Is Cross-lingual Evaluation Only About Cross-lingual?

Anonymous ACL submission

Abstract

Multilingual pre-trained language models (mPLMs) have achieved great success on various cross-lingual tasks. However, we find that the higher performance on these tasks cannot be regarded as the better cross-lingual ability because models' task-specific abilities can also influence the performance. In this work, we do a comprehensive study on two representative cross-lingual evaluation protocols: sentence retrieval and zero-shot transfer. We find that current cross-lingual evaluation results strongly depend on mPLMs' task-specific abilities so that the performance can be improved without any improvement in models' cross-lingual ability. To have more accurate comparisons of cross-lingual ability between mPLMs, we propose two new indexes based on the two evaluation protocols: calibrated sentence retrieval performance and transfer rate, and experimentally show that our proposed indexes effectively eliminate the effects of task-specific abilities on the cross-lingual evaluation.

1 Introduction

Multilingual pre-trained language models (mPLMs) have obtained remarkable achievements in the fields of multilingual and cross-lingual NLP (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Ouyang et al., 2021). Since mPLMs encode texts in different languages into a unified representation space, the models can generate powerful cross-lingual representations and support NLP research and application beyond English (Joshi et al., 2020), e.g., the transfer learning from high-resource to low-resource languages.

Researchers have constructed a variety of tasks to evaluate the cross-lingual ability of mPLMs, and the performance of mPLMs is increasing fast with more pre-training data, larger model size, and new pre-training objectives (Conneau and Lample, 2019; Chi et al., 2021b; Xue et al., 2021; Han et al., 2021). However, we find that higher-performing models do not always possess better cross-lingual ability because models' task-specific abilities also significantly contribute to the performance.

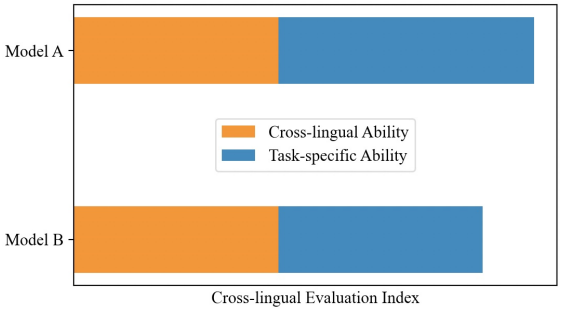

Figure 1: A visualization of how task-specific abilities influence cross-lingual evaluation.

In this work, we analyze the effects of task-specific abilities on two widely used evaluation protocols. The first one is cross-lingual sentence retrieval (Zweigenbaum et al., 2017; Artetxe and Schwenk, 2019a), which evaluates mPLMs' cross-lingual alignment ability by comparing the similarity between models' cross-lingual representations of sentences in different languages. The other is zero-shot cross-lingual transfer, which evaluates mPLMs' cross-lingual transferability by testing them on different languages with downstream tasks such as natural language inference (Conneau et al., 2018; Yang et al., 2019b) and question answering (Lewis et al., 2020b).

From the two protocols, researchers derive three indexes of cross-lingual ability: sentence retrieval performance, transfer gap, and zero-shot transfer performance. However, we experimentally find that all the three indexes are affected by models' task-specific abilities, making their evaluation results not only about models' cross-lingual ability. Figure 1 is an overview of our observations from experiments. It illustrates how these indexes assess models with the same cross-lingual ability differently: (i) enhancing a models' sentence embed-

dings makes higher sentence retrieval performance. (ii) improving a model's NLU ability results in a better transfer gap and a higher zero-shot transfer performance.

The differences in mPLMs' task-specific abilities hinder us to make fair comparisons between their cross-lingual ability using existing indexes, so we explore ways to eliminate the effects of task-specific abilities in cross-lingual evaluations:

(i) We find that the quality of mPLMs' monolingual sentence embeddings significantly affects their performance on cross-lingual sentence retrieval while most mPLMs do not possess good pre-trained sentence embedding. Thus, we propose to advance models' sentence embeddings on English data by contrastive learning before evaluation. We refer to the performance of models after fine-tuning as calibrated sentence retrieval performance.

(ii) We find that the monolingual NLU abilities of mPLMs are also much different. Fortunately, the translate-train performance can be used to measure these NLU abilities. Therefore, we propose a new index of cross-lingual ability, namely transfer rate, which is the ratio of the zero-shot transfer performance to the translate-train performance.

We examine the validity and rationality of our proposed indexes by experiments and show these indexes better reflect the cross-lingual ability of mPLMs than currently-used indexes. We hope this study will help future work to better analyze mPLMs' cross-lingual ability.

2 Background

In this section, we introduce current evaluation protocols, setups, indexes of the cross-lingual ability, and the mPLMs studied in this work.

2.1 Evaluation

A variety of tasks have been used to evaluate mPLMs, such as dependency parsing (Schuster et al., 2019), named entity recognition (Lin et al., 2019), sentiment analysis (Barnes et al., 2018), natural language inference (Conneau et al., 2018), document classification (Schwenk and Li, 2018), question answering (Liu et al., 2019; Lewis et al., 2020b; Artetxe et al., 2020; Clark et al., 2020), and cross-lingual sentence retrieval (Zweigenbaum et al., 2017; Artetxe and Schwenk, 2019a). We categorize these tasks into two evaluation protocols, cross-lingual sentence retrieval and zero-shot cross-lingual transfer.

Sentence Retrieval is to identify the translation of each sentence in a source language from sentences in another language through the sentence representations given by models. Pires et al. (2019) first evaluate mBERT on sentence retrieval to demonstrate its powerful cross-lingual alignment ability. They feed each sentence to mBERT without fine-tuning and use the average of all input tokens' hidden states from a specific layer as its sentence representation. Then, the sentence representations are used to retrieve the translation of each sentence by finding its nearest neighbor. Latter work (Hu et al., 2020; Dufter and Schütze, 2020; Lewis et al., 2020a; Chi et al., 2021c) evaluates the cross-lingual ability of mPLMs on sentence retrieval with the same evaluation setup.

Zero-shot Cross-lingual Transfer usually uses English as a source language for fine-tuning and evaluates the fine-tuned models on multilingual datasets. Based on this protocol, researchers develop two indexes of cross-lingual ability: zero-shot transfer performance and transfer gap. (i) **Zero-shot transfer performance** means the performance of an mPLM on target languages after fine-tuning on a source language. Multiple works use it to evaluate the cross-lingual ability of mPLMs. Pires et al. (2019); K et al. (2020); Ma et al. (2021) do ablation studies on factors that contribute to mPLMs' cross-lingual ability using the zero-shot transfer performance as an index. Huang et al. (2019); Chi et al. (2021b); Ahmad et al. (2021) compare the effectiveness of different cross-lingual pre-training tasks based on the zero-shot transfer performance on XNLI. (ii) **Transfer gap** is proposed by XTREME (Hu et al., 2020) to further analyze the cross-lingual transfer. When an mPLM is fine-tuned on one language and evaluated on other languages, there will be a gap between the performance of the model on the source language and the target languages. The transfer gap is the difference between the performance on the test sets of English and the average performance of other languages. They suppose a lower cross-lingual transfer gap indicates more task-related knowledge is transferred from English to target languages, so an mPLM with a perfect cross-lingual ability will have a transfer gap of 0. The transfer gap has been adopted by much recent work (Fang et al., 2021; Chi et al., 2021b,a; Ahmad et al., 2021; Zheng et al., 2021; Ouyang et al., 2021; Zhao et al., 2021) to measure the cross-lingual transferability of mPLMs.

