Analyzing Zero-shot Cross-lingual Transfer In Supervised NLP Tasks

SAMSUNG SDS

Hyunjin Choi, Judong Kim, Seongho Joe, Seungjai Min, Youngjune Gwon

Milan, Italy 10 | 15 January 2021

ICPR 2020

Motivation

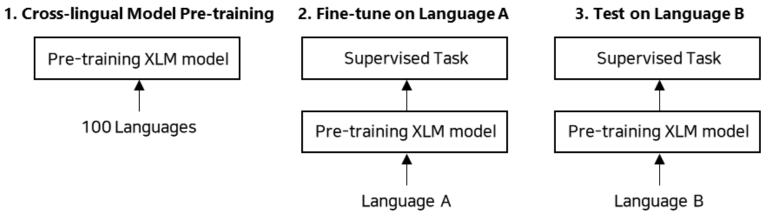
· **Zero-shot cross-lingual transfer**

- A supervised NLP task trained on a corpus in one language, or the “source,” is directly applied to another language or the “target” without any additional training
- Hypothesis: the **zero-shot transfer loss of performance** in a supervised NLP performance in the target language **is little** or none at all

· The purpose of this paper is to **empirically validate** such **hypothesis**

Framework

· Our experimental framework to validate zero-shot cross-lingual transfer using a supervised task



1. Cross-lingual Model Pre-training

Pre-training XLM model

100 Languages

2. Fine-tune on Language A

Supervised Task

Pre-training XLM model

Language A

3. Test on Language B

Supervised Task

Pre-training XLM model

Language B

Figure1: Zero-shot Cross-lingual Transfer Evaluation Framework

Experiments & Results

❶ Supervised NLP Tasks

· **Semantic Textual Similarity**

- Evaluate the **similarity between two sentences** (regression task)
- (English) Semantic Textual Similarity benchmark (**STSb**), Korean STS (KorSTS), SemEval-2017 Spanish, and SemEval-2017 Arabic
- Results

		Evaluation Language			
	Fine-tuning Task(s)	English	Korean	Spanish	Arabic
Zero-shot	STSb (English)	87.44 (87.43)	82.34 (82.27)	85.58 (87.02)	72.67 (70.54)
	KorSTS (Korean)	84.47 (84.40)	83.38 (83.16)	84.94 (85.00)	70.99 (69.66)
Mixed Launguage Fine-tuning	STSb → KorSTS	86.43 (86.47)	83.54 (83.42)	85.47 (86.05)	73.85 (73.39)
	KorSTS → STSb	88.33 (88.34)	85.12 (85.12)	86.77 (87.83)	73.37 (72.37)
	STSb + KorSTS	87.71 (87.84)	84.37 (84.48)	86.53 (86.99)	75.72 (75.22)

Table1: Evaluation on STS tasks.
Numbers represent spearman(pearson) correlations in percentile

- Presence of zero-shot cross-lingual transfer strong for STS : (English fine-tune→ Korean test) 1.24% decrease & (Korean fine-tune → English test) 3.40% decrease
- Low score for Arabic: relatively lower resource language compared to the others (XLM-R uses 28.0GB of Arabic, Korean 54.2GB, Spanish 53.3GB, English 300.8GB)

· **Machine Reading Comprehension (MRC)**

- Understand a paragraph and answer the question
- (English) Stanford Question Answering Dataset (SQuAD), Korean Question Answering Dataset (KorQuAD), and Spanish SQuAD (SQuAD-es)
- Results

		Evaluation Language		
	Fine-tuning Task(s)	English	Korean	Spanish
Zero-shot	SQuAD (Enlgish)	88.81 (81.68)	80.92 (45.08)	72.07 (53.18)
	KorQuAD (Korean)	72.03 (61.93)	89.58 (65.29)	58.65 (43.09)
	SQuAD-es (Spanish)	84.75 (74.51)	78.87 (42.76)	76.11 (59.68)
Mixed Language Fine-tuning	SQuAD → KorQuAD	85.81 (77.16)	90.17 (66.02)	70.54 (52.40)
	SQuAD → SQuAD-es	86.73 (76.78)	78.16 (36.87)	76.70 (59.87)
	KorQuAD → SQuAD	89.16 (82.20)	88.42 (62.83)	72.78 (53.92)
	SQuAD + KorQuAD	84.41 (75.93)	86.79 (62.45)	67.72 (48.49)
	SQuAD + KorQuAD + SQuAD-es	89.29 (81.98)	90.41 (66.36)	76.75 (59.66)

