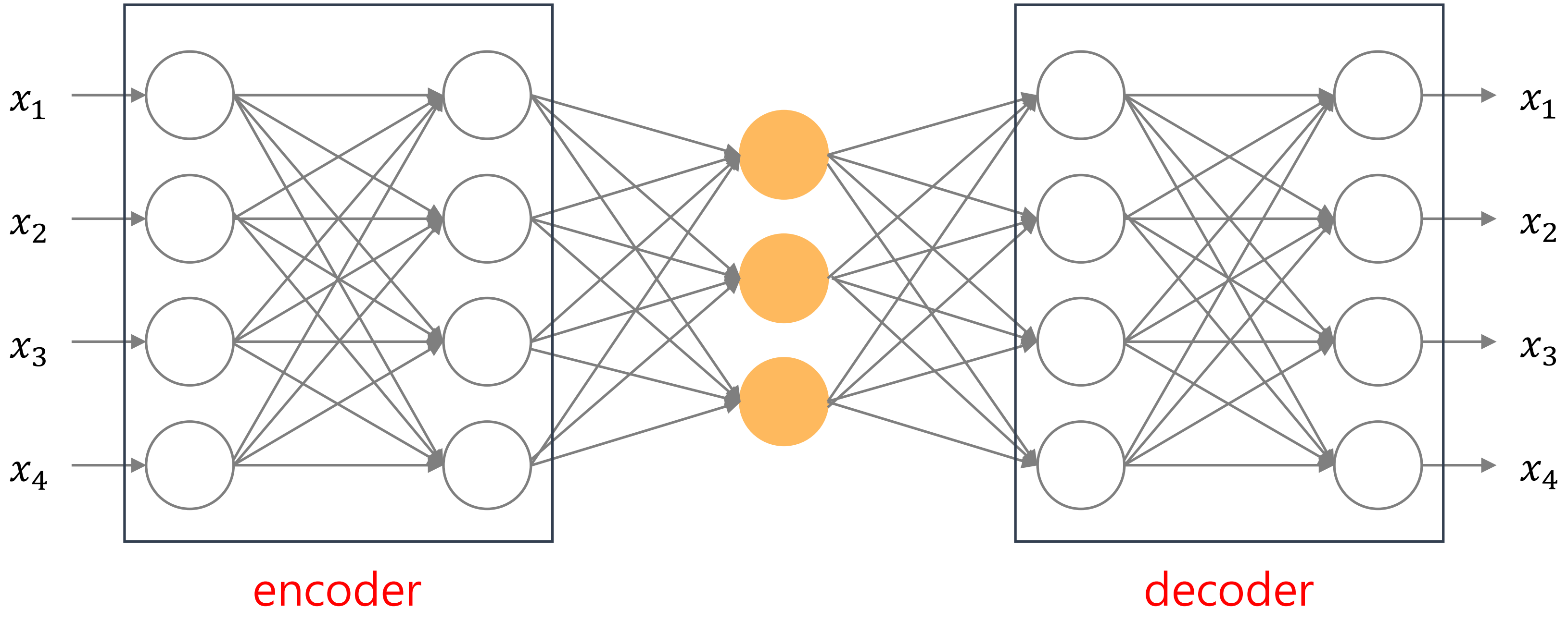


Back Translation

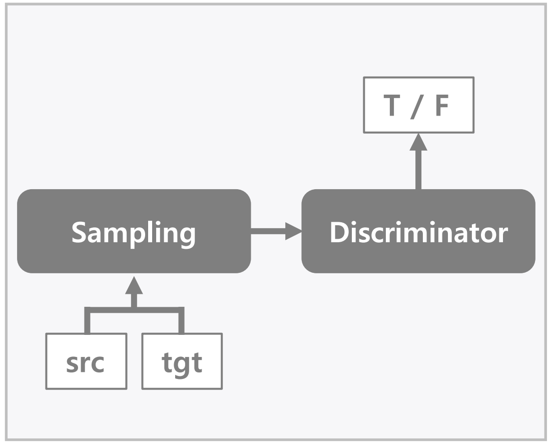
Before we start

Auto Encoder

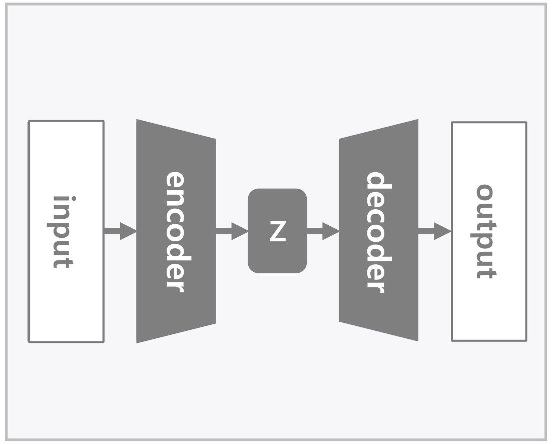
- 출력이 입력 데이터와 같아지도록 학습한 네트워크
- 차원 축소, noise 제거, 이상 데이터 검출, pre-train 등에 활용



