English Contrastive Learning Can Learn Universal Cross-lingual Sentence Embeddings

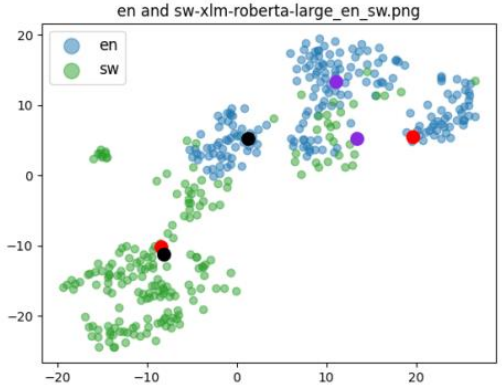
Authors: Yau-Shian Wang, Ashley Wu, Graham Neubig

Venue: EMNLP 2022

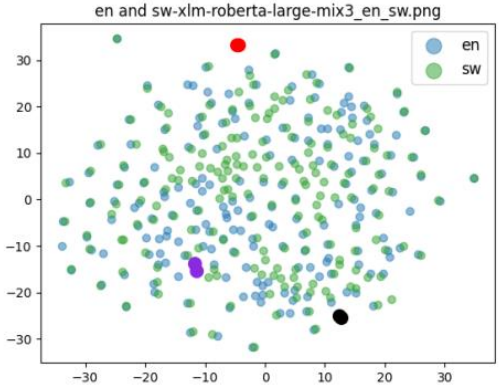
1. Introduction

- Universal cross-lingual sentence embedding은 여러 언어의 문장들을 공유 임베딩 공간으로 매핑함
  - 이 때, 언어 간에 의미적으로 유사한 문장은 서로 가까이에 있음
  - 다국어 문서 검색, 다국어 질의응답, 비지도적인 기계 번역, zero-shot transfer 에 활용됨

- 아래 그림 (a)에서 미세조정 하지 않은 XLM-R 은 임베딩 공간에서 각 언어의 임베딩을 서로 다른 클러스터로 분리함
- 그래서 이전 연구들에서는 다국어 언어 모델을 수십억 개의 병렬 데이터로 미세조정해서 분리했음 → 아래 그림 (b)
  - XLM-R은 사전 학습된 다국어 언어 모델
  - 병렬 데이터는 예를 들어 같은 의미를 가진 다른 언어 데이터 (ex. 영어-한국어)



(a) XLM-R without finetuning.



(b) XLM-R fintuned on English NLI data.

1. Introduction

- 하지만 모든 언어에 대해 수많은 병렬 데이터를 얻는 것이 중요한 것은 아님
  - 병렬 데이터의 필요성을 완화하기 위한 방법으로 문장 임베딩의 cross-lingual transfer을 향상시키면 됨

