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

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

Samsung SDS

Abstract—**In zero-shot cross-lingual transfer, a supervised NLP task trained on a corpus in one language is directly applicable to another language without any additional training. A source of cross-lingual transfer can be as straightforward as lexical overlap between languages (*e.g.*, use of the same scripts, shared subwords) that naturally forces text embeddings to occupy a similar representation space. Recently introduced cross-lingual language model (XLM) pretraining brings out neural parameter sharing in Transformer-style networks as the most important factor for the transfer. In this paper, we aim to validate the hypothetically strong cross-lingual transfer properties induced by XLM pretraining. Particularly, we take XLM-RoBERTa (XLM-R) in our experiments that extend semantic textual similarity (STS), SQuAD and KorQuAD for machine reading comprehension, sentiment analysis, and alignment of sentence embeddings under various cross-lingual settings. Our results indicate that the presence of cross-lingual transfer is most pronounced in STS, sentiment analysis the next, and MRC the last. That is, the complexity of a downstream task softens the degree of cross-lingual transfer. All of our results are empirically observed and measured, and we make our code and data publicly available.**

I. INTRODUCTION

Pretraining language models at a large scale has dramatically improved natural language understanding. According to a comprehensive analysis [1] on the limitations in pretraining a multi-lingual model, more languages lead to better cross-lingual performance for low-resource languages only up to a certain point when the number of languages increases. The phenomenon is dubbed the curse of multilinguality, which can only be freed up by scaling up the model size.

The recent experimental results show that multilingual models can outperform their monolingual counterparts. For a low-resource language that lacks in labeled examples, such results are an encouraging breakthrough for building an NLP application for low-resource languages. In cross-lingual language understanding, XLM by Conneau & Lample [2], despite being pretrained by only masked language modeling (MLM), has reported the state-of-the-art on downstream benchmarks. Shared lexical features (*e.g.*, subwords, scripts, anchor points) across languages have been suspected for the primary source of learning language-independent representation that leads to cross-lingual transfer. Recent studies, however, show that parameter sharing induced by the Transformer architecture is instead the most attributable factor for the transfer.

We are motivated by these progresses in language modeling. This work focuses on empirical analysis of cross-lingual transfer in supervised NLP tasks fine-tuned over XLM. In particular, we are interested in zero-shot transfer settings where no additional training is done using the target language examples after being fine-tuned in the source language. We experiment with XLM-RoBERTa (XLM-R) [1], a large XLM model with 550 million parameters and a 250k vocabulary size by extending semantic textual similarity, SQuAD [3] & KorQuAD [4] question answering, and sentiment classifications for various cross-lingual settings.

At last, beyond previous work that has attempted to align word embeddings across different languages [5], we compute a projection that directly maps sentence embeddings of one language to those of another. We then analyze the effect of fine-grained alignment of sentences across different languages to the quality of zero-shot cross-lingual transfer, manifested through the aforementioned NLP task performances measured empirically.

We make the following contributions.

- We provide rigorous results on cross-lingual transfer present in three important supervised NLP tasks that require high-level natural language understanding, namely STS, MRC, and sentiment classification.
- We propose to directly compute a cross-lingual mapping that aligns sentence embeddings of different languages whereas previous work has focused on word-level embeddings.
- We furthermore show benefits of the fine-grained cross-lingual sentence alignment that enables directly comparing sentences from different languages for sentence-pair regression tasks.

The rest of this paper is organized as follows. In Section II, we describe our approach by presenting the zero-shot cross-lingual evaluation framework. Section III discusses our experimental methodology and empirical results. Section IV concludes the paper.

II. OUR APPROACH

XLM pretraining is known to effectively promote cross-lingual transfer where a supervised model fine-tuned in one language is applied to another without additional training.

