Investigating the Potential of Task Arithmetic for Cross-Lingual Transfer

Marinela Parović Ivan Vulić Anna Korhonen
Language Technology Lab, University of Cambridge
{mp939,iv250,alk23}@cam.ac.uk

Abstract

Cross-lingual transfer has recently been tackled through modular, parameter-efficient fine-tuning methods which allow arbitrary combinations of language and task modules for transfer of any task to any language. Concurrently, task arithmetic has emerged as a powerful and modular tool for editing pretrained models using multiple full fine-tunings. In this work, we connect the paradigms of task arithmetic and cross-lingual transfer, demonstrating that modularity for cross-lingual transfer can be achieved even with full model fine-tuning. Our approach displays strong performance on a range of multilingual benchmarks encompassing both high-resource and low-resource languages.

1 Introduction

Massively multilingual Transformer-based models (MMTs) (Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2021, 2022; He et al., 2023) have shown impressive performance in cross-lingual transfer due to their ability to learn representations which have a degree of cross-lingual alignment, despite being trained using purely unsupervised objectives (e.g., masked language modeling). This allows an MMT to perform a task in a *target* language having seen labeled data only in a *source* language: the so-called zero-shot cross-lingual transfer (ZS-XLT).

The adaptation of MMTs to low-resource languages has been an attractive research area lately, stemming from a need to extend such models to under-represented and unseen languages (Wang et al., 2020; Muller et al., 2021; Ebrahimi and Kann, 2021). A particularly popular approach is based on modular and parameter-efficient (PEFT) adaptation of MMTs to particular languages and tasks, generally leading to improved ZS-XLT (Pfeiffer et al., 2020; Üstün et al., 2020; Parović et al., 2022; Ansell et al., 2022; Parovic et al., 2023).

While the PEFT methods are typically designed to facilitate modularity and module (re)combination, full fine-tuning appears to exhibit less flexibility in this regard. This has led to the development of techniques for merging multiple fine-tuned models (Wortsman et al., 2022; Matena and Raffel, 2022; Schmidt et al., 2023). One prominent approach to model merging is based on the so-called *task arithmetic*: Ilharco et al. (2023) have proposed editing monolingual and vision models using task vectors derived by subtracting the weights of the pretrained model from those of the *fully fine-tuned* model. Several such vectors can then be applied to the model through arithmetic operations such as addition and subtraction to steer its behaviour in a controlled manner (Daheim et al., 2023a,b).

In this work, we shed new light on the ability to maintain modularity even for fully fine-tuned MMTs in the context of ZS-XLT: we delve into the potential of full fine-tuning and task arithmetic for ZS-XLT. More precisely, starting from a pretrained MMT, we independently acquire language and task vectors, by fine-tuning the MMT on the language and task data, respectively. These vectors are subsequently combined with MMT through addition or subtraction to obtain the resulting, adapted model tailored for a specific language-task pair in a fully modular fashion. We extensively evaluate several promising variants of combining task and language vectors across a spectrum of multilingual benchmarks, encompassing both high-resource and low-resource languages. Our findings underscore the potency of task arithmetic for cross-lingual transfer and language adaptation, yielding notable performance gains over fully fine-tuned MMTs without task arithmetic and other strong ZS-XLT baselines, particularly prominent on benchmarks featuring low-resource languages. Our code is available at `https://github.com/parovicm/task-arithmetic`.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 124–137
March 17-22, 2024

2 Methodology

Background and Motivation. Prior work demonstrated that models which share a portion of the optimization path, typically through a common initialization, can be merged into a single model using weight interpolation while maintaining task accuracy (Ilharco et al., 2022; Wortsman et al., 2022; Choshen et al., 2022). Gueta et al. (2023) find that models trained on the same data or on different datasets of the same task tend to cluster together in the weight space. Daheim et al. (2023a) leverage the task arithmetic to address the challenges of hallucination within dialogue systems. They additionally employ Fisher information to weigh the importance of the parameters (Sung et al., 2021; Matena and Raffel, 2022) participating in the arithmetic. Inspired by the previous work on model merging in general and task arithmetic in particular, here we investigate its potential and benefits for modular ZS-XLT.

Task Arithmetic: Preliminaries. Given a pretrained model with the parameters $\theta_0 \in \mathbb{R}^d$ and the designated task T, the task-specific parameters $\theta_T \in \mathbb{R}^d$ can be derived by fine-tuning the pretrained model on T's task data. The task vector of T, denoted by $\boldsymbol{\tau}^T \in \mathbb{R}^d$, is defined as the difference in parameters before and after fine-tuning: $\boldsymbol{\tau}^T = \theta_T - \theta_0$. This vector characterizes the direction in the model's weight space, such that adjusting the parameters in this direction enhances task performance.

The acquired task vector can be integrated into the model by a simple addition and an optional scaling factor $\lambda \in \mathbb{R}$ governing its influence, yielding a new model with the following parameters:

$$\theta' = \theta_0 + \lambda \cdot \boldsymbol{\tau}^T. \tag{1}$$

Note that when $\lambda = 1$, then $\theta' = \theta_T$. Adding a task vector ($\lambda > 0$) has the effect of promoting a certain 'model behaviour', while subtracting it ($\lambda < 0$) 'suppresses' it. In a more general scenario, given n task vectors $\boldsymbol{\tau}^{T_1}, ..., \boldsymbol{\tau}^{T_n} \in \mathbb{R}^d$ along with their corresponding scaling coefficients $\lambda_{T_1}, ..., \lambda_{T_n} \in \mathbb{R}$, their application to the model yields the following:

$$\theta' = \theta_0 + \sum_{i=1}^{n} \lambda_{T_i} \cdot \boldsymbol{\tau}^{T_i}. \tag{2}$$