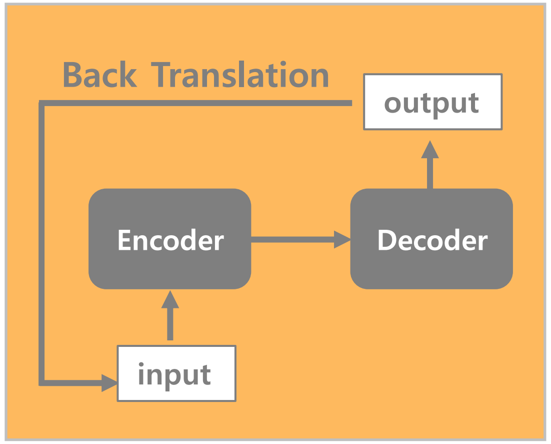
Before we start



Adversarial Learning



Auto Encoding



Back Translation

Before we start

Improving Neural Machine Translation Models with Monolingual Data

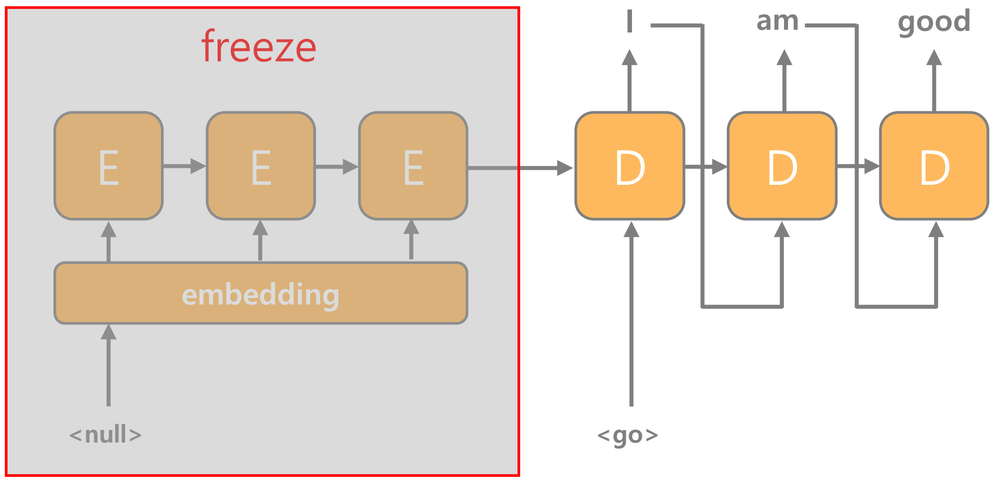
Rico Sennrich and **Barry Haddow** and **Alexandra Birch**
School of Informatics, University of Edinburgh
{rico.sennrich,a.birch}@ed.ac.uk,bhaddow@inf.ed.ac.uk

