Cross-lingual Few-Shot Learning on Unseen Languages

Genta Indra Winata[*1], **Shijie Wu**[*1], **Mayank Kulkarni**[*†2]
Thamar Solorio[‡1], **Daniel Preoţiuc-Pietro**[‡1]
[1]Bloomberg [2]Amazon Alexa AI
{gwinata,swu671,tsolorio,dpreotiucpie}@bloomberg.net, maykul@amazon.com

Abstract

Large pre-trained language models (LMs) have demonstrated the ability to obtain good performance on downstream tasks with limited examples in cross-lingual settings. However, this was mostly studied for relatively resource-rich languages, where at least enough unlabeled data is available to be included in pre-training a multilingual language model. In this paper, we explore the problem of cross-lingual transfer in unseen languages, where no unlabeled data is available for pre-training a model. We use a downstream sentiment analysis task across 12 languages, including 8 unseen languages, to analyze the effectiveness of several few-shot learning strategies across the three major types of model architectures and their learning dynamics. We also compare strategies for selecting languages for transfer and contrast findings across languages seen in pre-training compared to those that are not. Our findings contribute to the body of knowledge on cross-lingual models for low-resource settings that is paramount to increasing coverage, diversity, and equity in access to NLP technology. We show that, in few-shot learning, linguistically similar and geographically similar languages are useful for cross-lingual adaptation, but taking the context from a mixture of random source languages is surprisingly more effective. We also compare different model architectures and show that the encoder-only model, XLM-R, gives the best downstream task performance.

1 Introduction

The availability of large-scale multilingual pre-trained language models has enabled a more effective transfer of knowledge across languages (Conneau and Lample, 2019; Pires et al., 2019; Wu and Dredze, 2019a; Shliazhko et al., 2022; Lin et al., 2021), thus limiting the need to gather task-specific annotated data for a given target language. Recent research into few-shot learning approaches proposed methods that explicitly aim to improve performance when few annotated data points are available to perform a task (Brown et al., 2020; Lin et al., 2021; Srivastava et al., 2022), semantic parsing (Liu et al., 2021c), topic modeling (Bianchi et al., 2021). Further, cross-lingual few-shot learning uses multilingual models and few-shot learning methods to perform a task given limited training data in another language and has shown promise on several downstream tasks (Lauscher et al., 2020a; Liu et al., 2020; Zhao et al., 2021; Winata et al., 2021).

These studies have only looked at relatively resource-rich target languages, as they are part of the pre-training data for the multilingual language model, and even for these languages, the representation quality is not equal due to imbalanced corpus size (Wu and Dredze, 2020). Representation quality is expectedly lower for the vast majority of the spoken languages in the world, most of which are not part of the pre-training data in multilingual models, albeit being spoken by large populations. For example, Ngaju is the native language of over 890,000 people, yet there is no Wikipedia available for this language, which is a common source of data for pre-training. Cross-lingual few-shot learning methods are a promising avenue of research for enabling NLP technologies for such languages, especially as we can assume both unlabeled and, especially, labeled data for a given task are difficult to obtain at scale (Joshi et al., 2020; Lauscher et al., 2020b; Pfeiffer et al., 2020; Liu et al., 2021b; Winata et al., 2021; Aji et al., 2022).

This paper is the first to study cross-lingual few-shot learning methods in unseen languages at the pre-training stage. We focus mainly on how to most effectively train a model for a downstream classification task in an unseen language without having access to any labeled data in that language. We experiment with all three major types of multilin-

[*] The authors contributed equally. [†] The work was done while at Bloomberg. [‡] Senior authors.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 777–791
November 20–23, 2022.

gual pre-trained language model architectures, including the encoder-only XLM-R (Conneau et al., 2020a), the decoder-only XGLM (Lin et al., 2021) and the encoder-decoder mT5 (Xue et al., 2021) models. We combine these with different strategies for few-shot learning for a new language, including in-context learning, prompt-based fine-tuning, and encoder-based fine-tuning. We evaluate the effectiveness of these approaches under varying levels of available training data. We perform several analyses to understand aspects such as the performance gap between languages seen in pre-training compared to those unseen and which source languages are best suited for a target language.

We perform this study on the downstream task of sentiment analysis across 12 languages spoken in Indonesia plus English from the NusaX corpus (Winata et al., 2022). This dataset contains parallel sentences annotated for sentiment, which conveniently allows control for content drift when comparing transfer capabilities across languages.

Our contributions are as follows:

- The first study on cross-lingual few-shot learning on diverse low-resource languages not seen during pre-training across three model types and three few-shot learning strategies focusing on the task of sentiment prediction.
- Insights into the learning dynamics with varying amounts of training data.
- Analysis of various data mixing strategies for multi-source cross-lingual few-shot learning.
- Insights into transfer learning effectiveness across languages.

In sum, our work contributes new insights to the growing body of work in cross-lingual NLP for extremely low-resource languages, a critical step in increasing coverage and access to NLP technology.

