Adaptive Cross-lingual Text Classification through In-Context One-Shot Demonstrations

Anonymous ACL submission

Abstract

Zero-shot cross-lingual Transfer (ZS-XLT) utilizes a model trained in a source language to make predictions in a target language. However, this method often yields performance loss in the target language. To alleviate this loss, additional improvements can be achieved through subsequent fine-tuning using target demonstrations. In this paper, we exploit In-Context Tuning (ICT) for One-Shot cross-lingual transfer in the classification task by introducing In-Context Cross-lingual transfer (IC-XLT). The novel concept involves training a model to learn from context examples and subsequently adapting it at inference to a target language using One-Shot context demonstrations target language. Remarkably, this adaptation process requires no fine-tuning for reducing the performance gap with the source language. Our results show that IC-XLT successfully leverages these demonstrations to improve the cross-lingual capabilities of the evaluated mT5 model, outperforming prompt-based fine-tuned models in the Zero and One-shot scenarios. Moreover, we show that when source language data is limited, the fine-tuning framework employed for IC-XLT performs comparably to Prompt-based fine-tuning with significantly more training data in the source language. Hence, we also present a compelling alternative for One-Shot cross-lingual transfer in scenarios where computational resources or source-language data is constrained.

1 Introduction

The recent progress in the development of multilingual Language Models (LMs) has allowed for effective cross-lingual transfer (XLT) with minimal need for architectural modifications (Pires et al., 2019; Xue et al., 2020). By simply training a multilingual model in a language with abundant resources its acquired knowledge can be extended to target languages, in either zero-shot or few-shot scenarios. XLT is a significant topic as it addresses the prevalent challenge of data scarcity in languages other than widely resourced ones, such as English (Joshi et al., 2020). The ability to leverage the extensive linguistic resources available in high-resource languages to languages with limited training data enables the deployment of truly inclusive NLP systems.

Zero-shot Cross-lingual Transfer (ZS-XLT) involves transferring a model trained in a source language to a target language without any demonstration of target-language examples (Chen et al., 2021; Pires et al., 2019). This approach is highly modular, as it requires no adaptations specific to the target language. On the other hand, Few-shot Cross-lingual Transfer (FS-XLT) enhances target-language accuracy by further fine-tuning a model using labeled target data (Lauscher et al., 2020; Zhao et al., 2021). However, this improvement comes at the expense of additional computational resources and reduced modularity compared to the zero-shot approach.

Our perspective is that adapting to a target language should prioritize resource efficiency and modularity, where we can seamlessly deploy a single model trained in English (or another source language) across different languages without any fine-tuning. In this work, we aim to improve this aspect for the text classification task by eliciting a multilingual model's language-specific abilities by prepending One-Shot text-label target language demonstrations to the input text to predict the correct label. Specifically, we propose In-Context Cross-lingual transfer (IC-XLT), a simple yet effective method for One-Shot Cross-Lingual Transfer in Text Classification.

This novel approach employs In-Context Tuning (ICT) (Chen et al., 2022) to train an encoder-decoder model in the source language tasking it to predict input texts with information derived from context demonstrations. ICT is a meta-learning strategy that optimizes a model's ability to learn from in-context examples, originally designed for facilitating swift adaptation to new tasks by prepending target-task in-context demonstrations to the input during the adaptation process. To the best of our knowledge, the first study of ICT application in the context of cross-lingual transfer.

The proposed method is composed of a fine-tuning and an adaptation stage. Firstly, we fine-tune on the source language through ICT, where the model is trained for the classification task and also to learn from context demonstrations. Secondly, we adapt to the target language at inference by prepending One-Shot[1] demonstrations. Compared to other gradient-based FS-XLT techniques, this method is modular and cost-effective at the adaptation stage.

We evaluate IC-XLT on two multilingual text classification datasets, spanning five target languages, with English as the source language. We consider two distinct settings. First, we assume access to the entire source language training dataset. For the second setting, we deliberately constrain the amount of source training data available. This limitation aims to gauge the robustness of the proposed approach in scenarios where the availability of source data is restricted. We hypothesize that leveraging context information may prove particularly beneficial in tasks where source data is limited.

The contributions of this work are the following:

1. **IC-XLT as an effective strategy for One-Shot Cross-lingual transfer:** By comparing the reduction in the transfer gap of One-Shot IC-XLT against ZS-XLT –a standard cross-lingual approach– we present empirical evidence that training a model in a source language with In-Context Tuning allows it to leverage One-Shot demostrations through In-Context Learning to adapt to a target language. This results in a One-Shot XLT approach that requires no gradient update for language adaptation and can transfer at inference without modifying the model weights.

2. **ICT improves mT5 finetuning, especially when resources are limited.** We observe that for the evaluated tasks, ICT training yields better performance compared to traditional fine-tuning, especially when (source language) training data consists on few-shots per label. In particular IC-XLT models trained on this scenario (1) benefit from this behavior at the adaptation and (2) leverage target language in-context examples, achieving comparable performance to Prompt Tuning transfer methods with significantly less source language data.

[1] One-Shot per label

2 Related work

2.1 Zero and Few-Shot Cross-lingual Transfer

Multilingual transformers, such as mBERT (Devlin et al., 2018), XLMR (Conneau et al., 2019), and mT5 (Xue et al., 2020), have showcased notable ability in zero-shot cross-lingual transfer (ZS-XLT) (Pires et al., 2019). In this paradigm, these models are trained using abundant data in a source language and subsequently undergo evaluation in a target language without exposure to any training data in that specific language. However, this methodology is susceptible to significant performance variance (Keung et al., 2020), and the transfer performance gap is contingent upon the linguistic proximity between the source and target languages (Pires et al., 2019).

Furthermore, recent studies indicate that incorporating a small number of annotated examples in the target language can mitigate the performance gap between the source and target languages (Lauscher et al., 2020; Zhao et al., 2021; Schmidt et al., 2022). This methodology, termed few-shot cross-lingual transfer (FS-XLT), involves first fine-tuning a model on an extensive source dataset (as in ZS-XLT), and then subjecting it to a second fine-tuning on the reduced target language data, facilitating its adaptation to this target language. This approach yields a noticeable improvement in performance at a relatively low labeling cost across various NLP tasks (Lauscher et al., 2020).