A. Zero-shot Cross-lingual Evaluation Framework

We propose a simple approach to transfer a supervised model learned in one language to another for zero-shot cross-lingual evaluation as illustrated in Fig. 1. First, we place a pretrained XLM–for our case, the 550 million-parameter XLM-RoBERTa (XLM-R) [1] trained on 100 languages is used. We then fine-tune XLM-R for a downstream task using
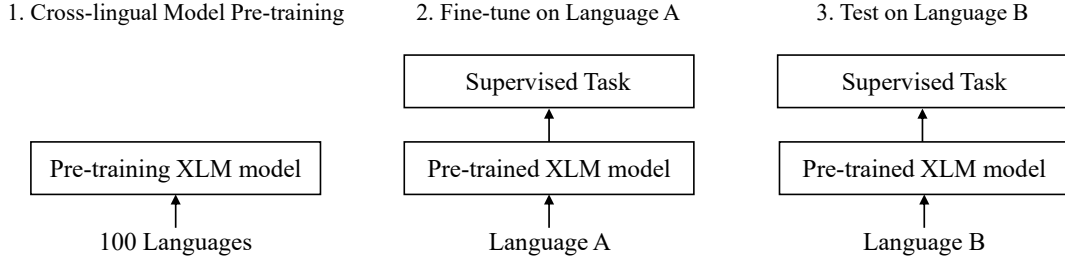
Fig. 1. Zero-shot cross-lingual evaluation

labeled data in language A. Lastly, we evaluate the fine-tuned downstream task in both languages A and B. Note that running a test set from language B to the fine-tuned task evaluates zero-shot cross-lingual transfer.

B. Sentence Embedding and Pair Modeling

Transformer [6] models such as BERT produce contextualized representations that are central to build a high-performance downstream task. XLM-R is a BERT variant whose output constitutes token embeddings (up to 512 token vectors of 768 dimensions each) for a given input. To produce fixed-size sentence embeddings necessary for a task like semantic textual similarity (STS), we average the token embedding output to obtain a single 768-dimensional pooled vector.

For text regression (or classification), one learns a function that maps sentence embeddings to a target value. Sentence-pair modeling gives an important primitive that underlies supervised NLP tasks such as STS. We adopt a siamese network architecture by Sentence-BERT [7] that avoids the combinatorial explosion to form sentence pairs. Fig. 2 depicts our sentence pair modeling for STS.
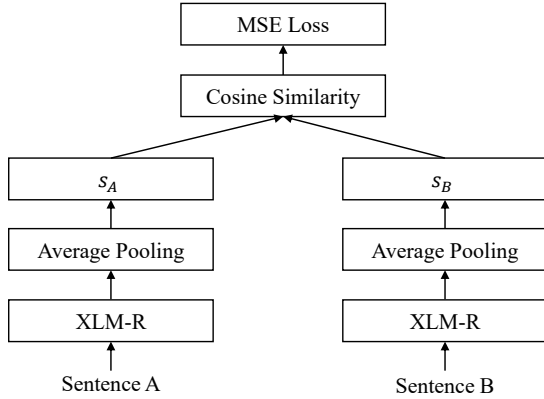

Fig. 2. Siamese net for sentence-pair modeling

C. Cross-lingual Mapping for Fine-grained Alignment of Sentence Embedddings

Cross-lingual mapping for word embeddings has been widely studied. Because context awareness is the key to language understanding, learning cross-lingual mapping for sentence-level transformations can be valuable. A sentence is less ambiguous than words since the words must be interpreted within a specific context.

We learn cross-lingual sentence mappings directly from sentence-pair examples. Note that sentence embeddings produced from contextualized cross-lingual word embeddings would imply loosely aligned sentences. Similar to the projection-based cross-lingual word embeddings framework [5], [8], we use linear algebraic methods to compute a projection matrix that achieves fine-grained alignment of sentence embeddings across different languages. We also use a single-layer neural net that can iteratively learn the same projection by gradient descent.

