Analyzing the Evaluation of Cross-Lingual Knowledge Transfer in Multilingual Language Models

Sara Rajaee and **Christof Monz**
Language Technology Lab, University of Amsterdam
{s.rajaee, c.monz}@uva.nl

Abstract

Recent advances in training multilingual language models on large datasets seem to have shown promising results in knowledge transfer across languages and achieve high performance on downstream tasks. However, we question to what extent the current evaluation benchmarks and setups accurately measure zero-shot cross-lingual knowledge transfer. In this work, we challenge the assumption that high zero-shot performance on target tasks reflects high cross-lingual ability by introducing more challenging setups involving instances with multiple languages. Through extensive experiments and analysis, we show that the observed high performance of multilingual models can be largely attributed to factors not requiring the transfer of actual linguistic knowledge, such as task- and surface-level knowledge. More specifically, we observe what has been transferred across languages is mostly data artifacts and biases, especially for low-resource languages. Our findings highlight the overlooked drawbacks of existing cross-lingual test data and evaluation setups, calling for a more nuanced understanding of the cross-lingual capabilities of multilingual models.[1]

1 Introduction

Massively Multilingual Transformers (MMTs) exhibit remarkable abilities in comprehending texts in multiple languages (Devlin et al., 2019; Conneau et al., 2020; Chi et al., 2022; Lin et al., 2022). Their performance on extensive multilingual benchmarks (Hu et al., 2020b; Ruder et al., 2021) including natural language inference (Conneau et al., 2018), paraphrase identification (Yang et al., 2019), question answering (Artetxe et al., 2020), and commonsense reasoning (Ponti et al., 2020), clearly indicates their usefulness for downstream tasks across different languages.

Since collecting labeled data for multiple languages is expensive, fine-tuning an MMT on only one language (usually English) and applying it to other languages is a common practice to expand multilingual models' applications to more languages. The success of employing this approach is primarily attributed to the cross-lingual ability of MMTs and knowledge transfer across languages (K et al., 2019; Keung et al., 2020; Fujinuma et al., 2022; Ebrahimi et al., 2022). However, in this work, we question this attribution and take a closer look at the cross-lingual evaluation pipelines used to assess MMTs' performance. Our study aims to broaden our horizons about the existing performance-based evaluation setups of MMTs and highlight their significant shortcomings.

To this end, we first provide insights into the definition of the cross-lingual ability in language models and the essential criteria for its evaluation. We challenge the assumption of *high* cross-lingual ability, i.e., requiring actual linguistic knowledge, in MMTs using three downstream tasks: Natural Language Inference (NLI), Paraphrase Identification (PI), and Question Answering (QA). Our experimental results demonstrate that multilingual models struggle with transferring linguistic knowledge across languages when the inputs involve multiple languages, such as in the NLI task with a premise in Arabic and a hypothesis in Spanish. Employing this setup, we show that, unlike previous assumptions, MMTs are not able to effectively connect the underlying semantics between languages in a zero-shot manner. It is worth mentioning that the evaluation of MMTs using multiple languages in the input offers both theoretical advantages and reflects real-world scenarios in NLP systems.

We extend our study to investigate if the failures come from the lack of *across*-language fine-tuning data. We find that even by fine-tuning MMTs on *across*-language data that involves two languages in an instance, they still can not successfully trans-

---

[1] The source code is available at: https://github.com/Sara-Rajaee/crosslingual-evaluation

fer knowledge between languages in a zero-shot setting.

Looking for the reasons behind the ineffectiveness of MMTs on the *across* setup, we examine the impact of individual samples on the cross-lingual performance and identify a specific subset where MMTs struggle. Our findings demonstrate that models achieve exaggerated high performance by strongly relying on spurious features and data artifacts. We notice that cross-lingual transfer primarily involves learned biases and shallow knowledge rather than linguistic knowledge. Notably, this phenomenon disproportionately affects low-resource languages, exacerbating the challenges faced by MMTs in achieving true cross-lingual competence.

As part of our methodology, we design control tasks (Hewitt and Liang, 2019) in which, during fine-tuning, the words within the instances have been randomly shuffled, and then, the model is evaluated on the original, i.e., not shuffled, test data. Surprisingly, our experiments show that although these new tasks do not provide the model with meaningful linguistic knowledge related to the target task, there is only a slight drop in their cross-lingual performance in both single and two-language evaluation settings. These results demonstrate that MMTs' understanding tends to be more reliant on surface-level patterns rather than linguistic comprehension.

Our experiments show from several angles that current MMTs' cross-lingual evaluation setups do not give us a clear and faithful picture of their cross-lingual ability. Our findings question the extent of high crosslinguality in language models and prompt us to pay more attention to the interpretation of knowledge transfer and cross-lingual ability in multilingual models solely based on their performance on downstream tasks.

2 Related Work

Massively Multilingual Transformers. Multilingual models have achieved success in understanding multiple languages without requiring language-specific supervision, thanks to only requiring unlabelled training data in multiple languages. Prominent examples include Multilingual BERT (mBERT) and XLM-R, which have been pre-trained on a diverse set of languages using the masked language modeling objective (Devlin et al., 2019; Conneau et al., 2020). Other models, such as XLM with Translation Language Modeling (TLM),

XLM-E with multilingual replaced token detection (MRTD) and translation replaced token detection (TRTD), and ALM with code-switched inputs, have employed similar strategies with different objectives (CONNEAU and Lample, 2019; Chi et al., 2022; Yang et al., 2020).

Analyzing MMT. The rise of multilingual models has sparked significant interest to understand their linguistic capabilities across different languages. Chi et al. (2020) have discovered subspaces within mBERT representations that can capture syntactic tree distances across different languages. Aligned with their finding, numerous studies have further explored the potential of multilingual models to capture both syntactic (Papadimitriou et al., 2021; Xu et al., 2022; Mueller et al., 2022; Ravishankar et al., 2021) and semantic knowledge (Foroutan et al., 2022; Vulić et al., 2020) for a wide range of languages.

Several studies have delved into understanding the factors influencing the cross-lingual ability of multilingual models. Pires et al. (2019) explored the effectiveness of multilingual BERT in transferring knowledge across languages using NER and POS tagging tasks, noting the impact of language similarity on performance. Chai et al. (2022) examined cross-linguality from a language structure perspective, emphasizing the significance of the composition property in facilitating cross-lingual transfer. Muller et al. (2021) analyzed representation similarities and discovered a strong connection between hidden cross-lingual similarity and the model's performance on downstream tasks. Building upon this finding, Deshpande et al. (2022) identified a correlation between token embedding alignment and zero-shot transfer across diverse tasks.

In the realm of multilingual models, both the design of new models and the examination of existing ones rely heavily on evaluating their performance in zero-shot cross-lingual scenarios. However, there remains the question of how accurately we can interpret their performance and its implications for the models' cross-lingual abilities. To shed light on this issue, our work focuses on investigating the faithfulness of the prevailing methods used to evaluate zero-shot cross-lingual performance in the literature. By doing so, we aim to provide deeper insights into the interpretation of model performance and its relationship to the cross-lingual capabilities of multilingual models.

3 Cross-lingual Evaluation

The ability of a multilingual model to effectively generalize across languages on downstream tasks is the key factor in determining its cross-linguality (Pires et al., 2019; Wu and Dredze, 2019; Artetxe et al., 2020). More specifically, a multilingual model is considered cross-lingual if it can successfully perform tasks in languages not seen during task fine-tuning.

