How to Translate Your Samples and Choose Your Shots? Analyzing Translate-train & Few-shot Cross-lingual Transfer

Anonymous ACL submission

Abstract

Translate-train or few-shot cross-lingual transfer can be used to improve the zero-shot performance of multilingual pretrained language models. Few-shot utilizes high-quality low-quantity samples (often manually translated from the English corpus). Translate-train employs a machine translation of the English corpus, resulting in samples with lower quality that could be scaled to high quantity. Given the lower cost and higher availability of machine translation compared to manual professional translation, it is important to systematically compare few-shot and translate-train, understand when few-shot is beneficial, and whether choosing the shots to translate increases the few-shot gain. This work aims to fill this gap: we compare and quantify the performance gain of few-shot vs. translate-train using a varying number of samples for three tasks/datasets (XNLI, PAWS-X, XQuAD) spanning 17 languages. We show that scaling up the training data using machine translation gives a larger gain compared to using the small-scale (higher-quality) few-shot data. When few-shot is beneficial, we show that there are random sets of samples that perform better across languages and that the performance on English and on the machine-translation of the samples can both be used to choose the shots to manually translate for an increased few-shot gain.[1]

1 Introduction

With the emergence of large-scale multilingual Pretrained Language Models like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), a significant amount of research went into exploring the cross-lingual transfer capabilities of these models, allowing for an easier adaptation to a task in many various languages. This is achieved through a number of approaches.

Zero-shot cross-lingual transfer has become a research focus, e.g. XTREME / XTREME-R benchmark (Hu et al., 2020; Ruder et al., 2021). In this approach, transfer to new languages is done by fine-tuning a multilingual PLM on the task at issue, using only an English corpus (source language) and reporting the performance on multiple target languages.

Few-shot cross-lingual transfer was recently shown to give an advantage over zero-shot cross-lingual transfer (Lauscher et al., 2020). In this approach, it's shown that fine-tuning the model using a small amount of target-language task data (few-shot) improves the performance, especially for low-resource languages.

Translate-train is another common approach to improve the performance. Here the full training dataset is machine translated to the target language and used for fine-tuning. There exists relatively good Machine Translation (MT) systems for the languages that are usually studied in the few-shot approach[2] that could be used in translate-train.

In the following, we use few-shot to refer to fine-tuning using fewer samples of high-quality professional manual translation. Translate-train is used to refer to fine-tuning using lower-quality machine translation that has the potential to be scaled to a larger number of samples. Although some research has dealt with few-shot cross-lingual transfer and analyzing it (Lauscher et al., 2020; Zhao et al., 2021), no systematic study was done to compare it to translate-train. Given that both zero-shot and few-shot cross-lingual transfer assume the availability of a large-scale English corpus of the task for source training, we hypothesize that the *translate-train approach might have an advantage over few-shot given the scale of data that would be available even if not at the best quality.*

[1] Our code will be published under: https://www.gihtub.com/***

[2] All target languages in the studied datasets are supported by e.g. Google Translate: https://cloud.google.com/translate/docs/languages

When there is, on the other hand, an actual need for or a benefit from doing few-shot cross-lingual transfer and therefore a need for professional translation of some samples for training, this usually entails significantly more effort and cost compared to using MT. It is then important to *find out which samples to manually translate given the high variance in performance depending on the choice of samples* (Zhao et al., 2021).

In this work we investigate both those research directions on 3 high-level semantic tasks and datasets, XNLI (Natural Language Inference), X-PAWS (Paraphrase Detection) and XQUAD (Question Answering), spanning 17 diverse languages. We investigate the following research questions:

Q1. How does the performance of few-shot cross-lingual transfer compare to that of translate-train? We show that there is a performance advantage for few-shot transfer over translate-train given the same number of samples, but that with the increase of samples used for translate-train, this gap shrinks, and using the full large-scale corpus in translate-train results in a clear advantage over few-shot. We show that at a scale of 10x-100x of machine-translation to manual-translation, quantity trumps quality and it's recommended in this case to use translate-train if MT is available for the language. Few-shot transfer still has an advantage when less source data is available and it's therefore not possible to benefit from the scale gain of using MT.

Q2. Are there sets of samples that have better few-shot performance if translated and how can those sets be identified? We show that when few-shot transfer is beneficial for the task, there are random sets of samples that perform better across most target languages and across different model initialization. We investigate using the performance on the English version of the samples and the machine-translated version to choose the best candidates to manually translate and use for few-shot transfer. We show that there is a correlation between the performance of the same set of shots across languages and that the few-shot samples that perform well on the source language, English, usually perform also better across languages and on average. The same correlation is seen between the performance of manual and machine translation. We show empirically that choosing the sets of samples for few-shot transfer using those heuristics or a model based on those features results in more bang for your shots.

2 Related Work

Cross-lingual transfer: The cross-lingual transfer capabilities of multilingual pretrained language models have led to major recent advances and a growing number of such models have been introduced, e.g., mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021) etc. The cross-lingual transfer is usually exploited in a zero-shot setup, and benchmarks are built based on this assumption e.g. XTREME/XTREME-R (Hu et al., 2020; Ruder et al., 2021). We conduct our experiments on 3 datasets from the XTREME benchmark and use the provided machine-translated training data for our translate-train experiments.

Few-shot: There has been recently some focus on few-shot cross-lingual transfer and its analysis. Lauscher et al. (2020) shows the effectiveness of few-shot compared to zero-shot cross-lingual transfer especially in lower-resource and distant languages, where zero-shot transfer is least effective and few-shot gives a bigger advantage. Zhao et al. (2021) further analyzes few-shot cross-lingual transfer emphasizing that the choice of shots has a significant effect on the performance. The experiments are conducted on few-shot at a smaller scale at around 10 samples. We conduct larger-scale few-shot experiments with a size up to hundreds of samples and focus on choosing the best-performing samples.

Translate-train: is commonly used to boost the performance for a target language using machine translation (Conneau et al., 2020; Hu et al., 2020) but no systematic study has been done on the effect of the size of the translated data and the interplay of data quality vs. quantity in this context.

3 Datasets

<table>
<tr><th>Dataset</th><th>|Train|</th><th>|mDev|</th><th>|mTest|</th><th>|Langs|</th><th>metric</th></tr>
<tr><td>XNLI</td><td>392,702</td><td>2,490</td><td>5,010</td><td>15</td><td>Acc</td></tr>
<tr><td>PAWS-X</td><td>49,401</td><td>2,000</td><td>2,000</td><td>7</td><td>Acc</td></tr>
<tr><td>XQuAD</td><td>87,599</td><td>261</td><td>1,190-261= 930</td><td>11</td><td>F1</td></tr>
</table>

Table 1: Datasets statistics. Train is the English training dataset. |mDev| and |mTest| are used to indicate the size of the multilingual split of the dataset.

We focus on high-level tasks and conduct our experiments on 2 classification tasks and a question answering task (Table 1). The details and properties of the languages can be found in Appendix Table 8. When attempting to choose the shots, we rely on measuring the performance of the same set of samples across different languages. This is why we are limited to datasets with parallel corpus, i.e. the target language corpus is created by translating the English corpus as opposed to collecting and annotating the target language corpus from scratch (refer to XTREME/XTREME-R for an overview).

XNLI: The Cross-lingual Natural Language Inference corpus (Conneau et al., 2018) is a translation of the dev and test set of the MultiNLI dataset (Williams et al., 2018) by professional translators into 14 languages. The dataset consists of pairs of sentences, a premise and a hypothesis, where the task is to predict whether the premise entails, contradicts, or is neutral to the hypothesis. The full English training data from MultiNLI is used for source training.

PAWS-X: The Cross-lingual Paraphrase Adversaries from Word Scrambling (Yang et al., 2019) dataset is a translation of the dev and test set of the PAWS dataset (Zhang et al., 2019) by professional translators into 6 languages. The dataset consists of pairs of sentences and the task is to predict whether those two sentences are paraphrases of each other. The full English training data from PAWS is used for source training.

XQuAD: The Cross-lingual Question Answering Dataset (Artetxe et al., 2020b) is a professional translation of the dev set from SQuAD v1.1 (Rajpurkar et al., 2016) into 10 other languages. The dataset consists of a paragraph and a set of questions. The task is to select the span of the paragraph that answers the question. We reserve 10 paragraphs from the multilingual split, similar to Lauscher et al. (2020), as our dev set resulting in a total of 261 question/answer samples. The rest is used as test set.

4 Experiments

For each task, we fine-tune XLM-R *base* on the source language (English) corpus for 5 epochs with early stopping using the loss on the English dev set. We then continue fine-tuning the model on the target language either in a few-shot or translate-train setup as explained in the following sections. For the two classification tasks, we use a maximum sequence length of 128. We limit hyperparmeter tuning to a search for the learning rate in $\{7e-6, 1e-5, 3e-5\}$ and use a batch size of 32. For Question Answering, we use a maximum sequence length of 384 with a paragraph slide of 128. We train using a learning rate of $3e-5$ and a batch size of 12 for 2 epochs. More details about training can be found in Appendix A.

4.1 Few-shot experiments

We use samples from the multilingual dev set as training samples. Few-shot fine-tuning is done as follows: for each language, we separately continue fine-tuning the source model for one epoch on $n \in \{10, 50, 100, 500, 1k\}$ samples from the target language corpus for the two classification tasks and for $n \in \{10, 50, 100, 250\}$ for the Question Answering task, given the smaller amount of data available for training in this case. We report the results on the test set for each target language.

For each n number of samples, the performance is averaged across 5 different sets of randomly-chosen samples using 5 different fine-tuned models with different random initialization, 25 runs in total. This is done to ensure more robust results when measuring the gain over zero-shot given the high variance across the different sets of samples (Zhao et al., 2021) as well as the variance in zero-shot performance across the random initialization (Keung et al., 2020)

For comparing the performance across shots, we make sure to use the same set of parallel samples across languages, using the sample ids, to compare how a set of samples performs when translated to different languages. This is possible due to our selection of tasks and datasets that have a parallel corpus for the various target languages.

