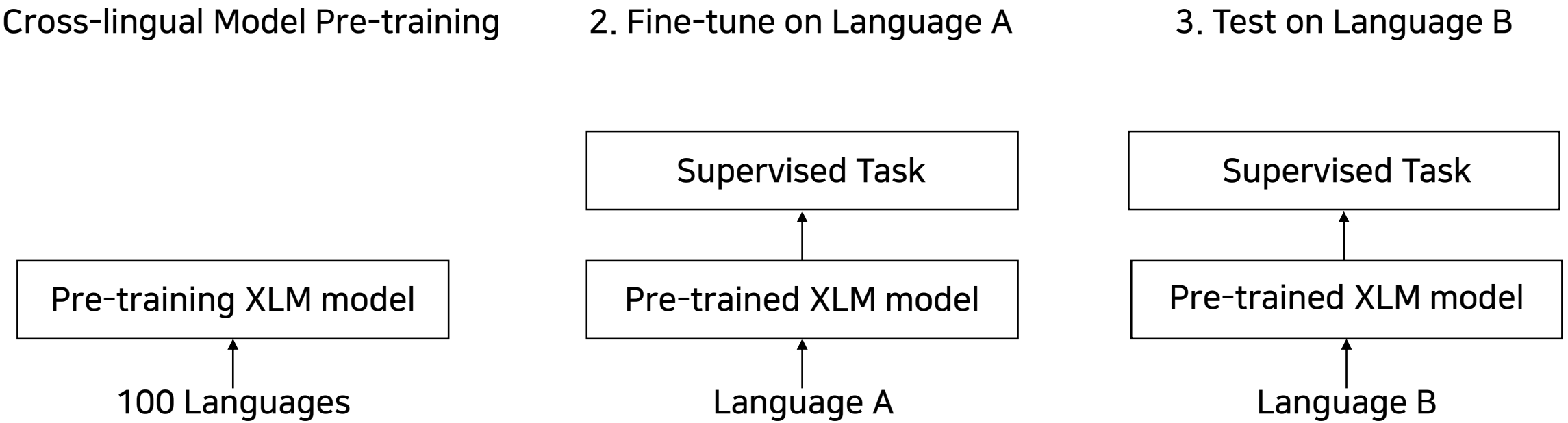
Analyzing Zero-shot Cross-lingual Transfer in Supervised NLP Tasks

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

Zero-shot Cross-lingual Transfer Evaluation Framework



Experiments

❶ Semantic Textual Similarity

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

TABLE I
EVALUATION ON STS TASKS. NUMBERS REPRESENT THE SPEARMAN (PEARSON) CORRELATIONS IN PERCENTILE.

		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 Laungauge 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)

Experiments

❷ Machine Reading Comprehension (MRC)

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

TABLE II
EVALUATION ON MRC TASKS. NUMBERS REPRESENT F1 SCORE, AND NUMBERS IN PARENTHESES ARE EXACT MATCHES.

		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)

Experiments

❸ Sentiment Analysis

- Large Movie Review Dataset (LMRD) and Naver Sentiment Movie Corpus (NSMC)

TABLE III
EVALUATION ON SENTIMENT CLASSIFICATION TASKS. THE NUMBERS REPRESENT CLASSIFICATION ACCURACY IN PERCENTAGE.

		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

Experiments

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

- Use linear algebraic methods to compute a projection matrix that achieves fine-grained alignment of sentence embeddings across different languages

Source language A → Target language B

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

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

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

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

Experiments

❹ Cross-lingual Mapping for Fine-grained Alignment of Sentence Embeddings - Results

Unaligned: 0.4636 (cosine similarity) → Aligned: 0.7131

Unaligned English (source) Aligned English via Φ^* Korean (Target)