2.1 Task Arithmetic for ZS-XLT

Given a source language L_s and a target language L_t, the 'task' vectors associated with these languages (i.e., *language vectors*), $\boldsymbol{\tau}^{L_s}$ and $\boldsymbol{\tau}^{L_t}$, can be obtained by fine-tuning a pretrained MMT on the respective unlabeled data. Furthermore, when presented with a specific task T and its corresponding dataset in the source language L_s, we can derive the task vector $\boldsymbol{\tau}^T$ by fine-tuning the model for task T. Then, the core idea is that the model designed to address the task T in the target language L_t can be formed through the arithmetic of the task vector $\boldsymbol{\tau}^T$ and the language vectors $\boldsymbol{\tau}^{L_s}$ and $\boldsymbol{\tau}^{L_t}$. There are multiple possible configurations based on addition and subtraction of the vectors; we motivate and describe those configurations in what follows.

First, inspired by the task analogy (Ilharco et al., 2023) which is applicable to tasks linked by the relation of the form *"A is to B as C is to D"*, we can define the model for the task T in language L_t as:

$$\theta' = \theta_0 + \lambda_T \cdot \boldsymbol{\tau}^T + \lambda_{L_t} \cdot \boldsymbol{\tau}^{L_t} - \lambda_{L_s} \cdot \boldsymbol{\tau}^{L_s}. \tag{3}$$

We denote this variant as –SRC+TGT.

Further, target language adaptation (without any intervention on the source language) is known to exhibit strong performance in cross-lingual transfer, particularly for low-resource languages (Pfeiffer et al., 2020; Ansell et al., 2022; Ebrahimi et al., 2022; Ansell et al., 2023). Inspired by this, we introduce +TGT variant, where alongside the task vector we only add the target language vector $\boldsymbol{\tau}^{L_t}$. Similarly, +SRC variant is obtained by adding the source language vector $\boldsymbol{\tau}^{L_s}$ only. This variant could be an insufficient adaptation method for low-resource languages, which necessitate target language-informed modelling.

Finally, we propose a variant which adds both $\boldsymbol{\tau}^{L_s}$ and $\boldsymbol{\tau}^{L_t}$ (+SRC+TGT). This variant hinges on the observation that knowledge of the source language is beneficial for a specific source-target transfer direction (Ansell et al., 2022), and subtraction of the source language vector done by the task analogy variant (-SRC+TGT) might suppress this valuable knowledge.

3 Experiments and Results

Tasks and Languages. We extensively evaluate our method on two classification tasks and four different datasets: 1) natural language inference (NLI) with (a) XNLI (Conneau et al., 2018) covering 14 high-resource and mid-resource languages, and (b) AmericasNLI (Ebrahimi et al., 2022) spanning 10 low-resource languages from the Americas; 2) sentiment classification (SA) with MARC (Keung et al., 2020) containing 5 high-resource languages

	MultiNLI	MARC	NusaX
Batch Size	32	32	16
Epochs	5	5	10
Learning Rate	$2 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$2 \cdot 10^{-5}$
Eval Freq. (steps)	625	625	250
Eval Metric	Acc	Acc	F1

Table 1: Hyperparameters with XLM-R$_{\text{BASE}}$.

and NusaX (Winata et al., 2023) consisting of 10 low-resource Indonesian languages. This totals 34 typologically diverse languages with different degrees of available resources.[1]

Pretrained MMT Models. Our primary MMT is XLM-R$_{\text{BASE}}$ (Conneau et al., 2020), and we also run a subset of experiments with XLM-R$_{\text{LARGE}}$.

Language Vectors are trained on unlabelled data of each language, primarily following the hyperparameters outlined in Pfeiffer et al. (2020). Details regarding the used monolingual corpora are provided in Appendix A. We train for $50,000$ steps (20,000 steps with XLM-R$_{\text{LARGE}}$), a batch size is 64, a learning rate is $5 \cdot 10^{-5}$ and a maximum sequence length is set to 256. We select the checkpoint that yields the lowest validation perplexity as the final language vector.

Task Vectors are trained on the corresponding task dataset in the source language (English for XNLI, AmericasNLI, and MARC; Indonesian for NusaX). The dataset used for obtaining the task vector for both XNLI and AmericasNLI is MultiNLI (Williams et al., 2018). Further details about the datasets and tasks are given in Appendix B. The hyperparameters are in Table 1 and Appendix G.[2]

Task-Arithmetic Variants. Our starting point, denoted as MODEL, is the pretrained model *fully fine-tuned* on the data of a particular task T. MODEL is subsequently applied to make predictions on data in different target languages, as in standard ZS-XLT. Further, it is then augmented with different task arithmetic variants discussed in §2.1. For example, +TGT variant outputs language-task specialized models in a modular fashion, by adding the corresponding target language vectors. For all the variants, we evaluate the configurations with different scaling factors for source and target language vectors (λ_{L_s}, λ_{L_t}). Task scaling factor λ_T is always set to 1. In the -SRC+TGT and +SRC+TGT variants, we use $\lambda_{L_s} = \lambda_{L_t}$. Following Ilharco et al. (2023), we consider scaling factors from the set $\{0.1, 0.2, ..., 1.0\}$ and choose the one with the highest average performance on the corresponding validation data. The scaling coefficients reaching the best performance are summarized in Appendix E.