4.2 Translate-train experiments

We train using a machine translation of the English train set to each target language[3] and adapt a similar setup as few-shot (Section 4.1): for each language, we separately continue fine-tuning on $n \in \{10, 50, 100, 500, 1k, 10k, |dataset|\}$ samples from the machine-translated train set and report the results on the test set of the target language.

5 Results

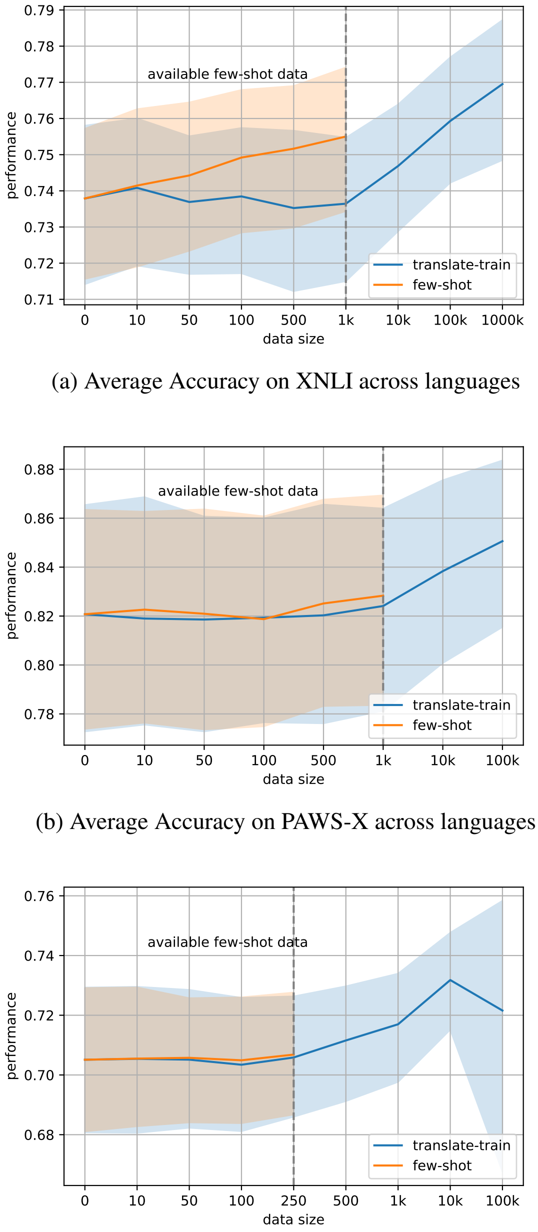
In the following sections, we present our findings while attempting to answer the following questions: 1. How to translate? Using manual or machine translation? (Section 5.1) 2. How to choose the best shots to manually translate? (Section 5.2)

[3] We use the Machine Translation provided by the XTREME Benchmark: https://console.cloud.google.com/storage/browser/xtreme_translations

5.1 How to translate your samples? Few-shot vs. translate-train

	XNLI	PAWS-X	XQuAD
English	84.04±0.65	93.99±0.35	83.10±0.29
cross-lingual transfer (average over all languages)			
zero-shot	-10.26±0.34	-11.92±0.92	-12.60±0.35
few-shot	-8.54±0.30	-11.16±0.52	-12.42±0.30
translate-train	**-7.09±0.32**	**-8.93±0.66**	**-10.95±0.16**

Table 2: **Gap to English performance**. Performance gap between the average performance for all languages compared to the performance for English



(a) Average Accuracy on XNLI across languages

(b) Average Accuracy on PAWS-X across languages

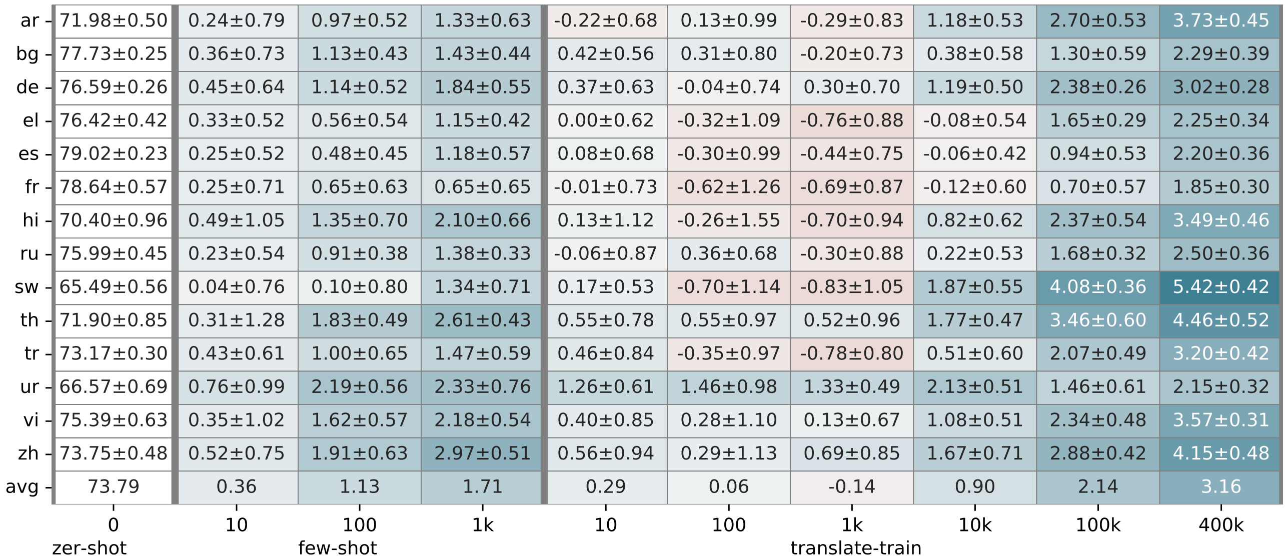
(c) Average F1 on XQuAD across languages

Figure 1: Average performance across languages for **translate-train vs. few-shot**. The biggest performance boost comes from using translate-train

To demonstrate the full potential for each approach, Table 2 shows a summary of the results for zero-shot, few-shot and translate-train when the maximum possible number of samples is used. **The gap to English performance** is the average of the gap between the target language performance and the performance on the English test set. Both few-shot and translate-train help bridge the gap, but using translate-train on a large scale has an advantage in further narrowing the gap as compared to the small scale that is possible with few-shot transfer. This results in a further gain for translate-train over few-shot by 1.45%, 2.22%, 1.47% on average for XNLI, PAWS-X and XQuAD.

To see the effect of the available dataset size in each scenario, Figure 1 shows the average performance across languages for **few-shot vs. translate-train** for various number of samples. We can see an advantage of having manual translation over machine translation resulting in a clear gap in performance in XNLI for the same number of samples as seen in Figure 1a. This gap increases with the increase of the size of training samples as seen at 1k samples. The availability of manual translation for few-shot transfer is limited though and starting from 10k-100k, the scale of translate-train has an advantage for all tasks. The performance on PAWS-X and XQuAD does not improve much with few-shot as shown in Figure 1b and Figure 1c , and the clear boost comes from using the large scale machine-translated dataset. We discuss the observed large variance on XQuAD across languages in the following paragraph.

The detailed results for zero-shot, few-shot and translate-train are shown for XNLI in Figure 2. We focus here on XNLI because it was the task with a clear few-shot gain (refer to Figure 5 and 6 in the Appendix for PAWS-X and XQuAD). We report the performance gains for few-shot and translate-train over zero-shot for each language across varying sizes of samples. We notice a performance advantage in general across the different sizes for both few-shot and translate-train (mostly starting from 1k). The gain is larger for low-resource languages like Swahili (sw) and non-European languages like Chinese (zh). Those languages also tend to have a larger zero-shot performance gap to English and this observation is seen for both



	zer-shot	few-shot			translate-train					
	0	10	100	1k	10	100	1k	10k	100k	400k
ar	71.98±0.50	0.24±0.79	0.97±0.52	1.33±0.63	-0.22±0.68	0.13±0.99	-0.29±0.83	1.18±0.53	2.70±0.53	3.73±0.45
bg	77.73±0.25	0.36±0.73	1.13±0.43	1.43±0.44	0.42±0.56	0.31±0.80	-0.20±0.73	0.38±0.58	1.30±0.59	2.29±0.39
de	76.59±0.26	0.45±0.64	1.14±0.52	1.84±0.55	0.37±0.63	-0.04±0.74	0.30±0.70	1.19±0.50	2.38±0.26	3.02±0.28
el	76.42±0.42	0.33±0.52	0.56±0.54	1.15±0.42	0.00±0.62	-0.32±1.09	-0.76±0.88	-0.08±0.54	1.65±0.29	2.25±0.34
es	79.02±0.23	0.25±0.52	0.48±0.45	1.18±0.57	0.08±0.68	-0.30±0.99	-0.44±0.75	-0.06±0.42	0.94±0.53	2.20±0.36
fr	78.64±0.57	0.25±0.71	0.65±0.63	0.65±0.65	-0.01±0.73	-0.62±1.26	-0.69±0.87	-0.12±0.60	0.70±0.57	1.85±0.30
hi	70.40±0.96	0.49±1.05	1.35±0.70	2.10±0.66	0.13±1.12	-0.26±1.55	-0.70±0.94	0.82±0.62	2.37±0.54	3.49±0.46
ru	75.99±0.45	0.23±0.54	0.91±0.38	1.38±0.33	-0.06±0.87	0.36±0.68	-0.30±0.88	0.22±0.53	1.68±0.32	2.50±0.36
sw	65.49±0.56	0.04±0.76	0.10±0.80	1.34±0.71	0.17±0.53	-0.70±1.14	-0.83±1.05	1.87±0.55	4.08±0.36	5.42±0.42
th	71.90±0.85	0.31±1.28	1.83±0.49	2.61±0.43	0.55±0.78	0.55±0.97	0.52±0.96	1.77±0.47	3.46±0.60	4.46±0.52
tr	73.17±0.30	0.43±0.61	1.00±0.65	1.47±0.59	0.46±0.84	-0.35±0.97	-0.78±0.80	0.51±0.60	2.07±0.49	3.20±0.42
ur	66.57±0.69	0.76±0.99	2.19±0.56	2.33±0.76	1.26±0.61	1.46±0.98	1.33±0.49	2.13±0.51	1.46±0.61	2.15±0.32
vi	75.39±0.63	0.35±1.02	1.62±0.57	2.18±0.54	0.40±0.85	0.28±1.10	0.13±0.67	1.08±0.51	2.34±0.48	3.57±0.31
zh	73.75±0.48	0.52±0.75	1.91±0.63	2.97±0.51	0.56±0.94	0.29±1.13	0.69±0.85	1.67±0.71	2.88±0.42	4.15±0.48
avg	73.79	0.36	1.13	1.71	0.29	0.06	-0.14	0.90	2.14	3.16

