XLM

Cross-lingual Language Model Pretraining

참여자 서덕진, 정환석, 최원석

Contents

01 Introduction

02 Cross-lingual language models

03 Experiments and results

문제점

2021년 은 집현전 에서 자연어 처리 공부 하는 해

→

2021년 은 [MASK] 에서 [MASK] 공부 하는 해

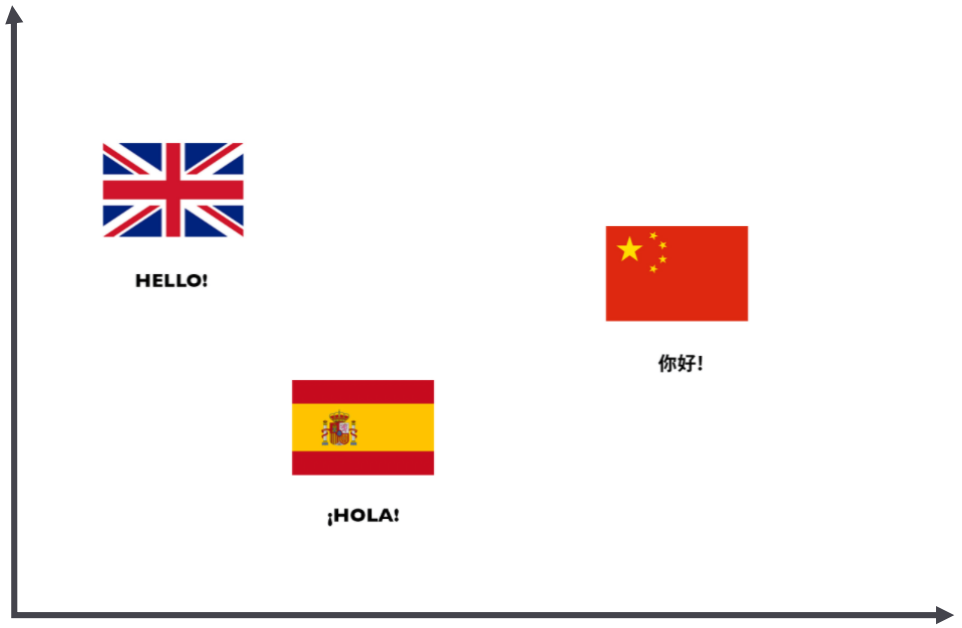
→

→

2021년 은 집현전 에서 자연어 처리 공부 하는 해

Bert 와 같은 기존의 pretraining model들이 좋은 성능이 보였지만, 대부분 영어를 중심으로 연구가 진행되었다.
이러한 문제를 해결하기 위해 다국어(multiple language)로 확장할 수 없을까?

해결책



이에 대한 해결책으로 여러 언어가 하나의 임베딩 공간을 공유(shared embedding space)하여,
어떤 문장이라도 해당 임베딩 공간에 인코딩할 수 있는 Universal cross-lingual encoder를 만들고자 하였다.

이 논문에서는...

여러 언어간 이해(XLU) 벤치마크 데이터셋에서 교차 언어모델의 사전 학습의 효과를 입증하고자 한다.
(We demonstrate the effectiveness of cross-lingual language model pretraining
on multiple cross-lingual understanding benchmarks)