System of least squares via normal equation. Suppose languages A and B that are the source and the target languages of the projection $\mathbf{\Phi}$. We seek the solution to the problem $\mathbf{S}_A \mathbf{\Phi} = \mathbf{S}_B$ with

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

$$
\mathbf{s}_A^{(i)} = \begin{bmatrix} a_1^{(i)} \\ a_2^{(i)} \\ \vdots \\ a_d^{(i)} \end{bmatrix}^\top, \ \mathbf{s}_B^{(i)} = \begin{bmatrix} b_1^{(i)} \\ b_2^{(i)} \\ \vdots \\ b_d^{(i)} \end{bmatrix}^\top \tag{1}
$$

where $\mathbf{S}_A$ and $\mathbf{S}_B$ are datasets that contain n sentence embeddings for languages A and B with each sentence $\mathbf{s} \in \mathbb{R}^d$. With $\mathbf{\Phi} = \begin{bmatrix} \phi^{(1)} & \phi^{(2)} & \dots & \phi^{(j)} & \dots & \phi^{(d)} \end{bmatrix}$ whose element $\phi^{(j)} \in \mathbb{R}^d$ is a column vector, each $\mathbf{S}_A \phi^{(j)} = [b_j^{(1)} b_j^{(2)} \dots b_j^{(n)}]$ gives a problem of the least squares. Since $j = 1, \dots, d$, we have a system of d least-square problems that can be solved linear algebraically via the normal equation: $\mathbf{\Phi}^* = \left(\mathbf{S}_A^\top \mathbf{S}_A \right)^{-1} \mathbf{S}_A^\top \mathbf{S}_B$.

Solving the Procrustes problem. Given two data matrices, a source $\mathbf{S}_A$ and a target $\mathbf{S}_B$, the orthogonal Procrustes problem [9] describes a matrix approximation searching for an orthogonal projection that most closely maps $\mathbf{S}_A$ to $\mathbf{S}_B$.

Formally, we write

$$\mathbf{\Psi}^* = \underset{\mathbf{\Psi}}{\arg\min} \left\| \mathbf{S}_A \mathbf{\Psi} - \mathbf{S}_B \right\|_{\mathrm{F}} \quad \text{s.t. } \mathbf{\Psi}^\top \mathbf{\Psi} = \mathbf{I} \quad (2)$$

The solution to Eq. (2) has the closed-form $\mathbf{\Psi}^* = \mathbf{U}\mathbf{V}^\top$ with $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathrm{SVD}(\mathbf{S}_B \mathbf{S}_A^\top)$, where SVD is the singular value decomposition.

Fully-connected single-layer neural net with linearly activated neurons. Contrasted to linear algebraic solutions $\mathbf{\Phi}^*$ and $\mathbf{\Psi}^*$, a neural net can be used to compute the projection matrix iteratively via gradient descent. We consider a fully-connected single hidden-layer neural net with linear activation functions as illustrated in Fig. 3. We use the neural net as an array of linear regressors with mean square error (MSE) objectives

$$\mathbf{S}_A \mathbf{W} = \mathbf{S}_B \text{ (feedforward) s.t. } \frac{1}{2} \left\| \mathbf{s}_B^{(i)} - \mathbf{S}_A \mathbf{w}^{(j)} \right\|_2^2 < \epsilon \ \ \forall i,j \tag{3}$$

where $\mathbf{W} = [\mathbf{w}^{(1)} \mathbf{w}^{(2)} \ldots \mathbf{w}^{(j)} \ldots \mathbf{w}^{(d)}]$ contains the weight parameters of the neural net. Instead of a cross-entropy loss, we impose the MSE loss function to optimize each $\mathbf{w}^{(j)}$ for stochastic gradient descent.
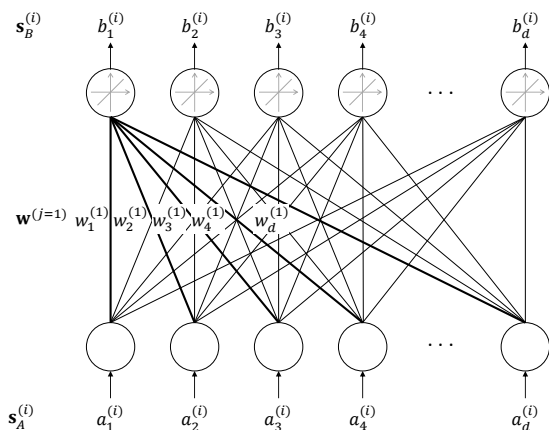

Fig. 3. Fully-connected single-layer neural net with neurons having linear activation functions.