Other Protocols. Apart from the above-mentioned protocols, there are other evaluation protocols which are sometimes used, including word retrieval (Dufter and Schütze, 2020), word alignment (Jalili Sabet et al., 2020), word translation (Gonen et al., 2020), machine translation (Conneau and Lample, 2019), and cross-lingual information retrieval (Sun and Duh, 2020).

2.2 Models

Recently, various multilingual models pre-trained on a wide range of languages have been proposed (Devlin et al., 2019; Conneau and Lample, 2019; Huang et al., 2019; Conneau et al., 2020; Siddhant et al., 2020; Chi et al., 2021b; Feng et al., 2020; Xue et al., 2021; Ouyang et al., 2021). From these models, we select three representative mPLMs pre-trained with different objectives, training data, and task-specific abilities.

mBERT (Devlin et al., 2019) is the first Transformer-based mPLM, which has achieved great success on amounts of cross-lingual tasks and has been widely used in cross-lingual research. mBERT is a 12-layer Transformer pre-trained on the Wikipedia dumps of 104 languages using Masked Language Model (MLM) and Next Sentence Prediction (NSP) objectives.

LaBSE (Feng et al., 2020) is a 12-layer Transformer using a dual-encoder architecture, which has powerful sentence representation ability and establishes new state-of-the-art performance on cross-lingual sentence retrieval. It is pre-trained on three pre-training tasks together, MLM, Translation Language Model (Conneau and Lample, 2019), and Translation Ranking (Yang et al., 2019a). Its training data consists of 17B monolingual sentences and 6B bilingual translation pairs over 109 languages.

XLM-R$_{\text{Base}}$ and XLM-R (Conneau et al., 2020) are 12-layer and 24-layer Transformers, which have better NLU ability than mBERT. Compared to mBERT, they are pre-trained on larger corpora, the filtered CommonCrawl data (Wenzek et al., 2020) of 100 languages using the MLM objective.

3 Sentence Retrieval

In this section, we analyze the effect of the task-specific ability on sentence retrieval performance, i.e., the monolingual sentence embedding quality. Previous studies have two observations: (i) PLMs pre-trained with only language modeling objectives have poor sentence embedding quality while PLMs with sentence-level pre-training tasks have good embedding quality (Gao et al., 2021); (ii) fine-tuning mPLMs on English intermediate-tasks, such as question answering and natural language inference, can significantly improve the performance of cross-lingual sentence retrieval (Phang et al., 2020; Ruder et al., 2021). Correspondingly, there are two research questions: (i) does the difference in mPLMs' sentence embedding quality influence their comparisons on cross-lingual sentence retrieval? (ii) does fine-tuning on English intermediate-tasks improve cross-lingual ability or sentence embedding quality? By answering these two questions, we explore to find out a better way to evaluate cross-lingual ability in sentence retrieval.

3.1 Datasets

We briefly introduce datasets we used for fine-tuning (SQuAD v1.1, AllNLI+STSb), validation (WMT20), and testing (Tatoeba, STS-2017, WikiANN-NER) in this experiment.

SQuADv1.1 (Rajpurkar et al., 2016) is a question-answering dataset with passages extracted from Wikipedia articles and crowdsourced question-answer pairs, where the answer to each question is a text span from the corresponding passage. The evaluation setup proposed by Ruder et al. (2021) involves the training set of SQuAD v1.1.

AllNLI+STSb consists of two sentence-pair datasets. AllNLI (Reimers and Gurevych, 2019) is an English natural language inference corpus that combines the training set of Stanford Natural Language Inference (Bowman et al., 2015) and Multi-Genre Natural Language Inference (Williams et al., 2018). The Semantic Textual Similarity Benchmark (STSb) (Cer et al., 2017) is a dataset that contains sentence pairs assigned with similarity scores. We use the training sets of these two datasets to enhance mPLMs' monolingual sentence embeddings.

Tatoeba (Artetxe and Schwenk, 2019b) comprises up to 1,000 English-aligned sentence pairs for 112 languages and is widely used for current sentence retrieval evaluation. We conduct sentence retrieval evaluation on parallel sentences of 36 different language pairs from it.

STS-2017 (Cer et al., 2017) is a multilingual Semantic Textual Similarity (STS) task with monolingual test data for en, ar, es, and cross-lingual test data for en-de, fr-en, it-en, and nl-en. We evaluate mPLMs' sentence embeddings on the monolingual test data for the three languages.

	No Fine-tuning			Question Answering Fine-tuning			Sentence Embedding Fine-tuning		
Task	CLSR	STS	NER	CLSR	STS	NER	CLSR	STS	NER
mBERT	37.5	50.7	62.4	40.8 (+3.3)	57.7 (+7.0)	61.1 (-1.3)	43.1 (+5.6)	70.9 (+20.2)	61.7 (-0.7)
XLM-R_{Base}	53.4	53.7	61.0	66.7 (+13.3)	56.4 (+2.7)	59.6 (-1.4)	74.7 (+21.3)	64.8 (+11.1)	58.5 (-2.5)
XLM-R	35.6	52.6	66.2	77.7 (+42.1)	65.8 (+13.2)	64.7 (-1.5)	83.0 (+47.7)	71.7 (+19.1)	65.7 (-0.5)
LaBSE	95.4	77.6	64.0	94.9 (-0.5)	69.3 (-8.3)	64.0 (-0.0)	95.2 (-0.2)	83.8 (+6.2)	63.2 (-0.8)

Table 1: Average scores of four models on cross-lingual sentence retrieval (CLSR), STS, and cross-lingual NER (NER). We report the results of the models without fine-tuning, and changes in the results after fine-tuning with the two different methods respectively. By convention, sentence retrieval, NER results are reported in accuracy, and STS results are reported in Spearman's correlation coefficient $\times$ 100. The full results can be found in the appendix A.

WikiANN-NER (Pan et al., 2017) is a cross-lingual named entity recognition dataset generated from Wikipedia covering 282 languages. As most mPLMs do not support all of these languages, our evaluations are restricted to 40 languages from XTREME. We use this dataset for cross-lingual transfer evaluation at word-level.

WMT20 (Barrault et al., 2020) is used as the validation set for sentence retrieval because the Tatoeba dataset has no validation data. Specifically, we use the test sets of 7 language pairs from WMT20 as our validation sets, from en to cs, de, ja, pl, ru, ta, zh.