Contribution 1

새로운 비지도 학습 방법 제시

Contribution 2

병렬데이터를 이용할 수 있는 지도 학습 방법 제시

Contribution 3

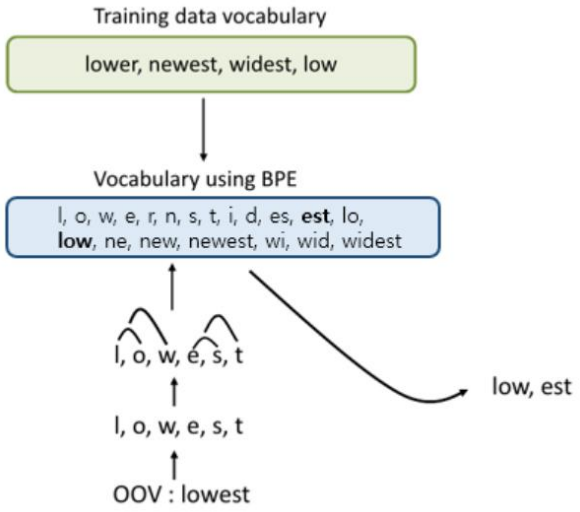
기계번역에서의 성능 평가

Contribution 4

Low resource 언어의 복잡도 개선

Shared sub-word vocabulary

Byte Pair Encoding(BPE)



$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}.$$

p_i : i번째 언어의 문장 개수 / 전체 언어 문장 개수의 합
q_i : i번째 언어의 샘플링할 확률
α : exponent subsampling factor (default : 0.5)

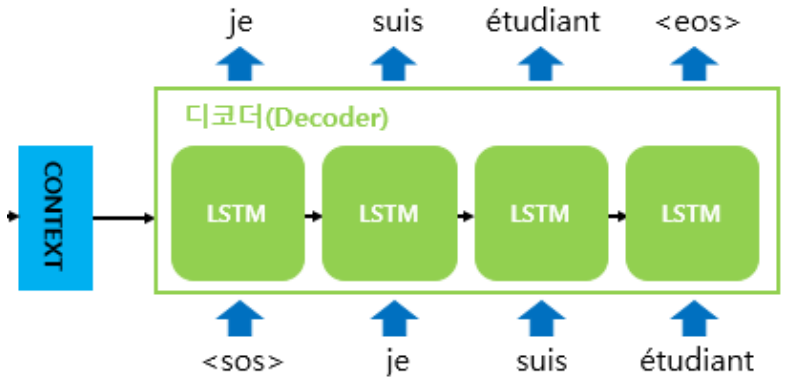
영어(e) : 3,600,000문장 프랑스어(f) : 10,000문장
일본어(j) : 200,000문장 한국어(k) : 713,000문장

p_e : 79.59	p_f : 0.22	p_j : 4.42	p_k : 15.76
q_e : 57.69	q_f : 3.04	q_j : 13.60	q_k : 25.67

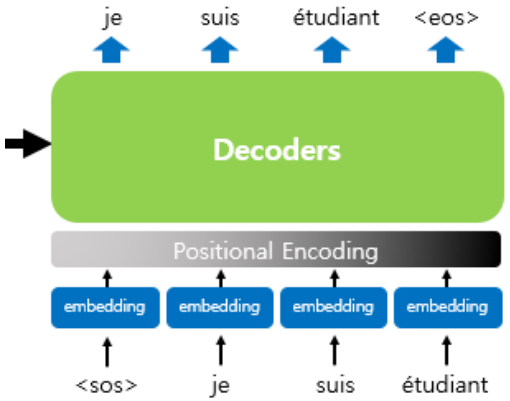
단일 말뭉치에서 무작위로 샘플링하여 BPE를 학습하며, 이 때 샘플링할 확률은 위 식에 근거한 다항분포를 따른다.
위와 같은 방법으로 샘플링 하면, low-resource와 관련된 token의 수가 증가하고, high-resource 언어와의 편향이 감소