III. Experiments

Throughout all our experiments, we use the pretrained XLM-RoBERTa (XLM-R) [1] downloaded from Hugging Face[1] [10] unmodified, upon which we build supervised NLP tasks and fine-tune. We focus on experimenting with sentence-level representations and their cross-lingual transfer quality evaluations when used in a downstream task under zero-shot settings. There is a considerable amount of the existing literature on evaluating the cross-lingual transfer quality of word representations, which we will not cover in this paper.

[1] https://huggingface.co/

A. Semantic Textual Similarity (STS)

Task & dataset. The first of our cross-lingual experiments are STS benchmark (STSb) [11], Korean STS (KorSTS) [12], SemEval-2017 Spanish, and SemEval-2017 Arabic. STSb is a set of English data originated for the STS task evaluations in the International Workshop on Semantic Evaluation (SemEval) [13]–[17] between 2012 and 2017. STSb is distributed as one of the four similarity and paraphrase tasks in the GLUE benchmark [18]. The STSb dataset includes 8,628 sentence pairs from image captions, news headlines, and user forums that are partitioned in train (5,749), dev (1,500) and test (1,379) sets.

The STSb sentence pairs are labeled with a similarity score ranging from 0 to 5 that indicates how similar the sentences are in terms of semantic relatedness. KorSTS is a translated dataset from STSb and has exactly the same structure. SemEval-2017 Spanish and Arabic are evaluation sets from SemEval-2017 Task 1 [19], which has 250 test pairs per each language.

Fine-tuning. We run the GLUE benchmark code as-is from Hugging Face to fine-tune STS tasks. This means that a text input to XLM-R is in the Sentence A–[SEP]–Sentence B format, which is the same as in pretraining. We use Rectified Adam (RAdam) optimizer with a linear learning rate warm-up for 10% of the training data and a learning rate of 4×10^{-5}. We have run 4 training epochs using a batch size of 32.

To evaluate zero-shot cross-lingual transfer, we fine-tune on the STSb train set and test using the STSb, KorSTS, SemEval-2017 Spanish, and SemEval-2017 Arabic test sets, and similarly for fine-tuning and testing on KorSTS. Furthermore, we carry out the following mixed instances: 1) fine-tune on STSb the first and KorSTS the next; 2) fine-tune on KorSTS the first and STSb the next; 3) fine-tune on sentence pair examples uniformly drawn from STSb and KorSTS.

Results. The upper portion of Table I reports the STS performances on zero-shot cross-lingual testing with 4 languages. We immediately find the presence of cross-lingual transfer strong for STS. When fine-tuned on English (the STSb train set), zero-shot testing with Korean results in 1.24% decrease in Spearman's rank correlation. On the other hand, when fine-tuned using the KorSTS train set, zero-shot testing with English results in 3.40% degradation.

For Spanish and Arabic, we observe better performance when fine-tuned on English. We find particularly low scores for Arabic and suspect that it is relatively lower resource language compared to the others. In fact, XLM-R uses 28.0GB of Arabic resources while for Korean 54.2GB is used, 53.3GB Spanish, and 300.8GB English [1].

The lower portion of Table I shows how two-stage fine-tuning mixed with two different languages affects the performance in each language. Although the performance numbers are similar regardless of the fine-tuning order, the last language fine-tuned slightly outperforms the others.

