

XLM

Cross-lingual Language Model Pretraining

Guillaume Lample*

Facebook AI Research

Sorbonne Universités

glample@fb.com

Alexis Conneau*

Facebook AI Research

Université Le Mans

aconneau@fb.com

[Advances in Neural Information Processing Systems 32
\(NeurIPS 2019\)](#)

2021.08.11

채혁주

목차

- 배경
- 핵심 아이디어 소개
- 모델 구조
- 결과
- 분석

배경

- Machine Translation 분야는 상대적으로 발전이 더디다.

배경지식 - dummy src sentence, BT

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`

accepted to ACL 2016; new section on effect of back-translation
quality

배경지식 – dummy src sentence, BT

- Dummy source sentence

- parallel data : mono data = 1 : 1
- mono data 에서는 encoder input 으로 <null> 토큰을 다 넣어줌
- encoder, attention parameter freeze 시키고 decoder 만 학습
- 얻는 성능 : trg 언어에 대한 fluency
- 문제점 :
 1. mono data의 비율을 늘릴 수가 없음
 2. 학습을 진행하고 난 뒤에 mono data로 fine tuning 불가능이유 : mono data비율이 지나치게 높으면 src data 에 대한 conditioning을 안하기 때문

배경지식 – dummy src sentence, BT

- Back Translation

- 두개의 mono data가 있음
- 각 각 기존의 번역 모델을 이용해서 반대의 언어로 번역
- mono_data : [언어1, 언어2] => synthetic_src : [언어2, 언어1]
- synthetic_src – mono_data pair를 이용하여 번역 모델 학습

name	BLEU	
	2014	2015
PBSMT (Haddow et al., 2015)	28.8	29.3
NMT (Gülçehre et al., 2015)	23.6	-
+shallow fusion	23.7	-
+deep fusion	24.0	-
parallel	25.9	26.7
+synthetic	29.5	30.4
+synthetic (ensemble of 4)	30.8	31.6

Table 5: German→English translation performance (BLEU) on WMT training/test sets (newstest2014; newstest2015).

배경지식 – Unsupervised MT

UNSUPERVISED MACHINE TRANSLATION USING MONOLINGUAL CORPORA ONLY

Guillaume Lample ^{† ‡}, **Alexis Conneau** [†], **Ludovic Denoyer** [‡], **Marc'Aurelio Ranzato** [†]