However, relying solely on the ultimate performance of multilingual models on a specific target task has significant drawbacks, as it can potentially create misconceptions about their true cross-lingual abilities. This evaluation approach lacks clarity in distinguishing between the extent of cross-lingual knowledge transfer and surface-level and non-linguistic features. It is possible for a multilingual model to achieve high performance on a task without possessing a deep semantic understanding of a language, instead mostly relying on language-independent and shallow knowledge, as previously reported for monolingual language models (Bhargava et al., 2021; Stacey et al., 2020; Gururangan et al., 2018; McCoy et al., 2019).

It is worth mentioning that fully disentangling linguistic and shallow (task-specific) knowledge is almost impossible for most NLP tasks. However, trying to separate these types of knowledge as much as possible gives us a clearer and more accurate perspective regarding the linguistic knowledge captured by models across languages.

In the following parts, we employ three different target tasks, namely multilingual Natural Language Inference, Paraphrase Identification, and Question Answering, to study the cross-lingual ability of multilingual models. We utilize multilingual BERT (Devlin et al., 2019, mBERT) and the base version of XLM-r (Conneau et al., 2020) for our experiments as they are among the most widely used multilingual models. We also employ INFOXLM (Chi et al., 2021) trained on a cross-lingual objective and parallel data to investigate the role of explicit pre-trained cross-lingual objective on the cross-lingual ability.[2]

3.1 Natural Language Inference

Natural language inference serves as a prominent task in the field of NLP for assessing the comprehension capabilities of language models (Bow-

man et al., 2015; Condoravdi et al., 2003). This task requires the prediction of the relationship between a given premise and hypothesis, where the model determines whether the premise entails the hypothesis, contradicts it, or remains undetermined (Williams et al., 2018, MNLI).

To study the cross-lingual ability of MMTs in the NLI task, we employ the multilingual NLI dataset (Conneau et al., 2018, XNLI), where the training set has 397k samples in English (adopted from the MNLI training set), and the test sets include 5k instances in fifteen different languages manually translated from the English data (Hu et al., 2020a).

We investigate the cross-linguality of MMTs in the context of NLI using two distinct evaluation settings. In the first setting, which we refer to as the *within language* setup, we fine-tune the model on English training data and assess its performance on NLI tasks across other languages, following previous studies (Artetxe and Schwenk, 2019; Lample and Conneau, 2019; Wu and Dredze, 2019; Qi et al., 2022; Chai et al., 2022; Huang et al., 2021). In the second setting, the *across language* evaluation, we assess cross-linguality by providing premise and hypothesis pairs in two different languages. To the best of our knowledge, none of the previous research has evaluated the cross-linguality of MMTs by employing instances involving multiple languages in downstream tasks. Since the XNLI test data is a fully parallel dataset, we can easily combine premises and hypotheses from different languages. We assert that this evaluation approach provides a more precise and reliable assessment of the models' cross-linguality. It asks the model to comprehend the underlying meaning of the input in two languages simultaneously, minimizing potential spurious correlations between examples and labels. It is worth mentioning that both evaluation approaches are conducted in a zero-shot manner.

The results are presented in Table 1. For a comprehensive breakdown of performance for each language pair, please refer to the Appendix. Aligned with previous research (Hu et al., 2020a), the results of the *within* language setting show the ability of MMTs to effectively generalize knowledge learned during fine-tuning from English to other languages when the premise and hypothesis are in the same language. Nevertheless, the extent of their success varies across languages, with lower accuracy observed for low-resource languages like Swahili compared to high-resource ones.

The results of the *across* language experiment,

---

	en	de	fr	ru	es	zh	vi	ar	tr	bg	el	ur	hi	th	sw	avg
mBERT																
within	81.5	70.6	73.5	68.6	68.2	68.6	69.9	64.2	62.0	68.7	67.5	58.7	60.5	52.3	50.3	65.7
across	61.3	57.7	59.5	57.2	59.2	55.2	56.0	54.3	51.1	55.9	54.3	50.7	52.3	47.2	45.6	54.5
XLM-r																
within	84.9	76.3	78.3	75.7	79.2	73.5	74.8	71.5	73.0	78.6	75.4	65.5	69.3	71.8	65.2	74.2
across	71.9	67.1	68.8	68.0	69.2	64.6	65.1	62.8	62.8	68.3	66.2	60.0	63.6	64.2	53.7	64.8
INFOXLM																
within	85.8	78.2	79.2	76.9	80.0	75.5	75.9	73.2	74.4	78.4	77.0	66.0	71.0	73.0	65.9	75.4
across	77.1	72.0	72.9	71.9	73.2	70.0	69.9	69.0	69.0	72.5	71.1	64.8	68.8	69.8	61.8	70.3

Table 1: The accuracy scores of mBERT, XLM-r, and INFOXLM for two evaluation settings on the NLI task: the *within* language setting, where both the hypothesis and premise are in the same language, and the *across* language evaluation, which involves two different languages. In the *across* evaluation, numbers represent the average performance when either the premise or the hypothesis (but not both) is in the given language. As the numbers show, MMTs have considerably lower performance in the *across* setting.

	en	de	fr	es	zh	ko	ja	avg
mBERT								
within	93.5	84.6	86.6	86.7	77.0	72.8	73.6	81.4
across	75.1	72.1	72.5	72.6	64.8	64.1	63.7	69.2
XLM-r								
within	94.4	87.8	89.5	89.1	81.9	76.3	77.3	85.7
across	76.4	72.1	72.2	72.2	63.6	65.2	62.9	69.2
INFOXLM								
within	94.0	88.4	90.0	90.2	83.0	78.7	78.9	86.2
across	84.6	79.9	80.3	80.4	75.2	72.4	73.3	78.0

Table 2: The accuracy of fine-tuned models on PAWS-X evaluated under the *within* and *across* language settings. For mBERT and XLM-r, there is, on average, a 17% drop in performance when the sentences are provided in two different languages. While the trend is similar for INFOXLM, the drop is less significant in this model, which can be attributed to its cross-lingual pretraining objective.

where the premise and hypothesis are in two different languages, show a comparatively lower level of cross-linguality in MMTs compared to the *within* setup. It raises concerns about the true extent of cross-lingual ability in language models. Even for high-resource languages within the same language family (e.g., English and German), the average performance declines by approximately 17% for mBERT and XLM-r. This drop is even more pronounced for low-resource languages such as Swahili, see Figures 2–4. As we expected, INFOXLM exhibits less of a performance drop in the *across* setup attributed to its cross-lingual pre-training objective. However, since English has been used as a pivot language in the pre-training data,

the cross-lingual objective has mostly helped the performance of pairs including English, and the trend of the other pairs is similar to mBERT and XLM-r, see Figure 4.

3.2 Paraphrase Identification

The Paraphrase Identification task evaluates a model's understanding of the semantic similarity between two sentences (Wang et al., 2018). Paraphrase Adversaries from Word Scrambling (PAWS) is a challenging dataset for this task, where both sentences in each example have high word overlap (Zhang et al., 2019). PAWS-X is a multilingual benchmark and extends this dataset to six languages beyond English using professionally translated validation and test sets (Yang et al., 2019).

We employ a similar evaluation setup as described for NLI to assess the cross-lingual capability of multilingual models on the semantic similarity task. Since not all instances are translated into all six other languages in the dataset, we only consider parallel sentences for the evaluation, resulting in the exclusion of a small number of examples (less than 0.5% on average) from the test sets.