기존 Machine Translation Model과 달리 Monolingual Data로 학습이 가능한 방법을 제안

- Dummy Source Sentence
- Synthetic Source Sentence → **Back Translation**

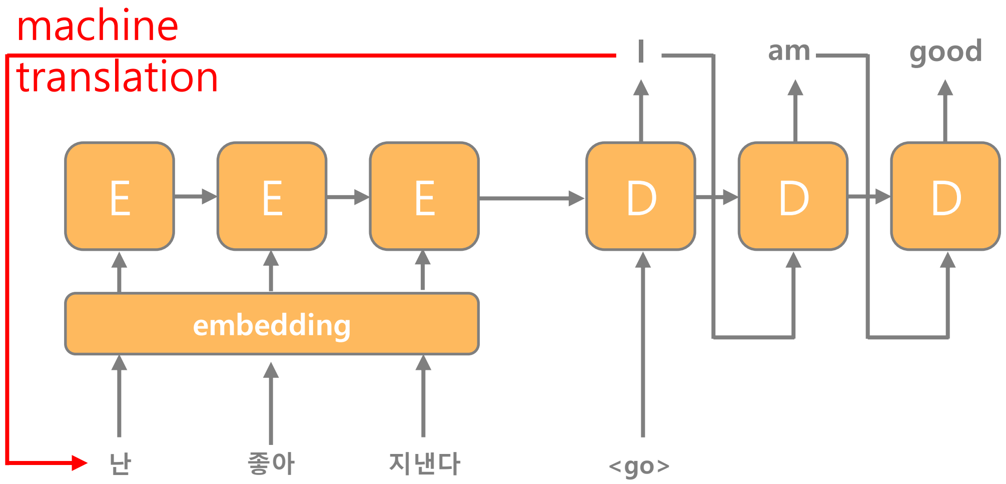
Before we start

Dummy Source Sentence



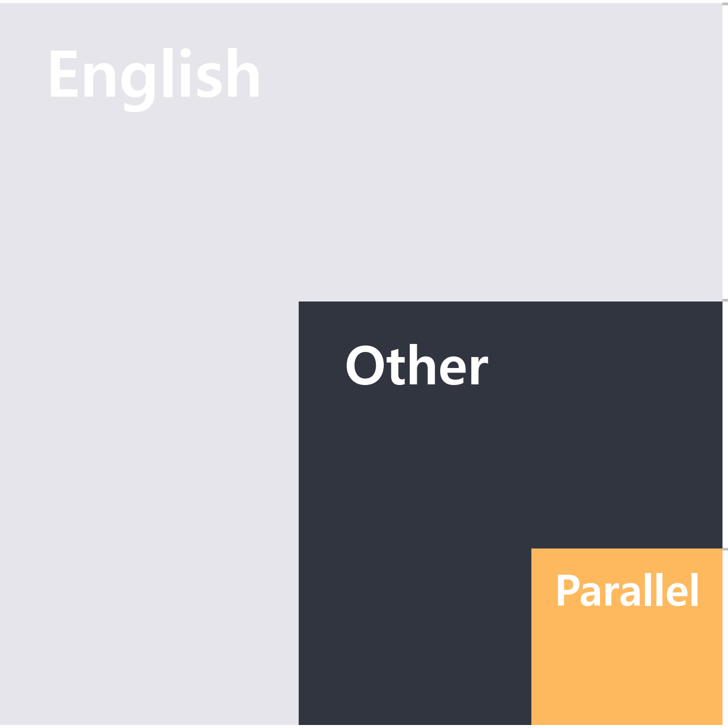
Encoder - dummy값, Decoder - target sentence가 되도록
한 뒤 Encoder의 parameter를 갱신되지 않도록 하며 학습

Synthetic Source Sentenece



Encoder에 인공적으로 만든 source sentence를 넣어 학습
인공데이터 생성 과정 = Back translation

Unsupervised Translation



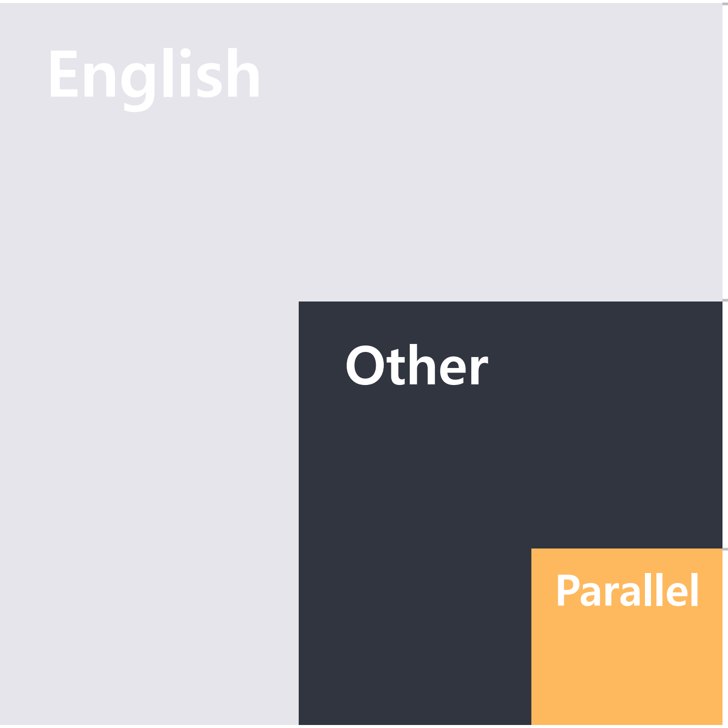
ENGLISH CORPUS

OTHER CORPUS

PARALLEL SENTENCE

- 구축에 전문 지식 필요
- 구축에 많은 비용 소요

Unsupervised Translation



ENGLISH CORPUS

OTHER CORPUS

- low resource 언어는 parallel corpus가 없음
- 단일 언어로 이루어진 corpus는 있음

PARALLEL SENTENCE

Method - Overview

- D_{src} : a dataset of sentences in the source domain
- D_{tgt} : another dataset in the target domain

- Naïve한 unsupervised translation model M
  → word-by-word translation

- 문장을 drop, swap하여 noise를 만듦
- D_{src}, D_{tgt}의 latent distribution을 align하기 위해 adversarial setting사용

- Noise가 있는 문장을 encoder, decoder로 번역
- Reconstruct 와 translation을 측정하는 object function이 최소화되도록 학습