Table2: Evaluation on MRC tasks.
Numbers represent F1 score (Exact match)

- Presence of zero-shot cross-lingual transfer exists for MRC tasks
- Compared to the performance on STS tasks, the degraded gap is higher for MRC tasks
- Fine-tuning with an additional language improves the MRC performance regardless of testing language
- Fine-tuning with all other languages yields the best MRC performance

· **Sentiment Analysis**

- (English) Large Movie Review Dataset (LMRD), and (Korean) Naver Sentiment Movie Corpus (NSMC)
- Results

		Evaluation Language	
	Fine-tuning Task(s)	English	Korean
Zero-shot	LMRD (English)	93.52	79.24
	NSMC (Korean)	86.38	90.10
Mixed Language Fine-tuning	LMRD → NSMC	90.65	90.12
	NSMC → LMRD	93.69	89.47
	LMRD + NSMC	93.80	90.24

Table3: Evaluation on sentiment classification tasks.
The numbers represent classification accuracy in percentage

- Presence of zero-shot cross-lingual transfer exists for Sentiment classification tasks
- We found cross-lingual transfer most pronounced in STS, the sentiment analysis the next, and MRC the last

❷ Cross-lingual Mapping for Fine-grained Alignment of Sentence Embeddings

· Compute a projection matrix that achieves fine-grained alignment of sentence embeddings across different languages

· System of least squares via normal equation

Source language A → Target language B

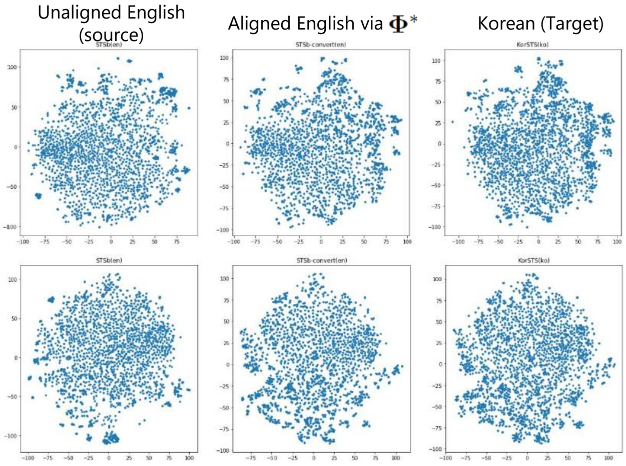
$$\mathbf{S}_A \mathbf{\Phi} = \mathbf{S}_B$$

$$\mathrm{S}_A = \begin{bmatrix} — & \mathrm{s}_A^{(1)} & — \\ — & \mathrm{s}_A^{(2)} & — \\ & \vdots & \\ — & \mathrm{s}_A^{(n)} & — \end{bmatrix}, \ \mathrm{S}_B = \begin{bmatrix} — & \mathrm{s}_B^{(1)} & — \\ — & \mathrm{s}_B^{(2)} & — \\ & \vdots & \\ — & \mathrm{s}_B^{(n)} & — \end{bmatrix},$$

$$\mathrm{s}_A^{(i)} = \begin{bmatrix} a_1^{(i)} \\ a_2^{(i)} \\ \vdots \\ a_d^{(i)} \end{bmatrix}^\top, \ \mathrm{s}_B^{(i)} = \begin{bmatrix} b_1^{(i)} \\ b_2^{(i)} \\ \vdots \\ b_d^{(i)} \end{bmatrix}^\top$$

$$\mathbf{\Phi}^* = \left(\mathbf{S}_A^\top \mathbf{S}_A\right)^{-1} \mathbf{S}_A^\top \mathbf{S}_B$$

· Unaligned: 0.4636 (cosine similarity) → Aligned: 0.7131



Unaligned English (source)

Aligned English via $\mathbf{\Phi}^*$

Korean (Target)

Figure2: t-SNE plots of English and Korean translated pairs from STSb and KorSTS

· Zero-shot Transfer vs. Cross-lingual Mapping

		Method	
		Zero-shot Transfer	Cross-lingual Mapping
Fine-tuning Task	STSb	49.03	59.16
	KorSTS	43.23	47.24