1. Causal Language Modeling : CLM

LSTM Decoder



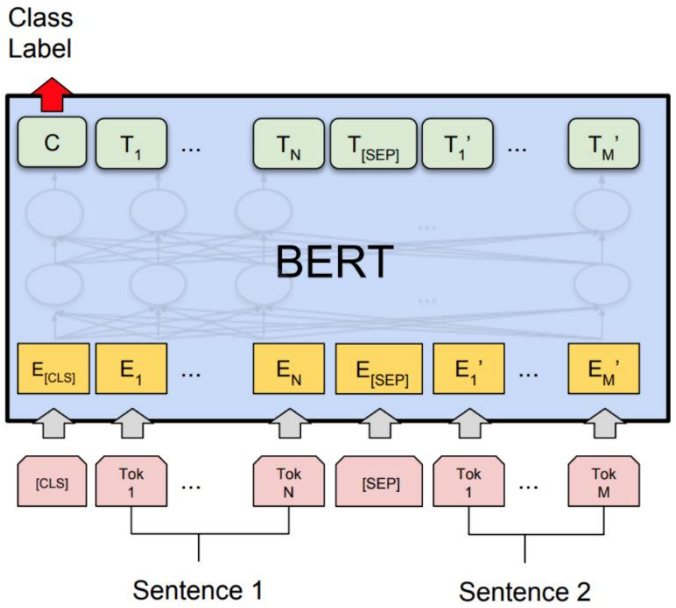
Transformer Decoder

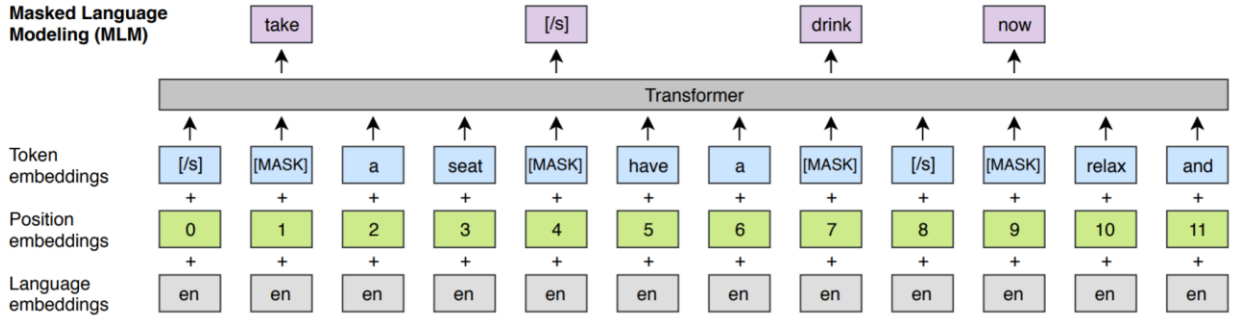


단일언어(monolingual)에 비지도학습 (unsupervised learning) 으로 학습시킨 언어모델

다음 단어를 예측하기 위해서 지금까지 예측했던 단어들과 이전 hidden state 의 값을 사용하는 방식

2. Masked Language Modeling : MLM

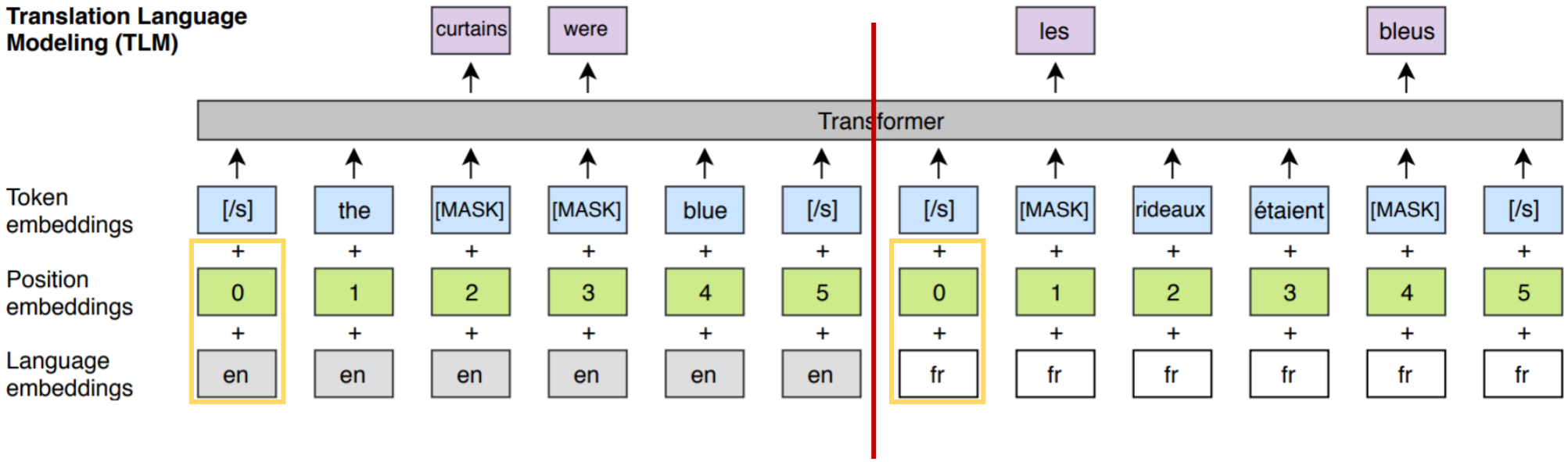




단일언어(monolingual)에 비지도학습 (unsupervised learning) 으로 학습시킨 언어모델

기존의 BERT와는 달리 문장의 쌍을 넣는 게 아니라 임의의 문장(arbitrary number of sentences)를 사용