- 이렇게 학습된 encoder와 decoder를 다음 iteration에서 사용

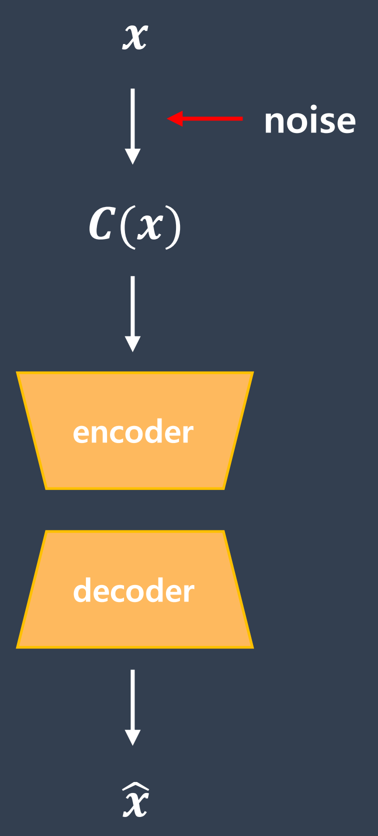
Method - Detail

Denoising Auto-Encoding

- D_{src}, D_{tgt} 간에 latent space를 만들기 위한 단계
- Source sentence x 에 noise를 주어 $C(x)$로 만듦
- $C(x)$를 encoder, decoder 통과 시켜 $\hat{x}$으로 만듦
- x와 $\hat{x}$을 latent space에 투영하여 거리를 가깝게 만듦

Object function

$$L_{auto}(\theta_{enc}, \theta_{dec}, Z, l) = E_{x\sim D_l, \hat{x}\sim d(e(C(x),l,l)}[\Delta(\hat{x}, x)]$$



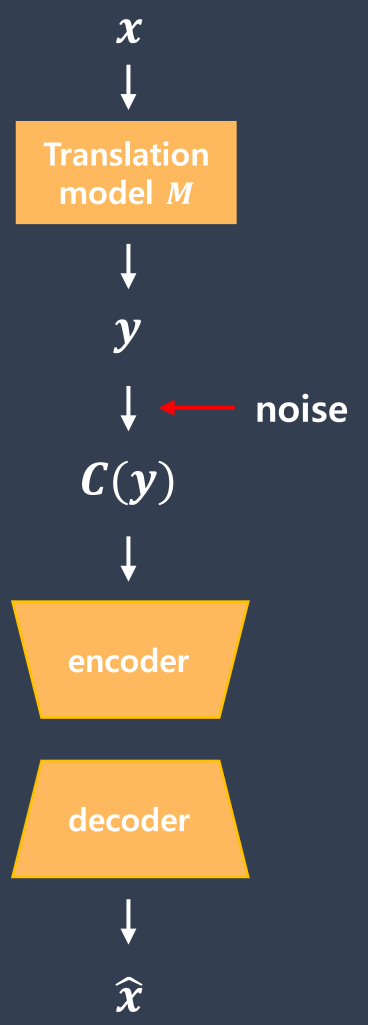
Method - Detail

Cross Domain Training

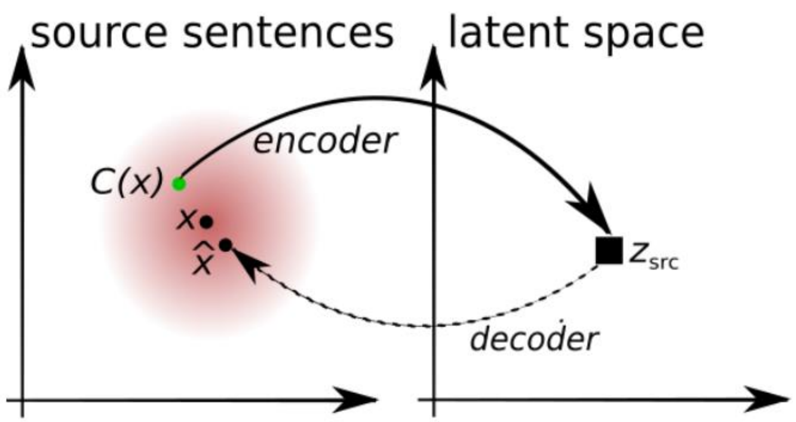
- D_{l_1}의 sentence를 D_{l_2}로 mapping하는 단계