Yet, according to (Schmidt et al., 2022), sequential FS-XLT can also exhibit unreliability in the few-shot scenario due to considerable variance in performance at different checkpoints during training. To address this issue, they propose jointly training the model using both source and target data in the adaptation stage of the process, which improves stability in the few-shot setting. This *fine-tuned* FS-XLT approach, however, has two notable drawbacks. Firstly, it lacks modularity, as the models are trained specifically for the selected target language during the adaptation stage. Secondly, there is a substantial increase in computational cost compared to zero-shot cross-lingual transfer due to the adaptation fine-tuning, whose cost scales with the size of the base model.

Moreover, existing methods predominantly address the XLT task under the assumption of abundant data in the source languages. Although this is a fair assumption for many cases, as in general it is much more likely to find labeled datasets in high resource languages, there are scenarios where the source domain itself is limited.

Instances of this include highly domain-specific tasks with a scarcity of annotated samples or tasks related to rapidly emerging trends and language patterns originated from social media, where labeled data may be scarce. In such cases, it might be more feasible to find labelers for high-resource languages, which can then be transferred to other languages.

Given these considerations, we believe it is pertinent to investigate how the XLT performance scales as the quantity of available source data is systematically reduced. The intuition behind this is that the introduction of target-language shots may alleviate the performance decrease associated with a reducing source training data.

2.2 In-Context Learning and Language Models

LMs have demonstrated an aptitude for learning from a small number of demonstrations through a method known as In-Context Learning (ICL) (Brown et al., 2020), where model is tasked with predicting an input prepended with labeled examples. Particularly, (Winata et al., 2021) observed that it is possible to achieve satisfactory performance in a cross-lingual setting when evaluating a mT5 model with a target-language input prefixed with labeled English demonstrations. This zero-shot approach, although efficient, can be suboptimal as it does not take fully advantage of resources in the source language due to the lack of fine-tuning.

Recent findings indicate that transformers (Vaswani et al., 2017) can perform *model selection* on functions encountered during pre-training through in-context demonstrations. Yet, they still find challenging in generalizing effectively to out-of-distribution classes, as highlighted by (Yadlowsky et al., 2023). Given that most pre-trained LMs have not been explicitly trained for ICL, they might exhibit sub-optimal behavior when presented with few-shot demonstrations. In response to this challenge, the authors of (Chen et al., 2022) introduce In-Context Tuning (ICT), a meta-learning[2] approach designed to train a model to effectively learn from in-context demonstrations[3]. ICT meta-trains a language model across a range of tasks, enhancing its ability to swiftly adapt to new tasks through ICL.

Still, In-Context Tuning has not yet been implemented for language transfer, as opposed to task transfer. We hypothesize that training a multilingual model concurrently for learning from input context and the classification task can leverage multilingual knowledge acquired during pretraining. This, we anticipate, will result in enhanced classification performance in a target language when provided with examples in that language. Therefore, in this study we showcase the efficacy of this idea for One-Shot Cross-lingual Transfer, particularly, for adapting to a target language through one-shot demonstrations in-context. This adaptation method proves effective in improving the classification performance and minimizing the transfer gap compared to the Zero-Shot setting. Moreover, we delve into the advantages of employing this approach in scenarios where source task data is not abundant.

3 Our proposed approach: In-Context Cross-Lingual Transfer

Our method aims to simultaneously train a pretrained multilingual encoder-decoder model for (1) a downstream text classification task, and (2) learning from context demonstrations. Then, we expect it to be able to generate predictions in a target language by including context demonstrations in this language. Therefore, we reframe the ICT meta-learning objective by focusing on the transfer between languages rather than tasks. As described above, our proposed procedure, called In-Context Cross-lingual Transfer (IC-XLT), is comprised of two stages:

In-Context Tuning During the meta-training stage, we fine-tune the base multilingual

[2]Meta-learning strategies aim to develop systems that rapidly adapt to new tasks using minimal data instances. In particular, model-based meta-learning focuses on training models to quickly learn from these demonstrations (Nooralahzadeh et al., 2020).

[3]Also, ICT consistently improves performance of ICL and is less sensitive to the shot selection when compared to raw, pre-trained LMs. (Chen et al., 2022)

model for a specific task using data from the source language. Let the set of pairs $D^{src} = \{(x_1^{src}, y_1^{src}), \ldots, (x_{|D|}^{src}, y_{|D|}^{src})\}$ represent the source-language training dataset. The objective is to train the model to predict the label y_i^{src} for a given text x_i^{src} with the following input⇒output format:

$$X^{src}, x_i^{src} \Rightarrow y_i^{src}$$

Here, $X^{src} = ((x_{j_1}, y_{j_1}), \ldots, (x_{j_M}, y_{j_M}))$ is a random sequence of M text-label pairs randomly sampled from D^{src} without replacement, which excludes the pair (x_i^{src}, y_i^{src}).

In-Context Learning At inference, we adapt to a target language by prepending the samples from the one-shot target language training dataset $\widetilde{D}^{tgt} = \{(\widetilde{x}_1^{tgt}, \widetilde{y}_1^{tgt}), \ldots, (\widetilde{x}_N^{tgt}, \widetilde{y}_N^{tgt})\}$ to each entry x_i^{tgt} of the test set to predict y_i^{tgt}. Consequently, the input format mirrors the structure observed in the ICT stage:

$$\widetilde{X}^{tgt}, x_i^{tgt} \Rightarrow y_i^{tgt}$$

Where the sequence $\widetilde{X}^{tgt}$ is a random permutation of $\widetilde{D}^{tgt}$ comprising the one-shot samples, prepended to each x_i^{tgt} entry at the inference stage.