Figure 2: **Detailed Results on XNLI**. Gains in performance over zero-shot for few-shot and translate-train. Low-resource languages like Swahili have the most gains in both cases

few-shot and translate-train. Once the full machine-translated training set is used, a clear advantage for translate-train is seen across almost all languages and in all tasks. We can see that the gain for Urdu (ur) is the highest up until 100k when it starts decreasing. We think this might be due to a lower-quality machine translation. The same effect is seen for Thai (th) on XQuAD with a significant performance degrade when the full training dataset is used (details in the Appendix in Figure 6). This is also the reason for the degrade and high variance seen at this point in Figure 9b.

We investigated whether training for more epochs would have changed the results and would have been beneficial, especially for the few-shot scenario where longer training on the manual high-quality translation might be beneficial. We split the available set of samples into train/dev and train for 10 epochs with early stopping on dev. Although some languages seem to benefit from this setup, it still yields comparable results and translate-train still has a clear advantage. (See Figure 7 and 8 in the Appendix for the detailed results).

5.2 How to choose your shots? Which samples to translate for few-shot?

Few-shot can still have an advantage over translate-train when the English dataset is not large enough to benefit from the scale effect of translate-train.

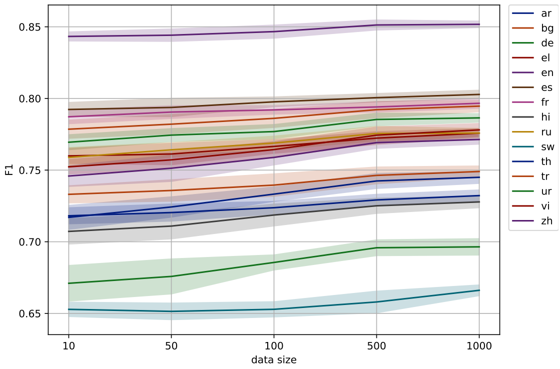


Figure 3: **XNLI accuracy variance on different shots**. High variance decreases with an increased data size

It can also be necessary when adapting to a target language that does not have an existing machine translation system or does not have a good one. Creating few-shot samples, in this case, can be done by collecting and labeling new samples or by translating samples from the available English dataset. The latter is a common method and 4 out of the 7 non-retrieval datasets in XTREME use manual professional translation to create samples in the target languages (all of which high-level semantic tasks). It is beneficial then to support in selecting the samples with higher-performance potential to translate and do few-shot training on.

	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh
ar	1.00	0.76	0.76	0.75	0.77	0.76	0.72	0.65	0.51	0.56	0.75	0.85	0.74	0.85	0.77
bg	0.76	1.00	0.74	0.65	0.59	0.76	0.59	0.63	0.59	0.69	0.73	0.65	0.62	0.82	0.64
de	0.76	0.74	1.00	0.69	0.62	0.77	0.50	0.69	0.64	0.61	0.77	0.72	0.63	0.80	0.71
el	0.75	0.65	0.69	1.00	0.69	0.74	0.56	0.66	0.45	0.69	0.65	0.77	0.55	0.73	0.67
en	0.77	0.59	0.62	0.69	1.00	0.73	0.79	0.66	0.15	0.62	0.57	0.79	0.38	0.65	0.55
es	0.76	0.76	0.77	0.74	0.73	1.00	0.55	0.60	0.51	0.71	0.66	0.75	0.50	0.77	0.66
fr	0.72	0.59	0.50	0.56	0.79	0.55	1.00	0.65	0.17	0.39	0.55	0.73	0.52	0.63	0.51
hi	0.65	0.63	0.69	0.66	0.66	0.60	0.65	1.00	0.53	0.62	0.79	0.70	0.63	0.77	0.74
ru	0.51	0.59	0.64	0.45	0.15	0.51	0.17	0.53	1.00	0.42	0.65	0.44	0.69	0.68	0.69
sw	0.56	0.69	0.61	0.69	0.62	0.71	0.39	0.62	0.42	1.00	0.63	0.60	0.35	0.65	0.54
th	0.75	0.73	0.77	0.65	0.57	0.66	0.55	0.79	0.65	0.63	1.00	0.74	0.74	0.83	0.85
tr	0.85	0.65	0.72	0.77	0.79	0.75	0.73	0.70	0.44	0.60	0.74	1.00	0.64	0.78	0.76
ur	0.74	0.62	0.63	0.55	0.38	0.50	0.52	0.63	0.69	0.35	0.74	0.64	1.00	0.77	0.79
vi	0.85	0.82	0.80	0.73	0.65	0.77	0.63	0.77	0.68	0.65	0.83	0.78	0.77	1.00	0.82
zh	0.77	0.64	0.71	0.67	0.55	0.66	0.51	0.74	0.69	0.54	0.85	0.76	0.79	0.82	1.00
avg	0.75	0.70	0.71	0.68	0.64	0.70	0.59	0.69	0.54	0.61	0.73	0.73	0.64	0.77	0.71

Table 3: XNLI Pearson correlation of the performance between languages

	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh
	0.65	0.86	0.89	0.75	0.88	0.88	0.80	0.75	0.72	0.85	0.86	0.61	0.89	0.89

Table 4: XNLI Pearson correlation between the performance of machine translation and manual translation

To emphasize the significance of choosing the samples to translate and how the performance is affected by this choice, we plot in Figure 3 the XNLI **performance variance on different shots** (using the same model initialization) across 20 sets of random few-shot samples varying in size from 10 to 1000 samples. We can see that the performance varies, sometimes significantly, depending on the set of samples used. Zhao et al. (2021) has shown similar variance observations on a smaller number of samples (around 10). We consider here a larger size range that is more representative of the expected data size if a manual translation is conducted. The performance variance across shots decreases with the increased number of shots in particular starting at 100 samples. This means that choosing the shots to translate is more important when smaller size of samples is used. (A somewhat similar observation is shown in the Appendix on PAWS-X and XQuAD in Figure 9a and 9b although for the latter the variance increases with the size).

In the following, we focus mainly on XNLI as the task that had the most few-shot gain. We investigate whether there are sets of samples that have a potential for better performance across languages and what could be an indication of that. For a set of shots, we consider two indicators: the performance of this set in another language, and the performance on the MT of the samples in the set.

5.2.1 Correlation between performance across languages

If the performance of a set of samples for one language can be an indication of its performance on another language, we expect to see a high correlation between the performance for both languages. To estimate this, we calculate the performance using the different manual translations across languages of the same set of training samples (shots). We then calculate the Pearson correlation of the performance across 5 random sets of samples (with varying sample-set sizes) using 5 models with different random initialization. As seen in Table 3, there is a high positive correlation between the performance on XNLI for the various languages. This is also the case, but to a lesser degree for PAWS-X as seen in the Appendix Table 10. XQuAD, on the other hand, has low and sometimes even negative correlation as seen in the Appendix Table 13. This might be due to the nature of the task or the fact that we have less data in this case for both training and test. It is also worth noting that the correlation is lower for both tasks, PAWS-X and XQuAD, which had low few-shot gain.

We can see that **English** has a good correlation for XNLI and PAWS-X, so we can consider the performance on English an indication of how well the set of samples would perform if translated to another language. A breakdown of the English

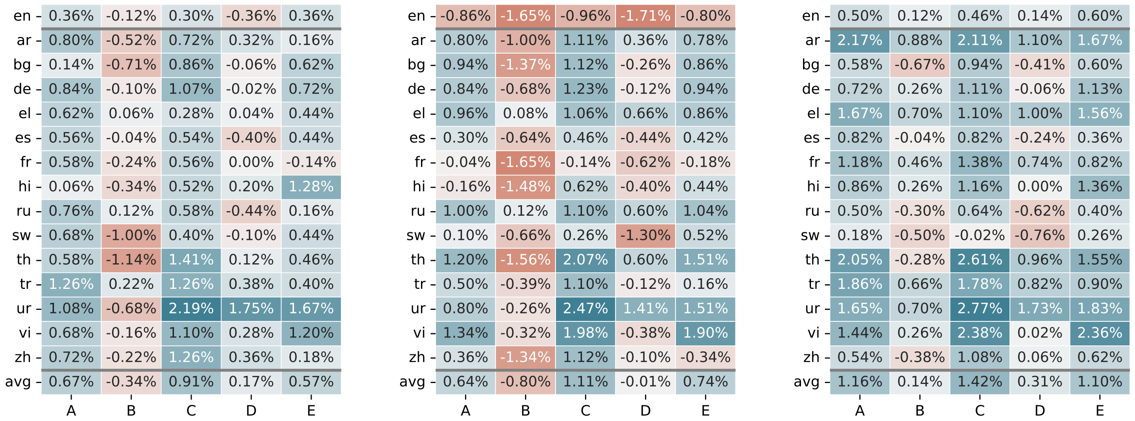


Figure 4: **XNLI few-shot gain** over zero-shot across 5 sets of samples (size=10) for 3 different model initializations. Sets C, A, and E yield better performance for the 3 different initializations. The English performance can be used as an indicator when choosing samples to translate

correlation based on data size is show in the Appendix Table 9 and 11. As an example of this, Figure 4 shows **XNLI few-shot gain** over zero-shot performance for 5 random sets of samples $\{A, B, C, D, E\}$ each containing 10 samples. The performance is shown for 3 different model initializations. We can see that the sets $\{A, C, E\}$ perform better than $\{B, D\}$ across target languages and on average as well as across different initializations. The performance on English can be used as an indicator of the best shots to choose and translate to a target language as seen when comparing the English performance (top) to the average (bottom, excluding the English performance). This is here the case even when further fine-tuning a model on English samples results in a decreased English performance as seen for the 2nd model initialization. The least negative sets of samples still correspond to the best performing shots. The results generalize for varying sizes of few-shot sets as seen for example in the case of 1000 samples in the Appendix Figure 10.