3. Translation Language Modeling : TLM



병렬문장(parallel sentences) 지도학습(supervised learning) 언어모델

=> 병렬 데이터 간 상호 보완적으로 언어모델을 학습

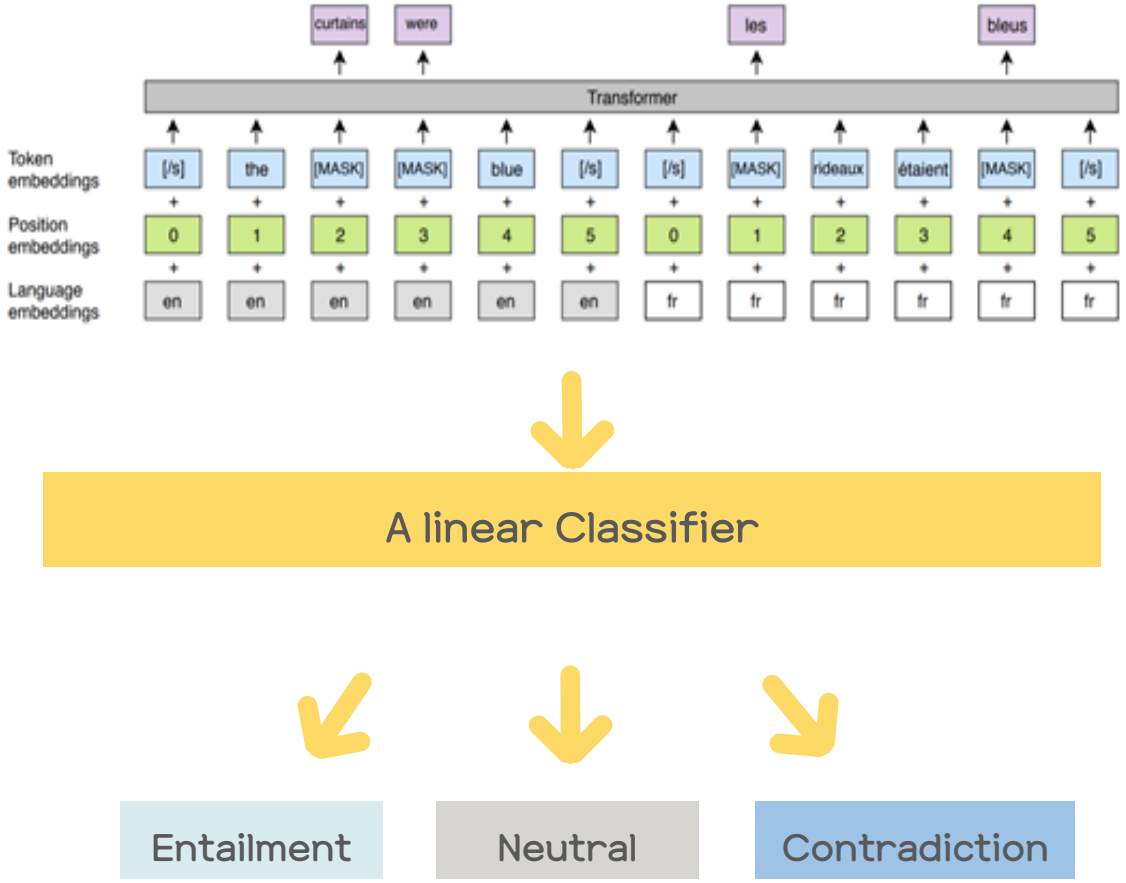
1. Cross-lingual classification : XNLI

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
Arabic	تحتاج الوكالات لأن تكون قادرة على قياس مستويات النجاح. لا يمكنأنالوكالات أ التعرف ما إذا كانت ناجحة أم لا	Nine-Eleven	Contradiction