B. Machine Reading Comprehension (MRC)

Task & dataset. Reading comprehension has been one of the most challenging tasks for machine, combining natural

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

	Fine-tuning Task(s)	Evaluation Language			
		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)

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

	Fine-tuning Task(s)	Evaluation Language		
		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)

language understanding and generation with knowledge about the world. We use Stanford Question Answering Dataset (SQuAD) [3], Korean Question Answering Dataset (KorQuAD) [4], and Spanish SQuAD (SQuAD-es) [20] for the cross-lingual transfer evaluation of machine reading comprehension (MRC) tasks. Both SQuAD and KorQuAD consist of crowdsourced question-answer pairs from English and Korean Wikipedia articles, respectively. SQuAD-es is a translated dataset of SQuAD for Spanish.

Using SQuAD v1.1, KorQuAD v1.0, and SQuAD-es v1.1, we do the following eight cross-lingual MRC tasks. We prepare three copies of XLM-R and fine-tune them on 1) SQuAD, 2) KorQuAD and 3) SQuAD-es for testing with SQuAD (English), KorQuAD (Korean) and SQuAD-es (Spanish) dev sets. We then fine-tune cross-lingually again using 4) KorQuAD on the SQuAD fine-tuned XLM-R, 5) SQuAD-es on the SQuAD fine-tuned XLM-R, and 6) SQuAD on the KorQuAD fine-tuned XLM-R for another round of testing with the dev sets. Additionally, we fine-tune XLM-R with 7) mixed set of SQuAD and KorQuAD and 8) mixed set of SQuAD, KorQuAD, and SQuAD-es.

Fine-tuning. We use RAdam optimizer with a linear learning rate warm-up for 10% of the training data and a learning rate of 2×10^{-5}. We have found that running just 3 training epochs with a batch size of 48 is sufficient.

Results. The upper portion of Table II reports the cross-lingual MRC performance evaluated on the SQuAD, KorQuAD, and SQuAD-es dev sets. For fine-tuned SQuAD, zero-shot testing with Korean and Spanish degrades 9.67% and 5.30% in F1 score. (Here, the compared baseline is KorQuAD dev set tested on KorQuAD train set fine-tuned XLM-R.) Fine-tuned on KorQuAD, however, zero-shot testing with English and Spanish degrades 18.89% and 22.94%, respectively. The results with SQuAD-es shows 4.57% and 11.96% decreases

for English and Korean. Compared to the performance on STS tasks, the degraded performance gap measured in F1 scores and exact match is much higher for MRC tasks.

The lower portion of Table II reports the cross-lingual MRC performance for mixed language fine-tuning cases. The result shows a similar trend as in STS tasks. In general, fine-tuning with an additional language seems to improve the MRC performance regardless of testing language. Fine-tuning with all other languages yields the best MRC performance as shown in the last row of Table II.

C. Sentiment Analysis

Task & dataset. For sentiment analysis, we use two datasets of the similar origin, namely Large Movie Review Dataset (LMRD) [21] and Naver Sentiment Movie Corpus (NSMC) [22]. LMRD is a movie review dataset in English. The dataset provides a set of 50,000 reviews with labels indicating whether a review is positive or negative. NSMC uses the same labeling system for movie reviews written in Korean language. The dataset consists of 200,000 reviews. Using LMRD and NSMC, we have experimented five cross-lingual evaluations: fine-tune using 1) LMRD, 2) NSMC, 3) NSMC on the LMRD fine-tuned XLM-R, 4) LMRD on the NSMC fine-tuned XLM-R, and 5) mixed set of LMRD and NSMC. All of these tasks are evaluated on the LMRD and NSMC test sets.

Fine-tuning. Again, using RAdam optimizer with a linear learning rate warm-up for 5% of the training data and a learning rate of 4×10^{-5}, we run 5 training epochs with a batch size of 48.

Results. The upper portion of Table III presents the zero-shot cross-lingual transfer results on sentiment analysis tasks. The numbers represent classification accuracy in percentage. Zero-shot testing with NSMC (Korean) on the LMRD

fine-tuned XLM-R results in 12.05% accuracy degradation, whereas zero-shot testing with English shows 7.63% decrease in classification accuracy.