[†] Facebook AI Research,

[‡] Sorbonne Universités, UPMC Univ Paris 06, LIP6 UMR 7606, CNRS

`{gl, aconneau, ranzato}@fb.com, ludovic.denoyer@lip6.fr`

Published as a conference paper at ICLR 2018

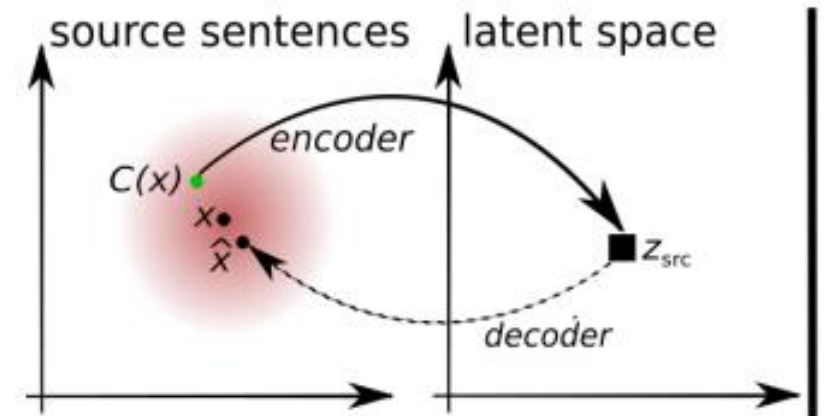
배경지식 – Unsupervised MT

- Denoising Auto-Encoding
- Cross Domain Training
- Adversarial Training

배경지식 – Unsupervised MT

- Denoising Auto-Encoding

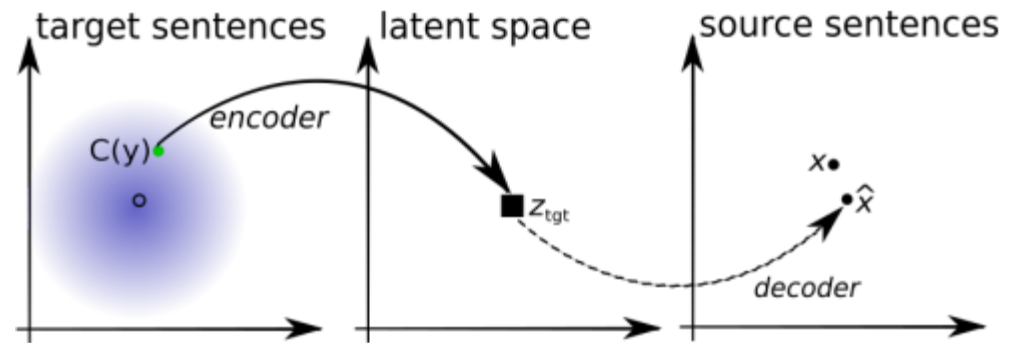
- 그냥 Auto Encoder만 사용하면 유용한 데이터의 의미 학습 없이 기계적으로 외워버림
- 랜덤 seq도 그냥 복사함
- Vincent et al., 2018 의 DAE를 적용



배경지식 – Unsupervised MT

- Cross Domain Training
 - 언어 : I_1, I_2
 - I_1 Corpus 에서 x 를 sampling
 - 현재의 I_1 to I_2 번역모델 M 을 이용하여 번역 $\Rightarrow M(x)$
 - corrupted version $C(y)$ sampling
 - $C(y)$ 를 이용해 x 를 복원하는 과정에서 학습

$$\mathcal{L}_{cd}(\theta_{\text{enc}}, \theta_{\text{dec}}, Z, \ell_1, \ell_2) = \mathbb{E}_{x \sim \mathcal{D}_{\ell_1}, \hat{x} \sim d(e(C(M(x)), \ell_2), \ell_1)} [\Delta(\hat{x}, x)]$$



배경지식 – Unsupervised MT

- Adversarial Training

- encoder-decoder MT모델의 decoder는 잘 훈련된 encoder output이 있거나, 그와 유사한 값이 있어야함
- encoder가 같은 space 에 존재하는 feature output 내놓으라고 함
- trg language 인지 무슨 언어인지는 상관없음
- Discriminator가 어떤 언어인지 classify

$$\mathcal{L}_{cd}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, \ell_1, \ell_2) = \mathbb{E}_{x \sim \mathcal{D}_{\ell_1}, \hat{x} \sim d(e(C(M(x))), \ell_2), \ell_1)} [\Delta(\hat{x}, x)]$$

$$\mathcal{L}_{adv}(\theta_{\text{enc}}, \mathcal{Z} | \theta_D) = -\mathbb{E}_{(x_i, \ell_i)} [\log p_D(\ell_j | e(x_i, \ell_i))]$$

$$\begin{aligned} \mathcal{L}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}) = & \lambda_{\text{auto}} [\mathcal{L}_{\text{auto}}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, \text{src}) + \mathcal{L}_{\text{auto}}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, \text{tgt})] + \\ & \lambda_{cd} [\mathcal{L}_{cd}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, \text{src}, \text{tgt}) + \mathcal{L}_{cd}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, \text{tgt}, \text{src})] + \\ & \lambda_{adv} \mathcal{L}_{adv}(\theta_{\text{enc}}, \mathcal{Z} | \theta_D) \end{aligned}$$