2 Methodology

We define our task as follows: Let θ be the LM and $\mathcal{T}_l$ be the dataset for language l consisting of N sentence and label pairs $\{x_1, y_1\}, \{x_2, y_2\}, ..., \{x_N, y_N\}$, where x_i, y_i, are the inputs and labels, respectively. In the **cross-lingual setting**, we take the source language l_{src} from a pool of languages L that does not include the target language l_{tgt}. In this work, we categorize languages as seen and unseen. The **unseen languages**, are those languages that were not present in the data used to pre-train the multilingual models, while the **seen languages** were included during pre-training. Our goal is to investigate what are the most successful strategies for cross-lingual transfer learning under extremely limited data settings. With this in mind, we want to answer the following questions:

- Multilingual models: *which model architecture is better for this scenario?*
- Few-shot learning: *different model architectures will require different learning, which is better?*
- Language selection: given that data is available for several source languages, *how should we select the languages to improve transfer to target languages?*

Next, we expand on the methods followed in order to answer the questions above.

2.1 Multilingual Language Models

We experiment with a model from each of the three major types of pre-trained language model architectures: encoder-only architectures such as BERT (Devlin et al., 2019), decoder-only architectures such as the GPT series (Brown et al., 2020) and encoder-decoder architectures such as T5 (Raffel et al., 2020). Pre-trained multilingual models, such as mBERT, significantly improve the ability to generate cross-lingual representations (Conneau and Lample, 2019; Pires et al., 2019; Wu and Dredze, 2019a), which led to the creation of multilingual variants for all architecture types. In this paper, we use XGLM (Lin et al., 2021), XLM-R (Conneau et al., 2020a), and mT5 (Xue et al., 2021).

2.2 Few Shot Learning Strategies

We explore multiple approaches to few-shot learning using LMs as follows:

2.2.1 Cross-lingual Few-shot Fine-tuning

Encoder-based Model Fine-tuning The common approach to applying a pre-trained LM to a downstream task involves fine-tuning the pre-trained model with a classification head on the labeled data. Given k training samples, we take them to fine-tune an encoder model θ (i.e., XLM-R). In this case, we fine-tune the model using the text samples as input and update all parameters of the encoder.

Prompt-based Fine-tuning For the XGLM and mT5 models, we conduct few-shot fine-tuning by casting the problem as text-to-text using a simple template $t = [x_i => y_i]$ as in Tab. 1. For mT5, the template is $t = ([x_i =>], [y_i])$. We fine-tune all pa-

Prompt	Example	Translation
$x_1 => y_1 \backslash n$	Susujih segar ngon sayur nyang bereh, nyum kuah mangat ngon peulayanan nyang ramah that**=>positive**	The milk is fresh with amazing vegetables and delicious soup flavour, complete with super nice service.**=>positive**
...	...	
$x_k => y_k \backslash n$	Menyeusai kupeugah bak kah, farrel.**=>negative**	I regret ever telling ye anything, Farrel.**=>negative**
$Q =>$	Ae beneh, iye sedeng nyaga warung**=>**	Yeah that's right, he's looking after the store now**=>**

Table 1: Cross-lingual prompt template. It shows the k-shot context in **Acehnese** and the query in **Balinese**.

rameters of the model to maximize $p_\theta(t)$. Instances in the template belong to the source language l_{src}. During inference, we compute the probability distribution of the label as the following:

$$\hat{y} = \arg\max_y P(y|x, \theta). \quad (1)$$

2.2.2 Cross-lingual In-context Learning

In-context learning is proposed as an alternative for few-shot learning in Brown et al. (2020). In this setting, we use a set of examples from a template to perform the downstream task directly without any gradient update.[1]

We set up our prompt $\mathcal{P} = (C, Q)$ as the concatenation of context C and query Q. The context C is generated by following a template shown in Tab. 1, and we sample k pairs of inputs and labels from l_{src} to fill the template. The query Q is the sentence from the test sample we want to evaluate. For each test sample, we compute the probability distribution of each label and take the highest score as the predicted label $\hat{y}$:

$$\hat{y} = \arg\max_y P(y|\mathcal{P}, \theta). \quad (2)$$

In the zero-shot in-context learning setting, the prompt $\mathcal{P}$ only consists of the query Q.

2.3 Language Sample Selection Methods

While many studies explore single- and multi-source transfer between languages seen during LM pre-training, to the best of our knowledge, there is no study covering the setup where languages are unseen during pre-training as both source and target languages. Given that existing labeled datasets only cover a small fragment of the languages worldwide, it would be helpful to be able to build NLP systems via cross-lingual transfer with as little labeled data in the target languages as possible.

We explore various methods for language selection for a multi-source transfer involving unseen languages, aiming to choose source languages that are likely to be useful for the target languages. We evaluate different mixing strategies based on the single-source performance of each target language, geographic vicinity, and linguistic language roots. Our goal is to understand whether mixed language prompts provide any advantage to unseen languages and to what extent they help alleviate the data scarcity problem in cross-lingual settings.