The intuitive idea for this approach is that, after the meta-training stage, we expect the model to understand both the classification task and the contextual relationships relevant to it. During the adaptation stage, the model leverages its multilingual pretraining to interpret context examples in the target language. Note that the adaptation to the target language in this context does not involve any gradient updates, as it occurs solely at the inference stage.

4 Experimental Methodology

In this section, we outline the methodology employed to evaluate the proposed approach. We assess IC-XLT effectiveness in adapting to a target language for the classification task and compare its performance in cross-lingual transfer under (1) full training data on the source language and (2) various source language data budgets. We conduct these limited data experiments to assess how much IC-XLT improves over a traditional fine-tuning method by leveraging the One-Shot demonstrations.

4.1 Data and Evaluation Metrics

We conduct evaluations on two mutlilingual text classification datasets. The first dataset is Aspect Category Detection (ACD) on Restaurant Reviews (Pontiki et al., 2016), a multi-label dataset comprising 12 classes representing different aspects mentioned in reviews. The second dataset is Domain Classification on assistant utterances from the MASSIVE dataset (FitzGerald et al., 2022), a single-label classification dataset with 18 possible domain classes. The datasets were chosen for their larger number of labels and their availability in multiple languages with shared labels. MASSIVE features parallel language splits, each comprising 11.5k samples in the training partition and 2.97k in the test partition.

	Train	Test
English	2000	676
Spanish	2070	881
French	1664	668
Turkish	1232	144
Russian	3655	1209
Dutch	1722	575

Table 1: Length of the training and test partitions in the Aspect Category Detection Dataset.

However, for the Aspect Category Detection dataset, which is non-parallel, the sample counts vary across languages. Detailed information on these counts is presented in Table 1.

We select F_1 micro as our evaluation metric, following (Pontiki et al., 2016). For both datasets, our model is trained in English as the source language, and its performance is evaluated across 5 target languages: *Dutch, Turkish, Russian, French,* and *Spanish* for ACD, and *Thai, Turkish, Russian, French,* and *Spanish* for MASSIVE.

To evaluate the performance of our proposed In-Context Cross-Lingual Transfer (IC-XLT) approach in a resource-constrained source scenario, we construct synthetically reduced datasets by sampling subsets of the training datasets following various k-shot configurations, specifically $K_{src} \in \{8, 16, 32, 64\}$. The objective of these evaluations is to assess IC-XLT's ability to leverage one-shot target demonstrations for enhancing performance in situations where the source language task has limited resources.

4.2 Shot selection

Similar to (Zhao et al., 2021), with "K-shot" we refer to selecting K examples for each of the N classes. The examples are randomly sampled from the training splits of the datasets. Note that the number of shots per label may not precisely be

K due to underrepresented classes in the training set. This holds true for the ACD dataset, where certain classes may have insufficient samples to meet the per-class K value. In such cases, the total number of shots per i-th class is determined as $\min(K, |C_i|)$, where $|C_i|$ represents the total number of samples for the i-th class in the dataset.

Furthermore, since the ACD task involves a multi-label dataset, multi-label examples may add to more than one of the N buckets simultaneously. Hence, the total number of examples in a k-shot dataset is $\leq (k \times |C|)$ where C is the number of classes.

4.3 Experimental Setting

As our multilingual base model, we utilize mT5 (1.2B) (Xue et al., 2020), an encoder-decoder model pre-trained on a diverse corpus encompassing over 100 languages. We employ LoRA (Hu et al., 2021) for fine-tuning the model on the source-language data with varying numbers of shots K_{src}. During the inference stage, label predictions are generated through text generation, which facilitates multi-label inference. We adopt a greedy decoding strategy as implemented in Wolf et al., (Wolf et al., 2020).

We train the ICT models in the source language with different number of context examples, especifically $M = 10$ and $M = 20$.

All models are trained on an NVIDIA Titan RTX GPU for 35 epochs employing a batch size of 8. We assessed learning rates within the range $\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$ for fine-tuning mT5, and selected 5×10^{-4} as it performed adequately for both evaluation datasets in the source language. The LoRA (Hu et al., 2021) parameters are $r = 16, \alpha = 32$, with dropout of 10%.

We conduct evaluations using two seeds for each of the following: the fine-tuning process, K_{src} shot selection, and K_{tgt} shot selection. Since zero-shot approaches do not require selecting target shots, we run a total of 4 and 8 runs for zero-shot and one-shot respectively. For the limited source data training runs, we utilized seeds within $\{1, 2\}$. For the models trained with full source-language data, we trained 5 models with seeds within $\{1, ..., 5\}$ and selected the best 3 in the English validation set.

4.4 Baselines

We benchmark our proposed approach against the following baseline methods, each exclusively utilizing either the source or target data:

(1S) One-shot Prediction Leveraging mT5's pretraining objective, we task the model with predicting the missing span corresponding to the correct label given an input text prepended with one-shot demonstrations. We expect the model to deduce label meanings from the examples without undergoing source-language fine-tuning. This experiment aims to assess the model's proficiency in one-shot prediction without any training, similar to the idea introduced in (Winata et al., 2021), serving as the lower bound when $K_{src} = 0$.

(ZS-XLT) Zero Shot XLT The standard Zero-Shot ($K_{tgt} = 0$) Cross-lingual Transfer approach, where the model is initially trained on a source language, and subsequent inference is conducted on the target language without any additional tuning. In this case, we train the mT5 model through *Prompt-based* fine-tuning (PFT), with the input-output form:

$$x_i \Rightarrow y_i$$

Hence, training is performed at the source and inference at target languages, with no access to source.

(1S-XLT) One Shot XLT Using the same training scheme (PFT), we continue fine-tuning on the checkpoints from the previous baseline, training with *One-Shot per label* in the target language. The training is conducted for 50 epochs with a learning rate of 5×10^{-6}. This approach is the standard gradient-based approach for adapting to a target language in Few-Shot Cross-Lingual Transfer (Lauscher et al., 2020). Although larger values for the number of target language shots K_{tgt} could be considered, it is outside the scope of this work, which is delimited to the One-shot setting.