배경지식 – Unsupervised MT / sum up

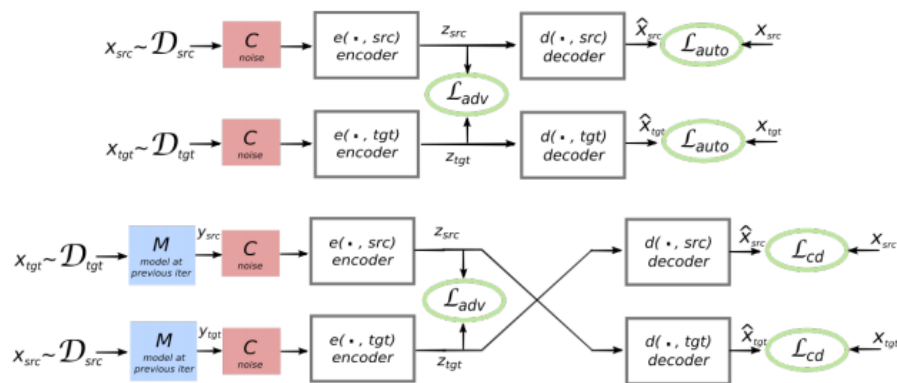


Figure 2: Illustration of the proposed architecture and training objectives. The architecture is a sequence to sequence model, with both encoder and decoder operating on two languages depending on an input language identifier that swaps lookup tables. Top (auto-encoding): the model learns to denoise sentences in each domain. Bottom (translation): like before, except that we encode from another language, using as input the translation produced by the model at the previous iteration (light blue box). The green ellipses indicate terms in the loss function.