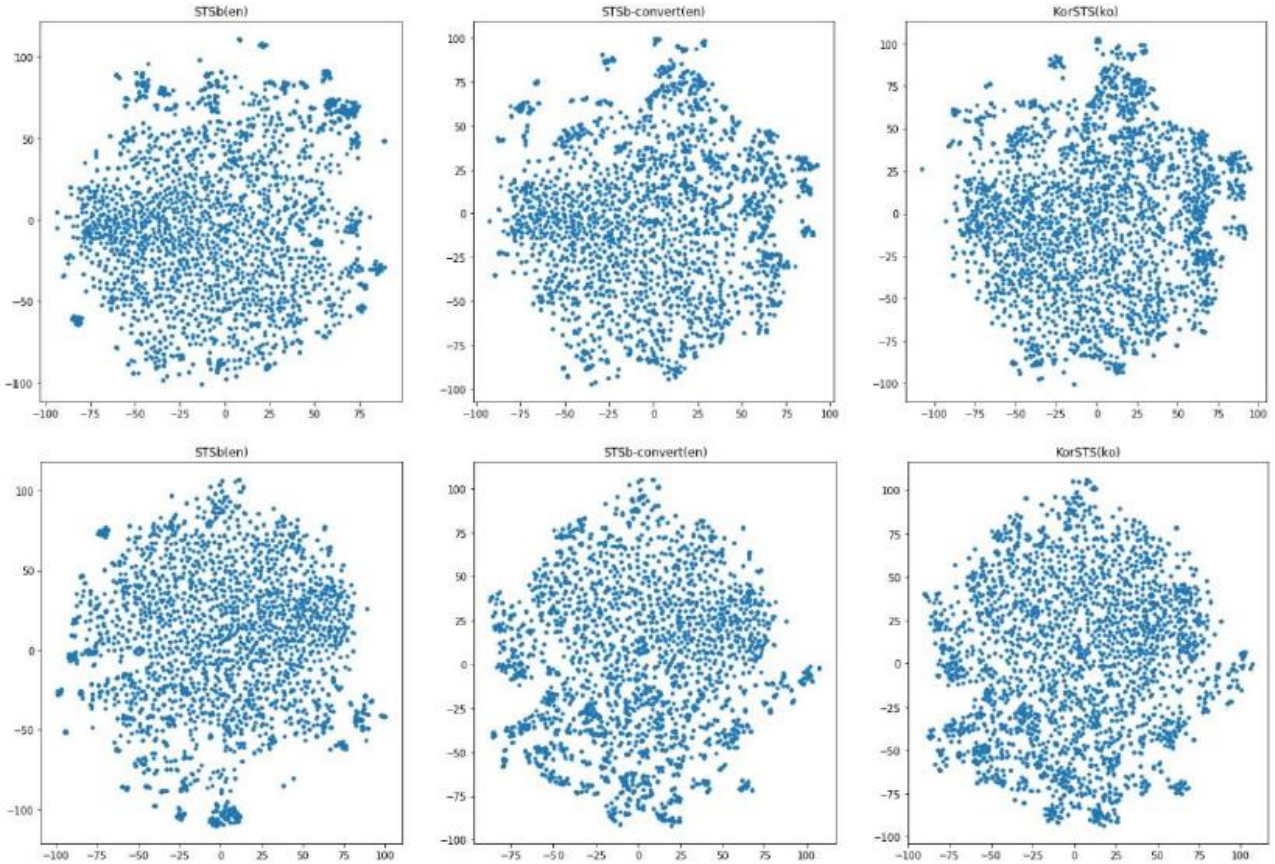


TABLE IV
STS EVALUATION WITH CROSS-LINGUAL SENTENCE PAIRS.

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

Conclusion

- This paper focuses on the empirical validation of the cross-lingual transfer properties induced by XLM pretraining
- Experiment with XLM-RoBERTa (XLM-R), a large cross-lingual language model
- Tasks including semantic textual similarity, machine reading comprehension, sentiment analysis
- Cross-lingual transfer be most pronounced in STS, the sentiment analysis the next, and MRC the last
- Compute matrix projections linear algebraically that directly map sentence embeddings of one language to another to analyze the effect of fine-grained alignment of sentences in zero-shot cross-lingual transfer