(IC-XLT$_{SRC}$) IC-XLT with source-language context We use the same models trained for IC-XLT, however, in this method In-Context examples are not drawn from the target language but from the source language used for their training. In essence, this can be considered a Zero-Shot baseline since no target language is involved for adaptation. Through this baseline we aim to evaluate the relevance of the *target* language One-Shot samples at the adaptation stage, assessing whether they are necessary for successful transfer to that target language.

5 Results and analysis

IC-XLT performance at Cross-lingual transfer For the first experiment, we compare our proposed

MASSIVE						
Method	ENG (SRC)	TUR	THA	SPA	FRA	RUS
1S	$39.41_{\pm 1.14}$	$33.74_{\pm 0.25}$	$33.5_{\pm 4.15}$	$32.09_{\pm 1.04}$	$30.64_{\pm 2.27}$	$26.04_{\pm 7.1}$
ZS-XLT	$86.57_{\pm 1.17}$	$64.23_{\pm 5.58}$	$70.09_{\pm 2.97}$	$72.49_{\pm 1.5}$	$74.94_{\pm 1.03}$	$74.64_{\pm 2.9}$
1S-XLT		$64.14_{\pm 5.06}$	$70.08_{\pm 2.87}$	$72.36_{\pm 1.51}$	$74.95_{\pm 0.8}$	$74.55_{\pm 2.73}$
IC-XLT$_{SRC}$	$89.45_{\pm 0.34}$	$69.39_{\pm 2.13}$	$78.02_{\pm 0.58}$	$77.55_{\pm 0.89}$	$79.96_{\pm 1.5}$	$82.63_{\pm 0.99}$
IC-XLT		$\mathbf{78.32_{\pm 2.41}}$	$\mathbf{78.87_{\pm 0.5}}$	$\mathbf{80.63_{\pm 1.76}}$	$\mathbf{83.47_{\pm 1.02}}$	$\mathbf{83.41_{\pm 1.1}}$
ACD						
	ENG (SRC)	TUR	NLD	SPA	FRA	RUS
1S	$37.38_{\pm 4.7}$	$19.52_{\pm 5.1}$	$20.51_{\pm 1.79}$	$34.76_{\pm 5.94}$	$31.84_{\pm 5.55}$	$34.02_{\pm 2.64}$
ZS-XLT	$76.6_{\pm 1.13}$	$61.72_{\pm 5.12}$	$66.37_{\pm 1.25}$	$65.96_{\pm 1.49}$	$65.42_{\pm 0.99}$	$68.88_{\pm 1.21}$
1S-XLT		$62.01_{\pm 4.7}$	$66.69_{\pm 1.19}$	$66.08_{\pm 1.28}$	$65.84_{\pm 0.57}$	$69.12_{\pm 1.0}$
IC-XLT$_{SRC}$	$81.68_{\pm 0.65}$	$70.14_{\pm 3.97}$	$70.05_{\pm 1.55}$	$72.83_{\pm 0.28}$	$73.57_{\pm 0.74}$	$75.67_{\pm 0.62}$
IC-XLT		$\mathbf{76.83_{\pm 1.66}}$	$\mathbf{71.5_{\pm 1.5}}$	$\mathbf{74.32_{\pm 0.32}}$	$\mathbf{74.88_{\pm 1.51}}$	$\mathbf{76.01_{\pm 1.02}}$

Table 2: Average F_1 micro in the two evaluated datasets, trained with full data in English, the source language. Here, $\pm$ is the standard deviation of the different runs. The ICT Methods (IC-XLT$_{SRC}$ and IC-XLT) are for $M = 20$.

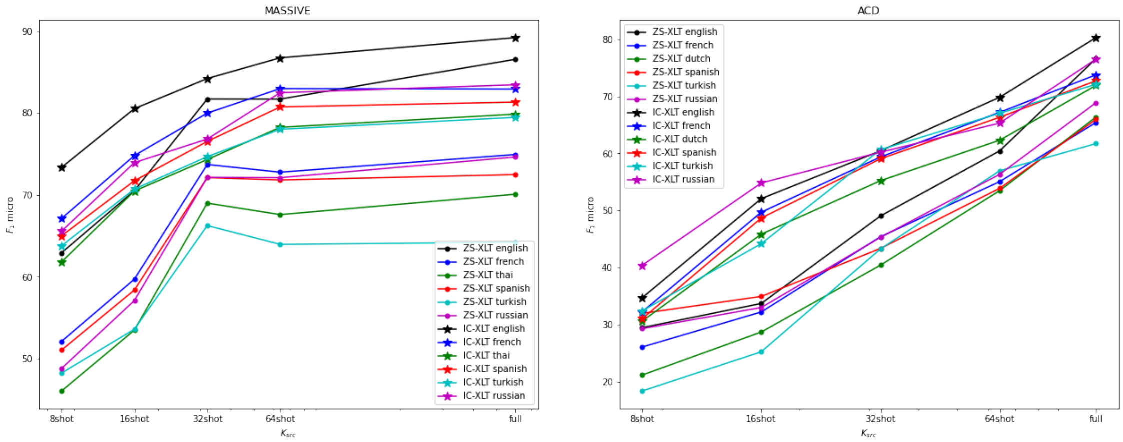
approach, IC-XLT, to the baselines detailed in Section 4.4 using the full training set in the source language. We observed a general trend where mT5 models trained with In-Context Tuning, which employs the input-output setting $\widetilde{X}, x_i \Rightarrow y_i$, consistently outperformed models subjected to Prompt-based fine-tuning with $x_i \Rightarrow y_i$ under the same training regimes, despite both models being trained for an equivalent number of steps and exact same data instances. We hypothesize that this superior performance may be attributed to the fact that the ICT-trained models *see* M randomly ordered input-output examples at each instance, even though they are tasked with predicting only x_i.