5.2.2 Correlation between manual and machine translation performance

Another possible indicator of the best performing set of samples could be the performance of the samples in the set when they are machine translated from English to the target language. Artetxe et al. (2020a) has shown that subtle patterns in the (machine or manual) translated samples can have a notable impact on the model performance, so it is important to empirically study the relation between both. Similar to the above, we calculate the correlation between the performance for both manual and machine translation of the same set of samples for each target language. As seen for XNLI in Table 4, there is an even higher correlation than with the English performance. A somewhat lower correlation is seen for PAWS-X in Appendix Table 12. Lower correlation might be a result of lower-quality machine translation or a result of the different patterns introduced by machine translation that affects the performance as mentioned before.

5.2.3 Gain from choosing shots and performance prediction

We show in Table 5 the **few-shot performance gain** resulting from choosing the shots with the highest English performance and the highest machine-translation performance. This is in comparison to the average few-shot gain across the different shots in *no choosing (avg)*, but also to the minimum in *no choosing (min)*, because an important aspect of choosing the shots is avoiding the worst-performing ones (Comparing to the average hides the fact that we might accidentally use a very bad set of shots). We exclude XQuAD from our results because of the low few-shot gains and the low correlations which resulted in no gains when choosing the shots (detailed results in Appendix Figure 12). We can see a clear gain when using *en performance* or *mt performance* for both XNLI and PAWS-X. The few-shot gain with chosen-shots is most significant at smaller data sizes where the gain is almost double that from *no choosing (avg)*.

		10	50	100	500	1000	avg
XNLI	no choosing (avg)	0.36	0.64	1.13	1.38	1.71	1.04
	no choosing (min)	0.04	-0.15	0.10	0.36	0.65	-2.26
	en performance	0.71	1.15	1.32	1.82	1.90	1.38
	mt performance	**0.88**	1.08	1.36	1.81	2.01	1.43
	en + mt model	0.85	1.11	1.42	**1.85**	2.01	1.45
	+ lang features	0.83	**1.13**	**1.44**	**1.85**	**2.03**	**1.46**
PAWS-X	no choosing (avg)	0.19	0.02	-0.20	0.44	0.76	0.24
	no choosing (min)	-0.34	-0.43	-1.05	-0.23	0.10	-4.70
	en performance	0.17	0.10	**0.23**	**0.53**	0.71	0.35
	mt performance	**0.38**	**0.19**	0.09	0.42	0.73	**0.36**
	en + mt model	0.32	0.09	0.13	0.44	0.76	0.35
	+ lang features	0.26	0.04	0.00	0.52	**0.84**	0.33

Table 5: **Chosen-shots performance gain**. Gain over zero-shot performance when choosing the best set of shots using a heuristic (en or mt performance) or a linear model that predicts the performance. No gain is observed for XQuAD

	XNLI		PAWS-X	
	MSE	RMSE	MSE	RMSE
avg (baseline)	1.05±0.56	0.99±0.26	1.26±0.76	1.08±0.34
model using features:				
en performance	0.68±0.41	0.80±0.23	1.08±0.92	0.97±0.42
mt performance	0.34±0.28	0.56±0.20	0.92±0.56	0.93±0.28
en + mt performance	0.33±0.26	0.55±0.18	0.91±0.56	0.92±0.28
+ lang features	**0.32±0.25**	**0.54±0.18**	**0.58±0.27**	**0.75±0.17**
only lang features	0.93±0.47	0.93±0.24	1.01±0.45	0.98±0.25

Table 6: **Performance prediction error**. Predicting the few-shot performance gain using models with the English and MT performance as features. *+lang features* further adds features from lang2vec

We investigate improving our choices by combining the various performance values as features to a linear model that predicts the performance gain. This is also helpful to avoid selecting any set of samples if none are expected to result in a positive and significant improvement. We use the performance metrics as a dataset based on 5 different random sets of samples for 5 different model initialization with varying sample sizes across all languages (excluding English). This results in 1750, 750, 1100 data points for XNLI, PAWS-X and XQuAD. For each language, we train the model using the data from all other languages and evaluate on the selected language. Cross-validation is done on the data after excluding the selected language to choose the best hyperparameters. For each language, the average performance gain for all other languages is used as a baseline.

The following **features** are considered: the English performance gain for the set of samples corresponding to each data point and/or the machine translation performance gain for the samples in the set. In all cases, we consider: the zero-shot performance (since the gain is usually larger when the zero-shot performance is lower), and the number of samples used for that data point. We also investigate whether adding **language features** can improve the prediction. We consider syntax, phonology, inventory, family and geographical location as features similar to the analysis by Lauscher et al. (2020). lang2vec [4] from Littell et al. (2017) is used to obtain the feature vectors for each language. The cosine similarity between the English vectors and the vectors for each language are added as 5 new scalar features (values are in Appendix Table 8). Those features can help the model better use the English performance depending on the similarity between the language and English.

The **prediction error** is reported in Table 6. Having a combination of English and MT performance with language features achieves the best results. We can also see in Table 5 that using the models further improves the chosen-shots performance gain for XNLI with the best result, as before, using a combination of all features. This is not the case for PAWS-X where the improvement in performance seems to be a general improvement and not specific to the different sets of samples. This could also be partially due to having a smaller performance data and fewer languages to train on (7 as compared to 15 languages for XNLI). The detailed results for the different languages are in the Appendix Figure 11. Choosing the shots improves the few-shot performance on XNLI for all languages across almost all sample sizes. For PAWS-X, there is mixed gain/loss but the improvement when using English performance at maximum size is concentrated in the European languages.

6 Conclusion and Future Work

In this work, we conducted a systematic comparison between translate-train and few-shot cross-lingual transfer. We quantified the performance gain for each and showed that starting from 1k samples, machine-translated data could be used to improve over zero-shot performance, and that at 10k-100k, there's an advantage for translate-train over few-shot. For the tasks that benefit from few-shot, we show that there are random sets of samples that perform better across languages and that the English performance of the samples in those sets can help us identify them. The performance of the machine translation of the samples can also be used as another indicator.

[4] https://github.com/antonisa/lang2vec

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. 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.

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, 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.

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask 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.

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.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. 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.

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.

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.

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 the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

A Training Details

Software: We use the Huggingface Transformers [5] for fine-tuning the pretrained language models. We use scikit-learn [6] to train the performance prediction models. Our code will be made publicly available [7]. **Hardware:** NVIDIA GeForce GTX 1080 Ti with 11G memory is used for most experiments. The linear model is trained locally on a CPU.
Model: XLM-R$_{base}$ has ∼270M parameters with 12-layers, 768-hidden-state, 3072 feed-forward hidden-state, 8-heads, and trained on on 2.5 TB of newly created clean CommonCrawl data in 100 languages[8].
Hyperparameters: The used learning rate along with the dev performance for a model with seed=42 is reported in Table 7. We use four other models fine-tuned on the English train split with $seed \in \{2, 4, 8, 16\}$

XNLI	PAWS-X	XQuAD
1e-5	7e-6	3e-5
84.82	92.45	89.10
Accuracy	Accuracy	F1

Table 7: learning rate and English dev performance

Training & Evaluation Runs: Starting from each of the 5 source fine-tuned models, we fine-tune on the target language for 5 different sets of samples. This is repeated for for each size resulting in 25 runs per size. The runtime for the target language fine-tuning varies based on the number of samples used and the number of languages in each dataset. For smaller sample sizes, most runtime is spent for the evaluation on the large test set.