The lower portion of Table III presents the cross-lingual sentiment analysis performance for mixed language fine-tuning cases. Here, the performance of the last language fine-tuned is improved while the first language fine-tuned degrades a little. When fine-tuned on the train set mixed with both languages, the sentiment analysis performance improves for both languages.

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

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

D. Cross-lingual Mapping for Fine-grained Alignment of Sentence Embeddings

Using the analytical findings of Section II.C, we have determined the cross-lingual mappings Φ^* and Ψ^* linear algebraically. We have applied the mappings to align the translated sentence pairs of STSb and KorSTS. Precisely, we set the source $\mathbf{S}_A$ English sentences from STSb, and the target $\mathbf{S}_B$ Korean sentences from KorSTS. The quality of alignment via linear projections Φ^* and Ψ^* is very similar. Based on the average cosine similarity of the translated sentence pairs, we find Φ^* slightly better than Ψ^*.

We determine $\mathbf{W}$ by stochastic gradient descent on the single-layer neural net of Fig. 3. Using the translated sentence pairs, we set the input $\mathbf{S}_A$ to the neural net English sentences from STSb, and the output $\mathbf{S}_B$ Korean sentences from KorSTS. The average cosine similarity of the translated sentence pairs after alignment via the Φ^* projection is 0.7131 whereas the average cosine similarity for the neural net is 0.7265. Without alignment by the projection matrix or the neural net, the average cosine similarity would have been 0.4636. Fig. 4 illustrates the t-SNE plots that visualize the effect of the sentence alignment. The top plots are unaligned English, aligned English and Korean sentences by the Φ^* projection, whereas the bottom plots represent unaligned and aligned English and Korean sentences via the neural net.

TABLE IV
STS EVALUATION WITH CROSS-LINGUAL SENTENCE PAIRS.

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

In Table IV, we compare the cosine similarity of aligned English and Korean translated sentence pairs of STSb and KorSTS through the fine-grained cross-lingual mapping to zero-shot transfer. Cross-lingual mapping that we compute linear algebraically or by the use of a neural net outperforms zero-shot cross-lingual transfer by 9.3–20% in cosine similarity matching of the translated sentences pairs of STSb and KorSTS.

E. Discussion

Generally, we find that cross-lingual transfer is present in important supervised NLP tasks that require high-level natural language understanding, namely STS, MRC, and sentiment classification. Our empirical evaluation suggests the presence of cross-lingual transfer be most pronounced in STS. The next is sentiment analysis, and MRC comes the last. It seems that more complex a task is, and the quality of cross-lingual transfer becomes less effective. For STS, we have observed the transfer quality in two different measures, the Spearman's rank and Pearson correlation coefficients, and found them concordant. For MRC, while zero-shot transfer performance measured by F1 score is reasonable, it suffers significantly more for the case of the exact match (EM) metric. Interestingly, if we fine-tune XLM-R with both source and target languages, the last language fine-tuned has the strongest impact on the performance.

IV. CONCLUSION

This paper focuses on the empirical validation of the cross-lingual transfer properties induced by XLM pretraining. We have experimented with XLM-RoBERTa (XLM-R), a large cross-lingual language model, and extended semantic textual similarity (STS), SQuAD and KorQuAD for machine reading comprehension (MRC), and sentiment analysis to cross-lingual settings. Our results suggest the presence of cross-lingual transfer be most pronounced in STS, the sentiment analysis the next, and MRC the last. We compute matrix projections linear algebraically that directly map sentence embeddings of one language to another for analyzing the effect of fine-grained alignment of sentences in zero-shot cross-lingual transfer. We have shown that such mapping can also be determined iteratively using a simple neural net. Our future work includes more systematic evaluations on broader range of low- and high-resource languages to generalize the quality of cross-lingual transfer manifested through important NLP tasks.