The significant increase in performance observed in the source language benefits evaluations in the target languages after the adaptation stage. We present the F_1 micro scores across five different languages on the Aspect Category Detection (ACD) and MASSIVE datasets in Table 2. We observe that IC-XLT effectively outperforms the evaluated baselines by a substantial margin in the evaluated datasets, greatly improving mT5 cross-lingual transfer performance. A crucial observation is that for both of the evaluated datasets there is a noticeable increase in performance from IC-XLT$_{SRC}$ to IC-XLT. This means that the proposed approach is effectively taking advantage of the One-Shot target language demonstrations for adapting to it *at inference* at the In-Context Learning stage.

On the other hand, the 1S-XLT approach, which is further fine-tuned on One-Shot target samples, did not improve over ZS-XLT by an important margin. While a small improvement is observable for the ACD task, there is also a minor performance decrease for the MASSIVE dataset. This result could be attributed to the limited number of samples available for the fine-tuning process, as only one shot per label is employed. Since we do not observe a noticeable improvement of 1S-XLT over ZS-XLT in the full training data experiments, and adapting the former requires further fine-tuning, we compare IC-XLT with ZS-XLT in the limited-resource scenario.

Performance with limited source-language data We conduct experiments to quantify the ability of IC-XLT to perform at scenarios with limited source language resources. For this we evaluate IC-XLT and ZS-XLT models trained with $K_{src} \in \{8, 16, 32, 64\}$. We noticed that models trained with the ICT framework generally perform better compared to PFT for low values of K_{src}. In Figures 1b and 1a, we illustrate the average performance per target language in the datasets at different source-language resource availability regimes. The plot shows evaluations for ZS-XLT (PFT training) and IC-XLT (ICT training). We can see that ICT makes better use of resources than Prompt-based fine-tuning specially at at smaller values for K_{src} for both datasets, although this is especially notable in MASSIVE. Furthermore, the performance difference with the source language (English) is visibly smaller for ICT training, more discussion on this can be found below.

The F_1-micro averages for the target languages are shown in Tables 3 and 4 for ACD and MASSIVE respectively. We can observe that models trained on limited data achieve competitive or superior performance compared to ZS-XLT models trained with full source datasets (See Appendix A



(a) MASSIVE performance with different souce data availability. IC-XLT trained with $M = 10$.

(b) ACD performance with different souce data availability. IC-XLT trained with $M = 10$.

Figure 1: Comparison of IC-XLT and ZS-XLT performance at different source language data budgets. We can observe that, in general, the IC-XLT models yield better performance compared to ZS-XLT. This is especially notable at lower resource scenarios.

for the complete tables with results in each target language). Given that the adaptation to each target language occurs at inference, the improvement over ZS-XLT comes at no extra computational cost and at a minimal data cost. This allows to achieve good performance with limited computational and data resources.

We find that, for models trained on full data, $M = 20$ (the number of in-context demonstrations during ICT training) performs slightly better on the Aspect Category Detection (ACD). For models trained with lower resources, $M = 20$ performs suboptimally compared to traditional fine-tuning and $M = 10$ in ACD, but achieves a better performance in MASSIVE. We believe that since ACD contains only 12 labels, a context length of 20 will inevitably prepend more repeated context examples than the MASSIVE dataset[4] when training with limited data. This reduced variability may hurt the model's performance compared to $M = 10$.

Measuring the transfer gap with the source language. By measuring the performance gap between the source language and the target language, we aim to quantify the contribution of the ICT training framework and One-Shot target demonstrations for mitigating this gap. As we provide the model with target language examples, we anticipate a *smaller* decrease in performance from the source language when adapting to a new language, compared to ZS-XLT. We can measure this by computing the average transfer gap $\bar{\Delta}\%$, which is the average percentage decrease in performance relative to the evaluations on the test set in the source language (English):

$$\bar{\Delta}\% = 100 \times \mathbf{E}\left[\frac{P_{tgt\ lang}}{P_{src\ lang}} - 1\right]$$

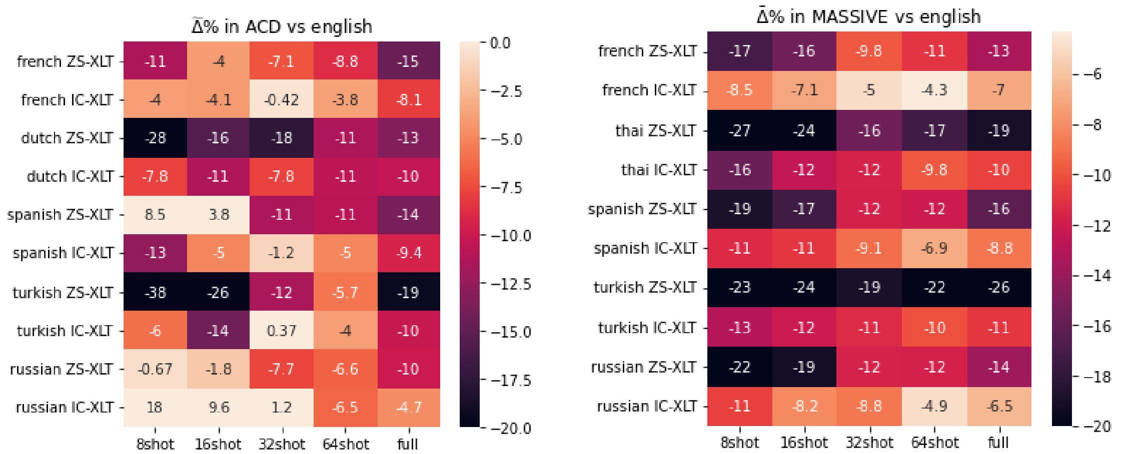
where $P_{tgt\ lang}$ and $P_{tgt\ lang}$ represent the evaluation performance of the exact same model on the target and source test sets, respectively. The performance gap values are shown in Figures 2a and 2b for ACD and MASSIVE respectively. We can observe that in almost all cases and all source language data budgets we obtain a reduced average transfer gap $\Delta\%$ through IC-XLT compared to ZS-XLT.