Table 1: Examples (premise and hypothesis) from various languages and genres from the XNLI corpus.

각 언어별 2개의 문장을 Entailment, Neutral, Contradiction 3개로 분류하는 Task

1. Cross-lingual classification : pretraining and fine-tuning



Pretraining

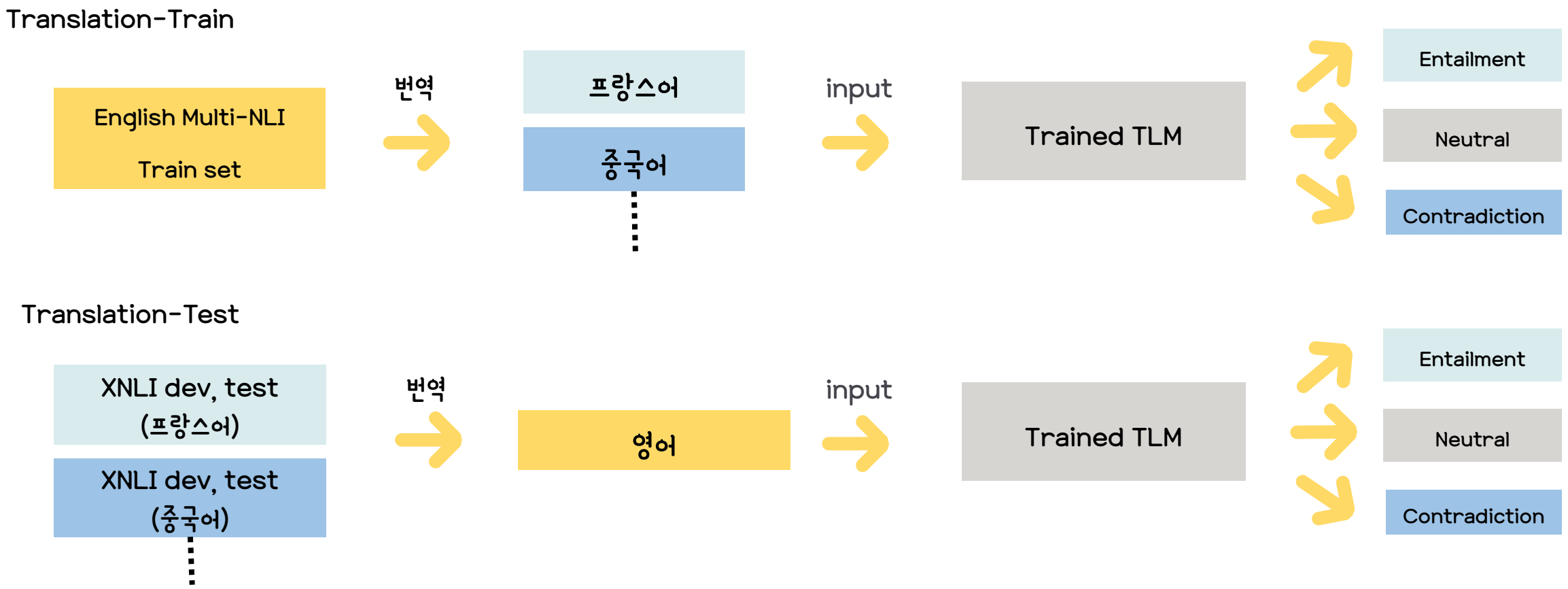
- Transformer 1024 hidden units, 8 heads
- GELU Activation, dropout rate 0.1
- ADAM optimizer, linear warm up
- TLM 학습에서 4000개의 토큰으로 이루어진 비슷한 문장을 포함
- 기계번역에는 6 layer를 사용, 그리고 2000개 토큰의 배치를 생성
- 12 layer 모델을 wikipedias XNLI languages로 학습