3.2 Experiment with Evaluation Setups

To investigate the effect of fine-tuning, we implement three variants: no fine-tuning, question answering fine-tuning, sentence embedding fine-tuning. Among them, sentence embedding fine-tuning is the first time to be used in cross-lingual sentence retrieval. The details are as follows:

No Fine-tuning. We directly use the mPLMs to encode each sentence and take the hidden states from the best-scoring layer of each model on the validation set for Tatoeba and STS evaluations. Specifically, we use the hidden states from the 8th layer for mBERT, the 7th layer for XLM-R_{Base}, the 13th layer for XLM-R, and the last layer for LaBSE. For NER evaluation, we fine-tune the models on the NER English training set and evaluate on the test sets of 40 languages from XTREME.

Question Answering Fine-tuning. We fine-tune the models on the training set of SQuAD v1.1 following (Ruder et al., 2021). We save and validate the training checkpoints every 500 training steps and pick the one with the highest sentence retrieval accuracy on our validation set (WMT20). For Tatoeba and STS evaluations, we extract the sentence representations from the same layer as we mentioned above for each model. For NER evaluation, we continue training the selected model on the NER training set.

Sentence Embedding Fine-tuning. To directly optimize sentence embeddings, we fine-tune the models with the siamese network structure proposed by Reimers and Gurevych (2019) on AllNLI and STS benchmark datasets, which are both English datasets. When fine-tuning on the NLI data, we use the Multiple Negatives Ranking Loss proposed by Henderson et al. (2017), which produces better sentence representations than the original softmax loss in (Reimers and Gurevych, 2020). When fine-tuning on the STS data, we use the cosine similarity loss. We save the training checkpoints when every 10% of training data is processed and pick the one with the highest sentence retrieval accuracy on the validation set. We evaluate the selected checkpoints on STS, Tatoeba, and NER with the identical evaluation setups as we described above (question answering fine-tuning).

To track the changes in monolingual sentence embedding quality and cross-lingual ability, we evaluate models on STS-2017 and WikiANN-NER, respectively.

3.3 Results and Discussions

We show the experimental results in Table 1. The followings are our observations from the results.

1. The huge differences in sentence embedding quality influence the comparisons of the performance of different mPLMs. mPLMs produce better sentence embeddings after adding sentence-level pre-training objectives and achieve better performance of cross-lingual sentence retrieval. For example, LaBSE is better than XLM-R on both CLSR and STS. However, LaBSE doesn't achieve better performance than XLM-R on NER. Hence, based on the result of CLSR, we cannot conclude that LaBSE has better cross-lingual ability than XLM-R. Furthermore, even if two mPLMs are pre-trained with the same objective on the same data, their evaluation results without fine-tuning

can still be inconsistent between CLSR and NER. In our experiment, XLM-R gets worse results on both STS and sentence retrieval than XLM-R_{Base}, but it outperforms XLM-R_{Base} on NER. Besides, the performance of CLSR is highly related to the performance of STS, which only measures the monolingual sentence embedding quality. In summary, sentence embedding quality significantly influences the performance of cross-lingual sentence retrieval. To make sentence retrieval results reflect the cross-lingual ability of mPLMs, we have to ensure the models can generate good sentence embeddings.

2. Fine-tuning improves models' sentence embeddings instead of cross-lingual ability. As shown in the table, both question answering fine-tuning and sentence embedding fine-tuning significantly improve the sentence retrieval performance on Tatoeba for all models except LaBSE, which is itself a good sentence embedding model. XLM-R benefits most from fine-tuning, where both fine-tuning methods lead to improvements of over 40 points on sentence retrieval. However, contrary to what Phang et al. (2020) might suggest, we argue that the huge improvement of mPLMs in sentence retrieval possibly comes from better monolingual sentence embeddings generated by the models rather than the higher cross-lingual ability. When looking at the NER results, we can only observe a slight decrease, indicating that fine-tuning does not actually improve the cross-lingual ability of the models. Meanwhile, we can see that the average STS performance across three languages is largely improved by fine-tuning. These results demonstrate that fine-tuning on English data enhances the sentence embeddings of mPLMs over all languages, which induces boosts in sentence retrieval performance. Hence, we can use fine-tuning to improve the sentence embeddings quality without influencing cross-lingual ability for fair evaluation.

3. Sentence embedding fine-tuning leads to larger improvements than question answering fine-tuning. Compared with Question Answering fine-tuning, the sentence embedding fine-tuning approach provides better sentence representations for all models. For mBERT and LaBSE, the STS results of the sentence embedding fine-tuning are more than 10 points higher than that of question answering fine-tuning. Similarly, we can see the same thing happens on the sentence retrieval performance, especially for XLM-R_{Base}, where the average sentence retrieval performance of sentence embedding fine-tuning is 8 points higher than that of question answering fine-tuning. These results suggest that models fine-tuned on SQuAD can not provide sentence representations that are good enough for sentence retrieval evaluation, so we propose to fine-tune models that generate poor sentence representations using the sentence embedding fine-tuning approach on AllNLI+STSb data.

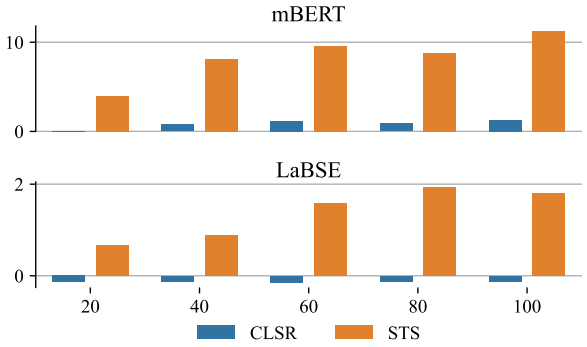

Figure 2: A visualization of the performance change of the two models on sentence retrieval and STS when fine-tuning them on different amounts of data. The x-axis is the percentage of training data used, and the y-axis is the difference in performance.

4. Sentence embedding fine-tuning provides good enough sentence representations for CLSR. According to the table, the STS results of mBERT and LaBSE increase a lot, but this does not lead to a significant change in their sentence retrieval results. We suppose that when a model is able to generate semantically meaningful sentence representations for each language, a further improvement on its sentence embeddings has little impact on its sentence retrieval results. To further validate our hypothesis, we divide the training data (AllNLI+STSb) into ten pieces and record the STS and sentence retrieval performance of the two models fine-tuned on 10% to 100% of the training data. We visualize the difference between the results of models fine-tuned on 10% of the training data and those more than 10% data in Figure 2. We can observe that the sentence retrieval performance of the two models stabilizes, whereas their STS results show a rising trend with more training data. It indicates that the sentence embedding fine-tuning approach can help mPLMs generate sentence embeddings that are good enough for sentence retrieval evaluation so the evaluation results can fully reflect the cross-lingual ability of models. We refer to the performance with sentence embedding fine-tuning

as **calibrated sentence retrieval performance**.