Loss

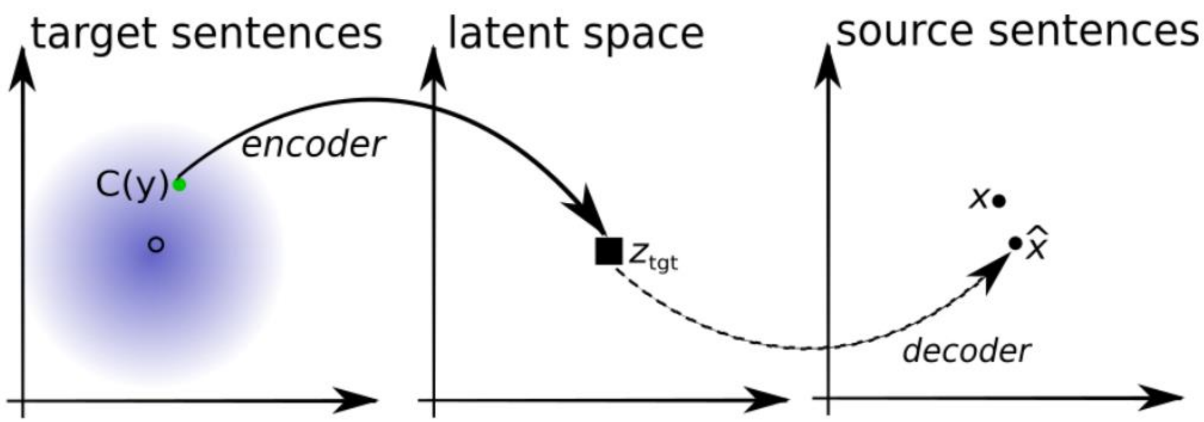
$$L_{cd}(\theta_{enc},\theta_{dec},Z,l_1,l_2) = E_{x\sim D_{l_1},\hat{x}\sim d(e(C(M(x)),l_1,)l_2)}[\Delta(\hat{x},x)]$$



Method - Detail



Denoising Auto Encoding



Cross Domain Training

Method - Detail

Adversarial Training

- Decoder에 encoder의 output과 비슷한 값을 input으로 줌
- Discriminator가 input문장이 D_{src}, D_{tgt} 중 어디에 속하는지 예측
- Encoder는 discriminator를 속이도록 학습

Objective Function

$$L(\theta_{enc}, \theta_{dec}, Z) = \lambda_{auto}[L_{auto}(\theta_{enc}, \theta_{dec}, Z, src) + L_{auto}(\theta_{enc}, \theta_{dec}, Z, tgt)] +$$

$$\lambda_{cd}[L_{cd}(\theta_{enc}, \theta_{dec}, Z, src, tgt) + L_{cd}(\theta_{enc}, \theta_{dec}, Z, tgt, src)] +$$