B Languages

code	language properties$^{(1)}$				cosine similarity to English$^{(2)}$					XNLI	PAWS-X	XQuAD
	name	size$^{(3)}$	script	language family	syntax	phonology	inventory	family	geo			
ar	Arabic	1.02	Arabic	Afro-Asiatic	0.65	0.70	0.71	0.00	0.97	x		x
vi	Vietnamese	1.24	Latin	Austro-Asiatic	0.66	0.78	0.75	0.00	0.85	x		x
de	German	2.37	Latin	IE: Germanic	0.90	0.81	0.76	0.54	1.00	x	x	x
en	English	5.98	Latin	IE: Germanic	1.00	1.00	1.00	1.00	1.00	x	x	x
el	Greek	0.17	Greek	IE: Greek	0.78	0.95	0.65	0.15	0.99	x		x
hi	Hindi	0.13	Devanagari	IE: Indo-Aryan	0.62	0.78	0.71	0.13	0.91	x		x
ur	Urdu	0.15	Perso-Arabic	IE: Indo-Aryan	0.62	0.86	0.72	0.13	0.93	x		
es	Spanish	1.56	Latin	IE: Romance	0.82	0.86	0.64	0.10	1.00	x	x	x
ro	Romanian	0.42	0.42	IE: Romance	0.80	0.90	0.73	0.12	0.99			x
fr	French	2.16	Latin	IE: Romance	0.81	0.75	0.74	0.10	1.00	x	x	
ru	Russian	1.58	Cyrillic	IE: Slavic	0.81	0.86	0.65	0.17	0.96	x		x
bg	Bulgarian	0.26	Cyrillic	IE: Slavic	0.86	0.86	0.68	0.14	0.99	x		
ja	Japanese	1.18	Ideograms	Japonic	0.50	0.67	0.65	0.00	0.86		x	
ko	Korean	0.47	Hangul	Koreanic	0.55	0.75	0.71	0.00	0.87		x	
th	Thai	0.13	Brahmic	Kra-Dai	0.64	0.78	0.75	0.00	0.85	x		x
sw	Swahili	0.05	Latin	Niger-Congo	0.46	0.91	0.76	0.00	0.92	x		
zh	Mandarin	1.09	Chinese ideograms	Sino-Tibetan	0.71	0.73	0.70	0.00	0.88	x	x	x
tr	Turkish	0.34	Latin	Turkic	0.51	0.82	0.67	0.00	0.98	x		x

(1) properties taken from XTREME
(2) similarity calculated using lang2vec
(3) size is the #wikipedia articles in millions

Table 8: Languages in the Datasets

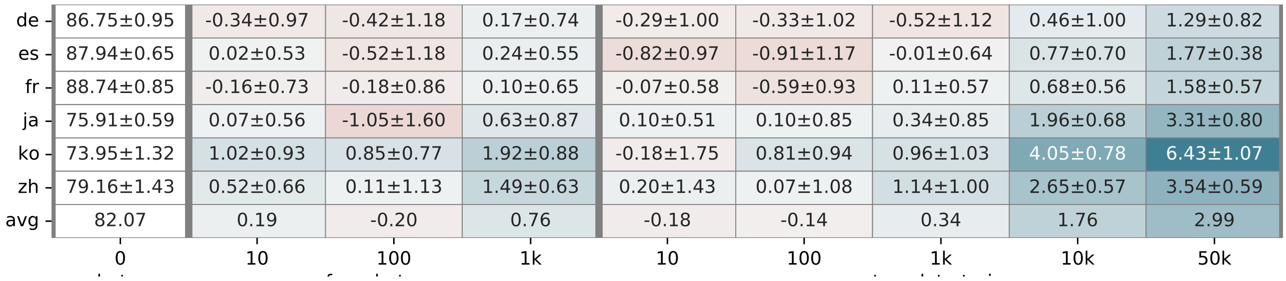
[5] https://github.com/huggingface/transformers

[6] https://github.com/scikit-learn/scikit-learn

[7] https://www.gihtub.com/***

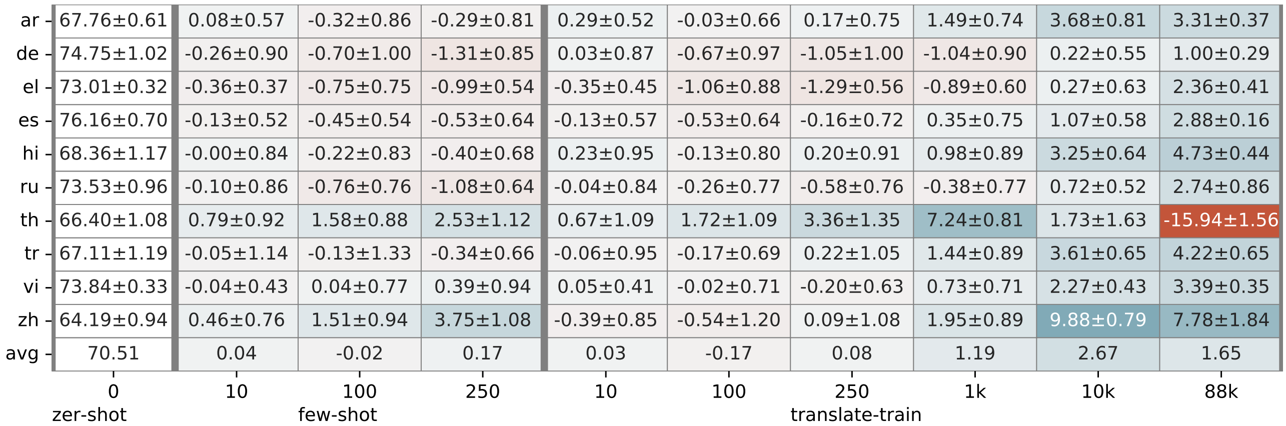
[8] from https://huggingface.co/transformers/pretrained_models.html

C More Results Details



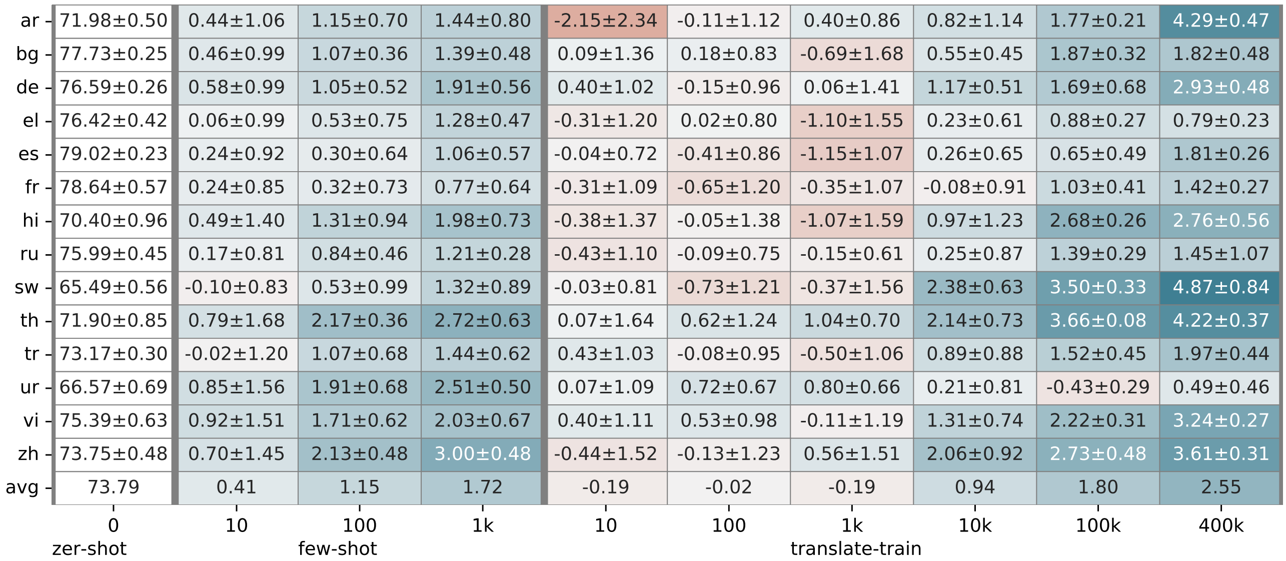
	0	10	100	1k	10	100	1k	10k	50k
de	86.75±0.95	-0.34±0.97	-0.42±1.18	0.17±0.74	-0.29±1.00	-0.33±1.02	-0.52±1.12	0.46±1.00	1.29±0.82
es	87.94±0.65	0.02±0.53	-0.52±1.18	0.24±0.55	-0.82±0.97	-0.91±1.17	-0.01±0.64	0.77±0.70	1.77±0.38
fr	88.74±0.85	-0.16±0.73	-0.18±0.86	0.10±0.65	-0.07±0.58	-0.59±0.93	0.11±0.57	0.68±0.56	1.58±0.57
ja	75.91±0.59	0.07±0.56	-1.05±1.60	0.63±0.87	0.10±0.51	0.10±0.85	0.34±0.85	1.96±0.68	3.31±0.80
ko	73.95±1.32	1.02±0.93	0.85±0.77	1.92±0.88	-0.18±1.75	0.81±0.94	0.96±1.03	4.05±0.78	6.43±1.07
zh	79.16±1.43	0.52±0.66	0.11±1.13	1.49±0.63	0.20±1.43	0.07±1.08	1.14±1.00	2.65±0.57	3.54±0.59
avg	82.07	0.19	-0.20	0.76	-0.18	-0.14	0.34	1.76	2.99

Figure 5: **Detailed Results on PAWS-X**. Gains in performance over zero-shot for few-shot and translate-train. Non-European language show the most gain especially Korean.