4 Zero-shot Cross-lingual Transfer

In this section, we examine the rationality of two evaluation indexes of cross-lingual ability, transfer gap and zero-shot transfer performance, and analyze the effect of task-specific abilities in zero-shot cross-lingual transfer.

4.1 Experiment with Transfer Gap

The basis of the transfer gap is that target languages must not outperform source languages in the cross-lingual transfer, so the difference between models' performance on English and other languages can measure the amount of knowledge transferred. Most mPLMs are pre-trained with the highest English resource, so it is not surprising that the models always perform higher on English than other languages. Nevertheless, it is worth studying that whether the transfer gap is still valid when English is a low-resource language in a model, that is the scenario of transferring knowledge from a low-resource language to high-resource languages. Hence, we choose a low-resource language, Urdu, as the source language in our experiment.

As stated in Hu et al. (2020), the transfer gap only applies to multilingual datasets with the same test sets across all languages (translated from English annotated data) because the zero-shot transfer performance can not be comparable across languages if test sets differ. Hence, we experiment with the transfer gap on XNLI, where the test sets for 14 languages are human-translated from its English test set. The XNLI training sets for 14 languages are machine translated from its English training set; meanwhile, the machine translation system adopted by Conneau et al. (2018) generates poor translations (low BLEU scores) from English to low-resource languages like Urdu. To exclude the potential effects of the translation quality, we additionally create a smaller version of XNLI by concatenating its human-translated validation sets and test sets of each language together and splitting them into *train, validation, test* sets, with a ratio of 8:1:1. We refer to this dataset as **XNLI Manual** and report the results of on both datasets.

4.2 Issues of Transfer Gap

Our experimental results are shown in Table 2. We have the following observations from the results.

1. Performance on the source language cannot be an upper bound for target languages. We can observe that most target languages in the cross-lingual transfer have better accuracy than the source language for all models, especially for LaBSE fine-tuned on XNLI, where no target language underperforms the source language. In more detail, the extremely low-resource language sw always underperforms ur, whereas high-resource languages such as en, es, fr, de significantly outperform the source language by approximately 10% on XNLI and 5% on the XNLI Manual. These results reveal two things about the cross-lingual transfer: (i) The performance of a model on a source language should not be an upper bound of its target languages' performance. (ii) The zero-shot transfer performance on a target language might relate to its pre-training resource. Hence, a transfer gap of zero does not amount to a perfect cross-lingual transfer.

2. Transfer gap can be negative. As shown in Table 2, mBERT fine-tuned on XNLI is the only model with a positive and transfer gap between the source and target languages, while all other models yield a negative transfer gap. These results show the possibility of obtaining a negative transfer gap is possible when transferring knowledge from low-resource to high-resource languages. We suppose that if an mPLM is pre-trained with the lowest English resources among all languages, it will also give a negative or close to zero transfer gap when fine-tuning on English training sets. A small transfer gap might come from bad source language performance rather than good cross-lingual transferability. Hence, the transfer gap cannot be a suitable index of mPLMs' cross-lingual ability.

4.3 Experiment with Zero-shot Transfer Performance

From the previous experiment, we observe that high-resource languages can always succeed in zero-shot transfer performance, no matter what the source language is. Hence, we suppose that the zero-shot transfer performance depends not only on mPLMs' cross-lingual representations but also on their NLU ability. To validate our supposition, we calculate the Pearson's correlation coefficients between the zero-shot transfer performance and the translate-train-all performance, the calibrated sentence retrieval performance.

The translate-train-all performance of a model is obtained by fine-tuning it on the concatenation of

Model	ur	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	avg	gap
Zero-shot Cross-lingual Transfer (Fine-tuned on XNLI Urdu training data)																	
mBERT	61.8	68.1	64.7	65.1	64.0	61.4	63.2	64.8	57.9	59.7	63.1	48.3	65.2	60.6	46.5	61.0	+0.9
XLM-R_{Base}	66.9	75.8	72.1	73.6	71.8	71.5	73.0	72.0	69.0	68.5	71.5	69.8	71.0	68.7	63.2	70.5	-3.9
XLM-R	73.5	88.6	83.2	84.1	82.8	82.6	83.3	80.7	79.2	79.2	79.7	77.2	80.0	76.6	71.8	80.1	-7.1
LaBSE	67.8	78.3	75.1	75.5	74.2	74.2	75.3	74.2	73.1	71.5	74.1	68.4	72.9	70.5	69.7	72.9	-5.5
Zero-shot Cross-lingual Transfer (Fine-tuned on XNLI Manual Urdu training data)																	
mBERT	54.5	59.8	57.6	59.4	59.4	56.7	59.0	59.5	53.7	55.3	58.0	49.9	57.0	57.4	44.8	56.1	-1.7
XLM-R_{Base}	60.3	67.9	64.7	64.4	62.7	63.5	65.8	63.4	60.9	59.9	64.0	61.6	62.1	62.0	55.3	62.5	-2.4
XLM-R	67.2	75.6	73.1	72.5	73.9	72.6	73.6	71.2	69.5	70.4	70.9	69.4	71.1	69.6	64.0	70.9	-4.0
LaBSE	62.2	66.6	65.6	67.8	65.7	64.5	66.4	64.3	65.1	62.6	64.9	57.4	62.9	64.0	60.3	64.0	-1.9

Table 2: The zero-shot transfer results of four models on XNLI and XNLI Manual. We report the accuracy on each of the 15 XNLI languages, the average accuracy, and the transfer gap. Note that target languages underperform the source language (Urdu) are underlined for each model.

Model	fr	de	ar	zh	sw	avg
Calibrated Sentence Retrieval (Tatoeba)						
mBERT	70.9	82.9	33.2	79.1	14.9	50.0
XLM-R_{Base}	87.0	96.6	62.6	87.1	34.9	79.4
XLM-R	91.8	97.4	77.7	93.0	35.4	86.5
LaBSE	96.0	99.2	90.1	96.6	88.2	95.8
Zero-shot Cross-lingual Transfer (XNLI)						
mBERT†	73.4	70.0	64.3	67.8	49.7	64.3
XLM-R_{Base}*	79.7	78.7	73.8	76.7	66.5	75.5
XLM-R*	84.1	83.9	79.8	80.2	73.9	80.3
LaBSE	81.0	79.3	75.4	77.0	71.8	76.5
Translate-Train-All (XNLI)						
mBERT†	77.8	77.6	73.8	77.6	70.5	74.6
XLM-R_{Base}*	81.4	80.3	77.3	80.2	73.1	78.7
XLM-R*	85.1	85.7	83.1	83.7	78.0	83.2
LaBSE	81.9	81.9	78.1	80.2	74.7	79.1
Correlation with Zero-shot Transfer Performance						
Sent Retrieval.	90.4	90.0	87.4	85.3	63.6	91.0
Translate-Train.	98.5	94.4	94.2	91.4	89.5	95.1

Table 3: The sentence retrieval results on Taoteba, zero-shot transfer, translate-train-all results on XNLI, and Pearson's correlation coefficients between the zero-shot transfer results and the translate-train-all, the sentence retrieval results. We report the results on five languages and the average results of the 14 XNLI languages (source language English is excluded). Results with †∗ are from Hu et al. (2020); Conneau et al. (2020). We boldface the best score of each column on each task. The full results can be found in the appendix A.

training sets from all languages and then evaluating on multilingual test sets. This evaluation setup does not examine the cross-lingual transferability of models, so we use models' translate-train-all performance to measure their NLU ability.