The performance of the fine-tuned models on PAWS-X is presented in Table 2, and the detailed results can be found in Figures 5–7. In the *within* language setup, where the sentences in every instance are from the same language, the models demonstrate the successful knowledge transfer across languages, as indicated by their relatively high performance compared to English. However, in the *across* language setting, which tests the models' cross-lingual ability in a more challenging scenario, similar to the findings in XNLI, there is

	en	de	ru	es	zh	vi	ar	tr	el	hi	th	avg
mBERT												
within	84.5	72.7	71.4	75.5	58.2	69.2	61.2	55.1	62.4	58.0	40.0	64.4
across	57.0	50.6	50.2	52.7	42.6	46.3	42.2	36.9	42.4	38.7	26.1	44.2
XLM-r												
within	84.2	75.2	74.5	77.1	63.7	74.3	66.3	68.1	73.8	68.3	66.5	72.0
across	58.1	44.1	42.9	43.4	28.1	33.9	25.8	31.3	34.7	32.3	29.6	36.8
INFOXLM												
within	85.1	76.0	75.0	77.8	66.4	75.6	70.3	69.9	74.8	71.6	69.9	73.8
across	73.8	66.6	66.7	68.1	61.4	64.3	61.0	61.4	64.2	61.9	60.1	64.5

Table 3: Zero-shot F1 scores of fine-tuned models for the QA task using the *within* and *across* language evaluation approaches. All the models struggle with the *across* setup, especially for mBERT and XLM-r, where we observe more than 50% drop in their performance, challenging the extent of their cross-linguality.

a significant drop in performance for the models. These results show that comprehending information from two different languages and making semantic-based decisions pose challenges for multilingual models, particularly for non-Latin script languages, where their performance is noticeably affected.

3.3 Question Answering

The question answering (QA) task challenges the reading comprehension ability in language models in which the model is asked to find the answer span to a question within the given context (Choi et al., 2018; Rajpurkar et al., 2016). While NLI and PI are similarity-based and classification tasks, QA offers an alternative perspective to evaluate the cross-lingual capacity of MMTs in a different type of target task. For this task, we employ XQuAD (Artetxe et al., 2020), a multilingual question-answering benchmark comprising 240 paragraphs and 1,190 question-answer pairs from the development set of SQuAD v1.1 (Rajpurkar et al., 2016). It includes professional translations into ten languages, making it a fully parallel dataset.

The results of the cross-lingual evaluation of the QA models are presented in Table 3 for the *within* and *across* language setups.[3] We report the exact match (EM) and F1 scores, which are commonly used metrics to evaluate question-answering systems. The reported results for the *across* setting are the average over cases that the context or question (not both of them) are in the listed language.

For the QA task, we observe a similar trend to the NLI and PI tasks. The numerical results indicate that multilingual models excel in generalizing across unseen languages when the context and question are in the same language, i.e., the *within* setting. However, when the context is in a different language than the question, the empirical results show a substantial decline in performance. This suggests that multilingual models face challenges in retrieving knowledge from several languages at the same time and bridging information across different language representation spaces.

Another interesting observation is that in the *across* language scenario, the models' performance is lower compared to both languages in the *within* language setup. Nonetheless, there is an exception when the question is in English in the *across* language setting. Regardless of whether the context language is low or high-resource language, the performance remains close to the *within* setup of the context language in mBERT and XLM-r (Figures 8 and 9). We speculate that this pattern may be linked to the utilization of Wikipedia articles in the pre-training data for both models, which were used in constructing the SQuAD dataset as well. However, we observe an opposite pattern for the INFOXLM model in that the model relies more on the context than the question that can be attributed to the English-centric pretraining data.

3.4 Discussion

The previous experimental results demonstrate that MMTs face difficulties in understanding and connecting multiple languages simultaneously. Although the cross-lingual pre-training objective of INFOXLM has enhanced the cross-lingual ability of this model compared to mBERT and XLM-r, the performance gap between *within* and *across*

---

[3]The language-pair results can be found in the appendix and see Table 8 for EM performance.

	en-de	ar-en	tr-el	es-de
$l1 - l2$	67.8 $_{64.1}$	75.9 $_{64.1}$	63.4 $_{52.9}$	76.4 $_{61.8}$
$l2 - l1$	58.7 $_{71.3}$	39.2 $_{56.5}$	41.7 $_{50.8}$	43.1 $_{65.3}$
$* - l2$	63.0 $_{59.4}$	59.6 $_{55.0}$	61.7 $_{52.7}$	63.2 $_{58.7}$
$l1 - *$	57.6 $_{58.7}$	41.6 $_{59.4}$	55.2 $_{55.0}$	43.8 $_{58.7}$
$* - *$	52.8 $_{55.3}$	54.9 $_{55.3}$	56.5 $_{55.3}$	56.4 $_{55.3}$

Table 4: mBERT's performance on the XNLI test set. The columns show the fine-tuning language pairs, and the rows show the evaluation pairs in which $l1$ and $l2$ represent the premise and hypothesis's languages, respectively, as given in the corresponding columns. The rows that include $*$ show the average performance over all languages. The smaller numbers present the baseline performance (fine-tuning on en-en) for the corresponding evaluation pairs.

setting is still considerable. A question raised here is whether the low performance comes from the lack of *across* language style fine-tuning data or a deeper incapability of cross-lingual knowledge transfer in MMTs. To answer this question, we fine-tune mBERT using *across* language setups on multiple language pairs on the NLI task. Table 9 breaks down the performance for the zero-shot and semi-zero-shot setups.[4] As can be observed, fine-tuning using *across* language style data does not help the model to generalize to other languages. Interestingly, the model cannot even achieve better performance on the same language pair when the premise and hypothesis languages have been swapped. Considering the results, we conclude that the lower performance on the *across* evaluation setup can not be attributed to the distinct fine-tuning and evaluation setups.

4 Breakdown Analysis

In this section, we delve into the reasons behind the lack of success exhibited by multilingual models in the *across* language setup. Our analysis begins by exploring the individual contributions of each class to the overall tasks' performance. By doing so, we aim to uncover the specific instances that present difficulties for multilingual models and subsequently lead to a decline in their performance.

NLI. In Table 5, we present the numerical results for the NLI task per label.[5] Following previous work (Yaghoobzadeh et al., 2021; Sanh et al., 2021),

---

[4]The results of other language pairs can be found in the appendix.

[5]Please refer to Table 10 and 11 for the XLM-r and IN-FOXLM results.

and for simplicity, we combine the *neutral* and *not-entailment* classes, considering them as the *not-entailment* class. Surprisingly, the contribution of the *entailment* and *not-entailment* classes to the overall performance in the *across* language setup is not equal, and this trend is consistent across all models. Notably, the drop in performance is primarily affected by the entailment class and is more pronounced for low-resource languages.

In relation to this behavior, we suspect that the observed patterns can be attributed to dataset artifacts, particularly the word overlap bias in the training set of XNLI (McCoy et al., 2019), and this bias is transferred to other languages. The word overlap bias refers to the tendency of NLI models trained on the MNLI dataset to favor the *entailment* label when there is a high word overlap between the premise and hypothesis. Moreover, previous studies have shown a strong correlation between low word overlap and the *not-entailment* label, which is referred to as reverse (word overlap) bias (Rajaee et al., 2022). Therefore, since in the *across* setting, the word overlap is minimized, the model is biased toward predicting the *not-entailment* label. Another possible explanation is that the multilingual models prioritize language similarity over semantic meaning, leading to a considerable drop in performance.

In the multilingual context, our findings show that, in the case of the NLI task, the transfer across languages primarily involves dataset artifacts and biases rather than linguistic knowledge. Especially in low-resource scenarios, where sufficient pre-training data is lacking, the model heavily relies on these shortcuts.