$$\lambda_{adv} L_{adv}(\theta_{enc}, Z | \theta_D)$$

Training Strategy

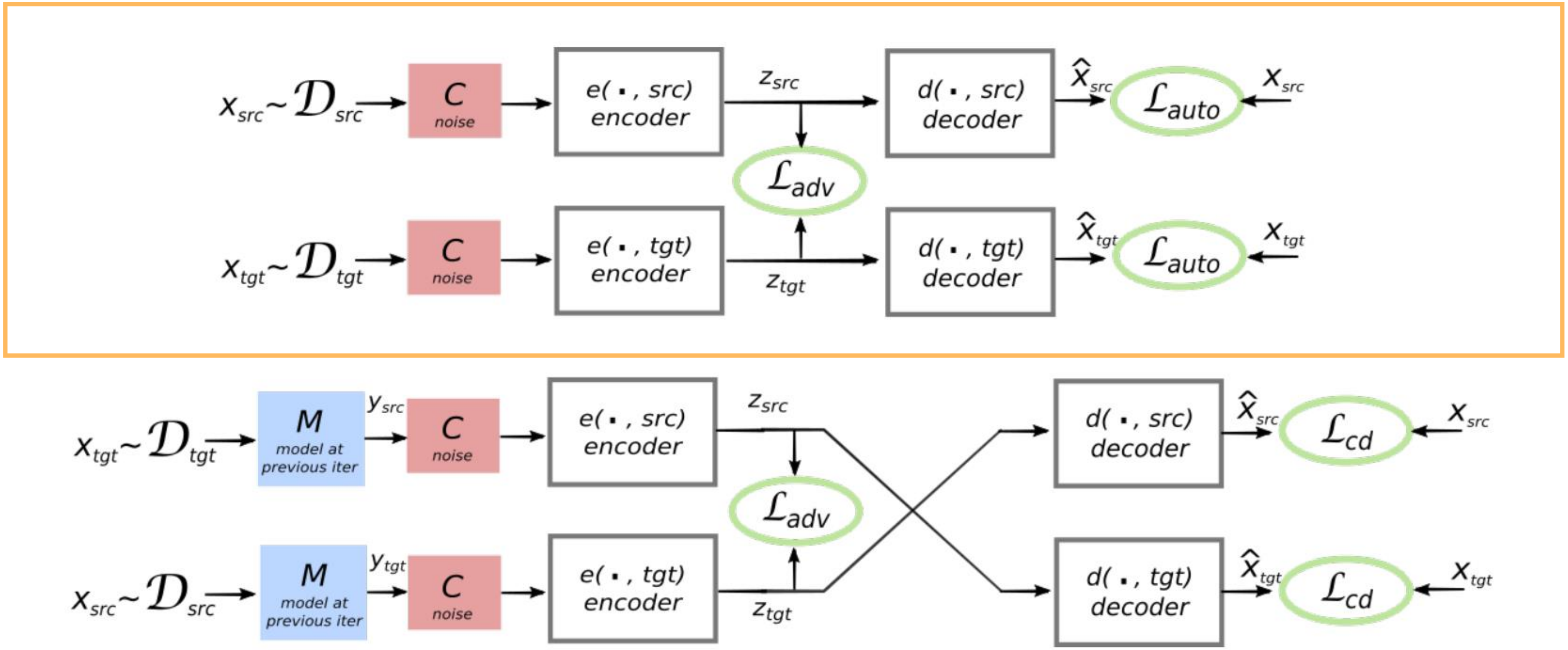
Interative Training

- Parallel word dictionary를 활용한 M의 번역이 input의 최소한의 정보를 담고 있다고 가정
- Noisy한 input이지만 denoising auto encoder이기에 latent feature space에 잘 mapping
- 이에 따라 decoder도 noiseless한 번역을 생성
- 같은 과정을 반복

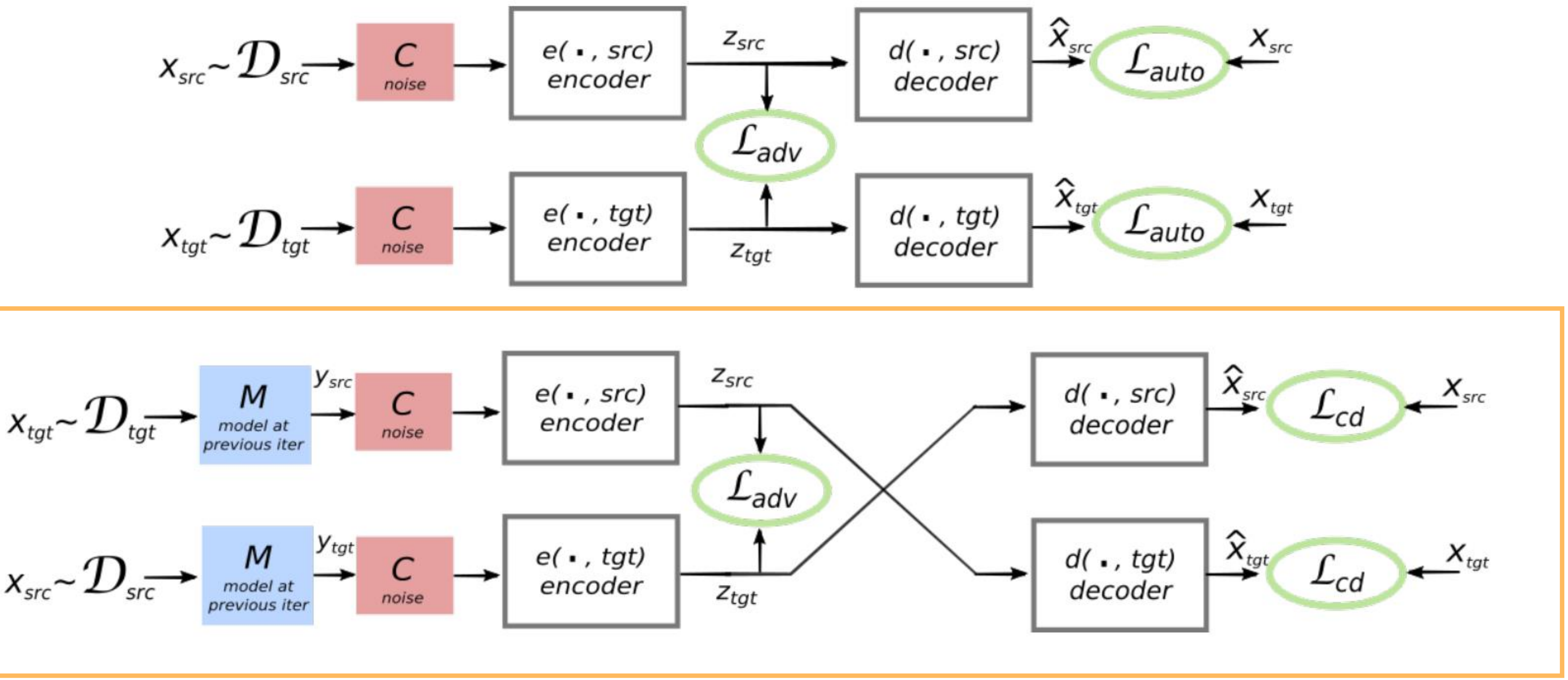
Algorithm 1 Unsupervised Training for Machine Translation

procedure TRAINING($\mathcal{D}_{src}, \mathcal{D}_{tgt}, T$)
  Infer bilingual dictionary using monolingual data (Conneau et al., 2017)
  $M^{(1)} \leftarrow$ unsupervised word-by-word translation model using the inferred dictionary
  for $t = 1, T$ **do**
    using $M^{(t)}$, translate each monolingual dataset
    // discriminator training & model training as in eq. 4
    $\theta_{\text{discr}} \leftarrow \arg\min \mathcal{L}_D, \quad \theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z} \leftarrow \arg\min \mathcal{L}$
    $M^{(t+1)} \leftarrow e^{(t)} \circ d^{(t)}$ // update MT model
  end for
  return $M^{(T+1)}$
end procedure