The calibrated sentence retrieval performance well reflects the cross-lingual ability of a model as we have shown in Sec. 3, so we use mPLMs' calibrated sentence retrieval performance to measure their cross-lingual representation quality.

4.4 Issues of Zero-shot Transfer Performance

We show the experimental results in Table 3. Based on these results, we analyze what affects the zero-shot performance and the potential dangers of using it as an index of cross-lingual ability.

1. NLU ability significantly affects the zero-shot transfer performance. According to the correlation coefficients, we can clearly see the correlation between the zero-shot transfer performance and translate-train-all performance is stronger than that of the calibrated sentence retrieval, whereas the translate-train results are not related to the cross-lingual ability of mPLMs. Moreover, there are two main inconsistencies between the zero-shot transfer performance and the cross-lingual alignment ability of the models. (i) LaBSE outperforms all other models on sentence retrieval, while XLM-R outperforms all other models on both zero-shot transfer and translate-train-all. (ii) LaBSE obtains more than 50 points than the other models on the en-sw sentence retrieval, but its zero-shot transfer performance on sw is lower than that of XLM-R. The strong correlation between the zero-shot transfer and translate-train results reveals that the zero-shot transfer performance is strongly affected by the NLU ability of models. This means a large model pre-trained on a huge amount of monolingual data can easily succeed at the zero-shot setting without an outstanding cross-lingual ability.

2. Measuring cross-lingual ability by the zero-shot transfer performance is problematic. Using the zero-shot transfer performance as an index of the cross-lingual ability not only compromises fair comparisons between models but also potentially leads to inadequate conclusions. (i) Many pre-training tasks like SOP and pre-training strategies such as n-gram masking (Raf-

Model	fr	de	ar	zh	sw	avg
Transfer Rate (XNLI)						
mBERT	94.3	90.2	87.1	87.4	70.5	86.0
XLM-R_{Base}	97.9	98.0	95.5	95.6	91.0	96.0
XLM-R	98.8	**97.9**	96.0	95.8	94.7	96.5
LaBSE	**98.9**	96.8	**96.5**	**96.0**	**96.1**	**96.7**
Correlation with Calibrated Sentence Retrieval						
Zero-shot.	90.4	90.0	87.4	85.3	63.6	91.0
Transfer Rate.	98.6	95.8	92.7	87.8	69.8	95.7

Table 4: The transfer results on XNLI, and Pearson's correlation coefficients between the calibrated sentence retrieval results and the zero-shot transfer, the transfer rate results. We report the results on five languages and the average results of the 14 XNLI languages. We boldfact the best score of each column on each task. The full results can be found in the appendix A.

fel et al., 2020), DeBERTa (He et al., 2021) have been shown to enhance monolingual language models' NLU ability effectively and improve the models' performance on downstream tasks including MNLI, SQuAD, so implementing the tasks and strategies on mPLMs possibly improve the zero-shot transfer performance on multilingual NLI and QA tasks. However, the improvement induced by better task-specific abilities will be regarded as an improvement in the cross-lingual ability of mPLMs when using the zero-shot performance as an index. (ii) Almost all mPLMs are pre-trained with different amounts of data and vocabulary sizes on each language, resulting in a large difference between their NLU abilities, which is ignored by existing studies (Pires et al., 2019; K et al., 2020) on the effect of language similarity in cross-lingual transfer. We notice that languages in the same language family as English, such as de, es, fr, ru, are four target languages with the highest resource in Wikipedia (Wu and Dredze, 2020). Thus, even if an equal amount of task knowledge is transferred from English to all other languages, the above-mentioned languages can still outperform others in terms of the zero-shot transfer. In other words, the effect of language similarity might be overestimated.

4.5 Transfer Rate

To eliminate the effect of NLU ability, we propose to use the translate-train-all performance as the estimation of NLU ability for each language, so the cross-lingual ability can be measured by the ratio of the zero-shot transfer performance to the translate-train performance. We refer to this index as "transfer rate", which measures the percentage of knowledge transferred from a source language to target languages.

We show the transfer rate of five languages of the models and Pearson's correlation coefficients between the transfer rate and sentence retrieval results in Table 4. The detailed results containing all languages can be found in the appendix. We can see the transfer rate scores better demonstrate the cross-lingual ability of models than the zero-shot scores by showing a stronger correlation with the sentence alignment ability of the models. LaBSE gets a higher average transfer rate than XLM-R, which suggests the cross-lingual pre-training objective could be an important step for mPLMs to obtain good cross-lingual ability.

Additionally, we notice that some multilingual datasets do not contain any training set for languages other than English, while an estimation of mPLMs' NLU ability on other languages is needed in the calculation of transfer rate. We hope that the research community can see the necessity of providing training sets for all languages when creating new multilingual datasets for cross-lingual evaluations.

5 Conclusion

In this work, we revisit two widely-used evaluation protocols of the cross-lingual ability of mPLMs, cross-lingual sentence retrieval and zero-shot cross-lingual transfer and find that the evaluations are not only about cross-lingual. Specifically, we observe that (i) better monolingual sentence embeddings can substantially boost the performance of models on sentence retrieval, which the current evaluation setups have ignored. (ii) the zero-shot transfer performance largely depends on the task-specific abilities of mPLMs, so the larger model with better NLU ability (XLM-R) can significantly outperform the model with better ability of cross-lingual alignment (LaBSE) on XNLI. Towards a better evaluation in the cross-lingual research, we propose two new indexes of cross-lingual ability: (i) Calibrated sentence retrieval performance, which is the performance of models after fine-tuning on the sentence embeddings objective. (ii) Transfer rate, which measures the percentage of task knowledge transferred from a source to target languages. We hope this study will enlighten future analyses on the cross-lingual ability and help the development of new mPLMs with better cross-lingual ability.

References

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of ACL*, pages 4538–4554.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of ACL*, pages 4623–4637.

Mikel Artetxe and Holger Schwenk. 2019a. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, 7:597–610.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, 7(0):597–610.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of ACL*, pages 2483–2493.