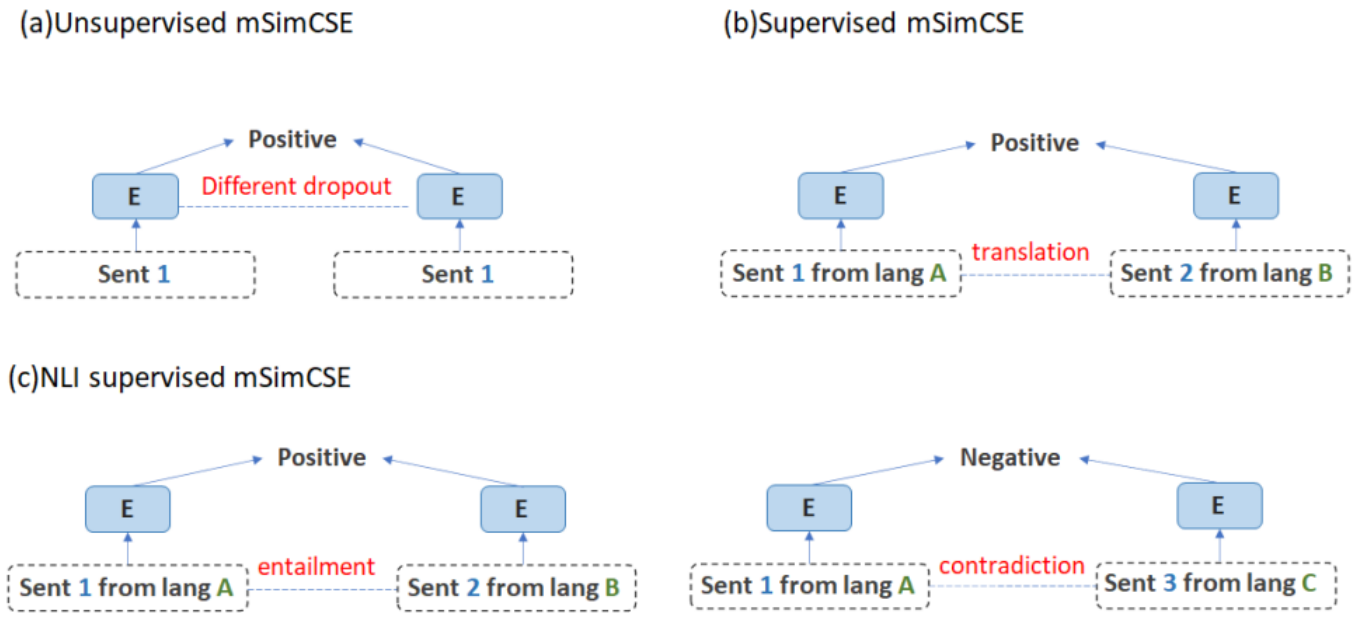
- Cross-lingual transfer?
  - labeled data가 풍부한 언어의 data를 이용해서 labeled data가 희소한 언어에 대한 성능을 보완하는 것
  - Source language로 미세조정한 모델이 target language에서 일반화가 가능하게 함

- 이전 연구에서 사전 학습된 다국어 언어 모델이 cross-lingual transfer에서 좋은 성능을 보였었음
  → downstream task에서 미세조정 된 표현이 다양한 언어에 보편적이라는 것을 의미함

- 이 논문에서의 제안
  - 영어 data만 사용하는 방법으로 문장 검색 task에서의 다양한 cross-lingual transfer 방법을 연구
  - SimCSE를 다국어로 확장해 cross-lingual transfer를 할 수 있도록 multilingual-SimCSE(mSimCSE)를 제안
    - SimCSE는 대조학습 프레임워크
    - 대조학습
      - 입력 문장(anchor)와 의미적으로 유사한 것(positive)과는 가깝게, 유사하지 않은 것 (negative)과는 멀어지게 학습하는 방법
    - SimCSE
      - 입력 문장(anchor)에 서로 다른 dropout mask를 적용하여 positive를 생성하여 높은 성능을 보임

2. Method

- 아래 그림은 mSimCSE를 학습하는 4가지 방법임
  - Unsupervised mSimCSE - (a)
  - English NLI supervised mSimCSE - (c)
  - Cross-lingual NLI supervised mSimCSE - (c)
  - Supervised mSimCSE - (b)



2. Method

- Unsupervised mSimCSE
  - Wikipedia data를 사용
  - SimCSE와 동일하게 anchor에 서로 다른 dropout mask를 적용하여 positive 생성

- English NLI supervised mSimCSE
  - NLI dataset 사용
  - ‘entailment’ label인 문장을 positive로, ‘contradiction’ label인 문장을 hard negative로 사용

- Cross-lingual NLI supervised mSimCSE
  - XNLI dataset 사용 (NLI dataset을 여러 언어로 번역한 dataset)

힌두어(hi)

premise (string)	hypothesis (string)	label (class label)
"Conceptually cream skimming has two basic dimensions - product and geography ."	"Product and geography are what make cream skimming work ."	1 (neutral)

영어(en)

premise (string)	hypothesis (string)	label (class label)
"Conceptually क्रीम एंजलिस में दो मूल आयाम हैं - उत्पाद और भूगोल ।"	"उत्पाद और भूगोल क्या क्रीम एंजलिस काम बनाते हैं ."	1 (neutral)

  - ‘entailment’ label인 문장을 positive로, ‘contradiction’ label인 문장을 hard negative로 사용
  - 하지만 랜덤하게 선택하기 때문에 언어가 다를 수도 있고 같을 수도 있음