Random Mixing We randomly sample instances from different languages for each target language, excluding the target language **(random-mix)**. For in-context learning, the prompt is then constructed using the instances. For fine-tuning, we treat the same set of instances as the training set.

Best Single-Source Languages Mixing We anticipate that selecting source languages using linguistic knowledge will give an advantage over the

[1]While there is no gradient update in in-context learning, we still refer to the act as "training" for writing simplicity.

Language	Language Root	Geographical Location	Availability in LM*
Acehnese (ace)	Malayo-Chamic	Sumatera	×
Balinese (ban)	Bali-Sasak-Sumbawa	Java†	×
Banjarese (bjn)	Malayo-Chamic	Borneo	×
Buginese (bug)	South Sulawesi	Sulawesi	×
English (eng)	Germanic	n/a	✓
Indonesian (ind)	Malayo-Chamic	‡	✓
Javanese (jav)	Javanese	Java	✓
Madurese (mad)	Madurese	Java	×
Minangkabau (min)	Malayo-Chamic	Sumatera	×
Ngaju (nij)	Greater Barito	Borneo	×
Sundanese (sun)	Sundanese	Java	✓
Toba Batak (bbc)	Northwest Sumatera	Sumatera	×

Table 2: Languages in the NusaX dataset. †We group Balinese to Java because it is located close to Java. *We check whether the language is part of the pre-training dataset of XLM-R, XGLM, and mT5. A language is considered "unseen" if it is not present in the pre-training data.

data split	positive	negative	neutral
train	189	192	119
valid	38	38	24
test	151	153	96

Table 3: The label distribution of the NusaX dataset splits.

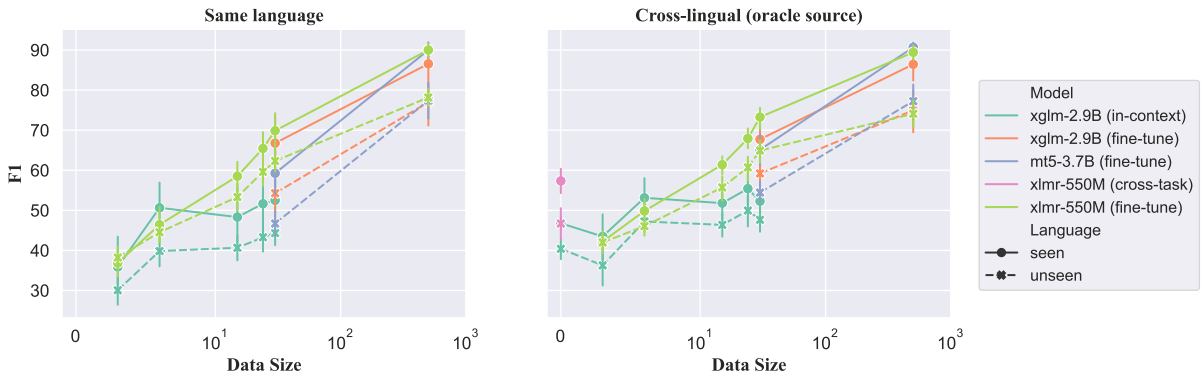

Figure 1: Experimental results on sentiment classification in F1 across various data sizes (X-axis), model types, and learning setups.

random and single source language settings. To evaluate this hypothesis, for each target language, we select the languages to be mixed based on their performance as a few-shot single source language. We fine-tune a multilingual encoder model (i.e., XLM-R). We take the best-performing source languages for each target language on the target validation set. We take the best 3 **(top-3)** and best 5 **(top-5)** languages.

Geographical Location We hypothesize that language proximity could be a good criterion for selecting source languages. In addition, we also verify the performance of the opposite strategy, selecting languages that are farthest from each other. Each language is part of only a single group, except for Indonesian, which has a high overlap with the two groups. We use the label **close-geo** for close languages and **far-geo** for distant languages based on the geographical location.

Language Roots We create two sets of languages based on their linguistic roots: languages belonging to the same language group, that we denote as **related-lang**, and all other languages being dissimilar from each other, denoted as **unrelated-lang**.

3 Experimental Setup

3.1 Data

We use the NusaX dataset (Winata et al., 2022), a parallel multilingual sentiment analysis dataset containing labeled data in 10 low-resource languages and their corresponding translations in English and Indonesian. The list of the languages can be found in Tab. 2 along with their language root and geographical location of the main body of speakers of the language. We highlight that 8 out of the 12 languages are not covered in pre-training by any of the three widely-used multilingual LMs that we considered. In this study, we are interested in quantifying the extent to which multilingual models generalize across languages. Given that the NusaX dataset is built from translating the original data to all languages, we expect there is little to no semantic drift across languages. The dataset contains 500 training, 100 validation, and 400 test samples for each language.