Baselines. Beyond comparing to the fully fine-tuned MODEL in all tasks, we compare our models against two strong ZS-XLT methods: 1) sparse fine-tuning (SFT) for cross-lingual transfer (Ansell et al., 2022) on AmericasNLI and NusaX, and 2) target language-ready (TLR) adapters (Parovic et al., 2023) on AmericasNLI, which both showed superiority over other established ZS-XLT variants with language adaptation such as MAD-X (Pfeiffer et al., 2020) in those tasks.[3] Note that these methods were created with the specific goal of enhancing ZS-XLT performance. Our primary goal, however, is to gain insight into the interaction between the task arithmetic and cross-lingual transfer. The scores of these baselines are inherited from prior work (Parovic et al., 2023; Ansell et al., 2023). We refrained from conducting experiments with these baselines on the XNLI and MARC datasets mainly for the following reasons: 1) these methods are tailored to low-resource languages, and exhibit the highest performance in such contexts, while XNLI and MARC feature high-resource languages; 2) the contributions of this paper do not hinge on direct comparisons with them. Instead, we position the task fine-tuned model as our principal baseline, and our goal lies in highlighting the effectiveness of language and task vector compositions relative to a simple task fine-tuning; 3) it is computationally expensive to train language modules for many languages which is necessary in these baselines.

3.1 Results and Discussion

Main Results. The main results for all tasks, languages, and configurations with XLM-R$_{\text{BASE}}$ are presented in Table 2. We find that task arithmetic can be very effective in improving ZS-XLT performance. For instance, our methods yield per-

[1]We exclude NIJ from our NusaX results since it does not have any unlabelled data available, and thus no language vector was trained for it.

[2]The hyperparameters for NusaX are different due to a significantly smaller training set (MultiNLI has 393k training examples, MARC has 160k, and the training set for NusaX (SMSA) has only 11k examples; see Table 6).

[3]We adhere to their suggested hyperparameters and adopt the strongest, ALL-MULTI variant of the TLR adapters, which is constructed by cycling over the language adapters of 36 languages during task adapter training; see Parovic et al. (2023) for further details.

Method	AR	BG	DE	EL	ES	FR	HI	RU	SW	TH	TR	UR	VI	ZH	avg
MODEL	72.22	77.52	76.55	75.15	78.38	78.08	69.88	75.19	64.45	71.84	72.38	64.91	74.15	73.13	73.13
MODEL + SRC	72.04	78.42	77.31	75.63	79.38	**78.80**	70.60	**76.81**	62.81	72.87	72.71	66.45	75.75	74.85	73.89
MODEL + TGT	72.55	78.22	77.41	76.47	79.86	78.76	**72.87**	76.25	**69.74**	72.42	**74.11**	67.88	76.05	74.51	74.79
MODEL + SRC + TGT	**73.71**	**78.90**	**77.66**	**76.81**	**80.02**	78.76	72.48	76.61	69.28	**73.25**	74.03	**68.56**	**76.61**	**75.57**	**75.16**
MODEL – SRC + TGT	72.24	77.17	76.71	75.11	78.24	78.02	69.90	74.87	66.83	71.78	72.00	65.03	73.99	72.75	73.19

(a) XNLI: accuracy

Method	AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR	avg
TLR ADAPTERS	53.47	42.27	47.73	57.47	41.47	49.73	40.91	58.80	50.27	40.93	48.31
SFT	**58.40**	**44.67**	47.60	**62.27**	**44.40**	**50.81**	**46.39**	60.40	49.47	43.07	**50.75**
MODEL	36.93	39.47	37.60	39.60	36.80	41.73	38.24	37.87	41.47	35.47	38.52
MODEL + SRC	36.67	39.07	38.80	37.87	35.33	41.06	37.03	37.73	40.13	38.27	38.20
MODEL + TGT	54.67	43.33	**48.27**	59.87	41.87	50.41	43.58	**64.93**	48.27	**45.33**	50.05
MODEL + SRC + TGT	46.40	43.33	46.27	56.27	38.67	49.05	40.37	62.53	**50.53**	44.53	47.80
MODEL – SRC + TGT	55.60	41.87	46.67	60.53	42.27	50.41	42.51	62.67	47.87	44.93	49.53

(b) AmericasNLI: accuracy

Method	ACE	BAN	BBC	BJN	BUG	JAV	MAD	MIN	SUN	avg
SFT	79.96	81.26	65.80	82.00	63.84	84.27	73.49	**86.60**	**84.36**	77.95
MODEL	70.84	72.16	47.76	76.88	42.83	81.01	70.34	81.54	78.12	69.05
MODEL + SRC	71.22	74.13	52.68	77.40	51.57	81.31	73.57	81.59	77.50	71.22
MODEL + TGT	81.18	**82.77**	74.22	**85.21**	69.26	**87.10**	75.46	85.66	83.00	**80.43**
MODEL + SRC + TGT	**82.68**	80.98	**77.51**	83.24	65.23	84.64	74.42	84.72	79.89	79.26
MODEL – SRC + TGT	76.24	81.13	73.48	80.30	**70.20**	86.66	**76.67**	86.38	82.63	79.30

(c) NusaX: F1

Table 2: Results of different methods on XNLI, AmericasNLI, and NusaX datasets with XLM-R_{BASE}. The last column is the average score over all languages. **Bold**: the best performing approach.

SF	XNLI	AmericasNLI	MARC	NusaX
0.1	73.88	39.66	78.93	74.11
0.2	74.51	40.02	**79.00**	74.86
0.3	74.89	40.47	78.95	76.31
0.4	74.85	42.51	78.85	78.80
0.5	**74.91**	44.57	78.55	79.85
0.6	74.66	46.57	78.15	80.04
0.7	74.07	48.08	77.74	**81.10**
0.8	72.88	**49.21**	77.28	79.91
0.9	70.96	48.58	76.64	79.92
1.0	68.50	47.78	76.10	79.13

Table 3: Effect of different scaling factors on the XLM-R_{BASE} performance with the +SRC+TGT variant. All scores are obtained on the validation sets; SF=Scaling Factor.