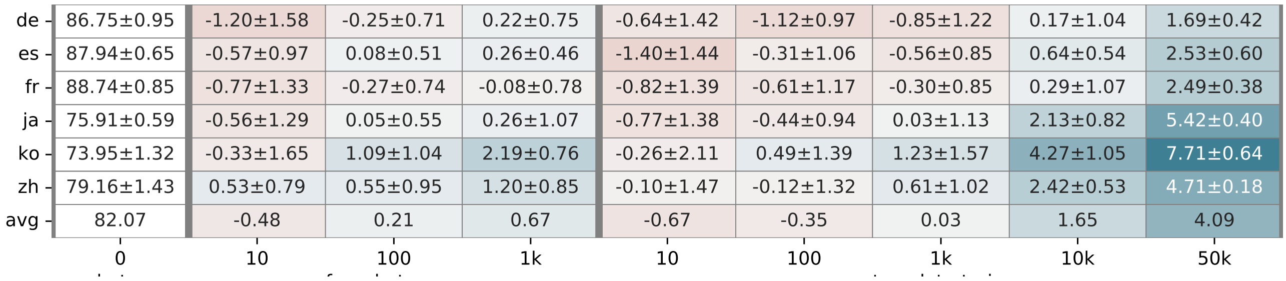
	0	10	100	250	10	100	250	1k	10k	88k
	zer-shot		few-shot				translate-train			
ar	67.76±0.61	0.08±0.57	-0.32±0.86	-0.29±0.81	0.29±0.52	-0.03±0.66	0.17±0.75	1.49±0.74	3.68±0.81	3.31±0.37
de	74.75±1.02	-0.26±0.90	-0.70±1.00	-1.31±0.85	0.03±0.87	-0.67±0.97	-1.05±1.00	-1.04±0.90	0.22±0.55	1.00±0.29
el	73.01±0.32	-0.36±0.37	-0.75±0.75	-0.99±0.54	-0.35±0.45	-1.06±0.88	-1.29±0.56	-0.89±0.60	0.27±0.63	2.36±0.41
es	76.16±0.70	-0.13±0.52	-0.45±0.54	-0.53±0.64	-0.13±0.57	-0.53±0.64	-0.16±0.72	0.35±0.75	1.07±0.58	2.88±0.16
hi	68.36±1.17	-0.00±0.84	-0.22±0.83	-0.40±0.68	0.23±0.95	-0.13±0.80	0.20±0.91	0.98±0.89	3.25±0.64	4.73±0.44
ru	73.53±0.96	-0.10±0.86	-0.76±0.76	-1.08±0.64	-0.04±0.84	-0.26±0.77	-0.58±0.76	-0.38±0.77	0.72±0.52	2.74±0.86
th	66.40±1.08	0.79±0.92	1.58±0.88	2.53±1.12	0.67±1.09	1.72±1.09	3.36±1.35	7.24±0.81	1.73±1.63	-15.94±1.56
tr	67.11±1.19	-0.05±1.14	-0.13±1.33	-0.34±0.66	-0.06±0.95	-0.17±0.69	0.22±1.05	1.44±0.89	3.61±0.65	4.22±0.65
vi	73.84±0.33	-0.04±0.43	0.04±0.77	0.39±0.94	0.05±0.41	-0.02±0.71	-0.20±0.63	0.73±0.71	2.27±0.43	3.39±0.35
zh	64.19±0.94	0.46±0.76	1.51±0.94	3.75±1.08	-0.39±0.85	-0.54±1.20	0.09±1.08	1.95±0.89	9.88±0.79	7.78±1.84
avg	70.51	0.04	-0.02	0.17	0.03	-0.17	0.08	1.19	2.67	1.65

Figure 6: **Detailed Results on XQuAD**. Gains in performance over zero-shot for few-shot and translate-train. Non-European languages show the most gain especially Chinese. Thai shows a significant degrade when using the full machine-translated dataset. This might be due to lower-quality machine translation for Thai.



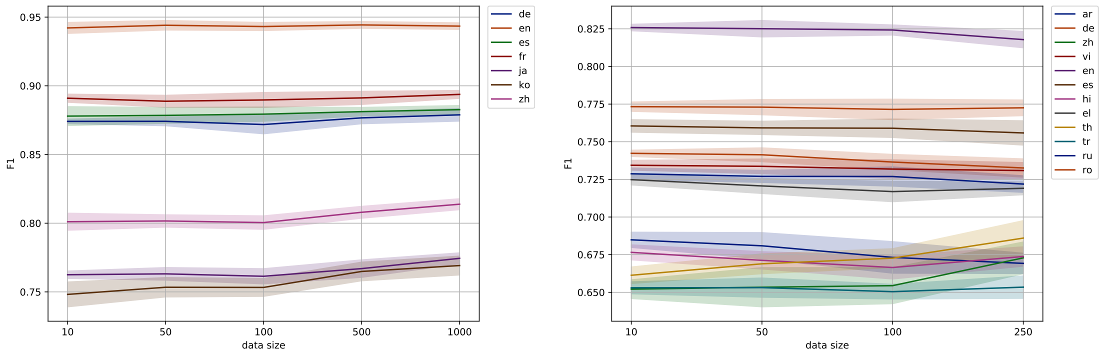
	0	10	100	1k	10	100	1k	10k	100k	400k
	zer-shot		few-shot				translate-train			
ar	71.98±0.50	0.44±1.06	1.15±0.70	1.44±0.80	-2.15±2.34	-0.11±1.12	0.40±0.86	0.82±1.14	1.77±0.21	4.29±0.47
bg	77.73±0.25	0.46±0.99	1.07±0.36	1.39±0.48	0.09±1.36	0.18±0.83	-0.69±1.68	0.55±0.45	1.87±0.32	1.82±0.48
de	76.59±0.26	0.58±0.99	1.05±0.52	1.91±0.56	0.40±1.02	-0.15±0.96	0.06±1.41	1.17±0.51	1.69±0.68	2.93±0.48
el	76.42±0.42	0.06±0.99	0.53±0.75	1.28±0.47	-0.31±1.20	0.02±0.80	-1.10±1.55	0.23±0.61	0.88±0.27	0.79±0.23
es	79.02±0.23	0.24±0.92	0.30±0.64	1.06±0.57	-0.04±0.72	-0.41±0.86	-1.15±1.07	0.26±0.65	0.65±0.49	1.81±0.26
fr	78.64±0.57	0.24±0.85	0.32±0.73	0.77±0.64	-0.31±1.09	-0.65±1.20	-0.35±1.07	-0.08±0.91	1.03±0.41	1.42±0.27
hi	70.40±0.96	0.49±1.40	1.31±0.94	1.98±0.73	-0.38±1.37	-0.05±1.38	-1.07±1.59	0.97±1.23	2.68±0.26	2.76±0.56
ru	75.99±0.45	0.17±0.81	0.84±0.46	1.21±0.28	-0.43±1.10	-0.09±0.75	-0.15±0.61	0.25±0.87	1.39±0.29	1.45±1.07
sw	65.49±0.56	-0.10±0.83	0.53±0.99	1.32±0.89	-0.03±0.81	-0.73±1.21	-0.37±1.56	2.38±0.63	3.50±0.33	4.87±0.84
th	71.90±0.85	0.79±1.68	2.17±0.36	2.72±0.63	0.07±1.64	0.62±1.24	1.04±0.70	2.14±0.73	3.66±0.08	4.22±0.37
tr	73.17±0.30	-0.02±1.20	1.07±0.68	1.44±0.62	0.43±1.03	-0.08±0.95	-0.50±1.06	0.89±0.88	1.52±0.45	1.97±0.44
ur	66.57±0.69	0.85±1.56	1.91±0.68	2.51±0.50	0.07±1.09	0.72±0.67	0.80±0.66	0.21±0.81	-0.43±0.29	0.49±0.46
vi	75.39±0.63	0.92±1.51	1.71±0.62	2.03±0.67	0.40±1.11	0.53±0.98	-0.11±1.19	1.31±0.74	2.22±0.31	3.24±0.27
zh	73.75±0.48	0.70±1.45	2.13±0.48	3.00±0.48	-0.44±1.52	-0.13±1.23	0.56±1.51	2.06±0.92	2.73±0.48	3.61±0.31
avg	73.79	0.41	1.15	1.72	-0.19	-0.02	-0.19	0.94	1.80	2.55

Figure 7: Detailed Results on **XNLI using a part of the available data as dev**. The few-shot performance only changes slightly with minor increases and decreases for across the languages. The highest increase on average is at 10 samples with an increase of 0.05%. Translate-train performance decreases for almost all languages and on average.



	0	10	100	1k	10	100	1k	10k	50k
de	86.75±0.95	-1.20±1.58	-0.25±0.71	0.22±0.75	-0.64±1.42	-1.12±0.97	-0.85±1.22	0.17±1.04	1.69±0.42
es	87.94±0.65	-0.57±0.97	0.08±0.51	0.26±0.46	-1.40±1.44	-0.31±1.06	-0.56±0.85	0.64±0.54	2.53±0.60
fr	88.74±0.85	-0.77±1.33	-0.27±0.74	-0.08±0.78	-0.82±1.39	-0.61±1.17	-0.30±0.85	0.29±1.07	2.49±0.38
ja	75.91±0.59	-0.56±1.29	0.05±0.55	0.26±1.07	-0.77±1.38	-0.44±0.94	0.03±1.13	2.13±0.82	5.42±0.40
ko	73.95±1.32	-0.33±1.65	1.09±1.04	2.19±0.76	-0.26±2.11	0.49±1.39	1.23±1.57	4.27±1.05	7.71±0.64
zh	79.16±1.43	0.53±0.79	0.55±0.95	1.20±0.85	-0.10±1.47	-0.12±1.32	0.61±1.02	2.42±0.53	4.71±0.18
avg	82.07	-0.48	0.21	0.67	-0.67	-0.35	0.03	1.65	4.09

Figure 8: Detailed Results on **PAWS-X using a part of the available data as dev**. The few-shot performance shows mixed gains decreasing by ∼0.60% for 10 samples, increasing by ∼0.40% at 100 then decreasing againg by ∼0.10%. Translate-train performance decreases util the full dataset is used where it increases by ∼1%.