Algorithm 1 Unsupervised Training for Machine Translation

```

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

I like studying ML

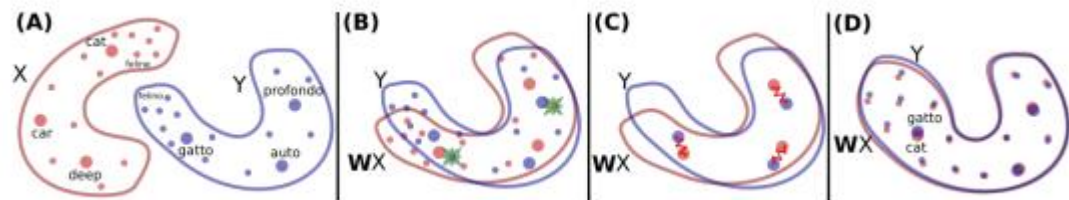


나 좋아한다 공부하기 머신러닝

ICLR 2018

WORD TRANSLATION WITHOUT PARALLEL DATA

Alexis Conneau^{*†‡}, Guillaume Lample^{*†§},
 Marc'Aurelio Ranzato[†], Ludovic Denoyer[§], Hervé Jégou[†]
 {aconneau, glample, ranzato, rvj}@fb.com
 ludovic.denoyer@upmc.fr



배경지식 – Unsupervised MT / results

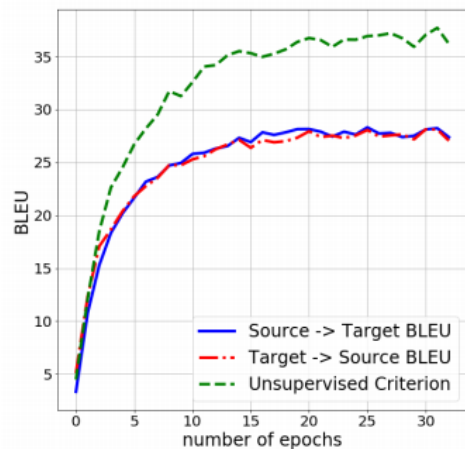


Figure 3: **Unsupervised model selection.** BLEU score of the source to target and target to source models on the Multi30k-Task1 English-French dataset as a function of the number of passes through the dataset at iteration $(t) = 1$ of the algorithm (training $M(2)$ given $M(1)$). BLEU correlates very well with the proposed model selection criterion, see Equation 5.

	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.**

배경지식

Phrase-Based & Neural Unsupervised Machine Translation

Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]

Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott

Facebook AI Research
myleott@fb.com

Alexis Conneau

Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]

Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato

Facebook AI Research
ranzato@fb.com

EMNLP 2018

배경지식

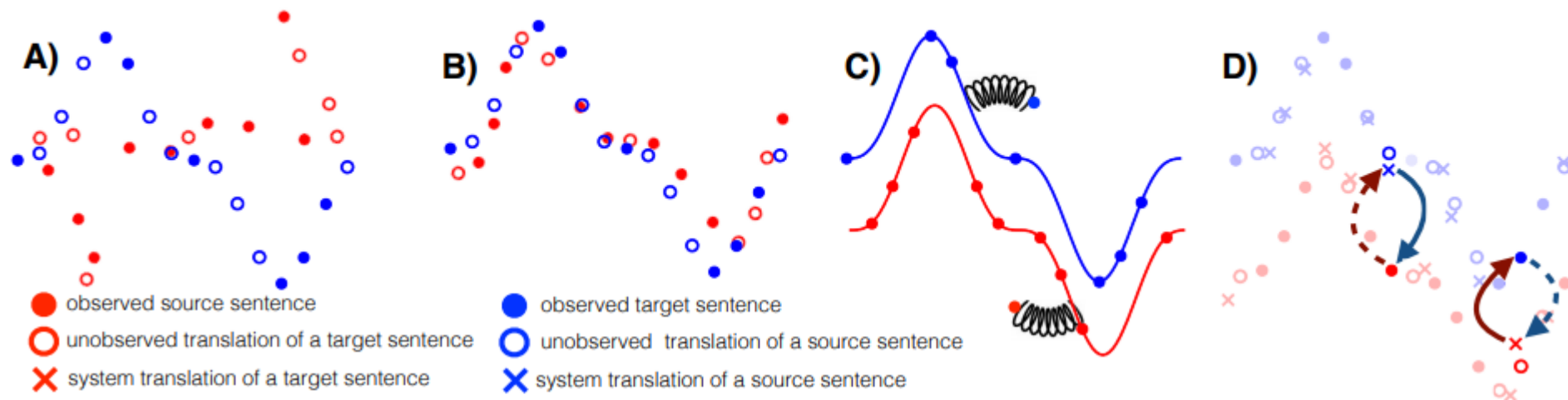


Figure 1: Toy illustration of the three principles of unsupervised MT. **A)** There are two monolingual datasets. Markers correspond to sentences (see legend for details). **B)** First principle: **Initialization**. The two distributions are roughly aligned, e.g. by performing word-by-word translation with an inferred bilingual dictionary. **C)** Second principle: **Language modeling**. A language model is learned independently in each domain to infer the structure in the data (underlying continuous curve); it acts as a data-driven prior to denoise/correct sentences (illustrated by the spring pulling a sentence outside the manifold back in). **D)** Third principle: **Back-translation**. Starting from an observed source sentence (filled red circle) we use the current source \rightarrow target