Loïc Barrault, Magdalena Biesialska, Ondrej Bojar, Marta Ruiz Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi kiu Lo, Nikola Ljubesic, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Findings of WMT*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SemEval*, pages 1–14.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. 2021a. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of EMNLP*, pages 1671–1683.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021b. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of NAACL*, pages 3576–3588.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021c. Xlm-e: Cross-lingual language model pre-training via electra. *arXiv preprint 2106.16138*.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *TACL*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, pages 7057–7067.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of EMNLP*, pages 4423–4437.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2021. Filter: An enhanced fusion method for cross-lingual language understanding. *Proceedings of the AAAI*, 35(14):12776–12784.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint 2007.01852*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, pages 6894–6910.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing word-level translations from multilingual BERT. In *Proceedings of BlackboxNLP*, pages 45–56.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-Trained models: Past, present and future. *arXiv preprint 2106.07139*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *Proceedings of ICLR*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint 1705.00652*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML*, pages 4411–4421.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of EMNLP-IJCNLP*, pages 2485–2494.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of EMNLP*, pages 1627–1643.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL*, pages 6282–6293.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *Proceedings of ICLR*.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. In *Proceedings of NIPS*, volume 33, pages 18470–18481.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020b. MLQA: evaluating cross-lingual extractive question answering. In *Proceedings of ACL*, pages 7315–7330.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL*, pages 3125–3135.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of ACL*, pages 2358–2368.

Weicheng Ma, Kai Zhang, Renze Lou, Lili Wang, and Soroush Vosoughi. 2021. Contributions of transformer attention heads in multi- and cross-lingual tasks. In *Proceedings of ACL*, pages 1956–1966.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of EMNLP*, pages 27–38.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL*, pages 1946–1958.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of AACL*, pages 557–575.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of ACL*, pages 4996–5001.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3980–3990.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint 2004.09813*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of EMNLP*, pages 10215–10245.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of NAACL-HLT*, pages 1599–1613.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of LREC*.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of AAAI*, pages 8854–8861.

Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of EMNLP*, pages 4160–4170.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of LREC*, pages 4003–4012.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL*, pages 1112–1122.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of RepL4NLP*, pages 120–130.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*, pages 483–498.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of IJCAI*, pages 5370–5378.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019b. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP-IJCNLP*, pages 3685–3690.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of STARSEM*, pages 229–240.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of ACL*, pages 3403–3417.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of BUCC@ACL*, pages 60–67.

A Appendix

A.1 All Language Results

Sentence Retrieval The full results of mBERT, XLM-R_{Base}, XLM-R, LaBSE under three different evaluation setups on CLSR, STS, and NER are shown in Table 5, Table 6, and Table 7 respectively.

Zero-shot Cross-lingual Transfer The 14 languages' results of calibrated sentence retrieval on Taoteba, zero-shot transfer, translate-train-all, transfer rate on XNLI can be found in Table 8. The Pearson's correlation coefficients between the zero-shot transfer performance and the translate-train-all, the calibrated sentence retrieval performance, and the Pearson's correlation coefficients between the transfer rate and the calibrated sentence retrieval performance can be seen at the bottom of the table. Additionally, the zero-shot transfer and translate-train-all performance of LaBSE of 15 XNLI languages on XNLI is shown in Table 9.

A.2 Hyperparameters

Question Answering Fine-tuning We fine-tune all models with a learning rate of 3e-5 and a batch size of 12 for 2 epochs.

Sentence Embedding Fine-tuning The hyperparameters for sentence embedding fine-tuning can be found in Table 10.

NER We search for the best learning rate out of [5e-6, 1e-5, 2e-5] and batch size out of [32, 128]. We train all models for 10 epochs and run validation on the English validation set every 300 training steps.

XNLI We search for the best learning rate out of [5e-6, 1e-5, 2e-5] and batch size out of [16, 32, 128]. We run validation when every 5% of training data is processed and pick the checkpoint with the best average performance across all languages on validation sets. We fine-tune each model for 10 epochs with five different seeds and report the mean performance on test sets across the five seeds.

No Fine-tuning

Lang.	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja
mBERT	38.9	24.5	48.8	17.0	75.4	29.8	64.1	28.1	25.5	41.2	39.0	64.3	40.1	34.8	36.9	53.5	57.3	40.9
XLM-R_{Base}	55.2	36.8	67.6	29.3	89.9	53.7	74.0	49.3	33.5	68.0	66.7	74.1	53.9	54.2	61.6	70.8	68.2	57.2
XLM-R	38.5	24.8	37.1	22.1	72.5	35.1	53.4	25.1	19.4	48.2	46.6	57.6	30.3	33.7	50.5	49.6	49.6	45.1
LaBSE	97.0	89.3	95.6	91.3	99.2	96.8	98.1	97.9	95.9	96.4	96.9	95.9	92.3	97.6	96.9	95.5	95.1	96.4

Lang.	jv	ka	kk	ko	ml	mr	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh
mBERT	17.6	19.6	27.1	36.0	17.9	20.1	63.7	68.4	59.4	10.8	13.4	14.1	13.7	16.0	32.9	30.8	61.0	68.6
XLM-R_{Base}	15.1	41.4	40.3	51.6	56.6	46.0	79.5	80.6	72.5	18.7	25.7	32.5	38.3	31.2	61.1	36.6	68.4	60.7
XLM-R	11.2	11.1	25.9	42.1	22.4	27.4	66.1	59.4	51.4	10.8	11.4	26.9	25.0	9.7	45.7	18.1	39.7	38.8
LaBSE	87.3	95.4	90.1	94.2	99.0	95.4	97.1	95.6	95.0	90.8	90.2	98.7	97.3	97.8	98.0	95.9	97.2	96.6

Question Answering Fine-tuning

Lang.	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja
mBERT	49.4	33.2	57.3	20.7	82.9	29.4	73.7	32.2	30.8	49.3	41.7	70.9	45.5	37.4	46.2	54.8	67.2	49.0
XLM-R_{Base}	74.8	62.6	83.1	59.0	96.6	76.6	90.0	67.9	53.1	86.1	85.2	87.0	75.5	84.7	82.3	89.0	80.8	80.3
XLM-R	80.0	77.7	89.6	76.1	97.4	84.8	94.9	72.1	59.5	92.3	90.0	91.8	84.7	94.0	87.8	92.8	87.5	89.9
LaBSE	96.8	90.1	95.1	91.2	99.2	96.4	97.9	96.9	94.7	96.1	96.6	96.0	92.0	98.1	96.3	95.9	95.3	96.5

Lang.	jv	ka	kk	ko	ml	mr	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh
mBERT	16.6	22.1	31.7	44.8	21.5	21.6	70.2	76.1	66.4	14.9	20.5	24.4	14.6	16.5	37.8	37.2	65.7	79.1
XLM-R_{Base}	29.8	65.4	58.6	76.9	79.5	75.1	90.7	90.5	87.9	34.9	53.4	70.1	79.9	53.7	81.6	70.5	89.7	87.1
XLM-R	34.1	79.4	69.9	86.1	92.7	84.2	95.0	93.3	90.9	35.4	80.5	88.5	91.8	61.9	91.2	84.3	94.0	93.0
LaBSE	83.9	95.3	89.4	93.8	98.5	95.2	97.4	96.3	94.8	88.2	91.5	97.4	96.5	96.9	98.0	95.8	98.0	96.6