Method	AmericasNLI	NusaX
MODEL	40.25	74.17
MODEL + SRC	40.38	75.36
MODEL + TGT	**52.46**	**83.43**
MODEL + SRC + TGT	51.36	80.30
MODEL - SRC + TGT	51.91	81.06

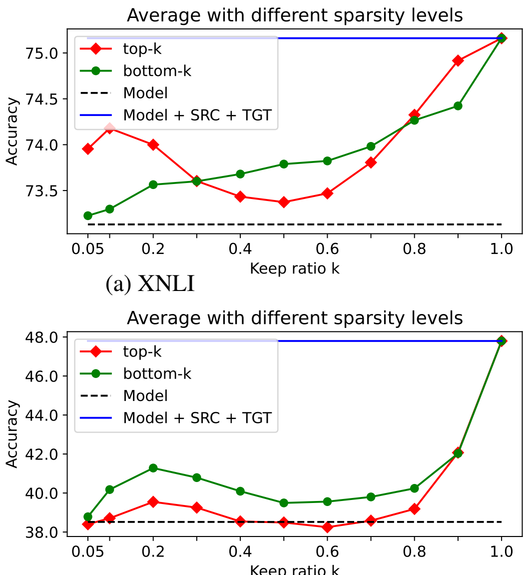
Table 4: Results with XLM-R_{LARGE}, averaged over languages. Full results are given in Appendix D.

formance gains ranging from 2 points on XNLI, with some gains observed even for high-resource languages such as Spanish and German, up to a substantial increase of 12 points on AmericasNLI and NusaX over MODEL.[4]

Low-Resource Languages in particular greatly benefit from language adaptation, as established in prior work (Pfeiffer et al., 2020; Ansell et al., 2021; Parovic et al., 2023; Ansell et al., 2023). Our results substantiate these trends. For instance, two of the low-resource languages in XNLI, SW and UR, meet gains of up to 4-5% while the remaining languages experience more moderate increases of ∼1-2%. This effect is more notably present on the two low-resource benchmarks, AmericasNLI and NusaX. There, the addition of the target language vectors results in an average gain of 12 points with +TGT variant, which outperforms other variants. Conversely, augmenting the model with the source language vectors leads to a performance improvement of 2 points on NusaX, while its impact on AmericasNLI is negligible. Similar trends are also observed with XLM-R_{LARGE} as the underlying model; cf., Table 4. This reaffirms that source language adaptation is insufficient in the context of low-resource languages.

Task Analogies. Our results reveal that the –SRC+TGT variant, which draws inspiration from

[4]The gains on the MARC dataset are relatively modest, which could be attributed to the nature of the task itself coupled with the high-resource nature of its target languages. We thus present the results on MARC in Appendix C.



(a) XNLI

(b) AmericasNLI

Figure 1: Averaged scores with different sparsity levels with the +SRC+TGT variant.

task analogies, lags slightly behind the best-performing variant on all tasks. While the exact reason behind this is unclear, we suspect it might be due to a different nature of language adaptation as opposed to other task or domain fine-tunings. Additionally, and as pointed out in §2.1, the knowledge of the source language is valuable for ZS-XLT (Ansell et al., 2022), while subtraction of the source language vector may suppress it.

Task Arithmetic vs Baselines. Interestingly, the proposed task arithmetic-based approach to ZS-XLT displays very competitive and even improved performance when compared against two state-of-the-art ZS-XLT methods: e.g., our most effective variant on AmericasNLI, MODEL + TGT, achieves 0.7% lower performance than SFTs and 1.7% higher than TLR adapters. Moreover, it outperforms SFTs by around 2.5% on the NusaX dataset. While the two techniques have been trained with different hyperparameter configurations, these results hold promise and warrant further in-depth exploration of task arithmetic in this particular context.

Effect of Scaling Factors. Our results reveal that scaling factors associated with language vectors have a significant impact on performance. Table 3 shows the scores on the validation sets of all datasets with different scaling factors attained with the +SRC+TGT variant. The observed variance in these scores could pose challenges in the wider application of task arithmetic for ZS-XLT, necessitating further investigation.

Analysis of Sparsity. In prior work, Ansell et al. (2022) elucidate that the right level of sparsity serves as a pivotal factor enabling both performance gains and modularity of SFTs. This is attributed to sparsity minimizing the parameter overlap between different fine-tunings; their analysis reveals a strong performance drop when the density level exceeds 30%, possibly due to interference during composition. Yadav et al. (2023) propose strategies to improve task arithmetic in the multi-task learning context, aiming to mitigate interference between different task vectors. They find that retaining only the top 20% of parameters with the highest magnitudes within a task vector does not result in performance degradation. Drawing inspiration from these works, we assess the effect of sparsity on the language vectors. Focusing on the +TGT and +SRC+TGT variants, we vary the proportion of kept parameters k from 5% to 90%, where we keep the parameters with largest magnitudes within the task vectors (*top-k*). As an ablation, we also present the scores obtained by keeping the k% parameters with the lowest magnitudes (*bottom-k*).

The plots on XNLI and AmericasNLI with +SRC+TGT are provided in Figure 1, with more results for other tasks and variants available in Appendix F. A general trend suggests that imposing higher degrees of sparsity is somewhat more detrimental for AmericasNLI. Retaining even 90% of parameters incurs a substantial drop of around $\sim 6\%$ on this dataset, as evident in both top-k and bottom-k variants. Notably, the top-k plots for both tasks suggest that the intermediate sparsity levels yield inferior performance, with some degree of recovery observed towards the higher sparsity end. This observation prompts further investigation on the interaction of sparsity levels and modularity of task arithmetic in cross-lingual transfer scenarios.