(a) **PAWS-X Performance variance on different shots.** Variance decreases with an increased data size

(b) **XQuAD Performance variance on different shots.** Variance increases with an increased data size

Figure 9: Performance variance on different shots

lang	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
10	0.64	0.42	0.42	0.51	0.63	0.75	0.49	-0.25	0.63	0.48	0.70	0.19	0.48	0.42	0.50
50	0.82	0.59	0.59	0.76	0.74	0.87	0.66	0.11	0.70	0.81	0.84	0.54	0.73	0.56	0.69
100	0.76	0.53	0.47	0.46	0.64	0.77	0.69	-0.50	0.58	0.58	0.64	0.06	0.73	0.33	0.52
500	0.84	0.64	0.77	0.79	0.73	0.81	0.77	0.18	0.67	0.71	0.88	0.54	0.77	0.73	0.72
1000	0.72	0.63	0.74	0.69	0.72	0.84	0.60	0.10	0.06	0.51	0.80	0.03	0.51	0.75	0.58
all	0.77	0.59	0.62	0.69	0.73	0.79	0.66	0.15	0.62	0.57	0.79	0.38	0.65	0.55	0.64

Table 9: **XNLI Pearson correlation** between the performance on **English** and the performance on other languages using the same set of samples.

	de	en	es	fr	ja	ko	zh
de	1.00	0.66	0.52	0.56	0.21	0.54	0.64
en	0.66	1.00	0.56	0.41	0.11	0.37	0.36
es	0.52	0.56	1.00	0.57	0.22	0.54	0.57
fr	0.56	0.41	0.57	1.00	0.03	0.59	0.55
ja	0.21	0.11	0.22	0.03	1.00	0.16	0.32
ko	0.54	0.37	0.54	0.59	0.16	1.00	0.54
zh	0.64	0.36	0.57	0.55	0.32	0.54	1.00
avg	0.59	0.50	0.57	0.53	0.29	0.54	0.57

Table 10: **PAWS-X Pearson correlation** of the performance between languages.

lang	de	es	fr	ja	ko	zh	avg
10	0.47	0.65	0.34	-0.22	0.53	0.56	0.48
50	0.81	0.56	0.57	-0.35	0.53	0.48	0.51
100	0.78	0.53	0.42	0.40	0.47	0.44	0.57
500	0.52	0.55	0.53	0.16	0.41	0.11	0.47
1000	0.75	0.77	0.30	-0.01	-0.02	0.35	0.45
all	0.66	0.56	0.41	0.11	0.37	0.36	0.50

Table 11: **PAWS-X Pearson correlation** between the performance on **English** and the performance on other languages using the same set of samples.

	de	es	fr	ja	ko	zh
	0.66	0.62	0.68	0.45	0.38	0.52

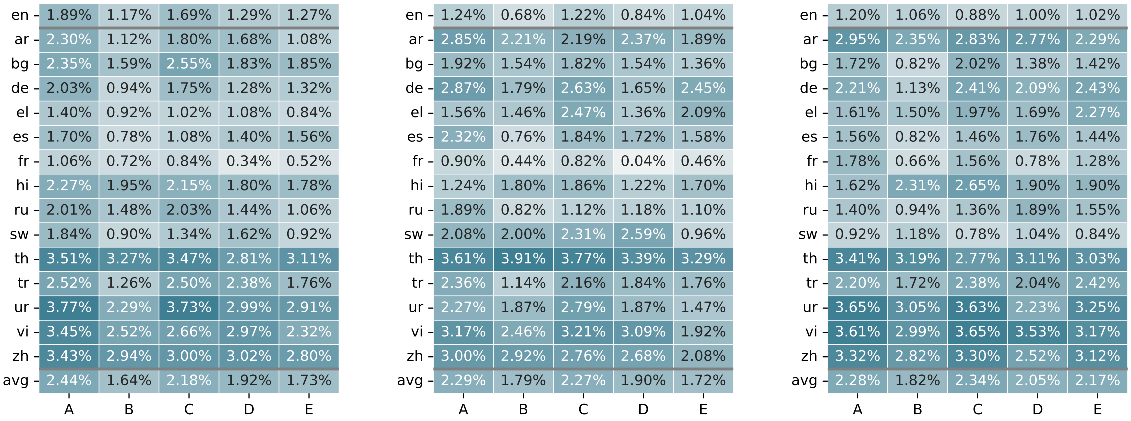
Table 12: **PAWS-X Pearson correlation** between the performance of **machine translation** and manual translation.

	ar	de	zh	vi	en	es	hi	el	th	tr	ru	ro
ar	1.00	-0.14	0.03	0.07	0.12	-0.02	0.01	-0.03	0.07	0.25	0.12	-0.06
de	-0.14	1.00	-0.54	-0.18	0.35	0.57	0.42	0.22	-0.26	0.40	-0.09	-0.00
zh	0.03	-0.54	1.00	0.16	-0.37	-0.38	-0.21	-0.41	0.55	-0.17	-0.24	-0.22
vi	0.07	-0.18	0.16	1.00	-0.08	-0.02	-0.08	-0.01	0.02	-0.18	-0.12	-0.26
en	0.12	0.35	-0.37	-0.08	1.00	0.46	0.08	0.07	-0.17	0.06	-0.04	-0.06
es	-0.02	0.57	-0.38	-0.02	0.46	1.00	0.10	0.02	-0.31	0.09	-0.29	-0.24
hi	0.01	0.42	-0.21	-0.08	0.08	0.10	1.00	0.18	0.06	0.37	0.27	0.18
el	-0.03	0.22	-0.41	-0.01	0.07	0.02	0.18	1.00	-0.15	0.01	0.34	0.13
th	0.07	-0.26	0.55	0.02	-0.17	-0.31	0.06	-0.15	1.00	0.17	0.07	0.10
tr	0.25	0.40	-0.17	-0.18	0.06	0.09	0.37	0.01	0.17	1.00	0.33	0.27
ru	0.12	-0.09	-0.24	-0.12	-0.04	-0.29	0.27	0.34	0.07	0.33	1.00	0.56
ro	-0.06	-0.00	-0.22	-0.26	-0.06	-0.24	0.18	0.13	0.10	0.27	0.56	1.00
avg	0.12	0.15	-0.07	0.03	0.12	0.08	0.20	0.11	0.10	0.22	0.16	0.12

Table 13: **XQuAD Pearson correlation** of the performance between languages.

lang	ar	de	zh	vi	es	hi	el	th	tr	ru	ro	avg
10	0.54	0.41	0.08	-0.40	0.30	0.05	-0.10	0.43	0.44	-0.25	-0.23	0.19
50	0.37	0.24	-0.28	0.11	-0.01	0.19	0.27	0.12	0.21	-0.04	-0.08	0.18
100	-0.37	0.35	-0.54	-0.03	0.71	0.02	0.08	-0.09	-0.08	-0.40	-0.12	0.05
250	0.08	0.20	-0.25	0.03	0.65	-0.16	-0.38	-0.31	-0.45	-0.33	-0.34	-0.02
all	0.12	0.35	-0.37	-0.08	0.46	0.08	0.07	-0.17	0.06	-0.04	-0.06	0.12

Table 14: **XQuAD Pearson correlation** between the performance on **English** and the performance on other languages using the same set of samples.

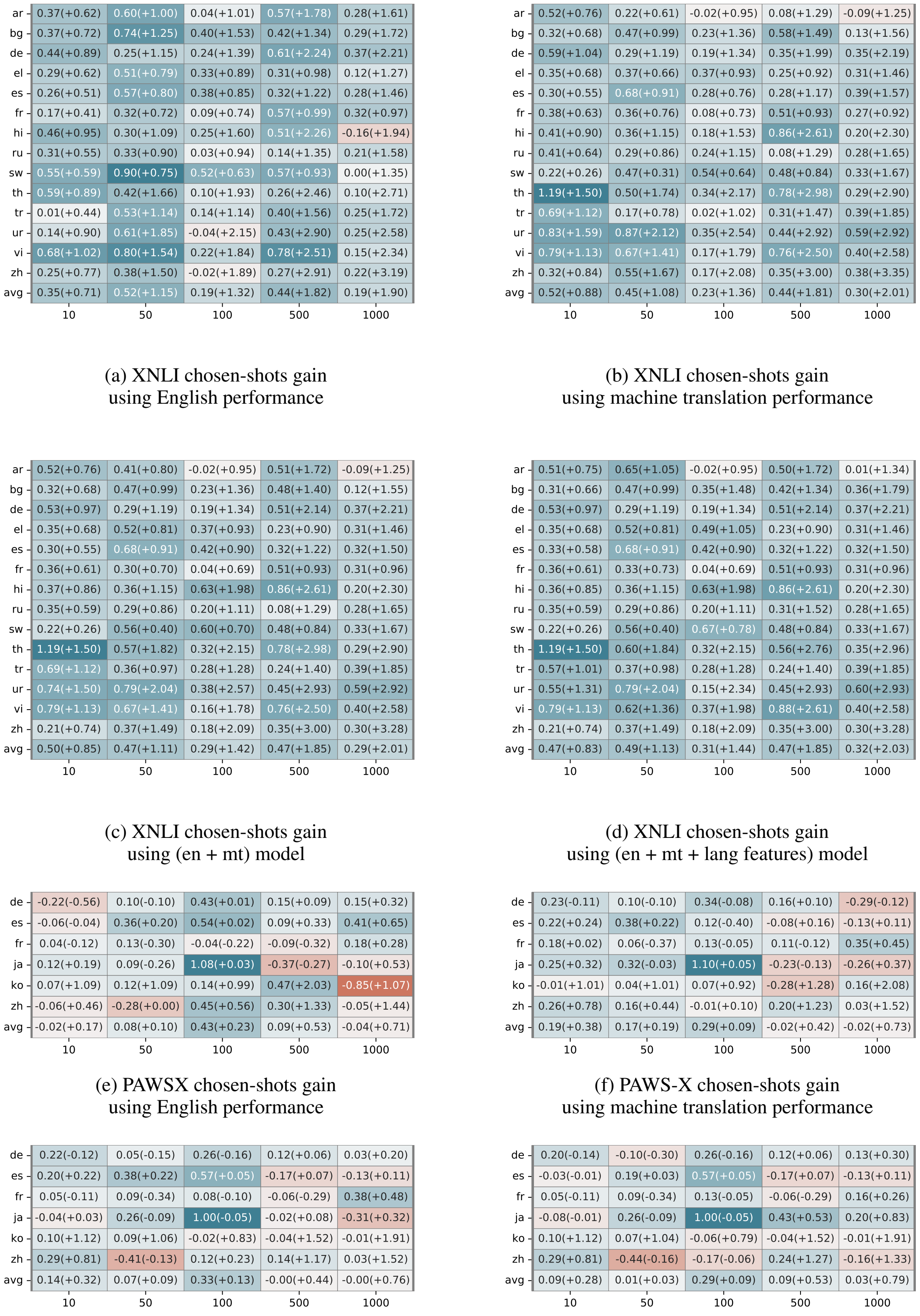


	A	B	C	D	E
en	1.89%	1.17%	1.69%	1.29%	1.27%
ar	2.30%	1.12%	1.80%	1.68%	1.08%
bg	2.35%	1.59%	2.55%	1.83%	1.85%
de	2.03%	0.94%	1.75%	1.28%	1.32%
el	1.40%	0.92%	1.02%	1.08%	0.84%
es	1.70%	0.78%	1.08%	1.40%	1.56%
fr	1.06%	0.72%	0.84%	0.34%	0.52%
hi	2.27%	1.95%	2.15%	1.80%	1.78%
ru	2.01%	1.48%	2.03%	1.44%	1.06%
sw	1.84%	0.90%	1.34%	1.62%	0.92%
th	3.51%	3.27%	3.47%	2.81%	3.11%
tr	2.52%	1.26%	2.50%	2.38%	1.76%
ur	3.77%	2.29%	3.73%	2.99%	2.91%
vi	3.45%	2.52%	2.66%	2.97%	2.32%
zh	3.43%	2.94%	3.00%	3.02%	2.80%
avg	2.44%	1.64%	2.18%	1.92%	1.73%