- Supervised mSimCSE
  - 1) 병렬 데이터를 positive로 사용 (English – Swahili)
  - 2) 추가적으로 영어 NLI dataset을 섞어 같이 사용한 모델도 실험 → 즉, (English, Swahili) + (English, English(with dropout mask))를 positive로 사용 → 기존 supervised와 달리 병렬 데이터가 상대적으로 적음

3. Experiments

- 실험 세팅
  - Learning rate: 1e-5, Epoch: 1, Batch size: 128
  - Backbone model: XLM-Roberta-large
  - 그 외는 SimCSE와 동일

3-1. Sentence Retrieval

- source language 문장과 target language 문장을 매칭하는 작업
- BUCC, Tatoeba dataset으로 평가

Models	BUCC	Tatoeba-14	Tatoeba-36
Unsupervised			
XLM-R	66.0	57.6	53.4
INFOXLM	-	77.8	67.3
DuEAM	77.2	-	-
XLM-E	-	72.3	62.3
HiCTL	68.4	-	59.7
$mSimCSE_{en}$	87.5	82.0	78.0
English NLI supervised			
(Phang et al., 2020)	71.9	-	81.2
$mSimCSE_{en}$	**93.6**	**89.9**	**87.7**
Cross-lingual NLI supervised			
$mSimCSE_{en,fr}$	94.2	90.8	88.8
$mSimCSE_{en,fr,sw}$	94.3	93.3	90.3
$mSimCSE_{all}$	**95.2**	93.2	91.4
DuEAM	81.7	-	-
Fully Supervised			
LASER	92.9	95.3	84.4
LaBSE	93.5	95.3	95.0
$mSimCSE_{sw}$	86.8	87.7	86.3
$mSimCSE_{fr}$	87.1	87.9	85.9
$mSimCSE_{sw,fr}$	88.8	90.2	88.3
$mSimCSE_{sw,fr}$+NLI	93.6	91.9	90.0

3. Experiments

3-1. Sentence Retrieval

- High resource language: Hindi(hi), French(fr), German(de), Afrikaans(af), Swahili(sw)
- Row resource language: Telugu(te), Tagalog(tl), Irish(ga), Georgian(ka), Amharic(am)

Models	hi	fr	de	af	te	tl	ga	ka	am	sw
Unsupervised										
CRISS	92.2	92.7	98.0	-	-	-	-	-	-	-
DuEAM	83.5	-	93.4	79.9	78.6	56.8	35.0	70.7	46.4	-
$mSimCSE_{en}$	86.9	87.2	94.1	76.0	78.8	49.7	39.2	75.2	48.8	29.4
English NLI supervised										
$mSimCSE_{en}$	**94.4**	**93.9**	**98.6**	**85.6**	**92.9**	**70.0**	**54.8**	**89.2**	**79.5**	**42.1**
Cross-lingual NLI supervised										
DuEAM	92.9	-	96.0	84.8	90.6	60.6	42.0	76.4	56.0	-
$mSimCSE_{en,fr}$	95.1	94.4	98.8	88.9	94.2	73.4	59.4	91.3	79.5	44.5
$mSimCSE_{en,fr,sw}$	95.7	94.2	98.4	87.9	94.4	75.6	62.1	90.5	82.7	**75.5**
$mSimCSE_{all}$	**96.2**	94.8	98.8	**90.6**	**96.2**	**80.9**	**65.1**	**92.4**	**82.4**	67.8
Fully supervised										
LASER	94.7	95.7	99.0	89.4	79.7	-	5.2	35.9	42.0	42.4
$mSimCSE_{sw}$	94.3	91.6	97.6	85.2	88.5	76.3	60.8	85.5	65.2	47.6
$mSimCSE_{fr}$	94.1	92.6	97.3	84.6	89.3	70.8	54.6	86.3	63.4	43.6
$mSimCSE_{sw,fr}$	95.1	93.8	97.8	86.1	91.2	75.8	59.6	88.9	74.4	51.5
$mSimCSE_{sw,fr}$+NLI	95.8	94.7	98.6	89.8	95.7	77.8	63.9	91.7	81.0	57.1

3. Experiments

3-2. Cross-lingual STS

- 두 문장 사이의 예측된 의미론적 유사성과 인간의 판단이 상관관계가 있는지를 평가함 (스피어만 상관계수)
- 두 문장은 같은 언어일수도 있고 다른 언어일수도 있음