4 Conclusion

We proposed the adoption of task arithmetic in the context of zero-shot cross-lingual transfer, investigating its potential for these transfer scenarios. Our approach involves independently creating and combining language and task vectors to attain models customized for specific language-task pairings. We empirically demonstrated the effectiveness of this technique across various multilingual benchmarks.

Limitations

As a short paper, this work is organically constrained by its content page constraints, which substantially impacts the extent and depth of the experiments and analysis. Keeping that in mind, we list some limitations of this work and outline several promising directions which could be explored as part of future work, but are out of scope of this particular project.

Due to a large number of languages and methods, we report all our results based on a single run. However, the large number of target languages and tasks we average over and the replication of the core findings with two MMTs enhances the confidence in the their correctness.

While in this work we consider encoder-only language models, our methodology can be readily applied for cross-lingual transfer with different model types, e.g., encoder-decoder models fine-tuned in a text-to-text fashion or through instruction tuning (Xue et al., 2021, 2022; Chung et al., 2022). Moreover, the proposed approach could also be applied to and evaluated in few-shot cross-lingual transfer scenarios (Lauscher et al., 2020; Ansell et al., 2023), which assume access to a small amount of supervised data in the target language. Ruder et al. (2023) introduce a benchmark XTREME-UP for few-shot learning and experiment with multilingual fine-tuning and in-language in-context learning to showcase the potency of large language models in understanding under-represented languages. Additionally, Asai et al. (2023) introduce BUFFET, another benchmark for few-shot learning in the cross-lingual transfer with all tasks cast into a text-to-text format. Future work could use our approach in synergy with these methods and benchmarks. Our core findings should hold regardless of the chosen model and cross-lingual transfer protocol.

We currently apply equal weighting to all parameters within the task and language vectors. However, the importance of individual parameters could vary depending on a task or language. Developing methods for more nuanced, per-parameter weighting is a potential avenue for future work. Prior work has proposed the Fisher information matrix to select (Sung et al., 2021) or weigh (Matena and Raffel, 2022; Daheim et al., 2023a) parameters effectively. Our preliminary results did not show significant gains with Fisher weighting, but this aspect could benefit from further exploration.

Finally, off-the-shelf application of sparsity on the language vectors has not been particularly effective. In order for it to outperform full language vectors, a more refined approach might be necessary. This could involve some form of re-training which would result in an approach akin to sparse fine-tuning (SFTs) (Ansell et al., 2022, 2024), or implementing a more sophisticated parameter selection mechanism beyond magnitude-based methods.

Acknowledgments

Marinela Parović is supported by Trinity College External Research Studentship. Ivan Vulić acknowledges the support of a personal Royal Society University Research Fellowship *'Inclusive and Sustainable Language Technology for a Truly Multilingual World'* (no 221137; 2022–). The work is also supported by the UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 (the UK government's funding guarantee for ERC Advanced Grants) awarded to Anna Korhonen at the University of Cambridge.

We thank Alan Ansell and Edoardo Maria Ponti for helpful feedback and ideas at the early stages of the project. We are also grateful to the three anonymous reviewers for their suggestions on how to further improve the presentation of the work.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Alan Ansell, Marinela Parović, Ivan Vulić, Anna Korhonen, and Edoardo Ponti. 2023. Unifying cross-lingual transfer across scenarios of resource scarcity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3980–3995, Singapore. Association for Computational Linguistics.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alan Ansell, Ivan Vulić, Hannah Sterz, Anna Korhonen, and Edoardo M. Monti. 2024. Scaling sparse fine-tuning to large language models. *CoRR*, abs/2401.16405.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. BUFFET: Benchmarking Large Language Models for Few-shot Cross-lingual Transfer.

David Brambila. 1976. *Diccionario Raramuri-Castellano: Tarahumar*.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Fajri Koto, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Ivan Halim Parmonangan, Ika Alfina, Muhammad Satrio Wicaksono, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaustubh D. Dhole, Arie Ardiyanti Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Ignatius, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Noor Fatyanosa, Ziwei Ji, Pascale Fung, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2022. NusaCrowd: Open Source Initiative for Indonesian NLP Resources.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. 2022. Fusing finetuned models for better pretraining.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Rubén Cushimariano Romano and Richer C. Sebastián Q. 2008. Ñaantsipeta asháninkaki birakochaki. diccionario asháninka-castellano. versión preliminar. http://www.lengamer.org/publicaciones/diccionarios/.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M Ponti. 2023a. Elastic Weight Removal for Faithful and Abstractive Dialogue Generation. *arXiv preprint arXiv:2303.17574*.

Nico Daheim, Thomas Möllenhoff, Edoardo Maria Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. 2023b. Model merging by uncertainty-based gradient matching. *CoRR*, abs/2310.12808.

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 and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. 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 4555–4567, Online. Association for Computational Linguistics.

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

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.

Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2023. Knowledge is a Region in Weight Space for Fine-tuned Language Models.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. 2022. Patching open-vocabulary models by interpolating weights. In *Advances in Neural Information Processing Systems*, volume 35, pages 29262–29277. Curran Associates, Inc.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 138–148, Hanoi, Vietnam. 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.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Michael S Matena and Colin A Raffel. 2022. Merging Models with Fisher-Weighted Averaging. In *Advances in Neural Information Processing Systems*, volume 35, pages 17703–17716. Curran Associates, Inc.

Elena Mihas. 2011. *Añaani katonkosatzi parenini, El idioma del alto Perené*. Milwaukee, WI: Clarks Graphics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

John E Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.

Marinela Parovic, Alan Ansell, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual transfer with target language-ready task adapters. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 176–193, Toronto, Canada. Association for Computational Linguistics.