3.2 Single Source Settings

Dataset Size We explore the impact of dataset size on the performance of within languages and cross-lingual transfer. We sample the dataset for *k-shot* training setups where $k \in \{0, 3, 6, 15, 24, 30, 500\}$. For $k < 500$, the samples are created with the same number of examples for each of the three labels. When $k = 500$, this is effectively training on all samples for the source language available. Tab. 3 shows the label distribution of the dataset.

Same Language Setting We conduct experiments where we use the training data from the same language as the target language.

Cross-lingual Transfer We conduct further experiments where we use training data from an **Oracle Source** language in a cross-lingual setting. This is determined, for each target language, as the source language with the best performance on the test set. We note this is an upper bound, given that in a realistic setting, we do not have access to test data to infer the best language.

Impact of Model Architecture As discussed in §2.1, we consider three multilingual LMs of different architecture types: XLM-R as an encoder

Target Lang.	Same language				Cross-lingual (oracle source)			
	XGLM (IC)	XGLM (FT)	mT5 (FT)	XLM-R (FT)	XGLM (IC)	XGLM (FT)	mT5 (FT)	XLM-R (FT)
Unseen Languages								
Acehnese	48.80	60.42	48.00	63.83	46.87	60.67	53.17	65.04
Balinese	45.03	57.33	54.08	63.61	50.68	61.83	55.50	68.39
Banjarese	40.44	68.17	48.83	68.03	53.89	65.83	59.92	74.77
Buginese	39.75	38.58	47.75	57.03	39.25	48.92	50.42	53.83
Madurese	45.18	51.08	45.17	58.74	47.29	59.25	55.08	64.91
Minangkabau	53.93	62.75	38.00	72.63	51.35	62.83	58.58	69.71
Ngaju	44.38	54.25	49.17	63.15	47.72	60.42	54.00	68.29
Toba Batak	37.06	41.67	42.75	51.59	44.06	53.92	48.83	54.42
avg.	44.32	54.28	46.72	**62.33**	47.64	59.21	54.44	**64.92**
Seen Languages								
English	58.02	76.67	57.33	78.07	53.80	70.58	72.83	70.43
Indonesian	56.64	78.33	71.50	73.07	56.17	75.92	62.25	75.83
Javanese	53.55	58.58	49.75	66.57	49.74	63.67	64.00	72.41
Sundanese	41.82	53.50	58.42	61.80	49.42	60.75	61.42	74.43
avg.	52.51	66.77	59.25	**69.88**	52.28	67.73	65.13	**73.28**

Table 4: Results on 30-shots on monolingual and cross-lingual transfer. In oracle source, we report the best source language for each target language. IC and FT denote in-context learning and fine-tuning, respectively. XGLM, mT5, and XLM-R refer to XGLM-2.9B, mT5-3.7B, and XLM-R_{LARGE} (550M), respectively.

Target Lang.	Single-source		Multi-source						
	mono	x-oracle	random-mix	top-3	top-5	close-geo	far-geo	related-lang	unrelated-lang
Unseen Languages									
Acehnese	63.83	65.04	55.83	58.41	58.35	46.62	52.32	54.25	55.00
Balinese	63.61	68.39	58.38	60.60	63.15	56.53	48.92	n/a	58.38
Banjarese	68.03	74.77	61.42	52.75	66.81	55.00	57.13	59.09	57.44
Buginese	57.03	53.83	37.37	44.60	50.33	n/a	n/a	n/a	37.37
Madurese	58.74	64.91	50.29	53.02	59.58	55.53	55.76	n/a	50.29
Minangkabau	72.63	69.71	58.40	53.33	60.50	54.75	60.74	62.23	59.93
Ngaju	63.15	68.29	50.90	48.00	57.28	59.70	49.73	n/a	50.90
Toba Batak	51.59	54.42	41.51	43.26	53.23	43.96	46.94	n/a	41.51
avg.	62.33	**64.92**	51.76	51.75	**58.65**	53.16*	53.08*	58.52*	51.35
Seen Languages									
English	78.07	70.43	35.39	49.60	57.03	n/a	n/a	n/a	35.39
Indonesian	73.07	75.83	49.86	58.07	68.39	51.46	53.85	45.43	53.74
Javanese	66.57	72.41	44.92	61.21	60.90	44.82	41.28	n/a	44.92
Sundanese	61.81	74.43	52.98	50.09	62.43	57.47	43.76	n/a	52.98
avg.	69.88	**73.28**	45.79	54.74	**62.19**	51.25*	46.30*	45.43*	46.76

Table 5: Results on 30-shots with multi-source cross-lingual mixing strategies via few-shot encoder-based fine-tuning using XLM-R_{LARGE}. Results marked with * are not directly comparable due to some results being n/a.

model, mT5 as an encoder-decoder model, and XGLM as a decoder model, and evaluate these models to determine which is most effective at cross-lingual transfer learning. Specifically, we consider the pre-trained versions $XGLM_{2.9B}$, XLM-$R_{0.5B}$, and $mT5_{3.7B}$ respectively.