PI. The performance of mBERT on the PAWS-X test sets, broken down by labels, is presented in Table 6 (see the other models' results in Table 12 and 13). Despite the adversarial construction of the PAWS dataset to prevent word overlap bias, with all instances having high word overlap between sentences, we observe a similar drop in performance in the cross-language setting, primarily originating from the *paraphrase* class. This suggests the possible presence of biases that have been transferred across languages rather than linguistic knowledge transfer.

To understand this behavior, we consider two potential reasons for the drop in *paraphrase* class performance. Firstly, the overlap bias may arise from the pre-training procedure of multilingual models.

	en	de	fr	ru	es	zh	vi	ar	tr	bg	el	ur	hi	th	sw	avg
Entailment																
within	76.9	61.2	66.8	60.4	66.6	54.8	61.7	56.6	62.7	63.0	62.0	46.3	51.9	60.1	68.1	61.3
across	32.2	28.1	31.3	28.4	30.5	21.0	24.0	22.2	16.2	26.0	22.1	15.2	19.5	10.3	8.3	22.3
NotEntailment																
within	83.9	75.3	76.9	72.8	77.7	75.0	73.5	68.0	61.6	71.6	70.2	64.9	64.8	48.5	41.4	67.3
across	75.9	72.5	73.7	71.6	73.5	72.2	71.6	70.3	68.5	70.7	70.4	68.4	68.6	65.7	64.2	70.5

Table 5: mBERT's accuracy scores on the XNLI test set separated based on the labels. While both classes have an almost equal contribution to the performance in the *within* setup, the performance on the *entailment* class significantly drops in the *across* setup.

	en	de	fr	es	zh	ko	ja	avg
Paraphrase								
within	94.9	89.0	90.1	88.8	76.5	55.7	67.4	81.4
across	57.6	54.7	54.9	53.3	35.5	33.0	31.2	45.7
NonParaphrase								
within	92.4	81.0	83.8	85.0	77.5	86.6	78.6	82.9
across	89.3	86.1	86.8	88.3	88.6	89.3	89.8	88.3

Table 6: The performance of mBERT for the *within* and *across* setting per label on the PI task. Most of the performance drop of the *across* setting occurs for the *paraphrase* class.

Additionally, there might be fine-grained biases related to bigram or trigram overlap in the PAWS training data. Investigating these intriguing patterns and biases in the dataset is an avenue for future research.

QA. For the question-answering task, which differs from a classification task, we adopt a distinct approach to uncover the obstacles impeding knowledge transfer across languages in the *across* setup.

Based on the nature of our *across* language setup that minimizes the overlap between the context and question and the results of the NLI analysis, we suspect that a similar bias exists for the QA task, which is easily transferable across languages. Previous studies have reported different biases in the SQuAD dataset, including answer-position, word overlap between the question and context, and type-matching (Ko et al., 2020; Sugawara et al., 2018; Weissenborn et al., 2017).

To capture the degree of overlap between the context and question, we introduce a new measurement technique. Instead of simple word count, we compute the average distance of each question word's occurrence in the context to the center of the answer span. Here, the difference between the shared word position index in the context and the center of the answer span is considered as the distance. If

there is no word overlap between the question and context, we assign the distance as the maximum length of context (which is a hyperparameter). We contend that the model's reliance on high word overlap cannot be simply regarded as a shortcut, and we advocate for the use of the distance metric as a more accurate measure of the possible bias evaluation.

To investigate a possible spurious correlation between the question and context overlap and the answer span, we calculate the average distance for the top and bottom 20% of samples, representing instances where the model achieves the highest and lowest F1 scores, respectively, across all individual languages.

The findings are presented in Figure 1.[6] It is evident that the average distance for the most challenging instances is twice that of the easiest ones, indicating a significant correlation between the concentration of shared words around the answer span and the model's performance. Furthermore, reliance on the concentration of shared words around the answer as a shortcut is consistently observed across different languages, as indicated by their corresponding performances.

Summary. Our analysis across different tasks highlights the predominant influence of dataset artifacts and reliance on shortcuts rather than robust cross-lingual knowledge transfer. The observed disability in performing tasks involving multiple languages shows that the models prioritize shallow knowledge over linguistic understanding.

5 Control Tasks

The idea of control tasks, proposed by Hewitt and Liang (2019), is aimed at enabling a meaningful and faithful interpretation of the linguistic knowledge encoded in language models' representations

---

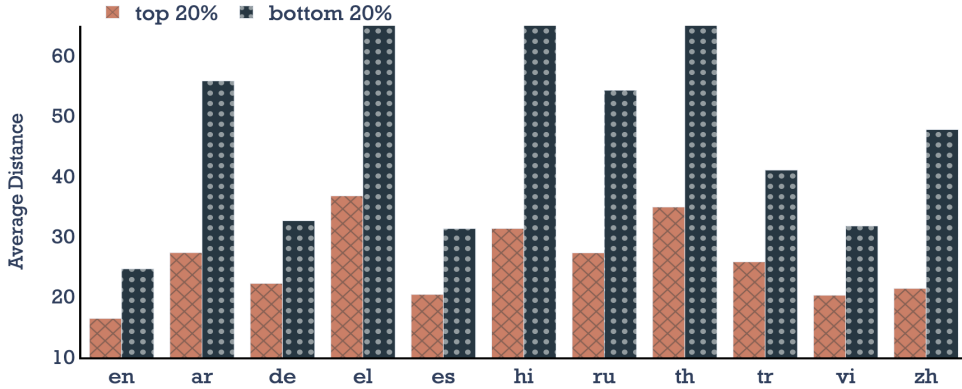[6]Please refer to Figures 11 and 12 for more results.

Figure 1: The average distance of the questions' words occurred in the context to the center of the answer span in the top 20% easiest and hardest instances for the mBERT fine-tuned on SQuAD evaluated on every language. As the average distance increases, the model's performance drops.

	en	de	fr	ru	es	zh	vi	ar	tr	bg	el	ur	hi	th	sw	ko	ja	avg	original
								XNLI											
within	77.9	68.7	71.3	66.2	71.0	66.3	67.0	62.4	61.3	65.9	64.1	56.0	59.0	51.8	48.8	-	-	63.8	65.7
across	57.8	55.1	56.4	55.0	56.6	52.5	53.4	52.4	49.8	53.4	52.0	49.1	50.6	46.9	44.8	-	-	52.4	54.5
								PI											
within	57.1	55.5	55.2	-	56.1	54.9	-	-	-	-	-	-	-	-	-	54.0	54.8	55.4	81.4
across	56.8	56.1	56.0	-	56.3	55.8	-	-	-	-	-	-	-	-	-	55.6	56.1	56.1	69.2
								QA											
within	82.5	68.6	-	70.0	72.9	58.0	67.6	56.8	55.1	-	60.0	-	56.0	43.3	-	-	-	62.8	64.4
across	54.0	46.4	-	46.5	48.8	38.4	41.6	37.7	34.9	-	37.3	-	35.5	25.7	-	-	-	40.6	44.2

Table 7: The performance of mBERT fine-tuned on the control tasks. We report accuracy for the NLI and PI tasks and the F1 score for QA. Although the fine-tuning data does not provide the model with any meaningful and task-related knowledge, the drop in performance is negligible for the NLI and QA tasks. The original column refers to the results in Tables 1–3.

during probing procedures. In the interpretability area, these tasks serve as baselines to measure the model's language understanding capabilities and ensure that the probe's high performance is not attributed to the linguistic knowledge learned by the probe itself and coming from the encoded linguistic knowledge in the representations.