We find that $\Delta\%$ for IC-XLT models can be reduced by a very significant margin especially in target languages linguistically distant from English such as Turkish or Thai. The obtained $\Delta\%$ values, as well as the performance improvement from IC-XLT$_{SRC}$ to IC-XLT shown in Table 2, under-

K_{src}	1S		
0	$28.13_{\pm 7.49}$		
	ZS-XLT	IC-XLT M = 10	IC-XLT M = 20
8	$25.4_{\pm 5.02}$	$33.34_{\pm 3.59}$	$16.64_{\pm 2.44}$
16	$30.84_{\pm 3.44}$	$48.66_{\pm 3.66}$	$47.04_{\pm 1.81}$
32	$43.56_{\pm 1.81}$	$58.91_{\pm 1.93}$	$61_{\pm 1.51}$
64	$55.15_{\pm 1.34}$	$65.64_{\pm 1.79}$	$65.28_{\pm 1.27}$
Full	$65.67_{\pm 2.3}$	$73.4_{\pm 1.68}$	$\mathbf{74.71_{\pm 1.83}}$

Table 3: Average F_1 micro across 5 target languages for Aspect Category Detection. In this table, $\pm$ refers to the standard deviation of the means of different language.

[4]Which contains 18.



(a) $\bar{\Delta}\%$ of the target languages vs English in the Aspect Category Detection dataset.

(b) $\bar{\Delta}\%$ of the target languages vs English in the MASSIVE domain detection dataset.

Figure 2: The average transfer gap $\bar{\Delta}\%$ of IC-XLT and ZS-XLT at different source language data budgets. (IC-XLT $M = 10$). We can observe that, for most cases, IC-XLT yields a smaller drop in performance after transfering to a target language compared to ZS-XLT.

K_{src}	1S		
0	$31.2_{\pm 3.14}$		
	ZS-XLT	IC-XLT M = 10	IC-XLT M = 20
8	$49.25_{\pm 2.14}$	$64.66_{\pm 1.78}$	$68.05_{\pm 1.39}$
16	$56.46_{\pm 2.52}$	$72.34_{\pm 1.74}$	$75.44_{\pm 1.55}$
32	$70.65_{\pm 2.68}$	$76.48_{\pm 2.02}$	$78.95_{\pm 1.78}$
64	$69.66_{\pm 3.38}$	$80.5_{\pm 2.07}$	$\mathbf{81.54_{\pm 1.52}}$
Full	$71.28_{\pm 3.93}$	$81.42_{\pm 1.59}$	$80.94_{\pm 2.18}$

Table 4: Average F_1 micro across 5 target languages for MASSIVE (Domain Classification). In this table, ± refers to the standard deviation of the means of different language.

score that introducing in-context target language examples through IC-XLT effectively mitigates the transfer gap.

6 Conclusion

In this paper, we investigated the application of In-Context Tuning for One-Shot Cross-lingual transfer, introducing In-Context Cross-lingual Transfer (IC-XLT). Our evaluations conducted on an mT5 model demonstrate the efficacy of the proposed method in effectively adapting at inference to target languages using only one-shot demonstrations in-context, all without incurring additional computational expenses. Furthermore, in comparison to ZS-XLT and 1S-XLT, IC-XLT exhibits a better performance and smaller transfer gap.

In scenarios with limited source-language training data, we provide empirical evidence that IC-XLT learns better the source language at the meta-training stage and demonstrates a smaller transfer gap at the adaptation stage with the one-shot demonstration, compared to ZS-XLT. This makes IC-XLT a valuable tool for cross-lingual transfer in resource-limited scenarios.

To the best of our knowledge, this is the first study on the application of In-Context Tuning to Cross-Lingual Transfer. For future work, we aim to explore the potential and limitations of this approach by evaluating its applicability to other architectures, such as decoder-only or encoder-only models, and examining the impact of training with a greater number of examples in-context.

7 Limitations

In this study, we implement our approach using an mT5-large encoder-decoder model. However, an evaluation of its applicability to encoder-only or decoder-only models remains unexplored and it is left for future work. Furthermore, due to storage constraints and the need to conduct experiments across diverse seeds and training data budgets, we opted to fine-tune the models using LoRA (Hu et al., 2021). While some variability compared to the fully trained model is expected with this architectural choice, empirical evidence from (Hu et al., 2021) suggests that its impact is minimal.

Finally, it is important to outline that due to the maximum input length of mT5 (1024), scaling IC-XLT is to a largett number of target language shots (e.g $K_{tgt} \in \{4, 8, 16\}$) may prove difficult using the current approach. This challenge is particularly pronounced in scenarios with a substantial number of labels, where input text may need to be truncated. Consequently, there is a need to devise a strategy to either reduce input length or integrate information from different example batches in order to address this limitation.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, Dublin, Ireland. 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. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, 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.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. 2023. Pretraining data mixtures enable narrow model selection capabilities in transformer models.

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 Appendix

A.1 Performance metrics per language on the limited data experiments.

In this section we show the complete results of evaluations in the different target languages with ZS-XLT and One-shot IC-XLT. Table 5 illustrates the cross-lingual transfer performance of the evaluated models with English as the source language. Similarly, the results on the MASSIVE dataset are shown in Table 6, also with English as the source language.

A.2 Evaluations in Russian and Turkish

Although the main focus of this work is to evaluate cross-lingual transfer with English as source language, we include –smaller– evaluations on the ACD dataset with Russian and Turkish as source languages. With these evaluations we aim to further demonstrate the effectiveness of our approach across languages and explore the potential for cross-lingual transfer in various language pairs. We evaluate $K_{tgt} = 64$ and full training data. In Table 7, we compare their performance in the source language with the average performance in the target languages on the Aspect Category Detection dataset. We also observe an important improvement compared to ZS-XLT and a reduction in the average transfer gap for most of the target languages when employing IC-XLT (See Figure 3). This reduction in the transfer gap, particularly pronounced in the case of $K_{src} = 64$, highlights the significance of target-shots, especially when working with limited source data. Also, we include the evaluations in Russian and Turkish in the ACD dataset, displayed in Table 7.