Fine-Tuning

- 8 or 16 batch size, 최대 256개 단어로 잘라냄
- Vocab size 95k, BPE 80k 사용
- 12 layer 모델을 English NLI로 fine tuning

다국어 병렬 데이터셋으로 TLM을 학습한 다음 영어 NLI dataset으로 fine tuning

1. Cross-lingual classification : baseline



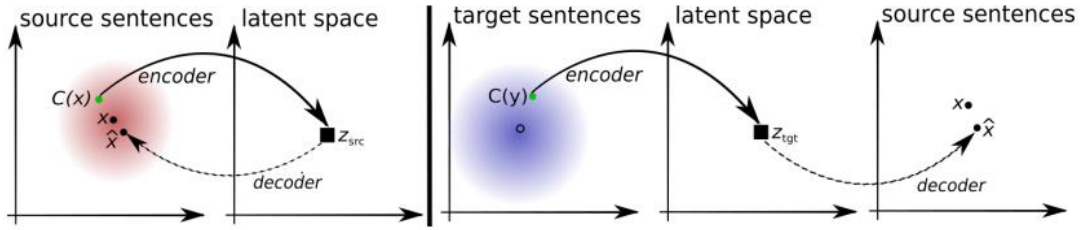
1. Cross-lingual classification : Result

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	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
Evaluation of cross-lingual sentence encoders																
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)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

- Unsupervised MLM은 zero-shot cross-lingual classification에서 sota 달성
- Unsupervised model인 MLM이 Supervised model인 Artetxe and schwenk 논문보다 더 좋은 성능
  - 이전 sota 기준 70.2% -> 71.5%로 1.3% 개선, TLM과 함께 사용시 71.5-> 75.1%로 3.6% 개선
- Low-resource 언어(sw, ur)에서 각각 6.2%, 6.3% 개선
- 영어만 평가 시 MLM보다 TLM을 결합했을 때 1.8% 개선

2. Unsupervised machine translation

Unsupervised Machine Translation using monolingual corpora only [CL/2018.04]



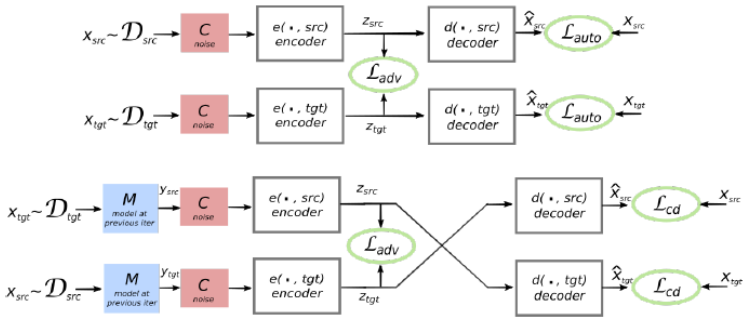


Figure 2: Illustration of the proposed architecture and training objectives. The architecture is a sequence to sequence model, with both encoder and decoder operating on two languages depending on an input language identifier that swaps lookup tables. Top (auto-encoding): the model learns to denoise 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.