Training Strategy We train models using in-context learning, prompt-based fine-tuning, and encoder-based model fine-tuning as described in §2.2 as different training strategies are afforded by each model architecture. XGLM is trained using both in-context learning and prompt-based fine-tuning. We note that XGLM cannot be trained with in-context learning with $k > 30$ as we are limited by the maximum sequence length of the positional embeddings. mT5 is trained with prompt-based fine-tuning. Finally, XLM-R is trained using encoder-based model fine-tuning.

Zero-Shot Cross-Task Finally, Winata et al. (2021) introduce zero-shot cross-lingual learning

with BERT fine-tuned on natural language entailment. Given a fine-tuned XLM-R with an entailment head θ_{TE}, a test sample as query Q, and all possible labels $\mathcal{Y}$. The model accepts two inputs, the query Q and label $y' \in \mathcal{Y}$, and generates the entailment score given any combinations of the hypothesis and label $P_\theta(y = \texttt{entail}|h, l)$:

$$\hat{y} = \arg\max_{y' \in \mathcal{Y}} P(y = \texttt{entail}|Q, y', \theta_{\text{TE}}) \quad (3)$$

We consider a zero-shot setup as cross-lingual as no real source language label was used.

3.3 Multi-Source Settings

Random Mixing As a baseline, we randomly mix the samples across different languages and we show the distribution of random mixing accumulated from three random seeds in Fig. 2.

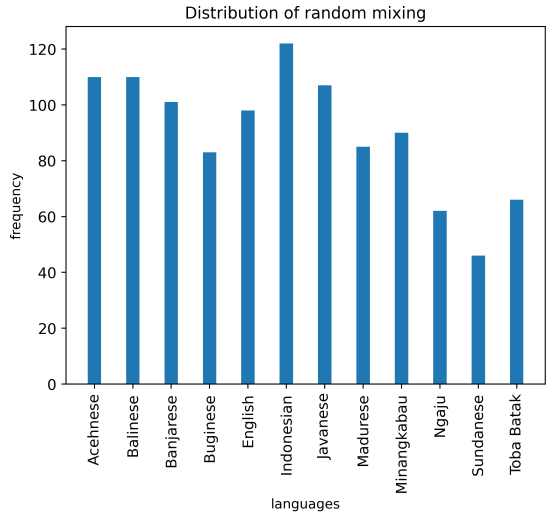

Figure 2: Language Distribution of Random Mixing

Geographical Location To evaluate this strategy, we form 5 groups of languages based on the geographical region as follows:

- **Sumatera Region**: Acehnese, Indonesian, Minangkabau, Toba Batak
- **Java Region**: Balinese, Javanese, Indonesian, Madurese, Sundanese
- **Kalimantan/Borneo Region**: Banjarese, Ngaju
- **Sulawesi Region**: Buginese
- **Non-regional**: English

Language Roots We look at grouping source languages based on their linguistic roots as described in Winata et al. (2022). Resulting in a grouping of Acehnese, Banjarese, Indonesian, and Minangkabau as related languages and all other languages as unrelated languages.

3.4 Label Translation

The labels in the NusaX dataset are in English. We explore the impact of translating labels to the target language. We choose a **seen** language, Indonesian, and an **unseen** language, Balinese, as our two target languages. The labels are translated by native speakers. The goal of this experiment is to assess whether the generative models can gain performance from leveraging semantic knowledge from the labels translated to the target language. We use the following translations for the labels of "positive", "negative" and "neutral" in the same order:

- **Indonesian**: positif, negatif, netral.
- **Balinese**: becik, jele, sedeng.

For Balinese, the native speaker was not able to identify a literal word-to-word translation for the labels and thus suggested words that, in their view, are closely related to the English labels.

3.5 Hyperparameters

All our experiments are reported across 3 runs with fixed seeds {42, 52, 62} for reproducibility, and we report error bars in figures to facilitate transparency. For fine-tuning using XLM-R, we use a batch size of 32, a learning rate of 1e-5, and a learning rate decay of 0.9. We apply early stopping with patience of 5. For XGLM and mT5 fine-tuning, we fine-tune the model with a constant learning rate of 1e-5. The batch size for XGLM and mT5 is 4 and 32, respectively. For XGLM, we fine-tune for 3 epochs when $k = 500$ and 6 epochs when $k = 30$. For mT5, we fine-tune for 24 epochs when $k = 500$ and 48 epochs when $k = 30$, keeping the same number of gradient updates as XGLM. Additionally, we use learning rate of 1e-4 for mT5 when $k = 30$. Due to the large model size, we use mixed precision and DeepSpeed (Rasley et al., 2020) for training. We utilize one V100 32GB GPU for XLM-R and two GPUs for XGLM and mT5.