Sentence Embedding Fine-tuning

Lang.	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja
mBERT	41.2	31.1	50.7	20.4	76.8	31.5	66.9	32.5	31.0	45.6	42.6	66.0	42.2	40.8	42.2	55.9	63.2	44.6
XLM-R_{Base}	63.9	53.2	77.1	46.8	91.7	69.3	81.1	57.4	46.3	79.5	77.4	80.2	69.2	74.3	72.7	82.7	76.0	68.7
XLM-R	77.8	69.2	85.2	65.8	96.6	78.5	90.7	72.9	57.4	88.6	88.0	88.9	79.5	91.0	85.6	90.6	82.9	86.3
LaBSE	97.0	89.4	91.1	90.8	99.4	96.4	98.1	98.2	95.0	95.8	97.4	95.1	92.6	97.7	96.4	95.5	95.2	95.7

Lang.	jv	ka	kk	ko	ml	mr	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh
mBERT	17.6	22.8	30.8	38.0	23.0	26.8	64.5	70.8	60.0	13.3	21.2	22.6	15.1	18.7	34.7	34.1	61.3	70.0
XLM-R_{Base}	26.8	56.4	52.3	67.2	75.3	65.3	85.5	84.5	80.5	27.2	50.5	63.2	68.8	46.7	70.8	56.7	81.2	74.7
XLM-R	28.8	73.7	65.9	81.6	88.5	78.0	92.0	90.8	88.5	30.3	58.0	72.2	77.2	59.6	89.5	62.2	92.5	91.0
LaBSE	85.4	94.6	91.0	92.7	98.8	95.0	97.4	95.9	95.3	88.2	90.6	97.0	96.5	97.1	97.5	95.8	97.7	96.0

Table 5: Full CLSR results of all models on 36 language-pairs under three evaluation setups.

	No Fine-tuning			**Question Answering Fine-tuning**			**Sentence Embedding Fine-tuning**		
Model	AR-AR	EN-EN	ES-ES	AR-AR	EN-EN	ES-ES	AR-AR	EN-EN	ES-ES
mBERT	49.28	49.44	53.53	49.85	57.24	66.05	58.23	76.28	78.28
XLM-R_{Base}	40.14	58.10	62.86	41.94	60.66	66.62	53.37	68.48	72.47
XLM-R	49.57	57.23	50.99	53.33	71.30	72.76	61.47	77.34	80.47
LaBSE	72.62	77.68	82.54	62.34	70.13	75.46	78.70	87.37	85.45

Table 6: The STS-2017 results of four models under three evaluation setups. All results are reported in Spearman's correlation coefficient $\times$ 100.

No Fine-tuning

Lang.	ar	he	vi	id	jv	ms	tl	eu	ml	ta	te	af	nl	en	de	el	bn	hi	mr	ur
mBERT	45.2	55.3	68.1	58.5	62.3	67.9	71.1	59.9	57.2	52.9	50.6	77.5	82.6	83.4	79.0	72.1	68.5	65.0	56.4	31.0
XLM-R_{Base}	47.0	53.2	67.3	49.0	59.5	55.7	72.4	58.9	62.1	56.2	48.3	75.8	80.9	82.7	74.6	74.8	70.4	69.1	62.6	56.8
XLM-R	53.0	56.8	79.1	54.3	61.9	69.0	74.3	68.6	64.1	61.8	54.7	78.6	84.4	84.1	80.0	79.4	79.9	73.4	65.2	55.8
LaBSE	44.6	56.7	68.5	49.1	66.7	69.9	75.1	63.6	66.9	56.4	53.4	76.7	81.4	83.3	76.8	71.8	73.6	68.5	54.7	54.9
Lang.	fa	fr	it	pt	es	bg	ru	ja	ka	ko	th	sw	yo	my	zh	kk	tr	et	fi	hu
mBERT	43.8	81.0	80.7	79.1	73.3	77.6	65.3	29.0	67.8	61.2	0.6	70.8	49.2	50.5	45.4	49.1	74.4	77.4	78.2	76.0
XLM-R_{Base}	51.1	77.0	78.6	78.0	73.5	77.5	64.8	20.5	66.8	51.5	3.9	69.3	33.8	49.8	28.4	40.8	74.2	72.1	76.0	76.8
XLM-R	67.3	80.8	81.8	82.9	76.3	82.7	71.9	18.5	71.0	59.6	2.3	69.2	43.0	53.3	28.8	54.8	82.5	81.2	81.0	81.9
LaBSE	48.2	77.9	79.7	78.1	70.9	78.8	65.8	24.7	68.6	56.8	2.3	75.6	75.3	61.1	27.8	50.0	76.7	74.8	77.0	77.3

Question Answering Fine-tuning

Lang.	ar	he	vi	id	jv	ms	tl	eu	ml	ta	te	af	nl	en	de	el	bn	hi	mr	ur
mBERT	49.4	55.3	71.2	61.7	56.8	68.2	72.6	61.9	55.7	50.0	49.2	77.1	82.1	83.6	78.0	66.2	68.1	65.6	54.5	36.0
XLM-R_{Base}	53.5	49.7	69.3	46.0	57.8	63.1	70.0	53.4	58.0	54.7	45.7	75.5	78.6	80.8	72.6	72.7	68.5	65.5	59.7	61.2
XLM-R	53.1	56.7	77.4	53.4	61.0	68.8	75.8	57.9	63.6	60.7	53.8	76.5	84.8	84.7	78.5	78.7	72.3	70.5	63.5	54.6
LaBSE	47.2	57.3	71.7	51.2	64.6	70.2	73.9	64.4	67.9	55.0	51.7	76.3	81.4	82.8	77.1	70.1	74.5	69.0	53.6	52.2
Lang.	fa	fr	it	pt	es	bg	ru	ja	ka	ko	th	sw	yo	my	zh	kk	tr	et	fi	hu
mBERT	42.7	77.7	80.6	77.0	66.8	76.3	64.4	29.3	63.8	57.6	0.1	67.1	47.1	43.2	43.1	44.2	69.7	77.1	77.4	74.3
XLM-R_{Base}	49.0	74.9	76.8	76.1	70.0	75.3	58.6	16.9	63.2	48.4	1.0	67.6	49.8	51.3	21.6	38.1	71.5	68.8	74.2	75.2
XLM-R	63.4	80.1	81.5	82.1	76.9	81.2	71.3	21.2	69.1	58.8	4.4	67.0	34.9	53.8	30.3	50.3	78.7	77.7	78.5	80.4
LaBSE	51.0	78.2	79.0	77.4	70.1	78.7	67.6	29.2	67.6	54.2	3.3	74.1	74.2	58.9	34.4	46.8	75.5	74.9	76.5	77.4