In this section, we borrow the idea of control tasks in probing and employ them to assess the cross-lingual abilities of MMTs. These tasks are indeed (partially) random sequences, allowing us to evaluate the models' performance when it does not receive linguistic cues. This unconventional approach provides valuable insights into the extent of the meaningful numerical performance of multilingual models on the current multilingual benchmarks.

To this aim, we randomly shuffle the inputs and fine-tune the models on these shuffled instances.

More specifically, for the NLI and PI tasks, we shuffle the words within every sentence, and for the QA task, we shuffle the question words and keep the original context.

It is evident that the designed tasks are only marginally related to human language comprehension. However, surprisingly, the performance of the models on these tasks, as shown in Table 7, indicates only a marginal decrease for both within and cross-language settings.

In the case of the NLI task, we observe a slight 4% drop in accuracy on the test sets when the model is fine-tuned on a nonsensical task. Similarly, the QA model exhibits a drop of approximately 3%. These findings suggest that the current test sets may not provide sufficient quality for effectively evaluating the cross-lingual capabilities of multilingual models. Out of all the tasks, the PI task stands out with its random performance, indicating

that the fine-tuned model with shuffled data lacks the necessary knowledge to complete the task successfully. This aligns with our expectations from a high-quality dataset.

6 Future Directions

In this work, we have demonstrated that current performance-based methods for evaluating cross-lingual abilities are insufficient. The main concern now is how to assess the cross-lingual capabilities of multilingual models.

Firstly, we should be aware of the shortcomings of existing multilingual benchmarks commonly used for evaluating cross-lingual knowledge transfer. As we described in our study, current tasks and datasets are prone to artifacts leading to exaggerated high performance. In addition to the random baseline, which is defined based on the number of labels for classification tasks, we suggest having a secondary baseline for every task. The secondary baseline could be considered the performance of a simple neural network on the task or the performance on the control tasks. We leave the exploration of this avenue to future work.

Another alternative could be our suggested setup, the *across* language approach, which involves multiple languages. The proposed evaluation setup is not only valuable for theoretical analysis but also reflects real-world scenarios where NLP systems need to integrate knowledge from different language sources.

7 Conclusion

In this paper, we take a fresh perspective on the cross-lingual ability of multilingual models. Through comprehensive experiments, we explored their capacity to simultaneously leverage knowledge from multiple languages. Our findings show that multilingual models struggle to establish connections between knowledge spaces across languages, resulting in subpar performance on cross-language task setups, i.e., when there are multiple languages in the input. Our results show that the previously reported high performance in the zero-shot setting predominantly stems from the transfer of shallow, language-independent knowledge. Surprisingly, we observed that dataset artifacts, rather than intrinsic linguistic features, are predominantly transferred across languages. This challenges the notion of relying solely on multilingual models' performance for assessing their true cross-lingual

capabilities. To assess the quality of existing multilingual benchmarks, we conducted fine-tuning experiments on control tasks with nonsensical input. Surprisingly, even when fine-tuned on meaningless tasks, the models demonstrated exceptionally high performance, prompting concerns regarding the quality of current multilingual datasets. In light of these insights, in the future, we plan to explore novel task and data-independent approaches to gain a more accurate understanding of multilingual models' true cross-lingual abilities.

Limitations

In our experiments, we primarily examined three widely used MMTs, namely mBERT, XLM-r, and INFOXLM. However, there is room for further expansion by incorporating a broader range of MMTs with diverse objectives and architectures to evaluate their cross-lingual ability. Additionally, while we assessed the cross-linguality of MMTs on three downstream tasks, there is potential for exploring additional target tasks and datasets.

8 Acknowledgements

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Yuan Chai, Yaobo Liang, and Nan Duan. 2022. Cross-lingual ability of multilingual masked language models: A study of language structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4702–4712, Dublin, Ireland. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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 the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1500–1512, Dublin, Ireland. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020a. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020b. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aaron Mueller, Yu Xia, and Tal Linzen. 2022. Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 95–109, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Sara Rajaee, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. Looking at the overlooked: An analysis on the word-overlap bias in natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10605–10616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. 2021. Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online. Association for Computational Linguistics.

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 the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8281–8291, Online. Association for Computational Linguistics.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Ningyu Xu, Tao Gui, Ruotian Ma, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2022. Cross-linguistic syntactic difference in multilingual BERT: How good is it and how does it affect transfer? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8073–8092, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Dongdong Zhang, ShuangZhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9386–9393.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A   Experimental Setups

We have used the HuggingFace library for fine-tuning multilingual models. For the NLI task, we have fine-tuned the models for 3 epochs with a batch size of 32, maximum length of 128, and learning rate of $2e-5$, using the last [CLS] token representation. For the PI task, we have considered the same hyperparameters but a batch size of 16. For the QA task, we have used the following setups: a maximum length of 384, batch size of 16, and learning rate of $2e-5$. All the reported numerical results are the average over three different random seeds. All models have been fine-tuned using one NVIDIA A6000 GPU.
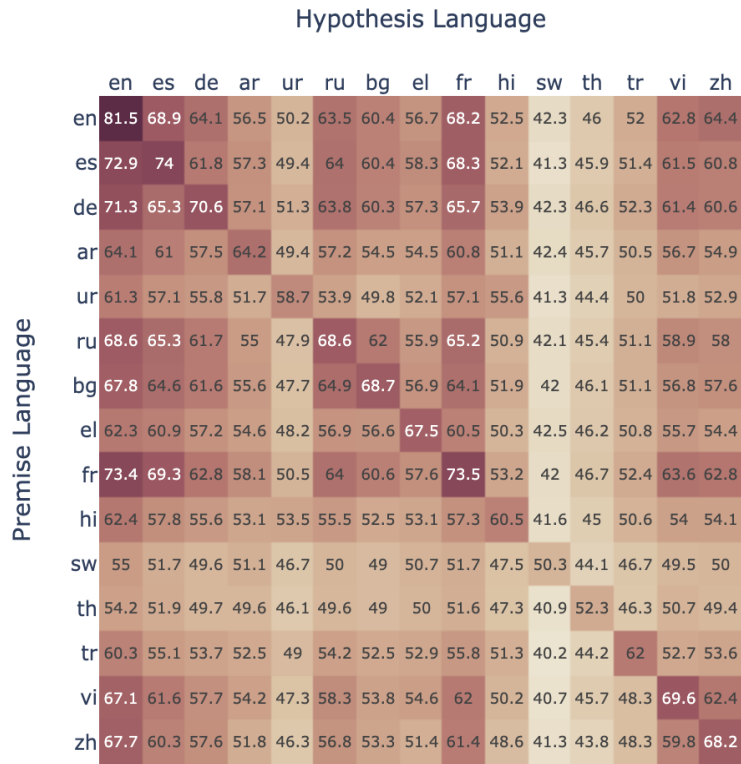
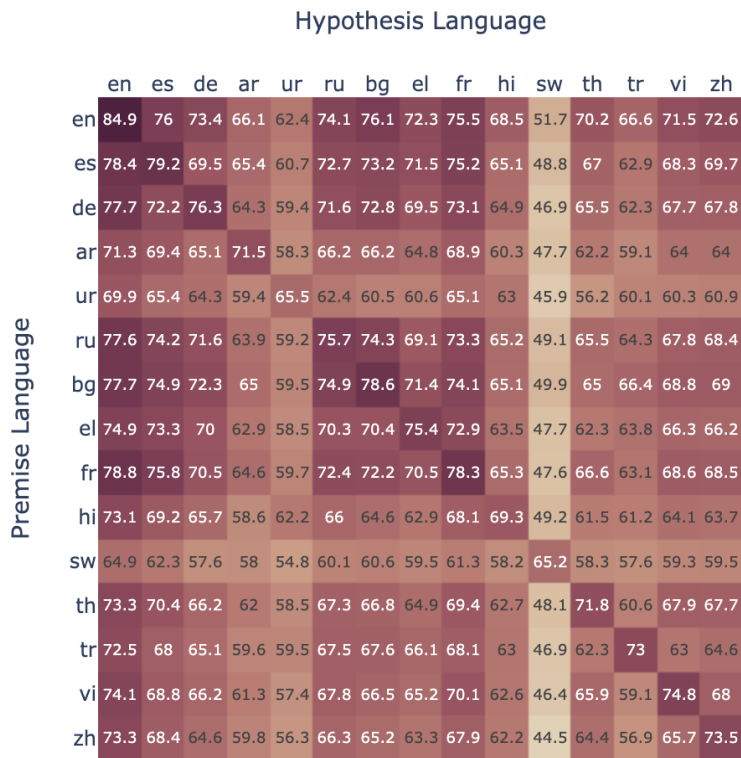Figure 2: Language-pairs accuracy scores for mBERT on the multilingual NLI task.