Models	ar-ar	ar-en	es-es	es-en	tr-en
Unsupervised					
XLM-R	53.5	26.2	68.1	10.7	10.5
mBERT	55.2	28.3	68.0	23.6	17.3
$mSimCSE_{en}$	72.3	48.4	83.7	57.6	53.4
English NLI supervised					
$mSimCSE_{en}$	**81.6**	71.5	**87.5**	**79.6**	71.1
Cross-lingual NLI supervised					
DuEAM	69.7	54.3	78.6	56.5	58.4
$mSimCSE_{all}$	79.4	72.1	85.3	77.8	74.2
Fully Supervised					
LASER	79.7	-	57.9	-	72.0
LaBSE	80.8	-	65.5	-	72.0
SP	76.7	78.4	85.6	77.9	79.5
$mSimCSE_{sw,fr}$+NLI	77.7	72.4	86.3	79.7	72.5

3. Experiments

3-3. Unsupervised Classification

- 영어 이외에 언어에서 의미론적으로 유사한 문서를 함께 클러스터 할 수 있는지 실험
- CLUE benchmark에 Tnews dataset 사용
  - Tnews dataset은 중국 뉴스 분류 데이터셋으로 15개의 category를 가지고 있음.
- Sentence embedding으로 k-means clustering을 진행
  - k는 category수만큼으로 지정
  - 각 클러스터의 평균 정확도를 측정

Models	Purity
Unsupervised	
Random	6.7
mBERT	15.2
XLM-R	13.7
$mSimCSE_{en}$	30.3
English NLI supervision	
$mSimCSE_{en}$	**40.3**
Cross-lingual NLI supervision	
$mSimCSE_{all}$	41.6
Supervised Classification Model	
BERT	56.6

3. Experiments

3-4. Zero-shot Cross-lingual Transfer of Sentence Classification

- Cross-lingual zero-shot transfer을 평가하기 위해 PAXS-X 문장 분류 task를 평가
  - PAXS-X는 두 문장이 paraphrase인지, 즉 다른 말로 바꾸어 표현한 것인지를 분류하는 작업

Models	Accuracy
Unsupervised	
mBERT	81.9
XLM-R	86.4
XLM-E	87.1
$mSimCSE_{en}$	**88.1**
English NLI supervised	
(Phang et al., 2020)	87.9
$mSimCSE_{en}$	88.2

4. Analysis

4-1. The effect of Parallel Sentences Number

- English-French dataset과 English NLI data를 섞음
- 점차 병렬 데이터의 수를 늘려가며 실험

Parallel data	BUCC	Tatoeba14	Tatoeba36
0	91.4	90.4	88.0
10k	92.6	90.6	88.5
100k	93.5	90.8	88.6
1M	94.4	90.6	88.5
5M	94.5	90.7	88.2

4-2. Can Contrastive Learning Removes Language Identity?

- 두 개의 언어 분류기 학습하여 Sentence embedding이 들어오면 어떤 언어인지 분류
- XNLI dataset 사용

Models	en,de,fr,hi	en,tr,ar,bg
XLM-R	99.2	99.8
$mSimCSE_{en}$	91.1	95.3

4. Analysis

4-3. Hyperparameters

- mSimCSE가 stable하고 hyperparameter에 민감하지 않다는 것을 보여줌

epoch	bs	lr	BUCC	Tatoeba-14	Tatoeba-36
1	128	1e-5	93.6	89.9	87.7
2	128	1e-5	94.4	90.0	87.8
3	128	1e-5	93.5	90.0	87.4
1	256	1e-5	93.3	90.0	87.9
1	128	2e-5	93.7	90.0	87.4

5. Conclusion

- SimCSE를 다국어로 확장한 mSimCSE를 제안함
- Unsupervised, English NLI supervised, Cross-lingual NLI supervised, supervised로 총 4가지 학습 방법의 mSimCSE를 제안함
- English NLI supervised 방법이 supervised 방법과 거의 동등한 수준의 성능을 달성함을 보임
- 오직 영어 데이터만 사용하는 것이 보편적인 sentence embedding을 학습하는데 효율적임을 보임
- 이 논문은 단일 언어 말뭉치에서 문장의 의미적 관계를 학습하는 대조학습을 사용하는 것이 유망한 방법이라는 것을 제안함

Thank You

감사합니다.