model to translate (dashed arrow), yielding a potentially incorrect translation (blue cross near the empty circle). Starting from this (back) translation, we use the target \rightarrow source model (continuous arrow) to reconstruct the sentence in the original language. The discrepancy between the reconstruction and the initial sentence provides error signal to train the target \rightarrow source model parameters. The same procedure is applied in the opposite direction to train the source \rightarrow target model.

이제 XLM 시작!

아이디어

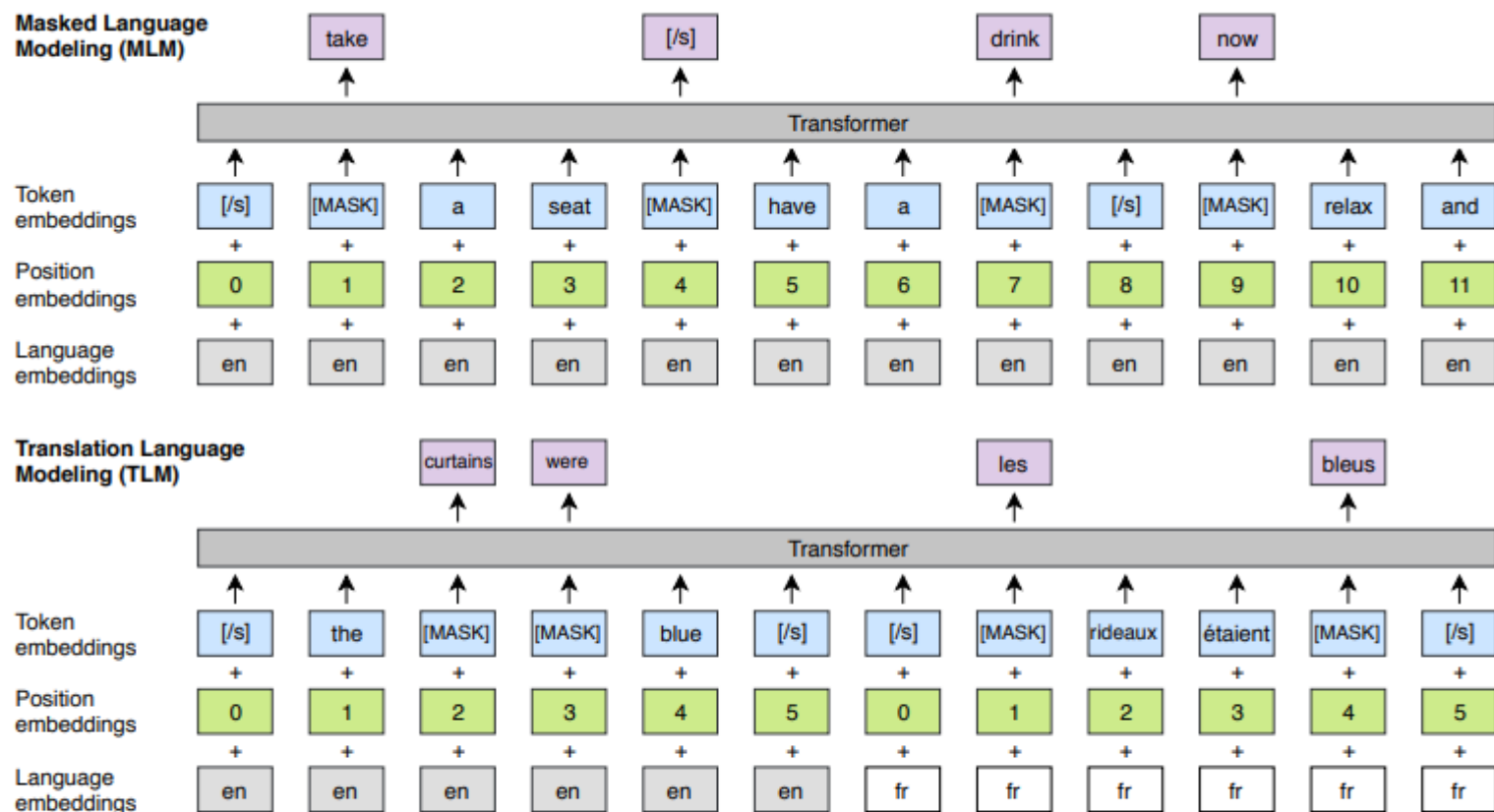
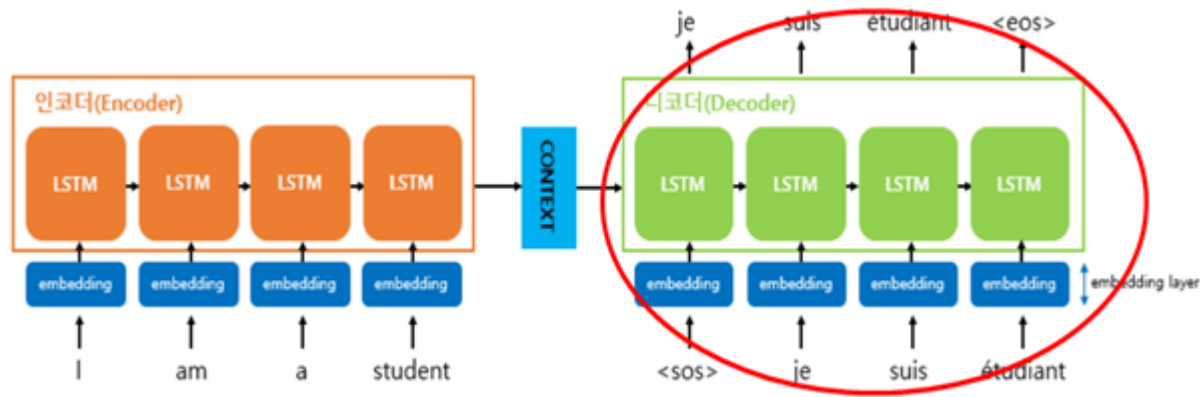
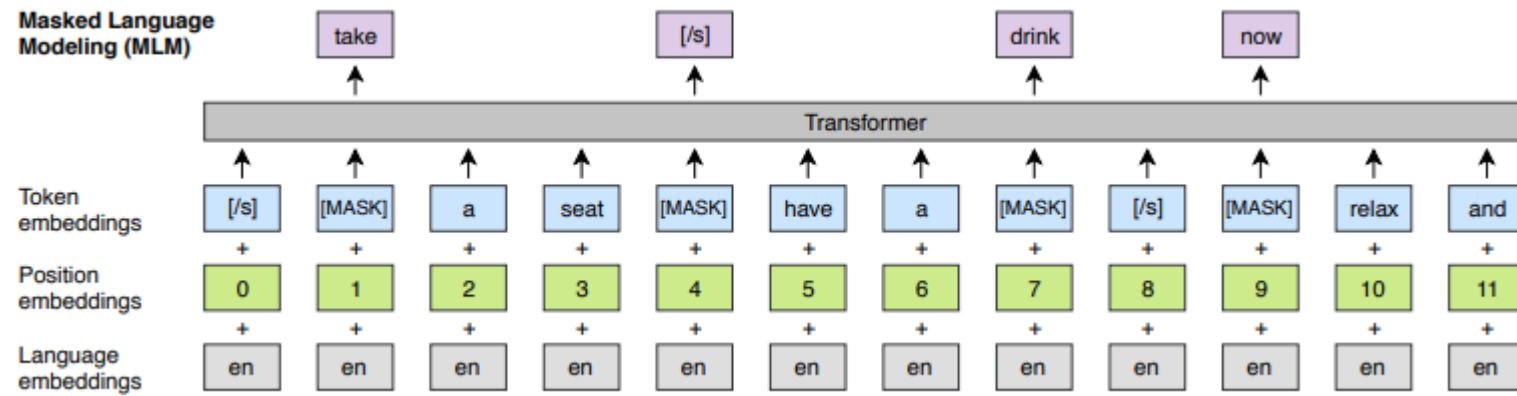


Figure 1: Cross-lingual language model pretraining. The MLM objective is similar to the one of [Devlin et al. \(2018\)](#), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

Causal Language Model



MLM



Cross-lingual language model

- pre train : CLM || MLM || MLM + TLM
- 장점
 - a better initialization of sentence encoders for zero-shot cross-lingual classification
 - a better initialization of supervised and unsupervised neural machine translation systems
 - language models for low-resource languages
 - unsupervised cross-lingual word embeddings

Cross-lingual classification

- XLM fine tune
- XNLI dataset 사용 (French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu, 14개 언어, 112.5k)
- pre-trained transformer의 첫번째 hidden state에 linear classifier 쌓아서 만듦

Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction

Unsupervised Machine Translation

- pretraining 이 매우 중요하다는 것을 이전 연구에서 밝힘
- cross-lingual word embedding
- UNMT = unsupervised neural machine translation
- WMT'14 en-fr, WMT'16 en-ge, WMT'16 en-ro 사용해서 test

Supervised Machine Translation

- WMT'16 ro-en 사용

Low-resource language modeling



- Wikipedia에는 네팔어 : 힌디어 = 1 : 6
- BPE vocab에서는 80%공유



English

↔

Nepali

My name is kevin × मेरो नाम केविन हो
Mērō nāma kēvina hō



 



English

↔

Hindi

My name is kevin × मेरा नाम केविन है
mera naam kevin hai

Unsupervised cross-lingual word embeddings

- MUSE = mono-lingual word embedding spaces with adversarial training
- shared vocab이 좋았다