	A	B	C	D	E
en	1.24%	0.68%	1.22%	0.84%	1.04%
ar	2.85%	2.21%	2.19%	2.37%	1.89%
bg	1.92%	1.54%	1.82%	1.54%	1.36%
de	2.87%	1.79%	2.63%	1.65%	2.45%
el	1.56%	1.46%	2.47%	1.36%	2.09%
es	2.32%	0.76%	1.84%	1.72%	1.58%
fr	0.90%	0.44%	0.82%	0.04%	0.46%
hi	1.24%	1.80%	1.86%	1.22%	1.70%
ru	1.89%	0.82%	1.12%	1.18%	1.10%
sw	2.08%	2.00%	2.31%	2.59%	0.96%
th	3.61%	3.91%	3.77%	3.39%	3.29%
tr	2.36%	1.14%	2.16%	1.84%	1.76%
ur	2.27%	1.87%	2.79%	1.87%	1.47%
vi	3.17%	2.46%	3.21%	3.09%	1.92%
zh	3.00%	2.92%	2.76%	2.68%	2.08%
avg	2.29%	1.79%	2.27%	1.90%	1.72%

	A	B	C	D	E
en	1.20%	1.06%	0.88%	1.00%	1.02%
ar	2.95%	2.35%	2.83%	2.77%	2.29%
bg	1.72%	0.82%	2.02%	1.38%	1.42%
de	2.21%	1.13%	2.41%	2.09%	2.43%
el	1.61%	1.50%	1.97%	1.69%	2.27%
es	1.56%	0.82%	1.46%	1.76%	1.44%
fr	1.78%	0.66%	1.56%	0.78%	1.28%
hi	1.62%	2.31%	2.65%	1.90%	1.90%
ru	1.40%	0.94%	1.36%	1.89%	1.55%
sw	0.92%	1.18%	0.78%	1.04%	0.84%
th	3.41%	3.19%	2.77%	3.11%	3.03%
tr	2.20%	1.72%	2.38%	2.04%	2.42%
ur	3.65%	3.05%	3.63%	2.23%	3.25%
vi	3.61%	2.99%	3.65%	3.53%	3.17%
zh	3.32%	2.82%	3.30%	2.52%	3.12%
avg	2.28%	1.82%	2.34%	2.05%	2.17%

Figure 10: **XNLI few-shot gain** over zero-shot across 5 sets of samples (**size=1000**) for 3 different model initalizations. Sets A and C yield better performance for the 3 different initalizations. The English performance can be used as an indicator.



(a) XNLI chosen-shots gain
using English performance

(b) XNLI chosen-shots gain
using machine translation performance

(c) XNLI chosen-shots gain
using (en + mt) model

(d) XNLI chosen-shots gain
using (en + mt + lang features) model

(e) PAWSX chosen-shots gain
using English performance

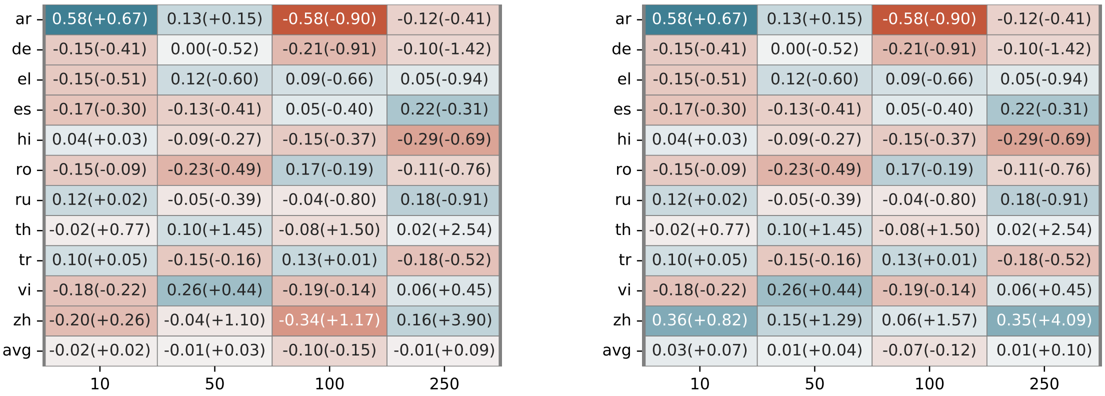
(f) PAWS-X chosen-shots gain
using machine translation performance

(g) PAWS-X chosen-shots gain
using (en + mt) model

de 0.20(-0.14) -0.10(-0.30) 0.26(-0.16) 0.12(+0.06) 0.13(+0.30)
es -0.03(-0.01) 0.19(+0.03) 0.57(+0.05) -0.17(+0.07) -0.13(+0.11)
fr 0.05(-0.11) 0.09(-0.34) 0.13(-0.05) -0.06(-0.29) 0.16(+0.26)
ja -0.08(-0.01) 0.26(-0.09) 1.00(-0.05) 0.43(+0.53) 0.20(+0.83)
ko 0.10(+1.12) 0.07(+1.04) -0.06(+0.79) -0.04(+1.52) -0.01(+1.91)
zh 0.29(+0.81) -0.44(-0.16) -0.17(-0.06) 0.24(+1.27) -0.16(+1.33)
avg 0.09(+0.28) 0.01(+0.03) 0.29(+0.09) 0.09(+0.53) 0.03(+0.79)
10 50 100 500 1000

(h) PAWS-X chosen-shots gain
using (en + mt + lang features) model

Figure 11: **Chosen-shots gain in performance**. The gain of choosing shots over the average of no-choosing (average over 5 random sets). The actual few-shot gain (compared to zero-shot) is shown in parenthesis as follows: *chosen-shots-gain (few-shot-gain)*. When *chosen-shots-gain* is positive (green), choosing the shots results in more gain. When negative (red), it hurts and results in less gain.



	10	50	100	250
ar	0.58(+0.67)	0.13(+0.15)	-0.58(-0.90)	-0.12(-0.41)
de	-0.15(-0.41)	0.00(-0.52)	-0.21(-0.91)	-0.10(-1.42)
el	-0.15(-0.51)	0.12(-0.60)	0.09(-0.66)	0.05(-0.94)
es	-0.17(-0.30)	-0.13(-0.41)	0.05(-0.40)	0.22(-0.31)
hi	0.04(+0.03)	-0.09(-0.27)	-0.15(-0.37)	-0.29(-0.69)
ro	-0.15(-0.09)	-0.23(-0.49)	0.17(-0.19)	-0.11(-0.76)
ru	0.12(+0.02)	-0.05(-0.39)	-0.04(-0.80)	0.18(-0.91)
th	-0.02(+0.77)	0.10(+1.45)	-0.08(+1.50)	0.02(+2.54)
tr	0.10(+0.05)	-0.15(-0.16)	0.13(+0.01)	-0.18(-0.52)
vi	-0.18(-0.22)	0.26(+0.44)	-0.19(-0.14)	0.06(+0.45)
zh	-0.20(+0.26)	-0.04(+1.10)	-0.34(+1.17)	0.16(+3.90)
avg	-0.02(+0.02)	-0.01(+0.03)	-0.10(-0.15)	-0.01(+0.09)

(a) XQuAD chosen-shots gain using English performance

	10	50	100	250
ar	0.58(+0.67)	0.13(+0.15)	-0.58(-0.90)	-0.12(-0.41)
de	-0.15(-0.41)	0.00(-0.52)	-0.21(-0.91)	-0.10(-1.42)
el	-0.15(-0.51)	0.12(-0.60)	0.09(-0.66)	0.05(-0.94)
es	-0.17(-0.30)	-0.13(-0.41)	0.05(-0.40)	0.22(-0.31)
hi	0.04(+0.03)	-0.09(-0.27)	-0.15(-0.37)	-0.29(-0.69)
ro	-0.15(-0.09)	-0.23(-0.49)	0.17(-0.19)	-0.11(-0.76)
ru	0.12(+0.02)	-0.05(-0.39)	-0.04(-0.80)	0.18(-0.91)
th	-0.02(+0.77)	0.10(+1.45)	-0.08(+1.50)	0.02(+2.54)
tr	0.10(+0.05)	-0.15(-0.16)	0.13(+0.01)	-0.18(-0.52)
vi	-0.18(-0.22)	0.26(+0.44)	-0.19(-0.14)	0.06(+0.45)
zh	0.36(+0.82)	0.15(+1.29)	0.06(+1.57)	0.35(+4.09)
avg	0.03(+0.07)	0.01(+0.04)	-0.07(-0.12)	0.01(+0.10)

(b) XQuAD chosen-shots gain using en performance model

Figure 12: **XQuAD chosen-shots gain in performance** (no gain!). The gain of choosing shots over the average of no-choosing (average over 5 random sets). The actual few-shot gain (compared to zero-shot) is shown in parenthesis as follows chosen-shots-gain (few-shot-gain). We can see that there is no gain in choosing the shots. Experiments with adding language features to the model further decrease the performance.