Sentence Embedding Fine-tuning

Lang.	ar	he	vi	id	jv	ms	tl	eu	ml	ta	te	af	nl	en	de	el	bn	hi	mr	ur
mBERT	47.3	55.5	69.4	58.9	64.9	67.6	73.0	57.5	56.1	50.9	50.3	75.2	81.8	83.8	78.1	69.6	69.7	65.3	56.7	33.4
XLM-R_{Base}	55.7	49.9	67.8	47.2	52.4	51.1	69.8	53.6	60.7	53.7	46.9	74.8	79.2	81.8	71.6	70.5	65.2	64.4	55.9	55.1
XLM-R	53.6	59.6	77.0	54.0	60.6	70.7	76.0	63.7	64.1	60.0	51.4	76.5	83.6	84.8	79.5	79.0	79.8	72.3	66.3	69.8
LaBSE	46.1	56.4	70.8	46.9	65.6	65.7	74.0	58.6	64.5	54.8	50.7	75.2	80.1	81.8	75.9	72.8	70.0	66.1	53.2	64.6
Lang.	fa	fr	it	pt	es	bg	ru	ja	ka	ko	th	sw	yo	my	zh	kk	tr	et	fi	hu
mBERT	40.4	79.2	80.9	78.3	70.3	78.1	65.5	27.6	66.3	59.1	0.4	67.5	49.6	49.4	42.3	46.6	73.6	75.9	76.3	74.8
XLM-R_{Base}	44.3	76.0	75.5	77.4	73.0	76.7	60.2	13.8	61.5	47.1	3.4	65.5	43.2	48.4	18.9	38.7	74.4	69.1	73.7	73.9
XLM-R	66.0	80.6	81.2	80.5	70.7	82.0	70.7	21.3	69.8	60.4	2.5	68.2	47.3	50.8	25.9	53.6	79.3	78.0	78.8	79.9
LaBSE	51.0	78.6	79.1	78.5	72.2	77.7	64.8	25.8	68.3	56.3	2.4	73.6	67.8	56.5	31.7	47.7	75.5	73.1	76.0	77.3

Table 7: Full NER results of all models on 40 languages under three evaluation setups.

Model	**fr**	**es**	**de**	**el**	**bg**	**ru**	**tr**	**ar**	**vi**	**th**	**zh**	**hi**	**sw**	**ur**	**avg**
Zero-shot Cross-lingual Transfer															
mBERT	73.4	73.5	70.0	65.3	68.0	67.8	60.9	64.3	69.3	54.1	67.8	58.9	49.7	57.2	64.3
XLM-R_{Base}	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	75.5
XLM-R	84.1	85.1	83.9	82.9	84.0	81.2	79.6	79.8	80.8	78.1	80.2	76.9	73.9	73.8	80.3
LaBSE	81.0	81.2	79.3	79.3	80.7	78.6	76.4	75.4	77.3	70.4	77.0	73.2	71.8	69.6	76.5
Calibrated Sentence Retrieval															
mBERT	70.9	73.7	82.9	29.4	57.3	66.4	37.8	33.2	65.7	14.6	79.1	37.4	14.9	37.2	50.0
XLM-R_{Base}	87.0	90.0	96.6	76.6	83.1	87.9	81.6	62.6	89.7	79.9	87.1	84.7	34.9	70.5	79.4
XLM-R	91.8	94.9	97.4	84.8	89.6	90.9	91.2	77.7	94.0	91.8	93.0	94.0	35.4	84.3	86.5
LaBSE	96.0	97.9	99.2	96.4	95.1	94.8	98.0	90.1	98.0	96.5	96.6	98.1	88.2	95.8	95.8
Translate-Train-All															
mBERT†	77.8	79.1	77.6	75.9	77.6	75.4	74.3	73.8	77.0	70.0	77.6	70.7	70.5	67.4	74.6
XLM-R_{Base}	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	78.7
XLM-R	85.1	86.6	85.7	85.3	85.9	83.5	83.2	83.1	83.7	81.5	83.7	81.6	78.0	78.1	83.2
LaBSE	81.9	82.9	81.9	81.5	82.4	80.8	78.5	78.1	80.8	75.1	80.2	76.3	74.7	72.1	79.1
Transfer Rate															
mBERT	94.3	92.9	90.2	86.0	87.6	89.9	82.0	87.1	90.0	77.3	87.4	83.3	70.5	84.9	86.0
XLM-R_{Base}	97.9	98.2	98.0	96.4	97.9	98.0	94.4	95.5	96.0	95.8	95.6	95.1	91.0	93.6	96.0
XLM-R	98.8	98.3	97.9	97.2	97.8	97.2	95.7	96.0	96.5	95.8	95.8	94.2	94.7	94.5	96.5
LaBSE	98.9	97.9	96.8	97.3	97.9	97.3	97.3	96.5	95.7	93.7	96.0	95.9	96.1	96.5	96.7
Pearson's Correlation Coefficient with the Zero-shot Transfer Performance															
Calibrated Sentence Retrieval.	90.4	90.0	90.0	93.8	94.4	95.3	96.6	87.4	91.4	92.1	85.3	96.8	63.6	89.9	91.0
Translate-Train-All.	98.5	96.7	94.4	96.1	94.3	96.5	91.6	94.2	96.6	96.3	91.4	93.5	89.5	93.9	95.1
Pearson's Correlation Coefficient with the Calibrated Sentence Retrieval Performance															
Transfer Rate.	98.6	93.9	95.8	97.6	95.6	95.5	99.6	92.7	95.8	96.2	87.8	97.9	69.8	98.0	95.7

Table 8: The calibrated sentence retrieval results on Taoteba and zero-shot transfer, translate-train-all results on XNLI, and Pearson's correlation coefficients on 14 XNLI languages.

Task	**en**	**fr**	**es**	**de**	**el**	**bg**	**ru**	**tr**	**ar**	**vi**	**th**	**zh**	**hi**	**sw**	**ur**	**avg**
Zero-shot.	86.0	81.0	81.2	79.3	79.3	80.7	78.6	76.4	75.4	77.3	70.4	77.0	73.2	71.8	69.6	77.2
Translate-train-all.	86.1	81.9	82.9	81.9	81.5	82.4	80.8	78.5	78.1	80.8	75.1	80.2	76.3	74.7	72.1	79.5

Table 9: Full zero-shot transfer and translate-train-all results of LaBSE.

Model	Learning Rate	Batch Size	Warmup Proportion	#Epochs (AllNLI)	#Epochs (STSb)
mBERT	2e-5	128	0.1	1	10
XLM-R_{Base}	2e-5	128	0.1	1	5
XLM-R	5e-6	64	0.1	1	5
LaBSE	2e-5	128	0.1	1	10

Table 10: Hyperparameters for the Sentence-BERT fine-tuning approach.