- Concat = fastText를 이용하여 monolingual corpora를 바로 붙인 data에 대해 shared vocab 만든 것
- MUSE / Concat / XLM 성능 비교(얼마나 latent space 에서 vector가 유사한지)

Training details

- Transformer
- 1024 hidden unit, 8 heads, GELU, 0.1 dropout rate, learned positional embedding, adam optimizer learning rate : $10^{-4} \sim 5 \cdot 10^{-4}$
- TLM에서는 4000token으로 길이 고정
- LM : 64 volta GPUs
- MT : 8 volta GPUs

A low-angle shot of graduates in silhouette against a bright, overcast sky. They are celebrating, with many holding up their black graduation caps and white diplomas. The scene is filled with the silhouettes of arms and heads, creating a sense of a large, joyful crowd.

Results

Cross-lingual classification

- unsupervised MLM zero-shot cross-lingual classification SOTA 달성
- TLM 이용했을때 3.6% 성능 향상

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
<i>Machine translation baselines (TRANSLATE-TRAIN)</i>																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	<u>85.0</u>	<u>80.2</u>	<u>80.8</u>	<u>80.3</u>	<u>78.1</u>	<u>79.3</u>	<u>78.1</u>	<u>74.7</u>	<u>76.5</u>	<u>76.6</u>	<u>75.5</u>	<u>78.6</u>	<u>72.3</u>	<u>70.9</u>	63.2	<u>76.7</u>
<i>Machine translation baselines (TRANSLATE-TEST)</i>																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	<u>85.0</u>	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
<i>Evaluation of cross-lingual sentence encoders</i>																
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	<u>85.0</u>	<u>78.7</u>	<u>78.9</u>	<u>77.8</u>	<u>76.6</u>	<u>77.4</u>	<u>75.3</u>	<u>72.5</u>	<u>73.1</u>	<u>76.1</u>	<u>73.2</u>	<u>76.5</u>	<u>69.6</u>	<u>68.4</u>	<u>67.3</u>	<u>75.1</u>

Table 1: **Results on cross-lingual classification accuracy.** Test accuracy on the 15 XNLI languages. We report results for machine translation baselines and zero-shot classification approaches based on cross-lingual sentence encoders. XLM (MLM) corresponds to our unsupervised approach trained only on monolingual corpora, and XLM (MLM+TLM) corresponds to our supervised method that leverages both monolingual and parallel data through the TLM objective. Δ corresponds to the average accuracy.

Unsupervised machine translation

- Lample et al.(2018b) 와 비교