비지도학습 기반 번역은 초기 cross-lingual word embedding의 품질이 중요 => initialized weight가 중요

두 언어의 동일한 단어에 대해서 encoder/decoder(back translation 역할)에서 동일한 잠재 공간을 가지도록 하는 것이 목표

2. Unsupervised machine translation

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
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
Our results for different encoder and decoder initializations							
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.

3. Supervised machine translation

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro → en	28.4	31.5	35.3
ro ↔ en	28.5	31.5	35.6
ro ↔ 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 ↔ en corresponds to models trained on both directions.

양방향 Back Translation 방식으로 학습

지도학습 기반 번역에서는 MLM과 Back Translation을 조합한 모델이 성능이 가장 좋음

4. Low-resource Language Model

네팔어 : समानान्तर कर्पस महत्त्वपूर्ण छ
힌디어 : समानांतर कॉर्पस महत्वपूर्ण है

- 위키피디아에는 네팔어가 100k 문장만 존재
- 힌디어는 네팔어보다 6배 많은 문장이 존재
- 또한 두 언어는 매우 유사해서, BPE vocabulary의 80% 정도를 공유

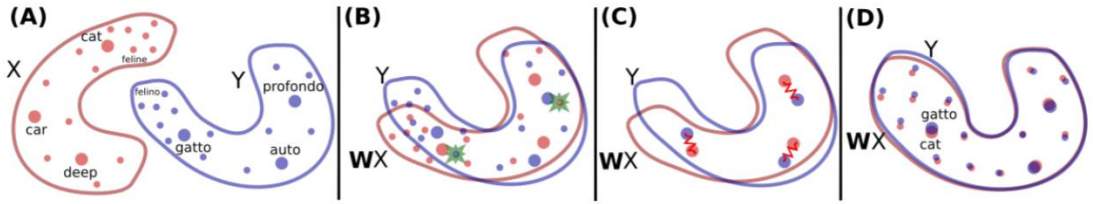
Training languages	Nepali perplexity
Nepali	157.2
Nepali + English	140.1
Nepali + Hindi	115.6
Nepali + English + Hindi	**109.3**

Low-resource 언어의 경우 유사한 다른 언어와 함께 학습 시켰을 때,
동일한 vocabulary에서 후보군을 결정하기 때문에 복잡도(Perplexity)를 줄일 수 있음

=> 다국어 n-gram의 대한 복잡성을 개선

5. Unsupervised cross-lingual word embedding

Word Translation without parallel data[CL/2018.01]



- 저자의 이전 논문 Word Translation without parallel data[CL/2018.01]에서 진행했던 연구의 확장
- 병렬 데이터 없이 같은 의미의 단어라면 다른 언어도 같은 공간에 맵핑하고자 함

	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).

3가지의 방법론을 가지고, 같은 의미의 다른 언어들을 동일한 공간에 맵핑시키고,
Cosine Similarity, L2 distance, 피어슨 상관계수(SemEval)로 얼만큼 잘 맵핑되어있는지 측정

Reference

[논문]

- Cross-lingual Language Model Pretraining [CL/2019.01]
- Word Translation without parallel data[CL/2018.01]
- XNLI: Evaluating Cross-lingual Sentence Representations[EMNLP/2018.09]
- Unsupervised Machine Translation using monolingual corpora only [CL/2018.04]
- Unsupervised Pretraining for Sequence to Sequence Learning [CL / 2018.02]

[블로그]

- https://yhdosu.tistory.com/entry/Cross-lingual-Language-Model-Pre-training
- https://keep-steady.tistory.com/41
- https://m.blog.naver.com/hist0134/221360305018

[유튜브]

- https://www.youtube.com/watch?v=aRwvwdfK0cA&t=405s
- https://www.youtube.com/watch?v=caZLVcJqsqo&t=968s

Thank You