4 Results

4.1 Single-Source Transfer

Fig. 1 plots the results of different models and training setups with varying amounts of training data. We observe a consistent trend in the same language than in the cross-lingual setting: in the extreme few shot setting, less than 15 examples, fine-tuning and in-context learning show comparable performance, although error bars for in-context learning show a large variance, a well-documented fact in recent

Figure 3: Relation between cross-lingual transfer and vocabulary overlap of different models.

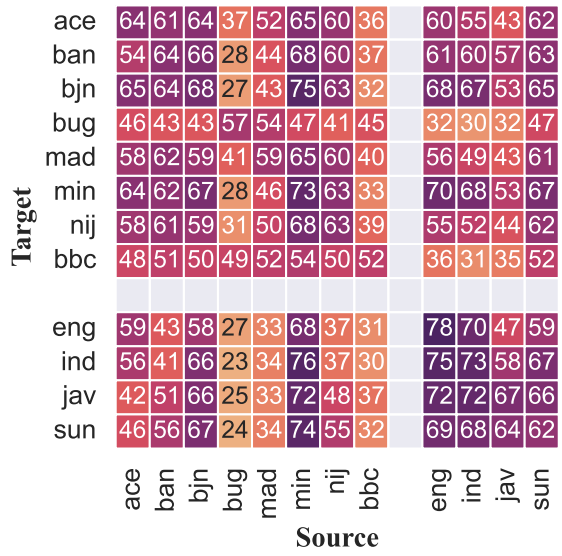

Figure 4: Results on 30-shot cross-lingual fine-tuning in the sentiment analysis task with XLM-R_{LARGE}. We separate seen and unseen languages with a clear row and column.

Source \ Target	**Indonesian (ind)** `l=eng`	`l=ind`	**Balinese (ban)** `l=eng`	`l=ban`
Unseen Languages				
Acehnese	62.08	**69.19**	**61.50**	43.55
Balinese	59.58	**65.54**	**57.33**	38.97
Banjarese	68.83	**73.65**	**59.83**	48.27
Buginese	42.42	**68.98**	32.00	**39.50**
Madurese	62.08	**70.53**	**50.75**	32.44
Minangkabau	72.42	**73.77**	**61.83**	49.36
Ngaju	62.33	**63.17**	**53.25**	39.54
Toba Batak	51.08	**59.55**	**47.75**	25.72
Seen Languages				
English	**75.92**	63.24	**53.83**	43.60
Indonesian	**78.33**	73.95	51.42	**54.34**
Javanese	**71.75**	68.25	**56.33**	46.57
Sundanese	**71.83**	69.43	**54.58**	34.92

Table 6: Single-source fine-tuning results with translated labels using XGLM. `l=ind` and `l=ban` denotes the labels are translated to Indonesian and Balinese, respectively.

work (Brown et al., 2020). As more labeled data becomes available, the best strategy is to use fine-tuning. Surprisingly, in the cross-lingual setting, the XLM-R cross-task baseline gives a very strong performance and seems like a better alternative in the case of having less than 15 labeled examples. As expected, when using all available training data, fine-tuning performs best. However, in smaller data regimes, XLM-R is the best approach.

Tab. 4 provides a window into the performance metrics in the 30-shot setting across all model architectures. We observe that XLM-R fine-tuning outperforms all other models by a considerable margin, both across unseen languages and seen languages. This demonstrates that fine-tuning methods leveraging an encoder-based model are the most effective at cross-lingual transfer for this task while having five times fewer parameters. In Fig. 3 we illustrate how token overlap correlates with model performance for unseen languages as the source. There is one very clear trend in these results: when the target language has not been seen by the model during pre-training, it is beneficial to choose a source

Source \ Target	Indonesian (ind)		Balinese (ban)	
	l=eng	l=ind	l=eng	l=ban
Unseen Languages				
acehnese	55.18	26.91	50.21	28.46
balinese	41.16	34.01	45.03	25.84
banjarese	44.96	23.60	38.08	28.07
buginese	52.02	30.39	46.63	24.68
madurese	51.53	29.77	43.29	28.04
minangkabau	55.70	31.71	50.10	26.04
ngaju	44.44	22.41	44.94	30.52
toba batak	41.95	30.73	40.40	25.23
avg.	**48.37**	28.69	**44.84**	27.11
Seen Languages				
english	49.85	19.98	37.41	33.41
indonesian	56.64	24.38	41.07	23.84
javanese	54.41	31.74	50.68	32.78
sundanese	56.17	21.82	49.70	29.01
avg.	**54.27**	24.48	**44.72**	29.76

Table 7: Single-source in-context learning results with translated labels using XGLM. `l=ind` and `l=ban` denotes the labels are translated to Indonesian and Balinese, respectively.

language with high token overlap with the target language.