- Lample씨 논문에서 구현한 성능보다 우리가 직접 구현한 성능이 더 좋음 ㅋㅋㅋ GPU 빨이라고 함(더 큰 batch size)
- 모든 language pair 에서 SOTA 달성

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
<i>Previous state-of-the-art - Lample et al. (2018b)</i>							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
<i>Our results for different encoder and decoder initializations</i>							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	27.0	33.2	31.8	30.5
MLM	CLM	33.4	32.3	24.9	32.9	31.7	30.4
MLM	MLM	33.4	33.3	26.4	34.3	33.3	31.8

Table 2: **Results on unsupervised MT.** BLEU scores on WMT’14 English-French, WMT’16 German-English and WMT’16 Romanian-English. For our results, the first two columns indicate the model used to pretrain the encoder and the decoder. “ - ” means the model was randomly initialized. EMB corresponds to pretraining the lookup table with cross-lingual embeddings, CLM and MLM correspond to pretraining with models trained on the CLM or MLM objectives.

Supervised MT

- prev SOTA 인 Sennrich et al.(2016) 과 비교(이전 SOTA 가 3년전???)
- BT = back translation
- SOTA 달성

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro \rightarrow en	28.4	31.5	35.3
ro \leftrightarrow en	28.5	31.5	35.6
ro \leftrightarrow en + BT	34.4	37.0	38.5

Table 3: **Results on supervised MT.** BLEU scores on WMT'16 Romanian-English. The previous state-of-the-art of [Sennrich et al. \(2016\)](#) uses both back-translation and an ensemble model. ro \leftrightarrow en corresponds to models trained on both directions.

Low-resource language model

Training languages	Nepali perplexity
Nepali	157.2
Nepali + English	140.1
Nepali + Hindi	115.6
Nepali + English + Hindi	109.3

Table 4: **Results on language modeling.** Nepali perplexity when using additional data from a similar language (Hindi) or a distant one (English).

Unsupervised cross-lingual w Emb

	Cosine sim.	L2 dist.	SemEval'17
MUSE	0.38	5.13	0.65
Concat	0.36	4.89	0.52
XLM	0.55	2.64	0.69

Table 5: **Unsupervised cross-lingual word embeddings** Cosine similarity and L2 distance between source words and their translations. Pearson correlation on SemEval'17 cross-lingual word similarity task of [Camacho-Collados et al. \(2017\)](#).