A.3 Licences of systems and datasets

In this work, the tools utilized include an mT5 model and the *transformers* library (Wolf et al., 2020), both of which use the Apache 2.0 license. The MASSIVE dataset, on the other hand, operates under a CC by 4.0 license. As for the Aspect Category Detection dataset, it employs a MS-NC-No ReD license, which limits its usage strictly to an academic scope. Since the aim of this work is to evaluate the performance of a proposed cross-lingual system, we adhere to all the licenses of the utilized material.

The research presented in this paper is intended for academic purposes, and therefore, we adhere to the licenses governing all utilized materials.

K_{src}	ENG	FRA	NLD	SPA	TUR	RUS
			Target Language			
			IC-XLT ($M = 10$)			
$K_{src} = 8$	$34.71_{\pm 7.33}$	$32.24_{\pm 4.16}$	$30.62_{\pm 6.04}$	$31.04_{\pm 12.53}$	$32.41_{\pm 8.23}$	$40.4_{\pm 7.85}$
$K_{src} = 16$	$52.08_{\pm 11.85}$	$49.71_{\pm 11.04}$	$45.86_{\pm 9.55}$	$48.69_{\pm 9.92}$	$44.2_{\pm 9.68}$	$54.83_{\pm 5.18}$
$K_{src} = 32$	$60.45_{\pm 8.84}$	$59.38_{\pm 5.55}$	$55.23_{\pm 5.6}$	$59.06_{\pm 5.06}$	$60.66_{\pm 9.33}$	$60.21_{\pm 2.98}$
$K_{src} = 64$	$69.84_{\pm 1.32}$	$67.2_{\pm 1.49}$	$62.32_{\pm 1.1}$	$66.33_{\pm 0.98}$	$67.04_{\pm 2.92}$	$65.32_{\pm 0.62}$
Full data	$80.28_{\pm 1.03}$	$73.76_{\pm 0.24}$	$71.91_{\pm 1.64}$	$72.73_{\pm 1.94}$	$72.1_{\pm 3.89}$	$76.51_{\pm 0.56}$
			IC-XLT ($M = 20$)			
$K_{src} = 8$	$23.66_{\pm 6.16}$	$17.12_{\pm 13.05}$	$15.17_{\pm 6.96}$	$13.49_{\pm 8.87}$	$20.83_{\pm 13.2}$	$16.58_{\pm 11.33}$
$K_{src} = 16$	$41.19_{\pm 9.22}$	$48.24_{\pm 4.75}$	$44.89_{\pm 6.13}$	$47.26_{\pm 8.59}$	$45.17_{\pm 8.6}$	$49.64_{\pm 4.95}$
$K_{src} = 32$	$63.25_{\pm 2.36}$	$61.37_{\pm 2.49}$	$58.01_{\pm 1.44}$	$61.74_{\pm 1.3}$	$61.85_{\pm 5.97}$	$62.05_{\pm 1.53}$
$K_{src} = 64$	$70.01_{\pm 1.77}$	$65.86_{\pm 1.28}$	$63.1_{\pm 0.79}$	$65.34_{\pm 1.18}$	$66.98_{\pm 2.66}$	$65.12_{\pm 0.96}$
Full data	$81.68_{\pm 0.65}$	$74.88_{\pm 1.51}$	$71.5_{\pm 1.5}$	$74.32_{\pm 0.32}$	$76.83_{\pm 1.66}$	$76.01_{\pm 1.02}$
			ZS-XLT			
$K_{src} = 8$	$29.49_{\pm 2.06}$	$26.1_{\pm 1.46}$	$21.19_{\pm 3.74}$	$31.99_{\pm 2.46}$	$18.4_{\pm 4.34}$	$29.32_{\pm 2.84}$
$K_{src} = 16$	$33.73_{\pm 4.12}$	$32.24_{\pm 3.63}$	$28.73_{\pm 6.76}$	$34.96_{\pm 4.48}$	$25.27_{\pm 7.68}$	$33.02_{\pm 3.22}$
$K_{src} = 32$	$49.05_{\pm 5.33}$	$45.41_{\pm 4.68}$	$40.44_{\pm 4.68}$	$43.37_{\pm 4.73}$	$43.28_{\pm 3.93}$	$45.31_{\pm 5.9}$
$K_{src} = 64$	$60.45_{\pm 3.3}$	$55.07_{\pm 2.46}$	$53.49_{\pm 1.88}$	$53.9_{\pm 3.4}$	$56.96_{\pm 3.02}$	$56.35_{\pm 1.23}$
Full data	$76.6_{\pm 1.13}$	$65.42_{\pm 0.99}$	$66.37_{\pm 1.25}$	$65.96_{\pm 1.49}$	$61.72_{\pm 5.12}$	$68.88_{\pm 1.21}$

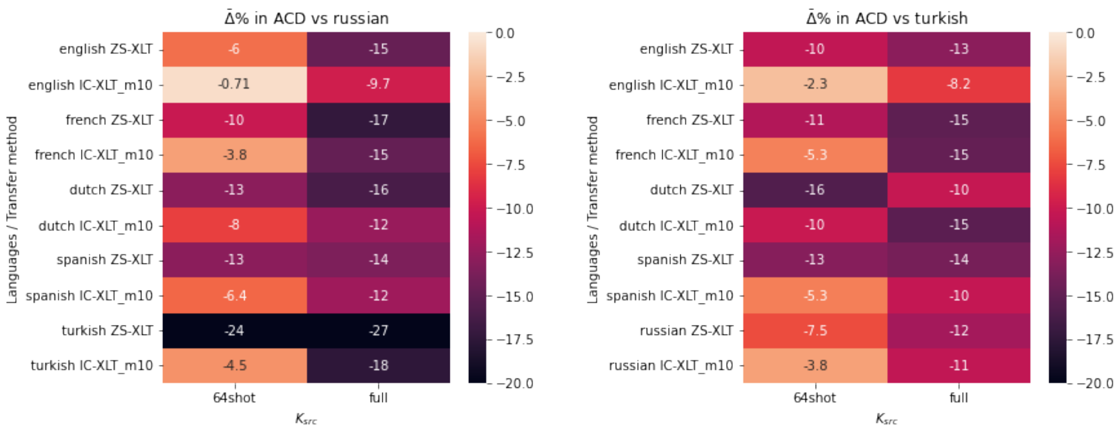
Table 5: Average per language across the different runs for evaluations under different resource budgets for the Aspect Category Detection dataset. In here, $\pm$ refers to the standard deviation of the performance on the conducted runs.