Label Translation We evaluate the effect of translated labels from English to target languages (Indonesian and Balinese) in the text-to-text framework. We use the label translations as described in §3.4 and Tab. 7 to translate the labels to target languages for each source language in the prompt-based fine-tuning and in-context learning, respectively. We use XGLM for our experiment as this supports both paradigms. For Indonesian, we observe that translated labels lead to significant improvement when source languages are unseen. However, these labels do not improve the performance when source languages are seen. As for Balinese, the translated labels lead to consistently worse performance, likely due to there not being direct translations for these labels in this language. This suggests more attention is needed when translating labels into target languages, and future work could consider cross-lingual transfer when the labels are in the corresponding languages instead of English.

4.2 Multi-Source Transfer

Fig. 4 shows that there could be more than a single good source language for a given target seen or unseen language. Moreover, as shown in Tab. 4, in many cases, the oracle source language outperforms using the target language as the source. One plausible explanation for why training on a source language can benefit a different target language could be its token overlap. Therefore, we perform experiments to explore the effectiveness of using multiple-source languages for cross-lingual transfer. We employ various multi-source language selection techniques as described in §2.3. In addition, we conduct experiments using XGLM in-context learning (Tab. 8) and XLM-R fine-tuning.

Tab. 5 shows the performance of the various language selection techniques when fine-tuning with XLM-R. We add “mono" (same language) and “x-oracle" (cross-lingual oracle source) as ceilings to compare against. We find that a nuanced selection of the source languages to mix is essential in obtaining competitive performance. We see that when randomly mixing all source languages or choosing languages that are unrelated linguistically to the target language, we obtain the worst performance in both seen and unseen languages. One challenge when using expert knowledge to select source languages such as geographical closeness or linguistic similarity is that there can be null sets for a given target language, denoted as n/a in Tab. 5. We observe that when these methods are applicable, they are effective techniques, obtaining performance that is largely better than random.

We propose to use the validation set to find the top-k most transferable source languages and use these for multi-source mixing. Here we find that when we add more languages to the mix based on this metric, performance improves. More concretely, using the top-5 transferable source languages for mixing is more effective than using the top-3. This is also a practical method as it induces some form of selection across languages but also scales to many languages in the source without needing detailed information about the language itself. Finally, we also observe that when using the top-5 mixing strategy, the gains compared to random are much more pronounced in the seen languages as compared to the unseen languages, as might be expected.

We also explore using constraints, such as forcing at least one example per label for any selected source language, with and without language replacement for language choice. However, we did not see noticeable trends and omitted these for brevity. In Fig. 3, we do not find significant differences in subword overlap between languages

and rule this out as an underlying cause for better source language performance.

5 Related Work

Language-Specific LM Self-supervised pre-trained LM methodologies leverage unlabeled data on low-resource languages (e.g., in French (Martin et al., 2020; Le et al., 2020), Indian languages (Kakwani et al., 2020), Indonesian (Wilie et al., 2020; Koto et al., 2020; Cahyawijaya et al., 2021), Korean (Park et al., 2021), Chinese (Xu et al., 2020), Italian (Polignano et al., 2019)). This has enabled transfer learning to low-resource languages. Another line of work is to train large multilingual languages models by taking hundreds of languages (e.g., mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020a), XGLM (Lin et al., 2021)). These models enable cross-lingual transfer when there are very limited in-language training samples available.

Cross-lingual Transfer The effectiveness of cross-language transfer with multilingual LMs has been extensively studied, focusing on languages that are seen during pre-training. Cross-lingual transfer learning has been applied to various downstream NLP and multimodal tasks, such as natural language understanding (Liu et al., 2019, 2020; Winata et al., 2021), named entity recognition (Liu et al., 2021a), textual entailment (Artetxe and Schwenk, 2019), entity linking (Rijhwani et al., 2019), hate speech detection (Nozza, 2021; Pamungkas et al., 2021), machine translation (Eriguchi et al., 2018), question answering (Zhou et al., 2021; Faisal and Anastasopoulos, 2021; Limkonchotiwat et al., 2022; Agarwal et al., 2022; Zhang and Wan, 2022), part-of-speech tagging (Wu and Dredze, 2019b; Ansell et al., 2021; Parović et al., 2022), sentiment analysis (Fei and Li, 2020; Ghasemi et al., 2022), text-to-image search (Huang et al., 2021), and information retrieval (Yarmohammadi et al., 2021). Malkin et al. (2022) show the effect of pre-trained language selection on the zero-shot setting by limiting the distribution of pre-trained data size to be balanced across all languages. Winata et al. (2021) conduct the first exploration on using English LM for cross-lingual transfer via in-context learning. For languages that are unseen during pre-training, Adelani et al. (2021) and Ebrahimi et al. (2022) explore the effectiveness of cross-lingual transfer in African and American languages, respectively. They found that fine-tuning the multilingual encoder model is an effective method for adapting to new languages. The difference between our study and theirs is we conducted a structured study on how to leverage the pre-trained LM in few-shot settings with various LM architectures (i.e., encoder and generative models). In another line of work, using more complex sampling strategies for few-shot multilingual transfer outperforms the random sampling (Kumar et al., 2022). Conneau et al. (2020b) explore factors on why multilingual models are effective for cross-lingual transfer.