Training Strategy



Training Strategy



Training Strategy

Criterion

- Parallel dataset이 아니기 때문에 번역의 품질 평가 어려움
- 따라서 input을 2step 번역을 통해 재구성하여 재구성한 문장과 input을 비교
- D_{src}, D_{tgt} 을 각각 비교하여 평균점수가 가장 높은 model 선택

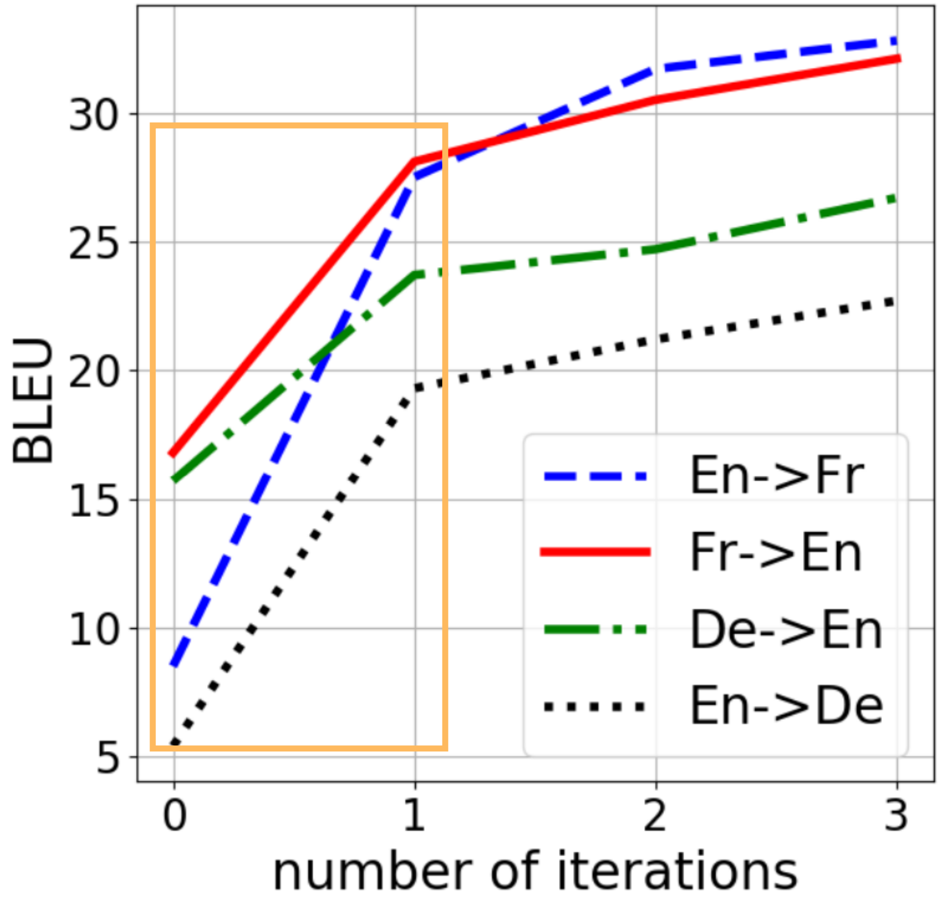
$$
\begin{aligned}
MS(e, d, \mathcal{D}_{src}, \mathcal{D}_{tgt}) \quad = \quad & \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{src}}\left[\mathrm{BLEU}(x, M_{src\to tgt}\circ M_{tgt\to src}(x))\right] + \\
& \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{tgt}}\left[\mathrm{BLEU}(x, M_{tgt\to src}\circ M_{src\to tgt}(x))\right]
\end{aligned}
$$