REFERENCES

[1] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[2] A. Conneau and G. Lample, "Cross-lingual Language Model Pretraining," in *Advances in Neural Information Processing Systems 32*, 2019.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[4] S. Lim, M. Kim, and J. Lee, "Korquad1.0: Korean QA Dataset for Machine Reading Comprehension," *arXiv preprint arXiv:1909.07005*, 2019.
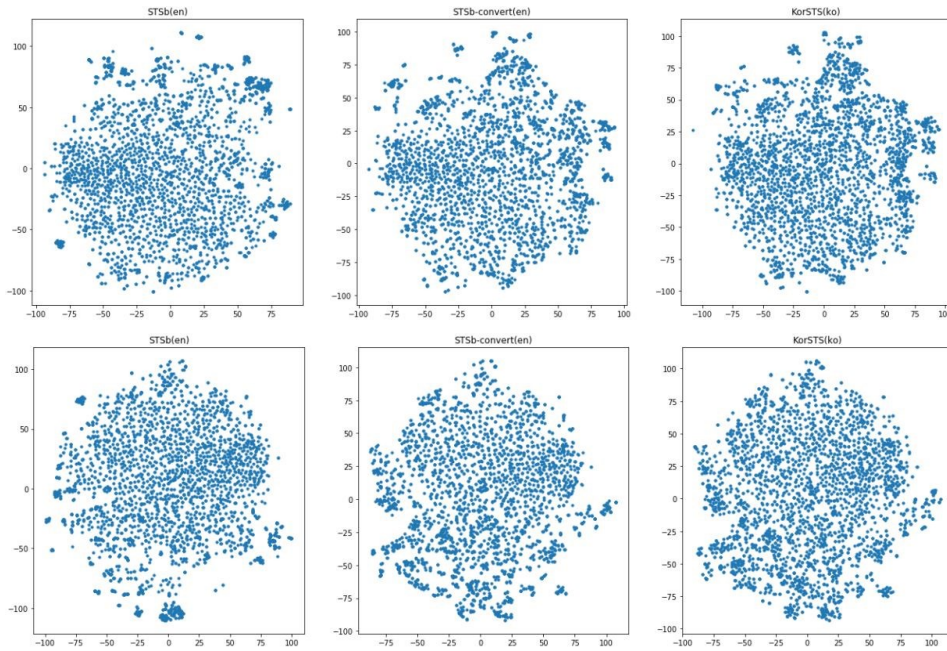
Fig. 4. t-SNE plots of English and Korean translated pairs from STSb and KorSTS. The leftmost plot on the top row is unaligned English sentences (source), and the middle represents aligned English via linear projection $\mathbf{\Phi}^*$, the rightmost Korean (target). The middle and the rightmost plots are aligned, showing similar patterns in t-SNE. The bottom plots are unaligned English, aligned English, and Korean sentences via the fully-connected single layer neural net whose weight parameters $\mathbf{W}$ are learned by stochastic gradient descent.

[5] G. Glavaš, R. Litschko, S. Ruder, and I. Vulić, "How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, 2017.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[8] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting Similarities among Languages for Machine Translation," *arXiv preprint arXiv:1309.4168*, 2013.

[9] P. Schönemann, "A Generalized Solution of the Orthogonal Procrustes Problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[10] Hugging Face, "Open Source NLP," https://huggingface.co, 2020.

[11] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017.

[12] J. Ham, Y. J. Choe, K. Park, I. Choi, and H. Soh, "KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding," *arXiv preprint arXiv:2004.03289*, 2020.

[13] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "SemEval-2012 task 6: A pilot on semantic textual similarity," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012.

[14] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "*SEM 2013 shared task: Semantic textual similarity," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, 2013.

[15] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2014 task 10: Multilingual semantic textual similarity," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014.

[16] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, and J. Wiebe, "SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015.

[17] E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016.

[18] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[19] D. M. Cer, M. T. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity, multilingual and cross-lingual focused evaluation," 2017.

[20] C. P. Carrino, M. R. Costa-jussà, and J. A. Fonollosa, "Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering," *arXiv preprint arXiv:1912.05200*, 2019.

[21] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[22] Naver Sentiment Movie Corpus (NSMC), "Naver Corp." https://github.com/e9t/nsmc, 2020.