K_{src}	ENG	FRA	THA	SPA	TUR	RUS
			Target Language			
			IC-XLT ($M = 10$)			
$K_{src} = 8$	$73.36_{\pm 0.92}$	$67.12_{\pm 1.62}$	$61.81_{\pm 2.62}$	$65_{\pm 1.37}$	$63.79_{\pm 2.42}$	$65.57_{\pm 2.6}$
$K_{src} = 16$	$80.54_{\pm 0.99}$	$74.81_{\pm 1.81}$	$70.48_{\pm 2.28}$	$71.74_{\pm 2.7}$	$70.72_{\pm 2.4}$	$73.95_{\pm 2.8}$
$K_{src} = 32$	$84.22_{\pm 0.62}$	$80_{\pm 0.73}$	$74.33_{\pm 1.03}$	$76.54_{\pm 0.66}$	$74.68_{\pm 0.97}$	$76.83_{\pm 0.94}$
$K_{src} = 64$	$86.75_{\pm 0.29}$	$82.99_{\pm 0.78}$	$78.26_{\pm 0.56}$	$80.75_{\pm 1.2}$	$78.02_{\pm 0.9}$	$82.49_{\pm 0.89}$
Full data	$89.22_{\pm 0.37}$	$82.93_{\pm 1.38}$	$79.87_{\pm 0.9}$	$81.34_{\pm 1.33}$	$79.48_{\pm 1.15}$	$83.46_{\pm 1}$
			IC-XLT ($M = 20$)			
$K_{src} = 8$	$73.24_{\pm 2.71}$	$67.26_{\pm 3.72}$	$66.53_{\pm 2.65}$	$67.04_{\pm 3.46}$	$70.03_{\pm 3.31}$	$69.41_{\pm 3.01}$
$K_{src} = 16$	$82_{\pm 1.37}$	$75.98_{\pm 1.83}$	$72.55_{\pm 0.81}$	$75.4_{\pm 1.5}$	$76.18_{\pm 1.43}$	$77.11_{\pm 1.51}$
$K_{src} = 32$	$85.03_{\pm 0.52}$	$80.06_{\pm 1.06}$	$76.1_{\pm 2.19}$	$78.68_{\pm 1.2}$	$78.46_{\pm 1.28}$	$81.43_{\pm 1.16}$
$K_{src} = 64$	$87.18_{\pm 0.66}$	$83.29_{\pm 0.79}$	$79.36_{\pm 0.97}$	$81.06_{\pm 1.33}$	$80.75_{\pm 1.52}$	$83.24_{\pm 0.8}$
Full data	$89.45_{\pm 0.34}$	$83.47_{\pm 1.02}$	$78.87_{\pm 0.5}$	$80.63_{\pm 1.76}$	$78.32_{\pm 2.41}$	$83.41_{\pm 1.1}$
			ZS-XLT			
$K_{src} = 8$	$62.93_{\pm 1.5}$	$52.11_{\pm 0.77}$	$46.05_{\pm 0.15}$	$51.05_{\pm 0.8}$	$48.24_{\pm 0.99}$	$48.8_{\pm 1.05}$
$K_{src} = 16$	$70.52_{\pm 7.24}$	$59.71_{\pm 7.75}$	$53.49_{\pm 7.93}$	$58.39_{\pm 7.04}$	$53.6_{\pm 5.47}$	$57.1_{\pm 7.63}$
$K_{src} = 32$	$81.72_{\pm 1.39}$	$73.72_{\pm 1.88}$	$69_{\pm 2.74}$	$72.12_{\pm 1.64}$	$66.26_{\pm 1.79}$	$72.15_{\pm 2}$
$K_{src} = 64$	$81.71_{\pm 2.81}$	$72.78_{\pm 5.1}$	$67.6_{\pm 5.01}$	$71.83_{\pm 4.43}$	$63.97_{\pm 5.46}$	$72.11_{\pm 5.17}$
Full data	$86.57_{\pm 1.17}$	$74.94_{\pm 1.03}$	$70.09_{\pm 2.97}$	$72.49_{\pm 1.5}$	$64.23_{\pm 5.58}$	$74.64_{\pm 2.9}$

Table 6: Average per language across the different runs for evaluations under different resource budgets in the MASSIVE Domain Classification Task. In here, $\pm$ refers to the standard deviation of the performance on the conducted runs.

		Russian as source		Turkish as Source	
Method	K_{src}	Russian	Avg target	Turkish	Avg target
ZS-XLT	64	$60.66_{\pm 4.65}$	$52.73_{\pm 3.64}$	$62.42_{\pm 3.12}$	$54.97_{\pm 1.73}$
IC-XLT	64	$68.33_{\pm 1.06}$	$65.12_{\pm 1.67}$	$67.45_{\pm 7.56}$	$63.69_{\pm 1.78}$
ZS-XLT	full	$74.55_{\pm 4.43}$	$61.31_{\pm 3.61}$	$63.46_{\pm 4.34}$	$55.25_{\pm 1.1}$
IC-XLT	full	$81.7_{\pm 1.17}$	$70.84_{\pm 2.17}$	$80.79_{\pm 2.5}$	$71.18_{\pm 2.21}$

Table 7: Average performance on the target languages on Turkish and Russian as source. For this experiments we set $M = 10$



(a) $\bar{\Delta}\%$ with Russian as source language.

(b) $\bar{\Delta}\%$ with Turkish as source language.

Figure 3: Average transfer gaps in Turkish and Russian.