Figure 3: Language-pairs accuracy scores for XLM-r on the multilingual NLI task.

Hypothesis Language

Premise Language	en	es	de	ar	ur	ru	bg	el	fr	hi	sw	th	tr	vi	zh
en	85.9	81.2	79.5	75.1	69.1	78.7	79.8	77.2	79.9	75.8	64.9	76.7	75.2	77.4	78.2
es	81.6	80	74.8	71.6	64.8	75.1	76.3	74.6	77.3	71.2	58.9	72.3	70.8	72.4	74.1
de	80.8	75.7	78.1	70.9	64.6	74.5	75.5	73.8	76	70.4	59.5	71.6	70.1	72.2	72.6
ar	76.4	73.1	70.5	73.3	61.8	70.5	71.1	69.8	72.1	66.6	60.1	67.2	68.7	68.8	68.4
ur	72.5	68.3	66.8	64.7	66	66.7	66.6	65.7	67.7	66.9	57.8	63.6	66.3	64	64.3
ru	79.8	75.9	74.5	69.5	63.5	76.9	75	71.9	75.5	70	60.5	70.5	71.1	71.2	72.4
bg	80.4	77.1	75	70.9	64.3	75.7	78.4	73.4	76.2	70	61.2	71.1	71.6	72.3	72.4
el	78.7	75.5	73.5	69.6	63.1	73	74.5	76.8	74.9	68.9	60.6	68.4	70.1	69.7	70.3
fr	81.4	78.4	74.7	70.8	64.7	75.5	75.8	73.9	79.3	71.5	58.9	72.5	70.5	73.2	73.9
hi	76.8	71.9	70	66.1	65.2	69.8	70.5	68.3	71.2	71	59.7	67	68.4	67.5	67.4
sw	69.5	64.9	63.3	62.9	59.1	64.9	65.6	64.3	64.8	63.2	66	62.9	61.1	62.4	62.8
th	77.5	73.1	71.4	67.8	63.1	71.6	72	70.2	72.4	68.2	59.5	73	69.1	71.8	70.5
tr	76.5	72.5	70.2	67.5	62.1	71.3	72.1	70.3	71.1	67.6	56	68.6	74.5	67.7	69.5
vi	78.2	73.3	71.1	68.3	62.5	71.7	71.9	69.7	72.7	67.8	57.3	70.6	67.9	76	72
zh	78.7	73.9	71.6	67.4	61.3	71.2	71.3	68.9	72.6	66.7	58.9	69	68.4	71.1	75.7

Figure 4: Language-pairs accuracy scores for INFOXLM on the multilingual NLI task.

	en	de	ru	es	zh	vi	ar	tr	el	hi	th	avg
mBERT												
within	73.0	56.6	54.5	57.0	48.1	50.1	44.6	40.1	45.2	43.9	31.3	49.5
across	43.0	36.5	36.0	37.1	30.0	31.2	28.9	24.8	28.9	26.2	17.3	30.9
XLM-r												
within	73.4	59.9	58.0	59.4	53.8	54.7	50.4	52.1	56.4	52.0	54.3	56.8
across	44.0	30.7	29.4	29.2	17.2	20.4	14.8	19.1	22.1	20.1	17.9	28.1
INFOXLM												
within	74.1	60.4	58.9	59.4	57.5	55.9	53.4	53.8	56.5	52.0	54.7	58.5
across	60.3	51.6	51.4	51.8	48.7	47.7	45.4	46.6	47.9	47.0	46.4	49.5

Table 8: Zero-shot EM scores of fine-tuned models for teh QA task using the *within* and *across* language evaluation approaches. There is a more than 50% drop in performance under the *across* evaluation method challenging the extent of cross-linguality of mBERT and XLM-r.
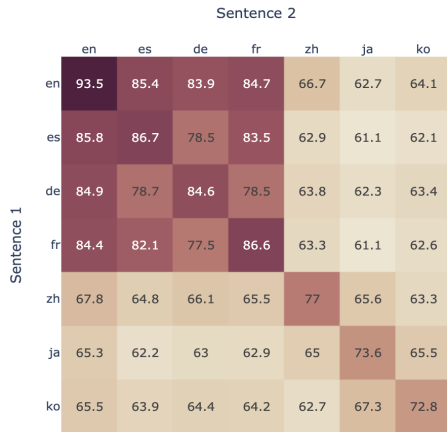
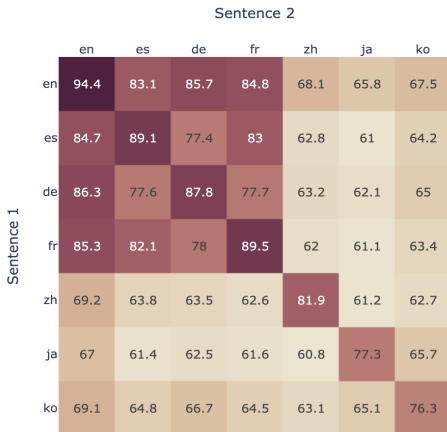Figure 5: Language-pairs accuracy scores for mBERT on the multilingual PI task.

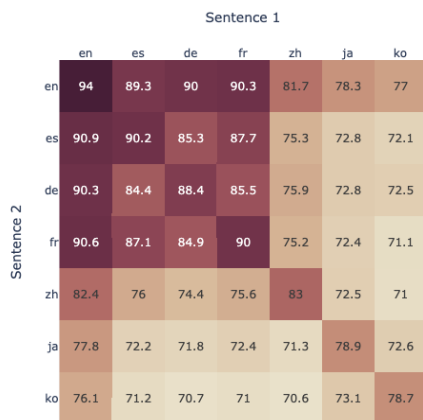Figure 6: Language-pairs accuracy scores for XLM-r on the multilingual PI task.

Figure 7: Language-pairs accuracy scores for INFOXLM on the multilingual PI task.