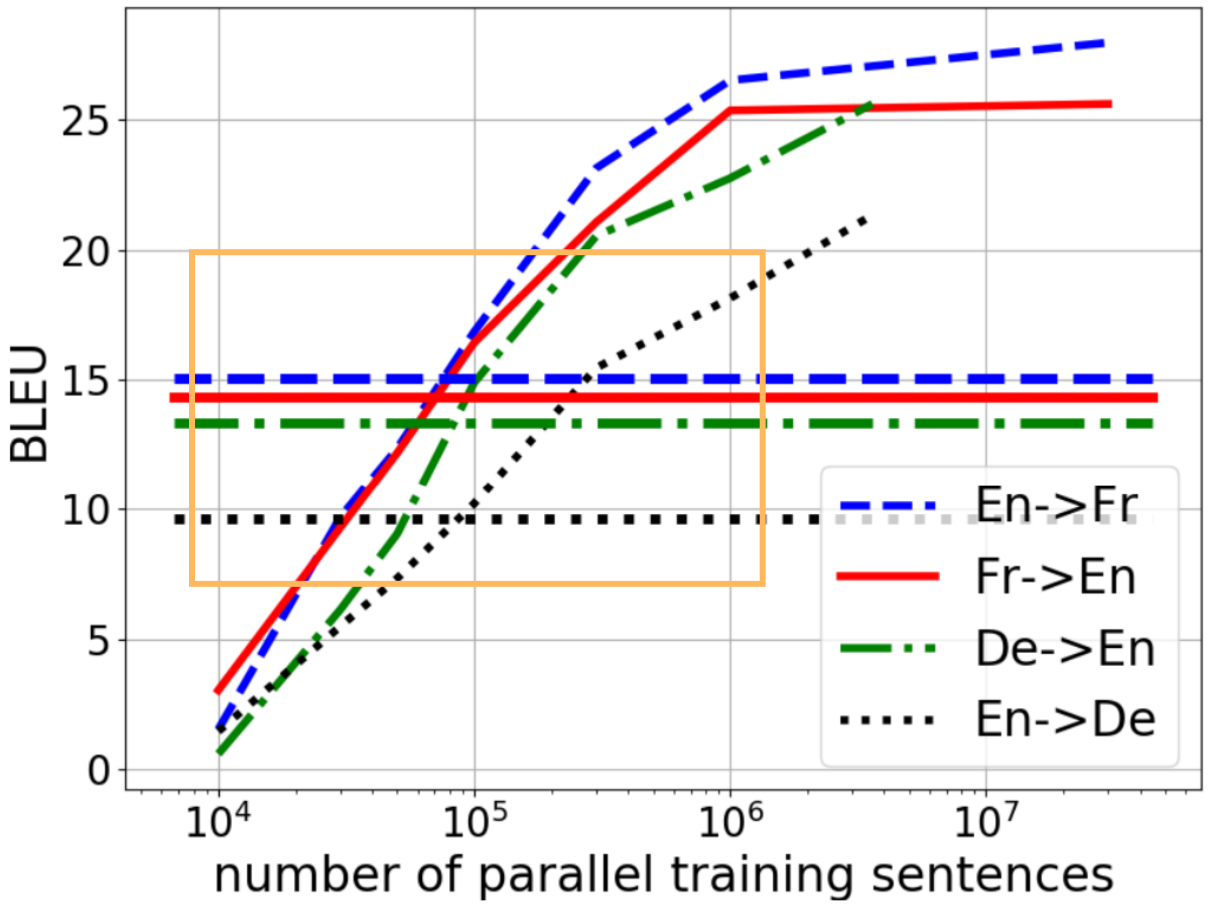
Experiment Results

	Multi30k-Task1				WMT			
	en-fr	fr-en	de-en	en-de	en-fr	fr-en	de-en	en-de
Supervised	56.83	50.77	38.38	35.16	27.97	26.13	25.61	21.33
word-by-word	8.54	16.77	15.72	5.39	6.28	10.09	10.77	7.06
word reordering	-	-	-	-	6.68	11.69	10.84	6.70
oracle word reordering	11.62	24.88	18.27	6.79	10.12	20.64	19.42	11.57
Our model: 1st iteration	27.48	28.07	23.69	19.32	12.10	11.79	11.10	8.86
Our model: 2nd iteration	31.72	30.49	24.73	21.16	14.42	13.49	13.25	9.75
Our model: 3rd iteration	32.76	32.07	26.26	22.74	15.05	14.31	13.33	9.64

Table 2: **BLEU score on the Multi30k-Task1 and WMT datasets** using greedy decoding.

Experiment Results





Ablation

- Word를 alignment하는 것
- Latent sentence representation의 분포(adversarial component)
- Input에 corruption을 주는 것

	en-fr	fr-en	de-en	en-de
$\lambda_{cd} = 0$	25.44	27.14	20.56	14.42
Without pretraining	25.29	26.10	21.44	17.23
Without pretraining, $\lambda_{cd} = 0$	8.78	9.15	7.52	6.24
Without noise, $C(x) = x$	16.76	16.85	16.85	14.61
$\lambda_{auto} = 0$	24.32	20.02	19.10	14.74
$\lambda_{adv} = 0$	24.12	22.74	19.87	15.13
Full	**27.48**	**28.07**	**23.69**	**19.32**

Table 4: **Ablation study on the Multi30k-Task1 dataset.**

Question and Answer

Thank You ☺