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, Seattle, United States. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Ayu Purwarianti and Ida Ayu Putu Ari Crisdayanti. 2019. Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A User-Centric Scarce-Data Benchmark for Under-Represented Languages.

Sakriani Sakti and Satoshi Nakamura. 2013. Towards language preservation: Design and collection of graphemically balanced and parallel speech corpora of Indonesian ethnic languages. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–5.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023. Free lunch: Robust cross-lingual transfer via model checkpoint averaging. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5712–5730, Toronto, Canada. Association for Computational Linguistics.

Herry Sujaini. 2020. Improving the role of language model in statistical machine translation (Indonesian-Javanese). *International Journal of Electrical and Computer Engineering*, 10:2102–2109.

Yi-Lin Sung, Varun Nair, and Colin A Raffel. 2021. Training Neural Networks with Fixed Sparse Masks. In *Advances in Neural Information Processing Systems*, volume 34, pages 24193–24205. Curran Associates, Inc.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Bayu Distiawan Trisedya and Dyah Inastra. 2014. Creating Indonesian-Javanese Parallel Corpora Using Wikipedia Articles. In *2014 International Conference on Advanced Computer Science and Information System*, pages 239–245.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 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, Suzhou, China. Association for Computational Linguistics.

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

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Cahya Wirawan. 2022. LibriVox-Indonesia. `https://huggingface.co/datasets/indonesian-nlp/librivox-indonesia`.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

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.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Resolving Interference When Merging Models. In *NeurIPS*, New Orleans, USA. Proceedings of Machine Learning Research.

A Languages

The complete overview of languages, their codes and families, together with the monolingual data sizes and resources is provided in Table 5.

B Tasks and Datasets

The details of tasks, languages and datasets are given in Table 6.

Following prior work (Keung et al., 2020; Asai et al., 2023), we consider a binarized version of the MARC dataset, which is obtained by discarding the neutral class (the reviews with a score of 3) and assigning reviews with scores of 4 and 5 to the positive class and reviews with scores of 1 and 2 to the negative class. We use the review body and title as input features since that yielded the best source language performance.

In addition, NusaX dataset is created through human translation of a subset of the SMSA dataset. We thus carefully remove every example from SMSA which appears in its original or modified form in the NusaX test set to avoid data leakage.

C Results on MARC Dataset

The results with XLM-R$_{\text{BASE}}$ on MARC are provided in Table 7.

D Per-Language Results with XLM-R$_{\text{LARGE}}$

The full per-language results with XLM-R$_{\text{LARGE}}$ on AmericasNLI and NusaX are provided in Table 8.

E Scaling Factors

The best-performing scaling factors used for all the reported results with XLM-R$_{\text{BASE}}$ and XLM-R$_{\text{LARGE}}$ are given in Table 9.

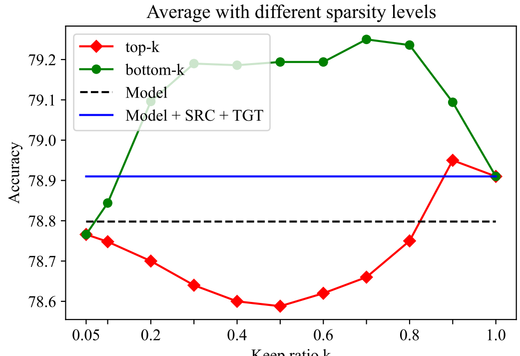
F Additional Sparsity Results

The sparsity results not covered in the main paper, with variants +SRC+TGT and +TGT are presented in Figures 2 and 3. We evaluate the top-k and bottom-k selections for all tasks, with k ranging between 5% and 90%.

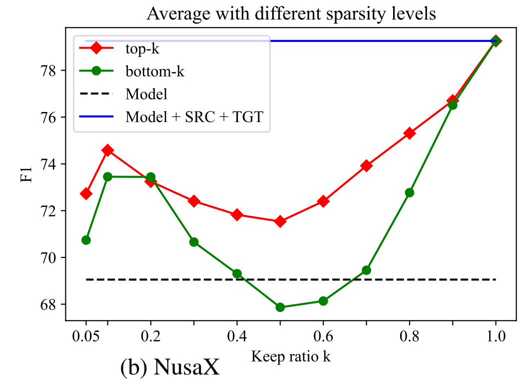
G Hyperparameters Details

All experiments were executed on a single RTX 3090 or RTX 600 Ada GPU. Training language vectors with both XLM-R models takes approximately 24 hours, while training of the task vectors takes several hours, depending on the task. As outlined in the limitations, all reported results are from a single run.

In addition to the hyperparameters summarized in Table 1 of the main paper, when training XLM-R$_{\text{LARGE}}$ model on the MultiNLI we introduce linear warmup for 6,000 steps to stabilize the training (this is approximately 10% of total training iterations). For training the XLM-R$_{\text{LARGE}}$ on the SMSA (source dataset of NusaX), we use a batch size of 32, and a learning rate of 10^{-5}. We train for 10 epochs and perform evaluation every 250 steps. We also introduce the linear warmup for 300 steps (roughly 10% of training) and an early stopping with a patience of 3 (i.e., we stop training when the F1 score does not increase for the three consecutive evaluation cycles).



(a) MARC



(b) NusaX

Figure 2: The average scores with different sparsity levels ranging from 5% to 90% with the MODEL + SRC + TGT variant.