	$l1-l2$	$l2-l1$	$*-l2$	$l1-*$	$*-*$
en-de	67.8 $_{64.1}$	58.7 $_{71.3}$	63.0 $_{59.4}$	57.6 $_{58.7}$	52.8 $_{55.3}$
en-tr	58.7 $_{52.0}$	40.0 $_{60.3}$	64.2 $_{59.4}$	51.6 $_{52.7}$	50.4 $_{55.3}$
en-el	62.0 $_{56.7}$	49.8 $_{62.3}$	62.8 $_{59.4}$	50.8 $_{55.0}$	53.1 $_{55.3}$
en-es	71.1 $_{68.9}$	41.7 $_{72.9}$	63.6 $_{59.4}$	56.3 $_{58.7}$	47.2 $_{55.3}$
en-ar	46.5 $_{56.5}$	68.5 $_{64.1}$	47.0 $_{59.4}$	57.5 $_{55.0}$	56.6 $_{55.3}$
de-en	78.3 $_{71.3}$	46.3 $_{64.1}$	62.0 $_{58.7}$	47.0 $_{59.4}$	56.6 $_{55.3}$
de-tr	56.5 $_{52.3}$	48.7 $_{53.7}$	62.7 $_{58.7}$	55.8 $_{52.7}$	54.3 $_{55.3}$
de-el	62.6 $_{57.3}$	55.6 $_{57.2}$	62.5 $_{58.7}$	56.1 $_{55.0}$	55.8 $_{55.3}$
de-es	69.0 $_{65.3}$	41.9 $_{61.8}$	61.8 $_{58.7}$	56.6 $_{58.7}$	50.0 $_{55.3}$
de-ar	58.9 $_{57.1}$	62.5 $_{57.5}$	57.3 $_{58.7}$	57.1 $_{55.0}$	56.9 $_{55.3}$
tr-en	76.1 $_{60.3}$	38.7 $_{52.0}$	61.1 $_{52.7}$	51.9 $_{59.4}$	56.9 $_{55.3}$
tr-de	72.3 $_{53.7}$	37.3 $_{52.3}$	62.0 $_{52.7}$	45.4 $_{58.7}$	57.4 $_{55.3}$
tr-el	63.4 $_{52.9}$	41.7 $_{50.8}$	61.7 $_{52.7}$	55.2 $_{55.0}$	56.5 $_{55.3}$
tr-es	64.4 $_{55.1}$	39.7 $_{51.4}$	59.9 $_{52.7}$	57.6 $_{58.7}$	53.5 $_{55.3}$
tr-ar	54.8 $_{52.5}$	53.1 $_{50.5}$	53.8 $_{52.7}$	55.9 $_{55.0}$	54.9 $_{55.3}$
el-en	76.0 $_{62.3}$	40.8 $_{56.7}$	60.1 $_{55.0}$	46.2 $_{59.4}$	54.9 $_{55.3}$
el-de	72.8 $_{57.2}$	39.8 $_{57.3}$	61.9 $_{55.0}$	45.5 $_{58.7}$	56.3 $_{55.3}$
el-tr	69.6 $_{50.8}$	40.1 $_{52.9}$	59.7 $_{55.0}$	40.9 $_{52.7}$	54.3 $_{55.3}$
el-es	67.1 $_{60.9}$	43.6 $_{58.3}$	60.4 $_{55.0}$	57.4 $_{58.7}$	54.1 $_{55.3}$
el-ar	57.1 $_{54.6}$	56.4 $_{54.5}$	57.6 $_{55.0}$	57.3 $_{55.0}$	56.0 $_{55.3}$
es-en	79.3 $_{72.9}$	48.4 $_{68.9}$	61.4 $_{58.7}$	46.3 $_{59.4}$	56.0 $_{55.3}$
es-de	76.4 $_{61.8}$	43.1 $_{65.3}$	63.2 $_{58.7}$	43.8 $_{58.7}$	56.4 $_{55.3}$
es-tr	71.9 $_{51.4}$	39.1 $_{55.1}$	62.4 $_{58.7}$	40.6 $_{52.7}$	55.4 $_{55.3}$
es-el	74.7 $_{58.3}$	38.1 $_{60.9}$	62.8 $_{58.7}$	38.1 $_{55.0}$	55.3 $_{55.3}$
es-ar	42.4 $_{57.3}$	70.1 $_{61.0}$	54.6 $_{58.7}$	59.6 $_{55.0}$	54.9 $_{55.3}$
ar-en	75.9 $_{64.1}$	39.2 $_{56.5}$	59.6 $_{55.0}$	41.6 $_{59.4}$	54.9 $_{55.3}$
ar-de	72.1 $_{57.5}$	38.2 $_{57.1}$	61.1 $_{55.0}$	41.5 $_{58.7}$	55.8 $_{55.3}$
ar-tr	69.3 $_{50.5}$	38.0 $_{52.5}$	58.9 $_{55.0}$	37.7 $_{52.7}$	54.2 $_{55.3}$
ar-el	71.2 $_{54.5}$	36.2 $_{54.6}$	61.0 $_{55.0}$	37.9 $_{55.0}$	55.5 $_{55.3}$
ar-es	73.6 $_{61.0}$	39.8 $_{57.3}$	60.8 $_{55.0}$	43.3 $_{58.7}$	54.6 $_{55.3}$

Table 9: mBERT's performance on the XNLI test set. The columns show the fine-tuning language pairs, and the rows show the evaluation pairs in which $l1$ and $l2$ represent the premise and hypothesis's languages, respectively, as given in the corresponding columns. The rows including $*$ show the average performance over all languages. The smaller numbers present the baseline performance (fine-tuning on en-en) for the corresponding evaluation pairs.

Question Language

	en	ar	de	el	es	hi	ru	th	tr	vi	zh
en	84.5	50.6	68.7	48.8	70	42.4	65.9	22.4	42.3	57.9	57.6
ar	58.7	61.2	49	39.1	52.4	35.4	50.6	21.2	32.7	43.1	43.7
de	72.9	43.9	72.7	45.3	62.6	38.8	62.6	20.2	37.4	51.8	49.3
el	62.8	42.3	52.9	62.4	54.7	34.6	53.4	19.9	33.1	45.3	43.6
es	75.9	50.5	64.2	48.4	75.5	40.2	64.4	23.3	37.7	53.5	53.7
hi	57.6	39.8	49.5	39.1	50.2	58	48.5	20.2	31.4	40.9	41.5
ru	67.8	46	58.8	44.8	62	38.5	71.4	21.1	36.8	49.6	49.3
th	38.1	30.5	33.2	29.6	35.3	26	34.4	40	24.1	31	30.9
tr	55.8	35.7	47.7	35.7	49.1	32.7	47.2	18.8	55.1	39.5	40.2
vi	68.2	42.1	56.8	41.6	58	36.6	55.8	23.4	32.4	69.2	54.2
zh	54.8	37	46.8	33.7	47.2	30.6	47.1	18.9	28.7	43.6	58.2

(Context Language — vertical axis)

Figure 8: Language-pairs F1 score for mBERT on the multilingual QA task.

Question Language

	en	ar	de	el	es	hi	ru	th	tr	vi	zh
en	84.2	38.1	64.1	53.1	62.6	45.7	67.2	34.4	40.9	49.1	43.8
ar	59.2	66.3	36.7	18.8	35.1	15.6	35	14.6	15.9	17.4	17.8
de	73.4	25.2	75.2	40.3	53.6	29.9	58	24.2	29.9	37.2	30.2
el	68.7	23.4	51.1	73.8	54.3	25.2	48.2	17.5	24.9	29.4	21.4
es	75	27.4	55.1	43.3	77.1	29.1	58.6	23.4	28.4	32	31.2
hi	66.1	25.1	46.7	24.7	46.2	68.3	43.6	18.4	28	30.7	32.1
ru	70.3	23.2	57.8	36.6	55.4	28.1	74.5	20.2	29.3	30.3	25.5
th	61.8	28.6	42.8	29.1	43.1	31.1	44.7	66.5	24.2	35.3	36.5
tr	63.3	20.8	46.7	37.5	42.3	29.4	48.8	21.2	68.1	28.7	26.5
vi	68.8	23	46.6	29.8	39.7	28.5	49.5	24.2	22.3	74.3	31
zh	58.2	16	32	17.5	31.6	22.2	28.2	17.6	17.6	24.1	63.7

(Context Language — vertical axis)

Figure 9: Language-pairs F1 score for XLM-r on the multilingual QA task.