6 Conclusion

We present the first comprehensive study to measure the effectiveness of few-shot in-context learning and fine-tuning approaches with multilingual LMs on languages that have never been seen during pre-training. We investigate the effectiveness of utilizing few-shot examples and present strategies and insights depending on the amount of labeled training data available. We find that fine-tuning the multilingual encoder model (i.e., XLM-R) is generally the most effective method when we have more than 15 samples; otherwise, zero-shot cross-task is preferable. We also observe that in-context learning has a relatively higher variance than fine-tuning, and mixing multiple source languages is a promising approach when the number of training examples in each language is limited.

Limitations

In this work, we only choose pre-trained models that are fit on maximum two V100 32GB GPUs for fine-tuning. To ensure the comparisons are fair, we choose generative models (i.e., XGLM and mT5) with similar sizes. It is possible to gain higher performance if we choose larger models and we leave this for future investigation.

Ethical Consideration

We didn't find any significant harms in applying in-context learning and fine-tuning on cross-lingual few-shot training. The methods we explore are general-purpose methods for low-resource language adaptation.

Acknowledgements

We are grateful to Shuyi Wang for their feedback on a draft of this manuscript. We sincerely thank the three anonymous reviewers for their insightful comments on our paper.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Sumit Agarwal, Suraj Tripathi, Teruko Mitamura, and Carolyn Rose. 2022. Zero-shot cross-lingual open domain question answering. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 91–99.

Alham Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. Mad-g: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781.

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

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, et al. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.

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

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Fahim Faisal and Antonios Anastasopoulos. 2021. Investigating post-pretraining representation alignment for cross-lingual question answering. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 133–148.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771.

Rouzbeh Ghasemi, Seyed Arad Ashrafi Asli, and Saeedeh Momtazi. 2022. Deep persian sentiment analysis: Cross-lingual training for low-resource languages. *Journal of Information Science*, 48(4):449–462.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander G Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2443–2459.

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.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770.

Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. "diversity and uncertainty in moderation" are the key to data selection for multilingual few-shot transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1042–1055, Seattle, United States. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020a. 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.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020b. 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.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Cl-relkt: Cross-lingual language knowledge transfer for multilingual retrieval question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2141–2155.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021a. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. 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 5834–5846.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. 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 1297–1303.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2021c. X2parser: Cross-lingual and cross-domain framework for task-oriented compositional semantic parsing. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 112–127.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 907–914.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. Bad-x: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

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.

Marco Polignano, Valerio Basile, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. Alberto: Modeling italian social media language with bert. *IJCoL. Italian Journal of Computational Linguistics*, 5(5-2):11–31.

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

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6924–6931.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, et al. 2022. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. *arXiv preprint arXiv:2205.15960*.

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.

Shijie Wu and Mark Dredze. 2019a. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019b. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772.

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.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, et al. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967.

Yunxiang Zhang and Xiaojun Wan. 2022. Birdqa: A bilingual dataset for question answering on tricky riddles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11748–11756.

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.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. 2021. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834.

A In-context Learning Results

We show detailed results of in-context learning with various multi-source mixing strategies in Tab. 8. In general, **random-mix** strategy outperforms other mixing strategies. This finding does not apply to few-shot fine-tuning experiments, where **random-mix** achieves worse performance compared to selecting top-k languages.

Target Lang.	random-mix	top-3	top-5	close geo	far geo	related lang.	unrelated lang.
Unseen Languages							
acehnese	57.03	46.68	34.19	41.47	48.33	41.09	37.77
balinese	58.52	44.83	45.10	47.72	49.45	n/a	58.52
banjarese	62.13	37.61	50.89	46.30	29.60	45.30	42.43
buginese	35.88	33.00	36.52	n/a	35.88	n/a	35.88
madurese	42.41	27.16	37.20	42.45	47.06	n/a	42.41
minangkabau	50.69	42.23	50.66	35.79	52.02	35.34	41.89
ngaju	46.96	30.68	35.54	35.37	25.41	n/a	46.96
toba batak	46.70	41.33	39.24	37.82	40.42	n/a	46.70
avg.	**50.04**	37.94	41.17	40.99	41.02	40.58	44.07
Seen Languages							
english	41.61	34.31	47.47	n/a	41.61	n/a	41.61
indonesian	53.58	45.87	60.44	49.66	49.10	43.78	48.63
javanese	51.95	45.44	45.69	46.33	53.16	n/a	51.95
sundanese	50.80	37.99	43.55	37.60	49.87	n/a	50.80
avg.	**49.49**	40.90	49.29	44.53	48.44	43.78	48.25

Table 8: Results on 30-shots with multi-source cross-lingual mixing strategies via in-context learning.