Task	Language	Code	Family	Corpus size (MB)	Corpus source(s)
Source	English	en	Indo-European, Germanic	13,860	Wikipedia
	Indonesian	id	Austronesian, Malayo-Sumbawan	600	
NLI	Aymara	aym	Aymaran	2.3	Tiedemann (2012); Wikipedia
	Asháninka	cni	Arawakan	1.4	Ortega et al. (2020); Cushimariano Romano and Sebastián Q. (2008); Mihas (2011); Bustamante et al. (2020)
	Bribri	bzd	Chibchan, Talamanca	0.3	Feldman and Coto-Solano (2020)
	Guarani	gn	Tupian, Tupi-Guarani	6.9	Chiruzzo et al. (2020); Wikipedia
	Náhuatl	nah	Uto-Aztecan, Aztecan	8.1	Gutierrez-Vasques et al. (2016); Wikipedia
	Otomí	oto	Oto-Manguean, Otomian	0.4	Hñähñu Online Corpus
	Quechua	quy	Quechuan	17	Agić and Vulić (2019); Wikipedia
	Rarámuri	tar	Uto-Aztecan, Tarahumaran	0.6	Brambila (1976)
	Shipibo-Konibo	shp	Panoan	2.1	Galarreta et al. (2017); Bustamante et al. (2020)
	Wixarika	hch	Uto-Aztecan, Corachol	0.5	Mager et al. (2018)
SA	Acehnese	ace	Austronesian, Malayo-Sumbawan	90	KoPI-NLLB (Cahyawijaya et al., 2022); LibriVox-Indonesia (Wirawan, 2022); NLLB-Seed (NLLB Team et al., 2022); Wikipedia
	Balinese	ban	Austronesian, Malayo-Sumbawan	42	INDspeech_NEWS_EthnicSR (Sakti and Nakamura, 2013), KoPI-NLLB (Cahyawijaya et al., 2022); LibriVox-Indonesia (Wirawan, 2022); NLLB-Seed (NLLB Team et al., 2022); Wikipedia
	Banjarese	bjn	Austronesian, Malayo-Sumbawan	28	KoPI-NLLB (Cahyawijaya et al., 2022); Korpus Nusantara (Sujaini, 2020); NLLB-Seed (NLLB Team et al., 2022); Wikipedia
	Buginese	bug	Austronesian, South Sulawesi	4.3	Korpus Nusantara (Sujaini, 2020); LibriVox-Indonesia (Wirawan, 2022); NLLB-Seed (NLLB Team et al., 2022); Wikipedia
	Javanese	jav	Austronesian, Javanese	49	Wikipedia
	Madurese	mad	Austronesian, Malayo-Sumbawan	0.8	Korpus Nusantara (Sujaini, 2020); Wikipedia
	Minangkabau	min	Austronesian, Malayo-Sumbawan	93	Indo Wiki Parallel Corpora (Trisedya and Inastra, 2014); KoPI-NLLB (Cahyawijaya et al., 2022); Korpus Nusantara (Sujaini, 2020); LibriVox-Indonesia (Wirawan, 2022); MinangNLP MT (Koto and Koto, 2020); Wikipedia
	Ngaju	nij	Austronesian, Barito	-	-
	Sundanese	sun	Austronesian, Malayo-Sumbawan	33	Wikipedia
	Toba Batak	bbc	Austronesian, Northwest Sumatra-Barrier Islands	0.4	Korpus Nusantara (Sujaini, 2020)

Table 5: Details of the languages and monolingual data used for training and evaluation of language vectors. The corpora of Bustamante et al. (2020) are available at `https://github.com/iapucp/multilingual-data-peru`; all other NLI corpora mentioned are available at `https://github.com/AmericasNLP/americasnlp2021`; all the SA corpora (Cahyawijaya et al., 2022) are available through `https://indonlp.github.io/nusa-catalogue/`. The remaining languages (those from XNLI and MARC datasets) utilize only the Wikipedia corpora.

Task	Source Dataset	Target Dataset	Target Languages
Natural Language Inference (NLI)	MultiNLI (tr: 393k / dev: 10k) (Williams et al., 2018)	AmericasNLI (test: 750) (Ebrahimi et al., 2022)	Aymara (AYM), Bribri (BZD), Asháninka (CNI), Guarani (GN), Wixarika (HCH), Náhuatl (NAH), Otomí (OTO), Quechua (QUY), Shipibo-Konibo (SHP), Rarámuri (TAR)
	MultiNLI (tr: 393k / dev: 10k) (Williams et al., 2018)	XNLI (test: 5k) (Conneau et al., 2018)	Arabic (AR)†, Bulgarian (BG)†, German (DE)†, Greek (EL)†, Spanish (ES)†, French (FR)†, Hindi (HI)†, Russian (RU)†, Swahili (SW)†, Thai (TH)†, Turkish (TR)†, Urdu (UR)†, Vietnamese (VI)†, Chinese (ZH)†
Sentiment Analysis (SA)	MARC (tr: 160k / dev: 4k) (Keung et al., 2020)	MARC (test: 4k) (Keung et al., 2020)	German (DE)†, Spanish (ES)†, French (FR)†, Japanese (JA)†, Chinese (ZH)†
	SMSA (tr: 11k / dev: 1.3k) (Purwarianti and Crisdayanti, 2019; Wilie et al., 2020)	NusaX-senti (test: 400) (Winata et al., 2023)	Acehnese (ACE), Balinese (BAN), Toba Batak (BBC), Banjarese (BJN), Buginese (BUG), Javanese (JAV)†, Madurese (MAD), Minangkabau (MIN), Sundanese (SUN)†

Table 6: Details of the tasks, datasets, and languages involved in our cross-lingual transfer experiments. †denotes languages seen during MMT pretraining; The source language is English for XNLI, AmericasNLI, and MARC, and Indonesian for the NusaX dataset.