Figure 10: Language-pairs F1 score for INFOXLM on the multilingual QA task.

Figure 11: The average distance of the questions' words occurred in the context to the center of the answer span in the top 20% easiest and hardest instances for XLM-r fine-tuned on SQuAD based on the test set of every language.
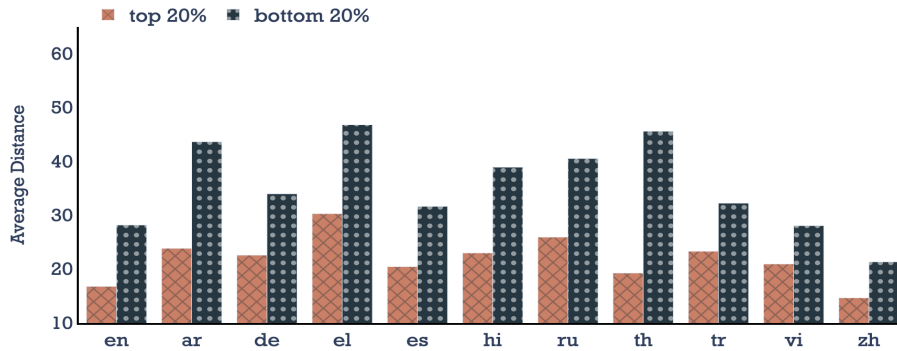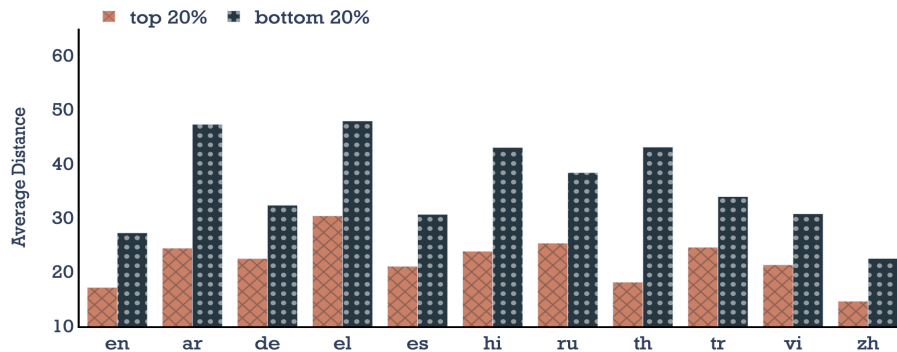

Figure 12: The average distance of the questions' words occurred in the context to the center of the answer span in the top 20% easiest and hardest instances for INFOXLM fine-tuned on SQuAD based on the test set of every language.

	en	de	fr	ru	es	zh	vi	ar	tr	bg	el	ur	hi	th	sw	avg
Entailment																
within	81.6	66.9	70.1	65.5	70.5	59.1	62.7	60.4	61.9	69.3	65.9	47.0	53.9	55.7	60.1	63.4
across	54.5	49.4	53.0	50.9	53.6	42.5	43.9	41.3	42.5	52.4	47.5	36.5	42.0	41.9	27.1	45.3
NotEntailment																
within	86.5	81.8	82.2	80.9	83.5	80.8	80.6	77.2	78.7	82.7	80.8	75.1	77.1	80.0	67.9	79.7
across	80.6	76.5	77.1	76.8	77.2	75.6	76.1	74.1	73.4	76.5	76.3	72.1	74.7	75.9	67.5	75.4

Table 10: Performance of XLM-r setting per label on the NLI task. Most of the performance drop in the *across* setting occurs for the *entailment* class.

	en	de	fr	ru	es	zh	vi	ar	tr	bg	el	ur	hi	th	sw	avg
Entailment																
within	82.4	68.0	70.8	66.1	72.0	61.7	63.5	60.5	62.0	68.9	66.4	45.9	55.2	56.5	53.0	63.5
across	64.8	54.6	56.4	54.8	57.0	49.2	49.5	48.2	49.3	55.8	51.8	39.4	48.4	48.7	34.4	50.8
NotEntailment																
within	87.7	83.1	83.6	82.3	84.0	82.7	82.2	79.7	80.7	83.2	82.0	76.1	78.9	81.3	72.5	81.3
across	83.2	80.7	81.1	80.4	81.3	80.4	80.1	79.2	78.9	80.8	80.3	77.3	78.9	80.2	75.2	79.9

Table 11: The performance of INFOXLM setting per label on the NLI task. Most of the performance drop in the *across* setting occurs for the *entailment* class.

	en	de	fr	es	zh	ko	ja	avg
Paraphrase								
within	95.8	87.7	89.6	88.8	80.9	63.5	71.8	82.6
across	58.4	47.3	47.8	47.9	26.5	32.3	25.2	40.8
NonParaphrase								
within	93.2	87.8	89.4	89.4	82.7	86.7	81.7	87.3
across	91.1	92.2	92.0	91.9	93.6	91.8	93.2	92.3

Table 12: Performance of XLM-r for the *within* and *across* setting per label on the PAWS-X test set. Most of the performance drop of the *across* setup setting originates from the drop in the *Paraphrase* class.

	en	de	fr	es	zh	ko	ja	avg
Paraphrase								
within	96.0	92.4	92.7	91.2	86.8	74.8	81.3	87.9
across	84.5	77.4	77.4	76.8	70.3	65.0	67.8	74.2
NonParaphrase								
within	92.9	85.1	88.3	90.0	79.8	82.2	77.3	85.1
across	85.5	81.6	82.8	83.2	78.8	78.4	77.5	81.1

Table 13: Performance of INFOXLM for the *within* and *across* setting per label on the PAWS-X test set. Most of the performance drop of the *across* setup setting originates from the drop in the *Paraphrase* class.

	en	de	fr	ru	es	zh	vi	ar	tr	bg	el	ur	hi	th	sw	ko	ja	avg
XNLI																		
within	81.3	74.1	75.8	72.9	76.0	71.0	71.0	69.5	71.3	75.2	72.9	63.8	66.7	68.8	62.6	-	-	71.3
across	68.1	63.7	65.2	65.1	65.2	61.6	60.4	59.9	60.1	64.9	63.4	57.1	60.2	61.8	51.7	-	-	61.9
PI																		
within	54.6	55.2	54.8	-	54.6	55.3	-	-	-	-	-	-	-	-	-	55.2	55.8	55.1
across	55.0	55.2	55.1	-	55.0	55.2	-	-	-	-	-	-	-	-	-	55.2	55.5	55.2
QA																		
within	82.8	73.6	-	72.8	76.1	62.7	72.7	63.9	66.1	-	71.4	-	66.6	66.7	-	-	-	70.5
across	56.6	44.1	-	43.5	42.9	30.2	34.7	27.8	34.7	-	37.5	-	34.7	34.1	-	-	-	38.3

Table 14: Performance of XLM-r fine-tuned on the control tasks. We report the accuracy score for the NLI and PI tasks and the F1 score for QA. Although the fine-tuning data does not train the model with any task-related knowledge, the drop in the performance is negligible.