Method	DE	ES	FR	JA	ZH	avg
MODEL	82.83	79.17	**79.77**	77.00	75.22	78.80
MODEL + SRC	82.75	79.50	79.73	77.60	75.30	78.98
MODEL + TGT	82.53	79.20	79.40	77.32	75.55	78.80
MODEL + SRC + TGT	82.73	79.40	79.25	77.55	**75.62**	78.91
MODEL − SRC + TGT	**82.85**	**79.57**	78.75	**78.55**	75.38	**79.02**

Table 7: Results on MARC dataset in accuracy with XLM-R$_{\text{BASE}}$.

Method	AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR	**avg**
MODEL	38.00	39.60	41.20	40.80	36.40	42.28	40.51	40.67	44.67	38.40	40.25
MODEL + SRC	38.27	39.60	40.80	41.07	36.53	44.04	39.97	40.00	45.20	38.27	40.38
MODEL + TGT	**63.47**	43.33	47.60	**64.93**	**44.00**	52.57	45.19	**66.53**	**51.07**	**45.87**	**52.46**
MODEL + SRC + TGT	59.20	42.27	46.00	64.80	43.60	51.22	**46.39**	64.53	50.40	45.20	51.36
MODEL – SRC + TGT	60.80	**43.47**	**48.80**	63.07	43.73	**54.61**	44.92	65.33	50.53	43.87	51.91

(a) AmericasNLI: accuracy

Method	ACE	BAN	BBC	BJN	BUG	JAV	MAD	MIN	SUN	**avg**
MODEL	69.89	77.67	55.78	84.56	55.46	86.54	71.83	79.60	86.16	74.17
MODEL + SRC	71.67	78.30	56.84	85.10	54.55	88.48	74.25	81.83	87.18	75.36
MODEL + TGT	**86.13**	**83.40**	**75.27**	**86.48**	**71.03**	89.75	**81.58**	**87.66**	**89.56**	**83.43**
MODEL + SRC + TGT	77.87	81.61	69.67	85.62	62.63	**90.15**	80.89	86.04	88.22	80.30
MODEL – SRC + TGT	80.08	80.35	74.38	82.57	70.01	89.05	81.10	84.06	87.97	81.06

(b) NusaX: F1

Table 8: Full per-language results with XLM-R_{LARGE} on AmericasNLI and NusaX.

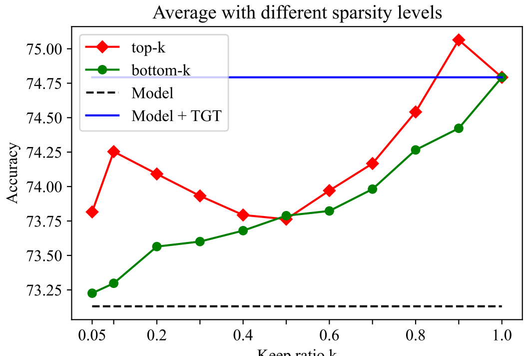
Method/Task	XNLI	AmericasNLI	MARC	NusaX
MODEL + SRC	0.5	0.7	0.2	0.3
MODEL + TGT	0.8	0.9	0.4	0.9
MODEL + SRC + TGT	0.5	0.8	0.2	0.7
MODEL – SRC + TGT	0.2	0.7	0.3	0.6

(a) XLM-R_{BASE}

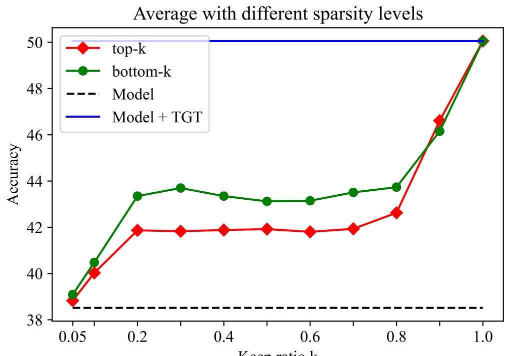
Method/Task	AmericasNLI	NusaX
MODEL + SRC	0.1	0.2
MODEL + TGT	0.8	0.6
MODEL + SRC + TGT	0.9	0.3
MODEL – SRC + TGT	0.8	0.5

(b) XLM-R_{LARGE}

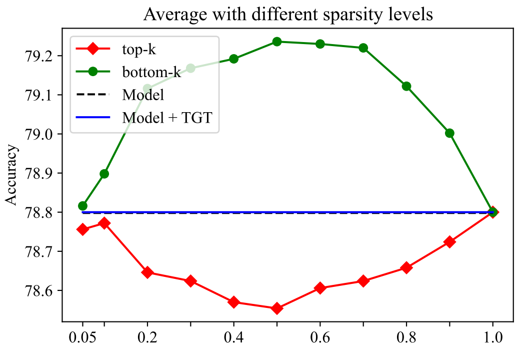
Table 9: Best scaling factors associated with the language vectors for different tasks with XLM-R_{BASE} and XLM-R_{LARGE}. They were chosen from the set $\{0.1, 0.2, ..., 1.0\}$ based on the best average performance on the validation sets.



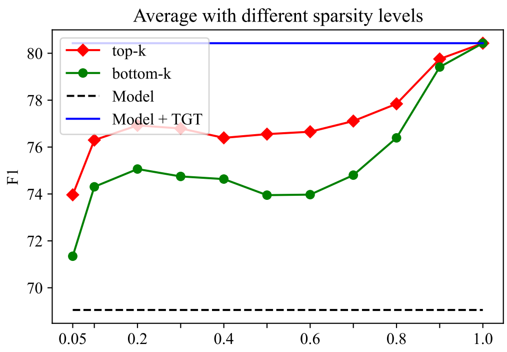
(a) XNLI



(b) AmericasNLI



(c) MARC



(d) NusaX

Figure 3: The average scores with different sparsity levels ranging from 5% to 90% with the MODEL + TGT variant.